Vysoká škola báňská – Technická univerzita Ostrava
Západočeská univerzita v Plzni

Úvod do statistiky (interaktivní učební text)

Martina Litschmannová

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Martina Litschmannová
Úvod do statistiky (interaktivní učební text)

ISBN

Předmluva

Mílí čtenáři,

skripta „Vybrané kapitoly z pravděpodobnosti“ a „Úvod do statistiky“ jsou určena pro studenty technických oborů vysoké školy. První díl těchto skript - „Vybrané kapitoly z pravděpodobnosti“ je koncipován tak, abyste si mohli učinit výchozí představu o základních pojmech a úlohách spadajících do oblasti pravděpodobnosti. Obtížnější části výkladu jsou prezentovány jen s nejnutnější mírou formálních prvků, mnohá odvození a důkazy jsou zařazeny pouze do kapitol určených pro zájemce o pozadí předkládaných vztahů. Přesto není předkládaný text lehké čtení. Prosím, počítejte s tím, že budete často muset usilovně přemýšlet, látku si postupně vyjasňovat a k mnoha tématům se opakovaně vracet. Při studiu Vám může pomoci řada animací (flash), appletů (java) a výpočetních programů (MS Excel), které budou v rámci pilotování výukových materiálů používány při výuce předmětů Statistika I., Biostatistika a Speciální analýza dat vyučovaných na VŠB-TU Ostrava a později se stanou součástí obrazovkové verze těchto materiálů.

V úvodu každé kapitoly jsou uvedeny cíle (konkrétní dovednosti a znalosti), kterých máte po prostudování této kapitoly dosáhnout. Náleduje vlastní výklad studované látky, zavedení nových pojmů a jejich vysvětlení, vše doprovázeno řešenými příklady. Množství řešených příkladů by Vám mělo umožnit aplikovat nabyté vědomosti při úlohách řešených v technické praxi. Hlavní pojmy, které si máte osvojit jsou na závěr kapitoly zopakovány v části Shrnutí. Pro ověření, zda jste dobře a úplně látku kapitoly zvládli, máte za každou kapitolou k dispozici několik testových otázek. Protože většina teoretických pojmů tohoto předmětu má bezprostřední význam a využití v praxi, jsou Vám rovněž předkládány i praktické úlohy k řešení. Schopnost aplikovat čerstvě nabyté znalosti při řešení reálných situací je hlavním cílem tohoto skripta. Výsledky testů a zadaných příkladů jsou uvedeny na konci každé kapitoly v Klíči k řešení. Používejte jej až po vlastním vyřešení testu a úloh, jen tak si samokontrolou ověříte, že jste obsah kapitoly skutečně úplně zvládli.

Úspěšné a příjemné studium s touto učebnicí Vám přeje,

V Ostravě 16. 1. 2011 Martina Litschmannová

Jak pracovat s testy?

Pro zahájení testu klikněte na tlačítko *Začátek testu*. Následně označte, resp. vyplňte, správné odpovědi. Za každou správnou odpověď obdržíte 1b, za chybnou Vám bude 1b odečten. (**POZOR**! Test může obsahovat i otázky s více správnými odpověďmi.) Pro ukončení testu klikněte na tlačítko *Konec testu*. Kliknete-li na tlačítko *Výsledky*, dojde k zobrazení správných výsledků testu. Správné číselné nebo slovní odpovědi se zobrazí tak, že u dané otázky kliknete na tlačítko *Odpověď*. Tato správná číselná, resp. slovní odpověď na danou otázku se zobrazí v rámečku v pravé šedé části obrazovky. Pro zobrazení správné číselné odpovědi u dalších otázek je potřeba u nich opět kliknout na tlačítko *Odpověď*.

Obsah

Kapitola 1

Explorační analýza proměnných

Cíle

Po prostudování této kapitoly budete znát

- základní pojmy explorační (popisné) statistiky
- typy datových proměnných
- statistické charakteristiky a grafickou demonstraci kvalitativních proměnných
- statistické charakteristiky a grafickou demonstraci kvantitativních proměnných

Původním posláním statistiky bylo zjišťování údajů o populaci na základě výběrového souboru. Pod pojmem **populace** přitom rozumějme množinu všech prvků, které sledujeme při statistickém výzkumu. Populace (základní soubor) bývá zadána buď výčtem prvků, nebo vymezením některých jejich společných vlastností. Například:

1. Provádíme-li stat. výzkum týkající se výšky 15-ti letých dívek, populaci tvoří všechny dívky, které mají 15 let.

2. Zkoumáme-li pevnost lan L50 vyrobených firmou LANOS, budeme za populaci považovat všechna lana L50 vyrobena firmou LANOS.

Vzhledem k tomu, že rozsah (počet prvků) populace (N) je obvykle vysoký, získáváme informace o populaci prostřednictvím statistického výzkumu. Nejběžnějším druhem statistického výzkumu je tzv. **výběrové šetření**, při němž je statistik pouze pasivním pozorovatelem – do průběhu šetření zasahuje co nejméně (ideálně vůbec ne). Zkoumaná část populace se nazývá **výběr**, popř. výběrový soubor. Počet prvků ve výběru označujeme n. Otázkou je jak stanovit takový výběr, aby byl skutečně reprezentativní, tj. aby charakteristiky výběru (např. průměr) dostatečně přesně reprezentovaly parametry populace. Jen si zkuste představit, k jakým výsledkům bychom došli při předvolebním průzkumu prováděném na vzorku voličů, který bychom získali pouze v domovech důchodců, popř. na schůzích mladých konzervativců. Existuje několik způsobu jak výběr provést (viz kapitola 9). Nejčastěji volíme **náhodný výběr**, v němž každý prvek populace má stejnou šanci být zařazen do výběru.

Je zřejmé, že výběrové šetření nemůže být nikdy tak přesné jako průzkum celé populace. Proč jej tedy preferujeme? Jmenujme tři nejdůležitější důvody.

1. <u>Úspora času a finančních prostředků</u> (zejména u rozsáhlé populace)

2. <u>Minimalizace ztrát v důsledku destruktivního testování</u> (některé testy – pevnost lan, životnost zářivek, obsah cholesterolu v krvi, atd. – vedou k destrukci zkoumaných prvků; zamyslete se sami, k čemu by vedlo testování celé populace)

3. <u>Nedostupnost celé populace</u> (při srovnávání působení faktorů okolí a dědičných znaků poskytují nejlepší informace jednovaječná dvojčata – jak je všechna najít a přesvědčit ke spolupráci?)

Přenášení závěrů z výběru na celou populaci je jedním z příkladů induktivního způsobu myšlení (**indukce = zevšeobecňování**). Mezi metody využívající statistickou indukci patří teorie odhadů a testování hypotéz. Jde o dvě rozsáhlé oblasti statistiky, v nichž budeme využívat poznatky získané analýzou výběru neboli explorační analýzou („exploratory data analysis“ – EDA).

Údaje, které u výběrového souboru sledujeme, nazýváme **proměnné** (znaky, veličiny) a jejich jednotlivé hodnoty **varianty** proměnné. **Explorační (popisná) statistika** bývá prvním krokem k odhalení informací skrytých ve velkém množství proměnných a jejich variant. To znamená uspořádání proměnných do názornější formy a jejich popis několika málo hodnotami, které by obsahovaly co největší množství informací obsažených v původním souboru. Vzhledem k tomu, že způsob zpracování proměnných závisí především na jejich typu, seznámíme se nyní se základním dělením proměnných do různých kategorií. Toto dělení je prezentováno na následujícím obrázku.

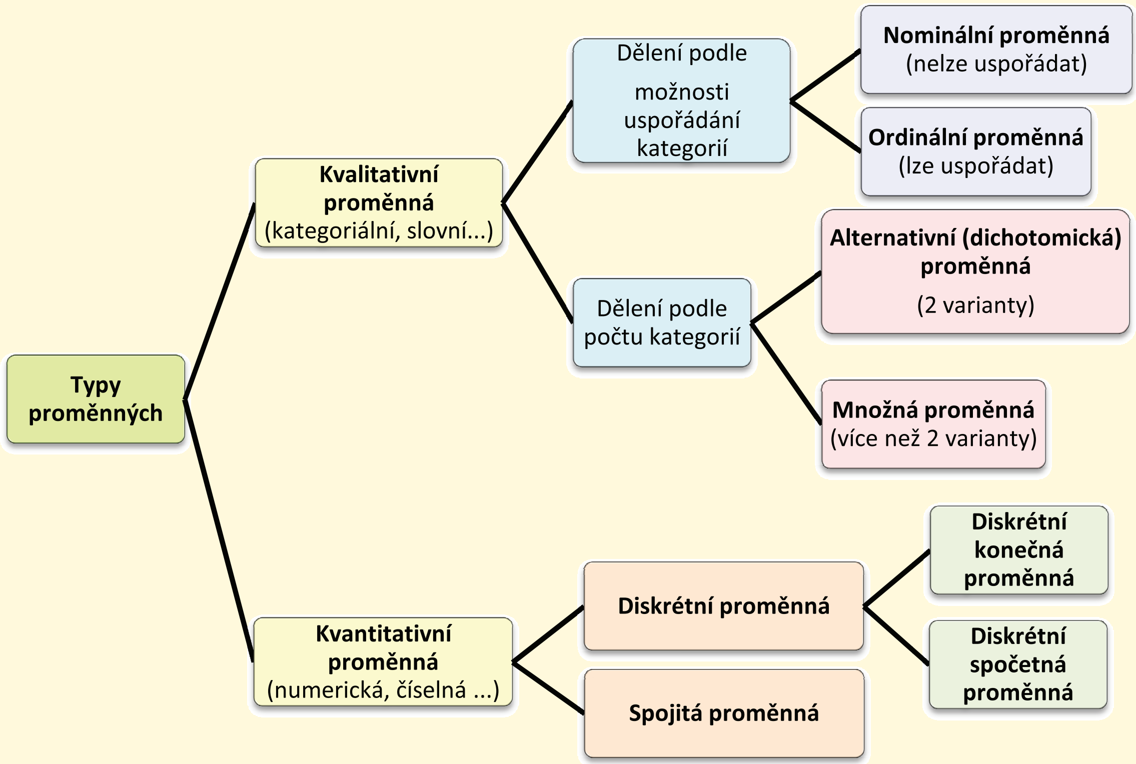
- **Proměnná kvalitativní** (kategoriální, slovní,...) je proměnná, kterou nemůžeme měřit, můžeme ji pouze zařadit do tříd. Varianty kvalitativní proměnné nazýváme kategoriemi, jsou vyjádřeny slovně a podle vztahu mezi jednotlivými kategoriemi se dělí na dvě základní podskupiny.
    - **Proměnná nominální** nabývá rovnocenných variant; nelze je smysluplně porovnávat ani seřadit (např. pohlaví, národnost, značka hodinek...)
    - **Proměnná ordinální** tvoří přechod mezi kvalitativními a kvantitativními proměnnými; jednotlivým variantám lze přiřadit pořadí a vzájemně je porovnávat nebo seřadit (např. známka ve škole, velikost oděvů (S, M, L))

  Jiným způsobem dělení kvalitativních proměnných je dělení podle počtu variant, jichž proměnné mohou nabývat.
    - **Proměnná alternativní** nabývá pouze dvou různých variant (např. pohlaví, zapnuto/vypnuto, živý/mrtvý...)
    - **Proměnná množná** nabývá více než dvou různých variant (např. vzdělání, jméno, barva očí...)
- **Proměnné kvantitativní** jsou proměnné měřitelné. Jsou vyjádřeny číselně a dělí se na
    - **Proměnné diskrétní** nabývající konečného nebo spočetného množství variant.
        - **Proměnné diskrétní konečné** – nabývají konečného počtu variant (např. známka z matematiky)
        - **Proměnné diskrétní spočetné** – nabývají spočetného množství variant (např. věk v letech, výška v centimetrech, váha v kilogramech...)

- **Proměnné spojité** nabývající libovolných hodnot z R nebo z nějaké podmnožiny R (např. výška, váha, vzdálenost měst...)

Průvodce studiem

*Tak, základní definice máme za sebou, proto můžeme přejít k věcem praktičtějším. Představte si situaci, že máte k dispozici statistický soubor o poměrně velkém rozsahu a stojíte před otázkou co s ním, jak jej co nejvýstižněji popsat a znázornit. Číselné hodnoty, kterými takovýto rozsáhlý soubor hodnot proměnné „nahradíme“, postihují základní vlastnosti tohoto souboru a my jim budeme říkat **statistické charakteristiky (statistiky)**. V následujících kapitolách se dozvíte, jak určit statistické charakteristiky pro různé typy proměnných a jak rozsáhlejší statistické soubory znázornit. Jdeme na to!*



Obr. 1.1: Demonstrace základních proměnných

1.1. Statistické charakteristiky kvalitativních proměnných

V tuto chvíli již víme, že kvalitativní proměnná má dva základní typy – nominální a ordinální.

1.1.1. Nominální proměnná

Nominální proměnná nabývá v rámci souboru různých, avšak rovnocenných kategorií. Počet těchto kategorií nebývá příliš vysoký, a proto první statistickou charakteristikou, kterou k popisu proměnné použijeme je četnost.

- **Četnost $\boldsymbol{n_i}$** (absolutní četnost, angl. „frequency“) je definována jako počet výskytu dané varianty kvalitativní proměnné.

  V případě, že kvalitativní proměnná ve statistickém souboru o rozsahu n hodnot nabývá k různých variant, jejichž četnosti označíme $n_1, n_2, \ldots, n_k$, musí zřejmě platit

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n.$$

Chceme-li vyjádřit, jakou část souboru tvoří proměnné s některou variantou, použijeme pro popis proměnné relativní četnost.

- **Relativní četnost $\boldsymbol{p_i}$** (angl. „relative frequency“) je definována jako

$$p_i = \frac{n_i}{n}, \qquad \text{popř. } p_i = \frac{n_i}{n} \cdot 100\,[\%].$$

(Druhý vzorec použijeme v případě, chceme-li relativní četnost vyjádřit v procentech.) Pro relativní četnosti musí platit

$$p_1 + p_2 + \ldots + p_k = \sum_{i=1}^{k} p_i = 1, \quad \text{popř. } 100\,\%.$$

Při zpracování kvalitativní proměnné je vhodné četnosti i relativní četnosti uspořádat do tzv. **tabulky rozdělení četnosti** (angl. „frequency table“) – Tab. 1.1.

Tab. 1.1: Tabulka rozdělení četností pro nominální proměnnou

TABULKA ROZDĚLENÍ ČETNOSTI		
Hodnoty x_i	**Absolutní četnosti**	**Relativní četnosti**
	n_i	p_i
x_1	n_1	p_1
x_2	n_2	p_2
x_k	n_k	p_k
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$

Poslední charakteristikou, kterou si pro popis nominální proměnné uvedeme, je modus.

- **Modus** definujeme jako název varianty proměnné vykazující nejvyšší četnost.

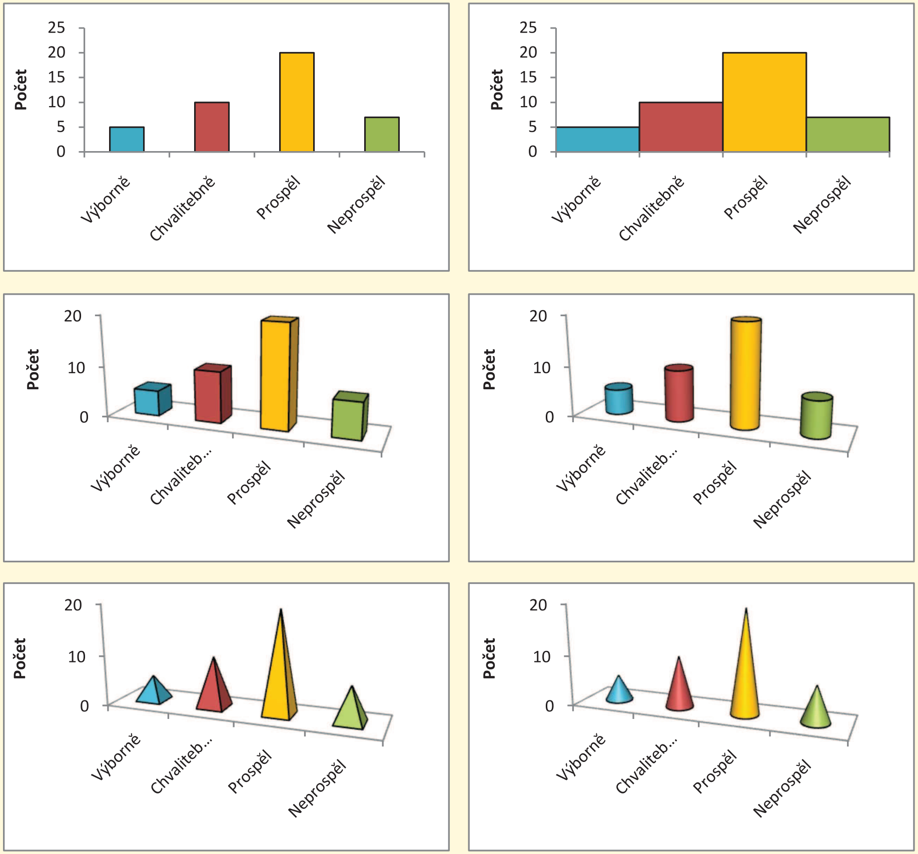
Modus tedy můžeme chápat jako typického reprezentanta souboru. V případě, že se ve statistickém souboru vyskytuje více variant s maximální četnosti, modus neurčujeme.

1.1.2. Grafické znázornění kvalitativní proměnné

Pro větší názornost analýzy proměnných se ve statistice často užívají **grafy**. Pro nominální proměnnou jsou to tyto dva typy:

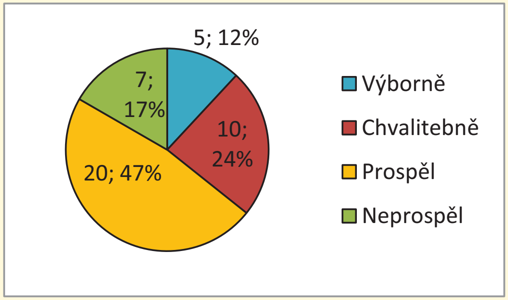
- **Histogram** (také sloupcový graf, angl. „bar chart“)
- **Výsečový graf** (také koláčový graf, angl. „pie chart“)

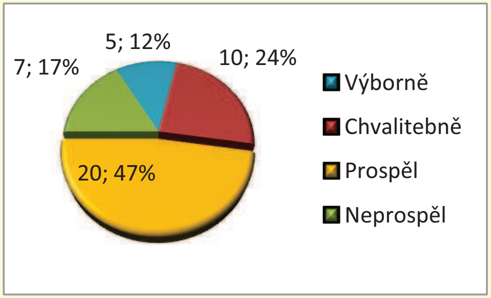
Histogram je klasickým grafem, v němž na jednu osu vynášíme varianty proměnné a na druhou osu jejich četnosti. Jednotlivé hodnoty četností jsou pak zobrazeny jako výšky sloupců (obdélníků, popř. hranolů, kuželů...)



Obr. 1.2: Ukázky histogramů

Výsečový graf prezentuje relativní četnosti jednotlivých variant proměnné, přičemž jednotlivé relativní četnosti jsou úměrně reprezentovány plochami příslušných kruhových výsečí. (Změnou kruhu na elipsu dojde k trojrozměrnému efektu.)



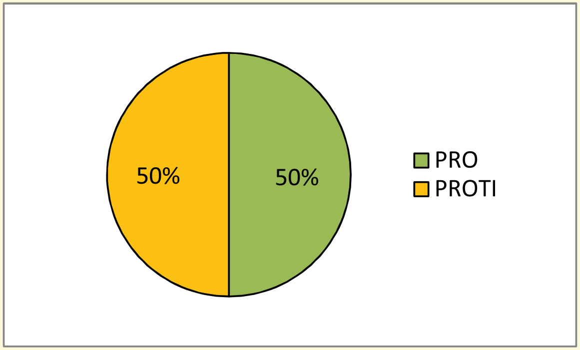


Obr. 1.3: Ukázky výsečových grafů

POZOR!!! V případě výsečového grafu si dejte zvláštní pozor na popis grafu. Jednotlivé výseče nestačí označit relativními četnostmi bez uvedení četnosti absolutních, popř. bez uvedení celkového počtu pozorování, to by mohlo vést k matení (ať už záměrnému nebo nechtěnému) toho, komu je graf určen. Zamyslete se nad následující ukázkou.

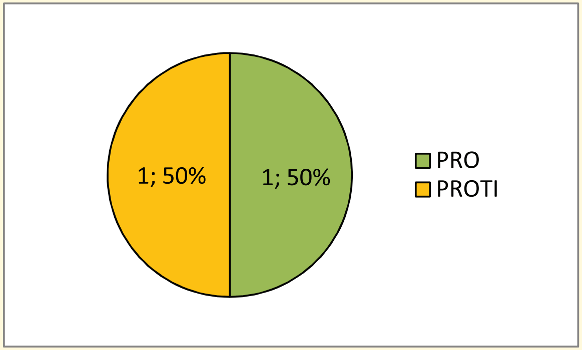
Příklad k zamyšlení: Minulý týden jsme zpracovali anketu týkající se názoru na zavedení školného na vysokých školách. Výsledky prezentuje následující graf.

Obsah
27. strana ze 525
Zavřít dokument
Celá obrazovka/Okno



Obr. 1.4: Chybná prezentace výsečového grafu

Co vy na to? Zajímavé výsledky, že? A věřte, nevěřte – pravdivé. A nyní graf doplníme tak, jak jsme doporučili.



Obr. 1.5: Správná prezentace výsečového grafu

Co si myslíte nyní? Z druhého grafu je patrné, že byli dotazování pouze dva lidé, jeden byl pro a druhý proti. Jaká je vypovídací schopnost takové ankety? Jaký je nyní Váš názor na prezentované výsledky? A závěr? Vytvářejte pouze takové grafy, jejichž interpretace je zcela jasná a je-li Vám výsečový graf bez uvedení absolutních četností předkládán, ptejte se vždy, zda je důvod v neznalosti autora nebo zda je to jeho záměr.

Průvodce studiem

Teď přišel čas na ověření, zda jste porozuměli předcházejícímu výkladu. Následující příklad se pokuste vyřešit samostatně, ukázkové řešení použijte ke kontrole svého postupu.

Příklad 1.1. Níže uvedená data představují částečný výsledek pozorování zaznamenaný při průzkumu zatížení jedné z ostravských křižovatek, a sice barvu projíždějících automobilů. Data vyhodnoťte a graficky znázorněte.

červená, modrá, zelená, modrá, červená, zelená, červená, červená, modrá, zelená, bílá, červená

Řešení 1.1.

1.1.3. Ordinální proměnná

Ordinální proměnná, stejně jako proměnná nominální, nabývá v rámci souboru různých slovních variant, avšak tyto varianty mají přirozené uspořádání, tj. můžeme určit, která je „menší“ a která „větší“.

Obsah

29. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Pro popis ordinální proměnné se používají stejné statistické charakteristiky a grafy jako pro popis proměnné nominální (četnost, relativní četnost, modus + histogram, výsečový graf), rozšířené o další dvě charakteristiky (kumulativní četnost, kumulativní relativní četnost), které berou v úvahu uspořádání ordinální proměnné.

- **Kumulativní četnost $\boldsymbol{m_i}$** (angl. „cumulative frequency“) definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné i-té variantě.

  Uvažte např. proměnnou „známka ze statistiky“, která nabývá variant: „výborně“, „velmi dobře“, „prospěl“, „neprospěl“, pak např. kumulativní četnost pro variantu „prospěl“ bude rovna počtu studentů, kteří ze statistiky získali známku „prospěl“ nebo lepší.

  Jsou-li jednotlivé varianty uspořádány podle své „velikosti“(„$x_1 < x_2 < \ldots < x_k$“), platí

  $$m_i = \sum_{j=1}^{i} n_j$$

  Je tedy zřejmé, že kumulativní četnost k-té („nejvyšší“) varianty je rovna rozsahu proměnné – $m_k = n$.

Druhou speciální charakteristikou určenou pouze pro ordinální proměnnou je kumulativní relativní četnost.

- **Kumulativní relativní četnost $\boldsymbol{F_i}$** (angl. „cumulative relative frequency“) vyjadřuje jakou část souboru tvoří hodnoty nabývající i-té a nižší varianty.

$$F_i = \sum_{j=1}^{i} p_j,$$

  což není nic jiného než relativní vyjádření kumulativní četnosti:

$$F_i = \frac{m_i}{n}.$$

Obdobně jako pro nominální proměnné, můžeme i pro proměnné ordinální prezentovat statistické charakteristiky pomocí tabulky rozdělení četnosti. Ta obsahuje ve srovnání s tabulkou rozdělení četností pro nominální proměnnou navíc hodnoty kumulativních a kumulativních relativních četností.

Tab. 1.2: Tabulka rozdělení četností pro ordinální proměnnou

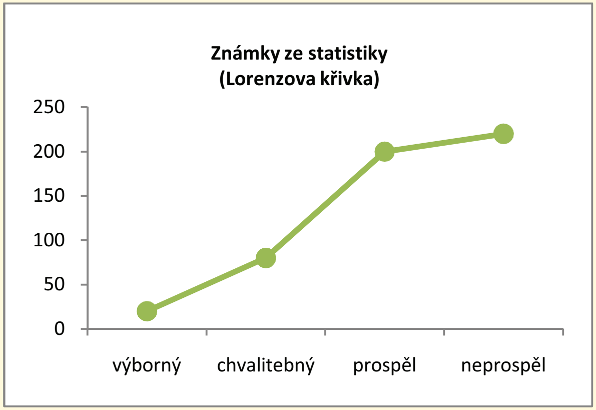
TABULKA ROZDĚLENÍ ČETNOSTÍ				
Hodnoty x_i	**Absolutní četnost**	**Relativní četnost**	**Kumulativní četnost**	**Kumulativní relativní četnost**
	n_i	**p_i**	**m_i**	**F_i**
x_1	n_1	p_1	$m_1 = n_1$	$F_1 = p_1$
x_2	n_2	p_2	$m_2 = n_1 + n_2 = m_1 + n_2$	$F_2 = p_1 + p_2 = F_1 + p_2$
x_k	n_k	p_k	$m_k = m_{k-1} + n_k = n$	$F_k = F_{k-1} + p_k = 1$
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	-----	-----

1.1.4. Grafické znázornění ordinální proměnné

Co se týče grafické prezentace ordinální proměnné, zmínili jsme histogram a výsečový graf. Ani jeden z těchto grafů však nezaznamenává uspořádání jednotlivých variant. K tomu nám slouží polygon kumulativních (resp. kumulativních relativních) četností, kterému se říká Lorenzova křivka, popř. Paretův graf.

Lorenzova křivka (polygon kumulativních četností, Galtonova ogiva, S křivka) je spojnicovým grafem, který získáme tak, že na vodorovnou osu vynášíme jednotlivé varianty proměnné v pořadí od „nejmenší“ do „největší“ a na svislou osu příslušné hodnoty kumulativních četností. Znázorněné body spojíme úsečkami.

Všimněte si, že směrnice (sklon) polygonu kumulativních četností je tím nižší, čím nižší je rozdíl mezi četnostmi jednotlivých variant.



Obr. 1.6: Lorenzova křivka

1.1.5. Paretova analýza

V různých odvětvích lidské činnosti (ekonomie, sociologie, řízení jakosti, ...) se setkáváme s Paretovým principem, který lze formulovat tak, že 80% následků pramení z 20% příčin (20% lidí vlastní 80% celkového bohatství, 80% závad je způsobeno 20% všech příčin, ...). V praxi pak bývá snahou nalézt toto malé spektrum příčin (životně důležitá menšina), které tak významně ovlivňuje výsledek. Tento postup, který si vysvětlíme na níže uvedeném příkladu, se nazývá Paretova analýza.

V následující animaci si můžete v krokovaném řešeném příkladu ověřit, zda dokážete aplikovat Paretovu analýzu v praxi.

Paretova analýza

Instrukce pro spuštění vložených animací naleznete na
http://mi21.vsb.cz/adobe-reader-multimedia.

Průvodce studiem

A znovu si můžete ověřit, zda dokážete správně aplikovat nabyté vědomosti.

Příklad 1.2. Následující data představují velikosti triček prodaných při výprodeji firmy TRIKO.

S, M, L, S, M, L, XL, XL, M, XL, XL, L, M, S, M, L, L, XL, XL, XL, L, M

a) Data vyhodnoťte a graficky znázorněte.
b) Určete kolik procent lidí si koupilo tričko velikosti nejvýše L.

Řešení 1.2.

1.2. Statistické charakteristiky numerických proměnných

Pro popis numerické proměnné můžeme použít většinu statistických charakteristik užívaných pro popis proměnné ordinální (četnost, relativní četnost, kumulativní četnost, kumulativní relativní četnost), což doplníme dalšími dvěma skupinami charakteristik - mírami polohy a mírami variability.

- **Míry polohy** určující typické rozložení hodnot proměnné (jejich rozmístění na číselné ose).
- **Míry variability** určující variabilitu (rozptyl) hodnot kolem své typické polohy.

1.2.1. Míry polohy a variability

Snad nejpoužívanějšími mírami polohy jsou průměry proměnných. Průměry představují průměrnou nebo typickou hodnotu výběrového souboru. Zřejmě nejznámějším průměrem pro kvantitativní proměnnou je

- **Aritmetický průměr $\bar{x}$** (angl. „mean“)

  Jeho hodnotu získáme pomocí známého vztahu

$$\bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n},$$

kde: x ... jednotlive hodnoty proměnné,
n ... rozsah výběrového souboru (počet hodnot proměnné).

Jsou-li hodnoty analyzované proměnné uspořádány do tabulky četností, používáme pro výpočet aritmetického průměru vztah

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum\limits_{i=1}^{k} x_i n_i}{\sum\limits_{i=1}^{k} n_i},$$

kde četnosti n_i představují váhu, která je přisuzována jednotlivým hodnotám proměnné x_i. Takto vypočítaný aritmetický průměr se nazývá **vážený aritmetický průměr**.

Známé jsou i **vlastnosti aritmetického průměru**.

1. $\sum\limits_{i=1}^{n}(x_i - \bar{x}) = 0,$

   neboli: součet všech odchylek hodnot proměnné od jejich aritmetického průměru je roven nule, což znamená, že aritmetický průměr kompenzuje vliv náhodných chyb na proměnnou.

2. $\forall a \in \mathbb{R} : \frac{\sum_{i=1}^{n}(a+x_i)}{n} = a + \bar{x},$

   neboli: přičteme-li ke všem hodnotám proměnné stejné číslo, zvětší se o toto číslo rovněž aritmetický průměr.

3. $\forall b \in \mathbb{R} : \frac{\sum_{i=1}^{n}(bx_i)}{n} = b\bar{x},$

   neboli: vynásobíme-li všechny hodnoty proměnné stejným číslem, zvětší se stejným způsobem rovněž aritmetický průměr.

Příklad 1.3. Učitel matematiky na gymnáziu přiřazuje jednotlivým výsledkům studentů váhy následujícím způsobem.

	Váha
Zkoušení a dílčí testy	1
Opakovací testy	2
Kompozice	3

U studenta Masaříka má učitel za 1. pololetí záznam:

Zkoušení:	2
Dílčí testy:	3, 2, 1, 3
Opakovací testy:	2, 3, 1
Kompozice:	3, 2

Určete výslednou průměrnou známku studenta.

Řešení 1.3.

Přestože to tak na první pohled vypadá, aritmetický průměr nemusí být vždy pro výpočet průměru výběrového souboru nejvhodnější.

- **Harmonický průměr**

  Pro výpočet průměru v případech, kdy proměnná má charakter části z celku (úlohy o společné práci, ...), používáme průměr harmonický, který je definován vztahem

$$\bar{x}_H = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}}.$$

  Máme-li údaje setříděné do tabulky četností, používáme dle níže uvedeného vztahu **vážený harmonický průměr**.

$$\bar{x}_H = \frac{\sum\limits_{i=1}^{k} n_i}{\sum\limits_{i=1}^{n} \frac{n_i}{x_i}}$$

Příklad 1.4. Totožná součástka se vyrábí na dvou automatech. Starší z nich vyrobí 1 kus každých 6 minut, nový každé 3 minuty. Jak dlouho trvá v průměru výroba jedné součástky?

Řešení 1.4.

- **Geometrický průměr**
  Pracujeme-li s kladnou proměnnou představující relativní změny (růstové indexy, cenové indexy...), používáme tzv. **geometrický průměr**, který je definován jako *n-tá* odmocnina ze součinu hodnot proměnné.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot\ x_2 \cdot\ \ldots\ \cdot x_n}$$

  Stejně jako v předchozích případech lze zapsat rovněž vzorec pro **vážený geometrický průměr**.

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot\ \ldots\ \cdot x_n^{n_k}},$$

  kde

$$n = \sum_{i=1}^{k} n_i.$$

Příklad 1.5. Předloni byla výše ročního platu zaměstnance ve firmě 200 000 Kč, loni 220 000 Kč a letos 250 000 Kč. Jaký je průměrný koeficient růstu jeho platu?

Řešení 1.5.

Obsah
39. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Vzhledem k tomu, že průměr se stanovuje ze všech hodnot proměnné, nese maximum informací o výběrovém souboru. Na druhé straně je však velmi citlivý na tzv. **odlehlá pozorování**, což jsou hodnoty, které se mimořádně liší od ostatních a dokážou proto vychýlit průměr natolik, že přestává daný výběr reprezentovat. K identifikaci odlehlých pozorování se vrátíme později.

Mezi míry polohy, které jsou na odlehlých pozorováních méně závislé, patří

- **Modus**

  Pozor! v případě modu budeme rozlišovat mezi diskrétní a spojitou kvantitativní proměnnou. **Pro diskrétní proměnnou** definujeme modus jako hodnotu nejčetnější varianty proměnné (podobně jako u kvalitativní proměnné).

  Naproti tomu **u spojité proměnné** považujeme za modus $\hat{x}$ hodnotu kolem níž je největší koncentrace hodnot proměnné. Mnohdy mluvíme o typické hodnotě proměnné. Pro určení této hodnoty využijeme tzv. **shorth** (čti „šórt“ a skloňuj podle hrad), což je nejkratší interval, v němž leží alespoň 50% hodnot proměnné (v případě výběru o rozsahu $n = 2k (k \in \mathbb{N})$ (sudý počet hodnot), leží v shorthu k hodnot–což je 50% $(n/2)$ hodnot proměnné, v případě výběru o rozsahu $n = 2k + 1 (k \in \mathbb{N})$ (lichý počet hodnot), leží v shorthu $k + 1$ hodnot - což je o 1 více než je 50% hodnot proměnné). **Modus** pak definujeme jako střed shorthu.

  Z předcházejících definic vyplývá, že délka shorthu (horní mez – dolní mez) je jednoznačně dána, to však nemusí platit pro jeho umístění a tudíž ani pro modus. Pokud lze modus určit jednoznačně, mluvíme o **unimodální proměnné**, má-li proměnná dva mody, nazýváme ji **bimodální**. Existence dvou a více modu ve výběru obvykle signalizuje nesourodost (heterogenitu) hodnot proměnné. Tuto nesourodost bývá možné odstranit rozdělením souboru na podsoubory - roztříděním podle některého jiného znaku (např. bimodální znak výška člověka lze roztřídit podle pohlaví na dva unimodální znaky - výška žen a výška mužů).

Průvodce studiem

Zdála se Vám pasáž o modu kvantitativní proměnné příliš složitá? Pokusíme se ji nyní osvětlit na jednoduchém příkladu, který Vám snad případné nejasnosti ozřejmí.

Příklad 1.6. Následující data představují věk hudebníků vystupujících na přehlídce dechových orchestrů. Proměnnou věk považujte za spojitou. Určete průměr, shorth a modus věku hudebníků.

22	82	27	43	19	47	41	34	34	42	35

Řešení 1.6.

Pro podrobnější vyjádření rozložení hodnot proměnné v rámci souboru slouží statistiky nazývané **výběrové kvantily.**

- **Výběrové kvantily** (angl. quantile, resp. percentile)

Výběrové kvantily jsou statistiky, které charakterizují polohu jednotlivých hodnot v rámci proměnné. Podobně jako modus, jsou i výběrové kvantily rezistentní (odolné) vůči odlehlým pozorováním. Obecně je výběrový kvantil (dále jen kvantil) chápán jako hodnota, která rozděluje výběrový soubor na dvě části – první z nich obsahuje hodnoty, které jsou menší než daný kvantil, druhá část obsahuje hodnoty, které jsou větší nebo rovny danému kvantilu. Pro určení kvantilu je proto nutné výběr uspořádat od nejmenší hodnoty k největší.

Kvantil proměnné x, který odděluje *100p*% menších hodnot od zbytku souboru, tj. od *100(1-p)*% hodnot, nazýváme ***100p* %-ním kvantilem** a značíme jej $\boldsymbol{x_p}$.
V praxi se nejčastěji setkáváme s následujícími kvantily:

- **Kvartily**

Dolní kvartil $\boldsymbol{x_{0,25}}$ = 25%-ní kvantil (rozděluje datový soubor tak, že 25% hodnot je menších než tento kvartil a zbytek, tj. 75% větších (nebo rovných))

Medián $\boldsymbol{x_{0,5}}$ = 50%-ní kvantil (rozděluje datový soubor tak, že polovina (50%) hodnot je menších než medián a polovina (50%) hodnot větších (nebo rovných))

Horní kvartil $\boldsymbol{x_{0,75}}$ = 75%-ní kvantil (rozděluje datový soubor tak, že 75% hodnot je menších než tento kvartil a zbytek, tj. 25% větších (nebo rovných))

Kvartily dělí výběrový soubor na 4 přibližně stejně četné části.

- **Decily**–$x_{0,1}; x_{0,2}; ...; x_{0,9}$

  Decily dělí výběrový soubor na 10 přibližně stejně četných části.

- **Percentily**–$x_{0,01}; x_{0,02}; \ldots; x_{0,99}$

  Percentily dělí výběrový soubor na 100 přibližně stejně četných části.

A nyní se dostáváme k tomu, **jak se kvantily určují.**

1. Výběrový soubor uspořádáme podle velikosti.
2. Jednotlivým hodnotám proměnné přiřadíme pořadí, a to tak, že nejmenší hodnota bude mít pořadí 1 a nejvyšší hodnota pořadí n (rozsah souboru).
3. 100p%- ní kvantil je roven hodnotě proměnné s pořadím z_p, kde

$$z_p = np + 0.5$$

   Není-li z_p celé číslo, pak daný kvantil určíme jako průměr prvků s pořadím $\lfloor z_p \rfloor$ a $\lceil z_p \rceil$.

POZOR! Zejména v souvislosti s hodnocením normovaných testů (SCIO testy, biometrické normy,...) se často setkáváme s vyjádřením „Patříte do p. percentilu“, přičemž p je celé číslo mezi 1 a 100. Je tím myšleno, že nejméně (p-1)% a zároveň méně než p% účastníků testu dosáhlo nižšího hodnocení než vy.
(Např. „Patříte do 80. percentilu“ znamená, že nejméně 79% (a nejvýše 80%) účastníku testu dosáhlo nižšího výsledku než vy.)

Obsah
43. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Za zmínku zajisté stojí i **vztah mezi kvantily a relativní kumulativní četnosti**. Zřejmě lze říci, že hodnota p udává relativní kumulativní četnost kvantilu x_p, tj. relativní četnost těch hodnot proměnné, které jsou menší než kvantil x_p. Kvantil a relativní kumulativní četnost jsou tedy inverzní pojmy. Grafické nebo tabulkové znázornění setříděné proměnné a příslušných kumulativních četností se označuje jako **distribuční funkce kumulativní četnosti**, popř. **empirická distribuční funkce**. Ujasněme si nyní, jak empirickou distribuční funkci pro kvantitativní proměnnou určit.

- **Empirická distribuční funkce F(x) pro kvantitativní proměnnou**

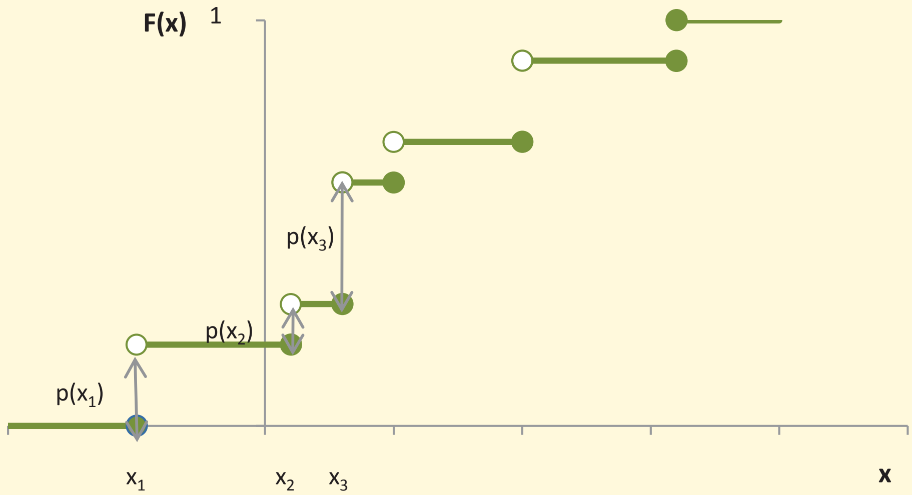
  Označme si $p(x_i)$ relativní četnost hodnoty x_i seřazeného výběrového souboru $x_1 < < x_2 < \ldots < x_n$. Pro empirickou distribuční funkci F(x) pak platí:

$$F(x) = \begin{cases} 0 & \text{pro } x \leqq x_i \\ \sum_{i=1}^{j} F(x) & \text{pro } x_j < x \leqq x_{j+1}, 1 \leqq j \leqq n-1 \\ 1, & \text{pro } x_n < x \end{cases}$$

  Empirická distribuční funkce je monotónně rostoucí, zleva spojitou funkcí, která „skáče“ podle relativních četností příslušných jednotlivým hodnotám proměnné. Zjevně tedy platí, že

$$p(x_i) = \lim_{x \to x_i} F(x) - F(x_i)$$

Prostřednictvím kvantilů jsou definovány i další dvě statistiky kvantitativní proměnné – interkvartilové rozpětí a MAD.



Obr. 1.7: Empirická distribuční funkce

- **Interkvartilové rozpětí IQR**

  Tato statistika je mírou variability souboru a je definována jako vzdálenost mezi horním a dolním kvartilem:

$$IQR = x_{0.75} - x_{0.25}$$

- **MAD**

Název MAD je zkratkou anglické definice – **m**edian **a**bsolute **d**eviation from the median, čili česky: medián absolutních odchylek od mediánu

Jak jej tedy určíme?

1. Výběrový soubor uspořádáme podle velikosti
2. Určíme medián souboru
3. Pro každou hodnotu souboru určíme absolutní hodnotu její odchylky od mediánu
4. Absolutní odchylky od mediánu uspořádáme podle velikosti
5. Určíme medián absolutních odchylek od mediánu, tj. MAD

Průvodce studiem

Zdá se Vám, že za sebou máte moc teorie? Abyste se ujistili, že nic není tak černé jak vypadá, zkuste pokračovat v předcházejícím řešeném příkladu.

Příklad 1.7. Pro data z řešeného příkladu 1.6 určete

a) všechny kvartily,

b) interkvartilové rozpětí,

c) MAD,

d) zakreslete empirickou distribuční funkci.

Řešení 1.7.

Průvodce studiem

Zvládli jste to? Gratuluji. Pokud jste s příkladem měli nějaké problémy, doporučuji vám, abyste pasáž o kvantilech a empirické distribuční funkci znovu důkladně prostudovali – není to naposled, co se s těmito pojmy setkáváte.

Až dosud jsme se zabývali převážně statistickými charakteristikami umožňujícími popis polohy proměnné, tj. mírami polohy. Průměry, modus, stejně jako medián vyjadřují pomyslný „střed“ proměnné, neříkají však nic o rozložení jednotlivých hodnot proměnné kolem tohoto „středu“, tj. o variabilitě proměnné. Je zřejmé, že čím větší je rozptýlenost hodnot proměnné kolem jejího pomyslného „středu“, tím menší je schopnost tohoto „středu“ reprezentovat proměnnou.

Následující statistické charakteristiky nám umožňují popis variability (rozptýlenosti) výběrového souboru, neboli popis rozptylu jednotlivých hodnot kolem středu proměnné – nazýváme je tedy mírami variability. Z dosud zmíněných statistických charakteristik zařazujeme mezi míry variability shorth a interkvartilové rozpětí.

- **Výběrový rozptyl** s^2 (čti „s kvadrát“, angl. sample variance) je nejrozšířenější mírou variability výběrového souboru. Určujeme jej podle vztahu

$$s^2 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  Vidíme, že výběrový rozptyl je dán podílem součtu kvadrátu odchylek jednotlivých hodnot od průměru a rozsahu souboru sníženého o jedničku.

  Mezi základní **vlastnosti výběrového rozptylu** patří:

Obsah
47. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

1. Výběrový rozptyl konstantního souboru je roven nule,což znamená, že jsou-li všechny hodnoty proměnné stejné, má soubor nulovou rozptýlenost.
2. $$\forall a \in \mathbb{R} : \left(\left(s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \right) \wedge (y_i = a + x_i) \right) \Rightarrow$$
$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}((a+x_i)-(a+\bar{x}))^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = s^2$$
což znamená, že přičteme-li ke všem hodnotám proměnné libovolnou konstantu, výběrový rozptyl proměnné se nezmění.
3. $$\forall b \in \mathbb{R} : ((s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \wedge y_i = bx_i)) \Rightarrow$$
$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}((bx_i)-(b\bar{x}))^2}{n-1} = \frac{\sum_{i=1}^{n}b^2(x_i-\bar{x})^2}{n-1} = b^2s^2$$
což znamená, že vynásobíme-li všechny hodnoty proměnné libovolnou konstantou (b), výběrový rozptyl proměnné se zvětší kvadrátem této konstanty (b^2 krát)

Nevýhodou použití výběrového rozptylu jakožto míry variability je to, že jednotka této charakteristiky je druhou mocninou jednotky proměnné. Např. je-li proměnnou denní tržba uvedena v Kč, bude výběrový rozptyl této proměnné vyjádřen v Kč^2. Následující míra variability tuto vlastnost nemá.

- **Výběrová směrodatná odchylka s** (angl. sample standard deviation) je definována jako kladná odmocnina výběrového rozptylu

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Nevýhodou výběrového rozptylu i výběrové směrodatné odchylky je skutečnost, že neumožňují porovnávat varibilitu proměnných vyjádřených v různých jednotkách. Která proměnná má větší variabilitu – výška nebo hmotnost dospělého člověka? Na tuto otázku nám dá odpověď tzv. variační koeficient.

- **Variační koeficient** V_x (angl. coefficient of variation)

  vyjadřuje relativní míru variability proměnné x. Podle níže uvedeného vztahu jej lze stanovit pouze pro proměnné, které nabývají výhradně kladných hodnot. Variační koeficient je bezrozměrný. Uvádíme-li jej v [%], hodnotu získanou z definičního vzorce vynásobíme 100%.

$$V_x = \frac{V}{\bar{x}}, \text{ popř. } V_x = \frac{V}{\bar{x}} \cdot 100[\%]$$

Příklad 1.8. Firma vyrábějící tabulové sklo vyvinula méně nákladnou technologii pro zlepšení odolnosti skla vůči žáru. Pro testování bylo vybráno 5 tabulí skla a rozřezáno na polovinu. Jedna polovina pak byla ošetřena novou technologií, zatímco druhá byla ponechána

Tab. 1.3: Tavná teplota skla při použití staré a nové technologie

Mezní teplota (sklo prasklo) [°C]	
Stará technologie x_i	**Nová technologie** y_i
475	485
436	390
495	520
483	460
426	488

jako kontrolní. Obě poloviny pak byly vystaveny zvyšujícímu se působení tepla, dokud nepraskly. Výsledky jsou uvedeny v Tab. 1.10. Porovnejte obě technologie pomocí základních charakteristik explorační statistiky (průměru a rozptylu, popř. směrodatné odchylky).

Řešení 1.8.

Vzpomínáte si ještě na zmínku o odlehlých pozorováních? Dozvěděli jste se, že za odlehlá pozorování považujeme ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. vypovídací hodnotu průměru. Nyní se dozvíte, jak odlehlé hodnoty identifikovat.

- **Identifikace odlehlých pozorování**(angl. outliers)

Ve statistické praxi se obvykle můžete setkat s několika způsoby identifikace odlehlých pozorování. My ukážeme tři z nich.

1. **Vnitřní hradby:** Za odlehlé pozorování lze považovat takovou hodnotu x_i, která je od dolního, resp. horního kvartilu vzdálená více než 1,5 násobek interkvartilového rozpětí. Tedy:

$$[(x_i < x_{0,25} - 1,5 \cdot IQR) \vee (x_i > x_{0,75} + 1,5 \cdot IQR)] \Rightarrow$$
$$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

2. **z-souřadnice (*z-skóre*):** Za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota z-souřadnice je větší než 3, tj. hodnota, která je od průměru vzdálenější než 3s. Tedy:

$$z - sk\acute{o}re_i = \frac{x_i - \bar{x}}{s}$$

$$|z - sk\acute{o}re_i| > 3 \Rightarrow \left|\frac{x_i - \bar{x}}{s}\right| > 3 \Rightarrow |x_i - \bar{x}| > 3s \Rightarrow$$
$$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

3. **$x_{0,5}$-souřadnice ($x_{0,5}$ − *skóre*):** Za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota mediánové souřadnice je větší než 3, tj. hodnota, která je od mediánu vzdálenější než $3 \cdot 1,483\cdot$ MAD. Tedy:

$$x_{0,5} - sk\acute{o}re_i = \frac{x_i - x_0,5}{1,483MAD}$$

$$|x_{0,5} - \mathit{skóre}_i| > 3 \Rightarrow \left|\frac{x_i - x_{0,5}}{1,483MAD}\right| > 3 \Rightarrow |x_i - x_{0,5}| > 3 \cdot 1,483MAD \Rightarrow$$

$$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

V konkrétním případě můžete pro identifikaci odlehlých pozorování zvolit libovolné z těchto tří pravidel. Za zmínku stojí, že z-souřadnice je „méně přísná" k odlehlým pozorováním než mediánová souřadnice. Je to proto, že z-souřadnice se určuje na základě průměru a výběrové směrodatné odchylky, jež jsou silně ovlivněny hodnotami odlehlých pozorování. Naproti tomu mediánová souřadnice se určuje na základě mediánu a MADu, které jsou vůči odlehlým pozorováním odolné.

Někteří statistici rozdělují odlehlá pozorování do dvou skupin – na **odlehlá pozorování** a **extrémní pozorování**. Pro toto rozlišení využívají pojmů vnitřní a vnější hradby. Definice hradeb vychází z pravidla pro identifikaci odlehlých pozorování pomocí IQR.

Vnitřní hradby:	dolní mez:	$h_D = x_{0,25} - 1,5IQR$
	horní mez:	$h_H = x_{0,75} + 1,5IQR$
Vnější hradby:	dolní mez:	$H_D = x_{0,25} - 3IQR$
	horní mez:	$H_H = x_{0,75} + 3IQR$

Pozorování ležící mimo vnější hradby pak nazýváme extrémní, pozorování ležící vně vnitřních hradeb, avšak uvnitř hradeb vnějších nazýváme odlehlá.

Pokud o některé hodnotě proměnné rozhodneme, že je odlehlým pozorováním, je nutné rozlišit o jaký typ odlehlosti se jedná. V případě, že odlehlost pozorování je způsobena:

- hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
- důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit z dalšího zpracování. V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

Dalšími charakteristikami popisujícími kvantitativní proměnnou jsou **výběrová šikmost** a **výběrová špičatost**. Vzorce podle nichž se určují tyto charakteristiky jsou poměrně složité a proto se podle nich „ručně" většinou nepočítá, jsou součástí většiny statistických programů.

- **Výběrová šikmost *a*** (angl. skewness)

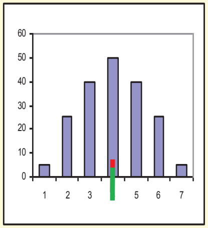
  vyjadřuje asymetrii rozložení hodnot proměnné kolem jejího průměru. Výběrová šikmost je definována vztahem:

$$a = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

A jak výběrovou šikmost interpretujeme?

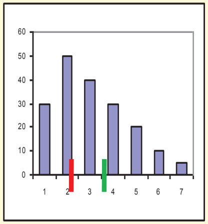
$a = 0$ ... hodnoty proměnné jsou kolem jejího průměru rozloženy symetricky
$a > 0$ ... u proměnné převažují hodnoty menší než průměr
$a < 0$ ... u proměnné převažují hodnoty větší než průměr

Symetrická data



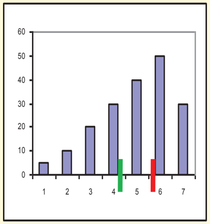
Průměr = **medián**

Pozitivně zešikmená data



Průměr > **medián**

Negativně zešikmená data



Průměr < **medián**

Souvislost mezi šikmostí a charakteristikami polohy

Symetrické rozdělení:	$\bar{x} = x_{0,5}$
Pozitivně zešikmené rozdělení:	$\bar{x} > x_{0,5}$
Negativně zešikmené rozdělení:	$\bar{x} < x_{0,5}$

- **Výběrová špičatost *b*** (angl. kurtosis)

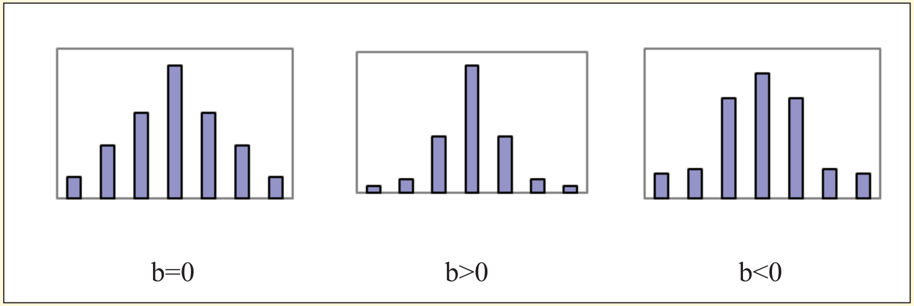
vyjadřuje koncentraci hodnot proměnné kolem jejího průměru. Výběrová špičatost je definována vztahem

$$b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}.$$

Obsah
54. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

A jak výběrovou výběrovou špičatost?

$b = 0$ ... špičatost odpovídá normálnímu rozdělení (bude definováno později)
$b > 0$ ... špičaté rozdělení proměnné
$b < 0$ ... ploché rozdělení proměnné



1.3. Přesnost statistických charakteristik kvantitativních proměnných

V této chvíli jste se seznámili s řadou statistických charakteristik. Vzniká otázka, s jakou přesností máme tyto číselné charakteristiky uvádět. Je zřejmé, že počet platných cifer by měl korespondovat s přesností měření. Víme-li, například, že nejistota měření určité proměnné je jeden kilogram, nemá smysl průměr této proměnné uvádět s přesností na gramy.

Platí jednoduché pravidlo.

Obsah
55. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme **nahoru** na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily…) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

Příklady chybně zapsaných hodnot číselných charakteristik vidíte v Tab. 1.4.

Tab. 1.4: Příklady chybného zápisu číselných charakteristik

	Délka [m]	**Váha [kg]**	**Teplota [^{0}C]**
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?	*Různý počet des. míst.*	*3 platné cifry u směrodatné odchylky.*	*Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky).*

Jak by měl zápis vypadat správně ukazuje Tab 1.5.

Tab. 1.5: Příklady správného zápisu číselných charakteristik

	Délka [m]	**Váha [kg]**	**Teplota [^{0}C]**
Průměr	2,26	128	14 600
Medián	2,68	118	13 700
Směrodatná odchylka	0,78	24	1 200

Obsah
56. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Průvodce studiem

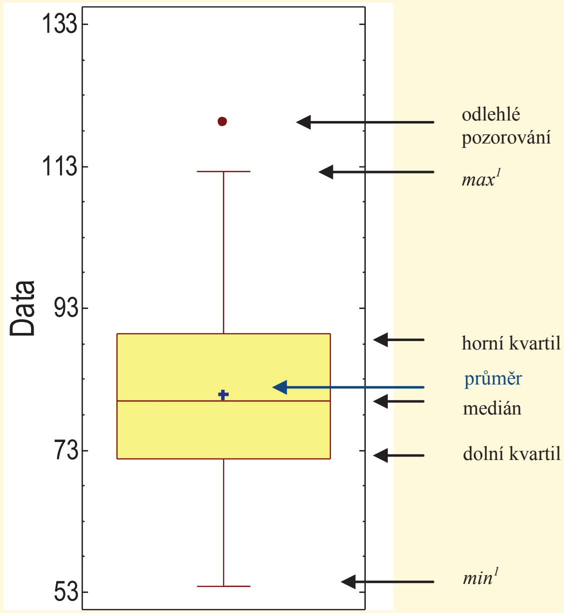
Tak, a máte to takřka vše za sebou – všechny číselné charakteristiky, které budete využívat pro popis kvantitativní proměnné jsou definovány. Zbývá nám jediné – ukázat si jak můžeme kvantitativní proměnnou znázornit graficky. Tak vzhůru do toho, neboť o nic složitého nejde.

1.3.1. Grafické znázornění kvalitativní proměnné

- Krabicový graf(angl. Box plot)

Krabicový graf se ve statistice využívá od roku 1977, kdy jej poprvé prezentoval americký statistik J. W. Tukey. Nazval jej „box with whiskers plot“ – krabicový graf s vousama. Grafická podoba tohoto grafu se v různých aplikacích mírně liší. Jednu z jeho verzí vidíte na uvedeném obrázku.

Odlehlá pozorování jsou znázorněna jako izolované body, konec horního (popř. konec dolního) vousu představují maximum (popř. minimum) proměnné po vyloučení odlehlých pozorování, „víko“ krabice udává horní kvartil, „dno“ dolní kvartil, vodorovná úsečka uvnitř krabice označuje medián.



Obr. 1.8: Krabicový graf

Z polohy mediánu vzhledem ke „krabici“ lze dobře usuzovat na symetrii vnitřních 50% dat a my tak získáváme dobrý přehled o středu a rozptýlenosti proměnné.

Pozn.: Z popisu krabicového grafu je zřejmé, že jeho konstrukci začínáme zakreslením odlehlých pozorování a až poté vyznačujeme ostatní číselné charakteristiky proměnné (min_1, max_1, kvartily a shorth).

Způsob výpočtu statistických charakteristik numerické proměnné a jejich souvislosti s histogramem a krabicovým grafem můžete pozorovat v java appletu Výběrové charakteristiky (790 KB).

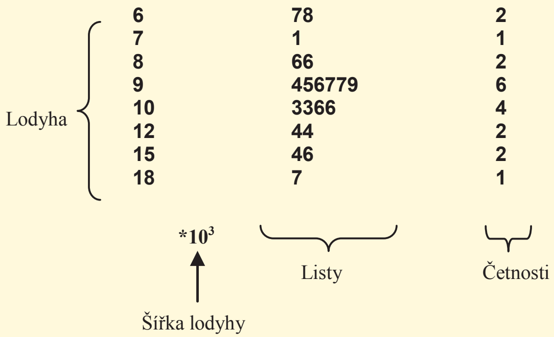
- **Číslicový histogram** (Lodyha s listy, angl. Stem and leaf plot)

Jak jsme si ukázali, výhodou krabicového grafu je jeho jednoduchost, někdy nám však chybí informace o konkrétních hodnotách proměnné. Chtěli bychom proto nějak přehledně zapsat číselné hodnoty výběru a k tomu nám slouží právě číslicový histogram. Navíc nám tento graf dává dobrou představu o šikmosti proměnné.

Představme si proměnnou představující průměrné měsíční platy zaměstnanců ve státní správě.

Průměrný měsíční plat [Kč]
10 654, 9 765, 8 675, 12 435, 9 675, 10 343, 18 786, 15 420, 8 675, 7 132, 6 732, 6 878, 15 657, 9 754, 9 543, 9 435, 10 647, 12 453, 9 987, 10 342.

A vy nyní stojíte před problémem jak tato data znázornit. Pokud se nad touto otázkou trochu zamyslíme, zjistíme, že pro naší informaci nejsou tak důležité koruny ani desetikoruny rozdílu. V tomto případě se nám jedná přinejmenším o stokoruny. Co kdybychom tedy informaci o „nedůležitých“ řádech zanedbali a znázornili setříděná data pouze na základě vyšších řádů? My jsme se rozhodli, že důležitý řád jsou pro nás stokoruny. Hodnoty stojící o řád výš (v našem případě tisíce) zapíšeme setříděné pod sebe, tak, že tvoří jakýsi stonek (**lodyhu**), přičemž pod graf uvedeme tzv. **šířku lodyhy**, která udává koeficient, jímž se hodnoty uvedené v grafu násobí.



Obr. 1.9: Číslicový histrogram

Druhý sloupec grafu, **listy**, budou tvořit číslice, reprezentující zvolený „důležitý“ řád, zapisované do příslušných řádků (opět seřazené podle velikosti). A konečně – třetí sloupec udává absolutní četnosti příslušné daným řádkům.

Jste ze slovního popisu poněkud zmateni? Prohlédněte si důkladně obrázek reprezentující číslicový histogram na Obr. 1.14. Např. první řádek reprezentuje dvě hodnoty – (6.7 a 6.8)*103 Kč, tj. 6700 Kč a 6800 Kč (koruny a desetikoruny jsme zanedbali), šestý řádek reprezentuje také dvě hodnoty – (12.4 a 12.4)*103 Kč, tj. dvě osoby s průměrným měsíčním příjmem 12400 Kč, atd. Už je to jasnější, dokázali byste tento graf sestrojit sami?

Existují různé modifikace číslicového histogramu. Např. zobrazované četnosti mohou být kumulativní, přičemž v řádku, v němž se nachází medián, se uvádí absolutní četnost (v závorce) a směrem k tomuto řádků se četnosti kumulují jednak od nejnižších hodnot, jednak od nejvyšších hodnot.

Konečně můžete namítnout, že způsobu konstrukce číslicového histogramu je pro jeden případ vždy několik. Nikde není dáno, který řád proměnné je pro zaznamenání důležitý a který už je zanedbatelný. (Srovnávali jsme platy dobře, když jsme je zaznamenali s přesností na stokoruny? Nestačilo znázornit číslicový histogram vzhledem k tisícikorunám?) Toto rozhodnutí leží vždy na tom, kdo data zpracovává. Můžeme uvést jen jedno pravidlo – dlouhé lodyhy s krátkými listy a krátké lodyhy s dlouhými listy svědčí o nevhodné volbě měřítka.

6	**78**	**2**
7	**1**	**3**
8	**66**	**5**
9	**456779**	**(6)**
10	**3366**	**9**
12	**44**	**5**
15	**46**	**3**
18	**7**	**1**

Lodyha

$*10^3$

Listy

Kumulativní četnosti

Šířka lodyhy

Obr. 1.10: Číslicový histrogram

0	66788999999	11
1	000022558	9

$*10^4$

Obr. 1.11: Nevhodná volba číslicového histogramu

Shrnutí:

K základní analýze jednorozměrných datových souborů můžete používat excelovský soubor Explorační analýza (700 KB).

Kvalitativní – Kategoriální proměnná

a) **Nominální proměnná** – nemá smysl uspořádání

Základní statistiky pro popis nominální proměnné:

- četnost
- relativní četnost
- modus

Grafické zobrazení nominální proměnné:

- histogram
- výsečový graf

b) **Ordinální proměnnná** – má smysl uspořádání

Základní statistiky pro popis ordinální proměnné:

- četnost
- relativní četnost

- kumulativní četnost
- relativní kumulativní četnost
- modus

Grafické zobrazení ordinální proměnné:

- histogram
- výsečový graf
- Lorenzova křivka
- Paretův graf

Paretův princip – 80% následků pramení z 20% příčin

Paretova analýza – postup vedoucí k nalezení „životně důležité menšiny“ (spektra příčin ovlivňujících rozhodujícím způsobem následky)

Kvantitativní – Numerická proměnná

Míry polohy

- Průměr $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Mopdus (střed shortu)
- Kvantily (dolní kvartil, medián, horní kvartil, ...)

Míry variability

- Variační rozpětí $x_{max} - x_{min}$
- Interkvartilové rozpětí $IQR = x_{0{,}75} - x_{0{,}25}$
- Výběrová směrodatná odchylka $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
- Variační koeficient $V_x = \frac{V}{\bar{x}},\ \textit{popř.}\ V_x = \frac{V}{\bar{x}} \cdot 100[\%]$

Míry šikmosti a špičatosti

- Výběrová šikmost $\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$
- Výběrová špičatost $\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme **nahoru** na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily ...) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

Identifikace odlehlých pozorování

- Vnitřní hradby: dolní mez: $h_D = x_{0,25} - 1,5IQR$
  horní mez: $h_H = x_{0,75} + 1,5IQR$
- Z – souřadnice $z - skóre_i = \frac{x_i - \bar{x}}{s}$
- Mediánová souřadnice $x_{0,5} - skóre_i = \frac{x_i - x_{0,5}}{1{,}483MAD}$

Grafické zobrazení numerické proměnné:

- Empirická distribuční funkce
- Krabicový graf (angl. Box plot)
- Číslicový histogram (lodyha s listy, angl. Stem and leaf)

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Test ze Statistiky píše velké množství studentů. Představte si, že každý z nich odpoví správně přesně na polovinu otázek. V tomto případě bude směrodatná odchylka počtu správných odpovědí

   (a) rovna průměru,
   (b) rovna mediánu,
   (c) rovna nule,
   (d) směrodatnou odchylku nelze určit bez dalších informací,
   (e) dvojnásobku módu.

2. (1b.) Největší kumulativní absolutní četnost v množině čísel se rovná

   (a) součtu všech absolutních četností,
   (b) 1,
   (c) dvojnásobku průměru,
   (d) dvojnásobku mediánu,
   (e) dvojnásobku módu.

3. (1b.) Několik studentů píše test ze Statistiky s 10-ti otázkami. Nejhorší výsledek jsou 3 správné odpovědi, nejlepší výsledek je 10 správných odpovědí. Jakou hodnotu má medián?

(a) $7\ (= 10 - 3)$

(b) $6{,}5(= \dfrac{3+10}{2})$

(c) Medián nelze určit, pokud neznáme konkrétní výsledky jednotlivých žáků.

4. (1b.) Představte si, že jste absolvovali normovaný test (např. SCIO test) a že Vám sdělili, že patříte do 91. percentilu. To znamená, že

(a) 90 žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy.

(b) 90 žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy.

(c) 90% žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy.

(d) 90% žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy.

5. (1b.) Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že

(a) medián mzdy je vyšší než průměrná mzda,

(b) medián mzdy je nižší než průměrná mzda,

(c) medián mzdy je stejný jako průměrná mzda,

(d) o vztahu mezi mediánem mzdy a průměrnou mzdou nelze rozhodnout.

6. (1b.) Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že

(a) mzdy mají kladnou šikmost,

(b) mzdy mají zápornou šikmost,

(c) mzdy mají kladnou špičatost,mzdy mají zápornou špičatost,
(d) vztah mezi průměrem a 60% kvantilem nevypovídá nic o šikmosti ani o špičatosti dat.

7. (1b.) Lékař Petře sdělil, že patří do 3. percentilu ohledně BMI (Body mass index – poměr váhy (kg) ke kvadrátu výšky (m)). Petra má pravděpodobně

(a) podváhu,
(b) normální váhu,
(c) nadváhu,
(d) bez dalších informací nelze usuzovat na Petřinu váhu.

8. (1b.) Představte si, že jste absolvovali normovaný test (např. SCIO test). Měl(a) jste lepší výsledek než 85 studentů ze 100. To znamená, že

(a) patříte do 99. decilu,
(b) patříte do 95. decilu,
(c) patříte do 10. decilu,
(d) patříte do 9. decilu,
(e) patříte do 2. kvartilu.

9. (1b.) Pro srovnání variability váhy a výšky je možné použít

(a) průměr,

(b) rozptyl,
(c) směrodatnou odchylku,
(d) variační koeficient,
(e) šikmost.

10. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, průměrný plat ve firmě se zvýší

(a) o 100,- Kč,
(b) o 1000,- Kč,
(c) průměrný plat se nezmění.

11. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, průměrný plat ve firmě se zvýší

(a) dvojnásobně,
(b) čtyřnásobně,
(c) průměrný plat se nezmění.

12. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, průměrný plat ve firmě se zvýší

(a) o 20%,
(b) o 400%,

(c) o 40%,
(d) o 44%,
(e) Průměrný plat se nezmění.

13. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, rozptyl platů ve firmě se zvýší

(a) o 100,- Kč,
(b) o 1000,- Kč,
(c) rozptyl platů se nezmění.

14. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, rozptyl platů ve firmě se zvýší

(a) dvojnásobně,
(b) čtyřnásobně,
(c) rozptyl platů se nezmění.

15. (1b.) Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, rozptyl platů ve firmě se zvýší

(a) o 20%,
(b) o 400%,
(c) o 40%,

(d) o 44%,
(e) Rozptyl platů se nezmění.

16. (1b.) Největší kumulativní relativní četnost se rovná

(a) dvojnásobku průměru,
(b) dvojnásobku mediánu,
(c) dvojnásobku módu,
(d) součtu všech jednotlivých hodnot absolutních četností,
(e) 1.

17. (3b.) Určete, zda jsou následující tvrzení pravdivá.

(a) Geometrický průměr je definován pro proměnné, které nabývají pouze kladných hodnot.Jedna čtvrtina hodnot je větší než 25% kvantil, zatímco tři čtvrtiny hodnot jsou menší.
(b) Mají-li dvě proměnné stejný průměr a stejný rozptyl, mají stejný variační koeficient.
(c) Mzdy v ČR mají kladnou šikmost. (V ČR mají zhruba 2/3 lidí podprůměrný plat.)
(d) Nejčetnější hodnota v souboru se nazývá medián.
(e) Rozptyl má vždy kladnou hodnotu.

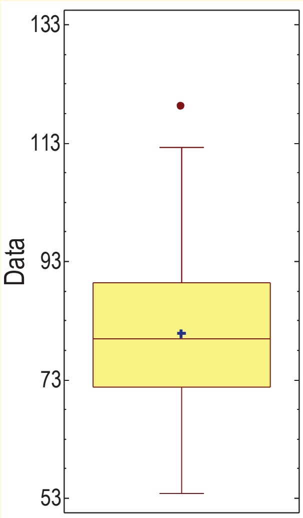
18. (1b.) V grafu na Obr. 1.12, modrý křížek označuje

(a) medián,

(b) průměr,

(c) modus,

(d) Interkvartilové rozpětí (IQR).



Obr. 1.12: Proměnná x

19. (2b.) Určete zda jsou následující tvrzení pravdivá. Proměnná znázorněna na Obr. 1.12

(a) neobsahuje odlehlá pozorování,

(b) má kladnou šikmost,
(c) je kladná,
(d) má více než polovinu hodnot větších než 83.

20. (1b.) Na atletických závodech mládeže žáci soutěžili ve 4 kategoriích. Určete, který výrok je nepravdivý.

(a) Na obrázku je znázorněn histogram a nejméně soutěžících bylo ve skoku do dálky.
(b) Celkem ve čtyřech kategoriích soutěžilo 80 žáků.
(c) Modus = hod koulí.
(d) Modus = 30.

Obr. 1.13: Zastoupení žáků na atletických závodech

21. (2b.) Následující graf Stem&leaf reprezentuje množství peněz, které studenti jedné třídy vybrali na humanitární účely.

0	11555889	8
1	112344555	(9)
2	005	6
3	025	3

Multiply by 10^2

Které z následujících výroků jsou určitě nepravdivé?

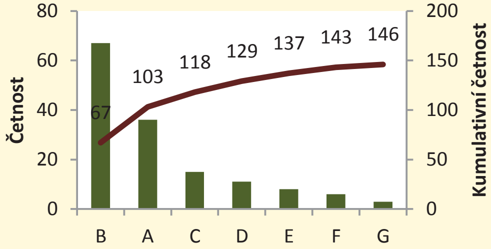
(a) 10 studentů věnovalo méně než 120 Kč.

(b) Medián vybrané částky činí 120 Kč.

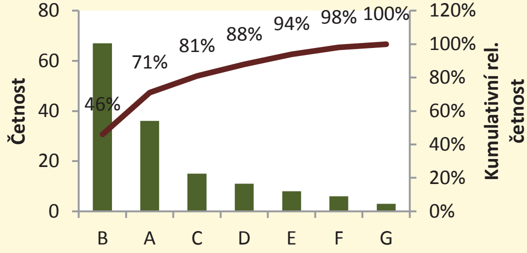
(c) Na humanitární účely přispělo v této třídě 23 studentů.

(d) Přispívající studenti věnovali na humanitární účely částky od 1,- Kč do 35,- Kč.

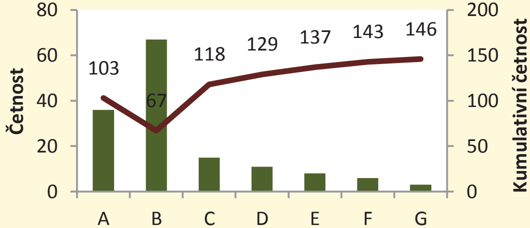
22. (1b.) Určete, na kterém obrázku je zobrazen Paretův graf.



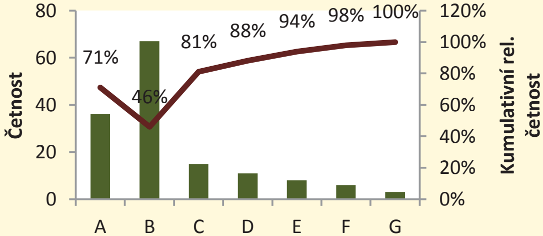
a)



b)



c)



d)

(a) A)
(b) B)
(c) C)

Obsah
75. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

(d) D)

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

Jak pracovat s testy?

Zacátek testu

1. Zemědělské družstvo dostalo 1 000 kuřat s průměrnou váhou 1,37 kg. Cena byla 50,- Kč za kilogram. Během dne se prodalo 300 kuřat za 24 000,- Kč. Jaká byla průměrná váha neprodaných kuřat?

   kg

2. V jisté společnosti je průměrný plat 13 500,- Kč. 30% pracovníků s nejnižším platem má průměrně 9 000,- Kč. Na začátku roku došlo ke zvýšení platů pracovníků této skupiny jednotně o 500,- Kč. O kolik % vzrostl průměrný plat v celé společnosti následkem uvedeného zvýšení platu?

   %

3. Petr, řidič zkušebního automobilu, jel z Ostravy do Olomouce rychlostí 70 km/h. Zpět jel rychlostí 90 km/h. Jaká byla průměrná rychlost zkušebního automobilu na trase Ostrava – Olomouc – Ostrava?

   km/h

4. V jistém supermarketu byla ve stejné chvíli na 8 pokladnách měřena doba, během které pokladní ověří platnost platební karty zákazníka v bance. U pěti zákazníků trvalo ověření 2 minuty, u zbývajících tří to byly 3 minuty. Určete průměrnou dobu potřebnou k ověření platnosti karty.

   min

5. Nákladní automobil jel z města A do města B rychlostí 40 km/h, z města B do města C rychlostí 50 km/h a z města C do města D rychlostí 60 km/h. Vypočítejte průměrnou rychlost, které dosáhl automobil na celé trase, víte-li, že:

(a) vzdálenost všech úseků je stejná – 5 km.
☐ km/h

(b) Vzdálenost z A do B je 15% trasy a vzdálenost z C do D je 60% trasy.
☐ km/h

6. Cena jedné akcie energetické společnosti vzrostla na burze XY v období od 13. do 15. března téhož roku z 952,50 Kč na 982,00 Kč. Jaký byl průměrný relativní přírůstek ceny této akcie?

☐ %

7. Při sledování proměnné x byl určen aritmetický průměr 110 a rozptyl 800. Dodatečně byly zjištěny chyby u dvou údajů. Místo 85 mělo být správně 95 a místo 120 má být 150. Ostatních 18 údajů bylo správných. Opravte vypočítané charakteristiky (průměr a rozptyl).

$\bar{x} =$ ☐

$s^2 =$ ☐

8. Ze čtyřiceti hodnot byl vypočítán aritmetický průměr 7,50 a rozptyl 2,25. Při kontrole bylo zjištěno, že chybí dvě hodnoty proměnné – 3,8 a 7. Opravte uvedené charakteristiky.

$\bar{x} =$ ☐

$s^2 = \boxed{}$

9. V důsledku výstavby satelitního městečka poklesl průměrný věk obyvatel vesnice o 19%, rozptyl věku vzrostl o 21%. Jak se změnil variační koeficient?
(vzrůst o – kladně, pokles o – záporně)

$\boxed{}$%

10. Ze známých dat byl určen rozptyl měsíčních mezd 250 000 $K\check{c}^2$. Určete směrodatnou odchylku mezd, zvýší-li se všechny měsíční mzdy

(a) o 150,- Kč $\boxed{}$

(b) 1,2 krát $\boxed{}$

(c) o 4%. $\boxed{}$

11. Máme n údajů o měření teploty ve ^oC. Průměrná teplota je 20^oC a rozptyl je 10^oC^2. Určete

(a) průměrnou teplotu ve stupních Fahrenheita, $\boxed{}\,^oF$

(b) rozptyl teploty ve stupních Fahrenheita, $\boxed{}\,^oF$

(c) variační koeficienty teploty ve stupních Celsia (^oC) a ve stupních Fahrenheita (^oF). (Vztah pro převod stupňů Celsia na stupně Fahrenheita: $T_{^oF} = 1,8 \cdot T_{^oC+32}$)

$V_{^oC} = \boxed{}$%

$V_{^oF} = \boxed{}$%

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Kapitola 2

Statistické šetření

Cíle

Po prostudování tohoto odstavce budete

- rozumět pojmům: základní soubor (populace), výběr, statistická jednotka, statistický znak, výběrové šetření,
- umět srovnat vyčerpávající a výběrové šetření,
- znát typy výběrových šetření,
- rozumět principům experimentu a pozorovací studie,
- znát možná rizika (chyby) výběrových šetření.

Motto:

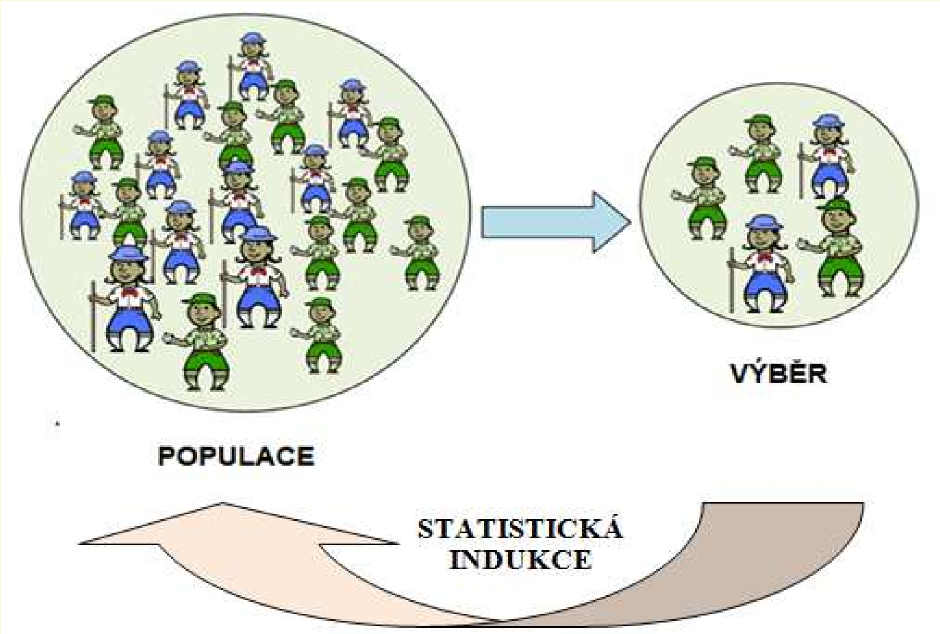
Chceme-li vědět, jak chutná víno v sudu, nemusíme vypít celý sud.
Stačí jenom malý doušek a víme, na čem jsme.

Statistika je věda o sběru, zpracování a vyhodnocování dat. V praxi většinou nemáme tolik času, energie a financí, abychom mohli pro učinění svého rozhodnutí prozkoumat všechny údaje vztahující se k analyzovanému problému. V mnoha oborech se proto setkáme s průzkumy opírajícími se o relativně malou část (**výběr, vzorek**) z dotčených dat (**základní soubor, populace**). Statistika pak používá postupy, pomocí nichž můžeme, sice s určitým (odhadnutelným) rizikem, na základě vlastností vzorku usuzovat na chování populace. Souboru metod, které umožňují usuzovat na vlastnosti populace z vlastností výběru se říká **statistická indukce**.
Provádění statistického průzkumu se většinou řídí následujícími čtyřmi kroky.

1. **Formulace problému** (co chceme zjistit, koho (resp. čeho) se daný problém týká).
2. **Sběr dat** (tzv. statistické šetření).
3. **Analýza** shromážděných dat vedoucí k získání potřebné informace.
4. **Vyhodnocecní** získané **informace**, tj. poznání.

V této kapitole budou zavedeny základní pojmy matematické statistiky a následně se zaměříme na druhy statistického šetření, tj. na způsoby sběru dat. V dalším kroku statistického průzkumu lze získaná data analyzovat metodami explorační analýzy. Statistická indukce, umožňující extrapolaci informací z výběru na celou populaci, je pak postupně popsána v kapitolách 8 až 14.

Obsah
82. strana ze 525
Zavřít dokument
Celá obrazovka/Okno



Obr. 2.1: Princip statistické indukce

2.1. Základní pojmy matematické statistiky

Je známo, že většina pozorování zaznamenaných v technické i ekonomické praxi, stejně jako v přírodních i humanitních vědách, vykazuje náhodné kolísání. Při opakovaných měřeních téže fyzikální veličiny (teploty, tlaku, ...), životnosti výrobků téhož typu, podobně jako při opakovaných měřeních biometrických údajů osob téhož pohlaví a věku nedostaneme stále stejné výsledky. Na zjištěná pozorování se pak díváme z pravděpodobnostního hlediska jako na výsledky náhodného pokusu prováděného na množině nějakých případů nebo předmětů.

Opakujeme-li n-krát nezávisle náhodný pokus, jehož výsledkem je hodnota náhodné veličiny X s distribuční funkci $F(x, \theta)$, kde θ je reálný parametr (resp. vektor parametrů) daného rozdělení pravděpodobnosti, pak pozorujeme náhodný vektor $\boldsymbol{X} = (X_1, ..., X_n)$, jehož složkami jsou nezávislé náhodné veličiny X_i se stejným rozdělením pravděpodobnosti. Náhodný vektor X se nazývá **náhodný výběr** (z náhodné veličiny X) a n je **rozsah** náhodného **výběru**.

Číselný vektor, který získáme jako realizaci (pozorovanou hodnotu) náhodného výběru budeme nazývat **statistický soubor**. Jeho prvky se nazývají **statistické jednotky**.

Soubor všech možných statistických jednotek, tj. obor hodnot náhodné veličiny X, se nazývá **základní soubor** (**populace**).

Na statistických jednotkách daného souboru pak sledujeme určitou vlastnost statistických jednotek (životnost výrobků, barvu laku, hmotnost, IQ, pohlaví, věk), kterou označujeme jako **statistický znak**.

Obsah
84. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

2.2. Způsoby statistického šetření

Pro většinu statistických souborů, s nimiž se v praxi setkáváme, je typický vysoký rozsah (počet zkoumaných jednotek). Jakmile jsme tedy postaveni před úkol provést určité šetření a analyzovat údaje z něj zjištěné, musíme nejprve rozhodnout, zda budeme toto šetření realizovat jako vyčerpávající nebo výběrové.

Vyčerpávající šetření (úplné šetření, census) - prošetření všech jednotek statistického souboru (populace). Příkladem je sčítání lidu, domů a bytů k určitému rozhodnému okamžiku a sledování demografických jevů, jako je narození nebo úmrtí. Zpravidla se jedná o záležitost velmi nákladnou (personálně, finančně, časově), mnohdy dokonce prakticky nerealizovatelnou (destrukční zkoušky). Pokud však toto šetření proběhne, mezi jeho nesporné výhody patří přesnost zjištěných charakteristik a detailnost informací o každé zkoumané jednotce. V praxi se, z výše uvedených důvodů, dává většinou přednost šetřením výběrovým.

Výběrové šetření (neúplné šetření) - ze základního souboru (populace) o rozsahu N vybereme jeho část, tzv. **výběrový soubor**, zkráceně **výběr**, o rozsahu n. Tento výběr zpracujeme a z výsledků pak usuzujeme na vlastnosti celé populace. Výběrová šetření se používají například při zjišťování jaká je podpora politických stran, při ověřování pevnosti trubek vyráběných určitým podnikem, apod. Mírou objektivnosti informací, které získáme, je kvalita provedení výběrového šetření. Podrobněji se typům výběrových šetření budeme věnovat v kapitole 7.3.

Zkoumají-li se kauzální závislosti, tedy vliv různých zásahů, používá se pro statistické zjišťování tzv. **experiment** (např. vyhodnocení účinnosti nového léku, zkoumání vlivu způsobu výuky čtení na kvalitu čtení na konci 1. třídy, ...). Experiment je většinou založen na tom, že některé náhodně vybrané prvky populace jsou podrobeny zásahu (intervenci), jejíž efekt se zkoumá, zatímco zbylé slouží jako kontrolní skupina. V ideálním případě by měli být pokusné subjekty i posuzovatelé experimentu drženi v nevědomosti ohledně zařazení subjektu do pokusné, resp. kontrolní skupiny. Je-li experimentem vyhodnocení účinnosti nového léku, může experiment narušit jak to, že pacient ví, do které skupiny byl zařazen (placebo efekt), tak i to, že tuto informaci má lékař (favorizování pokusných subjektů). Neví-li pokusný subjekt, do které skupiny je zařazen, mluvíme o utajeném pokusu, neví-li to ani posuzovatel, označujeme situaci jako dvojité utajení. **Znáhodněný a utajený pokus** zajišťuje, že obě skupiny jsou od počátku experimentu v zásadě rovnocenné a jako rovnocenné jsou i po celou dobu experimentu udržovány. Rozdíl mezi pokusnou skupinou (skupinou podrobenou zásahu) a kontrolní skupinou pak lze až na výběrovou chybu interpretovat jako vliv zásahu.

Posledním zmíněným způsobem statistického průzkumu je **pozorovací studie**. Podobně jako experiment, pozorovací studie umožňuje zkoumat kauzální závislosti. V případě pozorovací studie výzkumník do pokusu nezasahuje, pouze pozoruje, jak pokus probíhá u těch, kteří se jej účastní. Přestože tyto studie bývají často méně uspokojivé než znáhodněné experimenty, stává se, že jsou jediným způsobem, jak lze daný problém řešit. (Zkoumáme-li například vliv kojení na citovou vazbu matky a dítěte, probíhal by znáhodněný pokus tak, že by byly náhodně stanoveny matky, které budou své dítě kojit, a pak by se sledovalo, jak se vyvíjí citové vazby mezi matkami a jejich dětmi v průběhu deseti let. Protože nelze nařídit matkám, aby své dítě kojily (resp. nekojily), použijeme pozorovací studii.)

Obr. 2.2: Druhy statistického zjišťování

2.3. Typy výběrových šetření

Výběrová šetření dělíme do dvou základních skupin.

- **Náhodné výběry** (pravděpodobnostní výběry, angl. „probability samples“)
  V náhodných výběrech má každá jednotka populace známou (nenulovou) pravděpodobnost, že bude zařazena do výběru.
- **Nenáhodné výběry** (nepravděpodobnostní výběry, angl. „non-probability samples“)

V případě nenáhodných výběrů neznáme pravděpodobnost zařazení jednotlivých jednotek populace do výběru nebo si nemůžeme být jistí, zda je tato pravděpodobnost pro každou jednotku populace nenulová.

2.3.1. Nenáhodné výběry

Mezi hlavní druhy nenáhodných výběrů patří anketa, metoda základního masivu a záměrný výběr.

Anketa (angl. „voluntary sample“) oslovuje pouze nesystematicky vybranou část populace (osob, podniků, institucí). Dotazník s pečlivě sestavenými otázkami a se žádosti o jejich vyplnění a vrácení se k respondentům (dotazovaným) dostává prostřednictvím sdělovacích prostředků (anketa televizních diváků, anketa časopisu Mládí, ...) nebo je zaslán adresně, přičemž návratnost dotazníku je obvykle malá (odhaduje se, že 30 %). Výběr statistických jednotek je založený na rozhodnutí respondenta zúčastnit se průzkumu. Vzhledem k tomu, že nelze definovat populaci, ke které se nálezy ankety vztahují, nelze informace získané anketním šetřením zobecňovat.

Metoda základního masivu se používá v případech, kdy se základní soubor skládá z několika velkých jednotek a z většího počtu jednotek malých. Např. při šetření v oblasti hutnictví se můžeme podle této metody zaměřit na několik „obřích“ společností, tam provést šetření a „malé“ podniky vynechat. Výhody: menší pracnost a menší časová náročnost šetření. Nevýhody: zobecnění poznatků má menší platnost (nevystihuje specifika menších jednotek).

Záměrný (účelový, úsudkový) výběr spočívá v tom, že skupina odborníků na danou problematiku vybere podle svého nejlepšího uvážení ty jednotky, o nichž se lze domnívat, že ve svém souhrnu nejlépe umožní provést šetření. S tímto typem šetření se často setkáme například při průzkumech trhu a při průzkumech veřejného mínění. Záměrný výběr se provádí jako

- **výběr typický**, neboli výběr jednotek pro danou populaci typických (například zaměstnanci s platem blízkým průměrnému platu),
- **výběr konvenční**, kdy jsou do výběru zařazovány jednotky nejsnadněji dostupné – např. prvních 100 zákazníků prodejny, nebo
- **výběr kvótní**.

Kvótní výběr usiluje o strukturální shodu výběrového souboru se souborem základním (populaci). Je-li například v populaci 51 % žen, do výběru zařadíme 51 % žen, ... Používá se tehdy, když je známá struktura základního souboru, ale základní soubor je obtížně definovatelný jako soubor konkrétních jednotek (např. neexistuje jejich seznam). Výběr statistických jednotek do kvótního výběru probíhá na základě kritérií daných kvótou. Takovým kritériem může být například zastoupení jednotek podle pohlaví, věku, vzdělání... V praxi se používá maximálně 3 až 5 kritérií, která mohou být nezávislá nebo vzájemně provázána (kombinována).

Subjektivní přístup k záměrnému výběru zpochybňuje možnost zobecnění, a to i v případě kvótního výběru, který je reprezentativní pouze z hlediska znaků použitých ve kvótách.

2.3.2. Náhodné výběry

Pro náhodné výběry je charakteristické, že dobře reprezentují všechny známé i neznámé vlastnosti populace. Otázkou zatím zůstává jak náhodný výběr získat.

Prostý náhodný výběr (angl. „simple random sampling“)
V praxi nejpoužívanějším typem náhodného výběru je **prostý náhodný výběr**. Je to takový výběr o rozsahu n, při kterém mají všechny myslitelné n-členné kombinace jednotek základního souboru stejnou pravděpodobnost stát se výběrovým souborem. Při prostém náhodném výběru rozlišujeme mezi **výběrem s vracením** (každá jednotka je po výběru vracena zpět do základního souboru) a **výběrem bez vracení** (každá jednotka základního souboru může být do výběru zařazena nejvýše jednou). Připomeňme si, že z pravděpodobnostního hlediska má výběr s vracením charakter nezávislých pokusů (Bernoulliho pokusy, binomické rozdělení), zatímco výběr bez vracení má charakter pokusů závislých (hypergeometrické rozdělení). Je-li rozsah základního souboru mnohem větší (v praxi – alespoň dvacetkrát) než rozsah výběru, je rozdíl mezi výběry s vracením a bez vracení zanedbatelný.

Nejznámější technikou získání prostého náhodného výběru je **losování**. Při losování postupujeme tak, že každé jednotce základního souboru přiřadíme pořadové číslo. Soubor těchto „zástupců“ statistických jednotek (čísel, resp. značek) se obecně nazývá **opora výběru**. Tyto „zástupce“ napíšeme na lístečky a vložíme je do osudí. Osudí důkladně promícháme a vybereme tolik lístečků s čísly, jaký požadujeme rozsah výběru. (Provádí-li se výběr s vracením, je promíchání třeba opakovat po každém vracení.) V případě, že je základní soubor příliš rozsáhlý a losování se tak stává technicky neproveditelné, využíváme pro výběr z opory výběru **generátorů náhodných čísel** (agl. „random number generator“), které jsou dnes běžnou součástí statistického software.

Systematický výběr (angl. „systematic random sampling“) Jiným způsobem náhodného výběru je **výběr systematický**, kdy se první jednotka výběru vybere náhodně (metodou prostého náhodného výběru) a dále se vybírá každá k-tá jednotka základního souboru. Nevýhodou systematického výběru je skutečnost, že není zaručeno náhodné pořadí jednotek v základním souboru (může existovat skrytá pravidelnost v opoře výběru).

Kromě výše zmíněných přímých technik výběru používáme při některých zjišťováních složitější uspořádání výběru, které je založeno na dělení základního souboru na menší či větší podskupiny (může být provedeno ve vícero krocích), z nichž se teprve vybírají statistické jednotky. Takové dělení zajistí, aby nedocházelo k vytváření takových výběrových souborů, jež by dávaly silně nadhodnocené nebo podhodnocené odhady sledovaných skutečností.

Rozlišujeme dva základní způsoby složitějšího uspořádání náhodného výběru – náhodný stratifikovaný výběr a vícestupňový výběr.

Stratifikovaný výběr (angl. „stratified sampling“) V případě stratifikovaného výběru se snažíme o to, aby jednotlivé podskupiny obsahovaly jednotky stejných vlastností, tj. aby byly homogenní vzhledem k nějakému jasnému kritériu. Statistické jednotky jsou pak z podskupin, které bývají v tomto případě nazývány **oblastmi** (angl. „strata“), vybírány metodou prostého náhodného výběru. Oblastmi zde nemusí být pouze oblasti územní, mohou to být rovněž věkové kategorie, skupiny lidí s různým vzděláním, pohlavím, výrobky z různých výrobních linek, apod. (Například při zjišťování o studentech určité školy je vhodné jedince vybírat zvlášť z jednotlivých ročníků.)

Stratifikovaný výběr je oproti prostému náhodnému výběru náročnější na organizaci a zpracování výsledků. Je-li však správně proveden, pak jsou jednotlivé oblasti stejnorodějším celkem než původní základní soubor a stratifikovaný výběr nám tak umožní získat kvalitnější informace o základním souboru.

Vícestupňový výběr (angl. „cluster sampling“) V případě, že základní soubor je příliš rozsáhlý a prostorově rozptýlený, stoupá finanční, časová i personální náročnost prostého náhodného výběru. Překážkou pro provedení prostého náhodného výběru bývá rovněž, v praxi poměrně běžná, neexistence opory výběru (seznamu populace). V takovýchto případech přistupujeme k výběru vícestupňovému. U vícestupňového výběru jsou jednotlivé podskupiny, na rozdíl od stratifikovaného výběru, zastupitelné. Výběr statistických jednotek pak probíhá pouze z náhodně vybraných podskupin. (Příklad: Při předvolebním průzkumu vybíráme postupně okresy, v nich obce, v nich volební okrsky a v nich teprve respondenty.)

2.4. Chyby ve výběrových šetřeních

Připomeňte si, že výběrová šetření v podobě reprezentativních výběrů se používají proto, aby mohly být vytvářeny úsudky o základním souboru (populaci) jinak než na základě časově, finančně nebo personálně náročného vyčerpávajícího šetření. Je zřejmé, že i v případě, kdy je při výběrovém šetření použit náhodný výběr, nemusí tento výběr základní soubor reprezentovat zcela přesně. Rozdíl mezi naměřenou hodnotou hledaného populačního parametru (výběrovou charakteristikou) a jeho skutečnou hodnotou (populační charakteristikou) bývá v tomto případě označován jako **náhodná chyba výběru** (angl. „random error“). S rostoucím rozsahem výběru se náhodná chyba výběru obvykle snižuje.

Pokud se při výběrovém šetření neuplatní vhodné metody výběru, mohou být vykreslovány grafy, počítány číselné charakteristiky a vytvářeny závěry, ale všechny tyto informace budou zatíženy velkým rizikem zkreslení a vychýlení. Na co je třeba, zejména při průzkumech veřejného mínění, dávat pozor?

2.4.1. Výběrová chyba

Základním pravidlem dobře vedeného průzkumu je zásada, že výběr musí být reprezentativní, tzn. že všechny jednotky, z nichž se skládá populace, musí mít stejnou šanci na zařazení do zkoumaného výběru. Nedodržení tohoto pravidla vede k nejčastější a nejzávažnější chybě v průzkumech, které se říká **výběrová chyba** (angl. „selection bias“).

Pravděpodobně „nejslavnějším“ případem výběrové chyby je případ časopisu Literary Digest, který byl počátkem 20. století mimořádně populární v USA. V roce 1936 provedl časopis Literary Digest průzkum mezi 2,4 milióny respondentů o tom, zda v prezidentských volbách budou volit demokrata Franklina Rooswelta nebo republikána Alfreda Landona. Přestože většina (57 %) respondentů průzkumu uvedla, že by volila A. Landona, volby vyhrál F. D. Roosvelt s 62 % odevzdaných hlasů. Jak je možné, že takto rozsáhlé výběrové šetření vedlo k tak velké chybě? Chyba vznikla v důsledku konvenčního výběru. Redaktoři sice oslovili 2,4 miliónů respondentů, ty však oslovili na základě telefonních seznamů a seznamů klubových členství. Tento způsob výběru, bohužel, vyřadil z průzkumu občany z méně majetných vrstev, pro které nebylo v roce 1936 běžné ani vlastnictví telefonů ani členství v klubech. Právě tato část společnosti se v roce 1936 výrazně přiklonila k demokratům. Jde o ukázku toho, že i velký rozsah výběru, který není reprezentativní, může vést k chybným závěrům.

Speciálním případem výběrové chyby je chyba, která vzniká v důsledku toho, že oslovení respondenti průzkumu odmítnou odpovídat (angl. „nonresponse bias“). Například při telefonních průzkumech se často stává, že lidé jsou příliš zaměstnaní a příliš často jim volá někdo s obchodní nebo jinou nabídkou, než aby měli chuť a čas trávit půl hodiny na lince a odpovídat na dotazy tazatele. Situace je o to horší, oč se názory právě těchto lidí liší od názorů většinové populace.

2.4.2. Chyba v měření

Další častou chybou průzkumu veřejného mínění je tzv. **chyba v měření** (angl. „bias due to measurement error“). K této chybě dochází v případech, kdy samotná otázka (resp. množina odpovědí na otázku) má nežádoucí vliv na odpovědi respondentů. Každé slovo v otázce, stejně jako pořadí otázek, či intonace jakou se tazatel ptá, by mělo být pečlivě promyšleno. Uvedeme si dva příklady vedoucí k chybě v měření. První z nich je poměrně obecný.

Představme si průzkum spokojenosti zákazníků. Zákazník má zhodnotit míru své spokojenosti s produktem a má na výběr z možností: spokojen, nespokojen, velmi nespokojen. Je zřejmé, že respondent má pouze jednu možnost pro vyjádření spokojenosti a dvě možnosti pro vyjádření nespokojenosti. Průzkum tedy bude vychýlen k vyjádření nespokojenosti. (Zamyslete se nad tím, jaké možnosti odpovědi by měly být respondentovi nabídnuty.)

Další příklad je již konkrétní. V roce 1995 ohlásil Bill Clinton, že vyšle 20 000 amerických vojáků do Bosny. Následně byly zveřejněny výsledky několika průzkumů veřejného mínění.

- CNN: 46 % pro/ 14 % neví / 40 % proti,
- ABC: 39 % pro/ 4 % neví / 57 % proti,

- CBS: 33 % pro/ 9 % neví / 58 % proti.

Proč dopadl průzkum CNN výrazně lépe pro Clintona, než ostatní dva průzkumy? Přesně to nevíme, ale svůj podíl měly zřejmě dvě skutečnosti.

- V otázce CNN, na rozdíl od otázek ABC a CBS, nebyl uveden počet vojáků, kteří se měli mise zúčastnit.
- CNN vojáky popsala jako „mezinárodní mírové síly prosazující mírovou dohodu“, zatímco CBS volila příkřejší slova.

Shrnutí:

Statistika používá postupy pomocí nichž můžeme, sice s určitým rizikem (předem stanoveným), na základě části dotčených dat (**výběru**) usuzovat na chování celku (**populace**). Tomuto zobecňování říkáme **statistická indukce**.

Jakmile jsme postavení před úkol provést určité šetření a analyzovat údaje z něj zjištěné, musíme se obvykle nejprve rozhodnout, zda budeme toto **šetření** realizovat jako **vyčerpávající nebo výběrové**.

Vyčerpávající šetření – to je prošetření všech jednotek statistického souboru (populace).

Výběrové šetření – jde o prošetření vybraných jednotek statistického souboru (populace).

Zkoumají-li se kauzální závislosti, tedy vliv různých zásahů, používá se pro statistické zjišťování **experiment** nebo **pozorovací studie**.

Výběrová šetření dělíme do dvou základních skupin – na **výběry náhodné** a **výběry nenáhodné**.

Mezi nenáhodné výběry řadíme **anketu, metodu základního masivu** a **záměrný výběr**.

Základním typem náhodných výběrů je **prostý náhodný výběr**, kdy se výběr jednotek provádí nejčastěji **losováním** z kódů uvedených v **opoře výběru**. Není-li losování technicky možné, využívá se pro výběr statistických jednotek **generátoru náhodných čísel**.

V případě, že je zaručeno náhodné pořadí statistických jednotek v základním souboru (populaci), je vhodnou alternativou k prostému náhodnému výběru **výběr systematický**, kdy se první jednotka do výběru volí náhodně a dále se vybírá každá k-tá jednotka.

Při některých zjišťováních používáme složitější uspořádání výběru, které je založeno na dělení základního souboru na menší či větší podskupiny (může být provedeno ve vícero krocích), z nichž se teprve vybírají statistické jednotky. Rozlišujeme dva základní způsoby složitějšího uspořádání náhodného výběru – náhodný stratifikovaný výběr a vícestupňový výběr.

V případě **stratifikovaného výběru** se snažíme o to, aby jednotlivé podskupiny obsahovaly jednotky stejných vlastností, tj. aby byly homogenní vzhledem k nějakému jasnému kritériu. Statistické jednotky jsou pak z podskupin, které bývají v tomto případě nazývány **oblastmi** (angl. „strata“), vybírány metodou prostého náhodného výběru.

V případě, že základní soubor je příliš rozsáhlý a prostorově rozptýlený, stoupá finanční, časová i personální náročnost prostého náhodného výběru. Překážkou pro provedení prostého náhodného výběru bývá rovněž, v praxi poměrně běžná, neexistence opory výběru (seznamu populace). V takovýchto případech přistupujeme k **výběru vícestupňovému**. U vícestupňového výběru jsou jednotlivé podskupiny, na rozdíl od stratifikovaného výběru, navzájem zastupitelné.

Je zřejmé, že i v případě, kdy je při výběrovém šetření použit náhodný výběr, nereprezentuje většinou tento výběr základní soubor zcela přesně. Rozdíl mezi naměřenou hodnotou (výběrovou charakteristikou) a hodnotou skutečnou (populační charakteristikou) bývá v tomto případě označován jako **náhodná chyba výběru** (angl. „random error“). S rostoucím rozsahem výběru se náhodná chyba výběru snižuje.

Při statistickém zjišťování si musíme dávat pozor zejména na **výběrovou chybu**, tj. chybu, která vzniká v důsledku nereprezentativnosti výběru, a na **chybu v měření**, s níž se setkáváme zejména při dotazníkových šetřeních, kdy nevhodně položena otázka ovlivňuje odpověď respondenta.

Otázky k zamyšlení

1. Definujte pojmy
   a) náhodný výběr,
   b) statistická jednotka,
   c) základní soubor (populace),
   d) statistický znak.
2. V čem spočívá technika sběru dat nazývaná experiment?
3. Uveďte alespoň tři modelové situace, v nichž by bylo pro sběr dat vhodné použít experiment, resp. pozorovací studii.
4. Srovnejte výhody a nevýhody úplného a neúplného šetření.
5. Co musí splňovat výběr, aby mohl být označen za reprezentativní?
6. Popište základní způsoby nenáhodného výběru, tj. vysvětlete pojmy
   a) anketa,
   b) metoda masivního výběru,
   c) záměrný výběr (typický výběr, konvenční výběr, kvótní výběr).
7. Jakými způsoby lze získat prostý náhodný výběr? Co je to opora výběru?
8. V čem spočívá riziko (nevýhoda) systematického výběru?
9. Jaký je rozdíl mezi stratifikovaným a vícestupňovým výběrem?
10. Jaké chyby jsou spojeny se sběrem dat prostřednictvím dotazníkových šetření (průzkumu veřejného mínění, analýzy spokojenosti, průzkum trhu, ...)?

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Výběr, při kterém ze seřazeného základního souboru vybereme z prvních k prvků náhodně jeden prvek a od něho počítajíc vybereme $k-$tý, $2k-$tý, … prvek, nazýváme

   (a) stratifikovaný náhodný výběr,
   (b) reprezentativní výběr,
   (c) systematický výběr,
   (d) vícestupňový výběr.

2. (1b.) Anketu řadíme mezi

   (a) náhodné výběry,
   (b) nenáhodné výběry.

3. (1b.) Renovovaný supermarket chce zjistit spokojenost zákazníků s novým uspořádáním prodejny. Pracovníci managementu navrhují různé způsoby výběrového šetření (viz níže). Vyberte z navrhovaných způsobů šetření ten, který odpovídá představě náhodného výběru a zároveň je nejméně nákladný.

   (a) oslovit prvních 100 zákazníků, kteří přijdou do prodejny,
   (b) oslovit každého 20. zákazníka, který přijde do prodejny,

(c) vytvořit slosovatelné anketní lístky s otázkami a nechat je na viditelném místě u vchodu prodejny,
(d) obvolat náhodně z telefonního seznamu vybraných 200 lidí bydlících v okolí supermarketu.

4. (1b.) Metodu výběru, při které základní soubor rozdělíme do několika oblastí, ve kterých provedeme náhodný výběr, nazýváme

(a) stratifikovaný náhodný výběr,
(b) reprezentativní výběr,
(c) systematický výběr,
(d) vícestupňový výběr.

5. (1b.) Chyba vzniklá v důsledku toho, že velká část respondentů průzkumu odmítla odpovídat, je označována jako

(a) systematická chyba,
(b) výběrová chyba,
(c) chyba v měření,
(d) náhodná chyba výběru.

6. (1b.) Soubor „zástupců“ jednotek základního souboru nazýváme

(a) generátor náhodných čísel,

(b) základ losování,
(c) opora výběru,
(d) rozsah výběru.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Kapitola 3

Výběrové charakteristiky

Cíle

Po prostudování této kapitoly byste měli

- rozumět pojmům populační charakteristika a výběrová charakteristika,
- znát princip statistické indukce,
- znát a umět používat zákon velkých čísel a centrální limitní větu,
- znát rozdělení výběrového průměru a rozdílů dvou výběrových průměrů při dostatečně velkých výběrech, popř. výběrech z normálního rozdělení,
- znát rozdělení relativní četnosti a rozdílů dvou relativních četností při dostatečně velkých výběrech,
- znát speciální výběrová rozdělení - χ^2- rozdělení, Studentovo rozdělení a Fisherovo-

-Snedecorovo rozdělení,

- znát vlastnosti výše uvedených speciálních výběrových rozdělení, které umožňují popsat rozdělení průměru (resp. rozdílů průměrů) pro malé výběry a výběrového rozptylu (resp. poměru výběrových rozptylů) pro výběry z normálního rozdělení.

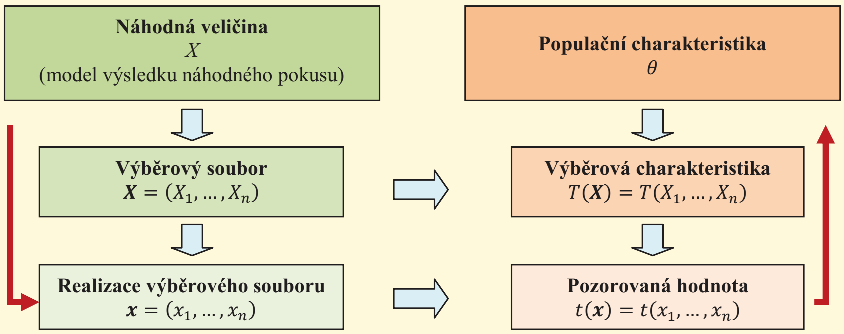
3.1. Parametry populace vs. výběrové charakteristiky

V předchozí kapitole jsme se zmínili o tom, že k modelování a zkoumání populace používáme výběrové soubory. Je-li výběr reprezentativní, dá se na jeho základě získat dobrá představa o vlastnostech populace.

Náhodnou veličinu X, jejíž hodnoty při realizaci náhodného pokusu pozorujeme, můžeme popsat pomocí různých číselných charakteristik. Ve statistice v souvislosti s náhodnou veličinou hovoříme častěji o **parametrech základního souboru (populace),** popř. o **parametrech rozdělení** náhodné veličiny. K parametrům základního souboru patří: střední hodnota μ, rozptyl σ^2, směrodatná odchylka σ, pravděpodobnost π, atd... Parametry populace jsou **konstantní hodnoty** (pro určitou náhodnou veličinu, v pevném čase). Neznáme-li však rozdělení pozorované náhodné veličiny, nedokážeme parametry populace většinou přesně určit.

Ve výběrovém souboru lze najít příslušné protějšky parametru populace. Říká se jim **výběrové charakteristiky** (resp. **statistiky**) a jsou definovány jako vhodné funkce náhodného výběru. Výběrové charakteristiky budeme obecně značit $T(\boldsymbol{X}) = T(X_1, ..., X_n)$. Možných výběrů ze základního souboru může být mnoho a výběrové charakteristiky budou proto nutně vykazovat proměnlivost (variabilitu). Hodnotu výběrové charakteristiky na konkrétním výběru nazýváme **empirická charakteristika** nebo **pozorovaná hodnota** výběrové charakteristiky $T(\boldsymbol{X})$. Z pravděpodobnostního hlediska mají výběrové charakteristiky charakter náhodných veličin a lze je tedy popsat nějakým rozdělením, mají také svou střední hodnotu, rozptyl a všechny ostatní charakteristiky.

Základní princip statistické indukce, který je schematicky znázorněn na obrázku 8.1, je pak založen na tom, že chceme-li získat informace o určitém parametru populace θ, pak analyzujeme takovou výběrovou charakteristiku T, která s velkou pravděpodobností nabývá hodnot blízkých neznámému parametru θ.



Obr. 3.1: Princip statistické indukce

Přehled nejpoužívanějších parametrů populace a příslušných výběrových charakteristik, včetně jejich značení je uveden v tabulce 8.1.
Jak již bylo řečeno, výběrové charakteristiky jsou náhodné veličiny, jejichž jednotlivé realizace lze získat výpočtem pozorovaných hodnot těchto charakteristik pro jednotlivé výběry o rozsahu n. (Např. Průměrný plat 20 občanů ČR je náhodná veličina. Výpočtem prů-

Obsah
105. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Tab. 3.1: Přehled základních parametrů populace a příslušných výběrových charakteristik

Základní soubor (populace)	střední hodnota $E(X)$, resp. μ	medián $x_{0,5}$	rozptyl $D(X)$, resp. σ^2	směrodatná odchylka σ	pravděpodobnost π
Výběrový soubor (výběr)	(výběrový) průměr $\bar{X}$	výběrový medián $\tilde{X}_{0,5}$	výběrový rozptyl S^2	výběrová směrodatná odchylka S	relativní četnost p

měrného platu konkrétních 20 občanů získáme jednu realizaci tohoto průměru, výpočtem průměrného platu jiného vzorku 20 občanů ČR získáme jinou realizaci průměru.) Pojmem **výběrová rozdělení** označujeme rozdělení pravděpodobností výběrových charakteristik.

3.2. Variabilita výběrových charakteristik

Vhodnou mírou variability výběrových charakteristik bývá často jejich rozptyl nebo jejich směrodatná odchylka. Variabilitu výběrových charakteristik přitom ovlivňují tři faktory:

- rozsah populace (N),
- rozsah výběru (n),
- způsob získání náhodného výběru.

Je-li rozsah populace mnohem větší než rozsah výběru ($N \gg n$), pak variabilita výběrových charakteristik je obvykle zhruba stejná jak pro výběry s opakováním, tak pro výběry bez opakování. Je-li však výběr významnou části populace (řekněme, $n \geqq 0,05N$), pak je variabilita výběrových charakteristik výrazně nižší, použijeme-li výběr bez opakování.

Následující výběrová rozdělení jsou odvozena pro případ, že rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru. Tuto podmínku budeme považovat za splněnou, pokud rozsah výběru nepřekročí 5% rozsahu populace, tj. pokud

$$\frac{n}{N} < 0,05.$$

3.3. Výběrový průměr (průměr, angl. „sample mean“)

Jednou z nejdůležitějších charakteristik náhodného výběru je výběrový průměr.

Mějme náhodný výběr $X_1, ..., X_n$ z náhodné veličiny X o rozdělení $F(x)$ (tzn. každá z veličin X_i má distribuční funkci $F(x)$ a všechny dvojice náhodných veličin X_i, X_j jsou nezávislé). Označme μ_X střední hodnotu a σ_X směrodatnou odchylku náhodné veličiny X_i. (Všechny náhodné veličiny X_i mají stejnou střední hodnotu i směrodatnou odchylku.)

Výběrovým průměrem náhodného výběru $X_1, ..., X_n$ rozumíme náhodnou veličinu

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Vlastnosti výběrového průměru

1. $E\left(\bar{X}\right) = E(X_i) = E(X) = \mu_X$

   Důkaz: $E\left(\bar{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot nE(X_i) = E(X_i) = \mu_X$

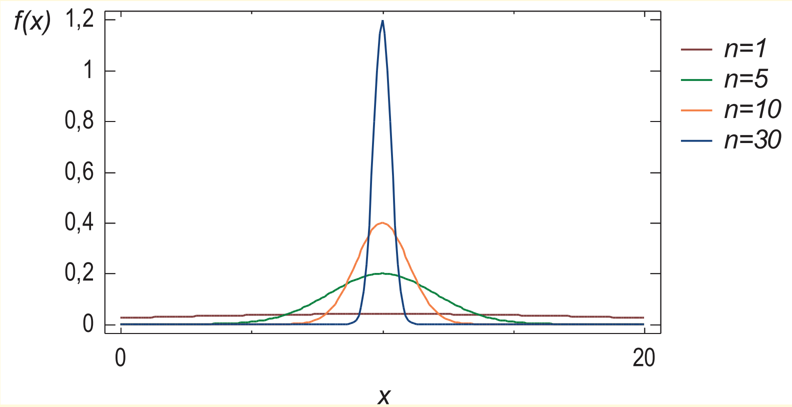
2. $D\left(\bar{X}\right) = \frac{1}{n} D(X_i) = \frac{\sigma_X^2}{n}$

   Důkaz: $D\left(\bar{X}\right) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot nD(X_i) = \frac{D(X_i)}{n} =$

   $$= \frac{\sigma_X^2}{n}$$

Poznámka: *Všimněte si (Obr. 8.2), že s rostoucím rozsahem výběru se snižuje variabilita výběrového průměru, tzn. pozorované hodnoty průměru se stále více koncentrují kolem střední hodnoty .*

3. Pochází-li náhodný výběr $X_1, ..., X_n$ z normálního rozdělení $N(\mu_X, \sigma_X^2)$, pak výběrový průměr má normální rozdělení s parametry $\mu_X, \frac{\sigma_X^2}{n}$,tj. $N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.



Obr. 3.2: Vliv rozsahu výběru na graf hustoty pravděpodobnosti výběrového průměru

3.4. Limitní věty

Nyní známe rozdělení výběrového průměru pro případ, že výběr pochází z normálního rozdělení. Další tvrzení o vlastnostech výběrového průměru, tentokrát pro případ dostatečně velkého rozsahu náhodného výběru, přináší limitní věty. Uvedeme si dvě nejdůležitější – zákon velkých čísel a centrální limitní větu.

3.4.1. Zákon velkých čísel

Ukázali jsme si, že pochází-li výběr z normálního rozdělení, pak s rostoucím rozsahem výběru se výběrový průměr stále silněji soustřeďuje kolem střední hodnoty. Obsahem zákona velkých čísel je zachování této vlastnosti i pro případ výběru z jiného než normálního rozdělení.

Vypočteme-li výběrový průměr z náhodného výběru o rozsahu rovném rozsahu populace, získáme střední hodnotu rozdělení, z něhož výběr pochází. Vypočteme-li výběrový průměr z náhodného výběru o rozsahu menším než je rozsah populace, nezískáme přesně střední hodnotu rozdělení, ale dostaneme číslo, které je skutečné střední hodnotě blízko.

Zákon velkých čísel má několik formulací. Uveďme přesnější formulaci tzv. **slabého zákona velkých čísel**:

Mějme nekonečný náhodný výběr $X_1, X_2, \ldots$ z rozdělení se střední hodnotou μ_X a konečným rozptylem σ_x^2, kde $X_1, X_2, \ldots$ jsou nekorelované náhodné veličiny. Potom platí, že výběrový průměr $\bar{X}_n$ vypočítaný z prvních n pozorování se pro $n \to \infty$ blíží ke střední hodnotě μ_X, což zapisujeme

$$\lim_{n\to\infty} \left[P\left(\left| \bar{X}_n - \mu_X \right| > \varepsilon \right) \right] = 0 \text{ pro každé } \varepsilon > 0.$$

3.4.2. Centrální limitní věta

Vlastnosti výběrového průměru říkají, že průměr $\bar{X}$ má střední hodnotu μ_X a rozptyl $\frac{\sigma_X^2}{n}$. Pocházejí-li X_i z normálního rozdělení, pak výběrový průměr rovněž podléhá normálnímu rozdělení. Centrální limitní věta, zkráceně CLV, tyto poznatky rozšiřuje o tvrzení, že

jsou-li X_i nezávislé náhodné veličiny s konečným rozptylem, pak výběrový průměr má při dostatečně velkém počtu pozorování přibližně normální rozdělení, ať už X_i pocházejí z libovolného rozdělení.

Centrální limitní větu zapisujeme

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \text{ nebo } \frac{\bar{X} - \mu_X}{\sigma_X}\sqrt{n} \sim N(0,1).$$

($X \sim N(\mu, \sigma^2)$ znamená, že X má přibližně normální rozdělení s parametry μ, σ^2.)

Ve statistické praxi vyvstává v souvislosti s použitím CLV otázka, kdy můžeme rozsah výběru považovat za „dostatečně velký“. Za dostatečně velké se běžně označují výběry o rozsahu 30 a větším. Zároveň se však ukazuje, že CLV platí, pokud je splněna libovolná z následujících podmínek.

- X_i pochází z normálního rozdělení.
- Výběrové rozdělení je symetrické, unimodální, výběr neobsahuje odlehlá pozorování a rozsah výběru je nejvýše 15.
- Výběrové rozdělení je symetrické nebo mírně zešikmené, unimodální, výběr neobsahuje odlehlá pozorování a rozsah výběru je 16 až 30.
- Výběr neobsahuje odlehlá pozorování a rozsah výběru je alespoň 30.

Důsledek CLV
Součet dostatečně velkého počtu nezávislých pozorování s konečným rozptylem má přibližně normální rozdělení s parametry $n\mu_X$ a $n\sigma_X^2$, což zapisujeme

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu_X, n\sigma_X^2\right).$$

Odvození: $\bar{X} = \dfrac{\sum\limits_{i=1}^{n} X_i}{n} \Rightarrow \sum\limits_{i=1}^{n} X_i = n\bar{X}$

$$E\left(\sum_{i=1}^{n} X_i\right) = nE\left(\bar{X}\right) = n\mu_X, \; D\left(\sum_{i=1}^{n} X_i\right) = nD\left(\bar{X}\right) = n^2\frac{\sigma_X^2}{n} = n\sigma_X^2.$$

$$\Rightarrow \sum_{i=1}^{n} X_i \sim N\left(n\mu_X, n\sigma_X^2\right).$$

Vliv tvaru rozdělení základního souboru a rozsahu výběru n na rozdělení průměru při k opakováních náhodného výběru můžete sledovat v appletu Rozdělení průměru (480 KB).

Příklad 3.1. Životnost elektrického holicího strojku EHS má exponenciální rozdělení se střední hodnotou 2 roky. Určete pravděpodobnost, že průměrná životnost 150 prodaných holicích strojků EHS bude vyšší než 27 měsíců.

Řešení 3.1

Příklad 3.2. Dlouhodobým průzkumem bylo zjištěno, že doba potřebná k objevení a odstranění poruchy stroje má střední hodnotu 40 minut a směrodatnou odchylku 30 minut. Jaká je pravděpodobnost, že doba potřebná k objevení a opravení 100 nezávislých poruch nepřekročí 70 hodin?

Řešení 3.2

Příklad 3.3. Výletní člun má nosnost 5000 kg. Hmotnost cestujících je náhodná veličina se střední hodnotou 70 kg a směrodatnou odchylkou 20 kg. Kolik cestujících může člunem cestovat, aby pravděpodobnost přetížení člunu byla menší než 0,001?

Řešení 3.3

Obsah
112. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

3.5. Relativní četnost

Uvažujme nějaký náhodný jev A vyskytující se s pravděpodobností π a předpokládejme, že provádíme opakovaná nezávislá pozorování tohoto jevu. Označme $X_i = 1$, pokud jev A při $i-$tém pozorování nastal a $X_i = 0$, pokud nenastal. Pak $X_1, X_2, ...$ je náhodný výběr z alternativního rozdělení $A(\pi)$, kde $E(X_i) = \pi, D(X_i) = \pi(1-\pi)$.

Výběrový průměr $\bar{X}$ vypočítaný z prvních n pozorování označujeme v tomto případě jako **relativní četnost** a značíme ji p.

$$\bar{X} = \frac{\sum\limits_{i=1}^{n}}{X_i} n = p$$

Vlastnosti relativní četnosti

1. $E(p) = \mu_p = \pi$

   Důkaz: $E(p) = \mu_p = E\left(\frac{\sum\limits_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E\left(\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot nE(X_i) = E(X_i) = \pi$

2. $D(p) = \sigma_p^2 = \frac{\pi(1-\pi)}{n}$

   Důkaz: $D(p) = \sigma_p^2 = D\left(\frac{\sum\limits_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} D\left(\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot nD(X_i) = \frac{D(X_i)}{n} =$

Obsah
113. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

$$= \frac{\pi(1-\pi)}{n}.$$

3. Podle zákona velkých čísel pak platí, že relativní četnost se pro $n \to \infty$ blíží střední hodnotě π, tj. pravděpodobnosti výskytu jevu A.

$$\lim_{n\to\infty} \left[P\left(|p-\pi| > \varepsilon\right)\right] = 0 \text{ pro každé } \varepsilon > 0.$$

Toto odpovídá intuitivnímu chápání pravděpodobnosti jako čísla, které udává relativní četnost výskytu sledovaného jevu.

Poznámka: *O zákonu velkých čísel vědí své všichni hráči a hlavně všichni majitelé kasin. J. S. Rosenthal ve své knize „Zasažen bleskem“ píše: „Je-li hra v průměru třeba jen sebenepatrněji vychýlená ve váš neprospěch a vy budete hrát dostatečně dlouho, můžete si být jisti, že prohrajete. I když každá jednotlivá partie hry probíhá nezávisle, bez ohledu na to, co se stalo předtím, tak přece jen jediné, na čem při dlouhém opakování záleží, je průměrné množství výher a proher... Zkrátka a dobře, k tomu, aby slušně vydělalo, nepotřebuje kasino štěstí, ale jen trpělivost. Zatímco hráči mohou své hráčské naděje zakládat na klamné představě, že mají „šťastnou ruku“ či „šťastné číslo“, nebo na postavení planet, kasino si může dovolit založit své naděje na něčem mnohem spolehlivějším: na zákonu velkých čísel.“*

Jelikož relativní četnost p je výběrovým průměrem náhodných veličin s alternativním rozdělením $A(\pi)$, můžeme poznatky o ní rozšířit aplikací CLV.

4. Relativní četnost p má při dostatečně velkém počtu pozorování přibližně normální rozdělení, ať už X_i pocházejí z libovolného rozdělení. Výběry jsou obvykle považovány za dostatečně velké v případě, že

$$n > \frac{9}{p(1-p)}.$$

$$p \sim N\left(\mu_p, \sigma_p^2\right), \text{ tj. } p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \Rightarrow \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

3.6. Rozdíl výběrových průměrů

Mějme náhodný výběr $X_{11}, \ldots, X_{1n_1}$ z rozdělení se střední hodnotou μ_1 a náhodný výběr $X_{21}, \ldots, X_{2n_2}$ z rozdělení se střední hodnotou μ_2. Dále nechť jsou splněny následující předpoklady.

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru $\left(\frac{n_i}{N_i} < 0,05\right)$.
- Výběry jsou nezávislé, tj. hodnoty pozorování z populace 1 nejsou ovlivněny hodnotami pozorování z populace 2, a naopak.
- Platí předpoklady CLV, zejména to, že každý z výběrů pochází z normálního rozdělení nebo je dostatečně velký (za dostatečně velké obvykle považujeme výběry s rozsahem větším než 30).

Jsou-li splněny výše uvedené předpoklady, pak má rozdíl výběrových průměrů následující vlastnosti.

1. $E\left(\bar{X}_1 - \bar{X}_2\right) = \mu_1 - \mu_2$

2. $D\left(\bar{X}_1 - \bar{X}_2\right) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

3. $\left(\bar{X}_1 - \bar{X}_2\right) \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ tj. $\dfrac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Důkaz:

Z vlastností výběrových průměrů je zřejmé, že $\bar{X}_1 \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right), \bar{X}_2 \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right)$.

$$E\left(\bar{X}_1 - \bar{X}_2\right) = E\left(\bar{X}_1\right) - E\left(\bar{X}_2\right) = \mu_1 - \mu_2,$$

$$\begin{aligned} D\left(\bar{X}_1 - \bar{X}_2\right) &= D\left(\bar{X}_1 + (-1)\bar{X}_2\right) = D\left(\bar{X}_1\right) + (-1)^2 D\left(\bar{X}_2\right) = D\left(\bar{X}_1\right) + D\left(\bar{X}_2\right) = \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

Vzhledem ke splnění předpokladů CLV, lze tvrdit, že

$$\left(\bar{X}_1 - \bar{X}_2\right) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Standardizací rozdílu náhodných veličin $\bar{X}_1$ a $\bar{X}_2$ dostaneme, že

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

3.7. Rozdíl relativních četností

Uvažujme nějaký náhodný jev A a předpokládejme, že provádíme opakovaná nezávislá pozorování tohoto jevu. Označme $X_{1i} = 1$, pokud jev A při $i-$tém pozorování nastal a $X_{1i} = 0$, pokud nenastal. Pak je náhodný výběr $X_{11}, \ldots$ z alternativního rozdělení $A(\pi_1)$, kde $E(X_{1i}) = \pi_1, D(X_{1i}) = \pi_1(1-\pi_1)$.

Dále uvažujme nějaký náhodný jev B a předpokládejme, že provádíme opakovaná nezávislá pozorování tohoto jevu. Označme $X_{2j} = 1$, pokud jev B při $j-$tém pozorování nastal a $X_{2j} = 0$, pokud nenastal. Pak je náhodný výběr $X_{21}, \ldots$ z alternativního rozdělení $A(\pi_2)$, kde $E(X_{2j}) = \pi_2, D(X_{2j}) = \pi_2(1-\pi_2)$.

Výběrový průměr $\bar{X}_1$ vypočítaný z prvních n_1 pozorování náhodného výběru 1 udává relativní četnost jevu A a značíme ji p_1. Obdobně výběrový průměr $\bar{X}_2$ vypočítaný z prvních n_2 pozorování náhodného výběru 2 udává relativní četnost jevu B a značíme ji p_2.

Dále nechť jsou splněny následující předpoklady.

- Rozsah každé z populací je dostatečně velký vzhledem k rozsahu příslušného výběru. (V tomto případě považujeme za dostatečně velkou populaci, jejíž rozsah je alespoň 10 násobkem rozsahu příslušného výběru.)
- Výběry z obou populací jsou dostatečně velké na to, aby pro modelování rozdílu mezi relativními četnostmi mohlo být použito normální rozdělení. Výběry jsou obvykle považovány za dostatečně velké v případě, že $\left(n_1 > \frac{9}{p_1(1-p_1)}\right) \wedge \left(n_2 > \frac{9}{p_2(1-p_2)}\right)$.
- Výběry jsou nezávislé, tzn. hodnoty pozorování z populace 1 nejsou ovlivněny hodnotami pozorování z populace 2, a naopak.

Jsou-li splněny výše uvedené předpoklady, pak má rozdíl relativních četností následující vlastnosti.

1. $E(p_1 - p_2) = \pi_1 - \pi_2$
2. $D(p_1 - p_2) = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}$
3. $(p_1 - p_2) \sim N\left(\pi_1 - \pi_2, \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}\right)$,

   tj. $\dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$

Důkaz:

Z vlastností relativních četností je zřejmé, že $p_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right)$, $p_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$.

$E(p_1 - p_2) = E(p_1) - E(p_2) = \pi_1 - \pi_2,$

$D(p_1 - p_2) = D(p_1) + D(p_2) = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}.$

Vzhledem ke splnění předpokladů CLV, lze tvrdit, že

$$(p_1 - p_2) \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Standardizaci rozdílu náhodných veličin p_1 a p_2 lze ukázat, že

$$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1).$$

Výše zmíněná výběrová rozdělení nacházejí uplatnění při odhadech střední hodnoty a pravděpodobnosti, resp. jejich rozdílů nebo při testování hypotéz o těchto parametrech. Při odhadech rozptylu, poměru rozptylů, odhadech střední hodnoty v případě, že máme k dispozici pouze malý výběr, který nepochází z normálního rozdělení, a v dalších metodách statistické indukce nacházejí uplatnění tři důležitá spojitá rozdělení (χ^2- rozdělení, Studentovo rozdělení, Fisherovo – Snedecorovo rozdělení), kterým bude věnován následující výklad. Jediným parametrem těchto rozdělení jsou tzv. ***stupně volnosti*** *(angl. „degrees of freedom“), v případě Fisherovo – Snedecorova rozdělení – dvojice stupňů volnosti.*

3.8. χ^2 - rozdělení (Pearsonovo rozdělení)

Mějme nezávislé náhodné veličiny $Z_1, Z_2, ..., Z_\nu$, z nichž každá má normované normální rozdělení. Součet čtverců těchto náhodných veličin, tj. náhodná veličina X má rozdělení χ^2 („čteme „chí-kvadrát“) s ν stupni volnosti, což značíme χ^2_ν.

$$\forall i = 1, ..., n : Z_i \to N(0, 1), \text{ pak } X = \sum_{i=1}^{\nu} Z_i^2 \to \chi^2_\nu$$

Počet stupňů volnosti označuje počet sčítaných nezávislých náhodných veličin a je jediným parametrem tohoto rozdělení. Z definice χ^2- rozdělení je zřejmé, že náhodná veličina s tímto rozdělením může nabývat pouze nezáporných hodnot.

Poznámka: *Někteří statistikové nazývají toto rozdělení Pearsonovým rozdělením.*

3.8.1. Vlastnosti rozdělení χ^2

1. Pro nezávislé náhodné veličiny s χ^2 - rozdělením se dá snadno ukázat, že jejich součet má opět χ^2 - rozdělení a počet stupňů volnosti je roven součtu stupňů volnosti ν_i jednotlivých veličin v součtu.
$$\text{Nechť } X_i \to \chi^2_{\nu_i} X \to \sum_{(i)} \chi^2_{\nu_i}, \text{ pak } X \to \chi^2_{\sum\limits_{(i)} \nu_i}.$$
2. Předpokládejme, že provedeme náhodný pokus spočívající v náhodném výběru o rozsahu n z populace **podléhající normálnímu rozdělení** s rozptylem σ^2. Pro uvedený výběr

Obsah
120. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

určíme výběrovou směrodatnou odchylku s. Lze ukázat, že náhodná veličina

$$\frac{(n-1)S^2}{\sigma^2}$$

má χ^2- rozdělení s $n-1$ stupni volnosti. Plyne to bezprostředně z toho, že tento výraz se dá převést na součet čtverců $(n-1)$ náhodných veličin s rozdělením $N(0,1)$.

Tuto skutečnost můžeme stručně zapsat takto:

$$\frac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}.$$

Nástin důkazu: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \Rightarrow \frac{S^2}{\sigma^2}\cdot(n-1) = \sum_{i=1}^{n}\frac{\left(X_i-\bar{X}\right)^2}{\sigma^2} =$$
$$= \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2.$$

Pomocí dalších úprav (zdlouhavé), které vedou na nahrazení průměru střední hodnotou, bychom zjistili, že

$$\frac{S^2}{\sigma^2}\cdot(n-1) = \sum_{i=1}^{n-1}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n-1} Z_i^2.$$

Nahrazení průměru střední hodnotou způsobí ztrátu jednoho stupně volnosti.

Zelená křivka na obrázku 8.3 ukazuje rozdělení náhodné veličiny $\frac{(n-1)S^2}{\sigma^2}$ vypočtené ze všech výběrů o rozsahu $3(\nu = n-1 = 3-1 = 2)$. Obdobně hnědá, resp. oranžová, křivka představují hustotu pravděpodobnosti této náhodné veličiny vypočtené ze všech výběrů o rozsahu 5, resp. 9.

Hustotu pravděpodobnosti v obecném tvaru (pro n stupňů volnosti) nebudeme pro značnou komplikovanost vztahu uvádět.

3. **Střední hodnota** náhodné veličiny X s rozdělením χ^2_ν je rovna počtu stupňů volnosti, tj. $E(X) = \nu$.

4. **Rozptyl** náhodné veličiny X s rozdělením χ^2_ν je roven dvojnásobku počtu stupňů volnosti, tj. $D(X) = 2\nu$.

5. Je-li počet stupňů volnosti rozdělení χ^2_ν větší nebo roven 2, pak modus náhodné veličiny mající toto rozdělení je $\nu - 2$.

6. **Kvantily** náhodné veličiny s rozdělením χ^2_ν jsou pro různé hodnoty ν a p tabelovány (viz příloha – Tabulka 3). Běžně lze také kvantily tohoto rozdělení určit pomocí statistického software.

7. Se vzrůstajícím počtem stupňů volnosti se χ^2_ν- rozdělení blíží normálnímu rozdělení $N(\nu, 2\nu)$.

Vliv počtu stupňů volnosti Pearsonova rozdělení na tvar křivek hustoty pravděpodobnosti a distribuční funkce můžete sledovat v excelovském souboru Spojitá rozdělení (870 KB).

3.8.2. Použití rozdělení χ^2

1. Vlastnosti, že
$$\frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}$$
se využívá k *testování toho, zda rozptyl základního souboru s normálním rozdělením je roven* σ_0^2 (viz kapitola 11).

2. χ^2- rozdělení se používá pro ověření nezávislosti kategoriálních proměnných (*test nezávislosti v kontingenční tabulce*), kterým se budeme zabývat v kapitole 14.

3. Pokud testujeme, zda náhodné veličiny (naměřená data) pocházejí z určitého rozdělení, můžeme také s úspěchem použít χ^2- rozdělení. Tento test je znám pod názvem „test dobré shody“ (viz kapitola 14).

Příklad 3.4. Firma Edison vyrábí žárovky Ed. Životnost těchto žárovek je průměrně 5 let se směrodatnou odchylkou 6 měsíců. Pro ověřování kvality výroby bude testováno 20 žárovek. Jaká je pravděpodobnost, že při tomto testu bude zjištěna směrodatná odchylka životnosti vyšší než 7 měsíců?

Řešení 3.4

Příklad 3.5. Odvoďte distribuční funkci a hustotu pravděpodobnosti náhodné veličiny X, která má χ^2- rozdělení s jedním stupněm volnosti.

Řešení 3.5

Obsah
123. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

3.9. Studentovo rozdělení (t rozdělení)

Dříve než přejdeme k popisu tohoto rozdělení, uveďme krátkou poznámku o jeho vzniku. Autorem Studentova rozdělení je irský chemik William Sealy Gosset (1876-1937), zaměstnanec pivovaru Guiness. Jedním Gossetových úkolů bylo posoudit kvalitu různých druhů vařených piv, přičemž k dispozici měl jen malý počet vzorků, často méně než 10. Gosset věděl, že použije-li pro odhad střední hodnoty při tak malých výběrových souborech běžně používané normální rozdělení, nalezený odhad skutečnou střední hodnotu podhodnotí. Proto se tímto problémem zabýval podrobněji a v roce 1908 publikoval postup, který měl poskytnout možnost získat i z malých vzorků použitelné závěry. (Jméno Gosset je už dnes téměř neznáme, neboť Gosset se pod svá průkopnická díla podepisoval pseudonymem Student, protože mu jeho firma z obavy, aby konkurence neodhalila tajemství jejich piva, nedovolila publikovat vědecké práce pod vlastním jménem.) Na práci Gosseta později navázalo množství dalších statistiků. Jmenujme alespoň R. A. Fishera, který se podílel téměř na všech směrech dalšího vývoje statistiky.
Po této krátké odbočce přejděme k popisu Studentova rozdělení.

Uvažujme dvě nezávislé náhodné veličiny: Z a V. Náhodná veličina Z má normované normální rozdělení, náhodná veličina V má χ^2- rozdělení s ν stupni volnosti. Potom náhodná veličina T,

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}},$$

má Studentovo t rozdělení s ν stupni volnosti, což značíme $T \to t_\nu$. Počet stupňů volnosti je jediný parametr tohoto rozdělení.

Pro $\nu \to \infty$ (vysoký počet stupňů volnosti, v praxi pro $\nu > 30$) se Studentovo t rozdělení blíží normovanému normálnímu rozdělení.

Hustotu pravděpodobnosti nebudeme ani v tomto případě pro složitost vztahu uvádět.

Vliv počtu stupňů volnosti Studentova rozdělení na tvar křivky hustoty pravděpodobnosti můžete sledovat v následující animaci nebo v excelovském souboru Spojitá rozdělení (870 KB).

Hustota studentova rozdělení

Instrukce pro spuštění vložených animací naleznete na
http://mi21.vsb.cz/adobe-reader-multimedia.

Střední hodnota: $E(T) = 0$ pro $\nu > 1$

Rozptyl: $D(T) = \dfrac{\nu}{\nu - 2}$ pro $\nu > 2$

100p% kvantily t_p:

Pro vybraná p a pro vybrané stupně volnosti ν jsou $100p\%$ kvantily tabelovány (například viz příloha – Tabulka 2). Většinou je tato tabelace provedena pouze pro $p < 0,5$. Kvantily t_p pro $p > 0,5$ získáme pomocí vztahu

$$t_p = -t_{1-p}.$$

Běžně se pro určování kvantilů využívá statistický software.

3.9.1. Vlastnosti Studentova t rozdělení

1. Pokud náhodné veličiny $X_1, X_2, ..., X_n$ mají normální rozdělení $N(\mu, \sigma^2)$ a jsou navzájem nezávislé, pak náhodná veličina definována jako

$$\frac{\bar{X} - \mu}{S}\sqrt{n}$$

má Studentovo t rozdělení s $(n-1)$ stupni volnosti, což značíme

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \to t_{n-1}.$$

Důkaz této vlastnosti je pro zájemce uveden v kapitole 8.11.

2. Mějme dva výběry z normálního rozdělení se stejným rozptylem.

   $\forall i = 1, 2, ..., n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu, \sigma^2)$,
   $\forall j = 1, 2, ..., n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \to N(\mu, \sigma^2)$.

   Nechť průměry $\bar{X}_1, \bar{X}_2$a výběrové rozptyly S_1^2, S_2^2 jsou náhodné veličiny definované jako

$$\bar{X}_1 = \frac{\sum\limits_{i=1}^{n_1} X_{1i}}{n_1}, \quad \bar{X}_2 = \frac{\sum\limits_{j=1}^{n_2} X_{2j}}{n_2}, \quad S_1^2 = \frac{\sum\limits_{i=1}^{n_1} \left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}, \quad S_2^2 = \frac{\sum\limits_{j=1}^{n_2} \left(X_{2j} - \bar{X}_2\right)^2}{n_2 - 1}.$$

   Pak

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \to t_{n_1 + n_2 - 2}.$$

3. Mějme dva výběry z normálního rozdělení s různými rozptyly.

   $\forall i = 1, 2, ..., n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu, \sigma_1^2)$,
   $\forall j = 1, 2, ..., n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \to N(\mu, \sigma_2^2)$.

   Nechť průměry $\bar{X}_1, \bar{X}_2$a výběrové rozptyly S_1^2, S_2^2 jsou náhodné veličiny definované jako

$$\bar{X}_1 = \frac{\sum\limits_{i=1}^{n_1} X_{1i}}{n_1}, \quad \bar{X}_2 = \frac{\sum\limits_{j=1}^{n_2} X_{2j}}{n_2}, \quad S_1^2 = \frac{\sum\limits_{i=1}^{n_1} \left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}, \quad S_2^2 = \frac{\sum\limits_{j=1}^{n_2} \left(X_{2j} - \bar{X}_2\right)^2}{n_2 - 1}.$$

Pak

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \to t_\nu,$$

kde

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1+1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2.$$

Důkaz vlastností 2 a 3 nebudeme provádět.

3.9.2. Použití Studentova t rozdělení

Studentovo t rozdělení má uplatnění zejména při modelování založeném na analýze malých výběrů. Uvedeme alespoň některé možnosti použití.

1. Užívá se k *testování hypotéz o střední hodnotě*, pokud je rozptyl základního souboru neznámý a výběr pochází z normálního rozdělení.
2. Užívá se k *testování hypotéz o shodě středních hodnot*, za předpokladu, že máme dispozici dva nezávislé výběry z normálních rozdělení, jejichž rozptyly jsou neznámé, ale shodné.
3. Rozdělení je vhodným prostředkem pro *analýzu výsledků regresní analýzy*.

3.10. Fisherovo-Snedecorovo rozdělení (F rozdělení)

Posledním spojitým rozdělením, kterým se budeme zabývat, je Fisherovo-Snedecorovo, čti Fišerovo-Snedekorovo, F rozdělení. Mějme dvě nezávislé náhodné veličiny V a W s rozdělením χ^2. První z nich má počet stupňů volnosti m, druhá má počet stupňů volnosti n (obecně mají různý počet stupňů volnosti). Pak má náhodná veličina

$$F = \frac{\frac{V}{m}}{\frac{W}{n}}$$

Fisherovo-Snedecorovo rozdělení o m a n stupních volnosti, což značíme $F \to F_{m,n}$. Fisherovo-Snedecorovo rozdělení má tedy dva parametry - počet stupňů volnosti v čitateli m a počet stupňů volnosti ve jmenovateli n.

Ani v tomto případě nebudeme uvádět vztah pro hustotu pravděpodobnosti (je značně složitý).

Vliv počtu stupňů volnosti Fisherova-Snedecorova rozdělení na tvar křivky hustoty pravděpodobnosti můžete sledovat v excelovském souboru Spojitá rozdělení (870 KB).

Střední hodnota: $E(F) = \dfrac{n}{n-2}$ pro $n > 2$

Rozptyl: $D(F) = \dfrac{2n^2\left(1+\frac{n-2}{m}\right)}{(n-2)^2(n-4)}$ pro $n > 4$

100p% kvantily - f_p:

Obsah
130. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Pro praktické aplikace jsou pro vybrané pravděpodobnosti ($p > 0,5$) a vybrané stupně volnosti m a n tabelovány kvantily f_p (viz příloha – Tabulka 4). Pro $p > 0,5$ se kvantily f_p určí ze vztahu

$$f_p = \frac{1}{f^*_{1-\nu}},$$

kde f_p je $100p\%$ kvantil Fisherova-Snedecorova rozdělení s m stupni volnosti pro čitatele a n stupni volnosti pro jmenovatele a f^*_{1-p} je je $100p\%$ kvantil Fisherova-Snedecorova rozdělení s n stupni volnosti pro čitatele a m stupni volnosti pro jmenovatele.

3.10.1. Vlastnosti Fisherova-Snedecorova rozdělení

Mějme dva výběry z normálního rozdělení.

$\forall i = 1, 2, ..., n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu, \sigma_1^2)$,
$\forall j = 1, 2, ..., n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \to N(\mu, \sigma_2^2)$.

Nechť výběrové rozptyly S_1^2 a S_2^2 jsou náhodné veličiny definované jako

$$S_1^2 = \frac{\sum\limits_{i=1}^{n_1} \left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} \quad \text{a} \quad S_2^2 = \frac{\sum\limits_{j=1}^{n_2} \left(X_{2j} - \bar{X}_2\right)^2}{n_2 - 1}.$$

Pak

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \to F_{n_1-1, n_2-2}.$$

Důkaz uvedené vlastnosti Fisherova-Snedecorova rozdělení je opět určen především čtenářům, kteří chtějí znát matematické pozadí uváděných vztahů a je uveden v části 8.11.

3.10.2. Použití Fischerova-Snedecorova rozdělení

Toto rozdělení má opět široké uplatnění, zejména při hodnocení výsledků statistických analýz. Používá se především

1. k testu o shodě rozptylů dvou základních souborů,
2. k testům o shodě středních hodnot více než dvou základních souborů, v tzv. analýze rozptylu,
3. k testům v regresní analýze.

Příklad 3.6. Vraťme se k řešenému příkladu 8.4. Firma Edison vyrábí žárovky Ed. Životnost těchto žárovek je průměrně 5 let se směrodatnou odchylkou 6 měsíců. Uvedené informace specifikujeme: Žárovky jsou vyráběny na dvou linkách. Předpokládejme, že obě linky mají srovnatelné parametry, tj. že průměrná životnost a variabilita životnosti žárovek Ed vyrobených ve firmě Edison nezávisí na tom, na jaké lince byly vyrobeny. Pro ověření kvality výroby bude testována životnost 20 žárovek z linky 1 a 30 žárovek z linky 2. Jaká je pravděpodobnost, že u vzorku z linky 1 bude zjištěn více než dvojnásobný rozptyl oproti rozptylu zjištěnému u vzorku z linky 2?

Řešení 3.6

3.11. Odvození vybraných vlastností Studentova a Fisherovo--Snedecorova rozdělení

Odstavec 8.11 je určen zájemcům o matematické odvození vztahů prezentovaných v této kapitole.

3.11.1. $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \to t_{n-1}$

Pokud náhodné veličiny $X_1, X_2, ..., X_n$ mají normální rozdělení $N(\mu, \sigma^2)$ a jsou navzájem nezávislé, pak lze snadno ukázat (viz kap. 3.4.2 Centrální limitní věta), že platí

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}.\right)$$

Vzhledem ke standardizaci (transformaci normální na normovanou normální náhodnou veličinu) platí

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \to N(0, 1).$$

Dále víme, že je-li

$$V = \frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1},$$

pak

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} \to t_{n-1}.$$

Po dosazení

$$\frac{\dfrac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}}}{\sqrt{\dfrac{\dfrac{(n-1)S^2}{\sigma^2}}{n-1}}} \to t_{n-1}$$

a po úpravě dostaneme

$$\frac{\dfrac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}}}{\sqrt{\dfrac{\dfrac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} \cdot \frac{\sigma}{S} = \frac{\bar{X}-\mu}{S} \cdot \sqrt{n} \to t_{n-1}.$$

3.11.2. $\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \to F_{n_1-1,n_2-1}$

Náhodná veličina

$$F = \frac{\dfrac{V}{m}}{\dfrac{W}{n}}$$

má Fisherovo-Snedecorovo rozdělení o m a n stupních volnosti, jsou-li V a W dvě nezávislé náhodné veličiny, přičemž

$$V \to \chi^2_m \text{ a } W \to \chi^2_n.$$

Z vlastností χ^2- rozdělení víme, že

$$\frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}.$$

Nechť

$$V = \frac{(n_1-1)S_1^2}{\sigma_1^2} \text{ a } W = \frac{(n_2-1)S_2^2}{\sigma_2^2}.$$

Je zřejmé, že $V \to \chi^2_{n_1-1}$ a $W \to \chi^2_{n_2-1}$.

Pak

$$F = \frac{\dfrac{V}{n_1-1}}{\dfrac{W}{n_2-1}} = \frac{\dfrac{\frac{(n_1-1)S_1^2}{\sigma_1^2}}{n_1-1}}{\dfrac{\frac{(n_2-1)S_2^2}{\sigma_2^2}}{n_2-1}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \to F_{n_1-1,n_2-1}.$$

Shrnutí:

K modelování a zkoumání populace používáme výběrové soubory. Je-li výběr reprezentativní, dá se na základě výběru získat určitá představa o populaci.

Výběrové charakteristiky jsou náhodné veličiny - jejich hodnoty se mění podle aktuálního výběru. Hodnotu výběrové charakteristiky na konkrétním výběru nazýváme **pozorovaná hodnota**.

Přehled nejpoužívanějších parametrů populace a příslušných výběrových charakteristik, včetně jejich značení je uveden v následující tabulce.

Základní soubor (populace)	střední hodnota μ $(E(X))$	medián $x_{0,5}$	rozptyl σ^2	směrodatná odchylka σ	pravděpodobnost π
Výběrový soubor (výběr)	(výběrový) průměr $\bar{X}$	výběrový medián $\tilde{X}_{0,5}$	výběrový rozptyl S^2	výběrová směrodatná odchylka S	relativní četnost p

Rozdělení pravděpodobností výběrových charakteristik označujeme pojmem **výběrová rozdělení**.

Důležitá tvrzení o vlastnostech výběrového průměru, pro případ dostatečně velkého rozsahu náhodného výběru, přináší limitní věty. Uvedli jsme si dvě nejdůležitější – zákon velkých čísel a centrální limitní větu.

Zákon velkých čísel říká, že s rostoucím rozsahem výběru se výběrový průměr stále silněji koncentruje kolem střední hodnoty.

Obsah
136. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Centrální limitní věta říká, že výběrový průměr má při *dostatečně velkém počtu pozorování* (v praxi pro $n > 30$) přibližně normální rozdělení, ať už X_i pocházejí z libovolného rozdělení.

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

Na základě CLV byla popsána rozdělení výběrového průměru při dostatečném rozsahu výběru, resp. při výběru z normálního rozdělení, rozdělení relativní četnosti při dostatečném rozsahu výběru, rozdělení rozdílu průměrů dvou nezávislých výběrů z normálního rozdělení a rozdílu relativních četností dvou dostatečně velkých nezávislých výběrů.

Při odhadech rozptylu, poměru rozptylů, odhadech střední hodnoty v případě, že máme k dispozici pouze malý výběr, který nepochází z normálního rozdělení, a v dalších metodách statistické indukce nacházejí uplatnění tři důležitá spojitá rozdělení - χ^2- rozdělení, Studentovo rozdělení a Fisherovo–Snedecorovo rozdělení.

Přehled nejpoužívanějších výběrových charakteristik a jejich rozdělení

Mějme náhodný výběr X z normálního rozdělení, tj.

$$\boldsymbol{X} = (X_1, ..., X_n), \forall i = 1, ..., n : X_i \to N\left(\mu, \sigma^2\right).$$

Mějme dostatečně velký náhodný výběr $\boldsymbol{X}$, tj.

$$n > \frac{9}{p(1-p)}.$$

Výběrová charakteristika	Rozdělení pravděpodobnosti	Poznámka
$\dfrac{\bar{X}-\mu}{\sigma}\sqrt{n}$	$N(0,1)$	viz CLV
$\dfrac{\bar{X}-\mu}{S}\sqrt{n}$	t_{n-1}	viz vlastnosti Studentova rozdělení
$\dfrac{S^2}{\sigma^2}(n-1)$	χ^2_{n-1}	viz vlastnosti χ^2- rozdělení

Výběrová charakteristika	Rozdělení pravděpodobnosti	Poznámka
$\dfrac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$	$N(0,1)$	viz vlastnosti relativní četnosti

Mějme dva nezávislé výběry z normálního rozdělení.

$\forall i = 1, 2, ..., n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu, \sigma_1^2)$,
$\forall j = 1, 2, ..., n_2$, kde n_2 je rozsah druhého výběru: $X_{2j} \to N(\mu, \sigma_2^2)$.

Výběrová charakteristika	Rozdělení pravděpodobnosti	Poznámka
$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$	viz CLV
$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2(n_1-1) + S_2^2(n_2-1)}}\sqrt{\dfrac{n_1 n_2(n_1+n_2-2)}{n_1+n_2}}$	$t_{n_1+n_2-2}$	viz vlastnosti Studentova rozdělení Předpoklad: $\sigma_1^2 = \sigma_2^2$
$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	t_ν $\nu = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\left(\dfrac{S_1^2}{n_1}\right)^2 \dfrac{1}{n_1+1} + \left(\dfrac{S_2^2}{n_2}\right)^2 \dfrac{1}{n_2+1}} - 2$	viz vlastnosti Studentova rozdělení Předpoklad: $\sigma_1^2 \neq \sigma_2^2$
$\dfrac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$	F_{n_1-1,n_2-1}	viz vlastnosti Fisherova – Snedecorova rozdělení

Mějme dostatečně velké náhodné výběry $\boldsymbol{X_1}$ a $\boldsymbol{X_2}$, tj.

$$\left(n_1 > \frac{9}{p_1(1-p_1)}\right) \wedge \left(n_2 > \frac{9}{p_2(1-p_2)}\right).$$

Výběrová charakteristika	Rozdělení pravděpodobnosti	Poznámka
$\dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}}$	$N(0{,}1)$	viz CLV

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Střední hodnota pevně zvolené náhodné veličiny je
   (a) náhodná veličina,
   (b) konstanta,
   (c) náhodný jev,
   (d) výběrová charakteristika.

2. (1b.) Výběrový průměr je
   (a) náhodná veličina,
   (b) konstanta,
   (c) náhodný jev,
   (d) populační charakteristika.

3. (1b.) S rostoucím rozsahem výběru se obvykle rozptyl průměru

(a) snižuje,
(b) zvyšuje,
(c) nemění.

4. (1b.) Statistická indukce je

(a) experiment,
(b) metoda, která umožňuje odhadnout vlastnosti výběru na základě znalostí vlastností populace,
(c) zobecnění statistických výsledků získaných zpracováním výběru na celou populaci,
(d) metoda sběru dat.

5. (1b.) Zákon velkých čísel v důsledku říká, že při dostatečném rozsahu výběru

(a) má průměr normální rozdělení,
(b) má průměr Studentovo rozdělení,
(c) se střední hodnota přibližuje teoretické hodnotě průměru,
(d) se relativní četnost přibližuje teoretické hodnotě pravděpodobnosti.

6. (1b.) Pro modelování průměru výběru dostatečně velkého rozsahu je vhodné použít rozdělení

(a) normální,

(b) Pearsonovo (χ^2),
(c) Studentovo,
(d) Fisherovo-Snedecorovo.

7. (1b.) Pro modelování průměru výběru malého rozsahu je vhodné použít rozdělení

(a) normální,
(b) Pearsonovo (χ^2),
(c) Studentovo,
(d) Fisherovo-Snedecorovo.

8. (1b.) Pro modelování relativní četnosti ve výběru o dostatečném rozsahu je vhodné použít rozdělení

(a) normální,
(b) Pearsonovo (χ^2),
(c) Studentovo,
(d) Fisherovo-Snedecorovo.

9. (1b.) Pro modelování rozptylu výběru z normálního rozdělení je vhodné použít rozdělení

(a) normální,
(b) Pearsonovo (χ^2),
(c) Studentovo,

(d) Fisherovo-Snedecorovo.

10. (1b.) Pro modelování poměru rozptylů dvou výběrů z normálního rozdělení je vhodné použít rozdělení

(a) normální,
(b) Pearsonovo (χ^2),
(c) Studentovo,
(d) Fisherovo-Snedecorovo.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

Jak pracovat s testy?

Zacátek testu

1. Farmář prodává brambory po koších. Váha koše má logaritmicko-normální rozdělení se střední hodnotou 17,80 kg a směrodatnou odchylkou 1,76 kg. Jaká je pravděpodobnost, že celková váha pěti košů brambor bude vyšší než 90 kg?

2. Zaměstnanci jistého podniku mají nárok na jeden den plně hrazené nemocenské měsíčně. Jestliže víme, že zaměstnanci si vybírají cca 0,78 dní měsíčně (na zaměstnance) a v podniku pracuje 220 zaměstnanců, jaká je pravděpodobnost, že si zaměstnanci příští měsíc budou nárokovat více než 195 dní?

3. V továrně na výrobu žárovek bylo při výstupní kontrole zjištěno, že životnost žárovky je (1600 ± 250) hodin. Jaká je pravděpodobnost, že vybereme-li náhodně 100 žárovek, tak jejich průměrná životnost bude nižší než 1560 hodin?

4. Majitel kiosku na tramvajové zastávce odhadnul, že 15 % zákazníků si kupuje hamburger. Ve středu nakupovalo v daném kiosku 375 zákazníků. Jaká je pravděpodobnost, že bylo prodáno více než 65 hamburgerů?

5. Místní firma kompletuje počítače PC. Průměrná doba potřebná k sestavení jednoho počítače je 35 minut. Ve firmě se kompletováním se pracuje 8 hodin denně, 20 dní měsíčně. Jaká je pravděpodobnost, že příští měsíc zaměstnanci sestaví:

(a) více než 300 počítačů,

(b) mezi 250 a 275 počítači (včetně)?

6. Firma XY se zabývá výrobou mobilních telefonů. 5 % výrobků je při výstupní kontrole vyřazeno v důsledku výrobních vad. Jaká je pravděpodobnost, že v kontrolní sérii 500 telefonů bude:

(a) méně než 30 vadných kusů,

(b) mezi 2,5 % a 7,5 % vadných kusů?

7. Před volbami je v populaci státu 52 % příznivců koaličních stran. Jaká je pravděpodobnost, že průzkum veřejnosti rozsahu $n = 1500$ ukáže nesprávně převahu opozice?

8. Pravděpodobnost zásahu letícího cíle střelcem je 0,95. Jaká je pravděpodobnost, že počet zásahu ve 100 pokusech bude alespoň 97?

9. Při zásahu jádra atomu určitého prvku dojde s pravděpodobností 10 % k vyzáření jisté částice.

(a) Kolem jaké střední hodnoty bude kolísat počet vyzářených částic při zásahu 100 jader?

$EX =$ ☐

$\sigma_X =$ ☐

(b) Odhadněte interval, v němž se bude pohybovat počet vyzářených částic při zásahu 100 jader s pravděpodobností 99,9 %.

$P(A < X < B) = 0,999$ $A =$ ☐

$B =$ ☐

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Kapitola 4

Úvod do teorie odhadu

Cíle

Po prostudování tohoto odstavce budete

- rozumět pojmům: bodový odhad, intervalový odhad,
- znát vlastnosti bodového odhadu,
- umět zkonstruovat intervalové odhady pro vybrané parametry normálního rozdělení: střední hodnotu, rozptyl, směrodatnou odchylku, relativní četnost (podíl), poměr dvou rozptylů (směrodatných odchylek), rozdíl dvou středních hodnot a rozdíl relativních četností (podílů).

Poznámka: *Pro porozumění základním principům uplatňovaným v teorii odhadu není nutné, abyste se vztahy pro meze intervalových odhadů jednotlivých parametrů učili zpaměti. Pro řešení konkrétních úloh budete moci využívat statistický software, resp. „tahák“, v němž* *budou potřebné vztahy uvedeny.*

Průvodce studiem

Metody statistické indukce jsou zaměřeny na řešení dvou základních úloh:

- *odhady populačních parametrů,*
- *testování statistických hypotéz o populačních parametrech a rozděleních populace.*

V této kapitole se zaměříme na první z uvedených úloh – na odhady parametrů populace. Na následujícím příkladu se pokusíme znovu ukázat rozdíl mezi výběrem (parametry výběru) a populací (parametry populace). Dále byste si na příkladu měli ujasnit, proč potřebujeme parametry populace odhadovat.

Denní produkce tyčí (o daném průměru) ocelářské firmy Tychom činí 600 ocelových tyčí. Naším cílem je určit střední hodnotu tažnosti těchto tyčí.

***Populace** je v tomto případě tvořena všemi tyčemi z denní produkce. Sledovaným statistickým znakem je jejich tažnost. k jejímu modelování slouží náhodná veličina X. Střední hodnota $E(X) = \mu$ (populační průměr) tažnosti je jeden z parametrů této populace. Je zřejmé, že požadovaný úkol, určení střední tažnosti, je prakticky neřešitelný – k jeho splnění bychom museli určit tažnost všech tyčí (destruktivní zkouška) a z naměřených hodnot určit průměr. To by bylo značně kontraproduktivní. Jediné možné řešení je – pokusit se o **odhad** tohoto parametru.*

Neznáme-li rozdělení náhodné veličiny X, pak

parametry náhodné veličiny X nelze většinou přesně určit, lze je jen odhadnout.

Jestliže vybereme náhodně například 10 tyčí (10 tyčí můžeme „obětovat“) a určíme jejich průměrnou tažnost, je zřejmé, že střední hodnota tažnosti bude ležet „blízko“ tohoto průměru. Hodnota průměru závisí na konkrétním výběru. Vybereme-li dalších 10 tyčí, jejich průměrná tažnost může být jiná než v předcházejícím případě. Průměr je ***výběrovou charakteristikou*** *denní produkce tyčí a je tedy* ***náhodnou veličinou****. Proto mu můžeme přiřadit nějaké* ***rozdělení*** *(viz kapitoly 8.4.2, 8.9). Známe-li rozdělení průměru, můžeme vytvářet různé úsudky o střední hodnotě původní náhodné veličiny. Např. dokážeme určit, jaká je pravděpodobnost, že střední hodnota tažnosti leží v námi zvoleném intervalu.*

V této kapitole se dozvíte, jak na základě znalosti výběrového souboru (a jeho charakteristik) najít co nejlepší odhad parametrů základního souboru. Nejdříve si však musíme ujasnit, co pod pojmem „nejlepší odhad“ rozumíme.

Z metodického hlediska používáme dva typy odhadů parametrů populace:

- **bodový odhad**, kdy parametr základního souboru aproximujeme jediným číslem,
- **intervalový odhad**, kdy tento parametr aproximujeme intervalem, v němž s velkou pravděpodobností příslušný populační parametr leží.

O tom, který z výše uvedených odhadů použijeme, rozhoduje konkrétní situace, v níž se nacházíme. Pokud potřebujeme hledaný parametr vyjádřit jedinou hodnotou (většinou v případech, kdy jej budeme používat v dalších výpočtech), použijeme bodový odhad. Potřebujeme-li znát přesnost nalezeného odhadu, použijeme intervalový odhad, najdeme tzv. interval spolehlivosti.

4.1. Bodové odhady

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$ z určitého rozdělení, které závisí na neznámém parametru Θ. Odhadem T parametru Θ je pak výběrová charakteristika $T(X_1, X_2, , \ldots, X_n)$, která nabývá hodnot „blízkých“ neznámému parametru Θ.

4.1.1. Vlastnosti „dobrého“ bodového odhadu

„Dobrý“ (věrohodný) odhad musí splňovat určité vlastnosti. Mezi základní vlastnosti věrohodných odhadů patří

- nestrannost (nevychýlenost, nezkreslenost),
- vydatnost (eficience),
- konzistence.

Protože odhad T je funkcí náhodných veličin $(X_1, X_2, \ldots, X_n)$, je také náhodnou veličinou. Řekneme, že odhad je nestranný, jestliže se jeho střední hodnota rovná hledanému parametru.

$$\boldsymbol{E(T) = \Theta}$$

Je-li odhad nestranný, pak systematicky nenadhodnocuje ani nepodhodnocuje odhadovaný parametr.

Nestrannost sama o sobě nezaručuje, že je odhad „dobrý“. Představte si, že máte k dispozici více nestranných odhadů parametru Θ. (Například k odhadu střední hodnoty lze použít nejen průměr, ale i medián nebo X_1 z výběrového souboru o rozsahu n.) Tyto konkurenční nestranné odhady lze porovnat podle velikosti kolísání kolem odhadované hodnoty.

Nestranný odhad, jehož rozptyl je nejmenší mezi rozptyly všech nestranných odhadů příslušného parametru, se nazývá **nejlepší nestranný (vydatný, eficientní)** odhad.

Někdy jsou vlastnosti odhadů zkoumány v závislosti na rozsahu výběru n. Žádoucí vlastností „dobrého" odhadu je pak konzistence. Odhad $T = T_n$ je **konzistentní**, pokud se s rostoucím rozsahem výběru zpřesňuje, k čemuž dochází pokud

- $\lim\limits_{n\to\infty} E(T_n) = \Theta$,
- $\lim\limits_{n\to\infty} D(T_n) = 0$,

tj. pokud se rozdělení odhadu T s rostoucím rozsahem výběru „zužuje" kolem hledaného parametru Θ.

4.1.2. Přesnost bodového odhadu

Připomeňme si, že bodový odhad je náhodná veličina. I v případě, kdy bude bodový odhad splňovat všechny výše uvedené požadavky je zřejmé, že jeho hodnota, vypočtena na základě jednoho výběru, bude obvykle odlišná od skutečné hodnoty parametru populace. Mírou této odlišnosti je tzv. **výběrová chyba** $(T - \Theta)$, která určuje velikost chyby, které se dopouštíme při odhadu na základě jednoho výběrového souboru. Je-li bodový odhad T nezkresleným odhadem parametru Θ, pak za měřítko přesnosti odhadu považujeme směrodatnou odchylku $\sigma_T = \sqrt{D(T)} = \sqrt{E(T-\Theta)^2}$, pro níž se často používá název **střední kvadratická chyba odhadu**. Střední kvadratická chyba odhadu udává „průměrnou" kvadratickou chybu odhadů určených z různých výběrových souboru daného rozsahu.

Příklad 4.1. Mějme náhodný výběr $(X_1, X_2, \dots, X_n)$ z normálního rozdělení se střední hodnotou μ a konečným rozptylem σ^2. Jako odhad rozptylu σ^2 se často využívá statistika S^2, kterou známe pod názvem výběrový rozptyl.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Dokažme, že tento odhad je

a) nestranný,

b) konzistentní.

Řešení 4.1

4.2. Intervalové odhady

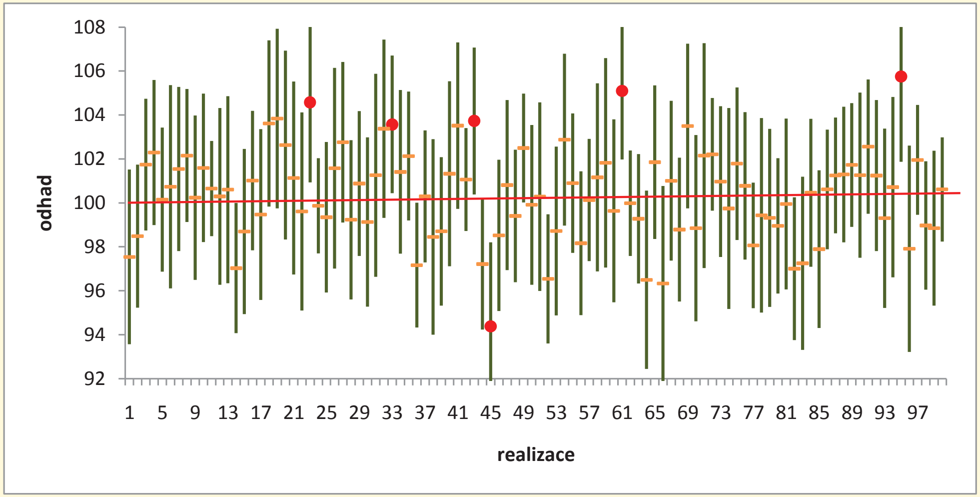
V praktických aplikacích často určujeme odhad příslušného parametru pomocí intervalového odhadu. Tento odhad je reprezentován intervalem $\langle t_D, t_H \rangle$, v němž hledaný parametr leží s předem určenou pravděpodobností (spolehlivostí), kterou označujeme $(1-\alpha)$.

Interval spolehlivosti (konfidenční interval) pro parametr Θ se spolehlivostí $1-\alpha$, kde $\alpha \in \langle 0; 1 \rangle$, je taková dvojice statistik (T_D, T_H), že

$$P(T_D \leqq \Theta \leqq T_H) = 1 - \alpha.$$

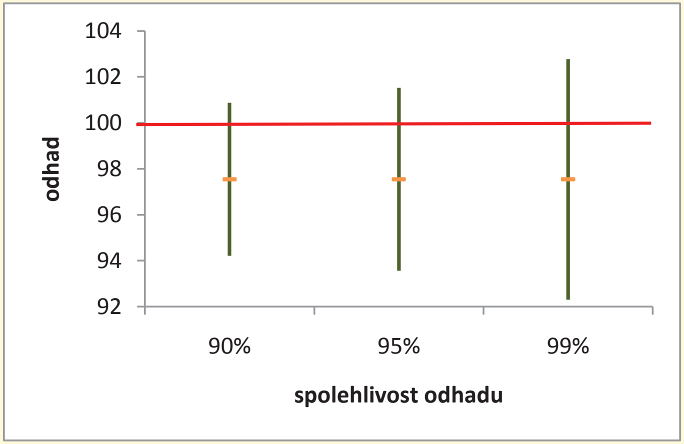
Intervalový odhad parametru Θ se spolehlivostí $1-\alpha$ je interval $\langle t_D, t_H \rangle$, kde t_D, t_H jsou hodnoty statistik T_D, T_H na daném statistickém souboru $(x_1, \dots, x_n)$. Intervalový odhad je tedy jednou z realizací intervalu spolehlivosti.

Spolehlivost odhadu $1-\alpha$ udává, že při opakovaných výběrech s konstantním rozsahem n z dané populace přibližně $100(1-\alpha)\%$ intervalových odhadů obsahuje skutečnou hodnotu odhadovaného parametru Θ a naopak $100\alpha\%$ intervalových odhadů skutečnou hodnotu odhadovaného parametru Θ neobsahuje. Simulace tohoto jevu je ilustrována na obrázku 4.1, který ukazuje 100 intervalových odhadů střední hodnoty (spolehlivost 0,95) získaných na základě opakovaných výběrů o rozsahu 30 z populace se střední hodnotou 100. Oranžové úsečky označují průměry jednotlivých výběrů. V případě, že nalezený intervalový odhad střední hodnoty neobsahuje skutečnou střední hodnotu (100), je průměr označen červeným puntíkem.



Obr. 4.1: Simulace intervalových odhadů střední hodnoty (spolehlivost 0,95) získaných na základě opakovaných výběrů o rozsahu 30 z populace se střední hodnotou 100. 6 intervalů ze 100 neobsahuje skutečnou střední hodnou.

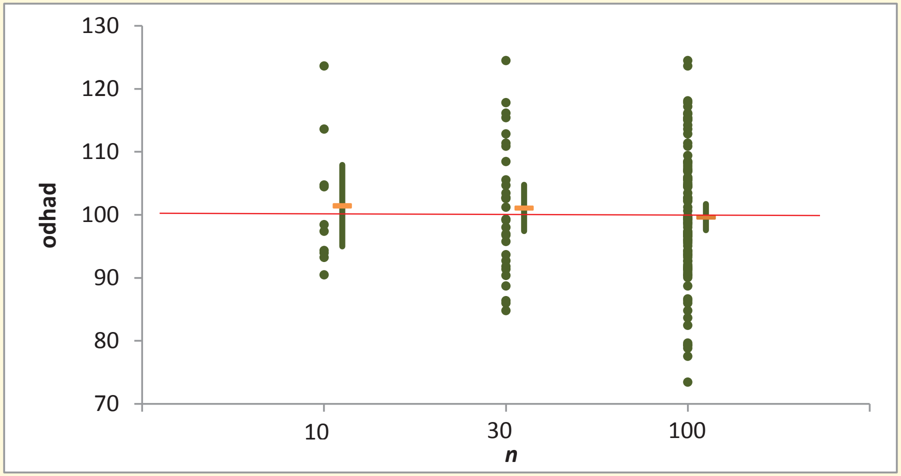
Spolehlivost odhadu $1-\alpha$ požadujeme blízkou jedné, resp. 100%, uvádíme-li ji v procentech. Je zřejmé, že čím vyšší spolehlivost odhadu požadujeme, tím širší intervalový odhad získáme (hledaná hodnota se v něm musí nacházet s vyšší pravděpodobností). Na obrázku 4.2 jsou pro jeden výběr z rozdělení se střední hodnotou rovnou 100 zkonstruovány intervalové odhady střední hodnoty se spolehlivostí 90%, 95% a 99%. Všimněte si, že všechny nalezené intervalové odhady jsou symetrické vzhledem k průměru (značen oranžovou úsečkou) a jejich šířka s rostoucí spolehlivostí roste.



Obr. 4.2: Intervalové odhady střední hodnoty se spolehlivostí 90%, 95% a 99% určené pro jeden výběr z populace se střední hodnotou 100.

Požadavek na spolehlivost odhadu bývá v aplikacích často stanoven předem. Chceme-li intervalový odhad zúžit („zpřesnit“), je proto vhodnější zajistit větší rozsah výběru n. s rostoucím rozsahem výběru se intervalový odhad populačních charakteristik zpřesňuje, tzn. šířka příslušných intervalových odhadů se zmenšuje a to úměrně $\sqrt{n}$ (viz obrázek 4.3).

Rostoucí šířka intervalového odhadu ubírá na jeho vypovídací schopnosti, jeho významnost klesá. (Uvědomte si, jaká je vypovídací schopnost informace, že průměrný věk všech lidí na zemi leží se spolehlivostí 100% v intervalu $\langle 0; 142\rangle$ let.) Proto v praxi vždy hledáme kompromis mezi spolehlivostí a **významností** odhadu. Označíme-li **spolehlivost odhadu** $1-\alpha$, pak α se nazývá **hladinou významnosti**. s rostoucí spolehlivostí odhadu klesá hladina významnosti. V technické praxi se spolehlivost odhadu volí nejčastěji 95% (hladina významnosti tedy bývá 5%).



Obr. 4.3: Intervalové odhady střední hodnoty získané na základě výběru o rozsahu n=10, 30, 100 z populace se střední hodnotou 100.

Intervaly spolehlivosti konstruujeme jako jednostranné (důležitá je pouze jedna mez, odhadujeme-li například délku života nějakého zařízení, je pro nás důležitá pouze dolní mez) nebo oboustranné.

Vliv rozsahu výběru a spolehlivosti odhadu na šířku intervalového odhadu můžete pozorovat v java appletu Intervalové odhady (515 KB).

4.2.1. Jednostranné intervaly spolehlivosti

U jednostranných intervalů spolehlivosti se udává pouze dolní mez (T_D) nebo pouze horní mez (T_H) intervalu.

Je-li dána pouze dolní mez intervalu T_D, mluvíme o **levostranném intervalu spolehlivosti** a platí pro něj

$$P(\Theta \geqq T_D) = 1 - \alpha.$$

Je-li dána pouze horní mez odhadu T_H, mluvíme o **pravostranném intervalu spolehlivosti** a platí pro něj

$$P(\Theta \leqq T_H) = 1 - \alpha.$$

4.2.2. Oboustranný interval spolehlivosti

Zajímají-li nás obě meze odhadu (dolní i horní), konstruujeme oboustranný interval spolehlivosti. Většinou tyto meze určujeme tak, aby platilo, že pravděpodobnost, že parametr populace leží pod dolní mezí byla stejná jako pravděpodobnost, že hledaný parametr leží nad horní mezí a byla rovna $\alpha/2$.

$$P(\Theta < T_D) = P(\Theta > T_H) = \frac{\alpha}{2}$$

Tyto dvě podmínky zaručují, že

$$P(T_D \leqq \Theta \leqq T_H) = 1 - \alpha.$$

Dvojice statistik T_D, T_H se pak nazývá **100(1 − α)% interval spolehlivosti pro parametr Θ**.

4.2.3. Jak najít intervalový odhad parametru Θ?

Připomeňte si, že $100p\%$ kvantil x_p je číslo, pro které platí, že pravděpodobnost, že náhodná veličina bude mít hodnoty menší než x_p je p.

$$P(X < x_p) = F(x_p) = p$$

Je-li X spojitá náhodná veličina, pak $P(X < x_p) = P(X \leqq x_p)$.

Pro libovolné $\alpha \in \langle 0; 1\rangle$ pak platí vztahy, z nichž budeme při odvozeních intervalových odhadů vycházet. Nechť x_p jsou kvantily výběrové charakteristiky $T(X)$, jejíž rozdělení známe. Pak

$$P\left(x_{\frac{\alpha}{2}} \leqq T(X) \leqq x_{1-\frac{\alpha}{2}}\right) = F(x_{1-\frac{\alpha}{2}}) - F(x_{\frac{\alpha}{2}}) = \left(1 - \frac{\alpha}{2} - \frac{\alpha}{2}\right) = 1 - \alpha,$$

$$P(T(X)) \leqq x_{1-\alpha}) = F(x_{1-\alpha}) = 1 - \alpha,$$
$$P(T(X)) \geqq x_{\alpha}) = 1 - F(x_{1-\alpha}) = 1 - \alpha.$$

Připomeňte si, že rozdělení výběrových charakteristik $T(X)$ byla odvozena (v kapitole 8) za předpokladu, že rozsah výběru nepřekročil 5% rozsahu populace, tj. pokud

$$n < 0,05N.$$

Pouze při splnění tohoto předpokladu lze dále uvedené vztahy pro intervalové odhady považovat za správné.

Obecné metody konstrukce intervalů spolehlivosti jsou značně náročné. Pro naše účely se omezíme na **intervaly spolehlivosti pro parametry normálního rozdělení**, které jsou dobře prozkoumané (i proto se tak často setkáváme s požadavkem na normalitu zpracovávaných dat). V případě, že základní soubor nemá normální rozdělení, musíme přistoupit k tzv. **neparametrickým (robustním) metodám odhadu**.

Poznámka: *Robustní statistické metody (useknuté průměry, pořádkové statistiky a windsorizované průměry, ale i Hodgesova-Lehmannova, Huberova, Tukeyova a Hempelova teorie, jak konstruovat robustní odhady v různém slova smyslu optimálně) nacházejí uplatnění všude tam, kde se vyskytují ojedinělé hrubé chyby při měření, a přesto jsme se rozhodli výběrový soubor (naměřené hodnoty) využít k odhadu populačních parametrů.*

4.3. Intervalový odhad střední hodnoty normálního rozdělení

Nejlepším **bodovým odhadem** střední hodnoty μ je průměr $\bar{x}$.

Intervalový odhad střední hodnoty μ se hledá jinak v případě, že známe rozptyl σ^2, resp. směrodatnou odchylku σ, populace (základního souboru) a jinak, když populační rozptyl σ^2, resp. směrodatnou odchylku σ, neznáme.

Obsah
159. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

4.3.1. Intervalový odhad střední hodnoty μ, známe-li směrodatnou odchylku σ

Předpokládejme, že sledovaná náhodná veličina X má normální rozdělení s neznámou střední hodnotou μ a známým rozptylem σ^2. Vyberme vzorek z dané populace. Nechť má tento výběrový soubor rozsah n a průměr $\bar{x}$.

Využijeme poznatku o asymptotickém rozdělení průměru (viz centrální limitní věta – kapitola 8.4.2). Víme, že pro dostatečně velký rozsah výběru lze rozdělení průměru aproximovat normálním rozdělením se střední hodnotou μ a rozptylem σ^2/n.

$$\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

Definujeme-li výběrovou statistiku $T(X)$ jako

$$T(X) = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n},$$

pak má $T(X)$ normované normální rozdělení.

$$T(X) \sim N(0; 1)$$

Nechť $z_{\frac{\alpha}{2}}$ a $z_{1-\frac{\alpha}{2}}$ jsou $100\frac{\alpha}{2}\%$ a $100\left(1 - \frac{\alpha}{2}\right)\%$ kvantily normovaného normálního rozdělení. Pak můžeme tvrdit, že

$$P\left(z_{\frac{\alpha}{2}} \leqq T(X) \leqq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

$$P\left(z_{\frac{\alpha}{2}} \leqq \frac{\bar{X}-\mu}{\sigma}\sqrt{n}, \leqq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Pro kvantily normovaného normálního rozdělení platí: $z_p = -z_{1-p}$. Proto

$$P\left(-z_{1-\frac{\alpha}{2}} \leqq \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leqq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Postupnými úpravami získáme oboustranný interval spolehlivosti pro střední hodnotu (při známém σ).

$$\begin{aligned} P\left(-\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \leqq -\mu \leqq -\bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ P\left(\bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \geqq -\mu \leqq \bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \leqq \mu \leqq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \end{aligned}$$

Oboustranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je tedy

$$\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle.$$

Využitím výběrové charakteristiky $T(X) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ a rovnosti $P(X < x_{1-\alpha}) = 1-\alpha$ získáme levostranný interval spolehlivosti.

$$
\begin{aligned}
P(T(X) \leqq z_{1-\alpha}) &= 1-\alpha \\
P\left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leqq z_{1-\alpha}\right) &= 1-\alpha \\
P\left(-\mu \leqq -\bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) &= 1-\alpha \\
P\left(\mu \geqq \bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) &= 1-\alpha
\end{aligned}
$$

Levostranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je tedy dán dolní mezí

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Jinými slovy, se spolehlivostí $1-\alpha$ je střední hodnota μ větší než $\bar{x} - \dfrac{\sigma}{\sqrt{n}} z_{1-\alpha}$.

Obdobně, dosadíme-li výběrovou charakteristiku $T(X) = \dfrac{\bar{X}-\mu}{\sigma}\sqrt{n}$ do rovnosti $P(X \geqq \geqq x_\alpha) = 1-\alpha$, získáme pravostranný interval spolehlivosti.

$$
\begin{aligned}
P(T(X) \geqq z_\alpha) &= 1-\alpha \\
P(T(X) \geqq -z_{1-\alpha}) &= 1-\alpha \\
P\left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \geqq -z_{1-\alpha}\right) &= 1-\alpha \\
P\left(-\mu \geqq -\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) &= 1-\alpha \\
P\left(\mu \leqq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) &= 1-\alpha
\end{aligned}
$$

Pravostranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je dán horní mezí

$$\bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

Jinými slovy, se spolehlivostí $1-\alpha$ je střední hodnota μ menší než $\bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$.

Přehled intervalových odhadů střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je uveden v tabulce 4.1.

Ve vztazích uvedených v Tab. 4.1 jsou z_p $100p\%$ kvantily normovaného normálního rozdělení. Příslušné kvantily najdete v Tabulce 1 v příloze nebo můžete pro jejich nalezení využít statistický software.

Tab. 4.1: odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2

Intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2	
Oboustranný	$\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \rangle$
Levostranný	$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
Pravostranný	$\bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \rangle$

Výše uvedené intervalové odhady používáme nejen v případech, kdy známe směrodatnou odchylku σ, ale i v případech, kdy máme dostatečně velký výběr ($n \geqq 30$) a směrodatnou odchylku σ neznáme. V těchto případech lze ve výše uvedených vzorcích nahradit směrodatnou odchylku σ výběrovou směrodatnou odchylkou s, aniž by tím vznikla významná chyba.

Odvození dále uvedených intervalových odhadů je založeno na obdobném postupu, proto vybraná odvození uvádíme pouze v kapitole 9.12, která je určena pro zájemce, popřípadě je ponecháváme jako cvičení.

4.3.2. Intervalový odhad střední hodnoty μ, neznáme-li směrodatnou odchylku σ

Podobně jako v kapitole 4.3.1, předpokládejme, že sledovaná náhodná veličina X má normální rozdělení s neznámou střední hodnotou μ. Rozptyl σ^2 náhodné veličiny X však, na rozdíl od kapitoly 4.3.1, neznámé. Vyberme vzorek z dané populace. Nechť má tento výběrový soubor rozsah n, průměr $\bar{x}$ a výběrovou směrodatnou odchylku s.

Přehled intervalových odhadů střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je uveden v tabulce 4.2. (Odvození můžete najít v kapitole 9.12.1.)

Tab. 4.2: Intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při neznámém rozptylu σ^2

Intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při neznámém rozptylu σ^2	
Oboustranný	$\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \rangle$
Levostranný	$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha}$
Pravostranný	$\bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}$

V uvedených vztazích jsou t_p $100\,p\%$ kvantily Studentova rozdělení s $n-1$ stupni volnosti. Příslušné kvantily najdete v Tabulce 2 v příloze nebo můžete pro jejich určení využít statistický software.

Příklad 4.2. Útvar kontroly podniku Edison testoval životnost žárovek. Kontroloři vybrali z produkce podniku náhodně 50 žárovek a došli k závěru, že průměrná doba života (přesněji řečeno výběrový průměr doby života) těchto 50 žárovek je 950 hodin a příslušná výběrová směrodatná odchylka doby života je 100 hodin. Se spolehlivostí 95% určete intervalový odhad střední životnosti žárovek firmy Edison. (Předpokládejte, že životnost žárovek lze modelovat normálním rozdělením.)

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady jednovýběrové (400 KB).

Řešení 4.2

Příklad 4.3. Obchodní řetězec TETO si v dubnu 2006 zadal studii týkající se počtu zákazníků v prodejně TETO Poruba v pátek odpoledne (od 12:00 do 18:00) hodin. Předpokládejme, že sledovaný počet zákazníků má normální rozdělení. Po jednom měsíci sledování prodejny jsme získali údaje uvedené v tabulce 4.3.

Tab. 4.3: Počet zákazníků v TETO Poruba

Datum	Počet zákazníků v TETO Poruba (12:00-18:00) hodin
2.5.2006	3756
9.5.2006	2987
16.5.2006	3042
23.5.2006	4206
30.5.2006	3597

a) Zamyslete se nad důvody, které výzkumníka vedly k analýze výběru o malém rozsahu (mnohem méně než 30 hodnot) a jaké jsou důsledky volby výběru o malém rozsahu.

b) Určete pro managment řetězce TETO intervalový odhad středního počtu zákazníků v prodejně TETO Poruba v pátek odpoledne (se spolehlivostí 95%).

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady jednovýběrové (400 KB).

Řešení 4.3

4.4. Robustní odhady střední hodnoty

Vztahy pro intervalové odhady střední hodnoty uvedené v kapitole 9.3 lze použít pouze v případě, že populace, kterou analyzujeme má normální rozdělení. V obecném případě, kdy neznáme typ rozdělení, používáme tzv. **robustní (neparametrické) postupy**. Robustní postupy hodnocení náhodné veličiny typicky používáme v případech, kdy

- výběrový soubor obsahuje odlehlá pozorování, která nemohou být opravena a není vhodné je vyloučit,
- výběrový soubor nepochází z normálního rozdělení,
- výběrový soubor má velké rozptýlení dat.

Dále popisované intervalové odhady mediánu a Gastwirthova mediánu řadíme mezi robustní intervalové odhady střední hodnoty. Uvedeme pouze jejich výpočetní vztahy pro spolehlivost 0,95.

4.4.1. Odhad mediánu

Medián je prostřední hodnotou uspořádaného datového souboru. Intervalový odhad se spolehlivostí 95% se odhaduje z interkvartilového rozpětí jako

$$\left\langle \hat{x}_{0,5} - 1,57\frac{(\hat{x}_{0,75} - \hat{x}_{0,25})}{\sqrt{n}}; \hat{x}_{0,5} + 1,57\frac{(\hat{x}_{0,75} - \hat{x}_{0,25})}{\sqrt{n}} \right\rangle,$$

kde $\hat{x}_p$ jsou $100p\%$ výběrové kvantily.

4.4.2. Odhad Gastwirthova mediánu

Rovněž Gastwirthův medián x_{GST} patří mezi robustní odhady střední hodnoty. Určuje se pomocí klasického výběrového mediánu, dolního a horního tercilu $(\hat{x}_{0,33}, \hat{x}_{0,67})$. Jeho bodový odhad je dán vztahem

$$\hat{x}_{GST} = 0,4 \cdot \hat{x}_{0,5} + 0,3 \cdot (\hat{x}_{0,33} + \hat{x}_{0,67}).$$

Intervalový odhad Gastwirthova mediánu se spolehlivostí 95% je pak dán jako

$$\left\langle \hat{x}_{GST} - 1,57\frac{(\hat{x}_{0,75} - \hat{x}_{0,25})}{\sqrt{n}}; \hat{x}_{GST} + 1,57\frac{(\hat{x}_{0,75} - \hat{x}_{0,25})}{\sqrt{n}} \right\rangle.$$

4.4.3. Bootstrap

Neznáme-li rozdělení studované populace, můžeme pro odhad střední hodnoty použít metodu bootstrap. Metodu bootstrap navrhl Efron v roce 1979. Základní myšlenka této metody spočívá v tom, že z výběrového souboru o rozsahu n budeme generovat M-tici náhodných výběru (s vracením), každý o stejném rozsahu n. V každém z generovaných výběrů (tzv. **bootstrap výběrů**) se tak libovolný prvek výběrového souboru může opakovat i několikrát (nebo v něm nemusí být obsažen vůbec).

Rozdělení bootstrap výběrů odpovídá rozdělení původního výběru. Z bootstrap výběrů se určí M-tice odhadů hledaného parametru $p_i = p(X)$. Z této M-tice hodnot pak lze určovat intervaly spolehlivosti pomocí celé řady metod. Jednou z nich je tzv. Studentizovaný odhad.

Studentizovaný odhad

Tento odhad vychází z jednoduché transformace vedoucí na náhodnou veličinu t_i, která má Studentovo rozdělení s $n-1$ stupni volnosti.

$$t_i = \frac{\bar{X}_i - \bar{X}}{S_i} \cdot \sqrt{n}, \quad i = 1, \dots M,$$

kde

$\bar{X}_i$ ... průměr i-tého bootstrap výběru,
S_i ... směrodatná odchylka i-tého bootstrap výběru,
$\bar{X}$ ... průměr původního výběru,
n ... rozsah původního výběru (i jednotlivých bootstrap výběrů)

Z rozdělení veličiny t_i můžeme snadno určit $100p\%$ kvantil veličiny t_i, jenž označíme t_{B_p}. Abychom obdrželi přesnější výsledek, museli bychom tento postup zopakovat celkem m krát a z těchto m $100p\%$ kvantilů bychom určili průměrný $100p\%$ kvantil. Zdůrazněme, že rozdělení veličin t_i nemusí být souměrné, tzn. že $100p\%$ kvantil a $100(1-p)\%$ kvantil nemusí mít stejné absolutní hodnoty.

Intervalový odhad s 95% spolehlivostí pro střední hodnotu pak určíme jako

$$\left\langle \bar{x} - t_{B_{0,975}} \cdot \frac{s}{\sqrt{n}}; \bar{x} - t_{B_{0,025}} \cdot \frac{s}{\sqrt{n}} \right\rangle .$$

4.5. Intervalový odhad rozptylu normálního rozdělení

Při modelování určité populace nás obvykle nezajímá pouze její střední hodnota μ, ale i její variabilita. Nejobvyklejšími mírami variability jsou rozptyl σ^2 a směrodatná odchylka σ.

Připomeňme, že nejlepším nestranným bodovým odhadem rozptylu σ^2 je výběrový rozptyl s^2.

Intervalový odhad rozptylu σ^2 se hledá jinak v případě, že známe střední hodnotu populace (základního souboru) a jinak, když tuto střední hodnotu neznáme. Protože znalost střední hodnoty μ při neznalosti rozptylu σ^2 není příliš obvyklá, omezíme se pouze na vztah popisující druhý případ.

Předpokládejme, že sledovaná náhodná veličina X má normální rozdělení s neznámou střední hodnotou μ a neznámým rozptylem σ^2. Zvolme výběrový soubor z dané populace. Nechť má tento výběrový soubor rozsah n a výběrový rozptyl s^2.

Přehled intervalových odhadů rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ je uveden v tabulce 4.4. (Odvození můžete najít v kapitole 9.12.2.) χ_p je $100p\%$ kvantil rozdělení χ^2 s $n-1$ stupni volnosti.

Tab. 4.4: Intervalový odhad rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ

Intervalový odhad rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ	
Oboustranný	$\left\langle \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}} ; \frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}} \right\rangle$
Levostranný	$\frac{(n-1)s^2}{\chi_{1-\alpha}}$
Pravostranný	$\frac{(n-1)s^2}{\chi_{\alpha}}$

4.6. Intervalový odhad směrodatné odchylky normálního rozdělení

Nejlepším nestranným **bodovým odhadem** směrodatné odchylky σ je **výběrová směrodatná odchylka** s.

Intervalový odhad směrodatné odchylky σ najdeme snadno, uvědomíme-li si, že směrodatná odchylka je odmocninou z rozptylu. Stačí tedy upravit intervalové odhady pro rozptyl.

Opět předpokládejme, že sledovaná náhodná veličina X má normální rozdělení s neznámou střední hodnotou μ a neznámým rozptylem σ^2. Zvolme výběrový soubor z dané populace. Nechť má tento výběrový soubor rozsah n a výběrovou směrodatnou odchylku s.

Přehled intervalových odhadů rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ je uveden v tabulce 4.5.

Tab. 4.5: Intervalový odhad směr. odchylky σ se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ

Intervalový odhad směr. odchylky σ se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ	
Oboustranný	$\left\langle \sqrt{\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}}}; \sqrt{\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}}} \right\rangle$
Levostranný	$\sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha}}}$
Pravostranný	$\sqrt{\frac{(n-1)s^2}{\chi_{\alpha}}}$

χ_p je $100p\%$ kvantil rozdělení χ^2 s $n-1$ stupni volnosti.

Příklad 4.4. Automat vyrábí pístové kroužky o daném průměru. Při kontrole kvality bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04 mm. Určete 95% levostranné intervalové odhady rozptylu a směrodatné odchylky průměru pístových kroužků. (Předpokládejte, že průměr pístových kroužku lze modelovat pomocí normálního rozdělení.)

Obsah
172. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady jednovýběrové (400 KB).

Řešení 4.4

4.7. Intervalový odhad relativní četnosti

Nejlepším nestranným **bodovým odhadem** relativní četnosti π je výběrová relativní četnost p.

Máme-li k dispozici výběrový soubor, jehož rozsah

- je dostatečně velký ($n > 30$),
- je menší než 5% rozsahu základního souboru ($\frac{n}{N} < 0,05$),
- splňuje podmínku $n > \frac{9}{p(1-p)}$,

pak lze relativní četnost pi odhadnout pomocí intervalů uvedených v tabulce 4.6. (Odvození můžete najít v kapitole 9.12.3.)

Poznámka: *Relativní četnost π je z intervalu $\langle 0; 1\rangle$. Je tedy zřejmé, že dolní mez intervalových odhadů relativní četnosti nemůže klesnout pod 0 a horní mez těchto odhadů nemůže být větší než 1!*

Příklad 4.5. Při kontrole data spotřeby určitého druhu masové konzervy ve skladech produktů masného průmyslu bylo náhodně vybráno 320 z 20 000 konzerv a zjištěno, že 59 z nich

Obsah
173. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Tab. 4.6: Intervalový odhad relativní četnosti π se spolehlivostí $1-\alpha$

Intervalový odhad relativní četnosti π se spolehlivostí $1-\alpha$ $\left(n > 30, \frac{n}{N} < 0{,}05, n > \frac{9}{p(1-p)}\right)$	
Oboustranný	$\left\langle p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right\rangle$
Levostranný	$p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$
Pravostranný	$p + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$

má prošlou záruční lhůtu. Stanovte se spolehlivostí 95% intervalový odhad podílu konzerv s prošlou záruční lhůtou.

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady jednovýběrové (400 KB).

Řešení 4.5

4.8. Odhad rozsahu výběru

Ještě před zahájením výběrového šetření musíme stanovit minimální velikost výběrového souboru. V kapitole 9.2 bylo ukázáno, že velikost výběru má přímý vliv na přesnost odhadu parametrů základního souboru - čím větší rozsah výběru, tím je intervalový odhad přesnější. V řešeném příkladu, který se věnoval studii pro obchodní řetězec TETO, jsme si však také

Obsah
174. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

ukázali, že ekonomické a časové důvody nás mnohdy nutí volit rozsah výběru co nejmenší. V praxi proto hledáme kompromis, který pro požadovanou přesnost výpočtu povede k co nejmenšímu rozsahu výběru.

V případě, že odhadujeme střední hodnotu nebo relativní četnost, je přesnost intervalového odhadu, tj. **chyba odhadu** Δ, rovna polovině šířky oboustranného intervalu spolehlivosti.

Požadovanou přesnost výpočtu vyjadřujeme pomocí tzv. **přípustné chyby odhadu** Δ_{max}. Jde o hodnotu, o kterou jsme ochotni se zmýlit oproti skutečné hodnotě odhadovaného parametru při dané spolehlivosti odhadu (hladině významnosti). To znamená, že požadujeme, aby chyba odhadu Δ nepřekročila přípustnou chybu odhadu Δ_{max}.

$$\Delta \leqq \Delta_{max}$$

Řešením této nerovnice získáme doporučený rozsah výběru (pro intervalové odhady střední hodnoty, popř. relativní četnosti), který bude postačující pro získání intervalových odhadů střední hodnoty (resp. relativní četnosti) s požadovanou spolehlivostí $1 - \alpha$ a požadovanou maximální přípustnou chybou Δ_{max}.

Odhadovaný rozsah výběru n je ve většině případů nejen funkcí přípustné chyby odhadu Δ_{max} a hladiny významnosti α, ale závisí také na některých dalších výběrových charakteristikách, které v případě, že ještě nemáme stanovený výběr, neznáme. Jejich hodnotu tedy také musíme odhadnout. Obvykle se pro tento účel provádí tzv. **předvýběr**, tj. výběr o malém rozsahu n_1. Pro předvýběr vypočteme požadované výběrové charakteristiky, které považujeme za odhad hledaných výběrových charakteristik. Po zjištění požadovaného rozsahu n pak stačí doplnit předvýběr o chybějících $(n - n_1)$ prvků a intervalový odhad pak provést z výběru o rozsahu n (iterační heuristická metoda).

Příslušná doporučení pro rozsah výběru jsou odvozena v kapitole 9.13 (pro zájemce) a uvedena v tabulce 4.7.

Tab. 4.7: Odhad rozsahu výběru

Odhad rozsahu výběru potřebného pro nalezení intervalového odhadu se spolehlivostí $1-\alpha$ a maximální přípustnou chybou Δ_{max}		
Odhadovaný populační parametr	Požadovaný rozsah výběru	Poznámka
Střední hodnota μ (známe σ)	$n \geq \left(\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}}\right)^2$	z_p je $100p\%$ kvantil normovaného normálního rozdělení
Střední hodnota μ (neznáme σ)	$n \geq \left(\frac{s_1}{\Delta_{max}} t_{1-\frac{\alpha}{2}}\right)^2$	t_p je $100p\%$ kvantil Studentova rozdělení s $n-1$ stupni volnosti, s_1 je výběrová směrodatná odchylka předvýběru
Relativní četnost π	$n \geq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{p_1(1-p_1)}{\Delta_{max}^2}$	z_p je $100p\%$ kvantil normovaného normálního rozdělení, p_1 je výběrová relativní četnost předvýběru
	$n \geq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{4\Delta_{max}^2}$	z_p je $100p\%$ kvantil normovaného normálního rozdělení, nemáme-li k dispozici předvýběr (předběžný odhad relativní četnosti), získáme „nejpřísnější“ odhad rozsahu výběru, dosadíme-li za p hodnotu 0,5.

Obsah
176. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Příklad 4.6. Výběrovým šetřením bychom chtěli odhadnout průměrnou mzdu pracovníků určitého výrobního odvětví. Z vyčerpávajícího šetření, které probíhalo před několika měsíci, víme, že směrodatná odchylka mezd byla 750,- Kč. Odhad chceme provést s 95% spolehlivostí a jsme ochotni připustit maximální chybu ve výši 50,-Kč. Jak velký musíme provést výběr, abychom zajistili požadovanou přesnost a spolehlivost?

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady jednovýběrové (400 KB).

Řešení 4.6

V následujících částech této kapitoly si ještě ukážeme, jak najít intervalové odhady poměru rozptylů dvou populací, rozdílu středních hodnot dvou populací a rozdílu relativních četností dvou populací. Princip odvození těchto odhadů je stejný jako u intervalových odhadů parametrů normálního rozdělení. Odvození těchto odhadů je proto zájemcům ponecháno jako cvičení.

4.9. Intervalový odhad poměru rozptylů dvou populací s normálním rozdělením

Mějme dva výběry z normálního rozdělení, tj.

$\forall i = 1, 2, \ldots, n_1$, kde n_1 je rozsah prvního výběru: $X_{1i} \to N(\mu_1; \sigma_1^2)$,
$\forall i = 1, 2, \ldots, n_2$, kde n_2 je rozsah prvního výběru: $X_{2j} \to N(\mu_2; \sigma_2^2)$.

Nechť výběrové rozptyly S_1^2 a S_2^2 jsou náhodné veličiny definované jako

$$S_1^2 = \frac{\sum\limits_{j=1}^{n_1} \left(X_{1j} - \bar{X}_1\right)^2}{n_1 - 1} \text{ a } S_2^2 = \frac{\sum\limits_{j=1}^{n_2} \left(X_{2j} - \bar{X}_2\right)^2}{n_2 - 1}$$

Z kapitoly 8.10 víme, že

$$T(X) = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \to F_{n_1-1,n_2-1}.$$

Aplikací postupu podrobně prezentovaného v kapitole 9.12 lze snadno odvodit intervalové odhady pro poměr rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$.

Tab. 4.8: Intervalový odhad poměru rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$

Intervalový odhad poměru rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$ se spolehlivostí $1-\alpha$	
Oboustranný	$\left\langle \frac{1}{f_{1-\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2}; \frac{1}{f_{\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2} \right\rangle$
Levostranný	$\frac{1}{f_{1-\alpha}}\frac{S_1^2}{S_2^2}$
Pravostranný	$\frac{1}{f_{\alpha}}\frac{S_1^2}{S_2^2}$

V tabulce f_p označují $100p\%$ kvantily Fisher-Snedecorova rozdělení s $n_1 - 1$ stupni volnosti v čitateli a $n_2 - 1$ stupni volnosti ve jmenovateli.

4.10. Intervalový odhad rozdílu středních hodnot dvou populací s normálním rozdělením

Obdobně jako u odhadu střední hodnoty jedné populace musíme i v tomto případě rozlišit situace, zda známe či neznáme směrodatné odchylky. Intervalový odhad rozdílu středních hodnot dvou populací s normálním rozdělením, z nichž byly pořízeny náhodné výběry, lze provádět za trojího předpokladu.

1. Známe rozptyly σ_1^2 a σ_2^2 obou populací.
2. Neznáme rozptyly obou populací, ale lze předpokládat, že jsou shodné.
3. Neznáme rozptyly obou populací a nelze předpokládat, že jsou shodné.

4.10.1. Intervalový odhad rozdílu středních hodnot dvou populací s normálním rozdělením známe-li jejich rozptyly σ_1^2 a σ_2^2

Mějme dvě populace s normálním rozdělením, jejichž rozptyly σ_1^2 a σ_2^2 známe. Z těchto populací jsme provedli dva nezávislé náhodné výběry o rozsahu n_1 a n_2 a určili jejich průměry $\bar{x}_1$ a $\bar{x}_2$.

V kapitole 8.6 bylo dokázáno, že

$$T(X) = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1).$$

Použitím stejného postupu jako v důkazech uvedených v kapitole 9.12 lze najít příslušné intervalové odhady rozdílu středních hodnot se spolehlivostí $1-\alpha$. Tyto odhady jsou uvedeny v tabulce 4.9.

Tab. 4.9: Intervalový odhad rozdílu středních hodnot $\mu_1 - \mu_2$ (známe σ_1, σ_2)

Intervalový odhad rozdílu středních hodnot $\mu_1 - \mu_2$ se spolehlivostí $1 - \alpha$ (známe σ_1, σ_2)	
Oboustranný	$\left\langle (\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\rangle$
Levostranný	$(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Pravostranný	$(\bar{x}_1 - \bar{x}_2) + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Obdobně jako v případě odhadu střední hodnoty pro jednu populaci, se v praxi většinou setkáváme pouze s případy, kdy neznáme směrodatné odchylky σ_1 a σ_2.

4.10.2. Intervalový odhad pro rozdíl středních hodnot dvou populací s normálním rozdělením neznáme-li jejich rozptyly $\boldsymbol{\sigma}_1^2$ a $\boldsymbol{\sigma}_2^2$, ale víme, že $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2$

Mějme dvě populace s normálním rozdělením, jejichž rozptyly neznáme. Z těchto populací jsme provedli dva nezávislé náhodné výběry o rozsahu n_1 a n_2 a určili jejich průměry $\hat{x}_1$ a $\hat{x}_2$ a výběrové směrodatné odchylky s_1 a s_2.

Je-li $\sigma_1^2 = \sigma_2^2$ (tento předpoklad bývá většinou nutné ověřit statistickým testem, který bude popsán v kapitole 10), pak lze pro nalezení příslušného intervalového odhadu použít statistiku $T(X)$, která má Studentovo rozdělení s $n_1 + n_2 - 2$ stupni volnosti. $T(X)$ je definována jako

$$T(X) = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \quad T(X) \to t(n_1+n_2-2)$$

Příslušné intervaly spolehlivosti pro rozdíl středních hodnot dvou populací s normálním rozdělením a shodnými rozptyly jsou uvedeny v tabulce 4.10.

Tab. 4.10: Intervalový odhad rozdílu středních hodnot $\mu_1 - \mu_2$ (neznáme σ_1, σ_2, ale víme, že $\sigma_1^2 = \sigma_2^2$)

Intervalový odhad rozdílu středních hodnot $\mu_1 - \mu_2$ se spolehlivostí $1-\alpha$ (neznáme σ_1^2, σ_2^2, ale víme, že $\sigma_1^2 = \sigma_2^2$)	
Oboustranný	$\left\langle (\bar{x}_1 - \bar{x}_2) - t_{1-\frac{\alpha}{2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}; (\bar{x}_1 - \bar{x}_2) + t_{1-\frac{\alpha}{2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right\rangle$
Levostranný	$(\bar{x}_1 - \bar{x}_2) - t_{1-\alpha}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
Pravostranný	$(\bar{x}_1 - \bar{x}_2) + t_{1-\alpha}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

t_p jsou $100p\%$ kvantily Studentova rozdělení s $n_1 + n_2 - 2$ stupni volnosti.

4.10.3. Intervalový odhad pro rozdíl středních hodnot dvou populací s normálním rozdělením neznáme-li jejich rozptyly σ_1^2 a σ_2^2, kde $\sigma_1^2 \neq \sigma_2^2$

Mějme dvě populace s normálním rozdělením, jejichž rozptyly neznáme. Z těchto populací jsme provedli dva nezávislé náhodné výběry o rozsahu n_1 a n_2 a určili jejich průměry $\bar{x}_1$ a $\bar{x}_2$ a výběrové směrodatné odchylky s_1 a s_2.

Byl-li statistickým testem zamítnut předpoklad, že $\sigma_1^2 = \sigma_2^2$, pak lze pro nalezení příslušného intervalového odhadu použít statistiku $T(X)$, která má Studentovo rozdělení

s $\dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1+1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2$ (zaokrouhleno na celé číslo) stupni volnosti.

$T(X)$ je definována jako

$$T(X)=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},T(X)\sim t_v,\ \text{kde } v\cong\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\frac{1}{n_1+1}+\left(\frac{s_2^2}{n_2}\right)^2\frac{1}{n_2+1}}-2$$

Příslušné intervaly spolehlivosti pro rozdíl středních hodnot dvou populací s normálním rozdělením a různými rozptyly jsou uvedeny v tabulce 4.11.

Tab. 4.11: Intervalový odhad rozdílu středních hodnot $\mu_1-\mu_2$ (neznáme σ_1,σ_2, ale víme, že $\sigma_1^2\neq\sigma_2^2$)

Intervalový odhad rozdílu středních hodnot $\mu_1-\mu_2$ se spolehlivostí $1-\alpha$ (neznáme σ_1^2,σ_2^2, že $\sigma_1^2\neq\sigma_2^2$)	
Oboustranný	$\langle(\bar{x}_1-\bar{x}_2)-t_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}};(\bar{x}_1-\bar{x}_2)+t_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\rangle$
Levostranný	$(\bar{x}_1-\bar{x}_2)-t_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
Pravostranný	$(\bar{x}_1-\bar{x}_2)+t_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{S_2^2}{n_2}}$

t_p jsou $100p\%$ kvantily Studentova rozdělení s $\dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\frac{1}{n_1+1}+\left(\frac{s_2^2}{n_2}\right)^2\frac{1}{n_2+1}}-2$ stupni volnosti.

4.11. Intervalový odhad pro rozdíl relativních četností dvou populací

Mějme dvě populace. Z těchto populací jsme provedli dva nezávislé náhodné výběry o rozsahu n_1 a n_2. Výběr z první populace obsahoval x_1 prvků se sledovanou vlastností, výběr z druhé populace obsahoval x_2 prvků se sledovanou vlastností. Výběrové relativní četnosti p_1, p_2 jsme pak určili dle vztahů

$$p_1 = \frac{x_1}{n_1}, p_1 = \frac{x_2}{n_2}.$$

Mají-li výběrové soubory rozsahy, které

- jsou dostatečně velké $(n_1 > 30, n_2 > 30)$,
- jsou menší než 5% rozsahu základního souboru $\left(\frac{n_1}{N_1} < 0,05, \frac{n_2}{N_2} < 0,05\right)$,
- splňují podmínky $n_1 > \frac{9}{p_1(1-p_1)}, n_2 > \frac{9}{p_2(1-p_2)}$,

pak má výběrová statistika

$$T(X) = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ kde } p = \frac{x_1 + x_2}{n_1 + n_2}$$

přibližně normované normální rozdělení $(T(X) \sim N(0;1))$.

Jednoduše lze ukázat, že rozdíl relativních četností $\pi_1 - \pi_2$ lze odhadnout pomocí intervalových odhadů uvedených v tabulce 4.12.

Tab. 4.12: Intervalový odhad rozdílu relativních četností $\pi_1 - \pi_2$

Intervalový odhad rozdílu relativních četností $\pi_1 - \pi_2$ se spolehlivostí $1-\alpha$ $\left(\forall i \in \{1{,}2\}: n_i > 30, \frac{n_i}{N_i} < 0{,}05, n_i > \frac{9}{p_i(1-p_i)}\right)$	
Oboustranný	$\left\langle (p_1 - p_2) - z_{1-\frac{\alpha}{2}}\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}; (p_1 - p_2) + z_{1-\frac{\alpha}{2}}\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right\rangle$
Levostranný	$(p_1 - p_2) - z_{1-\alpha}\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
Pravostranný	$(p_1 - p_2) + z_{1-\alpha}\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Poznámka: *Relativní četnosti π_1, π_2 jsou z intervalu $\langle 0; 1\rangle$. Je tedy zřejmé, že dolní mez intervalových odhadů rozdílu relativních četností nemůže klesnout pod -1 a horní mez těchto odhadů nemůže být větší než 1! Pokud meze intervalových odhadů nalezené pomocí vztahů uvedených v tabulce 9.11 tyto podmínky nesplňují, je třeba je upravit.*

Příklad 4.7. Diskety dvou velkých výrobců - DISK a EMEM byly podrobeny zkoušce kvality. Diskety obou výrobců jsou baleny po 20 kusech. Ve 40 balíčcích firmy DISK bylo nalezeno 24 vadných disket, ve 30 balíčcích EMEM bylo nalezeno 14 vadných disket. Se spolehlivostí 0,95 určete intervalový odhad rozdílu relativních četností (procent) vadných disket v celkové produkci firem DISK a EMEM.

Pro řešení tohoto příkladu můžete použít excelovský soubor Intervalové odhady rozdílu, resp. podílu (775 KB).

Řešení 4.7

4.12. Intervalové odhady parametrů normálního rozdělení – odvození

Odvození intervalových odhadů střední hodnoty náhodné veličiny X pro případ, že známe její rozptyl σ^2, bylo provedeno v kapitole 9.3.1. V této kapitole mohou zájemci o matematické pozadí uvedených vztahů nalézt odvození dalších intervalových odhadů parametrů normálního rozdělení.

4.12.1. Intervalový odhad střední hodnoty normálního rozdělení (neznáme $\boldsymbol{\sigma}$)

V praxi se většinou setkáváme s tím, že směrodatnou odchylku σ neznáme. Pokud nemáme ani dostatečný rozsah výběru ($n \geqq 30$), nemůžeme použít intervalové odhady střední hodnoty odvozené v kapitole 9.3.1. Je i v takovém případě možné najít intervalový odhad střední hodnoty?

S ohledem na zadání vezmeme opět vhodné výběrové rozdělení. Nyní to bude takové, které neobsahuje σ a přitom z něj můžeme získat interval spolehlivosti pro μ. Z kapitoly 8.9.1 víme, že pokud náhodné veličiny $X_1, X_2, \ldots, X_n$ mají normální rozdělení $N(\mu, \sigma^2)$ a jsou navzájem nezávislé, pak

$$\frac{\bar{X}-\mu}{S}\sqrt{n} \to t_{n-1}.$$

Nechť $T(X) = \frac{\bar{X}-\mu}{S}\sqrt{n}$. Pak $T(X) \to t_{n-1}$, $t_{\frac{\alpha}{2}}$ a $t_{1-\frac{\alpha}{2}}$ jsou $100\frac{\alpha}{2}\%$ a $100\left(1-\frac{\alpha}{2}\right)\%$ kvantily Studentova rozdělení s $n-1$ stupni volností. Můžeme tvrdit, že

$$\begin{aligned} P\left(t_{\frac{\alpha}{2}} \leqq T(X) \leqq t_{1-\frac{\alpha}{2}}\right) &= 1-\alpha. \\ P\left(t_{\frac{\alpha}{2}} \leqq \frac{\bar{X}-\mu}{S}\sqrt{n} \leqq t_{1-\frac{\alpha}{2}}\right) &= 1-\alpha. \end{aligned}$$

Pro kvantily Studentova rozdělení platí $t_p = t_{1-p}$. Proto

$$P\left(-t_{1-\frac{\alpha}{2}} \leqq \frac{\bar{X}-\mu}{S}\sqrt{n} \leqq t_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Postupnými úpravami získáme oboustranný interval spolehlivosti pro střední hodnotu (při neznámé hodnotě σ).

$$\begin{aligned} P\left(-\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}} \leqq -\mu \leqq -\bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ P\left(\bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}} \geqq \mu \geqq \bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ P\left(\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}} \leqq \mu \leqq \bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \end{aligned}$$

Oboustranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je proto

$$\left\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} ; \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \right\rangle .$$

Využitím výběrové charakteristiky $T(X) = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ a rovnosti $P(X \leqq x_{1-\alpha}) = 1-\alpha$ získáme levostranný interval spolehlivosti.

$$\begin{aligned} P(T(X) \leqq t_{1-\alpha}) &= 1-\alpha \\ P\left(\frac{\bar{X}-\mu}{S}\sqrt{n} \leqq -t_{1-\alpha}\right) &= 1-\alpha \\ P\left(-\mu \leqq -\bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}\right) &= 1-\alpha \\ P\left(\mu \geqq \bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha}\right) &= 1-\alpha \end{aligned}$$

Levostranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při neznámém rozptylu σ^2 je tedy

$$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha} .$$

Obdobně, dosadíme-li výběrovou charakteristiku $T(X) = \frac{\bar{X}-\mu}{S}\sqrt{n}$ do rovnosti $P(X \geqq x_{\alpha}) = 1-\alpha$, získáme pravostranný interval spolehlivosti.

Obsah
188. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

$$
\begin{aligned}
P(T(X) \geqq t_{\alpha}) &= 1-\alpha \\
P\left(T(X) \geqq -t_{1-\alpha}\right) &= 1-\alpha \\
P\left(\frac{\bar{X}-\mu}{S}\sqrt{n} \geqq -t_{1-\alpha}\right) &= 1-\alpha \\
P\left(-\mu \geqq -\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha}\right) &= 1-\alpha \\
P\left(\mu \leqq \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}\right) &= 1-\alpha
\end{aligned}
$$

Pravostranný intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 je tudíž

$$\bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}},$$

Víme, že pro $n \to \infty$ (vysoký počet stupňů volnosti n, v praxi pro $n \geqq 30$) se Studentovo t rozdělení blíží normovanému normálnímu rozdělení. Pro $n \geqq 30$ tedy můžeme kvantily Studentova rozdělení nahradit kvantily normovaného normálního rozdělení. Pak vztahy pro určení intervalů spolehlivosti střední hodnoty v případě neznámé směrodatné odchylky přecházejí ve vztahy pro určení intervalů spolehlivosti střední hodnoty v případě známé směrodatné odchylky, v nichž směrodatnou odchylku aproximujeme výběrovou směrodatnou odchylkou.

4.12.2. Intervalový odhad rozptylu normálního rozdělení (neznáme μ)

Předpokládejme, že sledovaná náhodná veličina X má normální rozdělení. Zvolme výběrový soubor z dané populace. Nechť má tento výběrový soubor rozsah n a výběrový rozptyl s^2.

Z vlastností rozdělení χ^2 (kap. 8.8) víme, že definujeme-li si výběrovou statistiku $T(X)$ jako

$$T(X) = \frac{(n-1)s^2}{\sigma^2},$$

pak má tato náhodná veličina rozdělení χ^2 s $n-1$ stupni volnosti.

$$T(X) \to \chi^2_{n-1}$$

Z toho plyne, že

$$P\left(\chi_{\frac{\alpha}{2}} \leqq T(X) \leqq \chi_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

kde χ_p označuje $100p\%$ kvantil rozdělení χ^2 s $n-1$ stupni volnosti. Postupnými úpravami získáme oboustranný interval spolehlivosti pro rozptyl.

$$\begin{aligned}
P\left(\chi_{\frac{\alpha}{2}} \leqq \frac{(n-1)s^2}{\sigma^2} \leqq \chi_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\
P\left(\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}} \leqq \sigma^2 \leqq \frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}}\right) &= 1-\alpha
\end{aligned}$$

Oboustranný intervalový odhad rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ je

$$\left\langle \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}}; \frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}} \right\rangle.$$

Obdobně lze odvodit levostranný a pravostranný interval spolehlivosti.

$$\begin{aligned} P(T(X) \leqq \chi_{1-\alpha}) &= 1-\alpha \\ P\left(\frac{(n-1)s^2}{\sigma^2} \leqq \chi_{1-\alpha}\right) &= 1-\alpha \\ P\left(\frac{(n-1)s^2}{\chi_{1-\alpha}} \leqq \sigma^2\right) &= 1-\alpha \end{aligned}$$

Levostranný intervalový odhad rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ je

$$\frac{(n-1)s^2}{\chi_{1-\alpha}}.$$

$$\begin{aligned} P(T(X) \geqq \chi_{\alpha}) &= 1-\alpha \\ P\left(\frac{(n-1)s^2}{\sigma^2} \geqq \chi_{\alpha}\right) &= 1-\alpha \\ P\left(\frac{(n-1)s^2}{\chi_{\alpha}} \geqq \sigma^2\right) &= 1-\alpha \\ P\left(\sigma^2 \leqq \frac{(n-1)s^2}{\chi_{\alpha}}\right) &= 1-\alpha \end{aligned}$$

Rozptyl σ^2 nemůže nabývat záporných hodnot, proto je **pravostranný intervalový odhad** rozptylu σ^2 se spolehlivostí $1-\alpha$ při neznámé střední hodnotě μ

$$\frac{(n-1)s^2}{\chi_\alpha}.$$

4.12.3. Intervalový odhad relativní četnosti

Mějme výběrový soubor, jehož rozsah

- je dostatečně velký $(n > 30)$,
- je menší než 5% rozsahu základního souboru $(\frac{n}{N} < 0,05)$,
- splňuje podmínku $n > \frac{9}{p(1-p)}$.

Je-li výběrová charakteristika T(X) definována jako

$$T(X) = \frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n},$$

pak má přibližně normované normální rozdělení (viz kapitola 8.5).

$$T(X) \sim N(0;1)$$

Nechť $z_{\frac{\alpha}{2}}$ a $z_{1-\frac{\alpha}{2}}$ jsou $100\frac{\alpha}{2}\%$ a $100(1-\frac{\alpha}{2})\%$ kvantily normovaného normálního rozdělení. Pak můžeme tvrdit, že

Obsah

192. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

$$\begin{aligned} P\left(z_{\frac{\alpha}{2}} \leqq T(X) \leqq z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha, \\ P\left(z_{\frac{\alpha}{2}} \leqq \frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} \leqq z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha. \end{aligned}$$

Další úpravy výše uvedeného výrazu by nám komplikovalo, že jmenovatel výrazu $\frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$ je funkcí odhadované relativní četnosti π. Relativní četnost π ve jmenovateli proto nahradíme jejím bodovým odhadem p.

$$P\left(z_{\frac{\alpha}{2}} \leqq \frac{p-\pi}{\sqrt{p(1-p)}}\sqrt{n} \leqq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Úpravou tohoto vztahu, při využití vlastnosti symetrie normovaného normálního rozdělení $z_{\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$ pak dostaneme požadovaný oboustranný interval spolehlivosti.

$$\begin{aligned} P\left(-z_{1-\frac{\alpha}{2}} \leqq \frac{p-\pi}{\sqrt{p(1-p)}}\sqrt{n} \leqq z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ P\left(-p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leqq -\pi \leqq -p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) &= 1-\alpha \\ P\left(p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leqq \pi \leqq p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) &= 1-\alpha \end{aligned}$$

Oboustranný intervalový odhad relativní četnosti π se spolehlivostí $1-\alpha$ je tedy

$$\left\langle p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right\rangle.$$

Relativní četnost π je z intervalu $\langle 0; 1\rangle$. Je tedy zřejmé, že relativní četnost nemůže klesnout pod 0 a nemůže být větší než 1. Pokud nalezené meze intervalových odhadů relativních četností nesplňují tyto podmínky, je vhodné je korigovat.

Obdobně bychom mohli ukázat, že **levostranný intervalový odhad** se spolehlivostí $1-\alpha$ je

$$p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$$

a pravostranný intervalový odhad se spolehlivostí $1-\alpha$ je

$$p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}.$$

4.13. Odhad rozsahu výběru - odvození

V této kapitole naleznete, v případě zájmu, odvození doporučení pro rozsah výběru potřebného pro stanovení intervalového odhadu střední hodnoty, resp. relativní četnosti, s požadovanou spolehlivostí a požadovanou přípustnou chybou.

4.13.1. Rozsah výběru při odhadu střední hodnoty

Obdobně jako při hledání intervalového odhadu střední hodnoty, musíme i zde rozlišit dva případy: situaci kdy známe populační směrodatnou odchylku a situaci, kdy tuto směrodatnou odchylku neznáme.

a) Známe populační směrodatnou odchylku σ

Oboustranný intervalový odhad je

$$\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle.$$

Interval je symetrický kolem průměru $\bar{x}$ a má šířku $2\frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$. Polovina šířky oboustranného intervalu spolehlivosti a tedy přípustná chyba odhadu je

$$\Delta = \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

Požadujeme-li, aby přípustná chyba odhadu Δ dosahovala při dané spolehlivosti odhadu maximálně určité přípustné hodnoty, pak rozsah výběru určíme jako funkci této chyby.

$$
\begin{aligned}
\Delta &\leqq \Delta_{max} \\
\frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} &\leqq \Delta_{max} \\
\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}} &\leqq \sqrt{n} \\
n &\geqq \left(\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}}\right)^2 \\
n &= \left\lceil \left(\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}}\right)^2 \right\rceil
\end{aligned}
$$

b) Neznáme populační směrodatnou odchylku $\boldsymbol{\sigma}$

Obdobně jako v předcházejícím případě bychom mohli ukázat, že přípustná chyba odhadu je

$$\Delta = \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}},$$

kde t_p je $100p\%$ kvantil Studentova rozdělení s $n-1$ stupni volnosti.

Přípustná chyba odhadu Δ je v tomto případě nejen funkcí hladiny významnosti α a rozsahu výběru n, ale závisí také na výběrové směrodatné odchylce s, kterou neznáme pokud ještě nemáme stanovený výběr. Její hodnotu tedy musíme odhadnout. Obvykle se pro tento účel provádí tzv. **předvýběr**, tj. výběr o malém rozsahu n_1. Pro předvýběr vypočteme výběrovou odchylku s_1, kterou považujeme za odhad výběrové směrodatné odchylky s. Pak určíme minimální rozsah výběru úpravou příslušného vztahu:

$$
\begin{aligned}
\Delta &\leqq \Delta_{max} \\
\frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} &\leqq \Delta_{max} \\
\frac{s_1}{\sqrt{n}} t_{1-\frac{\alpha}{2}} &\leqq \Delta_{max} \\
\frac{s_1}{\Delta_{max}} t_{1-\frac{\alpha}{2}} &\leqq \sqrt{n} \\
n &\geqq \left(\frac{s_1}{\Delta_{max}} t_{1-\frac{\alpha}{2}}\right)^2
\end{aligned}
$$

Po zjištění požadovaného rozsahu n pak stačí doplnit předvýběr o chybějících $(n-n_1)$ prvků a pak provést intervalový odhad z výběru o rozsahu n (iterační heuristická metoda).

4.13.2. Rozsah výběru při odhadu relativní četnosti (podílu)

Je-li rozsah výběru n

- dostatečně velký $(n > 30)$,
- menší než 5% rozsahu základního souboru $(\frac{n}{N} < 0,05)$,
- splňující podmínku $n > \frac{9}{p(1-p)}$.

pak oboustranný intervalový odhad relativní četnosti π je

$$
\left\langle p - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} ; p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right\rangle .
$$

Polovina šířky oboustranného intervalového odhadu relativní četnosti π a tedy přípustná chyba odhadu Δ je

$$\Delta = z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}.$$

Vidíme, že přípustná chyba odhadu závisí tentokrát na hladině významnosti α a na výběrové relativní četnosti, kterou neznáme. Nemáme-li žádné informace o výběrové relativní četnosti, můžeme dále postupovat dvěma způsoby.

a) Provedeme **předvýběr**, z něhož vypočteme výběrovou relativní četnost p_1, kterou budeme považovat za odhad výběrové relativní četnosti p. Pak odhadneme požadovaný rozsah výběru úpravou příslušného vztahu.

$$\begin{aligned}
\Delta &\leqq \Delta_{max} \\
z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} &\leqq \Delta_{max} \\
z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n}} &\leqq \Delta_{max} \\
z_{1-\frac{\alpha}{2}} \frac{\sqrt{p_1(1-p_1)}}{\Delta_{max}} &\leqq \sqrt{n} \\
n &\geqq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{p_1(1-p_1)}{\Delta_{max}^2}
\end{aligned}$$

Po zjištění požadovaného rozsahu n pak stačí doplnit předvýběr o chybějících $(n - n_1)$ prvků a pak provést intervalový odhad na základě výběru o rozsahu n.

b) Druhou možností je odhadnout výběrovou relativní četnost nejhorší možnou variantou, tj. maximální hodnotou rozptylu $p(1-p)$, které je dosaženo pro

$$p = 0,5.$$

Požadovaný rozsah výběru je pak zřejmě

$$n \geqq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{0,5(1-0,5)}{\Delta_{max}^2},$$

$$n \geqq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{4\Delta_{max}^2},$$

$$n = \left\lceil \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{4\Delta_{max}^2} \right\rceil.$$

Shrnutí:

V praktických případech většinou nedokážeme přesně určit **parametry základního souboru** (populace). k jejich odhadu používáme charakteristiky příslušného výběrového souboru – **výběrové charakteristiky**.

Z metodického hlediska používáme dva typy odhadů parametrů:

- **bodový odhad**, kdy parametr základního souboru aproximujeme jediným číslem,
- **intervalový odhad** (konfidenční interval), kdy tento parametr aproximujeme intervalem, v němž parametr leží s danou pravděpodobností. Této pravděpodobnosti říkáme **spolehlivost odhadu** a označujeme ji $1 - \alpha$, číslo α pak nazýváme **hladinou významnosti**.

„Dobrý“ (věrohodný) bodový odhad musí splňovat určité vlastnosti. Mezi základní vlastnosti věrohodných odhadů patří:

- **nestrannost** (nevychýlenost, nezkreslenost),
- **vydatnost** (eficience),
- **konzistence**.

Tab. 4.13: Intervaly spolehlivosti vybraných populačních parametrů

Odhadovaný parametr		Předpoklady	Meze oboustranného intervalového odhadu		Dolní mez levostranného intervalového odhadu	Horní mez pravostranného intervalového odhadu
			T_D	T_H	T_D	T_H
Míra polohy	$\boldsymbol{\mu}$	normalita, známe σ	$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$	$\bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$	$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$	$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
Míra polohy	$\boldsymbol{\mu}$	normalita, neznáme σ	$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}$	$\bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}$	$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha}$	$\bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha}$
Míry variability	$\boldsymbol{\sigma^2}$	normalita	$\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}}$	$\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}}$	$\frac{(n-1)s^2}{\chi_{1-\alpha}}$	$\frac{(n-1)s^2}{\chi_{\alpha}}$
Míry variability	$\boldsymbol{\sigma}$	normalita	$\sqrt{\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}}}$	$\sqrt{\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}}}$	$\sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha}}}$	$\sqrt{\frac{(n-1)s^2}{\chi_{\alpha}}}$
Relativní četnost	$\boldsymbol{\pi}$	$n > 30$, $\frac{n}{N} < 0{,}05$, $n > \frac{9}{p(1-p)}$	$p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$	$p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$	$p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$	$p + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$

V praktických aplikacích mnohdy určujeme intervalový odhad příslušného parametru. Tento odhad je reprezentován intervalem $t_D; t_H$, v němž hledaný parametr leží s předem určenou spolehlivostí $1 - \alpha$.

Intervalové odhady sestavujeme jako **jednostranné** nebo **oboustranné**. V následující tabulce najdete přehled intervalových odhadů pro vybrané populační parametry.

Ještě před zahájením výběrového šetření musíme stanovit velikost výběrového souboru. V případě, že odhadujeme střední hodnotu nebo relativní četnost, je přesnost intervalového odhadu, tj. **chyba odhadu** Δ, rovna polovině šířky oboustranného intervalu spolehlivosti.

Příslušná doporučení pro rozsah výběru jsou uvedena v tabulce 4.14.

Tab. 4.14: Doporučení pro rozsah výběru

Odhad rozsahu výběru potřebného pro nalezení intervalového odhadu se spolehlivostí $1-\alpha$ a maximální přípustnou chybou Δ_{max}		
Odhadovaný populační parametr	Požadovaný rozsah výběru	Poznámka
Střední hodnota μ (známe σ)	$n \geq \left(\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}}\right)^2$	
Střední hodnota μ (neznáme σ)	$n \geq \left(\frac{s_1}{\Delta_{max}} t_{1-\frac{\alpha}{2}}\right)^2$	s_1 je výběrová směrodatná odchylka předvýběru
Relativní četnost π	$n \geq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{p_1(1-p_1)}{\Delta_{max}^2}$	p_1 je výběrová relativní četnost předvýběru
	$n \geq \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{4\Delta_{max}^2}$	nemáme-li k dispozici předvýběr (předběžný odhad relativní četnosti), získáme „nejpřísnější“ odhad rozsahu výběru, dosadíme-li za p hodnotu 0,5.

Intervalové odhady můžeme použít také ke srovnávání středních hodnot, rozptylů (směrodatných odchylek), resp. relativních četností dvou populací. Příslušné oboustranné intervalové odhady jsou uvedeny v tabulce 4.15.

Tab. 4.15: Intervalové odhady rozdílu, resp. poměru parametrů normálního rozdělení

Odhadovaný vztah mezi parametry	Předpoklady	Oboustranný intervalový odhad	Poznámka
$\boldsymbol{\mu_1 - \mu_2}$	normalita obou populací, známe σ_1, σ_2	$\left\langle (\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\rangle$	
	normalita obou populací, neznáme σ_1^2, σ_2^2, $\sigma_1^2 = \sigma_2^2$	$\left\langle (\bar{x}_1 - \bar{x}_2) - t_{1-\frac{\alpha}{2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}};\ (\bar{x}_1 - \bar{x}_2) + t_{1-\frac{\alpha}{2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\rangle$	t_p je 100p% kvantil Studentova rozdělení s $n_1 + n_2 - 2$ stupni volnosti
	normalita obou populací, neznáme σ_1^2, σ_2^2, $\sigma_1^2 \neq \sigma_2^2$	$\left\langle (\bar{x}_1 - \bar{x}_2) - t_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}};\ (\bar{x}_1 - \bar{x}_2) + t_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\rangle$	t_p je 100p% kvantil Studentova rozdělení s $\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\frac{1}{n_1+1}+\left(\frac{s_2^2}{n_2}\right)^2\frac{1}{n_2+1}} - 2$ stupni volnosti
$\frac{\sigma_1^2}{\sigma_2^2}$	normalita obou populací	$\left\langle \frac{1}{f_{1-\frac{\alpha}{2}}}\frac{s_1^2}{s_2^2}; \frac{1}{f_{\frac{\alpha}{2}}}\frac{s_1^2}{s_2^2} \right\rangle$	f_p označují $100p\%$ kvantily Fisher-Snedecorova rozdělení s $n_1 - 1$ stupni volnosti pro čitatele a $n_2 - 1$ stupni volnosti pro jmenovatele.
$\frac{\sigma_1}{\sigma_2}$	normalita obou populací	$\left\langle \sqrt{\frac{1}{f_{1-\frac{\alpha}{2}}}\frac{s_1^2}{s_2^2}}; \sqrt{\frac{1}{f_{\frac{\alpha}{2}}}\frac{s_1^2}{s_2^2}} \right\rangle$	
$\pi_1 - \pi_2$	$\forall i \in \{1,2\}$: $n_i > 30$, $\frac{n_i}{N_i} < 0{,}05$, $n_i > \frac{9}{p_i(1-p_i)}$	$\left\langle (p_1 - p_2) - z_{1-\frac{\alpha}{2}}\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)};\ (p_1 - p_2) + z_{1-\frac{\alpha}{2}}\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right\rangle$	$p = \frac{x_1+x_2}{n_1+n_2}$

Obsah

204. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Chceme-li najít nejlepší možný odhad směrodatné odchylky vybrané vlastnosti nekonečné populace, měli bychom

(a) použít co možná největší výběrový soubor,

(b) použít co možná nejmenší výběrový soubor,

(c) zjistit hodnotu sledované vlastnosti u všech prvků populace,

(d) použít výběrový soubor o rozsahu nejvýše 10 000 prvků populace.

2. (1b.) Chceme-li najít nejlepší možný odhad směrodatné odchylky vybrané vlastnosti populace o rozsahu 50 000 jednotek (prvků), pak by rozsah výběru neměl překročit

(a) 49 999 jednotek,
(b) 10 000 jednotek,
(c) 5 000 jednotek,
(d) 2 500 jednotek,
(e) 1 000 jednotek.

3. (11b.) Doplňte:

(a) Průměr je . . . .

(a) náhodná veličina
(b) konstanta

(b) Střední hodnota je . . . charakteristika.

(a) výběrová
(b) populační

(c) Odhadujeme-li populační charakteristiku jedním číslem, hovoříme o ... odhadu.

(a) bodovém (b) intervalovém

(d) Řekneme, že odhad je ... , jestliže se jeho střední hodnota rovná hledanému parametru.

(a) nestranný (b) vydatný (c) konzistentní

(e) Nestranný odhad, jehož rozptyl je ... mezi rozptyly všech nestranných odhadů příslušného parametru, se nazývá nejlepší nestranný odhad.

(a) nejmenší (b) největší

(f) Mějme náhodný výběr. S rostoucí spolehlivostí odhadu $1-\alpha$ se obvykle intervalové odhady populačních parametrů ... .

(a) zužují (b) rozšiřují

(g) S rostoucí spolehlivostí odhadu $1-\alpha$ ... hladina významnosti α.

(a) roste (b) klesá

(h) Při dané spolehlivosti odhadu $1-\alpha$ se obvykle intervalové odhady populačních parametrů s rostoucím rozsahem výběru ... .

(a) zužují (b) rozšiřují

(i) V technické praxi se obvykle volí spolehlivost odhadu $1-\alpha$ rovna ... .

(a) 0,80 (b) 0,90 (c) 0,95 (d) 0,99
(e) 0,20 (f) 0,10 (g) 0,05 (h) 0,01

(j) V technické praxi se obvykle volí hladina významnosti α rovna ... .

(a) 0,80 (b) 0,90 (c) 0,95 (d) 0,99
(e) 0,20 (f) 0,10 (g) 0,05 (h) 0,01

(k) Horní mez pravostranného intervalového odhadu je ... horní mez příslušného oboustranného odhadu.

(a) stejná jako (b) menší než (c) větší než

4. (1b.) Výběrová charakteristika (náhodná veličina), která nabývá hodnot „blízkých“ hledanému parametru, se nazývá

(a) bodový odhad hledaného parametru,
(b) nestranný odhad hledaného parametru,
(c) konzistentní odhad hledaného parametru,
(d) vydatný odhad hledaného parametru.

5. (1b.) Interval, v němž skutečná hodnota hledaného parametru leží s pravděpodobností $1 - \alpha$ (α je hladina významnosti), se nazývá

(a) interval spolehlivosti, (b) intervalový odhad.

6. (1b.) Hladina významnosti α je pravděpodobnost toho, že skutečná hodnota hledaného parametru

(a) leží uvnitř intervalu spolehlivosti,
(b) neleží uvnitř intervalu spolehlivosti.

7. (1b.) Spolehlivost odhadu $1 - \alpha$ je pravděpodobnost toho, že skutečná hodnota hledaného parametru

(a) leží uvnitř intervalu spolehlivosti,

(b) neleží uvnitř intervalu spolehlivosti.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

Jak pracovat s testy?

Zacátek testu

1. (1b.) Náhodný výběr pěti států USA má následující rozlohy (v 1 000 čtverečních mil):

$$147, 84, 24, 85, 159$$

Se spolehlivostí 95% určete intervalový odhad střední rozlohy 50 států USA. (Předpokládejte, že pro modelování rozlohy států USA lze použít náhodnou veličinu s normálním rozdělením.)
(Interval zadejte ve tvaru: <x,x;x,x> *)*
tis. mil^2

2. (1b.) Z jedné studijní skupiny byli náhodně vybráni 4 studenti. Jejich výsledky u zkoušky byly: 64, 66, 89 a 77 bodů. Z druhé studijní skupiny byli vybráni 3 studenti a jejich výsledky byly: 56, 71 a 53 bodů. Se spolehlivostí 0,95 určete intervalový odhad rozdílu mezi středními výsledky obou skupin u zkoušky. (Předpokládejte, že výsledky jednotlivých skupin u zkoušky lze modelovat náhodnými veličinami s normálním rozdělením.)

3. (4b.) V náhodném výběru dětské obuvi 40% vzorků nevyhovuje novým požadavkům na kvalitu. Se spolehlivostí 95% určete intervalový odhad podílu nevyhovující dětské obuvi na trhu, jestliže rozsah výběru byl
   (a) $n = 40$,

(b) $n = 50$, $\boxed{}$

(c) $n = 100$, $\boxed{}$

(d) $n = 500$. $\boxed{}$

4. (1b.) Firma Sunoil se na vás obrátila s prosbou, zda byste nemohl(a) odhadnout, který z jeho benzínů dává lepší výkon (ujetá vzdálenost v km), zda A nebo B. Vybral(a) jste tedy náhodně 4 vozy a jel(a) jste s každým 2x po téže trase, jednou se 4 litry benzínu A v nádrži a podruhé se 4 litry benzínu B. (Předpokládejte, že počet ujetých km lze modelovat náhodnou veličinou s normálním rozdělením (pro oba typy benzínu).) Počet ujetých km je uveden v následující tabulce.

Počet ujetých km	
Benzín A	**Benzín B**
23	20
17	16
16	14
20	18

Se spolehlivostí 95% určete intervalový odhad rozdílu středních ujetých vzdáleností. $\boxed{}$

5. (1b.) Pro realizaci rozsáhlého šetření o diferenciaci mezd ve velkém průmyslovém podniku musíme velmi rychle získat určitou představu o průměrné odchylce mezd. Z celkového počtu 10.000 zaměstnanců jsme jich náhodně vybrali 40 a určili průměrnou mzdu 9.450,-Kč a směrodatnou odchylku ve výši 1.200,- Kč. V jakém intervalu lze s 95%

pravděpodobnosti očekávat směrodatnou odchylku mezd v celém podniku? (Předpokládáme, že mzdy v základním souboru všech pracovníků podniku mají normální rozdělení.) [ ]

6. (1b.) Jaký minimální rozsah výběru pro odhad podílu chybně zúčtovaných položek musíme navrhnout, chceme-li při 90% spolehlivosti zajistit přípustnou chybu ± 3%. O možném podílu chybných položek nemáme při prováděném auditu žádnou informaci. $n \geqq$ [ ]

7. (2b.) Hypermarket Hyper chce pro zkvalitnění služeb poskytovaných zákazníkům zkrátit dobu jejich čekání u pokladen. Náhodně bylo vybráno 10 zákazníků a byla změřena doba jejich čekání u pokladny. (Předpokládejte normalní rozdělení dob čekání). Výsledky šetření (v sekundách): 310, 225, 390, 265, 358, 255, 170, 265, 150, 240.

(a) V jakých mezích lze s pravděpodobnosti 0,95 očekávat průměrnou dobu čekání zákazníka na obsluhu? [ ] s

(b) Jaká je horní hranice doby čekání, která nebude s pravděpodobností 0,95 překročena? [ ] s

8. (1b.) Agentura provádějící průzkum veřejného mínění plánuje šetření, na základě kterého chce odhadnout, kolik procent voličů podporuje současnou vládní koalici. Předpokládejme (v praxi tomu tak ovšem není), že jsou dotazování vybírání zcela náhodně. Kolik dotazovaných by mělo být do výběru zařazeno, jestliže si vedení agentury přeje, aby se odhad z výběru nelišil od skutečného podílu příznivců koalice o více než 3%? (Volte hladinu významnosti 0,05.) $n \geqq$ [ ]

9. (2b.) Z 90 zkoušek meze kluzu konstrukční oceli z produkce určité ocelárny byl vypočten výběrový průměr 251,34 MPa a výběrový rozptyl 319,48 MPa^2. Najděte 80% intervalové odhady střední hodnoty a směrodatné odchylky meze kluzu. (Za předpokladu normality dat.)

(a) Intervalový odhad střední hodnoty: ______ MPa

(b) Intervalový odhad směrodatné odchylky: ______ MPa

Konec testu

Správně zodpovězené otázky: ______

Získané body: ______

Procento úspěšnosti: ______

Výsledky

Kapitola 5

Testování hypotéz - princip

Cíle

Po prostudování této kapitoly budete

- znát základní pojmy a principy testování hypotéz,
- znát koncepci klasického testu,
- umět rozhodovat o výsledku testu pomocí p - *hodnoty*,
- umět posoudit chybu při rozhodování,
- umět zkonstruovat operativní charakteristiku.

Průvodce studiem

Jak již víte, metody statistické indukce umožňují na základě výběrových dat usuzovat na obecnější skutečnosti týkající se základního souboru. V předcházející kapitole jsme se zabývali problémem, jak odhadnout prostřednictvím bodového, popř. intervalového odhadu, neznámý populační parametr θ. V této kapitole se seznámíte s principem testování hypotéz.

Cílem výzkumů mnohdy bývá srovnání účinnosti různých metod (např. srovnání úmrtnosti u klasických a laparoskopických operací) či srovnání výsledků různých skupin (např. porovnávání výsledků srovnávacích testů u absolventů odborných učilišť, středních průmyslových škol a gymnázií). Jinými slovy, cílem bývá prokázat nějaký rozdíl, tzv. ***efekt****, parametrů náhodných veličin (zkoumaného znaku). Náš předpoklad ohledně efektu, nazýváme* ***statistickou hypotézou*** *(například: mortalita je u laparoskopických operací nižší než u operací konvenčních, průměrné výsledky srovnávacích testů závisí na typu absolvované střední školy, ...).*

Je zřejmé, že o správnosti hypotézy by bylo možné teoreticky rozhodnout na základě vyčerpávajícího šetření celé dotčené populace. Takovéto vyčerpávající šetření je však, jak již víte z předcházejícího výkladu, většinou neekonomické nebo dokonce technicky neproveditelné. Pro ověření správnosti vyslovené hypotézy proto použijeme vhodný výběrový soubor. Proces ověřování správnosti statistické hypotézy pomocí výsledků získaných z výběrového šetření se nazývá ***testováním hypotéz****.*

5.1. Základní pojmy

5.1.1. Statistická hypotéza

Statistická hypotéza je výrok (tvrzení) o rozdělení pozorované náhodné veličiny zakládající se na předchozí zkušenosti, na rozboru dosavadních znalostí nebo na pouhé domněnce.

Pojednává-li statistická hypotéza o parametrech rozdělení náhodné veličiny (střední hodnotě, mediánu, rozptylu, ...), mluvíme o **parametrické hypotéze**, týká-li se jiných vlastností náhodné veličiny (typu rozdělení, nezávislosti výběru, ...), nazýváme ji **hypotézou neparametrickou**.

Parametrické hypotézy můžeme zapisovat jako

- rovnosti (resp. nerovnosti) mezi testovaným parametrem a jeho předpokládanou hodnotou, například:
  - střední hodnota obsahu cholesterolu v krvi je u české populace $4,7\,\mathrm{mmol}\cdot\mathrm{l}^{-1}$, tj. $\mu = 4,7$,
  - preference jisté politické strany klesly pod $20\,\%$, tj. $\pi < 0,2$.

nebo jako

- rovnosti (resp. nerovnosti) mezi testovanými parametry, například:

- průměrná cena výrobku se v krajích I, II, III neliší, tj. $\mu_I = \mu_{II} = \mu_{III}$,
- preference politické strany A jsou nižší než preference politické strany B, tj. $\pi_A < \pi_B$.

Příkladem neparametrických hypotéz pak mohou být tvrzení:

- výběrový soubor $x_1, x_2, \ldots, x_n$ je výběrem z normálního rozdělení,
- barva očí a barva vlasů u mužů jsou nezávislé znaky.

Jak jste si mohli na uvedených příkladech všimnout, statistické hypotézy lze dělit ještě dalšími způsoby, např. podle počtu šetřených populací (**hypotézy jednovýběrové, dvouvýběrové a vícevýběrové**) nebo podle toho, zda je hypotéza jednoduchým nebo složeným výrokem (**hypotézy jednoduché a složené**).

5.1.2. Nulová a alternativní hypotéza

Exaktním ověřováním správnosti hypotéz o rozdělení náhodné veličiny pomocí výsledků získaných náhodným výběrem, tzv. **testováním hypotéz**, se statistici začali zabývat krátce před vypuknutím druhé světové války. Jeho koncepci vytvořili Jerzy Neymant a Egon Pearson. Testování hypotéz pojali jako rozhodovací proces, v němž proti sobě stojí dvě tvrzení - nulová a alternativní hypotéza.

Nulová hypotéza $\boldsymbol{H_0}$ (někdy též **testovaná hypotéza**) představuje tvrzení, že sledovaný efekt je nulový a bývá vyjádřena rovností mezi testovaným parametrem θ a jeho očekávanou hodnotou θ_0.

$$H_0: \quad \theta = \theta_0$$

Poté, co zformulujeme nulovou hypotézu a získáme výběrový soubor, definujeme **alternativní hypotézu $\boldsymbol{H_A}$** (zkráceně alternativu, někdy označovanou též H_1), která nějakým způsobem popírá tvrzení dané nulovou hypotézou. V případě uvedené nulové hypotézy tak můžeme alternativní hypotézu zapsat pomocí jednoho ze čtyř možných zápisů:

$$
\begin{array}{ll}
\text{a) } H_A: & \theta = \theta_1, \\
\text{b) } H_A: & \theta \neq \theta_0, \\
\text{c) } H_A: & \theta < \theta_0, \\
\text{d) } H_A: & \theta > \theta_0.
\end{array}
$$

Formulaci alternativní hypotézy H_A ve tvaru a), tzv. **jednoduchou alternativní hypotézu**, používáme pouze v případě, kdy se rozhodujeme mezi dvěma hodnotami θ_0 a θ_1 . Dále uvedené alternativní **hypotézy** označujeme jako **složené**.

Zvolíme-li alternativní hypotézu ve tvaru b), pak alternativní hypotéza popírá platnost nulové hypotézy bez bližší specifikace. Tvrdí, že hodnota parametru je jiná než udává nulová hypotéza. Takto formulovaná **alternativní hypotéza** se nazývá **oboustranná**.

V případě c), resp. d), je formulovaná tzv. **jednostranná alternativní hypotéza**, která popírá platnost nulové hypotézy a zároveň tvrdí, že hodnota testovaného parametru je menší, resp. větší, než hodnota uvedená v nulové hypotéze.

Zatímco nulová hypotéza bývá stanovena jednoznačně (pomocí rovnosti, např. $\mu = 100$), pro stanovení alternativní hypotézy máme tři možnosti (např. $\mu < 100, \mu > 100, \mu \neq$ $\neq 100$). Obsahuje-li zadání problému vedoucího na testování hypotéz vztah jednostranné nerovnosti, volí se jako alternativa příslušná jednostranná hypotéza. V ostatních případech volíme oboustrannou alternativní hypotézu. Alternativní hypotéza by měla být v souladu s výběrovým souborem. Pokud tomu tak není, **přizpůsobujeme alternativní hypotézu závěrům získaným z výběrového souboru**.

Následující příklady konkrétních problémů vedoucích na testování hypotéz by Vám měly pomoci ujasnit si probranou terminologii.

1. **Zadání problému:** Ověřte, zda průměrný plat v ČR je větší než 24 000,- Kč.

Populace (základní soubor): všichni občané ČR pobírající mzdu
Sledovaný statistický znak (náhodná veličina): mzda
Nulová hypotéza H_0:$\mu = 24\,000$
Alternativní hypotéza H_A: $\mu > 24\,000$ (zadání obsahuje nerovnost v tomto tvaru)
Poznámka: Průměrný plat zjištěný z výběrového souboru by měl být větší než 24 000,- Kč. Pokud by tomu tak nebylo, měli bychom použít oboustrannou alternativní hypotézu.

2. **Zadání problému:** Ověřte, zda průměrné mzdy ve strojírenství a v hutnictví jsou stejné.

Populace 1 (základní soubor 1): všichni občané pracují ve strojírenství
Populace 2 (základní soubor 2): všichni občané pracují v hutnictví
Sledovaný statistický znak (náhodná veličina): mzda
Nulová hypotéza H_0: $\mu_S = \mu_H$, (kde μ_S, resp. μ_H označuje průměrnou mzdu ve strojírenství, resp. v hutnictví
Alternativní hypotéza H_A: $\mu_S \neq \mu_H$ (zadání problému neobsahuje jednostrannou nerovnost)

3. **Zadání problému:** Ověřte, zda použití bezpečnostních pásů.

   a) ovlivňuje úmrtnost při dopravních nehodách,

   b) snižuje úmrtnost při dopravních nehodách.

Populace 1 (základní soubor 1): účastníci dopravních nehod, kteří seděli na místech, na nichž je možno používat bezpečnostní pásy a byli připoutáni
Populace 2 (základní soubor 2): účastníci dopravních nehod, kteří seděli na místech, na nichž je možno používat bezpečnostní pásy a nebyli připoutáni
Sledovaný statistický znak (náhodná veličina): úmrtnost (relativní četnost zemřelých)
Nulová hypotéza H_0: $\pi_A = \pi_N$, (kde π_A, resp. π_N označuje úmrtnost účastníků dopravních nehod, kteří byli, resp. nebyli připoutáni
Alternativní hypotéza H_A:

a) $\pi_A \neq \pi_N$ (zadání problému neobsahuje jednostrannou nerovnost)

b) $\pi_A < \pi_N$ (zadání problému obsahuje nerovnost v uvedeném tvaru)

Poznámka: Při řešení problému b) by úmrtnost těch, co používají bezpečnostní pásy, měla být menší než úmrtnost těch, co bezpečnostní pásy nepoužívají (ve výběru z účastníků dopravních nehod). Pokud tomu tak není, měli bychom použít oboustrannou alternativní hypotézu.

5.1.3. Test statistické hypotézy

Testem statistické hypotézy rozumíme rozhodovací proces, při kterém na základě výběrového souboru provedeme rozhodnutí ve prospěch právě jedné z předkládaných hypotéz. Hypotézy tedy musí být formulovány tak, aby v daném okamžiku platila právě jedna.

Nulovou **hypotézu $\boldsymbol{H_0}$** přitom **považujeme za pravdivou až do okamžiku, kdy nás informace získané z výběrového souboru přesvědčí o opaku**. (*Srovnejte s principem presumpce neviny aplikovaným v soudnictví.*) Protože test statistické hypotézy můžeme provádět opakovaně, je zřejmé, že můžeme dospět pouze ke dvěma rozhodnutím.

a) Zamítáme hypotézu H_0 ve prospěch hypotézy H_A.
b) Nezamítáme H_0.

K jakému rozhodnutí se přiklonit? Obor hodnot testovaného parametru θ se dělí na dvě disjunktní množiny, které nazýváme **obor přijetí** (testované hypotézy H_0) $\boldsymbol{V}$ a **kritický obor** (obor zamítnutí hypotézy H_0) $\boldsymbol{W}$. Kritický obor W se stanovuje tak, aby pravděpodobnost výskytu pozorované hodnoty testovaného parametru θ v něm byla velmi malá. Hranice mezi kritickým oborem a oborem přijetí se nazývá **kritická hodnota testu** a označuje t_{krit}.

Padne-li tedy pozorovaná hodnota testovaného parametru θ do kritického oboru W, zamítáme H_0. Padne-li pozorovaná hodnota do oboru přijetí V, hypotézu H_0 nezamítáme.

Poznámka: *Všimněte si, že nikdy nelze říci, že jsme „přijali hypotézu H_0“ - nikdy nevíme, zda by informace z jiného výběru neumožnila hypotézu H_0 zamítnout.*

5.1.4. Testová statistika (testové kritérium)

Abychom mohli provést korektní test statistické hypotézy, musíme mít k dispozici nástroj, který nám to umožní. Tímto nástrojem nazývaným testovou statistikou, někdy také testovým kritériem, je výběrová charakteristika $T(X)$, která má vztah k nulové hypotéze, a jejíž rozdělení za předpokladu platnosti nulové hypotézy známe.

Kritický obor W lze často popsat prostřednictvím kritického oboru W^* testové statistiky $T(X)$. Padne-li pozorovaná hodnota testové statistiky $T(X)$ do kritického oboru W^*, zamítáme H_0. V opačném případě hypotézu H_0 nezamítáme.

5.1.5. Chyba I. a II. druhu

Při uvedeném způsobu rozhodování nastane vždy některý z případů, které popisuje Tab. 10.1.

Jestliže nulová hypotéza je ve skutečnosti platná a my ji přesto zamítneme, dopouštíme se chyby, označované jako **chyba I. druhu**. Pravděpodobnost, že k takovémuto pochybení dojde, nazýváme **hladina významnosti** a označujeme ji α. Platí-li nulová hypotéza a my jsme ji nezamítli, rozhodli jsme správně. Pravděpodobnost tohoto rozhodnutí označujeme $1-\alpha$ a nazýváme ji **spolehlivost testu**. Správným rozhodnutím je rovněž zamítnutí nulové hypotézy v případě, že je platná hypotéza alternativní. Tohoto rozhodnutí se dopouštíme s pravděpodobností $1-\beta$, což bývá označováno jako **síla testu**. **Chybou II. druhu** je

Tab. 5.1: Přehled výsledků testování hypotéz

		Výsledek testu	
		Nezamítáme H_0	Zamítáme H_0
Skutečnost	Platí H_0	Správné rozhodnutí $1-\alpha$ (spolehlivost testu)	**Chyba I. druhu** α (hladina významnosti)
	Platí H_A	**Chyba II. druhu** β	Správné rozhodnutí $1-\beta$ (síla testu)

nezamítnutí nulové hypotézy v případě, že je platná hypotéza alternativní. Pravděpodobnost této chyby označujeme β.

Pravděpodobnosti α a β, s nimiž chyby I. a II. druhu nastávají, rozhodují o kvalitě testu. Je-li test hypotézy $H_0 : \theta = \theta_0$ oproti alternativě $H_1 : \theta = \theta_1$ založený na testové statistice $T(X)$ s kritickým oborem W^*, pak

- $P\left(T(X)\right) \in W^*|H_0) = \alpha$
- $P\left(T(X)\right) \in V^*|H_A) = \beta$
- $P\left(T(X)\right) \in W^*|H_A) = 1 - \beta$

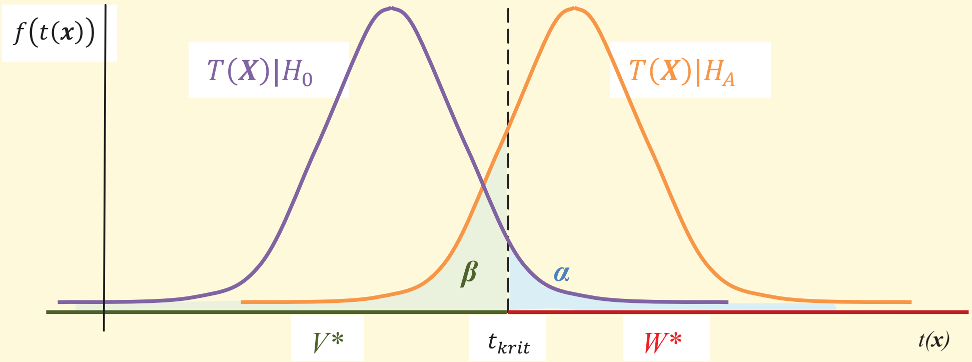
Při testování hypotéz se samozřejmě snažíme postupovat tak, abychom minimalizovali obě chyby, tj. dosáhnout vysoké síly testu (nízkého β) při co nejnižší hladině významnosti α.

Obsah

222. strana ze 525

Zavřít dokument

Celá obrazovka/Okno



Obr. 5.1:Demonstrace pravděpodobností chyb I. a II. druhu

To však není možné, neboť snížením β se zvýší hladina významnosti α a naopak. Proto je třeba najít kompromis mezi požadavky na α a β.

Ve statistice se volí jako rozhodující vstupní parametr testu pravděpodobnost chyby I. druhu – hladina významnosti α. V technických oblastech volíme obvykle hladinu významnosti $\alpha = 0,05$, ve speciálních případech (některé medicínské aplikace) nároky na pravděpodobnost chyby I. druhu ještě zvyšujeme (volíme $\alpha = 0,01$).

Chybu II. druhu β snižujeme volbou vhodného testu (pokud máme možnost výběru) popřípadě zvětšením rozsahu výběrového souboru, což je jediný způsob jak snížit pravděpodobnost chyby II. druhu β, aniž bychom tím zvýšili pravděpodobnost chyby I. druhu α.

V následující animaci si můžete ověřit, zda jste pochopili, co to jsou a čím jsou ovlivnitelné chyby I. a II. druhu.

Chyba prvního a druhého druhu

Instrukce pro spuštění vložených animací naleznete na *http://mi21.vsb.cz/adobe-reader-multimedia*.

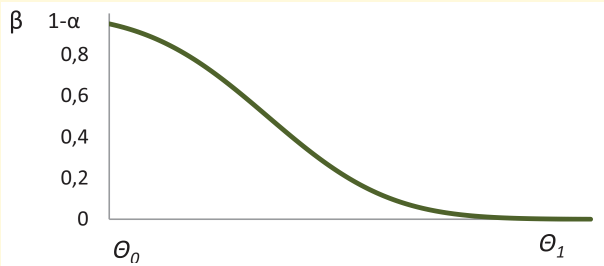
5.1.6. Operativní charakteristika

Proto, abychom určili pravděpodobnost chyby II. druhu β, musí být alternativní hypotéza dána jako hypotéza jednoduchá, tj.

$$H_A : \theta = \theta_1$$

V inženýrských aplikacích se pak mnohdy setkáváme s tzv. **operativní charakteristikou**, což je závislost pravděpodobnosti chyby II. druhu β na přesné specifikaci alternativní hypotézy.

Schematické znázornění operativní charakteristiky přináší následující obrázek:

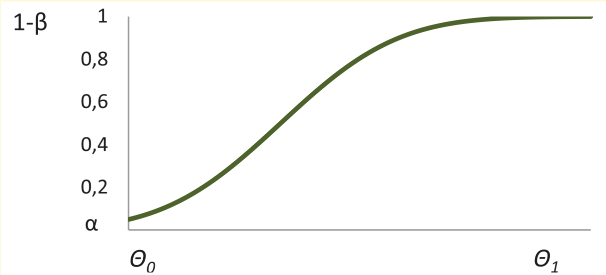


Obr. 5.2: Schematické znázornění operativní charakteristiky pro alternativu ve tvaru $\theta > \theta_0$

Z obrázku 10.2 je zřejmé, že vzdaluje-li se hodnota θ_1 testovaná v alternativní hypotéze od hodnoty θ_0 testované v nulové hypotézy, pravděpodobnost chyby II. druhu β klesá.

Obsah
225. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Místo operativní charakteristiky se mnohdy znázorňuje **křivka síly testu** (angl. „power curve“), tj. závislost síly testu $(1-\beta)$ na přesné specifikaci alternativní hypotézy.



Obr. 5.3: Schematické znázornění křivky síly testu pro alternativu ve tvaru $\theta > \theta_0$

V následující animaci si můžete vyzkoušet čím je ovlivnitelný tvar grafu operační charakteristiky, resp. tvar silofunkce.

Operacní charakteristika

Instrukce pro spuštění vložených animací naleznete na
http://mi21.vsb.cz/adobe-reader-multimedia.

5.2. Přístupy k testování hypotéz

Při testování hypotéz se běžně můžeme setkat se dvěma přístupy – klasickým testem a čistým testem významnosti. My se nejprve seznámíme obecně s oběma postupy a v dalším textu se pak zaměříme na čistý test významnosti.

5.2.1. Klasický test

Klasický test se skládá z několika kroků:

1. *Formulace nulové a alternativní hypotézy.*
2. *Volba testové statistiky (testového kritéria)* $T(X)$ – jde o výběrovou charakteristiku, na jejímž základě rozhodneme o pravdivosti nulové hypotézy. Pro další krok testu musíme znát rovněž rozdělení testové statistiky $T(X)$ při platnosti H_0 (nulové rozdělení) $F_0(x) = P\left(T(X) < x | H_0\right)$.
3. *Stanovení hladiny významnosti testu* α.
4. *Sestrojení kritického oboru* W^* *testové statistiky* $T(X)$.
   Konstrukce kritického oboru: Kritický obor W^* bude vymezen tak, aby pravděpodobnost, že testová statistika $T(X)$ leží v kritickém oboru W^* za předpokladu platnosti nulové hypotézy, byla rovna zvolené hladině významnosti α.

$$P\left(T(X) \in W^* | H_0\right) = \alpha$$

   Známe-li nulové rozdělení testové statistiky $T(X)$, není obtížné pro dané α stanovit kritický obor. (T_p značíme $100p\,\%$ kvantil nulového rozdělení testové statistiky $T(X)$).

a) Je-li **alternativní hypotéza** ve tvaru $\theta < \theta_0$ (ve prospěch alternativy svědčí nízké hodnoty testové statistiky), pak je kritický obor vymezen jako

$$W^* < T_\alpha$$

b) Je-li **alternativní hypotéza** ve tvaru $\theta > \theta_0$ (ve prospěch alternativy svědčí vysoké hodnoty testové statistiky), pak je kritický obor vymezen jako

$$W^* < T_{1-\alpha}$$

c) Je-li **alternativní hypotéza** ve tvaru $\theta \neq \theta_0$ (ve prospěch alternativy svědčí extrémně nízké nebo extrémně vysoké hodnoty testové statistiky), pak je kritický obor vymezen jako

$$W^* < T_{\frac{\alpha}{2}} \,\text{nebo}\, W^* > T_{1-\frac{\alpha}{2}}$$

5. *Výpočet pozorované hodnoty testové statistiky $T(X)$*
Předcházející kroky jsme mohli podniknout v rámci přípravy testu. V tomto kroku již musíme mít k dispozici výběrový soubor a pomocí něj určit konkrétní realizaci testové statistiky $T(X)$, kterou označíme x_{OBS}.

6. *Formulace závěru testu*
Jak již bylo zmíněno, každý test vede ke dvěma možným výsledkům.

a) Leží-li pozorovaná hodnota x_{OBS} v kritickém oboru W^*, **zamítáme nulovou hypotézu ve prospěch alternativní hypotézy.**

b) Neleží-li pozorovaná hodnota x_{OBS} v kritickém oboru W^*, **nulovou hypotézu nezamítáme.**

5.2.2. Čistý test významnosti

Jiným přístupem k testování hypotéz je tzv. čistý test významnosti. Oproti klasickému testu nepotřebujeme při čistém testu významnosti hladinu významnosti jako vstupní údaj. Jeho výsledek nám umožňuje rozhodnout, na jakých hladinách významnosti můžeme nulovou hypotézu zamítnout (resp. nezamítnout).

Čistý test významnosti se skládá z následujících kroků (všimněte si podobnosti s postupem při klasickém testu významnosti):

1. *Formulace nulové a alternativní hypotézy.*
2. *Volba testové statistiky (testového kritéria) $T(X)$.*
3. *Výpočet pozorované hodnoty x_{OBS} testové statistiky $T(X)$.*
4. *Výpočet p-hodnoty (angl. „p-value" nebo „significance level").*
   Je zřejmé, že čím nižší hladinu významnosti α, resp. čím vyšší spolehlivost $1-\alpha$, zvolíme, tím širší obor přijetí dostaneme a opačně - čím vyšší hladinu významnosti α , resp. čím nižší spolehlivost $1-\alpha$ zvolíme, tím užší obor přijetí dostaneme. Při určité hladině významnosti tedy kritická hodnota t_{krit} (hranice mezi oborem přijetí a kritickým oborem) splyne s pozorovanou hodnotou x_{OBS}. Tato hodnota hladiny významnosti se nazývá ***p-hodnota***. *P-hodnota* je tedy nejnižší hladina významnosti, na níž můžeme nulovou hypotézu zamítnout a zároveň nejvyšší hladiny významnosti, na níž se již nulová hypotéza nezamítá.

   Pozorovanou hodnotu statistiky *p-hodnota* vypočteme v závislosti na tvaru alternativní hypotézy podle jedné ze tří možných definic. Připomeňme, že je nutné, aby alternativní hypotéza korespondovala s výběrovým souborem.

Obsah
230. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

a) Je-li alternativa ve tvaru $\theta < \theta_0$, pak *p-hodnotu* určíme dle vztahu

$$p\text{-}hodnota = F_0\left(x_{OBS}\right).$$

Je-li alternativa v uvedeném tvaru, pak v neprospěch nulové hypotézy svědčí hodnoty příslušné výběrové charakteristiky významně nižší než testovaná hodnota θ_0. V tomto případě *p-hodnota* udává pravděpodobnost, že testovaný parametr populace bude nejvýše tak velký jako skutečně zjištěná příslušná výběrová charakteristika, za předpokladu, že H_0 je pravdivá.

b) Je-li alternativa ve tvaru $\theta > \theta_0$, pak *p-hodnotu* určíme dle vztahu

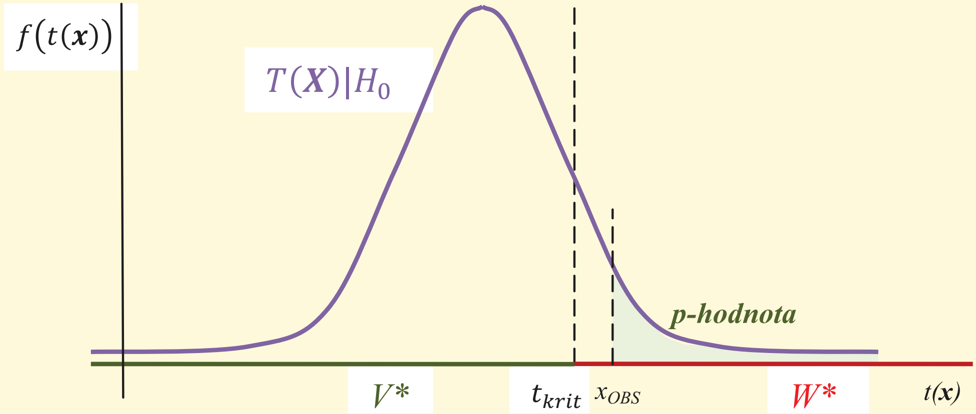
$$p\text{-}hodnota = 1 - F_0\left(x_{OBS}\right).$$

Je-li alternativa v uvedeném tvaru, pak v neprospěch nulové hypotézy svědčí hodnoty příslušné výběrové charakteristiky významně vyšší než testovaná hodnota θ_0. V tomto případě *p-hodnota* udává pravděpodobnost, že testovaný parametr populace bude alespoň tak velký jako skutečně zjištěná příslušná výběrová charakteristika, za předpokladu, že H_0 je pravdivá (viz Obr. 10.4).

c) Je-li alternativa ve tvaru $\theta \neq \theta_0$, pak *p-hodnotu* určíme dle vztahu

$$p\text{-}hodnota = 2min\left\{F_0\left(x_{OBS}\right); 1 - F_0\left(x_{OBS}\right)\right\}.$$

Je-li alternativa v uvedeném tvaru, pak v neprospěch nulové hypotézy svědčí hodnoty příslušné výběrové charakteristiky významně nižší nebo významně vyšší než testovaná hodnota θ_0. V tomto případě *p-hodnota* udává pravděpodobnost, že testovaný



Obr. 5.4: Ilustrace p-hodnoty pro alternativu ve tvaru $\theta > \theta_0$

parametr populace bude alespoň tak extrémní vzhledem k θ_0 jako skutečně zjištěná příslušná výběrová charakteristika, za předpokladu, že H_0 je pravdivá.

POZOR! *Tuto definici p-hodnoty lze použít pouze v případech,* ***kdy nulové rozdělení je symetrické*** *(tzn. nelze použít např. při testování rozptylu). p-hodnota je pak dvojnásobná vzhledem k jednostranným testům.*

Již víte co to je p-hodnota a jak se na jejím základě rozhoduje o výsledku testu? Ověřit si to můžete v následující animaci.

p-hodnota

Instrukce pro spuštění vložených animací naleznete na
http://mi21.vsb.cz/adobe-reader-multimedia.

5. *Rozhodnutí na základě p-hodnoty.*
   P-hodnota nám říká jaká je **minimální hladina významnosti, na níž bychom při daném výběrovém souboru mohli nulovou hypotézu zamítnout**. Například: je-li *p-hodnota* $= 0{,}006$, pak nulovou hypotézu H_0 můžeme zamítnout na hladinách významnosti 0,006 a vyšších. Jinak řečeno: nulovou hypotézu H_0 můžeme zamítnout se spolehlivostí nejvýše 0,994. Zvolíme-li si spolehlivost testu vyšší než 0,994, *p-hodnota* $=$ 0,006 nesvědčí pro zamítnutí nulové hypotézy.

   Je zřejmé, že čím menší je *p-hodnota*, tím silnější je výpověď náhodného výběru proti nulové hypotéze. Ale jak malá musí být *p-hodnota*, aby empirická výpověď byla dostatečně silná k zamítnutí nulové hypotézy? Výsledek testu obecně závisí na zvolené hladině významnosti α. Při známé *p-hodnotě* je rozhodnutí dáno tabulkou 10.2.

Tab. 5.2: Rozhodování na základě p-hodnoty

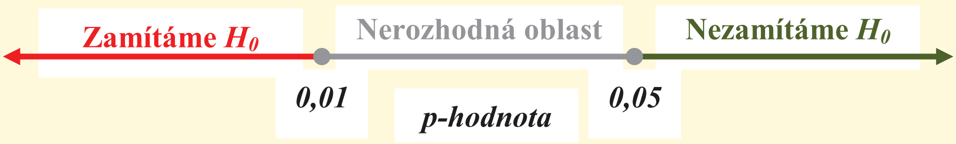
p-hodnota	**Rozhodnutí**
p-hodnota $< \alpha$	Zamítáme H_0 ve prospěch H_A.
p-hodnota $> \alpha$	Nezamítáme H_0.

Není-li při testování hypotéz specifikována hladina významnosti α, pak o zamítnutí nulové hypotézy rozhodujeme většinou na základě následujícího schématu (Tab. 10.3), které je založeno na nejběžněji používaných hladinách významnosti 0,01 a 0,05.

Tab. 5.3: Rozhodnutí na základě p-hodnoty, není-li specifikována hladina významnosti α

p-hodnota	**Rozhodnutí**
p-hodnota < 0,01	Zamítáme H_0 ve prospěch H_A.
0,01 < *p-hodnota* < 0,05	Většinou doporučujeme opakovat test s větším rozsahem výběru.
p-hodnota > 0,05	Nezamítáme H_0.

Je-li *p-hodnota* $< 0,01$, pak je také *p-hodnota* $< 0,05$ a na obou obvyklých hladinách významnosti nulovou hypotézu zamítáme. Je-li *p-hodnota* $> 0,05$, pak je taktéž *p-hodnota* $> 0,01$ a na obou obvyklých hladinách významnosti nulovou hypotézu nezamítáme. Je-li $0,01 <$*p-hodnota* $< 0,05$, pak na hladině významnosti 0,01 nulovou hypotézu nezamítáme, avšak na hladině významnosti 0,05 nulovou hypotézu zamítáme. V tomto případě je vhodné test opakovat s větším rozsahem výběru.



Obr. 5.5: Schéma pro rozhodování o správnosti nulové hypotézy (založeno na hladinách významnosti 0,01 a 0,05)

Příklad 5.1. Výšku asijských hybridů lilií lze modelovat náhodnou veličinou s normálním rozdělením $N(100; 144)$; tzn. průměrná výška μ tohoto druhů lilií je 100 cm a směrodatná odchylka výšky σ je 12 cm. Skupina 100 kusů těchto lilií byla pěstována za příznivějších podmínek, aby se zjistilo, zda se výška zvýší.

a) Určete kritickou hodnotu průměrné výšky tohoto vzorku, při jejímž překročení bude možno se spolehlivostí 0,95 tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

b) Průměrná výška testovaného vzorku lilií je 102,5 cm. Ověřte klasickým testem, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

c) Průměrná výška testovaného vzorku lilií je 102,5 cm. Ověřte čistým testem významnosti, zda lze se spolehlivostí 0,95, resp. 0,99, tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

d) Načrtněte příslušnou operativní charakteristiku.

Řešení 5.1

Shrnutí:

Pojmem testování statistických hypotéz označujeme rozhodování o pravdivosti **parametrických**, resp. **neparametrických hypotéz** o populaci. V tomto rozhodovacím procesu proti sobě stojí **nulová a alternativní hypotéza**. Naším cílem je rozhodnout, zda data z výběrového souboru $\boldsymbol{X}$ odpovídají nulové hypotéze.

Jelikož při rozhodování o nulové hypotéze vycházíme z výběrového souboru, který nemusí dostatečně přesně odpovídat vlastnostem základního souboru, můžeme se při rozhodování dopustit chyby. Při rozhodování mohou nastat situace, které popisuje Tab. 10.1, kterou zde pro přehlednost uvádíme znovu.

		Výsledek testu	
		Nezamítáme H_0	Zamítáme H_0
Skutečnost	Platí H_0	Správné rozhodnutí $1-\alpha$ (spolehlivost testu)	**Chyba I. druhu** α (hladina významnosti)
	Platí H_A	**Chyba II. druhu** β	Správné rozhodnutí $1-\beta$ (síla testu)

Pravděpodobnosti α a β, s nimiž chyby I. a II. druhu nastávají, rozhodují o kvalitě testu. Ve statistice se volí jako rozhodující vstupní parametr testu pravděpodobnost chyby I. druhu – hladina významnosti α. Chybu II. druhu β snižujeme volbou vhodného testu (pokud máme možnost výběru) popřípadě zvětšením rozsahu výběrového souboru.

Závislost pravděpodobnosti chyby II. druhu β na přesné specifikaci alternativní hypotézy je graficky interpretována **operativní charakteristikou**. Operativní charakteristika bývá v praxi taktéž nahrazována **křivkou síly testu**, což je graf závislosti síly testu $1-\beta$ na přesné specifikaci alternativní hypotézy.

Při testování hypotéz se běžně můžeme setkat se dvěma přístupy – klasickým testem a čistým testem významnosti.

Klasický test se skládá z několika kroků:

1. *Formulace nulové a alternativní hypotézy.*
2. *Volba testové statistiky (testového kritéria)* $T(X)$, tj. výběrové charakteristiky, která má vztah k nulové hypotéze. Je přitom nutné, abychom znali rozdělení $T(X)$ v případě platnosti nulové hypotézy.
3. *Sestrojení kritického oboru W a oboru přijetí V.* Kritický obor W přitom odpovídá hodnotám testového kritéria, které v případě platnosti nulové hypotézy nastávají s nízkou pravděpodobností α.
4. *Výpočet pozorované hodnoty testové statistiky* $T(X)$ značené x_{OBS}.
5. *Formulace závěru testu*- buď nulovou hypotézu zamítáme ve prospěch alternativy, nebo nulovou hypotézu nezamítáme.

Na rozdíl od klasického testu nemusíme pro čistý test významnosti znát hladinu významnosti α jako vstupní údaj. Jeho výsledek, *p-hodnota*, nám umožňuje rozhodnout, na jakých hladinách významnosti můžeme nulovou hypotézu zamítnout (resp. nezamítnout).

Čistý test významnosti se skládá z následujících kroků:

1. *Formulace nulové a alternativní hypotézy.*
2. *Volba testové statistiky (testového kritéria)* $T(X)$.
3. *Výpočet pozorované hodnoty testové statistiky* $T(X)$ značené x_{OBS}.
4. *Výpočet p-hodnoty.*
   p-hodnota je tedy nejnižší hladina významnosti, na níž můžeme nulovou hypotézu zamítnout a zároveň nejvyšší hladiny významnosti, na níž se již nulová hypotéza nezamítá. $P-hodnotu$ vypočteme podle jedné ze tří možných definic v závislosti na tvaru alternativní hypotézy. Je přitom nutné, aby alternativní hypotéza korespondovala s výběrovým souborem.

Tvar alternativní hypotézy H_A	***p-hodnota***
$\theta < \theta_0$	$p\text{-hodnota} = F_0(x_{OBS})$
$\theta > \theta_0$	$p\text{-hodnota} = 1 - F_0(x_{OBS})$
$\theta \neq \theta_0$	$p\text{-hodnota} = 2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$

5. *Rozhodnutí na základě p-hodnoty.* Rozhodujeme-li o správnosti nulové hypotézy se spolehlivostí $1-\alpha$, tj. na hladině významnosti α, pak je rozhodnutí dáno tabulkou 10. 2.

p-hodnota	**Rozhodnutí**
$p\text{-hodnota} < \alpha$	Zamítáme H_0 ve prospěch H_A.
$p\text{-hodnota} > \alpha$	Nezamítáme H_0.

V následujících kapitolách budeme pro rozhodování o statistických hypotézách používat výhradně čistý test významnosti.

Test

Jak pracovat s testy?

Zacátek testu

1. (11b.) Doplňte

(a) Statistická hypotéza je výrok o ______.

(b) Rozhodovací proces, který používáme k učinění závěrů o rozdělení náhodné veličiny na základě výběrového souboru a hypotéz se nazývá ______.

(c) Při testování hypotéz se rozhodujeme mezi ______ hypotézou.

(d) Obor hodnot testové statistiky (testového kritéria) lze rozdělit na dvě disjunktní množiny nazývané ______.

(e) Kritický obor se stanovuje tak, aby pravděpodobnost, že hodnota testové statistiky padne do kritického oboru byla v případě platnosti nulové hypotézy rovna ______.

(f) Pravděpodobnost chyby I. druhu i chyby II. druhu lze snížit, zvýšíme-li ______.

(g) Graf závislosti pravděpodobnosti chyby II. druhu β na konkrétní specifikaci alternativní hypotézy je nazýván ______.

(h) Přístup k testování hypotéz, který je založen na rozhodování pomocí kritického oboru bývá nazýván $\boxed{}$.

(i) Přístup k testování hypotéz, který je založen na rozhodování pomocí *p-hodnoty* bývá nazýván $\boxed{}$.

(j) Při testování hypotéz je možno učinit dvě rozhodnutí - $\boxed{}$.

(k) Je-li *p-hodnota* $= 0,03$, pak $\boxed{}$ nulovou hypotézu se spolehlivostí 0,95.

2. (1b.) Nezamítli-li jsme nulovou hypotézu, přestože ve skutečnosti neplatí, dopustili jsme se

(a) chyby I. druhu,
(b) chyby II. druhu.

3. (1b.) Zamítli-li jsme nulovou hypotézu, přestože ve skutečnosti platí, dopustili jsme se

(a) chyby I. druhu,
(b) chyby II. druhu.

4. (1b.) Chceme-li snížit pravděpodobnost chyby I. druhu na nejmenší možnou míru, jakou hodnotu hladiny významnosti je třeba volit?

(a) co nejmenší,

(b) co největší,
(c) volba hladiny významnosti nemá na pravděpodobnost chyby I. druhu vliv.

5. (1b.) Pravděpodobnost, že se při testování hypotéz dopustíme chyby I. druhu, se nazývá

(a) hladina významnosti,
(b) síla testu,
(c) spolehlivost testu.

6. (1b.) Pravděpodobnost, že se při testování hypotéz nedopustíme chyby II. druhu, se nazývá

(a) hladina významnosti,
(b) síla testu,
(c) spolehlivost testu.

7. (1b.) Leží-li pozorovaná hodnota testového kritéria v kritickém oboru, pak na dané hladině významnosti

(a) zamítáme nulovou hypotézu,
(b) nezamítáme nulovou hypotézu,
(c) přijímáme nulovou hypotézu,
(d) nedokážeme o výsledku testu rozhodnout.

8. (1b.) Neleží-li pozorovaná hodnota testového kritéria v kritickém oboru, pak na dané hladině významnosti
   (a) zamítáme nulovou hypotézu,
   (b) nezamítáme nulovou hypotézu,
   (c) přijímáme nulovou hypotézu,
   (d) nedokážeme o výsledku testu rozhodnout.

9. (1b.) Je-li $p - hodnota = 0,03$, pak na dané hladině významnosti 0,05
   (a) zamítáme nulovou hypotézu,
   (b) nezamítáme nulovou hypotézu,
   (c) přijímáme nulovou hypotézu,
   (d) nedokážeme o výsledku testu rozhodnout.

10. (1b.) Je-li $p - hodnota = 0,03$, pak na dané hladině významnosti 0,01
   (a) zamítáme nulovou hypotézu,
   (b) nezamítáme nulovou hypotézu,
   (c) přijímáme nulovou hypotézu,
   (d) nedokážeme o výsledku testu rozhodnout.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Kapitola 6

Jednovýběrové testy parametrických hypotéz

Cíle

Po prostudování tohoto odstavce budete umět testovat hypotézy

- o rozptylu a střední hodnotě normálního rozdělení,
- o mediánu (neparametrické testy o střední hodnotě),
- o parametru π alternativního rozdělení.

Jak již bylo uvedeno, hypotézy a jím příslušné testy dělíme podle počtu šetřených populací na jednovýběrové, dvouvýběrové a vícevýběrové. V této kapitole uvedeme často používané jednovýběrové testy parametrických hypotéz, tj. testy o parametrech jedné populace. Pro každý test budou popsány situace, v nichž se test používá, nulová a alternativní hypotéza a testové kritérium $T(X)$ včetně jejího nulového rozdělení. Při testování se zaměříme téměř výhradně na čistý test významnosti, tj. na testování s využitím p-hodnoty. (Postup uplatňovaný při čistém testu významnosti si můžete připomenout v kapitole 10.2.2.)

Poznámka: *V řešených příkladech byl pro výpočet p-hodnoty použit výpočetní applet vybrana_rozdeleni.xlsx, který je přílohou této učebnice.*

Častou statistickou úlohou je rozhodnout, zda neznámý parametr rozdělení populace (nejčastěji střední hodnota, rozptyl nebo relativní četnost) je roven nějaké konkrétní číselné hodnotě, resp. zda je neznámý parametr rozdělení populace větší či menší než nějaká konkrétní číselná hodnota. Rozhodovací proces, který je pro řešení těchto úloh používán, bývá označován jako jednovýběrový test. Testy o parametrech populace dělíme na

- parametrické,
- neparametrické (robustní).

Za parametrické označujeme testy, které předpokládají konkrétní rozdělení populace (nejčastěji normální rozdělení). Testy, které nepředpokládají konkrétní rozdělení populace, se nazývají neparametrické. Neparametrické testy se užívají zejména k analýze údajů, které nevyhovují požadavkům na rozdělení v parametrických testech, například jednovýběrovém, dvouvýběrovém, resp. párovém t testu.

6.1. Test o rozptylu normálního rozdělení

Předpokládejme, že máme normálně rozdělenou populaci se střední hodnotou μ a rozptylem σ^2 a žádný z parametrů μ, σ^2 neznáme. Na základě výběru $X_1, X_2, \ldots, X_n$ z dané populace chceme ověřit předpoklad, zda rozptyl populace σ^2 se rovná hodnotě σ_0^2.

Neznámý rozptyl σ^2 odhadneme výběrovým rozptylem s^2, který určíme z pozorovaných výběrových hodnot $x_1, x_2, \ldots, x_n$. Je zřejmé, že vypočtená a předpokládaná hodnota rozptylu (s^2 a σ_0^2) se mohou od sebe lišit. Rozdíl může být pouze nevýznamný a lze ho přičíst účinku náhodných vlivů, působících při výběru. Tento rozdíl však může být i nenáhodný (říkáme také **statisticky významný** nebo signifikantní). Test o rozptylu tak představuje ověření, zda se výběrový rozptyl s^2 a předpokládaný rozptyl σ_0^2 liší statisticky významně nebo pouze náhodně.

Nulovou hypotézu H_0 zvolíme ve tvaru $\sigma^2 = \sigma_0^2$. Zatímco volba nulové hypotézy je zřejmá, u alternativy H_A můžeme volit ze tří možností: $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$, $\sigma^2 \neq \sigma_0^2$. Jako testové kritérium použijeme výběrovou charakteristiku

$$T(X) = \frac{s^2}{\sigma^2}(n-1),$$

která má v případě platnosti nulové hypotézy χ^2 - rozdělení s $n-1$ stupni volnosti (kapitola 3.8.1). Dále pak pokračujeme podle obecného schématu čistého testu významnosti, tj. určíme pozorovanou hodnotu x_{OBS}, na základě tvaru alternativní hypotézy vypočteme *p-hodnotu* a pokud je *p-hodnota* menší než hladina významnosti α, zamítneme nulovou hypotézu. Všechny tři varianty testu o rozptylu, včetně předpokladu testu, jsou uvedeny v tabulce 6.1.

Dále popisované testy pak budou ve stručnosti uváděny pomocí obdobných tabulek.

Tab. 6.1: Test o rozptylu

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{S^2}{\sigma^2}(n-1)$	χ^2_{n-1}	$F_0(x_{OBS})$
	$\sigma^2 > \sigma_0^2$			$1 - F_0(x_{OBS})$
	$\sigma^2 \neq \sigma_0^2$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$
Předpoklad testu: Populace má normální rozdělení s neznámou střední hodnotou.				

Příklad 6.1. Hmotnost kulečníkové koule lze pokládat za náhodnou veličinu s rozdělením $N(\mu, \sigma^2)$. Hodnotíme-li kvalitu sady kulečníkových koulí, nezáleží ani tak na tom, kolik přesně jednotlivé koule váží, jako na tom, aby byly stejně těžké. Za kvalitní se považují koule, jejichž směrodatná odchylka hmotnosti nepřekračuje 2 gramy. Při zkoušce deseti náhodně vybraných koulí značky KULKOUL byly zjištěny následující hodnoty jejich hmotnosti $[g]$:

170 176 168 170 173 169 168 170 170 170

Ověřte, zda lze koule značky KULKOUL považovat za kvalitní.

Řešení 6.1

Obsah

249. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

6.2. Testy o střední hodnotě normálního rozdělení

Předpokládejme, že máme normálně rozdělenou populaci se střední hodnotou μ a rozptylem σ^2. Předpokládejme, že parametr μ neznáme. Na základě výběru X_1, X_2 až X_n chceme ověřit předpoklad, že se střední hodnota (populační průměr) μ rovná určité hodnotě μ_0.

Nejlepším bodovým odhadem neznámé střední hodnoty je výběrový průměr $\bar{x}$. Jde nám o ověření, zda se výběrový průměr ($\bar{x}$) a populační průměr (střední hodnota μ_0) liší statisticky významně nebo zda lze jejich rozdíl přisoudit náhodným vlivům. Testujeme nulovou hypotézu H_0: $\mu = \mu_0$ vůči alternativě $\mu < \mu_0$, $\mu > \mu_0$ nebo $\mu \neq \mu_0$. Volba testového kritéria závisí na tom, zda známe populační rozptyl σ^2.

6.2.1. Jednovýběrový z test

Má-li populace normální rozdělení o známém rozptylu σ^2, používáme tzv. **jednovýběrový z test**. Tento test (viz tab. 6.2) uvádíme pouze pro zajímavost - v praxi se obvykle nesetkáváme se situaci, kdy bychom znali rozptyl populace a neznali její střední hodnotu.

Tab. 6.2: Jednovýběrový z test

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$\mu = \mu_0$	$\mu < \mu_0$	$\dfrac{\bar{X} - \mu}{\sigma}\sqrt{n}$	$N(0;1)$ (viz kap. 3.4.2)	$F_0(x_{OBS})$
	$\mu > \mu_0$			$1 - F_0(x_{OBS})$
	$\mu \neq \mu_0$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$

Předpoklad testu: Populace má normální rozdělení se známým rozptylem σ^2.

6.2.2. Jednovýběrový t test

Máme-li normálně rozdělenou populaci s neznámou střední hodnotou μ a neznámým rozptylem σ^2, použijeme k ověření předpokladu, že se střední hodnota (populační průměr) μ rovná určité hodnotě μ_0 jednovýběrový t test.

Tab. 6.3: Jednovýběrový t test

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$\mu = \mu_0$	$\mu < \mu_0$	$\dfrac{\bar{X} - \mu}{S}\sqrt{n}$	t_{n-1} (viz kap. 3.9.1)	$F_0(x_{OBS})$
	$\mu > \mu_0$			$1 - F_0(x_{OBS})$
	$\mu \neq \mu_0$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$

Předpoklad testu: Populace má normální rozdělení s neznámým rozptylem.

Poznámka:
Jednovýběrový t test můžeme použít pouze v případě, má-li populace má normální rozdělení s neznámým rozptylem. V případě výrazné nenormality dáváme před t testem přednost některému z neparametrických testů, nejčastěji ***mediánovému testu*** *(kapitola 11.3) nebo* ***jednovýběrovému Wilcoxonovu testu*** *(kapitola 11.4).*

Příklad 6.2. Inteligenční kvocient (IQ) popisuje inteligenci jednotlivce v poměru k ostatní populaci, přičemž za střední hodnotu se považuje IQ 100 bodů. Je známo, že IQ má normální rozdělení. Při testu inteligence, kterého se zúčastnilo 10 náhodně vybraných studentů posledního ročníku výběrové školy ASNEM, byly naměřeny následující hodnoty IQ.

65	98	103	77	93	102	102	113	80	94

Ověřte čistým testem významnosti hypotézu, že na škole ASNEM je střední hodnota IQ studentů závěrečného ročníku školy ASNEM podprůměrná.

Řešení 6.2

6.3. Kvantilový test

Kvantilový test umožňuje na základě výběru $X_1, X_2, \ldots, X_n$ ověřit předpoklad, že se $100p\%$ kvantil x_p rovná určité hodnotě x_{p_0}. Tento test patří do skupiny neparametrických testů, tj. testů, které nepředpokládají určité rozdělení populace. Používáme jej zejména jako mediánový test v případech, kdy chceme testovat střední hodnotu populace, která má výrazně zešikmené rozdělení. Jelikož tento test má malou sílu (pravděpodobnost chyby II. druhu je velká ve srovnání s jinými testy), je vhodné mít k dispozici výběr o větším rozsahu.

V kvantilovém testu vycházíme z nulové hypotézy, že $100p\%$ kvantil spojité náhodné veličiny X je roven konstantě x_{p0}, tj. $x_p = x_{p0}$. Při volbě alternativní hypotézy máme tři možnosti: $x_p < x_{p0}$, $x_p > x_{p0}$, $x_p \neq x_{p0}$.

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$. Nechť náhodná veličina Y modeluje počet pozorování v náhodném výběru, u nichž je pozorovaná hodnota náhodné veličiny X menší než testována hodnota x_{p0}, tj. $x < x_{p0}$.

Je zřejmé, že platí-li nulová hypotéza, pak pravděpodobnost, že nějaké pozorování bude menší než x_{p0} je p. Počet pozorování v náhodném výběru, která jsou menší než x_{p0}, má proto, za předpokladu platnosti nulové hypotézy, binomické rozdělení $Bi(n; p)$.

Tab. 6.4: Kvantilový test

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$x_p = x_{p_0}$	$x_p < x_{p_0}$	Y, kde Y … počet pozorování v náhodném výběru, u nichž je $x < x_{p_0}$	$Bi(n;p)$	$F_0(x_{OBS})$
	$x_p > x_{p_0}$			$1 - F_0(x_{OBS})$
	$x_p \neq x_{p_0}$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$
Předpoklad testu: ---				

Poznámka: *V případě, že testujeme medián, tzn. pro* $p = 0,5$*, používáme pro tento test speciální označení -* ***mediánový test****. Mediánový test je alternativou jednovýběrového t testu v situaci, kdy nelze předpokládat normální rozdělení populace. V případě, že hodnoty analyzované náhodné veličiny* X *jsou rozdíly párových pozorování, užíváme pro mediánový test název* ***znaménkový test****.*

6.4. Jednovýběrový Wilcoxonův test

Dalším příkladem neparametrického testu je Wilcoxonův test. Mějme náhodný výběr X_1 až X_n ze spojitého rozdělení s hustotou f, která je symetrická kolem bodu a. Z toho plyne, že a musí být rovno mediánu $x_{0,5}$. Jednovýběrový Wilcoxonův test je určen k testování hypotézy $x_{0,5} = x_{0,5_0}$. Při volbě alternativní hypotézy máme opět tři možnosti: $x_{0,5} < x_{0,5_0}$, $x_{0,5} > x_{0,5_0}$, $x_{0,5} \neq x_{0,5_0}$.

Je-li některá z veličin $X_1, X_2, \ldots, X_n$ rovna testované hodnotě $x_{0,5_0}$, obvykle toto pozorování z výběrového souboru vypustíme. Položme $Y_i = X_i - x_{0,5_0}$, $i = 1, 2, \ldots, n$. Veličiny Y_i seřaďme vzestupně podle jejich absolutní hodnoty.

$$|Y_{(1)}| \leqq |Y_{(2)}| \leqq \ldots \leqq |Y_{(n)}|$$

Označme R_i^+ pořadí veličiny $|Y_{(i)}|$. Nechť

$$S^+ = \sum_{Y_i \geqq 0} R_i^+, \qquad S^- = \sum_{Y_i < 0} R_i^+.$$

Testové kritérium má tvar

$$T(X) = min(S^+; S^-).$$

Je-li alternativní hypotéza ve tvaru $x_{0,5} \neq x_{0,5_0}$, pak, dle klasického testu, nulovou hypotézu zamítneme na hladině významnosti α v případě, že pozorovaná hodnota testového kritéria je menší nebo rovna tabelované hodnotě $\omega_n\alpha$ (tabulka T6). Pro testování pak používáme klasický test, který je popsán v tabulce 6.5.

Tab. 6.5: Wilcoxonův test

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Kritický obor
$x_{0,5} = x_{0,5_0}$	$x_{0,5} \neq x_{0,5_0}$	$T(\boldsymbol{X}) = min(S^+; S^-)$ (viz výše)	$(0; \omega_n\alpha)$, kde $\omega_n\alpha$ najdete v tabulce T6
Předpoklad testu: symetrie hustoty *f* kolem mediánu			

Máme-li k dispozici výběr o dostatečně velkém rozsahu, využijeme toho, že S^+ má asymptoticky normální rozdělení s parametry

$$E(S^+) = \frac{1}{4}n(n+1), \qquad D(S^+) = \frac{1}{24}n(n+1)(2n+1).$$

Testové kritérium pak má tvar

$$T(X) = \frac{S^+ - E(S+)}{\sqrt{D(S^+)}}$$

a při platnosti nulové hypotézy má normované normální rozdělení $N(0; 1)$.

Tab. 6.6: Wilcoxonův test pro $n > 30$

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$x_{0,5} = x_{0,5_0}$	$x_{0,5} < x_{0,5_0}$	$\frac{S^+ - E(S^+)}{\sqrt{D(S^+)}}$ (viz výše)	$N(0;1)$	$F_0(x_{OBS})$
	$x_{0,5} > x_{0,5_0}$			$1 - F_0(x_{OBS})$
	$x_{0,5} \neq x_{0,5_0}$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$
Předpoklad testu: symetrie hustoty f kolem mediánu				

Poznámka: *Připomeňme, že předpokladem jednovýběrového Wilcoxonova testu je symetrie hustoty f kolem mediánu. K zamítnutí H_0 tak může dojít i tehdy je-li median roven $x_{0,5_0}$, ale hustota f je výrazně asymetrická.*

Příklad 6.3. U 10 náhodně vybraných osob byly zjištěny následující doby čekání [den] na preventivní prohlídku u paní zubařky Hrozné.

65 98 103 77 93 102 102 113 80 94

Paní zubařka Hrozná tvrdí, že polovina pacientů čeká na provedení preventivní prohlídky méně než 90 dnů od objednání. Ověřte čistým testem významnosti tvrzení paní zubařky Hrozné.

Řešení 6.3

Obsah
256. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

6.4.1. Test o parametru π alternativního rozdělení

Předpokládejme, že v sérii n nezávislých opakování pokusu se nějaký náhodný jev A, který má stálou, ale neznámou pravděpodobnost π, vyskytl X-krát. Náhodný výběr $X_1, \ldots, X_n$ lze považovat za výběr z alternativního rozdělení $A(\pi)$. Počet výskytu jevu A v takovéto skupině n opakování pokusu (náhodnou veličinu X) lze považovat za náhodnou veličinu s binomickým rozdělením $Bi(n; \pi)$. Na základě těchto údajů chceme ověřit předpoklad, že parametr π se rovná určité hodnotě π_0.

Neznámou pravděpodobnost π odhadujeme výběrovou relativní četností p výskytu jevu A, tzn. podílem X/n. Jde nám o ověření, zda se pozorovaná relativní četnost (p) a předpokládaná pravděpodobnost (π_0) liší statisticky významně nebo zda lze jejich rozdíl přisoudit náhodným vlivům. Pro provedení tohoto testu musíme mít k dispozici výběr o dostatečném rozsahu n, tj. $n > \dfrac{9}{p(1-p)}$.

Tab. 6.7: Test o parametru π alternativního rozdělení

Nulová hypotéza H_0	Alternativní hypotéza H_A	Testové kritérium $T(\boldsymbol{X})$	Nulové rozdělení	p-hodnota
$\pi = \pi_0$	$\pi < \pi_0$	$\dfrac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$	$N(0;1)$ (viz kap. 3.5)	$F_0(x_{OBS})$
	$\pi > \pi_0$			$1 - F_0(x_{OBS})$
	$\pi \neq \pi_0$			$2min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$
Předpoklad testu: $n > \frac{9}{p(1-p)}$				

Obsah
257. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Příklad 6.4. U 100 pojištěných aut bylo zjištěno, že 18 aut je starších než 7 let. Podle předpokladů a odhadů pojišťovny nemá podíl aut starších 7 let překračovat 25%. Ověřte, zda je podíl aut starších než 7 let skutečně nižší než 25%.

Řešení 6.4

Shrnutí:

Obvyklou statistickou úlohou je rozhodnout, zda neznámý parametr rozdělení populace (nejčastěji střední hodnota, rozptyl nebo relativní četnost) je roven nějaké konkrétní číselné hodnotě, resp. zda je neznámý parametr rozdělení populace větší či menší než nějaká konkrétní číselná hodnota. Rozhodovací proces, který je pro řešení těchto úloh používán je označován jako **jednovýběrový test** (parametrické hypotézy). Testy vyžadující znalost rozdělení populace označujeme jako **parametrické**. K analýze údajů, které nevyhovují požadavkům na rozdělení v parametrických testech, například v jednovýběrovém t testu, používáme testy **neparametrické**. Slabší předpoklady, které k neparametrickým testům neodmyslitelně patří, způsobují, že tyto testy nejsou tak silné, jako jejich parametrické protějšky.

Připomeňte si, že více informací než samotný test poskytují intervalové odhady populačních parametrů, které určují meze intervalu, v němž se populační parametry nacházejí s pravděpodobností $1-\alpha$ (obvykle $1-\alpha=0,95$).

Stručný přehled jednovýběrových testů, s nimiž jsme se seznámili

Jednovýběrové parametrické testy

Název testu	Testovaný parametr	Předpoklady testu	Testová statistika $T(X)$	Nulové rozdělení	Poznámka
Test o rozptylu	rozptyl σ^2 (směrodatná odchylka σ)	normalita populace, neznámé μ	$\frac{S^2}{\sigma^2}(n-1)$	χ^2_{n-1}	Při čistém testu významnosti nelze použít oboustrannou alternativu.
Jednovýběrový z test	střední hodnota μ	normalita populace, známé σ^2	$\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$	$N(0;1)$	
Jednovýběrový t test		normalita populace, neznámé σ^2	$\frac{\bar{X}-\mu}{S}\sqrt{n}$	t_{n-1}	

Test o rozptylu se používá k testování nulové hypotézy, která říká, že populační rozptyl *normálního* rozdělení je roven zadané hodnotě. Test tedy odpovídá na otázku, zda na základě náhodného výběru můžeme tvrdit, že se (neznámý) populační rozptyl rovná zadanému číslu (resp. zda je menší nebo větší než zadané číslo).

Pokud je *p-hodnota* menší než zvolená hladina významnosti α (obvykle 0,05), nulová hypotéza se zamítá a přikláníme se k alternativě. Znamená to, že rozdíl mezi zadanou hodnotou a rozptylem výběrového souboru je příliš velký na to, aby mohl být důsledkem náhodného výběru, je **statisticky významný**. Je-li *p-hodnota* větší než zvolená hladina významnosti, nulová hypotéza se nezamítá. Znamená to, že rozdíl mezi zadanou hodnotou a rozptylem výběrového souboru může být důsledkem náhodného výběru, je **statisticky nevýznamný**.

Jednovýběrové neparametrické testy

Název testu	Testovaný parametr	Předpoklady testu	Testová statistika $T(X)$	Nulové rozdělení	Poznámka
Test o parametru π alternativního rozdělení	Pravděpodobnost π	$n > \frac{9}{p(1-p)}$	$\frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$	$N(0;1)$	
Kvantilový test	$100p\%$ kvantil x_p		Y, kde Y modeluje počet pozorování v náhodném výběru, která jsou menší než x_{p_0}.	$Bi(n;p)$	V případě, že testujeme medián, tzn. pro $p = 0{,}5$, používáme pro tento test speciální označení - **mediánový test.**
Jednovýběrový Wilcoxonův test	medián $x_{0,5}$		$min(S^+;S^-)$, kde $S^+ = \sum_{Y_i \geq 0} R_i^+$, $S^- = \sum_{Y_i < 0} R_i^+$	Kritické hodnoty jsou tabelovány (Tab. T6)	Je-li pozorovaná hodnota testové statistiky menší nebo rovna kritické hodnotě, zamítáme H_0.
		$n > 30$	$\frac{S^+ - E(S^+)}{\sqrt{D(S^+)}}$, kde $E(S^+) = \frac{1}{4}n(n+1)$, $D(S^+) = \frac{1}{24}n(n+1)(2n+1)$	$N(0;1)$	

Jednovýběrový z **test** se používá k testování nulové hypotézy, která říká, že střední hodnota *normálního* rozdělení se *známým rozptylem* je rovna zadané hodnotě. Test tedy odpovídá na otázku, zda na základě náhodného výběru můžeme tvrdit, že se (neznámá) střední hodnota rovná zadanému číslu (resp. zda je menší nebo větší než zadané číslo). V praxi se se situací, kdy známe populační rozptyl a přitom neznáme střední hodnotu (populační průměr) setkáváme výjimečně. Mnohem častěji potřebujeme ověřit hypotézu o střední hodnotě *normálního* rozdělení s *neznámým rozptylem*. V této situaci používáme **jednovýběrový** t **test**. Jednovýběrový t test předpokládá normální rozdělení populace. Pokud je rozsah výběru malý a testy normality (budou uvedeny později) zamítnou normalitu, musíme použít neparametrické alternativy jednovýběrového t testu: **mediánový test**, popř. **Wilcoxonův test**, které testují nulovou hypotézu o shodě mediánu s konstantou.

Testujeme-li hypotézu, že pravděpodobnost výskytu určitého jevu v populaci je rovna nějakému číslu, použijeme **test o parametru π alternativního rozdělení**. Předpokladem pro použití tohoto testu je náhodný výběr dostatečného rozsahu.

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Zamítneme-li na základě t-testu nulovou hypotézu, pak lze tvrdit, že rozdíl mezi testovanou hodnotou a průměrem výběrového souboru
   (a) je na dané hladině významnosti statisticky významný,
   (b) není na dané hladině významnosti statisticky významný.

2. (2b.) Označte všechny parametrické testy, tj. testy vyžadující znalost rozdělení populace
   (a) test o střední hodnotě (t-test),
   (b) test o rozptylu (F-test),
   (c) mediánový test nebo Wilcoxonův test,
   (d) test o parametru π alternativního rozdělení.

3. (1b.) Neparametrické testy
   (a) nevyžadují splnění žádných předpokladů,
   (b) nevyžadují znalost rozdělení populace,
   (c) vyžadují znalost rozdělení populace.

4. (1b.) Neparametrické testy mají

(a) větší sílu testu než jejich parametrické protějšky,
(b) menší sílu testu než jejich parametrické protějšky.

5. (1b.) Předpokladem pro použití testu o parametru π alternativního rozdělení je

(a) normalita výběru,
(b) výběr ze spojitého rozdělení,
(c) dostatečný rozsah výběru ($n > \frac{9}{p(1-p)}$, kde p je relativní četnost výskytu sledovaného jevu).

6. (1b.) Předpokladem pro použití Wilcoxonova testu je

(a) normalita výběru,
(b) výběr ze spojitého rozdělení,
(c) dostatečný rozsah výběru ($n > \frac{9}{p(1-p)}$, kde p je relativní četnost výskytu sledovaného jevu).

7. (1b.) Chceme-li ověřit, zda lze výrobcem udávanou spotřebu 8,8 l/100km považovat za pravdivou (bylo testováno 11 automobilů, normalita výběru byla zamítnuta), použijeme

(a) test o střední hodnotě (t-test),
(b) test o rozptylu (F-test),
(c) mediánový test nebo Wilcoxonův test,
(d) test o parametru π alternativního rozdělení.

8. (1b.) Chceme-li ověřit, zda lze očekávat, že v prodejně je více než 5 % konzerv s prošlou záruční lhůtou (v kontrolním vzorku 100 konzerv bylo nalezeno 7 konzerv s prošlou záruční lhůtou), použijeme

(a) test o střední hodnotě (t-test),
(b) test o rozptylu (F-test),
(c) mediánový test nebo Wilcoxonův test,
(d) test o parametru π alternativního rozdělení.

9. (1b.) Chceme-li ověřit, zda je průměrná výška dospělé populace v ČR větší než 170 cm (rozsah výběru je 120, byla ověřena normalita výběru), použijeme

(a) test o střední hodnotě (t-test),
(b) test o rozptylu (F-test),
(c) mediánový test nebo Wilcoxonův test,
(d) test o parametru π alternativního rozdělení.

10. (1b.) Pro bavlněnou přízi je předepsaná horní mez variability pevnosti vlákna. Rozptyl pevnosti (která má normální rozdělení) nemá překročit 0,36. Chceme-li ověřit, zda je důvod k podezření na vyšší variabilitu než je stanoveno, použijeme

(a) test o střední hodnotě (t-test),
(b) test o rozptylu (F-test),
(c) mediánový test nebo Wilcoxonův test,

(d) test o parametru π alternativního rozdělení.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1. Firma FRIDGER pravidelně přijímá dodávky chladících jednotek pro své chladničky a za posledních 18 měsíců pouze 2% jednotek nedosahovaly požadovaných parametrů. Dodavatel však přešel na novou technologii a firma FRIDGER se obává možného zhoršení dodávek. Proto bylo náhodně vybráno 500 jednotek z následující dodávky a zjištěno, že 21 jednotek nesplňuje požadované parametry.

   a) Ověřte pomocí 95% intervalu spolehlivosti, zda došlo k zhoršení kvality

   b) Ověřte pomocí čistého testu významnosti, zda došlo k zhoršení kvality (na 5% hladině významnosti)

   c) Načrtněte křivku síly testu pro tento případ.

   Řešení příkladu

2. Výrobní proces produkuje milióny žárovek se střední životností 14 000 hodin. Novou technologií byl vyroben vzorek 25 žárovek s průměrnou životností 14 740 hodin a směrodatnou odchylkou 2 000 hodin. Ověřte čistým testem významnosti, zda nová technologie vedla ke zvýšení životnosti žárovek. (Předpokládejte, že životnost žárovek má normální rozdělení.)

   Řešení příkladu

3. Majitel rybníka ví z dlouhodobých záznamů, že střední váha kaprů z tohoto rybníka je 1,97 kg. V loňském roce majitel zkoušel nový způsob krmení ryb. Při minulém výlovu byla průměrná váha sta kaprů 1,99 kg se směrodatnou odchylkou 0,21 kg. Ověřte čistým testem významnosti, zda se při novém způsobu krmení:

   a) váha kaprů změnila

   b) váha kaprů zvýšila

Obsah

267. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Předpokládejte, že váha kaprů má normální rozdělení.

Řešení příkladu

4. U standardně vyráběného materiálu má mez pevnosti R_m lognormální rozdělení se střední hodnotou 640,0 MPa. Změnou posloupnosti tepelných úprav byl připraven nový materiál (předpokládáme stejný rozptyl), pro nějž bylo naměřeno R_m u deseti vzorků postupně

   651, 639, 645, 648, 650, 643, 652, 640, 644, 645.

   Ověřte, zda došlo po změně posloupnosti tepelných úprav ke zvýšení střední meze pevnosti.

   Řešení příkladu

5. Firma TT udává, že 1% jejich rezistorů nesplňuje požadovaná kritéria. V testované dodávce 1000ks bylo nalezeno 15 nevyhovujících rezistorů. Potvrzuje tento výsledek tvrzení TT? Ověřte čistým testem významnosti.

   Řešení příkladu

6. Výrobce garantuje, že jím vyrobené žárovky mají životnost v průměru 1.000 hodin. Aby útvar kontroly zjistil, zda tomuto konstatování odpovídá i v daném období vyrobená a expedovaná část produkce, vybral z připravené dodávky náhodně 50 žárovek a došel k závěru, že průměrná doba životnosti je 950 hodin a směrodatná odchylka doby životnosti pak 100 hodin. Je možné zjištěný rozdíl doby životnosti ve výběru připsat náhodě nebo je známkou nekvality produkce? Ověřte čistým testem významnosti. Předpokládejte, že životnost žárovek má normální rozdělení.

   Řešení příkladu

Obsah
268. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

7. Představenstvo velké akciové společnosti zvažuje odprodat část akcií zaměstnancům této společnosti. Odhaduje se, že zájem o nákup by mohlo projevit asi 20% z nich. Proto personální útvar připravil předběžný průzkum, v němž oslovil 400 náhodně vybraných pracovníků společnosti, z nichž zájem o nákup akcií projevilo 66 lidí. Je úvaha představenstva reálná? Ověřte čistým testem významnosti.

   Řešení příkladu

8. Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05mm. K ověření této informace bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04mm. Lze tento rozdíl považovat za významný ve smyslu zlepšení kvality produkce? Ověřte čistým testem významnosti. Předpokládejte, že průměr pístových kroužků má normální rozdělení.

   Řešení příkladu

9. Při analýze diferenciace mezd ve velkém podniku bylo zjištěno, že průměrná měsíční mzda činila 9.386,-Kč a směrodatná odchylka mezd 1.562,- Kč. Po rozsáhlých organizačních změnách bylo nutné rychle posoudit, zda došlo ke změnám v diferenciaci mezd. Náhodně bylo vybráno 30 pracovníků a byla zjištěna směrodatná odchylka mezd 1.708,-Kč. Je možné na 5% hladině významnosti tvrdit, že organizační změny prohloubily diferenciaci mezd? Předpokládejte, že mzdy mají normální rozdělení.

   Řešení příkladu

Obsah

269. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Kapitola 7

Dvouvýběrové testy parametrických hypotéz

Cíle

Po prostudování této kapitoly budete umět

- testovat hypotézy o shodě rozptylů dvou populací,
- testovat hypotézy o shodě středních hodnot dvou populací,
- testovat hypotézy o shodě mediánů dvou populací,
- testovat hypotézy o homogenitě dvou binomických rozdělení,
- používat párové testy.

Kromě testů o parametrech jedné populace je velmi často potřeba porovnat neznámé parametry dvou populací. V případě, že rozhodovací proces provádíme na základě dvou nezávislých výběrů, používáme tzv. dvouvýběrové testy.

Poznámka: *Nezávislost výběrů bývá v praxi zaručena tím, že každý výběr obsahuje znaky měřené na jiných statistických jednotkách.*

7.1. Test o shodě dvou rozptylů (F-test)

Při výběru testu vhodného pro ověření shody dvou středních hodnot (viz kap. 12. 2) hraje důležitou roli, zda jsou rozptyly srovnávaných populací stejné, či nikoliv. Předpoklad o shodě rozptylů lze na základě náhodných výběrů ověřit testem, který popíšeme v této kapitole.

Mějme dva **nezávislé** výběry $X_1, X_2, \ldots, X_{n_1}$ a $Y_1, Y_2, \ldots, Y_{n_2}$, které pocházejí z populací, které mají rozdělení $N\left(\mu_X; \sigma_X^2\right)$, resp. $N\left(\mu_Y; \sigma_Y^2\right)$. Parametry μ_X, σ_X^2, μ_Y, σ_Y^2 neznáme. Nejlepšími bodovými odhady neznámých rozptylů σ_X^2 a σ_Y^2 jsou výběrové rozptyly

$$S_X^2 = \frac{\sum\limits_{i=1}^{m}\left(X_i - \overline{X}\right)^2}{n_1 - 1} \qquad \text{a} \qquad S_Y^2 = \frac{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n_2 - 1}$$

Nulovou hypotézu formulujeme ve tvaru

$$H_0: \quad \sigma_1^2 = \sigma_2^2 \quad \text{neboli} \quad \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad \left(\sigma_2^2 \neq 0\right)$$

Při volbě alternativy máme tentokrát, podobně jako při testu o rozptylu (kapitola 11.1), pouze dvě možnosti. Oboustranné alternativě se v případě čistého testu významnosti vyhneme, protože definovaný výpočet $p-hodnoty$ pro oboustrannou alternativu je podmíněn tím, že nulové rozdělení testové statistiky je symetrické. Protože testová statistika používaná pro F-test má Fischer-Snedecorovo rozdělení a to není symetrické, není tato podmínka splněna.

$$H_A: \quad \sigma_X^2 < \sigma_Y^2 \quad \text{neboli} \quad \frac{\sigma_X^2}{\sigma_Y^2} < 1, \qquad (1)$$
$$\sigma_X^2 > \sigma_Y^2 \quad \text{neboli} \quad \frac{\sigma_X^2}{\sigma_Y^2} > 1, \qquad (2)$$

Volba vhodné alternativy je dána vztahem mezi výběrovými rozptyly jednotlivých výběrů. Je-li s_X^2 nižší než s_Y^2, volíme alternativu ve tvaru (1). Je-li s_X^2 vyšší než s_Y^2, volíme alternativu ve tvaru (2).

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\frac{s_X^2}{\sigma_X^2}}{\frac{s_Y^2}{\sigma_Y^2}},$$

která má za předpokladu platnosti nulové hypotézy Fisher-Snedecorovo rozdělení s $n_1 - 1$ stupni volnosti pro čitatele a $n_2 - 1$ stupni volnosti pro jmenovatele (kapitola 8.10.1).

Dále pokračujeme podle obecného schématu čistého testu významnosti.

Poznámka: *Pro shodu rozptylu používáme často termín* ***homoskedasticita****, různost rozptylů označujeme jako* ***heteroskedasticitu****.*

7.2. Testy o shodě dvou středních hodnot

Jde o jedny z nejpoužívanějších testů, které na základě porovnání dvou **nezávislých** výběrů umožňují porovnat neznámé střední hodnoty dvou populací.

Mějme dva **nezávislé** výběry $X_1, X_2, \ldots, X_{n_1}$ a $Y_1, Y_2, \ldots, Y_{n_2}$, které pochází z populace mající opět rozdělení $N\left(\mu_X; \sigma_X^2\right)$, resp. $N\left(\mu_Y; \sigma_Y^2\right)$.

Označme jednotlivé výběrové průměry

$$\overline{X} = \frac{\sum\limits_{i=1}^{m} X_i}{n_1}, \qquad \overline{Y} = \frac{\sum\limits_{i=1}^{n} Y_i}{n_2}$$

a výběrové rozptyly

$$S_X^2 = \frac{\sum\limits_{i=1}^{m}\left(X_i - \overline{X}\right)^2}{n_1 - 1} \qquad \text{a} \qquad S_Y^2 = \frac{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n_2 - 1}$$

Při volbě alternativy máme tři možnosti.

$$\begin{aligned} H_A: \quad & \mu_X < \mu_Y && \text{neboli} && \mu_X - \mu_Y < 0, && (1)\\ & \mu_X > \mu_Y && \text{neboli} && \mu_X - \mu_Y > 0, && (2)\\ & \mu_X \neq \mu_Y && \text{neboli} && \mu_X - \mu_Y \neq 0, && (3) \end{aligned}$$

Volba vhodné alternativy bývá v tomto případě dána vztahem mezi průměry jednotlivých výběrů. Je-li $\overline{x}$ výrazně nižší než $\overline{y}$, volíme alternativu ve tvaru (1). Je-li $\overline{x}$ výrazně vyšší než $\overline{y}$, volíme alternativu ve tvaru (2). Nachází-li se $\overline{x}$ v blízkosti $\overline{y}$, volíme alternativu ve tvaru (3).

Jak bylo zmíněno dříve, při výběru testu vhodného pro ověření shody dvou středních hodnot hraje důležitou roli, jaké máme informace o rozptylech populací, z nichž byly náhodné výběry pořízeny. Testové kritérium vybíráme na základě splnění některého ze tří předpokladů.

1) Známe rozptyly obou populací.
2) Rozptyly populací neznáme, ale předpokládáme, že jsou shodné.
3) Rozptyly populací neznáme a nemůžeme předpokládat, že jsou shodné.

7.2.1. Dvouvýběrový z test (známe rozptyly σ_X^2, σ_Y^2)

Známe-li rozptyly σ_X^2, σ_Y^2, použijeme jako testové kritérium statistiku

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}},$$

která má za předpokladu platnosti nulové hypotézy normované normální rozdělení (kapitola 8.6). Dále postupujeme dle čistého testu významnosti. Zdůrazněme, že podobně jako s jednovýběrovým z testem, ani s dvouvýběrovým z testem se v praxi běžně nesetkáváme.

7.2.2. Dvouvýběrový t test (neznáme rozptyly σ_X^2, σ_Y^2; $\sigma_X^2 = \sigma_Y^2$)

Pro porovnání středních hodnot dvou normálních populací s neznámými, avšak shodnými rozptyly používáme **dvouvýběrový t test**. Za testové kritérium volíme statistiku

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n_1-1)s_X^2+(n_2-1)s_Y^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

která má za předpokladu platnosti nulové hypotézy Studentovo rozdělení s $v = n_1 + n_2 - 2$ stupni volnosti. Dále postupujeme dle čistého testu významnosti.

7.2.3. Aspinové-Welchův test (neznáme rozptyly σ_X^2, σ_Y^2; $\sigma_X^2 \neq \sigma_Y^2$)

V případě, že rozptyly normálně rozdělených populací neznáme a nemůžeme předpokládat, že jsou shodné lze použít pro ověření shody středních hodnot například Aspinové-Welchův test (čti „aspinové-velčův“). Za testové kritérium volíme statistiku

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

která má za předpokladu platnosti nulové hypotézy Studentovo rozdělení s v stupni volnosti, kde

$$\nu \doteq \frac{\left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_X^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_Y^2}{n_2}\right)^2} \qquad (\nu \text{ je nutno zaokrouhlit na celé číslo}).$$

Dále postupujeme dle čistého testu významnosti.

Poznámky: *Předpoklad o rovnosti rozptylů můžeme otestovat pomocí F testu. Anděl v [1] uvádí, že se nedoporučuje rozhodovat o tom, zda použít dvouvýběrový t test, nebo nějakou jeho obdobu připouštějící nestejné rozptyly, až podle výsledku F testu. (F test by měl být použit pouze pro ověření předpokladu.)*

Splnění předpokladu nezávislosti náhodných výběrů je velmi podstatné, jeho porušení většinou způsobuje, že výsledky dvouvýběrových testů shody středních hodnot jsou silně zkreslené a nelze je použít. Není-li splněna podmínka nezávislosti náhodných výběrů, lze v případech „spárovaných“ náhodných výběrů použít tzv. párový t-test (kapitola 12.5).

Oproti tomu, mírné porušení předpokladu normality rozdělení zpravidla nemá na výsledky těchto testů podstatný vliv. V případě výrazné nenormality však raději použijeme některý neparametrický test (například Mannův-Whitneyův test (kapitola 12.3)).

Příklad 7.1. Předpokládejme, že obsah nikotinu v cigaretách má normální rozdělení. Tabáková firma TAB prohlašuje, že jejich cigarety mají nižší obsah nikotinu než cigarety NIK. Pro ověření tohoto prohlášení bylo náhodně vybráno z produkce TAB 20 krabiček cigaret (po 20 kusech) a v nich bylo zjištěno průměrně 42,6 mg nikotinu (v jedné cigaretě). Výběrová směrodatná odchylka obsahu nikotinu v testovaných cigaretách TAB byla 3,7 mg. Ve 25 krabičkách (po 20 kusech) cigaret NIK bylo zjištěno průměrně 48,9 mg nikotinu na cigaretu. Výběrová směrodatná odchylka obsahu nikotinu v testovaných cigaretách NIK byla 4,3 mg. Ověřte tvrzení firmy TAB čistým testem významnosti.

Řešení 7.1

Obsah
276. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

7.3. Mannův-Whitneyův test

Mannův-Whitneyův test je neparametrickým testem o shodě mediánů. Nechť X_1, X_2 až X_{n_1} a $Y_1, Y_2, \ldots, Y_{n_2}$ jsou dva nezávislé výběry ze spojitých rozdělení se stejným rozptylem a tvarem. Označení výběrů se volí tak, aby platilo $n_1 \geqq n_2$.

Testujeme nulovou hypotézu o shodě mediánů, tj.

$$H_0: \quad x_{0,5} = y_{0,5}$$

vůči alternativě v jednom z tvarů

$$\begin{aligned} H_A: \quad & x_{0,5} < y_{0,5}, & (1)\\ & x_{0,5} > y_{0,5}, & (2)\\ & x_{0,5} \neq y_{0,5}. & (3)\end{aligned}$$

Volba vhodné alternativy je v tomto případě dána vztahem mezi mediány jednotlivých výběrů. Je-li $\tilde{x}_0, 5$ jednoznačně nižší než $\tilde{y}_0, 5$, volíme alternativu ve tvaru (1). Je-li $\tilde{x}_0, 5$ jednoznačně vyšší než $\tilde{y}_0, 5$, volíme alternativu ve tvaru (2). Pohybuje-li se $\tilde{x}_0, 5$ v blízkosti $\tilde{y}_0, 5$, volíme alternativu ve tvaru (3).

Postup výpočtu testového kritéria:

- Všech $n_1 + n_2$ hodnot získaných z výběrů $X_1, X_2, \ldots, X_{n_1}$ a $Y_1, Y_2, \ldots, Y_{n_2}$ uspořádáme vzestupně a jednotlivým hodnotám přiřadíme pořadí. Nejnižší hodnotě je přiřazena hodnota 1, nejvyšší hodnotě je přiřazena hodnota $n_1 + n_2$, pokud soubor obsahuje několik pozorování se stejnou hodnotou, je těmto hodnotám přiřazeno tzv. průměrné pořadí.

- Označíme T_1 součet pořadí hodnot $X_1, X_2, \ldots, X_{n_1}$ a T_2 součet pořadí hodnot $Y_1, Y_2, \ldots, Y_{n_2}$. Platí, že $T_1 + T_2 = \frac{1}{0}2\,(n_1+n_2)\,(n_1+n_2+1)$.
- Vypočteme statistiky

$$U_1 = n_1 n_2 + \frac{n_1\,(n_1+1)}{2} - T_1, \qquad U_2 = n_1 n_2 + \frac{n_2\,(n_2+1)}{2} - T_2.$$

  (Platí, že $U_1 + U_2 = n_1 n_2$.)
- Testové kritérium pak určíme jako

$$T(\boldsymbol{X}, \boldsymbol{Y}) = min(U_1, U_2),$$

  které má za předpokladu platnosti H_0 rozdělení, jehož kritické hodnoty jsou tabelovány (Tabulka T7).
- Pokud je pozorovaná hodnota testového kritéria menší nebo rovna příslušné kritické hodnotě, nulová hypotéza se zamítá.

Pro velká n_1 a n_2 (v praxi pro $n_1 > 30$, $n_2 > 20$) lze použít testové kritérium

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\left(min\,(U_1, U_2) - \frac{n_1 n_2}{2}\right)}{\sqrt{\frac{1}{12} n_1 n_2\,(n_1+n_2+1)}},$$

které má za předpokladu platnosti nulové hypotézy normované normální rozdělení. Dále pak postupujeme dle obecného schématu čistého testu významnosti.

Příklad 7.2. Máme dvě skupiny studentů. První (kontrolní), v níž jsou studenti vyučováni tradičními metodami, a druhá, v níž jsou studenti vyučováni experimentálními metodami. V následujících tabulkách je uvedeno bodové hodnocení vybraných studentů u zkoušky. Na základě srovnání mediánu rozhodněte, zda studenti vyučováni experimentálním metodami dosahují lepších výsledků než studenti s klasickým vyučováním.

Výběr z první skupiny (klasická výuka)

60	49	52	68	68	45	57	52	13	40	33	30	28	30	48

Výběr z druhé skupiny (experimentální výuka)

38	18	68	84	72	48	36	92	6	54

Řešení 7.2

7.4. Test homogenity dvou binomických rozdělení

Jednou z nejstarších a ve statistice stále se velmi často vyskytujících úloh je srovnání homogenity dvou binomických rozdělení. Předpokládejme, že v sérii n_1 nezávislých opakování pokusu se nějaký náhodný jev A vyskytl X-krát. Pak se pokusy nezávisle opakují za jiných podmínek tak, že v sérii n_2 opakování pokusu se náhodný jev A vyskytne Y-krát. Počet výskytu jevu A ve skupině n_1 opakování pokusu (náhodnou veličinu X) lze považovat za náhodnou veličinu s rozdělením $Bi(n_1; \pi_1)$, počet výskytu jevu A ve skupině n_2 opakování pokusu (náhodnou veličinu Y) pak lze považovat za náhodnou veličinu s rozdělením $Bi(n_2; \pi_2)$, kde π_1, π_2 jsou neznámé pravděpodobnosti. Na základě těchto údajů chceme testovat hypotézu

$$H_0: \quad \pi_1 = \pi_2$$

proti jedné z alternativ

$$\begin{array}{llllr} H_A: & \pi_1 < \pi_2, & \text{resp.} & \pi_1 - \pi_2 < 0, & (1) \\ & \pi_1 > \pi_2, & \text{resp.} & \pi_1 - \pi_2 > 0, & (2) \\ & \pi_1 \neq \pi_2, & \text{resp.} & \pi_1 - \pi_2 \neq 0. & (3) \end{array}$$

Označme $p_1 = \frac{X}{n_1}$ bodový odhad pravděpodobnosti π_1 a $p_2 = \frac{Y}{n_2}$ bodový odhad pravděpodobnosti π_2. Volba vhodné alternativy je pak dána vztahem mezi relativními četnostmi jevu A v jednotlivých výběrech. Je-li p_1 výrazně nižší než p_2, volíme alternativu ve tvaru (1). Je-li p_1 výrazně vyšší než p_2, volíme alternativu ve tvaru (2). Nachází-li se p_1 v blízkosti p_2, volíme alternativu ve tvaru (3).

Pro provedení tohoto testu musíme mít k dispozici výběry o dostatečném rozsahu n_1, resp. n_2. Rozsahy jednotlivých výběrů lze považovat za dostatečné, pokud jsou splněny podmínky

$$n_1 > \frac{9}{p_1\,(1-p_1)} \qquad \text{a} \qquad n_2 > \frac{9}{p_2\,(1-p_2)}.$$

Testovým kritériem je statistika

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

která má v případě platnosti nulové hypotézy přibližně normované normální rozdělení $N(0;1)$ (viz 8.7).

Dále pokračujeme podle obecného schématu čistého testu významnosti.

Příklad 7.3. Byly testovány magnetofony od dvou výrobců – SONIE a PHILL. Firma SONIE prohlašuje, že jejich magnetofony mají nižší procento reklamací. Pro ověření tohoto prohlášení bylo dotazováno několik prodejců magnetofonů a bylo zjištěno, že z 300 prodaných magnetofonů firmy SONIE bylo v průběhu záruční doby reklamováno 10 výrobků a z 440 prodaných magnetofonů firmy PHILL bylo v záruční době reklamováno 18 výrobků. Otestujte pravdivost prohlášení firmy SONIE čistým testem významnosti.

Řešení 7.3

7.5. Párové testy

V předcházející kapitole jsme se věnovali dvouvýběrovým testům, které umožňují na základě dvou **nezávislých** výběrů porovnat neznámé parametry dvou populací. V praxi se však často stává také to, že u každé z n statistických jednotek zjišťujeme hodnoty nějakých dvou spolu souvisejících znaků (např. tlak krve před a po podání určitého léku, ostrost vidění levého a pravého oka, rychlost zavírání dveří automobilu měřena dvěma různýma metodami, ...). Výsledkem zjišťování jsou pak dvojice náhodných veličin $(X_1, Y_1), (X_2, Y_2)$ až (X_n, Y_n), které tvoří **páry závislých pozorování** (jde o veličiny zjišťované na stejné statistické jednotce).

Můžeme chtít ověřit, zda výběry $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ a $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ pocházejí z rozdělení se stejnými středními hodnotami μ_1 a μ_2, čili testovat hypotézu

$$H_0: \quad \mu_1 = \mu_2$$

vůči alternativě v jednom z tvarů

$$\begin{aligned} H_A: \quad & \mu_1 < \mu_2, \quad \text{resp.} \quad \mu_1 - \mu_2 < 0, \\ & \mu_1 > \mu_2, \quad \text{resp.} \quad \mu_1 - \mu_2 > 0, \\ & \mu_1 \neq \mu_2, \quad \text{resp.} \quad \mu_1 - \mu_2 \neq 0. \end{aligned}$$

Chceme-li například ověřit vliv určitého léku na tlak krve, budeme u každého pacienta pozorovat dvojici znaků (X_i, Y_i), kde X_i je tlak krve před podáním léku a Y_i je tlak krve po podání léku u i. pacienta. Pro ověření účinnosti léku nemá smysl zjišťovat, zda je statisticky významný rozdíl mezi průměrným tlakem všech pacientů před podáním léku a průměrným tlakem všech pacientů po podání léku. (Proč?) U každého pacienta určíme rozdíl tlaků krve po a před podáním léku a budeme zjišťovat, zda se tento rozdíl statisticky významně liší od nuly. Nebude-li prokázána statisticky významná odchylka od nuly, bude lék prohlášen za neúčinný.

Definujme soubor rozdílů (diferencí)

$$\boldsymbol{D} = (D_1, D_2, , \ldots, , D_n)\,, \quad \text{kde } D_i = X_i - Y_i.$$

Lze předpokládat, že náhodné veličiny $(D_1, D_2, \ldots, D_n)$ jsou nezávislé a že mají stejné rozdělení se střední hodnotou $\mu = \mu_1 - \mu_2$. Test o shodě dvou středních hodnot prováděný na základě dvou závislých výběrů můžeme převést na jednovýběrový test o střední hodnotě aplikovaný na soubor diferencí (rozdílů) $\boldsymbol{D}$, tzn. můžeme testovat hypotézu

$$H_0: \quad \mu = 0.$$

vůči alternativě v jednom z tvarů

$$H_A: \quad \mu < 0, \\ \mu > 0, \\ \mu \neq 0.$$

Lze-li předpokládat normální rozdělení veličin $(D_1, D_2, \ldots, D_n)$, můžeme použít jednovýběrový t test, nazývaný v tomto případě **párový *t* test**.

Mají-li veličiny $(D_1, D_2, \ldots, D_n)$ spojité rozdělení s hustotou symetrickou kolem mediánu, pak hypotézu o tomto mediánu můžeme testovat jednovýběrovým Wilcoxonovým testem (tzv. **párový Wilcoxonův test**), popřípadě mediánovým testem, kterému v případě párového testu říkáme **test znaménkový**.

Příklad 7.4. Předpokládejme, že ojetí předních pneumatik [mm] podléhá normálnímu rozdělení. U 6 aut bylo zjištěno ojetí předních pneumatik (viz tabulka).

Pravá	1,8	1,0	2,2	0,9	1,5	1,6
Levá	1,5	1,1	2,0	1,1	1,4	1,4

Ojíždějí se levá a pravá pneumatika stejně?

Řešení 7.4

Shrnutí:

Dvouvýběrové testy pro nezávislé výběry umožňují na základě dvou **nezávislých** výběrů porovnat neznámé parametry dvou populací.

Stručný přehled testových statistik, s nimiž jsme se seznámili

Dvouvýběrové parametrické testy pro nezávislé výběry

Název testu	Testované parametry	Předpoklady testu	Testová statistika $T(X,Y)$	Nulové rozdělení	Poznámka
test o shodě rozptylů	rozptyly σ_1^2, σ_2^2 (sm. odch. σ_1, σ_2)	nezávislé výběry, normalita populací, neznámé μ_1, μ_2	$\dfrac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}}$	F_{n_1-1,n_2-1}	Při čistém testu významnosti nelze použít oboustran. alternativu.
dvouvýběrový z test	střední hodnoty μ_1, μ_2	nezávislé výběry, normalita populací, známé σ_1^2, σ_2^2	$\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1}+\frac{\sigma_Y^2}{n_2}}}$	$N(0;1)$	
dvouvýběrový t test		nezávislé výběry, normalita populací, neznámé σ_1^2, σ_2^2, $\sigma_1^2=\sigma_2^2$	$\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{(n_1-1)s_X^2+(n_2-1)s_Y^2}{n_1+n_2-2}}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$t_{n_1+n_2-1}$	
Aspinové – Welchův test		nezávislé výběry, normalita populací, neznámé σ_1^2, σ_2^2, $\sigma_1^2\neq\sigma_2^2$	$\dfrac{(\bar{X}_{TAB}-\bar{X}_{NIK})-(\mu_{TAB}-\mu_{NIK})}{\sqrt{\frac{S_{TAB}^2}{n_{TAB}}+\frac{S_{NIK}^2}{n_{NIK}}}}$	t_ν kde, $\nu = \dfrac{\left(\frac{s_{TAB}^2}{n_{TAB}}+\frac{s_{NIK}^2}{n_{NIK}}\right)^2}{\frac{1}{n_{TAB}-1}\left(\frac{s_{TAB}^2}{n_{TAB}}\right)^2+\frac{1}{n_{NIK}-1}\left(\frac{s_{NIK}^2}{n_{NIK}}\right)^2}$	

Dvouvýběrové neparametrické testy pro nezávislé výběry

Název testu	Testovaný parametr	Předpoklady testu	Testová statistika	Nulové rozdělení	Poznámka
Mannův-Whitneyův test	mediány $x_{0,5}, y_{0,5}$	nezávislé výběry ze spojitých rozdělení se stejným rozptylem a tvarem.	$min(U_1, U_2)$, kde $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$, $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$	Kritické hodnoty rozdělení jsou uvedeny v tabulce	Označení výběrů se volí tak, aby platilo $n_1 \geq n_2$. Je-li pozorovaná hodnota testové statistiky menší nebo rovna kritické hodnotě, zamítáme H_0.
test homogenity dvou binomických rozdělení	pravděpodobnosti π_1, π_2	$n_1 > \frac{9}{p_1(1-p_1)}$, $n_2 > \frac{9}{p_2(1-p_2)}$	$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$	$N(0;1)$	

Dvouvýběrové párové testy

V praxi se často setkáváme se situací, kdy máme n měřených jednotek (či objektů), na nichž jsou provedena dvě pozorování, daná různými experimentálními podmínkami (např. působí či nepůsobí nějaký faktor, jehož účinky jsou předmětem šetření). Testování shody středních hodnot, resp. mediánů, provádíme tak, že vytvoříme jednu datovou hodnotu pro každou statistickou jednotku. V nejjednodušším datovém modelu bude touto hodnotou rozdíl získaných dvou pozorování pro danou i-tou statistickou jednotku. Dané rozdíly pak mohou být použity pro jednovýběrové testy o tom, zda sledovaný parametr je nula, což je ekvivalentní s tvrzením, že neexistují žádné rozdíly mezi experimentálními podmínkami (nebo že zkoumaný faktor je neúčinný).

Test

Jak pracovat s testy?

Zacátek testu

1. (3b.) 1. Označte všechny neparametrické (robustní) testy.
   (a) dvouvýběrový t-test,
   (b) párový t-test,
   (c) Aspinové-Welchův test,
   (d) Mannův-Whitneyův test,
   (e) znaménkový test nebo párový Wilcoxonův test,
   (f) dvouvýběrový F-test (test o shodě rozptylů),
   (g) test homogenity dvou binomických rozdělení.

2. (1b.) 2. Předpokladem pro použití Mannova-Whitneyova testu je
   (a) normalita obou výběrů,
   (b) normalita obou výběrů a homoskedasticita,
   (c) normalita obou výběrů a heteroskedasticita,
   (d) výběry ze spojitého rozdělení,
   (e) dostatečný rozsah obou výběrů ($n_i > \frac{9}{p_i(1-p_i}, i = 1, 2$, kde p_i je relativní četnost výskytu sledovaného jevu v $i-$tém výběru).

3. (1b.) 3. Předpokladem pro použití párového t- testu je

(a) normalita obou výběrů,
(b) normalita obou výběrů a homoskedasticita,
(c) normalita obou výběrů a heteroskedasticita,
(d) výběry ze spojitého rozdělení,
(e) dostatečný rozsah obou výběrů ($n_i > \frac{9}{p_i(1-p_i)}, i = 1, 2$, kde p_i je relativní četnost výskytu sledovaného jevu v $i-$tém výběru).

4. (1b.) 4. Předpokladem pro použití Aspinové-Welchova testu je

(a) normalita obou výběrů,
(b) normalita obou výběrů a homoskedasticita,
(c) normalita obou výběrů a heteroskedasticita,
(d) výběry ze spojitého rozdělení,
(e) dostatečný rozsah obou výběrů ($n_i > \frac{9}{p_i(1-p_i)}, i = 1, 2$, kde p_i je relativní četnost výskytu sledovaného jevu v $i-$tém výběru).

5. (1b.) 5. Neparametrickým protějškem Aspinové-Welchova testu je

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,

(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

6. (1b.) 6. Neparametrickým protějškem párového t-testu je

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

7. (1b.) 7. Neparametrickým protějškem dvouvýběrového t-testu je

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

8. (1b.) 8. Tabáková firma TAB prohlašuje, že jejich cigarety mají nižší obsah nikotinu než cigarety NIK. Obsah nikotinu byl změřen ve 100 cigaretách TAB a 100 cigaretách NIK. Na základě obou výběru byla ověřena homoskedasticita obsahů nikotinu v cigaretách TAB a NIK. Bylo ověřeno, že obsah nikotinu v cigaretách má normální rozdělení. Chceme-li ověřit, zda lze tvrzení firmy TAB prohlásit za nepravdivé, použijeme

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

9. (1b.) 9. Při testování ojetí [mm] pneumatik 11 automobilů určité značky byla zamítnuta normalita ojetí pneumatik [mm]. Chceme-li ověřit, zda se pravé a levé přední pneumatiky automobilů této značky ojíždějí srovnatelně, použijeme

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),

(g) test homogenity dvou binomických rozdělení.

10. (1b.) 10. Bylo ověřeno, že hmotnost balení cukru má normální rozdělení. Testujeme-li, zda seřízením výrobní linky došlo ke snížení kolísavosti hmotnosti balení cukru, použijeme

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

11. (1b.) 11. Určete, zda jsou následující tvrzení pravdivá. (ANO, NE)

(a) dvouvýběrový t-test,
(b) párový t-test,
(c) Aspinové-Welchův test,
(d) Mannův-Whitneyův test,
(e) znaménkový test nebo párový Wilcoxonův test,
(f) dvouvýběrový F-test (test o shodě rozptylů),
(g) test homogenity dvou binomických rozdělení.

12. (3b.) Určete, zda jsou následující tvrzení pravdivá. (ANO, NE)

(a) Při neparametrickém testu homogenity dvou binomických rozdělení nemusíme ověřovat žádné předpoklady o výběrech.

(a) ANO (b) NE

(b) Mannův-Whitneyův test se používá pro ověření shody úrovně ve dvou závislých výběrech.

(a) ANO (b) NE

(c) Každý test hypotézy $H_0 : \mu_1 = \mu_2$, tj. hypotézy o shodě dvou středních hodnot je testem párovým.

(a) ANO (b) NE

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1. Provozovatel čerpacích stanic chce postavit novou čerpací stanici na severním nebo jižním okraji menšího města. Projekt předpokládá, že bude vybrán ten výjezd z města, kde je vyšší intenzita provozu. Na severním výjezdu z města probíhalo šetření během 50 dní a byl zjištěn počet 4 000 projíždějících vozidel (denně, se směrodatnou odchylkou 70 vozidel). Na jižním výjezdu z města bylo za 45 dní zaznamenáno v průměru 3 900 projíždějících vozidel denně (směrodatná odchylka 60 vozidel). Lze rozhodnout, který výjezd je zatíženější? Předpokládejte, že počet vozidel projíždějících denně jednotlivými výjezdy lze modelovat normálním rozdělením.

   Řešení příkladu

2. Firma Modus zjišťovala v roce 2006 názory Čechů na bezpečnost jaderných elektráren. Ze 420 respondentů ve věku od 18 do 30 let považovalo 24% současná bezpečnostní opatření za postačující. Z 510 respondentů ve věku 30 až 50 let považovalo současná bezpečnostní opatření za postačující 34%. Ověřte čistým testem významnosti, zda má věk vliv na odpověď.

   Řešení příkladu

3. Byly testovány polovodičové součástky dvou výrobců – MM a PP. MM prohlašuje, že její výrobky mají nižší procento vadných kusů. Pro ověření tohoto tvrzení bylo z produkce MM náhodně vybráno 200 součástek, z nichž 14 bylo vadných. Podobný experiment byl proveden u firmy PP s výsledkem 10 vadných ze 100 náhodně vybraných součástek.

   a) Otestujte tvrzení firmy MM čistým testem významnosti.

   b) Otestujte tvrzení firmy MM prostřednictvím intervalového odhadu na hladině významnosti 0,05.

   c) Nalezněte 95% interval spolehlivosti pro počet vadných součástek firmy MM.

   Řešení příkladu

Obsah
292. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

4. Denní přírůstky váhy selat při krmení směsi A, resp. B jsou uvedeny v tabulce:

A	62	54	55	60	53	58
B	52	56	50	49	51	

Ovlivňuje výběr krmné směsi přírůstky váhy selat? (Bylo zjištěno, že denní přírůstky váhy selat mají lognormální rozdělení.)

Řešení příkladu

5. Na skupině dobrovolníků byl testován prostředek na snížení hmotnosti. Hmotnosti 12 testovaných lidí před a po dietní kůře jsou v níže uvedené tabulce. Určete na hladině významnosti 0,05, zda je prostředek účinný. Předpokládejte, že váha před i po dietní kůře má normální rozdělení.

hmotnost před dietou [kg]	85	75	90	65	150	80	110	56	88	73	67	134
hmotnost po dietě [kg]	76	75	81	64	155	72	99	45	89	66	56	110

Řešení příkladu

Kapitola 8

Vícevýběrové testy parametrických hypotéz

Cíle

Po prostudování tohoto odstavce budete

- umět testovat homoskedasticitu více než dvou souborů – budete znát Bartlettův, Leveneův, Hartleyův a Cochranův test,
- umět zvolit správný test pro ověření shody úrovně ve více než dvou souborech (ANOVA, Kruskalův-Wallisův test, Friedmanův test),
- umět provést post hoc analýzu pro vícevýběrové testy o shodě úrovně.

V této kapitole se budeme věnovat testům umožňujícím, na základě $k > 2$ náhodných výběrů, ověření shody k parametrů (rozptylů, středních hodnot, mediánů).

Označme:

celkový rozsah všech k výběrů: $n = \sum_{i=1}^{k} n_i,$

průměr i-tého výběru: $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},$

celkový průměr všech k výběrů: $\bar{\bar{X}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij},$

výběrový rozptyl i-tého výběru: $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2 .$

Výchozí situaci lze zachytit v následující tabulce.

Číslo skupiny	1	2	…	k
Náhodný výběr	X_{11} ⋮ X_{1n_1}	X_{21} ⋮ X_{2n_2}	⋮	X_{k1} ⋮ X_{kn_k}
Rozsah skupiny	n_1	n_2		n_k
Průměr skupiny	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_k$
Rozptyl skupiny	s_1^2	s_2^2		s_k^2

8.1. Testy shody rozptylů

Jedním z předpokladů analýzy rozptylu, testu umožňujícího na základě $k > 2$ náhodných výběrů ověření shody k středních hodnot, je shoda rozptylů (homoskedasticita) všech k normálních rozdělení, z nichž jsou výběry pořizovány. Předpoklad homoskedasticity se dá ověřit.

Předpokládejme, že máme $k > 2$ **nezávislých** výběrů z **normálního rozdělení**,

$$X_{11}, X_{12}, \ldots, X_{1n_1} \text{ je výběr z } N(\mu_1; \sigma_1^2),$$

atd. až

$$X_{k1}, X_{k2}, \ldots, X_{kn_1} \text{ je výběr z } N(\mu_k; \sigma_k^2),$$

Je třeba testovat hypotézu

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

proti alternativě, že se alespoň jedna dvojice rozptylů liší

$$H_A : \neg H_0.$$

K tomuto účelu se využívá například Bartlettův test.

8.1.1. Bartlettův test

Nechť

$$MS_{\mathrm{e}} = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2$$

(MS_{e} nazýváme reziduální rozptyl a je používán rovněž v analýze rozptylu),

$$C = 1 - \frac{1}{a(k-1)}\left(\frac{1}{n-k} - \sum_{i=1}^{k}\frac{1}{n_i-1}\right).$$

Platí-li nulová hypotéza, má testová statistika

$$B = \frac{1}{C}\left[(n-k)\ln MS_{\mathrm{e}} - \sum_{i=1}^{k}(n_i-1)\ln s_i^2\right]$$

přibližně χ^2 rozdělení s $n-k$ stupni volnosti. Pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $n-k$ stupni volnosti.

Bartlettův test je velmi **citlivý na porušení předpokladu normality**, nelze jej tedy použít, nepocházejí-li všechny porovnávané výběry z normálního rozdělení. V takovémto případě volíme pro ověření homoskedasticity raději tzv. Leveneův test.

8.1.2. Leveneův test

Tento test je ve srovnání s Barttletovým testem méně citlivý na porušení předpokladu normality. Nedošlo-li však k zamítnutí normality pro žádný ze sledovaných výběrů, volíme pro test homoskedasticity raději test Bartlettův, který má větší sílu testu.

Nechť $Z_{ij} = |X_{ij} - \bar{X}_i|$. Označme

$$\bar{Z}_i = \frac{\sum\limits_{j=1}^{n_i}}{n_i}, \qquad \bar{\bar{X}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{Z_{ij}}{n},$$

$$SS_{ZB} = \sum_{i=1}^{k} n_i\left(\bar{Z}_i - \bar{\bar{Z}}\right)^2, \qquad SS_{Ze} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \bar{Z}\right)^2.$$

Platí-li nulová hypotéza, pak má testová statistika

$$\frac{\dfrac{SS_{ZB}}{k-1}}{\dfrac{SS_{Ze}}{n-k}}$$

přibližně Fisher-Snedecorovo rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli. Pak

$$p\text{-hodnota} = 1 - F_0(x_{OBS}),$$

kde je $F_0(x)$ distribuční funkce Fisher-Snedecorova rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.

Pro jisté případy jsou navrženy i modifikace Leveneova testu. V případě, že výběrové soubory vykazují výraznou šikmost, lze použít $Z_{ij} = |X_{ij} - X_{i_{0,5}}|$, kde $X_{i_{0,5}}$ označuje medián i-tého výběru. Vykazují-li výběrové soubory výraznou špičatost, lze použít $Z_{ij} = |X_{ij} - \bar{X}_{i_{10}}|$, kde $\bar{X}_{i_{10}}$ označuje 10% useknutý průměr i-tého výběru, tj. průměr z výběru, z něhož bylo odstraněno 10% největších a 10% nejmenší hodnot.

Jsou-li rozsahy všech skupin stejné (říkáme, že třídění je vyvážené), tj. $n_1 = \ldots = n_k$, používá se k testování homoskedasticity také Hartleyův nebo Cochranův test.

8.1.3. Hartleyův test

Je zřejmé, že pokud nezjistíme statisticky významný rozdíl mezi největším a nejmenším výběrovým rozptylem, nebudou se statisticky významně lišit ani ostatní dvojice výběrových rozptylů. Hartleyův test je založen na testové statistice

$$F_{max} = \frac{\max s_i^2}{\min s_i^2}.$$

Nulová hypotéza se zamítá, je-li pozorovaná hodnota F_{max} větší nebo rovna kritické hodnotě $h_\alpha(k, n_1 - 1)$, která je tabelována ve speciálních tabulkách (tabulka T8).

8.1.4. Cochranův test

Tento test používá testovou statistiku

$$G_{max} = \frac{\max s_i^2}{s_1^2 + \ldots + s_k^2}.$$

K zamítnutí nulové hypotézy vedou vysoké pozorované hodnoty G_{max}. Kritické hodnoty $c_\alpha(k, n_1 - 1)$ jsou uvedeny v tabulce T9.

Příklad 8.1. Při sledování kvality pěnového polystyrénu (EPS) byla sledována hustota EPS $[kg/m^3]$ čtyř různých výrobců A, B, C, D. Hustota byla stanovena pro 7 produktů každého z výrobců. Výsledky byly vepsány do níže uvedené tabulky.

Výrobce	Objemová hmotnost EPS [kg/m^3]							Průměr [kg/m^3]	Výběrový rozptyl [kg^2/m^6]
A	14,3	13,0	17,6	16,9	16,1	20,0	18,4	16,61	5,73
B	19,1	22,5	21,2	21,0	20,3	17,4	22,7	20,60	3,52
C	19,7	16,8	15,8	20,1	18,2	18,6	18,9	18,30	2,36
D	13,2	12,6	12,9	13,7	17,3	11,2	15,0	13,70	3,83

Ověřte homoskedasticitu objemové hmotnosti EPS jednotlivých výrobců.

Řešení 8.1.

8.2. Jednofaktorová ANOVA

V kapitole 7 jsme se věnovali mimo jiné také dvouvýběrovému t testu, který na základě dvou nezávislých výběrů umožňuje porovnat střední hodnoty dvou normálně rozdělených populací. V mnoha případech však potřebujeme porovnat střední hodnoty více než dvou populací. Můžeme například zkoumat, zda

- typ absolvované střední školy ovlivňuje počet bodů dosažených studenty u přijímací zkoušky z matematiky,
- použitá medikace ovlivňuje krevní tlak pacientů,

Obsah
300. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

- typ použitého hnojiva ovlivňuje výnosy určité plodiny,
- pracovní výkon dělníka závisí na umístění stroje, apod.

8.2.1. Motivační příklad

Pro ilustraci si uveďme motivační příklad, jenž nás bude provázet touto kapitolou.

Naším úkolem je porovnat úspěšnost absolventů gymnázii, SPŠ a odborných učilišť s maturitou (OU) u přijímací zkoušky z matematiky. Dosažené výsledky náhodně vybraných dvaceti studentů jsou uvedeny v následující tabulce.

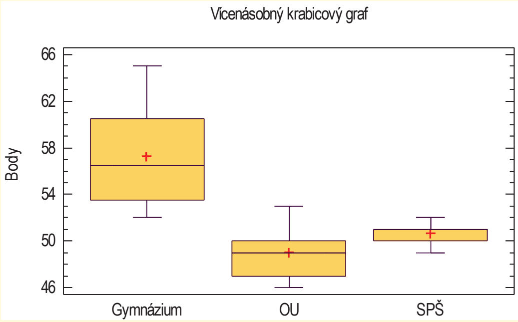
Gymnázium	SPŠ	OU
55	52	47
54	50	53
58	51	49
61	51	50
52	49	46
60		48
53		50
65		

***Poznámka**: Typ absolvované střední školy je vlastně kategoriální proměnnou, která od sebe rozlišuje jednotlivé porovnávané skupiny. Této rozlišující proměnné se říká **faktor**.*

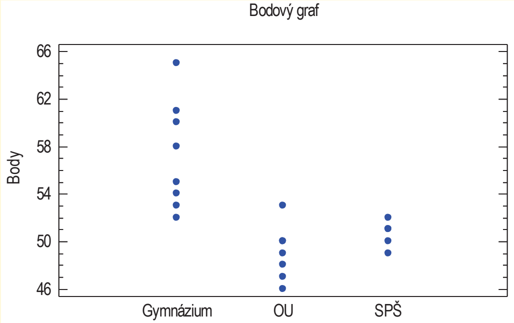
Protože tyto typy škol reprezentují studenti různých škol (není gymnázium jako gymnázium…), s různými studijními výsledky a různým nadáním na matematiku, a také vlivem dalších různých vlivů, bodové hodnocení zástupců jednotlivých typů škol značně kolísá.

8.2.2. Explorační analýza

Prvním krokem při analýze takovýchto dat je jejich vizualizace, popř. výpočet základních číselných charakteristik jednotlivých výběrů.



Obr. 8.1: Krabicový graf



Obr. 8.2: Bodový graf

Jsou-li analyzované výběry dostatečně malé, lze pro jejich vizualizaci použít bodový graf (Obr. 8.2). Dochází-li v bodovém grafu k překrývání jednotlivých bodů znesnadňujícímu interpretaci výsledků (typické pro rozsáhlejší výběry), používáme pro vizualizaci vícenásobný krabicový graf (Obr. 8.1).

Krabicový graf použijeme mimo jiné k identifikaci odlehlých pozorování, která obecně způ-

Tab. 8.1: Základní číselné charakteristiky

	Gymnázium	SPŠ	OU
rozsah	8	5	7
průměr	57,3	50,6	49,0
výběrový rozptyl	20,5	1,3	5,3

sobují selhání analýzy rozptylu. Pokud odlehlá pozorování vyskytující se v datech byla způsobena:

- hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
- důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlostí a předpokládáme-li, že již nenastane, vyloučíme je z dalšího zpracování. Jestliže odlehlá pozorování v datech ponecháme, použijeme raději Kruskalův-Wallisův test (kapitola 8.3).

V našem případě lze na základě krabicového grafu tvrdit, že skupiny neobsahují odlehlá pozorování. Zdá se, že mezi skupinami je rozdíl mezi získanými body – nejlepších průměrných výsledků dosáhli studenti gymnázií, výsledky absolventů SPŠ a OU se zdají srovnatelné. Nyní chceme zjistit, zda jsou výsledky výběrového šetření natolik „silné“, aby vedly k zamítnutí hypotézy o shodě středních hodnot, tj. k zamítnutí tvrzení, že typ absolvované střední školy nemá vliv na úspěšnost studentů při přijímací zkoušce z matematiky.

8.2.3. Předpoklady pro použití analýzy rozptylu

Jak porovnat průměry více než dvou výběrů? Zdánlivě by stačilo utvořit všechny dvojice náhodných výběrů a na všechny aplikovat dvouvýběrový t test. Jak již víte z kombinatoriky, těchto testů je $\binom{k}{2} = \frac{k(k-1)}{2}$. Kdyby byl každý z nich proveden na hladině významnosti α, byla by výsledná hladina významnosti testu mnohem vyšší než α. Tím by byl test zcela znehodnocen. Proto v roce 1925 vytvořil sir R. A. Fisher metodu nazývanou **analýza rozptylu**, resp. **ANOVA** (akronym z anglického „**AN**alysis **O**f **VA**riance“), která zachovává výslednou hladinu významnosti α a rozumnou sílu testu.

Na tomto místě je třeba zmínit požadavky parametrického testu, který budeme dále užívat.

Analýza rozptylu byla původně navržena pro stejný rozsah jednotlivých výběrů, což označujeme jako vyvážené třídění. V praxi bývá tento předpoklad málokdy splněn – platí však, že čím těsněji je toto pravidlo splněno, tím věrohodnější jsou výsledky testu.

Analýza rozptylu ve své parametrické podobě předpokládá

- nezávislost výběrů,
- normalitu rozdělení,
- homoskedasticitu (identické rozptyly).

Nezávislost výběrů je velmi důležitým předpokladem. Pokud není tento předpoklad splněn, můžeme získat užitím analýzy rozptylu zcela nesmyslné výsledky. Pro porovnání $k > 2$ závislých výběrů lze použít Friedmanův test (kapitola 8.4).

Na porušení normality není ANOVA příliš citlivá, zvlášť pokud mají všechny výběry rozsah větší než 30. Při výraznějším porušení normality (viz testy normality) se doporučuje použít neparametrickou obdobu analýzy rozptylu – Kruskalův - Wallisův test (kapitola 8.3).

Pro ověření homoskedasticity (shody rozptylů) lze použít například testy uvedené v kapitole 8.1. Při větším porušení homokedasticity se doporučuje, podobně jako při porušení normality, použít Kruskalův – Wallisův test (kapitola 8.3).

Předpokládejme, že máme $k < 2$ **nezávislých** výběrů z **normálního rozdělení**,

$$X_{11}, X_{12}, \ldots, X_{1n_1} \text{ je výběr z } N(\mu_1; \sigma_1^2),$$

$$\vdots$$

$$X_{k1}, X_{k2}, \ldots, X_{kn_1} \text{ je výběr z } N(\mu_k; \sigma_k^2),$$

Je třeba testovat hypotézu

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

proti alternativě, že se alespoň jedna dvojice středních hodnot liší

$$H_A : \neg H_0.$$

Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice μ_i, μ_j toto zamítnutí způsobily (kapitola 8.2.7).

8.2.4. Rozklad celkové variability

Proč se testu o shodě středních hodnot říká „analýza rozptylu“? Tento název zavedl její autor sir R. A. Fisher (1890-1962), aby postihl její charakter – úlohu o shodě $k > 2$ středních hodnot převedl na test shody dvou rozptylů, tzv. *F-test*, který již znáte z kapitoly 7.1.

Zabýváme se otázkou, zda se výsledky studentů opravdu liší podle toho, jaký typ střední školy absolvovali. Neboli – jsou průměry jednotlivých výběrů rozdílné vlivem různých středních hodnot příslušných populací, nebo lze rozdíly mezi průměry přičíst na vrub náhodnému kolísání?

Je třeba testovat hypotézu H_0: $\mu_G = \mu_{\text{SPŠ}} = \mu_{OU}$,

kde μ_G je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů gymnázia, $\mu_{\text{SPŠ}}$ je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů SPŠ, μ_{OU} je střední bodové hodnocení přijímacích zkoušek z matematiky absolventů OU

vůči alternativě: H_A: $\neg H_0$ (neplatí H_0).

Myšlenkou analýzy rozptylu je, že celkovou variabilitu závisle proměnné (výsledky přijímacího řízení z matematiky všech 20 studentů) rozdělíme do dvou částí, na variabilitu mezi skupinami a variabilitu uvnitř skupin.

Variabilitu jednotlivých pozorování kolem celkového průměru charakterizuje **celkový součet čtverců** (angl. „total sum of squares“),

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2,$$

resp. **celkový rozptyl** (angl. „mean of squares“)

$$MS_T = \frac{SS_T}{n-1}$$

kde $n-1$ je odpovídající počet stupňů volnosti df_T (z angl. „degrese of freedom“).

Vhodným kvantifikátorem meziskupinové variability (jinak řečeno efektu skupin či rozdílů mezi skupinovými průměry $\bar{X}_i$, v našem případě vlivu typu absolvované střední školy) je meziskupinový součet čtverců (angl. „sum of squares between groups“),

$$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2,$$

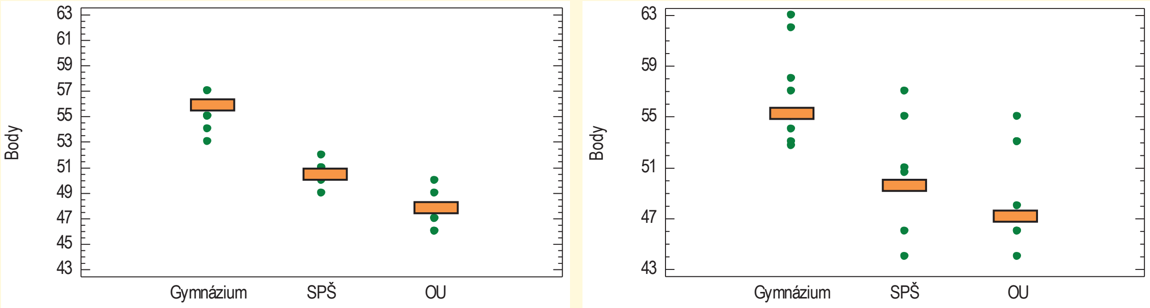
resp. **rozptyl mezi skupinami**

$$MS_B = \frac{SS_B}{k-1},$$

kde $k-1$ je odpovídající počet stupňů volnosti df_B.

Je zřejmé, že rozptyl mezi skupinami neposkytuje dostatečnou informaci o celkové variabilitě, neboť nepostihuje kolísání dat v jednotlivých skupinách.

Pro ujasnění si problému srovnejte dva následující grafy – graf na obr. 8.3a) uvádí bodové hodnocení náhodně vybraných studentů, graf na obr. 8.3b) taktéž, avšak výsledky prezentované v grafu na obr. 8.3b) vykazují značné kolísání v rámci jednotlivých typů škol. Vzhledem k tomu, že skupinové průměry (oranžové úsečky) dat prezentovaných v grafech na



Obr. 8.3: Srovnání datových souborů s nízkou a vysokou variabilitou uvnitř skupin

obr. 8.3a) i na obr. 8.3b) jsou stejné, jsou i rozptyly mezi skupinami pro data prezentována v jednotlivých grafech totožné!
Subjektivní vnímání studovaného problému je však rozdílné. Výsledky studentů prezentované v grafu na obr. 8.3 jsou v rámci jednotlivých skupin natolik rozkolísané oproti rozdílům mezi skupinovými průměry, že si dokážeme představit, že všechny tři výběry lze získat z jedné populace.

Variabilitu uvnitř skupin popisuje tzv. **reziduální součet čtverců** SS_{e} (angl. „sum of squares – errors“)

$$SS_{\mathrm{e}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2,$$

resp. **reziduální rozptyl**

$$MS_{\mathrm{e}} = \frac{SS_{\mathrm{e}}}{n-k}$$

Obsah
308. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

kde $n - k$ je odpovídající počet stupňů volnosti df_{e}.

Všimněte si, že reziduální součet čtverců lze vyjádřit pomocí výběrových rozptylů jednotlivých tříd.

$$SS_{\mathrm{e}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2 = \sum_{i=1}^{k}(n_i - 1)\sum_{j=1}^{n_i}\frac{(X_{ij} - \bar{\bar{X}})^2}{n_i - 1} = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

Lze dokázat, že

$$SS_T = SS_B + SS_{\mathrm{e}}.$$

Příklad 8.2. Rozdělte celkový rozptyl závisle proměnné z motivačního příkladu (výsledky přijímacího řízení z matematiky všech 20 studentů) na variabilitu mezi skupinami a variabilitu uvnitř skupin.

Řešení 8.2.

8.2.5. Testovací kritérium *F-poměr*

Připomeňme si, že se zabýváme otázkou, zda jsou průměry jednotlivých skupin rozdílné vlivem různých středních hodnot příslušných populací, nebo lze rozdíly mezi průměry přičíst na vrub náhodnému kolísání. Liší-li se průměry jednotlivých skupin vlivem různých středních hodnot příslušných populací, pak musí být rozptyl mezi třídami dostatečně velký vzhledem k rozptylu uvnitř tříd (viz obr. 8.3).

Běžně se zkoumá poměr, který se na počest Ronalda Fishera nazývá *F-poměr* (angl. „*F--ratio*“).

$$F - poměr = \frac{MS_B}{MS_{\mathrm{e}}}$$

Obsah
309. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

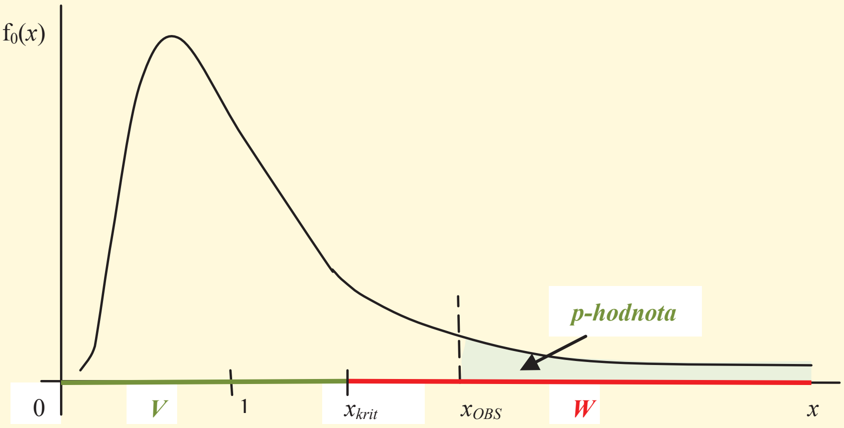
Není-li H_0 pravdivá (střední hodnoty nejsou stejné), pak variabilita mezi třídami SS_B bude relativně velká vůči variabilitě uvnitř tříd SS_e a *F-poměr* bude mnohem větší než 1. Čím větší je *F-poměr*, tím méně je H_0 pravděpodobná.

V případě platnosti nulové hypotézy má *F-poměr* Fisher – Snedecorovo rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.

Abychom test mohli dokončit, zbývá nám popsat způsob výpočtu *p-hodnoty*. Protože o zamítnutí H_0 vypovídají hodnoty kritéria *F-poměr* mnohem větší než 1, je zřejmé (viz obr. 8.4), že

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce Fisherova-Snedecorova rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.



Obr. 8.4: Ilustrace p-hodnoty pro testovou statistiku F-poměr

Pro úplnost lze dodat, že pokud bychom metodiku analýzy rozptylu uplatnili pro dvouvýběrový test shody středních hodnot, získali bychom výsledky stejné jako u oboustranného dvouvýběrového t testu. Metodou ANOVA však nelze provádět jednostranné testy shody středních hodnot, což dvouvýběrový t test umožňuje.

8.2.6. Tabulka ANOVA

Výsledky výpočtů se zapisují do tzv. tabulky jednofaktorové analýzy rozptylu.

Tab. 8.2: Tabulka jednofaktorové analýzy rozptylu

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Rozptyl (prům. součet čtverců)	*F*-poměr	*p*-hodnota
Skupinový (faktor)	$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$	$df_B = k - 1$	$MS_B = \frac{SS_B}{df_B}$	$\frac{MS_B}{MS_e}$	$1 - F_0(x_{OBS})$
Reziduální	$SS_e = \sum_{i=1}^{k} (n_i - 1)s_i^2$	$df_e = n - k$	$MS_e = \frac{SS_e}{df_e}$	---	---
Celkový	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$	$df_T = n - 1$	---	---	---

Příklad 8.3. Dokončete analýzu rozptylu pro motivační příklad.

Řešení 8.3.

Pomocné výpočty potřebné pro doplnění tabulky ANOVA a souvislosti mezi vztahem vnitrotřídní a mezitřídní variabilitou a rozhodnoutím o výsledku testu ANOVA můžete sledovat v java appletu ANOVA (460 KB).

8.2.7. Post hoc analýza aneb metody mnohonásobného porovnávání

V případě nezamítnutí nulové hypotézy je závěr jasný a testování končí. Pokud však zamítneme H_0 ve prospěch H_A, byla by naše analýza nekompletní, pokud bychom neidentifikovali, mezi kterými dvěma soubory existují statisticky významné rozdíly, kolik takových dvojic je a jaký je mezi nimi vztah. Tento další proces se nazývá post hoc analýza a spočívá v porovnávání středních hodnot všech dvojic populací, tzv. mnohonásobném porovnávání.

Obsah
312. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Metody mnohonásobného porovnávání středních hodnot vycházejí z testů shody dvou středních hodnot, které jste poznali v kapitole 7.2. Pro každou dvojici skupin I a J testujeme

$$H_0 : \mu_I = \mu_J$$

vůči alternativě

$$H_A : \mu_I \neq \mu_J$$

Zamítneme-li hypotézu H_0 znamená to, že skupiny I a J jsou rozlišitelné daným faktorem. Pro řešení problému mnohonásobného porovnávání existuje několik metod, jako například Fisherovo LSD (nejmenší významný rozdíl - Least Significant Difference), Bonferroniho, Scheffého a Tukeyova metoda. Cílem každé metody je udržet danou pravděpodobnost chyby prvního druhu α a v podstatě ji rozdělit mezi všechna porovnání.

Fisherovo LSD (metoda nejmenšího významného rozdílu)

Fisherovo LSD patří mezi nejstarší metody vícenásobného porovnávání. Jejím autorem se sir R. A. Fisher, autor analýzy rozptylu. Nulovou hypotézu zamítáme pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq LSD_{IJ},$$

kde LSD_{IJ} nazýváme nejmenší signifikantní diferencí (angl. Least Significant Difference) a určíme ji jako

$$LSD_{IJ} = t_{n-k}\left(1 - \frac{\alpha}{2}\right)\sqrt{MS_{\mathrm{e}}}\sqrt{\frac{1}{n_I} + \frac{1}{n_J}},$$

kde $t_{1-\frac{\alpha}{2}}(n-k)$ je $(1-\frac{\alpha}{2})$ kvantil Studentova rozdělení s $n-k$ stupni volnosti.

Nevýhodou metody je, že celková pravděpodobnost chyby I. druhu je vyšší (obvykle podstatně vyšší) než hladina významnosti α zvolená pro jednotlivá dílčí porovnávání dvojic. *(Jak určíme celkovou pravděpodobnost chyby prvního druhu, bude-li provedeno celkem* $\binom{k}{2}$ *porovnávání?*

Bonferroniho metoda aneb Fisherova metoda s Bonferroniho korekci

Italský matematik Bonferroni ukázal, že u Fisherova LSD s rostoucím počtem porovnávání roste pravděpodobnost, že se dopustíme chyby I. druhu. Aby bylo zajištěno, že celá post hoc analýza bude mít chybu I. druhu nejvýše α, je třeba v jednotlivých testech **upravenou hladinou významnosti $\boldsymbol{\alpha^*}$**. Tu získáme tak, že hladinu významnosti α vydělíme celkovým počtem $\binom{k}{2}$ porovnání, která chceme provést. Tato hodnota pak bude naší hladinou významnosti pro každý t test.

Nulovou hypotézu zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq t_{n-k}\left(1-\frac{\alpha^*}{2}\right)\sqrt{MS_\mathrm{e}}\sqrt{\frac{1}{n_I}+\frac{1}{n_J}},$$

kde α^* je upravená hladina významnosti, $\alpha^* = \frac{\alpha}{\binom{k}{2}}$,

$t_{1-\frac{\alpha^*}{2}}(n-k)$ je $(1-\frac{\alpha^*}{2})$ kvantil Studentova rozdělení s $n-k$ stupni volnosti.

Schefféhо metoda

Tato metoda je v praxi často preferována.

Nulovou hypotézu zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq \sqrt{MS_{\mathrm{e}}} \sqrt{F_{1-\alpha}(k-1, n-k)(k-1)\left(\frac{1}{n_I} + \frac{1}{n_J}\right)},$$

kde $F_{1-\alpha}(k-1, n-k)$ je $(1-\alpha)$ kvantil Fisherova-Snedecorova rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.

Tukeyho metoda

V případě **vyváženého třídění** (tj. stejného počtu pozorování u všech porovnávaných k skupin) lze pro post hoc analýzu použít Tukeyho metodu, která je sice méně obecnější než Scheffého metoda, ale zato citlivější.

Nulovou hypotézu zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq q_\alpha(k, n-k)\sqrt{MS_{\mathrm{e}}} \sqrt{\frac{1}{n_I}},$$

kde $q_\alpha(k, n-k)$ je α kvantil studentizovaného rozpětí, který je tabelován (tabulka T10).

V případě **nevyváženého třídění** lze použít modifikovaný Tukeyho test známý pod názvem **Tukey HSD**.

Nulovou hypotézu pak zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq q_\alpha(k, n-k)\sqrt{MS_\mathrm{e}}\sqrt{\frac{1}{2}\left(\frac{1}{n_I} + \frac{1}{n_J}\right)},$$

kde $q_\alpha(k, n-k)$ je α kvantil studentizovaného rozpětí, který je tabelován v T10.

8.2.8. Metody prezentace výsledků vícenásobného porovnávání

Pro souhrnnou a přehlednou prezentaci výsledků post hoc analýzy, zejména pro větší počet porovnávaných skupin, byly vyvinuty různé prostředky. S dvěma z nich se nyní seznámíme. Jsou to:

- znaménkové schéma,
- homogenní skupiny.

Znaménkové schéma (viz obr. 8.7) je tabulka $k \times k$, ve které každé porovnávané skupině odpovídá jeden řádek a jeden sloupec. V příslušném poli tabulky lze dohodnutým symbolem (tečka, křížek, hvězdička, ...) označit ty dvojice skupin, pro něž byl identifikován statisticky významný rozdíl mezi průměry. Chceme-li zdůraznit různé hladiny významnosti, na nichž lze rozdíl mezi průměry označit za statisticky významný, používáme obvykle pro různé hladiny významnosti různě velké skupiny znaků (např. jeden znak pro $\alpha = 0,05$, dva znaky pro $\alpha = 0,01$ a tři znaky pro $\alpha = 0,001$).

Jiným způsobem prezentace výsledků post hoc analýzy jsou tzv. **homogenní skupiny** (viz obr. 8.6). Jako homogenní označujeme ty skupiny, pro něž by v jednofaktorové analýze rozptylu nebyla zamítnuta hypotéza o shodě středních hodnot. Při tvorbě homogenních skupin se porovnávané skupiny seřadí do tabulky a to vzestupně podle výběrového průměru, tj. v prvním řádku bude skupina, jejíž průměr je nejmenší, v posledním řádku bude skupina s největším průměrem. Poté se pomocí vhodné metody mnohonásobného porovnávání ověřuje shoda mezi první z uvedených skupin a dalšími následujícími a to tak dlouho, dokud lze pro tyto hodnoty nezamítnout hypotézu o shodě středních hodnot. Tyto skupiny pak tvoří první homogenní skupinu. Dále se obdobným způsobem postupuje u dalších skupin v pořadí. Pokud by tímto postupem byla identifikována homogenní skupina, která je podmnožinou již vzniklé (větší) homogenní skupiny, pak se ve výsledku neuvažuje.

Poznámka: *Některé homogenní skupiny se mohou překrývat. Znamená to, že některé skupiny mohou mít vlastnosti blízké více homogenním skupinám současně.*

Příklad 8.4. Proveďte post hoc analýzu pro data z motivačního příkladu.

Řešení 8.4.

OU	x	
SPŠ	x	
Gymnázium		x

Obr. 8.5: Homogenní skupiny

	Gymnázium	SPŠ	OU
Gymnázium		x	x
SPŠ	x		
OU	x		

Obr. 8.6: Znaménkové schéma

Obsah
317. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Výsledky post hoc analýzy lze prezentovat pomocí znaménkového schématu (viz obr. 8.7) nebo pomocí homogenních skupin (viz obr. 8.6).

Na hladině významnosti 0,05 můžeme tvrdit, že absolventi gymnázií mají statisticky významně vyšší průměrné výsledky než studenti SPŠ a OU, jejichž průměrné výsledky jsou srovnatelné (viz obr. 8.6).

8.3. Kruskalův-Wallisův test

Tento test je neparametrickou obdobou jednofaktorové analýzy rozptylu, proto se mu někdy říká **neparametrická ANOVA**. Bývá používán tehdy, chceme-li srovnávat střední hodnoty více než dvou nezávislých souborů na základě výběrů nesplňujících předpoklady pro použití parametrické analýzy rozptylu (zejména normalitu).

Tak jako je analýza rozptylu vícevýběrovým testem shody středních hodnot, Kruskalův--Wallisův test je **vícevýběrovým testem shody mediánů**.

Nechť je dáno k nezávislých výběrů $X_{11}, X_{12}, \ldots, X_{1n_1}$ atd. až $X_{k1}, X_{k2}, \ldots, X_{kn_k}$ z rozdělení se spojitou distribuční funkci o rozsazích $n_1, n_2, \ldots, n_k$. Označme $n = n_1 + n_2 + \ldots + n_k$. Chceme testovat hypotézu o shodě mediánů

$$H_0: \quad x_{0,5_1} = x_{0,5_2} = \ldots = x_{0,5_k}$$

vůči alternativě, že H_0 neplatí.

Pro výpočet pozorované hodnoty testové statistiky se používá analogicky postup jako u Mannova-Whitneyova testu. Lze říci, že Kruskalů-Wallisův test je rozšířením Mannova-Whitneyova testu na více než 2 výběry. Všech n pozorovaných hodnot veličiny X_{ij} se seřadí do rostoucí posloupnosti a určí se jejich **pořadí** R_{ij}. Tato pořadí uspořádáme do tabulky a určíme tzv. **součty pořadí pro jednotlivé výběry** T_i.

Výběr	Pořadí veličin X_{ij} v uspořádané rostoucí posloupnosti				Součty pořadí
1	R_{11}	R_{12}	$\dots$	R_{1n_1}	T_1
2	R_{21}	R_{22}	$\dots$	R_{2n_2}	T_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	R_{k1}	R_{k2}	$\dots$	R_{kn_k}	T_k

Celkový součet všech pořadí je $T_1 + \ldots + T_k = \frac{n(n+1)}{2}$. Jako testová statistika se používá

$$Q = -3(n+1) + \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1).$$

Kritické hodnoty této statistiky jsou tabelovány ve speciálních tabulkách (nejsou součásti těchto skript). Jsou-li rozsahy jednotlivých výběru alespoň 5 prvků, má testová statistika Q v případě platnosti nulové hypotézy přibližně χ^2 rozdělení s $k-1$ stupni volnosti. Pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti.

8.3.1. Post hoc analýza pro Kruskalův-Wallisův test

Podobně jako u analýzy rozptylu, rovněž u Kruskalova-Wallisova testu nás v případě zamítnutí nulové hypotézy zajímá, která dvojice výběrů se od sebe statisticky významně liší. Pro mnohonásobné porovnávání se používá Dunnové metoda (viz Dunn, 1963).

Nechť průměrné pořadí i-té skupiny je $t_i = \frac{T_i}{n_i}$, $z_p \dots p$ kvantil normovaného normálního rozdělení, modifikovaná hladina významnosti je $\alpha^* = \frac{\alpha}{\binom{k}{2}}$. Jestliže

$$|t_I - t_J| \geqq \sqrt{\frac{1}{12}\left(\frac{1}{n_I} + \frac{1}{n_J}\right) n(n+1) z_{1-\alpha^*}},$$

pak se mediány I-tého a J-tého výběru statisticky významně liší.

V případě **vyváženého třídění** (všechny výběry mají týž rozsah, řekněme $n_1 = n_2 = \ldots = = n_k = m$), používáme pro post hoc analýzu Neméneyiovu metodu, která je citlivější než Dunnova metoda.

Neméneiova metoda (viz Neményi 1963 a Miller 1966)

Pro menší počty skupin k a rozsahy jednotlivých výběrů m jsou kritické hodnoty pro $|T_I - T_J|$ uvedeny v tabulce T11.

Je-li počet skupin $k > 10$ nebo rozsahy jednotlivých výběrů $m > 16$, užije se následující postup.

- Nechť $q_\alpha(k, \infty)$ je kritická hodnota rozpětí k nezávislých náhodných veličin s rozdělením $N(0; 1)$. Lze ji najít v posledním řádku tabulky …

- Řekneme, že se mediány I-tého a J-tého výběru statisticky významně liší, když

$$|t_I - t_J| \geqq q_\alpha(k,\infty)\sqrt{\frac{1}{12}k\ (km+1)}.$$

Výsledky post hoc analýzy Kruskalova-Wallisova testu lze prezentovat obdobně jako u parametrické jednofaktorové analýzy rozptylu, tj. pomocí znaménkového schématu, resp. pomocí homogenních skupin.

Pro výpočet Kruskalova-Wallisova testu, včetně post-hoc analýzy lze použít excelovský soubor Kruskalův-Wallisův test (4,4 MB). ***POZOR!!!*** *Statgraphics v. 5.0 post-hoc analýzu pro Kruskalův-Wallisův test neumožňuje provádět.*

Příklad 8.5. Analyzujte data z motivačního příkladu pomocí Kruskalova-Wallisova testu.

Řešení 8.5.

8.4. Friedmanův test

8.4.1. Motivační příklad

Basketbalové utkání je charakteristické plynulým průběhem hry s přechody z útoku do obrany a naopak. K testování výkonů basketbalistů slouží dané skupiny laboratorních i terénních testů. Při výzkumu byla sledována srdeční frekvence hráčů v průběhu utkání (viz tabulka 8.3). Zjistěte, zda se srdeční frekvence (tep) hráčů mění v průběhu utkání.

Tab. 8.3: Srdeční frekvence hráčů basketbalu v průběhu utkání

Srdeční frekvence [tep/min]				
Číslo hráče	Čtvrtina			
	1	2	3	4
1	163	166	177	183
2	160	170	180	180
3	189	180	188	190
4	182	180	183	185
5	170	175	177	190
6	153	169	166	180

Cílem této úlohy je porovnat úroveň spojité náhodné veličiny (srdeční frekvence) ve více než dvou (v našem případě ve čtyřech) výběrech. Je zřejmé, že analýza rozptylu není v tomto případě správnou volbou, neboť data, která máme analyzovat, jsou **závislá**. U každého hráče máme k dispozici uspořádanou čtveřici měření. K analýze úrovně spojité náhodné veličiny ve více než dvou závislých výběrech je určen Friedmanův test.

8.4.2. Friedmanův test

Friedmanův test, obdobně jako Kruskalův-Wallisův test, slouží k testování hypotézy o shodě mediánů více než dvou souborů. Na rozdíl od Kruskalova.Wallisova testu je však Friedmanův test určen pro porovnání výběrů **závislých**.

Nechť X_{IJ} jsou nezávislé náhodné veličiny se spojitými distribučními funkcemi F_{IJ} pro $i = 1, \ldots, m$, $j = 1, \ldots, k$. Nechť $x_{0{,}5_j}$ je medián j-té skupiny. Chceme testovat hypotézu

$$H_0 : x_{0,5_i} = \ldots = x_{0,5_k} \text{ neboli } F_{ij} \text{ nezávisí na } j$$

vůči alternativě

$$H_A : \neg H_0.$$

V našem případě tedy budeme testovat nulovou hypotézu, že srdeční tep se v průběhu utkání mění jen náhodně (zatímco u jednotlivých hráčů se může lišit) vůči alternativě, že nulová hypotéza neplatí.

Pro každé i zvlášť se určí pořadí R_{ij} veličiny X_{ij}. Jde tedy o pořadí mezi veličinami X_{i1} až X_{ik}. Označme součet pořadí j-tého výběru $R_j = \sum\limits_{i=1}^{m} R_{ij}$. Překročí-li pozorovaná hodnota testové statistiky

$$Q = -3m(k+1) + \frac{12}{mk(k+1)} \sum_{j=1}^{k} R_j^2$$

kritickou hodnotu (viz tabulka T12), zamítáme nulovou hypotézu. S rostoucím počtem porovnávaných skupin k a sledovaných objektů m(v praxi stačí, aby $\min(k; m) > 5$) lze nulové rozdělení testové statistiky Q aproximovat rozdělením χ^2 s $k-1$ stupni volnosti. Pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti.

Obsah
323. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

8.4.3. Post hoc analýza pro Friedmanův test

Zamítneme-li nulovou hypotézu, zajímá nás, pro které dvojice r a s se distribuční funkce F_{ir} a F_{is} významně liší.

Pro všechna $r < s$ testujeme hypotézu o rovnosti distribučních funkcí. Překročí-li $|R_r - R_s|$ kritickou hodnotu Friedmanova testu (tabulka T13), hypotézu o rovnosti $F_{ir} = F_{is}$ zamítneme.

Je-li počet porovnávaných skupin $k > 5$, lze kritické hodnoty Friedmanova testu určit jako

$$q_\alpha(k,\infty)\sqrt{\frac{1}{12}mk(k+1)},$$

kde $q_\alpha(k,\infty)$ je kritická hodnota rozpětí k nezávislých výběrů (kapitola 8.3.1) a lze ji najít v posledním řádku tabulky T10.

Pro výpočet Friedmanova testu, včetně post-hoc analýzy lze použít excelovský soubor Friedmanův test (200 KB). ***POZOR!!!*** *Statgraphic v. 5.0 post-hoc analýzu pro Friedmanův test neumožňuje provádět.*

Příklad 8.6. Při výzkumu byla sledována srdeční frekvence 6 hráčů basketbalu v průběhu utkání. Průměrné hodnoty srdeční frekvence [tep/min] v jednotlivých čtvrtinách utkání byly zaznamenány do tabulky 8.3, kterou zde pro přehlednost znovu uvedeme.

Srdeční frekvence [tep/min]				
Číslo hráče	Čtvrtina			
	1	2	3	4
1	163	166	177	183
2	160	170	180	180
3	189	180	188	190
4	182	180	183	185
5	170	175	177	190
6	153	169	166	180

Zjistěte, zda se srdeční frekvence (tep) hráčů mění v průběhu utkání.

Řešení 8.6.

Shrnutí:

Zobecněním dvouvýběrových t testů je analýza rozptylu neboli ANOVA (viz kapitola 8.2), která umožňuje srovnávat více než dvě střední hodnoty nezávislých náhodných výběrů. Analyzujeme tak vliv určitého faktoru A (nominální náhodné veličiny) na variabilitu pozorovaných hodnot spojité náhodné veličiny X.

Vstupem pro analýzu rozptylu je datová tabulka obsahující v j-tém sloupci vždy n_i pozorování X_{ij} ($i = 1, \ldots, n_i$, kde n_i je počet pozorování v jednotlivých výběrech, kterým se říká rovněž skupiny, resp. třídy. Přitom $j = 1, \ldots, k$, kde k je počet porovnávaných výběrů, neboli počet úrovní faktoru A).

Je třeba testovat hypotézu $H_0 : \mu_1 = \ldots = \mu_k$
vůči alternativě $H_A : \neg H_0$.

Už poctivou přípravou dat lze zajistit větší věrohodnost dosažených výsledků. ANOVA byla původně navržena pro stejný rozsah v jednotlivých výběrech. V praxi bývá tento předpoklad málokdy splněn - platí však, že čím více je zmíněné pravidlo naplněno, tím věrohodnější jsou výsledky.

Doporučený postup:

1) **Explorační analýza**: Prvním krokem při analýze rozptylu by měla být explorační analýza a s ní spojena vizualizace dat. Identifikujeme odlehlá pozorování, která obecně způsobují selhání analýzy rozptylu. Známe-li příčinu odlehlosti a předpokládáme-li, že již

nenastane, vyloučíme případná odlehlá pozorování z dalšího zpracování. Jestliže odlehlá pozorování v datech ponecháme, použijeme raději Kruskalův-Wallisův test.

2) **Ověření předpokladů**: Nestačí se soustředit na výsledky uvedené v tabulce ANOVA! Je třeba pečlivě ověřit splnění základních předpokladů pro použití analýzy rozptylu.

- **Nezávislost výběrů**: Pokud není tento předpoklad splněn, často dostaneme užitím analýzy rozptylu zcela nesmyslné výsledky. Pro porovnání $k > 2$ závislých výběrů lze použít Friedmanův test (viz kap. 8.4).
- **Normalita rozdělení**: Normalitu rozdělení lze ověřit pomocí některého z testů normality (kapitola 9)). Pokud data nemají ve všech výběrech normální rozdělení, je třeba použít vhodnou transformaci (mocninnou, logaritmickou). Vykazují-li data po transformaci normální rozdělení, přinese nám to větší důvěryhodnost výsledků. Na porušení normality není ANOVA příliš citlivá, zvlášť pokud mají všechny výběry rozsah větší než 30. Při výraznějším porušení normality (viz testy normality) se doporučuje použít neparametrickou obdobu analýzy rozptylu – Kruskalův - Wallisův test (kapitola 8.3).
- **Homoskedasticita** (identické rozptyly): Pro ověření homoskedasticity (shody rozptylů) lze použít například Barttletův nebo Leveneův test. (Pozor! Bartlettův test má větší sílu testu, je však citlivý na porušení normality. Proto v případě splnění předpokladu normality volíme Bartlettův test, v případě zamítnutí normality používáme test Leveneův.) V případě vyváženého třídění lze pro ověření homoskedasticity použít rovněž Hartleyův nebo Cochranův test (kapitola 8.1). Identifikujeme-li v datech heteroskedasticitu, pokusíme se rozptyl stabilizovat pomocí vhodné transformace (mocninné, logaritmické). Pokud dojde ke stabilizaci rozptylu, použijeme analýzu rozptylu na transformovaných datech. Při větším porušení homoskedasticity se doporučuje, podobně jako při porušení normality, použít Kruskalův – Wallisův test (kapitola 8.3).

3) **Post hoc analýza (vícenásobné porovnávání)**: Pokud při analýze rozptylu (popř. Kruskalově-Wallisově, resp. Friedmanově testu) došlo k zamítnutí nulové hypotézy, pokoušíme se pomocí vhodné metody vícenásobného porovnávání (kapitola 8.2.7, 8.3.1, 8.4.3) nalézt homogenní (srovnatelné) populace.

Testy o shodě rozptylů	
Název testu	Předpoklady testu
Bartlettův test	nezávislost a normalita výběrů
Leveneův test	nezávislost výběrů
Hartleyův test	nezávislost výběrů, vyváženost třídění
Cochranův test	nezávislost výběrů, vyváženost třídění

Testy o shodě úrovně			
Název testu	Předpoklady testu	Metoda vícenásobného porovnávání	Předpoklady pro použití metody vícenásobného porovnávání
Analýza rozptylu (ANOVA)	nezávislost, normalita a homoskedasticita výběrů (Pozor na odlehlá pozorování!)	Fisherovo LSD Bonferroniho metoda Schéffeho metoda Tukeyho metoda Tukey HSD	 vyváženost třídění
Kruskalův-Wallisův test	nezávislost výběrů	Dunnova metoda Neméneiova metoda	 vyváženost třídění
Friedmanův test	závislost výběrů	Friedmanova metoda	

Test

Jak pracovat s testy?

Zacátek testu

1. (8b.) Určete, zda jsou následující tvrzení pravdivá. (ANO, NE)

   (a) Jedním z předpokladů analýzy rozptylu je alespoň přibližná shoda rozptylů v jednotlivých skupinách.

   (a) ANO (b) NE

   (b) Reziduální rozptyl (v analýze rozptylu) lze určit jako aritmetické průměr rozptylů v jednotlivých skupinách.

   (a) ANO (b) NE

   (c) Post hoc analýza znamená, že stanovíme nejprve hypotézy H0, HA, a „následně“ provedeme řešení.

   (a) ANO (b) NE

   (d) Kruskalův-Wallisův test se nazývá rovněž neparametrická ANOVA.

   (a) ANO (b) NE

   (e) Hartleyův test homoskedasticity lze použít pouze v případě vyváženého třídění.

   (a) ANO (b) NE

   (f) Jediným předpokladem Leveneova testu je nezávislost výběrů.

   (a) ANO (b) NE

(g) Bartletův test je neparametrickým protějškem Leveneova testu.

(a) ANO (b) NE

(h) Friedmanův test je neparametrickou obdobou Kruskalova-Wallisova testu.

(a) ANO (b) NE

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1) Je třeba zjistit, zda se liší spotřeba automobilu při použití různých druhů benzínu. Zkouší se čtyři typy benzínu, jež se liší svým chemickým složením. Testovací jízdy se provádějí s 20 auty stejného modelu tak, že vždy pět aut použije stejný benzín. Výsledky měření spotřeby [l/100 km] při jednotlivých jízdách jsou zapsány v tabulce.

Typ benzínu			
A	B	C	D
6,7	7,1	7,3	9,1
7,4	8,0	8,3	9,4
6,9	6,9	6,5	9,7
7,5	7,2	7,6	9,7
6,9	7,6	8,5	9,3

Rozhodněte, zda složení benzínu ovlivňuje jeho spotřebu (na hladině významnosti 5%). Předpokládejte, že spotřeba benzínu má normální rozdělení.

Řešení příkladu

2) Byly srovnávány Lívance v prášku čtyř různých výrobců. Srovnávání probíhalo tak, že z každé směsi bylo upečeno 5 lívanců, které byly dány k ohodnocení 5-ti členné porotě. Výsledky hodnocení jsou uvedeny v tabulce.

Výrobce			
A	B	C	D
63	79	70	76
90	68	65	82
89	75	68	80
79	73	75	72

Rozhodněte, zda je rozdíl v kvalitě Lívanců v prášku od různých výrobců (na hladině významnosti 5%). Nelze předpokládat, že hodnocení poroty má normální rozdělení.

Řešení příkladu

3) Cílem experimentu je porovnat schopnost vidění v různých fázích dne. Náhodně bylo vybráno 11 osob a byly u nich provedeny zkoušky zrakových schopností ráno, v poledne, odpoledne a večer. Naměřené údaje byly zapsány do tabulky.

Id. číslo respondenta	Ráno	Poledne	Odpoledne	Večer
1	1	4	8	0
2	3	2	4	13
3	14	4	7	2
4	10	4	9	3
5	10	4	5	3
6	10	12	10	11
7	4	3	11	9
8	10	3	10	0
9	1	11	13	10
10	12	0	11	3
11	2	3	13	1

Zjistěte, zda se schopnost vidění v různých fázích dne mění.

Řešení příkladu

Kapitola 9

Testy dobré shody

Cíle

Po prostudování této kapitoly budete umět testovat shodu teoretického a empirického rozdělení, například normalitu.

9.1. Úvod

Domněnka o tom, že studovaná data (výběr) pocházejí z určitého teoretického (očekávaného) rozdělení bývá podložena buď informacemi o sledovaném jevu, nebo odhadem teoretického rozdělení na základě grafického zobrazení výběrového rozdělení. Náš odhad však nemusí být správný, a proto jej v praxi ověřujeme tzv. **testem dobré shody** (tj. shody mezi teoretickým a empirickým (pozorovaným, výběrovým) rozdělením. Nulovou a alternativní hypotézu můžeme v tomto případě formulovat:

H_0 : Teoretické a empirické rozdělení se **shoduje**.
H_A : Teoretické a empirické rozdělení se neshoduje.

Nejznámější z testů dobré shody, χ^2 - **test dobré shody** (angl. „Goodness of Fit test“), ověřuje, zda se empirické (pozorované, angl. „observed“) absolutní četnosti O_i jednotlivých variant náhodné veličiny shodují s očekávanými (angl. „expected“) absolutními četnostmi E_i, tj. četnostmi, které bychom očekávali v případě platnosti nulové hypotézy.

9.2. χ^2 - test dobré shody - ověření, zda jsou relativní četnosti jednotlivých variant rovny číslům $\pi_{0_1}; \ldots ;\pi_{0_k}$

V nejjednodušším případě lze konečnou populaci roztřídit podle nějakého znaku do k disjunktních skupin (tzv. variant) a my chceme na základě náhodného výběru ověřit, zda jsou relativní četnosti jednotlivých variant rovny číslům $\pi_{0_1}, \pi_{0_2}, \ldots, \pi_{0_k}$.

Jako testové kritérium se používá náhodná veličina

$$G = \frac{\sum\limits_{i=1}^{k}(O_i - E_i)^2}{E_i},$$

která má v případě platnosti nulové hypotézy a za předpokladu, že provádíme dostatečně velký výběr, přibližně χ^2 rozdělení s $k-1$ stupni volnosti.

Výběr považujeme za dostatečně velký, pokud jsou **všechny očekávané četnosti $\boldsymbol{E_i}$ větší než 5**. Pokud by předpoklad pro použití χ^2 testu dobré shody nebyl splněn, máme v podstatě dvě možnosti, jak mu vyhovět:

- můžeme rozšířit rozsah výběru tak, aby již byl tento předpoklad splněn,
- můžeme dodatečně sloučit varianty, které spolu věcně souvisí tak, aby nově vzniklé varianty již předpoklad testu splňovaly.

Je-li uvedený předpoklad splněn, pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti.

Příklad 9.1. Bylo provedeno šetření mezi ženami staršími 15 let. Mezi 246 náhodně oslovenými ženami bylo 80 (32,5%) svobodných, 110 (44,7%) vdaných, 30 (12,2%) rozvedených a 26 (10,6%) ovdovělých. Je známo (viz Český statistický úřad), že v ČR je mezi ženami staršími 15 let cca 24,8% svobodných, 49,0% vdaných, 12,6% rozvedených a 13,6% ovdovělých. Lze provedený výběr označit za reprezentativní?

Obsah
336. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Řešení 9.1.

9.3. χ^2 test dobré shody s očekávaným rozdělením

χ^2 test dobré shody nemusí být použit pouze pro ověření toho, zda jsou relativní četnosti jednotlivých variant rovny číslům $\pi_{0_1}, \pi_{0_2}, \dots, \pi_{0_k}$. Lze pomocí něj rovněž ověřit, zda výběr má rozdělení určitého typu (například normální). Připomeňme si, že chceme ověřovat nulovou hypotézu

H_0: Teoretické a empirické rozdělení se **shoduje**, neboli výběr **pochází** z určitého teoretického rozdělení.

vůči alternativě

H_A: Teoretické a empirické rozdělení se neshoduje, neboli není pravda, že výběr pochází z určitého teoretického rozdělení.

Chceme-li ověřovat, zda výběr **pochází z diskrétního rozdělení**, pak pro variantu x_i zjistíme empirickou četnost O_i a vypočteme pravděpodobnost π_{0_i}, že se náhodná veličina s pravděpodobnostní funkcí $P(x)$ odpovídající nulové hypotéze bude realizovat variantou x_i.

Ověřujeme-li, zda výběr **pochází z rozdělení spojitého**, pak je třeba nejprve testované rozdělení kategorizovat – tj. celý definiční obor testované náhodné veličiny rozdělit do k třídících intervalů a následně zjistit

- empirické četnosti O_i, tj. kolik realizací náhodné veličiny leží v daném intervalu,

- očekávané pravděpodobnosti π_{0_i}, tj. s jakou pravděpodobnosti bude za předpokladu platnosti nulové hypotézy náhodná veličina ležet v daném intervalu.

Očekávané četnosti jednotlivých variant, resp. třídících intervalů, pak určíme podle jednoduchého vztahu:

$$E_i = n\pi_{0_i},$$

kde n je rozsah výběru.

Pokud nulová hypotéza udává nejen typ rozdělení, ale i všechny jeho parametry, jde o **úplně specifikovaný test**. Příkladem úplně specifikovaného testu může být například ověření toho, zda výběr pochází z Poissonova rozdělení se střední hodnotou 10 (Poissonovo rozdělení má jeden parametr λt, který je roven střední hodnotě). V mnoha případech nás však zajímá pouze to, zda výběr pochází z určité třídy rozdělení – například z rozdělení normálního. Je-li v nulové hypotéze dán pouze typ rozdělení, resp. nejsou-li zadány všechny parametry rozdělení, mluvíme o neúplně specifikovaném testu. V případě **neúplně specifikovaného testu** je třeba nespecifikované parametry očekávaného rozdělení odhadnout pomocí náhodného výběru. Počet odhadovaných parametrů pak označíme h.

Jako testové kritérium používáme již známou náhodnou veličinu

$$G = \frac{\sum\limits_{i=1}^{k}(O_i - E_i)^2}{E_i},$$

která má v případě platnosti nulové hypotézy a za předpokladu, že provádíme dostatečně velký výběr (výběr považujeme za dostatečně velký, pokud jsou **všechny očekávané četnosti $\boldsymbol{E_i}$ větší než 5**) přibližně χ^2 rozdělení s $k - 1 - h$ stupni volnosti. *Všimněte si,* *že každý nespecifikovaný parametr rozdělení, který musíme odhadovat pomocí výběrového souboru, snižuje stupeň volnosti rozdělení testového kritéria o 1.*

Pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k - 1 - h$ stupni volnosti.

V následující animaci si pomocí krokovaného příkladu můžete ověřit, zda dokážete v praxi použít χ^2-test dobré shody. Následně se pokuste samostatně vyřešit příklady 9.2 a 9.3.

Testy dobré shody -Řešený příklad

Instrukce pro spuštění vložených animací naleznete na http://mi21.vsb.cz/adobe-reader-multimedia.

Příklad 9.2. Výrobní firma odhaduje počet poruch určitého zařízení během dne pomocí Poissonova rozdělení se střední hodnotou 1,2. Zaměstnanci zaznamenali pro kontrolu skutečné počty poruch celkem ve 150 dnech (výsledky jsou uvedeny v tabulce 9.3). Ověřte čistým testem významnosti, zda lze počet poruch daného zařízení během dne skutečně modelovat pomocí Poissonova rozdělení s parametrem $\lambda t = 1,2$.

Tab. 9.1: Pozorované četnosti počtu poruch během dne (za 150 dní celkem)

x_i – počet poruch během dne	0	1	2	3	4 a více
O_i – počet dní, v nichž byl pozorován počet poruch x_i	52	48	36	10	4

Řešení 9.2.

Příklad 9.3. Na dálnici byly v průběhu několika minut měřeny časové odstupy $[s]$ mezi průjezdy jednotlivých vozidel. Zjištěné hodnoty těchto odstupů jsou uvedeny v tabulce:

2,5	6,8	5,0	9,8	4,0	2,3	4,2	1,9	8,7	7,7	5,9	5,3	8,4	3,6	9,2
4,3	2,6	13,0	5,4	8,6	4,2	2,9	1,5	1,8	1,6	5,9	8,3	5,2	6,9	5,1
1,3	6,4	6,5	5,7	3,6	4,8	4,0	7,3	24,9	10,6	15,0	5,3	4,0	3,3	6,0
4,6	1,6	1,9	1,5	11,1	4,3	5,5	2,1	2,9	3,0	3,8	1,0	1,5	8,6	4,4
6,8	5,2	3,0	8,0	4,0	4,7	7,3	2,3	1,9	1,9	4,6	6,4	5,3	3,9	2,4
1,2	6,2	4,3	2,6	2,7	2,0	0,8	3,7	6,9	2,8	4,3	4,9	4,1	4,5	4,4
11,9	9,0	5,6	4,8	2,8	2,1	4,3	1,0	1,6	2,5	2,2	1,3	1,8	1,6	3,8
3,1	1,6	4,9	1,8	3,9	3,4	1,6	4,5	5,8	6,9	1,8	2,6	6,8	2,5	1,9
3,1	10,8	1,6	2,0	4,9	11,2	1,6	2,2	3,8	1,1	1,8	1,4			

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Obsah
341. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Ověřte čistým testem významnosti, zda lze časové odstupy mezi vozidly modelovat pomocí náhodné veličinu s normálním rozdělením.

Řešení 9.3.

9.4. Kolmogorovův – Smirnovův jednovýběrový test

Kolmogorovův – Smirnovův test se používá k ověření hypotézy, zda pořízený **výběr pochází z rozdělení se spojitou distribuční funkcí $\boldsymbol{F_0(x)}$**.

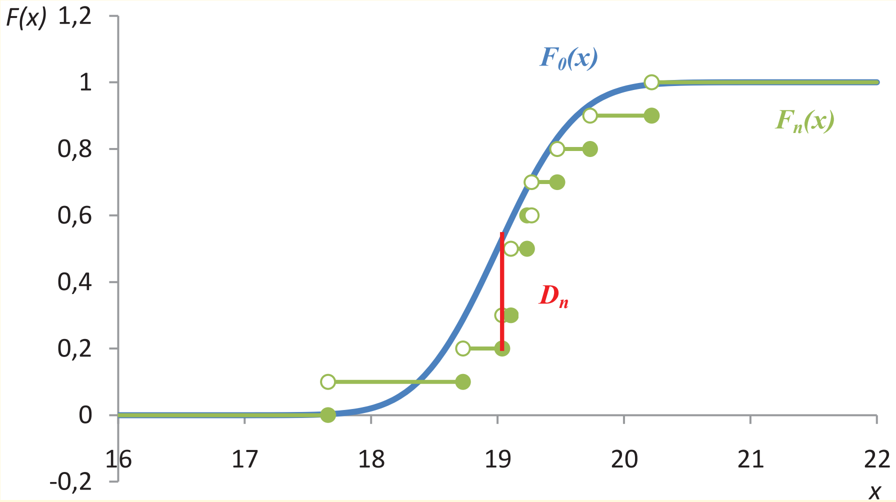
H_0 : Náhodný výběr **pochází** z rozdělení se spojitou distribuční funkcí $F_0(x)$.
H_A : Náhodný výběr nepochází z rozdělení se spojitou distribuční funkcí $F_0(x)$.

Mějme náhodný výběr $X_1, \ldots, X_n$ z rozdělení se spojitou distribuční funkcí. Nechť $X_{(1)} \leqq \leqq X_{(2)} \leqq \ldots \leqq X_{(n)}$ je tentýž náhodný výběr uspořádaný vzestupně podle velikosti. **Empirická (výběrová) distribuční funkce** $F_n(x)$ je pak dána vztahem

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ i/n, & X_{(i)} \leqq x \leqq X_{(i+1)}, i = 1, \ldots, n-1 \\ 1, & x \geqq X_{(n)}. \end{cases}$$

Jako testové kritérium použijeme statistiku D_n. Testová statistika D_n je definována jako maximální odchylka teoretické a empirické distribuční funkce (viz obr. 9.1).

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F_0(x)| = \max(D_1^*, D_2^*, \ldots, D_n^*),$$



Obr. 9.1: Grafická prezentace testové statistiky Kolmogorovova-Smirnovova testu

$$\text{kde } D_i^* = \max\left\{\left|F_0(x_i) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0(x)\right|\right\} \quad \text{pro } i = 1, 2, \ldots, n.$$

Nulovou hypotézu zamítáme, pokud pozorovaná hodnota testové statistiky D_n překročí kritickou hodnotu $D_{n(\alpha)}$. Je-li n malé, používáme pro určení kritických hodnot speciální tabulky kritických hodnot $D_{n(\alpha)}$. Při velkých hodnotách n se kritické hodnoty $D_{n(\alpha)}$ aproximují pomocí vztahu

$$D_{n(\alpha)} \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}.$$

Obsah
343. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

POZOR!
Je třeba zdůraznit, že nulová hypotéza H_0 musí distribuční funkci $F(x)$ určovat jednoznačně, včetně jejích případných parametrů. Říkáme, že ***distribuční funkce $F(x)$ musí být úplně specifikována****. Kolmogorovův-Smirnovův test tedy lze použít například k ověření, zda výběr pochází z rovnoměrného rozdělení $R(0;1)$, což se hodí například při testování generátorů náhodných čísel. Pokud však parametry distribuční funkce odhadujeme z výběru (testujeme-li například, zda výběr pochází z normálního rozdělení), změní se rozdělení testové statistiky D_n. Modifikované kritické hodnoty byly určeny pomocí simulačních studií, v těchto skriptech však nejsou uvedeny.*

Kolmogorovovu-Smirnovovu testu dáváme přednost před úplně specifikovaným χ^2 testem dobré shody. Má totiž větší sílu testu a v případě, že máme k dispozici pouze výběr malého rozsahu, vyhneme se komplikacím spojeným s omezujícím předpokladem χ^2 testu.

V následující animaci si pomocí krokovaného příkladu můžete ověřit, zda dokážete v praxi použít jednovýběrový Kolmogorovův-Smirnovův test. Následně se pokuste samostatně vyřešit příklady 9.4.

Kolmogorovův -Smirnovův test - Řešený příklad

Instrukce pro spuštění vložených animací naleznete na http://mi21.vsb.cz/adobe-reader-multimedia.

Příklad 9.4. V tabulce je 10 čísel generovaných jako hodnoty rozdělení $N(19; 0,49)$. Ověřte, zda generované hodnoty pocházejí z předpokládaného rozdělení.

Generované hodnoty x_i	19,732	19,108	19,234	19,038	19,270	19,105	19,473	17,660	20,219	18,727

Řešení 9.4.

Shrnutí:

Statistickou metodou umožňující ověřit, zda má náhodná veličina určité předem dané (tzv. teoretické) rozdělení pravděpodobnosti jsou **testy dobré shody**. Teoretické rozdělení může být dáno

- včetně parametrů (úplně specifikovaný test),
- s neznámými parametry (neúplně specifikovaný test, počet nespecifikovaných parametrů označujeme h).

Nulovou a alternativní hypotézu můžeme v tomto případě formulovat:

H_0: Teoretické a empirické (výběrové) rozdělení se **shoduje**.
H_A: Teoretické a empirické rozdělení se neshoduje.

Nejznámější z testů dobré shody, χ^2 - **test dobré shody**, používáme pro

- ověření, zda jsou relativní četnosti jednotlivých variant rovny číslům $\pi_{0_1}, \ldots, \pi_{0_k}$,
- ověření shody s očekávaným rozdělením.

Ověřujeme-li, zda výběr pochází z rozdělení spojitého, je třeba nejprve testované rozdělení kategorizovat – tj. celý definiční obor testované náhodné veličiny rozdělit do k třídících intervalů a následně zjistit

- empirické četnosti O_i,
- očekávané pravděpodobnosti π_{0_i}.

Očekávané četnosti jednotlivých variant, resp. třídících intervalů, pak určíme podle jednoduchého vztahu $E_i = n\pi_{0_i}$, kde n je rozsah výběru.

Jako testové kritérium používáme náhodnou veličinu

$$G = \frac{\sum\limits_{i=1}^{k}(O_i - E_i)^2}{E_i},$$

která má v případě platnosti nulové hypotézy a za předpokladu, že provádíme dostatečně velký výběr (výběr považujeme za dostatečně velký, pokud jsou **všechny očekávané četnosti větší než 5**) přibližně χ^2 rozdělení s $k-1-h$ stupni volnosti. Pak

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1-h$ stupni volnosti.

Před úplně specifikovaným χ^2 testem dobré shody se spojitým rozdělením dáváme přednost **Kolmogorovovu-Smirnovovu testu**. Má totiž větší sílu testu a v případě, že máme k dispozici pouze výběr malého rozsahu, vyhneme se komplikacím spojeným s omezujícím předpokladem χ^2 testu.

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Lze Kolmogorovův-Smirnovův test použít pro testování normality?
   (a) ANO (b) NE

2. (1b.) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika χ^2 rozdělení s
   (a) 4 stupni volnosti, (b) 5 stupni volnosti, (c) 6 stupni volnosti.

3. (1b.) Jak postupujeme v případě, kdy při použití χ^2 testu dobré shody vyšlo po rozdělení dat do dvaceti tříd 7 očekávaných třídních četností rovných 1?
   (a) V tomto případě nelze χ^2 testem dobré shody rozhodnout,
   (b) Prohlásíme předpoklady testu za splněné,
   (c) Sloučíme příslušné sousední třídy.

4. (1b.) Empirická distribuční funkce je funkce
   (a) diskrétní, (b) spojitá,
   (c) zleva spojitá, (d) zprava spojitá.

5. (1b.) Pomocí Kolmogorovova-Smirnovova testu lze testovat hypotézu, že náhodný výběr pochází z

I. normálního rozdělení,

II. Poissonova rozdělení,

III. normálního rozdělení se střední hodnotou 10,

IV. rovnoměrného rozdělení $R(0; 10)$.

Které z výše uvedených výroků jsou pravdivé?

(a) pouze I, (b) pouze II, (c) pouze III,
(d) pouze IV, (e) I a II, (f) III a IV,
(g) všechny .

6. (1b.) Pomocí χ^2 testu dobré shody lze testovat hypotézu, že náhodný výběr pochází z

I. normálního rozdělení,

II. Poissonova rozdělení,

III. normálního rozdělení se střední hodnotou 10,

IV. rovnoměrného rozdělení $R(0; 10)$.

Které z výše uvedených výroků jsou pravdivé?

(a) pouze I, (b) pouze II, (c) pouze III,
(d) pouze IV, (e) I a II, (f) III a IV,
(g) všechny .

7. (1b.) Testy normality (H_0 : Náhodný výběr pochází z normálního rozdělení.) patří mezi

(a) parametrické testy,
(b) úplně specifikované testy dobré shody,
(c) neúplně specifikované testy dobré shody.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1. Hodilo se 6000 krát hrací kostkou a zaznamenaly se počty padlých ok.

x_i (číslo které padlo)	1	2	3	4	5	6
n_i (četnost jeho výskytu)	979	1002	1015	980	1040	984

   Je možné na základě příslušného testu na hladině významnosti 0,05% spolehlivě tvrdit, že kostka není „férová“, tj. že pravděpodobnosti všech čísel na kostce nejsou stejné?

   Řešení příkladu

2. Pro ověření, zda generátor náhodných čísel z rovnoměrného rozdělení na intervalu $\langle 0; 1\rangle$ opravdu generuje výběr z tohoto rozdělení, bylo pomocí něj vygenerováno 1 000 čísel, která byla následně setříděna do deseti intervalů. Výsledky jsou v tabulce:

interval	četnost
$\langle 0{,}0; 0{,}1)$	89
$\langle 0{,}1; 0{,}2)$	91
$\langle 0{,}2; 0{,}3)$	74
$\langle 0{,}3; 0{,}4)$	97
$\langle 0{,}4; 0{,}5)$	99
$\langle 0{,}5; 0{,}6)$	106
$\langle 0{,}6; 0{,}7)$	123
$\langle 0{,}7; 0{,}8)$	100
$\langle 0{,}8; 0{,}9)$	110
$\langle 0{,}9; 1{,}0)$	111

Zjistěte, zda je možné na základě tohoto pokusu spolehlivě (na hladině významnosti 0,05) prohlásit, že generátor pracuje špatně, tj. že negeneruje náhodná čísla s rovnoměrným rozdělením na intervalu $\langle 0; 1\rangle$.

Řešení příkladu

3. Při testování nového typu výškoměru byly zaznamenávány chyby měření $[mm]$, tj. odchylky zjištěné a skutečné výšky. Přístrojem se opakovaně provedlo mnoho měření výšky jisté budovy. Výsledky jsou zaznamenány v následující tabulce.

interval	četnost
$(-\infty;-2)$	25
$\langle -2;-1)$	25
$\langle -1;0)$	40
$\langle 0;1)$	60
$\langle 1;2)$	20
$\langle 2;\infty)$	20

Ověřte na hladině významnosti 0,05, zda má chyba měření rozdělení dané hustotou pravděpodobnosti $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, x \in \mathbb{R}$.

Řešení příkladu

4. Při testování nového typu výškoměru byly zaznamenávány chyby měření $[mm]$, tj. odchylky zjištěné a skutečné výšky. Přístrojem se opakovaně provedlo mnoho měření výšky jedné budovy. Výsledky jsou zaznamenány v následující tabulce.

-1,7	0,8	0,6	-0,2	1,3	2,3	-2,1	0,5	-0,2	-1,1

Ověřte na hladině významnosti 0,05, zda má chyba měření rozdělení dané hustotou pravděpodobnosti $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, x \in \mathbb{R}$.

Řešení příkladu

Obsah
354. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Kapitola 10

Analýza závislostí

Cíle

Po prostudování této kapitoly budete umět analyzovat:

- závislost v kontingenčních a asociačních tabulkách,
- závislost v normálním rozdělení,
- závislost ordinálních veličin.

V praxi často u statistických jednotek (pozorovaných osob nebo jiných objektů) zjišťujeme současně řadu znaků. Například

- spotřeba, objem motoru, hmotnost a zrychlení automobilů,
- výše mzdy, velikost IQ, hmotnost a výška mužů,
- školní prospěch a pocit deprese u dětí, apod.

Jednotlivé znaky pak můžeme analyzovat metodami, s nimiž jsme se seznámili v předchozích kapitolách. Většinou však jednotlivé znaky nestudujeme jako takové, zajímají nás především jejich vazby k jiným znakům. Například nás může zajímat, zda existuje závislost mezi spotřebou automobilu a jeho hmotností, výši mzdy a velikostí IQ, pocitem deprese u dětí a školním prospěchu.

V případě, že znak X působí na znak Y, avšak znak Y již nepůsobí zpětně na znak X, mluvíme o **jednostranné závislosti**. Příkladem jednostranné závislosti může být vztah mezi typem absolvované střední školy a (bodovým) výsledkem přijímací zkoušky z matematiky nebo vztah mezi výškou a váhou.

Metody analýzy jednostranné závislosti popsané v tomto studijním materiálu jsou uvedeny v tabulce 10.
Pokud v analyzovaném vztahu nelze jednoznačně určit příčinu a důsledek, tzn. pokud znak X ovlivňuje znak Y a znak Y zpětně působí na znak X, hovoříme o **závislosti oboustranné**. (Například: vztah mezi výdaji domácností na oblečení a na potraviny.) V této kapitole se seznámíme se základními metodami analýzy oboustranné závislosti – vymezíme si metody pro analýzu síly vazeb mezi dvojicemi znaků, tj. metody pro analýzu síly závislostí dvojic náhodných veličin.

Tab. 10.1: Metody analýzy jednostranné závislosti

		Typ znaku Y (důsledek)	
		kategoriální	kvantitativní
Typ znaku X (příčina)	kategoriální		ANOVA (kapitola 13)
	kvantitativní		regresní a korelační analýza (kapitola 16)

Výběr vhodné metody závisí na typu analyzovaných veličin. V tabulce 10 jsou uvedeny jednotlivé metody analýzy závislostí pro různé typy dat.

Tab. 10.2: Metody analýzy oboustranné závislostí

		Typ znaku Y		
		kategoriální	ordinální	kvantitativní
Typ znaku X	kategoriální	analýza záv. v kontingenčních tabulkách, analýza záv. v asociačních tabulkách		
	ordinální		analýza závislostí ordinálních znaků	
	kvantitativní			analýza závislosti v normálním rozdělení

Obsah
357. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

10.1. Analýza závislostí v kontingenčních tabulkách

10.1.1. Motivační příklad

Analýzou dat v kontingenční tabulce nás provede následující příklad.

Příklad 10.1. Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Šetření se účastnilo 100 pracovníků z Prahy a 200 pracovníků z venkova. Výsledky šetření jsou v následující tabulce.

místo/stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
Praha	10	25	50	15
Venkov	20	10	130	40

Výsledky šetření analyzujte.

10.1.2. Základní pojmy

Výsledky šetření jsou uvedeny v tzv. kontingenční tabulce. **Kontingenční tabulka** vzniká setříděním prvků výběru podle variant dvou kategoriálních znaků, např. znaku X a znaku Y. Nechť znak X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a znak Y nabývá variant $y_{[1]}, \ldots, y_{[s]}$. V kontingenční tabulce jsou uspořádány absolutní četnosti n_{ij} dvojice variant $\left(x_{[i]}, y_{[j]}\right)$, přičemž názvy jednotlivých variant znaků X a Y jsou uvedeny v hlavičce tabulky.

Tab. 10.3: Schéma kontingenční tabulky

$X\backslash Y$	$y_{[1]}$	$y_{[2]}$	$\cdots$	$y_{[s]}$
$x_{[1]}$	n_{11}	n_{12}	$\cdots$	n_{1s}
$x_{[2]}$	n_{21}	n_{22}	$\cdots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$x_{[r]}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}

Pokud lze mezi analyzovanými znaky X a Y pozorovat kauzalitu (příčinnou souvislost), volíme označení X pro nezávislý znak a označení Y pro znak závislý. *(Všimněte si, že v motivačním příkladu jsme jako znak X, tj. znak jehož varianty jsou identifikátory řádků, zvolili umístění podniku...)*

Kontingenční tabulku často rozšiřujeme o další zajímavé číselné charakteristiky, jejichž výpočet pro data z motivačního příkladu můžete sledovat v tabulce 10.5.

- **Marginální četnosti**, které udávají celkové četnosti jednotlivých variant znaku X, resp. znaku Y. Marginální četnosti označujeme

  $$n_{(i\cdot)} \ldots \text{ součet všech četností v } i\text{-té řádce,}$$
  $$n_{(\cdot j)} \ldots \text{ součet všech četností v } j\text{-tém sloupci}$$

  a zapisujeme je na okraj kontingenční tabulky (viz tabulka 10.1.2).
- **Celkový rozsah výběru** n
- **Relativní četnosti**, které pro každé pole rozšířené kontingenční tabulky určíme jako podíl příslušné absolutní četnosti a celkového rozsahu výběru n. (Např.: Z celkového počtu

Tab. 10.4: Schéma rozšířené kontingenční tabulky

$X\backslash Y$	$y_{[1]}$	$y_{[2]}$	$\cdots$	$y_{[s]}$	Celkem
$x_{[1]}$	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$x_{[2]}$	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
Celkem	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	n

300 respondentů bylo 5,0 % velmi spokojených respondentů zaměstnaných v Praze.)

- **Řádkové rel. četnosti**, které udávají relativní četnosti znaku Y za předpokladu, že znak X nabývá určité varianty. Určujeme je jako podíl příslušné absolutní četnosti a marginální četnosti v odpovídajícím řádku. (Např.: Ze všech v Praze zaměstnaných respondentů bylo 10,0 % velmi nespokojených.)
- **Sloupcové rel. četnosti**, které udávají relativní četnosti znaku X za předpokladu, že znak Y nabývá určité varianty. Určujeme je jako podíl příslušné absolutní četnosti a marginální četnosti v odpovídajícím sloupci. (Např. Ze všech velmi spokojených respondentů je 20,0 % zaměstnaných na venkově.)

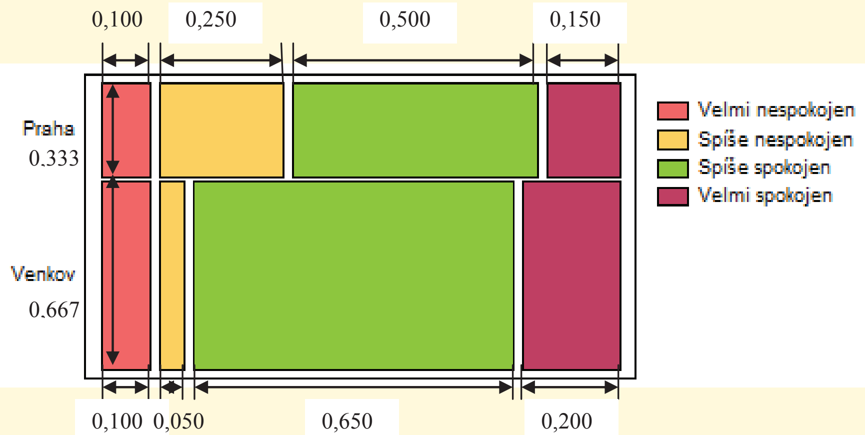
Grafickou obdobou kontingenční tabulky je **mozaikový graf**. Mozaikový graf se skládá z r řad obdélníků, přičemž r je počet variant (nezávislého) znaku X. (V našem případě $r = 2$.) Každá řada obsahuje s obdélníků, přičemž s je počet variant (závislého) znaku Y. (V našem případě $s = 4$.) Výšky jednotlivých řad obdélníků odpovídají příslušným marginálním relativním četnostem. Šířky obdélníků v jednotlivých řadách odpovídají příslušným řádkovým relativním četnostem (viz obr. 10.1.2).

Tab. 10.5: Rozšířená kontingenční tabulka pro data z motivačního příkladu (pozorované četnosti, celkový rozsah výběru, marginální četnosti, relativní četnosti, řádkové rel. četnosti, sloupcové rel. četnosti)

místo/stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen	celkem
Praha	10 0,033 (10/300) 0,100 (10/100) 0,333 (10/30)	25 0,083 (25/300) 0,250 (25/100) 0,714 (25/35)	50 0,167 (50/300) 0,500 (50/100) 0,278 (50/180)	15 0,050 (15/300) 0,150 (15/100) 0,273 (15/55)	100 0,333 (100/300)
venkov	20 0,067 (20/300) 0,100 (20/200) 0,667 (20/30)	10 0,033 (10/300) 0,050 (10/200) 0,286 (10/35)	130 0,433 (130/300) 0,650 (130/200) 0,722 (130/180)	40 0,133 (40/300) 0,200 (40/200) 0,727 (40/55)	200 0,067 (200/300)
celkem	30 0,100 (30/300)	35 0,117 (35/300)	180 0,600 (180/300)	55 0,183 (55/300)	300

Pokud by byl mozaikový graf v tomto případě tvořen svislými pruhy (jednotlivé obdélníky stejných barev by měly stejné šířky), znamenalo by to, že sledované znaky jsou nezávislé. Čím je mozaikový graf členitější, tím silnější závislost mezi znaky X a Y lze předpokládat. Dle obr. 10.1 lze předpokládat, že spokojenost v práci závisí na umístění závodu. *(Podívejte se znovu na obr. 10.1 a zvažte, jaký následek by mělo sloučení variant „spíše nespokojen“ a „spíše spokojen“.)*
Obdobou mozaikového grafu je **100% skládaný pruhový graf** (např. MS Excel). Od mozaikového grafu se tento graf liší tím, že šířky všech řádků jsou stejné, tzn. že tento typ grafu nezohledňuje řádkové marginální relativní četnosti.

Kromě mozaikového grafu se pro prezentaci dat zapsaných v kontingenční tabulce používají



Obr. 10.1: Mozaikový graf pro data z motivačního příkladu

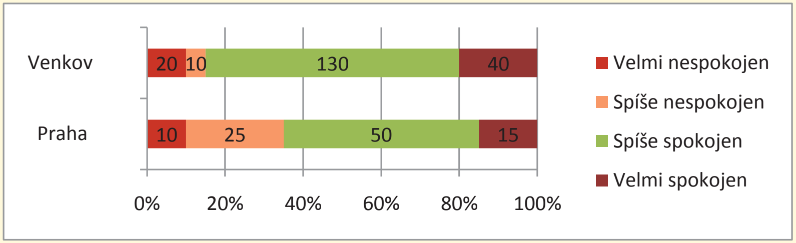
shlukový, popř. **kumulativní sloupcový graf** prezentované na obr. 10.3.

10.1.3. χ^2 test nezávislosti v kontingenční tabulce

Na základě explorační analýzy jsme v předcházející kapitole vyslovili domněnku, že stupeň spokojenosti v práci závisí na umístění podniku. Chceme-li takovou domněnku zobecnit na celou dotčenou populaci, lze testovat nulovou hypotézu

H_0 : Znaky X a Y v kontingenční tabulce jsou statisticky **nezávislé**

vůči alternativě



Obr. 10.2: 100% skládaný pruhový graf

H_A : Znaky X a Y v kontingenční tabulce jsou statisticky **závislé**.

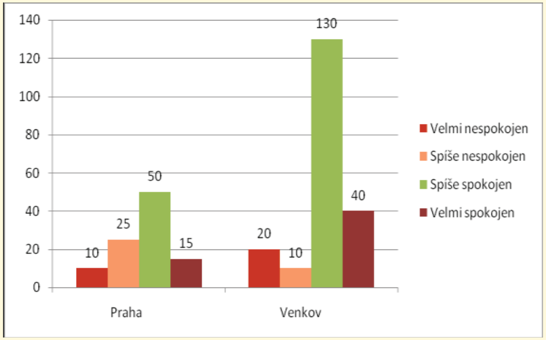
Pro ověření nezávislosti náhodných veličin X a Y (nezávislosti v kontingenční tabulce) používáme nejčastěji χ^2 test nezávislosti v kontingenční tabulce, který je, podobně jako χ^2 test dobré shody, založen na **porovnávání empirických** (pozorovaných) **četností s četnostmi teoretickými**, tj. takovými, které bychom očekávali v případě nezávislosti znaků X a Y.

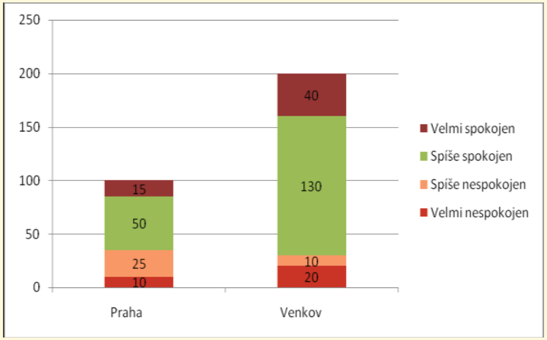
Označme empirické četnosti O_{ij}.

$$O_{ij} = n_{ij}$$

Očekávané četnosti E_{ij} určujeme jako četnosti odpovídající součinu příslušných marginálních relativních četností *(připomeňme si, že v případě, že jsou dvě diskrétní náhodné veličiny nezávislé, pak jejich sdružené pravděpodobnosti jsou rovny součinu příslušných marginálních pravděpodobností).*

$$E_{ij} = \left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right) \cdot n = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$





Obr. 10.3: Shlukový a kumulativní sloupcový graf

Jako testové kritérium používáme náhodnou veličinu

$$K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

která má v případě platnosti nulové hypotézy a za předpokladu splnění podmínek dobré aproximace přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

Podmínky dobré aproximace:

- žádná z očekávaných četností E_{ij} nesmí být menší než 2,
- alespoň 80 % očekávaných četností E_{ij} musí být větších než 5.

Obsah
364. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Jsou-li splněny podmínky dobré aproximace, pak

$$p\text{-}hodnota = 1 - F_0\left(x_{OBS}\right),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

10.1.4. Yatesova korekce χ^2 testu nezávislosti v kontingenční tabulce

V případě, že nejsou splněny podmínky dobré aproximace nutné pro použití χ^2 testu nezávislosti v kontingenční tabulce, tzn. že máme extrémně nízké očekávané četnosti, lze použít tzv. Yatesovu korekci. Efektem této korekce je snížení pozorované hodnoty testového kritéria, což znamená, že je obtížnější zamítnout nulovou hypotézu. Snížíme tak pravděpodobnost chyby I. druhu, chyba II. druhu se však zvýší – test tedy má menší sílu (oproti χ^2 testu nezávislosti).

Jako testové kritérium používáme náhodnou veličinu

$$K_{Yates} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij} - E_{ij} - 0,5\right)^2}{E_{ij}},$$

která má v případě platnosti nulové hypotézy přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti. Pak

$$p\text{-}hodnota = 1 - F_0\left(x_{OBS}\right),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

10.1.5. Měření síly závislosti

Musíme si uvědomit, že χ^2 test nezávislosti nevypovídá nic o síle vztahu, pouze zamítá, resp. nezamítá nulovou hypotézu o nezávislosti znaků X a Y. Pro zjištění síly vztahu používáme různé koeficienty. Mírou těsnosti závislosti obdobnou korelačnímu koeficientu je **koeficient kontingence**

$$CC = \sqrt{\frac{K}{K+n}}.$$

Koeficient kontingence se pro čtvercové kontingenční tabulky ($r = s$) může vyskytovat v intervalu $\langle 0; 1)$. Pro obdélníkové kontingenční tabulky ($r \neq s$) je však maximální hodnota koeficientu kontingence

$$CC_{max} = \sqrt{\frac{\min(r;s)-1}{\min(r;s)}},$$

proto se pro ně používá **korigovaný koeficient kontingence** (exaktní korekce do intervalu $\langle 0; 1 \rangle$)

$$CC_{cor} = \frac{CC}{CC_{max}}$$

Další často používanou mírou těsnosti závislosti je Cramerův koeficient nazývaný též Cramerovo V.

$$V = \sqrt{\frac{K}{n\left(\min(r;s)-1\right)}}$$

Rovněž Cramerův koeficient se může vyskytovat v intervalu $\langle 0; 1 \rangle$. Čím jsou tyto koeficienty blíže 1, tím je závislost mezi X a Y těsnější.

V následující animaci se pomocí krokovaného příkladu můžete přesvědčit, zda dokážete v praxi ověřit nezávislost dvou kategoriálních veličin. Následně se pokuste samostatně vyřešit příklad 10.2.

Analýza závislosti dvou kategoriálních veličin

Instrukce pro spuštění vložených animací naleznete na *http://mi21.vsb.cz/adobe-reader-multimedia*.

Příklad 10.2. Vraťme se nyní k našemu motivačnímu příkladu.
Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Výsledky šetření jsou v následující tabulce.

místo/stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
Praha	10	25	50	15
Venkov	20	10	130	40

Na základě explorační analýzy (rozšířená kontingenční tabulka, mozaikový graf) jsme vyslovili předpoklad, že spokojenost v práci závisí na umístění závodu. Ověřte tento předpoklad

Řešení 10.2.

10.2. Analýza závislostí v asociačních tabulkách

Speciálním typem kontingenčních tabulek jsou **tabulky asociační**, které používáme k sledování závislosti dvou dichotomických znaků, tj. kategoriálních znaků nabývajících pouze dvou variant. *(asociace = vztah dvou dichotomických znaků)* Většinou si můžeme představit, že náhodný pokus se provádí za dvojích různých okolností a může skončit buď úspěchem, nebo neúspěchem. Tradičně se pak u tohoto typu kontingenčních tabulek používáme zjednodušené označení: $n_{11} = a, n_{12} = b, n_{21} = c, n_{22} = d$.

Tab. 10.6: Schéma asociační tabulky rozšířené o marginální četnosti

X (*okolnosti*)\\Y(*výskyt události*)	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a+b$
$x_{[2]}$ (II.)	c	d	$c+d$
Celkem	$a+c$	$b+d$	n

Na asociační tabulku lze sice nahlížet jako na speciální případ kontingenčních tabulek a při analýze používat jejich aparát, nicméně vhodnější je využít specifické metody a charakteristiky asociace.

Dále uvedené míry asociace budeme prezentovat v souvislosti s medicínskými aplikacemi, v nichž obvykle sledujeme asociaci mezi sledovaným faktorem (nezávislý znak) a výskytem onemocnění (závislý znak).

Tab. 10.7: Rozšířená asociační tabulka v medicínské aplikaci

X (*sledovaný faktor*)\\Y(*výskyt onemocnění*)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a+b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c+d$
Celkem	$a+c$	$b+d$	n

10.2.1. Poměr šancí

Jako míru asociace často používáme charakteristiku nazývanou **poměr šancí** (angl. „odds ratio"). Pozorovaný poměr počtu úspěchů k počtu neúspěchů (tzv. pozorovaná **šance**) za okolností I. je $\frac{a}{c}$, za okolností II. $\frac{b}{d}$. Odhad poměru šancí je pak

$$\widehat{OR} = \frac{ad}{bc}.$$

V medicíně pak poměr šancí udává kolikrát je vyšší šance výskytu nemoci u exponované populace (tj. populace vystavené vlivu sledovaného faktoru) ve srovnání s neexponovanou populací. Někdy se můžeme s tímto ukazatelem setkat i pod označením **křížový poměr** (angl. „cross-product ratio").

OR (populační poměr šancí) nabývá kladných hodnot v intervalu $\langle 0; \infty)$. Při interpretaci poměru šancí je důležitá hodnota 1.

Tab. 10.8: Interpretace poměru šancí OR v medicínských aplikacích

$OR < 1$	U exponované populace (populace vystavené sledovanému faktoru) je nižší šance výskytu nemoci.
$OR = 1$	Šance výskytu onemocnění u exponované a neexponované populace jsou shodné.
$OR > 1$	U exponované populace je vyšší šance výskytu nemoci.

Je-li $OR \neq 1$, potřebujeme zpravidla ještě rozhodnout, zda je indikována asociace statisticky významná. Chceme tedy testovat nulovou hypotézu, že asociace neexistuje, proti alternativě, že asociace existuje. Hypotézu o nezávislosti znaků X a Y pak lze testovat pomocí $100(1 - \alpha)\,\%$ intervalu spolehlivosti pro OR. Meze intervalu spolehlivosti pro poměr šancí lze přímo určit pouze obtížně, a proto můžeme v literatuře nalézt jejich různé aproximace. Jednou z nich je Woolfova metoda (1955) založená na aproximaci normálním rozdělením. Podle této metody je $100(1 - \alpha)\,\%$ asymptotický intervalový odhad přirozeného logaritmu poměru šancí

$$\left\langle \ln \widehat{OR} - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}; \ln \widehat{OR} + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}} \right\rangle,$$

kde $z_{1-\frac{\alpha}{2}}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil normovaného normálního rozdělení.

Na základě znalosti $100(1 - \alpha)\,\%$ intervalového odhadu pro $\ln OR$ určíme $100(1 - \alpha)\,\%$ intervalový odhad OR

$$\left\langle \widehat{OR} \cdot \mathrm{e}^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}}; \widehat{OR} \cdot \mathrm{e}^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle.$$

Jestliže $100(1 - \alpha)\,\%$ intervalový odhad OR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

10.2.2. Relativní riziko

Jsou-li v medicíně záznamy z nějaké studie zapsány v asociační tabulce, uvádí se obvykle jako další popisné statistiky rovněž **absolutní rizika** výskytu události (onemocnění, úmrtí, ...)

v závislosti na okolnostech (přítomnosti sledovaného faktoru). Ve své podstatě jde o vybrané řádkové relativní četnosti. Je-li záznam ze studie dán tabulkou 10.2, pak

- odhad absolutního rizika onemocnění u exponovaných respondentů je $\frac{a}{a+b}$,
- odhad absolutního rizika onemocnění u neexponovaných respondentů je $\frac{c}{c+d}$.

Absolutní rizika mohou nabývat hodnot z intervalu $\langle 0; 1 \rangle$.

Jako míru asociace mezi sledovanými okolnostmi a výskytem události pak lze použít **relativní riziko** RR (angl. „relative risk“). Odhad relativního rizika RR získáme jako poměr odhadů absolutních rizik vzniku onemocnění u exponovaných a neexponovaných osob, tj.

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)}.$$

Z hlediska interpretace relativního rizika je, podobně jako u poměru šancí OR, důležitá hodnota 1.

Tab. 10.9: Interpretace relativního rizika RR v medicínských aplikacích

$RR < 1$	*Expozice snižuje riziko onemocnění.*
$RR = 1$	Mezi expozici a onemocněním neexistuje žádná asociace.
$RR > 1$	Expozice zvyšuje riziko onemocnění.

Podobně jako při interpretaci poměru šancí potřebujeme, je-li $RR \neq 1$, zpravidla ještě rozhodnout, zda je indikována asociace statisticky významná.

Stanovení přesných mezí intervalu spolehlivosti pro relativní riziko je složité a výpočetně náročné. Ukážeme si Katzovu metodu (1978) založenou na aproximaci normálním rozdělením. Podle této metody je $100(1-\alpha)\,\%$ asymptotický intervalový odhad přirozeného logaritmu relativního rizika

$$\left\langle \ln\widehat{RR} - \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}} ; \ln\widehat{RR} + \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}} \right\rangle,$$

kde $z_{1-\frac{\alpha}{2}}$ je $\left(1-\frac{\alpha}{2}\right)$ kvantil normovaného normálního rozdělení.

Na základě znalosti $100(1-\alpha)\,\%$ intervalového odhadu pro $\ln RR$ určíme $100(1-\alpha)\,\%$ intervalový odhad RR

$$\left\langle \widehat{RR}\cdot \mathrm{e}^{-\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}} ; \widehat{RR}\cdot \mathrm{e}^{\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}} \right\rangle.$$

Jestliže $100(1-\alpha)\,\%$ intervalový odhad RR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

Příklad 10.3. Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

Řešení 10.3.

Příklad 10.4. Někdy je třeba být při posuzování tabulek, které se skládají ze dvou či více skupin, opatrný.

V Horních Sádrovicích bylo hospitalizováno 600 „lehkých“ pacientů, z nichž 10 $(1,7\,\%)$ zemřelo a 400 „těžkých“ pacientů, z nichž zemřelo 190 $(47,5\,\%)$. Ve Staré Dláze bylo hospitalizováno 900 „lehkých“ pacientů, z nichž 30 $(3,2\,\%)$ zemřelo a 100 „těžkých“ pacientů, z nichž zemřelo 100 $(10,0\,\%)$.

Tab. 10.10: Kontingenční tabulky rozšířené o marginální četnosti a řádkové rel. četnosti

Horní Sádrovice			
stav pacienta při přijetí/úmrtnost	ANO	NE	celkem
lehký	10 0,017 (10/600)	590 0,983 (590/600)	600
těžký	190 0,475 (190/400)	210 0,525 (210/400)	400
celkem	200 0,200 (200/1000)	800 0,800 (800/1000)	1 000

Stará Dláha			
stav pacienta při přijetí/úmrtnost	ANO	NE	celkem
lehký	30 0,033 (30/900)	870 0,967 (870/900)	900
těžký	70 0, 700 (70/100)	30 0,300 (30/100)	100
celkem	100 0, 100 (100/1000)	900 0,900 (900/1000)	1 000

Je zřejmé, že u lehkých pacientů je nižší riziko úmrtí v Horních Sádrovicích $(0,017 < 0,033)$. Rovněž u těžkých pacientů je nižší riziko úmrtí v Horních Sádrovicích $(0,475 < 0,700)$. Očekáváte, že nemocnice v Horních Sádrovicích bude v žebříčku úmrtnosti na lepší pozici než nemocnice ve Staré Dláze? (Jinými slovy: Očekáváte, že riziko úmrtí je v Horních Sádrovicích nižší než ve Staré Dláze?) S překvapením konstatujeme, že tabulky ukazují opak. Riziko úmrtí v Horních Sádrovicích (0,200) je vyšší než riziko úmrtí ve Staré Dláze (0,100)! Jde o tzv. Simpsonův paradox.

(Zájemcům doporučujeme stručný článek na toto téma:
http://scienceworld.cz/psychologie/simpsonuv-paradox-a-problem-slucovani-dat-2198)

10.3. Analýza závislostí v normálním rozdělení

10.3.1. Pearsonův koeficient korelace

V teorii pravděpodobnosti byl jako míra lineární závislosti dvou složek spojitého náhodného vektoru zaveden Pearsonův korelační koeficient ρ.

$$\rho = \rho(X,Y) = \begin{cases} \dfrac{cov(X,Y)}{\sqrt{DX \cdot DY}} & DX, DY \neq 0, \\ 0 & \text{jinak.} \end{cases}$$

Připomeňme některé jeho vlastnosti:

1. $-1 \leqq \rho \leqq 1$, přičemž rovnosti je dosaženo pouze tehdy, je-li mezi náhodnými veličinami X a Y lineární závislost,
2. jsou-li X, Y nezávislé náhodné veličiny, pak $\rho = 0$,
3. je-li $\rho = 0$, říkáme, že X, Y jsou **nekorelované** náhodné veličiny,
4. je-li $\rho > 0$, říkáme, že X, Y jsou **pozitivně korelované** (s rostoucím X roste Y),
5. je-li $\rho < 0$, říkáme, že X, Y jsou **negativně korelované** (s rostoucím X klesá Y).

Je zřejmé, že Pearsonův korelační koeficient je vhodnou mírou lineární závislosti náhodných veličin X a Y.

10.3.2. Výběrový korelační koeficient

Pearsonův korelační koeficient ρ dokážeme určit pouze tehdy, známe-li sdružené rozdělení náhodného vektoru $(X; Y)$. V praxi však máme většinou k dispozici pouze výběr $(X_1; Y_1)$ až $(X_n; Y_n)$ z nějakého dvourozměrného rozdělení. Nechť

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,\ S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2,$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \frac{\sum\limits_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sqrt{\left(\sum\limits_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\left(\sum\limits_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right)}}$$

Je rozumné definovat výběrový korelační koeficient r pomocí vztahu analogického vzorci definujícímu Pearsonův korelační koeficient, v němž se neznámá (populační) kovariance a neznámé (populační) rozptyly nahradí jejich nestrannými odhady.

$$r = \begin{cases} \dfrac{S_{X,Y}}{\sqrt{S_X^2 \cdot S_Y^2}} & S_X^2, S_Y^2 \neq 0, \\ 0 & \text{jinak.} \end{cases}$$

10.3.3. Testování nezávislosti

Vlastnosti koeficientu korelace ρ se přenášejí i na výběrový korelační koeficient r. Zjistíme-li, že výběrový korelační koeficient $r \neq 0$, zpravidla nás zajímá, zda je indikovaná korelace statisticky významná. Chceme testovat nulovou hypotézu

$$H_0: \quad \rho = 0$$

vůči alternativě $H_A: \rho \neq 0$, resp. $\rho < 0$, resp. $\rho > 0$.

Nechť $(X_1; Y_1), \ldots, (X_n; Y_n)$ je výběr z dvourozměrného normálního rozdělení, tj. z rozdělení, jehož sdružená hustota pravděpodobnosti je dána vztahem

$$f(x;y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$$

Pak má za předpokladu platnosti nulové hypotézy testová statistika

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Obsah

378. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Studentovo rozdělení s $n-2$ stupni volnosti. Rozhodnutí o výsledku testu provedeme na základě standardně vypočtené $p-hodnoty$.

Poznámky:

- *Má-li náhodný vektor* $(X;Y)$ *dvourozměrné normální rozdělení, pak jeho složky, tj. náhodné veličiny* X *a* Y*, mají normální rozdělení* $N\left(\mu_X;\sigma_X^2\right)$*, resp.* $N\left(\mu_Y;\sigma_Y^2\right)$*. Předpoklad o sdruženém normálním rozdělení náhodných veličin* X *a* Y *se velmi těžko ověřuje. Normalita rozdělení obou sledovaných veličin* X *a* Y *je nutnou podmínkou pro to, aby měl náhodný vektor* $(X;Y)$ *dvourozměrné normální rozdělení. Není to však podmínka postačující. Ukazuje se však, že v praxi většinou zcela vyhovuje, omezíme-li se pouze na ověření této nutné podmínky.*
- *Jsou-li složky náhodného vektoru* $(X;Y)$ *s dvourozměrným normálním rozdělením nekorelované, jsou nezávislé. Ve sdruženém normálním rozdělení je tedy nekorelovanost ekvivalentní nezávislosti.* ***(POZOR! Obecně to neplatí.)***

Příklad 10.5. Máme k dispozici výsledky prvního a druhého zápočtového testu deseti studentů. Na hladině významnosti 0,05 testujte hypotézu, že výsledky zápočtových testů jsou kladně korelované.

X_i (1. test)	7	8	10	4	14	9	6	2	13	5
Y_i (2. test)	9	7	12	6	15	6	8	4	11	8

Řešení 10.5.

Obsah
379. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

10.4. Analýza závislostí ordinálních znaků

V předcházející kapitole jsme viděli, že hodnocení výběrového korelačního koeficientu r je vázáno na splnění předpokladu, že výběr pochází z dvourozměrného normálního rozdělení. Při porušení tohoto předpokladu, resp. v případě, že chceme analyzovat závislost dvou ordinálních znaků, můžeme použít například **Spearmanův koeficient korelace**.

10.4.1. Spearmanův korelační koeficient

Mějme náhodný výběr $(X_1; Y_1), \ldots, (X_n; Y_n)$ z dvourozměrného rozdělení. Nechť R_{X_1} až R_{X_n} jsou pořadí veličin $X_1, \ldots, X_n$ a nechť $R_{Y_1}, \ldots, R_{Y_n}$ jsou pořadí veličin $Y_1, \ldots, Y_n$.

Kdyby s rostoucími hodnotami X_i vzrůstaly i hodnoty Y_i, byla by zřejmě pořadí obou veličin shodná, tj. $R_{X_i} = R_{Y_i}$ pro $i = 1, \ldots, n$. Jestliže s rostoucími hodnotami X_i klesají hodnoty Y_i, jsou pořadí obou veličin právě opačná. Při nezávislosti veličin X a Y jsou pořadí zpřeházená zcela náhodně. Spearmanův korelační koeficient r_S se proto definuje pomocí diferencí pořadí $(R_{X_i} - R_{Y_i})$ jako

$$r_s = 1 - \frac{6}{n\,(n^2-1)} \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2 \,.$$

Při shodném pořadí nabývá koeficient r_S maximální hodnoty 1, při opačném pořadí minimální hodnoty -1. V ostatních případech je $-1 < r_S < 1$. Je-li hodnota Spearmanova korelačního koeficientu $r_S = 0$, pořadí veličin X a Y jsou náhodně zpřeházená, a mezi sledovanými veličinami tedy není závislost.

Pokud se v náhodných výběrech, z nichž je r_S počítán, vyskytuje mnoho shod (tj. stejně velkých pozorování), doporučuje se používat **korigovaný Spearmanův korelační koeficient** $r_{S_{korig}}$. Označme t_X počty stejně velkých X-ových hodnot. (Je-li mezi pozorovanými hodnotami náhodné veličiny X několik skupin stejně velkých pozorování, pak t_X jsou rozsahy těchto skupin.) Podobně definujme t_Y. Pak

$$r_{S_{korig}} = 1 - \frac{6}{n^3 - n - T_X - T_Y} \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2 ,$$

kde $T_X = \frac{1}{2} \sum \left(t_x^3 - t_x\right)$, $T_Y = \frac{1}{2} \sum \left(t_x^3 - t_x\right)$.

Je-li hodnota Spearmanova korelačního koeficientu r_S blízká nule, chceme zpravidla testovat, zda je odchylka koeficientu r_S od nuly náhodná či statisticky významná. Jsou-li odchylky Spearmanova korelačního koeficientu od nuly jen náhodné, jsou veličiny X a Y nezávislé.

H_0 : X, Y jsou **nezávislé** náhodné veličiny.
H_A : X, Y jsou **závislé** náhodné veličiny.

Testovou statistikou je Spearmanův korelační koeficient r_S. Nulovou hypotézu zamítáme pokud $|r_S| \geqq r_S^*(\alpha)$, kde $r_S^*(\alpha)$ je kritická hodnota Spearmanova korelačního koeficientu.

Pro rozsah výběru $\leqq 30$ a hladiny významnosti 0,05, resp. 0,01 jsou kritické hodnoty $r_S^*(\alpha; n)$ tabelovány (tabulka T16). Je-li rozsah výběru $n > 30$, pak

$$r_S^*(\alpha; n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}},$$

kde $z_{1-\frac{\alpha}{2}}$ je $\left(1-\frac{\alpha}{2}\right)$ kvantil normovaného normálního rozdělení.

Příklad 10.6. V tabulce 10.6 je uvedena spotřeba alkoholu a úmrtnost na cirhózu jater a alkoholismus ve vybraných zemích. Určete, zda úmrtnost na cirhózu jater a alkoholismus závisí na spotřebě alkoholu. (Zadání příkladu bylo převzato z [1]).

Tab. 10.11: Spotřeba alkoholu a úmrtnost na cirhózu jater ve vybraných zemích

země	spotřeba alkoholu [l/osoba]	úmrtnost na cirhózu jater a alkoholismus [počet zemřelých na 100 000 obyvatel]
Finsko	3,9	3,6
Norsko	4,2	4,3
Irsko	5,6	3,4
Holandsko	5,7	3,7
Švédsko	6,0	7,2
Anglie	7,2	3,0
Belgie	10,8	12,3
Rakousko	10,9	7,0
SRN	12,3	23,7
Itálie	15,7	23,6
Francie	24,7	46,1

Řešení 10.6.

Obsah
382. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Shrnutí:

Analýza závislosti v kontingenční tabulce

Na porovnávání empirických (pozorovaných) četností s četnostmi teoretickými je založen rovněž $\boldsymbol{\chi^2}$ **test nezávislosti v kontingenční tabulce**. Pomocí něj testujeme:

H_0 : Znaky X a Y v kontingenční tabulce jsou statisticky **nezávislé**.
H_A : Znaky X a Y v kontingenční tabulce jsou statisticky **závislé**.

Pro tabulku s r řádky a s sloupci používáme jako testové kritérium náhodnou veličinu

$$K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

která má v případě platnosti nulové hypotézy a za předpokladu splnění podmínek dobré aproximace přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

Podmínky dobré aproximace:

- žádná z očekávaných četností E_{ij} nesmí být menší než 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

χ^2 test nezávislosti nevypovídá nic o síle vztahu, pouze zamítá, resp. nezamítá nulovou hypotézu o nezávislosti znaků X a Y. Pro zjištění síly vztahu používáme různé koeficienty:

- koeficient kontingence $CC = \sqrt{\frac{K}{K+n}}$ (pro čtvercové kontingenční tabulky),
- korigovaný koeficient kontingence $CC_{cor} = \frac{CC}{CC_{max}}$, kde $CC_{max} = \sqrt{\frac{\min(r;s)-1}{\min(r;s)}}$ (pro obdélníkové kontingenční tabulky),
- Cramerův koeficient $V = \sqrt{\frac{K}{n(\min(r;s)-1)}}$.

Tyto koeficienty se mohou vyskytovat v intervalu $(0;1)$. Čím jsou blíže 1, tím je závislost mezi X a Y těsnější.

Analýza závislosti v asociační tabulce

Speciálním typem kontingenčních tabulek jsou **tabulky asociační**, které používáme k sledování závislosti dvou dichotomických znaků. Jako míru asociace používáme například:

- poměr šancí
- relativní riziko

Pozorovaný poměr počtu úspěchů k počtu neúspěchů (tzv. pozorovaná **šance**) za okolností I. je $\frac{a}{c}$, za okolností II. $\frac{b}{d}$. Odhad poměru šancí je pak

$$\widehat{OR} = \frac{ad}{bc}.$$

Intervalový odhad $\widehat{OR}$ je $\left\langle \widehat{OR}\cdot \mathrm{e}^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\cdot z_{1-\frac{\alpha}{2}}}; \widehat{OR}\cdot \mathrm{e}^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\cdot z_{1-\frac{\alpha}{2}}} \right\rangle$. Jestliže $100(1-\alpha)\%$ intervalový odhad OR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

Odhad relativního rizika RR získáme jako poměr odhadů absolutních rizik vzniku onemocnění u exponovaných a neexponovaných osob, tj. $\widehat{RR} = \frac{a(c+d)}{c(a+b)}$.

Intervalový odhad RR je $\left\langle \widehat{RR}\cdot \mathrm{e}^{-\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}}; \widehat{RR}\cdot \mathrm{e}^{\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}} \right\rangle$. Jestliže $100(1-\alpha)\%$ intervalový odhad RR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

Analýza závislosti v normálním rozdělení

Jsou-li složky náhodného vektoru $(X;Y)$ s dvourozměrným normálním rozdělením nekorelované, jsou nezávislé. Chceme-li tedy testovat nezávislost složek vektoru s dvourozměrným normálním rozdělením, můžeme testovat nulovou hypotézu

$$H_0: \quad \rho = 0$$

vůči alternativě $H_A: \rho \neq 0$, resp. $\rho < 0$, resp. $\rho > 0$.

Nechť je výběrový korelační koeficient r dán vztahem

$$r = \begin{cases} \frac{S_{X,Y}}{\sqrt{S_X^2\cdot S_Y^2}} & S_X^2, S_Y^2 \neq 0, \\ 0 & \text{jinak.} \end{cases}$$

Obsah
385. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Pak má za předpokladu platnosti nulové hypotézy testová statistika

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Studentovo rozdělení s $n-2$ stupni volnosti. Rozhodnutí o výsledku testu provedeme na základě standardně vypočtené $p - hodnoty$.

Analýza závislosti ordinálních veličin

Při porušení předpokladu, že výběr pochází z dvourozměrného normálního rozdělení resp. v případě, že chceme analyzovat závislost dvou ordinálních znaků, můžeme použít například **Spearmanův koeficient korelace**.

Mějme náhodný výběr $(X_1; Y_1), \ldots, (X_n; Y_n)$ z dvourozměrného rozdělení. Nechť R_{X_1} až R_{X_n} jsou pořadí veličin $X_1, \ldots, X_n$ a nechť $R_{Y_1}, \ldots, R_{Y_n}$ jsou pořadí veličin $Y_1, \ldots, Y_n$. Spearmanův korelační koeficient r_s se definuje jako

$$r_s = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} \left(R_{X_i} - R_{Y_i}\right)^2 .$$

Jsou-li odchylky Spearmanova korelačního koeficientu od nuly jen náhodné, jsou veličiny X a Y nezávislé.

H_0 : X, Y jsou **nezávislé** náhodné veličiny.
H_A : X, Y jsou **závislé** náhodné veličiny.

Testovou statistikou je Spearmanův korelační koeficient r_S. Nulovou hypotézu zamítáme pokud $|r_S| \geqq r_S^*(\alpha)$, kde $r_S^*(\alpha)$ je kritická hodnota Spearmanova korelačního koeficientu.

Pro rozsah výběru $\leqq 30$ a hladiny významnosti 0,05, resp. 0,01 jsou kritické hodnoty $r_S^*(\alpha; n)$ tabelovány (tabulka T16). Je-li rozsah výběru $n > 30$, pak $r_S^*(\alpha; n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}}$, kde $z_{1-\frac{\alpha}{2}}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil normovaného normálního rozdělení.

POZOR! *Při pozorování většiny události se obvykle vychází ze stanoviska, že každá událost (jev) ve světě vzniká jako následek nějaké jiné události, která je příčinou pozorovaného jevu, což označujeme jako kauzalitu. Zjistíme-li však mezi dvěma jevy korelaci, pak to nemusí nutně znamenat, že mezi nimi musí existovat vztah příčiny a následku. Korelace znamená v češtině souvztažnost. Je to stav, kdy změna hodnot jedné veličiny souvisí se změnou hodnot druhé veličiny. Zjištěná korelace mezi veličinami může znamenat, že existuje další, našemu pozorování dosud skrytá veličina, která působí jako příčina obou událostí. Mezi pozorovanými veličinami je pak tzv.* ***zdánlivá korelace*** *(viz známý příklad průkazné korelace mezi porodností a čapí populací v daném regionu z Disman (2002)).*

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Čím je mozaikový graf členitější, tím je pozorovaná závislost mezi veličinami v kontingenční tabulce

   (a) slabší, (b) silnější.

2. (2b.) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň

   (a) 4,
   (b) 10,
   (c) 16,
   (d) 20,

   očekávaných četností je větších než 5 a ostatní jsou rovny alespoň

   (a) 0,
   (b) 1,
   (c) 2.

3. (1b.) Koeficient kontingence

   (a) nabývá hodnot z intervalu $\langle 0; 1)$,

(b) nabývá hodnot z intervalu $\langle 0; 1\rangle$,
(c) může nabývat hodnot větších než 1.

4. (1b.) Cramerovo V

(a) nabývá hodnot z intervalu $(0; 1)$,
(b) nabývá hodnot z intervalu $\langle 0; 1\rangle$,
(c) může nabývat hodnot větších než 1.

5. (2b.)

(a) Kontingenční
(b) Asociační

tabulka je speciálním případem

(a) kontingenční
(b) asociační

tabulky.

6. (1b.) Je-li odhad relativního rizika $\widehat{RR} = 1,2$, pak

(a) mezi znaky v asociační tabulce existuje závislost,
(b) mezi znaky v asociační tabulce neexistuje závislost,
(c) o závislosti znaků v asociační tabulce musí rozhodnout test.

7. (1b.) Je-li odhad poměru šancí $\widehat{OR} = 10,2$, pak u exponované populace je

(a) nižší šance výskytu nemoci než u neexponované populace,
(b) vyšší šance výskytu nemoci než u neexponované populace,
(c) stejná šance výskytu nemoci jako u neexponované populace.

8. (1b.) Odhadujeme-li se spolehlivostí 0,95, že relativní riziko $RR \in (0,87; 1,45)$, pak na hladině významnosti 0,05

(a) nezamítáme hypotézu o nezávislosti znaků X a Y,
(b) zamítáme hypotézu o nezávislosti znaků X a Y,
(c) nelze o nezávislosti znaků X a Y rozhodnout.

9. (1b.) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je

(a) Spearmanův korelační koeficient,
(b) Pearsonův korelační koeficient,
(c) korigovaný koeficient kontingence,
(d) Cramerovo V.

10. (1b.) Hodnoty Pearsonova korelačního koeficientu blízké nule vypovídají o tom, že

(a) sledované veličiny X resp. Y nenesou prakticky žádnou informaci o Y resp. X,

(b) mezi sledovanými veličinami X a Y existuje silná lineární závislost,
(c) mezi sledovanými veličinami X a Y neexistuje silná lineární závislost,
(d) sledované veličiny X a Y jsou nezávislé.

11. (1b.) Hodnoty Pearsonova korelačního koeficientu blízké -1 vypovídají o tom, že

(a) sledované veličiny X resp. Y nenesou prakticky žádnou informaci o Y resp. X,
(b) na měřených objektech jsou nízké hodnoty veličiny X doprovázeny spíše vysokými hodnotami veličiny Y,
(c) na měřených objektech jsou nízké hodnoty veličiny X doprovázeny spíše nízkými hodnotami veličiny Y.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1. V tabulce je zaznamenáno dosažené vzdělání 100 párů snoubenců v den uzavření sňatku. Ověřte na hladině významnosti 0,10, zda existuje závislost mezi vzděláním nevěsty a ženicha a určete vhodnou míru závislosti.

ženich	**nevěsta**		
	základní	středoškolské	vysokoškolské
základní	24	12	3
středoškolské	7	24	3
vysokoškolské	3	9	15

Řešení příkladu

2. Níže uvedená tabulka uvádí data ze studie ověřující, zda je konzumace alkoholu faktorem, který ovlivňuje úspěšnost ukončení léčby odvykání kouření (Schiffman, 1982, Journal of Counseling and Clinical Psychology). Ověřte na hladině významnosti 0,05, zda existuje závislost mezi úspěšnosti ukončení léčby odvykání kouření a konzumaci alkoholu, určete poměr šancí na úspěšné ukončení léčby a relativní riziko neukončení léčby.

Konzumace alkoholu	**úspěšnost ukončení léčby – odvykání kouření**	
	kouří	nekouří
konzumuje	20	13
nekonzumuje	48	96

Řešení příkladu

3. V letech 1931-1961 byly měřeny průtoky v profilu nádrže Šance na Ostravici a v profilu nádrže Morávka na Morávce. Roční průměry v m^3/s jsou dány v následující tabulce:

rok	Šance	Morávka
1931	4,130	2,476
1932	2,386	1,352
1933	2,576	1,238
1934	2,466	1,725
1935	3,576	1,820
1936	2,822	1,913
1937	3,863	2,354
1938	3,706	2,268
1939	3,710	2,534
1940	4,049	2,308
1941	4,466	2,517
1942	2,584	1,726
1943	2,318	1,631
1944	3,721	2,028
1945	3,290	2,423

rok	Šance	Morávka
1946	2,608	1,374
1947	2,045	1,194
1948	3,543	1,799
1949	4,055	2,402
1950	2,224	1,019
1951	2,740	1,552
1952	3,792	1,929
1953	3,087	1,488
1954	1,677	0,803
1955	2,862	1,878
1956	3,802	1,241
1957	2,509	1,165
1958	3,656	1,872
1959	2,447	1,381
1960	2,717	1,679

Na hladině významnosti 0,05 ověřte, zda existuje závislost mezi ročními průměrnými průtoky v profilech nádrží Šance a Morávka.

Řešení příkladu

4. V rámci jisté studie byla u žáků základních škol sledována závislost agresivity jejich chování na školním prospěchu. Školní prospěch byl hodnocen nejhorší známkou na vysvědčení, agresivita jejich chování byla hodnocena posuzovací škálou (1–10). Na základě údajů uvedených v níže uvedené tabulce ověřte na hladině významnosti 0,05, zda existuje závislost mezi agresivitou chování a školním prospěchem.

Řešení příkladu

Obsah
393. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Identifikační číslo	1	2	3	4	5	6	7	8	9	10	11
Školní prospěch	1	4	2	3	1	2	3	5	3	1	3
Agresivita chování	1	5	5	6	2	4	8	10	7	3	9

Kapitola 11

Úvod do korelační a regresní analýzy

Cíle

Po prostudování této kapitoly budete

- rozumět základním pojmům regresní analýzy,
- znát zjednodušující předpoklady regresního modelu a umět je ověřit,
- umět používat metodu nejmenších čtverců pro odhad regresní funkce,
- umět posoudit vhodnost modelu pomocí indexu determinace,
- umět používat odhady střední hodnoty a individuální hodnoty závisle proměnné a budete si vědomi rizik spojených s extrapolaci.

11.1. Úvod

Regrese obecně znamená pohyb zpět, ústup nebo návrat. Do statistiky zavedl roku 1886 pojem regrese britský učenec Francis Galton v rámci spojení „regrese k průměru“. Tím označil fakt, že např. synové vysokých otců jsou obvykle nižší než byli jejich otcové, zatímco synové malých otců jsou vyšší než jejich rodiče. Podobně je tomu s jinými vlastnostmi, nejen u lidí. Galtonův název se z jeho výzkumů přenosu vlastností mezi generacemi rozšířil na jakékoliv zkoumání souvislostí mezi náhodnými veličinami a vznikla **regresní analýza**. Zatímco korelační analýza, jejíž základní pojmy jsme zavedli v kapitolách 10.3 a 10.4, se zabývá popisem síly závislosti, regresní analýza umožňuje získat informace o způsobu (tvaru) závislosti mezi kvantitativními znaky.

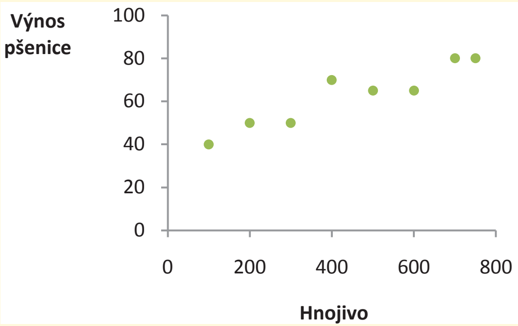
11.1.1. Motivační příklad

Základní pojmy a principy regresní analýzy budeme prezentovat v souvislosti s následujícím příkladem. V tabulce 11.1 jsou uvedeny pozorované hodnoty výnosů pšenice y [t/ha], množství hnojiva x_1 [kg/ha] a srážek x_2 [mm].
Vyneseme-li do grafů závislost výnosů pšenice (y) na množství hnojiva (x_1), resp. na srážkách (x_2), získáme následující bodové grafy označované také jako **korelační pole**.

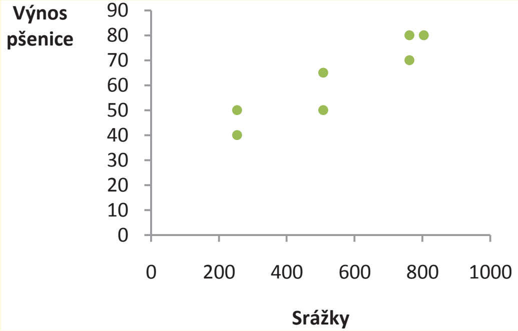
Tab. 11.1: Výnosy pšenice v závislosti na množství hnojiva a množství srážek

y - výnos pšenice [t/ha]	x_1 – hnojivo [kg/ha]	x_2 – srážky [mm]
40	100	254
50	200	508
50	300	254
70	400	762
65	500	508
65	600	508
80	700	762
80	750	804

Z grafů na obrázcích 11.1 a 11.2 a výběrových korelačních koeficientů ($r_{X_1,Y} = 0,939$, $r_{X_2,Y} = 0,911$) se zdá být zřejmé, že výnosy pšenice jsou ovlivněny jak množstvím použitého hnojiva, tak množstvím srážek. V této kapitole se naučíme, jak toto popsat pomocí vhodné funkce, jak nalezenou funkci používat k prognózám a jak vyhodnotit vhodnost volby typu této funkce.



Obr. 11.1: Výnosy pšenice v závislosti na množství použitého hnojiva



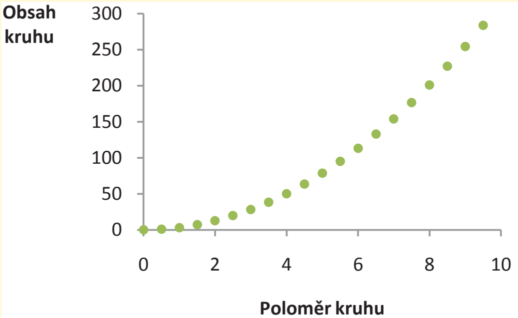
Obr. 11.2: Výnosy pšenice v závislosti na velikosti srážek

11.2. Základní pojmy

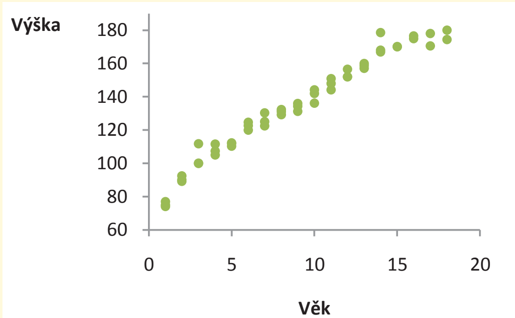
Řekněme, že se sledují dvě fyzikální veličiny Y a x, mezi nimiž existuje závislost $Y = f(x)$. Tento typ jednostranné závislosti označujeme jako tzv. **závislost jednoduchou**. (Např. závislost mezi množstvím použitého hnojiva a výnosy pšenice). Proměnná Y (výnosy pšenice), jejíž chování se snažíme vysvětlit, se označuje jako **závisle proměnná**, resp. jako **proměnná vysvětlovaná**. Proměnnou x (množství hnojiva), jejíž chování vysvětluje chování závisle proměnné Y, nazýváme **nezávisle proměnnou, proměnnou vysvětlující**, resp. **regresorem**.

Jestliže uvažujeme závislost proměnné Y na proměnných $x_1, x_2, \ldots, x_k$ (např. závislost mezi množstvím použitého hnojiva, výnosy pšenice a srážkami), hovoříme o **mnohonásobné (vícenásobné) závislosti**.

Závislost mezi kvantitativními proměnnými Y a $x_1, x_2, \ldots, x_k$ může být v zásadě dvojího typu: funkční a stochastická (volná). **Funkční závislost** (obr. 11.2) je charakteristická tím, že hodnotami nezávisle proměnných $x_1, \ldots, x_k$ je jednoznačně dána hodnota proměnné Y. Příkladem funkční závislosti může být závislost mezi poloměrem kruhu a jeho obsahem. Je zřejmé, že tímto typem závislosti se ve statistice zabývat nebudeme. Předmětem regresní analýzy je zkoumání tzv. **stochastických závislostí** (obr. 11.2), kdy závisle proměnná Y má charakter náhodné veličiny a nezávisle proměnné $x_1, \ldots, x_k$ mohou být jak nenáhodnými (pevnými), tak náhodnými veličinami (např.: závislost výšky na věku dítěte).



Obr. 11.3: Korelační pole pro funkční závislost



Obr. 11.4: Korelační pole pro stochastickou závislost

Stochastickou závislosti mezi náhodnou veličinou Y a proměnnými $x_1, x_2, \ldots, x_k$ rozumíme předpis, který každé uspořádané k-tici $x_1, x_2, \ldots, x_k$ přiřazuje podmíněné rozdělení náhodné veličiny Y. V praxi většinou rozdělení náhodné veličiny Y neznáme, máme k dispozici pouze náhodný výběr ve formě uspořádaných $(k+1)-$tic, $[x_1, x_2, \ldots, x_k, y]$. Na základě tohoto náhodného výběru a odborných informací provedeme výběr typu funkce, která má co nejlépe popisovat rozdělení všech údajů vztahujících se k analyzované závislosti. Tuto funkci nazýváme **regresní funkci** a uvádíme ji ve tvaru

$$E\left(Y|\boldsymbol{X}=\boldsymbol{x}\right)=f\left(x_1,\,\ldots,\,x_k;\beta_0,\,\ldots,\,\beta_p\right),$$

kde $\boldsymbol{x}=(x_1,\,\ldots,\,x_k)$ a $\beta_0,\,\ldots,\,\beta_p$ nazýváme **regresními koeficienty**. (Regresní funkce pro data z motivačního příkladu určuje střední výnosy pšenice při zvolených hodnotách množství hnojiva a srážek.) Regresní koeficienty mají povahu konstant, pokud však máme k dispozici pouze výběr, nedokážeme je přesně určit.

Nahradíme-li regresní koeficienty $\beta_0,\,\ldots,\,\beta_p$ jejich odhady $b_0,\,\ldots,\,b_p$, získáme **odhad regresní funkce**, tzv. vyrovnávací funkci

$$\widehat{Y}=f\left(x_1,\,\ldots,\,x_k;b_0,\,\ldots,\,b_p\right).$$

Odhady $b_0,\,\ldots,\,b_p$ musí být stanoveny tak, aby vyrovnávací funkce co nejlépe aproximovala pozorované hodnoty závislé veličiny Y.

V dalším textu se zaměříme na **lineární regresi**, tj. na případy, kdy je uvažovaná regresní funkce **lineární vzhledem k parametrům** $\beta_0,\,\ldots,\,\beta_k$ nebo se na takovou funkci dá převést. (Např.: $y=\beta_0+\beta_1x_1+\beta_2x_2$ nebo $y=\beta_0x_1^{\beta_1}x_2^{\beta_2}$, která se na funkci lineární vzhledem k parametrům dá převést logaritmováním).

11.3. Lineární regresní model

Hledáme-li při regresní analýze lineární regresní funkci, aplikujeme tzv. lineární regresní model, zkráceně lineární regresi, ve tvaru

$$y_i = \beta_0 + \beta_1 f_1 (x_{1i}) + \cdots + \beta_k f_k (x_{ki}) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

kde $n \ldots$ počet pozorování, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ jsou **náhodné chyby** popisující vliv neznámých nebo nepozorovaných regresorů a vliv náhody a $f_1 (x_{1i}), f_2 (x_{2i}), \ldots, f_k (x_{ki})$ jsou nějaké funkce jednotlivých regresorů. V dalším textu budeme používat zjednodušené označení $f_j (x_{ji}) = f_{ij}$.

Aby bylo možné pro odhad vektoru regresních parametrů použít metodu nejmenších čtverců, musí být splněny základní **předpoklady lineárního regresního modelu**:

1. Náhodné chyby ε_i mají normální rozdělení.
2. $E(\varepsilon_i) = 0$, tj. střední hodnota náhodné složky je nulová aneb náhodná složka nepůsobí systematickým způsobem na hodnoty vysvětlované proměnné Y.
3. $D(\varepsilon_i) = \sigma^2$, tj. rozptyl náhodné složky je konstantní aneb variabilita náhodné složky nezávisí na hodnotách vysvětlujících proměnných a tudíž i podmíněná variabilita vysvětlované proměnné nezávisí na hodnotách vysvětlujících proměnných a je rovna neznámé kladné konstantě σ^2.
4. $cov(\varepsilon_i, \varepsilon_j) = 0$, tj. hodnoty náhodné složky jsou nekorelované, z čehož vyplývá i nekorelovanost různých dvojic pozorování vysvětlované proměnné Y.
5. $h(\boldsymbol{X}) = k+1 < n$. Tato podmínka vyžaduje, aby mezi vysvětlujícími proměnnými nebyla funkční lineární závislost, tedy v matici $\boldsymbol{F}$ (viz kap. 11.4.2) nesmí existovat lineárně

závislé sloupce. Počet vysvětlujících proměnných nesmí být pochopitelně větší než počet pozorování. (V praxi by měl být počet pozorování výrazně větší než počet vysvětlujících proměnných.)

6. V případě vícenásobné regrese nesmí mezi vysvětlujícími proměnnými existovat silná korelace, tzv. multikolinearita, tj. mezi proměnnými f_{ij} pro $j = 1, 2, \ldots, k$ nesmí existovat lineární závislost.

Předpoklady, na nichž je model založen, ověřujeme většinou pomocí jednoduchých exploračních grafů, resp. pomocí známých testů (viz kapitola 11.8).

V některých dále uvedených odvozeních využijeme toho, že mají-li náhodné chyby ε_i rozdělení $N\left(0; \sigma^2\right)$, pak pro každé $i = 1, \ldots, n$:

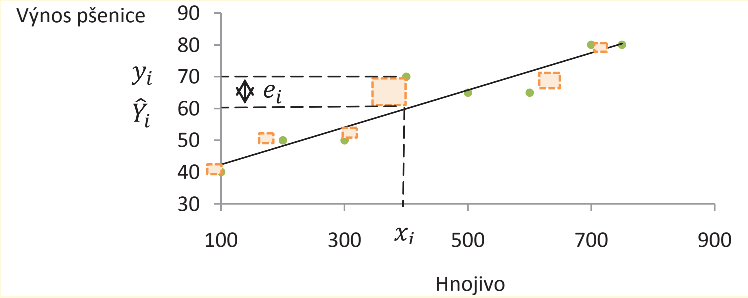
- y_i má normální rozdělení,
- $E(y_i) = \beta_0 + \beta_1 f_{i1} + \cdots + \beta_k f_{ik}$, tj. $E\left(Y_i\right)$ leží na přímce, o níž víme, že je skutečnou regresní přímkou,
- $D(y_i) = \sigma^2$.

11.4. Bodové odhady regresních koeficientů

Hledáme odhad regresní funkce ve tvaru

$$\widehat{Y} = b_0 + b_1 f_{i1} + \cdots b_k f_{ik}, \qquad i = 1, \ldots, n.$$

Jak již bylo zmíněno, pokud jsou splněny předpoklady lineárního regresního modelu, používáme pro jeho řešení nejčastěji **metodu nejmenších čtverců**, která slouží k nalezení takového řešení, aby součet druhých mocnin chyb nalezeného řešení byl minimální.



Obr. 11.5: Vizualizace principu metody nejmenších čtverců

Označme chyby nalezeného řešení $\mathrm{e}_i = y_i - \widehat{Y}_i$ a nazvěme je **rezidua**. Hledáme tedy minimum funkce

$$\varphi = \sum_{i=1}^{n} \mathrm{e}_i^2 .$$

Po dosazení získáme

$$\begin{aligned} \varphi &= \sum_{i=1}^{n} \mathrm{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 f_{i1} + \cdots + b_k f_{ik}))^2 = \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 f_{i1} - \cdots - b_k f_{ik})^2 . \end{aligned}$$

Požadujeme, aby součet čtverců reziduí byl minimální. Proto nejdříve určíme stacionární body, tj. body podezřelé z extrémů:

$$\frac{\partial \varphi}{\partial b_i} = 0, \qquad i = 0, \dots, k.$$

Po dosazení:

$$\begin{aligned}
-2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 f_{i1} - \dots - b_k f_{ik}\right) &= 0,\\
-2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 f_{i1} - \dots - b_k f_{ik}\right) f_{i1} &= 0,\\
&\vdots\\
-2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 f_{i1} - \dots - b_k f_{ik}\right) f_{ik} &= 0,
\end{aligned}$$

Po úpravě:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n b_0 - b_1 \sum_{i=1}^{n} f_{i1} - \dots - b_k \sum_{i=1}^{n} f_{ik},\\
\sum_{i=1}^{n} y_i f_1\left(x_i\right) &= b_0 \sum_{i=1}^{n} f_{i1} + b_1 \sum_{i=1}^{n} \left(f_{i1}\right)^2 + \dots + b_k \sum_{i=1}^{n} f_{i1} f_{ik},\\
&\vdots\\
\sum_{i=1}^{n} y_i f_k\left(x_i\right) &= b_0 \sum_{i=1}^{n} f_{i1} f_{ik} + b_1 \sum_{i=1}^{n} f_{i2} f_{ik} + \dots + b_k \sum_{i=1}^{n} \left(f_{ik}\right)^2 .
\end{aligned}$$

Poznámka: *Takto získanou soustavu označujeme jako* ***soustavu normálních rovnic****. Lze ukázat, že řešení této soustavy je jednoznačné, pokud je alespoň $k+1$ pozorování $[x_1$ až $x_k]$ navzájem různých.*

Poté pomocí klasických metod známých z matematické analýzy ověříme, zda se ve stacionárních bodech nachází minimum. Připomeňme, že řešením jsou čísla $b_0, \ldots, b_k$, která jsou bodovými odhady regresních koeficientů $\beta_0, \ldots, \beta_k$.

Jak dobře dokážete odhadnout regresní přímku bez použití matematických metod (tzv. od oka)? Vyzkoušejte si to v java appletu Regrese (460 KB).

11.4.1. Bodový odhad regresních koeficientů

Hledáme-li odhad regresní funkce ve tvaru

$$\widehat{Y}_i = b_0 + b_1 x_i,$$

hovoříme o **přímkové regresi**. Chceme-li minimalizovat součet čtverců reziduí, minimalizujeme v případě přímkové regrese funkci

$$\varphi = \sum_{i=1}^{n} \mathrm{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 .$$

Nejprve určíme soustavu normálních rovnic:

$$\begin{aligned} \frac{\partial \varphi}{\partial b_0} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) &= 0 \\ \frac{\partial \varphi}{\partial b_1} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\, x_i &= 0 \end{aligned}$$

Po úpravě získáme běžně uváděný tvar soustavy normálních rovnic pro přímkovou regresi.

$$\begin{aligned} \sum_{i=1}^{n} y_i &= n b_0 - b_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i x_i &= b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} (x_i)^2 \,. \end{aligned}$$

Z první rovnice vyjádříme odhad b_0 : $\quad b_0 = \dfrac{\sum_{i=1}^{n} y_i}{n} - b_1 \dfrac{\sum_{i=1}^{n} x_i}{n} = \overline{y} - b_1\overline{x}$. Ten dosadíme do druhé rovnice:

$$\sum_{i=1}^{n} y_i x_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} (x_i)^2$$

a z ní vyjádříme odhad b_1 : $\quad b_1 = \dfrac{\sum_{i=1}^{n} y_i x_i - \dfrac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} (x_i)^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \dfrac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$

Všimněte si, že odhad regresní přímky lze zapsat ve tvaru

$$\widehat{Y} = b_0 + b_1 x = \overline{y} - b_1 \overline{x} + b_1 x = \overline{y} + b_1 \left(x - \overline{x} \right).$$

Je tedy zřejmé, že regresní přímka prochází bodem $[\overline{x}; \overline{y}]$.

Poznámka: *Lze ukázat, že vztahy pro odhady koeficientů regresní přímky lze uvést rovněž v tzv. odchylkovém tvaru:*

$$b_1 = \frac{\sum\limits_{i=1}^{n} \left(x_i - \overline{x} \right) y_i}{\sum\limits_{i=1}^{n} \left(x_i - \overline{x} \right)^2}, \qquad b_0 = \overline{y} - b_1 \overline{x}.$$

Příklad 11.1. Metodou nejmenších čtverců najděte odhad lineární regresní funkce popisující závislost mezi výnosy pšenice a množstvím použitého hnojiva. Pozorované hodnoty k analyzované závislosti jsou uvedeny v tabulce 11.1.

Řešení 11.1

11.4.2. Maticové vyjádření regresního problému

Pro výpočty založené na výběrech o větším rozsahu a některé další úvahy týkající se lineární regrese je výhodné využít maticový způsob zápisu a výpočtu.

Obsah
407. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Lineární regresní model je dán předpisem

$$y_i = \beta_0 + \beta_1 f_1(x_{1i}) + \cdots + \beta_k f_k(x_{ki}) + \varepsilon_i \qquad i = 1, \ldots, n,$$

Pro n pozorování platí

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 f_{11} + \cdots + \beta_k f_{1k} + \varepsilon_1,\\
y_2 &= \beta_0 + \beta_1 f_{21} + \cdots + \beta_k f_{2k} + \varepsilon_2,\\
&\vdots\\
y_n &= \beta_0 + \beta_1 f_{n1} + \cdots + \beta_k f_{nk} + \varepsilon_n,
\end{aligned}$$

Soustavu tak můžeme zapsat v maticovém tvaru

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & f_{11} & \cdots & f_{1k} \\ 1 & f_{21} & \cdots & f_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & f_{n1} & \cdots & f_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{F\beta} + \boldsymbol{\varepsilon}.$$

Hledáme odhad regresní funkce ve tvaru

$$\widehat{Y}_i = b_0 + b_1 f_{i1} + \cdots + b_k f_{ik} \quad \text{pro každé } i = 1, \ldots, n,$$

to lze maticově zapsat jako

$$\widehat{\boldsymbol{Y}} = \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & f_{11} & \cdots & f_{1k} \\ 1 & f_{21} & \cdots & f_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & f_{n1} & \cdots & f_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} = \boldsymbol{Fb}.$$

Metoda nejmenších čtverců slouží k nalezení takového řešení, aby součet druhých mocnin chyb nalezeného řešení byl minimální. Chyby nalezeného řešení (rezidua) jsou definována jako

$$\mathrm{e}_i = y_i - \widehat{Y}_i \quad \text{pro každé } i = 1, \ldots, n,$$

neboli

$$\mathbf{e} = \begin{bmatrix} \mathrm{e}_1 \\ \mathrm{e}_2 \\ \vdots \\ \mathrm{e}_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & f_{11} & \cdots & f_{1k} \\ 1 & f_{21} & \cdots & f_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & f_{n1} & \cdots & f_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} =$$

$$= \boldsymbol{y} - \boldsymbol{F}\boldsymbol{b}.$$

Protože $\mathbf{e}$ je vektor, upravme požadavek na minimalizaci součtu čtverců reziduí tak, aby „součet čtverců jednotlivých odchylek (tedy složek vektoru $\mathbf{e}$) byl minimální“. Při takovém způsobu formulace kritéria se vlastně jedná minimalizaci skalárního součinu, který můžeme napsat

$$\varphi = \sum_{i=1}^{n} \mathrm{e}_i^2 = \mathbf{e}^T\mathbf{e} = (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{b}).$$

Po úpravě dostaneme $\varphi = \mathbf{e}^{\boldsymbol{T}}\mathbf{e} = (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{F}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{F}\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b}$.

Součin bude minimální tehdy, když jeho derivace podle proměnné $\boldsymbol{b}$ bude rovna nule.

$$\frac{\partial\varphi}{\partial\boldsymbol{b}} = 0 - \boldsymbol{F}^T\boldsymbol{y} - \left(\boldsymbol{y}^T\boldsymbol{F}\right)^T + \left(\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b} + \left(\boldsymbol{b}^T\boldsymbol{F}^T\boldsymbol{F}\right)^T\right) = 2\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b} - 2\boldsymbol{F}^T\boldsymbol{y} = 0$$

$\boldsymbol{F}^T\boldsymbol{y} = \boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b}$ je maticový zápis soustavy normálních rovnic, z něhož pak snadno určíme výsledný vzorec pro $\boldsymbol{b}$.

$$\boldsymbol{b} = \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{y}$$

Pro případ přímkové regrese, tj. $\widehat{Y} = b_0 + b_1 x$, dostaneme:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

$$\boldsymbol{F}^T\boldsymbol{F} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{bmatrix},$$

$$\boldsymbol{F}^T\boldsymbol{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{bmatrix},$$

$$\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b} = \begin{bmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} nb_0 + b_1 \sum\limits_{i=1}^{n} x_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 \end{bmatrix},$$

Maticový zápis soustavy normálních rovnic pro přímkovou regresi je

$$\boldsymbol{F}^T\boldsymbol{y} = \boldsymbol{F}^T\boldsymbol{F}\boldsymbol{b}, \text{ tj. } \begin{bmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} nb_0 + b_1 \sum\limits_{i=1}^{n} x_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 \end{bmatrix}.$$

(Srovnejte se soustavou normálních rovnic odvozenou v kapitole 11.4.)

Pro výpočet matice inverzní k matici $\boldsymbol{F}^T\boldsymbol{F}$ použijeme přímý postup pomocí determinantů a subdeterminantů, tj. pomocí determinantů adjungované matice (viz lineární algebra).

$$\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1} = \begin{bmatrix} \dfrac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} & \dfrac{-\sum\limits_{i=1}^{n} x_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} \\ \dfrac{-\sum\limits_{i=1}^{n} x_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} & \dfrac{n}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} \end{bmatrix} =$$

Obsah
411. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

$$= \begin{bmatrix} \dfrac{1}{n}\dfrac{\dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}}{\dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}-\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2} & \dfrac{\dfrac{-\sum\limits_{i=1}^{n} x_i}{n}}{\dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}-\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2} \\ \dfrac{1}{n}\dfrac{\dfrac{-\sum\limits_{i=1}^{n} x_i}{n}}{\dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}-\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2} & \dfrac{1}{n}\dfrac{1}{\dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}-\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2} \end{bmatrix} =$$

$$= \begin{bmatrix} \dfrac{1}{n}\dfrac{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}+\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2}{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}} & \dfrac{1}{n}\dfrac{-\overline{x}}{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}} \\ \dfrac{1}{n}\dfrac{-\overline{x}}{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}} & \dfrac{1}{n}\dfrac{1}{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}} \end{bmatrix} =$$

$$= \begin{bmatrix} \dfrac{\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n}+\left(\dfrac{\sum\limits_{i=1}^{n} x_i}{n}\right)^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \dfrac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \\ \dfrac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \dfrac{1}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n}+\dfrac{x^{-2}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \dfrac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \\ \dfrac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \dfrac{1}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix}.$$

$$\boldsymbol{b} = \left(\boldsymbol{F}^T \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T \boldsymbol{y} = \begin{bmatrix} \frac{1}{n} + \frac{x^{-2}}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} & \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \\ \\ \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} & \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \end{bmatrix} \begin{bmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{bmatrix}.$$

Příklad 11.2. Proveďte odhad koeficientů regresní přímky z řešeného příkladu pomocí maticového zápisu.

Řešení 11.2

11.4.3. Jaký je význam bodových odhadů jednotlivých koeficientů lineární regrese?

Všimněte si, že pomocí koeficientu b_0 lze odhadovat hodnotu závisle proměnné za předpokladu, že hodnoty všech regresorů jsou nulové. V našem případě, pokud by nebylo použito žádné hnojivo, očekáváme výnos pšenice ve výši $36,57\,\mathrm{t/ha}$.

Koeficienty $b_i, i = 1, \ldots, k$ pak udávají odhad závislé proměnné v případě, že se příslušný regresor x_i zvýší o 1 a ostatní regresory se nezmění. V našem případě jsme získali informaci, že pokud zvýšíme množství hnojiva o $1\,\mathrm{kg/ha}$, pak můžeme očekávat navýšení výnosů pšenice o $0,06\,\mathrm{t/ha}$.

11.5. Verifikace modelu

Výpočet konkrétního odhadu regresní funkce na základě výběru pochopitelně neumožňuje ztotožnit nalezený odhad s hypotetickou (populační) regresní funkcí. (*Proč?*) Potřebujeme najít odpověď na řadu otázek spojených s posouzením vhodnosti použití tohoto odhadu pro analýzu vnitřních souvislosti mezi veličinami a pro odhad vysvětlované proměnné při volbě libovolných kombinací vysvětlujících proměnných. Uveďme si zde některé z nich:

- Byl zvolen vhodný typ regresní funkce?
- Byl proveden správný výběr vysvětlujících proměnných?
- Jak lze hodnotit význam jednotlivých vysvětlujících proměnných zařazených do regresní funkce?
- Jak je nalezený odhad kvalitní?
- Bylo použití metody nejmenších čtverců oprávněné?

Podrobné odpovědi na tyto otázky najdete ve specializované literatuře, my se zaměříme pouze na základní verifikaci (ověření modelu):

- Ověření stability modelu pomocí celkového F-testu a dílčích t testů.
- Hodnocení odhadů regresních koeficientů pomocí intervalových odhadů.
- Hodnocení kvality modelu pomocí indexu determinace.
- Ověření předpokladů pro použití metody nejmenších čtverců pomocí analýzy reziduí.
- Ověření, zda mezi vysvětlujícími proměnnými neexistuje multikolinearita.

11.6. Ověřování stability modelu

Při aplikaci metody nejmenších čtverců platí vztah $SS_Y = SS_{\widehat{Y}} + SS_e$,

kde: $SS_Y = \sum_{i=1}^{n} (y_i - \overline{y})^2$ je celkový součet čtverců

$SS_{\widehat{Y}} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{y}\right)^2$ je součet čtverců modelu a

$SS_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2$ je reziudální součet čtverců.

U součtu čtverců modelu by se ve vzorci místo průměru $\overline{y}$ z napozorovaných hodnot měl spíše objevit průměr z hodnot odhadnutých, tj. $\widehat{\overline{Y}}$. Při aplikaci metody nejmenších čtverců se však dá odvodit, že tyto průměry jsou stejné, lze tedy psát

$$\overline{y} = \widehat{\overline{Y}}.$$

11.6.1. Odhad rozptylu náhodné složky

Abychom dokázali posoudit přesnost nalezeného odhadu regresní funkce, potřebujeme znát **rozptyl náhodné složky** σ^2.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (\varepsilon_i - E(\varepsilon_i))^2}{n} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}$$

Protože náhodné chyby $\varepsilon_1, \ldots, \varepsilon_n$ nelze zjistit, musíme se spokojit s jeho odhadem. Lze dokázat, že nevychýleným odhadem rozptylu σ^2 je statistika

$$S_\mathrm{e}^2 = \frac{\sum\limits_{i=1}^{n} \mathrm{e}_i^2}{n-(k+1)} = \frac{SS_\mathrm{e}}{n-(k+1)}$$

kde n je počet pozorování a k je počet regresorů.

11.6.2. Celkový F-test

Celkový F-test nám umožňuje zjistit, zda jsme zvolili správný typ regresní funkce. Slouží k testu hypotézy, zda hodnota vysvětlované proměnné závisí na lineární kombinaci vysvětlujících proměnných. Testujeme nulovou hypotézu

$$H_0: \quad \beta_1 = \cdots = \beta_k = 0$$

proti alternativě

$$H_0: \quad \overline{H_0}$$

Pokud bychom nulovou hypotézu nezamítli, znamenalo by to, že množina vysvětlujících proměnných je zvolena zcela špatně (říkáme, že **model je chybně specifikován**) a museli bychom najít jinou, lepší skladbu těchto proměnných. Poznamenejme, že nezamítnutí nulové hypotézy je jev velmi ojedinělý.

Obsah
416. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Testová statistika pro tento test má Fisherovo-Snedecorovo rozdělení s k stupni volnosti v čitateli a $n-(k+1)$ stupni volnosti ve jmenovateli a má tvar

$$F = \frac{\frac{SS_{\widehat{Y}}}{k}}{\frac{SS_{\mathrm{e}}}{n-(k+1)}},$$

kde výraz v čitateli označujeme jako průměrný čtverec modelu a výraz ve jmenovateli jako průměrný čtverec reziduí (nebo také reziduální rozptyl či odhad rozptylu náhodné složky).

$$p - hodnota = 1 - F_0\left(x_{OBS}\right),$$

kde $F_0(x)$ je distribuční funkce Fisherovo-Snedecorovo rozdělení s k stupni volnosti v čitateli a $n-(k+1)$ stupni volnosti ve jmenovateli. Výsledky celkového F-testu se zapisují do tabulky ANOVA.

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Rozptyl (prům. součet čtverců)	$F-poměr$	$p-hodnota$
Model	$SS_{\hat{Y}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2$	$df_{\hat{Y}} = k$	$\frac{SS_{\hat{Y}}}{df_{\hat{Y}}}$	$\frac{\frac{SS_{\hat{Y}}}{df_{\hat{Y}}}}{\frac{SS_e}{df_e}}$	$1 - F_0(x_{OBS})$
Reziduální	$SS_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{Y}_i)^2$	$df_e = n - (k+1)$	$\frac{SS_e}{df_e}$	---	---
Celkový	$SS_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$df_Y = n - 1$	---	---	---

Příklad 11.3. Pomocí celkového F-testu ověřte, zda lze výnosy pšenice odhadovat pomocí lineární závislosti na množství použitého hnojiva.

Řešení 11.3

11.6.3. Intervalové odhady regresních koeficientů

Vyjdeme-li z předpokladů lineárního regresního modelu $y = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, pak odhady regresních koeficientů b_i vypočítané z výběrových hodnot jsou náhodné veličiny s přibližně normálním rozdělením.

Střední hodnota regresních koeficientů

Lze jednoduše ukázat, že nalezené odhady regresních parametrů jsou nezkreslené, tj. nejsou zatíženy systematickou chybou.

$$E(\boldsymbol{b}) = \boldsymbol{\beta}$$

Pro zájemce

Důkaz.

V kapitole 11.4.2 jsme odvodili maticový zápis vzorce pro odhad vektoru regresních koeficientů: $\boldsymbol{b} = \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{y}$. Dosadíme-li do tohoto vztahu za regresní model $\boldsymbol{y}$ výraz $\boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, dostaneme

$$\boldsymbol{b} = \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}.$$

Obsah
418. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Pak

$$E(\boldsymbol{b}) = E\left(\boldsymbol{\beta} + \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}.$$

□

Rozptyl regresních koeficientů

Označme odhad rozptylu i-tého regresního koeficientu $s_{b_i}^2 (i = 0, 1, \ldots, k)$. Lze ukázat, že

$$s_{b_i}^2 = s_\mathrm{e}^2 x_{i+1,i+1},$$

kde s_e^2 je odhad rozptylu náhodné složky (viz kapitola **??**) a $x_{i+1,i+1}$ je prvek matice $(\boldsymbol{F}^T\boldsymbol{F})^{-1}$ na pozici $(i+1, i+1)$, tj. $i+1$-ní prvek na diagonále.

Jako míra přesnosti odhadu se používá směrodatná odchylka odhadu

$$s_{b_i} = s_\mathrm{e}\sqrt{x_{i+1,i+1}}.$$

Speciálně pro případ přímkové regrese bylo v kapitole 11.4.2 odvozeno, že

$$\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{x^{-2}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \\ \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \frac{1}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix}.$$

Vynásobíme-li reziduální rozptyl s_e^2 prvním prvkem diagonály této matice, získáme rozptyl koeficientu b_0

$$s_{b_0}^2 = s_\mathrm{e}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right).$$

Obsah

420. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Směrodatná odchylka odhadu pak je $s_{b_0} = s_{\mathrm{e}}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}}$.

Obdobně, vynásobíme-li reziduální rozptyl s_{e}^2 prvním prvkem diagonály této matice, získáme rozptyl koeficientu b_1

$$s_{b_1}^2 = \frac{s_{\mathrm{e}}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}$$

Směrodatná odchylka odhadu pak je $s_{b_1} = s_{\mathrm{e}}\sqrt{\frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}}$.

Pro zájemce

Důkaz.

Označme pro $i, j = 0, 1, \ldots, k, \quad i \neq j$

$$cov\,(b_i; b_j) = E\left((b_i - \beta_i)\,(b_j - \beta_j)\right)$$

kovariance mezi odhadovanými regresními koeficienty a

$$D\,(b_i) = cov\,(b_i; b_i) = E\left((b_i - \beta_i)\right)^2$$

rozptyly regresních koeficientů.

Pak

$$cov(\boldsymbol{b}) = \begin{bmatrix} D\left(b_0\right) & cov\left(b_0; b_1\right) & \cdots & cov\left(b_0; b_k\right) \\ cov\left(b_1; b_0\right) & D\left(b_1\right) & \cdots & cov\left(b_1; b_k\right) \\ \cdots & \cdots & \cdots & \cdots \\ cov\left(b_k; b_0\right) & cov\left(b_k; b_1\right) & \cdots & D\left(b_k\right) \end{bmatrix} = E\left(\left(\boldsymbol{b}-\boldsymbol{\beta}\right)\left(\boldsymbol{b}-\boldsymbol{\beta}\right)^T\right)$$

je kovarianční matice odhadu regresních koeficientů

V předcházejícím důkazu jsme odvodili vztah $\boldsymbol{b} = \boldsymbol{\beta} + \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}$. Dosadíme-li jej do

$$\begin{aligned} cov(\boldsymbol{b}) &= E\left(\left(\boldsymbol{b}-\boldsymbol{\beta}\right)\left(\boldsymbol{b}-\boldsymbol{\beta}\right)^T\right), \quad \text{platí} \\ cov(\boldsymbol{b}) &= E\left(\left(\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}\right)\left(\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}\right)^T\right) = \\ &= E\left(\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{F}\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\right) = \\ &= \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\boldsymbol{F}\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}. \end{aligned}$$

Podle předpokladů lineárního regresního modelu je $cov\left(\varepsilon_i, \varepsilon_j\right) = 0, E\left(\varepsilon_i\right) = 0$ a $D\left(\varepsilon_i\right) = \sigma^2$. Pak

$$
\begin{aligned}
cov(\boldsymbol{\varepsilon}) &= E\left((\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))^T\right) = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right), \\
cov(\boldsymbol{\varepsilon}) &= \begin{bmatrix} D(\varepsilon_0) & 0 & \cdots & 0 \\ 0 & D(\varepsilon_1) & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & D(\varepsilon_k) \end{bmatrix} = \sigma^2 I_{k+1},
\end{aligned}
$$

kde I_{k+1} je jednotková matice řádu $k+1$.

Dosadíme-li za $E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)$ výraz $\sigma^2 I_{k+1}$, dostaneme

$$
cov(\boldsymbol{\varepsilon}) = \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\sigma^2 I_{k+1}\boldsymbol{F}\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1} = \sigma^2\boldsymbol{F}^{-1}\boldsymbol{F}^T = \sigma^2\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}.
$$

Jak již bylo uvedeno v kapitole 11.6.1, rozptyl σ^2 náhodné složky musíme odhadnout pomocí statistiky s_{e}^2. Odhad kovarianční matice má proto tvar

$$
\widehat{cov}(\boldsymbol{b}) = s_{\mathrm{e}}^2\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}.
$$

Na hlavní diagonále kovarianční matice $\widehat{cov}(\boldsymbol{b})$ jsou odhady rozptylů odhadů regresních koeficientů. Označme je $s_{b_i}^2$.

$$
s_{b_i}^2 = s_{\mathrm{e}}^2 x_{i+1,i+1},
$$

kde $x_{i+1,i+1}$ je prvek matice $\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}$ na pozici $(i+1, i+1)$, tj. $i+1$- ní prvek na diagonále.

□

Příklad 11.4. Určete směrodatné odchylky parametrů b_0 a b_1 regresní přímky z řešeného příkladu 11.2.

Řešení 11.4

Intervalové odhady pro parametry regresní funkce

Z předcházejícího výkladu víme, že odhady regresních koeficientů b_i vypočítané z výběrových hodnot jsou náhodné veličiny s přibližně normálním rozdělením, střední hodnotou β_i a směrodatnou odchylkou σ_{b_i}.

$$b_i \to N\left(\beta_i; \sigma_{b_i}^2\right)$$

Je tedy zřejmé, že

$$\frac{b_i - \beta_i}{\sigma_{b_i}} \to N(0;1).$$

Směrodatnou odchylku σ_{b_i} neznáme, jejím odhadem je směrodatná odchylka s_{b_i}. Lze dokázat, že výběrová statistika

$$\frac{b_i - \beta_i}{s_{b_i}}$$

má Studentovo t rozdělení s $n-(k+1)$ stupni volnosti, kde n je počet pozorování a k je počet regresorů.

Pomocí této výběrové statistiky pak můžeme známým způsobem (kapitola 9) zkonstruovat intervalové odhady pro β_i. $100(1-\alpha)\,\%$ intervalový odhad koeficientu β_i pak je

$$\langle b_i - t_{1-\frac{\alpha}{2}} s_{b_i}; b_i + t_{1-\frac{\alpha}{2}} s_{b_i}\rangle,$$

kde $t_{1-\frac{\alpha}{2}}$ je $\left(1-\frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n-(k+1)$ stupni volnosti.

11.6.4. Testy hypotéz o koeficientech regresní funkce

Výběrovou statistiku

$$\frac{b_i - \beta_i}{s_{b_i}}$$

lze použít rovněž k testování hypotéz o koeficientech regresní funkce. Nalezli-li jsme odhad regresní funkce $\widehat{Y} = b_0 + b_1 f_1 + \cdots + b_k f_k$, pak nás zajímá, zda směrodatná chyba s_{b_i} odhadů některých koeficientů není natolik velká, že je možné příslušné regresní koeficienty β_i považovat za nulové a lze je z modelu vypustit (mezi Y a x_i není vztah daný funkcí f_i).

Testy nulové hypotézy

$$H_0: \quad \beta_i = 0$$

vůči alternativě $\quad H_A: \quad \beta_i \neq 0$

označujeme jako **dílčí t testy**. Jako testové kritérium používáme výběrovou statistiku

$$\frac{b_i - \beta_i}{s_{b_i}},$$

která má Studentovo rozdělení s $n - (k + 1)$ stupni volnosti. Nezamítneme-li nulovou hypotézu, znamená to, že příslušný regresní koeficient je na dané hladině významnosti statisticky nevýznamný a proto jej můžeme z modelu vypustit.

Příklad 11.5. Nalezněte 95 % intervalové odhady koeficientů regresní přímky z motivačního příkladu a pomocí dílčích t testů ověřte, zda lze nalezené odhady považovat za statisticky významné.

Řešení 11.5

Obsah
425. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

11.7. Testování reziduí

Další informace o vhodnosti modelu a o tom, zda jsou splněny předpoklady o náhodné složce ε_i učiněné pro klasický lineární model, můžeme získat pomocí testování reziduí e_i. V tuto chvíli tedy na rezidua pohlížíme jako na konkrétní hodnoty náhodné složky z regresního modelu.

11.7.1. Test normality reziduí

Ověření předpokladu, že náhodné chyby ε_i mají normální rozdělení, provádíme pomocí testu nulové hypotézy

$$H_0: \text{ rezidua mají normální rozdělení}$$

vůči alternativě, že tomu tak není. Při testu postupujeme standardním způsobem - používáme testy dobré shody. Testové statistiky konstruujeme obvyklým způsobem - buď použijeme χ^2-test dobré shody, modifikovaný Kolmogorovův-Smirnovův test nebo některý z dalších testů normality implementovaných ve statistickém softwaru.

11.7.2. Test nulovosti střední hodnoty reziduí

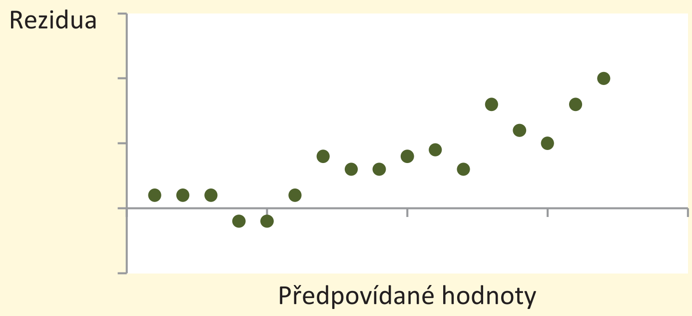
Porovnáme-li graficky rezidua s čímkoli dalším (pozorovanými hodnotami, odhadnutými hodnotami, hodnotami regresoru), pak jsou rezidua náhodně rozmístěna kolem nuly. Byla-li ověřena normalita reziduí, lze k ověření nulovosti střední hodnoty reziduí použít jeden z nejobvyklejších testů ve statistice, jednovýběrový t test.

11.7.3. Test homoskedasticity reziduí

Podstatou tohoto testu je ověření, zda rezidua mají stejný konstantní rozptyl. Konstrukce celého testu je poměrně složitou záležitostí a proto tento test ani nebývá běžně součástí komerčních statistických paketů. Pro orientační ověření homoskedasticity se často používá graf reziduí a odhadovaných hodnot $\widehat{Y}_i$ (angl. „predicted value“) závislé proměnné. Homoskedasticitní rezidua se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími odhadovanými hodnotami $\widehat{Y}_i$.

11.7.4. Autokorelace reziduí

Podle dalšího z předpokladů lineárního regresního modelu by náhodná složka ε_i měla mít charakter nekorelovaných náhodných veličin. Na grafu reziduí a předpovídaných hodnot $\widehat{Y}_i$ se autokorelace projeví tak, že se rezidua systematicky snižují nebo zvyšují, resp. můžeme mezi reziduí a předpovídanými hodnotami pozorovat nelineární závislost.



Při posuzování předpokladu o nekorelovanosti reziduí se obvykle vychází z autokorelační struktury prvního řádu:

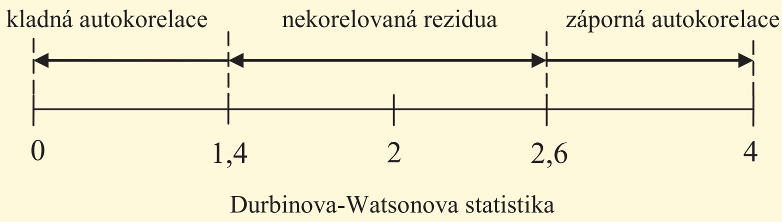
$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + u_i,$$

ve které $u_i \sim N(0; 1)$ a ρ_1 je neznámý parametr, tzv. autokorelační koeficient prvního řádu. Analogicky bychom sestrojili autokorelační strukturu druhého, třetího řádu atd. Autokorelace prvního řádu se však vyskytuje nejčastěji.

K testu se používá Durbinova-Watsonova statistika ve tvaru

$$D_W = \frac{\sum\limits_{i=2}^{n} (\mathrm{e}_i - \mathrm{e}_{i-1})}{\sum\limits_{i=1}^{n} \mathrm{e}_i^2} \doteq 2\left(1 - \widehat{\rho}_1\right).$$

(Všimněte si, že Durbinovu-Watsonovu statistiku lze použít k odhadu autokorelačního koeficientu ρ_1.) Hodnoty této statistiky se pohybují v intervalu $\langle 0; 4\rangle$. Pokud je tato statistika rovna číslu 2, rezidua nevykazují žádnou autokorelaci, hodnoty D_W menší než 2 značí pozitivní autokorelaci a hodnoty větší než 2 značí autokorelaci negativní. Kvantily této statistiky je obtížné vyjádřit explicitně, proto pro Durbinův–Watsonův test statistické programy běžně neposkytují u jiných testu obvyklý komfort, $p-hodnotu$. Při rozhodování lze pro hodnoty statistiky velmi blízké dvěma spoléhat na intuici a považovat rezidua za nekorelované. V praxi můžeme zjednodušeně postupovat podle schématu na obrázku.



Příklad 11.6. Proveďte analýzu reziduí pro model z řešeného příkladu 11.1.

Řešení 11.6

11.8. Multikolinearita

Pro jednoznačný odhad vektoru regresních koeficientů vícenásobných lineárních modelů je nezbytné, aby vysvětlující proměnné byly lineárně nezávislé, tedy aby žádná vysvětlující proměnná nebyla lineární kombinací ostatních regresorů. Tomuto požadavku lze vždy vyhovět, pokud jsou data získávána na základě plánovaných experimentů. V praxi se však

Obsah
429. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

obvykle pracuje s daty, jež mají neexperimentální charakter. V takových případech se v regresním modelu téměř vždy vyskytuje jistý stupeň multikolinearity, tzn., že jeho vysvětlující proměnné jsou určitým způsobem korelovány. Korelované vysvětlující proměnné poskytují podobnou, resp. nadbytečnou, informaci a při statistickém zpracování způsobují řadu obtíží, jež narůstají se stupněm (intenzitou) multikolinearity.

11.8.1. Příčiny multikolinearity

Mezi hlavní příčiny multikolinearity patří

- přeurčený regresní model, tj. model obsahující nadměrný počet vysvětlujících proměnných,
- nevhodný plán experimentu, tj. nevhodná volba kombinací hodnot vysvětlujících proměnných,
- fyzikální omezení v modelu nebo v datech, tj. věcně zdůvodněná závislost vzájemně propojených veličin.

11.8.2. Důsledky multikolinearity

- Multikolinearita zvyšuje rozptyly odhadů, což má za následek:
  a) Snížení přesnosti odhadů individuálních hodnot, tj. rozšíření predikčních intervalů – viz kapitola 11.10.
  b) Nízké hodnoty t_i pro dílčí t testy. To způsobuje, že některé (někdy dokonce všechny) regresní koeficienty se jeví statisticky nevýznamné i v případě jinak velmi kvalitního modelu. Může tak dojít k paradoxu, kdy výsledek celkového F testu je

statisticky významný, ačkoliv výsledky všech dílčích t testů jsou statisticky nevýznamné. (paradox - významný F-test, nevýznamné všechny dílčí t-testy).

c) Nestabilitu odhadů regresních koeficientů, které jsou velmi citlivé i na malé změny v datech a vykazují obvykle vysokou variabilitu. Bodové odhady regresních koeficientů se pro opakované výběry mohou podstatně lišit.

- Multikolinearita komplikuje rozumnou interpretaci individuálního vlivu jednotlivých vysvětlujících proměnných.
- Multikolinearita rovněž komplikuje a někdy zcela znemožňuje identifikaci a vyjádření odděleného působení jednotlivých vysvětlujících proměnných na závisle proměnnou.

11.8.3. Detekce multikolinearity

Pro zjišťování multikolinearity se v odborné literatuře uvádí řada pravidel a doporučení.

- Při silné vzájemné lineární závislosti vysvětlujících proměnných se determinant jejich korelační matice málo liší od nuly.
- Nízká hodnota nejmenšího charakteristického čísla korelační matice indikuje silnou korelaci vysvětlujících proměnných.
- Index podmíněnosti korelační matice (tj. odmocnina poměru největšího a nejmenšího charakteristického čísla) větší než 30 ukazuje na existenci multikolinearity.
- Hodnoty jednoduchých korelačních koeficientů dvojic vysvětlujících proměnných blízké 1 (v praxi větší než 0,8) naznačují multikolinearitu.

11.8.4. Možnosti odstranění multikolinearity

- V případě přeurčeného regresního modelu se snažíme identifikovat a vypustit nadbytečné vysvětlující proměnné.
- Je-li příčinou multikolinearity nevhodný plán experimentu, je možné nedostatky napravit a pořídit kvalitnější data.
- Nejkomplikovanější (a bohužel i nejčastější) případ multikolinearity je způsoben fyzikálními závislostmi v modelu. Vypuštění proměnných z modelu může vést k systematickým chybám a ani pořízení nových dat většinou nepomůže. Jediným rozumným řešením se ukazuje použití nelineárního regresního modelu. Popis tohoto modelu můžete najít například v [29].

11.9. Korelační analýza

Těsnost lineární závislosti mezi závisle proměnnou a regresory posuzujeme pomocí korelačních koeficientů. Posuzovaný vztah je tím silnější a odhad regresní funkce tím lepší, čím více jsou pozorované hodnoty vysvětlované proměnné soustředěné kolem odhadnuté regresní funkce, a naopak tím slabší, čím více jsou hodnoty y_i vzdáleny hodnotám vyrovnaným.

11.9.1. Index determinace

Při konstrukci míry ukazující na sílu závislosti vycházíme ze vztahu pozorovaných a vyrovnaných hodnot. Jak již víme, při aplikaci metody nejmenších čtverců platí vztah

$$SS_Y = SS_{\widehat{Y}} + SS_{\mathrm{e}},$$

$$\begin{aligned} \text{kde} \quad SS_Y &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad \text{je celkový součet čtverců,} \\ SS_{\widehat{Y}} &= \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{y}\right)^2 \quad \text{je součet čtverců modelu a} \\ SS_\mathrm{e} &= \sum_{i=1}^{n} \mathrm{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2 \quad \text{je reziduální součet čtverců.} \end{aligned}$$

Je zřejmé, že čím je model lepší, tím větších hodnot bude nabývat součet čtverců modelu a tím menší bude reziduální součet čtverců. Vydělíme-li rovnici $SS_Y = SS_{\widehat{Y}} + SS_\mathrm{e}$ celkovým součtem čtverců, převedeme ji na tvar

$$1 = \frac{SS_{\widehat{Y}}}{SS_Y} + \frac{SS_\mathrm{e}}{SS_Y}$$

Oba zlomky jsou kladné, jejich součet je roven jedničce, je tedy zřejmé, že každý ze zlomků nabývá hodnoty mezi nulou a jedničkou. Bude-li model dobře vystihovat závislost vysvětlované proměnné na regresorech, bude se hodnota prvního zlomku blížit k jedničce a hodnota druhého zlomku k nule. Bude-li model popisovat uvažovanou závislost špatně, bude tomu naopak. Ukazuje se jako logické použít první zlomek jako kritérium kvality modelu.

Označme tedy

$$R^2 = \frac{SS_{\widehat{Y}}}{SS_Y} = 1 - \frac{SS_\mathrm{e}}{SS_Y} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

a nazveme jej indexem determinace.

Index determinace R^2 udává kvalitu regresního modelu, přesněji řečeno udává, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno. Tento index nabývá hodnot od nuly do jedné (teoreticky i včetně těchto krajních mezí), přičemž hodnoty blízké nule značí špatnou kvalitu regresního modelu, hodnoty blízké jedné značí dobrou kvalitu regresního modelu, udává se většinou v procentech.

Je-li $R^2 = 1$, pak $SS_\mathrm{e} = 0$, což znamená, že regresní model vysvětluje závislost vysvětlované proměnné na regresorech úplně (tzv. dokonalá lineární závislost). Naopak, je-li $R^2 = 0$, pak model nevysvětluje nic, tedy $SS_\mathrm{e} = SS_T$, což nastane jen tehdy, když $b_1 = \cdots = b_k$ a $b_0 = \overline{y}$ (např. pro $k = 1$ je regresní přímka rovnoběžná s osou x v úrovni $b_0 = \overline{y}$).

POZOR! *Vyjde-li nízká hodnota indexu determinace, nemusí to ještě znamenat nízký stupeň závislosti mezi proměnnými, ale může to signalizovat chybnou volbu typu regresní funkce.*

Nevýhodou indexu determinace je skutečnost, že má tendenci nadhodnocovat podíl modelu na vysvětlení celkové variability závisle proměnné. Závisí totiž na počtu regresorů a s růstem jejich počtu narůstá i jeho hodnota. Proto se zavádí tzv. **modifikovaný (adjustovaný) index determinace** R^2_{adj}, který je „penalizovaný“ za nadbytečný počet vysvětlujících proměnných.

$$R^2_{adj} = 1 - \frac{\frac{SS_\mathrm{e}}{n-(k+1)}}{\frac{SS_Y}{n-1}} = 1 - \frac{n-1}{n-(k+1)}\left(1 - R^2\right)$$

Všimněte si, že $R^2_{adj} < R^2$. Rozdíl je výrazný, pokud je počet pozorování n jen o málo větší než počet regresorů k. Naopak, pokud je $n \ll k$, pak se hodnota R^2_{adj} hodnotě R^2 přibližuje.

V případě přímkové regrese je odmocnina z indexu determinace rovna výběrovému korelačnímu koeficientu $\left(\sqrt{R^2} = r\right)$. V případě mnohonásobné lineární regrese je odmocnina z indexu determinace rovna tzv. **koeficientu mnohonásobné korelace** $r_{Y\cdot x_1,x_2,\ldots,x_k}$, který udává míru lineární závislosti mezi závisle proměnnou Y a lineární kombinací regresorů $x_1, x_2, \ldots, x_k$.

$$r_{Y\cdot x_1,x_2,\ldots,x_k} = \sqrt{R^2}$$

Koeficient mnohonásobné korelace nabývá hodnot z intervalu $\langle 0; 1\rangle$, přičemž hodnoty 1 dosáhne v případě, že existuje funkční závislost

$$Y = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \cdots + \beta_k f_k(x_k).$$

11.9.2. Parciální korelační koeficienty

V případě mnohonásobné regrese, potřebujeme často určit také míru „čisté“ závislosti mezi závisle proměnnou a jedním z regresorů, bez vlivu regresorů ostatních. Toto nám umožňují parciální (dílčí) korelační koeficienty. Parciální korelační koeficient ve tvaru

$$\rho_{Y,x_1\cdot x_2,x_3,\ldots,x_k}$$

interpretujeme jako jednoduchý korelační koeficient mezi Y a x_1 při vyloučení vlivu x_2, x_3 až x_k. Tento koeficient je definován jako jednoduchý korelační koeficient náhodných složek ε^1 a ε^2 v regresních rovnicích

$$\begin{aligned} Y &= \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_k x_k + \varepsilon^1, \\ x_1 &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon^2. \end{aligned}$$

Odhad těchto koeficientů je možné počítat různými způsoby. Jednou z možností je výpočet z odhadu korelační matice vektoru náhodných veličin $Y, x_3, x_2, x_3, \ldots, x_k$, která má tvar

$$\boldsymbol{r} = \begin{bmatrix} 1 & r(Y,x_1) & r(Y,x_2) & \cdots & r(Y,x_k) \\ r(x_1,Y) & 1 & \cdots & \cdots & r(x_1,x_k) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r(x_k,Y) & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

Z této matice pak určíme odhad parciálního korelační koeficient jako

$$r_{Y,x_1 \cdot x_2,x_3,\ \ldots,\ x_k} = \frac{|r_{Y,x_1}|}{\sqrt{|r_{Y,Y}|\,|r_{x_1,x_2}|}},$$

kde $|r_{Y,x_1}|$ je determinant matice $\boldsymbol{r}$ zmenšené o první řádek (Y) a druhý sloupec (x_1),atd.

Vedle parciálního korelačního koeficientu $\rho_{Y,x_1 \cdot x_2,x_3,\ \ldots,\ x_k}$ bychom mohli uvažovat i parciální korelační koeficienty $\rho_{Y,x_2 \cdot x_1,x_3,\ \ldots,\ x_k}, \rho_{Y,x_3 \cdot x_1,x_2,x_4,\ \ldots,\ x_k}, \ldots$ Jejich odhad bychom obdrželi obdobně jako $r_{Y,x_1 \cdot x_2,x_3,\ldots,x_k}$.

Koeficient parciální korelace má podobné vlastnosti jako obyčejný korelační koeficient. Jsou-li splněny předpoklady lineárního regresního modelu, pak je možné testovat hypotézy o nulovosti koeficientu parciální korelace. Lze užívat metodu z kapitoly 15.3.3 s tím rozdílem, že testová statistika má Studentovo rozdělení s $n-(k+1)$ stupni volnosti.

Vzhledem k výpočetní náročnosti je potěšující, že výpočet parciálních korelačních koeficientů bývá standardně výbavou běžných statistických programů.

Příklad 11.7. Pomocí indexu determinace, resp. modifikovaného indexu determinace, určete kvalitu modelu nalezeného v řešeném příkladu 11.4.1.

Řešení 11.7

11.10. Využití úspěšně verifikovaných regresních modelů k predikci

Až dosud jsme studovali aspekty týkající se pozice celé regresní funkce. Nyní se zaměříme na odhad očekávané hodnoty závislé proměnné za dané úrovně regresorů.

Označme $\widehat{Y}_0 = \widehat{Y}\left(x_{1_0}, x_{2_0}, \ldots, x_{k_0}\right)$ odhadovanou hodnotu závislé proměnné y za daných hodnot regresorů $x_1, x_2, \ldots, x_k$. Následující úvahy budeme prezentovat na příkladu přímkové regrese $\widehat{Y} = b_0 + b_1 x_0$, v případě vícenásobné regrese bychom postupovali obdobně.

Odhad $\widehat{Y}_0 = \widehat{Y}\left(x_0\right)$ je přibližně normálně rozdělen se střední hodnotou

$$E\left(\widehat{Y}_0\right) = \beta_0 + \beta_1 x_0$$

a rozptylem

$$D\left(\widehat{Y}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right),$$

kde x_0 je daná hodnota regresoru x.

Pro zájemce

Důkaz.

$$E\left(\widehat{Y}_0\right) = E\left(b_0 + b_1 x_0\right) = E\left(b_0\right) + E\left(b_1\right) x_0 = \beta_0 + \beta_1 x_0$$

Pro nalezení rozptylu $D\left(\widehat{Y}_0\right)$ použijeme upravený předpis pro odhad závislé proměnné. Za b_0 dosadíme vztah $b_0 = \overline{y} - b_1 \overline{x}$ nalezený metodou nejmenších čtverců (kapitola).

$$\begin{aligned}
\widehat{Y}_0 &= b_0 + b_1 x_0 = \overline{y} - b_1 \overline{x} + b_1 x_0 = \overline{y} + b_1\left(x_0 - \overline{x}\right) \\
D\left(\widehat{Y}_0\right) &= D\left(\overline{y} + b_1\left(x_0 - \overline{x}\right)\right) = D\left(\overline{y}\right) + D\left(b_1\right)\left(x_0 - \overline{x}\right)^2 = \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\left(x_0 - \overline{x}\right)^2 = \sigma^2\left(\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right)
\end{aligned}$$

Jak již víme, střední hodnoty a rozptyly regresních koeficientů nedokážeme určit přesně (rozptyl σ^2 náhodné složky musíme odhadnout pomocí statistiky s_{e}^2), dokážeme je pouze odhadnout. Střední hodnotu $E\left(\widehat{Y}_0\right)$ odhadujeme

$$\widehat{E}\left(\widehat{Y}_0\right) = b_0 + b_1 x_0 = \widehat{Y}_0$$

a rozptyl $D\left(\widehat{Y}_0\right)$ odhadujeme

$$\widehat{D}\left(\widehat{Y}_0\right) = s_{\mathrm{e}}^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right) = s_{\widehat{Y}}^2.$$

□

11.10.1. Intervalový odhad střední hodnoty závislé proměnné $E(Y_0|x_0)$

Protože v případě přímkové regrese má

$$\frac{\widehat{E}\left(\widehat{Y}_0\right) - E\left(\widehat{Y}_0\right)}{S_{\widehat{Y}}} = \frac{\widehat{Y}_0 - E\left(\widehat{Y}_0\right)}{S_{\widehat{Y}}} = \frac{(b_0 + b_1 x_0) - E\left(\widehat{Y}_0\right)}{S_{\widehat{Y}}}$$

Studentovo rozdělení s $n-2$ stupni volnosti, lze jako intervalový odhad $E\left(\widehat{Y}_0\right)$ se spolehlivostí $1-\alpha$ použít

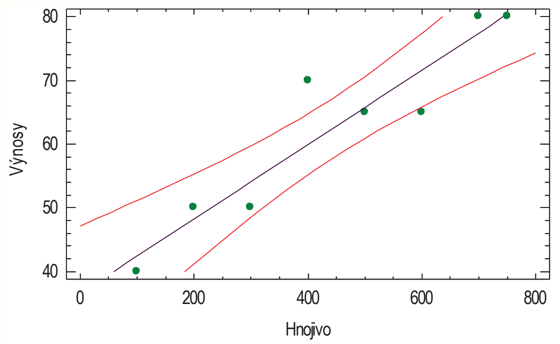
$$\left\langle (b_0 + b_1 x_0) - t_{1-\frac{\alpha}{2}} S_{\widehat{Y}}; (b_0 + b_1 x_0) + t_{1-\frac{\alpha}{2}} S_{\widehat{Y}} \right\rangle,$$

tj.

$$\left\langle (b_0 + b_1 x_0) - t_{1-\frac{\alpha}{2}} S_{\mathrm{e}} \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}; (b_0 + b_1 x_0) + t_{1-\frac{\alpha}{2}} S_{\mathrm{e}} \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}} \right\rangle,$$

kde $t_{1-\frac{\alpha}{2}}$ je $\left(1-\frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n-2$ stupni volnosti.

V praxi většinou není předem dáno, ve kterém bodě x_0 se bude tento interval potřebovat, proto se počítají jeho koncové body pro všechna $x_0 \in (\min x_i; \max x_i)$. Lze ukázat, že koncové body tvoří dvě větve hyperboly, které mezi sebou vytvářejí tzv. **pás spolehlivosti** kolem regresní přímky.



V některých aplikacích se můžeme setkat s otázkou, pro kterou volbu x_0 je pás spolehlivosti nejužší, a tudíž také odhad střední hodnoty $E\left(\widehat{Y}_0\right)$ nejpřesnější? Neboť šířka pásu spolehlivosti je závislá na hodnotě $S_{\widehat{Y}}$, je zřejmé, že na tuto otázku lze zodpovědět nalezením takového x_{opt}, které minimalizuje $S_{\widehat{Y}}$.

$$S_{\widehat{Y}} = S_e \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}} \Rightarrow x_{opt} = \overline{x}$$

Obsah

440. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Vidíme, že pás má nejmenší šířku pro $x_0 = \overline{x}$, a při změně x, ať už k větším či menším hodnotám, šířka pásu roste. Všimněte si, že šířku pásu lze do určité míry předem ovlivnit vhodnou volbou hodnot nezávisle proměnné $x_1, x_2, \ldots, x_n$. Čím větší rozptyl nezávisle proměnné, tím menší odhad rozptylu $\widehat{Y}_0$ $\left(s_{\widehat{Y}}\right)$ a tím přesnější odhad střední hodnoty $E\left(\widehat{Y}_0\right)$.

11.10.2. Intervalový odhad individuální hodnoty závislé proměnné

V praxi nám mnohdy nestačí znát chování střední hodnoty závisle proměnné při dané hodnotě regresorů, důležité je rovněž znát přímo chování závislé proměnné pro danou hodnotu regresorů. Odvození opět provedeme pouze pro přímkovou regresi.

Z předpokladů lineárního regresního modelu je známo, že závisle proměnná má přibližně normální rozdělení se střední hodnotou

$$E(y) = \beta_0 + \beta_1 x_0$$

a rozptylem

$$D(y) = \sigma^2$$

Z předchozí kapitoly víme, že odhad závisle proměnné $\widehat{Y}_0$ má rozdělení $N\left(Y_0; s^2_{\widehat{Y}_0}\right)$

Hodnota závislé proměnné Y_0 pro danou hodnotu nezávisle proměnné x_0 má přibližně normální rozdělení se střední hodnotou

$$E\left(Y_0\right) = \beta_0 + \beta_1 x_0$$

a rozptylem

$$D(Y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right).$$

Pro zájemce

Důkaz.

$$\begin{aligned} E(Y_0) &= E\left(\widehat{Y}_0 + \varepsilon\right) = E\left(\widehat{Y}_0\right) + E(\varepsilon) = E\left(\widehat{Y}_0\right) = \beta_0 + \beta_1 x_0 \\ D(Y_0) &= D\left(\widehat{Y}_0 + \varepsilon\right) = D\left(\widehat{Y}_0\right) + D(\varepsilon) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right) + \sigma^2 = \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right) \end{aligned}$$

Střední hodnoty a rozptyly regresních koeficientů nedokážeme určit přesně (rozptyl σ^2 náhodné složky musíme odhadnout pomocí statistiky s_e^2), dokážeme je pouze odhadnout. Proto střední hodnotu $E(Y_0)$ odhadujeme

$$\widehat{E}(Y_0) = \beta_0 + \beta_1 x_0 = \widehat{Y}_0$$

a rozptyl $D\left(Y_0\right)$ odhadujeme

$$\widehat{D}\left(Y_0\right) = s_{\mathrm{e}}^2 \left(1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\limits_{i=1}^{n} \left(x_i - \overline{x}\right)^2}\right)$$

Protože v případě přímkové regrese má

$$\frac{E\left(\widehat{Y}_0\right) - E\left(Y_0\right)}{\sqrt{\widehat{D}\left(Y_0\right)}} = \frac{\widehat{Y}_0 - E\left(Y_0\right)}{\sqrt{\widehat{D}\left(Y_0\right)}} = \frac{\left(b_0 + b_1 x_0\right) - E\left(Y_0\right)}{\sqrt{\widehat{D}\left(Y_0\right)}}$$

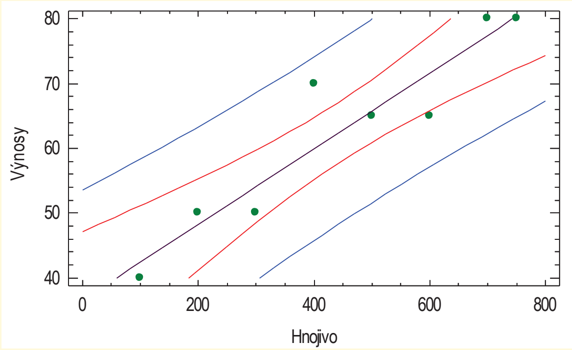
Studentovo rozdělení s $n-2$ stupni volnosti, lze jako intervalový odhad $E\left(Y_0\right)$ se spolehlivostí $1-\alpha$ použít

$$\left\langle \left(b_0 + b_1 x_0\right) - t_{1-\frac{\alpha}{2}} s_{\mathrm{e}} \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\limits_{i=1}^{n} \left(x_i - \overline{x}\right)^2}} ; \left(b_0 + b_1 x_0\right) + t_{1-\frac{\alpha}{2}} s_{\mathrm{e}} \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\limits_{i=1}^{n} \left(x_i - \overline{x}\right)^2}} \right\rangle,$$

kde $t_{1-\frac{\alpha}{2}}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n-2$ stupni volnosti.

$\square$

Obdobně jako v případě intervalového odhadu střední hodnoty závisle proměnné, ani v tomto případě není předem dáno, ve kterém bodě x_0 se bude tento interval potřebovat. Koncové body intervalu spolehlivosti pro individuální hodnotu závisle proměnné vypočtené pro všechna $x_0 \in (\min x_i ; \max x_i)$ tvoří dvě větve hyperboly, které mezi sebou vytvářejí tzv. **pás predikce** kolem regresní přímky. Všimněte si, že pás predikce je širší než pás spolehlivosti (výraz pod odmocninou se zvětšil o 1).



11.10.3. Rozšíření modelu

Odhad regresní funkce, intervalové odhady střední hodnoty a individuální hodnoty závisle proměnné nám umožňují předvídat závisle proměnnou při **libovolné** hodnotě x_0.

Je-li $x_0 \in \langle x_1; x_n \rangle$ (x_0 leží mezi pozorovanými hodnotami x_i), pak se proces předvídání nazývá **interpolace**. V opačném případě, tj. pokud $x_0 \notin \langle x_1; x_n \rangle$ (x_0 neleží mezi pozorovanými hodnotami x_i), se proces předvídání nazývá **extrapolace**. Vzhledem k tomu, že se jak intervalový odhad střední hodnoty závislé proměnné, tak i intervalový odhad individuální hodnoty, rozšiřují s rostoucí vzdáleností od $\overline{x}$, tak čím vzdálenější je x_0 od $\overline{x}$, tím větší riziko podstupujeme. Riziko výrazně roste v případě extrapolace. V podstatě platí, že vyrovnávací křivka proložená naměřenými body popisuje chování procesu pouze v rozsahu období, které je těmito body pokryto. Prodloužení vyrovnávací křivky mimo toto období (extrapolace) je možné, ale jen do jisté míry a jen s jistým stupněm důvěryhodnosti. My jsme se seznámili s metodami, které umožňují onu důvěryhodnost určit.

Příklad demagogie v regresi:
V civilizovaných zemích klesá dětská úmrtnost a v jistém období lze tento pokles graficky znázornit klesající přímkou. Je zřejmé, že takováto přímka nemůže být libovolně prodloužena. Procento úmrtí prostě nemůže být záporné. V jistém okamžiku se tedy příslušná přímka „zalomí“ v oblouk a časem se zhruba ustálí na nějaké téměř konstantní úrovni. V Británii nastal onen okamžik zlomu v době, kdy začalo hromadné očkování dětí. Pro odpůrce očkování a příslušníky různých extrémních sekt to byl dokonalý statistický důkaz škodlivosti očkování.

Příklad 11.8. S využitím odhadu regresního modelu (řešený příklad 11.4.2) pro data z motivačního příkladu odhadněte se spolehlivostí 0,95

a) střední výnos pšenice na polích, na nichž bylo použito 350 [kg/ha] hnojiva,

b) výnos pšenice na poli pana Nováka, který použil 350 [kg/ha] hnojiva.

Řešení 11.7

Shrnutí:

Statistika se zabývá analýzou stochastických závislosti, kdy závisle proměnná Y má charakter náhodné veličiny a nezávisle proměnné $x_1, \ldots, x_k$ mohou být jak nenáhodnými (pevnými), tak náhodnými veličinami. V rámci analýzy závislosti kvantitativních proměnných řešíme dvě základní úlohy.

- Informace o způsobu (tvaru) závislosti mezi kvantitativními znaky nám umožňuje získat regresní analýza.
- Popisem síly nalezené lineární závislosti se zabývá korelační analýza.

Doporučený postup při regresní a korelační analýze

1. Explorační analýza korelačního pole (případný odhad typu regresní funkce, identifikace vlivných bodů)
2. Odhad koeficientů regresní funkce (aplikace vyrovnávacího kritéria – např. metody nejmenších čtverců)
3. Verifikace modelu, tj. ověření předpokladů lineárního modelu
   a) *Celkový F-test* – testujeme, zda hodnota vysvětlované proměnné závisí na lineární kombinaci vysvětlujících proměnných, tj. testujeme nulovou hypotézu $H_0 : \beta_1 = \cdots = \beta_k$ vůči alternativě $H_A : \left(\overline{H}_0\right)$. Pokud bychom nulovou hypotézu nezamítli, znamenalo by to, že model je chybně specifikován.
   b) *Dílčí t-testy* - umožňují testovat oprávněnost setrvání vysvětlující proměnné v regresním modelu. Testujeme (postupně pro jednotlivá i) nulovou hypotézu ve tvaru

$H_0 : \beta_i = 0$ vůči alternativě $H_A : \beta_i \neq 0$ pro $i = 0, 1, \ldots, k$. Pokud pro konkrétní i nelze zamítnout nulovou hypotézu, je třeba zvážit setrvání příslušné vysvětlující proměnné v modelu.

c) *Analýza reziduí* – ověřujeme předpoklady pro použití lineárního regresního modelu.

- ověření normality reziduí - testy dobré shody,
- ověření nulovosti střední hodnoty - vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (rezidua musí kolísat kolem nuly) + dvouvýběrový t test,
- ověření homoskedasticity – vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (rezidua se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími odhadovanými hodnotami),
- ověření autokorelace reziduí - vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (autokorelace projeví tak, že se rezidua systematicky snižují nebo zvyšují, resp. můžeme mezi reziduí a předpovídanými hodnotami pozorovat nelineární závislost) + Durbinova-Watsonova statistika.

d) *Multikolinearita* – v případě vícenásobné regrese musíme ověřit, zda neexistuje multikolinearita mezi regresory.

e) *Ověření kvality modelu* – *index determinace* R^2 (udává kolik procent vysvětlované proměnné bylo vysvětleno modelem), *koeficient korelace* r (míra korelace mezi závisle proměnnou a regresorem v případě přímkové regrese), *koeficient vícenásobné korelace* $r_{(Y \cdot x_1, x_2, \ldots, x_k)}$ (míra korelace mezi závisle proměnnou na lineární kombinaci regresorů $x_1, x_2, \ldots, x_k$), koeficienty parciální korelace, např. $r_{(Y, x_1 \cdot x_2, \ldots, x_k)}$ (míra korelace mezi závisle proměnnou a jedním z regresorů při vyloučení vlivu ostatních regresorů).

4. **Využití verifikovaného modelu k predikci** – odhad střední hodnoty závisle proměnné při daných hodnotách regresorů (pás spolehlivosti), odhad individuální hodnoty závisle proměnné při daných hodnotách regresorů (pás predikce). **Pozor na extrapolaci!**

Test

Jak pracovat s testy?

Zacátek testu

1. (1b.) Regresní a korelační analýza umožňuje získat informace o
   (a) tvaru a síle závislosti mezi kvalitativními proměnnými,
   (b) tvaru a síle závislosti mezi kvantitativními proměnnými,
   (c) tvaru a síle závislosti mezi kvantitativními proměnnými, mezi nimiž je lineární vztah.

2. (1b.) V případě, že jsou splněny předpoklady lineárního regresního modelu, pak metoda nejmenších čtverců umožňuje nalézt
   (a) přesný funkční předpis hledané regresní funkce,
   (b) index determinace,
   (c) nejlepší odhad koeficientů hledané regresní funkce.

3. (1b.) Lze metodami lineární regrese nalézt regresní funkci ve tvaru mocninné funkce $\big(f : y = ax^b$, kde $a, b \in R \setminus \{0\}\big)$?
   (a) Ano, tato funkce je lineární vzhledem k parametrům.
   (b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
   (c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.

(d) Ne, toto lze řešit metodami nelineární regrese.

4. (1b.) Lze metodami lineární regrese nalézt regresní funkci ve tvaru polynomické funkce $\left(f : y = a_0 + a_1x^b + a_2x^2 + \ldots + a_nx^n\right.$, kde $\left.a, b \in \mathbb{R}, n \in \mathbb{N}\right)$?

(a) Ano, tato funkce je lineární vzhledem k parametrům.
(b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
(c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.
(d) Ne, toto lze řešit metodami nelineární regrese.

5. (1b.) Lze metodami lineární regrese nalézt regresní funkci $f : y = a_0 + a_1 e^{a_2 x}$, kde $a_0 \in \mathbb{R}$, $a_1, a_2 \in \mathbb{R} \setminus \{0\}$?

(a) Ano, tato funkce je lineární vzhledem k parametrům.
(b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
(c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.
(d) Ne, toto lze řešit metodami nelineární regrese.

6. (1b.) Koeficienty regresní funkce jsou

(a) konstanty,
(b) náhodné veličiny.

7. (1b.) Index determinace může nabývat hodnot z intervalu

(a) $\langle -1; 1 \rangle$,
(b) $\langle 0; 1 \rangle$,
(c) $\langle 0; \infty)$.

8. (1b.) Rezidua jsou odchylky

(a) pozorovaných a odhadovaných hodnot závislé proměnné,
(b) pozorovaných a odhadovaných hodnot nezávislé proměnné,
(c) pozorovaných a odhadovaných regresních funkcí.

9. (1b.) S rostoucím rozptylem reziduí se odhad rozptylu odhadů regresních koeficientů

(a) zvyšuje,
(b) snižuje.

10. (1b.) S rostoucím rozptylem jednotlivých regresorů se odhad rozptylu odhadů regresních koeficientů

(a) zvyšuje,
(b) snižuje.

11. (1b.) K ověření, zda hodnota vysvětlované proměnné závisí na lineární kombinaci všech vysvětlujících proměnných, používáme

(a) celkový F-test,
(b) dílčí t-testy,
(c) analýzu reziduí,
(d) index determinace.

12. (1b.) K testování oprávněnosti setrvání jednotlivých vysvětlujících proměnných v regresním modelu používáme

(a) celkový F-test,
(b) dílčí t-testy,
(c) analýzu reziduí,
(d) index determinace.

13. (1b.) Rezidua považujeme za nekorelované, pokud Durbin-Watsonova statistika leží v intervalu

(a) $\langle -1; 1\rangle$,
(b) $\langle 0; 1\rangle$,
(c) $\langle 0; \infty)$,
(d) $\langle -1; 1\rangle$,
(e) $\langle 1,4; 2,6\rangle$,
(f) $\langle 1,4; 2,6)$ nebo $(2,6; \infty)$.

14. (1b.) Pojmem multikolinearita označujeme

(a) lineární závislost mezi vysvětlovanou proměnnou a jednotlivými vysvětlujícími proměnnými,
(b) lineární závislost mezi vysvětlujícími proměnnými,
(c) lineární závislost mezi vysvětlovanými proměnnými,
(d) lineární závislost mezi jednotlivými regresními funkcemi.

15. (1b.) Pás spolehlivosti (odhad střední hodnoty závisle proměnné při daných hodnotách regresorů) je

(a) stejně široký jako,
(b) širší než,
(c) užší než.

16. (1b.) Odhad závislé proměnné pro hodnoty regresorů ležící mimo interval pozorovaných hodnot označujeme jako

(a) interpolaci,
(b) extrapolaci,
(c) korelaci.

Konec testu

Správně zodpovězené otázky:

Získané body:

Procento úspěšnosti:

Výsledky

Příklady k procvičení

1. Byla vyšetřována výška dvaceti 18letých mladíků y a výška jejich rodičů a prarodičů $(x_1, x_2, \ldots, x_7)$ a hledaná lineární závislost mezi závisle proměnnou y a nezávisle proměnnými $x_1, x_2, \ldots, x_7$. Všechny výšky jsou uvedeny v [cm].

Regresor	význam
x_1	výška matky v jejím věku 18 let
x_2	výška otce v jeho věku 18 let
x_3	výška babičky z matčiny strany v jejím věku 18 let
x_4	výška dědečka z matčiny strany v jeho věku 18 let
x_5	výška babičky z otcovy strany v jejím věku 18 let
x_6	výška dědečka z otcovy strany v jeho věku 18 let
x_7	výška 18-ti letého chlapce

a) Sestavte vhodný lineární model a testujte statistickou významnost parametrů $\beta_0, \ldots, \beta_7$.

b) Rozhodněte mezi dvěma navrženými regresními modely:
model A: $y = f(x_1, x_2, \ldots, x_7)$, model B: $y = f(x_1, x_2, x_3, x_4)$.

c) Verifikujte vybraný model (celkový F-test, dílčí t-testy, analýza reziduí).

d) Na základě informací o novorozeném Honzíkovi odhadněte jeho výšku v 18 letech. $x_1 = 50,8, x_2 = 152,4, x_3 = 182,9, x_4 = 154,9, x_5 = 180,3, x_6 = 157,7, x_7 = 177,8$. (Pro řešení použijte statistický software.)

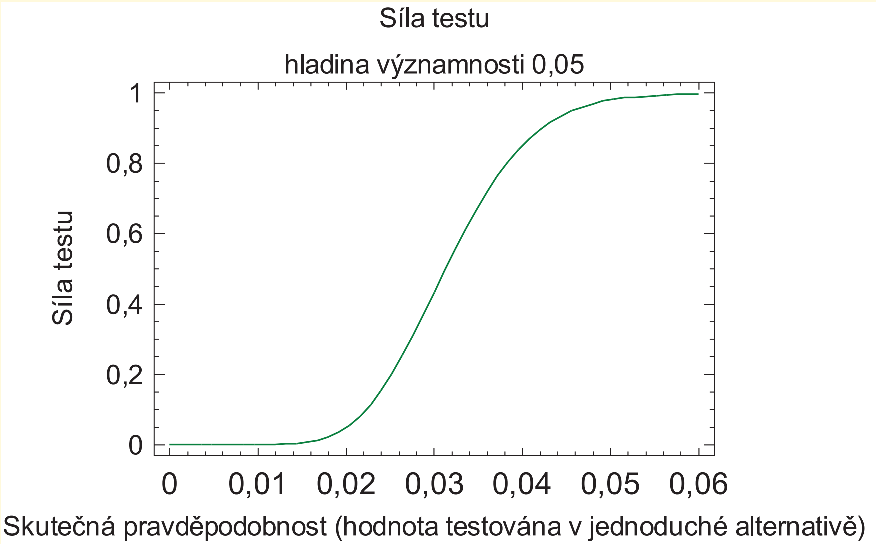
x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
50,00	153,70	178,60	166,90	176,00	166,90	170,90	170,70
49,80	164,80	178,80	159,00	176,80	164,10	168,70	175,50
49,30	166,10	167,10	168,10	174,80	162,60	176,30	170,20
49,30	161,00	182,60	154,20	172,70	164,80	170,40	183,90
50,00	165,40	165,40	166,40	166,40	157,00	180,10	161,50
49,80	165,60	180,60	161,30	168,10	170,90	174,20	184,70
50,30	163,30	172,50	158,50	181,40	161,00	176,30	174,00
50,00	165,90	174,80	156,20	167,60	158,50	172,00	177,00
50,00	163,80	174,50	162,30	174,80	158,20	174,80	173,70
50,50	161,00	178,60	167,40	175,30	161,80	165,40	178,80
48,00	160,80	178,80	161,80	175,80	168,10	174,00	171,50
52,80	168,10	178,30	166,10	169,20	156,70	162,60	186,20
51,60	164,80	174,80	165,60	178,30	158,50	170,20	177,80
50,00	161,30	178,60	160,30	163,60	165,40	170,20	177,30
50,50	157,50	166,40	162,80	172,00	157,70	168,90	161,50
49,80	161,30	165,60	162,30	177,80	163,10	163,80	163,30
54,10	167,90	166,10	164,60	173,70	168,70	179,80	174,00
51,10	164,60	178,30	165,90	166,40	161,80	169,90	179,10
51,30	159,00	174,20	161,80	177,30	169,40	172,70	173,00
48,80	158,00	170,90	161,50	180,10	161,50	169,40	167,90

Řešení příkladu

Obsah
456. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Klíč k příkladům k procvičení

1. $H_0 : \pi = 0,02, H_A : \pi > 0,02$ (minimální požadovaný rozsah výběru je 224)
   a) $0,02 \notin \langle 0,026; 0,063\rangle$, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme říci, že se kvalita chladících zařízení zhoršila.
   b) *p-hodnota* $= 0,007$, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme říci, že se kvalita chladících zařízení zhoršila.
   c)



Zpět na příklad

2. $H_0 : \mu = 14000, H_A : \mu > 14000$, *p-hodnota* $= 0,038$, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme říci, že nová technologie vedla ke zvýšení životnosti žárovek.

Zpět na příklad

3. a) $H_0 : \mu = 1,97, H_A : \mu \neq 1,97$, *p-hodnota* $=$ 0,34, proto na 5% hladině významnosti nezamítáme nulovou hypotézu, tzn. nemůžeme tvrdit, že nový způsob krmení vedl ke změně hmotnosti kaprů.

   b) $H_0 : \mu = 1,97, H_A : \mu > 1,97$, *p-hodnota* $=$ 0,17, proto na 5% hladině významnosti nezamítáme nulovou hypotézu, tzn. nemůžeme tvrdit, že nový způsob krmení vedl ke zvýšení hmotnosti kaprů.

Zpět na příklad

4. $H_0 : x_{0,5} = 640, H_A : x_{0,5} > 640$,
   mediánový test: *p-hodnota* $=$ 0,01,
   Wilcoxonův test: kritická hodnota jednovýběrového Wilcoxonova testu pro hladinu významnosti 0,05 ω_{10} (0,05) je 8. Pozorovaná hodnota (1) je menší než kritická hodnota (8).
   Na základě výsledku obou testu lze říci, že na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme tvrdit, že změna posloupnosti tepelných úprav ke zvýšení střední meze pevnosti.

Zpět na příklad

5. $H_0 : \pi = 0,01, H_A : \pi > 0,01$ (minimální požadovaný rozsah výběru je 610)
   p-hodnota $=$ 0,10, proto na 5% hladině významnosti nezamítáme nulovou hypotézu, tzn. daný výsledek potvrzuje tvrzení firmy TT.

Zpět na příklad

Obsah

458. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

6. $H_0 : \mu = 1000, H_A : \mu < 1000$, *p-hodnota* $=$ 0,0005, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme říci, že zjištěný rozdíl je známkou nekvality produkce.

   Zpět na příklad

7. $H_0 : \pi = 0,2, H_A : \pi < 0,2$ (minimální požadovaný rozsah výběru je 66)
   p-hodnota $=$ 0,03, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme tvrdit, že úvaha představenstva není reálná.

   Zpět na příklad

8. $H_0 : \sigma = 0,05, H_A : \sigma < 0,05$,
   p-hodnota $=$ 0,005, proto na 5% hladině významnosti zamítáme nulovou hypotézu, tzn. můžeme tvrdit, že došlo ke zlepšení kvality výroby.

   Zpět na příklad

9. $H_0 : \sigma = 1562, H_A : \sigma > 1562$, *p-hodnota* $=$ 0,22, proto na 5% hladině významnosti nezamítáme nulovou hypotézu, tzn. nelze tvrdit, že organizační změny prohloubily diferenciaci mezd.

   Zpět na příklad

Klíč k příkladům k procvičení

1. $H_0 : \sigma_S^2 = \sigma_J^2, H_A : \sigma_S^2 > \sigma_J^2, p\text{-}hodnota = 0,15 \Rightarrow$ nezamítáme hypotézu o shodě rozptylů $\Rightarrow$ pro ověření shody středních hodnot použijeme dvouvýběrový t test (máme k dispozici dva nezávislé výběry z normálního rozdělení).

   Zpět na příklad

   $H_0 : \mu_S = \mu_J, H_A : \mu_S > \mu_J, p\text{-}hodnota \doteq 0 \Rightarrow$ zamítáme hypotézu o shodě středních hodnot, tzn. lze tvrdit, že severní výjezd je zatíženější.

2. $H_0 : \pi_(18-30) = \pi_(30-50), H_A : \pi_(18-30) < \pi_(30-50)$, minimální požadované rozsahy: $n_{1_{min}} = 50, n_{2_{min}} = 41, p\text{-}hodnota = 0,004 \Rightarrow$ zamítáme hypotézu o homogenitě dvou binomických rozdělení, tzn. můžeme tvrdit, že lidé ve věku 18 až 30 let považují jaderné elektrárny za bezpečnější než lidé ve věku 30 až 50 let.

   Zpět na příklad

3. a) $H_0 : \pi_{MM} = \pi_{PP}, H_A : \pi_{MM} < \pi_{PP}$, minimální požadované rozsahy: $n_{1_{min}} = 139$, $n_{2_{min}} =$ $= 100$, $p\text{-}hodnota = 0,20 \Rightarrow$ nezamítáme hypotézu o homogenitě dvou binomických rozdělení, tzn. tvrzení firmy MM nelze označit za pravdivé.

   b) minimální požadované rozsahy: $n_{1_{min}} = 139, n_{2_{min}} = 100, P(\pi_{MM} - \pi_{PP} \in \langle -0,095; 0,035\rangle) =$ $= 0,95; 0 \in \langle -0,095; 0,035\rangle \Rightarrow$ nezamítáme hypotézu o homogenitě dvou binomických rozdělení, tzn. tvrzení firmy MM nelze označit za pravdivé.

   c) $P(\pi_{MM} \in \langle 0,035; 0,105\rangle = 0,95$

   Zpět na příklad

4. $H_0 : x_{0,5_A} = x_{0,5_B}, H_A : x_{0,5_A} > x_{0,5_B}$, pozorovaná hodnota (3) je menší nebo rovna kritické hodnotě (3) $\Rightarrow$ zamítáme hypotézu o shodě mediánů, tzn. lze tvrdit, že denní přírůstky vah selat

Obsah

460. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

jsou vyšší při použití krmné směsi A. (Mannův-Whitneyův test byl zvolen z důvodů porušení normality.)

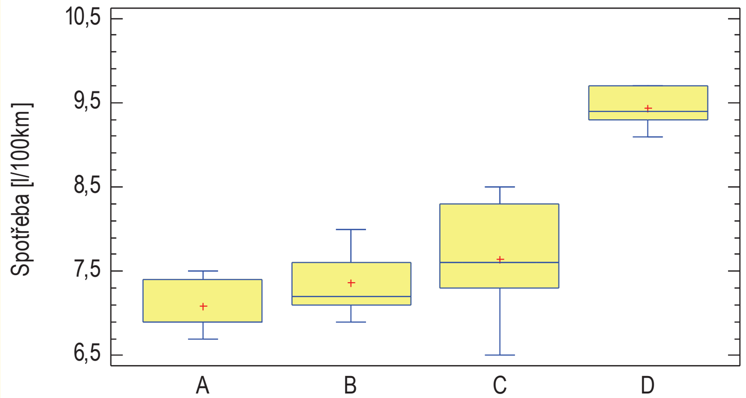
Zpět na příklad

5. Označme: $X \ldots$ hmotnost před dietou, $Y \ldots$ hmotnost po dietě.
   Párový t test, $D_i = Y_i - X_i, H_0 : \mu_D = 0, H_A : \mu_D < 0, p\text{-}hodnota = 0,004 \Rightarrow$ zamítáme nulovou hypotézu, tzn. lze tvrdit, že dietní přípravek je účinný (po dietě došlo ke statisticky významnému poklesu hmotnosti).

   Zpět na příklad

Klíč k příkladům k procvičení

1)



Ověření homoskedasticity
H_0: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$, $H_A : \neg H_0$
Bartlettův test: $x_{OBS} = 1,4$, *p-hodnota*$= 0,15 \Rightarrow$ Na hladině významnosti 0,05 nezamítáme H_0.
ANOVA, H_0: $\mu_A = \mu_B = \mu_C = \mu_D$, $H_A : \neg H_0$

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	17,008	3	5,66933	22,00	0,0000
Within groups	4,124	16	0,25775		
Total (Corr.)	21,132	19			

Obsah
462. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

p-hodnota$= 0,0000 \Rightarrow$ Na hladině významnosti 0,05 zamítáme H_0.

Post hoc analýza:

	Count	Mean	Homogeneous Groups
A	5	7,08	X
B	5	7,36	X
C	5	7,64	X
D	5	9,44	X

Contrast	Difference	+/- Limits
A - B	-0,28	0,919072
A - C	-0,56	0,919072
A - D	*-2,36	0,919072
B - C	-0,28	0,919072
B - D	*-2,08	0,919072
C - D	*-1,8	0,919072

* denotes a statistically significant difference.

Zpět na příklad

2) Kruskalův-Wallisův test, $H_0 : x_{0,5_R} = x_{0,5_P} = x_{0,5_O} = x_{0,5_V}$, $H_A : \neg H_0$
$x_{OBS} = 6,65$, *p-hodnota*$= 0,08 \Rightarrow$ Na hladině významnosti 0,05 nezamítáme H_0. Rozdíl v hodnocení produktů jednotlivých výrobců není statisticky významný.

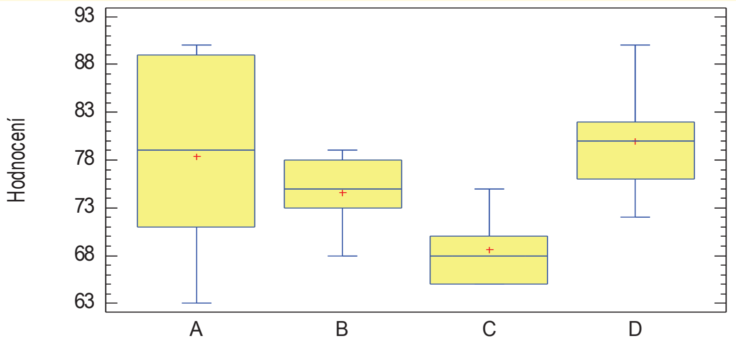
Zpět na příklad

Obsah
463. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

3) Friedmanův test, $H_0 : x_{0,5_R} = x_{0,5_P} = x_{0,5_O} = x_{0,5_V}$, $H_A : \neg H_0$
$x_{OBS} = 8,20$, *p-hodnota*$= 0,046 \Rightarrow$ Na hladině významnosti 0,05 zamítáme H_0. **Post hoc analýza:** Kritická hodnota Friedmanova testu: 15,6

$\vert R_I - R_J \vert$				
	Ráno	Poledne	Odpoledne	Večer
Ráno	-	6	6	10
Poledne		-	12	4
Odpoledne			-	16
Večer				-

Jako statisticky významný byl na hladině významnosti identifikován rozdíl ve schopnosti vidění odpoledne a večer.



Zpět na příklad

Obsah
464. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Klíč k příkladům k procvičení

1. H_0: Pravděpodobnost „počtu ok“ na kostce je dána následující tabulkou:

x_i (číslo které může padnout)	1	2	3	4	5	6
$\pi_{0,i}$ (nulová pravděpodobnost jeho výskytu)	1/6	1/6	1/6	1/6	1/6	1/6

H_A: $\neg H_0$, tj. pravděpodobnost „počtu ok“ na kostce je jiná, než je uvedeno ve výše uvedené tabulce.

χ^2 test dobré shody: $x_{OBS} = 2,93$, *p-hodnota*$= 0,71$ (viz vybrana_rozdeleni.xls)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, tj. nelze tvrdit, že kostka není „férová“.

Zpět na příklad

2.

H_0: Generovaný výběr pochází z rozdělení $R(0;1)$.
H_A: Generovaný výběr nepochází z rozdělení $R(0;1)$.

χ^2 test dobré shody: $x_{OBS} = 16,75$, *p-hodnota*$= 0,053$ (viz vybrana_rozdeleni.xls)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, tj. nelze tvrdit, že generátor negeneruje čísla z rozdělení $R(0;1)$.

Zpět na příklad

3.

H_0: Chyba měření má rozdělení dané hustotou pravděpodobnosti $f_0(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, $x \in \mathbb{R}$.
H_A: Chyba měření nemá rozdělení dané hustotou pravděpodobnosti $f_0(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, $x \in \mathbb{R}$.

$F_0(x) = \frac{1}{\pi} \cdot \operatorname{arctg}(x) + \frac{1}{2}, x \in \mathbb{R}$
χ^2 test dobré shody: $x_{OBS} = 8,70$, *p-hodnota*$= 0,12$ (viz vybrana_rozdeleni.xls)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, tj. nelze tvrdit, že chyba měření nemá rozdělení dané hustotou pravděpodobnosti $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, x \in \mathbb{R}$.

Zpět na příklad

4.
H_0: Chyba měření má rozdělení dané hustotou pravděpodobnosti $f_0(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, $x \in \mathbb{R}$.
H_A: Chyba měření nemá rozdělení dané hustotou pravděpodobnosti $f_0(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, $x \in \mathbb{R}$.

$F_0(x) = \frac{1}{\pi} \cdot \operatorname{arctg}(x) + \frac{1}{2}, x \in \mathbb{R}$

Kolmogorovův-Smirnovův test:

Seřazené hodnoty $x_{(i)}$	Pořadí i	$\frac{i-1}{n}$	$\frac{i}{n}$	$F_0(x_{(i)})$	$\left\lvert F_0(x_{(i)}) - \frac{i}{n} \right\rvert$	$\left\lvert F_0(x_{(i)}) - \frac{i-1}{n} \right\rvert$	D_i^*
-2,1	1	0,00	0,10	0,141	0,141	0,041	0,141
-1,7	2	0,10	0,20	0,169	0,069	0,031	0,069
-1,1	3	0,20	0,30	0,235	0,035	0,065	0,065
-0,2	4,5	0,30	0,40	0,437	0,137	0,037	0,137
-0,2	4,5	0,40	0,50	0,437	0,037	0,063	0,063
0,5	6	0,50	0,60	0,648	0,148	0,048	0,148
0,6	7	0,60	0,70	0,672	0,072	0,028	0,072
0,8	8	0,70	0,80	0,715	0,015	0,085	0,085
1,3	9	0,80	0,90	0,791	0,009	0,109	0,109
2,3	10	0,90	1,00	0,869	0,031	0,131	0,131

$x_{OBS} = 0,148,\ D_{10(0,05)} = 0,40925.$

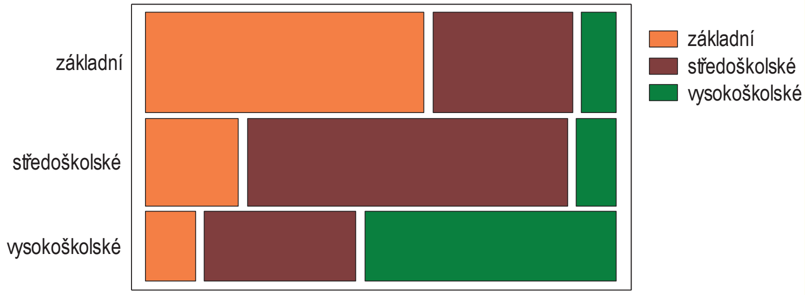
Pozorovaná hodnota $x_{OBS} = 0,148$ je menší než kritická hodnota $D_{10(0,05)} = 0,40925$, proto na hladině významnosti 0,05 nezamítáme nulovou hypotézu, tj. nelze tvrdit, že chyba měření nemá rozdělení dané hustotou pravděpodobnosti $f_0(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, x \in \mathbb{R}$.

Zpět na příklad

Klíč k příkladům k procvičení

1.

H_0 : Vzdělání nevěsty a ženicha jsou nezávislé veličiny.
H_A : Vzdělání nevěsty a ženicha nejsou nezávislé veličiny.



χ^2 test nezávislosti: Všechny očekávané četnosti jsou větší než 5.

$$x_{OBS} = 43,2; \quad p\text{-}hodnota \ll 0,001 \quad \text{(viz vybrana\_rozdeleni.xls)}$$

Na hladině významnosti 0,10 zamítáme nulovou hypotézu ve prospěch alternativy. Nelze tvrdit, že věk nevěsty a ženicha jsou nezávislé veličiny.

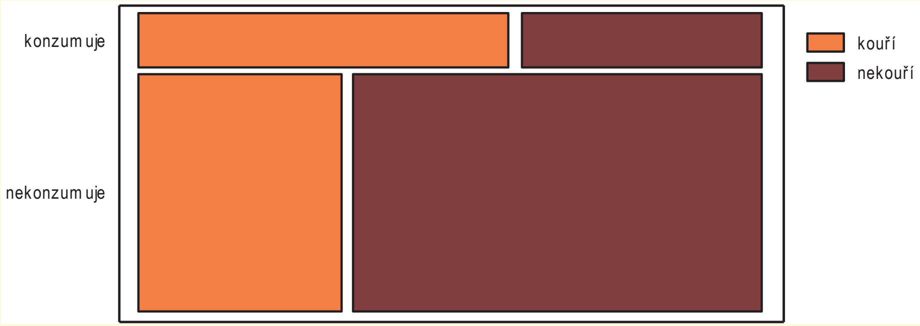
Na základě koeficientu kontingence ($CC = 0,55$) a Cramerova koeficientu ($V = 0,46$) lze usuzovat na poměrně silnou závislost mezi věkem nevěsty a ženicha.

Zpět na příklad

2.

H_0 : Ukončení léčby odvykání kouření a konzumace alkoholu jsou nezávislé veličiny.
H_A : Ukončení léčby odvykání kouření a konzumace alkoholu jsou závislé veličiny.

Obsah
468. strana ze 525
Zavřít dokument
Celá obrazovka/Okno



Odhad šance na úspěšné ukončení léčby u populace, která konzumuje alkohol je 0,65, tzn. že ve skupině pacientů konzumujících alkohol připadá cca 650 pacientů, kteří úspěšně ukončí léčbu odvykání kouření na 1000 pacientů, kteří léčbu neukončí.

Odhad šance na úspěšné ukončení léčby u populace, která nekonzumuje alkohol je 2,0, tzn. že ve skupině pacientů nekonzumujících alkohol připadají 2 pacienti, kteří úspěšně ukončí léčbu odvykání kouření na 1 pacienta, který léčbu neukončí.

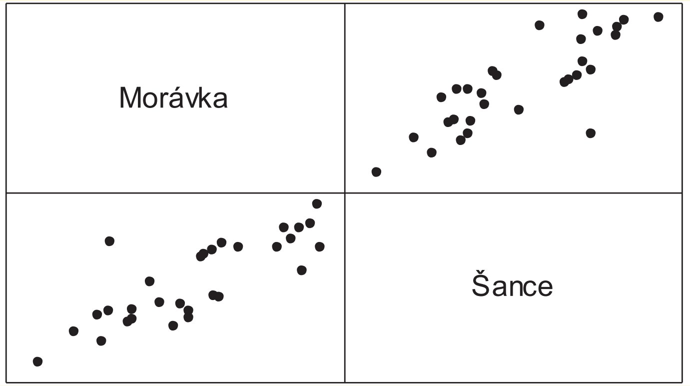
Poměr šancí odhadujeme na 0,325. Se spolehlivostí 95% lze očekávat poměr šancí v intervalu $\langle 0,17; 0,62\rangle$. Je zcela zřejmé, že konzumace alkoholu statisticky významně snižuje šanci na úspěšné ukončení léčby odvykání kouření $(1 \notin \langle 0,17; 0,62\rangle, OR < 1)$.

Obdobně: $\widehat{RR} = 1,8$, riziko, že pacient neukončí úspěšně léčbu odvykání kouření je 1,8x vyšší u pacientů konzumujících alkohol. 95% intervalový odhad RR je $\langle 1,3; 2,6\rangle$. Je zřejmé, že konzumace alkoholu statisticky významně zvyšuje riziko, že pacient neukončí úspěšně léčbu odvykání kouření $(1 \notin \langle 1,3; 2,6\rangle, RR > 1)$.

Zpět na příklad

Obsah
469. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

3.



H_0 : Analyzovaná data jsou výběrem z normálního rozdělení.
H_A : Analyzovaná data nejsou výběrem z normálního rozdělení.

χ^2 test dobré shody: $p\text{-}hodnota_{Morávka} = 0,13$; $p\text{-}hodnota_{Sance} = 0,055$

Na hladině významnosti 0,05 nelze zamítnout normalitu ročních průměrných průtoků profilem jak nádrže Morávka, tak nádrže Šance. Nutnou podmínku proto, aby výběr pocházel z dvourozměrného normálního rozdělení, lze považovat za splněnou.

$H_0 : \rho = 0, \qquad H_A : \rho > 0.$
$r = 0,81, x_{OBS} = 7,41, p\text{-}hodnota \ll 0,001$. Na hladině významnosti 0,05 lze zamítnout nulovou hypotézu ve prospěch alternativy, tzn. roční průtoky profily nádrží Morávka a Šance jsou kladně korelované.

Zpět na příklad

Obsah
470. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

4. Analyzujeme závislost ordinálních veličin, které obsahují mnoho shod, proto použijeme korigovaný Spearmanův korelační koeficient.

$T_X = 45, T_Y = 3, \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2 = 28,5, r_{Skorig} = 0,866.$

H_0 : Agresivita chování a školní prospěch jsou nezávislé veličiny.
H_A : Agresivita chování a školní prospěch nejsou nezávislé veličiny.

$r_S^*(0,05; 11) = 0,6091$

$|r_{Skorig}| \geqq r_S^*(0,05; 11)$, proto na hladině významnosti 0,05 zamítáme nulovou hypotézu ve prospěch alternativy. Agresivitu chování a školní prospěch nelze považovat za nezávislé veličiny.

Zpět na příklad

Obsah
471. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

Klíč k příkladům k procvičení

1. Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-193,625	68,6542	-2,8203	0,0155
X1	1,40221	0,529691	2,64723	0,0213
X2	0,772311	0,202763	3,80894	0,0025
X3	1,04776	0,136019	7,70301	0,0000
X4	-0,124488	0,173203	-0,71874	0,4861
X5	0,0718802	0,130565	0,550532	0,5921
X6	0,091777	0,162634	0,564317	0,5829
X7	-0,106723	0,156239	-0,68308	0,5075

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	864,102	7	123,443	18,61	0,0000
Residual	79,6096	12	6,63413		
Total (Corr.)	943,712	19			

R-squared = 91,5642 percent
R-squared (adjusted for d.f.) = 86,6433 percent
Standard Error of Est. = 2,57568
Mean absolute error = 1,44594
Durbin-Watson statistic = 2,47495 (P=0,1614)
Lag 1 residual autocorrelation = -0,276503

The StatAdvisor

a) Metodou nejmenších čtverců byl nalezen odhad regresní funkce ve tvaru

$$\begin{aligned} Y &= -193,625 + 1,40221 * X1 + 0,772311 * X2 + 1,04776 * X3 \\ &\quad -0,124488 * X4 + 0,0718802 * X5 + 0,091777 * X6 - 0,106723 * X7 \end{aligned}$$

Na hladině významnosti 0,05 zamítáme hypotézu $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$ $(p - hodnota \ll 0,001)$. Celkový F-test tak ukazuje na správnou specifikaci modelu.

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu $H_0 : \beta_i = 0$ pro $i = 3,4$ až 7 $(p - hodnota \ll 0,05)$. Dílčí t-testy ukazují na možnost zjednodušení modelu. Regresory $x_3, x_4, \ldots, x_7$ lze z modelu vypustit.

(Index determinace pro tento model je $86,6\,\%$.)

b)

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-199,722	33,79	-5,91069	0,0000
X1	1,3728	0,451027	3,04372	0,0077
X2	0,688085	0,161561	4,25898	0,0006
X3	1,10592	0,0994201	11,1237	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	853,698	3	284,566	50,58	0,0000
Residual	90,0145	16	5,6259		
Total (Corr.)	943,712	19			

R-squared = 90,4617 percent
R-squared (adjusted for d.f.) = 88,6732 percent
Standard Error of Est. = 2,3719
Mean absolute error = 1,63014
Durbin-Watson statistic = 2,36332 (P=0,1811)
Lag 1 residual autocorrelation = -0,241111

The equation of the fitted model is

Y = -199,722 + 1,3728*X1 + 0,688085*X2 + 1,10592*X3

Zjednodušený regresní model (model B) byl odhadnut ve tvaru

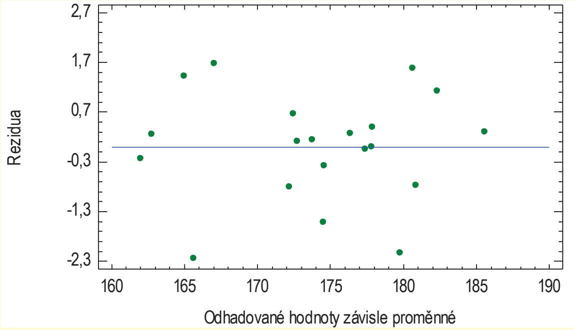
$$Y = -199,722 + 1,3728 * X1 + 0,688085 * X2 + 1,10592 * X3.$$

Na hladině významnosti 0,05 zamítáme hypotézu $H_0: \ \beta_1 = \beta_2 = \beta_3 = 0$ (p–*hodnota* $\ll 0,001$). Celkový F-test tak ukazuje na správnou specifikaci modelu.

Na hladině významnosti 0,05 zamítáme nulovou hypotézu $H_0: \ \beta_i = 0$ pro $i = 0, 1, 2, 3$ (p – – *hodnota* $\ll 0,05$). Model již nejde dále zjednodušit.

Modifikovaný index determinace pro tento model je $88,7\,\%$., tzn. že jej lze považovat za model kvalitnější než model A. Model B vysvětluje $90,5\,\%$ rozptylu závisle proměnné.

Analýza reziduí



c) a) normalita reziduí
H_0 : Rezidua mají normální rozdělení.
H_A : Rezidua nemají normální rozdělení.

$p - hodnota = 0,20$ (χ^2 test dobré shody)

Na hladině významnosti 0,05 nezamítáme normalitu reziduí.

b) nulovost střední hodnoty reziduí

$$\begin{array}{rcl} H_0: & E(\mathrm{e}_i) & = & 0 \\ H_A: & E(\mathrm{e}_i) & \neq & 0 \end{array}$$

$p - hodnota = 0,999$ (jednovýběrový t test)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, předpoklad o nulovosti střední hodnoty reziduí lze považovat za splněný.

c) homoskedasticita reziduí
Na grafu reziduí a odhadovaných hodnot závisle proměnné nepozorujeme zvyšování ani snižování rozptylu reziduí s rostoucími odhady závisle proměnné, předpoklad homoskedasticity proto považujeme za splněný.

d) autokorelace reziduí
Rezidua se systematicky nesnižují ani nezvyšují, mezi reziduí a předpovídanými hodnotami nepozorujeme ani nelineární závislost. Durbin-Watsonova statistika nabývá hodnoty $2,36 \in \langle 1,4; 2,6 \rangle$, rezidua můžeme považovat za nekorelovaná. (Všimněte si, že Statgraphics poskytuje rovněž $p - hodnotu$ pro test autokorelace reziduí.)

Multikolinearita:

Correlations

	X1	X2	X3
X1		0,4310	-0,1791
X2	0,4310		-0,0951
X3	-0,1791	-0,0951	

Absolutní hodnoty jednoduchých korelačních koeficientů žádné z dvojic regresorů nepřekročily hodnotu 0,8. Regresory lze považovat ze nekorelované.

Nalezený model splňuje předpoklady lineárního regresního modelu a je dostatečně kvalitní, proto jej lze použít pro predikci.

d) Se spolehlivostí 0,95 lze očekávat, že Honzík bude v 18 letech mít výšku z intervalu $\langle 170,5; 183,8\rangle$ cm.

Zpět na příklad

Obsah

477. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Kapitola 12

Statistické tabulky

T1. Distribuční funkce normovaného normálního rozdělení $\Theta(x)$ pro $x > 0$

$$\Theta(-x) = 1 - \Theta(x)$$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944

Obsah
479. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

x	0	1	2	3	4	5	6	7	8	9
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999
3,0	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,1	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,2	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

T2. Vybrané kvantily normovaného normálního rozdělení

$$z_{1-\alpha} = -z_\alpha$$

α	0,1000	0,0500	0,0250	0,0100	0,0050	0,0010	0,0005	0,0001
z_α	1,2816	1,6449	1,9600	2,3263	2,5758	3,0902	3,2905	3,7190

T3. Vybrané kvantily χ^2 rozdělení s v stupni volnosti

stupně volnosti ν	α							
	0,0001	0,0005	0,01	0,025	0,05	0,1	0,25	0,5
1	0,000	0,000	0,000	0,001	0,004	0,016	0,102	0,455
2	0,000	0,001	0,020	0,051	0,103	0,211	0,575	1,386
3	0,005	0,015	0,115	0,216	0,352	0,584	1,213	2,366
4	0,028	0,064	0,297	0,484	0,711	1,064	1,923	3,357
5	0,082	0,158	0,554	0,831	1,145	1,610	2,675	4,351
6	0,172	0,299	0,872	1,237	1,635	2,204	3,455	5,348
7	0,300	0,485	1,239	1,690	2,167	2,833	4,255	6,346
8	0,464	0,710	1,646	2,180	2,733	3,490	5,071	7,344
9	0,661	0,972	2,088	2,700	3,325	4,168	5,899	8,343
10	0,889	1,265	2,558	3,247	3,940	4,865	6,737	9,342
11	1,145	1,587	3,053	3,816	4,575	5,578	7,584	10,341
12	1,427	1,934	3,571	4,404	5,226	6,304	8,438	11,340
13	1,733	2,305	4,107	5,009	5,892	7,042	9,299	12,340
14	2,061	2,697	4,660	5,629	6,571	7,790	10,165	13,339
15	2,408	3,108	5,229	6,262	7,261	8,547	11,037	14,339
16	2,774	3,536	5,812	6,908	7,962	9,312	11,912	15,338
17	3,157	3,980	6,408	7,564	8,672	10,085	12,792	16,338
18	3,555	4,439	7,015	8,231	9,390	10,865	13,675	17,338

19	3,968	4,912	7,633	8,907	10,117	11,651	14,562	18,338
20	4,395	5,398	8,260	9,591	10,851	12,443	15,452	19,337
21	4,835	5,896	8,897	10,283	11,591	13,240	16,344	20,337
22	5,286	6,404	9,542	10,982	12,338	14,041	17,240	21,337
23	5,749	6,924	10,196	11,689	13,091	14,848	18,137	22,337
24	6,223	7,453	10,856	12,401	13,848	15,659	19,037	23,337
25	6,707	7,991	11,524	13,120	14,611	16,473	19,939	24,337
26	7,200	8,538	12,198	13,844	15,379	17,292	20,843	25,336
27	7,702	9,093	12,879	14,573	16,151	18,114	21,749	26,336
28	8,213	9,656	13,565	15,308	16,928	18,939	22,657	27,336
29	8,731	10,227	14,256	16,047	17,708	19,768	23,567	28,336
30	9,258	10,804	14,953	16,791	18,493	20,599	24,478	29,336
40	14,883	16,906	22,164	24,433	26,509	29,051	33,660	39,335
50	21,009	23,461	29,707	32,357	34,764	37,689	42,942	49,335
60	27,497	30,340	37,485	40,482	43,188	46,459	52,294	59,335
70	34,261	37,467	45,442	48,758	51,739	55,329	61,698	69,334
80	41,244	44,791	53,540	57,153	60,391	64,278	71,145	79,334
100	55,725	59,896	70,065	74,222	77,929	82,358	90,133	99,334
120	70,728	75,467	86,923	91,573	95,705	100,624	109,220	119,334

T3. Vybrané kvantily χ^2 rozdělení s v stupni volnosti (pokračování)

stupně volnosti ν	α						
	0,75	0,9	0,95	0,975	0,99	0,995	0,999
1	1,323	2,706	3,841	5,024	6,635	7,879	10,828
2	2,773	4,605	5,991	7,378	9,210	10,597	13,816
3	4,108	6,251	7,815	9,348	11,345	12,838	16,266
4	5,385	7,779	9,488	11,143	13,277	14,860	18,467
5	6,626	9,236	11,070	12,833	15,086	16,750	20,515
6	7,841	10,645	12,592	14,449	16,812	18,548	22,458
7	9,037	12,017	14,067	16,013	18,475	20,278	24,322
8	10,219	13,362	15,507	17,535	20,090	21,955	26,124
9	11,389	14,684	16,919	19,023	21,666	23,589	27,877
10	12,549	15,987	18,307	20,483	23,209	25,188	29,588
11	13,701	17,275	19,675	21,920	24,725	26,757	31,264
12	14,845	18,549	21,026	23,337	26,217	28,300	32,909
13	15,984	19,812	22,362	24,736	27,688	29,819	34,528
14	17,117	21,064	23,685	26,119	29,141	31,319	36,123
15	18,245	22,307	24,996	27,488	30,578	32,801	37,697
16	19,369	23,542	26,296	28,845	32,000	34,267	39,252
17	20,489	24,769	27,587	30,191	33,409	35,718	40,790
18	21,605	25,989	28,869	31,526	34,805	37,156	42,312

19	22,718	27,204	30,144	32,852	36,191	38,582	43,820
20	23,828	28,412	31,410	34,170	37,566	39,997	45,315
21	24,935	29,615	32,671	35,479	38,932	41,401	46,797
22	26,039	30,813	33,924	36,781	40,289	42,796	48,268
23	27,141	32,007	35,172	38,076	41,638	44,181	49,728
24	28,241	33,196	36,415	39,364	42,980	45,559	51,179
25	29,339	34,382	37,652	40,646	44,314	46,928	52,620
26	30,435	35,563	38,885	41,923	45,642	48,290	54,052
27	31,528	36,741	40,113	43,195	46,963	49,645	55,476
28	32,620	37,916	41,337	44,461	48,278	50,993	56,892
29	33,711	39,087	42,557	45,722	49,588	52,336	58,301
30	34,800	40,256	43,773	46,979	50,892	53,672	59,703
40	45,616	51,805	55,758	59,342	63,691	66,766	73,402
50	56,334	63,167	67,505	71,420	76,154	79,490	86,661
60	66,981	74,397	79,082	83,298	88,379	91,952	99,607
70	77,577	85,527	90,531	95,023	100,425	104,215	112,317
80	88,130	96,578	101,879	106,629	112,329	116,321	124,839
100	109,141	118,498	124,342	129,561	135,807	140,169	149,449
120	130,055	140,233	146,567	152,211	158,950	163,648	173,617

Obsah
485. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

T4. Vybrané kvantily Studentova rozdělení s v stupni volnosti

$$t_{1-\alpha} = -t_\alpha$$

stupně volnosti ν	α								
	0,75	0,9	0,95	0,975	0,99	0,995	0,9975	0,999	0,9995
1	1,000	3,078	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922

19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
70	0,678	1,294	1,667	1,994	2,381	2,648	2,899	3,211	3,435
80	0,678	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
100	0,677	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
120	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Obsah
487. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

T5. Vybrané kvantily Fisherova-Snedecorova rozdělení s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli

$$f_\alpha(m;n) = \frac{1}{f_{1-\alpha}(m;n)}$$

n	α	m								
		1	2	3	4	5	6	7	8	9
1	0,95	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54
	0,975	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28
	0,99	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47
2	0,95	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38
	0,975	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39
	0,99	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39
3	0,95	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
	0,975	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47
	0,99	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35
4	0,95	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
	0,975	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90
	0,99	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66
5	0,95	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
	0,975	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
	0,99	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16

	0,95	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
6	0,975	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
	0,99	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98
	0,95	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
7	0,975	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
	0,99	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72
	0,95	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
8	0,975	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
	0,99	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91
	0,95	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
9	0,975	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
	0,99	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35
	0,95	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
10	0,975	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
	0,99	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94
	0,95	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
11	0,975	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
	0,99	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63

T5. Vybrané kvantily Fisherova-Snedecorova rozdělení s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli (pokračování)

$$f_\alpha(m;n) = \frac{1}{f_{1-\alpha}(m;n)}$$

n	α	m									
		10	12	15	20	24	30	40	60	120	∞
1	0,95	241,88	243,91	245,95	248,01	249,05	250,10	251,14	252,20	253,25	254,31
	0,975	968,63	976,71	984,87	993,10	997,25	1001,41	1005,60	1009,80	1014,02	1018,25
	0,99	6055,85	6106,32	6157,28	6208,73	6234,63	6260,65	6286,78	6313,03	6339,39	6365,83
2	0,95	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
	0,975	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
	0,99	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	0,95	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
	0,975	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
	0,99	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	0,95	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	0,975	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
	0,99	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	0,95	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
	0,975	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
	0,99	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	0,95	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	0,975	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
	0,99	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88

	0,95	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
7	0,975	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
	0,99	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
	0,95	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
8	0,975	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
	0,99	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
	0,95	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
9	0,975	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
	0,99	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
	0,95	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
10	0,975	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
	0,99	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
	0,95	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
11	0,975	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
	0,99	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60

Obsah
491. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

T5. Vybrané kvantily Fisherova-Snedecorova rozdělení s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli (pokračování)

$$f_\alpha(m;n) = \frac{1}{f_{1-\alpha}(m;n)}$$

n	α	m = 1	2	3	4	5	6	7	8	9
12	0,95	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
	0,975	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
	0,99	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39
14	0,95	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
	0,975	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21
	0,99	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03
16	0,95	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
	0,975	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
	0,99	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78
18	0,95	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
	0,975	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
	0,99	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60
20	0,95	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
	0,975	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
	0,99	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46

24	0,95	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
	0,975	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
	0,99	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26
30	0,95	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
	0,975	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
	0,99	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07
40	0,95	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
	0,975	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45
	0,99	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89
60	0,95	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
	0,975	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
	0,99	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72
120	0,95	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96
	0,975	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
	0,99	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56
∞	0,95	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88
	0,975	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11
	0,99	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41

T5. Vybrané kvantily Fisherova-Snedecorova rozdělení s m stupni volnosti v čitateli a n stupni volnosti ve jmenovateli (pokračování)

$$f_\alpha(m;n) = \frac{1}{f_{1-\alpha}(m;n)}$$

n	α	m									
		10	12	15	20	24	30	40	60	120	∞
12	0,95	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
	0,975	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,73
	0,99	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
14	0,95	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
	0,975	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
	0,99	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
16	0,95	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
	0,975	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
	0,99	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
18	0,95	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
	0,975	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
	0,99	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
20	0,95	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
	0,975	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
	0,99	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42

24	0,95	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
	0,975	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
	0,99	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
30	0,95	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
	0,975	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
	0,99	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	0,95	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
	0,975	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
	0,99	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	0,95	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
	0,975	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
	0,99	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	0,95	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
	0,975	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
	0,99	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	0,95	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,01
	0,975	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,01
	0,99	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,01

T6. Kritické hodnoty jednovýběrového Wilcoxonova testu

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
6	0	-
7	2	-
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	3
20	52	37
21	58	42
22	65	48

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
36	208	171
37	221	182
38	235	194
39	249	207
40	264	220
41	279	233
42	294	247
43	310	261
44	327	276
45	343	291
46	361	307
47	378	322
48	396	339
49	415	355
50	434	373
51	453	390
52	473	408

Obsah
496. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

23	73	54
24	81	61
25	89	68
26	98	75
27	107	83
28	116	91
29	126	100
30	137	109
31	147	118
32	159	128
33	170	138
34	182	148
35	195	159

53	494	427
54	514	445
55	536	465
56	557	484
57	579	504
58	602	525
59	625	546
60	648	567
61	672	589
62	697	611
63	721	634
64	747	657
65	772	681

Zdroj: [1], tabulka T4

T7. Kritické hodnoty Mannova-Whitneyova testu

$\alpha = 0{,}05$	n																		
m	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	-	-	0																
5	-	0	1	2															
6	-	1	2	3	5														
7	-	1	3	5	6	8													
8	0	2	4	6	8	10	13												
9	0	2	4	7	10	12	15	17											
10	0	3	5	8	11	14	17	20	23										
11	0	3	6	9	13	16	19	23	26	30									
12	1	4	7	11	14	18	22	26	29	33	37								
13	1	4	8	12	16	20	24	28	33	37	41	45							
14	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		

19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	2	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	161
26	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200

Zdroj: [1], tabulka T8

T8. Kritické hodnoty $h_\alpha(k, v)$ Hartlyova testu

$\alpha = 0{,}05$	k										
stupně volnosti ν	2	3	4	5	6	7	8	9	10	11	12
2	39	87,5	142	202	266	333	403	475	550	626	704
3	15,4	27,8	39,2	50,7	62	72,9	83,5	93,9	104	114	124
4	9,6	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48	51,4
5	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
6	5,82	8,38	10,4	12,1	13,7	15	16,3	17,5	18,6	19,7	20,7
7	4,99	6,94	8,44	9,7	10,8	11,8	12,7	13,5	14,3	15,1	15,8
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7
9	4,03	5,34	6,31	7,11	7,8	8,41	8,95	9,45	9,91	10,3	10,7
10	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
12	3,28	4,16	4,79	5,3	5,72	6,09	6,42	6,72	7,00	7,25	7,48
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,4	5,59	5,77	5,93
20	2,46	2,95	3,29	3,54	3,76	3,94	4,1	4,24	4,37	4,49	4,59
30	2,07	2,4	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,3	2,33	2,36
∞	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

T8. Kritické hodnoty $h_\alpha(k, v)$ Hartlyova testu (pokračování)

$\alpha = 0{,}01$	*I*										
stupně volnosti ν	2	3	4	5	6	7	8	9	10	11	12
2	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
3	47,5	85	120	151	184	216	249	281	310	337	361
4	23,2	37	49	59	69	79	89	97	106	113	120
5	14,9	22	28	33	38	42	46	50	54	57	60
6	11,1	15,5	19,1	22	25	27	30	32	34	36	37
7	8,89	12,1	14,5	16,5	18,4	20	22	23	24	26	27
8	7,5	9,9	11,7	13,2	14,5	15,8	16,9	17,9	18,9	19,8	21
9	6,54	8,5	9,9	11,1	12,1	13,1	13,9	14,7	15,3	16	16,6
10	5,85	7,4	8,6	9,6	10,4	11,1	11,8	12,4	12,9	13,4	13,9
12	4,91	6,1	6,9	7,6	8,2	8,7	9,1	9,5	9,9	10,2	10,6
15	4,07	4,9	5,5	6	6,4	6,7	7,1	7,3	7,5	7,8	8
20	3,32	3,8	4,3	4,6	4,9	5,1	5,3	5,5	5,6	5,8	5,9
30	2,63	3	3,3	3,4	3,6	3,7	3,8	3,9	4	4,1	4,2
60	1,96	2,2	2,3	2,4	2,4	2.5	2,5	2,6	2,6	2,7	2,7
∞	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

Zdroj: [1], tabulka T13

T9. Kritické hodnoty $c_\alpha(k, v)$ Cochranova testu

$\alpha = 0{,}05$	k										
stupně volnosti ν	2	3	4	5	6	7	8	9	10	11	12
1	1,00	0,97	0,91	0,84	0,78	0,73	0,68	0,64	0,60	0,57	0,54
2	0,98	0,87	0,77	0,68	0,62	0,56	0,52	0,48	0,44	0,42	0,39
3	0,94	0,80	0,68	0,60	0,53	0,48	0,44	0,40	0,37	0,35	0,33
4	0,91	0,75	0,63	0,54	0,48	0,43	0,39	0,36	0,33	0,31	0,29
5	0,88	0,71	0,59	0,51	0,44	0,40	0,36	0,33	0,30	0,28	0,26
6	0,85	0,68	0,56	0,48	0,42	0,37	0,34	0,31	0,28	0,26	0,24
7	0,83	0,65	0,54	0,46	0,40	0,35	0,32	0,29	0,27	0,25	0,23
8	0,82	0,63	0,52	0,44	0,38	0,34	0,30	0,28	0,25	0,24	0,22
9	0,80	0,62	0,50	0,42	0,37	0,33	0,29	0,27	0,24	0,23	0,21
10	0,79	0,60	0,49	0,41	0,36	0,32	0,28	0,26	0,24	0,22	0,20
12	0,77	0,58	0,47	0,39	0,34	0,30	0,27	0,24	0,22	0,20	0,19
15	0,74	0,55	0,44	0,37	0,32	0,28	0,25	0,23	0,21	0,19	0,18
20	0,71	0,52	0,42	0,35	0,30	0,26	0,23	0,21	0,19	0,18	0,16
30	0,67	0,49	0,38	0,32	0,27	0,24	0,21	0,19	0,17	0,16	0,15
60	0,62	0,44	0,34	0,28	0,24	0,21	0,18	0,16	0,15	0,14	0,13
120	0,59	0,41	0,32	0,26	0,22	0,19	0,17	0,15	0,13	0,12	0,11

T9. Kritické hodnoty $c_\alpha(k, v)$ Cochranova testu (pokračování)

$\alpha = 0{,}01$	k										
stupně volnosti ν	2	3	4	5	6	7	8	9	10	11	12
1	1,00	0,99	0,97	0,93	0,88	0,84	0,79	0,75	0,72	0,68	0,65
2	1,00	0,94	0,86	0,79	0,72	0,66	0,62	0,57	0,54	0,50	0,48
3	0,98	0,88	0,78	0,70	0,63	0,57	0,52	0,48	0,45	0,42	0,39
4	0,96	0,83	0,72	0,63	0,56	0,51	0,46	0,43	0,39	0,37	0,34
5	0,94	0,79	0,68	0,59	0,52	0,47	0,42	0,39	0,36	0,33	0,31
6	0,92	0,76	0,64	0,55	0,49	0,43	0,39	0,36	0,33	0,31	0,29
7	0,90	0,73	0,61	0,53	0,46	0,41	0,37	0,34	0,31	0,29	0,27
8	0,88	0,71	0,59	0,50	0,44	0,39	0,35	0,32	0,29	0,27	0,25
9	0,87	0,69	0,57	0,49	0,42	0,38	0,34	0,31	0,28	0,26	0,24
10	0,85	0,67	0,55	0,47	0,41	0,36	0,32	0,30	0,27	0,25	0,23
12	0,83	0,65	0,53	0,44	0,39	0,34	0,30	0,28	0,25	0,23	0,22
15	0,80	0,61	0,50	0,42	0,36	0,32	0,28	0,26	0,23	0,22	0,20
20	0,77	0,58	0,46	0,39	0,33	0,29	0,26	0,23	0,21	0,20	0,18
30	0,72	0,53	0,42	0,35	0,30	0,26	0,23	0,21	0,19	0,17	0,16
60	0,66	0,47	0,37	0,30	0,26	0,22	0,20	0,18	0,16	0,15	0,14
120	0,62	0,43	0,33	0,27	0,23	0,20	0,17	0,16	0,14	0,13	0,12

Zdroj: [1], tabulka T14

Obsah
503. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

T10. Kritické hodnoty $q_\alpha(k, v)$ studentizovaného testu

$\alpha = 0{,}05$	k													
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	18	27	32,8	37,1	40,4	43,1	45,4	47,4	49,1	50,6	52	53,2	54,3	55,4
2	6,08	8,33	9,8	10,9	11,7	12,4	13	13,5	14	14,4	14,7	15,1	15,4	15,7
3	4,5	5,91	6,82	7,5	8,04	8,48	8,85	9,18	9,46	9,72	9,95	10,2	10,3	10,5
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,6	7,83	8,03	8,21	8,37	8,52	8,66
5	3,64	4,6	5,22	5,67	6,03	6,33	6,58	6,8	6,99	7,17	7,32	7,47	7,6	7,72
6	3,46	4,34	4,9	5,3	5,63	5,9	6,12	6,32	6,49	6,65	6,79	6,92	7,03	7,14
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,3	6,43	6,55	6,66	6,76
8	3,26	4,04	4,53	4,89	5,17	5,4	5,6	5,77	5,92	6,05	6,18	6,29	6,39	6,48
9	3,2	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	5,87	5,98	6,09	6,19	6,28
10	3,15	3,88	4,33	4,65	4,91	5,12	5,3	5,46	5,6	5,72	5,83	5,93	6,03	6,11
11	3,11	3,82	4,26	4,57	4,82	5,03	5,2	5,35	5,49	5,61	5,71	5,81	5,9	5,98
12	3,08	3,77	4,2	4,51	4,75	4,95	5,12	5,27	5,39	5,51	5,61	5,71	5,8	5,88
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79
14	3,03	3,7	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,64	5,71
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,2	5,31	5,4	5,49	5,57	5,65
16	3	3,65	4,05	4,33	4,56	4,74	4,9	5,03	5,15	5,26	5,35	5,44	5,52	5,59
17	2,98	3,63	4,02	4,3	4,52	4,7	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,54
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,43	5,5
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,31	5,39	5,46
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,9	5,01	5,11	5,2	5,28	5,36	5,43

24	2,92	3,53	3,9	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,1	5,18	5,25	5,32
30	2,89	3,49	3,85	4,1	4,3	4,46	4,6	4,72	4,82	4,92	5,0	5,08	5,15	5,21
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	4,82	4,9	4,98	5,04	5,11
60	2,83	3,4	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,0
120	2,8	3,36	3,68	3,92	4,1	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,9
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,74	4,8

Zdroj: [1], tabulka T11

T10. Kritické hodnoty $q_\alpha(k, v)$ studentizovaného testu (pokračování)

$\alpha = 0{,}01$	k													
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	90	135	164	186	202	216	227	237	246	253	260	266	272	277
2	14	19	22,3	24,7	26,6	28,2	29,5	30,7	31,7	32,6	33,4	34,1	34,8	35,4
3	8,26	10,6	12,2	13,3	14,2	15	15,6	16,2	16,7	17,1	17,5	17,9	18,2	18,5
4	6,51	8,12	9,17	9,96	10,6	11,1	11,5	11,9	12,3	12,6	12,8	13,1	13,3	13,5
5	5,7	6,97	7,8	8,42	8,91	9,32	9,67	9,97	10,2	10,5	10,7	10,9	11,1	11,2
6	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,1	9,3	9,49	9,65	9,81	9,95
7	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37	8,55	8,71	8,86	9	9,12
8	4,74	5,63	6,2	6,63	6,96	7,24	7,47	7,68	7,87	8,03	8,18	8,31	8,44	8,55
9	4,6	5,43	5,96	6,35	6,66	6,91	7,13	7,32	7,49	7,65	7,78	7,91	8,03	8,13
10	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21	7,36	7,48	7,6	7,71	7,81
11	4,39	5,14	5,62	5,97	6,25	6,48	6,67	6,84	6,99	7,13	7,25	7,36	7,46	7,56
12	4,32	5,04	5,5	5,84	6,1	6,32	6,51	6,67	6,81	6,94	7,06	7,17	7,26	7,36
13	4,26	4,96	5,4	5,73	5,98	6,19	6,37	6,53	6,67	6,79	6,9	7,01	7,1	7,19
14	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54	6,66	6,77	6,87	6,96	7,05
15	4,17	4,83	5,25	5,56	5,8	5,99	6,16	6,31	6,44	6,55	6,66	6,76	6,84	6,93
16	4,13	4,78	5,19	5,49	5,72	5,92	6,08	6,22	6,35	6,46	6,56	6,66	6,74	6,82
17	4,1	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27	6,38	6,48	6,57	6,66	6,73
18	4,07	4,7	5,09	5,38	5,6	5,79	5,94	6,08	6,2	6,31	6,41	6,5	6,58	6,65
19	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14	6,25	6,34	6,43	6,51	6,58
20	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09	6,19	6,29	6,37	6,45	6,52

Obsah
506. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

24	3,96	4,54	4,91	5,17	5,37	5,54	5,69	5,81	5,92	6,02	6,11	6,19	6,26	6,33
30	3,89	4,45	4,8	5,05	5,24	5,4	5,54	5,65	5,76	5,85	5,93	6,01	6,08	6,14
40	3,82	4,37	4,7	4,93	5,11	5,27	5,39	5,5	5,6	5,69	5,77	5,84	5,9	5,96
60	3,76	4,28	4,6	4,82	4,99	5,13	5,25	5,36	5,45	5,53	5,6	5,67	5,73	5,79
120	3,7	4,2	4,5	4,71	4,87	5,01	5,12	5,21	5,3	5,38	5,44	5,51	5,56	5,61
∞	3,64	4,12	4,4	4,6	4,76	4,88	4,99	5,08	5,16	5,23	5,29	5,35	5,4	5,45

Zdroj: [1], tabulka T11

T11. Kritické hodnoty vícenásobného porovnávání pomocí pořadí

$\alpha = 0{,}05$	k							
m	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9	10,5	12	13,5
2	8,8	12,6	16,5	20,5	24,7	28,9	33,1	37,4
3	15,7	22,7	29,9	37,3	44,8	52,5	60,3	68,2
4	23,9	34,6	45,6	57	68,6	80,4	92,4	104,6
5	33,1	48,1	63,5	79,3	95,5	112	128,8	145,8
6	43,3	62,9	83,2	104	125,3	147	169,1	191,4
7	54,4	79,1	104,6	130,8	157,6	184,9	212,8	240,9
8	66,3	96,4	127,6	159,6	192,4	225,7	259,7	294,1
9	78,9	114,8	152	190,2	229,3	269,1	309,6	350,6
10	92,3	134,3	177,8	222,6	268,4	315	362,4	410,5
11	106,3	154,8	205	256,6	309,4	363,2	417,9	473,3
12	120,9	176,2	233,4	292,2	352,4	413,6	476	539,1
13	136,2	198,5	263	329,3	397,1	466,2	536,5	607,7
14	152,1	221,7	293,8	367,8	443,6	520,8	599,4	679
15	168,6	245,7	325,7	407,8	491,9	577,4	664,6	752,8
16	185,6	270,6	358,6	449,1	541,7	635,9	732,0	829,2

$\alpha = 0{,}01$	k							
m	3	4	5	6	7	8	9	10
1	4,1	5,7	7,3	8,9	10,5	12,2	13,9	15,6
2	10,9	15,3	19,7	24,3	28,9	33,6	38,3	43,1
3	19,5	27,5	35,7	44	52,5	61,1	69,8	78,6
4	29,7	41,9	54,5	67,3	80,3	93,6	107	120,6
5	41,2	58,2	75,8	93,6	111,9	130,4	149,1	168,1
6	53,9	76,3	99,3	122,8	146,7	171	195,7	220,6
7	67,6	95,8	124,8	154,4	184,6	215,2	246,3	277,7
8	82,4	116,8	152,2	188,4	225,2	262,6	300,6	339
9	98,1	139,2	181,4	224,5	268,5	313,1	358,4	404,2
10	114,7	162,8	212,2	262,7	314,2	366,5	419,5	473,1
11	132,1	187,6	244,6	302,9	362,2	422,6	483,7	545,6
12	150,4	213,5	278,5	344,9	412,5	481,2	551	621,4
13	169,4	240,6	313,8	388,7	464,9	542,4	621	700,5
14	189,1	268,7	350,5	434,2	519,4	606	693,8	782,6
15	209,6	297,8	388,5	481,3	575,8	671,9	769,3	867,7
16	230,7	327,9	427,9	530,1	634,2	740,0	847,3	955,7

Zdroj: [1], tabulka T15

T12. Kritické hodnoty Friedmanova testu

$\alpha = 0{,}05$	k									
m	3	4	5	6	7	8	9	10	11	12
3	6	7,4	8,53	9,86	11,24	12,57	13,88	15,19	16,48	17,76
4	6,5	7,8	8,8	10,24	11,63	12,99	14,34	15,67	16,98	18,3
5	6,4	7,8	8,99	10,43	11,84	13,23	14,59	15,93	17,27	18,6
6	7	7,6	9,08	10,54	11,97	13,38	14,76	16,12	17,4	18,8
7	7,143	7,8	9,11	10,62	12,07	13,48	14,87	16,23	17,6	18,9
8	6,25	7,65	9,19	10,68	12,14	13,56	14,95	16,32	17,7	19
9	6,222	7,66	9,22	10,73	12,19	13,61	15,02	16,4	17,7	19,1
10	6,2	7,67	9,25	10,76	12,23	13,66	15,07	16,44	17,8	19,2
11	6,545	7,68	9,27	10,79	12,27	13,7	15,11	16,48	17,9	19,2
12	6,167	7,7	9,29	10,81	12,29	13,73	15,15	16,53	17,9	19,3
13	6	7,7	9,3	10,83	12,32	13,76	15,17	16,56	17,9	19,3
14	6,143	7,71	9,32	10,85	12,34	13,78	15,19	16,58	17,9	19,3
15	6,4	7,72	9,33	10,87	12,35	13,8	15,2	16,6	18	19,3
16	5,99	7,73	9,34	10,88	12,37	13,81	15,23	16,6	18	19,3
20	5,99	7,74	9,37	10,92	12,41	13,8	15,3	16,7	18	19,4
∞	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92	18,31	19,68

Obsah
510. strana ze 525
Zavřít dokument
Celá obrazovka/Okno

$\alpha = 0{,}01$	k									
m	3	4	5	6	7	8	9	10	11	12
3	-	9	10,13	11,76	13,26	14,78	16,28	17,74	19,19	20,61
4	8	9,6	11,2	12,59	14,19	15,75	17,28	18,77	20,24	21,7
5	8,4	9,96	11,43	13,11	14,74	16,32	17,86	19,37	20,86	22,3
6	9	10,2	11,75	13,45	15,1	16,69	18,25	19,77	21,3	22,7
7	8,857	10,371	11,97	13,69	15,35	16,95	18,51	20,04	21,5	23
8	9	10,35	12,14	13,87	15,53	17,15	18,71	20,24	21,8	23,2
9	8,667	10,44	12,27	14,01	15,68	17,29	18,87	20,42	21,9	23,4
10	9,6	10,53	12,38	14,12	15,79	17,41	19	20,53	22	23,5
11	9,455	10,6	12,46	14,21	15,89	17,52	19,1	20,64	22,1	23,6
12	9,5	10,68	12,53	14,28	15,96	17,59	19,19	20,73	22,2	23,7
13	9,385	10,72	12,58	14,34	16,03	17,67	19,25	20,8	22,3	23,8
14	9	10,76	12,64	14,4	16,09	17,72	19,31	20,86	22,4	23,9
15	8,933	10,8	12,68	14,44	16,14	17,78	19,35	20,9	22,4	23,9
16	8,79	10,84	12,72	14,48	16,18	17,81	19,4	20,9	22,5	24
20	8,87	10,94	12,83	14,6	16,3	18,0	19,5	21,1	22,6	24,1
∞	9,21	11,45	13,28	15,09	16,81	18,48	20,09	21,67	23,21	24,73

T13. Kritické hodnoty vícenásobného porovnávání u Friedmanova testu

$\alpha = 0{,}05$	k							
m	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9	10,5	12	13,5
2	4,7	6,6	8,6	10,7	12,7	14,8	17	19,2
3	5,7	8,1	10,6	13,1	15,6	18,2	20,8	23,5
4	6,6	9,4	12,2	15,1	18	21	24	27,1
5	7,4	10,5	13,6	16,9	20,1	23,5	26,9	30,3
6	8,1	11,5	14,9	18,5	22,1	25,7	29,4	33,2
7	8,8	12,4	16,1	19,9	23,9	27,8	31,8	35,8
8	9,4	13,3	17,3	21,3	25,5	29,7	34	38,3
9	9,9	14,1	18,3	22,6	27	31,5	36	40,6
10	10,5	14,8	19,3	23,8	28,5	33,2	38	42,8
11	11	15,6	20,2	25	29,9	34,8	39,8	44,9
12	11,5	16,2	21,1	26,1	31,2	36,4	41,6	46,9
13	11,9	16,9	22	27,2	32,5	37,9	43,3	48,8
14	12,4	17,5	22,8	28,2	33,7	39,3	45	50,7
15	12,8	18,2	23,6	29,2	34,9	40,7	46,5	52,5
16	13,3	18,8	24,4	30,2	36	42	48,1	54,2

$\alpha = 0{,}01$	k							
m	3	4	5	6	7	8	9	10
1	4,1	5,7	7,3	8,9	10,5	12,2	13,9	15,6
2	5,8	8	10,3	12,6	14,9	17,3	19,7	22,1
3	7,1	9,8	12,6	15,4	18,3	21,2	24,1	27
4	8,2	11,4	14,6	17,8	21,1	24,4	27,8	31,2
5	9,2	12,7	16,3	19,9	23,6	27,3	31,1	34,9
6	10,1	13,9	17,8	21,8	25,8	29,9	34,1	38,2
7	10,9	15	19,3	23,5	27,9	32,3	36,8	41,3
8	11,7	16,1	20,6	25,2	29,8	34,6	39,3	44,2
9	12,4	17,1	21,8	26,7	31,6	36,6	41,7	46,8
10	13	18	23	28,1	33,4	38,6	44	49,4
11	13,7	18,9	24,1	29,5	35	40,5	46,1	51,8
12	14,3	19,7	25,2	30,8	36,5	42,3	48,2	54,1
13	14,9	20,5	26,2	32,1	38	44	50,1	56,3
14	15,4	21,3	27,2	33,3	39,5	45,7	52	58,4
15	16	22	28,2	34,5	40,8	47,3	53,9	60,5
16	16,5	22,7	29,1	35,6	42,2	48,9	55,6	62,5

Zdroj: [1], tabulka T17

T14. Kritické hodnoty jednovýběrového Kolmogorova-Smirnovova testu

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	0,975	0,995
2	0,84189	0,92929
3	0,7076	0,829
4	0,62394	0,73424
5	0,56328	0,66853
6	0,51926	0,61661
7	0,48342	0,57581
8	0,45427	0,54179
9	0,43001	0,51332
10	0,40925	0,48893
11	0,39122	0,4677
12	0,37543	0,44905
13	0,36143	0,43247
14	0,3489	0,41762
15	0,3376	0,4042
16	0,32733	0,39201
17	0,31796	0,38086
18	0,30936	0,37062
19	0,30143	0,36117
20	0,29408	0,35241

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
31	0,23788	0,2853
32	0,23424	0,28094
33	0,23076	0,27677
34	0,22743	0,27279
35	0,22425	0,26897
36	0,22119	0,26532
37	0,21826	0,2618
38	0,21544	0,25843
39	0,21273	0,25518
40	0,21012	0,25205
41	0,2076	0,24904
42	0,20517	0,24613
43	0,20283	0,24332
44	0,20056	0,2406
45	0,19837	0,23798
46	0,19625	0,23544
47	0,1942	0,23298
48	0,19221	0,23059
49	0,19028	0,22828
50	0,18841	0,22604

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
61	0,17091	0,20506
62	0,16956	0,20343
63	0,16823	0,20184
64	0,16693	0,20029
65	0,16567	0,19877
66	0,16443	0,19729
67	0,16322	0,19584
68	0,16204	0,19442
69	0,16088	0,19303
70	0,15975	0,19167
71	0,15864	0,19034
72	0,15755	0,18903
73	0,15649	0,18776
74	0,15544	0,1865
75	0,15442	0,18528
76	0,15342	0,18408
77	0,15244	0,1829
78	0,15147	0,18174
79	0,15052	0,1806
80	0,1496	0,17949

21	0,28724	0,34427
22	0,28087	0,33666
23	0,2749	0,32954
24	0,26931	0,32286
25	0,26404	0,31657
26	0,25907	0,31064
27	0,25438	0,30502
28	0,24993	0,29971
29	0,24571	0,29466
30	0,2417	0,29987

51	0,18659	0,22386
52	0,18482	0,22174
53	0,18311	0,21968
54	0,18144	0,21768
55	0,17981	0,21574
56	0,17823	21384
57	0,17669	0,21199
58	0,17519	0,21019
59	0,17373	0,20844
60	0,17231	0,20673

81	0,14868	0,1784
82	0,14779	0,17732
83	0,14691	0,17627
84	0,14605	0,17523
85	0,1452	0,17421
86	0,14437	0,17321
87	0,14355	0,17223
90	0,14177	0,16938
95	0,13746	0,16493
100	0,13403	0,16081

Zdroj: [1], tabulka T18

Obsah

515. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

T15. Kritické hodnoty Spearmanova korelačního koeficientu

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	0,9	-
6	0,8286	0,9429
7	0,745	0,8929
8	0,6905	0,8571
9	0,6833	0,8167
10	0,6364	0,7818

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
11	0,6091	0,7545
12	0,5804	0,7273
13	0,5549	0,6978
14	0,5341	0,6747
15	0,5179	0,6536
16	0,5	0,6324
17	0,4853	0,6152
18	0,4716	0,5975
19	0,4579	0,5825
20	0,4451	0,5684

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
21	0,4351	0,5545
22	0,4241	0,5426
23	0,415	0,5306
24	0,4061	0,52
25	0,3977	0,51
26	0,3894	0,5002
27	0,3822	0,4915
28	0,3749	0,4828
29	0,3685	0,4744
30	0,362	0,4665

Zdroj: [1], tabulka T22

Obsah

516. strana ze 525

Zavřít dokument

Celá obrazovka/Okno

Literatura

[1] Anděl, J.: *Základy matematické statistiky*, MatFyzPress, Praha 2007, ISBN: 80-7378-003-8.

[2] Anděl, J.: *Statistické metody*, MatFyzPress, Praha 2007, ISBN: 80-7378-001-1.

[3] Briš R., Litschmannová M., *Statistika I. pro kombinované a distanční studium*, Ostrava 2004, dostupné na: www.am.vsb.cz/litschmannova.

[4] Budíková, M., Lerch, T., Mikoláš, Š.: *Základní statistické metody*, Brno 2005, ISBN: 80-210-3886-1.

[5] Budíková, M., Mikoláš, Š., Osecký, P.: *Teorie pravděpodobnosti a matematická statistika*, Brno 2007, ISBN: 80-210-3313-4.

[6] Dummer: *Introduction to statistical science*, VŠB-TU Ostrava, Ostrava, 1998.

[7] Dummer, Klímková: *Statistika I. (cvičení)*, VŠB-TU Ostrava, Ostrava, 1997.

[8] Friedrich, V.: *Statistika I. – vysokoškolská učebnice*, Plzeň 2002

[9] Friesl, M.: *Posbírané příklady z pravděpodobnosti a statistiky*, 2004, dostupné na: http://home.zcu.cz/ friesl/Archiv/PosbPsa.pdf.

[10] Gibilisco, S.: *Statistika bez předchozích znalostí*, Brno 2009, ISBN: 978-80-251-2465-9.

[11] Kazmier, L., J., Pohl, N., F. : *Basic Statistics for Business and Economics*, Second Edition. McGraw-Hill, Inc., New York,1984.

[12] Kohout, P.: *Příklady z teorie pravděpodobnosti*, dostupné na: http://www.kmt.zcu.cz/person/Kohout/info_soubory/exam1.htm.

[13] Kupka, K.: *Statistické řízení jakosti*, Trilobyte 1997, ISBN: 80-238-1818-X.

[14] Lane, D.: *HyperStat Online Statistics Textbook*, dostupné na: http://davidmlane.com/hyperstat.

[15] Likeš, J., Machek, J.: *Počet pravděpodobnosti*, SNTL, Praha, 1981

[16] Likeš, J., Machek, J.: *Matematická statistika*, SNTL, Praha, 1983

[17] Likeš, J., Laga: *Základní statistické tabulky*, Praha, 1978

[18] Litschmannová, M.: *Statistika I. - řešené příklady*, 2007, dostupné na: www.am.vsb.cz/litschmannova

[19] Otipka, P., Šmajstrla, V.: *Pravděpodobnost a statistika*, dostupné na: http://homen.vsb.cz/ oti73/cdpast1/index.htm.

[20] Plocki, A., Tlustý, P.: *Pravděpodobnost a statistika pro začátečníky a mírně pokročilé*, Prometheus, Praha 2007, ISBN: 978-80-7196-330-1.

[21] Rosenthal, J.: *Zasažen bleskem*, Academia, Praha 2008, ISBN: 978-80-200-1645-4.

[22] Seger, J., Hindls, R., Hronová, S.: *Statistika v hospodářství*, Manager – Podnikatel, Praha 1998.

[23] Schindler, M.: *Příklady*, dostupné na: http://artax.karlin.mff.cuni.cz/ schim9am/priklady06.pdf.

[24] Sternstein, M.: *Barrons AP Statistics*, Barron's Educational Series, 2010, ISBN: 0764140892.

[25] Triola, M., F. : *Elementary Statistics*, Fourth Edition. The Benjamin/Cummings Publishing Company, Inc., Redwood City, California,1989.

[26] Wonnacot, T. H., Wonnacot, R. J.: *Statistika pro obchod a hospodářství*, Victoria Publishing, Praha 1992.

[27] Zvára, K., Štěpán, J.: *Pravděpodobnost a matematická statistika*, MatFyzPress, Praha 2006, ISBN: 80-86732-71-1.

Rejstřík