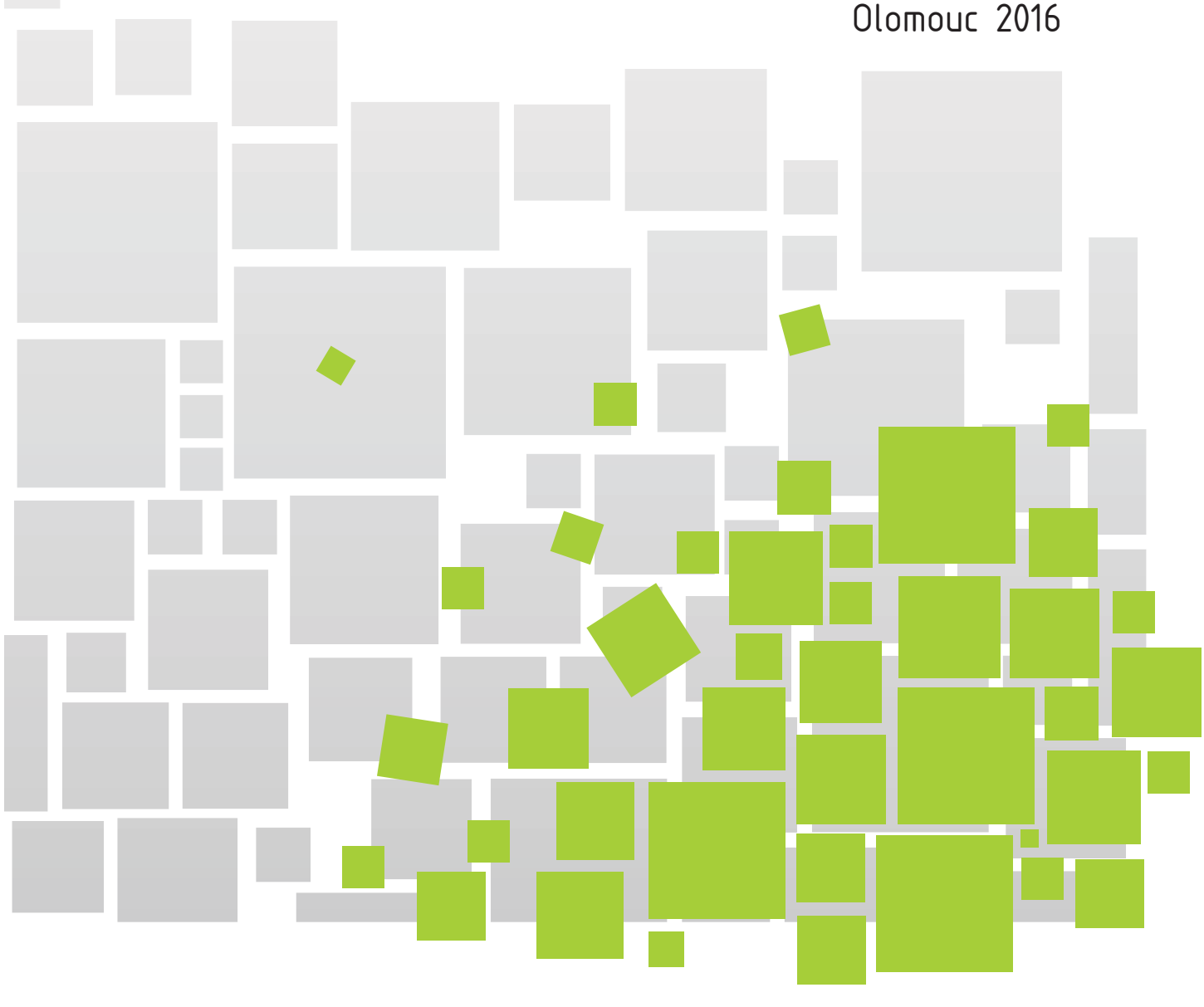
Fakulta
zdravotnických věd

Univerzita Palackého
v Olomouci

STATISTIKA PRO NELÉKAŘSKÉ ZDRAVOTNICKÉ OBORY

Eva Reiterová

Olomouc 2016



Fakulta zdravotnických věd
Univerzita Palackého v Olomouci

Statistika pro nelékařské zdravotnické obory

Eva Reiterová

Olomouc 2016

Oponenti:
PhDr. Martin Šamaj, MBA
doc. PhDr. Panajotis Cakirpaloglu, Dr.Sc.

Text neprošel jazykovou korekturou. Za obsahovou, jazykovou a stylistickou správnost odpovídá autor.

Studijní text byl zpracován s podporou rozvojových projektů v rámci institucionálního plánu Univerzity Palackého v Olomouci, projektu FRUP_2016_043 Inovace předmětů vědy a výzkumu, odborných ošetřovatelských předmětů a předmětů organizace a řízení ve studijním oboru Ošetřovatelská péče v interních oborech, včetně přípravy výukových materiálů.

ISBN 978-80-244-5082-7 (online: PDF)
DOI: 10.5507/fzv.16.24450827

První vydání

Obsah

Význam použitých ikon

Studijní cíle kapitoly 

Klíčová slova kapitoly 

Výklad – prezentace učiva 

Příklad 

Kontrolní otázky a úkoly 

Klíč k otázkám a úkolům 

Referenční seznam ke kapitole 

Úvod – z historie statistiky

Slovo „statistika" je odvozeno z latinského slova *status*, z něhož vzniklo italské slovo *statista*, které bylo poprvé použito v 16. století. Označovalo člověka, který se zabýval státními záležitostmi. Statistika tedy na počátku byla kvantitativním systémem sloužícím k popisu státních záležitostí a říkalo se jí „politická aritmetika". Poprvé ji použili v 17. století v Anglii londýnský obchodník John Graunt a irský přírodovědec William Petty. Na ně navázal skotský velkostatkář John Sinclair, který sepsal Statistickou zprávu o Skotsku. V ní se zabýval statistikou sociálních jevů a demografickými záležitostmi.

Ke konci 19. století se ze statistiky stává plnohodnotný akademický obor. Zabývá se shromažďováním, klasifikací, popisem a interpretací dat získaných při sociálních průzkumech, vědeckých experimentech a klinických zkouškách. Využívají se individuální odlišnosti ve skupině zachycením variability pomocí variačního rozpětí, rozptylu a směrodatné odchylky. Začínají se používat statistické testy významnosti k testování hypotéz. Matematická statistika má analytický charakter a může sloužit ke statistickým předpovědím.

V roce 1851 bylo v Anglii zahájeno první úplné sčítání lidu. Zaznamenával se věk, pohlaví, zaměstnání a místo narození, zjišťoval se i počet lidí slepých a hluchých. Toto první sčítání lidu přineslo podrobné informace o úmrtích na konkrétní nemoci a ukázalo špatné hygienické podmínky v přelidněných městech. Za hlavní problém Anglie se začal považovat hygienický stav měst. Za pomocí statistických údajů začaly v tomto období v Anglii rozsáhlé sanitární reformy a díky jejich úspěchům vzrostlo povědomí o významu shromažďování statistických dat. „Dáma s lampou" se říkalo britské ošetřovatelce Florence Nightingaleové, vášnivé statističce. Její zásluhou se z ošetřování nemocných stalo uznávané a vážené povolání. Za pomoci statistických výsledků se jí podařilo prosadit sanitární reformy a tím zlepšit poměry ve válečných nemocnicích a snížit úmrtnost pacientů. Za Krymské války se stala „vrchní inspektorkou zdravotních sester anglických všeobecných vojenských nemocnic v Turecku".

Úřední shromažďování velkého množství statistických dat umožnilo britským statistikům zacílit statistický systém tak, aby měřil zdraví obyvatelstva. Díky tomu pak došlo k politickým reformám a k přijetí zákonů o veřejné zdravotní péči.

Statistické metody v dnešním pojetí se původně vyvinuly pro výzkum v přírodních vědách. Ve stále větším rozsahu však vstupují i do metodiky zdravotnických a lékařských věd. Od začátku 20. století se statistika dostává stále více do oblasti medicíny, ekonomie i politiky a stává se součástí každodenního života. Statistické informace mají vliv na životy lidí, ovlivňují lékařské postupy a jiná důležitá rozhodnutí.

Znalost statistiky se dnes vyžaduje od všech pracovníků v těchto vědeckých oborech, kteří se podílejí na výzkumu. Zdravotníci a lékaři potřebují statistické metody k vyhodnocování svých empirických zkoumání. Jejich cílem je induktivně ověřit oprávněnost formulovaných hypotéz. Statistika jim umožňuje nejen plánování, ale také hodnocení vědeckých výzkumů a poskytuje postupy pro kvalitativní a kvantitativní popis nálezů zkoumání. Každý student by měl mít alespoň základní přehled a znalosti statistických metod, které jsou obsaženy v této příručce. V první části je důležité zopakovat a osvojit si základní matematické poučky a pojmy, se kterými se pak dále pracuje. Po této úvodní části začíná vlastní statistika, která je rozdělena na popisnou statistiku a na statistiku induktivní, která v sobě zahrnuje testování statistických hypotéz. Jsou zde obsaženy nejdůležitější statistické testy, parametrické i neparametrické, které studenti využijí při zpracovávání svých výzkumů. Na příkladech je ukázán praktický postup použití těchto testů a analýza dat v programu Microsoft Excel.

Doufám, že studentům tato příručka pomůže k lepšímu pochopení statistických metod a že ji využijí při zpracovávání výsledků svých diplomových prací.

1 Základní matematické pojmy a operace

Po prostudování této kapitoly by vám měly být jasné pojmy známé ze střední školy, které jsou důležité pro další studium tohoto textu. Vzhledem k tomu, že s uvedenými termíny se budete dále setkávat, je velmi důležité, abyste si byli jisti, že jim opravdu rozumíte. V případě, že máte pochybnosti, vraťte se k nejasným pojmům a znovu je nastudujte.

Studijní cíle
Cílem této kapitoly je připomenout a osvojit si vybrané matematické pojmy, se kterými statistika pracuje a které se budou objevovat v dalším textu.

Klíčová slova
Množina, operace se součty

1.1 Množina a její prvky

Na začátku je dobré osvěžit si základní matematické pojmy a operace, se kterými se ve statistice pracuje a které se budou neustále opakovat. Vychází se z pojmu ***množina***. Podle G. Cantora, zakladatele důležitého oboru současné matematiky, teorie množin, je množina souhrn objektů, které jsou přesně určené a rozlišitelné, a tvoří součást světa našich představ a myšlenek. Ve statistice množinou rozumíme souhrn rovnocenných jedinců (předmětů nebo událostí), u kterých je možné pozorovat jeden nebo více znaků. Tito jedinci (předměty nebo události) se nazývají ***prvky množiny***. Množiny označujeme velkými písmeny latinské abecedy, prvky množiny malými písmeny s indexy. Například množina M se skládá z prvků $x_1, x_2, \ldots x_n$, matematický zápis pak vypadá takto: $M = \{x_1, x_2, \ldots x_n\}$. O prvku x množiny M řekneme, že x patří do množiny M a symbolicky zapíšeme $x \in M$, jestliže prvek y nepatří do množiny M, píšeme $y \notin M$.

Množina

Prvky množiny

Některé množiny mají svá pevná označení:
N....množina přirozených čísel
R ....množina reálných čísel
C....množina komplexních čísel

1.2 Základní operace se součty

V matematické statistice pracujeme se vzorci, které obsahují součet hodnot znaku. Pro součet byl zaveden symbol Σ (suma).

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \,,\; i = 1, 2, \ldots\ldots, n$$

(Pokud nebude uvedeno jinak, pak zkrácený zápis bude vždy znamenat součet

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \,,\; i = 1, 2, \ldots\ldots, n\,).$$

Suma se ve statistice používá nejčastěji ve vzorci pro výpočet aritmetického průměru:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_i + \ldots + x_n)\,,\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad,\quad i = 1, 2, \ldots\ldots, n\,).$$

Suma součtu (rozdílu) x_i a y_i :

$$\sum_{i=1}^{n} (x_i \pm y_i) = (x_1 \pm y_1) + (x_2 \pm y_2) + \ldots + (x_i \pm y_i) + \ldots + (x_n \pm y_n) =$$
$$= (x_1 + x_2 + \ldots + x_i + \ldots + x_n) \pm \quad \pm\ (y_1 + y_2 + \ldots + y_i + \ldots + y_n) =$$
$$= \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$$

Suma součinů $c.x_i$, kde c je konstanta, kterou lze vytknout před symbol Σ :

$$\sum_{i=1}^{n} c \cdot x_i = c \cdot x_1 + c \cdot x_2 + \ldots + c \cdot x_i + \ldots + c \cdot x_n \;=$$
$$= c \cdot (x_1 + x_2 + \ldots + x_i + \ldots + x_n) = c \cdot \sum_{i=1}^{n} x_i$$

Speciální případ nastává, sčítáme-li pouze konstanty

$$\sum_{i=1}^{n} c = c + c + \ldots + c = n \cdot c$$

Součet součinů hodnot dvou proměnných veličin za předpokladu, že každá proměnná nabývá jiného počtu různých hodnot:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_i y_j = x_1 \cdot y_1 + x_1 \cdot y_2 + \ldots + x_1 \cdot y_j + \ldots$$
$$+ x_1 \cdot y_n + x_2 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_2 \cdot y_j + \ldots + x_2 \cdot y_n + \ldots$$
$$+ x_i \cdot y_1 + x_i \cdot y_2 + \ldots + x_i \cdot y_j + \ldots + x_i \cdot y_n + \ldots + x_m \cdot y_1$$
$$+ x_m \cdot y_2 + \ldots + x_m \cdot y_j + \ldots + x_m \cdot y_n$$

Sčítáme zde množiny všech kombinací i a j , jejichž počet je $m \cdot n$.

Příklad

Vypočítejte hodnotu z pro zadané hodnoty x_i a y_i.

$$z = \frac{\sum_{i=1}^{5}(x_i + y_i) - \sum_{i=1}^{5} x_i^{\,2}}{\sum_{i=1}^{5} x_i y_i} \qquad x_i = 1, 2, 3, 4, 5 \qquad y_i = 2, 4, 3, 1, 1$$

x_i	y_i	$x_i + y_i$	x^2_i	$x_i y_i$
1	2	3	1	2
2	4	6	4	8
3	3	6	9	9
4	1	5	16	4
5	1	6	25	5
Σ		26	55	28

$$z = \frac{\sum_{i=1}^{5}(x_i + y_i) - \sum_{i=1}^{5} x_i^{\,2}}{\sum_{i=1}^{5} x_i y_i} = \frac{26 - 55}{28} = \frac{-29}{28} = -1{,}04$$

Kontrolní otázky a úkoly

1. Co je to množina a kdo je zakladatel teorie množin?
2. Jak se pracuje se znakem Σ ?
3. Rozepište a vypočítejte výraz „z“ pro hodnoty:

$x_i = 2, 3, 4 \qquad y_i = 4, 3, 1$

$$z = \sum_{i=1}^{3} x_i + y_i$$

Klíč k otázkám a úkolům

Odpovědi k otázkám 1. a 2. najdete v textu.

$3 \cdot z = 17$

Referenční seznam

CYHELSKÝ, L., KAHOUNOVÁ, J., HINDLS, R., 2001. *Elementární statistická analýza*. Praha: Manegement Press. ISBN 80-7261-003-1.

REITEROVÁ, Eva, 2011. *Základy statistiky pro studenty psychologie*. Olomouc: UP. ISBN 978-80-244-2316-6.

2 Popisná statistika

K řešení různých problémů, ať už z oblasti zdravotnictví nebo z jiných vědních oblastí, většinou používáme číselné údaje, které pak zpracováváme pomocí metod matematické statistiky. Abychom mohli tyto metody použít, musíme si nejprve objasnit některé základní statistické pojmy.

Studijní cíle

Cílem popisné statistiky je určitým způsobem popsat a vyjádřit výsledky zkoumání. Není vhodné ani často možné jednotlivě zprostředkovat všechny naměřené hodnoty. Jde o to, aby se výsledky shrnuly do jasné a srozumitelné formy, která by vyjádřila podstatu věci.

Klíčová slova

Základní soubor, hromadný jev, kategorie znaku, tabulka četností

2.1 Základní statistické pojmy

Statistika se zabývá studiem situací, které se mohou opakovat. Toto opakování je dáno buď tím, že existuje reálná populace objektů daného typu (např. populace pacientů s určitým typem onemocnění) nebo populace opakování dané situace – hromadný jev. ***Hromadné jevy*** se v lidské společnosti se vyskytují v mnoha individuálních případech a jsou rozmístěny ve velkém prostoru a čase. Říkáme, že tyto jevy mají velkou variabilitu – proměnlivost - to znamená různý stupeň určité vlastnosti. Chceme-li postihnout všechny pravidelnosti a zákonitosti jevů ve společnosti, musíme provádět pozorování velkých skupin – celé ***populace***. Termín populace se používá ve statistice ve dvojím smyslu. V matematické statistice znamená statistický soubor a v demografii označuje obyvatelstvo.

Hromadné jevy

Populace

Statistický soubor je soubor, který se skládá ze statistických jednotek. Mohou to být rozmanité věci, jevy nebo procesy, podle toho, čím se statistika zabývá, a které vyhovují daným kritériím věcným, časovým nebo prostorovým. Volba vhodných statistických jednotek je důležitou součástí při vědeckém výzkumu. Statistickým souborem může být například soubor obyvatel České republiky k určitému datu nebo soubor osob, jejichž chování sledujeme, soubor studentů na vysoké škole atd.. Abychom získali nějaký statistický soubor, musíme pro-

Statistický soubor

vést řadu přímých či nepřímých pozorování. To je organizováno například jako terénní průzkum na základě nějaké statistické procedury.

Základní soubor je soubor všech statistických jednotek, který charakterizují určité znaky. ***Statistické znaky*** jsou veličiny, které vyjadřují úrovně nebo stavy vlastností a vztahy mezi nimi. Znaky, které měříme, můžeme rozdělit do několika skupin na znaky časové, prostorové a věcné. Mezi věcnými znaky rozlišujeme takové, které lze vyjádřit slovně. Takové znaky se nazývají ***kvalitativní*** nebo nominální znaky a ptáme se na ně otázkou „jaký" nebo „jak". Jestliže mají kvalitativní znaky dvě obměny, říká se jim alternativní (např. muž – žena, starý – mladý). Jestliže věcné znaky vyjadřujeme čísly, pak mluvíme o znacích kvantitativních a ptáme se na ně otázkou „kolik". ***Kvantitativní znaky*** mohou být buď spojité, nebo nespojité – diskrétní. Spojité znaky nabývají libovolných reálných hodnot a diskrétní znaky pouze izolovaných hodnot. Ke spojitým hodnotám znaku dospějeme většinou měřením a k diskrétním počítáním. Zvláštními kvantitativními znaky jsou znaky intervalové. Přechod mezi kvalitativními a kvantitativními znaky tvoří znaky ***ordinální*** neboli pořadové. Tyto znaky jsou věcně znaky kvalitativními, ale formálně mají podobu i vlastnosti znaků kvantitativních.

Základní soubor
Statistické znaky
Kvalitaitivní znaky
Kvantitativní znaky
Ordinální znaky

Kategorie znaku je skupina všech možností, variant a stavů znaku. Znak, který je určen svou kategorií se nazývá kategorizovaný znak. Kategorie mohou být charakterizovány názvem kategorie (textem), číslicemi, nebo různými symboly.

Kategorie znaku

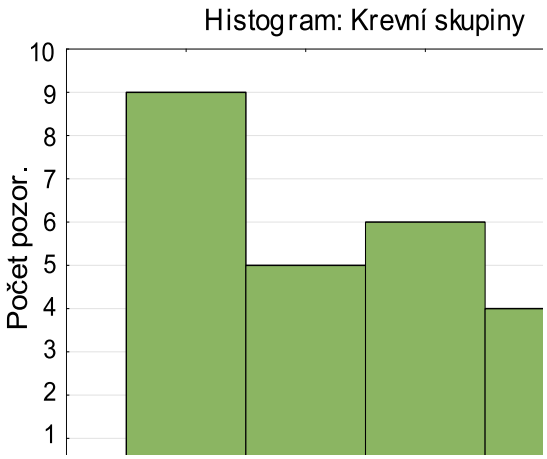
Statistický popis se provádí ve třech různých formách:
– v tabulkách,
– v grafickém znázornění,
– pomocí popisných statistických charakteristik (průměr, rozptyl).

Statistický popis

Příklad

Máme popsat rozdělení souboru 24 pacientů podle krevních skupin. Slovy bychom mohli říct: „Většina pacientů má krevní skupinu 0." Nebo „V tomto souboru má 9 pacientů krevní skupinu 0, 5 pacientů má skupinu A, 6 skupinu B a 4 skupinu AB". Přehledněji toto můžeme vyjádřit v tabulce nebo grafem. Grafické znázornění je přehlednější a jasněji ukazuje, jak jsou krevní skupiny v souboru rozdělené.

Sloupcový graf – graf četností



Tabulka četností

krevní skupiny	četnost
0	9
A	5
B	6
AB	4

2.2 Statistické třídění

Tabulka četností bude nepřehledná, jestliže máme zpracovat velké množství údajů. V tomto případě se snažíme zmenšit počet údajů tím, že shrneme vždy dvě nebo více sousedních hodnot do jedné třídy nebo do jednoho třídního intervalu. Takto se získá větší přehlednost a jednoduchost. Při skupinovém rozdělení četností volíme třídy, skupiny nebo intervaly, do nichž třídíme základní data. ***Třída*** je množina všech hodnot, které leží mezi určenými hranicemi třídního intervalu. ***Hranice třídy*** (hranice intervalu) tvoří nejnižší a nejvyšší hodnota v dané třídě. ***Šířka třídy*** (interval nebo rozpětí třídy) je u diskrétní proměnné počet měřitelných hodnot zahrnutých do třídy. Vypočítá se jako rozdíl dvou po sobě následujících středů tříd a označuje se h, $h > 0$. U třídních intervalů volíme hranice intervalů a jejich délku. Dolní (horní) hranice intervalu udává, kterou nejnižší (nejvyšší) hodnotu do intervalu ještě zařazujeme. Délkou intervalu označujeme kladný rozdíl dvou po sobě jdoucích dolních (dh), případně horních (hh) hranic intervalů. Délka intervalů má být pokud možno stejně velká a označuje se i. Třídní intervaly je nutno volit tak, aby každý prvek mohl být zařazen do jednoho třídního intervalu. Při volbě širokých třídních intervalů dostaneme malý počet tříd. Čím menší je šířka třídních intervalů, tím větší je jejich počet.

Třída
Hranice třídy

Šířka třídy

Příklad
Určete dolní a horní hranici intervalu (30–34) v tabulce:

Tabulka intervalového rozdělení četností

interval	četnost
40–44	4
35–39	2
30–34	6
25–29	0
20–24	3
15–19	1

Pro interval (30–34)
je $dh = 29{,}5$; $hh = 34{,}5$; $i = 5$.

2.3 Druhy četností

Absolutní četnost

1. ***Absolutní četnost*** f_i – udává počet prvků se stejnou obměnou statistického znaku nebo s hodnotami spadajícími do určité třídy nebo intervalu. Součtem všech absolutních četností dostaneme celkovou četnost v souboru – rozsah souboru

$$n = f_1 + f_2 + \ldots f_k = \sum_{i=1}^{k} f_i$$

Relativní četnost

2. ***Relativní četnost (procento výskytu)*** $\frac{f_i}{n}$ – udává poměr absolutní četnosti a rozsahu souboru. Součet všech relativních četností je roven jedné $\sum_{i=1}^{k} \frac{f_i}{n} = 1$.

Absolutní kumulativní četnost

3. ***Absolutní kumulativní četnost*** – vyjadřuje součet všech předcházejících absolutních četností.

Relativní kumulativní četnost

4. ***Relativní kumulativní četnost*** – vyjadřuje součet všech předcházejících relativních četností.

Nejjednodušším zpracováním neuspořádaných výsledků je jejich seřazení podle velikosti a přiřazení příslušné četnosti v tabulce.

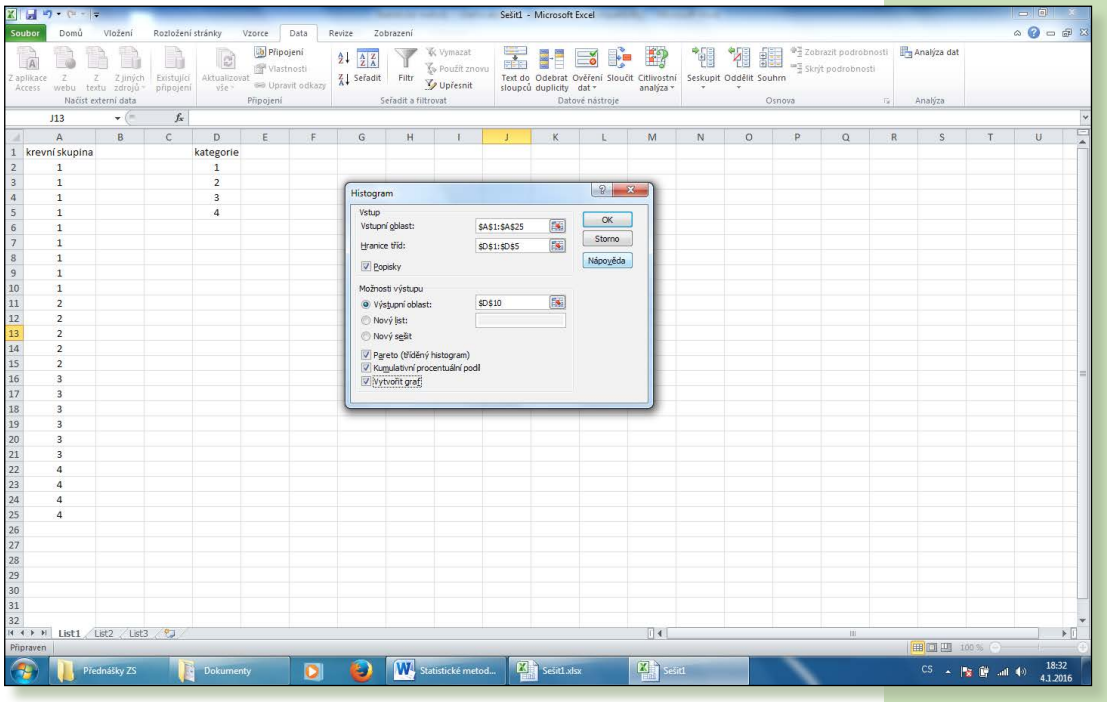
Tabulka četností s relativními a kumulativními četnostmi

Krevní skupiny	Krevní skupiny			
	Absolutní četnost	Kumulativní četnost	Relativní četnost	Kumulativní rel. četnost
0	9	9	37,50000	37,5000
A	5	14	20,83333	58,3333
B	6	20	25,00000	83,3333
AB	4	24	16,66667	100,0000

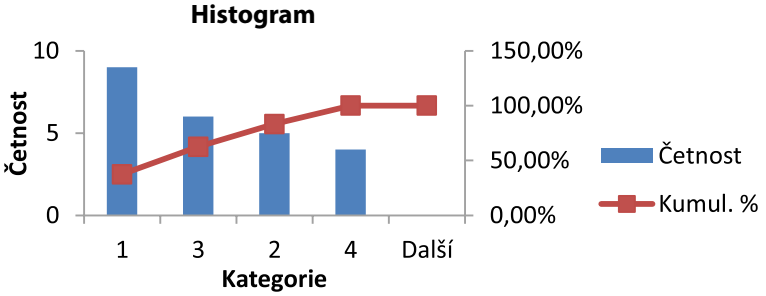
Příklad

Postup při vytvoření tabulky četností a sloupcového grafu (histogramu) v Excelu:

1. Krevní skupiny musíme nahradit čísly (0 = 1, A = 2, B = 3, AB = 4)
2. Do datové tabulky přidáme sloupec kategorie
3. Klikneme na Analýzu dat v záložce Data a vybereme Histogram, zadáme Vstupní oblast, Hranice tříd, Výstupní oblast a zatrhneme Popisky, Pareto, Kumulativní procentuální podíl a Vytvořit graf.



Kategorie	Četnost	Kumul. %	Kategorie	Četnost	Kumul. %
1	9	37,50%	1	9	37,50%
2	5	58,33%	3	6	62,50%
3	6	83,33%	2	5	83,33%
4	4	100,00%	4	4	100,00%
Další	0	100,00%	Další	0	100,00%



2.4 Normální – Gaussovo rozdělení

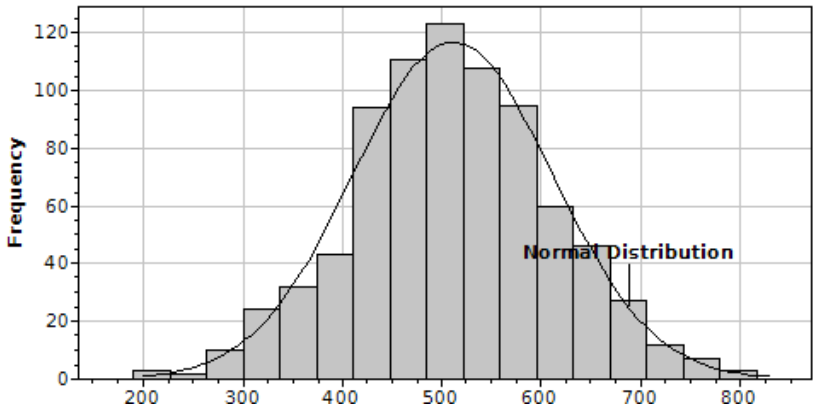
Normální – Gaussovo rozložení

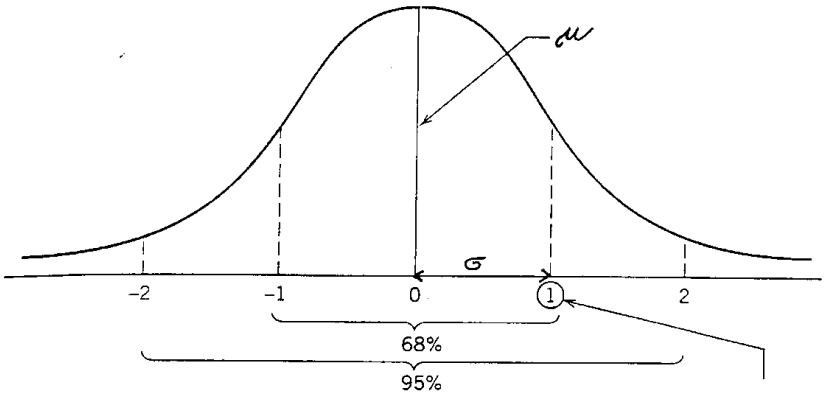
Při vyhodnocování kvantitativních znaků podle jejich rozložení se setkáváme při velkém počtu měření se zcela určitým druhem rozložení, které se dá snadno poznat ze svého grafického znázornění. Jedná se o ***Normální – Gaussovo rozložení*** spojité náhodné veličiny. Takovýmto rozložením četností se řídí především biologické jevy a jevy týkající se člověka. Projevy lidí jsou ovlivněny mnoha podmínkami, které jsou navzájem nezávislé a mohou se vyskytovat zcela náhodně. Rozložení takových znaků má pak tyto vlastnosti:

1. je symetrické – to znamená, že pozorované hodnoty se rozkládají stejnoměrně vlevo a vpravo od střední hodnoty.
2. je různě strmé a široké, ale vždy obdržíme stejný tvar rozdělení, jestliže měřítko na vodorovné ose změníme tak, aby standardní odchylka $\sigma = 1$ a průměr $\mu = 0$.

Gaussova křivka

Ke grafu ***Gaussovy křivky*** se dostaneme z histogramu četností:





jedna standardní odchylka od průměru ($\mu = 0, \sigma = 1$)

Tímto způsobem je tedy možno převést normální rozdělení na standardizované normální rozdělení, které se označuje $N(0,1)$. Pro standardizaci normální náhodné veličiny X s parametry μ a σ na náhodnou veličinu Z, která má parametry 0 a 1 se používá transformační rovnice tvaru $z = \frac{x-\mu}{\sigma}$.

Tato rovnice k jakékoliv hodnotě x náhodné veličiny X udává normovanou hodnotu z náhodné veličiny Z.
V grafu normálního rozdělení platí, že v intervalu (μ–σ, μ+σ) leží 68 % všech hodnot, v intervalu (μ–2σ, μ+2σ) leží 95,6 % všech naměřených hodnot a v intervalu (μ–3σ, μ+3σ) leží 99,7 % všech naměřených hodnot.
Má-li náhodná veličina X rozdělení N (μ,σ), pak jsou hodnoty její frekvenční funkce (hustoty pravděpodobnosti) vyjádřeny rovnicí

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

kde $\pi = 3{,}14$ a e je základ přirozených logaritmů.

Distribuční funkce $\Phi(Z)$ normálního standardizovaného rozdělení $N(0,1)$ má tvar

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx,$$ kde z leží v intervalu (-∞,∞).

Kontrolní otázky a úkoly

1. Definujte hromadný jev.
2. V jakých formách se provádí statistický popis?
3. Co je to histogram?
4. S jakými druhy četností popisná statistika pracuje?
5. Popište Gaussovu křivku

Klíč k otázkám a úkolům

Odpovědi najdete v textu.

Referenční seznam

KUNDEROVÁ, P. 2004. *Úvod do teorie pravděpodobnosti a matematické statistiky.* Olomouc: UP. ISBN 80-244-0843-0.
REITEROVÁ, E. 2011. *Základy statistiky pro studenty psychologie*. Olomouc: UP. ISBN 978-80-244-2316-6.
ŠŤASTNÝ, Z. 1999. *Matematické a statistické výpočty v Microsoft Excelu*.Brno: Computer Press. ISBN 80-7226-141-X.

3 Soubory dat

Sbíráme informace – data nejen o lidech a jejich vlastnostech, ale i o jejich činnostech, vztazích a jevech mezi nimi. Různým skupinám dat říkáme soubory dat. Máme například zjistit názory studentů na blokový systém studia. K tomu můžeme použít buď soubor všech studentů z celé univerzity nebo pouze části studentů, kterou vybereme podle předem stanoveného pravidla ze souboru studentů na celé univerzitě. Z tohoto dílčího souboru pak můžeme s určitou spolehlivostí vyslovit závěry nejen pro tento dílčí soubor, ale i pro soubor všech studentů na celé univerzitě.

Studijní cíle

V této kapitole bude po statistické stránce popsán základní soubor a soubor výběrový. Parametry základního souboru a výběrové charakteristiky sloužící k popisu jsou zde podrobně popsány a je vysvětlen jak jejich ruční výpočet, tak i postup výpočtu v programu Microsoft Excel.

Klíčová slova

Základní soubor, výběrový soubor, střední hodnoty, aritmetický průměr, medián, modus, míra variability, rozptyl, směrodatná odchylka, porovnání variability, kvantilové rozdělení, velikost výběru

3.1 Statistické charakteristiky a parametry základního souboru

Obecně rozlišujeme dva typy souborů:

1. ***Základní soubor (populace) – ZS*** je množina všech prvků, která je vymezena cílem výzkumu. Pro tento soubor vyslovujeme závěry z výzkumného šetření.
2. ***Výběrový soubor (výběr, vzorek) – VS*** je množina jednotek, které byly ze základního souboru vybrány podle předem stanovených pravidel. Pro tento výběrový soubor máme k dispozici data, která reprezentují soubor základní. To znamená, že výsledky zjištěné pro výběrový soubor můžeme zobecnit na soubor základní.

Základní soubor

Výběrový soubor

3.1.1 Střední hodnoty základního souboru

Pro popis rozložení naměřených údajů v základním souboru slouží charakteristiky polohy. Patří k nim aritmetický průměr, medián a modus.

Aritmetický průměr

Aritmetický průměr základního souboru se značí μ a vypočítá se podle vzorce $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, kde x_i je i-tá naměřená hodnota v základním souboru a n je rozsah základního souboru.

Medián

Další charakteristikou polohy je ***medián***, který se značí M_e. Je to naměřená hodnota, která se nachází ve středu řady všech hodnot základního souboru srovnaných podle velikosti. Při lichém počtu měření odpovídá medián skutečné prostřední hodnotě, při sudém počtu měření je medián průměr ze dvou prostředních členů řady. Jestliže například řada obsahuje 51 naměřených hodnot, medián je 26. naměřená hodnota. Jestliže má řada 100 hodnot, pak medián leží mezi 50. a 51. naměřenou hodnotou.

Příklad

V řadě jedenácti naměřených hodnot 125, 127, 128, 129, 129, 130, 131, 132, 133, 134, 134 je medián 6. člen řady tj. $M_e = 130$.

V řadě dvanácti naměřených hodnot 125, 128, 128, 129, 130, 130, 131, 132, 132, 132, 134, 135 je medián průměr ze dvou prostředních hodnot $M_e = \frac{130+131}{2} = \frac{261}{2} = 130{,}5$.

Modus

Na rozdíl od aritmetického průměru má použití mediánu jako charakteristiky polohy výhodu v tom, že se nemusí se počítat ze všech hodnot. Je nezávislý na maximální a minimální hodnotě základního souboru. ***Modus*** M_o je hodnota, která se v rozdělení četností vyskytuje nejčastěji. Pokud se v řadě hodnot budou stejně často vyskytovat hodnoty s maximální četností vedle sebe, modem bude jejich průměr. Jestliže v řadě existují dvě navzájem nesousedící hodnoty s maximálními četnostmi, pak se obě tyto hodnoty uvádí jako modus a rozdělení se nazývá bimodální (dvojvrcholové).

Příklad

V řadě 125, 128, 128, 129, 130, 130, 131, 132, 132, 132, 134, 135 je $M_o = 13$

V řadě 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8 je $M_o = 5{,}5$ (aritmetický průměr čísel 5 a 6).

V řadě 3, 4, 4, 5, 5, 5, 5, 6, 6, 7, 8, 8, 8, 8, 9 je $M_{o1} = 5$, $M_{o2} = 8$.

3.1.2 Míry variability základního souboru

Je třeba mít také míry pro individuální rozdíly – tzv. odchylky od střední hodnoty. Jestliže známe aritmetický průměr, pak se dá vypočítat odchylka od střední hodnoty pro každou naměřenou hodnotu $\delta = x_i - \mu$. Nejhrubší mírou, která se používá u malých souborů, je rozdíl mezi největší a nejmenší naměřenou hodnotou. Tomuto rozdílu se říká ***variační rozpětí*** a vypočítá se podle vzorce $R = max\ x_i - min\ x_i$. Variační rozpětí je však velmi závislé na náhodných vlivech, protože používá pouze dvou extrémních hodnot. Proto byly odvozeny míry rozptylu, které se počítají ze všech hodnot. Nejlepší mírou pro stupeň variability rozložení je ***směrodatná odchylka σ*** a její druhá mocnina σ^2, která se nazývá ***rozptyl*** neboli variance. Rozptyl se vypočítá jako průměr druhých mocnin odchylek všech naměřených hodnot od jejich aritmetického průměru $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$.

Variační rozpětí

Směrodatná odchylka

Rozptyl

Směrodatná odchylka je pak odmocnina z tohoto vzorce $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$.

Čím je tato hodnota větší, tím více je rozložení rozptýleno dál od průměru a čím je menší, tím více se naměřené hodnoty hromadí kolem průměru. Nejlépe se dá směrodatná odchylka a rozptyl vysvětlit na příkladě střeleckého terče. Je-li na terč vypáleno několik výstřelů „rozptýlí" se kolem středu. Střílel-li dobrý střelec, jsou vzdálenosti od středu skoro všechny malé – mají malý rozptyl. Jestliže však střílel špatný střelec, zásahy jsou od středu vzdáleny více – jejich rozptyl je velký.

Příklad

Máme dvě skupiny čísel, která mají stejný průměr. Rozdíl mezi těmito skupinami se vyjádří mírou rozptylu.

1. skupina (x): $x_1 = 7, x_2 = 8, x_3 = 9; \mu_1 = 8; n_1 = 3$

$$\sigma_1^2 = \frac{\sum_{i=1}^{3}(x_i - \mu_1)^2}{n_1} = \frac{(7-8)^2 + (8-8)^2 + (9-8)^2}{3} =$$

$$\frac{(-1)^2 + 0^2 + 1^2}{3} = \frac{2}{3} = 0{,}67$$

$$\Rightarrow \sigma_1 = \sqrt{\sigma_1^2} = \sqrt{0{,}67} = 0{,}82$$

2. skupina (y): $y_1 = 1, y_2 = 10, y_3 = 13; \mu_1 = 8; n_1 = 3$

$$\sigma_2^2 = \frac{\sum_{i=1}^{3}(y_i - \mu_2)^2}{n_2} = \frac{(1-8)^2 + (10-8)^2 + (13-8)^2}{3}$$

$$= \frac{(-7)^2 + 2^2 + 5^2}{3} = \frac{49+4+25}{3} = \frac{78}{3} = 26$$

$$\Rightarrow \sigma_2 = \sqrt{\sigma_2^2} = \sqrt{26} = 5{,}1$$

3.2 Výběrové soubory

Číselné údaje, které charakterizují celou populaci nebo celý základní soubor jsou střední hodnota μ, směrodatná odchylka σ a rozptyl σ^2. Tyto údaje se nazývají ***parametry základního souboru***. Ve většině případů nejsme schopni zachytit všechny jevy, které patří do základního souboru. Není to možné jednak teoreticky, ale také prakticky. Některá měření vedou ke zničení nebo znehodnocení měřeného předmětu. Rozsah základních souborů bývá hodně velký. Měření na všech prvcích velkého základního souboru jsou příliš nákladná ve srovnání s hodnotou a významem získaných výsledků. Někdy je nutné znát výsledky dříve, než lze všechna měření provést. To jsou hlavní důvody proč jsme v praxi odkázáni na zkoumání výběrů a zjišťování ***výběrových parametrů***, ze kterých se pak dají odhadnout parametry základního souboru. ***Výběrový průměr*** $\bar{x}$, ***výběrová směrodatná odchylka*** s a ***výběrový rozptyl*** s^2 jsou výběrové charakteristiky a slouží ke statistickému ověřování hypotéz vztahujících se na základní soubor. Rozsah výběru udává počet prvků obsažených ve výběrovém souboru.

Parametry základního souboru

Výběrové parametry

Výběrový průměr

Výběrová směrodatná odchylka

Výběrový rozptyl

Při provádění výběrů ze základního souboru se mohou vyskytnout dva druhy problémů:

1. jak musíme postupovat, abychom dostali výběr, který by byl reprezentativní vůči celému základnímu souboru;
2. jak je velká spolehlivost a přesnost výsledků získaných z „dobrého“ reprezentativního výběru.

Reprezentativní výběr lze ovšem sestavit jen na základě podrobných znalostí celého základního souboru, které zpravidla nemáme. Proto se používá tzv. náhodných výběrů, jejichž sestavení je založeno na podmínce, aby každý prvek základního souboru měl stejnou pravděpodobnost, že bude do výběru za-

Reprezentativní výběr

hrnut. Náhodné výběry můžeme vytvořit za použití náhodných čísel, která si můžeme vygenerovat na počítači nebo vybrat v tabulce náhodných čísel. Pokud nemáme k dispozici ani jeden z těchto způsobů očíslujeme si prvky základního souboru a pak losujeme. Proces, kdy ze základního souboru vybíráme vzorek (výběr), nazýváme výběrové statistické zjišťování. Proces opačný, kdy výsledky zjištěné ve výběrovém souboru přenášíme na soubor základní, nazýváme statistická indukce. Informace o základním souboru získáváme tedy pomocí výběrů, a to metodami tzv. statistické indukce.

Statistická indukce

Statistická indukce znamená rozšíření závěrů, získaných zpracováním určitého počtu výsledků, které tvoří statistický soubor, na soubor základní.

3.2.1 Druhy výběrů

Náhodný výběr

1. Náhodný výběr

Při náhodném výběru jednotlivých prvků ze základního souboru může dojít ke třem alternativám:

- prvky vybíráme po jednom a vybrané vracíme zpět do základního souboru. Tím je zaručeno, že počet prvků, ze kterých vybíráme se nemění. Každý prvek má stejnou pravděpodobnost vybrání. Tento výběr se označuje jako náhodný výběr s opakováním;
- před provedením výběru mají všechny prvky stejnou pravděpodobnost, že budou vybrány. Po vybrání určitého prvku se pravděpodobnost pro ostatní prvky zvětšuje. Tomuto výběru se říká náhodný výběr bez opakování;
- výběr s různými pravděpodobnostmi tzv. pravděpodobnostní výběr je založen na tom, že každý prvek má určitou předem danou pravděpodobnost vybrání, tyto pravděpodobnosti jsou mezi jednotlivými prvky různé.

Skupinový výběr

2. Skupinový výběr

Předpokladem použití tohoto výběru je to, že základní soubor je uspořádán do přirozených nebo umělých skupin. Skupiny ekonomické, územní, organizační jsou skupiny přirozené, skupinami umělými jsou například abecední seznamy, seznamy podle data narození, podle bydliště atd.. Vybírají se zde celé skupiny prvků bez zřetele k jejich velikosti. Například z abecedního seznamu se vyberou všechna příjmení začínající na písmeno D. Pak všechny osoby, kterým začíná příjmení na D, se stanou prvky výběru. Pro ostatní znaky jako datum narození, bydliště, zaměstnavatel atd. bude tento výběr velmi různorodý, protože se nepředpokládá, že by se tyto znaky vázaly ke jménu.

3. Mechanický výběr
Prvky základního souboru musí být uspořádány podle určitého znaku a postupuje se tak, že se rozdělí do stejně velkých podskupin podle rozsahu výběru, losováním se vybere jedno z čísel od 1 do hodnoty rozsahu podskupin a z každé podskupiny se vybere prvek, který má toto pořadí.

Mechanický výběr

4. Oblastní (stratifikační) výběr
Základní soubor je rozdělen do skupin – oblastí, které musí být svým obsahem co nejvíce různorodé. Z každé oblasti se pak vybere určité procento prvků. Výběr je pak úměrný velikosti oblasti.

Oblastní (stratifikační) výběr

5. Vícestupňový výběr
Vychází se ze skupin nejvyššího řádu a pokračuje se v několika stupních až k elementárním jednotkám.

Vícestupňový výběr

6. Záměrný výběr
Od jiných výběrů se liší tím, že o vybrání prvku ze základního souboru do vzorku rozhoduje každý výzkumník sám. Tento výběr je pak subjektivně ovlivněn. Jestliže výběry provede několik lidí nezávisle na sobě, budou se lišit výběrovými charakteristikami. Přesnost odhadovaných parametrů základního souboru pak záleží na odborném úsudku každého výzkumníka. Existují tři způsoby záměrného výběru:
- výběrové jednotky se dostávají do výběru samy (např. v anketách),
- záměrný výběr průměrných jednotek,
- výběr metodou kvót – zvolí se kontrolní znaky, které se vyskytují u každého prvku základního souboru a podle nich se výběr orientuje.

Záměrný výběr

3.2.2 Výběrové střední hodnoty

Výběrový aritmetický průměr

Výběrový medián

Výběrový modus

Vážený aritmetický průměr

Stejně jako u popisu středních hodnot základního souboru používáme k popisu výběrových středních hodnot ***výběrový aritmetický průměr*** $\bar{x}$, ***výběrový medián*** $\tilde{x}$ a ***výběrový modus*** $\hat{x}$. Někdy je potřebné shrnout závěry z více výběrů s různým rozsahem, abychom mohli charakterizovat soubor všech výběrů jedním ukazatelem – například ***váženým aritmetickým průměrem***, který označujeme $\bar{x}_v$ a vypočítáme jej z aritmetických průměrů jednotlivých výběrů podle vzorce

$$\bar{x}_v = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \ldots + \bar{x}_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{\sum_{i=1}^{k} n_i},$$

kde $\bar{x}_i$, $i = 1, 2, \ldots k$ je aritmetický průměr i-tého výběru a n_i je rozsah i-tého výběru.

Někdy je v praxi tendence počítat prostý aritmetický průměr z naměřených hodnot v různě velkých výběrech. To však vede k nesprávným a zkresleným výsledkům. Jestliže chceme shrnout výběry různého rozsahu do jednoho aritmetického průměru, musíme použít vážený aritmetický průměr.

Příklad

Ve dvou skupinách pacientů se zjišťovalo, kolik dní v určitém časovém období stráví v nemocnici. V první skupině byl průměr na jednoho pacienta 20 dní a ve druhé skupině 40 dní. Chceme zjistit průměrný počet dní hospitalizace v obou skupinách dohromady. Nejdříve budeme předpokládat stejný počet pacientů v obou skupinách a pak různě velké skupiny.

1. $n_1 = n_2 = 10$

Prostý i vážený aritmetický průměr se bude shodovat:

$$\bar{x} = \frac{20+40}{2} = 30 \qquad \bar{x}_v = \frac{20\cdot 10+40\cdot 10}{10+10} = \frac{200+400}{20} = \frac{600}{20} = 30$$

2. $n_1 = 100, n_2 = 10$

$$\bar{x}_v = \frac{20\cdot 100+40\cdot 10}{100+10} = \frac{2000+400}{110} = \frac{2400}{110} = 21{,}820 \approx 22$$

Průměrný počet dní hospitalizace v obou skupinách je 22 dní.

Prostřední hodnota členů výběru uspořádaného podle velikosti je ***medián*** $\tilde{x}$.

Medián

Modus $\hat{x}$ je hodnota, která se v rozdělení četností vyskytuje nejčastěji.

Modus

Při charakterizování či popisování výběrů se dává přednost výběrovému aritmetickému průměru ze dvou důvodů. Udává jednoznačnou hodnotu, která se dá lehce vypočítat a spolehlivě při dostatečném rozsahu výběru odhaduje střední hodnotu základního souboru.

Aritmetický průměr by se neměl používat u ***vícevrcholových (polymodálních) rozdělení***, při extrémně malých výběrech a asymetrických rozděleních. V těchto případech se doporučuje použít jinou charakteristiku polohy např. medián.

Vícevrcholová (polymodální) rozdělení

Medián je vhodný:
- při výběrech s malými četnostmi,
- při asymetrickém rozdělení.

Modus je vhodný při popisu vícevrcholových rozdělení.

3.2.3 Výběrové míry variability

Výběrové rozpětí

Výběrové rozpětí $R = max\ x_i - min\ x_i$. Udává rozdíl mezi největší a nejmenší naměřenou hodnotou. Je tedy určené extrémními hodnotami ve výběru.

Lépe charakterizují variabilitu míry, jejichž základ tvoří rozdíl každé naměřené hodnoty a průměru $x_i - \bar{x}$ (odchylka od průměru). Součet všech těchto odchylek ve výběru je roven nule a proto se používá tzv. průměrná odchylka, která se vypočítá jako průměr z absolutních hodnot odchylek všech naměřených hodnot od aritmetického průměru: $e = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$.

Ve výběru rozděleném do tříd odchylky od průměru vynásobíme příslušnými četnostmi: $e = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \bar{x}|$,

kde f_i je četnost i-té naměřené hodnoty (nebo i-tého středu třídy) v tabulce četností, $\bar{x}$ je aritmetický průměr, k je počet tříd.

Příklad

Vypočítejte průměrnou odchylku z následující tabulky:

x_i	f_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
3	1	–2,13	2,13	2,13
4	3	–1,13	1,13	3,39
5	4	–0,13	0,13	0,52
6	7	0,87	0,87	6,09
Σ	15			12,13

$$\bar{x} = \frac{1}{15}(1\cdot 3 + 3\cdot 4 + 4\cdot 5 + 7\cdot 6) = \frac{1}{15}(3 + 12 + 20 + 42) = \frac{77}{15} = 5{,}13$$

$$e = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \bar{x}| = \frac{1}{15}\cdot 12{,}13 = 0{,}81$$

Výběrový rozptyl

Výběrový rozptyl (variance) je součet druhých mocnin odchylek všech naměřených hodnot od aritmetického průměru vydělený jejich počtem $s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$.

Výběrová směrová odchylka

Výběrová směrodatná odchylka je odmocnina z výběrového rozptylu: $s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$.

Z každého výběrového souboru můžeme vypočítat odhad rozptylu základního souboru $\hat{\sigma}^2 = \frac{\sum (x_i^2 - \bar{x})}{n-1}$

a odhad směrodatné odchylky základního souboru

$$\hat{\sigma} = \sqrt{\frac{\sum (x_i^2 - \bar{x})}{n-1}}$$

K výpočtu výběrového rozptylu se dá podobně jako u rozptylu základního souboru použít vzorce, v němž se nemusí počítat odchylky od průměru:

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}$$

Vzorec pro odhad rozptylu základního souboru:

$$\hat{\sigma}^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

Příklad

Vypočítejte odhad rozptylu a výběrové směrodatné odchylky základního souboru z výběrového souboru naměřených hodnot u dvaceti pacientů.

Pacient	x_i	Pacient	x_i
S1	64	S11	60
S2	48	S12	43
S3	55	S13	67
S4	68	S14	70
S5	72	S15	65
S6	59	S16	55
S7	57	S17	56
S8	61	S18	64
S9	63	S19	61
S10	60	S20	60
Σ			1208

$\sum x_i = 1208$

$\left(\sum x_i\right)^2 = 1459264$

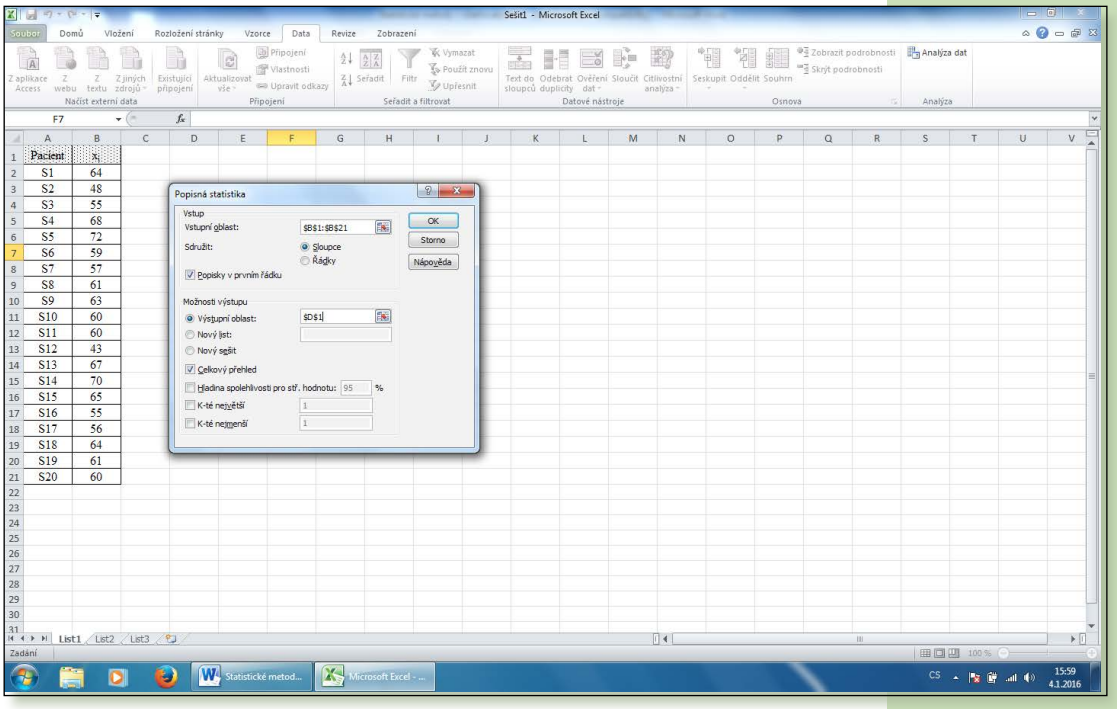
$\sum x_i^2 = 73894$

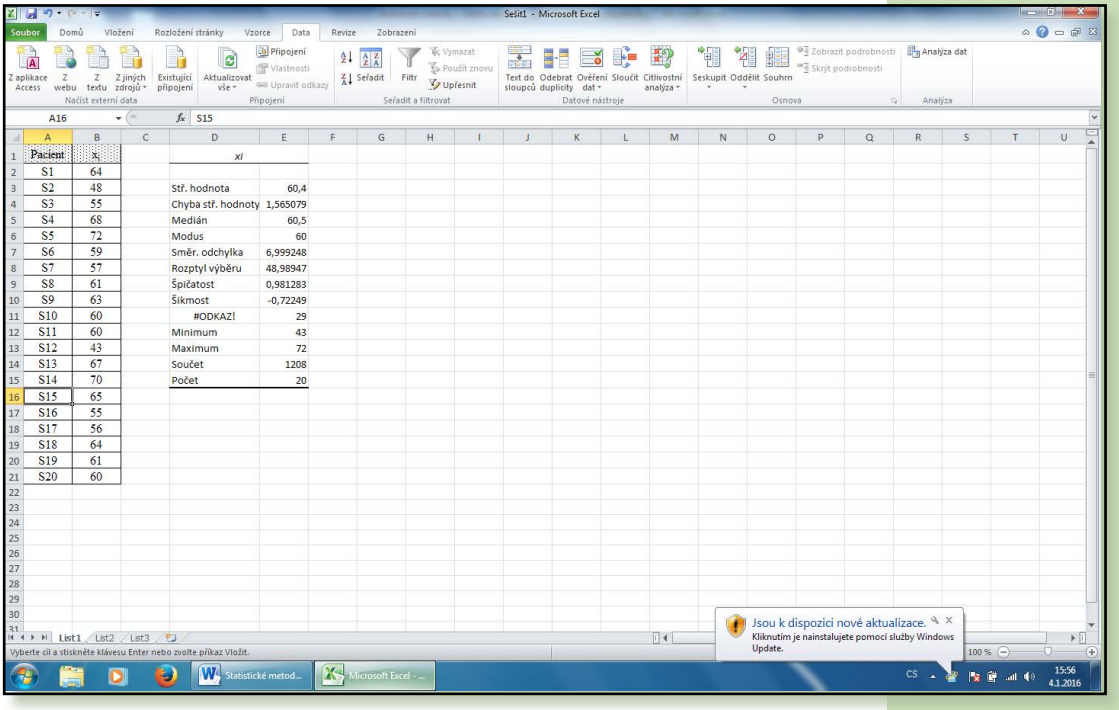
$N = 20$

$$\hat{\sigma}^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{73894 - \frac{1459264}{20}}{19} = \frac{73894 - 72963}{19} = \frac{931}{19} = 49$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{49} = 7$$

Výpočet popisných statistik v programu Microsoft Excel na datech z předchozího příkladu: Data → Analýza dat→ Popisná statistika





Pozn.: Hodnota u popisu #ODKAZ! znamená Variační rozpětí (= Maximum – Minimum)

V případě, že jsou naměřené hodnoty rozdělené do tříd s danými četnostmi, počítáme odhad rozptylu základního souboru podle vzorce: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2$,

kde x_i jsou naměřené hodnoty, f_i jsou jejich četnosti, $\bar{x}$ je aritmetický průměr naměřených hodnot a n je jejich počet.

Výběrový rozptyl a výběrová směrodatná odchylka nejsou ovlivněny náhodnými extrémními hodnotami ve výběru, počítají se ze všech hodnot, spolehlivě odhadují rozptyl základního souboru a používají se pro testování statistických hypotéz.

Vzorec pro odhad rozptylu ve tvaru $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2$ se používá pro malé výběry s již vypočítaným aritmetickým průměrem $\bar{x}$. Pak se údaje zapisují do následující tabulky:

x_i	f_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})$	$f_i (x_i - \bar{x})^2$
...	...	...	...	...
	$\sum_{i=1}^{k} f_i = n$			$\sum_{i=1}^{k} f_i (x_i - \bar{x})$

Příklad
Vypočítejte odhad rozptylu a odhad směrodatné odchylky z dat v tabulce:

x_i	f_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})$	$f_i (x_i - \bar{x})^2$
3	1	–2,13	4,54	4,54
4	3	–1,13	1,28	3,84
5	4	–0,13	0,02	0,08
6	7	0,87	0,76	5,32
Σ	15			13,78

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \frac{1}{14} \cdot 13{,}78 = 0{,}98 \doteq 1$$

$$s = \sqrt{s^2} = \sqrt{1} = 1$$

Pokud není nutné počítat aritmetický průměr, pak se pro odhad rozptylu používá vzorec

$$\hat{\sigma}^2 = \frac{n\sum_{i=1}^{k} f_i x_i^2 - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n(n-1)}$$

a údaje se zapisují do tabulky:

x_i	x_i^2	f_i	$f_i x_i$	$f_i x_i^2$
...	...	...	...	...
		$\sum_{i=1}^{k} f_i = n$	$\sum_{i=1}^{k} f_i x$	$\sum_{i=1}^{k} f_i x_i^2$

Příklad

x_i	f_i	x_i^2	f_i	$f_i x_i^2$
3	9	1	3	9
4	16	3	12	48
5	25	4	20	100
6	36	7	42	252
Σ		15	77	406

$$\hat{\sigma}^2 = \frac{n\sum_{i=1}^{k} f_i x_i^2 - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n(n-1)} =$$

$$= \frac{15 \cdot 409 - 77^2}{15 \cdot 14} = \frac{6135 - 5929}{210} =$$

$$= \frac{206}{210} = 0{,}98 \doteq 1$$

3.2.4 Porovnání variability

Jestliže chceme zjistit, zda je určitý znak v jednom výběru rozptýlenější než stejný znak ve druhém výběru, můžeme porovnat rozptyly a směrodatné odchylky těchto výběrů pouze v případě, že jsou výběry stejně velké a mají přibližně stejné průměry. Můžeme také porovnávat variační koeficienty, které vypočítáme u každého výběru zvlášť, pak výběry mohou mít různý rozsah a odlišný průměr. Proto byl zaveden

Pearsonův variační koeficient $V = \frac{100 \cdot s}{\bar{x}}$, který se udává v %, s je směrodatná výběrová odchylka a $\bar{x}$ je výběrový aritmetický průměr.

Pearsonův variační koeficient

Příklad

U sta pacientů se zjišťoval počet zásahů při experimentu na reakčním přístroji. Počet zásahů je mírou kvality výkonu. V tabulce jsou průměrné počty zásahů a směrodatné odchylky z 1., 5. a 10. pokusu. Chceme zjistit, zda se mění variabilita individuálních výkonů během experimentu.

Pořadí pokusu	$\bar{x}$ (zásahy)	s	V
1.	13,85	4.75	34,3
5.	22,6	4,65	20,6
10.	24,5	3,9	15,9

$$V_1 = \frac{100 \cdot 4{,}75}{13{,}85} = 34{,}3\ \%$$

$$V_2 = \frac{100 \cdot 4{,}65}{22{,}6} = 20{,}6\ \%$$

$$V_3 = \frac{100 \cdot 3{,}9}{24{,}5} = 15{,}9\ \%$$

Můžeme říct, že variabilita v počtu zásahů při experimentech klesá.

Stupně volnosti

Stupně volnosti – číslo $n - 1$ ve vzorci pro odhad rozptylu $(\hat{\sigma}^2 = \frac{1}{n-1}\sum_{n=1}^{k} f_i(x_i - \bar{x})^2)$ se nazývá počet stupňů volnosti.

Pokud máme k dispozici právě jedno pozorování ($n = 1$), pak toto pozorování (tato naměřená hodnota) bude výběrovým průměrem a poskytne nám určitou informaci o průměru celé populace. Pokud k této naměřené hodnotě nemáme žádnou charakteristiku variability výběru (rozptyl, směrodatnou odchylku, variační koeficient), nemáme ani žádnou informaci o variabilitě populace. Když změříme výšku jednoho šestiletého dítěte, pak tuto naměřenou hodnotu můžeme použít jako odhad průměrné výšky všech šestiletých dětí, ale nebudeme vědět, v jakých mezích se tato výška pohybuje (např. od 115 do 125 cm). Jedno pozorování k popisu výběrového souboru nestačí. Pouze v případě, že $n > 1$, je možné usuzovat na rozptýlenost výběru. Podmínkou tedy je, že pro výpočet variability musíme mít $n - 1$ jednotek informací (naměřených hodnot). Proto číslo $n - 1$ je dělitelem pro rozptyl a nazývá se počet stupňů volnosti.

3.3 Kvantilové rozdělení

Všechny výše uvedené střední hodnoty nám udávají obecnou velikost znaku, které nabývají prvky ve výběru. Můžeme ještě použít další charakteristiky, které sice nejsou středními hodnotami, ale přesto udávají svým způsobem polohu rozdělení četností a používají se k uspořádání výběrů podle velikostí. Říká se jim kvantily a dělí se na kvartily, decily a percentily.

Kvantil x_p (p-procentní kvantil) je hodnota znaku, pro kterou platí, že nejméně p procent prvků ve výběru má hodnotu menší nebo rovnu x_p a $100 - p$ procent prvků má hodnotu větší nebo rovnu x_p. Používají se tyto kvantily: medián x_{50}, ***dolní kvartil*** x_{25}, ***horní kvartil*** x_{75}, ***decily*** x_{10}, x_{20}, ..., x_{90} a ***percentily*** x_1, x_2, ..., x_{99}.

Kvantil

Dolní kvartil

Z tohoto popisu je vidět, že kvartily jsou dva, decilů je devět a percentilů devadesát devět, x_{100} udává maximální naměřenou hodnotu. Někdy se dává výrazům **kvartil**, **decil** a **percentil** poněkud odlišný obsah. Říká-li se, že se student umístil v horním kvartilu rozdělení, znamená to, že jeho kvalifikace (počet dosažených bodů) jej zařazuje mezi 25 % posluchačů s nejlepším výsledkem testu.

Kvartil

Decil

Percentil

Podobně jako medián dělí výběrový soubor na dvě poloviny, dolní a horní kvartil rozdělují výběrový soubor na čtyři stejně velké části. Dolnímu kvartilu se říká 25% kvantil a značí se Q_1, hornímu kvartilu 75% kvantil a značí se Q_2.

Ve výběru, kde nejsou naměřené hodnoty uspořádány do tříd, ale pouze podle velikosti, příslušný kvantil získáme jako pořadí k-té hodnoty, vypočítané ze vztahu $k = p \cdot u/100$, kde p je počet pozorování a u je úroveň kvantilu.

Poznámka

95% kvantil standardizovaného normálního rozdělení $N(0,1)$ je $z_{95} = 1{,}69$; 5% kvantil rozdělení $N(0,1)$ je $z_5 = -1{,}69$ (plyne ze symetrie normálního rozdělení).

Příklad

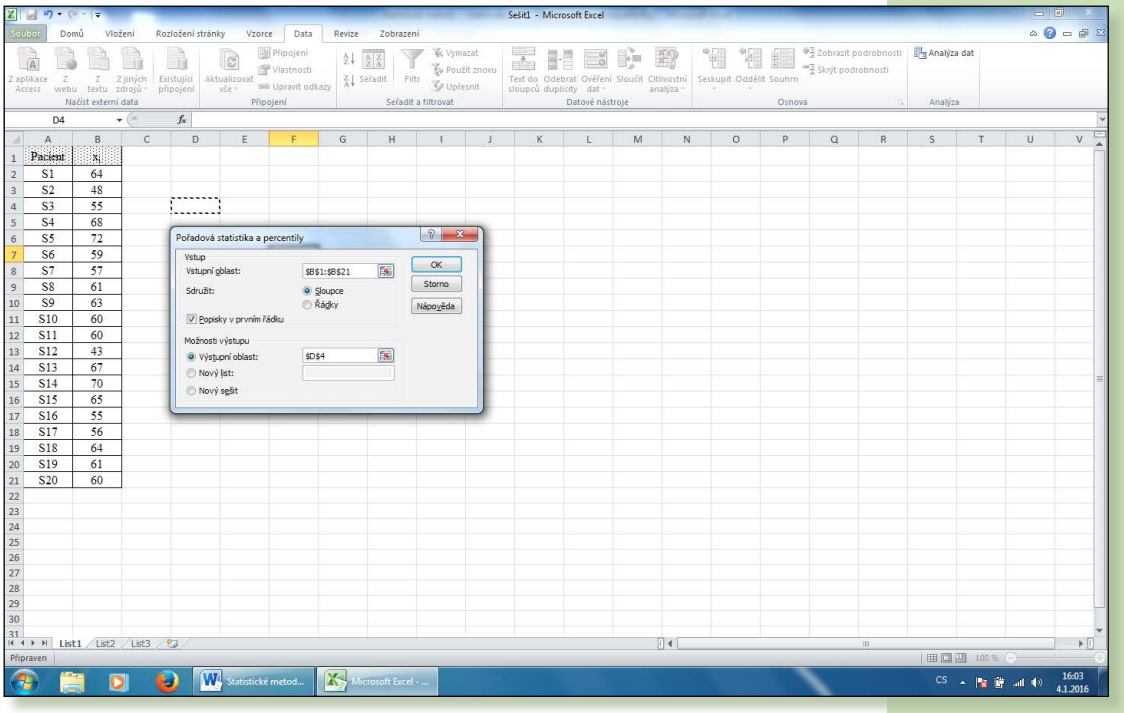
Mějme dána následující čísla $1, 2, 2, 1, 5, 4, 4, 2, 3, 1, 2, 2$.
Pro určení kvantilů uspořádáme hodnoty podle velikosti: $1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5$. Pořadí příslušného kvantilu pak vypočítáme podle výše uvedeného vzorce a zaokrouhlíme na celé číslo. Dolní kvartil je roven třetí hodnotě, protože $12 \cdot 25/100 = 3$ a je tedy 1. Horní kvartil je roven deváté hodnotě, protože $12 \cdot 75/100 = 9$ a je tedy 3. První decil je roven první hodnotě, protože $12 \cdot 10/100 = 1$ a je tedy 1.

Kvartilová odchylka $Q = Q_2 - Q_1$ (mezikvartilový interval) je interval ohraničený horním a dolním kvartilem. V této oblasti leží 50 % všech naměřených hodnot. Čím je kvartilová odchylka větší, tím jsou hodnoty více rozptýlené.

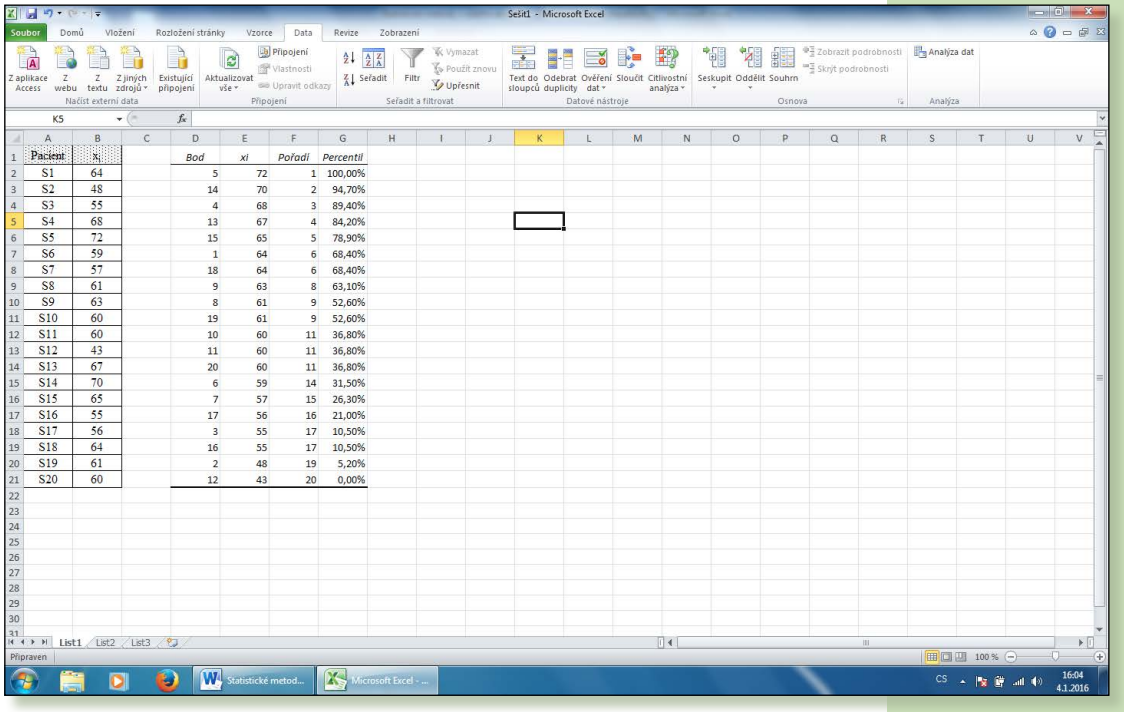
Kvartilová odchylka

Výpočet percentilů v programu Microsoft Excel

Data → Analýza dat → Pořadová statistika a percentily



	A	B
1	Pacient	xi
2	S1	64
3	S2	48
4	S3	55
5	S4	68
6	S5	72
7	S6	59
8	S7	57
9	S8	61
10	S9	63
11	S10	60
12	S11	60
13	S12	43
14	S13	67
15	S14	70
16	S15	65
17	S16	55
18	S17	56
19	S18	64
20	S19	61
21	S20	60



	A	B	D	E	F	G
1	Pacient	xi	Bod	xi	Pořadí	Percentil
2	S1	64	5	72	1	100,00%
3	S2	48	14	70	2	94,70%
4	S3	55	4	68	3	89,40%
5	S4	68	13	67	4	84,20%
6	S5	72	15	65	5	78,90%
7	S6	59	1	64	6	68,40%
8	S7	57	18	64	6	68,40%
9	S8	61	9	63	8	63,10%
10	S9	63	8	61	9	52,60%
11	S10	60	19	61	9	52,60%
12	S11	60	10	60	11	36,80%
13	S12	43	11	60	11	36,80%
14	S13	67	20	60	11	36,80%
15	S14	70	6	59	14	31,50%
16	S15	65	7	57	15	26,30%
17	S16	55	17	56	16	21,00%
18	S17	56	3	55	17	10,50%
19	S18	64	16	55	17	10,50%
20	S19	61	2	48	19	5,20%
21	S20	60	12	43	20	0,00%

Kontrolní otázky a úkoly

1. Charakterizujte základní a výběrový soubor
2. Definujte střední hodnoty základního a výběrového souboru
3. Jaké znáte míry variability?
4. K čemu slouží Pearsonův variační koeficient?
5. Které kvantily se používají nejčastěji?

Klíč k otázkám a úkolům

Odpovědi na otázky naleznete v textu

Referenční seznam

HANOUSEK, J., CHARAMZA, P. 1992. *Moderní metody zpracování dat*. Praha: Grada. ISBN 80-85623-31-5.

PROCHÁZKA, B. 2015. *Stručná biostatistika pro lékaře*. Praha: Karolinum. ISBN 978-80-246-2783-0.

4 Statistická závislost – korelace

Z přírodních věd známe závislosti, kdy určité hodnotě jedné veličiny odpovídá přesně daná hodnota druhé veličiny, a to pro každou její hodnotu. To je případ tzv. ***funkční závislosti***, neboť vztah mezi oběma veličinami se dá popsat matematickou funkcí. V praxi se však setkáváme se závislostmi, kde podobná jednoznačnost mezi veličinami neexistuje, ale hodnotám nezávisle proměnné odpovídají hodnoty závisle proměnné. Například proměnná X je tělesná výška dětí stejného věku (v cm) a proměnná Y je hmotnost těchto dětí (v kg). Je třeba zjistit souvislost mezi tělesnou výškou a hmotností měřených dětí. V tomto případě hovoříme o statistické závislosti a říkáme, že mezi veličinami existuje korelace. Podle toho, jakého typu jsou proměnné X a Y se rozhoduje, jaký druh statistického výpočtu korelace se může použít.

Funkční závislost

Studijní cíle

Cílem této kapitoly je seznámit studenty s výpočty korelačních koeficientů bez i s pomocí programu Excel. Jsou zde uvedeny nejpoužívanější korelační koeficienty – Pearsonův koeficient pro metrická data, Spearmanův pořadový koeficient, korelace pro čtyřpolní tabulku a pro vícepolní kontingenční tabulku.

Klíčová slova

Pearsonova korelace, Spearmannova pořadová korelace, Φ koeficient, C koeficient kontingence

4.1 Druhy korelací

V analýze závislostí mezi dvěma proměnnými se můžeme setkat s různými druhy korelací. Buď zjišťujeme závislost mezi dvěma naměřenými (metrickými) proměnnými, pak k výpočtu korelace použijeme ***Pearsonův korelační koeficient***. Pokud pracujeme s pořadovými hodnotami, nebo při malém počtu měření převádíme metrické proměnné na pořadové, pak k výpočtu korelace je vhodný ***Spearmanův pořadový korelační koeficient***. Pro práci s četnostmi u dvou alternativních proměnných použijeme ***čtyřpolní koeficient korelace Φ*** a u kategoriálních proměnných ***koeficient kontingence C*** pro vícepolní kontingenční tabulku.

Pearsonův korelační koeficient

Spearmanův pořadový korelační koeficient

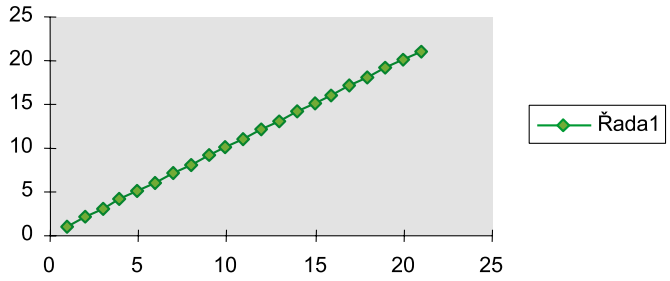
Čtyřpolní koeficient korelace

Koeficient kontingence

Dvojice korelovaných metrických proměnných lze pokládat za souřadnice bodů v rovině a náhodný výběr se pak znázorňuje tzv. tečkovým diagramem. Na tomto grafickém znázornění lze pak zhruba poznat, o jaký typ statistické závislosti se jedná.

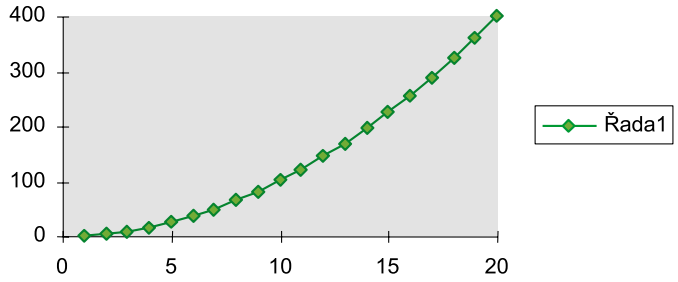
Lineární korelace

1. ***Lineární korelace*** (přímka se nazývá regresní přímka)



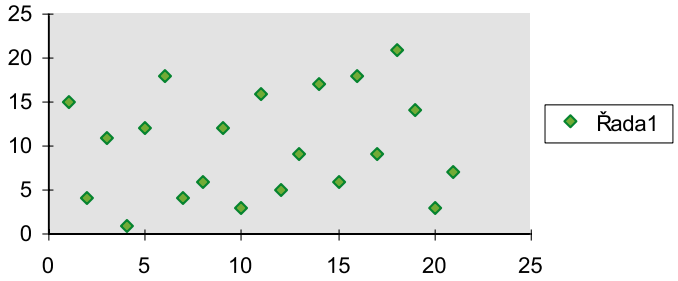
Nelineární korelace

2. ***Nelineární korelace*** (přímka se nazývá regresní křivka)



Statistická nezávislost

3. ***Případ statistické nezávislosti***



V analýze závislosti mezi dvěma nebo více proměnnými jde o dvě základní úlohy:

Regresní analýza

– stanovit charakter a průběh regresní čáry – to řeší ***regresní analýza***,

Korelační analýza

– určit těsnost zjištěného vztahu a posoudit jeho statistickou významnost – to řeší ***korelační analýza***.

Aby bylo možno použít koeficient korelace, musí být splněny dva základní předpoklady – musí se jednat o lineární regresi

a základní soubor musí mít dvojrozměrné normální rozložení četností.

Údaje pro dvojrozměrné rozdělení můžeme získat tak, že zjišťujeme na stejném prvku ve výběru dva znaky, a tak získáme dvě ***proměnné*** X a Y. Naměřené dvojice hodnot se zapisují do tabulky:

Proměnná X

Proměnná Y

Prvky výběru (pokusné osoby)	Proměnná X	Proměnná Y
A	x1	y1
B	x2	y2
C	x3	y3
.	.	.
.	.	.

Mezi dvěma proměnnými mohou existovat tyto souvislosti:

1. ***Shoda*** – velkým hodnotám X odpovídají velké hodnoty Y, malé hodnoty X se vyskytují společně s malými hodnotami Y. V tomto případě hovoříme o kladné korelaci – souvislosti mezi oběma proměnnými.
2. ***Protiklad*** – velkým hodnotám X odpovídají malé hodnoty Y a obráceně. Existuje zde záporná korelace mezi proměnnými X a Y.
3. ***Nezávislost*** – hodnoty X a Y sobě navzájem neodpovídají. V tomto případě neexistuje mezi proměnnými žádná souvislost a říkáme, že proměnné spolu nekorelují.

Shoda

Protiklad

Nezávislost

4.1.1 Pearsonova korelace pro metrická data

Korelaci pro metrická data používáme k určení, zda rozdíly v naměřených hodnotách dvou proměnných jsou ve vzájemném vztahu, zda korelují. Toto rozhodnutí je možné pomocí korelačního koeficientu, který se obvykle značí r. Korelační koeficient určuje stupeň vztahu mezi dvěma proměnnými a je vyjadřován hodnotou mezi 0 a 1 (–1). Žádný vztah znamená 0, úplná pozitivní závislost je označena 1, úplná negativní závislost –1. S růstem hodnoty r od 0 k 1 (–1) se míra vztahu zvětšuje. Absolutní hodnota korelačního koeficientu nám udává míru vztahu.

Jestliže máme pro obě proměnné X a Y k dispozici metrické údaje, můžeme popsat stupeň jejich závislosti pomocí ***Pearsonova korelačního koeficientu***. Označuje se r a nabývá hodnot v intervalu $\langle -1,+1 \rangle$.

Pearsonův korelační koeficient

Je-li $r = -1$, pak to znamená, že mezi oběma proměnnými je výrazně protikladný vztah – negativní korelace. Je-li $r = +1$, pak mezi proměnnými existuje pozitivní lineární souvislost. Po-

kud obě proměnné nejsou v žádné souvislosti, jsou rozptýlené nezávisle na sobě, korelační koeficient $r = 0$. Z velikosti r se dá zjistit těsnost zkoumaného vztahu.

Existuje několik různých vzorců pro výpočet Pearsonova korelačního koeficientu. Matematicky jsou si rovnocenné a dají se odvozovat jeden od druhého.

1. $r = \frac{s_{xy}}{s_x \cdot s_y}$, kde $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

se nazývá kovariance, s_x a s_y jsou výběrové směrodatné odchylky proměnných X a Y.

2. $$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right] \cdot \left[n\sum y_i^2 - (\sum y_i)^2\right]}},$$

kde x_i a y_i jsou hodnoty proměnných X a Y, n je počet naměřených dvojic X a Y. U tohoto vzorce není nutné počítat průměry X a Y a jejich směrodatné odchylky. Používá se pro malé n a nízké naměřené hodnoty.

3. Při velkém počtu měření používáme rozdělení do tříd s udanými četnostmi. Vzorec pro výpočet korelačního koeficientu pak vypadá takto:

$$r = \frac{n\sum f_{xy} x_i y_i - \sum f_x x_i \sum f_y y_i}{\sqrt{\left[n\sum f_x x_i^2 - (\sum f_x x_i)^2\right] \cdot \left[n\sum f_y y_i^2 - (\sum f_y y_i)^2\right]}},$$

kde f_x a f_y jsou četnosti naměřených hodnot a x_i a y_i je četnost s jakou vystupuje i-tá naměřená hodnota X s i-tou hodnotou Y.

Příklad

U pěti pacientů se měřil čas potřebný k jejich ošetření v prvním (X) a posledním (Y) dni hospitalizace. Chceme zjistit korelaci naměřených časů v těchto dnech. Naměřené hodnoty a pomocné výpočty jsou v tabulce:

Pacient	X	Y	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	x_i^2	y_i^2	$x_i y_i$
A	3	3	–7	49	–3	9	21	9	9	9
B	7	5	–3	9	–1	1	3	49	25	35
C	11	7	1	1	1	1	1	121	49	77
D	14	6	4	16	0	0	0	196	36	84
E	15	9	5	25	3	9	15	225	81	135
Σ	50	30		100		20	40	600	200	340

$\bar{x} = (3+7+11+14+15)/5 = 10 \qquad \bar{y} = (3+5+7+6+9)/5 = 6$

1. Použijeme vzorec

$$r = \frac{s_{xy}}{s_x \cdot s_y}, \text{ kde } s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{40}{5} = 8$$

$$s_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{5}100} = \sqrt{20} = 4{,}47$$

$$s_y = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})} = \sqrt{\frac{1}{5}20} = \sqrt{4} = 2$$

$$r = \frac{8}{4{,}47 \cdot 2} = \frac{8}{8{,}94} = 0{,}89$$

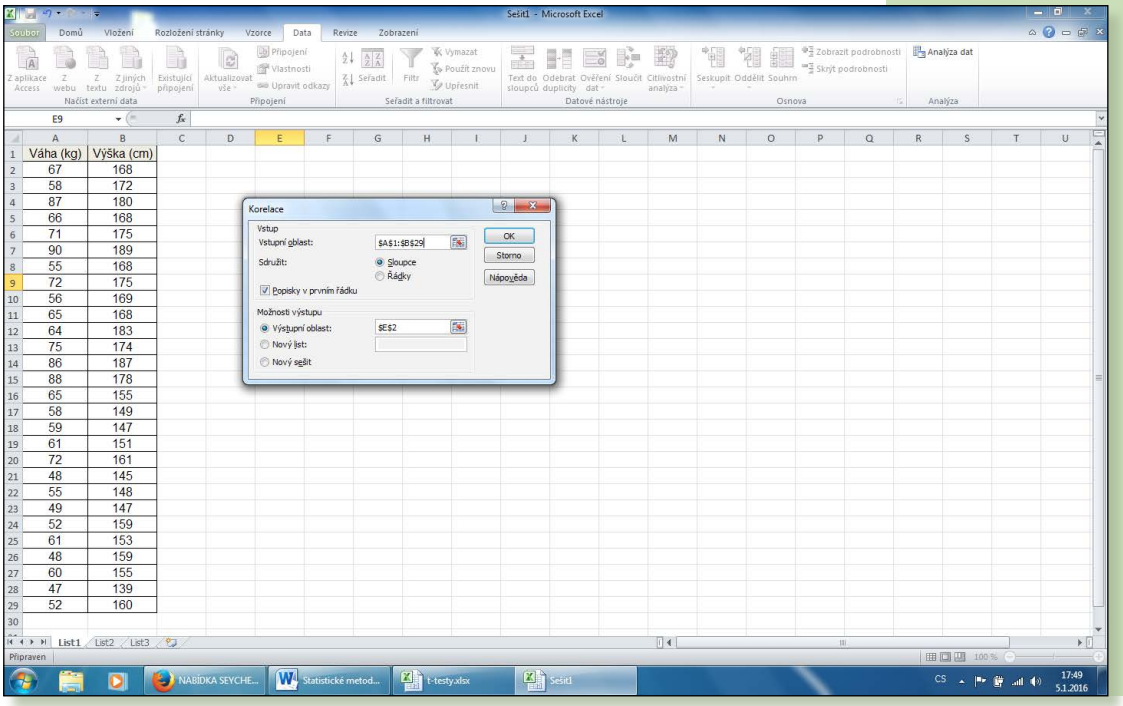
2. Použijeme vzorec $r = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right] \cdot \left[n\sum y_i^2 - (\sum y_i)^2\right]}}$,

který používá přímo naměřené hodnoty. $r = \dfrac{(50 \cdot 50 \cdot 30 - 50^2)}{\sqrt{(5 \cdot 600 - 50^2)(5 \cdot 200 - 30^2)}} =$

$$= \frac{1700 - 1500}{\sqrt{500 \cdot 100}} = \frac{200}{\sqrt{50000}} = \frac{200}{223{,}6} = 0{,}89$$

Výpočet Pearsonova korelačního koeficientu v programu Excel

Data → Analýza dat → Korelace → zadat vstupní a výstupní oblast → Ok



Výsledná tabulka

	Váha (kg)	Výška (cm)
Váha (kg)	1	
Výška (cm)	0,77858	1

Poznámka
Jedná se o Pearsonovu korelaci, jiný druh korelace Excel nepočítá.

4.1.2 Spearmanova pořadová korelace

Pořadová korelace je vhodná v případech, kdy můžeme pracovat s uspořádanými hodnotami. Jestliže je potřeba počítat korelaci z metrické proměnné X a ordinální proměnné Y, musí se vytvořit pořadí z naměřených metrických hodnot a toto pořadí pak porovnávat s proměnnou Y.
Stupeň souvislosti mezi oběma pořadími určuje ***Spearmanův koeficient*** pořadové korelace. Označuje se R a nabývá hodnot z intervalu $\langle -1,+1 \rangle$.

Spearmanův koeficient

Tento koeficient se vypočítá podle vzorce $R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$,

kde d_i je diference dvojice pořadí (rozdíl naměřených hodnot x_i a y_i) a n je počet pořadí.

Příklad

Máme zjistit, zda existuje souvislost mezi klinickým stavem pacienta (X – pořadí jeho stavu) a hodnotou sedimentace červených krvinek seřazenou do pořadí (Y). Hodnocení je v tabulce:

x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	2	–1	1
2	1	1	1
3	5	–2	4
4	3	1	1
5	7	–2	4
6	6	0	0
7	9	–2	4
8	8	0	0
9	4	5	25
Σ			40

$$R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 40}{9(81-1)} =$$

$$= 1 - \frac{240}{980} = 1 - 0{,}33 = 0{,}67$$

Z výsledku je zřejmý vyšší stupeň závislosti mezi klinickým stavem pacientů a pořadím sedimentace.

4.1.3 Závislost mezi alternativními znaky

Alternativní (dichotomická) proměnná

Alternativní (dichotomická) ***proměnná*** je taková proměnná, která může nabývat pouze dvou hodnot – odpovědi ano/ne, dobrý/špatný, muž/žena atd. Závislost mezi dvěma alternativními proměnnými se počítá pomocí čtyřpolního koeficientu korelace z kontingenční tabulky 2 × 2 (čtyřpolní tabulka). V jednotlivých polích tabulky jsou dány četnosti alternativních znaků:

		Znak I +	Znak I –	
Znak II	+	a	b	a + b
	–	c	d	c + d
		a + c	b + d	a + b + c + d

\+ přítomnost znaku u sledované osoby
– nepřítomnost znaku u sledované osoby

Čtyřpolní koeficient korelace

Čtyřpolní koeficient korelace se značí Φ ($\Phi \in \langle -1,+1 \rangle$) a vypočítá se podle vzorce $\Phi = \dfrac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$,

kde a, b, c, d jsou četnosti v jednotlivých polích tabulky.

Příklad

Šedesát pacientů (31 mužů a 29 žen) bylo podrobeno preventivní prohlídce. Podle hodnoty BMI byli pacienti rozděleni na pacienty s normální váhou a s nadváhou. Existuje souvislost mezi pohlavím a nadváhou? Příslušné četnosti jsou v tabulce.

	normální váha	nadváha	
muži	a = 14	b = 17	a + b = 31
ženy	c = 10	d = 19	c + d = 29
	a + c = 24	b + d = 36	a + b + c + d = 60

$$\Phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{14\cdot 19 - 17\cdot 10}{\sqrt{31\cdot 29\cdot 24\cdot 36}} = \frac{266-170}{\sqrt{776736}} = \frac{96}{881} = 0{,}1$$

Koeficient souvislosti mezi pohlavím a nadváhou je 0,1. Jelikož je koeficient malá hodnota, můžeme říct, že mezi pohlavím a nadváhou v našem vzorku souvislost neexistuje.
Pozn. Signifikantnost koeficientu se určuje pomocí testu χ^2 pro čtyřpolní tabulku (viz kap. 6.2)

Čtyřpolní tabulka v Excelu
Vložení → Kontingenční tabulka → zadat oblast dat a výstupní oblast
Celý postup viz Test nezávislosti χ^2 pro čtyřpolní tabulku.

Kontrolní otázky a úkoly

1. Pro jaký typ proměnných používáme Pearsonův a Spearmanův korelační koeficient?
2. Který druh korelace je vhodný k výpočtu závislostí mezi alternativními proměnnými?

Klíč k otázkám a úkolům

Odpovědi na otázky najdete v textu.

Referenční seznam

CHRÁSKA, M. 2007. *Metody pedagogického výzkumu. Základy kvantitativního výzkumu.* Praha: Grada. ISBN 978-80-247-1369-4.

MELOUN, M., MILITKÝ, J. 2004. *Statistická analýza experimentálních dat.* Praha: Academia. ISBN 80-200-1254-0.

REITEROVÁ, E. 2011. *Základy statistiky pro studenty psychologie*. Olomouc: UP. ISBN 978-80-244-2316-6.

5 Testování statistických hypotéz

Popisná statistika charakterizuje výběry pomocí kvantitativních charakteristik (středních hodnot, rozptylu, korelačních koeficientů, atd.). Testovací statistika určuje, zda se tyto ukazatele odlišují „reálně" nebo „náhodně". O skutečný rozdíl se jedná tehdy, jestliže se charakteristiky dvou nebo více výběrů natolik liší, že vedou k odhadu různých parametrů. Rozdíl mezi výběrovými ukazateli je náhodný, jestliže je slučitelný s předpokladem, že příslušné výběry jsou z jednoho a toho samého základního souboru. Jestliže provedeme určitý výzkum a zjistíme charakteristiky jednoho nebo více výběrů, pak pomocí těchto charakteristik zjišťujeme, zda výběry pocházejí ze stejného základního souboru nebo z různých základních souborů. Ptáme se, zda základní soubor má normální (Gaussovo) rozdělení četností nebo, zda je možné považovat sledovaný základní soubor za náhodně uspořádaný. Na tyto otázky odpovídáme pomocí statistických testů. Testují se účinky nových léků, úspěšnost léčby pacientů, hledají se souvislosti mezi dávkou léku a velikostí odezvy, zjišťuje se, která ze dvou nebo více léčebných metod je nejúčinnější.

Studijní cíle

Cílem této kapitoly je seznámit studenty s postupem, který se používá při testování statistických hypotéz a tento postup pak aplikovat na parametrické a neparametrické statistické testy. Studenti by měli být schopni ze zádání poznat, jaký statistický test na svá data použít.

Klíčová slova

Hypotéza, statistická hypotéza, parametrický test, neparametrický test, analýza rozptylu, Studentovy t-testy, X^2 testy

5.1 Základní pojmy

Hypotéza

Hypotéza je tvrzení nebo předpoklad, jímž se v rámci dané teorie vyjádří určitá představa. Oprávněnost hypotéz prověřují pozorování a experimenty. Existují velké hypotézy – o vzniku sluneční soustavy a celého vesmíru, o existenci mimozemských civilizací, ale také hypotézy, které tvrdí, že lék nebo terapie má větší účinnost než jiný lék nebo jiná terapeutická metoda.

Statistická hypotéza

Statistická hypotéza je tvrzení o statistických objektech, a protože předmětem zájmu ve statistice jsou soubory a zejména rozdělení četností znaků sledovaných v těchto souborech, bývá statistická hypotéza tvrzením o těchto rozděleních. Statistickou hypotézou tedy rozumíme jakýkoliv výrok nebo tvrzení o typu rozdělení jedné nebo více náhodných veličin. Statistická hypotéza je vyjádřena smysluplnou oznamovací větou, o jejíž míře pravdivosti můžeme usuzovat ze zjištěných hodnot. Úlohou teorie testování statistických hypotéz je vytváření vhodných metod, pomocí nichž je možné rozhodnout, zda je daná hypotéza pravdivá nebo ne. Jednou z forem hodnocení číselných dat je testování statistických hypotéz na základě teorie vypracované matematiky Neumannem a Pearsonem.

Pokud chceme formulovat statistickou hypotézu, musíme mít o zkoumané populaci určité základní informace. Například předpokládáme, že daná populace má normální rozdělení četností. V takovém případě se statistická hypotéza vztahuje pouze na hodnoty dvou parametrů normálního rozdělení – střední hodnoty (průměru μ) a standardní odchylky (σ). V jiných případech víme o základním souboru – populaci – jen to, že má spojité rozdělení.

Mezi některé základní typy statistických hypotéz patří tato tvrzení:

1. zkoumaný výběr pochází z populace, která má určité teoretické rozdělení;
2. dva zkoumané výběry pocházejí ze stejného základního souboru;
3. existuje lineární závislost mezi dvěma nebo více náhodnými veličinami;
4. jedna nezávisle proměnná ovlivňuje sledovanou závisle proměnnou více než druhá.

Pokud chceme zjistit, zda je daná hypotéza správná, je třeba vytvořit pravidlo, pomocí kterého, na základě výběrového souboru, získaného z populace, rozhodneme, zda hypotéza může být přijata nebo zda je třeba ji zamítnout. Jinými slovy, je třeba vytvořit pravidlo, které by každému výběrovému souboru přiřazovalo jedno ze dvou možných rozhodnutí: hypotézu buď přijmout, nebo zamítnout. Toto pravidlo se nazývá statistickým testem.

Statistický test

Statistický test pro každý výběrový bod určí, máme-li testovanou hypotézu zamítnout nebo nikoliv. To znamená, že bude jednoznačně určena jistá hodnota, která výběrový prostor (množinu možných rozhodnutí) rozdělí na dvě disjunktní části, které nazýváme kritický obor (ozn. W) nebo též obor zamítnutí a doplněk kritického oboru ($\overline{W}$), kterému se také někdy říká obor přijetí. Hodnota (bod, mez), která rozděluje výběrový prostor na tyto dvě části, se nazývá kritická hodnota.

Nulová hypotéza

Důležitým pojmem je tzv. ***nulová hypotéza*** (označuje se H_0). Nulovou hypotézu formulujeme na začátku každého testu tak, že například tvrdíme: srovnávané parametry, které odhadujeme z výběrových charakteristik jsou stejné; nebo výběrové hodnoty patří k populaci o jistém rozdělení.
Symbolicky zapisujeme nulové hypotézy takto:

1. $H_0: \mu = \mu_0$ Populační průměr se rovná počátečnímu populačnímu průměru.
2. $H_0: \mu_1 = \mu_2$ Průměrná hodnota populace, z níž byl pořízen první výběr je rovna průměrné hodnotě populace, z níž byl pořízen druhý výběr.
3. $H_0: H_0: \mu_1 = \mu_2 = \ ... \ = \mu_k - = \mu_k$ Populační průměry z k populací jsou si rovny

Při testování statistických hypotéz se předpokládá, že může platit buď nulová hypotéza H_0, nebo k ní alternativní hypotéza H_1. Například k hypotéze $H_0: \mu = \mu_0$ existuje alternativní hypotéza $H_1: \mu \neq \mu_0$ tzv. dvoustranná hypotéza nebo dvě alternativní jednostranné hypotézy $H_1: \mu > \mu_0$, $H_2: \mu < \mu_0$.

Sestrojit test, který by ověřil správnost hypotézy H_0 pak znamená najít princip, jak rozdělit výběrový prostor – tj. množinu všech v úvahu přicházejících výběrů na dvě disjunktní, navzájem se nepřekrývající podmnožiny tak, aby jedna z nich zahrnula ty výběry, které lze očekávat za platnosti hypotézy H_1, a druhá, aby naopak zahrnovala ty výběry, jejichž výskyt lze očekávat spíše za platnosti hypotézy H_0. První podmnožina se nazývá kritický obor a druhá doplněk kritického oboru. Kritický obor může být podle typu alternativní hypotézy jednostranný nebo dvoustranný. Statistický test je pak prováděn podle rozhodovacího pravidla:

- padne-li výběr, který dostaneme jako výsledek konkrétního pokusu do kritického oboru, zamítá se nulová hypotéza H_0 jako nesprávná a jako správná se přijímá alternativa H_1;
- padne-li výběr mimo kritický obor, přijme se hypotéza H_0 jako správná.

Toto rozhodovací pravidlo přiřazuje každému experimentálnímu výsledku (tj. každému výběru) jedno ze dvou možných rozhodnutí. Situace, které mohou nastat shrnuje následující tabulka:

Rozhodnutí (test)

Skutečnost	H_0 je správná +	H_0 je nesprávná –
H_0 je správná +	Test odhaduje skutečnost správně +	Test odhaduje skutečnost nesprávně – (chyba 1. druhu α)
H_0 je nesprávná –	Test odhaduje skutečnost nesprávně – (chyba 1. druhu β)	Test odhaduje skutečnost správně +

Ze čtyř možných situací jsou dvě žádoucí a dvě nežádoucí. Pravděpodobnosti chyb 1. a 2. druhu se nazývají rizika chyb a značí se α a β. Požadujeme, aby pravděpodobnost výskytu obou nežádoucích situací byla minimální.

Chyba 1. druhu α znamená pravděpodobnost zamítnutí správné hypotézy H_0. Nazývá se ***hladina významnosti*** zvoleného testu.

Chyba 2. druhu β znamená pravděpodobnost přijetí nesprávné hypotézy – neplatné H_0.

Hladina významnosti

Hodnotě $1-\beta$ říkáme síla nebo mohutnost testu a je to pravděpodobnost s jakou rozpoznáme nepravdivou hypotézu H_0.

Snížení chyby jednoho druhu má za následek zvýšení chyby druhého druhu. Snížení hodnot chyb 1. a 2. druhu dosáhneme jedině zvětšením rozsahu výběru. Protože se však většinou změna rozsahu výběru nedá uskutečnit, volí se chyba 1. druhu α pevně podle povahy daného experimentu buď $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$. Riziko chyb α a β musí být v rovnováze s rozsahem výběru.. Podle toho, jak se dosahuje této rovnováhy, lze rozlišovat dvě třídy testů testy s pevným rozsahem výběru a testy sekvenční. U testů s pevným rozsahem výběru se tento rozsah stanoví předem, před provedením pokusu spolu s rizikem chyb α a β. Druhou třídou testů jsou testy sekvenční, kdy se předepíše riziko chyby α a β a pokus se provádí tak dlouho, až se dospěje k rozhodnutí o oprávněnosti či neplatnosti nulové hypotézy. Riziko chyby α se nazývá hladina významnosti (signifikance) testu. Celá testovací statistika je postavena na této myšlence: Hypotézu odmítneme pouze tehdy, jestliže výběry dávají výsledky, které jsou při platnosti výchozí hypotézy nepravděpodobné. Při rozhodování o platnosti hypotézy si můžeme stanovit různě přísná kritéria. Pokud stačí, že v průměru 5 ze sta případů bude úsudek nesprávný, tak se rozhodneme pro pravděpodobnost chyby $\alpha = 0{,}05 = 5$ %. Pro zbylých 95 % případů bude výsledek statisticky významný (signifikantní). Jestliže kritérium ještě zpřísníme a budeme požadovat, aby pouze pro jeden ze sta případů byl úsudek nesprávný, pak se rozhodneme pro pravděpodobnost chyby $\alpha = 0{,}01 = 1$ % a výsledek bude statisticky významný pro 99 % případů.

Jestliže statistický test zamítne nulovou hypotézu H_0 jako nesprávnou, označí se výsledek jako statisticky významný (signifikantní), v opačném případě jako statisticky nevýznamný (nesignifikantní).

V tabulce je udána výrazová symbolika pro hladiny významnosti $\alpha = 0{,}05$; 0,01 a 0,001.

Pravděpod. chyby	> 0,05	≤ 0,05	≤ 0,01	≤ 0,001
Slovní vyjádření	nesignifikantní	signifikantní	vysoce signifikantní	velmi vysoce signifikantní
Písmenová symbolika	n.s.	s.	v.s.	v.v.s
Grafická symbolika		*	**.	***

Postup používaný při testování nulových hypotéz:

1. Formulace nulové a příslušné alternativní hypotézy.
2. Volba odpovídajícího testového kritéria (F-test, t-test).
3. Volba hladiny významnosti α (0,01 nebo 0,05).
4. Určení počtu stupňů volnosti v.
5. Výpočet hodnoty testového kritéria (podle vzorce).
6. Nalezení kritické (tabulkové) hodnoty k dané hladině významnosti a k danému počtu stupňů volnosti.
7. Porovnání hodnoty vypočítaného testového kritéria s kritickou hodnotou.
8. Rozhodnutí o zamítnutí či přijetí nulové hypotézy na základě tohoto porovnání.

Příklad

Výzkumný úkol: Zjistit rozdíl mezi muži a ženami v kvalitě jejich života pomocí dotazníku WHOQOL-BREF, který se skládá ze čtyř domén: PR – prostředí, SV – sociální vztahy, P – prožívání, F – fyzické zdraví.

Domény	Muži průměr	Ženy průměr	t	p	významnost
PR	14,2	16,7	3,28	< 0,05	sgn
SV	15,6	14,9	1,56	> 0,05	nsg
P	15,2	17,4	2,89	< 0,05	sgn
F	16,7	15,8	1,94	> 0,05	nsg

Pro formulaci hypotéz jsou dvě možnosti:

a) Zformulovat jednu hypotézu pro celý dotazník

H_0: Mezi muži a ženami v dotazníku kvality života není signifikantní rozdíl.

H_A: Mezi muži a ženami v dotazníku kvality života je signifikantní rozdíl.

Závěr – V doménách PR a P byl zjištěn signifikantní rozdíl mezi muži a ženami. Pro tyto domény H_0 zamítáme (H_A přijímáme). Pro domény SV a F H_0 přijímáme (H_A zamítáme).

b) Rozložit na čtyři samostatné hypotézy

H_{01}: V doméně PR není mezi muži a ženami signifikantní rozdíl.
H_{A1}: V doméně PR je mezi muži a ženami signifikantní rozdíl.
H_{02}: V doméně SV není mezi muži a ženami signifikantní rozdíl.
H_{A2}: V doméně SV je mezi muži a ženami signifikantní rozdíl.
H_{03}: V doméně P není mezi muži a ženami signifikantní rozdíl.
H_{A3}: V doméně P je mezi muži a ženami signifikantní rozdíl.
H_{04}: V doméně F není mezi muži a ženami signifikantní rozdíl.
H_{A4}: V doméně F je mezi muži a ženami signifikantní rozdíl.

Závěr – V doménách PR a P byl zjištěn signifikantní rozdíl mezi muži a ženami – zamítáme H_{01} a H_{03} (Přijímáme H_{A1} a H_{A3}).

5.2 Parametrické testy

Parametrický test vyžaduje určité podmínky, týkající se parametrů populace, ze které je pořízen výběr. Zpravidla se jedná o populaci s normálním rozložením četností zkoumaného znaku. Grafem rozložení četností musí být Gaussova křivka normálního rozložení.

V praxi se často setkáváme s problémem ověření určitého předpokladu. Chceme se například přesvědčit, zda zkoumaný výběr pochází z daného základního souboru, nebo zda dva náhodné výběry pocházejí ze stejného nebo z různých základních souborů, nebo zda je možné považovat studovaný soubor za náhodně uspořádaný atd.

Na tyto otázky odpovídáme pomocí parametrických testů významnosti. Je to skupina testů, ve které se ověřuje významnost rozdílu mezi dvěma veličinami. Při testování záleží na volbě hladiny významnosti (podle přesnosti jaké chceme v testu dosáhnout se volí $\alpha = 0{,}05$ nebo 0,01), na formulaci nulové hypotézy H_0 a na volbě vhodného testového kritéria (F-test, t-test).

5.2.1 Fisherův F-test

Tento parametrický test významnosti testuje hypotézy o populačním rozptylu. Umožňuje určit jak signifikantní je rozdíl mezi dvěma rozptyly. Například, když máme zjistit, zda jsou dvě skupiny osob homogenní nebo heterogenní při experimentálním zásahu. To znamená, že máme určit, zda pokusné osoby reagují na stejný zákrok velmi podobně nebo hodně rozdílně. **Homogenní soubor** má malou variabilitu (soubor je stejnorodý), v **heterogenním souboru** (nestejnorodém) jsou naměřená data značně rozptýlena. Je zde velká variabilita těchto dat.

Homogenní soubor

Heterogenní soubor

V souborech experimentálně naměřených hodnot je variabilita způsobována dvěma zdroji:

1. přirozenou rozdílností objektů v populaci (biologickou variabilitou),
2. chybami měření, které způsobují, že i při opakovaném měření jakékoliv veličiny na stejném objektu můžeme dostávat rozdílné hodnoty.

Buď můžeme porovnávat přirozenou variabilitu ve dvou populacích – např. vyrovnanost sportovního výkonu u sedmi a devítiletých dětí nebo homogennost reakce na daný podnět u mužů a žen nebo porovnáváme variabilitu způsobenou dvěma různými experimentálními zásahy. Například se porovnávají dvě vyšetření u stejných pacientů. Zjišťuje se pak variabilita dat při opakovaném měření na stejném subjektu. Měla by být minimální.

Tento problém lze pak formulovat jako statistický test hypotézy o rovnosti dvou rozptylů, který testuje nulovou hypotézu $H_0: \sigma_1^2 = \sigma_2^2$. Musí se však předpokládat, že měřená veličina má normální Gaussovo rozdělení.

Mějme dva výběry s rozsahy n_1 a n_2 a charakteristikami $\bar{x}_1, s_1$ a $\bar{x}_2, s_2$, které byly odebrány ze dvou základních souborů s parametry μ_1, σ_1 a μ_2, σ_2. U Fisherova F-testu se srovnávají dva rozptyly. K výpočtu ***testového kritéria*** F potřebujeme odhady rozptylů ZS a to odhad $\hat{\sigma}_1^2 = S_1^2$ a odhad $\hat{\sigma}_2^2 = S_2^2$.

Hodnota F se pak vypočítá jako poměr $F = \frac{S_1^2}{S_2^2}$,

Testové kritérium

přičemž do čitatele vkládáme větší z obou odhadů rozptylů, abychom pro veličinu F získali hodnotu $F \geq 1$.
Ke zjištění tabulkové hodnoty F na dané hladině významnosti potřebujeme znát počty stupňů volnosti v_1 a v_2. V tomto případě je $v_1 = n_1 - 1$ a $v_2 = n_2 - 1$. Ve statistických tabulkách najdeme kritickou hodnotu $F_\alpha(v_1, v_2)$, se kterou porovnáme vypočítané F.
Je-li $F \geq F_\alpha(v_1, v_2)$, zamítáme nulovou hypotézu a můžeme tvrdit, že mezi rozpyly obou výběrů je signifikantní rozdíl.

Postup při testování:

1. Předpokládáme, že platí nulová hypotéza $H_0: \sigma_1^2 = \sigma_2^2$ o rovnosti rozptylů.
2. Zvolíme si hladinu významnosti α (buď 0,05, nebo 0,01).
3. Vypočítané testové kritérium $F = \frac{S_1^2}{S_2^2}$ porovnáme s kritickou hodnotou $F_\alpha(v_1, v_2)$, kterou najdeme ve statistických tabulkách. Pozor – musí platit $S_1^2 > S_2^2$.

4. Jestliže zjistíme, že vypočítaná hodnota $F < F_{\alpha}(\nu_1,\nu_2)$, nastává případ, který jsme očekávali a nulovou hypotézu o rovnosti rozptylů nezamítáme.
6. Pokud $F \geq F_{\alpha}(\nu_1,\nu_2)$ nebo $P < 0{,}05$ (v Excelu), zamítáme nulovou hypotézu a tvrdíme, že rozdíl mezi rozptyly je statisticky významný na hladině významnosti α, a tedy můžeme přijmout alternativní hypotézu $H_0 : \sigma_1^2 \neq \sigma_2^2$.

Příklad

Posuďte, zda výběry pocházejí ze stejné populace. Odhady výběrových rozptylů jsou $S_1^2 = 64$, $S_2^2 = 25$ při počtech pozorování $n_1 = 30$ v prvním výběru a $n_2 = 20$ ve druhém výběru

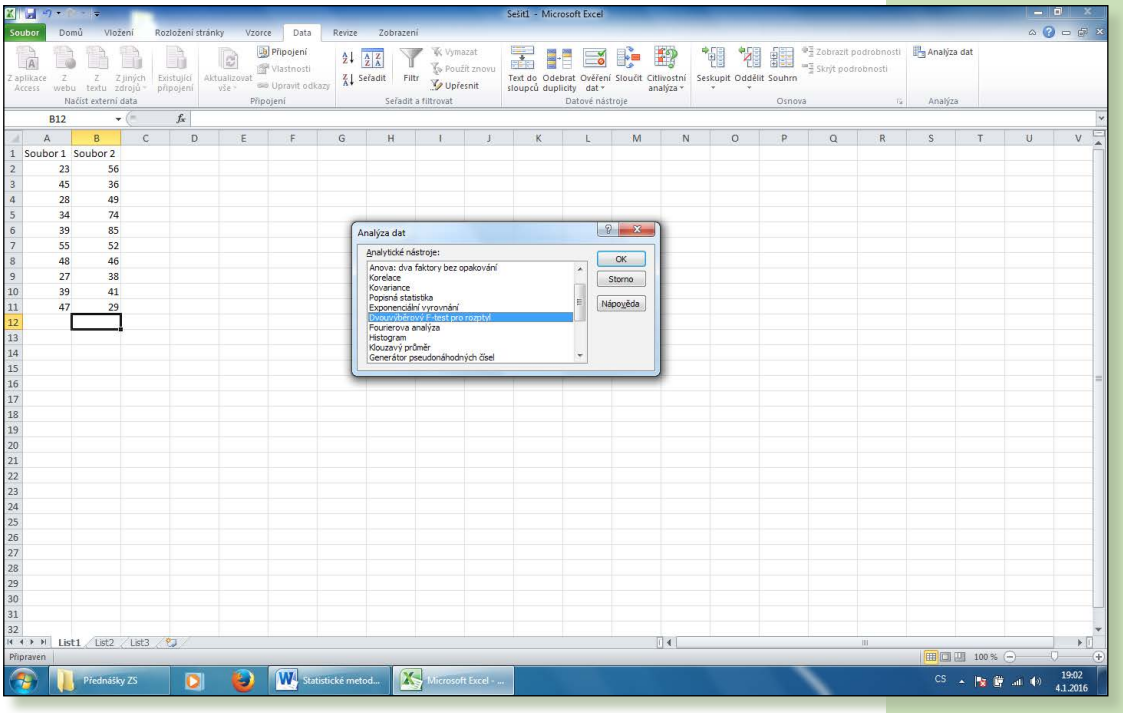
$F = \frac{S_1^2}{S_2^2} = \frac{64}{25} = 2{,}56$. Tabulková hodnota F_{α} na hladině významnosti $\alpha = 0{,}05$ je pro stupně volnosti $\nu_1 = 29$ a $\nu_2 = 19$ rovna

$F_{\alpha}(n_1 - 1, n_2 - 1) = F_{0{,}05}(29{,}19) = 2{,}039$.

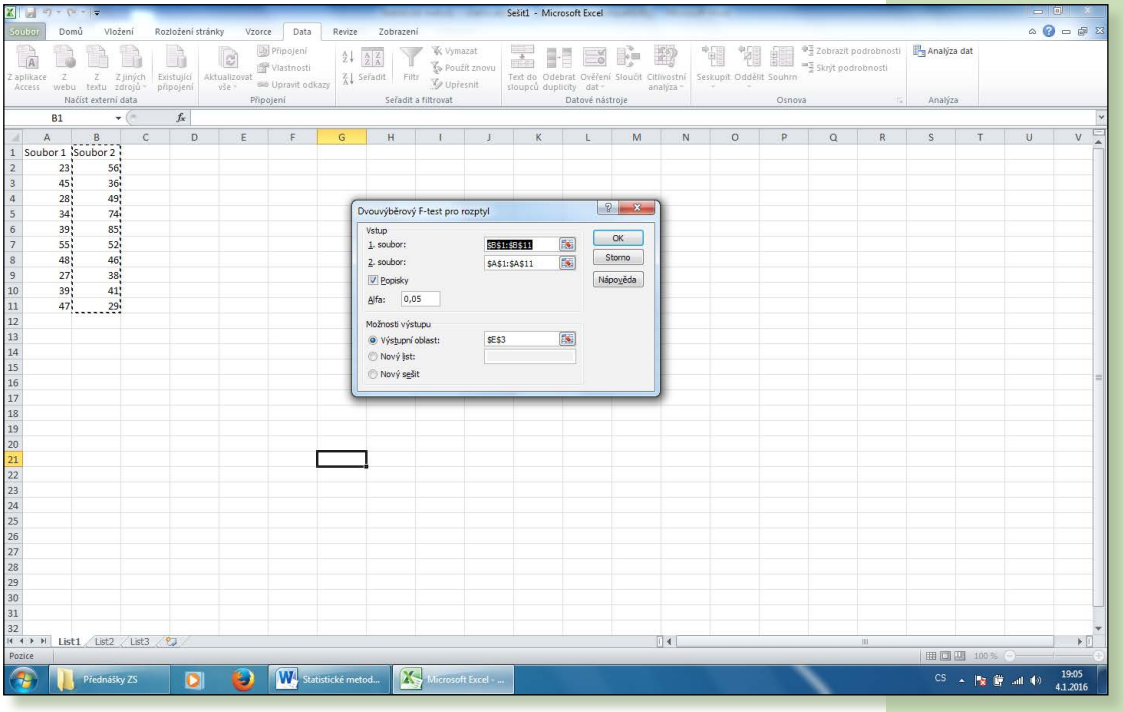
Jelikož $F > F_{\alpha}$, zamítá se nulová hypotéza o rovnosti rozptylů a můžeme říct, že mezi rozptyly je statisticky významný rozdíl.

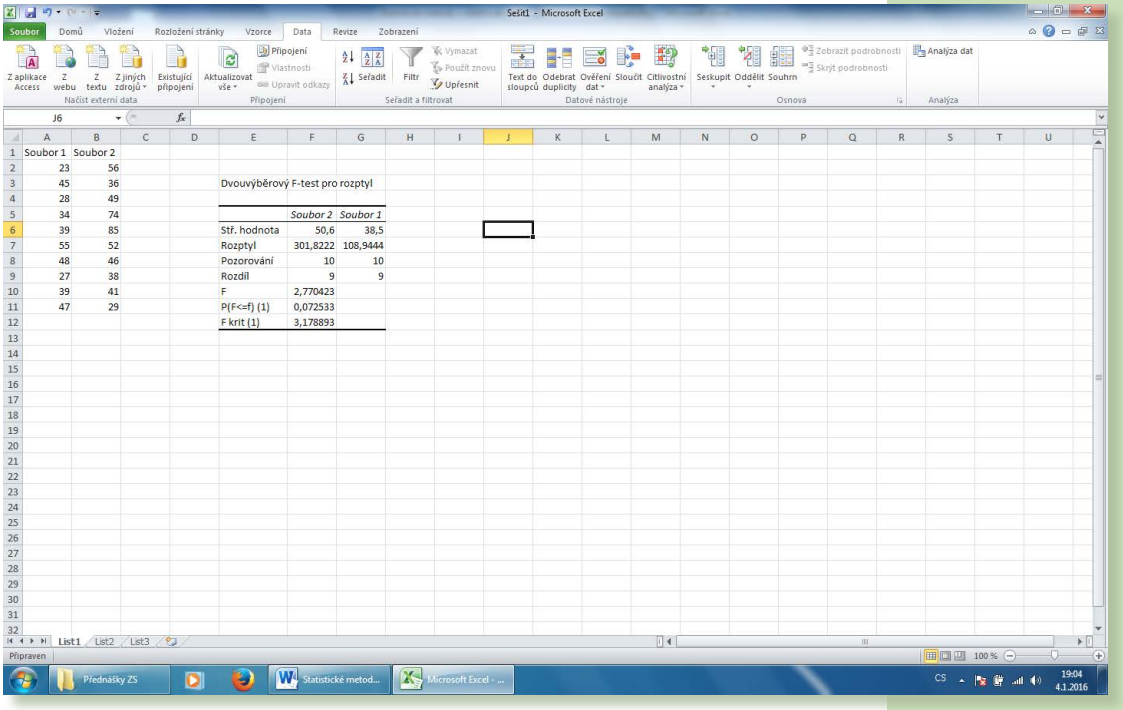
Fisherův F-test v programu Excel

Data → Analýza dat → Dvouvýběrový F-test pro rozptyl



Do vstupní oblasti se zadává každý výběrový soubor zvlášť. Pokud chceme mít ve výsledcích i popis sledovaných výběrů, zatrhneme Popisky. U F-testu se musí dávat pozor na pořadí zadávání souborů. Jako první musí být zadán soubor s větším rozptylem.





Soubor 1	Soubor 2
23	56
45	36
28	49
34	74
39	85
55	52
48	46
27	38
39	41
47	29

Dvouvýběrový F-test pro rozptyl

	Soubor 2	*Soubor 1*
Stř. hodnota	50,6	38,5
Rozptyl	301,8222	108,9444
Pozorování	10	10
Rozdíl	9	9
F	2,770423	
P(F<=f) (1)	0,072533	
F krit (1)	3,178893	

Ve výsledcích se objeví aritmetický průměr, rozptyl, počet pozorování, rozdíl – zde znamená počet stupňů volnosti, Fisherovo F, P je pravděpodobnost, kterou srovnáváme s hladinou významnosti α. Pokud $P < 0{,}05$, pak je potvrzen signifikantní rozdíl mezi rozptyly sledovaných výběrových souborů. Signifikantní rozdíl můžeme také zjistit porovnáním F a F krit (1).
Je-li $F > F$ krit (1), pak je potvrzen signifikantní rozdíl mezi rozptyly sledovaných výběrových souborů.

5.2.2 Studentovy t-testy

Studentův t-test

Studentovy t-testy jsou testy významnosti rozdílu dvou průměrů. Podobně jako lze hodnotit rozdíl mezi dvěma rozptyly F-testem, můžeme testovat i významnost rozdílu jiných dvou veličin, například průměrů.

Pomocí ***Studentova t-testu*** se nejčastěji řeší úloha, která se nazývá experiment, to znamená, že všichni pacienti, se kterými se pracuje, jsou zcela rovnocenní a liší se pouze náhodně. Například má být porovnán účinek dvou experimentálních zásahů, srovnává se účinek dvou různých léků nebo terapií, porovnává se efektivnost dvou různých léčebných programů, obtížnost dvou úkolů, výkonnost v nějaké sportovní disciplíně atd. Důležitým rysem, společným pro tyto situace je, že přiřazení vlastního experimentálního zásahu pokusnému objektu je provedeno náhodně.

Podle typu alternativní hypotézy můžeme t-testy dělit na jednostranné a dvoustranné. U dvoustranného testu se předpokládá, že nulovou hypotézu zamítáme, je-li odchylka od ní nepravděpodobně velká na jednu či druhou stranu. Alternativní hypotéza se pak dá rozložit na dvě jednostranné hypotézy. Testy, o kterých zde bude řeč, se používají u dvou výběrů a jsou si dosti podobné. Je proto nutné mezi nimi pečlivě volit. Záleží na počtu dat – Studentovo rozdělení je vhodnější pro výběry o menším rozsahu, jinak se používá rozdělení Gaussovo. Záleží také na tom, zda se rozptyly výběrů statisticky významně liší nebo nikoliv (zda jsou data heterogenní nebo homogenní), zda jsou naměřené hodnoty spárovány a také na tom, zda jsou nebo nejsou navzájem korelovány (zda se jedná o závislé nebo nezávislé výběry). Při testu významnosti rozdílu dvou průměrů je nulovou hypotézou rovnost průměrů.

Podle porovnávaných parametrů rozlišujeme pak tyto typy t-testů:

1. t-test rozdílu výběrového průměru a známého průměru základního souboru;
2. t-test pro srovnání rozdílu dvou středních hodnot:
   a) t-test pro rozdíl dvou výběrových průměrů, jestliže F-testem ověříme, že $\sigma_1^2 = \sigma_2^2$;

b) t-test pro rozdíl dvou výběrových průměrů, jestliže F-testem ověříme, že $\sigma_1^2 \neq \sigma_2^2$;
3. párový t-test;
4. test rozdílu dvou relativních hodnot.

5.2.2.1 Jednovýběrový t-test

Tento test ověřuje, zda platí nulová hypotéza H_0: $\bar{x} = \mu$. V praxi se setkáváme s případy, kdy známe určitou konstantní hodnotu, kterou můžeme považovat za průměr základního souboru μ. Této hodnotě se říká ***referenční konstanta***. Jestliže provedeme výběr ze základního souboru a vypočítáme výběrový průměr $\bar{x}$, pak se ptáme, zda je nebo není statisticky významný (signifikantní) rozdíl mezi výběrovým průměrem a průměrem základního souboru. Nulová hypotéza nám tvrdí, že průměr základního souboru a výběrový průměr se shodují (H_0: $\bar{x} = \mu$).

Referenční konstanta

Testovým kritériem je veličina při $\nu = n - 1$stupních volnosti. $t = \frac{|\bar{x} - \mu| . \sqrt{n}}{s}$

Vypočítané t pak porovnáváme s kritickou hodnotou $t_\alpha(\nu)$. Je- li $t > t_\alpha(\nu)$ nebo $P < 0{,}05$ (v Excelu) zamítáme nulovou hypotézu a můžeme říct, že výběrový průměr se na zvolené hladině významnosti statisticky významně liší od známé hodnoty průměru základního souboru. V opačném případě se nám potvrdí nulová hypotéza o rovnosti výběrového průměru a průměru základního souboru. Předpokládá se, že základní soubor má normální rozdělení četností.

Příklad

Zjistěte, zda výběr dvaceti pacientů, u nichž byly naměřeny hodnoty dané v tabulce, pochází ze základního souboru s průměrem $\mu = 58$.

1. Zvolíme hladinu významnosti $\alpha = 0{,}05$
2. Výběrová směrodatná odchylka $s = 7$, výběrový průměr

$$\bar{x} = \frac{1208}{20} = 60{,}4$$

Pacient	x_i	Pacient	x_i
S1	64	S11	60
S2	48	S12	43
S3	55	S13	67
S4	68	S14	70
S5	72	S15	65
S6	59	S16	55
S7	57	S17	56
S8	61	S18	64
S9	63	S19	61
S10	60	S20	60
Σ			1208

3. Ověříme, zda je signifikantní rozdíl mezi výběrovým průměrem a průměrem ZS pomocí testového kritéria

$$t = \frac{|\bar{x} - \mu| \cdot \sqrt{n}}{s} = \frac{|60{,}4 - 58| \cdot \sqrt{20}}{7} = \frac{2{,}4 \cdot 4{,}47}{7} = \frac{10{,}73}{7} = 1{,}53 .$$

4. Tabulková hodnota $t_\alpha(\nu)$ je $t_{0,05}(19) = 2{,}093$.
5. Porovnáme vypočítanou hodnotu t s tabulkovou hodnotou $t_\alpha(\nu)$: $t = 1{,}53 < t_\alpha(\nu) = 2{,}093$ pak nezamítáme nulovou hypotézu H0: a můžeme říct, že není statisticky významný rozdíl mezi výběrovým průměrem a průměrem ZS. Zjistili jsme, že soubor dvaceti subjektů je vybrán ze základního souboru s průměrem $\mu = 58$.

5.2.2.2 Dvouvýběrový t-test pro homogenní soubory

Budeme předpokládat, že platí nulová hypotéza, která říká, že rozdíly mezi průměry dvou výběrů jsou způsobeny pouze náhodou: $H_0 : \bar{x}_1 = \bar{x}_2$ a že rozptyly obou základních souborů, ze kterých pocházejí výběry jsou stejné (musí se potvrdit F-testem). Oba výběry tedy budou pocházet z jediného základního souboru s průměrem μ a rozptylem σ^2. Za tohoto předpokladu budeme očekávat, že při častějším opakování pokusu se dvěma libovolnými výběry bude jednou průměr jedné skupiny větší než průměr skupiny druhé a obráceně. Protože pozitivní a negativní rozdíly jsou stejně pravděpodobné, měl by průměr rozdílů mezi vždy dvěma výběrovými průměry být roven nule.

Skutečně se zjistilo, že rozdíly $\bar{x}_1 - \bar{x}_2, \bar{x}_1 - \bar{x}_3, \bar{x}_2 - \bar{x}_3, \ldots$ mezi průměry dvou náhodných výběrů z jednoho ZS mají normální rozložení s očekávaným průměrem μ_d ($d = \bar{x}_1 - \bar{x}_2$ ...diference) neboli $\mu_d = \mu_{\bar{x}_1 - \bar{x}_2} = 0$. Chybový rozptyl rozdílu průměrů je roven při nezávislých výběrech součtu rozptylů obou výběrových průměrů:

$$\sigma_d^2 = \sigma_{\bar{x}_1 - \bar{x}_2}^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Mají-li oba výběry stejný rozsah, pak jsou i $\sigma_{\bar{x}_1}^2$ a $\sigma_{\bar{x}_2}^2$ stejně velké a vzorec se zjednoduší na tvar:

$$\sigma_d^2 = 2\sigma_{\bar{x}}^2 = \frac{2\sigma^2}{n}$$

Výběrová chyba rozdílu průměrů při výběrech o různém rozsahu

je pak $\sigma_d = \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$.

Výběrová chyba rozdílu průměrů při výběrech o stejném rozsahu je $\sigma_d = \sqrt{\frac{2\sigma^2}{n}}$.

V konkrétním případě však nikdy neznáme rozptyl ZS, ale pouze dva jeho odhady – rozptyly obou výběrů:

$$s_1^2 = \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x}_1)^2}{n_1 - 1} \quad \text{a} \quad s_2^2 = \hat{\sigma}^2 = \frac{\sum (x_j - \bar{x}_2)^2}{n_2 - 1} .$$

Použijeme-li jako nejlepšího odhadu pro σ^2 váženého aritmetického průměru z těchto obou rozptylů, dostaneme odhad

$\sigma^2 = \frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$. Tuto hodnotu dosadíme do vzorce $\sigma_d = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$ a dostaneme konečný vzorec pro výběrovou chybu rozdílů průměrů dvou výběrů jednoho souboru: $\sigma_d = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n_1} + \sum x_j^2 - \frac{(\sum x_j)^2}{n_2}}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

(Pro ulehčení výpočtu se odchylky od průměru počítají přímo z naměřených hodnot:

$\sum (x_i - \bar{x}_1)^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n_1}$; $\sum (x_j - \bar{x}_2)^2 = \sum x_j^2 - \frac{(\sum x_j)^2}{n_2}$.)

V jednotkách Studentova *t*-rozložení bude

$t = \frac{X - \mu}{\sigma} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sigma_d}$, jelikož $\mu_d = \mu_{\bar{x}_1 - \bar{x}_2} = 0$,

pak bude mít testové kritérium Studentova t-testu pro dva nezávislé výběry tvar

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n_1} + \sum x_j^2 - \frac{(\sum x_j)^2}{n_2}}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Toto testové kritérium se používá v případě, že se F-testem prokáže rovnost rozptylů základních souborů.
Vypočítanou hodnotu t dále porovnáme s tabulkovou hodnotou $t_\alpha(\nu)$ na hladině významnosti α = 0,05 nebo α = 0,01 a při počtu stupňů volnosti $\nu = n_1 + n_2 - 2$. Je-li $t < t_\alpha(\nu)$ přijímáme nulovou hypotézu $H_0 : \bar{x}_1 = \bar{x}_2$ a můžeme říct, že mezi výběrovými průměry není statisticky významný rozdíl.

Je-li $t > t_\alpha(\nu)$, zamítáme H_0 a tvrdíme, že platí alternativní hypotéza H_1: $\bar{x}_1 \neq \bar{x}_2$. V tomto případě je mezi výběrovými průměry $\bar{x}_1, \bar{x}_2$ statisticky významný (signifikantní) rozdíl.

Příklad

Máme zjistit, zda je statisticky významný rozdíl mezi výsledky vyšetření mužů a žen.

Muži	skór-x_i	x_i^2	Ženy	skór-x_j	x_j^2
M1	15	225	F1	20	400
M2	26	676	F2	22	484
M3	32	1024	F3	27	729
M4	48	2304	F4	31	961
M5	52	2704	F5	38	1444
M6	63	3969	F6	44	1936
M7	72	5184	F7	46	2116
M8	80	6400	F8	57	3249
M9	86	7396	F9	59	3481
M10	85	7225	F10	65	4225
M11	89	7921	F11	74	5476
			F12	77	5929
Σ	648	45028		560	30430

1. F-testem prokážeme rovnost rozptylů ZS.
   Vypočítáme odhady rozptylů ZS:

$$S_1^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n_1}}{n_1 - 1} = \frac{45028 - \frac{648^2}{11}}{10} = 685{,}5$$

$$S_2^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n_2}}{n_2 - 1} = \frac{30430 - \frac{560^2}{12}}{11} = 390{,}6$$

$$F = \frac{S_2^2}{S_1^2} = \frac{685{,}5}{390{,}6} = 1{,}755$$

$$F_\alpha(\nu_1, \nu_2) = F_{0{,}05}(10{,}11) = 2{,}854$$

Jelikož $F < F_\alpha(\nu_1, \nu_2)$, přijímá se nulová hypotéza o rovnosti rozptylů ZS.

2. Vypočítáme výběrové průměry a doplníme sumy v tabulce.

$$\bar{x}_1 = \frac{648}{11} = 58{,}9 \quad , n_1 = 11,\ n_2 = 12$$

$$\bar{x}_2 = \frac{560}{12} = 46{,}7$$

Dosadíme do vzorce pro výpočet testového kritéria t Studentova t-testu:

$$t = \frac{|58{,}9 - 46{,}7|}{\sqrt{\frac{45028 - \frac{648^2}{11} + 30430 - \frac{560^2}{12}}{11+12-2}\left(\frac{1}{11}+\frac{1}{12}\right)}} = \frac{12{,}2}{\sqrt{90{,}27}} = \frac{12{,}2}{9{,}5} = 1{,}28$$

Tuto vypočítanou hodnotu porovnáme s tabulkovou hodnotou $t_\alpha(\nu)$, $\nu = 21$. Pro $\alpha = 0{,}05$ je $t_{0{,}05}(21) = 2{,}08$.
Jelikož je $t = 1{,}28 < t_{0{,}05}(21) = 2{,}08$ můžeme říct, že není statisticky významný rozdíl mezi průměry skórů skupin mužů a žen.

5.2.2.3 Dvouvýběrový t-test pro nehomogenní soubory

Je-li splněn požadavek, aby základní soubory měly alespoň přibližně normální rozdělení, a zjistí-li se F-testem, že mezi rozptyly je statisticky významný rozdíl, používá se pro testování významnosti rozdílu průměrů při nulové hypotéze $H_0 : \bar{x}_1 = \bar{x}_2$ testovacího kritéria $t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$.

Tuto hodnotu opět porovnáváme s kritickou hodnotou, která se pro tento případ označí t_α^* a musí se vypočítat podle vzorce

$$t_\alpha^* = \frac{t'_\alpha \frac{s_1^2}{n_1 - 1} + t''_\alpha \frac{s_2^2}{n_2 - 1}}{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}},$$

kde t'_α značí kritickou hodnotu t rozdělení pro $\nu_1 = n_1 - 1$ stupňů volnosti a t''_α kritickou hodnotu pro $\nu_2 = n_2 - 1$. Obě hodnoty se vyhledají ve statistických tabulkách.

Postup při testování významnosti rozdílu průměrů, jestliže $\sigma_1^2 \neq \sigma_2^2$:
a) zvolíme hladinu významnosti α;
b) z obou výběrů vypočítáme charakteristiky $\bar{x}_1, s_1^2, \bar{x}_2, s_2^2$;
c) F-testem prokážeme, že zamítáme nulovou hypotézu H_0: $\sigma_1^2 = \sigma_2^2$, to znamená, že přijímáme alternativní hypotézu H_1: $\sigma_1^2 \neq \sigma_2^2$;

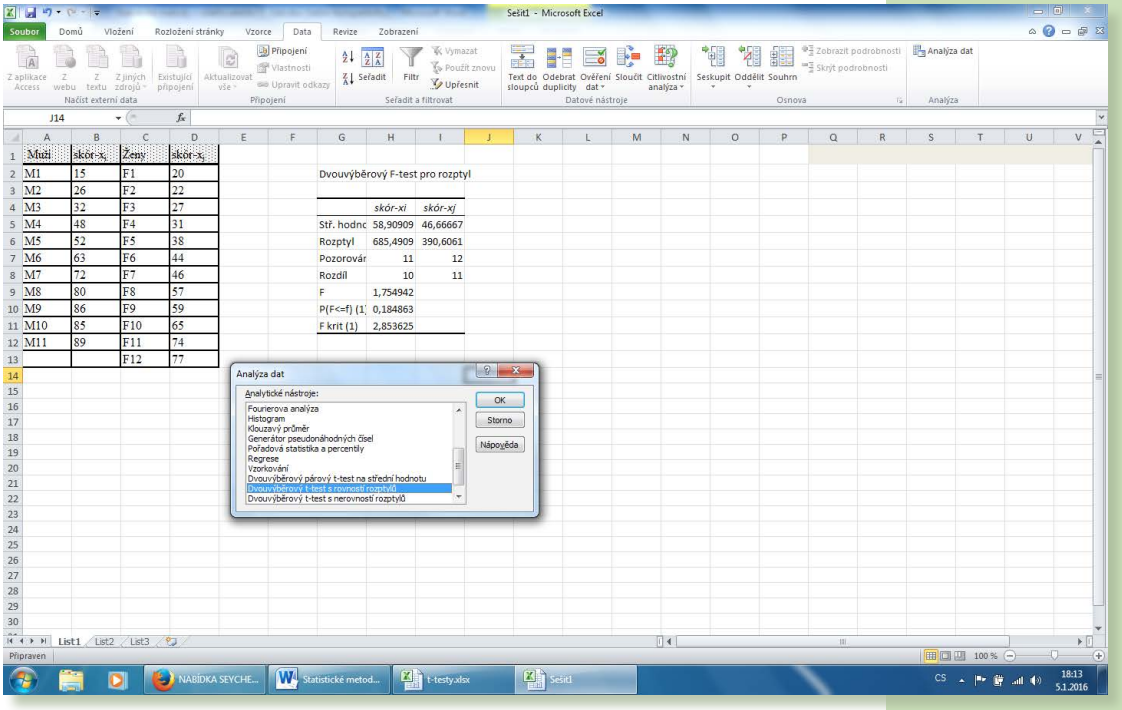
d) vypočítáme hodnotu testovacího kritéria t;
e) ve statistických tabulkách vyhledáme hodnoty t'_α a t''_α pro $\nu_1 = n_1 - 1$ a $\nu_2 = n_2 - 1$ pro stupňů volnosti;
f) vypočítáme kritickou hodnotu t_α^*;
g) nulovou hypotézu $H_0 : \bar{x}_1 = \bar{x}_2$ zamítáme, když $t > t_\alpha^*$; v tom případě tvrdíme, že rozdíl průměrů dvou výběrů je statisticky významný na hladině významnosti α.

Studentovy t-testy v Excelu

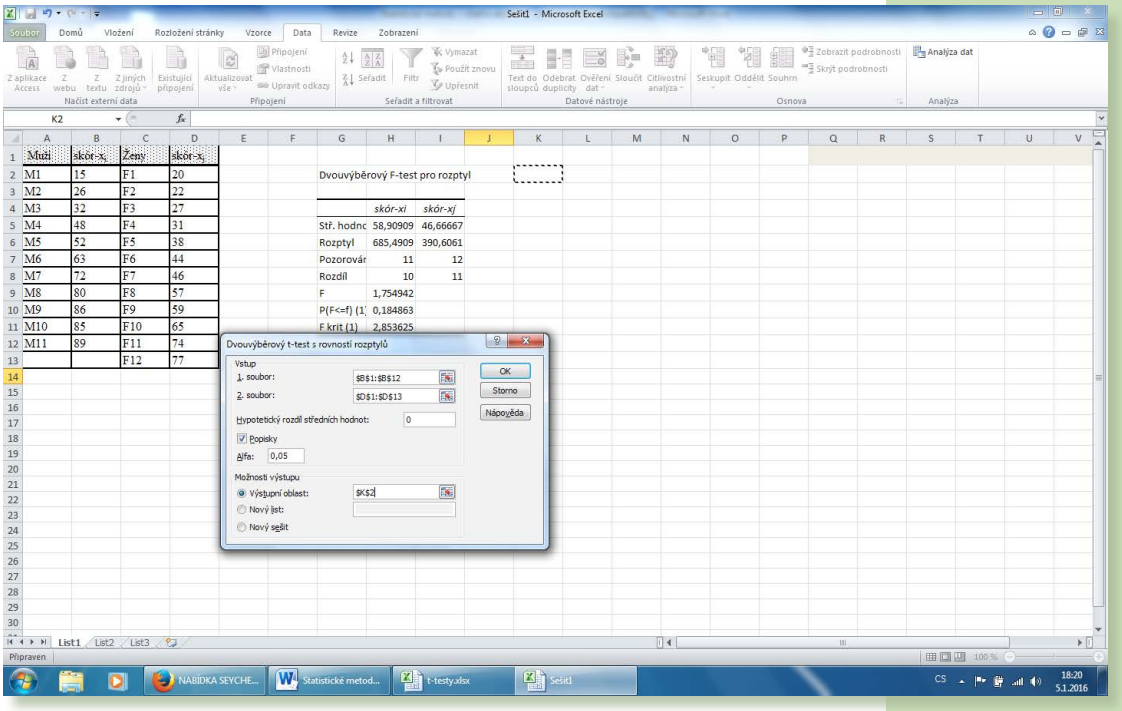
Před Studentovým t-testem nejprve ověříme rozptyly sledovaných výběrů.

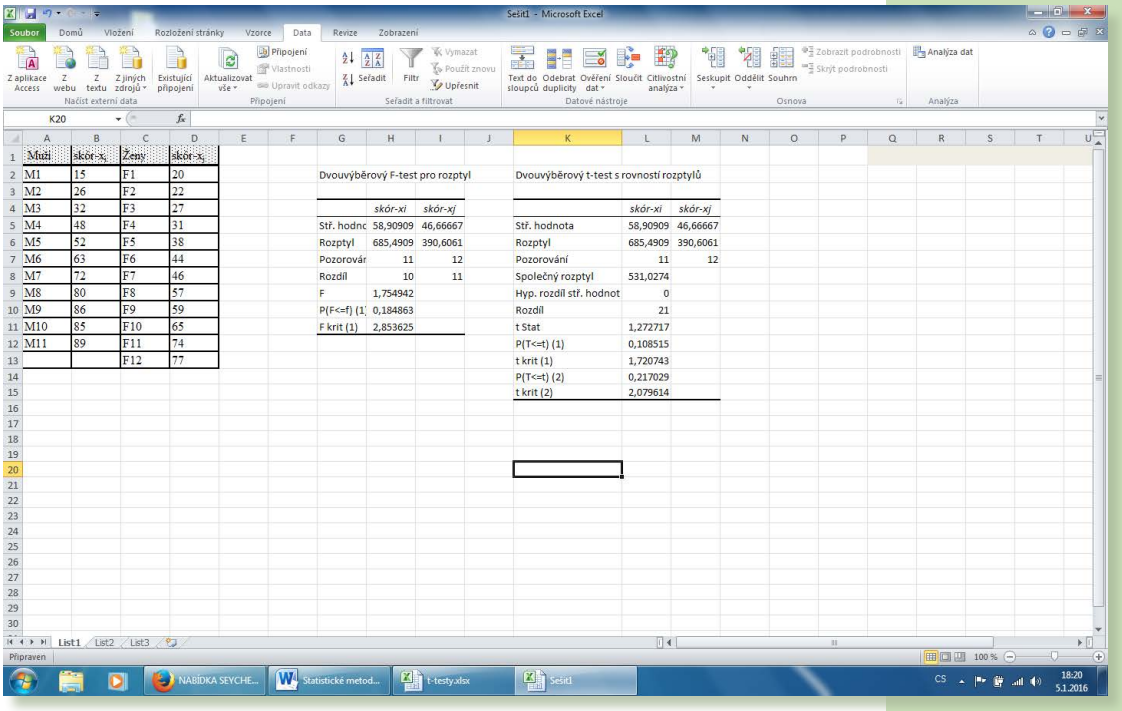
Data → Analýza dat → Dvouvýběrový F-test pro rozptyl

Na základě výsledku F-testu určíme typ Studentova t-testu, v našem případě není signifikantní rozdíl mezi rozptyly – vybereme Dvouvýběrový t-test s rovností rozptylů.



Zadávání dat je podobné jako u F-testu. Do pole pro hypotetický rozdíl středních hodnot se vždy zadává nula.





Ve výsledcích je kromě aritmetického průměru, rozptylů obou souborů a počtu pozorování uveden tzv. rozdíl, který ve skutečnosti znamená počet stupňů volnosti. Vypočítanou hodnotu Studentova t – t-Stat v absolutní hodnotě porovnáme s hodnotou t krit (2), která představuje tabulkovou kritickou hodnotu. Je-li $|t\ \text{stat}| \leq t\ \text{krit}\ (2)$, pak přijímáme nulovou hypotézu a můžeme říct, že není signifikantní rozdíl v průměrech sledovaných

souborů. Zda je nebo není statisticky významný rozdíl mezi průměry poznáme také podle velikosti pravděpodobnosti $P\ (T \le t)$ (2), kterou porovnáme s hladinou významnosti $\alpha = 0{,}05$. Je-li $P\ (T \le t)\ (2) > 0{,}05$, pak výsledek není signifikantní – není signifikantní rozdíl v průměrech sledovaných souborů.

V případě, že F-test potvrdí signifikantní rozdíl mezi rozptyly sledovaných souborů, použijeme Dvouvýběrový t-test s nerovností rozptylů.

5.2.2.4 Párový t-test

Tento t-test se používá při testování hypotézy o tom, zda při měření nějaké veličiny na subjektech nebo objektech došlo ke změně. Jedná se tedy o měření, která provádíme na jednom subjektu nebo objektu dvakrát, obvykle na začátku a na konci určitého procesu, nebo v případě, že každý prvek jednoho výběru tvoří pár s jedním zcela určitým prvkem výběru druhého. V těchto případech se bere místo dvou výběrů po n prvcích n párů měření. Protože ***párová měření*** jsou na sobě závislá, mluví se také o dvou závislých výběrech. U nich nemá význam počítat průměr hodnot jednoho výběru před sledovaným procesem a srovnávat jej s průměrem hodnot druhého výběru po tomto procesu, ale je nutno stanovit rozdíly naměřených hodnot v každém páru $x_{1i} - x_{2i}$ a s nimi dále pracovat jako s náhodnou veličinou. Tyto rozdíly se značí d_i a nazývají se ***diference***. Ptáme se pak, zda je průměrná hodnota vypočítaných rozdílů statisticky významně odlišná od nuly. Nulová hypotéza tedy zní: „Průměr rozdílu naměřených hodnot ve dvou výběrech je nula." (H_0: $\bar{d} = 0$).

Párová měření

Diference

Předpokládejme, že jsme ze dvou normálně rozdělených základních souborů s parametry μ_1, σ_1 a μ_2, σ_2 odebrali po jednou výběru. Rozsahy obou výběrů jsou stejné (n = n_1 = n_2) a jejich prvky tvoří páry.

Nulová hypotéza je v tomto případě H_0: $\bar{d} = 0$ (ekviv. H_0: $\bar{x}_1 = \bar{x}_2$).

Tuto hypotézu pak testujeme pomocí testového kritéria $t = \dfrac{|\bar{d}| \cdot \sqrt{n}}{s_d}$, kde $\bar{d} = \dfrac{1}{n}\sum_{i=1}^{n} d_i$

a s_d je směrodatná odchylka pro rozdíly naměřených hodnot. Vypočítanou hodnotu t pak porovnáváme s kritickou hodnotou $t_\alpha(\nu)$, kterou pro zvolenou hladinu významnosti α a počet stupňů volnosti ν = n – 1 najdeme ve statistických tabulkách. Je-li $t < t_\alpha(\nu)$, pak přijímáme nulovou hypotézu H_0: $\bar{d} = 0$.

V opačném případě H_0 zamítáme a můžeme říct, že je mezi výběrovými průměry statisticky významný rozdíl.

Postup:

1. Zvolíme hladinu významnosti α.
2. Vypočítáme rozdíly d_i mezi naměřenými párovými hodnotami. Vypočítáme průměr $\bar{d}$ a směrodatnou odchylku s_d těchto rozdílů.
3. Vypočítáme hodnotu testového kritéria t a stanovíme počet stupňů volnosti ν.
4. Ve statistických tabulkách vyhledáme pro dané νa zvolené α kritickou hodnotu $t_\alpha(\nu)$.
5. Na základě porovnání vypočítané hodnoty t a kritické hodnoty t_α nebo hodnoty P s hladinou významnosti 0,05 (v Excelu) ne/zamítneme nulovou hypotézu H_0.

Příklad

Zjistěte, zda redukční dieta způsobila signifikantní úbytek váhy u deseti pacientů. Údaje o hmotnosti na začátku diety a na jejím konci jsou v tabulce.

Pacient	Začátek diety	Konec diety	d_i	d_i^2
A	68	65	3	9
B	71	69	2	4
C	69	64	5	25
D	88	80	8	64
E	85	81	4	16
F	77	74	3	9
G	80	75	5	25
H	72	71	1	1
I	67	65	2	4
J	75	70	5	25
Σ			38	182

$$\alpha = 0{,}05, \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = 3{,}8$$

$$s_d = \sqrt{\frac{\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}}{n-1}} = \sqrt{\frac{182 - \frac{38^2}{10}}{9}} = \sqrt{\frac{182 - 144{,}4}{9}} = \sqrt{\frac{37{,}6}{9}} = \sqrt{4{,}17} = 2$$

$$t = \frac{|\bar{d}|.\sqrt{n}}{s_d} = \frac{3{,}8.\sqrt{10}}{2} = \frac{3{,}8.3}{2} = \frac{11{,}4}{2} = 5{,}7$$

$t = 5{,}7 > t_\alpha(\nu) = t_{0{,}01}(9) = 3{,}24$

Na základě porovnání t a t_α zamítáme nulovou hypotézu a můžeme tvrdit, že dieta způsobila signifikantní úbytek váhy.

Párový t-test v Excelu
Data → Analýza dat → Dvouvýběrový párový t-test na střední hodnotu.
Zadávání dat a interpretace výsledků je podobná jako u předcházejícího testu.

5.2.2.5 Studentův t-test rozdílu dvou relativních hodnot

V praxi se často řeší otázka, zda se určitý jev vyskytuje v jednom výběru častěji než ve druhém. Údaje bývají uváděny formou ***relativních hodnot***. Ptáme se tedy, zda je ve výskytu nějakého jevu ve dvou výběrech statisticky významný rozdíl. Nejdříve stanovíme ***relativní četnost*** výskytu sledovaného jevu v prvním výběru $f_1 = \frac{m_1}{n_1}$, kde m_1 je počet případů, u nichž se zkoumaný jev vyskytl a n_1 je počet všech sledovaných případů v prvním výběru. Podobně stanovíme hodnotu $f_2 = \frac{m_2}{n_2}$, kde m_2 je počet případů, u nichž se zkoumaný jev vyskytl a n_2 je počet všech sledovaných případů ve druhém výběru. V tomto případě testujeme nulovou hypotézu o tom, že oba výběry pocházejí ze stejného základního souboru. Neznámou hodnotu relativního výskytu sledovaného jevu v základním souboru nahrazujeme jejím odhadem z četností obou výběrů podle vztahu

$$\hat{f} = \frac{m_1 + m_2}{n_1 + n_2}.$$

Nulovou hypotézu $H_0: f_1 = f_2$ ověřujeme pomocí testového kritéria

$$t = \frac{|f_1 - f_2|}{\sqrt{\hat{f}(1-\hat{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Tomuto testovému kritériu přísluší při dostatečně velkých n_1 a n_2 (větších než 30) normální rozdělení četností s nulovým průměrem a jednotkovou směrodatnou odchylkou. Testové kritérium tedy můžeme porovnat s kritickými hodnotami z_α=1,96 (pro $\alpha = 0{,}05$) a z_α=2,58 (pro $\alpha = 0{,}01$).
Jestliže $t > z_\alpha$, zamítneme nulovou hypotézu a tvrdíme, že se výskyt sledovaného jevu ve dvou výběrech statisticky významně liší na zvolené hladině významnosti, tedy jinými slovy, že oba výběry nepocházejí z téhož základního souboru.

Příklad

Byla řešena otázka, zda se určité onemocnění vyskytuje častěji u mužů než u žen. Po příslušném vyšetření bylo nalezeno 25 onemocnění u 54 mužů a 20 onemocnění u 47 žen z předem vybrané lokality. Prokažte, zda je vyšší relativní četnost onemocnění u mužů proti ženám statisticky významná.

1. $\alpha = 0{,}05$
2. $f_1 = \frac{m_1}{n_1} = \frac{25}{54} = 0{,}46$; $f_2 = \frac{m_2}{n_2} = \frac{20}{47} = 0{,}42$
3. $\hat{f} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{25+20}{54+47} = \frac{45}{101} = 0{,}45$
4. $t = \frac{|f_1 - f_2|}{\sqrt{\hat{f}(1-\hat{f}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0{,}46 - 0{,}42}{\sqrt{0{,}45(1-0{,}45) \cdot \left(\frac{1}{54} + \frac{1}{47}\right)}} = \frac{0{,}04}{0{,}99} = 0{,}04$
5. Jelikož $t = 0{,}04 < z_\alpha = 1{,}96$ pro $\alpha = 0{,}05$, můžeme říct, že ve výskytu sledovaného onemocnění není statisticky významný rozdíl mezi muži a ženami.

5.2.2.6 Studentův t-test pro signifikantnost korelačního koeficientu

Statistickou významnost korelačního koeficientu je možno ověřovat Studentovým t-testem podle vzorce

$$t = r\sqrt{\frac{(n-2)}{(1-r^2)}},$$

kde r je ověřovaný korelační koeficient, n je počet párových měření.

Vypočítanou hodnotu t porovnáváme s tabulkovou hodnotou $t_\alpha(\nu)$, kde $\nu = n - 2$ je počet stupňů volnosti. Pokud Studentovo t přesahuje tabulkovou hodnotu na hladině významnosti $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$, je korelační koeficient signifikantní.

Statistickou významnost korelace lze také vyhledat přímo ve statistických tabulkách, kde pro počet párových měření n a hladinu významnosti α nalezneme nejnižší hodnotu, které by měl korelační koeficient dosáhnout, aby byl signifikantní (viz Tabulka kritických hodnot korelačního koeficientu).

5.2.2.7 Procentový z-test

Používá se pro zjištění statistické významnosti rozdílu mezi procenty výskytu sledovaného jevu u dvou výběrových souborů.

P_1… procento výskytu sledovaného jevu v prvním výběru
P_2… procento výskytu sledovaného jevu ve druhém výběru
n_1…rozsah prvního výběru
n_2…rozsah druhého výběru
$Q_1 = 100 - P_1$
$Q_2 = 100 - P_2$

$$z = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

Vypočítaná hodnota z se porovnává se standardizovanou hodnotou z_α:
$z_{0,05} = 1,96$
$z_{0,01} = 2,58$
$z_{0,001} = 3,29$

Je-li $z \geq z_\alpha$, pak je signifikantní rozdíl mezi procenty sledovaného jevu ve dvou výběrech.

Příklad
Ve dvou nemocnicích se sledovalo procento výskytu určité nemoci. V první nemocnici z celkového počtu 300 pacientů onemocnělo 15 %, ve druhé z celkového počtu 200 pacientů onemocnělo 35 %. Zjistěte, zda je signifikantní rozdíl v procentech výskytu onemocnění mezi nemocnicemi.

$P_1 = 15\ \%$
$P_2 = 35\ \%$
$n_1 = 300\ (100\ \%)$
$n_2 = 200\ (100\ \%)$
$Q_1 = 100 - P_1 = 100 - 15 = 85$
$Q_2 = 100 - P_2 = 100 - 35 = 65$

$$z = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} = \frac{|15 - 35|}{\sqrt{\frac{15 \cdot 85}{300} + \frac{35 \cdot 65}{200}}} =$$

$$= \frac{|-20|}{\sqrt{4,25 + 11,375}} = \frac{20}{3,95} = 5,06$$

Vypočítaná hodnota z se porovnává se standardizovanou hodnotou $z_{0,001} = 3,29$; $z \geq z_{0,001}$

Existuje velmi vysoce signifikantní rozdíl mezi procenty sledovaného jevu ve sledovaných výběrech.

Kontrolní otázky a úkoly

1. Definujte pojem hypotéza a statistická hypotéza
2. Jaký je rozdíl mezi parametrickými a neparametrickými testy?
3. Které parametry testují Studentovy *t*-testy?

Klíč k otázkám a úkolům

Odpovědi na otázky najdete v textu

Referenční seznam

HENDL, J. 2004. *Přehled statistických metod zpracování dat.* Praha: Portál. ISBN 80-7178-820-1.

WALKER, I. 2013. *Výzkumné metody a statistika*. Praha: Grada. ISBN 978-80-247-3920-5.

ZVÁROVÁ, J. 2001. *Základy statistiky pro biomedicínské obory.* Praha: Karolinum. ISBN 80-7184-786-0.

5.3 Testy χ^2

Testy χ^2 patří do samostatné skupiny testů, které tvoří přechod mezi parametrickými a neparametrickými metodami. Používají se při testování významnosti rozdílu mezi četnostmi v kategoriálních datech.

5.3.1 Test shody χ^2

Test shody χ^2

Test shody χ^2 řeší otázku shody rozdělení. Tímto testem se ověřuje nulová hypotéza o tom, že empirická pozorování jsou v souladu s předpoklady o pravděpodobnostním rozdělení určitého znaku. Oprávněnost hypotézy se ověřuje testovým kritériem, které zjišťuje rozdíly mezi empiricky pozorovaným rozdělením četností a teoretickým rozdělením pravděpodobností. Toto kritérium zavedl Pearson – proto se někdy tento test označuje jako ***Pearsonův*** χ^2.

Pearsonův χ^2

Při výpočtu χ^2 předpokládáme, že výsledky pozorování roztřídíme určitým způsobem (např. skupinovým rozdělením četností). Tak získáme v jednotlivých třídách počty hodnot, které se označí jako ***experimentální četnosti*** O_i, protože podávají informaci, ke které jsme dospěli experimentální cestou. Dále si musíme zvolit určité rozdělení, které budeme považovat za model pro náš výběr. Pomocí tohoto rozdělení stanovíme tzv. ***očekávané*** (modelové) ***četnosti*** E_i. Smysl testu je pak v tom, že hodnotíme rozdíly mezi jednotlivými četnostmi experimentálními a očekávanými, tj rozdíly $O_i - E_i$. Za nulovou hypotézu nám pak slouží

Experimentální četnosti

Očekávané četnosti

předpoklad, že se experimentální a očekávané četnosti liší pouze náhodně, tj. že mezi nimi není statisticky významný rozdíl.
Postup:

1. Zvolíme hladinu významnosti α.
2. Výsledky výběrového šetření roztřídíme do zvolených skupin.
3. Stanovíme hodnoty očekávaných četností v jednotlivých skupinách (nejčastěji se bere průměr experimentálních četností).
4. V každé skupině vypočítáme $\frac{(O_i - E_i)^2}{E_i}$ a tyto hodnoty sečteme.
5. Ve statistických tabulkách vyhledáme kritickou hodnotu $\chi^2_\alpha(\nu)$ pro $\nu = k - 1$ stupňů volnosti.
6. Vypočítanou hodnotu testového kritéria porovnáme s tabulkovou hodnotou a na základě tohoto porovnání přijmeme nebo zamítneme nulovou hypotézu.
   Pokud $\chi^2 > \chi^2_\alpha$ ($P < 0,05$ v Excelu), zamítáme nulovou hypotézu a můžeme tvrdit, že mezi četnostmi experimentálními a očekávanými je statisticky významný rozdíl.

Jednotlivé mezivýsledky se zapisují do tabulky:

i-tá třída	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1					
...					
k					

Příklad

60krát házíme kostkou a výsledky zapisujeme do tabulky. Zjistěte, zda je kostka falešná.

Číslo na kostce	Počet padnutí O_i experim. četnost	očekávaná četnost E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	8	10	-2	4	0,4
2	9	10	-1	1	0,1
3	10	10	0	0	0
4	10	10	0	0	0
5	9	10	-1	1	0,1
6	14	10	4	16	1,6

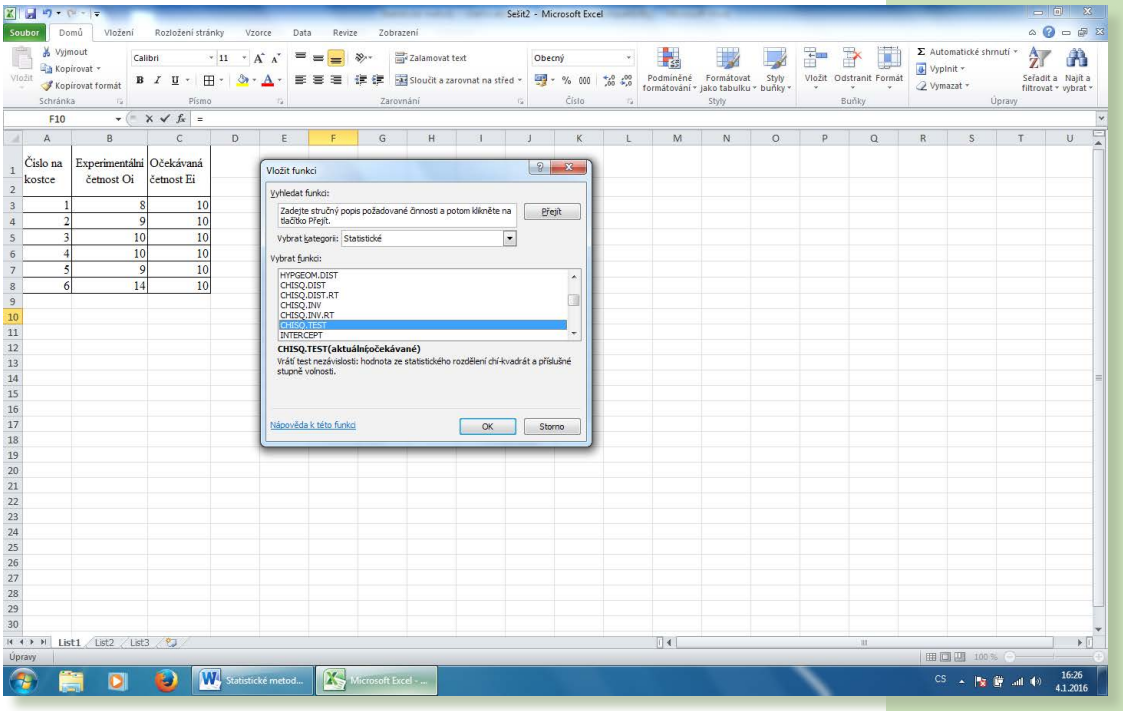
$\Sigma = 2,1 = \chi^2$

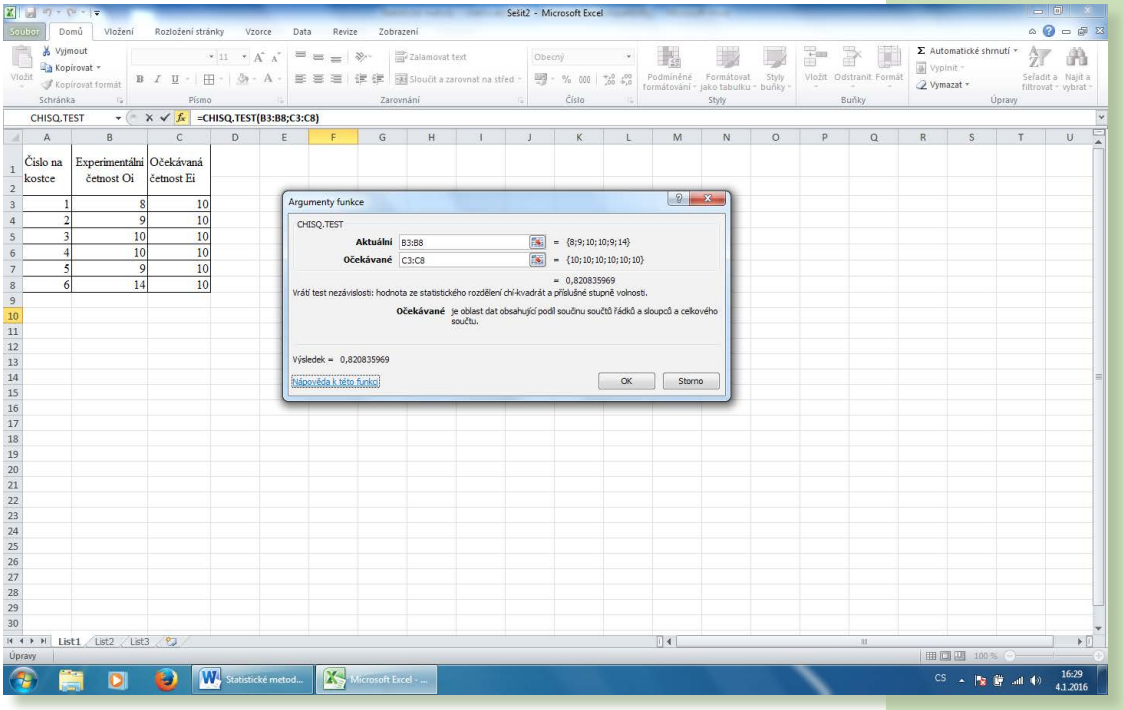
$\chi^2_{0,05}(5) = 11,1$
$\chi^2 < \chi^2_{0,05}(5)$

Není signifikantní rozdíl v četnostech, kostka není falešná.

Test shody χ^2 v programu Excel

Vložit funkci f_x → Statistické → Ok → CHISQ.TEST





Tato funkce nevrací výslednou hodnotu χ^2, ale pouze pravděpodobnost, kterou porovnáme s hladinou významnosti $\alpha = 0{,}05$

$P = 0{,}8 > 0{,}05$ výsledek není signifikantní.

5.3.2 Test nezávislosti χ^2 pro čtyřpolní tabulku (kontingenční tabulku 2 × 2) a Φ koeficient

Test nezávislosti

Test nezávislosti χ^2 se používá v případě, když máme rozhodnout, zda existuje významná souvislost mezi dvěma alternativními jevy, tj. jevy, které mohou nabývat jen dvou možných hodnot (např.: pohlaví – muž × žena, stav – svobodný × ženatý, léčba – úspěšná × neúspěšná, odpověď – ano × ne).

Čtyřpolní tabulka

Čtyřpolní tabulka vypadá takto:

		Jev X		
		x_1	x_2	Řádkový součet
Jev Y	y_1	a	b	$a + b$
	y_2	c	d	$c + d$
Sloupcový součet		$a + c$	$b + d$	$(a + b) + (c + d) =$ $= (a + c) + (b + d) = n$

a, b, c, d jsou příslušné četnosti
a + b, c + d, a + c, b + d jsou okrajové – marginální součty

$$\chi^2 = n \cdot \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Vypočítanou hodnotu χ^2 porovnáme s tabulkovou hodnotou $\chi^2_\alpha(\nu)$, kde $\nu = 1$ (počet řádků zmenšený o 1 vynásobený počtem sloupců zmenšeným o 1) a na základě porovnání přijmeme nebo zamítneme nulovou hypotézu o tom, že výběry pocházejí z téže populace. Pokud je výsledek signifikantní, pak má smysl vypočítat míru souvislosti – koeficient Φ. Vztah mezi koeficientem Φ a χ^2 se dá vyjádřit jako $\Phi = \sqrt{\frac{\chi^2}{n}}$.

Příklad

Ve výběru 100 pacientů (50 mužů a 50 žen) se zjišťovalo, kdo byl hospitalizován déle než 7 dní. Máme zjistit, zda existuje souvislost mezi pohlavím pacientů a délkou jejich hospitalizace a určit těsnost této souvislosti pomocí Φ koeficientu. Jednotlivé četnosti jsou dány v čtyřpolní tabulce.

		Hospitalizace		
		< 7 dní	> 7 dní	Σ
Pohlaví	Muži	40	10	50
	ženy	20	30	50
	Σ	60	40	100

Pomocí testu χ^2 zjistíme, zda je mezi oběma proměnnými nenáhodný vztah (signifikantní rozdíl mezi danými četnostmi) a tedy, že výběry pocházejí z rozdílných populací.

$$\chi^2 = n \cdot \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} = 100 \cdot \frac{(40\cdot 30 - 10\cdot 20)^2}{60\cdot 40\cdot 50\cdot 50} = 16{,}67$$

Tabulková hodnota je $\chi^2_\alpha(\nu) = \chi^2_{0,05}(1) = 3{,}84 \Rightarrow \chi^2 > \chi^2_{0,05}(1) \Rightarrow$ je signifikantní rozdíl mezi skupinou mužů a žen $\Rightarrow$ existuje souvislost mezi pohlavím a hospitalizací. Těsnost této souvislosti se určí pomocí Φ koeficientu.

$$\Phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{40\cdot 30 - 10\cdot 20}{\sqrt{50\cdot 50\cdot 60\cdot 40}} = \frac{1000}{2449{,}5} = 0{,}408$$

5.3.3 Test nezávislosti χ^2 pro kontingenční tabulku větší než 2 × 2 (obecně pro tabulku r × k) a kontingenční koeficient C

Vícepolní kontingenční tabulka

Při zachycování jevů pomocí dotazníku nebo rozhovoru se používá ***vícepolní kontingenční tabulka***. Opět se rozhoduje o tom, zda existuje významná souvislost mezi dvěma jevy – máme tedy opět dvě proměnné X a Y, ale každá z nich je ještě dělená do tříd. Četnosti uvedené v kontingenční tabulce jsou experimentální (pozorované) četnosti – O_i. Očekávané četnosti – E_i se musí vypočítat ze řádkových a sloupcových (tzv. marginálních) součtů:

$E_i = \frac{\sum \text{řádek} \times \sum \text{sloupec}}{n}$, kde n je počet naměřených hodnot.

Testové kritérium χ^2 se pak vypočítá jako $\sum_{j=1}^{r}\sum_{i=1}^{k} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$.

Koeficient kongingence

Vztah mezi proměnnými X a Y určuje ***koeficient kontingence*** C, který se vypočítá podle vzorce $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$.

Tento koeficient může nabývat hodnot mezi 0 a +1. Na rozdíl od koeficientu Φ se koeficient C používá u libovolných čtvercových nebo obdélníkových tabulek např. 4 × 3, 5 × 7 atd.

Příklad

Bylo zkoumáno, zda existuje souvislost mezi vzděláním a bydlištěm pacientů. Výsledky dotazníkového šetření, které bylo provedeno u 50 pacientů, uvádí následující tabulka:

		Stupeň vzdělání			
		Základní 1	Střední 2	Vysokoškolské 3	Σ
Bydliště	Malá vesnice 1	7	6	2	15
	Velká vesnice 2	5	4	6	15
	Malé město 3	7	4	2	13
	Velké město 4	3	2	2	7
	Σ	22	16	12	50

K tomu, abychom mohli vypočítat testové kritérium χ^2, musíme vytvořit novou tabulku s očekávanými četnostmi. Tyto četnosti vypočítáme z okrajových součtů, které přísluší každé experimentální četnosti v poli tabulky.

$$E_i = \frac{\sum \text{řádek} \times \sum \text{sloupec}}{n}$$

Pro experimentální četnost v 1. poli tabulky $O_1 = 7$ bude očekávaná četnost $E_1 = \frac{22 \times 15}{50} = \frac{330}{50} = 6{,}6$.

Stejným způsobem vypočítáme ostatní četnosti a zapíšeme je do nové tabulky:

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
7	6,6	0,4	0,16	0,02
6	4,8	1,2	1,44	0,30
2	3,6	-1,6	2,56	0,71
5	6,6	-1,6	2,56	0,39
4	4,8	-0,8	0,64	0,13
6	3,6	2,4	5,76	1,60
7	5,72	1,28	1,64	0,29
4	4,16	-0,16	0,03	0,01
2	3,12	-1,12	1,25	0,40
3	3,08	-0,08	0,01	0,00
2	2,24	-0,24	0,06	0,03
2	1,68	0,32	0,10	0,06

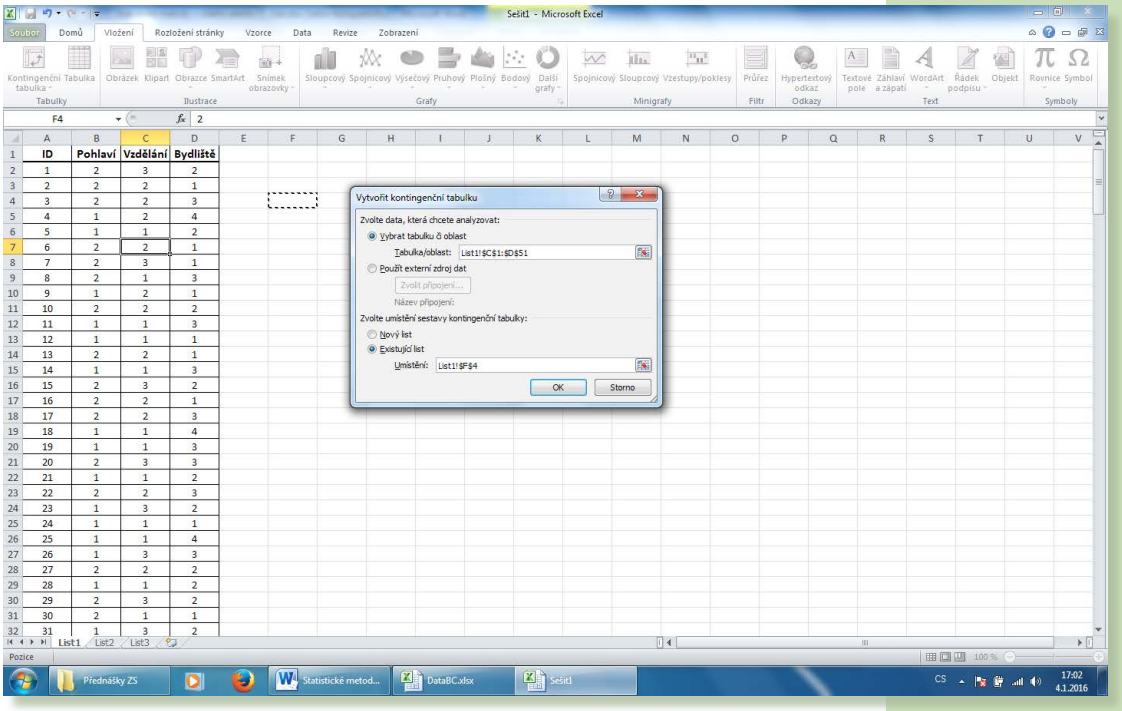
$\Sigma = \chi^2 = 3{,}94$

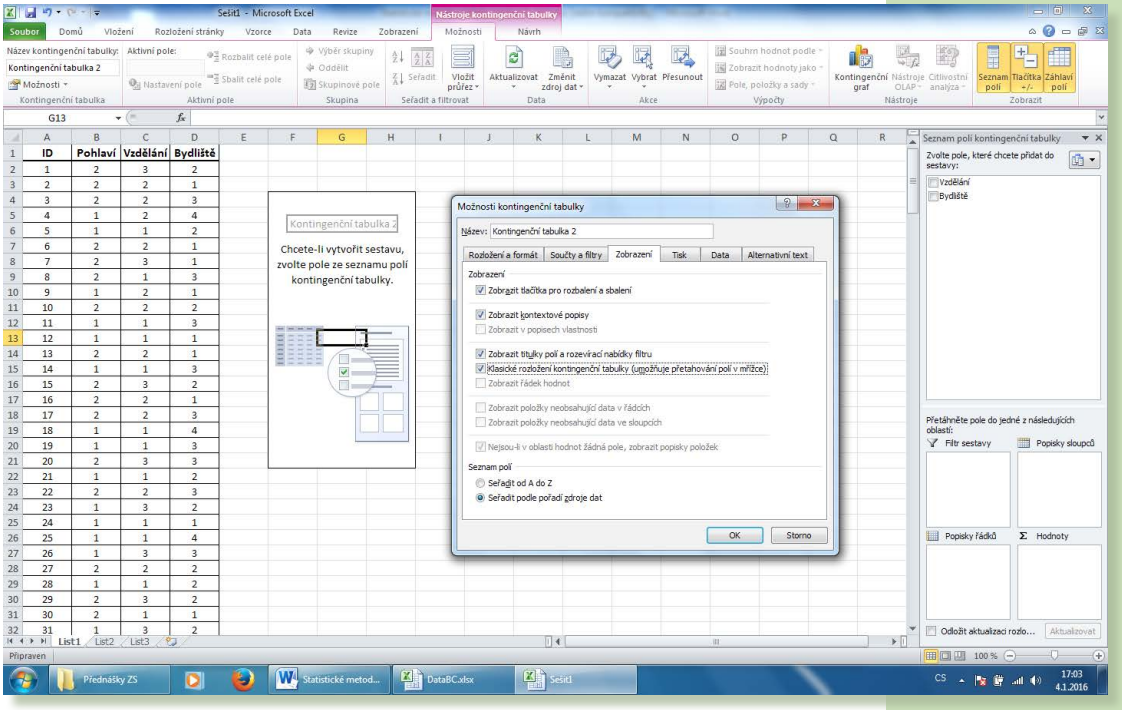
Vypočítanou hodnotu χ^2 porovnáme s tabulkovou hodnotou na hladině významnosti $\alpha = 0{,}05$ s počtem stupňů volnosti $\nu = (k - 1)(r - 1) = (4 - 1)(3 - 1) = 6$ což je 12,592. Porovnáním zjistíme, že vypočítaný $\chi^2 = 3{,}94 < \chi^2_{0{,}05}(6) = 12{,}592$. Výsledná hodnota testového kritéria χ^2 není signifikantní, není nutné počítat koeficient C. Můžeme tedy konstatovat, že mezi vzděláním a bydlištěm pacientů není statisticky významná souvislost.

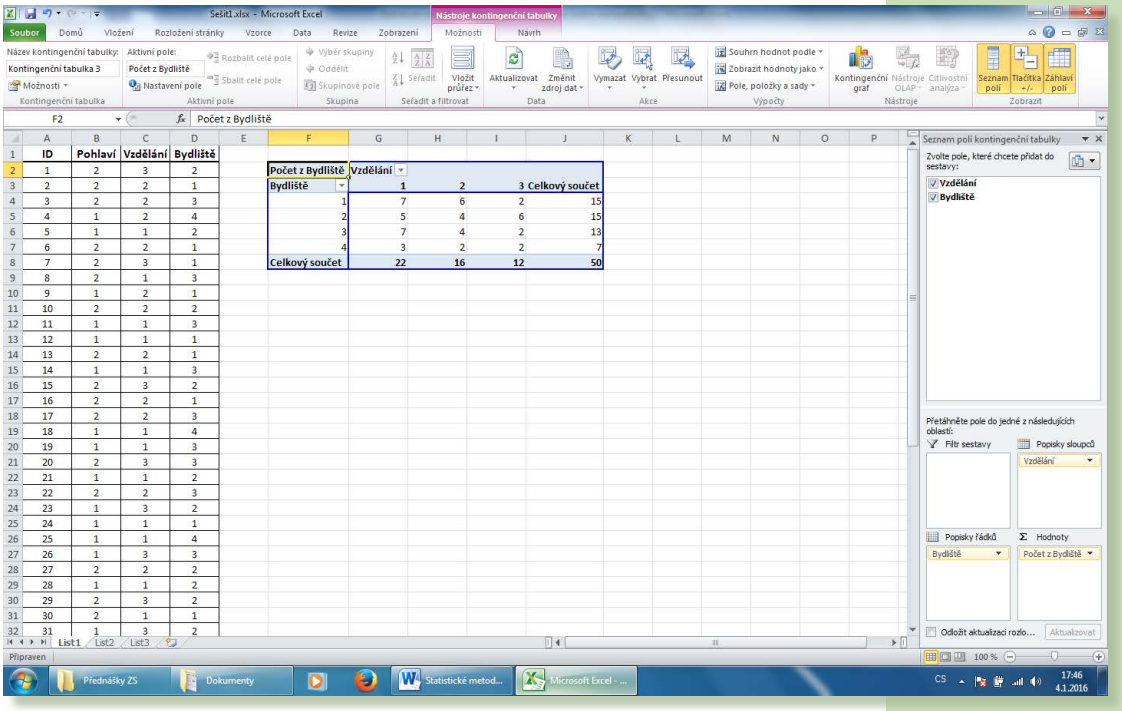
Tvorba kontingenční tabulky v Excelu

Vložení → Kontingenční tabulka → Zadat oblast dat → Existující list → Ok → Do vzniklého obdélníku kliknout pravým tlačítkem myši → Možnosti kontingenční tabulky → Zobrazení → zatrhnout Klasické rozložení → Ok

Pozn.: Výsledný χ^2 a C koeficient z Excelu nevypočítáme, abychom získali p hodnotu, musíme dopočítat očekávané četnosti a zadat je do funkce f_x.


Sešit1 - Microsoft Excel
Vytvořit kontingenční tabulku
Zvolte data, která chcete analyzovat:
Vybrat tabulku či oblast
Tabulka/oblast: List1!$C$1:$D$51
Použít externí zdroj dat
Zvolit připojení...
Název připojení:
Zvolte umístění sestavy kontingenční tabulky:
Nový list
Existující list
Umístění: List1!$F$4
OK
Storno


Sešit1 - Microsoft Excel
Nástroje kontingenční tabulky
Kontingenční tabulka 2
Chcete-li vytvořit sestavu, zvolte pole ze seznamu polí kontingenční tabulky.
Možnosti kontingenční tabulky
Název: Kontingenční tabulka 2
Rozložení a formát
Součty a filtry
Zobrazení
Tisk
Data
Alternativní text
Zobrazení
Zobrazit tlačítka pro rozbalení a sbalení
Zobrazit kontextové popisy
Zobrazit v popisech vlastnosti
Zobrazit titulky polí a rozevírací nabídky filtru
Klasické rozložení kontingenční tabulky (umožňuje přetahování polí v mřížce)
Zobrazit řádek hodnot
Zobrazit položky neobsahující data v řádcích
Zobrazit položky neobsahující data ve sloupcích
Nejsou-li v oblasti hodnot žádná pole, zobrazit popisky položek
Seznam polí
Seřadit od A do Z
Seřadit podle pořadí zdroje dat
OK
Storno
Seznam polí kontingenční tabulky
Zvolte pole, které chcete přidat do sestavy:
Vzdělání
Bydliště
Přetáhněte pole do jedné z následujících oblastí:
Filtr sestavy
Popisky sloupců
Popisky řádků
Hodnoty



Kontrolní otázky a úkoly

1. S jakými četnostmi pracují testy χ^2?
2. Jak vypadá čtyřpolní tabulka a co zpracovává?
3. Z jakých proměnných se počítá koeficient kontingence C?

Klíč k otázkám a úkolům

Odpovědi na otázky najdete v textu.

Referenční seznam

DUPAČ, V., HUŠKOVÁ, M. 2013. *Pravděpodobnost a matematická statistika*. Praha: Karolinum. ISBN 978-80-246-2208-8.

ZVÁRA, K. 2013. *Základy statistiky v prostředí R*. Praha: Karolinum. ISBN 978-80-246-2245-3

5.4 Neparametrické testy

Nejčastějším problémem statistických metod je rozhodnout, zda určitý náhodný výběr je výběrem ze základního souboru se známým rozložením nebo zda dva či více náhodných výběrů pochází ze stejné populace. Klasické parametrické metody jsou použitelné jen za určitých předpokladů. Nelze je použít, jestliže nastane alespoň jeden z následujících případů:

- stupnice, na níž je založeno měření je pouze pořadová, to znamená, že vyjadřuje jenom vzestupné nebo sestupné uspořádání pořadových hodnot;
- teoretické rozložení proměnné v populaci je neznámé a výběry jsou příliš malé na to, abychom mohli dostatečně spolehlivě odhadnout jeho typ;
- rozložení proměnné nelze převést žádnou vhodnou transformací na normální.

V těchto případech pak přistupujeme k testům neparametrickým, protože nepracují s parametry rozložení základního souboru ani s parametry výběrů (průměr, směrodatná odchylka, rozptyl, atd.). Tyto metody nepotřebují předpoklad o konkrétním typu rozdělení. Jsou slabší než odpovídající testy parametrické, protože často využívají jen část informace obsažené v pozorovaných datech (některé používají jen znamének diferencí, pořadových čísel skutečných hodnot znaku atd.). Přesto však bývají za uvedených podmínek jedinými použitelnými statistickými testy vůbec. Kromě toho neparametrické testy vyžadují zpravidla jen jednoduché numerické výpočty, takže se hodí pro rychlou orientaci v experimentálních výsledcích.

Studijní cíle

Cílem této kapitoly je naučit studenty pracovat s neparametrickými ekvivalenty metod parametrických v případech, kdy nejsou pro jejich použití splněny některé předpoklady. Nejčastěji to bývá předpoklad normality nebo malé výběrové soubory. Také se může jednat při statistickém zpracování o data jiná než metrická.

Klíčová slova

Neparametrické metody, ordinální data, alternativní data, kategoriální data

5.4.1 McNemarův test

Je vhodný pro srovnávání dvou korelovaných výběrů. Ověřuje, zda se poměr počtu případů v alternativních kategoriích signifikantně mění v důsledku nějakého experimentu. Data sestavíme do tabulky 2×2 (čtyřpolní tabulka), ve které jednotlivá pole označíme a, b, c, d. Je důležité, aby jednotlivá pole byla označována vždy stejným způsobem. Při jiném označení by vzorce neplatily.

		Po experimentu: +	Po experimentu: –	Po experimentu: Σ
Před experimentem	+	a	b	$a + b$
	–	c	d	$c + d$
	Σ	$a + c$	$b + d$	n

U osob, které jsou zařazeny do polí a a d, došlo v důsledku experimentálního zásahu ke změně (u osob a ke změně z výsledku + na –, u osob d ke změně z – na +). Osoby zařazené v polích b a c nezaznamenaly žádnou změnu. Tyto osoby ve vlastních výpočtech neuvažujeme. Celkový počet osob, jejichž výsledky se změnily je $a + d$.

Testujeme nulovou hypotézu, která říká, že polovina změn, tj. $(a + d)/2$ bude v jednom směru a polovina v druhém. To znamená, že změny se vzájemně vyrovnávají a experimentální zásah nemá na výsledek podstatný vliv. Teoretické četnosti v případě platnosti nulové hypotézy polí a a d jsou tedy obě rovny $(a + b)/2$. Pro srovnání experimentálních a hypotetických četností použijeme χ^2 – ***test dobré shody***. Testová charakteristika je rovna $\chi^2 = \frac{(a-d)^2}{a+d}$.

Test dobré shody

Výběrová distribuce je přibližně χ^2 s jedním stupněm volnosti. Tato distribuce je však spojitá a my počítáme pouze s jedním stupněm volnosti a s daty diskrétními. Chyba vzniklá aproximací rozložení diskrétní veličiny spojitým rozložením by zde byla dosti značná, užíváme proto Yatesovy korekce pro spojitost. Výsledný vzorec pro testovou charakteristiku je pak $\chi^2 = \frac{(|a-d|-1)^2}{a+d}$.

McNemarova testu je možno použít ve všech případech, kdy sledujeme rozdíl mezi dvěma šetřeními, jejichž výsledky se rozpadají do dvou vzájemně disjunktních tříd. Vzorec pro χ^2 bez Yatesovy korekce je použitelný v případě větších výběrů a vzorec s Yatesovou korekcí platí pro malé výběrové soubory. Test by se neměl vůbec používat, jsou-li očekávané četnosti příliš malé tj. $(a + d)/2 = 5$.

Příklad

Sledujeme dlouhodobý účinek farmakologického preparátu na motoriku dětí. Máme možnost sledovat děti bezprostředně po podání preparátu a 6 měsíců po jeho podání. Motorickou úroveň dětí můžeme hodnotit jen podle kritéria: zlepšení či zhoršení stavu oproti stavu, který byl před podáním farmaka. Sledujeme 50 dětí. Získaná data jsou uspořádána do tabulky.

		Po 6 týdnech		
		zhoršení	zlepšení	Σ
Bezprostředně po podání	zlepšení	23	12	35
	zhoršení	8	7	15
	Σ	31	19	50

Podle nulové hypotézy předpokládáme, že po 6 měsících dojde u stejného počtu dětí, které se původně zlepšily, ke zhoršení a u stejného počtu původně horších dětí dojde ke zlepšení.

$$\chi^2 = \frac{(|a-d|-1)^2}{a+d} = \frac{(|23-7|-1)^2}{23+7} = 7{,}5$$

$\chi^2_{0{,}01}(1) = 6{,}63$; $\chi^2 > \chi^2_{0{,}01}(1) \Rightarrow$ zamítáme H_0

Z tabulek distribuce χ^2 zjistíme, že výsledek 7,5 má při jednom stupni volnosti pravděpodobnost výskytu menší než 0,01. Zamítáme proto nulovou hypotézu a na základě rozboru dat můžeme konstatovat, že působení preparátu je pouze krátkodobé. Více zlepšení nastalo bezprostředně po podání preparátu.

5.4.2 Bowkerův test symetrie

Test zjišťuje, zda jsou četnosti v tabulce symetricky rozložené podle hlavní diagonály. Nulová hypotéza říká, že při opakovaném měření nedochází v tabulce ke vzniku asymetrie (četnosti nejsou asymetricky rozložené kolem hlavní diagonály). Jedná se o rozšíření testu McNemara pro tabulku větší než 2 × 2, obecně pro tabulku r × r.

Tabulka je ve tvaru:

		2. měření					
		x_1	x_2	…		x_r	Σ
1. měření	x_1	n_{11}	n_{12}	…		n_{1r}	
	x_2	n_{21}	n_{22}	…		n_{1r}	
	.	.				.	
	.	.				.	
	.	.				.	
	x_r	n_{r1}	n_{r2}	…		n_{rr}	

Testové kritérium B vypočítáme podle vzorce

$$B = \sum_{i=1}^{r} \sum_{j>i} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

i…počet řádků
j…počet sloupců

Toto kritérium má za platnosti nulové hypotézy rozdělení χ^2 s $r(r-1)/2$ stupni volnosti.

Příklad

Byla porovnávána úspěšnost léčby dvěma metodami A a B. Oprávněnost nulové hypotézy (obě metody mají podobnou účinnost) se ověřovala na datech získaných jako dvojice pozorování u každého z $n = 100$ pacientů. Úspěšnost léčby byla registrována na tří bodové škále: neúčinná, málo účinná, hodně účinná. Četnosti pacientů jsou v tabulce.

		Metoda B			
		neúčinná	málo účinná	hodně účinná	Σ
Metoda A	neúčinná	6	10	13	29
	málo účinná	11	23	7	41
	hodně účinná	8	12	10	30
	Σ	25	45	30	100

$$B = \sum_{i=1}^{r} \sum_{j>i} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} = \frac{(11-10)^2}{11+10} + \frac{(8-13)^2}{8+13} + \frac{(12-7)^2}{12+7} =$$

$$= \frac{1}{21} + \frac{25}{21} + \frac{25}{19} = 0{,}05 + 1{,}19 + 1{,}32 = 2{,}56$$

Ve statistických tabulkách vyhledáme kritické hodnoty rozdělení χ^2.
Počet stupňů volnosti $\nu = r(r-1)/2 = 4 \cdot 3/2 = 6$
Jelikož $B = 2{,}56 > \chi^2_{0{,}05}(6) = 12{,}6$

nezamítáme nulovou hypotézu a můžeme říct, že léčby A a B mají podobnou účinnost.

5.4.3 Mann-Whitneyův U-test

Používá se pro dva nezávislé výběry a je jedním z nejsilnějších neparametrických testů. Vychází z pořadových hodnot a ověřuje, že dva náhodné výběry byly pořízeny z téže populace nebo ze dvou identických základních souborů. Skóry obou pozo-

Testové kritérium

rovaných výběrů o velikostech n_1 a n_2 musí být možné seřadit do společné vzestupné posloupnosti o rozsahu $n_1 + n_2$. ***Testové kritérium*** zjišťuje počet inverzí a označuje se U. Pro každý prvek z první skupiny musíme zjistit, kolik pozorování z druhé skupiny mu v sestaveném pořadí předchází. V závislosti na tom, který z obou výběrů vezmeme jako první nebo druhý, dostáváme dva údaje o počtu inverzí , U a U'. Jako testové kritérium se volí menší číslo.

Při výpočtu U seřadíme skóry obou výběrů vzestupně podle velikosti do jedné posloupnosti, přidělíme jim pořadová čísla a ta vrátíme zpět do původní skupiny. Sečteme je pro každou skupinu zvlášť a označíme symboly T_1 a T_2.

$$U = n_1.n_2 + [n_1(n_1+1)/2] - T_1$$

$$U' = n_1.n_2 + [n_2(n_2+1)/2] - T_2$$

Veličiny U a U' spolu souvisejí vztahem $U + U' = n_1 \cdot n_2$.

Mann a Whitney přesně odvodili za předpokladu platnosti nulové hypotézy distribuci U až do rozsahu n_2 = 8. (str. 71). Pro výběry rozsahu většího než 20 se může použít normální aproximace. Za platnosti nulové hypotézy má veličina U normální distribuci. Pro $n_2 > 20$ je tedy veličina

rozložena přibližně normálně s průměrem 0 a rozptylem 1. Pro test významnosti se používají tabulky kvantilů normálního rozložení.

$$z = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Příklad a)

Máme porovnat motorický vývoj dětí různě dlouho hospitalizovaných v důsledku nějaké somatické nemoci. Jeden výběr tvoří děti hospitalizované dva měsíce, druhý výběr je z dětí hospitalizovaných čtyři měsíce. Kromě této skutečnosti se děti od sebe neliší. Všechny děti byly vyšetřeny několika zkouškami motorického vývoje a výsledky byly vyjádřeny celkovým počtem bodů u jednotlivých dětí.

Výsledky dětí v první skupině (děti hospitalizované 2 měsíce): $22, 24, 29, 30, 36, 39, 45$ $n_1 = 7$

Výsledky dětí ve druhé skupině (děti hospitalizované 4 měsíce): $18, 21, 31, 33, 35, 38$ $n_2 = 6$

Hodnoty obou výběrů spojíme vzestupně do jedné posloupnosti

18	21	22	24	29	30	31	33	35	36	38	39	45
II	II	I	I	I	I	II	II	II	I	II	I	I
1	2	3	4	5	6	7	8	9	10	11	12	13

V prvním řádku jsou uvedeny výběrové hodnoty, ve druhém řádku je římskou číslicí označeno, z kterého výběru každá hodnota pochází, třetí řádek zaznamenává pořadí hodnot.

Součet pořadí pro první výběr je

$$T_1 = 3 + 4 + 5 + 6 + 10 + 12 + 13 = 53,$$

součet pořadí pro druhý výběr je

$$T_2 = 1 + 2 + 7 + 8 + 9 + 11 = 38.$$

Hodnotu testové charakteristiky vypočítáme ze vzorce

$$U = n_1 \cdot n_2 + \left[n_1(n_1+1)/2\right] - T_1 = 6 \cdot 7 + 56/2 - 53 = 17$$

$$U' = n_1 \cdot n_2 + \left[n_2(n_2+1)/2\right] - T = 6 \cdot 7 \ + 42/2 - 38 = 25$$

Pro menší z vypočítaných hodnot $U = 17$ najdeme v tabulkách pro $n_1 = 6$, $n_2 = 7$ pravděpodobností hodnotu $\alpha = 0{,}314$. Což je dostatečně vysoká pravděpodobnost pro nezamítnutí nulové hypotézy. Hranice, do které ještě nulovou hypotézu přijmeme, je 0,05 ($\alpha > 0{,}05$). Rozdíl mezi výsledky dětí hospitalizovaných 2 měsíce a 4 měsíce můžeme prohlásit za nevýznamný.

Příklad b)

U dvou vyrovnaných skupin sportovců sledujeme dva různé druhy tréninkových metod. Chceme zjistit, zda obě metody tréninku jsou stejně účinné, či zda se od sebe liší. Zlepšení ve výkonu sportovce po čtyřměsíčním tréninku vyjadřujeme v procentech. Jako nulovou hypotézu si stanovíme tvrzení, že obě metody jsou stejně účinné. V tabulce jsou uvedeny výsledky sportovců obou skupin a jejich pořadí v posloupnosti hodnot obou výběrů.

Tréninková metoda A				Tréninková metoda B			
% zlepšení	poř. číslo	% zlepšení	poř. číslo	% zlepšení	poř. číslo	% zlepšení	poř. číslo
0,5	1	12,4	18	2,6	5	26,4	39
1,0	2	14,0	21	6,4	9	27,0	40
1,7	3	15,6	24	9,5	14	28,2	41
2,0	4	17,2	26	10,8	16	29,9	42
4,3	6	17,9	27	12,5	19	31,5	43
5,1	7	18,0	28	13,5	20	31,7	44
5,9	8	18,8	30	14,7	22	32,0	45
6,5	10	20,3	31	14,9	23	34,7	46
8,0	11	22,6	33	16,3	25	38,0	47
8,7	12	23,0	34	18,7	29	38,4	48
8,9	13	23,1	35	20,5	32	39,1	49
10,0	15	24,8	37	24,3	36	40,2	50
11,2	17			25,0	38	42,6	51

Součty pořadí $T_1 = 453$, $T_2 = 873$ použijeme pro výpočet U. Rozsahy obou výběrů jsou dostatečně velké pro použití normální aproximace. Ve vzorci pro výpočet z použijeme menší hodnotu U.

$$U = n_1 \cdot n_2 + \left[n_1(n_1+1)/2\right] - T_1 = 25 \cdot 26 + (25 \cdot 26/2) - 453 =$$
$$= 650 + 325 - 453 = 522$$
$$U' = n_1 \cdot n_2 + \left[n_2(n_2+1)/2\right] - T_2 = 25 \cdot 26 + (26 \cdot 27/2) - 873 =$$
$$= 650 + 351 - 873 = 128$$

Pro výpočet testového kritéria z použijeme menší z obou U.

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{128 - \frac{25 \cdot 26}{2}}{\sqrt{\frac{25 \cdot 26(25 + 26 + 1)}{12}}} = -3{,}71$$

$|z| > z_{0,01} = 2{,}58 \Rightarrow$ zamítáme nulovou hypotézu a můžeme prohlásit, že tréninkové metody se velmi významně od sebe liší (metoda B je lepší než metoda A).

5.4.4 Mediánový test

Jednou z možností, jak rozdělit data výběru do dvou alternativních tříd, je jejich roztřídění na hodnoty větší a menší než je medián těchto dat.

Představme si, že chceme zjistit, zda dva nezávislé výběry pocházejí z populací se stejným mediánem. Jde nám v podstatě o zjištění, zda se výběry liší v centrální tendenci. Pocházejí-li výběry ze stejné populace, nebo z populací se stejným mediánem, očekáváme, že asi polovina dat prvního výběru bude větší než společný medián obou výběrů a polovina dat bude menší. Totéž platí také pro druhý výběr.

Mediánový test

Na tomto předpokladu je založen ***mediánový test.*** Pracuje se dvěma nezávislými výběry, jejichž rozsah nemusí být stejný. Data obou výběrů sloučíme, uspořádáme podle velikosti a stanovíme jejich společný medián. Četnosti hodnot nad a pod mediánem zapíšeme do čtyřpolní tabulky. Výpočet provedeme buď Fischerovým testem nebo testem χ^2 pro čtyřpolní tabulku. Při platnosti nulové hypotézy by mělo platit $a = c$ a $b = d$.

Mediánový test se může použít i pro více výběrů, zjišťování signifikantnosti se pak provádí testem χ^2 pro vícepolní kontingenční tabulku.

Příklad
U dvou skupin osob byly ve stejném testu naměřeny tyto výsledky:
A: 10, 15, 18, 12, 14, 10, 9
B: 8, 10, 16, 12, 15, 9, 10
Určete pomocí mediánového testu, zda je mezi nimi signifikantní rozdíl.

Data srovnáme podle velikosti do pořadí a určíme medián:
8, 9, 9, 10, 10, 10, 10, 12, 12, 14, 15, 15, 16, 18
$M_e = (10 + 12)/2 = 22/2 = 11$

	Skupina A	Skupina B	Σ
Nad mediánem	4	3	7
Pod mediánem	3	4	7
Σ	7	7	14

$$\chi^2 = n\frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} =$$

$$= 14\frac{(4\cdot 4 - 3\cdot 3)^2}{7\cdot 7\cdot 7\cdot 7} = \frac{2\cdot 49}{49\cdot 7} = \frac{2}{7} = 0{,}29$$

$\chi^2 < \chi^2_{0,05}(1) \Rightarrow$ Není signifikantní rozdíl v počtu četností nad a pod mediánem. $\chi^2_{0,05}(1) = 3{,}84$

5.4.5 Wilcoxonův pořadový test pro párované hodnoty

Slouží k ověření hypotézy o tom, zda je signifikantní rozdíl mezi opakovaným měřením na stejných subjektech. V datech musí být možné stanovit pořadí pro všechna měření dohromady. Jedná se o test významnosti diferencí. Podobně jako u parametrického párového t-testu počítáme i zde rozdíly $d_i = x_i - y_i$ mezi naměřenými hodnotami, které tvoří pár, tj. které si navzájem odpovídají. Některé hodnoty d_i jsou kladné, některé záporné. Bez ohledu na znaménko stanovíme jejich pořadí. K pořadovým číslům pak znovu doplníme znaménka a provedeme součet všech kladných a všech záporných pořadových čísel. Nulová hypotéza nám říká, že obě vyšetřované skupiny jsou ekvivalentní, tj. že součty kladných a záporných pořadí jsou v absolutní hodnotě přibližně stejné. Testovým kritériem T je ten součet pořadí se stejným znaménkem, který je menší. Pro výběry, jejichž rozsah je menší nebo roven 25, nalezneme kritickou hodnotu $T_\alpha(n)$ pro danou hladinu významnosti $\alpha = 0{,}05$ a $\alpha = 0{,}01$ ve statistických tabulkách a porovnáme ji s vypočítanou hodnotou T. Je-li $T < T_\alpha(n)$, pak zamítáme nulovou hypotézu a můžeme tvrdit,

že oba výběry pocházejí z různých základních souborů. V opačném případě nulovou hypotézu přijmeme. Pro výběry o rozsahu větším než 25 musíme vypočítat hodnotu veličiny z, která má za platnosti H_0 přibližně normální distribuci s průměrem

$$\mu = \frac{n(n+1)}{4} \quad \text{a rozptylem} \quad \sigma = \frac{n(n+1)(2n+1)}{24}.$$

Ze vzorce pro standardizaci plyne $z = \frac{x-\mu}{\sigma} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$.

Kritické hodnoty pro jsou $z_{0,05} = 1{,}96$ a pro $z_{0,01} = 2{,}58$. Je-li $z < z_{0,05}$, pak zamítáme H_0 na hladině významnosti $\alpha = 0{,}05$ a je-li $z < z_{0,01}$, pak zamítáme H_0 na hladině významnosti $\alpha = 0{,}01$.

Příklad

U deseti pacientů byl zjišťován počet chyb ve dvou testech. Určete, zda se chybovost v testech od sebe signifikantně odlišuje. Jako nulovou hypotézu stanovíme předpoklad, že je počet chyb v obou testech stejný. Výsledky jsou zaznamenány v tabulce:

Pacient	1. test	2. test	d_i	pořadí +	pořadí −
1	12	10	2	4	
2	14	8	6	10	
3	9	7	2	4	
4	15	11	4	8,5	
5	10	13	−3		6,5
6	11	9	2	4	
7	7	8	−1		1,5
8	10	6	4	8,5	
9	11	14	−3		6,5
10	13	12	1	1,5	
Σ				40,5	**14,5**

Menší je součet pořadí záporných diferencí, proto testové kritérium $T = 14{,}5$. Tabulková hodnota pro $n = 10$ a $\alpha = 0{,}05$ je $T_\alpha = 8$. Jelikož je $T > T_\alpha$ přijímáme nulovou hypotézu a můžeme říct, že chybovost v obou testech je přibližně stejná.

Příklad
Předpokládejme, že tentýž test byl proveden s padesáti pacienty. Postup testování bude stejný až na to, že kritickou hodnotu musíme vypočítat. Předpokládejme, že testová charakteristika T počítaná stejným způsobem jako v předcházejícím příkladě má nyní hodnotu $T = 98$. Kritickou hodnotou je veličina z, která má za platnosti nulové hypotézy normální distribuci s parametry 0 a 1 a pro $\alpha = 0{,}05$ nabývá hodnoty $z = 1{,}96$ a pro $\alpha = 0{,}01$ má hodnotu $z = 2{,}58$. Testové kritérium pak vypočítáme podle vzorce

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{98 - \frac{50 \cdot 51}{4}}{\sqrt{\frac{50 \cdot 51 \cdot 101}{24}}} = 0{,}89.$$

Porovnáním z a z_α ($|z| \prec z_\alpha$) zjistíme, že nezamítáme nulovou hypotézu. Není tedy signifikantní rozdíl v chybovosti testů.

5.4.6 Znaménkový test

Používá se podobně jako Wilcoxonův test v případě dvou opakovaných měření na stejných subjektech. Hodnoty naměřené u jednoho subjektu ve výběru tvoří vždy páry. Je použitelný tehdy, jestliže dokážeme říct, jaký směr změna měla, zda kladný nebo záporný. Test tedy nepoužívá přímo naměřené hodnoty, ale pouze znaménka. Je založen na úvaze, že v případě, že by nebyl mezi oběma měřeními žádný rozdíl, měla by se obě znaménka vyskytovat se stejnou pravděpodobností, tj. měl by jich být stejný počet. Rozdíl mezi měřením se pak projeví tím, že začnou převažovat buď znaménka kladná nebo záporná. Menší počet znamének se označí S. V tabulkách vyhledáme hodnotu $S_\alpha(n)$, která určuje, kolikrát se může méně se vyskytující znaménko objevit, aby byl rozdíl statisticky významný. To je v případě, že $S \leq S_\alpha(n)$.

Příklad

Zjistěte, zda týdenní plavecký kurz významně zlepšuje výkony jeho účastníků v plavání na 100 m. Na začátku kurzu byl změřen 1. časový údaj u dvaceti účastníků. Po týdnu výcviku bylo měření času v plavání na 100 m zopakováno. Výsledky jsou v tabulce:

Účastník	1. čas	2. čas	změna
1	10,5	9,6	+
2	9,1	9,9	–
3	8,9	7,5	+
4	9,7	9,6	+
5	11,1	9,8	+
6	8,4	6,6	+
7	10,1	9,9	+
8,5	9,4	10,3	–
7,1	7,9	7,3	+
10	8,9	6,5	+

Účastník	1. čas	2. čas	změna
11	8,0	7,6	+
12	9,8	8,9	+
13	9,5	9,6	–
14	11,3	9,4	+
15	10,7	8,5	+
16	9,1	7,0	+
17	8,1	8,3	–
18	9,9	9,2	+
19	10,4	9,8	+
20	11,9	10,1	+

H_0: Mezi opakovaným měřením není signifikantní rozdíl.

1. Z tabulky určíme, které znaménko se vyskytuje méně. V tomto případě je to –. Vyskytuje se 4× ($S = 4$).
2. Ze statistických tabulek zjistíme, že při dvaceti naměřených dvojicích hodnot by byl signifikantní výsledek 5 a méně (znamének –) ($S_{0,05}(n) = 5$). Na základě porovnání $S = 4 < S_{0,05}(n) = 5$ zamítáme nulovou hypotézu a můžeme říct, že plavecký kurz signifikantně zlepšil výkony účastníků.

Kontrolní otázky a úkoly

1. Které neparametrické metody z výše probraných jsou ekvivalentem Studentova dvouvýběrového t-testu?
2. Které neparametrické metody z výše probraných jsou ekvivalentem Párového t-testu?
3. Který z neparametrických dvouvýběrových testů je nejsilnější?

Klíč k otázkám a úkolům

Odpovědi na otázky naleznete v textu.

Referenční seznam

HENDL, J. 2004. *Přehled statistických metod zpracování dat*. Praha: Portál. ISBN 80-7178-820-1.

REITEROVÁ, E. 2008. *Základy psychometrie*. Olomouc: UP. ISBN 978-80-244-2065-3.

5.5 Analýza rozptylu

Analýza rozptylu, řečeno velmi zjednodušeně, nám říká, zda existují statisticky významné rozdíly mezi více než dvěma sledovanými skupinami. Pokud chceme analyzovat dvě skupiny např. pacientů, pak použijeme Studentův t-test. Jestliže budeme zkoumat rozdíly mezi více jak dvěma skupinami a rozhodneme se testovat vždy každé dvě skupiny zvlášť, můžeme použít t-test, ale musíme snížit hladinu významnosti α tak, že ji vydělíme počtem sledovaných skupin. Tím se α sníží a většinou signifikantní výsledky nedostaneme. Proto je lepší ke zjišťování významnosti rozdílů mezi více skupinami požívat analýzu rozptylu.

Studijní cíle
Cílem této kapitoly je objasnění jednofaktorové parametrické a neparametrické analýzy rozptylu.

Klíčová slova
Parametrická analýza rozptylu, celkový součet čtverců, součty čtverců, Friedmannova analýza rozptylu, Kruskal-Wallisova analýza rozptylu

5.5.1 Parametrická analýza rozptylu

Analýza rozptylu je statistická metoda, která umožňuje zjišťovat rozdíl v naměřených hodnotách u více sledovaných skupin pacientů. Jejím speciálním případem pro dvě skupiny je dvouvýběrový Studentův t-test. Jako svého základního nástroje používá Fisher-Snedecorova F-rozdělení. Řeší otázku, zda je statisticky významný rozdíl mezi naměřenými hodnotami u více výběrů. V jednoduché analýze rozptylu se předpokládá, že výsledky pozorování proměnné veličiny X můžeme rozdělit do několika skupin podle působení faktoru Y. Výsledky pozorování se budou lišit jednak uvnitř každé skupiny a jednak mezi skupinami navzájem. Variabilita výsledků celého pozorování je způsobena jednak faktorem Y, který pozorujeme a jednak neznámými – náhodnými vlivy. Tyto vlivy se nazývají náhodný rozptyl, který slouží jako odhad neznámého rozptylu základního souboru. K výpočtu rozptylu potřebujeme znát součty čtverců odchylek jednotlivých pozorování od aritmetického průměru:

1. ***celkový součet čtverců*,** kde x_i jsou jednotlivé naměřené hodnoty a n je celkový počet pozorování. $S_{ČT}$ se dělí na dvě složky: $SČ_T = SČ_{BG} + SČ_{WG}$

$$SČ_T = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

Celkový součet čtverců

2. ***součet čtverců mezi skupinami*,** kde n_j, j = 1, 2, ... k jsou rozsahy jednotlivých skupin a n je celkový rozsah.

$$SČ_{BG} = \frac{(\sum x_j)^2}{n_j} - \frac{(\sum x_i)^2}{n}$$

Součet čtverců mezi skupinami

3. ***součet čtverců uvnitř skupin***

$$SČ_{WG} = SČ_T - SČ_{BG}$$

Součet čtverců uvnitř skupin

K odpovědi na otázku, zda se údaje v jednotlivých skupinách statisticky významně liší, potřebujeme vypočítat rozptyly mezi skupinami a uvnitř skupin a z nich pak testové kritérium F. Rozptyly se vypočítají jako poměr příslušného součtu čtverců a počtu stupňů volnosti $(\nu_{BG} = k-1, \nu_{WG} = k(n_j - 1), j = 1,2,...,k)$.

Rozptyl mezi skupinami se vypočítá jako $s_{BG}^2 = \frac{SČ_{BG}}{k-1}$.

Rozptyl uvnitř skupin je roven $s_{WG}^2 = \frac{SČ_{WG}}{k(n_j - 1)}$.

Poměrem těchto dvou rozptylů pak získáme testové kritérium $F = \frac{s_{BG}^2}{s_{WG}^2}$.

K této vypočítané hodnotě najdeme ve statistických tabulkách kritickou hodnotu $F_\alpha(\nu_{BG}, \nu_{BG})$ a obě porovnáme. Je-li $F > F_\alpha$, zamítáme nulovou hypotézu o homogenitě sledovaného souboru (rozptyly se sobě nerovnají) a můžeme tvrdit, že se projevil vliv rozdílného působení sledovaného faktoru v jednotlivých skupinách.

Výsledky jednoduché analýzy rozptylu zapisujeme do následující tabulky:

Zdroj rozptylu	součet čtverců	stupně volnosti	rozptyl	F
mezi skupinami	S_{BG}	$k-1$	s_{BG}^2	s_{BG}^2 / s_{WG}^2
uvnitř skupin	S_{WG}	$k(n_j-1)$	s_{WG}^2	–
celkem	S_T	$n_j \cdot k - 1 = n - 1$	–	–

Poznámka: Pro dvě sledované skupiny dává analýza rozptylu stejný výsledek jako t-test. Platí, že $F = t^2 \Rightarrow t = \sqrt{F}$. Pro více skupin se musí počítat pouze F. (F je obecný test a t-test je jeho zvláštní případ).

Příklad

Zjistěte, zda jsou signifikantní rozdíly ve skórech testu mezi čtyřmi věkovými skupinami pacientů.

Věkové skupiny

	A	B	C	D
	63	74	86	95
	59	59	75	84
	68	68	69	87
	79	49	67	69
	48	85	81	58
	69	71	79	79
	55	68	58	82
	66	63	88	93
	72	55	91	91
Σ	55	64	71	80

1. Vypočítáme celkový součet čtverců

$$S\check{C}_T = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 63^2 + 74^2 + \ldots + 80^2 - \frac{2873^2}{40} = 6151{,}775$$

2. Vypočítáme součet čtverců mezi skupinami

$$S\check{C}_{BG} = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3^2)}{n_3} + \frac{(\sum x_4)^2}{n_4} - \frac{(\sum x_i)^2}{n} =$$

$$= \frac{634^2}{10} + \frac{656^2}{10} + \frac{765^2}{10} + \frac{818^2}{10} - \frac{2873^2}{40} = 2310{,}88$$

3. Z celkového součtu čtverců a ze součtu čtverců mezi skupinami vypočítáme součet čtverců uvnitř skupin

$$S\check{C}_{WG} = S\check{C}_T - S\check{C}_{BG} = 3840{,}9$$

4. Vypočítáme rozptyly mezi skupinami a uvnitř skupin

$$s_{BG}^2 = \frac{S\check{C}_{BG}}{\nu_1} = 770{,}29 \; ; \; s_{WG}^2 = \frac{S\check{C}_{WG}}{\nu_2} = 106{,}69$$

5. Vypočítáme Fisherovo F: $F = \frac{s_{BG}^2}{s_{WG}^2} = \frac{770{,}29}{106{,}69} = 7{,}22$

6. Vypočítanou hodnotu F porovnáme s kritickou hodnotou, kterou najdeme ve statistických tabulkách pro hladinu významnosti $\alpha = 0{,}001$ a počty stupňů volnosti: $\nu_1 = 3$, $\nu_2 = 36$: $F = 7{,}22 > F_{0{,}001}(3, 36) = 2{,}886$. Protože je vypočítaná hodnota větší než tabulková, zamítáme nulovou hypotézu o rovnosti rozptylů a můžeme říct, že mezi výsledky jednotlivých skupin žáků jsou statisticky významné rozdíly na hladině významnosti $\alpha = 0{,}001$. Tímto končí jednofaktorová analýza rozptylu, jejíž výsledky zapíšeme do následující tabulky:

Zdroj rozptylu	součet čtverců	ν	s^2	F
mezi skupinami	$S\check{C}_{BG}$ = 2310,875	4 – 1 = 3	770,29	7,22
uvnitř skupin	$S\check{C}_{WG}$ = 3840,9	40 – 4 = 36	106,69	–
celkem	$S_{\check{C}T}$ = 108,5	40 – 1 = 39	–	–

Analýza rozptylu v Excelu

Data ⇒ Analýza dat ⇒ Anova: jeden faktor

Výsledky jednofaktorové analýzy rozptylu v Excelu:

Faktor

Výběr	Počet	Součet	Průměr	Rozptyl
A	10	634	63,4	86,04444
B	10	656	65,6	103,1556
C	10	765	76,5	108,9444
D	10	818	81,8	128,6222

ANOVA

Zdroj variability	SS	Rozdíl	MS	S	Hodnota P	F krit
Mezi výběry	2310,875	3	770,2917	7,219792	0,000649	2,866266
Všechny výběry	3840,9	36	106,6917			
Celkem	6151,775	39				

F = 7,21979 > F krit=2,86626 → F je velmi vysoce signifikantní na α = 0,001 (vypočítaná hodnota P = 0,00064 < 0,001).

5.5.2 Neparametrická analýza rozptylu

Pokud pracujeme s malými soubory nebo se soubory, které nevykazují normální rozložení četností, nebo s ordinálními daty, použijeme ke zjištění rozdílu mezi soubory neparametrickou alternativu analýzy rozptylu.

5.5.2.1 Friedmannova analýza rozptylu pro k závislé výběry

Ordinální data

Závislé výběry

Tato metoda je neparametrickou obměnou parametrické analýzy rozptylu pro ***ordinální data*** (tj. pořadové hodnoty). Pracuje s k ***závislými výběry***, které jsou všechny stejného rozsahu. Jsou to nějčastěji výsledky stejných pokusných osob za k různých

podmínek. Výhodou pořadové Friedmannovy analýzy rozptylu je to, že se nemusíme zajímat o normalitu rozdělení, které je třeba brát v úvahu při její parametrické obdobě. Protože neparametrický postup této statistiky vychází z pořadových hodnot, je aplikovatelný na mnohem širší okruh výzkumných problémů. Výhoda jednoduchosti výpočtů kompenzuje jedinou nevýhodu tohoto postupu, totiž ztrátu informace, která vzniká tím, že místo původních naměřených hodnot používáme jejich pořadí.

Tabulka dvojného třídění

Základem je ***tabulka dvojného třídění***, která má n řádků a k sloupců, v ní n označuje výběrový rozsah stejný pro všechny vyšetřované skupiny a k je počet výběrů. Nulová hypotéza předpokládá, že všechny výběry pocházejí z jediné populace, bereme-li v úvahu n pokusných osob za k různých podmínek, a distribuce výsledků zkoumaných osob je za všech podmínek stejná. Nulovou hypotézu testujeme takto: nejprve přiřadíme prvkům každého jednotlivého řádku pořadí od 1 do k a pak sestavíme novou tabulku dvojného třídění, v níž původní hodnoty zastupují pořadová čísla od nejnižšího po nejvyšší. Platí-li nulová hypotéza, pocházejí-li tedy všechny výběry skutečně z jediné populace, budou pořadí v každém řádku rozdělena náhodně, a tedy i pořadí v každém sloupci budou tvořit náhodný výběr rozsahu n.

Testová charakteristika je založena na srovnání průměrných pořadí jednotlivých sloupců s populačním průměrem a má rozdělení χ^2 s $k-1$ stupni volnosti. Vypočítá se podle následující rovnice:

$$\chi_r^{\,2} = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 - 3n(k+1)$$

kde r_{ij} je pořadí, které je zapsáno v i-tém řádku a j-tém sloupci. Toto testové kritérium porovnáme s tabulkovou hodnotou rozdělení χ^2.

Příklad

Zjistěte, zda 4 sady (I, II, III, IV) úkolů pro výzkum logického myšlení jsou stejně obtížné, či zda se od sebe co do obtížnosti liší. Zkouška byla provedena tak, že celkem čtyři vybrané skupiny osob (A, B, C, D) prošly všemi čtyřmi sadami úkolů. Výsledky jsou shrnuty do tabulky, ve které je zachycen průměrný počet vyřešených úkolů u jednotlivých skupin (maximálně bylo možno vyřešit 10 úkolů).

Úkoly

Skupiny	I	II	III	IV
A	9	7	4	5
B	6	5	2	8
C	8	3	5	7
D	9	6	1	4

Nyní hodnotám každého řádku přiřadíme pořadí a sestavíme tak novou tabulku:

Úkoly

Skupiny	I	II	III	IV
A	1	2	4	3
B	2	3	4	1
C	1	4	3	2
D	1	2	4	3
Σ	5	11	15	9

Nezáleží na tom, zda řadíme hodnoty v jednotlivých řadách vzestupně či sestupně, musíme však stejný postup dodržet ve všech řádcích. Nulovou hypotézou je předpoklad, že všechny sady úkolů jsou stejně těžké, tj. že pořadí v každém sloupci je rozloženo náhodně. Hodnotu testové charakteristiky vypočítáme podle vzorce

$$\chi_r^{\,2} = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 - 3n(k+1)$$

$$= \frac{12}{4 \cdot 4 \cdot 5} \left(5^2 + 11^2 + 15^2 + 9^2 \right) - 3 \cdot 4 \cdot 5 = 7{,}8$$

Z tabulky hodnot χ^2 zjistíme, že hodnota 7,8 je na hranici významnosti.
$\chi^2(4-1) = 7{,}81 \quad \chi^2 < \chi^2(3) \Rightarrow$
nezamítáme nulovou hypotézu a můžeme říct, že jednotlivé sady úkolů jsou stejně obtížné.

5.5.2.2 Kruskal-Wallisova analýza rozptylu

Touto metodou se ověřuje homogenita rozdělení vyšetřovaného znaku v několika populacích, z nichž máme k dispozici ***nezávislé náhodné výběry***. Nulová hypotéza nám říká, že všechny výběry pocházejí z téže populace. Obvyklou metodou pro ověřování této hypotézy je parametrická analýza rozptylu, Její použití však vyžaduje, aby výběry pocházely z populací s normální distribucí. Tento předpoklad je v praxi zřídka splněn, proto je vítáno zeslabení předpokladů, které přináší Kruskal-Wallisův

Nezávislé náhodné výběry

test. Výhodou tohoto testu oproti klasické analýze rozptylu je podstatné zjednodušení výpočtů.

Pracujeme s k nezávislými výběry $n_1, n_2, n_3, \ldots n_k$, kde n_i je velikost výběru z i-té populace a k je počet srovnávaných výběrů. Všechny naměřené hodnoty seřadíme vzestupně podle velikosti do jediné posloupnosti a každé z nich přiřadíme pak pořadí, které jí přísluší v celkovém uspořádání. Pro každý výběr sečteme pořadové hodnoty všech jeho členů a jejich součet v i-tém výběru označíme R_i. Platí-li nulová hypotéza, potom průměry pořadí jednotlivých výběrů (R_i / n_i) by se neměly signifikantně odlišovat a pozorované odchylky by měly vyjadřovat pouze náhodné kolísání. Budou-li se vysoká a nízká pořadí objevovat převážně v některých výběrech a v jiných ne, můžeme usuzovat na neplatnost nulové hypotézy. Za testovou charakteristiku navrhli Kruskal a Wallis veličinu H ve tvaru

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1), \qquad N = \sum_{i=1}^{k} n_i$$

kde R_i je součet pořadových hodnot v i-tém výběru a N je součet rozsahů jednotlivých výběrů, se kterými pracujeme a n_i je velikost i-tého výběru.

Za předpokladu, že rozsahy výběrů nejsou příliš malé (ne menší než 5), má veličina H přibližně χ^2 – rozdělení s $k - 1$ stupni volnosti. Proto pro test významnosti můžeme použít obvyklých tabulek kritických hodnot χ^2 – distribuce.

Příklad

Máme tři skupiny pacientů s různými diagnózami. Předkládáme jim určitý úkol a měříme čas, který potřebují pro jeho splnění. Data jsou sestavena v následující tabulce:

Výsledky pacientů při řešení úkolu

Skupiny	Čas v sekundách	Počet
A	96, 123, 85, 67, 112	$n_1 = 5$
B	79, 127, 139, 142	$n_2 = 4$
C	119, 153, 162, 149	$n_3 = 4$

Jako nulovou hypotézu si stanovíme předpoklad, že se pacienti jednotlivých skupin neliší ve výkonu, tj. že jejich časy můžeme pokládat za náhodný výběr z jediné populace.
Data nejprve sestavíme do pořadí, tato znovu uspořádáme do tabulky a sečteme je pro každý ze tří výběrů (skupin):

Pořadové hodnoty

Skupiny	Pořadí	R_i
A	4, 7, 3, 1, 5	20
B	2, 8, 9, 10	29
C	6, 12, 13, 11	42

Hodnoty dosadíme do vzorce pro H

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) =$$

$$= \frac{12}{13\cdot 14}\left[\frac{20^2}{5} + \frac{29^2}{4} + \frac{42^2}{4}\right] - 3\cdot 14 = 6{,}21$$

Z tabulky plyne, že vypočítaná hodnota $H > H_\alpha$ (= 5,6172) pro $\alpha = 0{,}05$. Nulovou hypotézu tady zamítáme a můžeme konstatovat, že při řešení daného úkolu je mezi zkoumanými diagnostickými skupinami významný rozdíl.

Kontrolní otázky a úkoly

1. K čemu slouží analýza rozptylu?
2. Z čeho se skládá závěrečná tabulka analýzy rozptylu?
3. V čem se liší použití Friedmannovy a Kruskal-Wallisovy analýzy rozptylu?

Klíč k otázkám a úkolům

Odpovědi na otázky naleznete v textu.

Referenční seznam

MELOUN, M., MILITKÝ, J., HILL, M. 2012. *Statistická analýza vícerozměrných dat v příkladech*. Praha: Academia. ISBN 978-80-200-2071-0.

REITEROVÁ, E. 2008. *Základy psychometrie.* Olomouc: UP. ISBN 978-80-244-2065-3.

6 Velikost výběrového souboru

Power analýza

Odhad optimální velikosti výběrového souboru v daném výzkumném šetření řeší ***Power analýza*** – analýza síly testu. Jejím cílem je zajistit, aby nedocházelo k plýtvání zdroji tím, že se budou realizovat výzkumy, které mají jen malou šanci odhalit statisticky významný efekt. Rozhodování o velikosti výběru je založeno na designu výzkumu, hlavním cíli výzkumu (hlavní výzkumné otázce), na odhadovaném počtu případů (pacientů), které jsme schopni v daném časovém období sehnat, na dostupných zdrojích a na matematickém výpočtu.

Nevhodná velikost souboru

Nevhodná velikost souboru – pokud máme v souboru příliš málo případů, tak soubor neposkytuje spolehlivou, jasnou a přesvědčivou odpověď na výzkumnou otázku, výsledky nejsou dostatečně přesné. Malá velikost souboru může vést ke zbytečnému zamítnutí přístupu, který by byl pro pacienty přínosný. Naopak příliš velký výběrový soubor znamená plýtvání zdroji a prostředky a chybná interpretace výsledků může vést k přeceňování statistické významnosti tam, kde výsledek není klinicky významný.

Vhodná velikost souboru

Vhodná velikost souboru – při určování velikosti výběrového souboru musíme vědět, zda nás zajímají výsledky za celkový soubor, nebo zda budeme pracovat pouze s částmi souboru (muži/ženy, věkové skupiny, regiony apod.). Obecně platí, že čím větší je výběrový soubor, tím jsou výsledky přesnější, ale nikoliv lineárně. Výběrové soubory přes 1000 respondentů se pro populaci ČR využívají méně často, protože dodatečné náklady na realizaci výzkumu N = 5000 při porovnání přesnosti s variantou N = 1000 nejsou ve většině případů efektivní. Parametry základního souboru se odhadují se z výběrových charakteristik. Čím větší máme výběrový soubor, tím přesnější získáme odhad parametru základního souboru.

Matematický výpočet velikosti souboru

Při výpočtu vhodné velikosti výběrového souboru pro případ testování statistických hypotéz musíme brát v úvahu následující předpoklady:

- metodu analýzy dat – pro situaci danou typem výzkumu je třeba volit nejsilnější statistickou metodu, pro niž jsou splněny podmínky k použití (např. parametrické × neparametrické testy);
- zvolená hladina významnosti α (standardně $\alpha = 0{,}05$; což znamená 5% hladinu významnosti);
- požadovaná síla testu (1–β) – nejčastější volba = 0,8 (80 %), někdy se také zadává 0,85 nebo 0,9;

– další nutné vstupy závisí na metodě analýzy dat, odhad směrodatné odchylky se nejčastěji pořizuje z dřívějších studií nebo z pilotního výzkumu.

Nejčastějšími případy situací, které se vyskytují v klinickém výzkumu, jsou:
a) porovnání proporcí (%) ve dvou skupinách;
b) porovnání průměrů ve dvou skupinách.

Počet případů (pacientů) v každé skupině vypočítáme podle vzorce $$n = \frac{2*(z_{(1-\alpha/2)} + z_{(1-\beta)})^2}{\Delta^2}$$

α ..... hladina významnosti
$1-\beta$ ..... síla (mohutnost) testu, viz kapitola Testování statistických hypotéz
z ..... příslušný kvantil standardizovaného normálního rozdělení
Δ ..... standardizovaná diference

a) pro porovnání proporcí $\Delta = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})}}$, kde $\bar{p} = \frac{p_1 + p_2}{2}$

b) pro porovnání průměrů $\Delta = \frac{\mu_1 - \mu_2}{s}$.

Tabulka Nejčastěji používané hodnoty kvantilů

Hladina významnosti α			Síla testu $1-\beta$			
0,05	0,01	0,001	0,8	0,85	0,9	0,95
$z(1-\alpha/2)$			$z(1-\beta)$			
1,96	2,58	3,29	0,84	1,04	1,29	1,64

Příklad

Porovnání procent ve dvou nezávislých skupinách

Při standardní terapii dosáhlo 40 % pacientů příznivého výsledku, při nové terapii se očekává zlepšení na 50 %. Jaký je potřebný počet případů ve skupinách, aby byl odhalen efekt léčby na 5% hladině významnosti a s 80% sílou testu?

$$\bar{p} = \frac{p_1 + p_2}{2} = \frac{0{,}5 + 0{,}4}{2} = 0{,}45$$

$$\Delta = \frac{p_1 - p_2}{\sqrt{\bar{p}\cdot(1-\bar{p})}} = \frac{0{,}5 - 0{,}4}{\sqrt{0{,}45\cdot(1-0{,}45)}} = 0{,}201$$

$$n = \frac{2\cdot(z_{(1-\alpha/2)} + z_{(1-\beta)})^2}{\Delta^2} = \frac{2\cdot(1{,}96 + 0{,}84)^2}{0{,}201^2} = 388{,}5$$

V každé skupině je potřeba nejméně 389 pacientů, tedy dohromady 778 pacientů.

Příklad

Porovnání průměrů ve dvou nezávislých skupinách

Jaký je potřebný počet pacientů s mírnou hypertenzí, aby byl odhalen rozdíl 5 mm Hg systolického krevního tlaku mezi kontrolní a experimentální skupinou? Předpokládejme, že směrodatná odchylka systolického tlaku je 10 mm Hg, hladinu významnosti zvolíme 5% a sílu testu 90%.

$$\Delta = \frac{5}{10} = 0{,}5$$

$$n = \frac{2 \cdot (z_{(1-\alpha/2)} + z_{(1-\beta)})^2}{\Delta^2} = \frac{2 \cdot (1{,}96 + 1{,}28)^2}{0{,}5^2} = 84{,}1$$

V každé skupině je potřeba nejméně 85 pacientů, tedy dohromady 170 pacientů.

Poznámka: hodnoty n zaokrouhlujeme vždy nahoru.

Tabulka

Orientační velikost výběrového souboru při dané velikosti základního souboru

Velikost ZS	VS	%
100	70	70,0
1 000	350	35,0
10 000	1 000	10,0
50 000	1 250	2,5
100 000	1 500	1,5
1 000 000	5 000	0,5

Kontrolní otázky a úkoly

1. Co je cílem Power analýzy?
2. Jaká je nevhodná velikost souboru?

Klíč k otázkám a úkolům

Odpovědi na otázky naleznete v textu.

Referenční seznam

ŽIAKOVÁ, K. et al., 2009. *Ošetrovatelstvo – Teória a vedecký výskum*. Bratislava: Osveta, ISBN 978-80-8063-304-2.

http://www.ouh.nhs.uk/researchers/planning/is-it-research/documents/medical-statistics-online-help.pdf

7 Příklady k procvičení

I. Základní operace se součty

1. Vypočítejte: $\sum_{i=1}^{3} 5.\frac{x_i}{y_i + 1} =$

$x_1 = 2$

$x_2 = 3$

$x_3 = 4$

$y_1 = 1$

$y_2 = 2$

$y_3 = 3$

2. Dosaďte za x a y, vypočítejte z:

$$z = \frac{\sum_{i=1}^{5}(x_i + y_i) - \sum_{i=1}^{5} x_i^{\,2}}{\sum_{i=1}^{5} x_i y_i}$$

$x_i = 1, 2, 3, 4, 5$ $\quad$ $y_i = 2, 4, 3, 1, 1$

II. Popisná statistika

3. Doplňte další druhy četností:

Interval	Absolutní četnost	Relativní četnost	Absolutní kumulativní četnost	Relativní kumulativní četnost
11–15	5			
16–20	3			
21–25	7			
26–30	2			

4. Určete všechny střední hodnoty (aritmetický průměr, medián a modus) u souboru dat:

22, 24, 16, 18, 12, 30, 25, 22, 31, 19, 18, 15, 21, 22, 31, 16, 14, 13, 20, 26, 31, 16

5. Vypočítejte míry variability pro tyto naměřené hodnoty:

5, 8, 9, 10, 13, 14, 16, 19, 22, 24

6. Vypočítejte na zadaném souboru dat všechny popisné statistiky v Excelu:

Věk	Váha	Výška
35	92,8	175
45	78	174
47	100,5	183
54	64	167
75	47	154
69	80,7	171,5
33	87,2	172,5
42	97,9	186
32	60	167,8
42	73,5	182
24	85,6	175,5
44	57	170
69	98,3	174,5
54	89,2	167,5
25	93,3	174
22	95,1	176,5
34	95,9	167,5
59	48	155
55	53	176,5
64	98,6	169
60	68	166
66	85,3	169
66	55	174

7. Vypočítejte v Excelu Pearsonovu korelaci mezi věkem a váhou, věkem a výškou, váhou a výškou z výše uvedených dat.

8. Vypočítejte Spearmanův koeficient pořadové korelace z dat v tabulce:

x_i	y_i
1	3
2	5
3	7
4	2
5	4
6	1
7	6
8	9
9	8

9. Vytvořte v Excelu tabulku četností a histogram ze zadaných dat:

Pohlaví	Vzdělání	Bydliště
2	3	2
2	2	1
2	2	3
1	2	4
1	1	2
2	2	1
2	4	1
2	4	3
1	4	1
2	2	2
1	1	3
1	1	1
2	2	1
1	1	3
2	3	2
2	4	1
2	2	3

10. Vytvořte v Excelu kontingenční tabulky z proměnných pohlaví × vzdělání, pohlaví × bydliště, vzdělání × bydliště

11. Vypočítejte ze zadané tabulky čtyřpolní koeficient korelace:

14	17	Σ
10	19	Σ
Σ	Σ	

12. Vypočítejte Pearsonovu korelaci z proměnných X a Y:

X: 3, 7, 11, 14, 15
Y: 3, 5, 7, 6, 9

III. Testování statistických hypotéz

13. Zjistěte pomocí testu χ^2 zda jsou počty pacientů, kteří se dostavili na zubní pohotovost rovnoměrně rozloženy na jednotlivé dny v týdnu, nebo zda se v některé dny počet pacientů signifikantně odlišuje od očekávané četnosti.
po – 8, út – 9, st – 15, čt – 20, pá – 17, so – 5, ne – 6

14. Určete testem χ^2 pro čtyřpolní tabulku, zda se od sebe liší účinnost dvou léků A a B.

	Lék A	Lék B
Úspěch	23	12
Neúspěch	8	7

15. Zjistěte, zda existuje souvislost mezi spokojeností s pobytem v nemocnici a dosaženým stupněm vzdělání. Výsledky dotazníkového šetření, které bylo provedeno u 400 pacientů, uvádí následující tabulka:

		Spokojenost			
		Plně spokojen	Vcelku spokojen	nespokojen	Σ
Stupeň vzdělání	Základní	22	15	133	
	Stř. všeobecné	25	11	74	
	Stř. odborné	20	8	42	
	Vysokoškolské	30	7	13	
	Σ				

16. Určete míru souvislosti mezi kvantitativními proměnnými v tabulce:

věk	cholesterol	LDL	HDL
25	4,56	2,36	1,33
30	3,81	2,15	1,69
35	4,98	0,89	1,12
38	4,55	0,96	1,25
49	4,96	3,22	1,05
52	5,01	3,42	1,01
55	5,68	2,11	0,69
58	5,98	2,45	0,97
60	5,97	2,65	0,89
65	6,05	3,21	0,78

17. Posuďte, zda se od sebe odlišují v rozptylech dva výběrové soubory. Odhady výběrových rozptylů jsou 64 (při $n = 30$) a 25 (při $n = 20$).

18. Zjistěte, zda výběr dvaceti subjektů, u nichž byly naměřeny hodnoty dané v tabulce, pochází ze základního souboru s průměrem $\mu = 58$.

Subjekt	x_i	Subjekt	x_i
S1	64	S11	60
S2	48	S12	43
S3	55	S13	67
S4	68	S14	70
S5	72	S15	65
S6	59	S16	55
S7	57	S17	56
S8	61	S18	64
S9	63	S19	61
S10	60	S20	60

19. Zjistěte, zda je statisticky významný rozdíl mezi výsledky testu mužů a žen.

Subjekt – muž	skór-x_i	Subjekt – žena	skór-x_i
S1	15	S12	20
S2	26	S13	22
S3	32	S14	27
S4	48	S15	31
S5	52	S16	38
S6	63	S17	44
S7	72	S18	46
S8	80	S19	57
S9	86	S20	59
S10	85	S21	65
S11	89	S22	74
		S23	77

20. Zjistěte, zda redukční dieta způsobila signifikantní úbytek váhy u deseti osob. Údaje o hmotnosti na začátku diety a na jejím konci jsou v tabulce.

Osoba	Začátek diety	Konec diety
A	59	57
B	68	64
C	65	61
D	86	85
E	88	84
F	79	76
G	62	60
H	57	55
I	64	61
J	74	72

21. Byla řešena otázka, zda se určité onemocnění vyskytuje častěji u mužů než u žen. Po příslušném vyšetření bylo nalezeno 23 onemocnění u 58 mužů a 28 onemocnění u 43 žen z předem vybrané lokality. Prokažte, zda je vyšší relativní četnost onemocnění u mužů proti ženám statisticky významná.

Výsledky příkladů

1. $$\sum_{i=1}^{3} 5.\frac{x_i}{y_i+1} = 5.\sum_{i=1}^{3} \frac{x_i}{y_i+1} = 5.\left(\frac{x_1}{y_1+1}+\frac{x_2}{y_2+1}+\frac{x_3}{y_3+1}\right)=$$
$$5.\left(\frac{2}{1+1}+\frac{3}{2+1}+\frac{4}{3+1}\right)=5.\left(\frac{2}{2}+\frac{3}{3}+\frac{4}{4}\right)=5.3=\underline{\underline{15}}$$

2. $$z=\frac{\sum_{i=1}^{5}(x_i+y_i)-\sum_{i=1}^{5}x_i^{\,2}}{\sum_{i=1}^{5}x_i y_i}=\frac{26-55}{28+31}=\frac{-29}{59}=-0{,}49$$

3.

Interval	Absolutní četnost	Relativní četnost	Absolutní kumulativní četnost	Relativní kumulativní četnost
11–15	5	0,29	5	0,29
16–20	3	0,18	8	0,47
21–25	7	0,41	15	0,88
26–30	2	0,12	17	1

4. $\mu = 21 \quad M_e = 20{,}5 \quad M_{o1} = 16 \quad M_{o2} = 22 \quad M_{o3} = 31$

5. $\sigma^2 = 35{,}2 \quad \sigma = 5{,}93 \quad R = 19$

6.

Věk		*Váha*		*Výška*	
Stř. hodnota	48,52174	Stř. hodnota	78,43043	Stř. hodnota	171,6435
Chyba stř. hodnoty	3,332597	Chyba stř. hodnoty	3,778729	Chyba stř. hodnoty	1,570929
Medián	47	Medián	85,3	Medián	172,5
Modus	54	Modus	########	Modus	174
Směr. odchylka	15,98257	Směr. odchylka	18,12215	Směr. odchylka	7,533913
Rozptyl výběru	255,4427	Rozptyl výběru	328,4122	Rozptyl výběru	56,75984
Špičatost	-1,17066	Špičatost	-1,28392	Špičatost	1,146504
Šikmost	-0,0915	Šikmost	-0,46738	Šikmost	-0,54728
#ODKAZ!	53	#ODKAZ!	53,5	#ODKAZ!	32
Minimum	22	Minimum	47	Minimum	154
Maximum	75	Maximum	100,5	Maximum	186
Součet	1116	Součet	1803,9	Součet	3947,8
Počet	23	Počet	23	Počet	23

7.

	Věk	*Váha*	*Výška*
Věk	1		
Váha	-0,35417	1	
Výška	-0,41771	0,552533	1

8. $R = 0{,}84$

9.

kat.pohlaví	*Četnost*	*Kumul. %*	*kat. pohlaví*	*Četnost*	*Kumul. %*
1	6	35,29 %	2	11	64,71%
2	11	100,00 %	1	6	100,00 %
Další	0	100,00 %	Další	0	100,00 %

kat.vzdělání	*Četnost*	*Kumul. %*	*kat. vzdělání*	*Četnost*	*Kumul. %*
1	4	23,53 %	2	7	41,18 %
2	7	64,71 %	1	4	64,71 %
3	2	76,47 %	4	4	88,24 %
4	4	100,00 %	3	2	100,00 %
Další	0	100,00 %	Další	0	100,00 %

kat.bydliště	*Četnost*	*Kumul. %*	*kat. bydliště*	*Četnost*	*Kumul. %*
1	7	41,18 %	1	7	41,18 %
2	4	64,71 %	3	5	70,59 %
3	5	94,12 %	2	4	94,12 %
4	1	100,00 %	4	1	100,00 %
Další	0	100,00 %	Další	0	100,00 %

10.

Kontingeční tabulka (List 1 v Kont. tabulky)
Četnost označených buněk > 10
(Marginální součty nejsou označeny)

Pohlaví	Vzdělání 1	Vzdělání 2	Vzdělání 3	Vzdělání 4	Řádk. součty
1	4	1	0	1	6
2	0	6	2	3	11
vš. skup.	4	7	2	4	17

Kontingeční tabulka (List 1 v Kont. tabulky)
Četnost označených buněk > 10
(Marginální součty nejsou označeny)

Pohlaví	Bydliště 1	Bydliště 2	Bydliště 3	Bydliště 4	Řádk. součty
1	2	1	2	1	6
2	5	3	3	0	11
vš. skup.	7	4	5	1	17

Kontingeční tabulka (List 1 v Kont. tabulky) Četnost označených buněk > 10 (Marginální součty nejsou označeny)					
Pohlaví	Bydliště 1	Bydliště 2	Bydliště 3	Bydliště 4	Řádk. součty
1	1	1	2	0	4
2	3	1	2	1	7
3	0	2	0	0	2
4	3	0	1	0	4
vš. skup.	7	4	5	1	17

11. $\Phi = 0{,}1$

12. $r = 0{,}89$

13. $\chi^2 = 10$, nesignifikantní výsledek, není sgn rozdíl mezi četnostmi osob ve dnech v týdnu

14. $\chi^2 = 16{,}67$, léky se od sebe co do účinnosti signifikantně odlišují

15. $\chi^2 = 53{,}6,\ C = 0{,}34$

Mezi spokojeností s pobytem v nemocnici a dosaženým stupněm vzdělání existuje souvislost, míra této souvislosti je dána koeficientem C.

16. Pearsonovy korelace

	Věk	*Cholesterol*	*LDL*	*HDL*
věk	1			
cholesterol	0,888729	1		
LDL	0,542504	0,34000277	1	
HDL	-0,84675	-0,9097442	-0,32624	1

17. Mezi rozptyly je signifikantní rozdíl, $F = 2{,}56$

18. $t = 1{,}53$ nesignifikantní výsledek, výběrový soubor pochází ze základního souboru s průměrem 58

19. $t = 1{,}28$ nesignifikantní výsledek, není signifikantní rozdíl mezi skóry mužů a žen

20. $t = 8{,}06 > t_\alpha(\nu) = t_{0{,}01}(9) = 3{,}24$

Dieta způsobila signifikantní úbytek váhy.

21. $t = 2{,}52$ V četnostech onemocnění mezi muži a ženami je signifikantní rozdíl

Výtah ze statistických tabulek

Kritické hodnoty rozdělení F

$\alpha = 0{,}05$

ν_2/ν_1	9	10	12	15	20	24	30	40	60	120
9	3,179	3,137	3,073	3,006	2,937	2,901	2,864	2,826	2,787	2,748
10	3,020	2,978	2,913	2,845	2,774	2,737	2,7	2,661	2,621	2,58
11	2,896	2,854	2,788	2,719	2,646	2,609	2,571	2,531	2,49	2,448
14	2,646	2,602	2,534	2,463	2,388	2,349	2,308	2,266	2,223	2,178
19	2,423	2,378	2,308	2,234	2,156	2,114	2,071	2,026	1,98	1,93
20	2,393	2,348	2,278	2,203	2,124	2,083	2,039	1,994	1,946	1,896
21	2,366	2,321	2,25	2,176	2,096	2,054	2,01	1,965	1,917	1,866
24	2,3	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,79
29	2,223	2,177	2,105	2,028	1,945	1,901	1,854	1,806	1,754	1,698
30	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,74	1,684
40	2,124	2,077	2,004	1,925	1,839	1,793	1,744	1,693	1,637	1,577
60	2,04	1,993	1,917	1,836	1,748	1,7	1,649	1,594	1,534	1,467
120	1,959	1,911	1,834	1,751	1,659	1,608	1,554	1,495	1,429	1,352

$\alpha = 0{,}01$

ν_2/ν_1	9	10	12	15	20	24	30	40	60	120
9	5,351	5,257	5,111	4,962	4,808	4,729	4,649	4,567	4,483	4,398
10	4,942	4,849	4,706	4,405	4,405	4,327	4,247	1,165	4,082	3,997
11	4,632	4,539	4,397	4,099	4,099	4,021	3,941	3,86	3,776	3,690
14	4,03	3,939	3,8	3,656	3,505	3,427	3,348	3,266	3,181	3,094
19	3,523	3,434	3,297	3,153	3,003	2,925	2,844	2,761	2,674	2,584
20	3,457	3,368	3,231	3,088	2,938	2,859	2,779	2,695	2,608	2,517
21	3,398	3,31	3,173	3,03	2,880	2,801	2,72	2,636	2,548	2,457
24	3,256	3,168	3,032	2,889	2,738	2,659	2,577	2,492	2,404	2,31
29	3,092	3,005	2,869	2,726	2,574	2,495	2,412	2,325	2,234	2,138
30	3,067	2,979	2,843	2,7	2,549	2,469	2,386	2,299	2,208	2,111
40	2,888	2,801	2,665	2,522	2,369	2,288	2,203	2,114	2,019	1,917
60	2,719	2,632	2,496	2,352	2,198	2,115	2,029	1,936	1,836	1,726
120	2,559	2,472	2,336	2,192	2,035	1,95	1,86	1,763	1,656	1,533

Kritické hodnoty rozdělení t

$\nu \backslash \alpha$	0,05	0,01
1	12,706	63,657
2	4,3027	9,9248
3	3,1825	5,8409
4	2,7764	4,6041
5	2,5706	4,0321
6	2,4469	3,7074
7	2,3646	3,4995
8	2,3060	3,3554
9	2,2622	3,2498
10	2,2281	3,1693
11	2,2010	3,1058
12	2,1788	3,0545
13	2,1604	3,0123
14	2,1448	2,9768
15	2,1315	2,9467
16	2,1199	2,9208
17	2,1098	2,8982
18	2,1009	2,8784
19	2,0930	2,8609
20	2,0860	2,8453
21	2,0796	2,8314
22	2,0739	2,8188
23	2,0687	2,8073
24	2,0639	2,7969
25	2,0595	2,7874
26	2,0555	2,7787
27	2,0518	2,7707
28	2,0484	2,7633
29	2,0452	2,7564
30	2,0423	2,7500
40	2,0211	2,7045
60	2,0003	2,6603
120	1,9799	2,6174
∞	1,9600	2,5758

Kritické hodnoty rozdělení χ^2

	Hladina významnosti	
Počet stupňů volnosti ν	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	3,84	6,63
2	5,99	9,21
3	7,81	11,3
4	9,49	13,3
5	11,1	15,1
6	12,6	16,8
7	14,1	18,5
8	15,5	20,1
9	16,9	21,7
10	18,3	23,2
11	19,7	24,7
12	21,0	26,2
13	22,4	27,7
14	23,7	29,1
15	25,0	30,6
16	26,3	32,0
17	27,6	33,4
18	28,9	34,8
19	30,1	36,2
20	31,4	37,6
21	32,7	38,9
22	33,9	40,3
23	35,2	41,6
24	36,4	43,0
25	37,7	44,3
26	38,9	45,6
27	40,1	47,0
28	41,3	48,3
29	42,6	49,6
30	43,8	50,9
31	45,0	52,2
32	46,2	53,5
33	47,4	54,8
34	48,6	56,1
35	49,8	57,3
36	51,0	58,6

Kritické hodnoty pro Wilcoxonův test

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	9
14	21	12
15	25	15
16	30	18
17	35	22
18	40	26
19	46	30
20	52	36
21	59	41
22	66	47
23	73	53
24	81	59
25	89	66

Kritické hodnoty S pro znaménkový test

počet dvojic	počet znamének	počet dvojic	počet znamének
14	2	21	5
15	3	22	5
16	3	23	6
17	4	24	6
18	4	25	7
19	4	26	7
20	5	27	8

Kritické hodnoty korelačního koeficientu (signifikantnost korelace)

n/r_α	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	0,997	1,000
2	0,950	0,990
3	0,878	0,959
4	0,811	0,917
5	0,754	0,874
6	0,707	0,834
7	0,666	0,798
8	0,632	0,765
9	0,602	0,735
10	0,576	0,707
11	0,553	0,684
12	0,532	0,661
13	0,514	0,641
14	0,497	0,623
15	0,482	0,605
16	0,468	0,590
17	0,456	0,575
18	0,444	0,561
19	0,433	0,549
20	0,422	0,536
21	0,413	0,526
22	0,404	0,515
23	0,396	0,505
24	0,388	0,496
25	0,380	0,486
26	0,374	0,478
27	0,367	0,470
28	0,361	0,463
29	0,355	0,456
30	0,349	0,448
35	0,324	0,418
40	0,304	0,393
50	0,273	0,354
60	0,250	0,324
70	0,231	0,301
80	0,217	0,283
90	0,205	0,267
100	0,194	0,254

Je-li $r > |r_\alpha|$, pak r je signifikantní na hladině významnosti α.

Referenční seznam

CHRÁSKA, M. 2007. *Metody pedagogického výzkumu. Základy kvantitativního výzkumu.* Praha: Grada. ISBN978-80-247-1369-4.

CYHELSKÝ, L., KAHOUNOVÁ, J., HINDLS, R. 2001. *Elementární statistická analýza*. Praha: Management Press. ISBN 80-7261-003-1.

DUPAČ, V., HUŠKOVÁ, M. 2013. *Pravděpodobnost a matematická statistika*. Praha: Karolinum. ISBN 978-80-246-2208-8.

HANOUSEK, J., CHARAMZA, P. 1992. *Moderní metody zpracování dat*. Praha: Grada. ISBN 80-85623-31-5.

HENDL, J. 2004. *Přehled statistických metod zpracování dat.* Praha: Portál. ISBN 80-7178-820-1.

KUNDEROVÁ, P. 2004. *Úvod do teorie pravděpodobnosti a matematické statistiky*. Olomouc: UP. ISBN 80-244-0843-0.

MELOUN, M., MILITKÝ, J. 2004. *Statistická analýza experimentálních dat*. Praha: Academia. ISBN 80-200-1254-0.

MELOUN, M., MILITKÝ, J., HILL, M. 2012. *Statistická analýza vícerozměrných dat v příkladech*. Praha: Academia. ISBN 978-80-200-2071-0.

PROCHÁZKA, B. 2015. *Stručná biostatistika pro lékaře*. Praha: Karolinum. ISBN 978-80-246-2783-0.

REITEROVÁ, E. 2011. *Základy statistiky pro studenty psychologie*. Olomouc: UP. ISBN 978-80-244-2316-6.

REITEROVÁ, E. 2008. *Základy psychometrie*. Olomouc: UP. ISBN 978-80-244-2065-3.

ŠŤASTNÝ, Z. 1999. *Matematické a statistické výpočty v Microsoft Excelu*. Brno: Computer Press. ISBN 80-7226-141-X.

WALKER, I. 2013. *Výzkumné metody a statistika*. Praha: Grada. ISBN 978-80-247-3920-5.

ZVÁRA, K. 2013. *Základy statistiky v prostředí R*. Praha: Karolinum. ISBN 978-80-246-2245-3.

ZVÁROVÁ, J. 2001. *Základy statistiky pro biomedicínské obory*. Praha: Karolinum. ISBN 80-7184-786-0.

ŽIAKOVÁ, K. et al., 2009. *Ošetrovatelstvo – Teória a vedecký výskum*. Bratislava: Osveta. ISBN 978-80-8063-304-2.

http://www.ouh.nhs.uk/researchers/planning/is-it-research/documents/medical-statistics-online-help.pdf

RNDr. Eva Reiterová, Ph.D.

Statistika pro nelékařské zdravotnické obory

Určeno pro studenty

Výkonný redaktor Mgr. Šárka Vévodová, Ph.D.
Odpovědná redaktorka Mgr. Jana Kreiselová
Technická redakce Mgr. Šárka Rýznarová
Zpracování obálky Ivana Perůtková

Vydala Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.vydavatelstvi.upol.cz
www.e-shop.upol.cz
vup@upol.cz

1. vydání

Olomouc 2016

Edice – Skripta

ISBN 978-80-244-5082-7 (online: PDF)
DOI: 10.5507/fzv.16.24450827

VUP 2016/0392