Masarykova univerzita
Ekonomicko–správní fakulta

Statistika II

distanční studijní opora

Marie Budíková

Brno 2006

Tento projekt byl realizován za finanční podpory Evropské unie v rámci programu SOCRATES — Grundtvig.

Za obsah produktu odpovídá výlučně autor, produkt nereprezentuje názory Evropské komise a Evropská komise neodpovídá za použití informací, jež jsou obsahem produktu.

This project was realized with financial support of European Union in terms of program SOCRATES — Grundtvig.

Author is exclusively responsible for content of product, product does not represent opinions of European Union and European Commission is not responsible for any uses of informations, which are content of product

Recenzoval:

Statistika II
Vydala Masarykova univerzita
Ekonomicko–správní fakulta

Vydání první
Brno, 2006

ISBN

Identifikace modulu

Znak

- CN_KMSTII, KMSTII

Název

- Statistika II

Určení

- Celoživotní magisterské studium, kombinované magisterské studium

Autor

- RNDr. Marie Budíková, Dr.

Garant

- doc. RNDr. Jaroslav Michálek, CSc.

Cíl

Vymezení cíle

Cílem kurzu je naučit studenty základní techniky matematické statistiky pro analýzu reálných ekonomických dat a zároveň je připravit pro studium dalších statistických metod používaných v ekonomii.

Dovednosti a znalosti získané po studiu textů

Studenti se seznámí s podstatou řady užitečných statistických metod a naučí se tyto metody aplikovat na reálná data. Přitom budou využívat softwarový produkt STATISTICA. Získají znalosti, které jim umožní uspořádat experiment tak, aby bylo možno statisticky korektně vyhodnotit jeho výsledky, naučí se posuzovat vlastnosti dat pomocí diagnostických grafů, zvládnou řešení úloh o jednom, dvou a více nezávislých náhodných výběrech z normálních rozložení a dozví se, jak analyzovat závislost dvou veličin.

Časový plán

Rozsah předmětu je dán akreditací a je rozdělen do pěti tutoriálů po čtyřech hodinách.

V 1. tutoriálu jsou zařazena témata

Základní pojmy matematické statistiky
Náhodný výběr a statistiky odvozené z náhodného výběru
Bodové a intervalové odhady parametrů a parametrických funkcí
Úvod do testování hypotéz
Uspořádání pokusů
Jednoduché pozorování
Dvojné pozorování
Mnohonásobné pozorování
Základní informace o statistickém programovém systému STATISTICA

Ve 2. tutoriálu jsou zařazena témata

Diagnostické grafy a testy normality dat
Krabicový diagram, normální pravděpodobnostní graf, kvantil–kvantilový graf, histogram, dvourozměrný tečkový diagram
Kolmogorovův-Smirnovův test normality
Shapirův-Wilksův test normality
V systému STATISTICA je ukázáno
jak konstruovat uvedené typy diagnostických grafů
jak provádět uvedené testy normality

Ve 3. tutoriálu jsou zařazena témata

Úlohy o jednom náhodném výběru z normálního rozložení
Rozložení statistik odvozených z výběrového průměru a výběrového rozptylu
Intervaly spolehlivosti pro střední hodnotu a rozptyl a testování hypotéz o těchto parametrech (jednovýběrový t-test, test o rozptylu)
Náhodný výběr z dvourozměrného normálního rozložení, párový t-test
Úlohy o dvou nezávislých náhodných výběrech z normálních rozložení
Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů
Intervaly spolehlivosti pro rozdíl středních hodnot a podíl rozptylů a testování hypotéz o těchto parametrických funkcích (dvouvýběrový t-test, F-test)
V systému STATISTICA je ukázáno
jak získat meze intervalů spolehlivosti pro parametry normálního rozložení
jak provádět testy hypotéz o parametrech normálního rozložení

Ve 4. tutoriálu jsou zařazena témata

Analýza rozptylu jednoduchého třídění
Testování hypotézy o shodě středních hodnot aspoň tří nezávislých náhodných výběrů z normálních rozložení
Testy shody rozptylů (Levenův test, Bartlettův test)
Metody mnohonásobného porovnávání (Tukeyova metoda, Scheffého metoda)
Význam předpokladů v analýze rozptylu
Pořadové testy o mediánech
Pojem pořadí
Jednovýběrové a párové pořadové testy
Dvouvýběrové pořadové testy
Neparametrické obdoby analýzy rozptylu jednoduchého třídění
V systému STATISTICA je ukázáno
jak získat tabulku analýzy rozptylu a interpretovat ji
jak provádět metody mnohonásobného porovnávání
jak provádět neparametrické testy o mediánech

V 5. tutoriálu je zařazeno téma

Analýza závislosti dvou náhodných veličin
Testování nezávislosti veličin nominálního typu (chí-kvadrát test, podmínky dobré aproximace, Cramérův koeficient, Fisherův přesný test ve čtyřpolní tabulce, podíl šancí)

Testování nezávislosti veličin ordinálního typu (Spearmanův koeficient pořadové korelace)
Testování nezávislosti veličin intervalového či poměrového typu (výběrový koeficient korelace, jeho vlastnosti, testování hypotézy o nezávislosti veličin s dvourozměrným normálním rozložením)
V systému STATISTICA je ukázáno
jak získat kontingenční tabulku, vypočítat Cramérův koeficient, ověřit podmínky dobré aproximace, provést chí-kvadrát test nezávislosti
jak pro čtyřpolní tabulku provést Fisherův přesný test
jak vypočítat Spearmanův koeficient pořadové korelace a s jeho pomocí testovat hypotézu o nezávislosti
jak orientačně ověřit dvourozměrnou normalitu dat, jak vypočítat výběrový koeficient korelace a jak testovat hypotézu o nezávislosti

Časová náročnost

- prezenční část 20 hodin
- samostudium 115 hodin
- POT 9 hodin

Celkový studijní čas

- 144 hodin

Harmonogram

Září
- 4. týden první tutoriál, seznámení s kursem a požadavky, zadání POT – 4 hodiny

Říjen
- 1. a 2. týden samostudium a příprava na tutoriál – 20 hodin
- 3. týden druhý tutoriál – 4 hodiny
- 4. týden samostudium a práce s PC – 10 hodin

Listopad
- 1. týden samostudium a příprava na tutoriál – 10 hodin
- 2. týden třetí tutoriál – 4 hodiny
- 3. týden samostudium a příprava na tutoriál – 10 hodin
  řešení prvních dvou úkolů z POTu – 3 hodiny
- 4. týden čtvrtý tutoriál – 4 hodiny

Prosinec
- 1. týden samostudium a příprava na tutoriál– 10 hodin
  řešení třetího a čtvrtého úkolu z POTu – 3 hodiny
- 2. týden pátý tutoriál – 4 hodiny
  řešení pátého, šestého a sedmého úkolu z POTu – 3 hodiny
- 3. a 4. týden samostudium a odeslání POTu tutorovi – 15 hodin

Leden
- Příprava na zkoušku a případné opravy POTu – 40 hodin

Způsob studia

Studijní pomůcky

Základní literatura

- Budíková, M.: *Statistika II*. Distanční studijní opora
- Hanousek, J. a Charamza, P.: *Moderní metody zpracování dat – matematická statistika pro každého*. EDUCA 1992. ISBN 80-85623-31-5
- Hindls, R., Hronová, S. a Seger, J.: *Statistika pro ekonomy*. Professional Publishing 2002. ISBN 80-86419-26-6
- Osecký P.: *Statistické vzorce a věty*. ESF MU, Brno 1999 ISBN 80-210-2057-1

Doplňková literatura

- Budíková, M., Mikoláš, Š., Osecký, P.: *Teorie pravděpodobnosti a matematická statistika – sbírka příkladů*. Brno, 1998. ISBN 80-210-1832-1
- Hendl, J.: *Přehled statistických metod zpracování dat. Analýza a metaanalýza dat*. 1. vydání, 2004. ISBN 80-7178-820-1
- Wonnancot, T. H. a Wonnancot, R. J.: *Statistika pro obchod a hospodářství*. Praha, Victoria Publishing 1993. ISBN 80-85605-09-0

Vybavení

- PC
- CD-ROM se systémem STATISTICA

Návod práce se studijním textem

Text je rozdělen do osmi kapitol a dvou příloh. První příloha obsahuje statistické tabulky, druhá zadání POT. Studium textu předpokládá znalost základních pojmů popisné statistiky a počtu pravděpodobnosti v rozsahu distanční studijní opory Statistika autorky Marie Budíkové a také schopnost pracovat se systémem STATISTICA.

V úvodní části každé kapitoly je vymezen její cíl a je uveden přibližný čas, který budete potřebovat ke zvládnutí příslušného tématu. V každé kapitole je zařazeno několik vzorových příkladů s podrobným návodem, jak je vyřešit pomocí systému STATISTICA. Kapitoly jsou uzavřeny stručným shrnutím probrané látky a poté následují kontrolní otázky, autokorekční test a neřešené příklady s výsledky. Kromě příkladů, k jejichž řešení potřebujete systém STATISTICA, jsou v učebním textu též příklady teoretického charakteru nebo výpočetně jednoduché příklady, u nichž vystačíte s kapesním kalkulátorem. Rovněž tyto příklady jsou důležité, protože při jejich podrobném řešení dobře pochopíte podstatu určité statistické metody.

Tabulková příloha obsahuje vybrané statistické tabulky, a to jak pro parametrické, tak pro neparametrické metody. Druhá příloha je věnována zadání samostatné práce – POT. Hlavním cílem POTu je naučit vás nejen používat statistické metody při zpracování rozsáhlého datového souboru, ale také správně interpretovat výsledky těchto metod a prezentovat je v přehledné a srozumitelné podobě.

Obsah

Stručný obsah

Kapitola 1
Základní pojmy matematické statistiky

Zavádí pojem náhodného výběru z jednorozměrného a vícerozměrného rozložení a pojem statistiky jako transformace náhodného výběru. Uvádí příklady důležitých statistik. Ukazuje, jak na základě znalosti náhodného výběru bodově či intervalově odhadnout parametry rozložení, z něhož náhodný výběr pochází. Zabývá se otázkou, jak na dané hladině významnosti testovat hypotézy o parametrech rozložení, z něhož daný náhodný výběr pochází. Popisuje tři způsoby, jak testovat nulovou hypotézu proti alternaivní hypotéze.

Kapitola 2
Uspořádání pokusů

Vysvětluje rozdíl mezi jednoduchým, dvojným a mnohonásobným pozorováním, přičemž v rámci dvojného pozorování rozlišuje dvouvýběrové a párové porovnávání a v rámci mnohonásobného pozorování rozlišuje mnohovýběrové a blokové porovnávání.

Kapitola 3
Diagnostické grafy a testy normality dat

Zabývá se způsobem konstrukce krabicového diagramu, normálního pravděpodobnostního grafu, kvantil–kvantilového grafu, histogramu a dvourozměrného tečkového diagramu. Popisuje Kolmogorovův-Smirnovův test normality a Shapirův-Wilksův test normality a ukazuje, jak uvedené grafy zkonstruovat v systému STATISTICA a jak provést testy normality v tomto systému.

Kapitola 4
Úlohy o jednom náhodném výběru z normálního rozložení

Věnuje se vlastnostem statistik odvozených z náhodného výběru z normálního rozložení. Ukazuje, jak bodově a intervalově odhadnout střední hodnotu a rozptyl normálního rozložení, z něhož daný náhodný výběr pochází a jak testovat hypotézy o těchto parametrech. Popisuje způsob, jak pomocí systému STATISTICA získat meze intervalů spolehlivosti pro parametry normálního rozložení a provádět testy hypotéz o těchto parametrech.

Kapitola 5
Úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Věnuje se vlastnostem statistik odvozených ze dvou nezávislých náhodných výběrů z normálních rozložení. Ukazuje, jak bodově a intervalově odhadnout rozdíl středních hodnot a podíl rozptylů dvou normálních rozložení, z nichž dané náhodné výběry pocházejí a jak testovat hypotézy o těchto parametrických funkcích. Popisuje způsob, jak pomocí systému STATISTICA získat meze intervalů spolehlivosti pro rozdíl středních hodnot a podíl rozptylů a provádět testy hypotéz o těchto parametrických funkcích.

Kapitola 6
Analýza rozptylu jednoduchého třídění

Zabývá se situací, kdy se hodnotí vliv faktoru o aspoň třech úrovních na variabilitu hodnot sledované náhodné veličiny. Popisuje dvě metody mnohonásobného porovnávání, které umožní identifikovat dvojice náhodných výběrů, které se významně liší střední hodnotou. Věnuje pozornost významu jednotlivých předpokladů v analýze rozptylu a ukazuje, jak tuto analýzu provést v systému STATISTICA.

Kapitola 7

Pořadové testy o mediánech

Popisuje testy hypotéz o mediánu jednoho spojitého rozložení a ukazuje, jak hodnotit shodu dvou nezávislých náhodných výběrů ze spojitých rozložení lišících se posunem či hodnotit shodu aspoň tří nezávislých náhodných výběrů ze spojitých rozložení lišících se posunem a identifikovat dvojice významně odlišných náhodných výběrů. Popisuje způsob provedení pořadových testů o mediánech v systému STATISTICA.

Kapitola 8

Analýza závislosti dvou náhodných veličin

Vysvětluje, jak provádět test nezávislosti v kontingenční tabulce a jak hodnotit intenzitu závislosti dvou náhodných veličin nominálního typu pomocí Cramérova koeficientu. Popisuje rovněž Fisherův přesný test ve čtyřpolní kontingenční tabulce. Věnuje se testování pořadové nezávislosti dvou náhodných veličin ordinálního typu pomocí Spearmanova koeficientu pořadové korelace a testování hypotézy o nezávislosti dvou náhodných veličin intervalového či poměrového typu, které se řídí dvourozměrným normálním rozložením. Ukazuje použití systému STATISTICA při analýze závislosti.

Úplný obsah

Obsah

- Motivace
- Náhodný výběr a statistiky odvozené z náhodného výběru
- Bodové a intervalové odhady parametrů a parametrických funkcí
- Úvod do testování hypotéz

1. Základní pojmy matematické statistiky

Cíl kapitoly

Po prostudování této kapitoly budete

- rozumět pojmu „náhodný výběr“
- znát vlastnosti důležitých statistik odvozených z náhodného výběru
- znát vlastnosti bodových a intervalových odhadů parametrů a parametrických funkcí
- umět formulovat nulovou a alternativní hypotézu o parametru či parametrické funkci
- znát tři způsoby, jak testovat nulovou hypotézu proti alternativní hypotéze na dané hladině významnosti

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 18 hodin studia.

1.1 Motivace

Při aplikaci metod popisné statistiky dospíváme pomocí zjištěných dat k závěrům, které se týkají pouze výběrového souboru. Naproti tomu matematická statistika nám umožňuje na základě znalosti náhodného výběru a statistik z něj odvozených (tj. např. výběrového průměru, výběrového rozptylu, výběrového koeficientu korelace, hodnoty výběrové distribuční funkce apod.) učinit závěry o parametrech nebo tvaru rozložení, z něhož daný náhodný výběr pochází. Často se jedná o bodové či intervalové odhady parametrů a parametrických funkcí a testování hypotéz o nich.

1.2 Náhodný výběr a statistiky odvozené z náhodného výběru

1.2.1 Pojem náhodného výběru

Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$. Řekneme, že $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozložení $L(\vartheta)$. (Číselné realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ uspořádané do sloupcového vektoru představují datový soubor.)

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1, Y_1), \ldots, (X_n, Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1, y_1), \ldots, (x_n, y_n)$ náhodného výběru $(X_1, Y_1), \ldots, (X_n, Y_n)$ uspořádané do matice typu $n \times 2$ představují dvourozměrný datový soubor.)

Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\vartheta)$.

1.2.2 Pojem statistiky, příklady důležitých statistik

Libovolná funkce $T = T(X_1, \dots, X_n)$ náhodného výběru $X_1, \dots, X_n$ (resp. p-rozměrného náhodného výběru) se nazývá statistika.

a) Nechť $X_1, \dots, X_n$ je náhodný výběr, $n \geq 2$.

Statistika $M = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ se nazývá výběrový průměr,

Statistika $S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - M)^2$ výběrový rozptyl,

Statistika $S = \sqrt{S^2}$ výběrová směrodatná odchylka.

Pro libovolné, ale pevně zvolené reálné číslo x je statistikou též hodnota výběrové distribuční funkce

$$F_n(x) = \frac{1}{n}\,\mathrm{card}\{i;\ X_i \leq x\}.$$

b) Nechť $X_{11}, \dots, X_{1n_1}, \dots, X_{p1}, \dots, X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \dots, n_p \geq 2$. Celkový rozsah je $n = \sum\limits_{j=1}^{p} n_j$. Označme $M_1, \dots, M_p$ výběrové průměry a $S_1^2, \dots, S_p^2$ výběrové rozptyly jednotlivých výběrů. Nechť $c_1, \dots, c_p$ jsou reálné konstanty, aspoň jedna nenulová.

Statistika $\sum\limits_{j=1}^{p} c_j M_j$ se nazývá lineární kombinace výběrových průměrů.

Statistika $S_*^2 = \dfrac{\sum\limits_{j=1}^{p}(n_j - 1)S_j^2}{n-p}$ se nazývá vážený průměr výběrových rozptylů.

c) Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení. Označme $M_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$.

Statistika $S_{12} = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ je výběrová kovariance,

statistika

$$R_{12} = \begin{cases} \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\frac{X_i - M_1}{S_1}\cdot\frac{Y_i - M_2}{S_2} & \text{pro } S_1 S_2 \neq 0, \\ 0 & \text{jinak.} \end{cases}$$

výběrový koeficient korelace.

(Číselné realizace m, s^2, s, s_{12}, r_{12} statistik M, S_2, S, S_{12}, R_{12} odpovídají číselným charakteristikám znaků v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n-1}$, nikoli $\frac{1}{n}$, jak tomu bylo v popisné statistice.)

1.3 Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru $X_1,\dots,X_n$ z rozložení $L(\vartheta)$, které závisí na parametru ϑ. Množinu všech přípustných hodnot tohoto parametru označíme Ξ. Parametr ϑ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\vartheta)$).

Bodovým odhadem parametrické funkce $h(\vartheta)$ je statistika $T_n = T(X_1,\dots,X_n)$, která nabývá hodnot blízkých $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

Intervalovým odhadem parametrické funkce $h(\vartheta)$ rozumíme interval (D,H), jehož meze jsou statistiky $D = D(X_1,\dots,X_n)$, $H = H(X_1,\dots,X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv.

1.3.1 Typy bodových odhadů

Nechť $X_1,\dots,X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, $T, T_1, T_2, \dots$ jsou statistiky.

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\vartheta)$, jestliže
$$\forall\vartheta \in \Xi : E(T) = h(\vartheta).$$
(Význam nestrannosti spočívá v tom, že odhad T nesmí parametrickou funkci $h(\vartheta)$ systematicky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li T_1, T_2 nestranné odhady téže parametrické funkce $h(\vartheta)$, pak řekneme, že T_1 je lepší odhad než T_2, jestliže
$$\forall\vartheta \in \Xi : D(T_1) < D(T_2).$$

c) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost asymptoticky nestranných odhadů parametrické funkce $h(\vartheta)$, jestliže
$$\forall\vartheta \in \Xi : \lim_{n\to\infty} E(T_n) = h(\vartheta).$$
(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu.)

d) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost konzistentních odhadů parametrické funkce $h(\vartheta)$, jestliže
$$\forall\vartheta \in \Xi\ \forall\varepsilon > 0 : \lim_{n\to\infty} P\big(|T_n - h(\vartheta)| > \varepsilon\big) = 0.$$
(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat daleko od parametrické funkce $h(\vartheta)$.)

Lze dokázat, že z nestrannosti odhadu vyplývá jeho asymptotická nestrannost a z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů odhadu konverguje k nule.

1.3.2 Vlastnosti důležitých statistik

a) Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané $x \in \mathbb{R}$ $F_n(x)$ hodnotu výběrové distribuční funkce.

   Pak M_n je nestranným odhadem μ (tj. $E(M_n) = \mu$) s rozptylem $D(M) = \dfrac{\sigma^2}{n}$, S_n^2 je nestranným odhadem σ^2 (tj. $E(S_n^2) = \sigma^2$), ať jsou hodnoty parametrů μ, σ^2 jakékoli. Dále platí, že pro libovolné, ale pevně dané $x \in \mathbb{R}$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$ (tj. $E(F_n(x)) = \Phi(x)$) s rozptylem $D(F_n(x)) = \Phi(x)(1-\Phi(x))/n$, ať je hodnota distribuční funkce $\Phi(x)$ jakákoliv.

   Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů μ. $\left\{S_n^2\right\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů σ^2 . Pro libovolné, ale pevně dané $x \in \mathbb{R}$ je $\{F_n(x)\}_{n=1}^{\infty}$ posloupnost konzistentních odhadů $\Phi(x)$.

b) Nechť $X_{11}, \dots, X_{1n_1}, \dots, X_{p1}, \dots, X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \dots, n_p \geq 2$ z rozložení se středními hodnotami $\mu_1, \dots, \mu_p$ a rozptylem σ^2. Celkový rozsah je $n = \sum\limits_{j=1}^{p} n_j$. Nechť $c_1, \dots, c_p$ jsou reálné konstanty, aspoň jedna nenulová. Pak lineární kombinace výběrových průměrů $\sum\limits_{j=1}^{p} c_j M_j$ je nestranným odhadem lineární kombinace středních hodnot $\sum\limits_{j=1}^{p} c_j \mu_j$, ať jsou střední hodnoty $\mu_1, \dots, \mu_p$ jakékoli a vážený průměr výběrových rozptylů $S_*^2 = \dfrac{\sum\limits_{j=1}^{p} (n_j - 1) S_j^2}{n-p}$ je nestranným odhadem rozptylu σ^2, ať je rozptyl σ^2 jakýkoliv.

c) Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Pak výběrová kovariance S_{12} je nestranným odhadem kovariance σ_{12}, ať je kovariance σ_{12} jakákoli, avšak $E(R_{12})$ je rovno ρ pouze přibližně (shoda je vyhovující pro $n > 30$), ať je korelační koeficient ρ jakýkoli.

1.3.3 Pojem intervalu spolehlivosti

Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, $\alpha \in (0,1)$, $D = D(X_1, \dots, X_n)$, $H = H(X_1, \dots, X_n)$ jsou statistiky.

a) Interval (D, H) se nazývá $100(1-\alpha)\%$ (oboustranný) interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi : P(D < h(\vartheta) < H) \geq 1 - \alpha$.

b) Interval (D, ∞) se nazývá $100(1-\alpha)\%$ levostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi : P(D < h(\vartheta)) \geq 1 - \alpha$.

c) Interval $(-\infty, H)$ se nazývá $100(1-\alpha)\%$ pravostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi : P(h(\vartheta) < H) \geq 1-\alpha$.

d) Číslo α se nazývá riziko (zpravidla $\alpha = 0{,}05$, méně často 0,1 či 0,01), číslo $1-\alpha$ se nazývá spolehlivost.

1.3.4 Postup při konstrukci intervalu spolehlivosti

a) Vyjdeme ze statistiky V, která je nestranným bodovým odhadem parametrické funkce $h(\vartheta)$.

b) Najdeme tzv. pivotovou statistiku W, která vznikne transformací statistiky V, je monotónní funkcí $h(\vartheta)$ a přitom její rozložení je známé a na $h(\vartheta)$ nezávisí. Pomocí známého rozložení pivotové statistiky W najdeme kvantily $w_{\alpha/2}$, $w_{1-\alpha/2}$ tak, že platí:

$$\forall \vartheta \in \Xi : P\big(w_{\alpha/2} < W < w_{1-\alpha/2}\big) \geq 1-\alpha.$$

c) Nerovnost $w_{\alpha/2} < W < w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost

$$D < h(\vartheta) < H.$$

d) Statistiky D, H nahradíme jejich číselnými realizacemi d, h a získáme tak $100(1-\alpha)\%$ empirický interval spolehlivosti, o němž prohlásíme, že pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1-\alpha$. (Tvrzení, že (d, h) pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1-\alpha$ je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\vartheta)$ a pomocí každé této realizace sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\vartheta)$, pak podíl počtu těch intervalů, které pokrývají $h(\vartheta)$ k počtu všech sestrojených intervalů bude přibližně $1-\alpha$.)

1.3.5 Příklad

Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde $n \geq 2$ a rozptyl σ^2 známe. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro neznámou střední hodnotu μ.

Řešení:

V tomto případě parametrická funkce $h(\vartheta) = \mu$. Nestranným odhadem střední hodnoty je výběrový průměr (viz 1.3.2 (a)) $M = \frac{1}{n}\sum_{i=1}^{n} X_i$. Protože M je lineární kombinací normálně rozložených náhodných veličin, bude mít také normální rozložení se střední hodnotou $E(M) = \mu$ a rozptylem $D(M) = \dfrac{\sigma^2}{n}$. Pivotovou statistikou W bude standardizovaná náhodná veličina $U = \dfrac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Kvantil $w_{\alpha/2} = u_{\alpha/2} = -u_{1-\alpha/2}$, $w_{1-\alpha/2} = u_{1-\alpha/2}$.

$$\forall \vartheta \in \Xi : 1-\alpha \leq P(-u_{1-\alpha/2} < U < u_{1-\alpha/2}) =$$
$$P\left(-u_{1-\alpha/2} < \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\alpha/2}\right) = P\left(M - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} < \mu < M + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right).$$

Meze $100(1-\alpha)\%$ intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 tedy jsou:

$$D = M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \quad H = M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}.$$

Při konstrukci jednostranných intervalů spolehlivosti se riziko nepůlí, tedy $100(1-\alpha)\%$ levostranný interval spolehlivosti pro μ je $\left(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty\right)$ a pravostranný je $\left(-\infty, M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right)$.

Dosadíme-li do vzorců pro dolní a horní mez číselnou realizaci m výběrového průměru M, dostaneme $100(1-\alpha)\%$ empirický interval spolehlivosti.

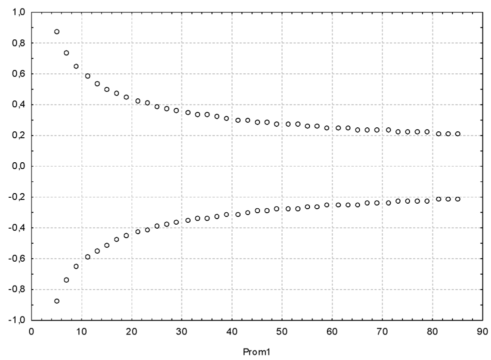
1.3.6 Šířka intervalu spolehlivosti

Nechť (d, h) je $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\vartheta)$ zkonstruovaný pomocí číselných realizací $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\vartheta)$.

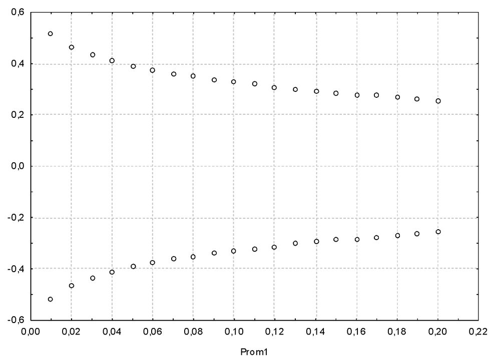
a) Při konstantním riziku klesá šířka *h–d* s rostoucím rozsahem náhodného výběru.

b) Při konstantním rozsahu náhodného výběru klesá šířka *h–d* s rostoucím rizikem.

Závislost dolní a horní meze na rozsahu výběru (při konst. riziku)



Závislost dolní a horní meze na riziku (při konst. rozsahu výběru)



1.3.7 Příklad

Využití bodu 1.3.6 (a) při stanovení minimálního rozsahu výběru z normálního rozložení: Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Jaký musí být minimální rozsah výběru n, aby šířka $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu μ nepřesáhla číslo Δ?

Řešení:

Požadujeme, aby $\Delta \geq h - d = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} - \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right) = \frac{2\sigma}{\sqrt{n}} u_{1-\alpha/2}$.

Z této podmínky dostaneme, že $n \geq \dfrac{4\sigma^2 u_{1-\alpha/2}^2}{\Delta^2}$. Za rozsah výběru zvolíme nejmenší přirozené číslo vyhovující této podmínce.

1.4 Úvod do testování hypotéz

Nulovou hypotézou rozumíme nějaké tvrzení o parametrech nebo typu rozložení, z něhož pochází náhodný výběr. Nulová hypotéza vyjadřuje nějaký teoretický předpoklad, často skeptického rázu a uživatel ji musí stanovit předem, bez přihlédnutí k datovému souboru. Proti nulové hypotéze stavíme alternativní hypotézu, která říká, co platí, když neplatí nulová hypotéza. Např. nulová hypotéza tvrdí, že střední hodnota hmotnosti balíčků cukru balených na automatické lince se nezměnila seřízením automatu, zatímco alternativní hypotéza tvrdí opak. Postup, který je založen na daném náhodném výběru a s jehož pomocí rozhodneme o zamítnutí či nezamítnutí nulové hypotézy, se nazývá testování hypotéz.

1.4.1 Nulová a alternativní hypotéza

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, kde parametr $\vartheta \in \Xi$ neznáme. Nechť $h(\vartheta)$ je parametrická funkce a c daná reálná konstanta.

a) Oboustranná alternativa: Tvrzení H_0: $h(\vartheta) = c$ se nazývá jednoduchá nulová hypotéza. Proti nulové hypotéze postavíme složenou alternativní hypotézu H_1: $h(\vartheta) \neq c$.

b) Levostranná alternativa: Tvrzení H_0: $h(\vartheta) \geq c$ se nazývá složená pravostranná nulová hypotéza. Proti jednoduché nebo složené pravostranné nulové hypotéze postavíme složenou levostrannou alternativní hypotézu H_1: $h(\vartheta) < c$.

c) Pravostranná alternativa: Tvrzení H_0: $h(\vartheta) \leq c$ se nazývá složená levostranná nulová hypotéza. Proti jednoduché nebo složené levostranné nulové hypotéze postavíme složenou pravostrannou alternativní hypotézu H_1: $h(\vartheta) > c$.

Testováním H_0 proti H_1 rozumíme rozhodovací postup založený na náhodném výběru $X_1, \ldots, X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

1.4.2 Chyba 1. a 2. druhu

Při testování H_0 proti H_1 se můžeme dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že H_0 zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že H_0 nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

skutečnost	rozhodnutí	
	H_0 nezamítáme	H_0 zamítáme
H_0 platí	správné rozhodnutí	chyba 1. druhu
H_0 neplatí	chyba 2. druhu	správné rozhodnutí

Pravděpodobnost chyby 1. druhu se značí α a nazývá se hladina významnosti testu (většinou bývá $\alpha = 0{,}05$, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí β. Číslo $1-\beta$ se nazývá síla testu a vyjadřuje pravděpodobnost, s jakou test vypoví, že H_0 neplatí.

1.4.3 Testování pomocí kritického oboru

Najdeme statistiku $T_0 = T_0(X_1, \ldots, X_n)$, kterou nazveme testovým kritériem (testovou statistikou). Množina všech hodnot, jichž může testové kritérium nabýt, se rozpadá na obor nezamítnutí nulové hypotézy (značí se V) a obor zamítnutí nulové hypotézy (značí se W a nazývá se též kritický obor). Tyto dva obory jsou odděleny kritickými hodnotami (pro danou hladinu významnosti α je lze najít ve statistických tabulkách).

Jestliže číselná realizace t_0 testového kritéria T_0 padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti α a znamená to skutečné vyvrácení testované hypotézy. Jestliže t_0 padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Pravděpodobnosti chyb 1. a 2. druhu nyní zapíšeme takto:

$$P(T_0 \in W / H_0 \text{ platí}) = \alpha, \qquad P(T_0 \in V / H_1 \text{ platí}) = \beta.$$

Stanovení kritického oboru pro danou hladinu významnosti α:
Označme $t_{\min}$ (resp. $t_{\max}$) nejmenší (resp. největší) hodnotu testového kritéria. Kritický obor v případě oboustranné alternativy má tvar

$$W = (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max}),$$

kde $K_{\alpha/2}(T)$ a $K_{1-\alpha/2}(T)$ jsou kvantily rozložení, jímž se řídí testové kritérium T_0, je-li nulová hypotéza pravdivá.

Kritický obor v případě levostranné alternativy má tvar:

$$W = (t_{\min}, K_{\alpha}(T)\rangle.$$

Kritický obor v případě pravostranné alternativy má tvar:

$$W = \langle K_{1-\alpha}(T), t_{\max}).$$

Doporučuje se dodržovat následující postup:

- Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
- Zvolíme hladinu významnosti α. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.
- Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
- Stanovíme kritický obor.

– Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti α. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α.

1.4.4 Testování pomocí intervalu spolehlivosti

Sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. Pokryje-li tento interval hodnotu c, pak H_0 nezamítáme na hladině významnosti α, v opačném případě H_0 zamítáme na hladině významnosti α.

- Pro test H_0 proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.
- Pro test H_0 proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.
- Pro test H_0 proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.

1.4.5 Testování pomocí *p*-hodnoty

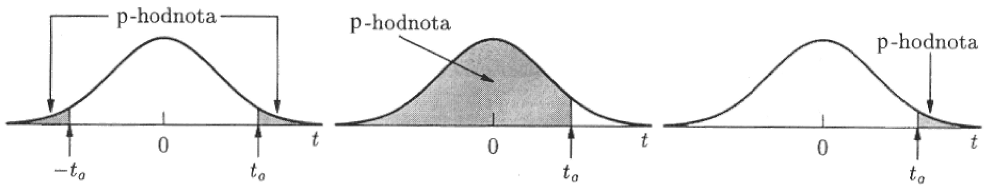
p-hodnota udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je-li p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.

Způsob výpočtu p-hodnoty:

- Pro oboustrannou alternativu $p = 2\min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$.
- Pro levostrannou alternativu $p = P(T_0 \leq t_0)$.
- Pro pravostrannou alternativu $p = P(T_0 \geq t_0)$.

p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ podporují H_0, je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium T_0, je-li H_0 pravdivá. Vzhledem k tomu, že v běžných statistických tabulkách jsou uvedeny pouze hodnoty distribuční funkce standardizovaného normálního rozložení, bez použití speciálního software jsme schopni vypočítat p-hodnotu pouze pro test hypotézy o střední hodnotě normálního rozložení při známém rozptylu.

Ilustrace významu p-hodnoty pro test nulové hypotézy proti oboustranné, levostranné a pravostranné alternativě:



(Zvonovitá křivka reprezentuje hustotu rozložení, kterým se řídí testové kritérium, je-li nulová hypotéza pravdivá.)

1.4.6 Příklad

$10\times$ nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly:

$$2 \quad 1{,}8 \quad 2{,}1 \quad 2{,}4 \quad 1{,}9 \quad 2{,}1 \quad 2 \quad 1{,}8 \quad 2{,}3 \quad 2{,}2.$$

Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \ldots, X_{10}$ z rozložení $N(\mu, 0{,}04)$. Nějaká teorie tvrdí, že $\mu = 1{,}95$. Proti nulové hypotéze H_0: $\mu = 1{,}95$ postavíme oboustrannou alternativu H_1: $\mu \neq 1,95$. Na hladině významnosti 0,05 testujte H_0 proti H_1 všemi třemi popsanými způsoby.

Řešení:

$$m = \frac{1}{10}(2 + \cdots + 2{,}2) = 2{,}06, \quad \sigma^2 = 0{,}04, \quad n = 10, \quad \alpha = 0{,}05, \quad c = 1{,}95.$$

a) Test provedeme pomocí kritického oboru.
Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku $U = \dfrac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ (viz 1.3.5). Testové kritérium tedy bude $T_0 = \dfrac{M - c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení $N(0,1)$, pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria:

$$t_0 = \frac{2{,}06 - 1{,}95}{\frac{0{,}2}{\sqrt{10}}} = 1{,}74.$$

Stanovíme kritický obor:

$$\begin{aligned} W &= (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max}) = (-\infty, u_{\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = \\ &= (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{0{,}975}\rangle \cup \langle u_{0{,}975}, \infty) = \\ &= (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty) \end{aligned}$$

Protože $1{,}74 \notin W$, H_0 nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 jsou (viz 1.3.5):

$$(d, h) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right).$$

V našem případě dostáváme:

$$d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} \cdot u_{0{,}975} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} \cdot 1{,}96 = 1{,}936, \quad h = 2{,}184.$$

Protože $1{,}95 \in (1{,}936; 2{,}184)$, H_0 nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.
Protože proti nulové hypotéze stavíme oboustrannou alternativu, použijeme vzorec

$$
\begin{aligned}
p &= 2\min\{P(T_0 \le t_0), P(T_0 \ge t_0)\} = 2\min\{P(T_0 \le 1{,}74), P(T_0 \ge 1{,}74)\} = \\
&= 2\min\{\Phi(1{,}74), 1-\Phi(1{,}74)\} = 2\min\{0{,}95907, 1-0{,}95907\} = \\
&= 0{,}08186.
\end{aligned}
$$

Jelikož $0{,}08186 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Shrnutí kapitoly

Ústředním pojmem matematické statistiky je pojem *náhodného výběru*, a to jednorozměrného i vícerozměrného. Transformací jednoho nebo více náhodných výběrů vzniká náhodná veličina zvaná *(výběrová) statistika*. K nejdůležitějším statistikám patří *výběrový průměr*, *výběrový rozptyl*, *výběrová směrodatná odchylka*, *hodnoty výběrové distribuční funkce*, *výběrová kovariance*, *výběrový koeficient korelace*.

Jelikož statistika je náhodná veličina, má smysl počítat její střední hodnotu a rozptyl. Ukázali jsme si vlastnosti střední hodnoty a rozptylu výběrového průměru či hodnoty výběrové distribuční funkce a střední hodnoty výběrového rozptylu, výběrové kovariance a výběrového koeficientu korelace.

Na základě znalosti náhodného výběru aproximujeme neznámou hodnotu parametru či parametrické funkce *bodovým odhadem parametrické funkce*. Zpravidla požadujeme, aby tento odhad měl jisté žádoucí vlastnosti. K těm patří *nestrannost*, resp. *asymptotická nestrannost* či *konzistence*, pokud pracujeme s posloupností bodových odhadů téže parametrické funkce. Bodové odhady však mají jednu značnou nevýhodu – nevíme, s jakou pravděpodobností odhadují hodnotu neznámé parametrické funkce. Tuto nevýhodu odstraňují *intervalové odhady parametrické funkce*: jsou to intervaly, jejichž meze jsou statistiky a které s předem danou dostatečně velkou pravděpodobností pokrývají hodnotu neznámé parametrické funkce. Pokud do vzorců pro meze $100(1-\alpha)\%$ intervalu spolehlivosti pro danou parametrickou funkci dosadíme číselné realizace náhodného výběru, dostaneme $100(1-\alpha)\%$ *empirický interval spolehlivosti*.

Tvrzení o parametrech rozložení, z něhož pochází daný náhodný výběr, nazýváme *nulovou hypotézou*. Proti nulové hypotéze stavíme *alternativní hypotézu*, která říká, co platí, když neplatí nulová hypotéza. Při testování nulové hypotézy proti alternativní hypotéze se můžeme dopustit buď *chyby 1. druhu* (nulovou hypotézu zamítneme, ač ve skutečnosti platí) nebo *chyby 2. druhu* (nulovou hypotézu nezamítneme, ač ve skutečnosti neplatí). Pravděpodobnost chyby 1. druhu se značí α a nazývá se hladina významnosti testu.

Klasický přístup k testování hypotéz spočívá v nalezení *vhodného testového kritéria*. Množina hodnot, jichž může testové kritérium nabýt, se rozpadá *na obor nezamítnutí nulové hypotézy* a na *kritický obor*. Tyto dva neslučitelné obory jsou odděleny *kritickými hodnotami*. Pokud se testové kritérium realizuje v kritickém oboru, nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α. Tím jsme ovšem neprokázali její pravdivost, můžeme pouze říci, že naše data nejsou natolik průkazná, abychom mohli nulovou hypotézu zamítnout.

Test nulové hypotézy proti alternativní hypotéze lze též provést pomocí intervalu spolehlivosti.

Máme-li k dispozici statistický software, můžeme vypočítat *p-hodnotu* jako nejmenší možnou hladinu významnosti pro zamítnutí nulové hypotézy.

Kontrolní otázky

1. Vysvětlete pojem „náhodný výběr“ a „statistika“ a uveďte příklady důležitých statistik.
2. K čemu slouží bodový odhad parametrické funkce a jaké typy bodových odhadů znáte?
3. Definujte interval spolehlivosti a popište způsob jeho konstrukce.
4. Jaký vliv na šířku intervalu spolehlivosti má riziko a jaký vliv má rozsah výběru?
5. Co rozumíme pojmem „testování hypotéz“?
6. Popište nulovou a alternativní hypotézu.
7. Vysvětlete rozdíl mezi chybou 1. a 2. druhu.
8. Popište tři způsoby testování hypotéz.

Autokorekční test

1. Která z následujících tvrzení jsou pravdivá?
   a) Náhodným výběrem rozumíme objekty základního souboru, které byly vybrány do výběrového souboru náhodně, např. losováním.
   b) Náhodným výběrem rozumíme posloupnost stochasticky nezávislých a stejně rozložených náhodných veličin či vektorů.
   c) Číselné realizace náhodného výběru uspořádané do vektoru či matice tvoří datový soubor.
2. Která z následujících tvrzení jsou pravdivá?
   a) Výběrový rozptyl je aritmetickým průměrem kvadrátů centrovaných složek náhodného výběru.
   b) Číselné realizace výběrového průměru se mohou výběr od výběru lišit.
   c) V definici váženého průměru výběrových rozptylů hrají roli vah rozsahy jednotlivých náhodných výběrů.
3. Která z následujících tvrzení jsou pravdivá?
   a) Statistika je nestranným odhadem parametrické funkce, pokud její střední hodnota je rovna této parametrické funkci, ať je hodnota parametru jakákoliv.
   b) Posloupnost statistik je posloupností konzistentních odhadů parametrické funkce, pokud s rostoucím rozsahem náhodného výběru roste

pravděpodobnost, že odhady se budou realizovat daleko od parametrické funkce, ať je hodnota parametru jakákoliv.

c) Máme-li dva nestranné odhady téže parametrické funkce, tak za lepší považujeme ten, který má větší rozptyl, ať je hodnota parametru jakákoliv.

4. Která z následujících tvrzení jsou pravdivá?

a) Výběrový průměr je nestranným odhadem střední hodnoty.

b) Výběrová směrodatná odchylka je nestranným odhadem směrodatné odchylky.

c) Výběrový koeficient korelace je nestranným odhadem koeficientu korelace.

5. Která z následujících tvrzení jsou pravdivá?

a) Při konstrukci intervalu spolehlivosti pro parametrickou funkci musíme znát statistiku, která je nestranným bodovým odhadem této parametrické funkce.

b) Empirický $100(1-\alpha)\%$ interval spolehlivosti slouží jako odhad neznámého parametrické funkce v tomto smyslu: pravděpodobnost, že tento interval pokrývá skutečnou hodnotu parametrické funkce, je aspoň $1-\alpha$.

c) Při konstantním riziku α klesá šířka empirického $100(1-\alpha)\%$ intervalu spolehlivosti s rostoucím rozsahem náhodného výběru.

6. Která z následujících tvrzení jsou pravdivá?

a) Kritický obor a obor nezamítnutí nulové hypotézy jsou vždy disjunktní.

b) Pravděpodobnost chyby 2. druhu lze určit na základě znalosti rizika α.

c) Pokud byla nulová hypotéza zamítnuta na hladině významnosti 0,01, byla by zamítnuta i na hladině významnosti 0,05.

Správné odpovědi: 1b), c) 2b) 3a) 4a) 5a), b), c) 6a), c)

Příklady

1. Nezávisle opakovaná laboratorní měření určité konstanty jsou charakterizována náhodným výběrem $X_1,\dots,X_n$ z rozložení se střední hodnotou μ a rozptylem σ^2. Uvažme statistiky

$$M=\frac{1}{n}\sum_{i=1}^{n}X_i,\qquad L=\frac{X_1+X_n}{2}.$$

Dokažte, že M a L jsou nestranné odhady konstanty μ a zjistěte, který z nich je lepší.

Výsledek:

Výpočtem zjistíme, že $E(M)=\mu$, $E(L)=\mu$, tudíž statistiky M a L jsou nestranné odhady konstanty μ. Pro posouzení kvality vypočteme $D(M)=\frac{\sigma^2}{n}$, $D(L)=\frac{\sigma^2}{2}$. Vidíme tedy, že pro $n\geq 3$ je lepším odhadem výběrový průměr M.

2. Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení $N(\mu; 0{,}04)$. Jaký musí být nejmenší rozsah náhodného výběru, aby šířka 95% empirického intervalu spolehlivosti pro neznámou střední hodnotu μ nepřesáhla číslo 0,16?

Výsledek: 25

3. Nechť $X_1, \dots, X_9$ je náhodný výběr z rozložení $N(\mu; 0{,}01)$. Realizace výběrového průměru je $m = 3$. Sestrojte $100(1-\alpha)$% empirický interval spolehlivosti pro neznámou střední hodnotu μ, je-li
   a) $\alpha = 0{,}01$, b) $\alpha = 0{,}05$, c) $\alpha = 0{,}1$.

Výsledek:

ad a) $2{,}914 < \mu < 3{,}086$ s pravděpodobností aspoň 0,99.
ad b) $2{,}935 < \mu < 3{,}065$ s pravděpodobností aspoň 0,95.
ad c) $2{,}945 < \mu < 3{,}055$ s pravděpodobností aspoň 0,90.

Vidíme, že s rostoucím rizikem klesá šířka intervalu spolehlivosti.

4. Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení $N(\mu; 0,01)$. Realizace výběrového průměru je $m = 3$. Sestrojte 95% empirický interval spolehlivosti pro neznámou střední hodnotu μ, je-li
   a) $n = 4$, b) $n = 9$, c) $n = 16$.

Výsledek:

ad a) $2{,}902 < \mu < 3{,}098$ s pravděpodobností aspoň 0,95.
ad b) $2{,}935 < \mu < 3{,}065$ s pravděpodobností aspoň 0,95.
ad c) $2{,}951 < \mu < 3{,}049$ s pravděpodobností aspoň 0,95.

Vidíme, že s rostoucím rozsahem výběru klesá šířka intervalu spolehlivosti.

5. Je známo, že výška hochů ve věku 9,5 až 10 let má normální rozložení s neznámou střední hodnotou μ a známým rozptylem $\sigma^2 = 39{,}112$ cm^2. Dětský lékař náhodně vybral 15 hochů uvedeného věku, změřil je a vypočítal realizaci výběrového průměru $m = 139{,}13$ cm. Podle jeho názoru by výška hochů v tomto věku neměla přesáhnout 142 cm s pravděpodobností aspoň 0,95. Lze tvrzení lékaře akceptovat?

Výsledek:

Testujeme H_0: $\mu \leq 142$ proti H_1: $\mu > 142$ na hladině významnosti 0,05.

Testování pomocí kritického oboru: $W = \langle 1{,}6449, \infty)$, realizace testového kritéria je $-1{,}7773$. Protože testové kritérium se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Testování pomocí intervalu spolehlivosti: 95% empirický levostranný interval spolehlivosti pro střední hodnotu μ je $(136{,}47; \infty)$. Protože číslo 142 patří do tohoto intervalu, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Testování pomocí p-hodnoty: $p = 0{,}9622$. Protože p-hodnota je větší než hladina významnosti 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

- Motivace
- Jednoduché pozorování
- Dvojné pozorování
- Mnohonásobné pozorování

2. Uspořádání pokusů

Cíl kapitoly

Po prostudování této kapitoly budete

- schopni správně naplánovat pokus
- rozeznávat jednoduché, dvojné a mnohonásobné pozorování
- v rámci dvojného pozorování rozlišovat dvouvýběrové a párové porovnávání
- v rámci mnohonásobného pozorování rozlišovat mnohovýběrové a blokové porovnávání

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 2 hodiny studia.

2.1 Motivace

Abychom mohli správně vyhodnotit výsledky pokusu, musí být pokus dobře naplánován. V závislosti na záměrech experimentátora rozeznáváme několik typů uspořádání pokusů: jednoduché pozorování (zkoumají se hodnoty náhodné veličiny pozorované za týchž podmínek), dvojné pozorování (zkoumá se rozdílnost hodnot náhodné veličiny pozorované za dvojích různých podmínek) a mnohonásobné pozorování (zkoumá se rozdílnost hodnot náhodné veličiny pozorované za $r \geq 3$ různých podmínek). Podle typu uspořádání pokusu pak volíme vhodnou statistickou metodu.

V této kapitole probereme pouze ty nejjednodušší typy uspořádání pokusů. V praxi (např. v medicínském nebo zemědělském výzkumu) používají vědci často velmi složité plány experimentů. V doporučené literatuře [HENDL] je plánování experimentů věnována podkapitola 2.4.

V následujícím textu se zaměříme na situaci, kdy zkoumáme hmotnostní přírůstky stejně starých selat téhož plemene při různých výkrmných dietách. Určitou výkrmnou dietu aplikujeme např. po dobu půl roku. Každý den zjišťujeme hmotností přírůstky každého selete a po uplynutí půl roku vypočteme pro každé sele průměrný hmotnostní přírůstek.

2.2 Jednoduché pozorování

Náhodná veličina je pozorována za týchž podmínek. Situace je charakterizována jedním náhodným výběrem $X_1, \dots, X_n$. (Náhodně vybereme n stejně starých selat téhož plemene, podrobíme je jediné výkrmné dietě a zjistíme hmotnostní přírůstky. Tak dostaneme realizaci jednoho náhodného výběru.)

Pokud lze očekávat, že náhodný výběr pochází z normálního rozložení, můžeme např. konstruovat interval spolehlivosti pro neznámou střední hodnotu, neznámý rozptyl či směrodatnou odchylku průměrných denních hmotnostních přírůstků nebo testovat hypotézu, že střední hodnota průměrných denních hmotnostních přírůstků neklesne pod určitou hranici. (Tyto úkoly budeme řešit ve 4. kapitole.)

2.3 Dvojné pozorování

Zkoumá se rozdílnost hodnot náhodné veličiny pozorované za dvojích různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

2.3.1 Dvouvýběrové porovnávání

Situace je charakterizována dvěma nezávislými náhodnými výběry $X_{11}, \ldots, X_{1n_1}$ a $X_{21}, \ldots, X_{2n_2}$. (Z populace všech dostupných stejně starých selat téhož plemene náhodně vybereme $n_1 + n_2$ jedinců. Náhodně je rozdělíme na dva soubory o rozsazích n_1 a n_2, první podrobíme výkrmné dietě č. 1 a druhý výkrmné dietě č. 2. Tak dostaneme realizace dvou nezávislých náhodných výběrů.)

Za předpokladu, že dané náhodné výběry pocházejí z normálních rozložení, lze např. konstruovat interval spolehlivosti pro rozdíl středních hodnot či podíl rozptylů průměrných denních hmotnostních přírůstků nebo testovat hypotézu o stejné účinnosti obou výkrmných diet. (Tyto úkoly budeme řešit v 5. kapitole.)

2.3.2 Párové porovnávání

Situace je charakterizována jedním náhodným výběrem $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ z dvourozměrného rozložení. Párem se rozumí dvojice (X_{i1}, X_{i2}), $i = 1, \ldots, n$. Úloha se zpravidla převádí na jednoduché pozorování náhodného výběru rozdílů $X_{i1} - X_{i2}$, kde $i = 1, \ldots, n$. (Náhodně vybereme n vrhů stejně starých selat téhož plemene a z nich vždy dva sourozence a náhodně jim přiřadíme 1. a 2. výkrmnou dietu. Tak dostaneme realizaci náhodného výběru z dvourozměrného rozložení.)

Lze-li dvourozměrný náhodný výběr považovat za výběr z dvourozměrného normálního rozložení, budeme se zabývat konstrukcí intervalu spolehlivosti pro rozdíl středních hodnot průměrných denních hmotnostních přírůstků nebo testovat hypotézu o stejné účinnosti obou výkrmných diet. (Řešení úkolů tohoto typu je popsáno ve 4. kapitole.)

2.4 Mnohonásobné pozorování

Zkoumá se rozdílnost hodnot náhodné veličiny pozorované za $r \geq 3$ různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

2.4.1 Mnohovýběrové porovnávání

Situace je charakterizována r nezávislými náhodnými výběry $X_{11}, \ldots, X_{1n_1}, \ldots, X_{r1}, \ldots, X_{rn_r}$. (Z populace všech dostupných stejně starých selat téhož plemene náhodně vybereme $n_1 + n_2 + \cdots + n_r$ jedinců. Náhodně je rozdělíme na r souborů o rozsazích $n_1, n_2, \ldots, n_r$. Selata z prvního souboru podrobíme výkrmné dietě č. 1, ..., selata z r-tého souboru podrobíme výkrmné dietě č. r. Tak dostaneme realizace r nezávislých náhodných výběrů.)

Za předpokladu, že všechny náhodné výběry se řídí normálním rozložením s týmž rozptylem, můžeme testovat hypotézu o stejné účinnosti všech r výkrmných diet. (Tomuto problému je věnována 6. kapitola.)

2.4.2 Blokové porovnávání

Situace je charakterizována jedním náhodným výběrem $(X_{11}, X_{12}, \ldots, X_{1r}), \ldots, (X_{n1}, X_{n2}, \ldots, X_{nr})$ z r-rozměrného rozložení. Blokem se rozumí r-tice $(X_{i1}, X_{i2}, \ldots, X_{ir})$, $i = 1, \ldots, n$. (Náhodně vybereme n vrhů starých selat téhož plemene a z nich vždy r sourozenců a náhodně jim přiřadíme 1. až r-tou výkrmnou dietu. Tak dostaneme realizaci náhodného výběru z r-rozměrného rozložení.)

Vyhodnocení výsledků při blokovém porovnávání se provádí např. pomocí Friedmanova testu. Jeho popis se již vymyká náplni předmětu Statistika II. Poučení lze nalézt v doporučené literatuře [HENDL] na str. 360.

Shrnutí kapitoly

Existují tři základní způsoby uspořádání pokusů:

- *jednoduché pozorování* (náhodná veličina je pozorována za týchž podmínek),
- *dvojné pozorování* (náhodná veličina je pozorována za dvojích různých podmínek, přičemž lze použít buď *dvouvýběrové porovnávání* – výsledkem jsou dva nezávislé náhodné výběry nebo *párové porovnávání* – výsledkem je jeden náhodný výběr z dvourozměrného rozložení)
- *mnohonásobné pozorování* (náhodná veličina je pozorována za $r \geq 3$ různých podmínek, přičemž lze použít buď *mnohovýběrové porovnávání* – výsledkem je $r \geq 3$ nezávislých náhodných výběrů nebo *blokové porovnávání* – výsledkem je jeden náhodný výběr z r-rozměrného rozložení).

Správnému uspořádání pokusů je zapotřebí věnovat patřičnou pozornost, neboť při nevhodném uspořádání nelze efektivně využít informace obsažené v datech a prostředky vynaložené na jejich získání jsou znehodnoceny.

Kontrolní otázky

1. Popište tři způsoby plánování pokusů.
2. Jak se liší dvouvýběrové a párové porovnávání?
3. Jak se liší mnohovýběrové a blokové porovnávání?

Autokorekční test

1. Z následujících tří možností vyberte správnou:
   Pokud u několika osob měříme krevní tlak před zátěží a po zátěži, jedná se o
   a) jednoduché pozorování
   b) dvouvýběrové porovnávání
   c) párové porovnávání.
2. Z následujících tří možností vyberte správnou:
   Náhodně vybereme dostatečný počet rodin s dětmi a zkoumáme, zda počet dětí ovlivňuje průměrné roční výdaje rodiny na průmyslové zboží. V tomto případě se jedná o
   a) párové porovnávání
   b) mnohovýběrové porovnávání
   c) blokové porovnávání.

3. Z následujících tří možností vyberte správnou:
Náhodně vybereme dostatečný počet mužů a žen se stejným pracovním zařazením. Zkoumáme, zda pohlaví má vliv na výši průměrného ročního platu. Pro tuto situaci využijeme

a) blokové porovnávání
b) párové porovnávání
c) dvouvýběrové porovnávání.

Správné odpovědi: 1c) 2b) 3c)

- Motivace
- Krabicový diagram
- Normální pravděpodobnostní graf (N–P plot)
- Kvantil–kvantilový graf (Q–Q plot)
- Histogram
- Dvourozměrný tečkový diagram
- Kolmogorovův-Smirnovův test normality dat
- Shapirův-Wilksův test normality dat
- Vzorový příklad

3. Diagnostické grafy a testy normality dat

Cíl kapitoly

Po prostudování této kapitoly budete

- znát způsob konstrukce krabicového diagramu, normálního pravděpodobnostního grafu, kvantil–kvantilového grafu, histogramu a dvourozměrného tečkového diagramu a budete umět tyto grafy vytvořit v systému STATISTICA
- schopni pomocí těchto diagnostických grafů orientačně posoudit povahu dat
- umět v systému STATISTICA provádět testy normality dat

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 20 hodin studia.

3.1 Motivace

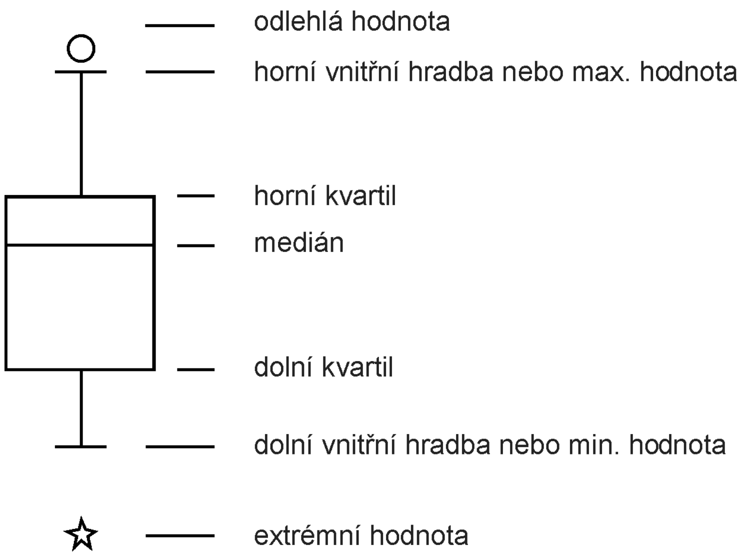
Diagnostické grafy slouží především k tomu, aby nám pomohly orientačně posoudit povahu dat a určit směr další statistické analýzy. Při zpracování dat se často předpokládá splnění určitých podmínek. V případě jednoho náhodného výběru je to především normalita (posuzujeme ji pomocí N–P plotu, Q–Q plotu, histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram neboli box plot). U dvou či více nezávislých náhodných výběrů sledujeme kromě normality též shodu středních hodnot nebo shodu rozptylů – homoskedasticitu (porovnáváme vzhled krabicových diagramů). V případě jednoho dvourozměrného náhodného výběru často posuzujeme dvourozměrnou normalitu dat (použijeme dvourozměrný tečkový diagram s proloženou $100(1-\alpha)\%$ elipsou konstantní hustoty pravděpodobnosti).

Vzhledem k důležitosti předpokladu normality se vedle grafického posouzení doporučuje též použití některého testu normality, např. Kolmogorovova-Smirnovova testu nebo Shapirova-Wilksova testu. K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně $n > 30$) a test zamítne na obvyklé hladině významnosti 0,01 nebo 0,05 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.

3.2 Krabicový diagram

3.2.1 Popis diagramu

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot. Způsob konstrukce je zřejmý z obrázku:



Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1,5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1,5q)$. Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

3.2.2 Příklad

U 30 domácností byl zjišťován počet členů.

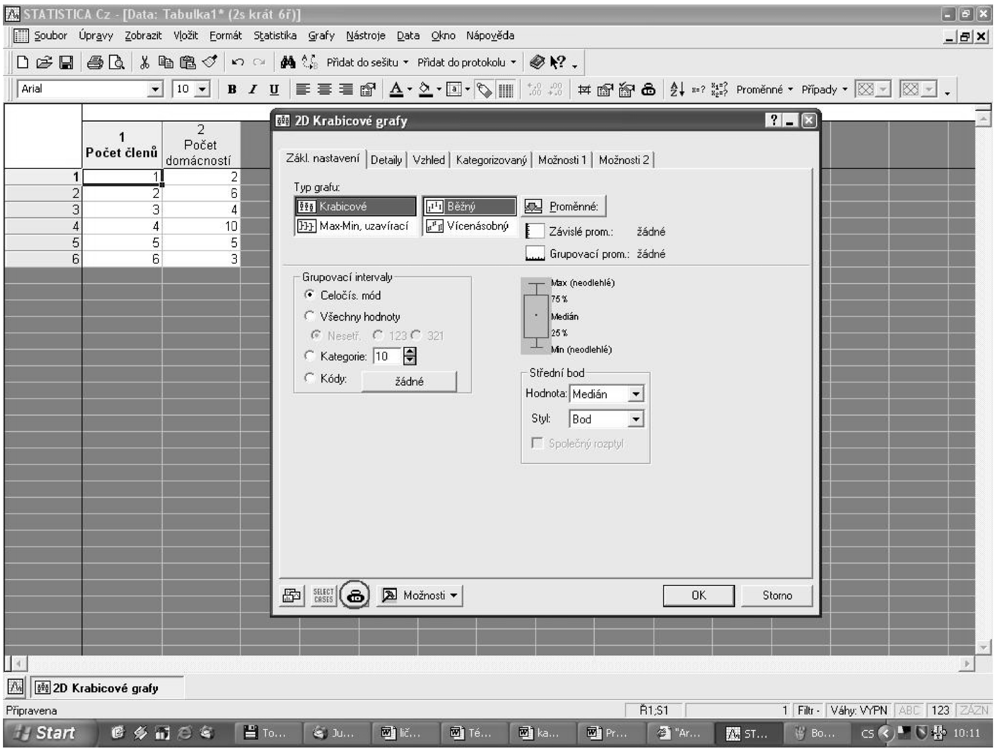
Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

Pro tyto údaje sestrojte krabicový diagram.

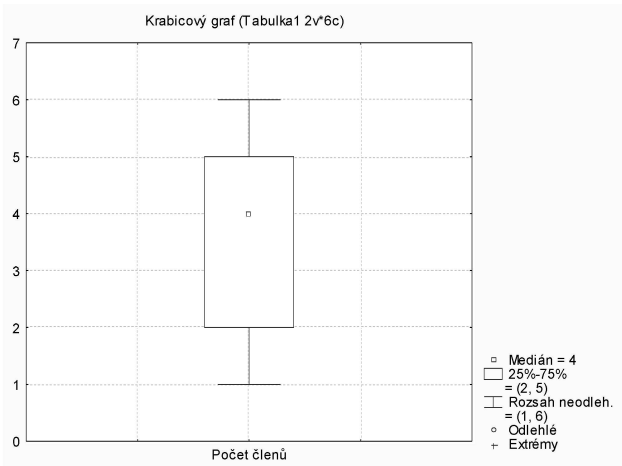
Řešení pomocí systému STATISTICA:

Data zapíšeme do datového okna programu STATISTICA. Po spuštění programu zadáme Soubor – Nový – Počet proměnných 2, Počet případů 6, OK. První proměnnou přejmenujeme na Počet členů, druhou na Počet domácností. (Přejmenování uskutečníme tak, že 2× klikneme myší na název proměnné a tím se otevře okno se specifikacemi proměnné.)

Vytvoření krabicového diagramu: Grafy – 2D Grafy – Krabicové grafy. Abychom systému STATISTICA sdělili, že pracujeme s údaji, pro které známe absolutní četnosti, klikneme myší na tlačítko s obrázkem závaží – na obrázku je v kroužku.



V okénku Váhy případů pro analýzu/graf zaškrtneme Status Zapnuto a zadáme Proměnná vah Počet domácností, OK. Na panelu 2D Krabicové grafy zadáme Proměnné – Závisle proměnné Počet členů, OK. Dostaneme krabicový digram



Z obrázku lze vyčíst, že medián je 4 (aspoň polovina domácností má aspoň 4 členy), dolní kvartil 2 (aspoň čtvrtina domácností má aspoň 2 členy), horní kvartil 5 (aspoň tři čtvrtiny domácností mají aspoň 5 členů), minimum 1, maximum 6. Kvartilová odchylka je $5 - 2 = 3$. Datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně zešikmen. Odlehlé ani extrémní hodnoty se nevyskytují.

3.3 Normální pravděpodobnostní graf (N–P plot)

Před popisem tohoto grafu se musíme seznámit s pojmem pořadí čísla v posloupnosti čísel: Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.

a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.
b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

3.3.1 Příklad

a) Jsou dána čísla 9, 4, 5, 7, 3, 1.
b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.

Stanovte pořadí těchto čísel.

Řešení

ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

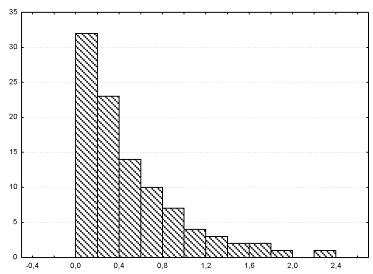
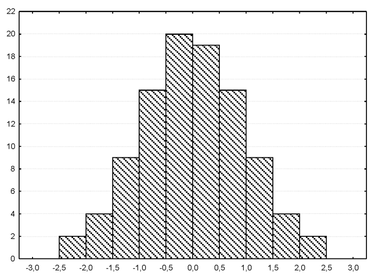
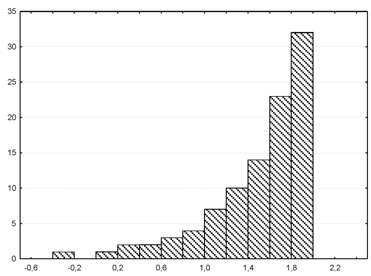
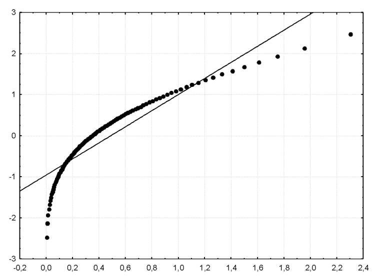
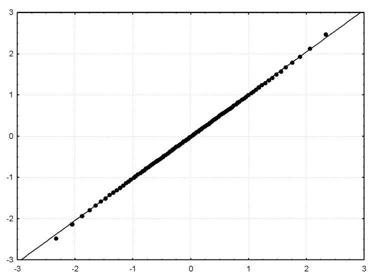
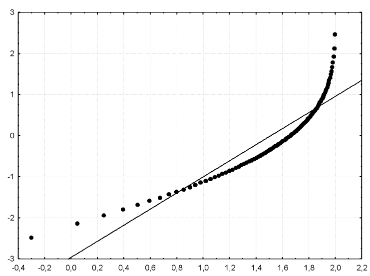
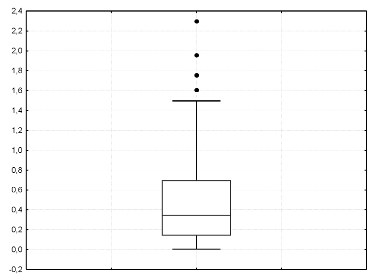
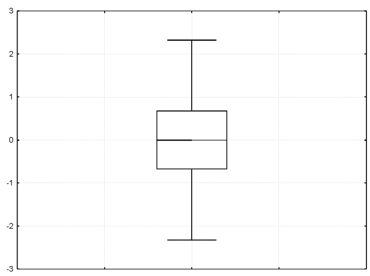
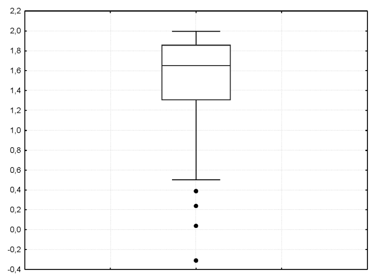
usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,25	2,25	2,25	2,25	5,5	5,5	7	8,5	8,5	10

3.3.2 Popis grafu

N–P plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

Způsob konstrukce:
na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde $\alpha_j = \dfrac{3j-1}{3n+1}$, přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince). Pocházejí-li data z normálního rozložení, pak všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konkávní křivky, zatímco pro data z rozložení se zápornou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konvexní křivky.

Rozložení s kladnou šikmostí	Normální rozložení	Rozložení se zápornou šikmostí
Histogram 	Histogram 	Histogram 
N–P plot 	N–P plot 	N–P plot 
Krabicový diagram 	Krabicový diagram 	Krabicový diagram 

3.3.3 Příklad

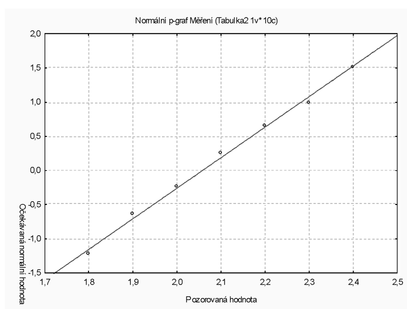
Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření:

2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2.

Pomocí normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

Po zapsání dat do proměnné nazvané Měření zvolíme Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné Měření, OK.



Protože dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

3.4 Kvantil–kvantilový graf (Q–Q plot)

3.4.1 Popis grafu

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. systém STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo). Pro nás je nejdůležitější právě normální rozložení.

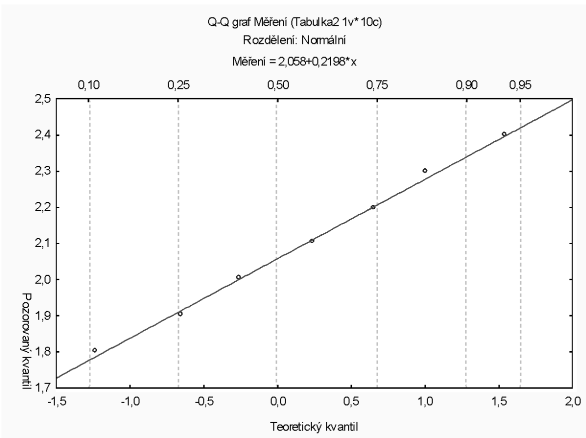
Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde $\alpha_j = \dfrac{j - r_{\text{adj}}}{n + n_{\text{adj}}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0{,}5$, implicitně $r_{\text{adj}} = 0{,}375$ a $n_{\text{adj}} = 0{,}25$. (Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel. Body $\left(K_{\alpha_j}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

3.4.2 Příklad

Pro data z příkladu 3.3.3 posuďte pomocí kvantil–kvantilového grafu, zda pocházejí z normálního rozložení.

Řešení:

Zvolíme Grafy – 2D Grafy – Grafy typu Q–Q – ponecháme implicitní nastavení na normální rozložení (pokud bychom chtěli změnit nastavení na jiný typ rozložení, zvolili bychom ho na záložce Detaily) – Proměnné Měření, OK.



Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

3.5 Histogram

3.5.1 Popis grafu

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se může zakreslit tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě osmi typů rozložení uvedených u Q–Q plotu umožňuje STATISTICA použít ještě další čtyři rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

3.5.2 Příklad

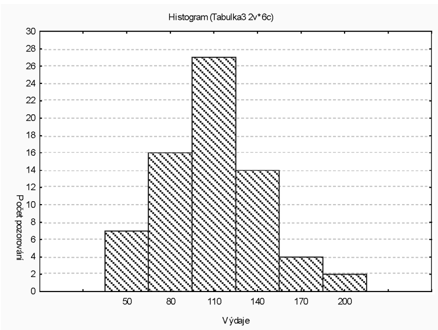
U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	$\langle 35,65\rangle$	$\langle 65,95\rangle$	$\langle 95,125\rangle$	$\langle 125,155\rangle$	$\langle 155,185\rangle$	$\langle 185,215\rangle$
Počet dom.	7	16	27	14	4	2

Nakreslete histogram

Řešení pomocí systému STATISTICA:

Vytvoříme nový datový soubor s dvěma proměnnými Výdaje a Počet domácností. Do proměnné Výdaje zapíšeme středy třídicích intervalů, do proměnné Počet domácností odpovídající absolutní četnosti třídicích intervalů. V menu zvolíme Grafy – Histogramy – pomocí tlačítka s obrázkem závaží zadáme proměnnou vah Počet domácností – OK, Proměnná Výdaje – zapneme volbu Všechny hodnoty – OK. Dostaneme histogram:



Vidíme, že tvar histogramu není symetrický. Malé hodnoty jsou četnější než velké – datový soubor je kladně zešikmen.

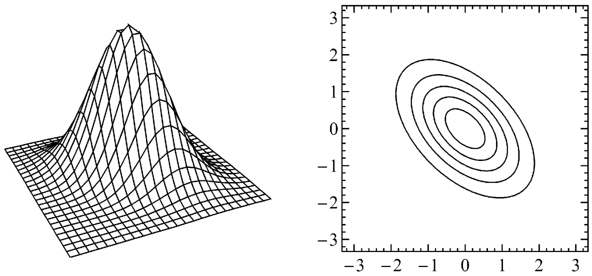
3.6 Dvourozměrný tečkový diagram

3.6.1 Popis diagramu

Máme dvourozměrný datový soubor $(x_1,y_1),\ldots,(x_n,y_n)$, který je realizací dvourozměrného náhodného výběru $(X_1,Y_1),\ldots,(X_n,Y_n)$ z dvourozměrného rozložení.

Na vodorovnou osu vyneseme hodnoty x_j, na svislou hodnoty y_k a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dvojice (x_j, y_k). Jedná-li se o náhodný výběr z dvourozměrného normálního rozložení, měly by tečky zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy – viz následující obrázek.

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}5$:



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než $100\alpha\%$ teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

3.6.2 Příklad

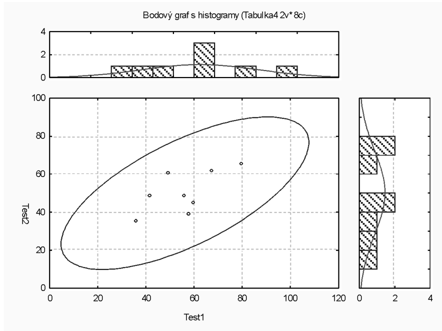
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti a histogramy pro počty bodů v 1. a 2. testu posuďte, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Řešení pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými *Test1* a *Test2* a osmi případy. Nyní nakreslíme dvourozměrný tečkový diagram: Grafy – 2D Grafy – Bodové grafy s histogramy. V typu proložení pro bodový graf vypneme lineární proložení. Proměnné – X – Test1, Y – Test2 – OK. Dostaneme dvourozměrný tečkový diagram pro vektorovou proměnnou (*Test1*, *Test2*) a histogramy pro *Test1* a *Test2*. Nyní do diagramu zakreslíme 95% elipsu konstantní hustoty pravděpodobnosti: 2× klikneme na pozadí grafu a otevře se okno s názvem Vš. možnosti. Vybereme Graf: Elipsa, zvolíme Přidat novou elipsu. Po vykreslení elipsy změníme měřítko: na vodorovné ose bude minimum 0, maximum 120, na svislé ose bude minimum 0, maximum 100. (Stačí 2× kliknout na číselný popis osy a na záložce Měřítka vybrat manuální mód.)



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti, tzn., že u studentů, kteří měli vysoký resp. nízký počet bodů v 1. testu, lze očekávat vysoký resp. nízký počet bodů ve 2. testu.

3.7 Kolmogorovův-Smirnovův test normality dat

3.7.1 Popis testu

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozložení s parametry μ a σ^2. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup\limits_{-\infty < x < \infty} |F_n(x) - \Phi_T(x)|$. Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$.

V případě, že neznáme parametry μ a σ^2 normálního rozložení, změní se rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

3.7.2 Poznámka ke K-S testu ve STATISTICE

Test normality poskytuje hodnotu testové statistiky (ozn. d) a dvě p-hodnoty. První se vztahuje k případu, kdy μ a σ^2 známe předem, druhá (ozn. Liliefors p) se vztahuje k případu, kdy μ a σ^2 neznáme. Objeví-li se ve výstupu $p =$ n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti $0{,}05$.

3.7.3 Příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K-S testu zjistěte na hladině významnosti $0{,}05$, zda tato data pocházejí z normálního rozložení.

Řešení pomocí systému STATISTICA:

Vytvoříme nový datový soubor o jedné proměnné nazvané X a pěti případech. Do proměnné X zapíšeme uvedené hodnoty. V menu vybereme Statistika – Základní statistiky/tabulky – Popisné statistiky – OK, Proměnné X – OK. Na záložce zvolíme

Normalita a vybereme buď Tabulky četností nebo histogramy. Zvolíme např. tabulku četností:

	Tabulka četností:X (Tabulka5) K-S d=,22409, p> .20; Lilliefors p> .20					
Kategorie	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
7,000000<x<=8,000000	1	1	20,00000	20,0000	20,00000	20,0000
8,000000<x<=9,000000	1	2	20,00000	40,0000	20,00000	40,0000
9,000000<x<=10,00000	1	3	20,00000	60,0000	20,00000	60,0000
10,00000<x<=11,00000	0	3	0,00000	60,0000	0,00000	60,0000
11,00000<x<=12,00000	1	4	20,00000	80,0000	20,00000	80,0000
12,00000<x<=13,00000	0	4	0,00000	80,0000	0,00000	80,0000
13,00000<x<=14,00000	0	4	0,00000	80,0000	0,00000	80,0000
14,00000<x<=15,00000	0	4	0,00000	80,0000	0,00000	80,0000
15,00000<x<=16,00000	1	5	20,00000	100,0000	20,00000	100,0000
ChD	0	5	0,00000		0,00000	100,0000

(V posledním řádku symbol ChD znamená chybějící data – v našem případě se v souboru nevyskytují).

Vidíme, že testová statistika K-S testu je $d = 0{,}22409$, odpovídající Lilieforsova p-hodnota je větší než 0,2, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05. Stejný výsledek dostaneme, pokud necháme vykreslit histogram.

3.8 Shapirův-Wilksův test normality dat

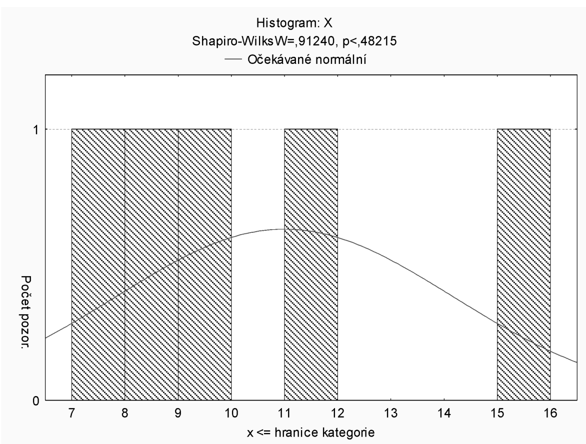
Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$. Test je založen na zjištění, zda body v kvantil–kvantilovém grafu jsou významně odlišné od regresní přímky proložené těmito body. S-W test se používá především pro výběry menších rozsahů, $n < 50$.

3.8.1 Příklad

Pro data z příkladu 3.7.3 proveďte S-W test.

Řešení pomocí systému STATISTICA:

Postupujeme stejně jako v předešlém příkladě, ale na záložce Normalita zaškrtneme Shapiro-Wilksův W test. Vykreslíme histogram:



Testová statistika S-W testu je $W = 0{,}9124$, odpovídající p-hodnota je $0{,}48215$, tedy hypotézu o normalitě nezamítáme na hladině významnosti $0{,}05$.

3.9 Vzorový příklad

Zadání příkladu:

Vedení pojišťovny (zaměřené na pojištění automobilů) požádalo manažera oddělení marketingového výzkumu o provedení průzkumu, který by ukázal názory zákazníků na uvažovaný nový systém pojištění aut.

Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

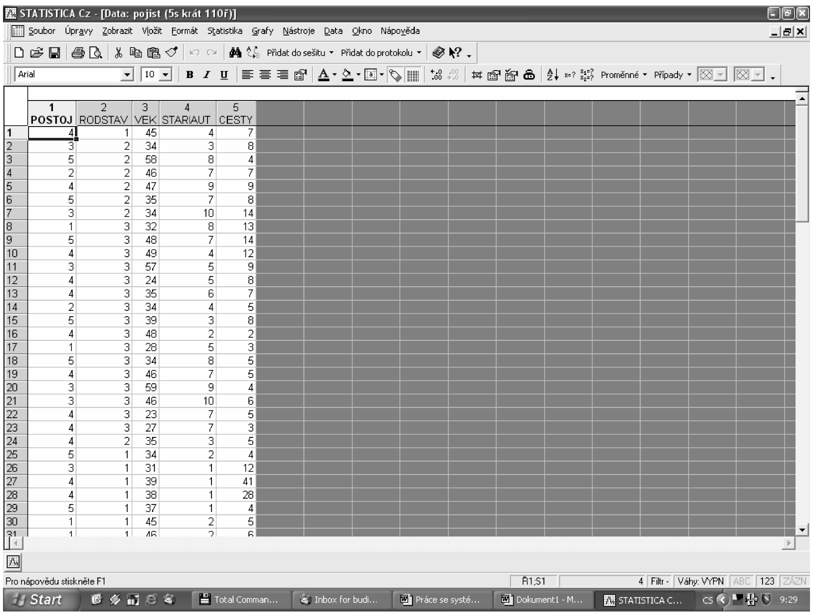
„Naše pojišťovna nabízí nový systém pojištění aut výhradně pro cesty nad 300 km. Za roční poplatek 12 tisíc Kč budete pojištěni pro případ libovolných potíží s autem při všech cestách nad 300 km. V případě nehody pojišťovna uhradí opravu, cestovní náklady a popř. i některé další výlohy, jako je ubytování a stravování v hotelu, telefon atd.

Stupnicí od 1 (jednoznačný nezájem) do 5 (jednoznačný zájem) laskavě vyjádřete svůj postoj k nabízenému novému typu pojištění. Dále uveďte svůj věk, počet cest nad 300 km v loňském roce, stáří vašeho auta a váš rodinný stav. Děkujeme.“

Získané odpovědi byly zaznamenány do datového souboru a zakódovány takto:

POSTOJ ... postoj k novému typu pojištění (jednoznačný nezájem = 1, lehký nezájem = 2, neutrální postoj = 3, lehký zájem = 4, jednoznačný zájem = 5).
RODSTAV ... rodinný stav (svobodný = 1, rozvedený, ovdovělý = 2, ženatý = 3).
VEK ... věk v dokončených letech.
STARIAUT ... stáří auta v letech.
CESTY ... počet cest nad 300 km v předešlém roce.

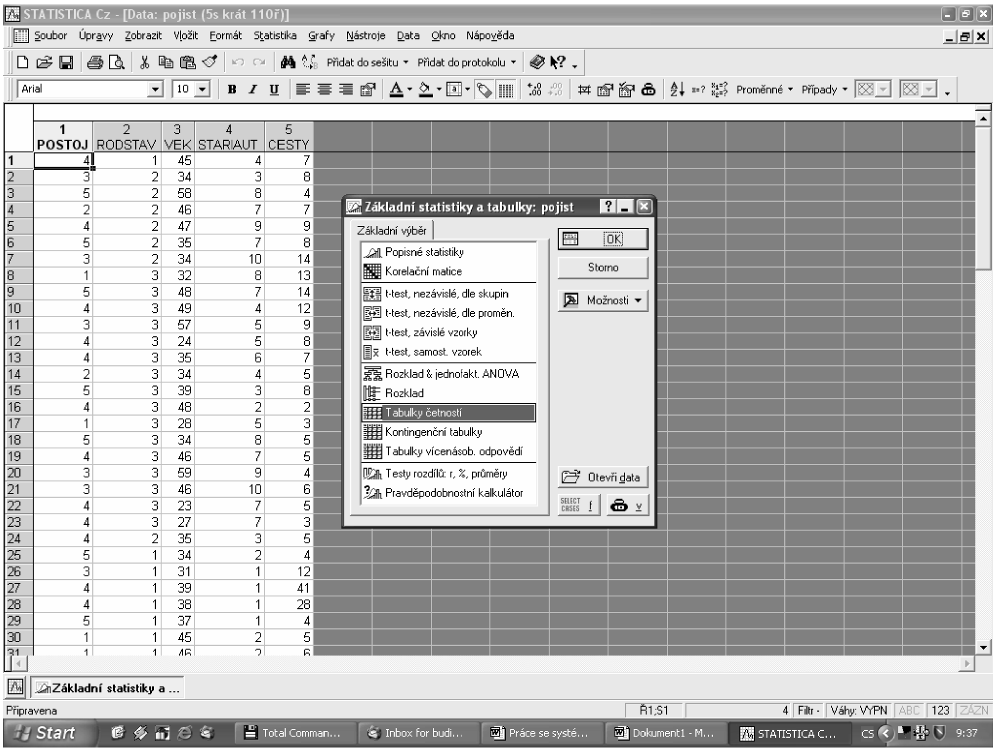
Ukázka části datového souboru:



	1 POSTOJ	2 RODSTAV	3 VEK	4 STARIAUT	5 CESTY
1	4	1	45	4	7
2	3	2	34	3	8
3	5	2	58	8	4
4	2	2	46	7	7
5	4	2	47	9	9
6	5	2	35	7	8
7	3	2	34	10	14
8	1	3	32	8	13
9	5	3	48	7	14
10	4	3	49	4	12
11	3	3	57	5	9
12	4	3	24	5	8
13	4	3	35	6	7
14	2	3	34	4	5
15	5	3	39	3	8
16	4	3	48	2	2
17	1	3	28	5	3
18	5	3	34	8	5
19	4	3	46	7	5
20	3	3	59	9	4
21	3	3	46	10	6
22	4	3	23	7	5
23	4	3	27	7	3
24	4	2	35	3	5
25	5	1	34	2	4
26	3	1	31	1	12
27	4	1	39	1	41
28	4	1	38	1	28
29	5	1	37	1	4
30	1	1	45	2	5
31	1	1	46	2	6

Úkol 1. Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných *POSTOJ* a *RODSTAV*.

Návod:
V menu zvolíme položku Statistika – Základní statistiky/tabulky – Tabulky četností – OK.



Pro analýzu vybereme proměnné *POSTOJ*, *RODSTAV* – OK. Zvolíme Výpočet: Tabulky četností. Získáme tabulku četností pro *POSTOJ*

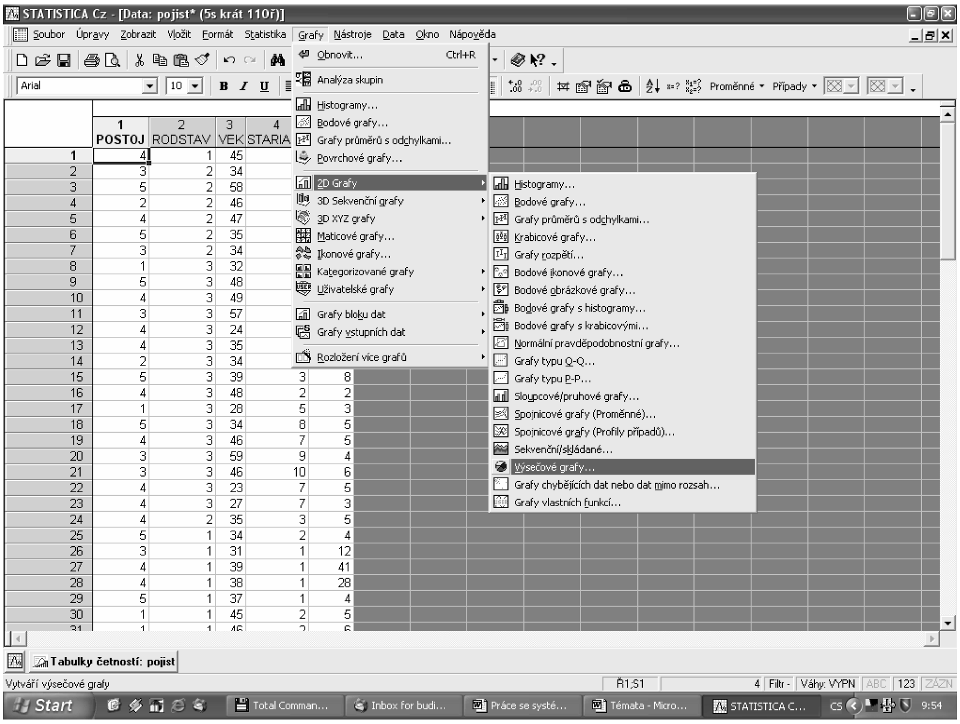
	Tabulka četností:POSTOJ: Postoj k novému typu pojištění (pojist)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
jednoznačný nezájem	8	8	7,27273	7,2727
lehký nezájem	21	29	19,09091	26,3636
neutrální postoj	23	52	20,90909	47,2727
lehký zájem	34	86	30,90909	78,1818
jednoznačný zájem	24	110	21,81818	100,0000
ChD	0	110	0,00000	100,0000

a pro *RODSTAV*

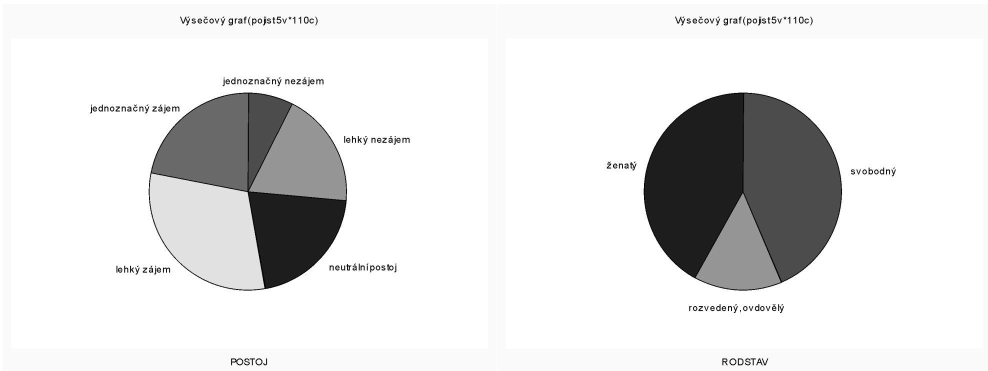
	Tabulka četností:RODSTAV: Rodinný stav (pojist)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
svobodný	48	48	43,63636	43,6364
rozvedený, ovdovělý	16	64	14,54545	58,1818
ženatý	46	110	41,81818	100,0000
ChD	0	110	0,00000	100,0000

Úkol 2. Absolutní četnosti proměnných *POSTOJ* a *RODSTAV*. Znázorněte graficky pomocí výsečového diagramu.

Návod:
V menu zvolíme Grafy – 2D grafy – Výsečové grafy.



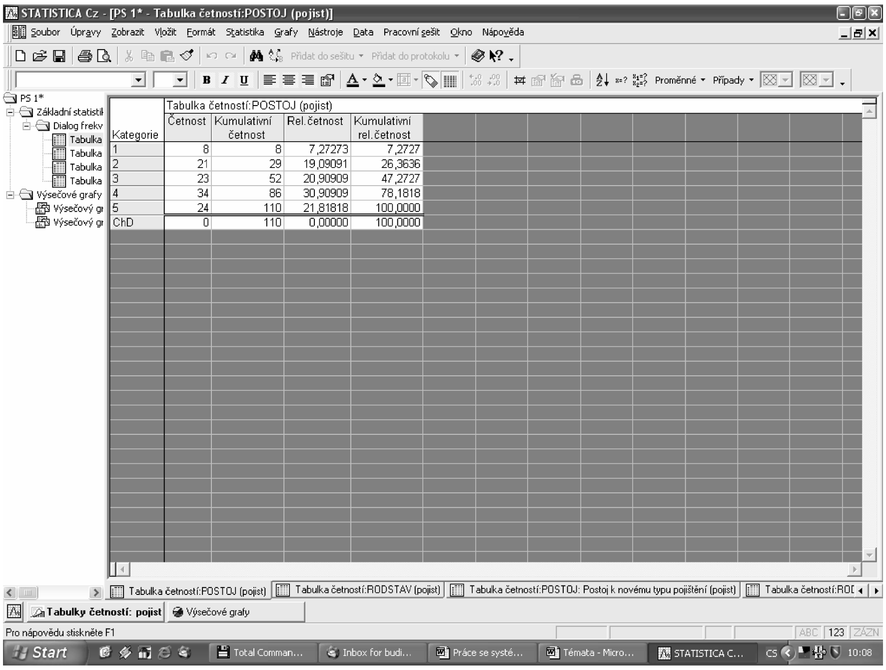
Vybereme proměnné *POSTOJ*, *RODSTAV* a dostaneme následující grafy:



Z prvního diagramu je zřejmé, že nejméně zákazníků projevilo jednoznačný nezájem o nový typ pojištění. Ostatní varianty jsou zastoupeny vcelku rovnoměrně.

Co se týká rodinného stavu zákazníků, vidíme, že v daném souboru jsou s přibližně stejnou četností zastoupeni ženatí a svobodní zákazníci. Rozvedených či ovdovělých je nejméně.

Všechny tabulky a grafy se ukládají do pracovního sešitu. Listovat v nich lze pomocí stromové struktury v levém okně.



Úkol 3. Vypočtěte následující číselné charakteristiky:

a) *POSTOJ* (ordinální proměnná) – modus, medián, dolní a horní kvartil, kvartilová odchylka.

b) *RODSTAV* (nominální proměnná) – modus.

c) *VEK*, *STARIAUT*, *CESTY* (poměrové proměnné) – průměr, směrodatná odchylka, šikmost, špičatost.

Návod:

ad a) Statistika – Základní statistiky/tabulky – Popisné statistiky – Proměnné *POSTOJ* – OK. Na záložce Detaily vybereme Medián, Modus, Dolní & horní kvartily, Kvartilové rozpětí – Souhrn. Dostaneme tabulku

	Popisné statistiky (pojist)					
Proměnná	Medián	Modus	Četnost modu	Spodní kvartil	Horní kvartil	Kvartilové rozpětí
POSTOJ	4,000000	4,000000	34	2,000000	4,000000	2,000000

Vidíme, že medián, modus a horní kvartil jsou stejné – je to varianta 4 „lehký zájem“. Dolním kvartilem je varianta 2 „lehký nezájem“.

ad b) V tabulce Popisné statistiky změníme proměnnou na *RODSTAV* – OK. Na záložce Detaily vybereme Modus – Souhrn. Dostaneme tabulku

	Popisné statistiky (pojist)	
Proměnná	Modus	Četnost modu
RODSTAV	1,000000	48

V našem datovém souboru je nejčetnější variantou rodinného stavu varianta 1 „svobodný“.

ad c) V tabulce Popisné statistiky změníme proměnné na *VEK*, *STARIAUT*, *CESTY* – OK. Na záložce Detaily vybereme Průměr, Směrodat. odchylka, Šikmost, Špičatost – Souhrn. Dostaneme tabulku

Proměnná	Popisné statistiky (pojist)			
	Průměr	Sm. odch.	Šikmost	Špičatost
VEK	39,58182	8,823844	0,191625	-0,59532
STARIAUT	4,16364	2,359938	0,905405	0,35924
CESTY	7,16364	5,304537	3,150711	15,99807

Průměrný věk zákazníků je 39,6 roku, směrodatná odchylka věku činí 8,8 roku. Rozložení věku vykazuje kladnou šikmost (podprůměrné hodnoty věku jsou četnější než nadprůměrné) a zápornou špičatost (rozložení věku je plošší než normální rozložení).

Průměrné stáří auta je 4,2 roku se směrodatnou odchylkou 2,4 roku. Rozložení stáří aut je kladně zešikmené a špičatější než normální rozložení.

Průměrný počet cest v předešlém roce činil 7,2 se směrodatnou odchylkou 5,3. Rozložení počtu cest je značně kladně zešikmené a podstatně špičatější než normální rozložení.

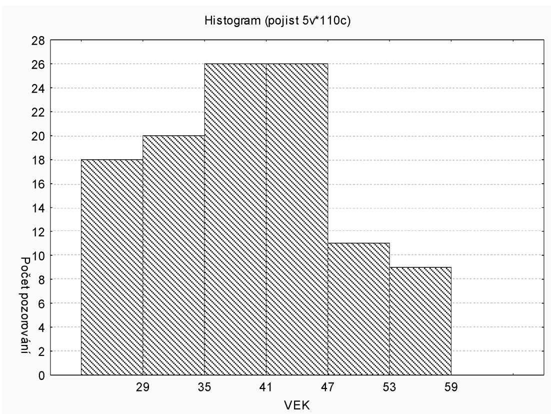
Poznámka: Pokud bychom chtěli porovnat variabilitu uvedených tří proměnných, mohli bychom vypočítat koeficienty variace (koeficient variace je podíl směrodatné odchylky a průměru). Do tabulky s vypočítanými číselnými charakteristikami přidáme další proměnnou nazvanou CV: Proměnné – Přidat – Kolik 1 – Za Špičatost – Jméno CV – do okénka Dlouhé jméno napíšeme =v2/v1 – OK. Dostaneme tabulku

Proměnná	Popisné statistiky (pojist)				
	Průměr	Sm. odch.	Šikmost	Špičatost	CV
VEK	39,58182	8,823844	0,191625	-0,59532	0,222927
STARIAUT	4,16364	2,359938	0,905405	0,35924	0,566797
CESTY	7,16364	5,304537	3,150711	15,99807	0,740481

Vidíme, že nejvyšší variabilitu má proměnná *CESTY*, nejnižší *VEK*.

Úkol 4. Vytvořte histogram proměnné *VEK* se šesti třídicími intervaly $\langle 23, 29\rangle$, $\langle 29, 35\rangle$, $\langle 35, 41\rangle$, $\langle 41, 47\rangle$, $\langle 47, 53\rangle$, $\langle 53, 59\rangle$.

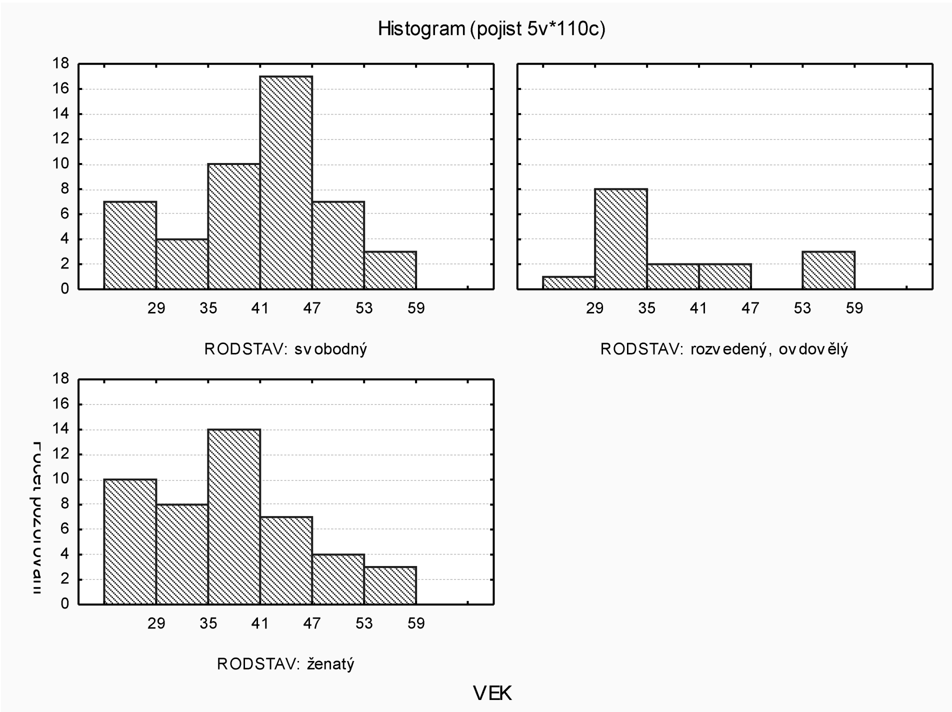
Návod:
V menu vybereme Grafy – Histogramy – Proměnné *VEK*, OK. Odškrtneme Typ proložení: Normální. V záložce Detaily vybereme Hranice – Určit hranice – zadáme horní meze intervalů, tj. 29 35 41 47 53 59, OK, OK. Dostaneme histogram – v tomto tvaru:



Ze vzhledu histogramu lze soudit, že v souboru zákazníku jsou nejvíce zastoupeni lidé od 35 do 47 let. Soubor vykazuje kladné zešikmení, protože mladší věkové kategorie jsou zastoupeny s vyšší četností než starší věkové kategorie.

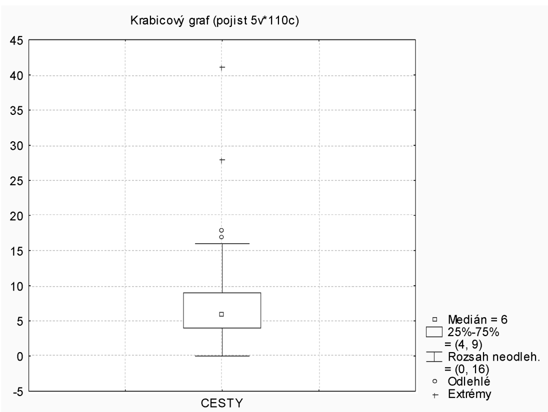
Úkol 5. Vytvořte kategorizovaný histogram proměnné *VEK* podle proměnné *RODSTAV*.

Návod:
Postupujeme stejně jako v předešlém případě, jenom na záložce Kategorizovaný zvolíme Kategorie *X* – Zapnuto, Změnit proměnnou – *RODSTAV*, OK, OK Dostaneme tři histogramy:



Úkol 6. Sestrojte krabicový diagram proměnné *CESTY*. S jeho pomocí zjistěte, zda proměnná *CESTY* obsahuje odlehlé či extrémní hodnoty.

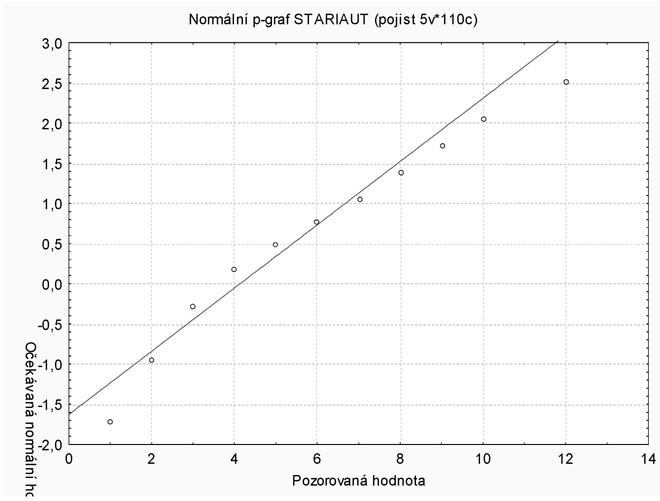
Návod:
V menu Grafy zvolíme 2D Grafy – Krabicové grafy – Proměnné – Závisle proměnné – *CESTY* – OK, OK.



Medián je posunut k dolnímu kvartilu, což svědčí o kladně zešikmeném rozložení. Vyskytují se odlehlé i extrémní hodnoty, jedná se tedy o špičaté rozložení.

Úkol 7. Pro proměnnou *STARIAUT* sestrojte N–P graf a s jeho pomocí posuďte normalitu této proměnné.

Návod:
Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné *STARIAUT* – OK.



Tečky v NP grafu se značně odchylují od zakreslené přímky a řadí se do konkávního tvaru. Datový soubor vykazuje kladné zešikmení, nejedná se tedy o normální rozložení.

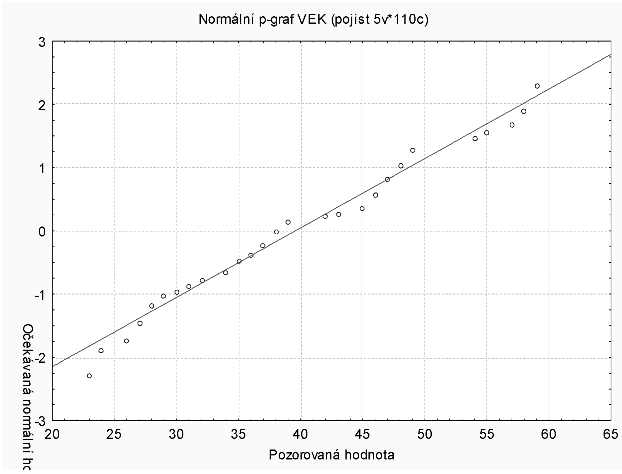
Úkol 8. Rozhodněte pomocí K-S testu a S-W testu na hladině významnosti 0,05, zda lze údaje o věku zákazníků považovat za realizace náhodného výběru z normálního rozložení.

Návod:

Statistika – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné *VEK* – OK. Na záložce zaškrtneme K-S & Lilieforsův test normality a Shaphiro-Wilksův W test – zvolíme Tabulky četností.

	Tabulka četností:VEK: Věk zákazníka (v letech) (pojist) K-S d=,11222, p<,15 ; Lilliefors p<,01 Shapiro-WilksW=,96695, p<,00783					
Kategorie	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
15,00000<x<=20,00000	0	0	0,00000	0,0000	0,00000	0,0000
20,00000<x<=25,00000	4	4	3,63636	3,6364	3,63636	3,6364
25,00000<x<=30,00000	15	19	13,63636	17,2727	13,63636	17,2727
30,00000<x<=35,00000	19	38	17,27273	34,5455	17,27273	34,5455
35,00000<x<=40,00000	26	64	23,63636	58,1818	23,63636	58,1818
40,00000<x<=45,00000	8	72	7,27273	65,4545	7,27273	65,4545
45,00000<x<=50,00000	29	101	26,36364	91,8182	26,36364	91,8182
50,00000<x<=55,00000	3	104	2,72727	94,5455	2,72727	94,5455
55,00000<x<=60,00000	6	110	5,45455	100,0000	5,45455	100,0000
ChD	0	110	0,00000		0,00000	100,0000

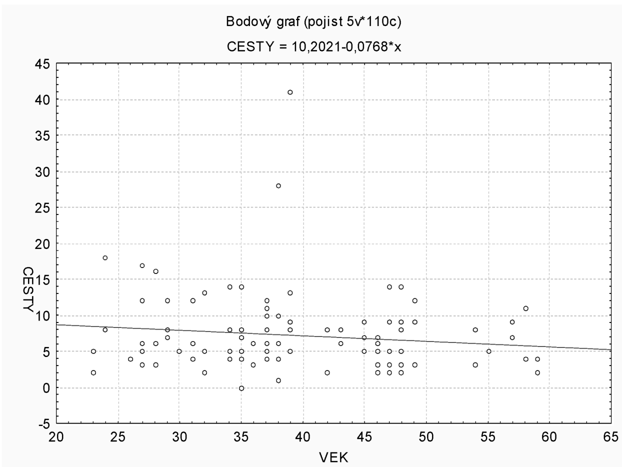
Ve výstupu se objeví tabulka, v níž je uvedena hodnota testové statistiky pro K-S test ($d = 0{,}11222$) a S-W test ($W = 0{,}96695$) a odpovídající p-hodnoty. U K-S testu uvažujeme Lilieforsovo p, které je počítáno na základě parametrů odhadnutých z dat. V našem případě $p < 0{,}01$ a pro S-W test $p = 0{,}00783$, tedy oba testy zamítají na hladině významnosti 0,05 hypotézu o normalitě. Výpočet je vhodné doplnit N–P grafem:



Úkol 9. Pomocí dvourozměrného tečkového diagramu posuďte, zda mezi věkem zákazníka a počtem cest nad 300 km v předešlém roce existuje nějaká lineární závislost.

Návod:

Grafy – Bodové grafy – Proměnné *X* – *VEK*, *Y* – *CESTY* – OK. OK. Dostaneme tento graf:



Vidíme, že s rostoucím věkem zákazníka poněkud klesá počet cest, mezi proměnnými *VEK* a *CESTY* tedy existuje dosti slabá nepřímá lineární závislost.

Shrnutí kapitoly

Při určení směru statistické analýzy dat používáme diagnostické grafy, které umožní posoudit

– normalitu dat či tvar rozložení (*N–P plot*, *Q–Q plot*, *histogram*)
– existenci odlehlých či extrémních hodnot (*krabicový graf*)
– dvourozměrnou normalitu dat (*dvourozměrný tečkový diagram*)

Kromě grafického znázornění dat používáme *testy normality dat*, např. *Kolmogorovův–Smirnovův test* (ve většině reálných situací jeho variantu poskytující *Lilieforsovu p-hodnotu*) nebo *Shapirův-Wilksův* test. Musíme si být ovšem vědomi toho, že pro výběry větších rozsahů (orientačně $n > 30$) i malé odchylky od normality mohou být statisticky významné, i když věcně nikoliv. V takovém případě se nedopustíme závažné chyby, pokud použijeme metodu založenou na předpokladu normality dat.

Kontrolní otázky

1. K čemu slouží diagnostické grafy?

2. Popište způsob konstrukce krabicového diagramu.

3. Jak budete interpretovat situaci, kdy v krabicovém diagramu je medián posunut směrem k dolnímu kvartilu?

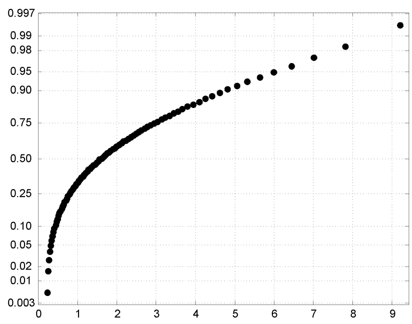
4. V dvourozměrném tečkovém diagramu jsou tečky zhruba rovnoměrně rozptýleny uvnitř kruhového obrazce. Co lze říci o vztahu veličin X a Y?

5. Jak se liší provedení K-S testu normality dat v případě, kdy známe parametry normálního rozložení od případu, kdy je neznáme?

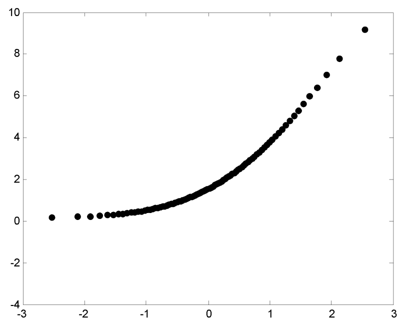
6. Jak souvisí S-W test normality dat s kvantil–kvantilovým grafem?

7. Pro datový soubor o rozsahu $n = 50$ byl vytvořen normální pravděpodobnostní graf a kvantil–kvantilový graf. Pomocí těchto grafů posuďte, zda se data mohou řídit normálním rozložením.

N–P plot



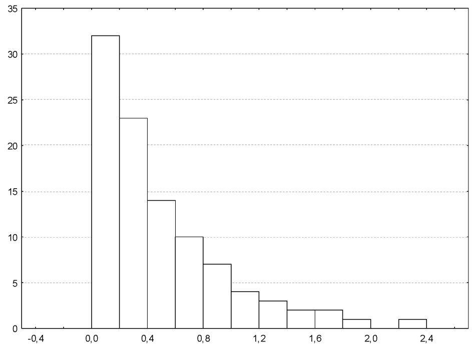
Q–Q plot

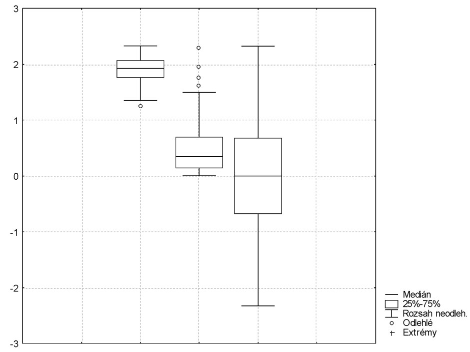


Výsledek: Data nepocházejí z normálního rozložení, vzhled obou diagramů svědčí o značném kladném zešikmení.

Autokorekční test

1. Z 99 hodnot byl sestrojen histogram. Určete, který ze tří uvedených krabicových diagramů byl sestrojen ze stejných hodnot.





a) První krabicový diagram.
b) Druhý krabicový diagram.
c) Třetí krabicový diagram.

2. Určete, která tvrzení jsou pravdivá:

a) Odlehlá hodnota v datovém souboru leží za vnějšími hradbami.
b) Extrémní hodnota v datovém souboru leží mezi vnitřními a vnějšími hradbami.
c) Extrémní hodnota je více vzdálena od mediánu než odlehlá hodnota.

3. Určete, která tvrzení jsou pravdivá:

a) Pocházejí-li data z normálního rozložení, budou se tečky v normálním pravděpodobnostním grafu řadit do přímky.
b) Pocházejí-li data z rozložení s kladnou šikmostí, budou se tečky v normálním pravděpodobnostním grafu řadit do konvexní křivky.
c) Pocházejí-li data z rozložení se zápornou šikmostí, budou se tečky v normálním pravděpodobnostním grafu řadit do konkávní křivky.

4. Určete, která tvrzení jsou pravdivá:

a) Pokud se v dvourozměrném tečkovém diagramu seskupují tečky do elipsovitého útvaru, jehož hlavní osa je přímka s kladnou směrnicí, lze usoudit, že mezi veličinami X a Y existuje určitý stupeň přímé lineární závislosti.
b) Pokud se v dvourozměrném tečkovém diagramu seskupují tečky do kruhovitého útvaru, lze usoudit, že mezi veličinami X a Y existuje určitý stupeň nelineární závislosti.
c) Pokud v dvourozměrném tečkovém diagramu leží všechny tečky na přímce se zápornou směrnicí, lze usoudit, že mezi veličinami X a Y existuje úplná nepřímá lineární závislost.

Správné odpovědi: 1b) 2c) 3a) 4a), c)

Příklady

1. Během semestru se studenti podrobili písemnému testu z matematiky, v němž bylo možno získat 0 až 10 bodů. Výsledky jsou uvedeny v tabulce:

Počet bodů	0	1	2	3	4	5	6	7	8	9	10
Počet studentů	1	4	6	7	11	15	19	17	12	6	3

Pro počet bodů sestrojte krabicový diagram. Je počet bodů symetricky rozložen kolem mediánu? Vyskytují se v datech odlehlé nebo extrémní hodnoty?

Výsledek: $x_{0,25} = 1$, $x_{0,50} = 6$, $x_{0,75} = 7$, medián je posunut k hornímu kvartilu, data vykazují zápornou šikmost. Odlehlé ani extrémní hodnoty se nevyskytují.

2. Pro počet bodů z 1. příkladu sestrojte normální pravděpodobnostní graf.

3. Pro počet bodů z 1. příkladu sestrojte kvantil–kvantilový graf pro normální rozložení.

4. Pro počet bodů z 1. příkladu testujte pomocí K-S testu na hladině významnosti 0,05 hypotézu, že se řídí normálním rozložením. Zjistěte hodnotu testové statistiky a odpovídající p-hodnotu.

Výsledek:
Testová statistika $= 0{,}12895$, Liliefors $p < 0{,}01$, hypotézu o normalitě zamítáme na hladině významnosti 0,05.

5. Pro počet bodů z 1. příkladu testujte pomocí S-W testu na hladině významnosti 0,05 hypotézu, že se řídí normálním rozložením. Zjistěte hodnotu testové statistiky a odpovídající p-hodnotu.

Výsledek:
Testová statistika $= 0{,}96906$, $p < 0{,}01784$, hypotézu o normalitě zamítáme na hladině významnosti 0,05.

6. Na 10 automobilech stejného typu se testovaly dva druhy benzínu lišící se oktanovým číslem. U každého automobilu se při průměrné rychlosti 90 km/h měřil dojezd (tj. dráha, kterou ujede na dané množství benzínu) při použití každého z obou druhů benzínu. Výsledky:

číslo auta	1	2	3	4	5	6	7	8	9	10
benzín A	17,5	20,0	18,9	17,9	16,4	18,9	17,2	17,5	18,5	18,2
benzín B	17,8	20,8	19,5	18,3	16,6	19,5	17,5	17,9	19,1	18,6

Pro uvedená data sestrojte dvourozměrný tečkový diagram se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti. Mohou data pocházet z dvourozměrného normálního rozložení?

Výsledek: ano.

- Motivace
- Rozložení statistik odvozených z výběrového průměru a výběrového rozptylu
- Intervaly spolehlivosti pro parametry μ, σ^2
- Testování hypotéz o parametrech μ, σ^2
- Náhodný výběr z dvourozměrného normálního rozložení

4. Úlohy o jednom náhodném výběru z normálního rozložení

4. Úlohy o jednom náhodném výběru z normálního rozložení

Cíl kapitoly

Po prostudování této kapitoly budete

– znát vlastnosti pivotových statistik odvozených z náhodného výběru z normálního rozložení a budete je umět použít pro řešení konkrétních úloh
– umět sestrojit intervaly spolehlivosti pro střední hodnotu a rozptyl normálního rozložení
– provádět testy hypotéz o střední hodnotě a rozptylu normálního rozložení

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 5 hodin studia.

4.1 Motivace

Mnoho náhodných veličin, s nimiž se setkáváme ve výzkumu i praxi, se řídí normálním rozložením. Za jistých předpokladů obsažených v centrální limitní větě se dá rozložení jiných náhodných veličin aproximovat normálním rozložením. Proto je zapotřebí věnovat velkou pozornost právě náhodným výběrům z normálního rozložení.

Normální rozložení je charakterizováno dvěma parametry – střední hodnotou μ a rozptylem σ^2. Budeme tedy řešit úlohy, které se týkají těchto parametrů. Jedná se především o jednovýběrový t-test či test o rozptylu. Seznámíme se rovněž se situací, kdy máme k dispozici jeden náhodný výběr z dvourozměrného normálního rozložení a posuzujeme rozdílnost středních hodnot obou náhodných veličin. K řešení tohoto problému slouží párový t-test.

4.2 Rozložení statistik odvozených z výběrového průměru a výběrového rozptylu

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$. Pak platí

a) Výběrový průměr M a výběrový rozptyl S^2 jsou stochasticky nezávislé.

b) $M \sim N(\mu, \frac{\sigma^2}{n})$, tedy $U = \dfrac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.
(Pivotová statistika U slouží k řešení úloh o μ, když σ^2 známe.)

c) $K = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.
(Pivotová statistika K slouží k řešení úloh o σ^2, když μ neznáme.)

d) $\dfrac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$.
(Tato pivotová statistika slouží k řešení úloh o σ^2, když μ známe.)

e) $T = \dfrac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$.
(Pivotová statistika T slouží k řešení úloh o μ, když σ^2 neznáme.)

4.2.1 Příklad

Na výrobní lince jsou automaticky baleny balíčky rýže o deklarované hmotnosti 1 000 g. Působením náhodných vlivů hmotnost balíčků kolísá. Lze ji považovat za náhodnou veličinu, která se řídí normálním rozložením se střední hodnotou 996 g a směrodatnou odchylkou 18 g. Jaká je pravděpodobnost, že náhodně vybraný balíček rýže neprojde výstupní kontrolou, jestliže je povolená tolerance ±30 g od deklarované hmotnosti 1 000 g?

Řešení:

Použijeme pivotovou statistiku U z bodu (b).

$$X \sim N(996, 18^2), \quad U = \frac{X - 996}{18} \sim N(0,1)$$

$$\begin{aligned} P(X \notin \langle 970, 1\,030\rangle) &= 1 - P(970 < X < 1\,030) = \\ &= 1 - \left(\frac{970-996}{18} < U < \frac{1\,030-996}{18} \right) = \\ &= 1 - \Phi(1{,}89) + \Phi(-1{,}44) = 2 - 0{,}971 - 0{,}925 = 0{,}104 \end{aligned}$$

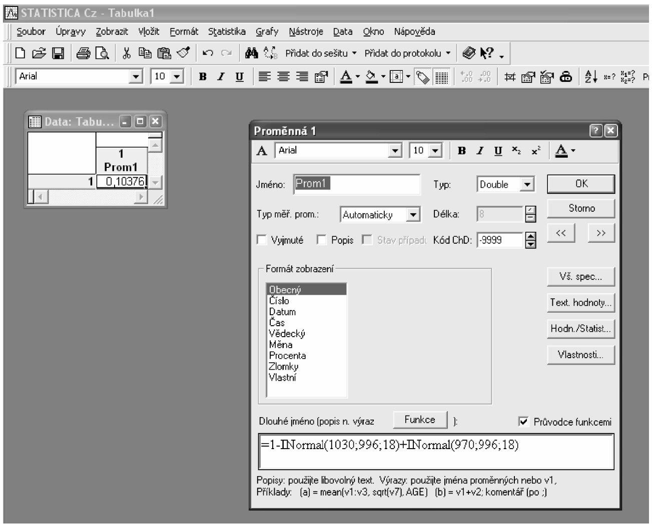
Řešení pomocí systému STATISTICA:

Využijeme toho, že STATISTICA pomocí funkce INormal(x;mu;sigma) umí vypočítat hodnotu distribuční funkce normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma. Tedy

$$\begin{aligned} P(X \notin \langle 970, 1\,030\rangle) &= 1 - P(970 < X < 1\,030) = 1 - [\Phi(1\,030) - \Phi(970)] = \\ &= 1 - \Phi(1\,030) + \Phi(970), \end{aligned}$$

kde Φ je distribuční funkce rozložení $N(996, 18^2)$.

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Dvakrát klikneme na název proměnné *Prom1*. Do Dlouhého jména této proměnné napíšeme =1-INormal(1030;996;18)+INormal(970;996;18). V proměnné *Prom1* se objeví hodnota 0,10376.



4.3 Intervaly spolehlivosti pro parametry μ, σ^2

V kapitole 1 jsme se seznámili s pojmem intervalu spolehlivosti pro parametrickou funkci $h(\vartheta)$. Nyní se budeme zabývat speciálními případy, kdy za parametrickou funkci $h(\vartheta)$ považujeme střední hodnotu μ nebo rozptyl σ^2 normálního rozložení. V příkladu 1.3.5. jsme si ukázali způsob, jak zkonstruovat interval spolehlivosti pro střední hodnotu μ, když rozptyl σ^2 známe. Odvození intervalu spolehlivosti pro další tři situace (tj. pro μ, když σ^2 neznáme, pro σ^2, když μ neznáme a konečně pro σ^2, když μ známe) provádět nebudeme, uvedeme jen přehled vzorců pro meze $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro tyto parametry .

4.3.1 Přehled vzorců

a) Interval spolehlivosti pro μ, když σ^2 známe (využití pivotové statistiky $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$)

Oboustranný: $(d,h) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right)$

Levostranný: $(d,\infty) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right)$

b) Interval spolehlivosti pro μ, když σ^2 neznáme (využití pivotové statistiky $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$)

Oboustranný: $(d,h) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1), m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d,\infty) = \left(m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1), \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$

c) Interval spolehlivosti pro σ^2, když μ neznáme (využití pivotové statistiky $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$)

Oboustranný: $(d,h) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right)$

Levostranný: $(d,\infty) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)}\right)$

d) Interval spolehlivosti pro σ^2, když μ známe (využití pivotové statistiky $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$)

$$\text{Oboustranný: } (d,h) = \left(\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \right)$$

$$\text{Levostranný: } (d,\infty) = \left(\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty,h) = \left(-\infty, \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)} \right)$$

4.3.2 Příklad

10krát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly:

2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2.

Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1,\dots,X_{10}$ z rozložení $N(\mu,\sigma^2)$, kde parametry μ, σ^2 neznáme. Najděte 95% empirický interval spolehlivosti pro μ, a to

a) oboustranný,
b) levostranný,
c) pravostranný.

Řešení:

Vypočteme realizaci výběrového průměru: $m = 2{,}06$, výběrového rozptylu: $s^2 = 0{,}0404$ a výběrové směrodatné odchylky: $s = 0{,}2011$. Riziko α je 0,05. Jde o situaci popsanou v bodě (b), kde využíváme pivotovou statistiku T, která se řídí Studentovým rozložením $t(9)$. V tabulkách najdeme kvantil $t_{0{,}975}(9) = 2{,}2622$ pro oboustranný interval spolehlivosti a kvantil $t_{0{,}95}(9) = 1{,}8331$ pro jednostranné intervaly spolehlivosti.

ad a)
$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}} 2{,}2622 = 1{,}92$$
$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}} 2{,}2622 = 2{,}20$$
$1{,}92 < \mu < 2{,}20$ s pravděpodobností aspoň 0,95.

ad b)
$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 1{,}94$$
$1{,}94 < \mu$ s pravděpodobností aspoň 0,95.

ad c)
$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}} 1{,}8331 = 2{,}18$$
$\mu < 2{,}18$ s pravděpodobností aspoň 0,95.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné (nazveme ji Měření) a 10 případech. Do této proměnné zapíšeme výsledky měření.

ad a) Meze $100(1-\alpha)\%$ empirického oboustranného intervalu spolehlivosti pro střední hodnotu při neznámém rozptylu vypočteme takto: Statistika – Základní statistiky/tabulky – Popisné statistiky – OK, Proměnné – Měření – OK. Na záložce Detaily vybereme Meze spolehl. prům. a ponecháme implicitně nastavenou hodnotu 95 %. Po kliknutí na Souhrn dostaneme tabulku

	Popisné statistiky (Tabulka2)	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. +95,000%
Měření	1,916136	2,203864

Po zaokrouhlení na dvě desetinná místa dostaneme výsledek $1{,}92 < \mu < 2{,}20$ s pravděpodobností aspoň 0,95.

ad b), c) U volby Meze spolehl. prům. změníme hodnotu na 90 %. Dostaneme tabulku

	Popisné statistiky (Tabulka2)	
Proměnná	Int. spolehl. -90,000%	Int. spolehl. +90,000%
Měření	1,943421	2,176579

Odtud získáme dolní mez 95% empirického levostranného intervalu spolehlivosti pro střední hodnotu: $1{,}94 < \mu$ s pravděpodobností aspoň 0,95 a horní mez 95% empirického pravostranného intervalu spolehlivosti pro střední hodnotu: $\mu < 2{,}18$ s pravděpodobností aspoň 0,95.

4.4 Testování hypotéz o parametrech μ, σ^2

a) Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\mu = c$ proti H_1: $\mu \neq c$ se nazývá z-test.

b) Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 neznáme. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\mu = c$ proti H_1: $\mu \neq c$ se nazývá jednovýběrový t-test.

c) Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde μ neznáme. Nechť $n \geq 2$ a c je konstanta. Test H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 \neq c$ se nazývá test o rozptylu.

4.4.1 Provedení testů o parametrech μ, σ^2 pomocí kritického oboru

V kapitole 1 byly uvedeny tři způsoby testování hypotéz – pomocí kritického oboru, pomocí intervalu spolehlivosti a pomocí p-hodnoty. V tomto odstavci si ukážeme, jak testovat hypotézy o střední hodnotě μ a rozptylu σ^2 pomocí kritického oboru.

a) Provedení z-testu

H_0: $\mu = c$ proti H_1: $\mu \neq c$ (resp. H_1: $\mu < c$ resp. H_1: $\mu > c$) zamítáme na hladině významnosti α, jestliže

$$\left|\frac{m-c}{\frac{\sigma}{\sqrt{n}}}\right| \geq u_{1-\alpha/2} \quad \left(\text{resp. } \frac{m-c}{\frac{\sigma}{\sqrt{n}}} \leq -u_{1-\alpha} \quad \text{resp. } \frac{m-c}{\frac{\sigma}{\sqrt{n}}} \geq u_{1-\alpha}\right).$$

b) Provedení jednovýběrového t-testu

H_0: $\mu = c$ proti H_1: $\mu \neq c$ (resp. H_1: $\mu < c$ resp. H_1: $\mu > c$) zamítáme na hladině významnosti α, jestliže

$$\left|\frac{m-c}{\frac{s}{\sqrt{n}}}\right| \geq t_{1-\alpha/2}(n-1)$$

$$\left(\text{resp. } \frac{m-c}{\frac{s}{\sqrt{n}}} \leq -t_{1-\alpha}(n-1) \quad \text{resp. } \frac{m-c}{\frac{s}{\sqrt{n}}} \geq t_{1-\alpha}(n-1)\right).$$

c) Provedení testu o rozptylu

H_0: $\sigma^2 = c$ proti H_1: $\sigma^2 \neq c$ (resp. H_1: $\sigma^2 < c$ resp. H_1: $\sigma^2 > c$) zamítáme na hladině významnosti α, jestliže

$$\frac{(n-1)s^2}{c} \leq \chi^2_{\alpha/2}(n-1) \quad \text{nebo} \quad \frac{(n-1)s^2}{c} \geq \chi^2_{1-\alpha/2}(n-1)$$

$$\left(\text{resp. } \frac{(n-1)s^2}{c} \leq \chi^2_{\alpha}(n-1), \quad \text{resp. } \frac{(n-1)s^2}{c} \geq \chi^2_{1-\alpha}(n-1)\right).$$

Před provedením kteréhokoli z uvedených testů je zapotřebí ověřit normalitu dat pomocí diagnostických grafů a testů normality popsaných v kapitole 3. Zjistíme-li u jednovýběrového t-testu, že rozsah souboru je malý ($n < 30$) a porušení normality je výraznější, doporučuje se přejít k neparametrickému jednovýběrovému Wilcoxonovu testu (viz kapitola 7). Pro výběry větších rozsahů není mírné porušení normality na překážku použití uvedených testů.

4.4.2 Příklad

Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo, že jejich průměrná hmotnost je 122 g a směrodatná odchylka $8,6$ g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti 0,01 považovat stížnosti kupujících za oprávněné?

Řešení:

$X_1, \ldots, X_{50}$ je náhodný výběr z $N(\mu, \sigma^2)$. Testujeme hypotézu H_0: $\mu = 125$ proti levostranné alternativě H_1: $\mu < 125$. Protože neznáme rozptyl σ^2, použijeme jednovýběrový t-test. Realizace testového kritéria

$$t_0 = \frac{m-c}{\frac{s}{\sqrt{n}}} = \frac{122-125}{\frac{8{,}6}{\sqrt{50}}} = -2{,}4667.$$

Hodnotu t_0 porovnáme s opačnou hodnotou kvantilu $t_{0,99}(49) = 2{,}4049$. Jelikož $-2{,}4667 \leq -2{,}4049$, zamítáme nulovou hypotézu na hladině významnosti 0,01. Stížnosti kupujících tedy lze považovat za oprávněné (s rizikem omylu nejvýše 1 %).

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnými (nazveme je t_0 a $kvantil$) a o jednom případu. Do Dlouhého jména proměnné t_0 napíšeme =(122-125)*sqrt(50)/8,6 (tím vypočteme realizaci testového kritéria) a do Dlouhého jména proměnné kvantil napíšeme =-VStudent(0,99;49) (tím vypočteme opačnou hodnotu kvantilu $t_{0,99}(49)$.

Protože $-2{,}46665 \leq -2{,}40489$, zamítáme nulovou hypotézu na hladině významnosti 0,01.

4.5 Náhodný výběr z dvourozměrného normálního rozložení

Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$, přičemž $n \geq 2$. Označíme $\mu = \mu_1 - \mu_2$ a zavedeme rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$. Vypočteme

$$M = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - M)^2.$$

4.5.1 Interval spolehlivosti pro parametr μ

Pro výpočet mezí $100(1-\alpha)$% empirického intervalu spolehlivosti pro střední hodnotu μ použijeme vzorec uvedený v 4.3.1 (b).

4.5.2 Párový t-test

Testujeme H_0: $\mu_1 - \mu_2 = 0$ (tj. $\mu = 0$) proti H_1: $\mu_1 - \mu_2 \neq 0$ (tj. $\mu \neq 0$). Přechodem k rozdílovému náhodnému výběru převedeme párový t-test na jednovýběrový t-test, jehož provedení je popsáno v 4.4.1 (b).

Před provedením párového t-testu je zapotřebí se aspoň orientačně přesvědčit o dvourozměrné normalitě dat pomocí dvourozměrného tečkového diagramu. Je-li rozsah výběru malý ($n < 30$) a porušení normality je výraznější, je zapotřebí místo párového testu použít neparametrický párový Wilcoxonův test (viz kapitola 7). Pro výběry větších rozsahů, které vykazují jen mírné porušení normality, můžeme použít párový t-test.

4.5.3 Příklad

Na 10 automobilech stejného typu se testovaly dva druhy benzínu lišící se oktanovým číslem. U každého automobilu se při průměrné rychlosti 90 km/h měřil dojezd (tj.

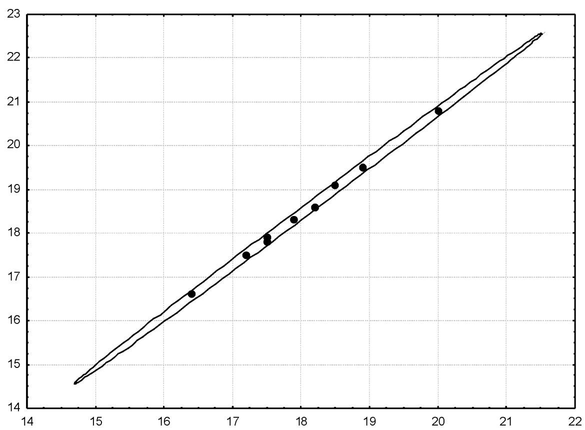
dráha, kterou ujede na dané množství benzínu) při použití každého z obou druhů benzínu. Výsledky:

Číslo auta	1	2	3	4	5	6	7	8	9	10
benzín A	$17,5$	$20,0$	$18,9$	$17,9$	$16,4$	$18,9$	$17,2$	$17,5$	$18,5$	$18,2$
benzín B	$17,8$	$20,8$	$19,5$	$18,3$	$16,6$	$19,5$	$17,5$	$17,9$	$19,1$	$18,6$

Za předpokladu, že dojezd se řídí normálním rozložením, testujte na hladině významnosti 0,05 hypotézu, že rozdíl středních hodnot dojezdu při dvou druzích benzínu se neliší.

Řešení:

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti posoudíme oprávněnost předpokladu o dvourozměrné normalitě dat.



Vidíme, že tečky se řadí do velmi úzkého elipsovitého obrazce. Data můžeme považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Přejdeme k rozdílovému náhodnému výběru. Označíme $\mu = \mu_1 - \mu_2$. Testujeme hypotézu H_0: $\mu = 0$ proti H_1: $\mu \neq 0$ na hladině významnosti 0,05. Vypočteme $m = -0{,}46$, $s = 0{,}1838$ a testové kritérium $t_0 = -7{,}9148$. Absolutní hodnotu testového kritéria porovnáme s kvantilem $t_{0{,}975}(9) = 2{,}2622$. Protože $7{,}9148 \geq 2{,}2622$, zamítáme nulovou hypotézu na hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnými benzín A, benzín B a o deseti případech. Do těchto proměnných zapíšeme zjištěné hodnoty. Dvourozměrný tečkový diagram vytvoříme podobně jako v příkladu 3.6.2. Nyní provedeme párový t-test: Statistika – Základní statistiky/tabulky – t-test, závislé vzorky – OK, Proměnné – 1. seznam proměnných benzín A, benzín B – OK – Souhrn. Dostaneme tabulku

Proměnná	t-test pro závislé vzorky (priklad453) Označ. rozdíly jsou významné na hlad. p < ,05000							
	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
benzín A	18,10000	1,028483						
benzín A	18,10000	1,028483	10	0,000000	0,000000	0,00000	9	1,000000
benzín A	18,10000	1,028483						
benzín B	18,56000	1,207569	10	-0,460000	0,183787	-7,91484	9	0,000024
benzín B	18,56000	1,207569						
benzín A	18,10000	1,028483	10	0,460000	0,183787	7,91484	9	0,000024
benzín B	18,56000	1,207569						
benzín B	18,56000	1,207569	10	0,000000	0,000000	0,00000	9	1,000000

Vidíme, že testová statistika se realizovala hodnotou $-7{,}91484$, počet stupňů volnosti $= 9$, odpovídající p-hodnota $= 0{,}000024 \leq 0{,}05$, tedy nulovou hypotézu zamítáme na hladině významnosti 0,05.

Shrnutí kapitoly

V praxi se často setkáváme s náhodným výběrem z normálního rozložení. Toto rozložení je charakterizováno střední hodnotou μ a rozptylem σ^2. Při řešení úloh o těchto dvou parametrech používáme čtyři pivotové statistiky, které jsou odvozeny z výběrového průměru M a výběrového rozptylu S^2. Jsou zavedeny ve 4.2. Pro výpočet mezí $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro μ či pro σ^2 slouží vzorce uvedené ve 4.3.1. Meze lze počítat též pomocí systému STATISTICA, jak je uvedeno v příkladu 4.3.2.

Testování hypotéz o střední hodnotě a rozptylu je popsáno ve 4.4 včetně způsobu, jak při těchto testech využít systém STATISTICA. Jedná se o *jednovýběrový z-test*, *jednovýběrový t-test* a *test o rozptylu*. V situaci, kdy máme k dispozici jeden náhodný výběr z dvourozměrného normálního rozložení a posuzujeme rozdílnost středních hodnot obou náhodných veličin, použijeme *párový t-test* popsaný ve 4.5.

Při ověřování předpokladu normality se opíráme o diagnostické grafy či o testy normality dat popsané ve 3. kapitole.

Kontrolní otázky

1. Jaké pivotové statistiky odvozené z výběrového průměru M a výběrového rozptylu S^2 používáme při řešení úloh o střední hodnotě μ a rozptylu σ^2 normálního rozložení?
2. Jak vypadají meze $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro neznámou střední hodnotu μ, když rozptyl σ^2 není znám?
3. Jaké testy o parametrech normálního rozložení znáte?
4. V jaké situaci a za jakých podmínek použijete jednovýběrový t-test?
5. V jaké situaci a za jakých podmínek použijete párový t-test?

Autokorekční test

1. Máme-li sestrojit interval spolehlivosti pro střední hodnotu normálního rozložení a neznáme rozptyl, použijeme pivotovou statistiku, která se řídí

a) standardizovaným normálním rozložením,
b) Pearsonovým chí-kvadrát rozložením,
c) Studentovým rozložením.

2. Která z následujících tvrzení jsou pravdivá?

a) $100(1-\alpha)\%$ empirický interval spolehlivosti pro neznámou směrodatnou odchylku normálního rozložení při neznámé střední hodnotě má meze

$$\left(\sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}},\sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)}}\right).$$

b) $100(1-\alpha)\%$ empirický interval spolehlivosti pro neznámou střední hodnotu normálního rozložení při známém rozptylu má meze

$$\left(m-\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, m+\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right).$$

c) $100(1-\alpha)\%$ empirický interval spolehlivosti pro neznámý rozptyl normálního rozložení při známé střední hodnotě má meze

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)},\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right).$$

3. Jednovýběrový t-test slouží k testování hypotézy

a) o střední hodnotě normálního rozložení při neznámém rozptylu,
b) o směrodatné odchylce normálního rozložení při neznámé střední hodnotě,
c) o střední hodnotě normálního rozložení při známém rozptylu.

4. Nechť je dán náhodný výběr rozsahu n z rozložení $N(\mu,\sigma^2)$, kde rozptyl σ^2 známe. Jak musíme změnit rozsah náhodného výběru, chceme-li, aby šířka $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro neznámou střední hodnotu μ klesla na polovinu?

a) Rozsah zvětšíme $2\times$.
b) Rozsah zvětšíme $4\times$.
c) Rozsah zmenšíme na polovinu.

5. Nechť je dán náhodný výběr rozsahu n z rozložení $N(\mu,\sigma^2)$, kde parametry μ,σ^2 neznáme. Dále je dána reálná konstanta c. Testujeme nulovou hypotézu H_0: $\sigma^2=c$ proti levostranné alternativě H_1: $\sigma^2<c$. Kritický obor pro tento test má tvar

a) $W=\left(0,\chi^2_{1-\alpha}(n-1)\right)$
b) $W=\left(0,\chi^2_{\alpha}(n-1)\right)$
c) $W=\left(\chi^2_{1-\alpha}(n-1),\infty\right)$

Správné odpovědi: 1c) 2b) 3a) 4b) 5b)

Příklady

1. Lze předpokládat, že hmotnost pomerančů dodávaných do obchodní sítě se řídí normálním rozložením se střední hodnotou 170 g a směrodatnou odchylkou 12 g. Jaká je pravděpodobnost, že celková hmotnost devíti náhodně vybraných pomerančů balených do síťky překročí 1,5 kg?

Výsledek: Hledaná pravděpodobnost je 0,797.

2. Počet bodů v testu inteligence je náhodná veličina, která se řídí rozložením $N(100, 225)$. Jaká je pravděpodobnost, že průměr v náhodně vybrané skupině 20 osob bude větší než 105 bodů?

Výsledek: Hledaná pravděpodobnost je 0,06811.

3. Při provádění určitého pokusu bylo zapotřebí udržovat v laboratoři konstantní teplotu 26,5 °C. Teplota byla v jednom pracovním týdnu 46× namátkově kontrolována v různých denních a nočních hodinách. Z výsledků měření byly vypočteny realizace výběrového průměru a výběrové směrodatné odchylky: $m = 26{,}33$ °C, $s = 0{,}748$ °C. Za předpokladu, že výsledky měření teploty se řídí rozložením $N(\mu, \sigma^2)$, vypočtěte 95% empirický interval spolehlivosti

a) pro střední hodnotu μ

b) pro směrodatnou odchylku σ.

Výsledek:

ad a) Dosazením do vzorce 4.3.1 (b) dostaneme $26{,}11\ ^\circ\mathrm{C} < \mu < 26{,}55\ ^\circ\mathrm{C}$ s pravděpodobností aspoň 0,95.

ad b) Dosazením do vzorce 4.3.1 (d), kde meze odmocníme, dostaneme $0{,}62\ ^\circ\mathrm{C} < \sigma < 0{,}94\ ^\circ\mathrm{C}$ s pravděpodobností aspoň 0,95.

4. U 25 náhodně vybraných dvoulitrových lahví s nealkoholickým nápojem byl zjištěn přesný objem nápoje. Výběrový průměr činil $m = 1{,}99$ l a výběrová směrodatná odchylka $s = 0{,}1$ l. Předpokládejme, že objem nápoje v láhvi je náhodná veličina s normálním rozložením.

a) Na hladině významnosti 0,05 ověřte tvrzení výrobce, že zákazník není znevýhodněn.

b) Na hladině významnosti 0,05 ověřte tvrzení výrobce, že směrodatná odchylka je 0,08 l.

Výsledek:

ad a) Testujeme hypotézu H_0: $\mu = 2$ proti levostranné alternativě H_1: $\mu < 2$ pomocí jednovýběrového t-testu (viz 4.4.1 (b)). Jelikož hodnota testového kritéria $-0{,}5$ neleží v kritickém oboru $(\infty; -2{,}064\rangle$, nezamítáme nulovou hypotézu na hladině významnosti 0,05.

ad b) Testujeme hypotézu H_0: $\sigma = 0{,}08$ proti oboustranné alternativě H_1: $\sigma \neq 0{,}08$ pomocí testu o rozptylu (viz 4.4.1 (c)). Jelikož hodnota testového kritéria 37,5 neleží v kritickém oboru $(0; 12{,}4\rangle \cup \langle 39{,}4; \infty)$, nejsme oprávněni na hladině významnosti 0,05 zamítnout tvrzení výrobce.

5. Bylo vybráno šest nových vozů téže značky a po určité době bylo zjištěno, o kolik mm se sjely jejich levé a pravé přední pneumatiky. Výsledky:

$$(1{,}8; 1{,}5), \quad (1{,}0; 1{,}1), \quad (2{,}2; 2{,}0), \quad (0{,}9; 1{,}1), \quad (1{,}5; 1{,}4), \quad (1{,}6; 1{,}4).$$

Za předpokladu, že uvedené dvojice tvoří náhodný výběr z dvourozměrného normálního rozložení s vektorem středních hodnot (μ_1, μ_2), testujte na hladině významnosti 0,05 hypotézu, že obě pneumatiky se sjíždějí stejně rychle.

Výsledek: Vzhled dvourozměrného tečkového diagramu není v rozporu s předpokladem o dvourozměrném normálním rozložení. Přejdeme k rozdílovému náhodnému výběru a testujeme nulovou hypotézu H_0: $\mu = 0$ proti oboustranné alternativě H_1: $\mu \neq 0$ pomocí párového t-testu. Hodnota testového kritéria $= 1{,}0512$, počet stupňů volnosti $= 5$. Protože odpovídající p-hodnota $= 0{,}3411$ je větší než hladina významnosti 0,05, nelze na hladině významnosti 0,05 zamítnout nulovou hypotézu. Ke stejnému rozhodnutí dospějeme, pokud stanovíme kritický obor: $W = (-\infty; -2{,}571) \cup (2{,}571; \infty)$. Testové kritérium se nerealizuje v kritickém oboru, tedy nelze na hladině významnosti 0,05 zamítnout nulovou hypotézu.

6. Uměle připravený vzorek minerálu obsahoval 10 % křemene a byl 12krát proměřen. Výsledky měření byly:

8,7 10,2 10,07 9,75 9,65 10,37 10,14 10,5 9,48 11,22 9,49 9,86

Na hladině významnosti 0,05 testujte hypotézu, že obsah křemene byl stanoven správně.

Výsledek: K-S test ani S-W test nezamítají na hladině významnosti 0,05 normalitu dat. Testujeme nulovou hypotézu H_0: $\mu = 10$ proti oboustranné alternativě H_1: $\mu \neq 10$. Úloha vede na jednovýběrový t-test. Realizace testového kritéria $= -0{,}262$, počet stupňů volnosti $= 9$. Protože odpovídající p-hodnota $= 0{,}7981$ je větší než hladina významnosti 0,05, nelze na hladině významnosti 0,05 zamítnout nulovou hypotézu.

- Motivace
- Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů
- Intervaly spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$, σ_1^2/σ_2^2
- Testování hypotéz o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2/σ_2^2

5. Úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

5. Úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Cíl kapitoly

Po prostudování této kapitoly budete

- znát vlastnosti pivotových statistik odvozených ze dvou nezávislých náhodných výběrů z normálních rozložení a budete je umět použít pro řešení konkrétních úloh
- umět sestrojit intervaly spolehlivosti pro rozdíl středních hodnot a podíl rozptylů dvou normálních rozložení
- provádět testy hypotéz o rozdílu středních hodnot a podílu rozptylů dvou normálních rozložení

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 5 hodin studia.

5.1 Motivace

V tomto případě je naším úkolem porovnat střední hodnoty či rozptyly dvou normálních rozložení na základě znalosti dvou nezávislých náhodných výběrů pořízených z těchto rozložení. Zpravidla konstruujeme intervaly spolehlivosti pro rozdíl středních hodnot nebo podíl rozptylů respektive hodnotíme shodu středních hodnot pomocí dvouvýběrového t-testu či dvouvýběrového z-testu a shodu rozptylů pomocí F-testu.

5.2 Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů

Nechť $X_{11},\dots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1,\sigma_1^2)$ a $X_{21},\dots,X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2,\sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Označme M_1, M_2 výběrové průměry a S_1^2, S_2^2 výběrové rozptyly. Pak platí:

a) Statistiky $M_1 - M_2$ a $S_*^2 = \dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ jsou stochasticky nezávislé.

b) $M_1 - M_2 \sim N\left(\mu_1-\mu_2, \dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$, tedy

$$U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

(Pivotová statistika U slouží k řešení úloh o $\mu_1-\mu_2$, když σ_1^2 a σ_2^2 známe.)

c) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $K = \dfrac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$.

(Pivotová statistika K slouží k řešení úloh o neznámém společném rozptylu σ^2.)

d) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $T = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$.

(Pivotová statistika T slouží k řešení úloh o $\mu_1 - \mu_2$, když σ_1^2 a σ_2^2 neznáme, ale víme, že jsou shodné.)

e) $F = \dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$.

(Pivotová statistika F slouží k řešení úloh o σ_1^2/σ_2^2.)

5.2.1 Příklad

Nechť jsou dány dva nezávislé náhodné výběry, první pochází z rozložení $N(2, 3/2)$ a má rozsah 10, druhý pochází z rozložení $N(3, 4)$ a má rozsah 5. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude menší než výběrový průměr 2. výběru?

Řešení:

Statistika $M_1 - M_2$ se podle 5.2 (b) řídí rozložením $N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$, kde $\mu_1 - \mu_2 = 2 - 3 = -1$, $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} = \dfrac{1{,}5}{10} + \dfrac{4}{5} = 0{,}95$, tj. $M_1 - M_2 \sim N(-1; 0{,}95)$

Tedy statistika $U = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{M_1 - M_2 + 1}{\sqrt{0{,}95}}$

Dostáváme

$$P(M_1 < M_2) = P(M_1 - M_2 < 0) = P\left(U < \frac{0+1}{\sqrt{0{,}95}}\right) = \Phi(1{,}026) = 0{,}8475.$$

S pravděpodobností přibližně 84,8 % je výběrový průměr 1. výběru menší než výběrový průměr 2. výběru.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Dvakrát klikneme na název proměnné Prom1. Do Dlouhého jména této proměnné napíšeme
= INormal(0;-1;sqrt(0,95)).

V proměnné Prom1 se objeví hodnota 0,847549.

5.3 Intervaly spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$, σ_1^2/σ_2^2

Budeme zabývat speciálními případy, kdy za parametrickou funkci $h(\vartheta)$ považujeme rozdíl středních hodnot $\mu_1 - \mu_2$ nebo podíl rozptylů σ_1^2/σ_2^2 dvou normálních rozložení. Při konstrukci intervalu spolehlivosti pro rozdíl středních hodnot buď rozptyly známe nebo neznáme a víme, že jsou shodné či nikoliv. Shodu rozptylů ověřujeme pomocí F-testu. Uvedeme jen přehled vzorců pro meze $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro parametrické funkce $\mu_1-\mu_2$, σ_1^2/σ_2^2.

5.3.1 Přehled vzorců

a) Interval spolehlivosti pro $\mu_1-\mu_2$, když σ_1^2, σ_2^2 známe

(využití pivotové statistiky $U = \dfrac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$)

Oboustranný: $(d,h) =$
$$= \left(m_1-m_2-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{1-\alpha/2}, m_1-m_2+\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{1-\alpha/2}\right)$$

Levostranný: $(d,\infty) = \left(m_1-m_2-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{1-\alpha}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m_1-m_2+\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{1-\alpha}\right)$

b) Interval spolehlivosti pro $\mu_1-\mu_2$, když σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné

(využití pivotové statistiky $T = \dfrac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$).

Oboustranný: $(d,h) = \Big(m_1-m_2-s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{1-\alpha/2}(n_1+n_2-2),$
$$m_1-m_2+s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{1-\alpha/2}(n_1+n_2-2)\Big)$$

Levostranný: $(d,\infty) = \left(m_1-m_2-s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{1-\alpha}(n_1+n_2-2), \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m_1-m_2+s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{1-\alpha}(n_1+n_2-2)\right)$

c) Interval spolehlivosti pro společný neznámý rozptyl σ^2

(využití pivotové statistiky $K = \dfrac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$)

Oboustranný: $(d,h) = \left(\dfrac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha/2}(n_1+n_2-2)}, \dfrac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha/2}(n_1+n_2-2)}\right)$

Levostranný: $(d,\infty) = \left(\dfrac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha}(n_1+n_2-2)}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, \dfrac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha}(n_1+n_2-2)}\right)$

d) Interval spolehlivosti pro podíl rozptylů $\dfrac{\sigma_1^2}{\sigma_2^2}$

(využití pivotové statistiky $F = \dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$)

Oboustranný: $(d,h) = \left(\dfrac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)}, \dfrac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)}\right)$

$$\text{Levostranný: } (d, \infty) = \left(\frac{s_1^2/s_2^2}{F_{1-\alpha}(n_1-1, n_2-1)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left(-\infty, \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1, n_2-1)} \right)$$

Upozornění: Není-li v 5.3.1 (b) splněn předpoklad o shodě rozptylů, lze sestrojit aspoň přibližný $100(1-\alpha)\%$ interval spolehlivosti pro $\mu_1 - \mu_2$. V tomto případě má statistika T přibližně rozložení $t(\nu)$, kde počet stupňů volnosti

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1-1} + \dfrac{(s_2^2/n_2)^2}{n_2-1}}.$$

Není-li ν celé číslo, použijeme v tabulkách kvantilů Studentova rozložení lineární interpolaci.

5.3.2 Příklad

Ve dvou nádržích se zkoumal obsah chlóru (v g/l). Z první nádrže bylo odebráno 25 vzorků, z druhé nádrže 10 vzorků. Byly vypočteny realizace výběrových průměrů a rozptylů: $m_1 = 34{,}48$, $m_2 = 35{,}59$, $s_1^2 = 1{,}7482$, $s_2^2 = 1{,}7121$. Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení:

Úloha vede na vzorec 5.3.1 (b). Vypočteme vážený průměr výběrových rozptylů a najdeme odpovídající kvantily Studentova rozložení:

$$s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{24 \cdot 1{,}7482 + 9 \cdot 1{,}7121}{33} = 1{,}7384,$$
$$t_{0{,}975}(33) = 2{,}035.$$

Dosadíme do vzorců pro dolní a horní mez intervalu spolehlivosti:

$$d = m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2) =$$
$$= 34{,}48 - 35{,}59 - \sqrt{1{,}7384} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -2{,}114$$
$$h = m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2) =$$
$$= 34{,}48 - 35{,}59 + \sqrt{1{,}7384} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -0{,}106$$

Zjistili jsme, že $-2{,}114$ g/l $< \mu_1 - \mu_2 < -0{,}106$ g/l s pravděpodobností aspoň 0,95.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor o jednom případu a dvou proměnných, které nazveme *dm* a *hm*. Do Dlouhého jména proměnné *dm* napíšeme =34,48-35,59-sqrt((24*

1,7482+9*1,7121)/33)*sqrt(1/25+1/10)*VStudent(0,975;33). Dostaneme výsledek $-2{,}11368$. (Přitom funkce VStudent(x;sv) poskytuje x% kvantil Studentova rozložení s počtem stupňů volnosti sv.)

Do Dlouhého jména proměnné *hm* napíšeme =34,48-35,59+sqrt((24*1,7482+9*1,7121)/33)*sqrt(1/25+1/10)*VStudent(0,975;33). Dostaneme výsledek $-0{,}10632$.

5.3.3 Příklad

V příkladu 5.3.2 nyní předpokládáme, že dané dva náhodné výběry pocházejí z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

Řešení:

Úloha vede na vzorec 5.3.1 (d).

$$d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0{,}975}(24{,}9)} = \frac{1{,}7482/1{,}7121}{3{,}6142} = 0{,}28$$

$$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0{,}025}(24{,}9)} = \frac{1{,}7482/1{,}7121}{1/2{,}7027} = 2{,}76$$

Dostáváme, že $0{,}28 < \dfrac{\sigma_1^2}{\sigma_2^2} < 2{,}76$ s pravděpodobností aspoň 0,95.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor o jednom případu a dvou proměnných, které nazveme *dm* a *hm*. Do Dlouhého jména proměnné *dm* napíšeme =(1,7482/1,7121)/VF(0,975;24;9). Dostaneme výsledek 0,282521. (Přitom funkce VF(x;ný;omega) poskytuje x% kvantil Fisherova–Snedecorova rozložení s počtem stupňů volnosti čitatele ný a jmenovatele omega.)

Do Dlouhého jména proměnné *hm* napíšeme =(1,7482/1,7121)/VF(0,025;24;9). Dostaneme výsledek 2,759698.

5.4 Testování hypotéz o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2/σ_2^2

5.4.1 Přehled testů

a) Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$, $n_2 \geq 2$ a σ_1^2, σ_2^2 známe. Nechť c je konstanta. Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový z-test.

b) Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta. Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový t-test.

c) Nechť $X_{11},\dots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1,\sigma_1^2)$ a $X_{21},\dots,X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2,\sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Test H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$ se nazývá F-test.

5.4.2 Provedení testů o parametrických funkcích $\mu_1-\mu_2$, σ_1^2/σ_2^2 pomocí kritického oboru

a) Provedení dvouvýběrového z-testu

Hypotézu H_0: $\mu_1-\mu_2 = c$ proti H_1: $\mu_1-\mu_2 \neq c$ (resp. H_1: $\mu_1-\mu_2 < c$, resp. H_1: $\mu_1-\mu_2 > c$) zamítáme na hladině významnosti α, jestliže

$$\left|\frac{m_1-m_2-c}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| \geq u_{1-\alpha/2}$$

$$\left(\text{resp.}\quad \frac{m_1-m_2-c}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq -u_{1-\alpha},\quad \text{resp.}\quad \frac{m_1-m_2-c}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geq u_{1-\alpha}\right).$$

b) Provedení dvouvýběrového t-testu

Hypotézu H_0: $\mu_1-\mu_2 = c$ proti H_1: $\mu_1-\mu_2 \neq c$ (resp. H_1: $\mu_1-\mu_2 < c$ resp. H_1: $\mu_1-\mu_2 > c$) zamítáme na hladině významnosti α, jestliže

$$\left|\frac{m_1-m_2-c}{s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| \geq t_{1-\alpha/2}(n_1+n_2-2)$$

$$\left(\text{resp.}\quad \frac{m_1-m_2-c}{s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -t_{1-\alpha}(n_1+n_2-2),\right.$$

$$\left.\text{resp.}\quad \frac{m_1-m_2-c}{s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq t_{1-\alpha}(n_1+n_2-2)\right).$$

c) Provedení F-testu

Hypotézu H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$ (resp. H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} < 1$, resp. H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} > 1$) zamítáme na hladině významnosti α, jestliže

$$\frac{s_1^2}{s_2^2} \leq F_{\alpha/2}(n_1+n_2-2) \quad \text{nebo} \quad \frac{s_1^2}{s_2^2} \geq F_{1-\alpha/2}(n_1+n_2-2)$$

$$\left(\text{resp.}\quad \frac{s_1^2}{s_2^2} \leq F_{\alpha}(n_1+n_2-2),\quad \text{resp.}\quad \frac{s_1^2}{s_2^2} \geq F_{1-\alpha}(n_1+n_2-2)\right).$$

Podobně jako v kapitole 4 musíme ověřit normalitu dat. Pokud výběry menších rozsahů (pod 30) vykazují výraznější odchylky od normality, doporučuje se místo dvouvýběrového t-testu použít neparametrický dvouvýběrový Wilcoxonů test (viz kapitola 7).

Před provedením dvouvýběrového t-testu bychom se měli F-testem přesvědčit o shodě rozptylů. Zamítne-li F-test na dané hladině významnosti hypotézu o shodě rozptylů, musíme pro testování hypotézy o shodě středních hodnot použít speciální variantu dvouvýběrového t-testu, tzv. dvouvýběrový t-test se separovanými odhady rozptylů.

Musíme si být vědomi rozdílu mezi dvouvýběrovým t-testem a párovým t-testem. Dvouvýběrový t-test je založen na předpokladu nezávislosti daných dvou výběrů. Pokud v situaci, která vede na párový test, použijeme dvouvýběrový t-test, můžeme dostat nepravdivé výsledky. Naopak, mají-li dva nezávislé výběry stejný rozsah a my použijeme párový t-test místo dvouvýběrového t-testu, nedopustíme se hrubé chyby, pouze méně efektivně využijeme informaci obsaženou v datech.

5.4.3 Příklad

V restauraci „U bílého koníčka“ měřili ve 20 případech čas obsluhy zákazníka. Výsledky v minutách: $6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5$. V restauraci „Zlatý lev“ bylo dané pozorování uskutečněno v 15 případech s těmito výsledky: $9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8, 7$.

Za předpokladu, že uvedené hodnoty pocházejí ze dvou normálních rozložení, na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty doby obsluhy jsou v obou restauracích stejné.

Řešení:

Na hladině významnosti 0,05 testujeme nulovou hypotézu $H_0\colon \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_1\colon \mu_1 - \mu_2 \neq 0$. Je to úloha na dvouvýběrový t-test. Před provedením tohoto testu je však nutné pomocí F-testu testovat shodu rozptylů. Na hladině významnosti 0,05 tedy testujeme $H_0\colon \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti $H_1\colon \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$. Podle 5.4.2 (c) nulovou hypotézu zamítáme na hladině významnosti α, jestliže

$$\frac{s_1^2}{s_2^2} \leq F_{\alpha/2}(n_1-1, n_2-1) \quad \text{nebo} \quad \frac{s_1^2}{s_2^2} \geq F_{1-\alpha/2}(n_1-1, n_2-1).$$

Vypočteme $m_1 = 8{,}25$, $m_2 = 8{,}13$, $s_1^2 = 6{,}307$, $s_2^2 = 9{,}41$.

V našem případě $\dfrac{s_1^2}{s_2^2} = \dfrac{6{,}307}{9{,}41} = 0{,}6702$. V tabulkách najdeme

$$F_{\alpha/2}(n_1-1, n_2-1) = F_{0{,}025}(19, 14) = \frac{1}{F_{0{,}975}(14, 19)} = \frac{1}{2{,}6469} = 0{,}3778,$$
$$F_{1-\alpha/2}(n_1-1, n_2-1) = F_{0{,}975}(19, 14) = 2{,}8607.$$

Protože 0,6702 nepatří do kritického oboru $\langle 0; 0{,}3778\rangle \cup \langle 2{,}8607; \infty)$, hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

Nyní se vrátíme k dvouvýběrovému t-testu.

a) Testování pomocí kritického oboru: Podle 5.4.2 (b) nulovou hypotézu zamítáme na hladině významnosti α, když absolutní hodnota realizace testové statistiky

$$|t_0| = \frac{|m_1 - m_2 - c|}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{1-\alpha/2}(n_1 + n_2 - 2),$$

kde

$$s_*^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{19 \cdot 8{,}13 + 14 \cdot 9{,}41}{33} = 7{,}623.$$

V tabulkách najdeme $t_{0{,}975}(33) = 1{,}96$. Dosadíme do vzorce pro výpočet absolutní hodnoty realizace testové statistiky

$$\frac{|m_1 - m_2 - c|}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|8{,}25 - 8{,}13|}{\sqrt{7{,}623} \cdot \sqrt{\frac{1}{20} + \frac{1}{15}}} = 0{,}124$$

Protože $0{,}124 < 1{,}96$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

b) Testování pomocí intervalu spolehlivosti: Podle 5.3.1 (b) máme

$$(d,h) = \Bigg(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2), \\ m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2)\Bigg)$$

V tabulkách najdeme $t_{0{,}975}(33) = 1{,}96$.

$$d = 8{,}25 - 8{,}13 - \sqrt{7{,}623} \cdot \sqrt{\frac{1}{20} + \frac{1}{15}} \cdot 1{,}96 = -1{,}73,$$
$$h = 8{,}25 - 8{,}13 + \sqrt{7{,}623} \cdot \sqrt{\frac{1}{20} + \frac{1}{15}} \cdot 1{,}96 = 1{,}97.$$

Protože $0 \in (-1{,}73; 1{,}97)$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

c) Testování pomocí p-hodnoty: Podle 1.4.6 (c) dostáváme

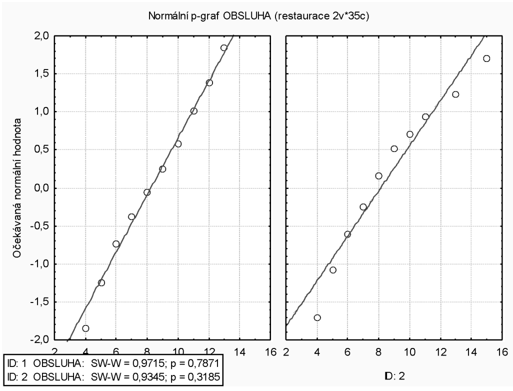
$$\begin{aligned} p &= 2\min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\} = \\ &= 2\min\{P(T_0 \leq 0{,}124), P(T_0 \geq 0{,}124)\} = \\ &= 2\min\{\Phi(0{,}124), 1 - \Phi(0{,}124)\}, \end{aligned}$$

kde $\Phi(x)$ je distribuční funkce Studentova rozložení s počtem stupňů volnosti 33. Pomocí statistického software získáme $\Phi(0{,}124) = 0{,}549$, tedy $p = 2 \cdot (1 - 0{,}549) = 0{,}902$. Protože $0{,}902 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

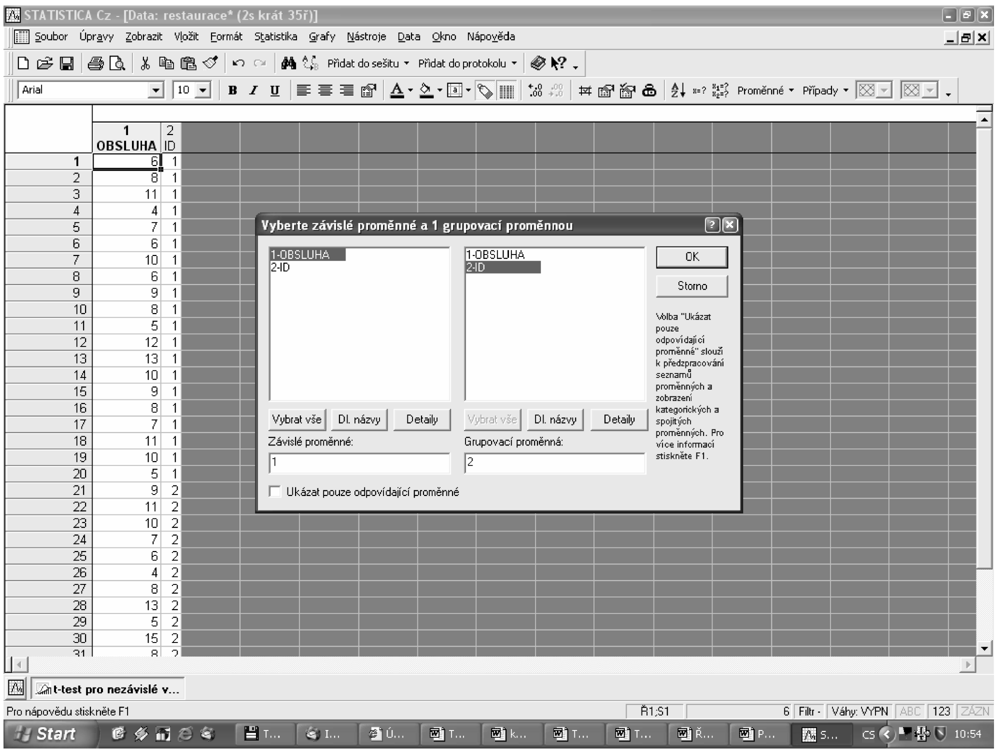
Otevřeme nový datový soubor o dvou proměnných a 35 případech. První proměnnou nazveme *OBSLUHA*, druhou *ID*. Do proměnné *OBSLUHA* napíšeme nejprve doby obsluhy v první restauraci a poté doby obsluhy ve druhé restauraci. Do proměnné *ID*, která slouží k rozlišení první a druhé restaurace, napíšeme 20krát jedničku a 15krát dvojku.

Pomocí NP-grafu a S-W testu ověříme normalitu dat v obou skupinách. Grafy – 2D Grafy – Normální pravděpodobnostní grafy – zaškrtneme S-W test, Proměnné *OBSLUHA*, OK, Kategorizovaný – Kategorie X, zaškrtneme Zapnuto, Změnit proměnnou – *ID*, OK. Dostaneme graf



V obou případech se tečky odchylují od přímky jenom málo. Rovněž p-hodnoty S-W testu jsou v obou případech větší než 0,05, tedy hypotézy o normalitě nezamítáme na hladině významnosti 0,05.

Nyní provedeme dvouvýběrový t-test současně s testem o shodě rozptylů: Statistika – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné – Závislé proměnné *OBSLUHA*, Grupovací proměnná *ID* – OK.



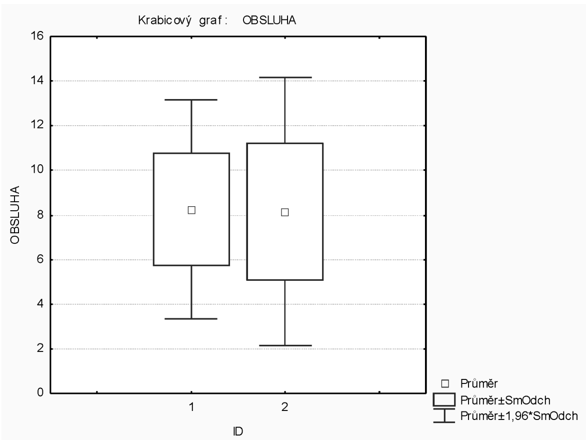
Po kliknutí na tlačítko Souhrn dostaneme tabulku

Proměnná	t-testy; grupováno: ID (restaurace) Skup. 1: 1 Skup. 2: 2 Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr rozptyly	p rozptyly
OBSLUHA	8,250000	8,133333	0,123730	33	0,902279	20	15	2,510504	3,067495	1,492952	0,410440

Vidíme, že testová statistika pro test shody rozptylů se realizuje hodnotou 1,492952 (je to převrácená hodnota k číslu 0,6702, které jsme vypočítali při ručním postupu), odpovídající *p*-hodnota je 0,41044, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů. (Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Dále z tabulky plyne, že testová statistika pro test shody středních hodnot se realizuje hodnotou 0,12373, počet stupňů volnosti je 33, odpovídající *p*-hodnota 0,902279, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se neprokázal rozdíl ve středních hodnotách dob obsluhy v restauracích „U bílého koníčka“ a „Zlatý lev“.

Tabulku ještě doplníme krabicovými diagramy. Na záložce Detaily zaškrtneme krabicový graf a vybereme volbu Průměr/SmOdch/1,96*SmOdch.



Z grafu je vidět, že průměrná doba obsluhy v první restauraci je nepatrně delší a má menší variabilitu než ve druhé restauraci. Extrémní ani odlehlé hodnoty se zde nevyskytují.

Shrnutí kapitoly

V této kapitole jsme porovnávali střední hodnoty či rozptyly dvou normálních rozložení na základě znalosti dvou nezávislých náhodných výběrů pořízených z těchto rozložení. Vzorce pro výpočet mezí $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$ či $\frac{\sigma_1^2}{\sigma_2^2}$ jsou uvedené v 5.3.1. Meze lze počítat též pomocí systému STATISTICA, jak je uvedeno v příkladech 5.3.2 a 5.3.3.

Testování hypotéz o rozdílu středních hodnot a podílu rozptylu je popsáno ve 5.4 včetně způsobu, jak při těchto testech využít systém STATISTICA. Jedná se o *dvouvýběrový z-test*, *dvouvýběrový t-test* a *F-test*. Provedení dvouvýběrového t-testu a F-testu v systému STATISTICA je popsánu v příkladu 5.4.3.

Kontrolní otázky

1. Které pivotové statistiky používáme při řešení úloh o rozdílu středních hodnot a podílu rozptylů dvou normálních rozložení?
2. Jaké meze má $100(1-\alpha)\%$ empirický interval spolehlivosti pro podíl směrodatných odchylek dvou normálních rozložení?
3. V čem spočívá rozdíl mezi dvouvýběrovým z-testem a dvouvýběrovým t-testem?
4. V jakých situacích používáme dvouvýběrový t-test a v jakých a párový t-test?
5. K čemu slouží F-test?

Autokorekční test

1. Na základě znalosti dvou nezávislých náhodných výběrů o rozsazích n_1 a n_2 ze dvou normálních rozložení se shodným rozptylem máme sestrojit interval spolehlivosti pro rozdíl středních hodnot. Použijeme pivotovou statistiku, která se řídí
   a) standardizovaným normálním rozložením
   b) Fisherovým–Snedecorovým rozložením $F(n_1-1, n_2-1)$
   c) Studentovým rozložením $t(n_1+n_2-1)$
2. Na základě znalosti dvou nezávislých náhodných výběrů o rozsazích n_1 a n_2 ze dvou normálních rozložení s neznámými středními hodnotami máme sestrojit interval spolehlivosti pro podíl rozptylů. Použijeme pivotovou statistiku, která se řídí
   a) standardizovaným normálním rozložením
   b) Fisherovým - Snedecorovým rozložením $F(n_1-1, n_2-1)$
   c) Studentovým rozložením $t(n_1+n_2-1)$
3. Testujeme-li hypotézu o shodě středních hodnot dvou normálních rozložení se shodným, ale neznámým rozptylem na základě znalosti dvou nezávislých náhodných výběrů, použijeme
   a) dvouvýběrový t-test
   b) dvouvýběrový z-test
   c) F-test
4. Testujeme-li hypotézu o shodě rozptylů dvou normálních rozložení na základě znalosti dvou nezávislých náhodných výběrů, použijeme
   a) dvouvýběrový t-test
   b) dvouvýběrový z-test
   c) F-test

Správné odpovědi: 1c) 2b) 3a) 4c)

Příklady

1. Bylo vylosováno 11 stejně starých selat téhož plemene. Šesti z nich byla předepsána výkrmná dieta č. 1 a zbylým pěti výkrmná dieta č. 2. Průměrné denní přírůstky v Dg za dobu půl roku jsou následující:

   dieta č. 1: 62, 54, 55, 60, 53, 58
   dieta č. 2: 52, 56, 49, 50, 51.

Zjištěné hodnoty považujeme za realizace dvou nezávislých náhodných výběrů pocházejících z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů a 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Výsledek:

$0{,}1872 < \dfrac{\sigma_1^2}{\sigma_2^2} < 12{,}9541$ s pravděpodobností aspoň 0,95.

$0{,}99$ Dg $< \mu_1 - \mu_2 < 9{,}81$ Dg s pravděpodobností aspoň 0,95.

2. Pro údaje z příkladu 1. testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv na hmotnostní přírůstky selat.

Výsledek: Testujeme hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$

1. způsob – pomocí intervalu spolehlivosti. 95% empirický interval spolehlivosti pro $\mu_1 - \mu_2$ je interval $(0{,}99; 9{,}81)$. Neobsahuje nulu, proto H_0 zamítáme na hladině významnosti 0,05.

2. způsob – pomocí kritického oboru. Protože testové kritérium se realizuje hodnotou 2,771, která patří do kritického oboru $(-\infty; -2{,}2622\rangle \cup \langle 2{,}2622; \infty)$, H_0 zamítáme na hladině významnosti 0,05.

3. Máme k dispozici realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$ o rozsazích $n_1 = 10$, $n_2 = 15$. Výběrové průměry se realizovaly hodnotami $m_1 = 120{,}56$, $m_2 = 124{,}13$, výběrové rozptyly hodnotami $s_1^2 = 9{,}14$, $s_2^2 = 8{,}95$. Lze na základě těchto výsledků zamítnout na hladině významnosti 0,1 nulovou hypotézu H_0: $\mu_1 - \mu_2 = 0$ ve prospěch oboustranné alternativy H_1: $\mu_1 - \mu_2 \neq 0$?

Výsledek: Nulovou hypotézu zamítáme na hladině významnosti 0,1.

4. Výrobce limonád chtěl zjistit, zda změna technologie výroby se projeví v prodeji limonád. Proto sledoval po 14 náhodně vybraných dnů před zavedením nových limonád tržby v určitém regionu a zjistil, že za den utržil v průměru 39 600 Kč se směrodatnou odchylkou 5 060 Kč. Po zavedení nových limonád prověřil stejným způsobem tržby v 11 náhodně vybraných dnech v témž regionu a zjistil průměrný příjem 41 200 Kč se směrodatnou odchylkou 4 310 Kč. Předpokládejte, že tržby za starý typ limonád se řídí rozložením $N(\mu_1, \sigma_1^2)$ a tržby za nový typ limonád se řídí rozložením $N(\mu_2, \sigma_2^2)$.

a) Na hladině významnosti 0,05 testujte hypotézu H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$.

b) Na hladině významnosti 0,05 testujte hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$.

Výsledek:

ad a) Úloha vede na F-test. Vypočteme realizaci testového kritéria:

$$\frac{s_1^2}{s_2^2} = \frac{5\,060^2}{4\,310^2} = 1{,}3783,$$

dále najdeme příslušné kvantily:

$$F_{\alpha/2}(n_1-1, n_2-1) = F_{0{,}025}(13, 10) = 0{,}3077,$$
$$F_{1-\alpha/2}(n_1-1, n_2-1) = F_{0{,}975}(13, 10) = 3{,}5832.$$

Protože testové kritérium $\frac{s_1^2}{s_2^2} = 1{,}3783$ se nerealizuje v kritickém oboru $W = (0; 0{,}3077\rangle \cup \langle 3{,}5832; \infty)$, nelze na hladině významnosti 0,05 zamítnout hypotézu o shodě rozptylů.

ad b) Úloha vede na dvouvýběrový t-test. Protože jsme na hladině významnosti 0,05 nezamítli hypotézu o shodě rozptylů, můžeme rozptyly σ_1^2, σ_2^2 považovat za shodné a za jejich odhad vezmeme vážený průměr výběrových rozptylů

$$s_*^2 = \frac{13 \cdot 5\,060^2 + 10 \cdot 4\,310^2}{23} = 22\,548\,165{,}217.$$

Vypočteme realizaci testového kritéria:

$$\frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{39\,600 - 41\,200}{\sqrt{22\,548\,165{,}217} \cdot \sqrt{\frac{1}{14} + \frac{1}{11}}} = -0{,}8363$$
$$t_{1-\alpha/2}(n_1 + n_2 - 2) = t_{0{,}975}(23) = 2{,}0687$$

Protože testové kritérium $-0{,}8363$ se nerealizuje v kritickém oboru $W = (\infty; -2{,}0687\rangle \cup \langle 2{,}0687; \infty)$, na hladině významnosti 0,05 nelze zamítnout hypotézu o shodě středních hodnot.

- Motivace
- Označení
- Testování hypotézy o shodě středních hodnot
- Testy shody rozptylů
- Metody mnohonásobného porovnávání
- Příklad
- Význam předpokladů v analýze rozptylu

6. Analýza rozptylu jednoduchého třídění

Cíl kapitoly

Po prostudování této kapitoly budete umět

- hodnotit vliv faktoru o $r \geq 3$ úrovních na variabilitu hodnot sledované náhodné veličiny
- sestrojit tabulku analýzy rozptylu
- identifikovat dvojice náhodných výběrů, které se významně liší střední hodnotou
- provést test shody rozptylů
- znázornit rozložení dat v daných $r \geq 3$ náhodných výběrech graficky pomocí kategorizovaných krabicových diagramů

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 5 hodin studia.

6.1 Motivace

Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny X, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina X).

Předpokládáme, že faktor A má $r \geq 3$ úrovní a i-té úrovni odpovídá n_i výsledků $X_{i1}, \ldots, X_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $X_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$. Výsledky lze zapsat do tabulky

faktor A	výsledky
úroveň 1	$X_{11}, \ldots, X_{1n_1}$
úroveň 2	$X_{21}, \ldots, X_{2n_2}$
$\vdots$	$\vdots$
úroveň r	$X_{r1}, \ldots, X_{rn_r}$

Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné proti alternativní hypotéze, která tvrdí, že aspoň jedna dvojice středních hodnot se liší. Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $\binom{r}{2}$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Tento postup však nelze použít, neboť nezaručuje splnění podmínky, že pravděpodobnost chyby 1. druhu je nejvýše α. Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.

6.2 Označení

V analýze rozptylu jednoduchého třídění se používá následující označení.

$n = \sum_{i=1}^{r} n_i$ .... celkový rozsah všech r výběrů

$X_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}$ ...... součet hodnot v i-tém výběru

$M_{i\bullet} = \frac{1}{n_i} X_{i\bullet}$ .... výběrový průměr v i-tém výběru

$X_{\bullet\bullet} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$ ... součet hodnot všech výběrů

$M_{\bullet\bullet} = \frac{1}{n} X_{\bullet\bullet}$ .... celkový průměr všech r výběrů

6.3 Testování hypotézy o shodě středních hodnot

Náhodné veličiny X_{ij} se řídí modelem

$$M0 : X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{pro } i = 1, \dots, r, \quad j = 1, \dots, n_i,$$

přičemž

ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,
μ je společná část střední hodnoty závisle proměnné veličiny,
α_i je efekt faktoru A na úrovni i.

Parametry μ, α_i neznáme. Požadujeme, aby platila tzv. reparametrizační rovnice:

$$\sum_{i=1}^{r} \alpha_i = 0.$$

Zavedeme součty čtverců

$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{\bullet\bullet})^2$ ... celkový součet čtverců (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru), má počet stupňů volnosti $f_T = n - 1$,

$S_A = \sum_{i=1}^{r} (M_{i\bullet} - M_{\bullet\bullet})^2$ ...... skupinový součet čtverců (charakterizuje variabilitu mezi jednotlivými náhodnými výběry), má počet stupňů volnosti $f_A = r - 1$,

$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i\bullet})^2$ ... reziduální součet čtverců (charakterizuje variabilitu uvnitř jednotlivých výběrů), má počet stupňů volnosti $f_E = n - r$.

Lze dokázat, že $S_T = S_A + S_E$. Sčítanec $(M_{i\bullet} - M_{\bullet\bullet})$ představuje bodový odhad efektu α_i.

Kdyby nezáleželo na faktoru A, platila by hypotéza $\alpha_1 = \cdots = \alpha_r = 0$ a dostali bychom model

$$M1 : X_{ij} = \mu + \varepsilon_{ij}.$$

Rozdíl mezi modely $M0$ a $M1$ ověřujeme pomocí testové statistiky $F_A = \dfrac{S_A/f_A}{S_E/f_E}$, která se řídí rozložením $F(r-1, n-r)$, je-li model $M1$ správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti α, když platí:

$$F_A \geq F_{1-\alpha}(r-1, n-r).$$

Výsledky výpočtů zapisujeme do tabulky analýzy rozptylu jednoduchého třídění.

Zdroj variability	součet čtverců	stupně volnosti	podíl	F_A
skupiny	S_A	$f_A = r-1$	S_A/f_A	$F_A = \dfrac{S_A/f_A}{S_E/f_E}$
reziduální	S_E	$f_E = n-r$	S_E/f_E	—
celkový	S_T	$f_T = n-1$	—	—

6.4 Testy shody rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech.

6.4.1 Levenův test

Položme $Z_{ij} = \left|X_{ij} - M_{i\bullet}\right|$. Označíme

$$M_{Zi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}, \quad M_Z = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Z_{ij},$$

$$S_{ZE} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Z_{ij} - M_{Zi})^2, \quad S_{ZA} = \sum_{i=1}^{r} n_i(M_{Zi} - M_Z)^2.$$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \sim F(r-1, n-r).$$

H_0 tedy zamítáme na hladině významnosti α, když $F_{ZA} \geq F_{1-\alpha}(r-1, n-r)$.

6.4.2 Bartlettův test

Platí-li hypotéza o shodě rozptylů, pak statistika

$$B = \frac{1}{C}\left((n-r)\ln S_*^2 - \sum_{i=1}^{r}(n_i - 1)\ln S_i^2\right)$$

má přibližně rozložení $\chi^2(r-1)$, kde

$$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{n-r}\right),$$

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - M_{i.})^2 \text{ je výběrový rozptyl } i\text{-tého výběru,}$$

$$S_*^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i - 1)S_i^2 = \frac{S_E}{n-r} \text{ je vážený průměr výběrových rozptylů.}$$

H_0 zamítáme na přibližné hladině významnosti α, když $B \geq \chi^2_{1-\alpha}(r-1, n-r)$. Bartlettův test lze použít, pokud rozsahy všech výběrů jsou aspoň 7.

6.5 Metody mnohonásobného porovnávání

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α.

6.5.1 Tukeyova metoda

Mají-li všechny výběry týž rozsah p (říkáme, že třídění je vyvážené), použijeme Tukeyovu metodu: rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq q_{1-\alpha}(r, n-r)\frac{S_*}{\sqrt{p}},$$

kde hodnoty $q_{1-\alpha}(r, n-r)$ jsou kvantily studentizovaného rozpětí a najdeme je ve statistických tabulkách.

6.5.2 Scheffého metoda

Nemají-li všechny výběry stejný rozsah, použijeme Scheffého metodu: rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right)F_{1-\alpha}(r-1, n-r)}$$

Pozor, může nastat situace, kdy při zamítnutí H_0 nenajdeme významný rozdíl u žádné dvojice středních hodnot. Pak je významně rozdílná některá složitější kombinace středních hodnot, tzv. kontrast.

6.6 Příklad

U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg):

odrůda	hmotnost				
A	0,9	0,8	0,6	0,9	
B	1,3	1,0	1,3		
C	1,3	1,5	1,6	1,1	1,5
D	1,1	1,2	1,0		

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení:

Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.

$$M_{1.} = 0{,}8, \quad M_{2.} = 1{,}2, \quad M_{3.} = 1{,}4, \quad M_{4.} = 1{,}1, \quad M_{..} = 1{,}14,$$
$$S_E = 0{,}3, \quad S_A = 0{,}816, \quad S_T = 1{,}116, \quad F_A = 9{,}97, \quad F_{0{,}95}(3,11) = 3{,}59.$$

Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

Výsledky zapíšeme do tabulky ANOVA

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	F_A
skupiny	$S_A = 0{,}816$	$f_A = 3$	$S_A/3 = 0{,}272$	$\frac{S_A/f_A}{S_E/f_E} = 9{,}97$
reziduální	$S_E = 0{,}3$	$f_E = 11$	$S_E/11 = 0{,}02727$	—
celkový	$S_T = 1{,}116$	$f_T = 14$	—	—

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

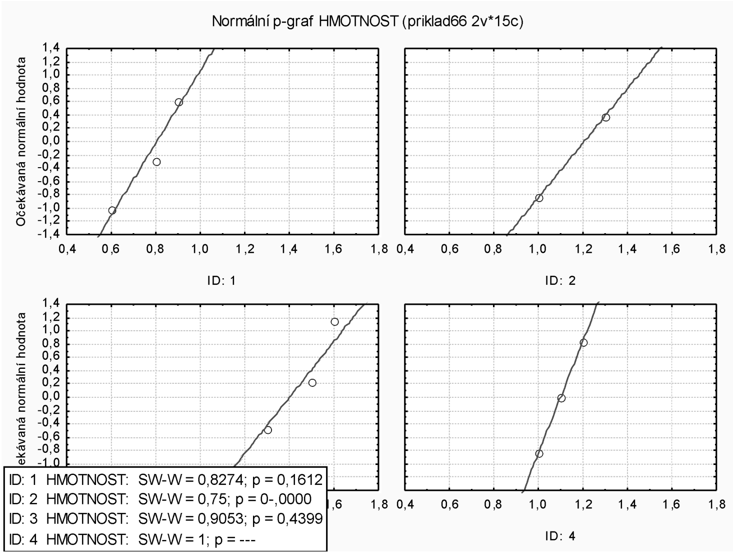
Srovnávané odrůdy	Rozdíly $\lvert M_{k.} - M_{l.}\rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy A a C.

Řešení pomocí systému STATISTICA:

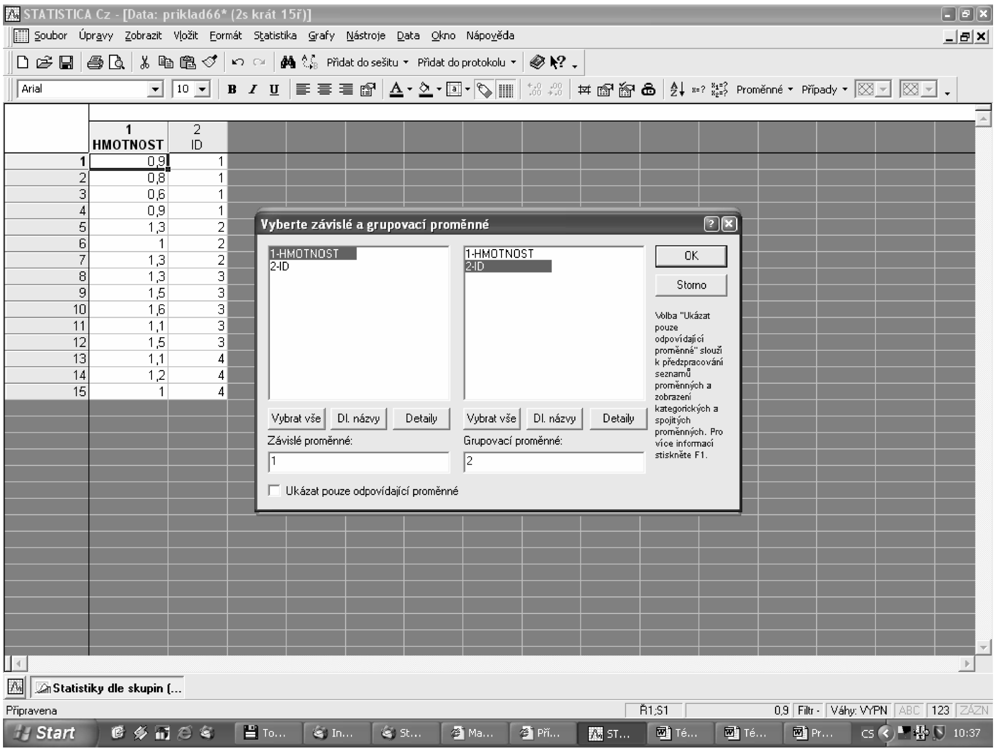
Otevřeme nový datový soubor o dvou proměnných a 15 případech. První proměnnou nazveme *HMOTNOST*, druhou *ID*. Do proměnné *HMOTNOST* zapíšeme zjištěné hmotnosti, do proměnné *ID*, která slouží k rozlišení odrůd, zapíšeme 4krát jedničku, 3krát dvojku, 5krát trojku a 3krát čtyřku.

Pomocí NP-grafu a S-W testu ověříme normalitu dat v daných čtyřech skupinách. Grafy – 2D Grafy – Normální pravděpodobnostní grafy – zaškrtneme S-W test, Proměnné *HMOTNOST*, OK, Kategorizovaný – Kategorie X, zaškrtneme Zapnuto, Změnit proměnnou – *ID*, OK. Dostaneme graf



Vidíme, že ve všech čtyřech případech jsou odchylky teček od přímky jenom malé a data tedy lze považovat za realizace náhodných výběrů z normálních rozložení.

Nyní budeme na hladině významnosti 0,05 testovat hypotézu o shodě rozptylů: Statistika – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK, Proměnné – Závislé proměnné *HMOTNOST*, Grupovací proměnná *ID* – OK.



Na záložce ANOVA & testy vybereme Leveneovy testy. Ve výstupu dostaneme tabulku

Proměnná	Leveneův test homogenity rozpylů (priklad66) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
HMOTNOST	0,018667	3	0,006222	0,065333	11	0,005939	1,047619	0,410027

Testová statistika Levenova testu se realizuje hodnotou 1,047619, počet stupňů volnosti čitatele je 3, jmenovatele 11, odpovídající p-hodnota je 0,410027, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů.

Dále budeme na hladině významnosti 0,05 testovat hypotézu o shodě středních hodnot. Na záložce ANOVA & testy vybereme Analýza rozptylu. Ve výstupu dostaneme tabulku

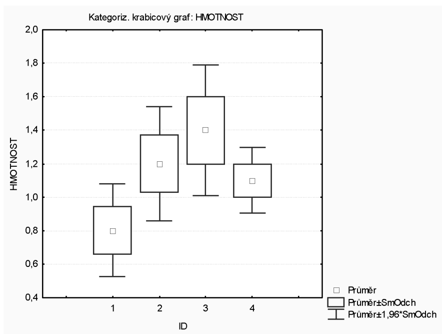
Proměnná	Analýza rozptylu (priklad66) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
HMOTNOST	0,816000	3	0,272000	0,300000	11	0,027273	9,973333	0,001805

Testová statistika F_A se realizuje hodnotou 9,97333, počet stupňů volnosti čitatele je 3, jmenovatele 11, odpovídající p-hodnota je 0,001805, tedy na hladině významnosti 0,05 zamítáme hypotézu o shodě středních hodnot.

Vytvoříme ještě tabulku s hodnotami výběrových průměrů a výběrových směrodatných odchylek tak, že na záložce Popisné statistiky zvolíme Výpočet: Tabulka statistik.

Rozkladová tabulka popisných statistik (priklad66) N=15 (V seznamu záv. prom. nejsou ChD)			
ID	HMOTNOST průměr	HMOTNOST N	HMOTNOST Sm.odch.
1	0,800000	4	0,141421
2	1,200000	3	0,173205
3	1,400000	5	0,200000
4	1,100000	3	0,100000
Vš.skup.	1,140000	15	0,282337

Rovněž sestrojíme krabicové diagramy tak, že na záložce Popisné statistiky zvolíme Kategoriz. krabicový graf. Vybereme typ Průměr/SmOdch/1.96SmOdch.



Vidíme, že nejnižší průměrnou hmotnost má odrůda A, nejnižší variabilitu hmotnosti vykazuje odrůda D.

Abychom zjistili, které dvojice odrůd se liší na hladině významnosti 0,05, na záložce Posthoc vybereme Schefféův test.

ID		Scheffeho test; proměn.:HMOTNOST (priklad66) Označ. rozdíly jsou významné na hlad. p < ,05000			
		{1} M=,80000	{2} M=1,2000	{3} M=1,4000	{4} M=1,1000
1	{1}		0,059165	0,001950	0,190463
2	{2}	0,059165		0,464537	0,905502
3	{3}	0,001950	0,464537		0,163499
4	{4}	0,190463	0,905502	0,163499	

V tabulce jsou uvedeny p-hodnoty pro testování hypotéz o shodě dvojic středních hodnot. Pouze jediná z těchto p-hodnot je menší nebo rovna 0,05, tedy na hladině významnosti 0,05 se liší odrůdy A a C.

6.7 Význam předpokladů v analýze rozptylu

a) Nezávislost jednotlivých náhodných výběrů – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.

b) Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení normality se doporučuje Kruskalův-Wallisův test.

c) Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův-Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

Shrnutí kapitoly

Analýza rozptylu jednoduchého třídění slouží k hodnocení vlivu faktoru o $r \geq 3$ úrovních na variabilitu hodnot sledované náhodné veličiny s normálním rozložením. Test hypotézy o shodě středních hodnot odvodil *R. A. Fisher*. Výpočty spojené s testováním této hypotézy se zapisují do *tabulky ANOVA*. Dojde-li na dané hladině významnosti α k zamítnutí nulové hypotézy, použijeme některou z *metod mnohonásobného porovnávání* (např. *Schefféého* nebo *Tukeyovu metodu*), abychom identifikovali dvojice náhodných výběrů, které přispěly k zamítnutí nulové hypotézy. ANOVA předpokládá shodu rozptylů. Hypotézu o shodě rozptylů můžeme testovat pomocí *Bartlettova testu* nebo *Levenova testu*. Vlastnosti rozložení dat v daných $r \geq 3$ náhodných výběrech graficky znázorňujeme pomocí *kategorizovaných krabicových diagramů* typu průměr–směrodatná odchylka–1,96 směrodatná odchylka.

Kontrolní otázky

1. Jaký problém řeší analýza rozptylu jednoduchého třídění?
2. Jak je definován celkový, skupinový a reziduální součet čtverců a co tyto součty čtverců charakterizují?
3. Popište způsob testování hypotézy o shodě středních hodnot.
4. Jak se testuje hypotéza o shodě rozptylů?

5. Které metody mnohonásobného porovnání se používají v analýze rozptylu jednoduchého třídění?
6. Pojednejte o významu předpokladů v analýze rozptylu jednoduchého třídění.

Autokorekční test

1. Z následujících tří možností vyberte tu správnou:
   Analýza rozptylu jednoduchého třídění slouží k vyhodnocení dat, které vznikly:
   a) při párovém uspořádání pokusu
   b) při blokovém uspořádání pokusu
   c) při mnohovýběrovém uspořádání pokusu.
2. Která z následujících tvrzení jsou pravdivá?
   Analýza rozptylu jednoduchého třídění vyžaduje, aby
   a) jednotlivé náhodné výběry byly stochasticky nezávislé
   b) jednotlivé náhodné výběry pocházely z binomického rozložení
   c) jednotlivé náhodné výběry měly stejný rozptyl.
3. Která z následujících tvrzení jsou pravdivá?
   a) Celkový součet čtverců charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru.
   b) Skupinový součet čtverců charakterizuje variabilitu jednotlivých pozorování kolem skupinových průměrů.
   c) Reziduální součet čtverců charakterizuje variabilitu skupinových průměrů kolem celkového průměru.
4. Nulovou hypotézu o shodě středních hodnot zamítáme na hladině významnosti α, když testové kritérium F se realizuje v kritickém oboru
   a) $W = \big(0, F_{\alpha}(r-1, n-r)\big)$
   b) $W = \big(0, F_{1-\alpha}(r-1, n-r)\big)$
   c) $W = \big(F_{1-\alpha}(r-1, n-r), \infty\big)$
5. Pokud zamítneme hypotézu o shodě středních hodnot a všechny výběry mají stejný rozsah, pak pro zjištění, které dvojice středních hodnot se liší na zvolené hladině významnosti, použijeme
   a) Tukeyovu metodu mnohonásobného porovnávání
   b) Bartlettův test
   c) Levenův test

Správné odpovědi: 1c) 2a), c) 3a) 4c) 5a)

Příklady

1. Jsou známy měsíční tržby (v tisících Kč) tří prodavačů za dobu půl roku.

1. prodavač:	12	10	9	10	11	9
2. prodavač:	10	12	11	12	14	13
3. prodavač:	19	18	16	16	17	15

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty tržeb všech tří prodavačů jsou stejné. Pokud zamítneme nulovou hypotézu, zjistěte, tržby kterých dvou prodavačů se liší na hladině významnosti 0,05.

Výsledek:

$$M_{1\bullet} = 10{,}17, \quad M_{2\bullet} = 12, \quad M_{3\bullet} = 16{,}83, \quad M_{\bullet\bullet} = 13,$$
$$S_E = 27{,}7, \quad S_A = 142{,}3, \quad S_T = 170,$$
$$F_A = 38{,}58, \quad F_{0{,}975}(2,15) = 3{,}6823,$$

H_0 tedy zamítáme na hladině významnosti 0,05.

Výsledky zapíšeme do tabulky ANOVA

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	F_A
skupiny	$S_A = 142{,}3$	$f_A = 2$	$S_A/f_A = 71{,}17$	$\dfrac{S_A/f_A}{S_E/f_E} = 38{,}58$
reziduální	$S_E = 27{,}7$	$f_E = 15$	$S_E/f_E = 1{,}84$	—
celkový	$S_T = 170$	$f_T = 17$	—	—

Nyní pomocí Tukeyovy metody zjistíme, které dvojice prodavačů se liší na hladině významnosti 0,05.

Srovnávaní prodavači	Rozdíly $\lvert M_{k\bullet} - M_{l\bullet}\rvert$	Pravá strana vzorce
1, 2	1,83	2,03
1, 3	$6{,}67^*$	2,03
2, 3	$4{,}83^*$	2,03

Pravá strana:

$$q_{1-\alpha}(r, n-r)\frac{S_*}{\sqrt{p}} = q_{0{,}95}(3,15)\frac{\sqrt{1{,}84}}{\sqrt{6}} = 4{,}83\frac{\sqrt{1{,}84}}{\sqrt{6}} = 2{,}03,$$

kde $S_*^2 = \dfrac{S_E}{n-r} = 1{,}84$.

Na hladině významnosti 0,05 se liší tržby prodavačů 1, 3 a 2, 3.

2. Je dáno pět nezávislých náhodných výběrů o rozsazích 5, 7, 6, 8, 5, přičemž i-tý výběr pochází z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \dots, 5$. Byl vypočten celkový součet čtverců $S_T = 15$ a reziduální součet čtverců $S_E = 3$. Na hladině významnosti 0,05 testujte hypotézu o shodě středních hodnot.

Výsledek:

$$n = 5+7+6+8+5 = 31, \quad r = 5, \quad S_A = S_T - S_E = 15 - 3 = 12$$
$$F = \frac{S_A/(r-1)}{S_E/(n-r)} = \frac{12/4}{3/26} = 26, \quad F_{0{,}95}(4,26) = 2{,}9752$$

Protože $F \geq F_{0{,}95}(4,26)$, H_0 zamítáme na hladině významnosti 0,05.

3. Je dána neúplná tabulka ANOVA. Místo otazníků doplňte chybějící čísla.

zdroj variability	součet čtverců	stupně volnosti	podíl	F
skupiny	?	2	?	?
reziduální	16,033	?	?	—
celkový	17,301	35	—	—

Výsledek:

zdroj variability	součet čtverců	stupně volnosti	podíl	F
skupiny	**1,268**	2	**0,634**	**1,304**
reziduální	16,033	**33**	**0,486**	—
celkový	17,301	35	—	—

3. Studenti byli vyučováni předmětu za využití pěti pedagogických metod: tradiční způsob, programová výuka, audiotechnika, audiovizuální technika a vizuální technika. Z každé skupiny byl vybrán náhodný vzorek studentů a všichni byli podrobeni témuž písemnému testu. Na hladině významnosti 0,05 testujte hypotézu, že znalosti všech studentů jsou stejné a nezávisí na použité pedagogické metodě. V případě zamítnutí nulové hypotézy zjistěte, které výběry se liší na hladině významnosti 0,05.

metoda	počet bodů							
tradiční	76,2	48,3	85,1	63,7	91,6	87,2		
programová	85,2	74,3	76,5	80,3	67,4	67,9	72,1	60,4
audio	67,3	60,1	55,4	72,3	40,0			
audiovizuální	75,8	81,6	90,3	78,0	67,8	57,6		
vizuální	50,5	70,2	88,8	67,1	77,7	73,9		

Výsledek:
Všech pět náhodných výběrů má rozložení blízké normálnímu rozložení. Levenův test shody rozptylů má testové kritérium 0,819, počet stupňů volnosti je 4 a 26, odpovídající p-hodnota je 0,5248, tedy na hladině významnosti 0,05 hypotézu o shodě rozptylů nezamítáme. Analýza rozptylu má testové kritérium 1,6236, počet stupňů volnosti je 4 a 26, odpovídající p-hodnota je 0,1983, tedy na hladině významnosti 0,05 hypotézu o shodě středních hodnot nezamítáme. Znamená to, že na hladině významnosti 0,05 se neprokázaly odlišnosti ve znalostech studentů.

4. Pan Novák může cestovat z místa bydliště do místa pracoviště třemi různými způsoby: tramvají (způsob A), autobusem (způsob B) a metrem s následným přestupem na tramvaj (způsob C). Máme k dispozici jeho naměřené časy cestování do práce v době ranní špičky (včetně čekání na příslušný spoj) v minutách.

Způsob A:	32,	39,	42,	37,	34,	38	
Způsob B:	30,	34,	28,	26,	32		
Způsob C:	40,	37,	31,	39,	38,	33,	34

Pro všechny tři způsoby dopravy vypočtěte průměrné časy cestování. Na hladině významnosti 0,05 testujte hypotézu, že doba cestování do práce nezávisí na způsobu dopravy. V případě zamítnutí nulové hypotézy zjistěte, které způsoby dopravy do práce se od sebe liší na hladině významnosti 0,05.

Výsledek:
Průměrné časy cestování pro tři způsoby dopravy jsou 37 min, 30 min, 36 min.

Všechny tři náhodné výběry mají rozložení blízké normálnímu rozložení. Levenův test shody rozptylů má testové kritérium 0,1054, počet stupňů volnosti je 2 a 15, odpovídající p-hodnota je 0,9007, tedy na hladině významnosti

0,05 hypotézu o shodě rozptylů nezamítáme. Analýza rozptylu má testové kritérium 6,7151, počet stupňů volnosti je 2 a 15, odpovídající p-hodnota je 0,0083, tedy na hladině významnosti 0,05 hypotézu o shodě středních hodnot zamítáme. Scheffého metoda mnohonásobného porovnávání prokázala na hladině významnosti 0,05 rozdíl mezi způsoby A a B a mezi způsoby B a C.

- Motivace
- Jednovýběrové pořadové testy
- Dvouvýběrové pořadové testy
- Kruskalův-Wallisův test a mediánový test

7. Pořadové testy o mediánech

Cíl kapitoly

Po prostudování této kapitoly budete umět

- provádět testy hypotéz o mediánu jednoho spojitého rozložení
- hodnotit shodu dvou nezávislých náhodných výběrů ze spojitých rozložení lišících se posunem
- hodnotit shodu aspoň tří nezávislých náhodných výběrů ze spojitých rozložení lišících se posunem a identifikovat dvojice významně odlišných náhodných výběrů

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 5 hodin studia.

7.1 Motivace

Při používání t-testů či analýzy rozptylu by měl být splněn předpoklad normality dat. Pro výběry větších rozsahů ($n \geq 30$) nemá mírné porušení normality závažný dopad na výsledky. Někdy se však setkáváme s výběry malých rozsahů, které pocházejí z výrazně nenormálních rozložení. Pro práci s nimi byly vytvořeny tzv. neparametrické testy, které nevyžadují konkrétní typ rozložení (např. normální), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

Tyto neparametrické testy se rovněž používají v situacích, kdy zkoumaná data nemají intervalový či poměrový charakter, ale pouze ordinální charakter.

Ve srovnání s klasickými parametrickými testy jsou však neparametrické testy slabší, tzn., že nepravdivou hypotézu zamítají s menší pravděpodobností než testy parametrické.

V této kapitole se omezíme na ty neparametrické testy, které jsou založeny na pořadí a týkají se mediánů. Nazývají se pořadové testy.

7.2 Jednovýběrové pořadové testy

Jde o neparametrické obdoby jednovýběrového t-testu a párového t-testu.

7.2.1 Jednovýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozložení s hustotou $\varphi(x)$, která je symetrická kolem mediánu $x_{0,50}$, tj. $\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$. Nechť c je reálná konstanta. Testujeme hypotézu H_0: $x_{0,50} = c$ proti oboustranné alternativě H_1: $x_{0,50} \neq c$ (resp. proti levostranné alternativě H_1: $x_{0,50} < c$, resp. proti pravostranné alternativě H_1: $x_{0,50} > c$).

Utvoříme rozdíly $Y_i = X_i - c$, $i = 1, \ldots, n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

Absolutní hodnoty $|Y_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.

Zavedeme statistiku $S_W^+ = \sum\limits_{Y_i>0} R_i^+$, což je součet pořadí přes kladné hodnoty Y_i. Analogicky zavedeme statistiku $S_W^- = \sum\limits_{Y_i<0} R_i^-$, což je součet pořadí přes záporné hodnoty Y_i. Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$. Za platnosti H_0 statistika S_W^+ má střední hodnotu $E(S_W^+) = n(n+1)/4$ a rozptyl $D(S_W^+) = n(n+1)(2n+1)/24$.

H_0 zamítáme na hladině významnosti α, když testová statistika je menší nebo rovna tabelované kritické hodnotě. Testová statistika $= \min(S_W^+, S_W^-)$ pro oboustrannou alternativu, $= S_W^+$ pro levostrannou alternativu, $= S_W^-$ pro pravostrannou alternativu.

Pro $n \geq 30$ lze využít asymptotické normality statistiky S_W^+. Platí-li H_0, pak

$$U_0 = \frac{S_W^+ - E\left(S_W^+\right)}{\sqrt{D\left(S_W^+\right)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx N(0,1).$$

Kritický obor pro oboustrannou alternativu má tvar:

$$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty).$$

(Analogicky pro jednostranné alternativy.) H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Wilcoxonův test se hodí jen pro výběr ze symetrického rozložení. Není-li tento předpoklad splněn, lze použít např. znaménkový test (viz [HENDL], str. 193).

7.2.2 Příklad

U 12 náhodně zemí bylo zjištěno procento populace starší 60 let:

4,9   6,0   6,9   17,6   4,5   12,3   5,7   5,3   9,6   13,5   15,7   7,7.

Na hladině významnosti 0,05 testujte hypotézu, že medián procenta populace starší 60 let je 12 proti oboustranné alternativě.

Řešení:

Vypočteme rozdíly pozorovaných hodnot od čísla 12:

$-7{,}1 \quad -6{,}0 \quad -5{,}1 \quad 5{,}6 \quad -7{,}5 \quad 0{,}3 \quad -6{,}3 \quad -6{,}7 \quad -2{,}4 \quad 1{,}5 \quad 3{,}7 \quad -4{,}3.$

Absolutní hodnoty těchto rozdílů uspořádáme vzestupně podle velikosti. Kladné rozdíly přitom označíme tučně:

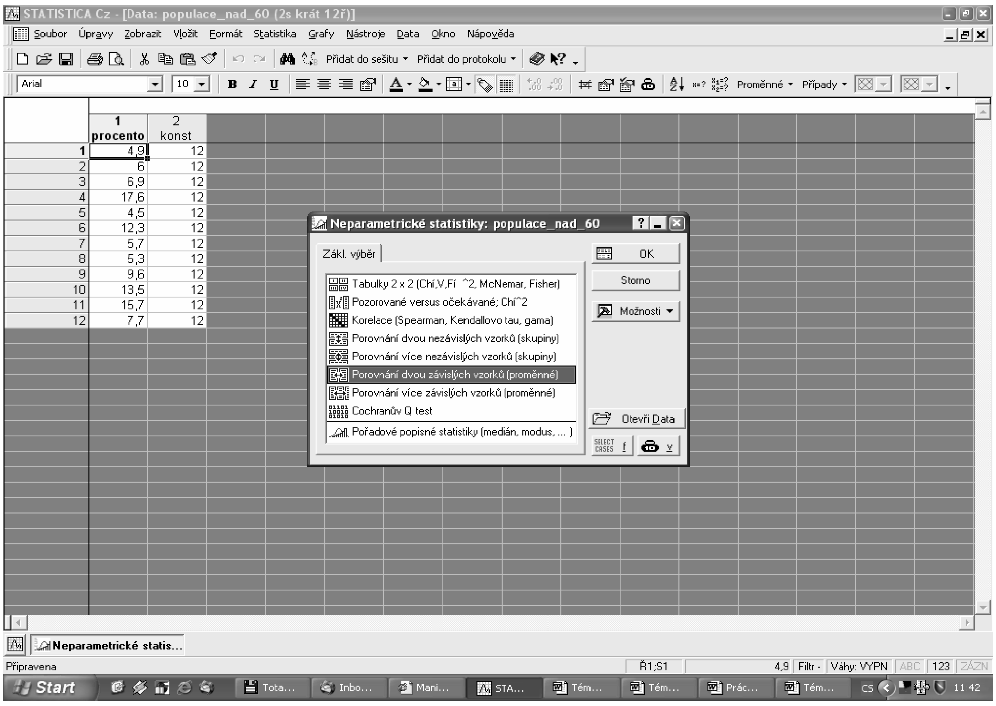
usp. $\lvert x_i - 12\rvert$	**0,3**	**1,5**	2,4	**3,7**	4,3	5,1	**5,6**	6	6,3	6,7	7,1	7,5
pořadí	**1**	**2**	3	**4**	5	6	**7**	8	9	10	11	12

$S_W^+ = 14$, $S_W^- = 64$, $n = 12$, $\alpha = 0{,}05$, tabelovaná kritická hodnota $= 13$, testová statistika $= \min(S_W^+, S_W^-) = \min(14, 64) = 14$. Protože $14 > 13$, H_0 nezamítáme na hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnou a dvanácti případy. První proměnnou nazveme *PROCENTA*, druhou *KONSTANTA*. Do proměnné *PROCENTA* napíšeme zjištěná procenta populace starší 60 let a proměnnou *KONSTANTA* vyplníme čísly 12 (do Dlouhého jména proměnné *KONSTANTA* napíšeme =12).

Statistika – Neparametrická statistika – Porovnání dvou závislých vzorků (proměnné) – OK.



Proměnné – 1. seznam proměnných – *PROCENTA*, 2. seznam proměnných – *KONSTANTA*, OK, Wilcoxonův párový test. Dostaneme tabulku

	Wilcoxonův párový test (populace_nad_60) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
procento & konst	12	14,00000	1,961161	0,049861

V této tabulce je symbolem T označena testová statistika $\min(S_W^+, S_W^-)$, symbolem Z realizace asymptotické testové statistiky U_0. Uvedená p-hodnota je vypočítána pro realizaci asymptotické testové statistiky U_0. Protože $p \leq 0{,}05$, hypotézu H_0: $x_{0,50} = 12$ zamítáme na asymptotické hladině významnosti 0,05. Pokud bychom chtěli provést přesný test a nikoliv pouze asymptotický, vyhledali bychom ve statistických tabulkách kritickou hodnotu jednovýběrového Wilcoxonova testu pro $n = 12$, $\alpha = 0{,}05$ (viz výše). Protože tato hodnota je 13, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

7.2.3 Párový Wilcoxonův test

Nechť $(X_1,Y_1),\ldots,(X_n,Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozložení. Testujeme H_0: $x_{0,50} - y_{0,50} = c$ proti H_1: $x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám). Utvoříme rozdíly $Z_i = X_i - Y_i$, $i = 1,\ldots,n$ a testujeme hypotézu o mediánu $z_{0,50}$, tj. H_0: $z_{0,50} = c$ proti H_1: $z_{0,50} \neq c$.

7.2.4 Příklad

K zjištění cenových rozdílů mezi určitými dvěma druhy zboží bylo náhodně vybráno 15 prodejen a byly zjištěny ceny zboží A a ceny zboží B: $(11,10)$, $(14,11)$, $(11,9)$, $(13,9)$, $(11,9)$, $(10,9)$, $(12,10)$, $(10,8)$, $(12,11)$, $(11,9)$, $(13,10)$, $(14,10)$, $(14,12)$, $(19,15)$, $(14,12)$. Na hladině významnosti 0,05 je třeba testovat hypotézu, že medián cenových rozdílů činí 3 Kč.

Řešení:

Jedná se o párový test. Vypočteme rozdíly mezi cenou zboží A a cenou zboží B, čímž úlohu převedeme na jednovýběrový test. Výpočty uspořádáme do tabulky:

č. prodejny	cena zboží A	cena zboží B	rozdíl	$\lvert$rozdíl $-$ medián$\rvert$	pořadí
1	11	10	1	2	12
2	14	11	3	0	—
3	11	9	2	1	5,5
4	13	9	4	1	**5,5**
5	11	9	2	1	5,5
6	10	9	1	2	12
7	12	10	2	1	5,5
8	10	8	2	1	5,5
9	12	11	1	2	12
10	11	9	2	1	5,5
11	13	10	3	0	—
12	14	10	4	1	**5,5**
13	14	12	2	1	5,5
14	19	15	4	1	**5,5**
15	14	12	2	1	5,5

Tučně jsou vytištěna pořadí pro kladné hodnoty rozdíl $-$ medián.

$S_W^+ = 16{,}5$, $S_W^- = 74{,}5$, $n = 13$, $\alpha = 0{,}05$, tabelovaná kritická hodnota $= 17$, testová statistika $= \min(S_W^+, S_W^-) = \min(16{,}5; 74{,}5) = 16{,}5$. Protože $16{,}5 \leq 17$, H_0 zamítáme na hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se čtyřmi proměnnými a 15 případy. První proměnnou nazveme *CENA A*, druhou *CENA B*, třetí *ROZDÍL* a čtvrtou *KONSTANTA*. Do proměnných *CEANA A* a *CENA B* zapíšeme ceny zboží A a B, do Dlouhého jména proměnné *ROZDÍL* napíšeme = v1-v2 a proměnnou *KONSTANTA* vyplníme samými trojkami. Nyní provedeme párový Wilcoxonův test:

Statistika – Neparametrická statistika – Porovnání dvou závislých vzorků (proměnné) – OK. Proměnné – 1. seznam proměnných – *ROZDÍL*, 2. seznam proměnných – *KONSTANTA*, OK, Wilcoxonův párový test. Dostaneme tabulku

	Wilcoxonův párový test (priklad734) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
rozdíl & konst	15	16,50000	2,026684	0,042696

Podobně jako v příkladu 7.2.2 je symbolem T označena testová statistika $\min(S_W^+, S_W^-)$, symbolem Z realizace asymptotické testové statistiky U_0. Uvedená p-hodnota je vypočítána pro realizaci asymptotické testové statistiky U_0. Protože $p \leq 0{,}05$, hypotézu H_0: $z_{0,50} = 3$ zamítáme na asymptotické hladině významnosti 0,05. Pokud bychom chtěli provést přesný test a nikoliv pouze asymptotický, vyhledali bychom ve statistických tabulkách kritickou hodnotu jednovýběrového Wilcoxonova testu pro $n = 13$, $\alpha = 0{,}05$ (viz výše). Protože tato hodnota je 17 a testová statistika 16,5, nulovou hypotézu zamítáme na hladině významnosti 0,05.

7.3 Dvouvýběrové pořadové testy

Jedná se o neparametrickou obdobu dvouvýběrového t-testu.

7.3.1 Dvouvýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme $x_{0,50}$ medián prvního rozložení a $y_{0,50}$ medián druhého rozložení. Testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné.

Všech $n+m$ hodnot $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ uspořádáme vzestupně podle velikosti. Zjistíme součet pořadí hodnot $X_1, \ldots, X_n$ a označíme ho T_1. Součet pořadí hodnot $Y_1, \ldots, Y_m$ označíme T_2. Vypočteme statistiky

$$U_1 = mn + n(n+1)/2 - T_1, \quad U_2 = mn + m(m+1)/2 - T_2.$$

Přitom platí $U_1 + U_2 = mn$. Pokud $\min(U_1, U_2) \leq$ tabelovaná kritická hodnota (pro dané rozsahy výběrů m, n a dané α), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti α.

Pro velká n, m (prakticky $n, m > 30$) lze využít asymptotické normality statistiky U_1. V případě platnosti H_0 má statistika

$$U_0 = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

asymptoticky rozložení $N(0,1)$. Kritický obor pro oboustrannou alternativu má tvar: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. (Analogicky pro jednostranné alternativy.) H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Dvouvýběrový Wilcoxonův test se používá v situacích, kdy distribuční funkce rozložení, z nichž dané dva nezávislé náhodné výběry pocházejí, se mohou lišit pouze posunutím.

7.3.2 Příklad

Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba zjistit, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.

hektarové výnosy při novém způsobu:	51	52	49	55		
hektarové výnosy při starém způsobu:	45	54	48	44	53	50

Řešení:

usp. hodnoty	44	45	48	**49**	50	51	**52**	53	54	**55**
pořadí *x*-ových hodnot				4		6	7			10
pořadí *y*-ových hodnot	1	2	3		5			8	9	

$$T_1 = 4+6+7+10 = 27, \quad T_2 = 1+2+3+5+8+9 = 28$$
$$U_1 = 4\cdot 6+4\cdot 5/2-27 = 7, \quad U_2 = 4\cdot 6+6\cdot 7/2-28 = 17$$

Kritická hodnota pro $\alpha = 0{,}05$, $\min(4,6) = 4$, $\max(4,6) = 6$ je 2. Protože $\min(7,17) > 2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

Řešení pomocí systému STATISTICA:

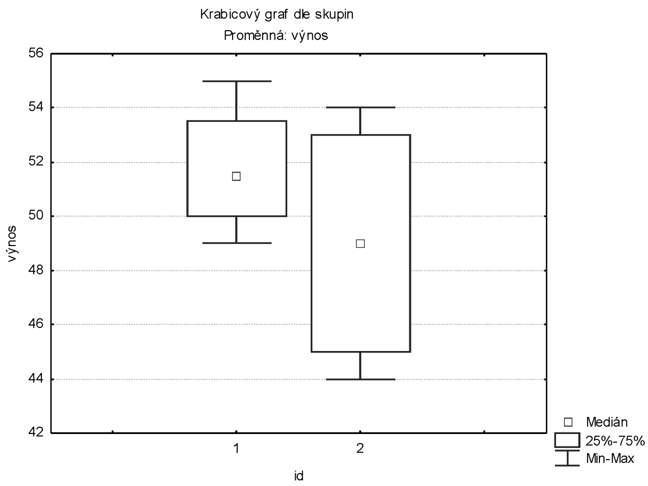
Otevřeme nový datový soubor se dvěma proměnnými *VÝNOS* a *ID* a 10 případy. Do proměnné *VÝNOS* zapíšeme hektarové výnosy pšenice a do proměnné *ID*, která slouží k rozlišení nového a starého způsobu hnojení, napíšeme 4krát jedničku a 6krát dvojku. Nyní provedeme dvouvýběrový Wilcoxonův test, který je ve STATISTICE uveden pod názvem Mannův – Whitneyův test:

Statistika – Neparametrická statistika – Porovnání dvou nezávislých vzorků (skupiny) – OK. Proměnné – Seznam závislých proměnných – *VÝNOS*, Nezáv. (grupov.) proměnné – *ID* – OK, Mann–Whitneyův U test. Dostaneme tabulku

Proměnná	Mann-Whitneyův U test (priklad742) Dle proměn. id Označené testy jsou významné na hladině p <,05000 Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
výnos	27,00000	28,00000	7,000000	1,066004	0,286423	1,066004	0,286423	4	6	0,352381

Zde je symbolem U označena testová statistika $\min(U_1, U_2)$. V našem případě $U = 7$, odpovídající p-hodnotu najdeme v posledním sloupci pod označením 2*1 str. přesné p. Protože $0{,}352381 > 0{,}05$, nezamítáme na hladině významnosti 0,05 nulovou hypotézu.

Výpočet ještě doplníme krabicovým diagramem. Na záložce Zákl. výsledky vybereme Krabicový graf dle skupin, OK, proměnná *VÝNOS*, OK. Dostaneme graf



Je zřejmé, že medián hektarových výnosů při starém způsobu hnojení je menší než při novém způsobu a také vidíme, že variabilita hektarových výnosů při starém způsobu hnojení je větší než při novém způsobu.

7.4 Kruskalův-Wallisův test a mediánový test (neparametrické obdoby analýzy rozptylu jednoduchého třídění)

7.4.1 Formulace problému

Nechť je dáno $r \geq 3$ nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_r$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \cdots + n_r$. Chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

7.4.2 Kruskalův-Wallisův test

Všech n hodnot seřadíme do rostoucí posloupnosti a určíme pořadí každé hodnoty. Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, $j = 1, \ldots, r$ (kontrola: musí platit $T_1 + \cdots + T_r = n(n+1)/2$). Testová statistika má tvar:

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1).$$

Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(r-1)$. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(r-1)$.

7.4.3 Mediánový test

Testová statistika má tvar $Q_M = 4 \sum_{j=1}^{r} \frac{P_j^2}{n_j} - n$, kde P_j je počet hodnot v j-tém výběru, které jsou větší nebo rovny mediánu vypočtenému ze všech n hodnot. Platí-li H_0, má statistika Q_M asymptoticky rozložení $\chi^2(r-1)$. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q_M \geq \chi^2_{1-\alpha}(r-1)$.

7.4.4 Metody mnohonásobného porovnávání

Zamítneme-li H_0, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti.

a) Neményiho metoda
Používá se v případě, že všechny výběry mají týž rozsah p. Je-li $|T_l - T_k| \geq$ tabelovaná kritická hodnota (pro dané p, r, α), pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

b) Obecná metoda mnohonásobného porovnávání
Jestliže

$$|T_l - T_k| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_1} + \frac{1}{n_k}\right) n(n+1) h_{KW}(\alpha)},$$

pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení. Kritickou hodnotu $h_{KW}(\alpha)$ najdeme ve speciálních statistických tabulkách. Při větších rozsazích výběrů je možno ji nahradit kvantilem $\chi^2_{1-\alpha}(r-1)$.

7.4.5 Příklad

V roce 1980 byly získány tři nezávislé výběry obsahující údaje o průměrných ročních příjmech (v tisících dolarů) čtyř sociálních skupin ve třech různých oblastech USA.

jižní oblast: 6 10 15 29
pacifická oblast: 11 13 17 131
severovýchodní oblast: 7 14 28 25

Na hladině významnosti 0,05 testujte hypotézu, že příjmy v těchto oblastech se neliší. Zamítnete-li nulovou hypotézu, vyšetřete, které dvojice výběrů se od sebe liší na hladině významnosti 0,05.

Řešení:

Kruskalův-Wallisův test

Usp. hodnoty	6	7	10	11	13	14	15	17	25	28	29	131
Pořadí 1. výběru	1		3				7				11	
Pořadí 2. výběru				4	5			8				12
Pořadí 3. výběru		2				6			9	10		

$$T_1 = 22, \quad T_2 = 29, \quad T_3 = 27,$$

$$Q = \frac{12}{12 \cdot 13}\left(\frac{22^2}{4} + \frac{29^2}{4} + \frac{27^2}{4}\right) - 3 \cdot 13 = 0{,}5, \quad \chi^2_{0{,}95}(2) = 5{,}991.$$

Protože $Q < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05. Rozdíly mezi průměrnými ročními příjmy v uvedených třech oblastech se neprokázaly.

Mediánový test
Medián všech 12 hodnot je 14,5. V 1. výběru leží nad mediánem 2 hodnoty, ve 2. výběru 2 hodnoty, ve 3. výběru 2 hodnoty. Testová statistika

$$Q_M = 4\left[\frac{1}{4}(2^2+2^2+2^2)\right] - 12 = 0,$$

odpovídající kvantil $\chi^2_{0,95}(2) = 5{,}991$. Protože $Q_M < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnými *PŘÍJEM* a *ID* a s 12 případy. Do proměnné *PŘÍJEM* zapíšeme hodnoty příjmu, do proměnné *ID*, která slouží jako identifikátor oblasti, napíšeme 4krát jedničku, 4krát dvojku a 4krát trojku. Nyní provedeme Kruskalův-Wallisův a mediánový test.

Statistika – Neparametrická statistika – Porovnání více nezávislých vzorků (skupiny) – OK.

Proměnné – Závisle proměnné – *PŘÍJEM*, Nezáv. (grupov.) proměnná – *ID* – OK, Shrnutí: Kruskal-Wallis ANOVA a mediánový test, Výpočet. Pro K-W test dostaneme tabulku

Kruskal-Wallisova ANOVA založ. na poř.; příjem (priklad755)
Nezávislá (grupovací) proměnná :ID
Kruskal-Wallisův test: H (2, N= 12) =,5000000 p =,7788

Závislá: příjem	Kód	Počet platných	Součet pořadí
1	1	4	22,00000
2	2	4	29,00000
3	3	4	27,00000

Testová statistika se realizuje hodnotou 0,5, počet stupňů volnosti je 2, odpovídající p-hodnota $= 0{,}7788$, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o shodě mediánů.

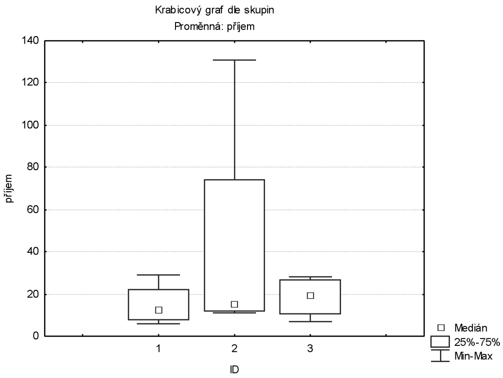
Pro mediánový test máme tabulku

Mediánový test, celk. medián = 14,5000; příjem (priklad755)
Nezávislá (grupovací) proměnná :ID
Chi-Kvadr. = 0,000000 df = 2 p = 1,000

Závislá: příjem	1	2	3	Celkem
<= Medián: pozorov.	2,000000	2,000000	2,000000	6,00000
očekáv.	2,000000	2,000000	2,000000	
poz.-oč.	0,000000	0,000000	0,000000	
> Medián: pozorov.	2,000000	2,000000	2,000000	6,00000
očekáv.	2,000000	2,000000	2,000000	
poz.-oč.	0,000000	0,000000	0,000000	
Celkem: oček.	4,000000	4,000000	4,000000	12,00000

Realizace testové statistiky $= 0$, počet stupňů volnosti $= 2$, odpovídající p-hodnota $= 1$, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o shodě mediánů.

Výpočet ještě doplníme krabicovým diagramem. Na záložce Zákl. výsledky vybereme Krabicový graf, proměnná *PŘÍJEM*, OK, Typ krabicového grafu Medián/Kvartily/Rozpětí, OK. Dostaneme graf



Vidíme, že mediány se liší jenom nepatrně, ale druhá skupina (tj. pacifická oblast) vykazuje velkou variabilitu. Je to díky hodnotě 131.

Poznámka:

Pokud bychom na dané hladině významnosti α zamítli hypotézu o shodě mediánů, mohli bychom pomocí metody mnohonásobného porovnívání zjistit, které dvojice skupin se liší na hladině významnosti α. Na záložce Základní výsledky stačí zaškrtnout Vícenás. porovnání průměrného pořadí pro vš. sk. Výstupní tabulka obsahuje p-hodnoty pro test shody mediánů dvojic skupin.

Shrnutí kapitoly

V některých situacích se setkáváme s náhodnými výběry malých rozsahů, které pocházejí z výrazně nenormálních rozložení. V takových případech nelze použít klasické testy založené na předpokladu normality, které byly popsány ve 4., 5. a 6. kapitole. Místo nich používáme neparametrické testy, které nepotřebují splnění předpokladu normality, stačí např. předpokládat spojitost distribuční funkce rozložení, z něhož daný náhodný výběr pochází.

Pro testování hypotézy o mediánu používáme *jednovýběrový* či *párový Wilcoxonův test*, což je neparametrická obdoba jednovýběrového či párového t.testu. Máme-li testovat hypotézu o shodě mediánů dvou rozložení, která se mohou lišit jen posunutím (tj. testujeme hypotézu o shodě těchto dvou rozložení), aplikujeme *dvouvýběrový Wilcoxonův test* – neparametrickou obdobu dvouvýběrového t-testu.

Jako neparametrická obdoba analýzy rozptylu jednoduchého třídění slouží *Kruskalův-Wallisův test* nebo *mediánový test*. Při zamítnutí nulové hypotézy identifikujeme dvojice odlišných výběrů pomocí *metod mnohonásobného porovnávání*, a to buď *obecnou metodu mnohonásobného porovnávání* nebo *Neméniyho metodu*.

Při provádění neparametrických testů potřebujeme speciální tabulky kritických hodnot. Jsou obsaženy v příloze A tohoto učebního textu. Všechny uvedené testy jsou implementovány v systému STATISTICA.

Kontrolní otázky

1. V jakých situacích používáme neparametrické testy?
2. Jaká je nevýhoda neparametrických testů oproti testům parametrickým?
3. Jak vypočítáme pořadí čísla v dané posloupnosti čísel?
4. Popište rozdíl mezi jednovýběrovým a párovým Wilcoxonovým testem.
5. Jaké podmínky musí být splněny pro dvouvýběrový Wilcoxonův test?
6. K čemu slouží Kruskalův-Wallisův test?
7. Jak provedeme mediánový test?
8. Které metody mnohonásobného porovnávání znáte?

Autokorekční test

1. Máme za úkol zjistit, zda tři nezávislé výběry pocházejí z téhož rozložení. Přitom všechny mají malý rozsah (menší než 30) a vykazují odchylky od normálního rozložení. Jaký test použijeme?
   a) Analýzu rozptylu jednoduchého třídění,
   b) mediánový test,
   c) Kruskalův-Wallisův test.
2. Testujeme hypotézu, že dva nezávislé náhodné výběry pocházejí z téhož rozložení. Oba výběry mají malý rozsah (menší než 30) a diagnostické grafy i testy normality poukazují na závažnější odchylky od normálního rozložení. Jaký test použijeme?
   a) Párový Wilcoxonův test,
   b) dvouvýběrový t-test,
   c) dvouvýběrový Wilcoxonův test.
3. Pomocí K-W testu testujeme na asymptotické hladině významnosti 0,05 hypotézu, že pět nezávislých náhodných výběrů o rozsazích 4, 7, 5, 4, 5 pochází z téhož rozložení. Kritický obor má tvar:
   a) $W = \langle 9{,}488; \infty)$,
   b) $W = \langle 0{,}711; \infty)$,
   c) $W = \langle 0; 9{,}488)$.
4. Máme dvourozměrný náhodný výběr z dvourozměrného rozložení, které se výrazně liší od normálního rozložení. K testování hypotézy, že mediány obou složek tohoto rozložení jsou stejné, použijeme
   a) jednovýběrový t-test,
   b) dvouvýběrový Wilcoxonův test,
   c) párový Wilcoxonův test.

Správné odpovědi: 1b), c) 2c) 3a) 4c)

Příklady

1. U 10 náhodně vybraných vzorků benzínu byly zjištěny následující hodnoty oktanového čísla:

   98,2 96,8 96,3 99,8 96,9 98,6 95,6 97,1 97,7 98,0.

Na hladině významnosti 0,05 testujte hypotézu, že medián oktanového čísla je 98 proti oboustranné alternativě.

Výsledek:
Použijeme jednovýběrový Wilcoxonův test. Testová statistika se realizuje hodnotou 12, tabelovaná kritická hodnota pro $\alpha = 0{,}05$ a $n = 9$ je 5. Protože $12 > 5$, H_0 nezamítáme na hladině významnosti 0,05.

2. Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologiemi. Rozhodující je procentní obsah určité látky.

1. technologie:	1,52	1,57	1,71	1,34	1,68			
2. technologie:	1,75	1,67	1,56	1,66	1,72	1,79	1,64	1,55

Na hladině významnosti 0,05 posuďte pomocí dvouvýběrového Wilcoxonova testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.

Výsledek:
Testová statistika se realizuje hodnotou 12, tabelovaná kritická hodnota pro $\alpha = 0{,}05$, $\min(5,8) = 5$, $\max(5,8) = 8$ je 6. Protože $\min(28,12) > 2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že obě technologie poskytují stejné procento účinné látky.

3. Výrobce koláčů v prášku má 4 nové recepty a chce zjistit, zda se jejich kvalita liší. Upekl proto 5 koláčů z každého druhu a dal je porotě k ohodnocení.

recept A:	72	88	70	87	71,
recept B:	85	89	86	82	88,
recept C:	94	94	88	87	89,
recept D:	91	93	92	95	94.

Na hladině významnosti 0,05 testujte hypotézu, že recepty se neliší.

Výsledek:
Použijeme Kruskalův-Wallisův test. Všech 20 hodnot uspořádáme vzestupně podle velikosti a stanovíme součet pořadí pro recepty A, B, C, D: $T_1 = 23{,}5$, $T_2 = 37{,}5$, $T_3 = 66$, $T_4 = 83$. Testová statistika:

$$Q = \frac{12}{20 \cdot 21}\left(\frac{23{,}5^2}{5} + \frac{37{,}5^2}{5} + \frac{66^2}{5} + \frac{83^2}{5}\right) - 3 \cdot 21 = 12{,}45,$$

$\chi^2_{0,95}(3) = 7{,}81$. Protože $Q \geq 7{,}81$, H_0 zamítáme na asymptotické hladině významnosti 0,05. Neményiho metoda prokázala, že na hladině významnosti 0,05 se liší recepty A a D.

4. U osmi osob byl změřen systolický krevní tlak před pokusem a po něm.

č. osoby	1	2	3	4	5	6	7	8
tlak před	130	185	162	136	147	181	128	139
tlak po	139	190	175	135	155	175	158	149

Na hladině významnosti 0,05 testujte hypotézu, že pokus neovlivní systolický krevní tlak

Výsledek:
Párový Wilcoxonův test poskytl p-hodnotu 0,04995, tedy H_0 zamítáme na hladině významnosti 0,05.

5. Majitel obchodu chtěl zjistit, zda velikost nákupů (v dolarech) placených kreditními kartami Master/EuroCard a Visa jsou přibližně stejné. Náhodně vybral 7 nákupů placených Master/EuroCard:

42 77 46 73 78 33 37

a 9 placených Visou:

39 10 119 68 76 126 53 79 102.

Lze na hladině významnosti 0,05 tvrdit, že mediány nákupů placených těmito dvěma typy karet se shodují?

Výsledek:
Dvouvýběrový Wilcoxonův test poskytl p-hodnotu 0,2523, H_0 tedy nezamítáme na hladině významnosti 0,05.

6. Z produkce tří podniků vyrábějících televizory bylo vylosováno 10, 8 a 12 kusů. Byly získány následující výsledky zjišťování citlivosti těchto televizorů v mikrovoltech:

1. podnik:	420	560	600	490	550	570	340	480	510	460		
2. podnik:	400	420	580	470	470	500	520	530				
3. podnik:	450	700	630	590	420	590	610	540	740	690	540	670

Ověřte na hladině významnosti 0,05 hypotézu o shodě úrovně citlivosti televizorů v jednotlivých podnicích.

Výsledek:
K-W test poskytl testovou statistiku 3,2043, počet stupňů volnosti $= 2$, odpovídající p-hodnota $= 0{,}0165$, H_0 zamítáme na asymptotické hladině významnosti 0,05. Liší se výrobky podniků 2 a 3.

- Motivace
- Testování nezávislosti nominálních veličin
- Testování nezávislosti ordinálních veličin
- Testování nezávislosti intervalových či poměrových veličin

8. Analýza závislosti dvou náhodných veličin

Cíl kapitoly

Po prostudování této kapitoly budete umět

- provádět test nezávislosti v kontingenční tabulce
- hodnotit intenzitu závislosti dvou náhodných veličin nominálního typu pomocí Cramérova koeficientu
- provádět Fisherův přesný test ve čtyřpolní kontingenční tabulce a počítat podíl šancí na úspěch za dvojích různých podmínek
- provádět test pořadové nezávislosti dvou náhodných veličin ordinálního typu pomocí Spearmanova koeficientu pořadové korelace
- testovat hypotézu o nezávislosti dvou náhodných veličin intervalového či poměrového typu, které se řídí dvourozměrným normálním rozložením

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 15 hodin studia.

8.1 Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá nebo zda počet dnů absence a věk pracovníka jsou nezávislé. Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové.

Při zkoumání závislosti je nesmírně důležité provést logický rozbor problému. Nemá smysl se zabývat hledáním závislosti v případech, když

- z logických důvodů nemůže existovat,
- závislost je způsobena formálními vztahy mezi veličinami,
- soubor dvourozměrných dat je nehomogenní,
- závislost je způsobena společnou příčinou.

Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

8.2 Testování nezávislosti nominálních veličin

8.2.1 Popis testu

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]},\ldots,x_{[r]}$ a Y nabývá variant $y_{[1]},\ldots,y_{[s]}$. Pořídíme dvourozměrný náhodný výběr rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X,Y). Zjištěné absolutní četnosti n_{jk} dvojice variant $(x_{[j]},y_{[k]})$ uspořádáme do kontingenční tabulky:

y / x n_{jk}	$y_{[1]}$	$\dots$	$y_{[s]}$	$n_{j\bullet}$
$x_{[1]}$	n_{11}	$\dots$	n_{1s}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$x_{[r]}$	n_{r1}	$\dots$	n_{rs}	$n_{r\bullet}$
$n_{\bullet k}$	$n_{\bullet 1}$	$\dots$	$n_{\bullet s}$	n

Testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j\bullet} n_{\bullet k}}{n}\right)^2}{\frac{n_{j\bullet} n_{\bullet k}}{n}}.$$

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$. Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

8.2.2 Podmínky dobré aproximace

Výraz $\frac{n_{j\bullet} n_{\bullet k}}{n}$ se nazývá teoretická četnost. Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti aspoň v 80 % případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20 % neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

8.2.3 Měření síly závislosti

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.

8.2.4 Příklad

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

Typ školy	Sociální skupina				$n_{j\bullet}$
	I	II	III	IV	
univerzitní	50	30	10	50	140
technický	30	50	20	10	110
ekonomický	10	20	30	50	110
$n_{\bullet k}$	90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení:

$$\frac{n_{1\bullet}n_{\bullet 1}}{n} = \frac{140\cdot 90}{360} = 35, \quad \frac{n_{1\bullet}n_{\bullet 2}}{n} = \frac{140\cdot 100}{360} = 38{,}9,$$
$$\frac{n_{1\bullet}n_{\bullet 3}}{n} = \frac{140\cdot 60}{360} = 23{,}3, \quad \frac{n_{1\bullet}n_{\bullet 4}}{n} = \frac{140\cdot 110}{360} = 42{,}8,$$

$$\frac{n_{2\bullet}n_{\bullet 1}}{n} = \frac{110\cdot 90}{360} = 27{,}5, \quad \frac{n_{2\bullet}n_{\bullet 2}}{n} = \frac{110\cdot 100}{360} = 30{,}6,$$
$$\frac{n_{2\bullet}n_{\bullet 3}}{n} = \frac{110\cdot 60}{360} = 18{,}3, \quad \frac{n_{2\bullet}n_{\bullet 4}}{n} = \frac{110\cdot 110}{360} = 33{,}6,$$

$$\frac{n_{3\bullet}n_{\bullet 1}}{n} = \frac{110\cdot 90}{360} = 27{,}5, \quad \frac{n_{3\bullet}n_{\bullet 2}}{n} = \frac{110\cdot 100}{360} = 30{,}6,$$
$$\frac{n_{3\bullet}n_{\bullet 3}}{n} = \frac{110\cdot 60}{360} = 18{,}3, \quad \frac{n_{3\bullet}n_{\bullet 4}}{n} = \frac{110\cdot 110}{360} = 33{,}6,$$

$$K = \frac{(50-35)^2}{35} + \frac{(30-38{,}9)^2}{38{,}9} + \cdots + \frac{(50-33{,}6)^2}{33{,}6} = 76{,}84,$$
$$r = 3, \quad s = 4, \quad \chi^2_{0{,}95}(6) = 12{,}6.$$

Protože $K \geq 12{,}6$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient:

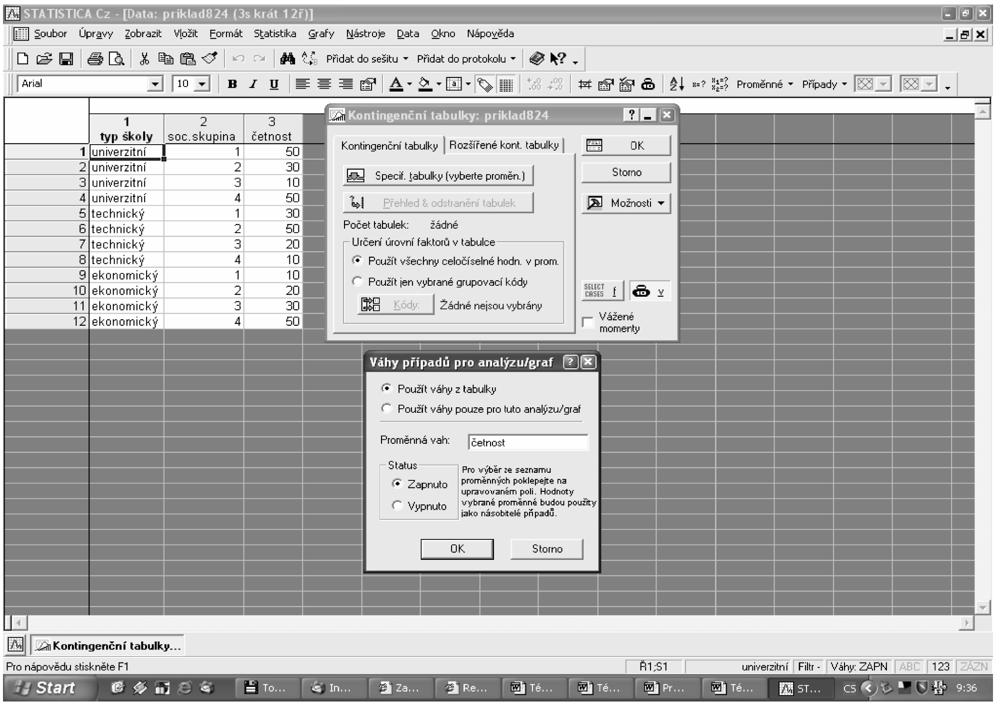
$$V = \sqrt{\frac{76{,}4}{360\cdot 2}} = 0{,}3267.$$

Podmínky dobré aproximace jsou splněny, protože všechny teoretické četnosti jsou větší než 5.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor o 12 případech a třech proměnných *TYP ŠKOLY*, *SOC. SKUPINA*, *ČETNOST*). Do proměnné *TYP ŠKOLY* napíšeme varianty typu školy $x_{[1]} = 1$ (univerzitní), $x_{[2]} = 2$ (technický), $x_{[3]} = 3$ (ekonomický), přičemž každá varianta se objeví čtyřikrát pod sebou. Do proměnné *SOC. SKUPINA* napíšeme třikrát pod sebe všechny varianty $y_{[1]} = 1$, $y_{[2]} = 2$, $y_{[3]} = 3$, $y_{[4]} = 4$. Do proměnné *ČETNOST* napíšeme absolutní četnosti jednotlivých dvojic variant $(x_{[j]}, y_{[k]})$.

Statistika – Základní statistiky/tabulky – Kontingenční tabulky – OK – klikneme myší na tlačítko s obrázkem závaží – Status zapnuto – Proměnná vah *ČETNOST* – OK – Specif. tabulky – List 1 *TYP ŠKOLY* – List 2 *SOC. SKUPINA* – OK.



Přesvědčíme se o splnění podmínek dobré aproximace. Na záložce Možnosti zaškrtneme Očekávané četnosti, zvolíme Výpočet. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (priklad824)
Četnost označených buněk > 10
Pearsonův chí-kv. : 76,8359, sv=6, p=,000000

typ školy	soc.skupina 1	soc.skupina 2	soc.skupina 3	soc.skupina 4	Řádk. součty
univerzitní	35,00000	38,8889	23,33333	42,7778	140,0000
technický	27,50000	30,5556	18,33333	33,6111	110,0000
ekonomický	27,50000	30,5556	18,33333	33,6111	110,0000
Vš.skup.	90,00000	100,0000	60,00000	110,0000	360,0000

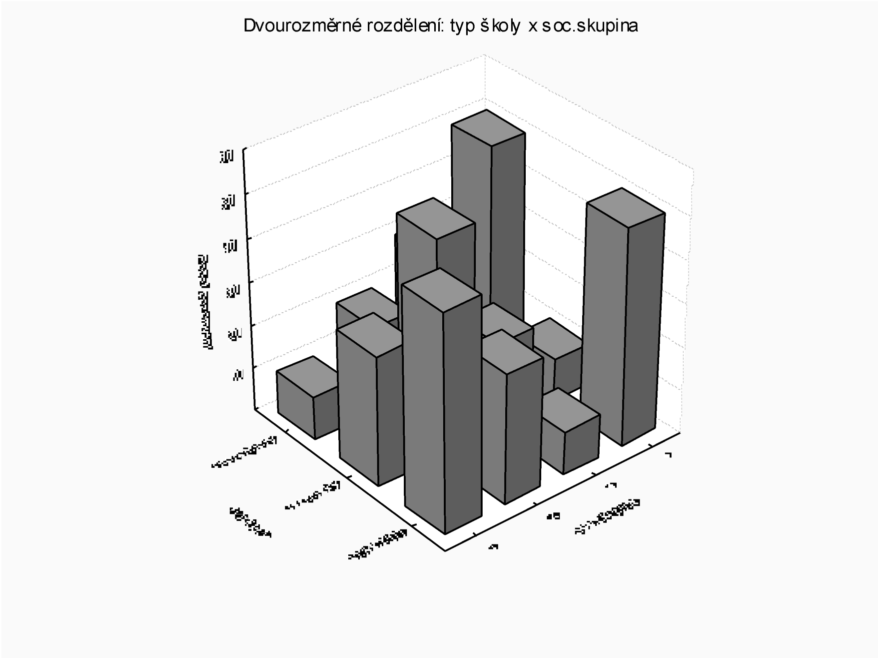
Vidíme, že všechny teoretické četnosti jsou dostatečně velké, větší než 5. V tabulce je rovněž uvedena realizace testové statistiky $K = 76{,}8359$, počet stupňů volnosti $= 6$. Odpovídající p-hodnota je blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny, z níž uchazeč pochází.

Dále vypočteme Cramérův koeficient. Na záložce Možnosti zaškrtneme Fí & Cramérovo C&V. Přejdeme na záložku Detailní výsledky a vybereme Detailní 2-rozm. tabulky.

	Statist. : typ školy(3) x soc.skupina(4) (priklad824)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	76,83589	df=6	p=,00000
M-V chí-kvadr.	84,53528	df=6	p=,00000
Fí	,4619881		
Kontingenční koeficient	,4193947		
Cramér. V	,3266749		

V této tabulce najdeme Cramérův koeficient $V = 0{,}3266749$ a také hodnotu testové statistiky K s počtem stupňů volnosti 6 a odpovídající p-hodnotou blízkou 0.

Výpočet ještě doplníme grafickým znázorněním simultánních absolutních četností proměnných *TYP ŠKOLY* a *SOC. SKUPINA*. Na záložce Detailní výsledky zvolíme 3D histogramy.



Poznámka:

Graf lze různě natáčet, stačí v menu vybrat Formát – Vš. možnosti – Zorný bod.

8.2.5 Čtyřpolní tabulky

Nechť $r = s = 2$. Pak hovoříme o čtyřpolní kontingenční tabulce a používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

X	Y		$n_{j\bullet}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	$a+b$
$x_{[2]}$	c	d	$c+d$
$n_{\bullet k}$	$a+c$	$b+d$	n

Pro tuto tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test. (Je popsán např. v knize ZVÁRA, K.: *Biostatistika*, Karolinum, Praha 1998.) STATISTICA poskytuje p-hodnotu pro tento test. Jestliže vyjde $p \leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \dfrac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j\bullet}$
	I	II	
úspěch	a	b	$a+b$
neúspěch	c	d	$c+d$
$n_{\bullet k}$	$a+c$	$b+d$	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za prvních okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$. Pomocí $100(1-\alpha)\%$ asymptotického intervalu spolehlivosti pro podíl šancí lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický $100(1-\alpha)\%$ interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:

$$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2}.$$

Jestliže po odlogaritmování nezahrne interval spolehlivosti 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

8.2.6 Příklad

U 125 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu a dojem u přijímací zkoušky jsou nezávislé veličiny.

přijetí	dojem		$n_{j\bullet}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{\bullet k}$	56	69	125

Řešení:

$$OR = \frac{ad}{bc} = \frac{17\cdot 58}{11\cdot 39} = 2{,}298, \quad \ln OR = 0{,}832,$$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0{,}439, \quad u_{0{,}975} = 1{,}96,$$

$$\ln dm = 0{,}832 - 0{,}439\cdot 1{,}96 = -0{,}028, \quad \ln hm = 0{,}832 + 0{,}439\cdot 1{,}96 = 1{,}692,$$

$$dm = e^{-0{,}028} = 0{,}972, \quad hm = e^{1{,}692} = 5{,}433.$$

Protože interval $(0{,}972; 5{,}433)$ obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

Řešení pomocí systému STATISTICA:

Otevřeme nový datový soubor se třemi proměnnými *PŘIJETÍ*, *DOJEM*, *ČETNOST* a se čtyřmi případy. Proměnná *PŘIJETÍ* má varianty 1 (přijat), 2 (nepřijat), proměnná

DOJEM má varianty 1 (dobrý), 2 (špatný). Způsob zadání dat je podobný jako v příkladu 8.2.4. Nesmíme zapomenout zadat váhovou proměnnou *ČETNOST*.

Provedeme Fisherův přesný test. Na záložce Možnosti zaškrtneme Fisher exakt., přejdeme na záložku Detailní výsledky a vybereme Detailní 2-rozm. tabulky.

	Statist. : přijetí(2) x dojem(2) (priklad826)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	3,695333	df=1	p=,05457
M-V chí-kvadr.	3,687112	df=1	p=,05484
Yatesův chí-kv.	2,912566	df=1	p=,08790
Fisherův přesný, 1-str.			p=,04413
2-stranný			p=,08331
McNemarův chí-kv. (A/D)	21,33333	df=1	p=,00000
(B/C)	14,58000	df=1	p=,00013

Vidíme, že p-hodnota Fisherova přesného testu je 0,08331, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že dojem a přijetí na fakultu jsou nezávislé veličiny.

8.3 Testování nezávislosti ordinálních veličin

8.3.1 Popis testu

Nechť X,Y jsou dvě ordinální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání). Pořídíme dvourozměrný náhodný výběr $(X_1,Y_1),\ldots,(X_n,Y_n)$ z rozložení, jímž se řídí náhodný vektor (X,Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, $i = 1,\ldots,n$. Testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny (resp. proti levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost resp. proti pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

Testová statistika se nazývá Spearmanův koeficient pořadové korelace a má tvar:

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2.$$

H_0 zamítáme na hladině významnosti α

a) ve prospěch oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha}(n)$

b) ve prospěch levostranné alternativy, když $r_S \leq -r_{S,1-\alpha}(n)$

c) ve prospěch pravostranné alternativy, když $r_S \geq r_{S,1-\alpha}(n)$,

kde $r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro $\alpha = 0{,}05$ nebo 0,01 a $n \leq 30$ najdeme v tabulkách. Pro $n > 30$ H_0 zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když $|r_S| \geq \frac{u_{1-\alpha/2}}{\sqrt{n-1}}$ (analogicky pro jednostranné alternativy).

Spearmanův koeficient r_S současně měří sílu pořadové závislosti náhodných veličin X, Y. Nabývá hodnot z intervalu $\langle -1, 1\rangle$. Čím je jeho hodnota bližší -1 (resp. 1), tím je silnější nepřímá (resp. přímá) pořadová závislost veličin X, Y. Čím je jeho hodnota bližší 0, tím je slabší pořadová závislost veličin X, Y.

8.3.2 Příklad

Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení:

$$r_S = 1 - \frac{6}{1(7^2-1)}\big[(4-4)^2 + (1-2)^2 + (6-5)^2 + (5-6)^2 + \\ + (3-1)^2 + (2-3)^2 + (7-7)^2\big] = 0{,}857.$$

Kritická hodnota: $r_{S,0{,}95}(7) = 0{,}745$. Protože $0{,}857 \geq 0{,}745$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

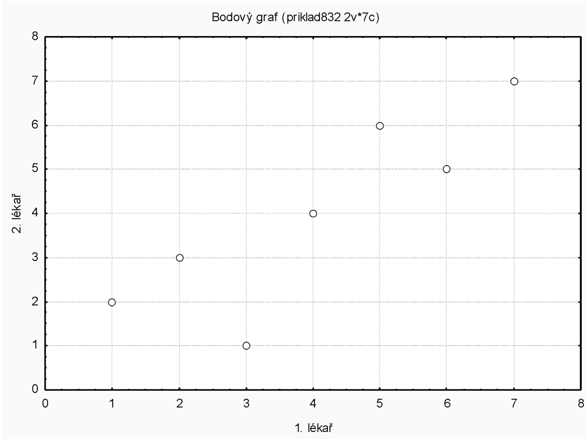
Řešení pomocí sytému STATISTICA:

Otevřeme nový datový soubor se dvěma proměnnými *1.LÉKAŘ*,. *2.LÉKAŘ* a sedmi případy. Do těchto proměnných zapíšeme zjištěná hodnocení. Statistika – Neparametrická statistika – Korelace – OK, Vytvořit Detailní report, Proměnné – 1. seznam proměnných *1.LÉKAŘ*,. 2. seznam proměnných *2.LÉKAŘ* – OK – Spearman R.

	Spearmanovy korelace (priklad832) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
1. lékař & 2. lékař	7	0,857143	3,721042	0,013697

Spearmanův koeficient korelace nabyl hodnoty 0,857143, asymptotická testová statistika se realizovala číslem 3,721042, odpovídající p-hodnota je 0,013697, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení obou lékařů. Pokud bychom chtěli provést přesný test, nikoliv asymptotický test, museli bychom použít statistické tabulky a vyhledat v nich kritickou hodnotu $r_{S,0{,}95}(7)$ – viz výše.

Výpočet ještě doplníme dvourozměrným tečkovým diagramem. Grafy – Bodové grafy – vypneme Typ proložení – Proměnné – *X 1.LÉKAŘ*, *Y 2.LÉKAŘ*, OK, OK.



Vidíme, že s rostoucím hodnocením 1. lékaře roste hodnocení 2. lékaře a naopak. Tedy mezi oběma proměnnými existuje určitý stupeň přímé pořadové závislosti.

8.4 Testování nezávislosti intervalových či poměrových veličin

8.4.1 Pearsonův koeficient korelace

V teorii pravděpodobnosti byl zaveden Pearsonův koeficient korelace náhodných veličin X, Y (které jsou aspoň intervalového charakteru) vztahem

$$R(X,Y) = \begin{cases} \dfrac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)}\sqrt{D(Y)} > 0, \\ 0 & \text{jinak} \end{cases}$$

Připomeneme jeho vlastnosti:

a) $R(X,X) = 1$
b) $R(X,Y) = R(Y,X)$
c) $R(a+bX, c+dY) = \operatorname{sgn}(bd)R(X,Y)$
d) $-1 \leq R(X,Y) \leq 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, $b \neq 0$ tak, že $P(Y = a+bX) = 1$, přičemž $R(X,Y) = 1$ pro $b > 0$ a $R(X,Y) = -1$ pro $b < 0$.

Z těchto vlastností plyne, že $R(X,Y)$ je vhodnou mírou těsnosti lineárního vztahu náhodných veličin X, Y.

8.4.2 Výběrový koeficient korelace

$R(X,Y)$ většinou nemůžeme počítat přímo, protože to vyžaduje znalost simultánního rozložení náhodného vektoru (X,Y). V praxi jsme většinou odkázáni na náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry $\quad M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i,$

výběrové rozptyly $\quad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2, \quad S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2,$

výběrovou kovarianci $\quad S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$

a s jejich pomocí zavedeme výběrový koeficient korelace $R_{12} = \dfrac{S_{12}}{S_1 S_2}$ (pro $S_1 S_2 > 0$).

Vlastnosti a), b), c), d) koeficientu korelace se přenášejí i na výběrový koeficient korelace.

8.4.3 Koeficient korelace dvourozměrného normálního rozložení

Nechť náhodný vektor (X,Y) má dvourozměrné normální rozložení s hustotou

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]},$$

přičemž $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = D(X)$, $\sigma_2^2 = D(Y)$, $\rho = R(X,Y)$.

Marginální hustoty jsou:

$$\varphi_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad \varphi_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

Je-li $\rho = 0$, pak pro $\forall(x,y) \in \mathbb{R}^2$: $\varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti.

Je-li $\rho \neq 0$, jsou náhodné veličiny X, Y stochasticky závislé. Je-li $\rho > 0$, říkáme, že jsou kladně korelované, je-li $\rho < 0$, říkáme, že jsou záporně korelované.

Upozornění:

V dalším textu budeme předpokládat, že náhodný výběr $(X_1,Y_1),\ldots,(X_n,Y_n)$ pochází z dvourozměrného normálního rozložení s parametry μ_1, μ_2, σ_1^2, σ_2^2, ρ.

8.4.4 Testování hypotézy o nezávislosti

Testujeme H_0: $\rho = 0$ proti oboustranné alternativě H_1: $\rho \neq 0$ (resp. proti levostranné alternativě H_1: $\rho < 0$ resp. proti pravostranné alternativě H_1: $\rho > 0$). Testová statistika má tvar:

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$$

Platí-li nulová hypotéza, pak $T \sim t(n-2)$. Kritický obor pro test H_0 proti oboustranné alternativě: $W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty)$, proti levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2)\rangle$ a proti pravostranné alternativě: $W = \langle t_{1-\alpha}(n-2), \infty)$. H_0 zamítáme na hladině významnosti α, když $T \in W$.

Není-li splněn předpoklad dvourozměrné normality, použijeme Spearmanův koeficient pořadové korelace.

8.4.5 Příklad

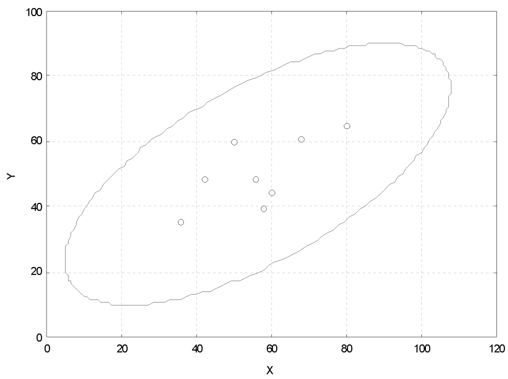
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení:

Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec.



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Testujeme H_0: $\rho = 0$ proti pravostranné alternativě H_1: $\rho > 0$.

Výpočtem zjistíme: $R_{12} = 0{,}6668$, $T = 2{,}1917$. V tabulkách najdeme $t_{0,95}(6) = 1{,}9432$. Kritický obor: $W = \langle 1{,}9432; \infty)$. Protože $T \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

Řešení pomocí sytému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných *1.TEST* a *2.TEST* a osmi případech. Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné *X 1.TEST*, *Y 2.TEST* – OK . Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu (viz obrázek výše)

Formát – Vš. Možnosti – Osa: Měřítka – Osa X – automatický mód změníme na manuální s minimem 0 a maximem 120. Totéž pro osu Y, ale stačí maximum 100.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky – Korelační matice – OK – 1.seznam proměnných *1.TEST*, *2.TEST*, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

	Korelace (priklad845) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD)										
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r^2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
1. test	56,25000	13,99745									
1. test	56,25000	13,99745	1,000000	1,000000			8	0,00000	1,000000	0,00000	1,000000
1. test	56,25000	13,99745									
2. test	50,00000	10,92834	0,666802	0,444625	2,191693	0,070909	8	20,71637	0,520598	13,54665	0,854067
2. test	50,00000	10,92834									
1. test	56,25000	13,99745	0,666802	0,444625	2,191693	0,070909	8	13,54665	0,854067	20,71637	0,520598
2. test	50,00000	10,92834									
2. test	50,00000	10,92834	1,000000	1,000000			8	0,00000	1,000000	0,00000	1,000000

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} ($r = 0{,}666802$, tzn. že mezi X a Y existuje nepříliš silná přímá lineární závislost), realizaci testové statistiky $t = 2{,}191693$ a p-hodnotu pro test hypotézy o nezávislosti ($p = 0{,}070909$, H_0 tedy nelze zamítnout na hladině významnosti 0,05).

Poznámka:

Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru. Statistika – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Počítat ρ pomocí r – Výpočet.

Shrnutí kapitoly

Při testování hypotézy o nezávislosti dvou náhodných veličin nominálního typu vycházíme z *kontingenční tabulky* sestrojené na základě znalosti náhodného výběru rozsahu n z dvourozměrného rozložení. Používáme testovou statistiku, která se za splnění *podmínek dobré aproximace* asymptoticky řídí Pearsonovým χ^2-rozložením. Intenzitu závislosti daných dvou veličin hodnotíme pomocí *Cramérova koeficientu*.

Máme-li dvě náhodné veličiny ordinálního typu, pak testujeme hypotézu o pořadové nezávislosti těchto dvou veličin pomocí *Spearmanova koeficientu pořadové korelace*, který slouží zároveň jako testová statistika i jako míra intenzity pořadové závislosti daných veličin. Pro menší rozsahy výběrů (orientačně $n < 30$) porovnáváme tento koeficient s tabelovanou kritickou hodnotou, pro větší rozsahy výběrů využijeme jeho asymptotické normality.

Při testování hypotézy o nezávislosti dvou náhodných veličin intervalového či poměrového typu, které se řídí dvourozměrným normálním rozložením, využijeme skutečnosti, že v tomto případě je stochastická nezávislost ekvivalentní nekorelovanosti těchto dvou veličin. Testová statistika vznikne transformací *výběrového koeficientu korelace* a v případě platnosti nulové hypotézy se řídí Studentovým rozložením.

Při zkoumání závislosti dvou náhodných veličin aspoň ordinálního typu je vhodné vytvořit dvourozměrný tečkový diagram a s jeho pomocí posoudit intenzitu a směr závislosti, případně orientačně ověřit dvourozměrnou normalitu dat.

Všechny popsané testy jsou implementovány v systému STATISTICA.

Kontrolní otázky

1. Jak testujeme nezávislost nominálních veličin? Jaké podmínky musí být splněny?
2. K čemu slouží Cramérův koeficient?
3. K čemu slouží Spearmanův koeficient pořadové korelace?
4. Uveďte vlastnosti výběrového koeficientu korelace.
5. Jak se na vzhledu dvourozměrného tečkového diagramu projeví, jsou-li náhodné veličiny X, Y kladně korelovány?
6. Pro náhodný výběr z dvourozměrného normálního rozložení popište test hypotézy o nezávislosti veličin X, Y.

Autokorekční test

1. Nechť $(X_1, Y_1), \dots, (X_{16}, Y_{16})$ je náhodný výběr z dvourozměrného normálního rozložení. Výběrový koeficient korelace $R(X,Y)$ nabyl hodnoty $-0{,}87$. Jestliže provedeme transformaci $U_i = 1 + 3X_i$, $V_i = -3 - Y_i$, $i = 1, \dots, 16$, jakou hodnotu nabude výběrový koeficient korelace $R(U,V)$?
   a) $-0{,}61$
   b) $0{,}87$
   c) $-0{,}87$
2. Pro 12 náhodně vybraných ojetých automobilů byl vypočten výběrový koeficient korelace mezi jejich stářím v měsících a počtem najetých kilometrů. Nabyl hodnoty 0,831. Předpokládáme, že data pocházejí z dvourozměrného normálního rozložení. Jaká je hodnota testové statistiky pro test nezávislosti obou veličin?
   a) 4,724
   b) 0,831
   c) 6,392
3. Ve čtyřpolní kontingenční tabulce jsou uvedeny tyto absolutní četnosti: $a = 5$, $b = 3$, $c = 6$, $d = 4$. Podíl šancí je
   a) 1,11
   b) 0,625
   c) 0,9
4. Pro dvourozměrný náhodný výběr rozsahu $n = 10$ z dvourozměrného normálního rozložení byl vypočten výběrový koeficient korelace. Nabyl hodnoty $-0{,}94$. Co lze usoudit o vztahu náhodných veličin X a Y?
   a) S růstem hodnot jedné náhodné veličiny hodnoty druhé náhodné veličiny lineárně rostou.
   b) Veličiny X a Y jsou nezávislé.
   c) S růstem hodnot jedné náhodné veličiny hodnoty druhé náhodné veličiny lineárně klesají.
5. Nechť dvourozměrný náhodný výběr pochází z dvourozměrného rozložení, které je výrazně odlišné od normálního. Chceme-li testovat hypotézu, že náhodné veličiny X a Y, které jsou poměrového typu, jsou nezávislé, použijeme testovou statistiku, která je založena na

a) Cramérově koeficientu
b) Spearmanově koeficientu pořadové korelace
c) výběrovém koeficientu korelace.

Správné odpovědi: 1c) 2a) 3a) 4c) 5b)

Příklady

1. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

Výsledek:
Podmínky dobré aproximace jsou splněny, pouze jedna teoretická četnost klesne pod 5. Testová statistika se realizuje hodnotou 3,5, počet stupňů volnosti $= 2$, kritický obor je $W = \langle 5{,}991; \infty)$. Hypotézu o nezávislosti pohlaví a pedagogické hodnosti tedy nezamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient $V = 0{,}187$.

2. Dvanáct různých softwarových firem nabízí programy pro vedení účetnictví. Programy byly posouzeny odbornou komisí a komisí složenou z profesionálních účetních. Výsledky v 1. a 2. komisi: $(6,4)$, $(7,5)$, $(1,2)$, $(8,10)$, $(4,6)$, $(2{,}5;1)$, $(9,7)$, $(12,11)$, $(10,8)$, $(2{,}5;3)$, $(5,12)$, $(11,9)$. Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu o nezávislosti pořadí v obou komisích.

Výsledek:
Spearmanův koeficent pořadové korelace je 0,715, kritická hodnota pro $n = 12$ a $\alpha = 0{,}05$ je 0,576. H_0 zamítáme na hladině významnosti 0,05 ve prospěch oboustranné alternativy.

3. V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc (veličina X) a počet zhotovených výrobků (veličina Y). Orientačně ověřte dvourozměrnou normalitu dat, vypočtěte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti veličin X a Y.

X	20	21	18	17	20	18	19	21	20	14	16	19	21	15	15
Y	92	93	83	80	91	85	82	98	90	60	73	86	96	64	81

Výsledek:
Vzhled dvourozměrného tečkového diagramu svědčí o tom, že předpoklad dvourozměrné normality je oprávněný. Výběrový koeficient korelace je 0,927, testová statistika se realizuje hodnotou 8,597, kritický obor je $W = (-\infty, -3{,}012\rangle \cup \langle 3{,}012, \infty))$. Hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01.

4. 100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

pohlaví	nápoj	
	A	*B*
muž	20	30
žena	30	20

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Výsledek:
V našem případě se jedná o jednostrannou závislost, zajímáme se tedy o Fisher exact, one tailed. Ta je 0,03567. Protože p-hodnota je menší nebo rovna 0,05, zamítáme na hladině významnosti hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

5. V následující tabulce jsou uvedeny číselné realizace a absolutní četnosti náhodného výběru $(X_1,Y_1),(X_1,Y_2),\ldots,(X_{62},Y_{62})$ z dvourozměrného rozložení:

x	y						
	1	3	5	7	9	11	13
15	0	0	0	0	1	2	1
25	0	0	0	5	4	2	0
35	0	0	5	8	2	0	0
45	0	5	6	4	0	0	0
55	3	5	3	0	0	0	0
65	4	2	0	0	0	0	0

Podle vzhledu dvourozměrného tečkového diagramu orientačně posuďte dvourozměrnou normalitu dat. Vypočtěte výběrový koeficient korelace a interpretujte ho. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X a Y.

Výsledek:
Protože tečky v dvourozměrném tečkovém diagramu vytvářejí elipsovitý obrazec, lze připustit dvourozměrnou normalitu. Výběrový koeficient korelace nabývá hodnoty $-0{,}899$, což znamená, že mezi veličinami X a Y existuje dosti silná nepřímá lineární závislost. Testová statistika se realizuje hodnotou $-13{,}6613$, odpovídající p-hodnota je velmi blízká 0, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Příloha A – Statistické tabulky

Distribuční funkce standardizovaného normálního rozložení

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
0,00	0,50000	0,50	0,69146	1,00	0,84134	1,50	0,93319
0,01	0,50399	0,51	0,69497	1,01	0,84375	1,51	0,93448
0,02	0,50798	0,52	0,69847	1,02	0,84614	1,52	0,93574
0,03	0,51197	0,53	0,70194	1,03	0,84850	1,53	0,93699
0,04	0,51595	0,54	0,70540	1,04	0,85083	1,54	0,93822
0,05	0,51994	0,55	0,70884	1,05	0,85314	1,55	0,93943
0,06	0,52392	0,56	0,71226	1,06	0,85543	1,56	0,94062
0,07	0,52790	0,57	0,71566	1,07	0,85769	1,57	0,94179
0,08	0,53188	0,58	0,71904	1,08	0,85993	1,58	0,94295
0,09	0,53586	0,59	0,72240	1,09	0,86214	1,59	0,94408
0,10	0,53983	0,60	0,72575	1,10	0,86433	1,60	0,94520
0,11	0,54380	0,61	0,72907	1,11	0,86650	1,61	0,94630
0,12	0,54776	0,62	0,73237	1,12	0,86864	1,62	0,94738
0,13	0,55172	0,63	0,73565	1,13	0,87076	1,63	0,94845
0,14	0,55567	0,64	0,73891	1,14	0,87286	1,64	0,94950
0,15	0,55962	0,65	0,74215	1,15	0,87493	1,65	0,95053
0,16	0,56356	0,66	0,74537	1,16	0,87698	1,66	0,95154
0,17	0,56749	0,67	0,74857	1,17	0,87900	1,67	0,95254
0,18	0,57142	0,68	0,75175	1,18	0,88100	1,68	0,95352
0,19	0,57535	0,69	0,75490	1,19	0,88298	1,69	0,95449
0,20	0,57926	0,70	0,75804	1,20	0,88493	1,70	0,95543
0,21	0,58317	0,71	0,76115	1,21	0,88686	1,71	0,95637
0,22	0,58706	0,72	0,76424	1,22	0,88877	1,72	0,95728
0,23	0,59095	0,73	0,76730	1,23	0,89065	1,73	0,95818
0,24	0,59483	0,74	0,77035	1,24	0,89251	1,74	0,95907
0,25	0,59871	0,75	0,77337	1,25	0,89435	1,75	0,95994
0,26	0,60257	0,76	0,77637	1,26	0,89617	1,76	0,96080
0,27	0,60642	0,77	0,77935	1,27	0,89796	1,77	0,96164
0,28	0,61026	0,78	0,78230	1,28	0,89973	1,78	0,96246
0,29	0,61409	0,79	0,78524	1,29	0,90147	1,79	0,96327
0,30	0,61791	0,80	0,78814	1,30	0,90320	1,80	0,96407
0,31	0,62172	0,81	0,79103	1,31	0,90490	1,81	0,96485
0,32	0,62552	0,82	0,79389	1,32	0,90658	1,82	0,96562
0,33	0,62930	0,83	0,79673	1,33	0,90824	1,83	0,96638
0,34	0,63307	0,84	0,79955	1,34	0,90988	1,84	0,96712
0,35	0,63683	0,85	0,80234	1,35	0,91149	1,85	0,96784
0,36	0,64058	0,86	0,80511	1,36	0,91309	1,86	0,96856
0,37	0,64431	0,87	0,80785	1,37	0,91466	1,87	0,96926
0,38	0,64803	0,88	0,81057	1,38	0,91621	1,88	0,96995
0,39	0,65173	0,89	0,81327	1,39	0,91774	1,89	0,97062
0,40	0,65542	0,90	0,81594	1,40	0,91924	1,90	0,97128
0,41	0,65910	0,91	0,81859	1,41	0,92073	1,91	0,97193
0,42	0,66276	0,92	0,82121	1,42	0,92220	1,92	0,97257
0,43	0,66640	0,93	0,82381	1,43	0,92364	1,93	0,97320
0,44	0,67003	0,94	0,82639	1,44	0,92507	1,94	0,97381
0,45	0,67364	0,95	0,82894	1,45	0,92647	1,95	0,97441
0,46	0,67724	0,96	0,83147	1,46	0,92785	1,96	0,97500
0,47	0,68082	0,97	0,83398	1,47	0,92922	1,97	0,97558
0,48	0,68439	0,98	0,83646	1,48	0,93056	1,98	0,97615
0,49	0,68793	0,99	0,83891	1,49	0,93189	1,99	0,97670

$$\Phi(-u) = 1 - \Phi(u)$$

Distribuční funkce standardizovaného normálního rozložení

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
2,00	0,97725	2,50	0,99379	3,00	0,99865	3,50	0,99977
2,01	0,97778	2,51	0,99396	3,01	0,99869	3,51	0,99978
2,02	0,97831	2,52	0,99413	3,02	0,99874	3,52	0,99978
2,03	0,97882	2,53	0,99430	3,03	0,99878	3,53	0,99979
2,04	0,97932	2,54	0,99446	3,04	0,99882	3,54	0,99980
2,05	0,97982	2,55	0,99461	3,05	0,99886	3,55	0,99981
2,06	0,98030	2,56	0,99477	3,06	0,99889	3,56	0,99981
2,07	0,98077	2,57	0,99492	3,07	0,99893	3,57	0,99982
2,08	0,98124	2,58	0,99506	3,08	0,99897	3,58	0,99983
2,09	0,98169	2,59	0,99520	3,09	0,99900	3,59	0,99983
2,10	0,98214	2,60	0,99534	3,10	0,99903	3,60	0,99984
2,11	0,98257	2,61	0,99547	3,11	0,99906	3,61	0,99985
2,12	0,98300	2,62	0,99560	3,12	0,99910	3,62	0,99985
2,13	0,98341	2,63	0,99573	3,13	0,99913	3,63	0,99986
2,14	0,98382	2,64	0,99585	3,14	0,99916	3,64	0,99986
2,15	0,98422	2,65	0,99598	3,15	0,99918	3,65	0,99987
2,16	0,98461	2,66	0,99609	3,16	0,99921	3,66	0,99987
2,17	0,98500	2,67	0,99621	3,17	0,99924	3,67	0,99988
2,18	0,98537	2,68	0,99632	3,18	0,99926	3,68	0,99988
2,19	0,98574	2,69	0,99643	3,19	0,99929	3,69	0,99989
2,20	0,98610	2,70	0,99653	3,20	0,99931	3,70	0,99989
2,21	0,98645	2,71	0,99664	3,21	0,99934	3,71	0,99990
2,22	0,98679	2,72	0,99674	3,22	0,99936	3,72	0,99990
2,23	0,98713	2,73	0,99683	3,23	0,99938	3,73	0,99990
2,24	0,98745	2,74	0,99693	3,24	0,99940	3,74	0,99991
2,25	0,98778	2,75	0,99702	3,25	0,99942	3,75	0,99991
2,26	0,98809	2,76	0,99711	3,26	0,99944	3,76	0,99992
2,27	0,98840	2,77	0,99720	3,27	0,99946	3,77	0,99992
2,28	0,98870	2,78	0,99728	3,28	0,99948	3,78	0,99992
2,29	0,98899	2,79	0,99736	3,29	0,99950	3,79	0,99992
2,30	0,98928	2,80	0,99744	3,30	0,99952	3,80	0,99993
2,31	0,98956	2,81	0,99752	3,31	0,99953	3,81	0,99993
2,32	0,98983	2,82	0,99760	3,32	0,99955	3,82	0,99993
2,33	0,99010	2,83	0,99767	3,33	0,99957	3,83	0,99994
2,34	0,99036	2,84	0,99774	3,34	0,99958	3,84	0,99994
2,35	0,99061	2,85	0,99781	3,35	0,99960	3,85	0,99994
2,36	0,99086	2,86	0,99788	3,36	0,99961	3,86	0,99994
2,37	0,99111	2,87	0,99795	3,37	0,99962	3,87	0,99995
2,38	0,99134	2,88	0,99801	3,38	0,99964	3,88	0,99995
2,39	0,99158	2,89	0,99807	3,39	0,99965	3,89	0,99995
2,40	0,99180	2,90	0,99813	3,40	0,99966	3,90	0,99995
2,41	0,99202	2,91	0,99819	3,41	0,99968	3,91	0,99995
2,42	0,99224	2,92	0,99825	3,42	0,99969	3,92	0,99996
2,43	0,99245	2,93	0,99831	3,43	0,99970	3,93	0,99996
2,44	0,99266	2,94	0,99836	3,44	0,99971	3,94	0,99996
2,45	0,99286	2,95	0,99841	3,45	0,99972	3,95	0,99996
2,46	0,99305	2,96	0,99846	3,46	0,99973	3,96	0,99996
2,47	0,99324	2,97	0,99851	3,47	0,99974	3,97	0,99996
2,48	0,99343	2,98	0,99856	3,48	0,99975	3,98	0,99997
2,49	0,99361	2,99	0,99861	3,49	0,99976	3,99	0,99997

$$\Phi(-u) = 1 - \Phi(u)$$

Kvantily standardizovaného normálního rozložení

α	u_α	α	u_α	α	u_α	α	u_α
0,500	0,00000	0,850	1,03643	0,930	1,47579	0,965	1,81191
0,510	0,02507	0,860	1,08032	0,931	1,48328	0,966	1,82501
0,520	0,05015	0,870	1,12639	0,932	1,49085	0,967	1,83842
0,530	0,07527	0,880	1,17499	0,933	1,49851	0,968	1,85218
0,540	0,10043	0,890	1,22653	0,934	1,50626	0,969	1,86630
0,550	0,12566	0,900	1,28155	0,935	1,51410	0,970	1,88079
0,560	0,15097	0,901	1,28727	0,936	1,52204	0,971	1,89570
0,570	0,17637	0,902	1,29303	0,937	1,53007	0,972	1,91104
0,580	0,20189	0,903	1,29884	0,938	1,53820	0,973	1,92684
0,590	0,22754	0,904	1,30469	0,939	1,54643	0,974	1,94313
0,600	0,25335	0,905	1,31058	0,940	1,55477	0,975	1,95996
0,610	0,27932	0,906	1,31652	0,941	1,56322	0,976	1,97737
0,620	0,30548	0,907	1,32251	0,942	1,57179	0,977	1,99539
0,630	0,33185	0,908	1,32854	0,943	1,58047	0,978	2,01409
0,640	0,35846	0,909	1,33462	0,944	1,58927	0,979	2,03352
0,650	0,38532	0,910	1,34076	0,945	1,59819	0,980	2,05375
0,660	0,41246	0,911	1,34694	0,946	1,60725	0,981	2,07485
0,670	0,43991	0,912	1,35317	0,947	1,61644	0,982	2,09693
0,680	0,46770	0,913	1,35946	0,948	1,62576	0,983	2,12007
0,690	0,49585	0,914	1,36581	0,949	1,63523	0,984	2,14441
0,700	0,52440	0,915	1,37220	0,950	1,64485	0,985	2,17009
0,710	0,55338	0,916	1,37866	0,951	1,65463	0,986	2,19729
0,720	0,58284	0,917	1,38517	0,952	1,66456	0,987	2,22621
0,730	0,61281	0,918	1,39174	0,953	1,67466	0,988	2,25713
0,740	0,64335	0,919	1,39838	0,954	1,68494	0,989	2,29037
0,750	0,67449	0,920	1,40507	0,955	1,69540	0,990	2,32635
0,760	0,70630	0,921	1,41183	0,956	1,70604	0,991	2,36562
0,770	0,73885	0,922	1,41865	0,957	1,71689	0,992	2,40892
0,780	0,77219	0,923	1,42554	0,958	1,72793	0,993	2,45726
0,790	0,80642	0,924	1,43250	0,959	1,73920	0,994	2,51214
0,800	0,84162	0,925	1,43953	0,960	1,75069	0,995	2,57583
0,810	0,87790	0,926	1,44663	0,961	1,76241	0,996	2,65207
0,820	0,91537	0,927	1,45381	0,962	1,77438	0,997	2,74778
0,830	0,95417	0,928	1,46106	0,963	1,78661	0,998	2,87816
0,840	0,99446	0,929	1,46838	0,964	1,79912	0,999	3,09023

Kvantily Pearsonova rozložení

			α		
n	0,001	0,005	0,010	0,025	0,050
	0,001	0,005	0,010	0,025	0,050
1	0,000	0,000	0,000	0,001	0,004
2	0,002	0,010	0,020	0,051	0,103
3	0,024	0,072	0,115	0,216	0,352
4	0,091	0,207	0,297	0,484	0,711
5	0,210	0,412	0,554	0,831	1,145
6	0,381	0,676	0,872	1,237	1,635
7	0,598	0,989	1,239	1,690	2,167
8	0,857	1,344	1,646	2,180	2,733
9	1,152	1,735	2,088	2,700	3,325
10	1,479	2,156	2,558	3,247	3,940
11	1,834	2,603	3,053	3,816	4,575
12	2,214	3,074	3,571	4,404	5,226
13	2,617	3,565	4,107	5,009	5,892
14	3,041	4,075	4,660	5,629	6,571
15	3,483	4,601	5,229	6,262	7,261
16	3,942	5,142	5,812	6,908	7,962
17	4,416	5,697	6,408	7,564	8,672
18	4,905	6,265	7,015	8,231	9,390
19	5,407	6,844	7,633	8,907	10,117
20	5,921	7,434	8,260	9,591	10,851
21	6,447	8,034	8,897	10,283	11,591
22	6,983	8,643	9,542	10,982	12,338
23	7,529	9,260	10,196	11,689	13,091
24	8,085	9,886	10,856	12,401	13,848
25	8,649	10,520	11,524	13,120	14,611
26	9,222	11,160	12,198	13,844	15,379
27	9,803	11,808	12,879	14,573	16,151
28	10,391	12,461	13,565	15,308	16,928
29	10,986	13,121	14,256	16,047	17,708
30	11,588	13,787	14,953	16,791	18,493
35	14,688	17,192	18,509	20,569	22,465
40	17,916	20,707	22,164	24,433	26,509
45	21,251	24,311	25,901	28,366	30,612
50	24,674	27,991	29,707	32,357	34,764
55	28,173	31,735	33,570	36,398	38,958
60	31,738	35,534	37,485	40,482	43,188
65	35,362	39,383	41,444	44,603	47,450
70	39,036	43,275	45,442	48,758	51,739
75	42,757	47,206	49,475	52,942	56,054
80	46,520	51,172	53,540	57,153	60,391
85	50,320	55,170	57,634	61,389	64,749
90	54,155	59,196	61,754	65,647	69,126
95	58,022	63,250	65,898	69,925	73,520
100	61,918	67,328	70,065	74,222	77,929

Kvantily Pearsonova rozložení

	α				
n	0,950	0,975	0,990	0,995	0,999
1	3,841	5,024	6,635	7,879	10,828
2	5,991	7,378	9,210	10,597	13,816
3	7,815	9,348	11,345	12,838	16,266
4	9,488	11,143	13,277	14,860	18,467
5	11,070	12,833	15,086	16,750	20,515
6	12,592	14,449	16,812	18,548	22,458
7	14,067	16,013	18,475	20,278	24,322
8	15,507	17,535	20,090	21,955	26,124
9	16,919	19,023	21,666	23,589	27,877
10	18,307	20,483	23,209	25,188	29,588
11	19,675	21,920	24,725	26,757	31,264
12	21,026	23,337	26,217	28,300	32,909
13	22,362	24,736	27,688	29,819	34,528
14	23,685	26,119	29,141	31,319	36,123
15	24,996	27,488	30,578	32,801	37,697
16	26,296	28,845	32,000	34,267	39,252
17	27,587	30,191	33,409	35,718	40,790
18	28,869	31,526	34,805	37,156	42,312
19	30,144	32,852	36,191	38,582	43,820
20	31,410	34,170	37,566	39,997	45,315
21	32,671	35,479	38,932	41,401	46,797
22	33,924	36,781	40,289	42,796	48,268
23	35,172	38,076	41,638	44,181	49,728
24	36,415	39,364	42,980	45,559	51,179
25	37,652	40,646	44,314	46,928	52,620
26	38,885	41,923	45,642	48,290	54,052
27	40,113	43,195	46,963	49,645	55,476
28	41,337	44,461	48,278	50,993	56,892
29	42,557	45,722	49,588	52,336	58,301
30	43,773	46,979	50,892	53,672	59,703
35	49,802	53,203	57,342	60,275	66,619
40	55,758	59,342	63,691	66,766	73,402
45	61,656	65,410	69,957	73,166	80,077
50	67,505	71,420	76,154	79,490	86,661
55	73,311	77,380	82,292	85,749	93,168
60	79,082	83,298	88,379	91,952	99,607
65	84,821	89,177	94,422	98,105	105,988
70	90,531	95,023	100,425	104,215	112,317
75	96,217	100,839	106,393	110,286	118,599
80	101,879	106,629	112,329	116,321	124,839
85	107,522	112,393	118,236	122,325	131,041
90	113,145	118,136	124,116	128,299	137,208
95	118,752	123,858	129,973	134,247	143,344
100	124,342	129,561	135,807	140,169	149,449

Kvantily Studentova rozložení

	α					
n	0,900	0,950	0,975	0,990	0,995	0,999
1	3,0777	6,3138	12,7062	31,8205	63,6567	318,3088
2	1,8856	2,9200	4,3027	6,9646	9,9248	22,3271
3	1,6377	2,3534	3,1824	4,5407	5,8409	10,2145
4	1,5332	2,1318	2,7764	3,7469	4,6041	7,1732
5	1,4759	2,0150	2,5706	3,3649	4,0321	5,8934
6	1,4398	1,9432	2,4469	3,1427	3,7074	5,2076
7	1,4149	1,8946	2,3646	2,9980	3,4995	4,7853
8	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008
9	1,3830	1,8331	2,2622	2,8214	3,2498	4,2968
10	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437
11	1,3634	1,7959	2,2010	2,7181	3,1058	4,0247
12	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296
13	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520
14	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874
15	1,3406	1,7531	2,1314	2,6025	2,9467	3,7328
16	1,3368	1,7459	2,1199	2,5835	2,9208	3,6862
17	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458
18	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105
19	1,3277	1,7291	2,0930	2,5395	2,8609	3,5794
20	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518
21	1,3232	1,7207	2,0796	2,5176	2,8314	3,5272
22	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050
23	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850
24	1,3178	1,7109	2,0639	2,4922	2,7969	3,4668
25	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502
26	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350
27	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210
28	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082
29	1,3114	1,6991	2,0452	2,4620	2,7564	3,3962
30	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852
∞	1,2816	1,6449	1,9600	2,3263	2,5758	3,0000

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}95$

n_2	n_1 = 1	2	3	4	5	6	7
1	161,4500	199,5000	215,7074	224,5832	230,1619	233,9860	236,7684
2	18,5128	19,0000	19,1643	19,2468	19,2964	19,3295	19,3532
3	10,1280	9,5521	9,2766	9,1172	9,0135	8,9406	8,8867
4	7,7086	6,9443	6,5914	6,3882	6,2561	6,1631	6,0942
5	6,6079	5,7861	5,4095	5,1922	5,0503	4,9503	4,8759
6	5,9874	5,1433	4,7571	4,5337	4,3874	4,2839	4,2067
7	5,5914	4,7374	4,3468	4,1203	3,9715	3,8660	3,7870
8	5,3177	4,4590	4,0662	3,8379	3,6875	3,5806	3,5005
9	5,1174	4,2565	3,8625	3,6331	3,4817	3,3738	3,2927
10	4,9646	4,1028	3,7083	3,4780	3,3258	3,2172	3,1355
11	4,8443	3,9823	3,5874	3,3567	3,2039	3,0946	3,0123
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,9134
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,8321
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,8477	2,7642
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,7066
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,7413	2,6572
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,6987	2,6143
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,6613	2,5767
19	4,3807	3,5219	3,1274	2,8951	2,7401	2,6283	2,5435
20	4,3512	3,4928	3,0984	2,8661	2,7109	2,5990	2,5140
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4876
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,5491	2,4638
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,5277	2,4422
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,4226
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,4904	2,4047
26	4,2252	3,3690	2,9752	2,7426	2,5868	2,4741	2,3883
27	4,2100	3,3541	2,9604	2,7278	2,5719	2,4591	2,3732
28	4,1960	3,3404	2,9467	2,7141	2,5581	2,4453	2,3593
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,4324	2,3463
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,4205	2,3343
40	4,0847	3,2317	2,8387	2,6060	2,4495	2,3359	2,2490
60	4,0012	3,1504	2,7581	2,5252	2,3683	2,2541	2,1665
80	3,9604	3,1108	2,7188	2,4859	2,3287	2,2142	2,1263
120	3,9201	3,0718	2,6802	2,4472	2,2899	2,1750	2,0868
∞	3,8415	2,9957	2,6049	2,3719	2,2141	2,0986	2,0096

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}95$

	n_1						
n_2	8	9	10	11	12	13	14
1	238,8827	240,5433	241,8818	242,9835	243,9060	244,6899	245,3640
2	19,3710	19,3848	19,3959	19,4050	19,4125	19,4189	19,4244
3	8,8452	8,8123	8,7855	8,7633	8,7446	8,7287	8,7149
4	6,0410	5,9988	5,9644	5,9358	5,9117	5,8911	5,8733
5	4,8183	4,7725	4,7351	4,7040	4,6777	4,6552	4,6358
6	4,1468	4,0990	4,0600	4,0274	3,9999	3,9764	3,9559
7	3,7257	3,6767	3,6365	3,6030	3,5747	3,5503	3,5292
8	3,4381	3,3881	3,3472	3,3130	3,2839	3,2590	3,2374
9	3,2296	3,1789	3,1373	3,1025	3,0729	3,0475	3,0255
10	3,0717	3,0204	2,9782	2,9430	2,9130	2,8872	2,8647
11	2,9480	2,8962	2,8536	2,8179	2,7876	2,7614	2,7386
12	2,8486	2,7964	2,7534	2,7173	2,6866	2,6602	2,6371
13	2,7669	2,7144	2,6710	2,6347	2,6037	2,5769	2,5536
14	2,6987	2,6458	2,6022	2,5655	2,5342	2,5073	2,4837
15	2,6408	2,5876	2,5437	2,5068	2,4753	2,4481	2,4244
16	2,5911	2,5377	2,4935	2,4564	2,4247	2,3973	2,3733
17	2,5480	2,4943	2,4499	2,4126	2,3807	2,3531	2,3290
18	2,5102	2,4563	2,4117	2,3742	2,3421	2,3143	2,2900
19	2,4768	2,4227	2,3779	2,3402	2,3080	2,2800	2,2556
20	2,4471	2,3928	2,3479	2,3100	2,2776	2,2495	2,2250
21	2,4205	2,3660	2,3210	2,2829	2,2504	2,2222	2,1975
22	2,3965	2,3419	2,2967	2,2585	2,2258	2,1975	2,1727
23	2,3748	2,3201	2,2747	2,2364	2,2036	2,1752	2,1502
24	2,3551	2,3002	2,2547	2,2163	2,1834	2,1548	2,1298
25	2,3371	2,2821	2,2365	2,1979	2,1649	2,1362	2,1111
26	2,3205	2,2655	2,2197	2,1811	2,1479	2,1192	2,0939
27	2,3053	2,2501	2,2043	2,1655	2,1323	2,1035	2,0781
28	2,2913	2,2360	2,1900	2,1512	2,1179	2,0889	2,0635
29	2,2783	2,2229	2,1768	2,1379	2,1045	2,0755	2,0500
30	2,2662	2,2107	2,1646	2,1256	2,0921	2,0630	2,0374
40	2,1802	2,1240	2,0772	2,0376	2,0035	1,9738	1,9476
60	2,0970	2,0401	1,9926	1,9522	1,9174	1,8870	1,8602
80	2,0564	1,9991	1,9512	1,9105	1,8753	1,8445	1,8174
120	2,0164	1,9588	1,9105	1,8693	1,8337	1,8026	1,7750
∞	1,9384	1,8799	1,8307	1,7886	1,7522	1,7202	1,6918

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}95$

n_2	n_1 = 15	16	17	18	19	20	25
1	245,9499	246,4639	246,9184	247,3232	247,6861	248,0131	249,2601
2	19,4291	19,4333	19,4370	19,4402	19,4431	19,4458	19,4558
3	8,7029	8,6923	8,6829	8,6745	8,6670	8,6602	8,6341
4	5,8578	5,8441	5,8320	5,8211	5,8114	5,8025	5,7687
5	4,6188	4,6038	4,5904	4,5785	4,5678	4,5581	4,5209
6	3,9381	3,9223	3,9083	3,8957	3,8844	3,8742	3,8348
7	3,5107	3,4944	3,4799	3,4669	3,4551	3,4445	3,4036
8	3,2184	3,2016	3,1867	3,1733	3,1613	3,1503	3,1081
9	3,0061	2,9890	2,9737	2,9600	2,9477	2,9365	2,8932
10	2,8450	2,8276	2,8120	2,7980	2,7854	2,7740	2,7298
11	2,7186	2,7009	2,6851	2,6709	2,6581	2,6464	2,6014
12	2,6169	2,5989	2,5828	2,5684	2,5554	2,5436	2,4977
13	2,5331	2,5149	2,4987	2,4841	2,4709	2,4589	2,4123
14	2,4630	2,4446	2,4282	2,4134	2,4000	2,3879	2,3407
15	2,4034	2,3849	2,3683	2,3533	2,3398	2,3275	2,2797
16	2,3522	2,3335	2,3167	2,3016	2,2880	2,2756	2,2272
17	2,3077	2,2888	2,2719	2,2567	2,2429	2,2304	2,1815
18	2,2686	2,2496	2,2325	2,2172	2,2033	2,1906	2,1413
19	2,2341	2,2149	2,1977	2,1823	2,1683	2,1555	2,1057
20	2,2033	2,1840	2,1667	2,1511	2,1370	2,1242	2,0739
21	2,1757	2,1563	2,1389	2,1232	2,1090	2,0960	2,0454
22	2,1508	2,1313	2,1138	2,0980	2,0837	2,0707	2,0196
23	2,1282	2,1086	2,0910	2,0751	2,0608	2,0476	1,9963
24	2,1077	2,0880	2,0703	2,0543	2,0399	2,0267	1,9750
25	2,0889	2,0691	2,0513	2,0353	2,0207	2,0075	1,9554
26	2,0716	2,0518	2,0339	2,0178	2,0032	1,9898	1,9375
27	2,0558	2,0358	2,0179	2,0017	1,9870	1,9736	1,9210
28	2,0411	2,0210	2,0030	1,9868	1,9720	1,9586	1,9057
29	2,0275	2,0073	1,9893	1,9730	1,9581	1,9446	1,8915
30	2,0148	1,9946	1,9765	1,9601	1,9452	1,9317	1,8782
40	1,9245	1,9037	1,8851	1,8682	1,8529	1,8389	1,7835
60	1,8364	1,8151	1,7959	1,7784	1,7625	1,7480	1,6902
80	1,7932	1,7716	1,7520	1,7342	1,7180	1,7032	1,6440
120	1,7505	1,7285	1,7085	1,6904	1,6739	1,6587	1,5980
∞	1,6640	1,6435	1,6228	1,6038	1,5865	1,5705	1,5061

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}95$

n_2	n_1 30	40	60	80	120	∞
1	250,0952	251,1432	252,1957	252,7237	253,2529	254,3100
2	19,4624	19,4707	19,4791	19,4832	19,4874	19,4960
3	8,6166	8,5944	8,5720	8,5607	8,5494	8,5264
4	5,7459	5,7170	5,6877	5,6730	5,6581	5,6281
5	4,4957	4,4638	4,4314	4,4150	4,3985	4,3650
6	3,8082	3,7743	3,7398	3,7223	3,7047	3,6689
7	3,3758	3,3404	3,3043	3,2860	3,2674	3,2298
8	3,0794	3,0428	3,0053	2,9862	2,9669	2,9276
9	2,8637	2,8259	2,7872	2,7675	2,7475	2,7067
10	2,6996	2,6609	2,6211	2,6008	2,5801	2,5379
11	2,5705	2,5309	2,4901	2,4692	2,4480	2,4045
12	2,4663	2,4259	2,3842	2,3628	2,3410	2,2962
13	2,3803	2,3392	2,2966	2,2747	2,2524	2,2064
14	2,3082	2,2664	2,2229	2,2006	2,1778	2,1307
15	2,2468	2,2043	2,1601	2,1373	2,1141	2,0658
16	2,1938	2,1507	2,1058	2,0826	2,0589	2,0096
17	2,1477	2,1040	2,0584	2,0348	2,0107	1,9604
18	2,1071	2,0629	2,0166	1,9927	1,9681	1,9168
19	2,0712	2,0264	1,9795	1,9552	1,9302	1,8780
20	2,0391	1,9938	1,9464	1,9217	1,8963	1,8432
21	2,0102	1,9645	1,9165	1,8915	1,8657	1,8117
22	1,9842	1,9380	1,8894	1,8641	1,8380	1,7831
23	1,9605	1,9139	1,8648	1,8392	1,8128	1,7570
24	1,9390	1,8920	1,8424	1,8164	1,7896	1,7330
25	1,9192	1,8718	1,8217	1,7955	1,7684	1,7110
26	1,9010	1,8533	1,8027	1,7762	1,7488	1,6906
27	1,8842	1,8361	1,7851	1,7584	1,7306	1,6717
28	1,8687	1,8203	1,7689	1,7418	1,7138	1,6541
29	1,8543	1,8055	1,7537	1,7264	1,6981	1,6376
30	1,8409	1,7918	1,7396	1,7121	1,6835	1,6223
40	1,7444	1,6928	1,6373	1,6077	1,5766	1,5089
60	1,6491	1,5943	1,5343	1,5019	1,4673	1,3893
80	1,6017	1,5449	1,4821	1,4477	1,4107	1,3247
120	1,5543	1,4952	1,4290	1,3922	1,3519	1,2539
∞	1,4591	1,3940	1,3180	1,2735	1,2214	1,0000

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}975$

n_2	n_1 = 1	2	3	4	5	6	7
1	647,7890	799,5000	864,1630	899,5833	921,8479	937,1111	948,2169
2	38,5063	39,0000	39,1655	39,2484	39,2982	39,3315	39,3552
3	17,4434	16,0441	15,4392	15,1010	14,8848	14,7347	14,6244
4	12,2179	10,6491	9,9792	9,6045	9,3645	9,1973	9,0741
5	10,0070	8,4336	7,7636	7,3879	7,1464	6,9777	6,8531
6	8,8131	7,2599	6,5988	6,2272	5,9876	5,8198	5,6955
7	8,0727	6,5415	5,8898	5,5226	5,2852	5,1186	4,9949
8	7,5709	6,0595	5,4160	5,0526	4,8173	4,6517	4,5286
9	7,2093	5,7147	5,0781	4,7181	4,4844	4,3197	4,1970
10	6,9367	5,4564	4,8256	4,4683	4,2361	4,0721	3,9498
11	6,7241	5,2559	4,6300	4,2751	4,0440	3,8807	3,7586
12	6,5538	5,0959	4,4742	4,1212	3,8911	3,7283	3,6065
13	6,4143	4,9653	4,3472	3,9959	3,7667	3,6043	3,4827
14	6,2979	4,8567	4,2417	3,8919	3,6634	3,5014	3,3799
15	6,1995	4,7650	4,1528	3,8043	3,5764	3,4147	3,2934
16	6,1151	4,6867	4,0768	3,7294	3,5021	3,3406	3,2194
17	6,0420	4,6189	4,0112	3,6648	3,4379	3,2767	3,1556
18	5,9781	4,5597	3,9539	3,6083	3,3820	3,2209	3,0999
19	5,9216	4,5075	3,9034	3,5587	3,3327	3,1718	3,0509
20	5,8715	4,4613	3,8587	3,5147	3,2891	3,1283	3,0074
21	5,8266	4,4199	3,8188	3,4754	3,2501	3,0895	2,9686
22	5,7863	4,3828	3,7829	3,4401	3,2151	3,0546	2,9338
23	5,7498	4,3492	3,7505	3,4083	3,1835	3,0232	2,9023
24	5,7166	4,3187	3,7211	3,3794	3,1548	2,9946	2,8738
25	5,6864	4,2909	3,6943	3,3530	3,1287	2,9685	2,8478
26	5,6586	4,2655	3,6697	3,3289	3,1048	2,9447	2,8240
27	5,6331	4,2421	3,6472	3,3067	3,0828	2,9228	2,8021
28	5,6096	4,2205	3,6264	3,2863	3,0626	2,9027	2,7820
29	5,5878	4,2006	3,6072	3,2674	3,0438	2,8840	2,7633
30	5,5675	4,1821	3,5894	3,2499	3,0265	2,8667	2,7460
40	5,4239	4,0510	3,4633	3,1261	2,9037	2,7444	2,6238
60	5,2856	3,9253	3,3425	3,0077	2,7863	2,6274	2,5068
80	5,2184	3,8643	3,2841	2,9504	2,7295	2,5708	2,4502
120	5,1523	3,8046	3,2269	2,8943	2,6740	2,5154	2,3948
∞	5,0239	3,6889	3,1161	2,7858	2,5665	2,4082	2,2875

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}975$

n_2	n_1 8	9	10	11	12	13	14
1	956,6562	963,2846	968,6274	973,0252	976,7080	979,8368	982,5278
2	39,3730	39,3869	39,3980	39,4071	39,4146	39,4210	39,4265
3	14,5399	14,4731	14,4189	14,3742	14,3366	14,3045	14,2768
4	8,9796	8,9047	8,8439	8,7935	8,7512	8,7150	8,6838
5	6,7572	6,6811	6,6192	6,5678	6,5245	6,4876	6,4556
6	5,5996	5,5234	5,4613	5,4098	5,3662	5,3290	5,2968
7	4,8993	4,8232	4,7611	4,7095	4,6658	4,6285	4,5961
8	4,4333	4,3572	4,2951	4,2434	4,1997	4,1622	4,1297
9	4,1020	4,0260	3,9639	3,9121	3,8682	3,8306	3,7980
10	3,8549	3,7790	3,7168	3,6649	3,6209	3,5832	3,5504
11	3,6638	3,5879	3,5257	3,4737	3,4296	3,3917	3,3588
12	3,5118	3,4358	3,3736	3,3215	3,2773	3,2393	3,2062
13	3,3880	3,3120	3,2497	3,1975	3,1532	3,1150	3,0819
14	3,2853	3,2093	3,1469	3,0946	3,0502	3,0119	2,9786
15	3,1987	3,1227	3,0602	3,0078	2,9633	2,9249	2,8915
16	3,1248	3,0488	2,9862	2,9337	2,8890	2,8506	2,8170
17	3,0610	2,9849	2,9222	2,8696	2,8249	2,7863	2,7526
18	3,0053	2,9291	2,8664	2,8137	2,7689	2,7302	2,6964
19	2,9563	2,8801	2,8172	2,7645	2,7196	2,6808	2,6469
20	2,9128	2,8365	2,7737	2,7209	2,6758	2,6369	2,6030
21	2,8740	2,7977	2,7348	2,6819	2,6368	2,5978	2,5638
22	2,8392	2,7628	2,6998	2,6469	2,6017	2,5626	2,5285
23	2,8077	2,7313	2,6682	2,6152	2,5699	2,5308	2,4966
24	2,7791	2,7027	2,6396	2,5865	2,5411	2,5019	2,4677
25	2,7531	2,6766	2,6135	2,5603	2,5149	2,4756	2,4413
26	2,7293	2,6528	2,5896	2,5363	2,4908	2,4515	2,4171
27	2,7074	2,6309	2,5676	2,5143	2,4688	2,4293	2,3949
28	2,6872	2,6106	2,5473	2,4940	2,4484	2,4089	2,3743
29	2,6686	2,5919	2,5286	2,4752	2,4295	2,3900	2,3554
30	2,6513	2,5746	2,5112	2,4577	2,4120	2,3724	2,3378
40	2,5289	2,4519	2,3882	2,3343	2,2882	2,2481	2,2130
60	2,4117	2,3344	2,2702	2,2159	2,1692	2,1286	2,0929
80	2,3549	2,2775	2,2130	2,1584	2,1115	2,0706	2,0346
120	2,2994	2,2217	2,1570	2,1021	2,0548	2,0136	1,9773
∞	2,1918	2,1136	2,0483	1,9927	1,9447	1,9027	1,8656

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}975$

n_2	n_1 15	16	17	18	19	20	25
1	984,8668	986,9187	988,7331	990,3490	991,7973	993,1028	998,0808
2	39,4313	39,4354	39,4391	39,4424	39,4453	39,4479	39,4579
3	14,2527	14,2315	14,2127	14,1960	14,1810	14,1674	14,1155
4	8,6565	8,6326	8,6113	8,5924	8,5753	8,5599	8,5010
5	6,4277	6,4032	6,3814	6,3619	6,3444	6,3286	6,2679
6	5,2687	5,2439	5,2218	5,2021	5,1844	5,1684	5,1069
7	4,5678	4,5428	4,5206	4,5008	4,4829	4,4667	4,4045
8	4,1012	4,0761	4,0538	4,0338	4,0158	3,9995	3,9367
9	3,7694	3,7441	3,7216	3,7015	3,6833	3,6669	3,6035
10	3,5217	3,4963	3,4737	3,4534	3,4351	3,4185	3,3546
11	3,3299	3,3044	3,2816	3,2612	3,2428	3,2261	3,1616
12	3,1772	3,1515	3,1286	3,1081	3,0896	3,0728	3,0077
13	3,0527	3,0269	3,0039	2,9832	2,9646	2,9477	2,8821
14	2,9493	2,9234	2,9003	2,8795	2,8607	2,8437	2,7777
15	2,8621	2,8360	2,8128	2,7919	2,7730	2,7559	2,6894
16	2,7875	2,7614	2,7380	2,7170	2,6980	2,6808	2,6138
17	2,7230	2,6968	2,6733	2,6522	2,6331	2,6158	2,5484
18	2,6667	2,6404	2,6168	2,5956	2,5764	2,5590	2,4912
19	2,6171	2,5907	2,5670	2,5457	2,5265	2,5089	2,4408
20	2,5731	2,5465	2,5228	2,5014	2,4821	2,4645	2,3959
21	2,5338	2,5071	2,4833	2,4618	2,4424	2,4247	2,3558
22	2,4984	2,4717	2,4478	2,4262	2,4067	2,3890	2,3198
23	2,4665	2,4396	2,4157	2,3940	2,3745	2,3567	2,2871
24	2,4374	2,4105	2,3865	2,3648	2,3452	2,3273	2,2574
25	2,4110	2,3840	2,3599	2,3381	2,3184	2,3005	2,2303
26	2,3867	2,3597	2,3355	2,3137	2,2939	2,2759	2,2054
27	2,3644	2,3373	2,3131	2,2912	2,2713	2,2533	2,1826
28	2,3438	2,3167	2,2924	2,2704	2,2505	2,2324	2,1615
29	2,3248	2,2976	2,2732	2,2512	2,2313	2,2131	2,1419
30	2,3072	2,2799	2,2554	2,2334	2,2134	2,1952	2,1237
40	2,1819	2,1542	2,1293	2,1068	2,0864	2,0677	1,9943
60	2,0613	2,0330	2,0076	1,9846	1,9636	1,9445	1,8687
80	2,0026	1,9741	1,9483	1,9250	1,9037	1,8843	1,8071
120	1,9450	1,9161	1,8900	1,8663	1,8447	1,8249	1,7462
∞	1,8326	1,8028	1,7759	1,7515	1,7291	1,7085	1,6259

Kvantily Fisherova-Snedecorova rozložení pro $\alpha = 0{,}975$

	n_1					
n_2	30	40	60	80	120	∞
1	1001,4140	1005,5980	1009,8000	1011,9080	1014,0200	1018,3000
2	39,4646	39,4729	39,4812	39,4854	39,4896	39,4980
3	14,0805	14,0365	13,9921	13,9697	13,9473	13,9020
4	8,4613	8,4111	8,3604	8,3349	8,3092	8,2573
5	6,2269	6,1750	6,1225	6,0960	6,0693	6,0153
6	5,0652	5,0125	4,9589	4,9318	4,9044	4,8491
7	4,3624	4,3089	4,2544	4,2268	4,1989	4,1423
8	3,8940	3,8398	3,7844	3,7563	3,7279	3,6702
9	3,5604	3,5055	3,4493	3,4207	3,3918	3,3329
10	3,3110	3,2554	3,1984	3,1694	3,1399	3,0798
11	3,1176	3,0613	3,0035	2,9740	2,9441	2,8828
12	2,9633	2,9063	2,8478	2,8178	2,7874	2,7249
13	2,8372	2,7797	2,7204	2,6900	2,6590	2,5955
14	2,7324	2,6742	2,6142	2,5833	2,5519	2,4872
15	2,6437	2,5850	2,5242	2,4930	2,4611	2,3953
16	2,5678	2,5085	2,4471	2,4154	2,3831	2,3163
17	2,5020	2,4422	2,3801	2,3481	2,3153	2,2474
18	2,4445	2,3842	2,3214	2,2890	2,2558	2,1869
19	2,3937	2,3329	2,2696	2,2368	2,2032	2,1333
20	2,3486	2,2873	2,2234	2,1902	2,1562	2,0853
21	2,3082	2,2465	2,1819	2,1485	2,1141	2,0422
22	2,2718	2,2097	2,1446	2,1108	2,0760	2,0032
23	2,2389	2,1763	2,1107	2,0766	2,0415	1,9677
24	2,2090	2,1460	2,0799	2,0454	2,0099	1,9353
25	2,1816	2,1183	2,0516	2,0169	1,9811	1,9055
26	2,1565	2,0928	2,0257	1,9907	1,9545	1,8781
27	2,1334	2,0693	2,0018	1,9665	1,9299	1,8527
28	2,1121	2,0477	1,9797	1,9441	1,9072	1,8291
29	2,0923	2,0276	1,9591	1,9232	1,8861	1,8072
30	2,0739	2,0089	1,9400	1,9039	1,8664	1,7867
40	1,9429	1,8752	1,8028	1,7644	1,7242	1,6371
60	1,8152	1,7440	1,6668	1,6252	1,5810	1,4821
80	1,7523	1,6790	1,5987	1,5549	1,5079	1,3997
120	1,6899	1,6141	1,5299	1,4834	1,4327	1,3104
∞	1,5660	1,4835	1,3883	1,3329	1,2684	1,0000

Kritické hodnoty jednovýběrového Wilcoxonova testu pro $n = 6, 7, \dots, 30$, $\alpha = 0{,}05$ a $\alpha = 0{,}01$

n	$\alpha = 0{,}05$ kritická hodnota	$\alpha = 0{,}01$ kritická hodnota
6	0	—
7	2	—
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68
26	98	75
27	107	83
28	116	91
29	126	100
30	137	109

Zdroj: ANDĚL, J.: *Matematická statistika*. (Tabulka XVIII.9).

Kritické hodnoty dvouvýběrového Wilcoxonova testu pro $m = 1, 2, \ldots, 30$, $n = 1, 2, \ldots, 20$, $\alpha = 0{,}05$

	n																			
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–																			
2	–	–																		
3	–	–	–																	
4	–	–	–	0																
5	–	–	0	1	2															
6	–	–	1	2	3	5														
7	–	–	1	3	5	6	8													
8	–	0	2	4	6	8	10	13												
9	–	0	2	4	7	10	12	15	17											
10	–	0	3	5	8	11	14	17	20	23										
11	–	0	3	6	9	13	16	19	23	26	30									
12	–	1	4	7	11	14	18	22	26	29	33	37								
13	–	1	4	8	12	16	20	24	28	33	37	41	45							
14	–	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	–	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	–	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	–	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	–	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	–	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	–	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	–	2	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	–	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	–	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	–	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	–	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	161
26	–	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	–	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	–	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	–	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	–	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200

Zdroj: ANDĚL, J.: *Matematická statistika*. (Tabulka XVIII.10a).

Kritické hodnoty a modifikované kritické hodnoty Kolmogorovova-Smirnovova testu pro $n = 5, \ldots, 30$, $\alpha = 0{,}05$

n	$D_n(\alpha)$	Modif. $D_n(\alpha)$
5	0,563	0,343
6	0,519	0,319
7	0,483	0,300
8	0,454	0,285
9	0,430	0,271
10	0,409	0,258
11	0,391	0,249
12	0,375	0,242
13	0,361	0,234
14	0,349	0,227
15	0,338	0,220
16	0,327	0,213
17	0,318	0,206
18	0,309	0,200
19	0,301	0,195
20	0,294	0,190
21	0,287	0,187
22	0,281	0,183
23	0,275	0,180
24	0,242	0,176
25	0,238	0,173
26	0,233	0,171
27	0,229	0,168
28	0,225	0,166
29	0,221	0,163
30	0,218	0,161

Zdroj: SPRENT, P.: *Nonparametric Statistical Method*. Second edition. (Table IV)

Kritické hodnoty pro Spearmanův koeficient pořadové korelace pro $n = 5, 6, \ldots, 30$, $\alpha = 0{,}05$

n	kritická hodnota
5	0,900
6	0,829
7	0,745
8	0,691
9	0,683
10	0,636
11	0,609
12	0,580
13	0,555
14	0,534
15	0,518
16	0,500
17	0,485
18	0,472
19	0,458
20	0,445
21	0,435
22	0,424
23	0,415
24	0,406
25	0,398
26	0,389
27	0,382
28	0,375
29	0,369
30	0,362

Zdroj: ANDĚL, J.: *Matematická statistika*, Tab. XVIII.6.

Kritické hodnoty Neményiho metody, $r = 3, 4, \ldots, 10$, $n = 1, 2, \ldots, 25$, $\alpha = 0{,}05$

	r							
n	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9,0	10,5	12,0	13,5
2	8,8	12,6	16,5	20,5	24,7	28,9	33,1	37,4
3	15,7	22,7	29,9	37,3	44,8	52,5	60,3	68,2
4	23,9	34,6	45,6	57,0	68,6	80,4	92,4	104,6
5	33,1	48,1	63,5	79,3	95,5	112,0	128,8	145,8
6	43,3	62,9	83,2	104,0	125,3	147,0	169,1	191,4
7	54,4	79,1	104,6	130,8	157,6	184,9	212,8	240,9
8	66,3	96,4	127,6	159,6	192,4	225,7	259,7	294,1
9	75,9	114,8	152,0	190,2	229,3	269,1	309,6	350,6
10	92,3	134,3	177,8	222,6	268,4	315,0	362,4	410,5
11	106,3	154,8	205,0	256,6	309,4	363,2	417,9	473,3
12	120,9	176,2	233,4	292,2	352,4	413,6	476,0	539,1
13	136,2	198,5	263,0	329,3	397,1	466,2	536,5	607,7
14	152,1	221,7	293,8	367,8	443,6	520,8	599,4	679,0
15	168,6	245,7	325,7	407,8	491,9	577,4	664,6	752,8
16	185,6	270,6	358,6	449,1	541,7	635,9	732,0	829,2
17	203,1	296,2	392,6	491,7	593,1	696,3	801,5	907,9
18	221,2	322,6	427,6	535,5	646,1	758,5	873,1	989,0
19	239,8	349,7	463,6	580,6	700,5	822,4	946,7	1072,4
20	258,8	377,6	500,5	626,9	756,4	888,1	1022,3	1158,1
21	278,4	406,1	538,4	674,4	813,7	955,4	1099,8	1245,9
22	298,4	435,3	577,2	723,0	872,3	1024,3	1179,1	1335,7
23	318,9	465,2	616,9	772,7	932,4	1094,8	1260,3	1427,7
24	339,8	495,8	657,4	823,5	993,7	1166,8	1343,2	1521,7
25	361,1	527,0	698,8	875,4	1056,3	1240,4	1427,9	1611,6

Zdroj: BLATNÁ, D.: *Neparametrické metody*. Tabulka T21/1.

Příloha B – Zadání POT

Zadání POT

Popis situace: Na fakultu speciálních studií ve městě N. se v minulém školním roce dostavilo 341 uchazečů k přijímacímu řízení. Podrobili se písemné přijímací zkoušce, z níž bylo možno získat maximálně 80 bodů. Jelikož fakulta nemá k dispozici takovou prostorovou kapacitu, aby všichni uchazeči mohli vykonat zkoušku naráz, byli rozděleni na tři skupiny, které skládaly zkoušku postupně v 9 h, 12 h a 15 h. O uchazečích jsou k dispozici následující údaje:

- Pohlaví (1 muž, 2 žena) . . . proměnná *SEX*
- Forma studia (1 denní studium, 2 kombinované studium, 3 celoživotní studium) . . . proměnná *FS*
- Doba konání zkoušky (9 h, 12 h, 15 h) . . . proměnná *CAS*
- Průměr známek ze střední školy . . . proměnná *SS_PRUMER*
- Počet bodů získaných z písemné přijímací zkoušky . . . proměnná *BODY*
- Informace o přijetí na fakultu (0 ne, 1 ano) . . . proměnná *PRIJETI*

Úkol 1. Sestrojte empirický 95% interval spolehlivosti pro střední hodnotu proměnné *BODY*, a to

a) pro všechny uchazeče
b) pro muže
c) pro ženy
d) pro uchazeče o denní studium
e) pro uchazeče o kombinované studium
f) pro uchazeče o celoživotní studium
g) pro uchazeče, kteří konali zkoušku v 9 h
h) pro uchazeče, kteří konali zkoušku ve 12 h
i) pro uchazeče, kteří konali zkoušku v 15 h.

Upozornění: ve všech případech ověřte pomocí K-S testu či S-W testu a pomocí N–P grafu normalitu proměnné *BODY*.

Úkol 2. Na hladině významnosti 0,05 testujte hypotézu, že se neliší střední hodnota proměnné *BODY* pro muže a ženy. Nakreslete krabicové diagramy.

Úkol 3. Na hladině významnosti 0,05 proveďte analýzu rozptylu proměnné *BODY* pro faktor *FS* (forma studia). V případě zamítnutí nulové hypotézy aplikujte Scheffého metodu mnohonásobného porovnávání. Pro všechny úrovně faktoru nakreslete krabicové diagramy.

Úkol 4. Na hladině významnosti 0,05 testujte hypotézu, že rozložení proměnné *BODY* je stejné ve skupinách uchazečů, kteří konali přijímací zkoušku v 9 h, 12 h, 15 h. V případě zamítnutí nulové hypotézy zjistěte, které dvojice skupin uchazečů se liší na hladině významnosti 0,05. Nakreslete krabicové diagramy.

Úkol 5. Sestavte kontingenční tabulku proměnných *SEX* a *FS* a simultánní četnosti znázorněte též graficky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že forma studia nezávisí na pohlaví uchazeče. Vypočtěte Cramérův koeficient.

Úkol 6.

a) Pomocí Fisherova přesného testu testujte na hladině významnosti 0,05 hypotézu, že přijetí na fakultu speciálních studií nezávisí na pohlaví uchazeče. Vypočtěte též podíl šancí na přijetí pro muže a pro ženy a sestrojte asymptotický 95% interval spolehlivosti pro podíl šancí.

b) Na asymptotické hladině významnosti 0,05 testujte hypotézu, že podíly uchazečů přijatých do denního studia, kombinovaného studia a celoživotního studia jsou stejné.

Úkol 7. Budeme se zabývat vztahem mezi průměrným prospěchem na střední škole (proměnná *SS_PRUMER*) a počtem bodů dosaženým u přijímací zkoušky (proměnná *BODY*).

a) Na hladině významnosti 0,05 ověřte pomocí K-S testu, S-W testu a N–P grafu, zda proměnná *SS_PRUMER* se řídí normálním rozložením, a to pro všechny uchazeče a pak zvlášť pro muže a pro ženy.

b) Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou orientačně ověřte dvourozměrnou normalitu proměnných *SS_PRUMER* a *BODY*, a to pro všechny uchazeče a pak zvlášť pro muže a pro ženy.

c) Vypočtěte koeficient korelace proměnných *SS_PRUMER* a *BODY*, a to pro všechny uchazeče a pak zvlášť pro muže a pro ženy. Na hladině významnosti 0,05 testujte v těchto třech případech hypotézu o nezávislosti proměnných *SS_PRUMER* a *BODY*.

Rejstřík