Vysoká škola báňská – Technická univerzita Ostrava
Západočeská univerzita v Plzni

Úvod do statistiky (interaktivní učební text) - Řešené příklady

Martina Litschmannová

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

Obsah

Kapitola 1

Explorační analýza proměnných - řešené příklady

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.

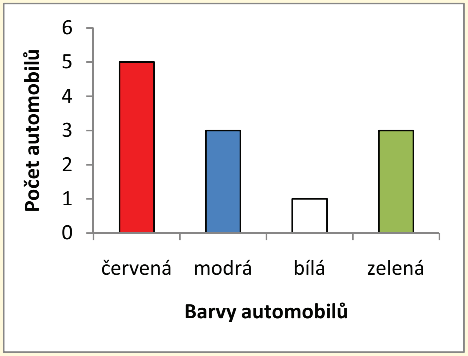
červená, modrá, zelená, modrá, červená, zelená, červená, červená, modrá, zelená, bílá, červená

Řešení. Je zřejmé, že se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že barvy automobilů nemá smysl seřazovat, víme, že se jedná o proměnnou nominální. Pro její popis proto zvolíme tabulku četností, určíme modus a barvu projíždějících automobilů

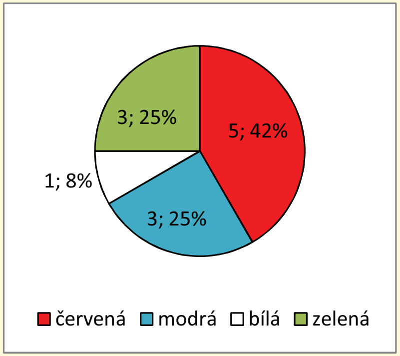
Tab. 1.1: Tabulka rozdělení četností pro pozorované barvy automobilů

TABULKA ROZDĚLENÍ ČETNOSTI		
Barvy projíždějících automobilů	**Absolutní četnost** $\mathbf{n_i}$	**Relativní četnost** $\mathbf{p_i}$
červená	5	$5/12 = 0{,}42$
modrá	3	$3/12 = 0{,}25$
bílá	1	$1/12 = 0{,}08$
zelená	3	$3/12 = 0{,}25$
Celkem	12	1,00

znázorníme prostřednictvím histogramu a výsečového grafu.
Modus = červená (tj. v zaznamenaném vzorku se vyskytlo nejvíce červených automobilů)



Obr. 1.1: Pozorované barvy automobilů - histogram



Obr. 1.2: Pozorované barvy automobilů - výsečový graf

Celkem bylo pozorováno 12 automobilů. ▲

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.

ad a) Zřejmě se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že velikosti triček lze seřadit, jde o proměnnou ordinální. Pro její popis proto použijeme tabulku četností pro ordinální proměnnou, v níž varianty velikosti triček budou seřazeny od nejmenší po největší (S, M, L, XL) a modus.

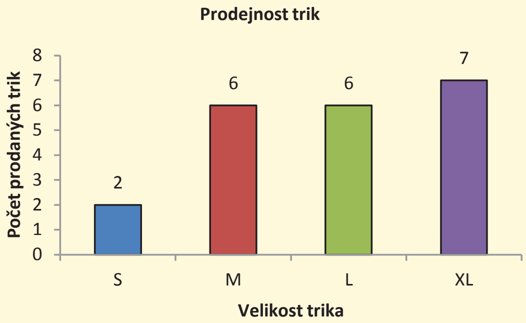
Tab. 1.2: Tabulka rozdělení četností prodejnosti triček podle velikosti

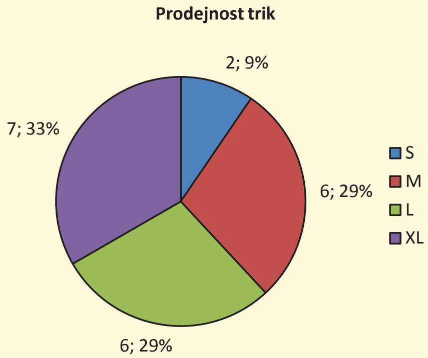
TABULKA ROZDĚLENÍ ČETNOSTÍ				
Velikosti triček	**Absolutní četnost**	**Relativní četnost**	**Kumulativní četnost**	**Kumulativní relativní četnost**
	$\mathbf{n_i}$	$\mathbf{p_i}$	$\mathbf{m_i}$	$\mathbf{F_i}$
S	3	$3/22 = 0{,}14$	3	$3/22 = 0{,}14$
M	6	$6/22 = 0{,}27$	$3+6=9$	$9/22 = 0{,}41$
L	6	$6/22 = 0{,}27$	$9+6=15$	$15/22 = 0{,}68$
XL	7	$7/22 = 0{,}32$	$15+7=22$	$22/22 = 1{,}00$
Celkem	22	1,00	-----	-----

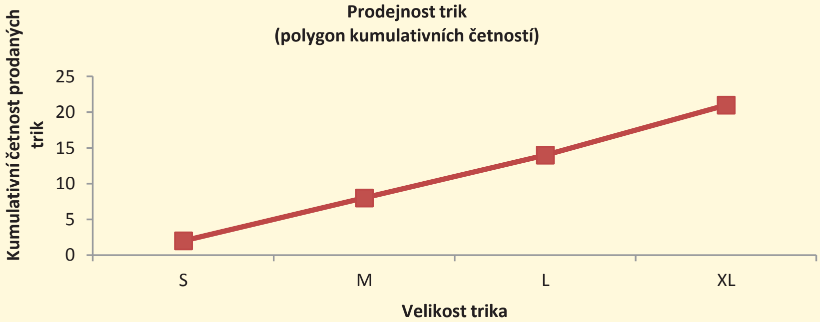
Modus = XL (nejvíce lidí si koupilo tričko velikosti XL)

Grafický výstup bude tvořit histogram, výsečový graf a Lorenzova křivka. Jelikož nechceme používat Paretův princip, Paretův graf vytvářet nebudeme.

Grafický výstup:







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

ad b) Na tuto otázku nám dá odpověď relativní kumulativní četnost pro variantu L, která určuje jaká část prodaných triček byla velikosti L a nižších. Tj. 68% zákazníků si koupilo tričko velikosti L a menší.

▲

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. Jde o klasický případ užití váženého průměru, kdy význam jednotlivých známek je oceněn jejich váhami.

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

$$\bar{x} = \frac{2 \cdot 1 + 3 \cdot 1 + 2 \cdot 1 + 1 \cdot 1 + 3 \cdot 1 + 2 \cdot 2 + 3 \cdot 2 + 1 \cdot 2 + 3 \cdot 3 + 2 \cdot 3}{1 + 1 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3} = \frac{38}{17} \doteq 2,2$$

Vzhledem k tomu, že vážený průměr známek studenta Masaříka je 2,2, měl by tento student na pololetní vysvědčení dostat z matematiky 2.

▲

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. Jde o typickou úlohu o společné práci. Pro určení průměrné doby trvání výroby součástky proto použijeme harmonický průměr.

$$\bar{x}_H = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} = \frac{2}{\frac{1}{6} + \frac{1}{3}} = 4\ [min]$$

Výroba jedné součástky trvá průměrně 4 minuty.

▲

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. **Koeficient růstu** k_t je podíl dvou hodnot kladné proměnné.

$$k_t = \frac{x_t}{x_{t-1}},$$

kde x_t ... hodnota proměnné x v aktuálním období t,
x_{t-1} ... hodnota proměnné x v předchozím období $t-1$.

Často se koeficient růstu uvádí v procentech, pak hovoříme o **relativním přírůstku** $\boldsymbol{\sigma_t}$.

$$\sigma_t = (k_t - 1) \cdot 100 = \frac{x_t - x_{t-1}}{x_{t-1}} \cdot 100 \ [\%]$$

	Plat [Kč]	Koeficient růstu	Relativní přírůstek [%]
předloni	200 000		
loni	220 000	$\frac{220\,000}{200\,000} = 1{,}100$	10,0%
letos	250 000	$\frac{250\,000}{220\,000} = 1{,}136$	13,6%

Koeficient růstu představuje relativní změnu, pro výpočet průměru proto použijeme geometrický průměr.

$$\bar{k}_t = \sqrt{1,100 \cdot 1,136} = 1,118$$

Plat zaměstnance během posledních třech let rostl průměrně o 11,8% ročně.

▲

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. **a) Určení průměru:**

V tomto případě jednoznačně použijeme aritmetický průměr (proměnná věk nepředstavuje ani část celku ani relativní změnu).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{22+82+27+43+19+47+41+34+34+42+35}{11} = 38,7 \textit{ let}$$

Průměrný věk hudebníka vystupujícího na přehlídce dechových orchestrů je 38,7 let.

Prohlédněte si ještě jednou zadaná data a promyslete si nakolik je průměrný věk reprezentativní statistikou daného výběru (pozor na odlehlá pozorování).

b) Určení shorthu:

Náš výběrový soubor má 11 hodnot, z čehož vyplývá, že v shorthu bude ležet 6 z nich (rozsah souboru je 11 (lichý počet hodnot), 50% z toho je 5,5 *(5,5 hodnoty se špatně určuje, že?)* a nejbližší vyšší přirozené číslo je 6 – neboli: $\lceil \frac{n}{2} \rceil = \lceil \frac{11}{2} \rceil = \lceil 5,5 \rceil = 6$).

A další postup?

- Hodnoty proměnné seřadíme.
- Určíme délky všech 6-ti členných intervalů, v nichž $x_1 < x_{i+1} < \ldots < x_{i+5}$ pro $i = 1, 2, \ldots, n-5$.
- Nejkratší z těchto intervalů prohlásíme za shorth (délka intervalu $= x_{i+5} - x_i$)

Originální data	Seřazená data	Délky 6-ti členných intervalů
22	19	16 (= 35 – 19)
82	22	19 (= 41 – 22)
27	27	15 (= 42 – 27)
43	***34***	***9*** (= 43 – 34)
19	***34***	13 (= 47 – 34)
47	***35***	47 (= 82 – 35)
41	***41***	
34	***42***	
34	***43***	
42	47	
35	82	

Z tabulky je zřejmé, že nejkratší interval má délku 9, čemuž odpovídá jediný interval: $\langle 34; 43\rangle$.

Shorth $= \langle 34; 43\rangle$, což můžeme interpretovat např. tak, že polovina hudebníků je ve věku 34 až 43 let (jde přitom o nejkratší interval ze všech možných).

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

c)Určení modu:

Modus je definován jako střed shortu.

$$\widehat{x} = \frac{34 + 43}{2} = 38,5 \; let$$

Modus = 38,5 let, tj. typický věk hudebníka vystupujícího na této přehlídce dechových orchestrů je 38,5 let.

▲

Příklad 1.7. Pro data z řešeného příkladu 1.7 určete

a) všechny kvartily,

b) interkvartilové rozpětí,

c) MAD,

d) zakreslete empirickou distribuční funkci.

Řešení. ad a)Naším úkolem je určit dolní kvartil $x_{0,25}$, medián $x_{0,5}$ a horní kvartil $x_{0,75}$. Budeme dodržovat postup doporučený pro určování kvantilů, to znamená – data seřadit a přiřadit jim pořadí. Výsledek prvních dvou bodů postupu ukazuje Tab. 1.3.

Tab. 1.3: Přiřazení pořadí hodnotám proměnné

Originální data	**Seřazená data**	**Pořadí**
22	19	1
82	22	2
27	**27**	**3**
43	**34**	**4**
19	34	5
47	**35**	**6**
41	41	7
34	**42**	**8**
34	**43**	**9**
42	47	10
35	82	11

A můžeme přejít k bodu 3, tj. stanovit pořadí hodnot proměnné pro jednotlivé kvartily a tím i jejich hodnoty.

Dolní kvartil $x_{0,25}$: $p = 0,25; n = 11 \Rightarrow z_p = 11 \cdot 0,25 + 0,5 = 3,25$,
Dolní kvartil je tedy průměrem prvků s pořadím 3 a 4. $x_{0,25} = \dfrac{27 + 34}{2} = 30,5$ let,
tj. 25% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 30,5 let (75% z nich má 30,5 let a více).

Medián $x_{0,5}$: $p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow x_{0,5} = 35\ let$,
tj. polovina hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 35 let (50% z nich má 35 let a více).

Horní kvartil $x_{0,75}$: $p = 0,75; n = 11 \Rightarrow z_p = 11 \cdot 0,75 + 0,5 = 8,75$
Horní kvartil je tedy průměrem prvků s pořadím 8 a 9.$x_{0,75} = \dfrac{42 + 43}{2} = 42,5$ let, tj. 75% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 42,5 let (25% z nich má 42,5 let a více).

ad b) **Interkvartilové rozpětí IQR:** $\text{IQR} = x_{0,75} - x_{0,25} = 43 - 27 = 16$.

Jak již bylo zmíněno, praktická interpretace IQR neexistuje.

ad c) **MAD** Chceme-li určit tuto statistiku, budeme postupovat přesně podle toho, co

skrývá zkratka v názvu – medián absolutních odchylek od mediánu. Provedení uvedeného postupu ukazuje Tab 1.4.

Tab. 1.4: Postup při výpočtu statistiky MAD

Originální data x_i	**Seřazená data y_i**	Absolutní hodnoty odchylek seřazených dat od jejich mediánu $\lvert y_i - x_{0,5} \rvert$	Seřazené absolutní hodnoty odchylek seřazených dat od jejich mediánu **M_i**
22	19	$16 = \lvert 19 - 35 \rvert$	0
82	22	$13 = \lvert 22 - 35 \rvert$	1
27	27	$8 = \lvert 27 - 35 \rvert$	1
43	34	$1 = \lvert 34 - 35 \rvert$	6
19	34	$1 = \lvert 34 - 35 \rvert$	7
47	35	$0 = \lvert 35 - 35 \rvert$	**8**
41	41	$6 = \lvert 41 - 35 \rvert$	8
34	42	$7 = \lvert 42 - 35 \rvert$	12
34	43	$8 = \lvert 43 - 35 \rvert$	13
42	47	$12 = \lvert 47 - 35 \rvert$	16
35	82	$47 = \lvert 82 - 35 \rvert$	47

$x_{0,5}$=35

$MAD = M_{0,5}$,

$p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow M_{0,5} = 8$, (MAD je medián absolutních odchylek od mediánu, tj. 6. hodnota seřazeného souboru absolutních odchylek od mediánu).MAD = 8.

ad d) Zbývá poslední úkol – sestrojit **empirickou distribuční funkci**. Připomeňme si proto její definici a postupujme podle ní.

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

Do tabulky si zapíšeme seřazené hodnoty proměnné, jejich četnosti, relativní četnosti a z nich odvodíme empirickou distribuční funkci.

Z definice emp. dist. funkce *F(x)* tedy plyne, že pro všechna x menší než 19 je *F(x)* rovna nule, pro x větší než 19 a menší nebo rovna 22 je *F(x)* rovna 1/11, pro x větší než 22 a menší nebo rovna 27 je *F(x)* rovna 1/11 + 1/11, atd. Pro $x > 82$ je *F(x)*=1. Shrneme do Tab. 1.6.

Obsah

23. strana ze 159

Zavřít dokument

Celá obrazovka/Okno

Tab. 1.5: Postup výpočtu empirické distribuční funkce

Originální data x_i
22
82
27
43
19
47
41
34
34
42
35

Seřazené hodnoty x_i	Absolutní četnosti seřazených hodnot n_i	Relativní četnosti seřazených hodnot p_i	Empirická dist. funkce $F(x_i)$
19	1	1/11	0
22	1	1/11	1/11
27	1	1/11	2/11
34	2	2/11	3/11
35	1	1/11	5/11
41	1	1/11	6/11
42	1	1/11	7/11
43	1	1/11	8/11
47	1	1/11	9/11
82	1	1/11	10/11

Tab. 1.6: Empirická distribuční funkce

x	$(-\infty;19\rangle$	$(19;22\rangle$	$(22;27\rangle$	$(27;34\rangle$	$(34;35\rangle$
F(x)	0	1/11	2/11	3/11	5/11

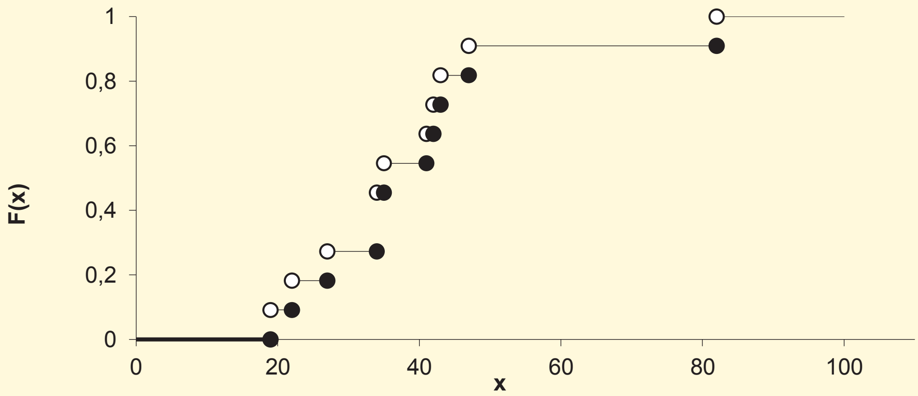
x	$(35;41\rangle$	$(41;42\rangle$	$(42;43\rangle$	$(43;47\rangle$	$(47;82\rangle$	$(82;\infty)$
F(x)	6/11	7/11	8/11	9/11	10/11	11/11

Obsah

24. strana ze 159

Zavřít dokument

Celá obrazovka/Okno



Obr. 1.3: Empirická distribuční funcke-graf

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 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).

Tab. 1.7: 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

Řešení. Nejprve se pokusíme porovnat obě technologie pouze za pomocí průměru. Vzhledem k tomu, že proměnná „mezní teplota“ nevyjadřuje ani část celku ani relativní změny, volíme průměr aritmetický.

Průměr pro starou technologii vychází

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{475 + 436 + \ldots + 426}{5} \doteq 463\,[^oC]$$

Průměr pro novou technologii:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{485 + 390 + \ldots + 488}{5} \doteq 469\,[^oC]$$

Na základě vypočtených průměrů bychom mohli říci, že novou technologii doporučujeme, poněvadž mezní teplota je při nové technologii o 6^oC vyšší.

A jaký závěr vyvodíme, doplníme-li k základním informacím míry variability?

Stará technologie:

Výběrový rozptyl:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1} = \frac{(475-463)^2 + (436-463)^2 + \ldots + (426-463)^2}{5-1} \doteq 916\,[^oC^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}} = \sqrt{\frac{(475-463)^2 + \ldots + (426-463)^2}{5-1}} \doteq 31\,[^oC].$$

Nová technologie:

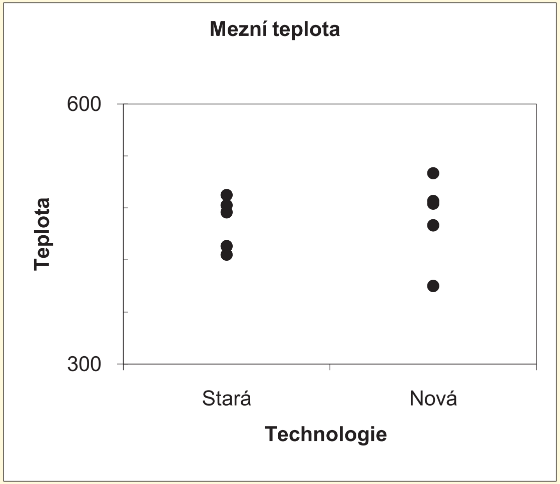
Výběrový rozptyl:

$$s_y^2 = \frac{\sum\limits_{i=1}^{n}(y_i - y)^2}{n-1} = \frac{(485-469)^2 + (390-469)^2 + \ldots + (488-469)^2}{5-1} \doteq 2384\,[^oC^2]$$

Výběrová směrodatná odchylka:

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - y)^2}{n-1}} == \sqrt{\frac{(485-469)^2 + \ldots + (488-469)^2}{5-1}} \doteq 49\,[^oC].$$

Výběrový rozptyl (výběrová směrodatná odchylka) vyšel pro novou technologii mnohem vyšší než pro technologii starou. Co to znamená? Podívejte se na grafické znázornění naměřených dat na Obr. 1.4.



Obr. 1.4: Srovnání technologií teplot pro starou a novou technologii

Mezní teploty pro novou technologii jsou mnohem rozptýlenější, tzn. že tato technologie není ještě dobře zvládnutá a její použití nám nezaručí zkvalitnění výroby. V tomto případě může dojít k silnému zvýšení, ale také k silnému snížení mezní teploty – proto by se měla nová technologie ještě vrátit do vývoje.

Zdůrazněme, že tyto závěry jsou stanoveny pouze na základě explorační analýzy. Pro rozhodnutí takovýchto případů nám statistika nabízí exaktnější metody (testování hypotéz), s nimiž se seznámíte později.

▲

Kapitola 2

Statistické šetření - řešené příklady

Kapitola 3

Výběrové charakteristiky - řešené příklady

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.

X_i... životnost $i-$tého holícího strojku EHS

$$X_i \to Exp\left(\frac{1}{2}\right) \Rightarrow E(X_i) = \mu_X = \frac{1}{\lambda} = \ 2 \text{ roky } \Rightarrow \lambda = \frac{1}{2}\text{rok}^{-1} \Rightarrow D(X_i) = \sigma_X^2 = \frac{1}{\lambda^2} =$$

$$= \ 4 \text{ rok}^2$$

$\bar{X}$... průměrná životnost 150-ti strojků EHS

$$\bar{X} = \frac{\sum_{i=1}^{150} X_i}{150} = \frac{1}{150}\sum_{i=1}^{150} X_i$$

Neboť testovaný vzorek holících strojků byl dostatečně velký (150 strojků), byly splněny předpoklady CLV a tudíž platí, že $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.

V našem případě: $\bar{X} \sim N\left(2; \dfrac{4}{150}\right)$

Nyní, když známe rozdělení průměrné životnosti 150 holicích strojků EHS, můžeme řešení dokončit (27 měsíců = 2,25 roků):

$$P\left(\bar{X} > 2,25\right) = 1 - F(2,25) = 1 - \Phi\left(\frac{2,25-2}{\sqrt{\frac{4}{150}}}\right) = 1 - \Phi(1,53) \doteq 1 - 0,937 = 0,063$$

Pravděpodobnost, že průměrná životnost 150 prodaných holicích strojků EHS bude vyšší než 27 měsíců je 0,063.

▲

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.

X_i... doba potřebná k objevení a odstranění $i-$té poruchy

Víme, že $E(X_i) = \mu_X = 40\,\text{minut}$ a $D(X_i) = \sigma_X^2 = 30^2\,\text{minut}^2$, přičemž rozdělení náhodné veličiny X_i neznáme.

Nechť náhodná veličina X modeluje celkovou dobu do objevení sté poruchy.

$$X = \sum_{i=1}^{100} X_i$$

Na základě CLV víme, že součet n náhodných veličin se stejným rozdělením (nemusíme vědět jakým), stejnými středními hodnotami a stejnými rozptyly můžeme aproximovat normálním rozdělením s parametry $n\mu_X$ a $n\sigma_X^2$. (Vzhledem k tomu, že $n > 30$, předpokládáme předpoklady CLV za splněné.)

$$X = \sum_{i=1}^{100} X_i \sim N\left(100 \cdot 40, 100 \cdot 30^2\right)$$

Nyní již není problém určit hledanou pravděpodobnost (nesmíme jen zapomenout na užívání stejných jednotek, v našem případě minut ($70\text{ h} = 4200\text{ minut}$).

$$P\left(X < 4200\right) = F(4200) = \Phi\left(\frac{4200 - 4000}{\sqrt{90000}}\right) = \Phi(0,67) \doteq 0,749$$

Pravděpodobnost, že doba potřebná k objevení a opravení 100 nezávislých poruch nepřekročí 70 hodin, je 0,749.

▲

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.

Nechť X_i je náhodná veličina popisující hmotnost jednotlivých cestujících,
kde $E(X_i) = \mu_X = 70\,\mathrm{kg}$ a $D(X_i) = \sigma_X^2 = 20^2\,\mathrm{kg}^2 = 400\,\mathrm{kg}^2$.

Označme X náhodnou veličinu modelující celkovou hmotnost všech cestujících. Na základě CLV (předpoklady CLV považujeme za splněné $(n > 30)$) lze tvrdit, že

$$X = \sum_{i=1}^{n} X_i \sim N\left(n \cdot 70, n \cdot 400\right).$$

Člun má nosnost 5000 kg. Pravděpodobnost jeho přetížení má být menší než 0,001, což zapíšeme

$$P(X > 5000) < 0,001.$$

Po dosazení:

$$
\begin{aligned}
1 - F(5000) &< 0,001 \\
1 - \Phi\left(\frac{5000 - 70n}{\sqrt{400n}}\right) &< 0,001 \\
0,999 &< \Phi\left(\frac{5000 - 70n}{\sqrt{400n}}\right) \\
60\sqrt{n} &< \frac{5000 - 70n}{\sqrt{400n}} \\
3600n &< 4900n^2 - 700000n + 25000000 \\
0 &< 49n^2 - 7036n + 250000
\end{aligned}
$$

Řešení kvadratické nerovnice je $n \in \mathbb{N} : (n < 64,5) \cup (n > 79)$.

Je tedy zřejmé, že člunem může cestovat maximálně 64 osob.

▲

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.

Jak již víte, výběrová směrodatná odchylka S je náhodná veličina. Je zřejmé, že nedošlo-li k žádné změně při výrobě žárovek Ed, tj. střední životnost těchto žárovek μ je stále 5 let a směrodatná odchylka životnosti μ je 6 měsíců, pak výběrová směrodatná odchylka S se bude pohybovat **„kolem“** 6 měsíců.

Víme, že bude testováno 20 žárovek Ed a máme zjistit, jaká je pravděpodobnost, že bude zjištěna výběrová směrodatná odchylka životnosti S vyšší než 7 měsíců.

$$P(S > 7) =?$$

Protože neznáme rozdělení náhodné veličiny S, využijeme znalosti rozdělení náhodné veličiny $\frac{(n-1)S^2}{\sigma^2}$.

Předpokládejme, že **životnost žárovek Ed podléhá normálnímu rozdělení**. (Ověření toho, zda testovaný vzorek je výběrem z normálního rozdělení se naučíte provádět v kapitole 14)

Z vlastností χ^2- rozdělení víte, že $\frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}$.

Zavedeme-li substituci $X = \frac{(n-1)S^2}{\sigma^2}$, kde $n = 20$ (počet testovaných žárovek) a $\sigma = 6$ [měsíc], tj. $X = \frac{(20-1)S^2}{6^2} = \frac{19S^2}{36}$, pak náhodná veličina X má χ^2- rozdělení s 19 stupni volnosti, což značíme

$$X \to \chi^2_{19}.$$

Je-li $\frac{19S^2}{36}$, pak je zřejmé, že $(S > 7) \Leftrightarrow \left(X > \frac{19\cdot 7^2}{36}\right)$, tj. $(X > 25,86)$.

Této ekvivalence využijeme při určení hledané pravděpodobnosti.

$$P(S > 7) = P(X > 25,86) = 1 - F_{\chi^2_{19}}(25,86) = 0,134,$$

kde $F_{\chi^2_\nu}(x)$ značíme distribuční funkci náhodné veličiny s χ^2- rozdělením s ν stupni volnosti. (*Pro určení* $F_{\chi^2_{19}}(25,86)$ *lze použít statistický software, MS Excel, tabulky...*).

Pravděpodobnost, že při testu 20 žárovek bude zjištěna směrodatná odchylka životnosti větší než 7 měsíců je přibližně 0,134.

▲

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

Řešení.

Z definice χ^2-rozdělení je zřejmé, že náhodná veličina X, která má χ^2-rozdělení s jedním stupněm volnosti je rovna kvadrátu náhodné veličiny Z, která má normované normální rozdělení.

$$X = Z^2$$

$$Z \to N(0;1) \Rightarrow X \to \chi^2_1$$

Náhodná veličina X je funkcí náhodné veličiny Z a proto budeme při hledání její distribuční funkce dále postupovat již známým způsobem (pouze vezmeme v úvahu, že náhodná veličina s rozdělením χ^2 nabývá pouze nezáporných hodnot).

pro $x > 0$:

$$\begin{aligned} F(x) &= P(X < x) = P\left(Z^2 < x\right) = P\left(-\sqrt{x} < Z < \sqrt{x}\right) = \Phi\left(\sqrt{x}\right) - \Phi\left(-\sqrt{x}\right) = \\ &= \Phi\left(\sqrt{x}\right) - \left[1 - \Phi\left(\sqrt{x}\right)\right] = 2\Phi\left(\sqrt{x}\right) - 1 = \frac{2}{\sqrt{2\pi}} \int\limits_0^{\sqrt{x}} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t - 1 = \\ &= \sqrt{\frac{2}{\pi}} \cdot \int\limits_0^{\sqrt{x}} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t - 1 \end{aligned}$$

pro $x \leqq 0$:

$$F(x) = 0$$

Hustotu pravděpodobnosti pak určíme jednoduše jako derivaci distribuční funkce.

pro $x > 0$:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = 2 \cdot \frac{1}{2\sqrt{x}} \cdot \varphi\left(\sqrt{x}\right) = \frac{1}{\sqrt{x}} \cdot \varphi\left(\sqrt{x}\right) = \frac{1}{\sqrt{2\pi x}} \mathrm{e}^{-\frac{x}{2}}$$

pro $x \leqq 0$:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = 0$$

Hustota pravděpodobnosti náhodné veličiny X je tedy

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}} \mathrm{e}^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leqq 0. \end{cases}$$

▲

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.

Označme S_1^2 rozptyl životnosti zjištěný u vzorku z linky 1 a S_2^2 rozptyl životnosti zjištěný u vzorku z linky 2.

Hledáme pravděpodobnost, že $S_1^2 > 2S_2^2$, tj. pravděpodobnost, že $\frac{S_1^2}{S_2^2} > 2$.

$$P\left(S_1^2 > 2S_2^2\right) = P\left(\frac{S_1^2}{S_2^2} > 2\right) = ?$$

Za předpokladu, že **oba vzorky jsou výběrem z normálního rozdělení** (ověřovat tento předpoklad se naučíte v kapitole 14), platí

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

Dle zadání předpokládáme, že rozptyl životnosti žárovek vyrobených na jednotlivých linkách je stejný, tj.

$$\sigma_1^2 = \sigma_2^2.$$

Pak

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

V našem případě bude testováno 20 žárovek z linky 1 ($n_1 = 20$) a 30 žárovek z linky 2 ($n_2 = 30$), proto

$$\frac{S_1^2}{S_2^2} \to F_{19,29}.$$

$$P\left(\frac{S_1^2}{S_2^2} > 2\right) = 1 - F_{F_{19,29}}(2) \doteq 0,045,$$

kde $F_{F_{m,n}}(x)$ označuje distribuční funkci náhodné veličiny s Fisher–Snedecorovým rozdělením s n stupni volnosti pro čitatele a m stupni volnosti pro jmenovatele. (Hodnotu distribuční funkce tohoto rozdělení lze určit pomocí statistického software, pomocí MS Excel nebo lze pro určení přibližné hodnoty této funkce použít příslušné tabulky.)

Pravděpodobnost, že u vzorku z linky 1 bude zjištěn více než dvojnásobný rozptyl oproti rozptylu zjištěnému u vzorku z linky 2 je přibližně 0,045.

▲

Kapitola 4

Úvod do teorie odhadu - řešené příklady

Příklad 4.1. Mějme náhodný výběr $(X_1, X_2, \ldots, X_n)$ 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.

ada)

Nejprve odvodíme vztah $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$, který využijeme při důkazu nestrannosti odhadu.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}\left((X_i - \bar{X}) + (\bar{X} - \mu)\right)^2 \\
&= \sum_{i=1}^{n}\left((X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)\right)^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) + \sum_{i=1}^{n}(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 0 + n(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2
\end{aligned}$$

Dále si připomeňme, že rozptyl populace o rozsahu N je dán vztahem $\sigma^2 = D(X) = E\left((X - \mu)^2\right)$ a rozptyl výběrového průměru lze určit dle vztahu $D(\bar{X}) = E\left((\bar{X} - E(\bar{X}))^2\right) = E\left((\bar{X} - \mu)^2\right)$.

Důkaz:

Odhad je nestranný právě když

$$E(S^2) = \sigma^2.$$

$$
\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right) = \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) - \frac{n}{n-1}E\left((\bar{X}-\mu)^2\right) = \\
&= \frac{1}{n-1}\sum_{i=1}^{n}E\left((X_i-\mu)^2\right) - \frac{n}{n-1}E\left((\bar{X}-\mu)^2\right) = \\
&= \frac{n}{n-1}D(X) - \frac{n}{n-1}D(\bar{X}) = \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} = \frac{n-1}{n-1}\sigma^2 = \sigma^2
\end{aligned}
$$

Výběrový rozptyl S^2 je proto nestranným odhadem rozptylu σ^2.

Poznámka: *Mimochodem, právě jsme ukázali, proč není výběrový rozptyl definován jako* $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$. *(Takto definovaný výběrový rozptyl by nebyl nestranným odhadem rozptylu.)*

adb)
Odhad S^2 je konzistentní, pokud se s rostoucím rozsahem výběru zpřesňuje, k čemuž dochází pokud

- $\lim\limits_{n\to\infty} E(S^2) = \sigma^2$,
- $\lim\limits_{n\to\infty} D(S^2) = 0$,

Důkaz:

Pro první část důkazu využijeme nestrannosti odhadu S^2 odvozené v bodě a) této úlohy.

$$\lim_{n \Rightarrow \infty} E(S^2) = \lim_{n \to \infty} \sigma^2 = \sigma^2$$

Pro druhou část důkazu využijeme znalosti vlastností rozdělení χ^2(kap. 8.8.1).

$$\text{Je-li } X = \frac{(n-1)s^2}{\sigma^2}, \text{ pak } X \to \chi^2_{n-1} \text{ a } D(X) = 2(n-1).$$

$$X = \frac{(n-1)s^2}{\sigma^2} \Rightarrow S^2 = \frac{\sigma^2}{n-1} X, \text{ pak } D(S^2) = \left(\frac{\sigma^2}{n-1}\right)^2 D(X) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot 2(n-1) = \\ = \frac{2\sigma^4}{n-1}$$

$$\lim_{n \to \infty} D(S^2) = \lim_{n \to \infty} \frac{2\sigma^4}{n-1} = 0$$

Tímto jsme dokázali, že $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ je nestranným konzistentním odhadem rozptylu σ^2.

Zájemci se mohou pokusit dokázat, že odhad $S_*^2 = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X})^2$ je nejen vychýlený, ale že taktéž $D(S_*^2) > D(S^2)$.

▲

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.)

Řešení.

Chceme najít 95% intervalový odhad střední hodnoty životnosti žárovek firmy Edison, přičemž neznáme směrodatnou odchylku životnosti těchto žárovek. Máme k dispozici informace pocházející z výběru o rozsahu 50 žárovek, tj. rozsah výběru je vyšší než 30. Životnost žárovek lze modelovat normálním rozdělením. Jde tedy o intervalový odhad střední hodnoty normálního rozdělení pro známé σ, kde směrodatnou odchylku životnosti σ odhadneme výběrovou směrodatnou odchylkou s.

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

spolehlivost intervalového odhadu $\quad 1-\alpha = 0,95$

$\Rightarrow$ hladina významnosti $\quad \alpha = 1 - 0,95 = 0,05$

$\Rightarrow$ $\frac{\alpha}{2} = 0,025;\ 1-\frac{\alpha}{2} = 0,975$

$\Rightarrow$ $z_{0,975} = 1,96$ (viz Tabulka 1)

Výběrový soubor: $\quad \bar{x} = 950$ hodin

$s = 100$ hodin

$$n = 50$$

$n \geqq 30 \Rightarrow \sigma \doteq s$

Zjištěné hodnoty dosadíme do předpisu pro meze oboustranného intervalového odhadu střední hodnoty se spolehlivostí 0,95.

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

$$\mu \in \left\langle 950 - \frac{100}{\sqrt{50}} \cdot 1,96; 950 + \frac{100}{\sqrt{50}} \cdot 1,96 \right\rangle \text{ hodin}$$

$$\mu \in \langle 922,3; 977,7 \rangle \text{ hodin}$$

Střední životnost žárovek firmy Edison se se spolehlivostí 0,95 pohybuje v rozmezí 922 hodin 18 minut až 977 hodin 42 minut.

▲

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 4.1.

Tab. 4.1: 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%).

Řešení.

ada) Pro získání výběru o rozsahu minimálně 30 hodnot bychom museli danou prodejnu sledovat minimálně 30 pátku (tj. déle než půl roku), což by vedlo jak k zvýšení finanční náročnosti studie, tak k vysoké časové náročnosti průzkumu. Z těchto důvodu byl zvolen menší rozsah výběru ($n = 5$) odpovídající měsíčnímu sledování prodejny. Nevýhodou malého rozsahu výběru je nízká přesnost odhadu (poměrně široký intervalový odhad).

adb) Určujeme intervalový odhad střední hodnoty s neznámou směrodatnou odchylkou a malým rozsahem výběru, proto pro jeho výpočet použijeme předpis

$$\left\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}; \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \right\rangle$$

spolehlivost intervalového odhadu $1 - \alpha = 0,95$
hladina významnosti $\alpha = 1 - 0,95 = 0,05$
$\frac{\alpha}{2} = 0,025$; $1 - \frac{\alpha}{2} = 0,975$
$t_{0,975} = 2,78$ (viz Tabulka 2, máme 4(=5-1) stupně volnosti)

Výběrový soubor:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{3756 + 2987 + 3042 + 4206 + 3597}{5} = 3517,6$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(3756 - 3517,6)^2 + \ldots + (3597 - 3517,6)}{4} = 261191,3 \Rightarrow$$

$\Rightarrow s = 511,1$

$n = 5$

Zjištěné hodnoty dosadíme do předpisu pro meze intervalového odhadu střední hodnoty se spolehlivostí 0,95.

$$\mu \in \left\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}; \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \right\rangle$$

$$\mu \in \left\langle 3517,6 - \frac{511,1}{\sqrt{5}} \cdot 2,78; 3517,6 + \frac{511,1}{\sqrt{5}} \cdot 2,78 \right\rangle$$

$$\mu \in \langle 2882,2; 4153,0 \rangle$$

Se spolehlivostí 0,95 se střední návštěvnost TETO Poruba v pátek v odpoledních hodinách bude pohybovat v rozmezí 2882 až 4153 zákazníků.

▲

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.)

Řešení.

Vzhledem k tomu, že naším úkolem je určit levostranné intervalové odhady rozptylu a směrodatné odchylky normálního rozdělení, využijeme vztahy uvedené v kapitolách **??** a **??**.

Levostranný intervalový odhad rozptylu normálního rozdělení je $\dfrac{(n-1)s^2}{x_{1-\alpha}}$.

Spolehlivost intervalového odhadu: $1-\alpha = 0,95 \Rightarrow \quad x_{0,95} \doteq 100,7$ (Tabulka 3, počet stupňů volnosti je $n-1$, tj. 79)

Výběrový soubor: $s^2 = (0,04)^2\, mm^2 = 0,0016\, mm^2$
$n = 80$

Po dosazení:

$$\frac{(80-1)0,0016}{100,7} \doteq 0,0013$$

S 95% spolehlivostí je rozptyl průměru pístových kroužků větší než 0,0013 mm^2.

Jednoduchou úpravou pak získáme 95% levostranný intervalový odhad směrodatné odchylky normálního rozdělení.

$$\sqrt{0,0013} \doteq 0,035$$

S 95% spolehlivostí tedy můžeme tvrdit, že směrodatná odchylka průměru pístových kroužků je větší než 0,035 mm.

▲

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

Řešení.

Výběrový soubor $n = 320,$

$$p = \frac{59}{320} \doteq 0,018,$$

$$\frac{9}{p(1-p)} \doteq 60,$$

$$\frac{n}{N} = \frac{320}{20000} = 0,016.$$

Rozsah výběru je dostatečně velký ($n > 30, n > \frac{9}{p(1-p)}$) a nepřevyšuje 5% rozsahu populace ($\frac{n}{N} < 0,05$). Intervalový odhad podílu (relativní četnosti) konzerv s prošlou záruční lhůtou lze tedy stanovit jako

$$\left\langle p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right\rangle$$

Spolehlivost intervalového odhadu: $1-\alpha = 0,95$

$\Rightarrow$ Hladina významnosti: $\alpha = 1 - 0,95 = 0,05$

$\Rightarrow$ $\frac{\alpha}{2} = 0,025;\ 1 - \frac{\alpha}{2} = 0,975$

$\Rightarrow \qquad z_{0,975} = 1,96 \quad$ (viz Tabulka 1)

Po dosazení:

$$\left\langle 0,018 - 1,96\sqrt{\frac{0,018(1-0,018)}{320}}; 0,018 + 1,96\sqrt{\frac{0,018(1-0,018)}{320}} \right\rangle$$

$\langle 0,138; 0,222 \rangle$

S 95% spolehlivostí můžeme tvrdit, že mezi masovými konzervami se v daném skladu nachází mezi 13,8% a 22,2% konzerv s prošlou záruční lhůtou.

▲

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?

Řešení.

Chceme odhadnout rozsah výběru pro intervalový odhad střední hodnoty, známe-li směrodatnou odchylku σ (vyčerpávající šetření = zkoumání celého základního souboru (populace)).

Dle tabulky **??** je doporučený rozsah výběru

$$n \geqq \left(\frac{\sigma}{\Delta_{max}} z_{1-\frac{\alpha}{2}} \right)^2 .$$

Ze zadání víme, že
σ =750 Kč
Δ_{max} =50Kč

$1 - \alpha = 0,95 \Rightarrow \alpha = 0,05 \Rightarrow 1 - \frac{\alpha}{2} = 0,975,\ z_{0,975} = 1,96$ (viz Tabulka 1)

Rozsah výběru proto odhadneme jako

$$n \geqq \left(\frac{750}{50} \cdot 1,96 \right)^2, \text{ tj. } n \geqq 864,4.$$

Chceme-li dosáhnout přípustné chyby ve výši maximálně 50,- Kč, musíme pro nalezení intervalového odhadu průměrného platu se spolehlivostí 95% provést výběrové šetření na výběrovém souboru o rozsahu minimálně 865 pracovníků.

▲

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.

Řešení.

Uvědomte si, že ze zadání příkladu jste získali informace o podílech vadných disket v náhodných výběrech z celkové produkce firem DISK a EMEM. Vaším úkolem je odhadnout, jak se liší podíl vadných disket v celkové produkci těchto dvou výrobců.

Označme si procento vadných disket v produkci firmy DISK π_D a procento vadných disket v produkci firmy EMEM π_E.

Z výběrového šetření víme, že bylo testováno 800 ($= 40 \cdot 20$) disket firmy DISK, přičemž 24 z nich bylo vadných.

$$\left.\begin{array}{rcl} x_D & = & 24 \\ n_D & = & 800 \end{array}\right\} \Rightarrow p_D = \frac{24}{800} = 0,030,$$

tzn., že mezi testovanými disketami firmy DISK bylo 3,0% vadných disket.

Obdobně lze ukázat, že mezi 600 ($= 30 \cdot 20$) testovanými disketami firmy EMEM bylo 14, tj. 2,3% vadných:

$$\left.\begin{array}{rcl} x_E & = & 14 \\ n_E & = & 600 \end{array}\right\} \Rightarrow p_E = \frac{14}{600} = 0,023.$$

Víme, že v testovaných výběrech se ukázaly kvalitnější diskety EMEM. (Testovaný vzorek disket EMEM obsahoval o 0,7% ($= 3,0\% - 2,3\%$) méně vadných disket než vzorek disket DISK.) Pokud byly výběry provedeny skutečně náhodně, je zřejmé, že se v celkové produkci firem DISK a EMEM bude rozdíl mezi podílem vadných disket pohybovat „kolem" 0,7%. V jakém rozmezí lze rozdíl mezi podílem vadných disket obou firem očekávat nám ukáže intervalový odhad.

- Oba výběry mají rozsah větší než 30,
- lze předpokládat, že rozsahy jednotlivých výběrů nepřekročily 5% celkové produkce firem,
- $\frac{9}{p_D(1-p_D)} \doteq 309 \Rightarrow n_D > \frac{9}{p_D(1-p_D)}$, $\frac{9}{p_E(1-p_E)} \doteq 395 \Rightarrow n_E > \frac{9}{p_E(1-p_E)}$,

proto lze se spolehlivostí $1-\alpha$ stanovit oboustranný intervalový odhad rozdílu relativních četností stanovit jako

$$\left\langle (p_D - p_E) - z_{1-\frac{\alpha}{2}} \sqrt{p(1-p)\left(\frac{1}{n_D} + \frac{1}{n_E}\right)}; (p_D - p_E) + z_{1-\frac{\alpha}{2}} \sqrt{p(1-p)\left(\frac{1}{n_D} + \frac{1}{n_E}\right)} \right\rangle.$$

Zvolíme-li $1-\alpha = 0,95$, pak $1-\frac{\alpha}{2} = 0,975$. Za pomocí Tabulky 1 nebo statistického softwaru určíme příslušný kvantil normovaného normálního rozdělení: $z_{0,975} = 1,96$.

Dále určíme $p = \frac{x_D + x_E}{n_D + n_E} = \frac{24+14}{800+600} = \frac{38}{1400} = 0,027$.

Po dosazení zjistíme, že se spolehlivostí 95% se rozdíl podílu vadných disket DISK a EMEM ($\pi_D - \pi_E$) nachází v intervalu

$$\langle 0,007 - 0,017; 0,007 + 0,017 \rangle\,,$$

$$\langle -0,010; 0,024 \rangle\,,\ \text{tj.}\ \langle -1,0\%; 2,4\% \rangle\,.$$

Jakou informaci jsme získali? Pokud by diskety firem DISK a EMEM byly stejně kvalitní, pak by podíly vadných disket v jejích produkcích byly stejné, neboli rozdíl v podílech vadných disket v jednotlivých produkcích by byl 0.

$$\pi_D = \pi_E,\ \text{tj.}\ \pi_D - \pi_E = 0.$$

Ukázali jsme, že intervalový odhad rozdílu podílu vadných disket obsahuje 0.

$$0 \in \langle -0,010; 0,024 \rangle$$

Se spolehlivostí 95% lze tedy tvrdit, že diskety obou výrobců jsou stejně kvalitní. Zamyslete se nad tím, jak by musel vypadat nalezený intervalový odhad, abychom mohli tvrdit, že diskety firmy 5M jsou kvalitnější. Ale to už jsme se dostali k testování hypotéz, jimž se budeme zabývat v kapitole 10.

▲

Kapitola 5

Testování hypotéz, princip - řešené příklady

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. Ze zadání úlohy je zřejmé, že máme rozhodovat o střední hodnotě výšky rostliny, přičemž směrodatnou odchylku výšky lze považovat za známou.

ada)

V této části úlohy máme zadánu spolehlivost testu $1-\alpha = 0,95$ a tím i pravděpodobnost chyby I. druhu $\alpha = 0,05$. Pokud by byly nové pěstební podmínky účinné, mělo by dojít ke zvýšení průměrné výšky lilií μ. Nulovou a alternativní hypotézu proto stanovíme ve tvaru

$$\begin{aligned} H_0: &\quad \mu = 100, \\ H_A: &\quad \mu > 100. \end{aligned}$$

V dalším kroku bychom měli najít vhodné testové kritérium $T(X)$, tzn. výběrovou charakteristiku, která má vztah k nulové hypotéze a jejíž rozdělení za předpokladu platnosti nulové hypotézy známe.

V tomto případě lze jako testové kritérium zvolit průměrnou výšku 100 lilií $\overline{X}_{100}$, která má, dle centrální limitní věty, za předpokladu platnosti nulové hypotézy H_0, normální rozdělení se střední hodnotou $\mu = 100$ cm a rozptylem $\frac{\sigma^2}{n} = \frac{144}{100} = 1,44\ [cm^2]$.

$$\begin{aligned} T(X) &= \overline{X}_{100} \\ \overline{X}_{100} &\to N(100; 1,44) \end{aligned}$$

Podle tvaru alternativní hypotézy je zřejmé, že v neprospěch nulové hypotézy budou vypovídat vysoké hodnoty průměrné výšky zkoumaného vzorku lilií. Kritickou hodnotu $\overline{X}_{krit}$ průměrné výšky určíme z podmínky uvedené v zadání. Pravděpodobnost, že průměrná výška zkoumaného vzorku překročí kritickou hodnotu $\overline{X}_{krit}$, tj. pravděpodobnost chyby I. druhu, má být 0,05.

$$P\left(\overline{X}_{100} > \overline{X}_{krit}\right) = 0,05$$

Označme$F_{\overline{X}}(x)$ distribuční funkci náhodné veličiny $\overline{X}_{100}$ za předpokladu platnosti H_0. Pak

$$1 - F_{\overline{X}}\left(\overline{X}_{krit}\right) = 0,05.$$

Postupnými úpravami určíme $\overline{X}_{krit}$.

$$F_{\overline{X}}\left(\overline{X}_{krit}\right) = 0,95$$
$$\Phi\left(\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}}\right) = 0,95$$
$$\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}} = z_{0,95}$$
$$\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}} = 1,645\,(\text{viz Tabulka1})$$
$$\overline{X}_{krit} \cong 102,0\,\text{cm},\ \text{tj.}\ W > 102,0\,\text{cm}$$

Kritický obor W je pro tento test vymezen hodnotami průměrné výšky $\overline{X}_{100}$ vyššími než 102,0 cm. Tzn., bude-li průměrná výška 100 rostlin vyšší než 102,0 cm, můžeme na hladině významnosti 0,05 zamítnout nulovou hypotézu ve prospěch alternativy a tvrdit, že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

adb)

Klasický test provádíme tak, že ověříme, zda příslušná výběrová charakteristika, resp. pozorovaná hodnota vhodného testového kritéria, leží v kritické oblasti W, resp. v kritické oblasti testového kritéria W^*, určeného pro příslušnou spolehlivost testu.

Nulová a alternativní hypotéza byly stanoveny ve tvaru

$$\begin{aligned} H_0: &\quad \mu = 100, \\ H_A: &\quad \mu > 100. \end{aligned}$$

Pro spolehlivost testu 0,95 (hladinu významnosti 0,05) byl v otázce a) stanoven kritický obor $W > 102,0$ cm. Je zřejmé, že průměrná výška $\overline{X}_{100} = 102,5$ cm sledovaného vzorku lilií leží v kritickém oboru W.

Se spolehlivostí 0,95 lze tedy tvrdit, že zamítáme H_0 ve prospěch H_A, tzn., že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Chcete-li o správnosti nulové hypotézy rozhodnout s jinou spolehlivostí, musíte určit znovu kritický obor W. Máte-li rozhodovat se spolehlivostí 0,99, pak pravděpodobnost chyby I. druhu α, tj. pravděpodobnost překročení kritické hodnoty průměrné výšky $\overline{X}_{krit}$ při platnosti nulové hypotézy H_0, je 0,01.

$$P\left(\overline{X}_{100} > \overline{X}_{krit}\right) = 0,01$$

Označme $F_{\overline{X}}(x)$ distribuční funkci náhodné veličiny $\overline{X}_{100}$ za předpokladu platnosti H_0. Pak

$$1 - F_{\overline{X}}\left(\overline{X}_{krit}\right) = 0,01$$

Postupnými úpravami určíme $\overline{X}_{krit}$.

$$F_{\overline{X}}\left(\overline{X}_{krit}\right) = 0,99$$
$$\Phi\left(\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}}\right) = 0,99$$
$$\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}} = z_{0,99}$$
$$\frac{\overline{X}_{krit} - 100}{\sqrt{1,44}} = 2,326\,(\text{viz Tabulka1})$$
$$\overline{X}_{krit} \cong 102,8\,\text{cm},\ \text{tj.}\ W > 102,8\,\text{cm}$$

Pro spolehlivost testu 0,99 (hladinu významnosti 0,01) je zřejmé, že průměrná výška $\overline{X}_{100} = 102,5$ cm sledovaného vzorku lilií neleží v kritickém oboru W.

Všimněte si, že rozhodnutí o výsledku testu je vázáno na zvolenou spolehlivost testu, tj. na zvolenou pravděpodobnost chyby I. druhu α. Zvýšení spolehlivosti testu z 0,95 na 0,99 vedlo k rozšíření oboru přijetí V (zúžení kritického oboru W), tzn., že k zamítnutí nulové hypotézy bylo zapotřebí zjistit „extrémnější“ hodnoty příslušné výběrové charakteristiky – v našem případě vyšší průměrnou výšku sledované skupiny lilií.

adc)

Rozhodnutí v čistém testu významnosti je prováděno na základě *p-hodnoty*.

Nulová a alternativní hypotéza byly stanoveny ve tvaru

$$\begin{aligned} H_0: &\quad \mu = 100, \\ H_A: &\quad \mu > 100. \end{aligned}$$

Jako testové kritérium $T(X)$ jsme zvolili průměrnou výšku $\overline{X}_{100}$ sledovaného vzorku lilií, která má v případě platnosti nulové hypotézy rozdělení

$$\overline{X}_{100} \to N\,(100; 1,44)$$

Pro daný tvar alternativy je

$$p\text{-}hodnota = 1 - F_0\,(x_{OBS})$$

kde x_{OBS} je pozorovaná hodnota průměrné výšky lilií (102,5 cm) a $F_0(x)$ je distribuční funkce testového kritéria v případě platnosti nulové hypotézy. V našem případě je $F_0(x)$ distribuční funkci rozdělení $N\,(100; 1,44)$.

$$p\text{-}hodnota = 1 - F_0\,(102,5) = 1 - \Phi\left(\frac{102,5 - 100}{\sqrt{1,44}}\right) = 1 - 0,981 = 0,019$$

Je zřejmé, že nulovou hypotézu H_0 lze zamítnout na hladině významnosti 0,019 a vyšších, tj. se spolehlivostí 0,981 a nižší.

Se spolehlivostí 0,95 lze tedy tvrdit, že zamítáme H_0, tzn., že nové pěstební podmínky vedly ke zvýšení střední výšky asijských hybridů lilií.

Se spolehlivostí 0,99 lze tedy tvrdit, že nezamítáme H_0, tzn., že nové pěstební podmínky nevedly ke zvýšení střední výšky asijských hybridů lilií.

add)

Operativní charakteristika je závislosti pravděpodobnosti chyby II. druhu β na konkrétních hodnotách alternativy (při pevně zvolené hodnotě α). Abychom mohli načrtnout operativní charakteristiku, stanovíme si proto hodnoty pravděpodobnosti chyby II. druhu (β) pro několik různých hodnot specifikovaných v jednoduché alternativě (např. 100,5 cm; 101,0 cm; 101,5 cm; 102,0 cm; 103,0 cm a 104,0 cm).

Připomeňte si, že pravděpodobnost chyby II. druhu je

$$P\left(T(X) \in V^*|H_A\right) = \beta,$$

kde V^* označuje obor přijetí.

Zvolíme-li pravděpodobnost chyby I. druhu $\alpha = 0,05$, pak k nezamítnutí nulové hypotézy dojde tehdy, nepřekročí-li průměr $\overline{X}_{100}$ hodnotu 102,0 cm (viz úloha a), tj.

$$P\left(\overline{X}_{100} < 102,0|H_A\right) = \beta$$

Nulovou a jednoduché alternativní hypotézy stanovíme ve tvaru

$$\begin{aligned} H_0: &\quad \mu = 100, \\ H_{A_i}: &\quad \mu = \mu_i, \qquad \forall i = 1,2,\ldots,6 \end{aligned}$$

kde $\mu_1 = 100,5; \mu_2 = 101,0; \mu_3 = 101,5; \mu_4 = 102,0; \mu_5 = 103,0; \mu_6 = 104,0.$

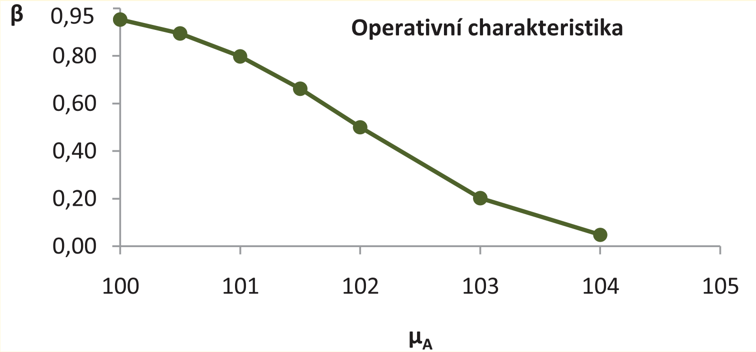
Je zřejmé, že platí-li H_A, pak

$$\overline{X}_{100} \to N\left(\mu_i; 1,44\right).$$

Označme $F_{\overline{x}_{Ai}}$ distribuční funkci náhodné veličiny $\overline{X}_{100}$ za předpokladu platnosti H_A.

Po dosazení dostaneme

$$\begin{aligned}
\beta(\mu_1) &= P\left(\overline{X}_{100} < 102,0|H_{A_1}\right) = F_{\overline{X}_{A_1}}(102,0) = \Phi\left(\frac{102,0-100,5}{\sqrt{1,44}}\right) = \\
&= \Phi(1,25) = 0,894 \\
\beta(\mu_2) &= P\left(\overline{X}_{100} < 102,0|H_{A_2}\right) = F_{\overline{X}_{A_2}}(102,0) = \Phi\left(\frac{102,0-101,0}{\sqrt{1,44}}\right) = \\
&= \Phi(0,83) = 0,798 \\
\beta(\mu_3) &= P\left(\overline{X}_{100} < 102,0|H_{A_3}\right) = F_{\overline{X}_{A_3}}(102,0) = \Phi\left(\frac{102,0-101,5}{\sqrt{1,44}}\right) = \\
&= \Phi(0,42) = 0,662 \\
\beta(\mu_4) &= P\left(\overline{X}_{100} < 102,0|H_{A_4}\right) = F_{\overline{X}_{A_4}}(102,0) = \Phi\left(\frac{102,0-102,0}{\sqrt{1,44}}\right) = \\
&= \Phi(0,00) = 0,5 \\
\beta(\mu_5) &= P\left(\overline{X}_{100} < 102,0|H_{A_5}\right) = F_{\overline{X}_{A_5}}(102,0) = \Phi\left(\frac{102,0-103,0}{\sqrt{1,44}}\right) = \\
&= \Phi(-0,83) = 0,202 \\
\beta(\mu_6) &= P\left(\overline{X}_{100} < 102,0|H_{A_6}\right) = F_{\overline{X}_{A_6}}(102,0) = \Phi\left(\frac{102,0-104,0}{\sqrt{1,44}}\right) = \\
&= \Phi(-1,67) = 0,050
\end{aligned}$$


Operativní charakteristika
β
0,95
0,80
0,60
0,40
0,20
0,00
100
101
102
103
104
105
μ_A

Kapitola 6

Jednovýběrové testy parametrických hypotéz - řešené příklady

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.
Měřítkem kvality kulečníkových koulí je směrodatná odchylka jejich hmotností. Chceme-li testovat směrodatnou odchylku, převedeme daný problém na test rozptylu. Za kvalitní se považují koule, jejichž směrodatná odchylka σ hmotnosti nepřekračuje 2 g, tj. koule, jejichž rozptyl hmotnosti σ^2 nepřekračuje 4 g^2.

Budeme testovat nulovou hypotézu

$$H_0: \quad \sigma^2 = 4.$$

Rozptyl s^2 hmotností $n = 10$ testovaných koulí určíme jako $s^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{n-1}$, kde $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n}$.

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{10} x_i}{n} = \frac{170 + 176 + \ldots + 170}{10} = 170,3\,g \\
s^2 &= \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{n-1} = \frac{(170 - 170,3)^2 + (176 - 170,3)^2 + \ldots + (170 - 170,3)^2}{10 - 1} = \\
&= 5,3\,g^2
\end{aligned}$$

Zajímá nás, zda rozptyl hmotnosti koulí překračuje $4\,g^2$. Vzhledem k tomu, že výběr není v rozporu s tímto očekáváním (výběrový rozptyl s^2 je větší než testovaná hodnota rozptylu ($4\,g^2$)), zvolíme alternativní hypotézu ve tvaru

$$H_A: \quad \sigma^2 > 4.$$

Pro test o rozptylu normálního rozdělení používáme testové kritérium

$$T(X) = \frac{s^2}{\sigma^2}(n-1).$$

mající v případě platnosti nulové hypotézy χ^2 - rozdělení s $n-1$ stupni volnosti. Jelikož v zadání příkladu je uvedeno, že lze předpokládat normalitu hmotnosti kulečníkových koulí, nemusíme normalitu ověřovat.

Pozorovaná hodnota testového kritéria je

$$x_{OBS} = T(X)|_{H_0} = \frac{5,3}{4}(10-1) = 11,88.$$

Vzhledem k tvaru alternativní hypotézy určíme *p-hodnotu* podle vztahu

$$p\text{-}hodnota = 1 - F_0(x_{OBS}),\ (\text{viz tab. } \textbf{??})$$

kde $F_0(x)$ je distribuční funkce χ^2 - rozdělení s 9 stupni volnosti.

$$p\text{-}hodnota = 1 - F_0(11,88) = 0,22\ (\text{viz vybrana\_rozdeleni.xlsx}),$$

p-hodnota je větší než 0,05. Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, rozdíl mezi předpokládaným populačním rozptylem σ_0^2 a zjištěným výběrovým rozptylem (s^2) je statisticky nevýznamný (způsobený náhodným kolísáním). Nelze tedy tvrdit, že rozptyl hmotností kulečníkových koulí je větší než 4 g^2. Sadu kulečníkových koulí značky KULKOUL lze označit za kvalitní.

▲

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.
Budeme testovat nulovou hypotézu

$$H_0: \quad \mu = 100.$$

Průměrné IQ 10 testovaných studentů je

$$\bar{x} = \frac{\sum\limits_{i=1}^{10} x_i}{n} = \frac{65 + 98 + \ldots + 94}{10} \doteq 92,7.$$

Zjištěné průměrné IQ (92,7) je menší než testovaná hodnota (100), což je v souladu s očekáváním, že IQ studentů bude nižší než IQ dospělé populace. Alternativní hypotézu proto zvolíme ve tvaru

$$H_A: \quad \mu < 100.$$

Pro jednovýběrový t test, tj. test o střední hodnotě normálního rozdělení s neznámým rozptylem, používáme testové kritérium

$$T(X) = \frac{\bar{x} - \mu}{s}\sqrt{n},$$

mající v případě platnosti nulové hypotézy Studentovo rozdělení s $n - 1$ stupni volnosti. Jelikož je v zadání příkladu uvedeno, že lze předpokládat normalitu IQ, nemusíme normalitu ověřovat.

Proto, abychom mohli určit pozorovanou hodnotu testového kritéria, musíme nejdříve vypočítat výběrovou směrodatnou odchylku s.

$$s = \sqrt{\frac{\sum\limits_{i=1}^{10}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(65-93)^2+(98-93)^2+\ldots+(94-93)^2}{10-1}} \doteq 14,5$$

Pak

$$x_{OBS} = T(X)|_{H_0} = \frac{92,7 - 100}{14,5}\sqrt{10} = -1,59.$$

Vzhledem ke tvaru alternativní hypotézy určíme *p-hodnotu* podle vztahu

$$p\text{-}hodnota = F_0(x_{OBS}),$$

kde $F_0(x)$ je distribuční funkce Studentova rozdělení s 9 stupni volnosti.

$$p\text{-}hodnota = F_0(-1,59) = 0,073 \text{ (viz vybrana\_rozdeleni.xlsx)}$$

p-hodnota je větší než 0,05. Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, nelze tedy tvrdit, že střední hodnota IQ studentů závěrečného ročníku školy ASNEM je podprůměrná. Jinak řečeno, rozdíl mezi předpokládanou střední hodnotou IQ a pozorovaným průměrným IQ je statisticky nevýznamný.

▲

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.
Ukážeme si řešení pomocí obou výše zmíněných testů hypotéz o mediánu. První krok, tj. stanovení nulové a alternativní hypotézy, je v obou případech stejný.

Data seřadíme a určíme výběrový medián.

65	77	80	93	94	98	102	102	103	113

$$\tilde{x}_{0,5} = \frac{94+98}{2} = 96$$

Budeme testovat nulovou hypotézuu

$$H_0: \quad x_{0,5} = 90$$

vůči alternativě

$H_A: \quad x_{0,5} > 90$ (výběrový soubor ukazuje na to, že je možné, že tvrzení doktorky Hrozné nemusí být pravdivé).

Mediánový (kvantilový) test

Označme Y počet pozorování v náhodném výběru o rozsahu 10, která jsou menší než testovaná hodnota mediánu, tj. 90. Testové kritérium $T(X) = Y$ má za předpokladu platnosti nulové hypotézy binomické rozdělení $Bi(10; 0,5)$. Pozorovaná hodnota testového kritéria $x_{OBS} = 3$ (ve výběru jsou 3 hodnoty menší než 90).

Protože nulové rozdělení je rozdělení diskrétní a v neprospěch nulové hypotézy svědčí nízké hodnoty testového kritéria, určíme *p-hodnotu* jako pravděpodobnost, že testové kritérium nabude hodnoty nejvýše rovné pozorované hodnotě.

$$p\text{-}hodnota = P(T(X) \leqq 3|_{H_0}) = \sum_{k=0}^{3} \binom{10}{k} 0,5^k (1-0,5)10-k \doteq 0,17$$

Vzhledem k pozorované *p-hodnotě* (0,17) nulovou hypotézu nezamítáme.

Jednovýběrový Wilcoxonův test

Pokud by medián rozdělení byl $x_{0,5_0} = 90$ dnů, pak jsou náhodné veličiny $Y_i = X_i - 90$ rovny

$$-25 \quad 8 \quad 13 \quad -13 \quad 3 \quad 12 \quad 12 \quad 23 \quad -10 \quad 4.$$

Seřadíme je vzestupně podle jejich absolutních hodnot, čímž získáme

$$3 \quad 4 \quad 8 \quad -10 \quad 12 \quad 12 \quad -13 \quad 13 \quad 23 \quad -25.$$

Jednotlivým hodnotám přiřadíme pořadí. Nejnižší hodnotě y_i je přiřazena hodnota 1, nejvyšší hodnotě y_i je přiřazena hodnota n. Pokud soubor obsahuje několik pozorování se stejnou absolutní hodnotou, je těmto hodnotám přiřazeno tzv. průměrné pořadí. Např. pozorování -13 a 13 mají stejnou absolutní hodnotu, v seřazeném souboru mají pořadí 7 a 8, proto je oběma těmto hodnotám přiřazeno průměrné pořadí 7,5.)

y_i	3	4	8	-10	12	12	-13	13	23	-25.
r_i^+	1	2	3	4	5,5	5,5	7,5	7,5	9	10

Testové kritérium má tvar

$$T(X) = min(S^+; S^-), \text{ kde } S^+ = \sum_{Y_i \geqq 0} R^+{}_i, S^- = \sum_{Y_i < 0} R^+{}_i.$$

Určíme pozorovanou hodnotu testovacího kritéria.

$$s^+ = \sum_{y_i \geqq 0} r^+{}_i = 1 + 2 + 3 + 5,5 + 5,5 + 7,5 + 9 = 33,5$$

$$s^- = \sum_{y_i < 0} r^+{}_i = 4 + 7,5 + 10 = 21,5$$

$$x_{OBS} = min(s^+; s^-) = 21,5$$

Kritická hodnota jednovýběrového Wilcoxonova testu pro hladinu významnosti 0,05 $\omega_{10}(0,05)$ je 8 (viz tabulka T6). Pozorovaná hodnota (21,5) je větší než kritická hodnota (8), proto nulovou hypotézu nezamítáme.

Obsah

78. strana ze 159

Zavřít dokument

Celá obrazovka/Okno

Považovali-li bychom rozsah výběru za dostatečný (to bychom však měli dělat pouze v případě, že $n > 30$), mohli bychom jako testové kritérium použít

$$T(X) = \frac{S^+ - E(S^+)}{\sqrt{S(S^+)}},$$

kde $E(S^+) = \frac{1}{4}n(n+1), D(S^+) = \frac{1}{24}n(n+1)(2n+1)$. Testové kritérium má při platnosti nulové hypotézy normované normální rozdělení $N(0;1)$

$$E(S^+) = \frac{1}{4}n(n+1) = \frac{1}{4} \cdot 10 \cdot 11 \doteq 27,5$$

$$D(S^+) = \frac{1}{24}n(n+1)(2n+1) = \frac{1}{24} \cdot 10 \cdot 11 \cdot 21 \doteq 96,3$$

$$x_{OBS} = \frac{s^+ - E(S^+)}{\sqrt{D(S^+)}} = \frac{33,5 - 27,5}{\sqrt{96,3}} \doteq 0,61$$

$$p\text{-}hodnota = 1 - \Phi(x_{OBS}) = 1 - \Phi(0,61) \doteq 0,27$$

I při tomto přístupu k testu (připomeňme, že vzhledem k nízkému rozsahu výběru je zde tento přístup uveden jen pro demonstraci postupu) jsme došli k závěru, že nezamítáme nulovou hypotézu.

▲

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.
Na základě výběru $X_1, X_2, \ldots, X_{100}$ (100 pojištěných aut) chceme ověřit předpoklad, že podíl aut starších 7 let (π) je roven 0,25 (π_0). Připomeňme si, že v nulové hypotéze testujeme vždy „rovnost“. Tvrzení, jehož pravdivost chceme ověřit, uvádíme obvykle v alternativě.

Podmínkou pro použití statistického testu je, aby rozsah výběru byl dostatečný, tj. aby byla splněna podmínka

$$n > \frac{9}{p(1-p)}, \text{ tj. } n > 60,98 \left(= \frac{9}{\frac{18}{100}\left(1 - \frac{18}{100}\right)} \right).$$

Abychom mohli ověřit odhad, který uvádí pojišťovna, musíme mít k dispozici výsledky výběrového šetření o rozsahu alespoň 61 pojištěných aut. Toto je splněno. V analyzovaném výběru 100 pojištěných aut bylo zjištěno 18 aut starších než 7 let, tzn.

$$p = \frac{18}{100} = 0,18.$$

Nulovou hypotézu stanovíme ve tvaru

$$H_0: \quad \pi = 0,25.$$

Výběrová relativní četnost p aut starších než 7 let je menší než pravděpodobnost π_0 odhadovaná pojišťovnou, proto alternativu volíme ve tvaru

$$H_A: \quad \pi < 0,25.$$

Testovým kritériem je statistika

$$T(X) = \frac{p - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n},$$

která má v případě platnosti nulové hypotézy normované normální rozdělení $N(0;1)$.

Stanovíme pozorovanou hodnotu testové statistiky a na základě tvaru alternativy vypočteme *p-hodnotu*.

$$x_{OBS} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} = \frac{0,18 - 0,25}{\sqrt{0,25(1-0,25)}}\sqrt{100} \doteq -1,617$$

$$p\text{-}hodnota = F_0(-1,617) = \Phi(-1,617) \doteq 0,053$$

Na hladině významnosti 0,05 nulovou hypotézu nezamítáme, nelze tedy tvrdit, že podíl aut starších 7 let je nižší než 25%. (Všimněte si, že pokud bychom se spokojili s vyšší pravděpodobnosti chyby I. druhu (např. 0,06), nulovou hypotézu bychom zamítli a bylo by možné prohlásit, že podíl aut starších 7 let je nižší než 25%.)

▲

Kapitola 7

Dvouvýběrové testy parametrických hypotéz - řešené příklady

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.

Chceme porovnávat střední obsah nikotinu v cigaretách TAB a NIK, směrodatnou odchylku obsahu nikotinu v cigaretách neznáme, lze předpokládat, že není stejná. Předpoklad normality je splněn, předpoklad o shodě rozptylů obsahu nikotinu v cigaretách TAB a NIK vyvrátíme F-testem.

$$\begin{aligned} H_0: &\quad \sigma^2_{TAB} = \sigma^2_{NIK} \quad \text{neboli} \quad \frac{\sigma^2_{TAB}}{\sigma^2_{NIK}} = 1 \\ H_A: &\quad \sigma^2_{TAB} < \sigma^2_{NIK} \quad \left(s^2_{TAB} = 3,7^2 \text{ je menší než } s^2_{NIK} = 4,3^2\right) \end{aligned}$$

$$x_{OBS} = \left.\frac{\frac{s^2_{TAB}}{\sigma^2_{TAB}}}{\frac{s^2_{NIK}}{\sigma^2_{NIK}}}\right|_{H_0} = \left.\frac{\frac{s^2_{TAB}}{s^2_{NIK}}}{\frac{\sigma^2_{TAB}}{\sigma^2_{NIK}}}\right|_{H_0} = \frac{\frac{3{,}7^2}{4{,}3^2}}{1} \doteq 0,74$$

$$p\text{-}hodnota = F_0(0,74),$$

kde $F_0(x)$ je distribuční funkce Fisher-Snedecorova rozdělení s $n_{TAB} - 1 = 399$ stupni volnosti pro čitatele a $n_{NIK} - 1 = 499$ stupni volnosti pro jmenovatele.

$$p\text{-}hodnota = 0,0008$$

Nulovou hypotézu zamítáme, předpoklad o různosti rozptylů byl potvrzen. Pro ověření shody středních hodnot proto zvolíme **Aspinové-Welchův test**.

$$\begin{aligned} H_0: &\quad \mu_{TAB} = \mu_{NIK} \\ H_A: &\quad \mu_{TAB} < \mu_{NIK} \quad (\overline{x}_{TAB} = 42,6 \text{ je menší než } \overline{x}_{NIK} = 48,9) \end{aligned}$$

Testové kritérium

$$T\left(\boldsymbol{X},\boldsymbol{Y}\right)=\frac{\left(\overline{X}_{TAB}-\overline{Y}_{NIK}\right)-\left(\mu_{TAB}-\mu_{NIK}\right)}{\sqrt{\frac{s_{TAB}^2}{n_{TAB}}+\frac{s_{NIK}^2}{n_{NIK}}}}$$

má za předpokladu platnosti nulové hypotézy Studentovo rozdělení s v stupni volnosti, kde

$$\nu=\frac{\left(\frac{s_{TAB}^2}{n_{TAB}}+\frac{s_{NIK}^2}{n_{NIK}}\right)^2}{\frac{1}{n_{TAB}-1}\left(\frac{s_{TAB}^2}{n_{TAB}}\right)^2+\frac{1}{n_{NIK}-1}\left(\frac{s_{NIK}^2}{n_{NIK}}\right)^2}=\frac{\left(\frac{3{,}7^2}{400}+\frac{4{,}3^2}{500}\right)^2}{\frac{1}{399}\left(\frac{3{,}7}{400}\right)^2+\frac{1}{499}\left(\frac{4,3^2}{500}\right)^2}\doteq 893$$

$$x_{OBS}=\frac{\left(\overline{x}_{TAB}-\overline{x}_{NIK}\right)-\left(\mu_{TAB}-\mu_{NIK}\right)}{\sqrt{\frac{s_{TAB}^2}{n_{TAB}}+\frac{s_{NIK}^2}{n_{NIK}}}}=\frac{(42,6-48,9)-(0)}{\sqrt{\frac{3{,}7^2}{400}+\frac{4{,}3^2}{500}}}=-23,6$$

$$p\text{-hodnota}=F_0(-23,6),$$

kde $F_0(x)$ je distribuční funkce Studentova rozdělení s 893 stupni volnosti.

$$p\text{-hodnota}\doteq 0$$

Zamítáme nulovou hypotézu (na hladině významnosti 0,05), tvrzení firmy TAB lze považovat za pravdivé.

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.

Označme $x_1, x_2, \ldots, x_{15}$ výběr studentů, kteří absolvovali klasickou výuku a $y_1, y_2, \ldots, y_{10}$ výběr studentů, kteří absolvovali výuku experimentální. (Označení výběrů bylo provedeno v souladu s požadavkem, aby $n_1 \geqq n_2$.)

Budeme testovat nulovou hypotézu

$$H_0: \quad x_{0,5} = y_{0,5},$$

vůči proti alternativě $H_A: \quad x_{0,5} < y_{0,5} \qquad (\tilde{x}_{0,5} = 48, \tilde{y}_{0,5} = 51)$

Nyní vypočteme pozorovanou hodnotu testové statistiky. Nejdříve přiřadíme pořadí hodnotám z obou výběrů seřazeným podle velikosti.

Skupina	Y	X	Y	X	X	X	X	Y	Y	X	X	X	Y	X	X	X	Y	X	X	X	X	Y	Y	Y	Y
Výsledek	6	13	18	28	30	30	33	36	38	40	45	48	48	49	52	52	54	57	60	68	68	68	72	84	92
Pořadí	1	2	3	4	5,5	5,5	7	8	9	10	11	12,5	12,5	14	15,5	15,5	17	18	19	21	21	21	23	24	25

Rozsah prvního výběru $n_1 = 15$, rozsah druhého výběru $n_2 = 10$.

Nyní určíme:
součet pořadí prvního výběru $T_1 = 2 + 4 + \cdots + 21 = 181,5$,
součet pořadí druhého výběru $T_2 = 1 + 3 + \cdots + 25 = 143,5$.

Pak $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1 = 88,5$, $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2 = 61,5$. Pro kontrolu numerické správnosti výpočtu lze ověřit, že $U_1 + U_2 = n_1 n_2$.

$$T(\boldsymbol{X},\boldsymbol{Y}) = min\,(U_1, U_2) = 61,5$$

Kritická hodnota uvedena v tabulce T7 je 39. Protože pozorovaná hodnota testové statistiky $61,5 > 39$, na hladině významnosti 0,05 nezamítáme nulovou hypotézu, že způsob výuky nemá vliv na studijní výsledky.

Kdybychom pro ilustraci použili postup pro velká n_1 a n_2, pak bychom dostali

$$T(\boldsymbol{X},\boldsymbol{Y}) = \frac{\left(min\,(U_1,U_2) - \frac{n_1 n_2}{2}\right)}{\sqrt{\frac{1}{12} n_1 n_2 \,(n_1 + n_2 + 1)}} \doteq -0,748, \ p\text{-}hodnota = \Phi(-0,748) = 0,23.$$

Je zřejmé, že ani při tomto přístupu bychom nulovou hypotézu nezamítli.

▲

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

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.

Chceme porovnávat podíl reklamovaných výrobků u obou firem. Volíme tedy test homogenity dvou binomických rozdělení. Nejdříve ověříme, zda pro provedení testu máme k dispozici výběry dostatečného rozsahu.

Označme relativní četnost reklamovaných magnetofonů SONIE p_S a relativní četnost reklamovaných magnetofonů PHILL p_P.

$$p_S = \frac{10}{300} \doteq 0,033, \qquad p_P = \frac{18}{440} \doteq 0,041.$$

Pro splnění výše uvedených kritérií zaručujících korektnost testu musí být testováno alespoň $\frac{9}{p_S(1-p_S)} \doteq 280$ magnetofonů firmy SONIE a $\frac{9}{p_P(1-p_P)} \doteq 230$ magnetofonů firmy PHILL. To je splněno ($n_S = 300, n_P = 440$).

Budeme testovat nulovou hypotézu

$$H_0: \quad \pi_S = \pi_P$$

vůči alternativě $H_A: \quad \pi_S < \pi_P$.

(Uvědomte si, proč byla zvolena alternativa v tomto tvaru.)

Pozorovaná hodnota testového kritéria je

$$x_{OBS} = \left.\frac{(p_S - p_P) - (\pi_S - \pi_P)}{\sqrt{\frac{p_S(1-p_S)}{n_S} + \frac{p_P(1-p_P)}{n_P}}}\right|_{H_0} = \frac{(0{,}033 - 0{,}041) - (0)}{\sqrt{\frac{0{,}033(1-0{,}033)}{300} + \frac{0{,}041(1-0{,}041}{440}}} = 0,54.$$

Nulové rozdělení testového kritéria je normované normální a alternativa je ve tvaru $\pi_S < \pi_P$, proto

$$p\text{-}hodnota = \Phi(-0,54) \doteq 0,290.$$

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu ($p\text{-}hodnota > 0,05$), tvrzení firmy SONIE o nižším procentu reklamací tedy nelze považovat za oprávněné.

▲

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.

Je zřejmé, že máme k dispozici páry závislých pozorování, proto přistoupíme k párovému t testu. Nemá smysl porovnávat průměrné ojetí pravých a levých pneumatik. Budeme zjišťovat, jaká je střední hodnota rozdílu ojetí pravé a levé pneumatiky.

Označme X_i ojetí i-té pravé pneumatiky a Y_i ojetí i-té levé pneumatiky. Pak $D_i = X_i - Y_i$ udává rozdíl v ojetí pravé a levé pneumatiky u i-tého automobilu.

Pravá X	1,8	1,0	2,2	0,9	1,5	1,6
Levá Y	1,5	1,1	2,0	1,1	1,4	1,4
Pravá-Levá D	0,3	-0,1	0,2	-0,2	0,1	0,2

Rozdíl v ojetí pravé a levé pneumatiky [mm] má normální rozdělení. Proto lze pro srovnání ojetí předních pneumatik použít párový t test.

Označme $\mu = E(D)$. Budeme testovat nulovou hypotézu

$$H_0: \quad \mu = 0.$$

Průměrný rozdíl ojetí pravé a levé pneumatiky je

$$\overline{d} = \frac{\sum\limits_{i=1}^{n} d_i}{n} = \frac{0{,}3+(-0{,}1)+\cdots+0{,}2}{6} \doteq 0,08.$$

Zjištěný průměrný rozdíl v ojetí pneumatik (0,08) je větší než testovaná hodnota (0). Výběr ukazuje na to, že by se mohly pravé pneumatiky ojíždět více než levé. Alternativní hypotézu proto zvolíme ve tvaru $H_A: \mu > 0$.

Pro párový t test používáme testové kritérium $T(D) = \frac{\overline{d}-\mu}{S_D}\sqrt{n}$ mající v případě platnosti nulové hypotézy Studentovo rozdělení s $n-1$ stupni volnosti.

$$s_D = \frac{\sqrt{\sum\limits_{i=1}^{n}\left(d_i - \overline{d}\right)^2}}{n-1} \doteq \sqrt{\frac{(0{,}3-0{,}08)^2+\cdots+(0{,}2-0{,}08)^2}{6-1}} \doteq 0,19$$

Pak $x_{OBS} = T(D)|_{H_O} = \frac{0{,}08-0}{0{,}19}\sqrt{6} = 1,05.$

Vzhledem k tvaru alternativní hypotézy určíme $p-hodnotu$ podle vztahu

$$p\text{-}hodnota = 1 - F_0\left(x_{OBS}\right),$$

kde $F_0(x)$ je distribuční funkce Studentova rozdělení s 5 stupni volnosti.

$$p\text{-}hodnota = F_0(1,05) = 1 - F_0(1,05) = 0,17 \quad \text{(viz vybrana\_rozdeleni.xlsx)}$$

p-hodnota je větší než 0,05. Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, která říká, že pozorovaný rozdíl v ojetí pneumatik není statisticky významný. Nelze tvrdit, že se přední pneumatiky ojíždějí různě.

▲

Kapitola 8

Vícevýběrové testy parametrických hypotéz - řešené příklady

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.

Máme 4 nezávislé výběry. Je třeba testovat hypotézu

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$

proti alternativě, že se alespoň jedna dvojice rozptylů liší

$$H_A : \neg H_0.$$

Bartlettův test

$$s_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)s_i^2 = 3,86,$$

$$C = 1 - \frac{1}{a(k-1)}\left(\frac{1}{n-k} - \sum_{i=1}^{k}\frac{1}{n_i-1}\right) = 1,069.$$

$$x_{OBS} = \frac{1}{c}\left[(n-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2\right] = 1,106.$$

p-hodnota $= 1 - F_0(1,106)$, kde $F_0(x)$ je distribuční fuknce χ^2 rozdělení s 24 stupni volnosti.

p-hodnota $\doteq 1$

Protože *p-hodnota* $\doteq 1$ nelze zamítnout nulovou hypotézu. Protože nemáme informaci o normalitě jednotlivých výběrů, provedeme Leveneův test. (Barttletův test je citlivý na porušení normality!)

Leveneův test

Nechť $Z_{ij} = |X_{ij} - \bar{X}_i|$.

Výrobce	Z_{ij} [kg/m^3]							Průměr $\bar{Z}_i$ [kg/m^3]
A	2,3	3,6	1,0	0,3	0,5	3,4	1,8	1,8
B	1,5	1,9	0,6	0,4	0,3	3,2	2,1	1,4
C	1,4	1,5	2,5	1,8	0,1	0,3	0,6	1,2
D	0,5	1,1	0,8	0,0	3,6	2,5	1,3	1,4

Pak

$$\bar{\bar{Z}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{Z_{ij}}{n} = 1,46,$$

$$SS_{ZB} = \sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{\bar{Z}})^2 = 1,63,$$

$$SS_{Ze} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2 = 31,34,$$

$$x_{OBS} = \frac{\frac{SS_{ZB}}{k-1}}{\frac{SS_{Ze}}{n-k}} = 0,42.$$

p-hodnota$= 1 - F_0(0,42)$,

kde $F_0(x)$ je distribuční funkce Fisherova-Snedecorova rozdělení s 3 stupni volnosti v čitateli a 24 stupni volnosti ve jmenovateli.

p-hodnota = 0,74

Protože *p-hodnota* = 0,74, nelze homoskedasticitu zamítnout ani na základě Leveneova testu.

Vzhledem k vyváženosti třídění lze pro ověření homoskedasticity použít rovněž Hartleyův a Cochranův test.

Hartleyův test

Hartleyův test je založen na testové statistice

$$F_{max} = \frac{\max s_i^2}{\min s_i^2}.$$

Pozorovaná hodnota $x_{OBS} = 2,43 (= 5,73/2,36)$. Pozorovaná hodnota nepřekročila kritickou hodnotu $h_{0,05}(4,6) = 10,4$ (tabulka T8), proto na hladině významnosti 0,05 nezamítá homoskedasticitu ani tento test.

Cochranův test

Tento test používá testovou statistiku

$$G_{max} = \frac{\max s_i^2}{s_1^2 + \ldots + s_k^2}.$$

Pozorovaná hodnota $x_{OBS} = 0,37(= 5,73/(5,73+3,52+2,36+3,83))$. Pozorovaná hodnota nepřekročila kritickou hodnotu $c_{0,05}(4,6) = 0,56$ (tabulka T9), proto na hladině významnosti 0,05 nezamítáme nulovou hypotézu.

▲

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.

Dílčí výpočty zaznamenáme do tabuly.

	Skupina			
	Gymnázium 1	SPŠ 2	OU 3	
	55	52	47	
	54	50	53	
	58	51	49	
	61	51	50	
	52	49	46	
	60		48	
	53		50	
	65			
Rozsah	8	5	7	$n = 20$
Průměr $\bar{X}_i$	57,3	50,6	49,0	$\bar{\bar{X}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{X_{ij}}{n} = 52{,}7$
$(\bar{X}_i - \bar{\bar{X}})$	4,6	-2,1	-3,7	
$n_i(\bar{X}_i - \bar{\bar{X}})^2$	165,62	22,05	95,83	$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 = 283{,}5$
Výběrový rozptyl s_i^2	20,5	1,3	5,3	

Celková variabilita je dána celkovým součtem čtverců SS_T, resp. celkovým rozptylem MS_T.

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = (55 - 52,7)^2 + \ldots + (50 - 52,7)^2 = 464,2$$

$$MS_T = \frac{SS_T}{n-1} = \frac{464,2}{20-1} = 24,4$$

Variabilita mezi třídami je dána součtem čtverců mezi třídami SS_B, resp. rozptylem mezi třídami MS_B.

$$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 = 283,5$$

$$MS_B = \frac{SS_B}{k-1} = \frac{283,5}{3-1} = 141,8$$

Variabilita uvnitř tříd je dána reziduálním součtem čtverců SS_e, resp. reziduálním rozptylem MS_e.

$$SS_\mathrm{e} = \sum_{i=1}^{k} (n_i - 1)s_i^2 = 180,7$$

$$MS_\mathrm{e} = \frac{SS_\mathrm{e}}{n-k} = \frac{180,7}{20-3} = 10,6$$

▲

Příklad 8.3. Dokončete analýzu rozptylu pro motivační příklad.

Řešení.

Z předcházejícího řešeného příkladu převezmeme veškeré dílčí výsledky, určíme pozorovanou hodnotu testového kritéria a určíme *p-hodnotu*. Postupně vyplňujeme tabulku analýzy rozptylu.

$$x_{OBS} = \frac{MS_B}{MS_{\mathrm{e}}} = \frac{141,8}{10,6} = 13,3$$

p-hodnota$= 1 - F_0(x_{OBS}) = 1 - F_0(13,3)$,

kde $F_0(x)$ je distribuční funkce Fisherova-Snedecorovo rozdělení s 2 stupni volnosti v čitateli a 17 stupni volnosti ve jmenovateli.

p-hdonota = 0,0003 (viz vybrana_rozdeleni.xls)

Na hladině významnosti 0,05 zamítáme nulovou hypotézu o shodě středních hodnot. Lze tedy tvrdit, že typ absolvované střední školy má vliv na výsledek přijímací zkoušky z matematiky.

Připomeňme si, že výsledek analýzy rozptylu nám pouze říká, že průměry nejsou stejné. Je třeba provést další analýzu, abychom zjistili, jak se liší. Absolventi, jakého typu střední školy mají statisticky významně lepší (resp. horší) šanci na lepší výsledek? Odpověď na tuto otázku nám dá tzv. post hoc analýza neboli mnohonásobné porovnávání.

▲

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	283,5	2	141,75	13,34	0,0003
Within groups	180,7	17	10,63		
Total (Corr.)	464,2	19			

Obr. 8.1: Ukázka výstupu metody ANOVA (software Statgraphics)

Příklad 8.4. Proveďte post hoc analýzu pro data z motivačního příkladu.

Řešení.

Výsledkem analýzy rozptylu bylo zamítnutí nulové hypotézy, zajímá nás tedy odpověď na otázku „Absolventi, jakého typu střední školy mají statisticky významně lepší (resp. horší) šanci na lepší výsledek?“

Připomeňme si potřebné dílčí výsledky získané v průběhu analýzy rozptylu.

	Skupina			
	Gymnázium 1	SPŠ 2	OU 3	
Rozsah	8	5	7	$n = 20$
Průměr $\bar{X}_i$	57,3	50,6	49,0	$\bar{\bar{X}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{X_{ij}}{n} = 52{,}7$

$$MS_e = 10,6$$

Testujeme $H_0 : \mu_I = \mu_J$ vůči alternativě $H_A : \mu_I \neq \mu_J$.

Fisherovo LSD

Nulovou hypotézu zamítáme pokud $|\tilde{x}_I - \tilde{x}_J| \geqq LSD_{IJ}$, kde LSD_{IJ} určíme jako

$$LSD_{IJ} = t_{1-\frac{\alpha}{2}}(n-k)\sqrt{MS_e}\sqrt{\frac{1}{n_I} + \frac{1}{n_J}}.$$

$t_{1-\frac{\alpha}{2}}(n-k) = t_{0,975}(17) = 2,1 \Rightarrow LSD_{IJ} = 2,1\sqrt{10,6}\sqrt{\frac{1}{n_I} + \frac{1}{n_J}} = 6,837\sqrt{\frac{1}{n_I} + \frac{1}{n_J}}$

	$\lvert\bar{x}_I - \bar{x}_J\rvert$	LSD_{IJ}
Gymnázium – SPŠ*	6,7	3,898
Gymnázium – OU*	8,3	3,539
SPŠ - OU	1,6	4,003

Fisherovo LSD identifikovalo jako statisticky významné rozdíly mezi průměrným hodnocením absolventů gymnázií a SPŠ a gymnázií a OU. Lze tedy tvrdit, že absolventi gymnázií

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

mají statisticky významně vyšší průměrné výsledky než studenti SPŠ a OU, jejichž průměrné výsledky jsou srovnatelné.

Bonferroniho metoda

Nulovou hypotézu zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq t_{1-\frac{\alpha^*}{2}}(n-k)\sqrt{MS_{\mathrm{e}}}\sqrt{\frac{1}{n_I}+\frac{1}{n_J}}$$

kde α^* je upravená hladina významnosti, $\alpha^2 = \frac{\alpha}{\binom{k}{2}}$.

$\alpha^* = \frac{\alpha}{\binom{k}{2}} = \frac{0{,}05}{\binom{3}{2}} = 0,0167,\ t_{1-\frac{\alpha^*}{2}}(n-k) = t_{0{,}99165}(17) = 2,65$

$t_{\alpha^*}(n-k)\sqrt{MS_{\mathrm{e}}}\sqrt{\frac{1}{n_I}+\frac{1}{n_J}} = 2,65\sqrt{10,6}\sqrt{\frac{1}{n_I}+\frac{1}{n_J}} = 8,628\sqrt{\frac{1}{n_I}+\frac{1}{n_J}}$

	$\lvert \bar{x}_I - \bar{x}_J \rvert$	Kritická hodnota
Gymnázium – SPŠ*	6,7	4,919
Gymnázium – OU*	8,3	4,465
SPŠ - OU	1,6	5,052

Bonferroniho metoda poskytla stejné výsledky jako Fisherovo LSD.

Schefféo metoda

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 $F_{1-\alpha}(k-1, n-k)(k-1)$ je $(1-\alpha)$ kvantil Fisher-Snedecorova rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.

$F_{1-\alpha}(k-1, n-k) = F_{0{,}98}(2, 17) = 3{,}59$

$$\sqrt{MS_\mathrm{e}}\sqrt{F_{1-\alpha}(k-1, n-k)(k-1)\left(\tfrac{1}{n_I} + \tfrac{1}{n_J}\right)} = \sqrt{10{,}6}\sqrt{3{,}59 \cdot 2\left(\tfrac{1}{n_I} + \tfrac{1}{n_J}\right)} =$$
$$= 8{,}72\sqrt{\left(\tfrac{1}{n_I} + \tfrac{1}{n_J}\right)}$$

	$\lvert \bar{x}_I - \bar{x}_J \rvert$	Kritická hodnota
Gymnázium – SPŠ*	6,7	4,973
Gymnázium – OU*	8,3	4,515
SPŠ - OU	1,6	5,108

Rovněž Scheffého metoda identifikovala „Gymnázium“ jako skupinu, která se statisticky významně liší od ostatních.

Neboť rozsahy jednotlivých výběrů nejsou stejné, nelze pro post hoc analýzu použít Tukeyho metodu.

Tukey HSD

Nulovou hypotézu pak zamítáme, pokud

$$|\tilde{x}_I - \tilde{x}_J| \geqq q_\alpha(k, n-k)\sqrt{MS_e}\sqrt{\frac{1}{2}\left(\frac{1}{n_I} + \frac{1}{n_J}\right)},$$

kde $q_\alpha(k, n-k)$ je α kvantil studentizovaného rozpětí, který je tabelován.

$q_\alpha(k, n-k) = q_{0,05}(3, 17) = 3,63$ (viz tabulka T10)

$q_\alpha(k, n-k)\sqrt{MS_e}\sqrt{\frac{1}{2}\left(\frac{1}{n_I} + \frac{1}{n_J}\right)} = 3,63\sqrt{10,6}\sqrt{\frac{1}{2}}\sqrt{\left(\frac{1}{n_I} + \frac{1}{n_J}\right)} = 8,357\sqrt{\left(\frac{1}{n_I} + \frac{1}{n_J}\right)}$

	$\left\|\bar{x}_I - \bar{x}_J\right\|$	Kritická hodnota
Gymnázium – SPŠ*	6,7	4,764
Gymnázium – OU*	8,3	4,325
SPŠ - OU	0,4	4,893

Výsledky post hoc analýzy získané metodou Tukey HSD jsou v souladu s výsledky získanými pomocí Fisherova LSD, resp. pomocí Bonferroniho metody.

▲

Příklad 8.5. Analyzujte data z motivačního příkladu pomocí Kruskalova-Wallisova testu.

Řešení.

Chceme testovat hypotézu o shodě mediánů

$$H_0: \quad x_{0,5_G} = x_{0,5_{\text{SPŠ}}} = x_{0,5_{OU}}$$

vůči alternativě, že H_0 neplatí.

Všech n pozorovaných hodnot seřadíme do rostoucí posloupnosti a určíme jejich **pořadí** R_i. Tato pořadí uspořádáme do tabulky a určíme tzv. **součty pořadí pro jednotlivé výběry** T_i.

Data		
Gymnázium 1	SPŠ 2	OU 3
55	52	47
54	50	53
58	51	49
61	51	50
52	49	46
60		48
53		50
65		

	Pořadí R_{ij}			
	Gymnázium 1	SPŠ 2	OU 3	
	16	11,5	2	
	15	7	13,5	
	17	9,5	4,5	
	19	9,5	7	
	11,5	4,5	1	
	18		3	
	13,5		7	
	20			
Rozsah výběru n_i	8	5	7	$n = \sum_{i=1}^k n_i$
Součty pořadí T_i	130	42	38	$\sum_{i=1}^k T_i = 210$
$t_i = \frac{T_i}{n_i}$	16,25	8,40	5,43	
$\frac{T_i^2}{n_i}$	2112,5	352,8	206,3	$\sum_{i=1}^k \frac{T_i^2}{n_i} = 2671{,}6$

Všimněte si, že $\sum_{i=1}^{k} T_i = \frac{n(n+1)}{2} = \frac{20 \cdot 21}{2} = 210.$

Pozorovaná hodnota $x_{OBS} = -3(n+1) + \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} = 13,3.$

p-hodnota $= 1 - F_0(13,3)$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s 2 stupni volnosti.

p-hodnota $= 0{,}001$

Zamítáme nulovou hypotézu o shodě mediánů. Proto provedeme post hoc analýzu. Protože analyzujeme výběry o různém rozsahu, použijeme pro post hoc analýzu Dunnové test.

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.

$z_{1-\alpha^*} = z_{1-\frac{\alpha}{\binom{k}{2}}} = z_{1-\frac{0,05}{\binom{3}{2}}} = z_{0,9833} = 2,13$ (viz vybrana_rozdeleni.xls)

$$\sqrt{\frac{1}{12}\left(\frac{1}{n_I}+\frac{1}{n_J}\right)n(n+1)z_{1-\alpha^*}} = \sqrt{\frac{1}{12}\left(\frac{1}{n_I}+\frac{1}{n_J}\right)20\cdot 21\cdot 2,13} = 8,634\sqrt{\left(\frac{1}{n_I}+\frac{1}{n_J}\right)}$$

	$\lvert t_I - t_J \rvert$	Kritická hodnota
Gymnázium – SPŠ*	7,85	4,922
Gymnázium – OU*	10,82	4,469
SPŠ - OU	2,97	5,056

Na základě post hoc analýzy lze na hladině významnosti 0,05 tvrdit, že absolventi gymnázií mají statisticky významně vyšší průměrné výsledky než studenti SPŠ a OU, jejichž průměrné výsledky jsou srovnatelné.

▲

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.

Chceme porovnat srdeční frekvenci hráčů v jednotlivých čtvrtinách utkání. Pro každého hráče máme čtveřici pozorování, je tedy zřejmé, že chceme analyzovat shodu úrovně ve 4 závislých výběrech. Pro takovouto analýzu je určen Friedmanův test, kterým vyšetříme, zda se tep v průběhu utkání mění jen náhodně nebo zda se do jeho změn promítá nějaký systematický vliv času.

Chceme testovat hypotézu

$$H_0 : x_{0,5_1} = x_{0,5_2} = x_{0,5_3} = x_{0,5_4}$$

vůči alternativě

$$H_A : \neg H_0.$$

U každého sledovaného hráče nahradíme zjištěné výsledky jejich pořadím (viz tabulka 8.4).

Tab. 8.1: Tabulka pořadí

Pořadí R_{ij}				
Číslo hráče	Čtvrtina			
	1	2	3	4
1	1	2	3	4
2	1	2	3,5	3,5
3	3	1	2	4
4	2	1	3	4
5	1	2	3	4
6	1	3	2	4
$R_j = \sum_{i=1}^{m} R_{ij}$	9	11	16,5	23,5

Počet sledovaných objektů $m = 6$, počet porovnávaných skupin $k = 4$. Protože $\min(k; m) >$ > 5 lze nulové rozdělení testové statistiky

$$Q = \frac{12}{mk(k+1)} \sum_{j=1}^{2} -3m(k+1)$$

aproximovat rozdělením χ^2 s $k-1$ stupni volnosti. Proto *p-hodnota*$= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti.

$$x_{OBS} = \frac{12}{6 \cdot 4(4+1)}(9^2 + 11^2 + 16,5^2 + 23,5^2) - 3 \cdot 6 \cdot (4+1) = 12,65$$

p-hodnota$= 1 - F_0(12,65) = 0,0005$ (viz vybrana_rozdeleni.xlsx)

Na hladině významnosti 0,05 zamítáme nulovou hypotézu. Lze tedy tvrdit, že v průběhu utkání dochází ke změnám srdeční frekvence hráčů.

Post hoc analýza

Vypočteme rozdíly mezi součty pořadí $|R_r - R_s|$ pro všechny dvojice $r < s$ a srovnáme je s příslušnou tabelovanou kritickou hodnotou 11,5 (viz tabulka T13).

$\lvert R_r - R_s \rvert$				
	1	2	3	4
1	-	2	7,5	14,5
2		-	5,5	12,5
3			-	7
4				-

Kritickou hodnotu překračují $|R_1 - R_4|$ a $|R_2 - R_4|$. Tím je prokázán signifikantní rozdíl mezi srdeční frekvenci v 1. a ve 4. čtvrtině a v 2. a ve 4. čtvrtině.

▲

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

Kapitola 9

Testy dobré shody - řešené příklady

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?

Řešení.
Chceme zjistit (na hladině významnosti 0,05), zda je výběr reprezentativní, tj. zda lze odchylky mezi zjištěnými a očekávanými četnostmi jednotlivých kategorií označit za náhodné. Nulovou hypotézu proto formulujeme:

H_0: Provedený výběr **je** výběrem z populace, v níž jsou relativní četnosti jednotlivých variant dány tabulkou 9.1.

Tab. 9.1: Očekávané relativní četnosti jednotlivých kategorií rodinného stavu žen starších 15 let

Stav	svobodná	vdaná	rozvedená	ovdovělá
relativní četnost π_{0_i}	0,248	0,490	0,126	0,136

Alternativu stanovíme jako negaci nulové hypotézy.

H_A: $\neg H_0$, tj. provedený výběr není výběrem z populace, v níž jsou relativní četnosti jednotlivých variant dány tabulkou 9.1.

Jako testové kritérium používáme náhodnou veličinu

$$G = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i},$$

která má v případě platnosti nulové hypotézy a za předpokladu, že provádíme dostatečně velký výběr, přibližně χ^2 rozdělení s $k-1$ stupni volnosti.

Empirické četnosti O_i jsou dány v zadání příkladu, očekávané četnosti E_i (tj. zastoupení žen v jednotlivých kategoriích očekávané v případě platnosti nulové hypotézy) určíme jako

$$E_i = n\pi_{i_0},$$

kde n je rozsah výběru, v našem případě 246. Například: pokud by platila nulová hypotéza, pak by v uskutečněném výběru mělo být $E_1 = 246 \cdot 0,248 \doteq 61$ svobodných žen. Pozorované a očekávané četnosti jednotlivých variant jsou uvedeny v tabulce 9.2.

Tab. 9.2: Pozorované a očekávané četnosti jednotlivých kategorií rodinného stavu žen starších 15 let

Stav	svobodná	vdaná	rozvedená	ovdovělá
pozorované četnosti O_i	80	110	30	26
očekávané četnosti E_i	61,0	120,5	31,0	33,5

Předpokladem pro použití χ^2- testu dobré shody je, aby očekávané četnosti E_i byly větší než 5. Je zřejmé, že tento předpoklad lze považovat za splněný.

Pozorovaná hodnota testového kritéria

$$x_{OBS} = \frac{\sum_{i=1}^{4}(O_i - E_i)^2}{E_i} = \frac{(80-61,0)^2}{61,0} + \frac{(110-120,5)^2}{120,5} + \frac{(30-31,0)^2}{31,0} + \\ + \frac{(26-33,5)^2}{33,5} = 8,53$$

Všimněte si, že čím větší jsou odchylky pozorovaných a očekávaných četností, tím větší je pozorovaná hodnota x_{OBS}. Čím větší je pozorovaná hodnota x_{OBS}, tím silnější je výpověď výběru proti nulové hypotéze.

Předpoklad testu je splněn, *p-hodnota* $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s 3 (=4-1) stupni volnosti.

$p\text{-}hodnota = 1 - F_0(8,53) = 0,036$ (viz vybrana_rozdeleni.xls)

$p\text{-}hodnota < 0,05$, proto na hladině významnosti 0,05 zamítáme nulovou hypotézu ve prospěch alternativy. Výběr nelze označit za reprezentativní.

▲

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. 9.3: 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.

Definujeme-li si náhodnou veličinu X jako počet poruch daného zařízení během jednoho dne, pak nulovou a alternativní hypotézu formulujeme ve tvaru:

H_0: Počet poruch daného zařízení během jednoho dne (náhodná veličina X) má Poissonovo rozdělení s parametrem $\lambda t = 1,2$, neboli výběr pochází z Poissonova rozdělení s parametrem $\lambda t = 1,2$.

H_A : $\neg H_0$, tj.není pravda, že počet poruch daného zařízení během jednoho dne má Poissonovo rozdělení s parametrem $\lambda t = 1,2$.

Poissonovo rozdělení má pouze jediný parametr λt. Tento parametr je specifikován v nulové hypotéze, tzn. jde o **úplně specifikovaný test** (počet odhadovaných parametrů $h = 0$).

Poissonovo rozdělení je rozdělením diskrétním, proto pro každou variantu x_i vypočteme pravděpodobnost π_{0_i}, že se náhodná veličina X s pravděpodobnostní funkcí $P(x)$ odpovídající nulové hypotéze bude realizovat variantou x_i. (Empirické četnosti 0_i jsou dány v zadání příkladu.)

Platí-li nulová hypotéza, pak má náhodná veličina X (počet poruch daného zařízení během jednoho dne) Poissonovo rozdělení s parametrem $\lambda t = 1,2$. Pravděpodobnostní funkce Poissonova rozdělení je dána vztahem

$$P(x) = \frac{(\lambda t)^x}{x!}\mathrm{e}^{-\lambda t}.$$

V našem případě $P(x) = \frac{(1{,}2)^x}{x!}\mathrm{e}^{-1{,}2}$. Nyní můžeme určit očekávané pravděpodobnosti π_{0_i}. Například: Očekávaná pravděpodobnost π_{0_1}, že během jednoho dne nedojde k žádné poruše (počet poruch bude 0) je

$$\pi_{0_1} = P(X=0) = P(0) = \frac{(1,2)^0}{0!}\mathrm{e}^{-1,2} = 0,301.$$

Obdobně:
$\pi_{0_2} = P(X=1) = P(1) = \frac{(1{,}2)^1}{1!}\mathrm{e}^{-1{,}2} = 0,361,$

$\pi_{0_3} = P(X=2) = P(2) = \frac{(1{,}2)^2}{2!}\mathrm{e}^{-1{,}2} = 0,217,$

$\pi_{0_4} = P(X=3) = P(3) = \frac{(1{,}2)^3}{3!}\mathrm{e}^{-1{,}2} = 0,087,$

$\pi_{0_5} = P(X \geqq 4) = 1 - P(X<4) = 1 - \sum_{i=0}^{3} \frac{(1{,}2)^i}{i!}\mathrm{e}^{-1{,}2} = 0,034.$

Očekávané četnosti pak určíme podle vztahu $E_i = n\pi_{0_i}$, kde n je rozsah výběru (v našem případě $n = 150$). Například: platí-li nulová hypotéza, pak by během 150 dnů v cca $E_1 = 150 \cdot 0,301 = 45,2$ dnech nemělo dojít k žádné poruše.

Tab. 9.4: 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 – **pozorovaná četnost**	52	48	36	10	4
π_{0_i} – **pozorovaná pravděpodobnost**	0,301	0,361	0,217	0,087	0,034
E_i – **očekávaná četnost**	45,2	54,2	32,6	13,1	5,1

Všechny očekávané četnosti E_i jsou větší než 5, tudíž rozsah výběru je dostatečný proto, abychom mohli použít testovou statistiku

$$G = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i}.$$

Pozorovaná hodnota $x_{OBS} = \frac{\sum_{i=1}^{5} (O_i - E_i)^2}{E_i} = \frac{(52-45,2)^2}{45,2} + \ldots + \frac{(4-5,1)^2}{5,1} = 3,13.$

Testové kritérium G má χ^2 rozdělení s $4 = (k - 1 - h)$ stupni volnosti. (Počet variant $k = 5$, počet odhadovaných parametrů $h = 0$.)

p-hodnota $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s 4 stupni volnosti.

p-$hodnota = 1 - F_0(3,13) = 0,54$ (viz vybrana_rozdeleni.xls)

p-$hodnota > 0,05$, proto nezamítáme nulovou hypotézu, tzn. nemáme námitek proti použití Poissonova rozdělení s parametrem 1,2 pro odhad počtu poruch daného zařízení během jednoho dne.

▲

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			

Ověřte čistým testem významnosti, zda lze časové odstupy mezi vozidly modelovat pomocí náhodné veličinu s normálním rozdělením.

Řešení.

Nechť je náhodná veličina X definována jako časový odstup mezi průjezdy jednotlivých vozidel.

Nulovou a alternativní hypotézu formulujeme ve tvaru:

H_0 : Časové odstupy mezi průjezdy jednotlivých vozidel **mají** normální rozdělení.

H_A : Časové odstupy mezi průjezdy jednotlivých vozidel nemají normální rozdělení.

Normální rozdělení má dva parametry: μ a σ^2. Ani jeden z nich není v nulové hypotéze specifikován, tzn. jde o **neúplně specifikovaný test** (počet odhadovaných parametrů $h = 2$).

Nejdříve pomocí výběru (o rozsahu $n = 132$) odhadneme parametry očekávaného (normálního) rozdělení. Nejlepším odhadem střední hodnoty μ je výběrový průměr $\bar{x}$, nejlepším odhadem rozptylu σ^2 je výběrový rozptyl s^2.

$$\hat{\mu} = \bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \frac{\sum\limits_{i=1}^{132} x_i}{132} = 4,6, \quad \hat{\sigma}^2 = s^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \tilde{x})^2}{n-1} = \frac{\sum\limits_{i=1}^{1} 32(x_i - 4,6)^2}{131} = 10,9$$

Ověřujeme, zda výběr pochází z rozdělení normálního, tj. spojitého, proto je třeba nejprve testované rozdělení **kategorizovat**.

Pokusíme se tedy rozdělit data do k třídících intervalů, určíme empirické četnosti O_i a najdeme očekávané pravděpodobnosti π_{0_i} pro příslušné třídící intervaly.

Poznámka:
Třídící intervaly se volí většinou pouze na základě vlastní úvahy. Jejich počet se snažíme volit v „rozumných“ mezích. Počet intervalů nemá být ani příliš malý (kategorizace spojitého rozdělení snižuje vypovídací schopnost o tomto rozdělení), ani příliš velký (čím větší počet třídících intervalů, tím menší očekávané četnosti v těchto intervalech – limitujícím předpokladem pro použití χ^2 testu dobré shody je, aby očekávané četnosti byly větší než 5). Obvykle se považuje za vhodné volit 5 až 15 třídících intervalů.

- Definiční obor náhodné veličiny rozdělíme například do 13 třídících intervalů.

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

- Empirické četnosti O_i určíme jako počet pozorování, které leží v příslušném intervalu.
- Platí-li nulová hypotéza, pak náhodná veličina X má rozdělení $N(\hat{\mu}; \hat{\sigma}^2)$, přičemž parametry tohoto rozdělení jsme odhadli. Očekávané pravděpodobnosti π_{0_i} pak určíme jako pravděpodobnosti výskytu náhodné veličiny X s rozdělením $N(\hat{\mu}; \hat{\sigma}^2)$ na příslušném intervalu.

  V našem případě: Platí-li H_0, pak $X \to N(4,6; 10,9)$.
  $P(X \in (-\infty; 1,5\rangle)) = P(X \leqq 1,5) = F(1,5) = \Phi(\frac{1{,}5-4{,}6}{\sqrt{10{,}9}}) = \Phi(-0,94) = 0,174$,
  $P(X \in (1,5; 1,8\rangle)) = P(1,5 < X \leqq 1,5) = F(1,8) - F(1,5) = \Phi(\frac{1{,}8-4{,}6}{\sqrt{10{,}9}}) - \Phi(\frac{1{,}5-4{,}6}{\sqrt{10{,}9}}) = \Phi(-0,85) - \Phi(-0,94) = 0,024$,
  atd.

Očekávané četnosti jednotlivých třídících intervalů pak určíme podle již známého vztahu $E_i = n\pi_{0_i}$, kde n je rozsah výběru (v našem případě $n = 132$).

Veškeré zjištěné hodnoty zapíšeme do tabulky.

i	Třídící interval [s]	Empirické četnosti O_i	Očekávané pravděpodobnosti $\pi_{0,i}$	Očekávané četnosti E_i
1	$(-\infty; 1{,}5\rangle$	11	0,174	22,9
2	$(1{,}5; 1{,}8\rangle$	13	0,024	3,2
3	$(1{,}8; 2{,}0\rangle$	7	0,017	2,3
4	$(2{,}0; 2{,}5\rangle$	10	0,047	6,2
5	$(2{,}5; 2{,}9\rangle$	8	0,041	5,4
6	$(2{,}9; 3{,}6\rangle$	8	0,078	10,3
7	$(3{,}6; 4{,}0\rangle$	10	0,047	6,2
8	$(4{,}0; 4{,}4\rangle$	10	0,048	6,3
9	$(4{,}4; 4{,}9\rangle$	10	0,060	8,0
10	$(4{,}9; 5{,}8\rangle$	12	0,106	14,0
11	$(5{,}8; 6{,}8\rangle$	10	0,106	13,9
12	$(6{,}8; 8{,}7\rangle$	12	0,145	19,2
13	$(8{,}7; \infty)$	11	0,107	14,1
Celkem	-	132	1,000	-

Pohledem na očekávané četnosti zjistíme, že jsme třídící intervaly zvolili poměrně dobře – pouze 2. a 3. intervalu přísluší očekávané četnosti nižší než 5 (to odporuje předpokladu pro použití χ^2 testu dobré shody). Tento nedostatek snadno napravíme tím, že tyto intervaly sloučíme.

i	Třídící interval [s]	Empirické četnosti O_i	Očekávané pravděpodobnosti $\pi_{0,i}$	Očekávané četnosti E_i
1	$(-\infty; 1{,}5\rangle$	11	0,174	22,9
2	$(1{,}5; 2{,}0\rangle$	20	0,041	5,5
3	$(2{,}0; 2{,}5\rangle$	10	0,047	6,2
4	$(2{,}5; 2{,}9\rangle$	8	0,041	5,4
5	$(2{,}9; 3{,}6\rangle$	8	0,078	10,3
6	$(3{,}6; 4{,}0\rangle$	10	0,047	6,2
7	$(4{,}0; 4{,}4\rangle$	10	0,048	6,3
8	$(4{,}4; 4{,}9\rangle$	10	0,060	8,0
9	$(4{,}9; 5{,}8\rangle$	12	0,106	14,0
10	$(5{,}8; 6{,}8\rangle$	10	0,106	13,9
11	$(6{,}8; 8{,}7\rangle$	12	0,145	19,2
12	$(8{,}7; \infty)$	11	0,107	14,1
Celkem	-	132	1,000	-

Nyní jsou předpoklady pro použití χ^2 testu dobré shody splněny. Můžeme použít testovou statistiku

$$G = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i}.$$

Pozorovaná hodnota $x_{OBS} = \frac{\sum_{i=1}^{12} (O_i - E_i)^2}{E_i} = \frac{(11-22{,}9)^2}{22{,}9} + \ldots + \frac{(11-14{,}1)^2}{14{,}1} = 59,7.$

Testové kritérium G má χ^2 rozdělení s $9 (= k - 1 - h)$ stupni volnosti. (Počet třídících intervalů $k = 12$, počet odhadovaných parametrů $h = 2$.)

p-hodnota $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s 9 stupni volnosti.

p-hodnota $= 1 - F_0(59,7) < 0,001$ (viz vybrana_rozdeleni.xls)

p-hodnota $< 0,05$, proto zamítáme nulovou hypotézu ve prospěch alternativy, tzn. časové odstupy mezi průjezdy jednotlivých vozidel nemají normální rozdělení.

▲

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.

Chceme testovat nulovou hypotézu

H_0: Výběr pochází z rozdělení $N(19; 0,49)$

vůči alternativě

H_A: $\neg H_0$, tj. výběr nepochází z rozdělení $N(19; 0,49)$.

Vzhledem k tomu, že máme k dispozici výběr pouze velmi malého rozsahu ($n = 10$), nelze použít úplně specifikovaný χ^2 test dobré shody (očekávané četnosti v třídících intervalech by nepřekročily požadovanou hodnotu 5). Jedinou možností tak je Kolmogorovův-Smirnovův test.

Testovým kritériem je náhodná veličina

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)| = \max(D_1^*, D_2^*, \dots, D_n^*),$$

kde $F_0(x) \dots$ distribuční funkce testovaného rozdělení,

$$D_i^* = \max\left\{\left|F_0(x_i) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0(x)\right|\right\} \quad \text{pro } i = 1, 2, \ldots, n.$$

Výpočty potřebné pro stanovení pozorované hodnoty jsou uvedeny v tabulce 9.5, kde $F_0(x_{(i)}) = \Phi\left(\frac{x_{(i)}-19}{\sqrt{0{,}49}}\right)$.

Tab. 9.5: Pomocné výpočty pro určení pozorované hodnoty testové statistiky D_n

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^*
17,660	1	0,00	0,10	0,03	0,07	0,03	0,07
18,727	2	0,10	0,20	0,35	0,15	0,25	0,25
19,038	3	0,20	0,30	0,52	0,22	0,32	0,32
19,105	4	0,30	0,40	0,56	0,16	0,26	0,26
19,108	5	0,40	0,50	0,56	0,06	0,16	0,16
19,234	6	0,50	0,60	0,63	0,03	0,13	0,13
19,270	7	0,60	0,70	0,65	0,05	0,15	0,15
19,473	8	0,70	0,80	0,75	0,05	0,05	0,05
19,732	9	0,80	0,90	0,85	0,05	0,05	0,05
20,219	10	0,90	1,00	0,96	0,04	0,06	0,06

Pozorovaná hodnota $x_{OBS} = 0,32$.

Kritická hodnota testové statistiky $D_{10(0,05)} = 0,40925$.

Pozorovaná hodnota $x_{OBS} = 0,32$ je menší než kritická hodnota $D_{10(0,05)} = 0,40925$, proto nezamítáme nulovou hypotézu, tzn. nelze tvrdit, že získaná data nepodléhají rozdělení

$N(19; 0,49)$.

▲

Kapitola 10

Analýza závislosti - řešené příklady

Příklad 10.1. 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.

H_0 : Spokojenost v práci **nesouvisí** s umístěním závodu.
H_A : Spokojenost v práci **souvisí** s umístěním závodu.

Pro test nezávislosti v kontingenční tabulce lze v případě splnění podmínek dobré aproximace použít χ^2 test nezávislosti. Nutno ověřit, zda očekávané četnosti neklesly pod 2 a zda alespoň 80 % z nich je větších než 5.

Nejdříve si tedy pomocí rozšířené kontingenční tabulky určíme očekávané četnosti. Očekávané četnosti E_{ij} určujeme jako četnosti odpovídající součinu příslušných marginálních relativních četností.

$$E_{ij} = \left(\frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}\right) \cdot n = \frac{n_{i.} \cdot n_{.j}}{n}$$

Všechny očekávané četnosti jsou větší než 5 (viz tabulka 10.1), podmínky dobré aproximace lze tedy považovat za splněné.

Tab. 10.1: Kontingenční tabulka rozšířená o marginální a očekávané četnosti

místo\stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen	celkem
Praha	10 10,00 $\left(\frac{100\cdot 30}{300}\right)$	25 11,67 $\left(\frac{100\cdot 35}{300}\right)$	50 60,00 $\left(\frac{100\cdot 180}{300}\right)$	15 18,33 $\left(\frac{100\cdot 55}{300}\right)$	100
venkov	20 20,00 $\left(\frac{200\cdot 30}{300}\right)$	10 23,33 $\left(\frac{200\cdot 35}{300}\right)$	130 120,00 $\left(\frac{200\cdot 180}{300}\right)$	40 36,67 $\left(\frac{200\cdot 55}{300}\right)$	200
celkem	30	35	180	55	300

Pozorovaná hodnota testové statistiky K

$$x_{OBS} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(10-10,00)^2}{10,00} + \frac{(20-20,00)^2}{20,00} + \cdots +$$

$$+ \frac{(40-36,67)^2}{36,67} = 27,0.$$

Podmínky dobré aproximace jsou splněny, proto

$$p\text{-}hodnota = 1 - F_0\left(x_{OBS}\right),$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r-1)(s-1) = (2-1)(4-1) = 3$ stupni volnosti.

$p\text{-}hodnota = 1 - F_0(27,0)' < 0,001$ (viz vybrana_rozdeleni.xls)

$p\text{-}hodnota < 0,05$, proto zamítáme nulovou hypotézu ve prospěch alternativy, tj. spokojenost v práci souvisí s umístěním závodu. (Uvědomte si, že test nijak neověřoval kauzalitu závislosti!)

Zbývá určit, jaká je těsnost identifikované závislosti. Vzhledem k tomu, že analyzujeme obdélníkovou tabulku ($r = 2; s = 4$), můžeme použít korigovaný koeficient kontingence nebo Cramerův koeficient.

$$CC = \sqrt{\frac{K}{K+n}} = \frac{27,0}{27,0+300} = 0,287;$$

$$CC_{max} = \sqrt{\frac{\min(r;s)-1}{\min(r;s)}} = \sqrt{\frac{2-1}{2}} = 0,707;$$

$$CC_{cor} = \frac{CC}{CC_{max}} = 0,406;$$

$$V = \sqrt{\frac{K}{n\left(\min(r;s) - 1\right)}} = \sqrt{\frac{27,0}{300(2-1)}} = 0,3$$

Jak podle koeficientu kontingence, tak podle Cramerova koeficientu lze závislost mezi umístěním závodu a stupněm spokojenosti v práci označit za silnou.

▲

Příklad 10.2. 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.

Data jsou zapsána v asociační tabulce, proto je vhodné použít speciální metody určené pro analýzu asociací.

Odhad šance novorozeneckého úmrtí u dětí s nízkou porodní váhou je

$$\frac{a}{b} = \frac{618}{4597} = 0,134,$$

což odpovídá přibližně 134 novorozeneckým úmrtím na 1 000 přeživších novorozenců s nízkou porodní váhou. Obdobně odhadneme šanci novorozeneckého úmrtí u dětí s normální porodní váhou.

$$\frac{c}{d} = \frac{422}{67093} = 0,006$$

Lze očekávat přibližně 6 novorozeneckých úmrtí na 1 000 přeživších novorozenců s normální porodní hmotností.

Odhadněme poměr šancí novorozeneckého úmrtí u dětí s nízkou a normální porodní váhou.

$$\widehat{OR} = \frac{ad}{bc} = \frac{618 \cdot 67093}{4597 \cdot 422} \cong 21,4$$

Odhad udává, že šance novorozeneckého úmrtí je 21,4 krát vyšší u novorozenců s nízkou porodní váhou než u novorozenců s normální porodní váhou.

95% intervalový odhad OR je dán vztahem

$$\left\langle \widehat{OR} \cdot \mathrm{e}^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{0,975}}; \widehat{OR} \cdot \mathrm{e}^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{0,975}} \right\rangle.$$

$z_{0,975} = 1,64$ (viz vybrana_rozdeleni.xls)

Po dosazení: 95% intervalový odhad OR je $\langle 19,2; 23,8 \rangle$. Je zcela zřejmé, že šance novorozeneckého úmrtí závisí na porodní váze ($1 \notin \langle 19,2; 23,8 \rangle$).

Jiným přístupem je analyzovat asociaci pomocí relativního rizika.

Odhad absolutního rizika novorozeneckého úmrtí u dětí s nízkou porodní hmotností je $\frac{a}{a+b} = \frac{618}{5215} = 0,119$ (tj. novorozenecké úmrtí lze očekávat u cca 119 z 1 000 novorozenců s nízkou porodní váhou), u dětí s normální porodní hmotností $\frac{c}{c+d} = \frac{422}{67515} = 0,006$ (tj. novorozenecké úmrtí lze očekávat u cca 6 z 1 000 novorozenců s normální porodní váhou).

Odhad relativního rizika novorozeneckého úmrtí

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)} = \frac{0,119}{0,006} = 19,0.$$

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

Tento výsledek ukazuje, že ve sledovaném období bylo u dětí s nízkou porodní váhou 19 krát vyšší riziko novorozeneckého úmrtí než u dětí s normální porodní váhou.

95% intervalový odhad RR je dán vztahem

$$\left\langle \widehat{RR} \cdot \mathrm{e}^{-\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}} \cdot z_{0,975}}; \widehat{RR} \cdot \mathrm{e}^{\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}} \cdot z_{0,975}} \right\rangle .$$

$z_{0,975} = 1,64$ (viz vybrana_rozdeleni.xls)

Po dosazení: 95% intervalový odhad RR je $\langle 17,1; 21,0 \rangle$. Je zcela zřejmé, že riziko novorozeneckého úmrtí závisí na porodní váze ($1 \notin \langle 17,1; 21,0 \rangle$).

▲

Příklad 10.3. 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.

Nejdříve je nutné ověřit, zda výběr, který máme k dispozici, pochází z dvourozměrného normálního rozdělení. Jak bylo zmíněno, v praxi většinou zcela vyhovuje, omezíme-li se pouze na ověření normality rozdělení obou sledovaných veličin X a Y. Pro ověření normality použijeme Kolmogorovův-Smirnovův test používající modifikované kritické hodnoty implementovaný v softwaru Statgraphics.

H_0 : Výběr z náh. veličiny X, resp. Y, pochází z normálního rozdělení.
H_A : Výběr z náh. veličiny X, resp. Y, nepochází z normálního rozdělení.

$$p\text{-}hodnota_X > 0,10, \text{ resp. } p\text{-}hodnota_Y > 0,10 \,(\text{dle Statgraphics})$$

Na hladině významnosti 0,05 nelze zamítnout nulovou hypotézu, že výběr z náh. veličiny X, resp. Y, pochází z normálního rozdělení.

Jak již víme, ve sdruženém normálním rozdělení je nekorelovanost ekvivalentní nezávislosti. Chceme tedy testovat hypotézu

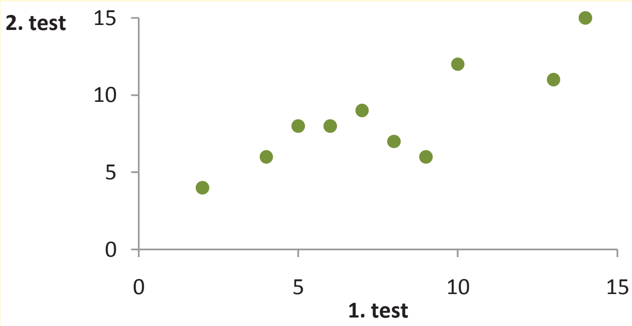
$$H_0: \quad \rho = 0, \text{ tj. výsledky 1. a 2. zápočtového testu jsou nezávislé.}$$

vůči alternativě

$$H_A: \quad \rho > 0, \text{ tj. výsledky 1. a 2. zápočtového testu jsou kladně korelované.}$$

Nejdříve určíme výběrový korelační koeficient r.

Obr. 10.1: Korelační pole pro výsledky 1. a 2. testu



$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 7,8; \quad \overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i = 8,6;$$

$$S_X^2 = \frac{1}{9}\sum_{i=1}^{10}\left(X_i - \overline{X}\right)^2 = \frac{131{,}6}{9} = 14,6; \quad S_Y^2 = \frac{1}{9}\sum_{i=1}^{10}\left(Y_i - \overline{Y}\right)^2 = \frac{96{,}4}{9} = 10,7;$$

$$S_{XY} = \frac{1}{9}\sum_{i=1} 10\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \frac{95{,}2}{9} = 10,6$$

Tab. 10.2: Pomocné výpočty pro určení výběrového korelačního koeficientu r

											součet
X_i (1. test)	7	8	10	4	14	9	6	2	13	5	78
Y_i (2. test)	9	7	12	6	15	6	8	4	11	8	86
$(X_i-\bar{X})^2$	0,64	0,04	4,84	14,44	38,44	1,44	3,24	33,64	27,04	7,84	131,6
$(Y_i-\bar{Y})^2$	0,16	2,56	11,56	6,76	40,96	6,76	0,36	21,16	5,76	0,36	96,4
X_iY_i	63	56	120	24	210	54	48	8	143	40	766
$(X_i-\bar{X})(Y_i-\bar{Y})$	-0,32	-0,32	7,48	9,88	39,68	-3,12	1,08	26,68	12,48	1,68	95,2

$$r = \begin{cases} \frac{S_{X,Y}}{\sqrt{S_X^2 \cdot S_Y^2}} & S_X^2, S_Y^2 \neq 0, \\ 0 & \text{jinak.} \end{cases}$$

$r = 0,845$

Jak je zřejmé, na základě bodového grafu a hodnoty výběrového korelačního koeficientu lze očekávat zamítnutí nulové hypotézy.

Pozorovaná hodnota $x_{OBS} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 4,47$.

Vzhledem k tvaru alternativy: *p-hodnota* $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce Studentova rozdělení s $n-2=8$ stupni volnosti.

$p - hodnota = 1 - F_0(4,47) = 0,001$

Na hladině významnosti 0,05 zamítáme nulovou hypotézu ve prospěch alternativy, tj. výsledek 1. a 2. zápočtového testu je kladně korelovaný.

▲

Příklad 10.4. V tabulce 10.3 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. 10.3: 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.

Označme:
X ...spotřeba alkoholu,
Y ...úmrtnost na cirhózu jater.

Chceme testovat:

H_0 : X, Y jsou **nezávislé** náhodné veličiny.
H_A : X, Y jsou **závislé** náhodné veličiny.

Nejdříve ověříme, zda náhodný výběr pochází z dvourozměrného normálního rozdělení. Nutnou podmínkou tohoto předpokladu je, aby náhodná veličina X i náhodná veličina Y měly normální rozdělení. K ověření těchto podmínek jsme použili v softwaru Statgraphics aplikovaný χ^2 test dobré shody.

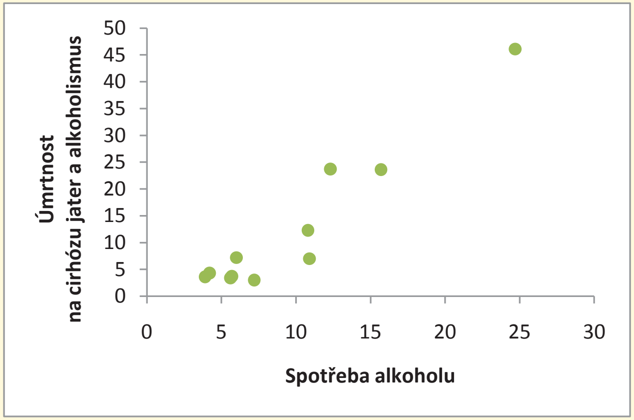
$p\text{-}hodnota_X = 0,336$, $p\text{-}hodnota_Y = 0,001$ (dle Statgraphics)

Je zřejmé, že na hladině významnosti 0,05 lze zamítnout normalitu náhodné veličiny Y (tj. úmrtnosti na cirhózu jater a alkoholismus). Jako míru korelace mezi spotřebou alkoholu a úmrtnosti na cirhózu jater a alkoholismus proto volíme Spearmanův koeficient korelace.

Tabulku 10.4 rozšíříme o pořadí veličin X_i a Y_i, jejich diference a kvadráty diferencí.

Tab. 10.4: Pomocné výpočty pro výpočet Spearmanova korelačního koeficientu

země	X_i	Y_i	R_{X_i}	R_{Y_i}	$R_{X_i} - R_{Y_i}$	$(R_{X_i} - R_{Y_i})^2$
Finsko	3,9	3,6	1	3	-2	4
Norsko	4,2	4,3	2	5	-3	9
Irsko	5,6	3,4	3	2	1	1
Holandsko	5,7	3,7	4	4	0	0
Švédsko	6,0	7,2	5	7	-2	4
Anglie	7,2	3,0	6	1	5	25
Belgie	10,8	12,3	7	8	-1	1
Rakousko	10,9	7,0	8	6	2	4
SRN	12,3	23,7	9	10	-1	1
Itálie	15,7	23,6	10	9	1	1
Francie	24,7	46,1	11	11	0	0
Součet	-	-	-	-	-	50



$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2 = 1 - \frac{6}{11(11^2-1)} \cdot 50 = 0,773$$

Kritická hodnota $r_S^*(0,05;11) = 0,6091$ (viz tabulka T15).

$|r_S| \geqq r_S^*(0,05;11)$, proto na hladině významnosti 0,05 zamítáme nulovou hypotézu, že spotřeba alkoholu a úmrtnost na cirhózu jater a alkoholismus jsou nezávislé veličiny.

Poznámka: *Všimněte si, že nesprávně použitý Pearsonův výběrový korelační koeficient* $(r = 0,956)$ *by ukazoval na mnohem těsnější závislost.*

▲

Kapitola 11

Úvod do korelační a regresní analýzy - řešené příklady

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 **??**.

Řešení. Hledáme odhad regresní přímky ve tvaru $\widehat{Y} = b_0 + b_1 x$. Ukázali jsme si, že odhady regresních koeficientů určíme dle

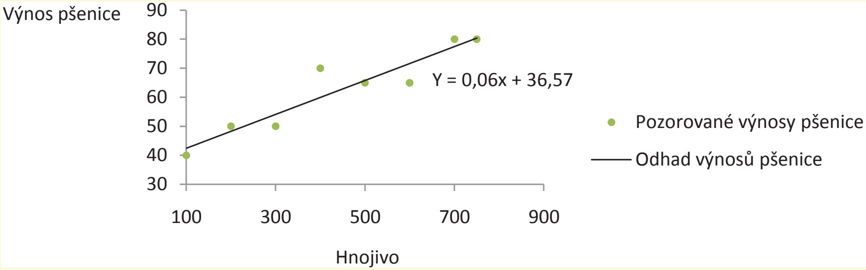
$$b_1 = \frac{n \sum\limits_{i=1}^{n} y_i x_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i}{n \sum\limits_{i=1}^{n} (x_i)^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2}, \qquad b_0 = \overline{y} - b_1 \overline{x}.$$

Pomocné výpočty uvádíme v tabulce 11.1.

Tab. 11.1: Pomocné výpočty pro výpočet odhadu regresních koeficientů

ident. číslo	y- výnos pšenice [t/ha]	x – hnojivo [kg/ha]	yx	x^2
1	40	100	4 000	10 000
2	50	200	10 000	40 000
3	50	300	15 000	90 000
4	70	400	28 000	160 000
5	65	500	32 500	250 000
6	65	600	39 000	360 000
7	80	700	56 000	490 000
8	80	750	60 000	562 500
Celkem	500	3 550	244 500	1 962 500

Po dosazení: $b_1 = 0,06, b_0 = 36,57.$



Pokud jsou splněny předpoklady lineárního regresního modelu, můžeme výnosy pšenice odhadovat na základě množství použitého hnojiva pomocí funkce $\widehat{Y} = 36,57 + 0,06x$. (Ověření předpokladů se budeme věnovat v kapitole **??**.)

▲

Příklad 11.2. Proveďte odhad koeficientů regresní přímky z řešeného příkladu pomocí maticového zápisu.

Řešení.

Hledáme odhad regresní přímky ve tvaru

$$\widehat{Y} = b_0 + b_1 x, \text{ tj. } \widehat{\boldsymbol{Y}} = \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \boldsymbol{F}\boldsymbol{b}.$$

Potřebné údaje zjistíme v tabulce 11.2.

$\overline{x} = \frac{3550}{8} = 443,75 \quad n = 8,$

$$\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{x^{-2}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \\ \frac{-\overline{x}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} & \frac{1}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \end{bmatrix} = \begin{bmatrix} 0,634 & -0,001 \\ -0,001 & 2,58 \cdot 10^{-6} \end{bmatrix},$$

$$\boldsymbol{F}^T\boldsymbol{y} = \begin{bmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 500 \\ 244500 \end{bmatrix},$$

Obsah

145. strana ze 159

Zavřít dokument

Celá obrazovka/Okno

Tab. 11.2: Pomocné výpočty pro odhad koeficientů regresní přímky pomocí maticového zápisu

ident. číslo	y- výnos pšenice [t/ha]	x – hnojivo [kg/ha]	xy	$x-\bar{x}$	$(x-\bar{x})^2$
1	40	100	4000	-343,75	118164,1
2	50	200	10000	-243,75	59414,06
3	50	300	15000	-143,75	20664,06
4	70	400	28000	-43,75	1914,063
5	65	500	32500	56,25	3164,063
6	65	600	39000	156,25	24414,06
7	80	700	56000	256,25	65664,06
8	80	750	60000	306,25	93789,06
Celkem	500	3 550	244500		387187,5

$$\boldsymbol{b} = \left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1}\boldsymbol{F}^T\boldsymbol{y} = \begin{bmatrix} 0,634 & -0,001 \\ -0,001 & 2,58\cdot 10^{-6} \end{bmatrix} \begin{bmatrix} 500 \\ 244500 \end{bmatrix} = \begin{bmatrix} 36,57 \\ 0,06 \end{bmatrix}.$$

Vyrovnávací přímka má tedy tvar $\widehat{Y} = 36,57 + 0,06x$, což je výsledek shodný s výsledkem získaným řešením bez použití maticového zápisu.

▲

Příklad 11.3. Pomocí celkového F-testu ověřte, zda lze výnosy pšenice odhadovat pomocí lineární závislosti na množství použitého hnojiva.

Řešení.

Regresní funkce obsahuje pouze jeden regresor, proto chceme testovat nulovou hypotézu

$$H_0: \quad \beta_1 = 0$$

proti alternativě

$$H_A: \quad \beta_1 \neq 0$$

Pomocné výpočty pro součet čtverců modelu $SS_{\widehat{Y}}$ a reziduální součet čtverců SS_e zaznamenáme do tabulky. $\left(\overline{y} = \frac{500}{8} = 62,5\right)$

$$\begin{aligned}
SS_{\widehat{Y}} &= 1322,19; \quad SS_\mathrm{e} = 177,93; \quad SS_Y = SS_{\widehat{Y}} + SS_\mathrm{e} = 1500,12;\\
\frac{SS_{\widehat{Y}}}{k} &= \frac{1322,19}{1} = 1322,19; \quad \frac{SS_\mathrm{e}}{n-(k+1)} = \frac{177,93}{8-(1+1)} = 29,66;\\
x_{OBS} &= \frac{\frac{SS_{\widehat{Y}}}{k}}{\frac{SS_\mathrm{e}}{n-(k+1)}} = \frac{1322,19}{29,66} = 44,59; \quad p - hodnota = 1 - F_0(44,59) = 0,0005;
\end{aligned}$$

kde $F_0(x)$ je distribuční funkce Fisherovo-Snedecorovo rozdělení s 1 stupněm volnosti v čitateli a 6 stupni volnosti ve jmenovateli.

(Pro výpočet p-hodnoty byl použít applet vybrana_rozdeleni.xls.)

Obsah

147. strana ze 159

Zavřít dokument

Celá obrazovka/Okno

Tab. 11.3: Pomocné výpočty pro konstrukci celkového F-testu

ident. číslo	y- výnos pšenice [t/ha]	x – hnojivo [kg/ha]	$\hat{Y} = 36{,}57 + 0{,}06x$	$\hat{Y} - \bar{y}$	$(\hat{Y} - \bar{y})^2$	$e = y - \hat{Y}$	e^2
1	40	100	42,41	-20,09	403,61	-2,41	5,82
2	50	200	48,26	-14,24	202,78	1,74	3,04
3	50	300	54,10	-8,40	70,56	-4,10	16,81
4	70	400	59,94	-2,56	6,55	10,06	101,13
5	65	500	65,79	3,29	10,82	-0,79	0,62
6	65	600	71,63	9,13	83,36	-6,63	43,96
7	80	700	77,47	14,97	224,10	2,53	6,38
8	80	750	80,40	17,90	320,41	-0,40	0,16
Celkem	500	---	---		1322,19	---	177,93

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Rozptyl (prům. součet čtverců)	x_{OBS}	$p - hodnota$
Model	1 322,19	1	1 322,19	44,59	0,0005
Reziduální	177,93	6	29,66	---	---
Celkový	1 500,12	7	---	---	---

Na hladině významnosti 0,05 lze zamítnout nulovou hypotézu, zvolený model je statisticky významný.

▲

Příklad 11.4. Určete směrodatné odchylky parametrů b_0 a b_1 regresní přímky z řešeného příkladu 11.2.

Řešení.

V řešeném příkladu **??** jsme našli odhad regresní přímky ve tvaru $\widehat{Y} = 36,57 + 0,06x$.

Směrodatné odchylky parametrů b_0 a b_1 regresní přímky jsou dány předpisem

$$s_{b_i} = s_{\mathrm{e}}\sqrt{x_{i+1,i+1}}.$$

Rozptyl náhodné složky

$$s_{\mathrm{e}}^2 = \frac{\sum\limits_{i=1}^{n} \mathrm{e}_i^2}{n-(k+1)}$$

jsme určili již v řešeném příkladu **??**.

$$s_{\mathrm{e}}^2 = 29,66, s_{\mathrm{e}} = 5,446$$

Z řešeného příkladu **??** víme, že

$$\left(\boldsymbol{F}^T\boldsymbol{F}\right)^{-1} = \begin{bmatrix} 0,634 & -0,001 \\ -0,001 & 2,58 \cdot 10^{-6} \end{bmatrix}.$$

Nyní můžeme určit směrodatné odchylky odhadů.

$$\begin{aligned} s_{b_0} &= s_{\mathrm{e}}\sqrt{x_{1,1}} = 5,446 \cdot \sqrt{0,634} = 4,336 \\ s_{b_1} &= s_{\mathrm{e}}\sqrt{x_{2,2}} = 5,446 \cdot \sqrt{2,58 \cdot 10^{-6}} = 0,009 \end{aligned}$$

Obsah

149. strana ze 159

Zavřít dokument

Celá obrazovka/Okno

Je zřejmé, že čím větší je směrodatná odchylka s_{b_i} vzhledem k bodovému odhadu b_i regresního koeficientu, tím je tento odhad méně spolehlivý. (Srovnejte s_{b_i} a b_i.)

▲

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.

V předcházejících řešených příkladech jsme nalezli odhad regresní přímky ve tvaru

$$\widehat{Y} = 36,57 + 0,06x,$$

tj. $b_0 = 36,57, b_1 = 0,06$

Směrodatné odchylky odhadů jsou $s_{b_0} = 4,336, s_{b_1} = 0,009$.

$100\,(1-\alpha)$ % 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 $C-(k+1)$ stupni volnosti.

V našem případě $\alpha = 0,05$, počet pozorování $n = 8$, počet regresorů (nezávisle proměnných) $k = 1$. Pak $t_{0,975} = 2,45$ (viz vybrana_rozdeleni.xls, 0,975 kvantil Studentova rozdělení s 6 stupni volnosti).

Po dosazení do vzorce pro intervalový odhad koeficientu β_i dostaneme:

- 95 % intervalový odhad koeficientu β_0 je $\langle 25,95; 47,19 \rangle$,

- 95 % intervalový odhad koeficientu β_1 je $\langle 0,04; 0,08\rangle$.

Dílčí t testy

$$\begin{aligned} H_0: &\quad \beta_0 = 0 \\ H_A: &\quad \beta_0 \neq 0 \end{aligned}$$

$$x_{OBS} = \left.\frac{b_0-\beta_0}{s_{b_0}}\right|_{H_0} = \frac{36{,}57-0}{4{,}336} = 8,43$$

$$p - hodnota = 2\,\min\{F_0\,(x_{OBS})\,; 1 - F_0\,(x_{OBS})\},$$

kde $F_0(x)$ je distribuční funkce Studentova rozdělení s 6 stupni volnosti.

$$p - hodnota \doteq 0,002$$

Na hladině významnosti 0,05 zamítáme nulovou hypotézu, parametr β_0 je statisticky významný, nelze jej z modelu vypustit.

$$\begin{aligned} H_0: &\quad \beta_1 = 0 \\ H_A: &\quad \beta_1 \neq 0 \end{aligned}$$

$$x_{OBS} = \left.\frac{b_1-\beta_1}{s_{b_1}}\right|_{H_0} = \frac{0{,}06-0}{0{,}009} = 6,67$$

$$p - hodnota = 2\,\min\{F_0\,(x_{OBS})\,; 1 - F_0\,(x_{OBS})\},$$

kde $F_0(x)$ je distribuční funkce Studentova rozdělení s 6 stupni volnosti.

$$p - hodnota \doteq 0,005$$

Na hladině významnosti 0,05 zamítáme nulovou hypotézu, parametr β_1 je statisticky významný, nelze jej z modelu vypustit. (Všimněte si, že oba dílčí t testy jsme mohli provést rovněž pomocí nalezených intervalových odhadů.)

▲

Příklad 11.6. Proveďte analýzu reziduí pro model z řešeného příkladu 11.1.

Řešení.

Rezidua verifikovaného modelu jsou vypočtena například v tabulce . Pro jejich testování využijeme statistický software Statgraphics v.5.0. Nejdříve ověříme normalitu reziduí.

H_0 : Rezidua mají normální rozdělení.
H_A : Rezidua nemají normální rozdělení.

$p - hodnota > 0,10$ (modifikovaný Kolmogorovův-Smirnovův test, Statgraphics)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, předpoklad o normalitě reziduí můžeme považovat za splněný.

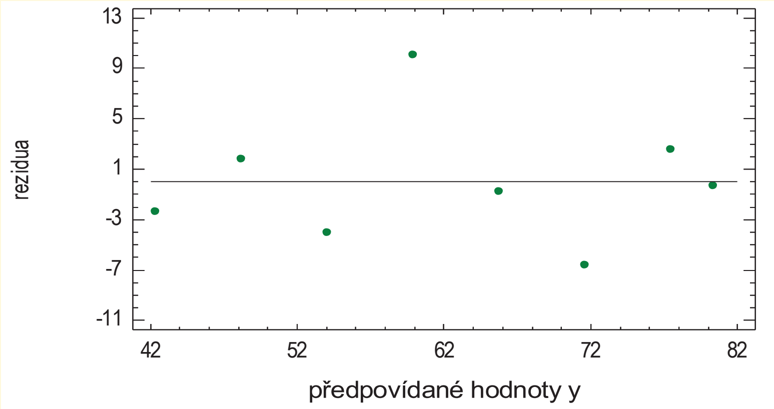
Nyní můžeme pro ověření nulovosti střední hodnoty reziduí použít jednovýběrový t test.

$$H_0 : \quad E\left(\mathrm{e}_i\right) = 0$$
$$H_A : \quad E\left(\mathrm{e}_i\right) 0$$

$p - hodnota \doteq 1,0$ (Statgraphics)

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, předpoklad o nulovosti střední hodnoty reziduí můžeme považovat za splněný.

Pro orientační vyhodnocení homoskedasticity a autokorelace reziduí použijeme graf reziduí a předpovídaných hodnot závislé proměnné.



Rezidua jsou náhodně rozmístěna kolem nuly a nemají žádný zřejmý vztah k předpovídaným hodnotám: ani se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími předpovídanými hodnotami a není zde ani náznak nelineárního vztahu.

Předpoklad homoskedasticity reziduí tedy považujeme za splněný. Předpoklad o nekorelovanosti reziduí ověříme alespoň orientačně pomocí Durbinovy-Watsonovy statistiky.

$D_W = 2,79$

Protože pozorovaná hodnota statistiky D_W překročila hodnotu 2,6, musíme označit rezidua za slabě záporně korelovaná. Autokorelace může být zapříčiněna chybnou specifikací modelu, měli bychom uvažovat o zařazení dalších vysvětlujících proměnných do modelu.

Pozor! *Porušení předpokladů může způsobit vychýlenost odhadů rozptylů regresních koeficientů a tím i chybné určení intervalových odhadů regresních koeficientů.*

▲

Příklad 11.7. Pomocí indexu determinace, resp. modifikovaného indexu determinace, určete kvalitu modelu nalezeného v řešeném příkladu **??**.

Řešení.

V Tabulce Anova, kterou jsme získali jako součást řešení příkladu **??**, nalezneme jak celkový, tak i reziduální součet čtverců.

$SS_{\mathrm{e}} = 177,93; \quad SS_Y = 1500,12; \quad n = 8; \quad k = 1$

Pak index determinace $R^2 = 1 - \frac{SS_{\mathrm{e}}}{SS_Y} = 0,881$ a modifikovaný index determinace $R^2_{adj} = = 1 - \frac{n-1}{n-(k+1)}\left(1 - R^2\right) = 0,862.$

Model vysvětluje více než 86 % celkového rozptylu závisle proměnné, proto jej lze označit za velmi kvalitní.

▲

Příklad 11.8. S využitím odhadu regresního modelu (řešený příklad **??**) 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.

a) Pro odhad středního výnosu pšenice na polích, na nichž bylo použito 350 [kg/ha] hnojiva použijeme předpis pro intervalový odhad střední hodnoty závisle proměnné.

$$\left\langle (b_0 + b_1 x_0) - t_{1-\frac{\alpha}{2}} s_\mathrm{e} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}} ; (b_0 + b_1 x_0) + t_{1-\frac{\alpha}{2}} s_\mathrm{e} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}} \right\rangle ,$$

kde $t_{1-\frac{\alpha}{2}}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n - 2$ stupni volnosti.

Hledáme 95 % intervalový odhad v $x_0 = 350$ [kg/ha], proto určíme 0,975 kvantil Studentova rozdělení s $6(= 8 - 2)$ stupni volnosti.

$t_{0,975} = 2,45$ (dle vybrana_rozdeleii.xls)

Další potřebné údaje zjistíme z předcházejících řešených příkladů.

$n = 8, b_0 = 36,57, b_1 = 0,06$ (příklad **??**), $s_\mathrm{e} = 5,446$ (příklad **??**),
$\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 = 387187,5$ (Tab. 11.2)

Po dosazení do předpisu pro intervalový odhad střední hodnoty závisle proměnné zjistíme, že

$$P\left(E\left(Y|\, x_0\right) \in \langle 51,9; 62,1\rangle\right) = 0,95.$$

Se spolehlivostí 0,95 lze očekávat střední výnos pšenice na polích hnojených 350 [kg/ha] v intervalu $\langle 51,9; 62,1\rangle$ [t/ha].

b) Pro odhad výnosu pšenice na poli pana Nováka, který použil 350 [kg/ha] hnojiva, použijeme předpis pro intervalový odhad individuální hodnoty závisle proměnné.

$$\left\langle (b_0 + b_1 x_0) - t_{1-\frac{\alpha}{2}} s_\mathrm{e} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}}; (b_0 + b_1 x_0) + t_{1-\frac{\alpha}{2}} s_\mathrm{e} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}} \right\rangle,$$

kde $t_{1-\frac{\alpha}{2}}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n-2$ stupni volnosti.

Po dosazení údajů uvedených v řešení otázky a) dostaneme
$P\left(E\left(Y|\, x_0\right) \in \langle 42,7; 71,3\rangle\right) = 0,95.$

Se spolehlivostí 0,95 lze výnos pšenice na poli pana Nováka očekávat v intervalu $\langle 42,7; 71,3\rangle$ [t/ha]. Vzhledem k tomu, že odhad regresního modelu byl verifikován (celkový F-test, dílčí t-testy, analýza reziduí) a oba odhady jsou interpolací, lze nalezené odhady považovat za důvěryhodné.

▲