UČEBNÍ TEXTY OSTRAVSKÉ UNIVERZITY

Přírodovědecká fakulta

ANALÝZA DAT

2. upravené vydání

Josef Tvrdík

OSTRAVSKÁ UNIVERZITA 2008

OBSAH:

1 Úvod

Tento text slouží jako opora pro předmět Analýza dat. Navazuje na kurs Základy matematické statistiky. Cílem kursu je aplikovat základní statistické znalosti v relativně jednoduchých úlohách, s nimiž se velmi často setkáváme při analýze dat.

I když je text napsán s co největší snahou vysvětlovat nutné pojmy i jejich aplikaci jednoduše bez zbytečných a z pohledu využití statistických metod okrajových podrobností, počítejte s tím, že text nebude oddechová četba a že spoustu věcí bude potřeba důkladně promýšlet a opakovaně se k nim vracet, někdy i s opakováním pojmů z předmětu Základy matematické statistiky.

Časovou náročnost zvládnutí tohoto textu a vyřešení zadaných příkladů lze odhadnout na přibližně 80 až 100 hodin.

V některých příkladech, jejichž řešení je uvedeno v učebním textu, se užívají data ze souborů BI97.ASC. Pokud si chcete uvedená řešení sami ověřit a zopakovat, tato data si můžete stáhnout z webových stránek autora textu, http://albert.osu.cz/tvrdik/down/vyuka.html.

Hlavní úlohou, kterou byste měli osvědčit poznatky získané v tomto kursu, je analýza vámi vybraného souboru dat z vašeho okolí. Proto se poohlédněte po datech, které byste chtěli statisticky zpracovat, a kde jste zvědavi na výsledky této analýzy. Případné nejasnosti včas konzultujte s vyučujícím. Výsledky analýzy bude pak potřeba předložit formou vytištěné stručné a přehledné zprávy, pokud možno v rozsahu max. 3 strany. Před přípravou zprávy si prostudujte kap. 7 o prezentaci výsledků.

Ostatní korespondenční úlohy budou zadány na začátku semestru.

2 Parametrické testy o shodě středních hodnot

2.1 Jednovýběrový t-test

Jednovýběrový oboustranný *t*-test byl podrobně vysvětlen v učebním textu Základy matematické statistiky. Doporučujeme se k tomu vrátit a základy testování hypotéz si znovu připomenout.

Máme náhodný výběr $(X_1, X_2, \ldots, X_n)$ nezávislých náhodných veličin normálně rozdělených, tj. $X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n$. Testujeme hypotézu, že střední hodnota rozdělení populace, z níž máme výběr, tj. μ je rovna nějaké dané hodnotě μ_0. proti alternativě, že $\mu \neq \mu_0$ Za platnosti nulové hypotézy má statistika T rozdělení podle následujícího vztahu

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

a při oboustranné alternativě $\mu \neq \mu_0$ je kritický obor

$$W \equiv \left(-\infty, t_{n-1}(\alpha/2)\right] \cup \left[t_{n-1}(1-\alpha/2), +\infty\right)$$

Pokud hodnota T patří do kritického oboru, tak nulovou hypotézu $\mu = \mu_0$ pro dané α zamítáme.

Oboustranná alternativa $H_1 : \mu \neq \mu_0$ však není jediná možná formulace alternativní hypotézy. Máme-li k dispozici nějakou *apriorní informaci* o střední hodnotě populace, ze které je realizován výběr, můžeme zformulovat alternativu jednostranně:

$H_0 : \quad \mu = \mu_0$

$H_1 : \quad \mu > \mu_0$ (tzv. *pravostranná* alternativa)

Další postup testu bude zcela analogický jako u oboustranného testu, pouze kritický obor bude jiný, totiž $W \equiv \left[t_{n-1}(1-\alpha), +\infty\right)$. Nulovou hypotézu můžeme zamítnout ve prospěch této alternativy tehdy, když výběrový průměr $\bar{X}$ je o hodně větší než μ_0, přesněji vyjádřeno, když pro hodnotu testového kritéria platí

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \geq t_{n-1}(1-\alpha).$$

Vidíme, že pravděpodobnost neoprávněného zamítnutí nulové hypotézy je opět rovna hladině významnosti α. Tím, že jsme alternativu formulovali s využitím nějaké apriorní informace, stačí k zamítnutí nulové hypotézy, aby hodnota testového kriteria T byla alespoň $t_{n-1}(1-\alpha)$. U oboustranné alternativy by to bylo $t_{n-1}(1-\alpha/2)$.

Zcela analogicky, pokud bychom měli k tomu důvod, můžeme formulovat i *levostrannou* alternativu H_1: $\mu < \mu_0$. Pak kritický obor je $W \equiv \left(-\infty, t_{n-1}(\alpha)\right]$.

Obecně při užívání testů, zejména jednostranných, je vhodné nejdříve formulovat alternativu ve tvaru obsahujícím tvrzení, které bychom chtěli „prokázat“. Pak pokud nulovou hypotézu zamítneme, máme téměř jistotu (s rizikem rovným α), že tvrzení vyjádřené alternativní hypotézou je pravdivé.

2.2 Dvouvýběrový t-test

Předpokládáme, že máme dva nezávislé výběry o rozsahu n_1, resp. n_2, ze dvou normálně rozdělených populací, první populace má rozdělení $N(\mu_1, \sigma_1^2)$, druhá $N(\mu_2, \sigma_2^2)$.

Z kapitoly 4.1 v textu pro Základy matematické statistiky víme (viz rov. 4.1-10), že když neznámé parametry σ_1^2, σ_2^2 můžeme považovat za shodné, tedy $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (rozptyl v obou populacích je shodný), pak pro náhodnou veličinu T platí

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \ .$$

Chceme-li testovat hypotézu, že střední hodnoty v obou populacích jsou shodné, tj.

$H_0: \quad \mu_1 = \mu_2$

proti některé z alternativ

$H_1: \quad \mu_1 \neq \mu_2$ (oboustranná alternativa)

$H_1: \quad \mu_1 < \mu_2$ (levostranná alternativa)

$H_1: \quad \mu_1 > \mu_2$ (pravostranná alternativa)

užijeme testovou statistiku

$$T_{eq} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (1)$$

která má za platnosti nulové hypotézy Studentovo *t*-rozdělení s $n_1 + n_2 - 2$ stupni volnosti.

Pokud rozptyly v obou populacích shodné nejsou, tj. $\sigma_1^2 \neq \sigma_2^2$, užívá se pro test hypotézy o shodě středních hodnot statistika

$$T_{noneq} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \qquad (2)$$

která má přibližně *t*-rozdělení s ν stupni volnosti, kde počet stupňů volnosti ν se určí podle vztahu

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Znamená to tedy, že při testování nulové hypotézy o shodě středních hodnot se musíme rozhodnout, zda je nebo není splněn předpoklad o shodě rozptylů, tj. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ a podle toho volit testové kriterium dané výrazem (1) nebo (2). Toto rozhodnutí provedeme testem hypotézy H_0: $\sigma_1^2 = \sigma_2^2$ proti alternativě H_1: $\sigma_1^2 \neq \sigma_2^2$.

Pokud naše výběry o rozsazích n_1 , n_2 jsou z normálně rozdělených populací, $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, platí (viz vztah 4.1-5, Základy matematické statistiky)

$$\frac{(n_1 - 1)\, s_1^{\,2}}{\sigma_1^{\,2}} \sim \chi^2_{n_1 - 1} \quad \text{a} \quad \frac{(n_2 - 1)\, s_2^{\,2}}{\sigma_2^{\,2}} \sim \chi^2_{n_2 - 1}$$

a tedy také platí

$$\frac{s_1^{\,2} / \sigma_1^{\,2}}{s_2^{\,2} / \sigma_2^{\,2}} \sim F_{n_1 - 1, n_2 - 1}$$

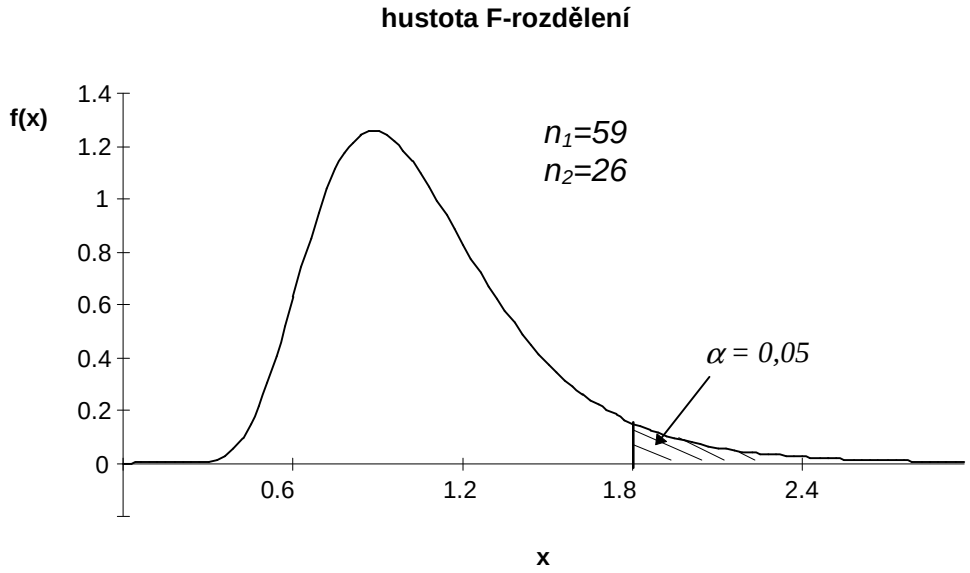
Za platnosti nulové hypotézy $\sigma_1^2 = \sigma_2^2$ má testová statistika $F = s_1^{\,2} / s_2^{\,2}$ Fisher-Snedecorovo rozdělení s parametry $n_1 - 1, n_2 - 1$,

$$F = \frac{s_1^{\,2}}{s_2^{\,2}} \sim F_{n_1 - 1,\, n_2 - 1} \qquad (3)$$

Lze se dohodnout, že indexování výběrů zvolíme tak, aby platilo $s_1^{\,2} \geq s_2^{\,2}$. Prakticky to znamená. ve jmenovateli bude menší z obou výběrových rozptylů. Pak kritickým oborem bude

$$W = \left[F_{n_1 - 1, n_2 - 1}(1 - \alpha), +\infty \right), \qquad (4)$$

jinými slovy, hypotézu o shodě rozptylů $\sigma_1^2 = \sigma_2^2$ zamítneme, když poměr výběrových rozptylů $s_1^{\,2} / s_2^{\,2}$ bude podstatně větší než jedna. Situaci ilustruje následující obrázek, $F_{59,26}(0{,}95) = 1{,}804$.



Při testování hypotéz obvykle používáme statistický software. Při dvouvýběrovém *t*-testu prováděném v Excelu nejdříve otestujeme hypotézu o shodě rozptylů (v doplňku Analýza dat funkce s názvem Dvouvýběrový *F*-test pro rozptyl) a podle jeho výsledku se rozhodneme, zda máme užít funkci Dvouvýběrový *t*-test s rovností rozptylů nebo Dvouvýběrový *t*-test s nerovností rozptylů.

V NCSS je ve výsledcích vyhodnocena jak testová statistika (1) pro rovnost rozptylů, tak kritérium (2) pro neshodu rozptylů. Je na nás, abychom si vybrali správnou část výsledku pro interpretaci. Postup si ukážeme na příkladu.

Příklad 1: Máme posoudit, zda střední hodnota veličiny K1 (data BI97) jsou stejné v populaci odrůdy 1 i odrůdy 2.

Použijeme program NCSS, z menu *Analysis* vybereme *T-Tests*, z nich *Two-sample*. Zadáme *k1* jako *Response variable* a veličinu *Odruda* jako *Group variable* (tato veličina rozděluje pozorování do dvou skupin) a dostaneme výstup, který zde uvedeme ve zkrácené podobě.

Variable k1

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation	95% LCL of Mean	95% UCL of Mean
odruda=1	60	13.84833	3.45197	12.95659	14.74007
odruda=2	27	12.17778	2.767717	11.08291	13.27265

Equal-Variance T-Test Section

Alternative Hypothesis	T-Value	Prob Level	Decision (5%)	Power (Alpha=.05)	Power (Alpha=.01)
Difference <> 0	2.2127	0.029602	Reject Ho	0.590054	0.342280
Difference < 0	2.2127	0.985199	Accept Ho	0.000061	0.000003
Difference > 0	2.2127	0.014801	Reject Ho	0.708885	0.440816

Difference: (odruda=1)-(odruda=2)

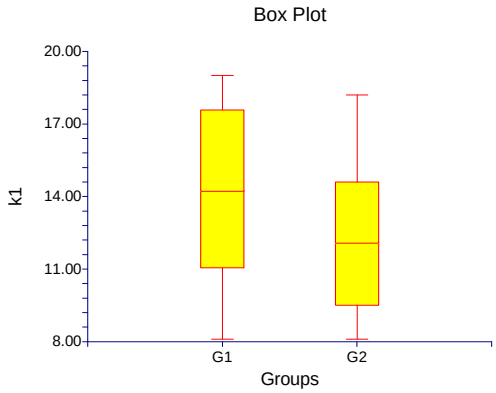
Aspin-Welch Unequal-Variance Test Section

Alternative Hypothesis	T-Value	Prob Level	Decision (5%)	Power (Alpha=.05)	Power (Alpha=.01)
Difference <> 0	2.4054	0.019160	Reject Ho	0.658359	0.407180
Difference < 0	2.4054	0.990420	Accept Ho	0.000029	0.000001
Difference > 0	2.4054	0.009580	Reject Ho	0.768562	0.510535

Difference: (odruda=1)-(odruda=2)

Tests of Assumptions Section

Assumption	Value	Probability	Decision(5%)
Skewness Normality (odruda=1)	-0.2373	0.812435	Cannot reject normality
Skewness Normality (odruda=2)	0.7455	0.455956	Cannot reject normality
Variance-Ratio Equal-Variance Test	1.5556	0.189787	Cannot reject equal variances



I zkrácený výstup je dosti obsažný a napoprvé nám dá trochu práce se v něm orientovat a správně interpretovat výsledky. Naším úkolem je testovat nulovou hypotézu o shodě středních hodnot proti oboustranné alternativě, tj.

$H_0: \quad \mu_1 = \mu_2$

$H_1: \quad \mu_1 \neq \mu_2$

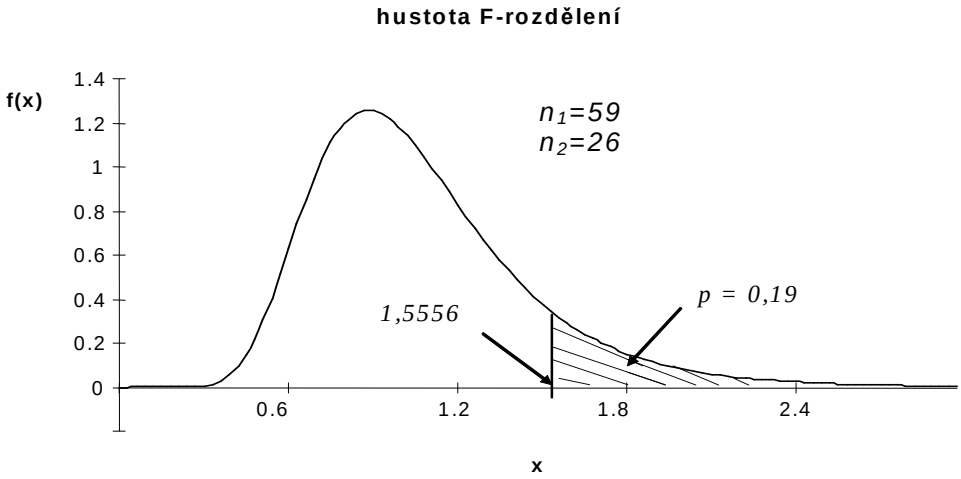
Stejnou nulovou i alternativní hypotézu můžeme formulovat i takto:

$H_0: \quad \mu_1 - \mu_2 = 0$

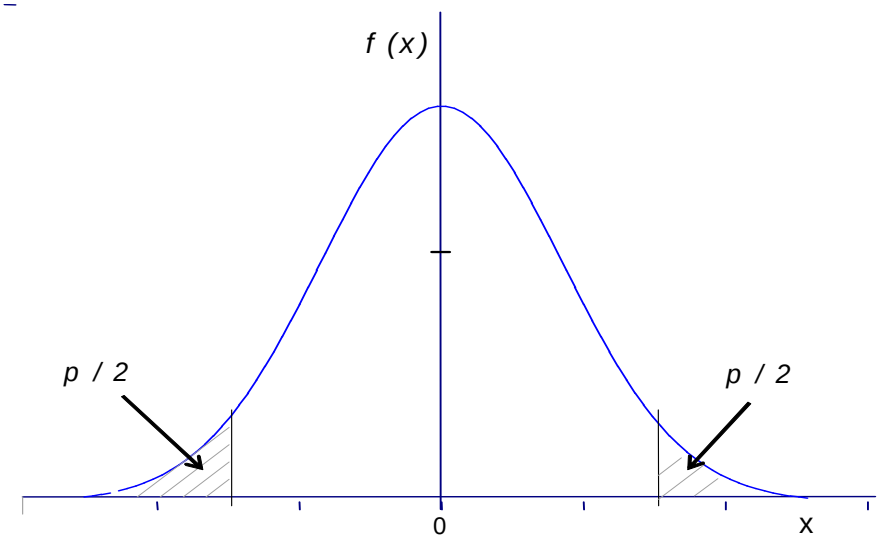
$H_1: \quad \mu_1 - \mu_2 \neq 0$

Této formulaci odpovídá forma výsledků, kde se objevuje rozdíl středních hodnot (*difference*). Ještě se musíme rozhodnout, zda máme pro naše rozhodování užít statistiku T_{eq} definovanou rov. (1) nebo statistiku T_{noneq} definovanou rov. (2), čili který odstavec z výsledků se nás týká, zda *Equal variances section* nebo *Unequal variances section*. Musíme rozhodnout, zda můžeme považovat za splněný předpoklad o shodě rozptylů v obou populacích či nikoliv. K tomuto rozhodnutí nám poslouží test hypotézy $H_0: \sigma_1^2 = \sigma_2^2$ proti alternativě $H_1: \sigma_1^2 \neq \sigma_2^2$. Jeho výsledky nalezneme v odstavci testů předpokladů (*Tests of Assumptions*) na řádku *Variance-Ratio Equal-Variance Test*. Tam nalezneme hodnotu testové statistiky spočtené podle vztahu (3) a kromě toho také tzv. dosaženou úroveň významnosti této hodnoty, která je uvedena ve sloupci *Probability*. Tato významnost (*probability*, někdy označovaná také *p-value*, *prob-level* nebo krátce *p*) je často užívanou charakteristikou, která usnadňuje interpretaci výsledků. V případě *jednostranného* testu, a to tento test je, viz kritický obor daný vztahem (4), *p* udává pravděpodobnost, že za platnosti nulové hypotézy bude mít testová

statistika hodnotu větší než hodnotu spočítanou z výběru, tedy v našem příkladu $p = P(X \geq 1{,}5556) \cong 0{,}19$. Smysl p v tomto příkladu i v jiných jednostranných testech vysvětluje následující obrázek.



Je zřejmé, že pokud platí $p \leq \alpha$, nulovou hypotézu zamítáme, jinak nezamítáme. Jelikož v našem příkladu vyšlo $p \cong 0{,}19$, tedy větší než obvykle volená hladina významnosti $\alpha = 0{,}05$, přijímáme představu o shodě rozptylů v obou populacích, $\sigma_1^2 = \sigma_2^2$. Proto statistika pro test hypotézy o rovnosti středních hodnot obou populací je statistika T_{eq} definovaná rovnicí (1). Její hodnotu nalezneme ve výsledcích v odstavci *Equal-Variance T-Test*. Její hodnota je 2,2127 a u ní je uvedena i odpovídající hodnota *p*. Jelikož ale v tomto případě se jedná o oboustranný test, p udává pravděpodobnost, že za platnosti nulové hypotézy bude absolutní hodnota testové statistiky větší nebo rovna absolutní hodnotě statistiky spočítané z výběru, tedy v našem příkladu $p = P(|X| \geq 2{,}2127) \cong 0{,}03$. Jednoduše řečeno, u oboustranných testů zamítáme nulovou hypotézu, je-li hodnota testové statistiky buď velmi velká nebo velmi malá. Opět pokud platí, že $p \leq \alpha$, nulovou hypotézu zamítáme. Názorně situaci vidíme na následujícím obrázku.



Jelikož v uvedeném příkladu je $p \cong 0{,}03$, hypotézu o shodě středních hodnot, tedy $\mu_1 - \mu_2 = 0$, na hladině významnosti $\alpha = 0{,}05$ zamítáme. Pokud bychom předem z nějakých důvodů zvolili hladinu významnosti $\alpha = 0{,}01$, naše výběrová data by nám neposkytovala důvod nulovou hypotézu zamítnout.

Obecně můžeme říci, že počítačové výstupy výsledků statistických testů s uvedenými hodnotami *p* usnadňují interpretaci v tom, že nepotřebujeme pro určování kritického oboru statistické tabulky. To, zda vypočtená statistika je či není v kritickém oboru, poznáme bezprostředně z hodnoty *p*: Je-li $p \le \alpha$, víme, že hodnota testového kriteria je v kritickém oboru, pokud $p > \alpha$, hodnota testového kriteria v kritickém oboru není.

V uvedeném dvouvýběrovém *t*-testu se vychází z předpokladu, že oba výběry jsou z normálně rozdělených populací. Splnění tohoto předpokladu není tak důležité, pokud rozsahy obou výběrů jsou dost velké. Jak víme z odstavce o centrální limitní větě, při dostatečně velkém počtu pozorování má testové kriterium

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (5)$$

normované normální rozdělení *N(0,1)* a při velkém počtu stupňů volnosti se tvar *t* - rozdělení přibližuje rozdělení *N(0,1).* Pro velké rozsahy výběrů hodnoty testových statistik (1) a (2) se přibližují hodnotě dané rov. (5) a statistiku *U* můžeme pak použít i pro test hypotézy o shodě středních hodnot dvou populací libovolného rozdělení.

2.3 Párový t-test

Dalším často užívaným *t*-testem je tzv. *párový t-test*. Obecně o párových testech hovoříme tehdy, když máme pro vybrané objekty změřeny dvojice hodnot, např. délka levé a pravé končetiny, krevní tlak před a po podání léku, stupeň opotřebení pravé a levé pneumatiky atd. Ve statistice je tato situace označována jako dva závislé výběry stejného rozsahu *n*.

Máme-li tedy dva závislé náhodné výběry $(X_1, X_2, \ldots, X_n)$, $(Y_1, Y_2, \ldots, Y_n)$, můžeme zjistit rozdíly těchto hodnot: $D_i = X_i - Y_i$ .a spočítat výběrové statistiky, průměr $\bar{D}$ a rozptyl $s_D{}^2$.

Při testu hypotézy o shodě středních hodnot veličin *X* a *Y*, tedy $H_0: \mu_1 - \mu_2 = 0$ vlastně testujeme, zda střední hodnota veličiny D je nulová. To je situace, kterou už známe z jednovýběrového *t*-testu. Testovým kriteriem pro test této hypotézy je

$$T_p = \frac{\bar{D}}{s_D / \sqrt{n}}, \qquad (6)$$

která má rozdělení t_{n-1}. Podobně jako u jednovýběrového testu může být alternativní hypotéza formulována jako oboustranná nebo jednostranná.

Při párovém testu můžeme nulovou hypotézu formulovat nejen tak, že střední hodnoty obou veličin jsou shodné, ale i tak, že jejich rozdíl je roven hodnotě a, $H_0 : \mu_1 - \mu_2 = a$. Pak testovou statistikou je

$$T_p = \frac{\bar{D} - a}{s_D / \sqrt{n}}, \qquad (7)$$

která opět za platnosti nulové hypotézy má rozdělení t_{n-1}.

Souhrn:

- Statistický test hypotézy se užívá k rozhodování za nejistoty.
- Rozhodujeme mezi nulovou hypotézou a alternativou.
- Jsou dva druhy chybného rozhodnutí.
- Pravděpodobnost chyby I. druhu při testu volíme předem (hladina významnosti).
- Test hypotézy je analogický rozhodování soudu, ale rozdíl je v tom, že pravděpodobnost chyby prvního druhu je u statistických testů známa, dokonce ji zvolíme.
- Kritický obor test závisí na tom, jak je zformulována alternativa.

Kontrolní otázky:

1. *Proč testy o parametrech jsou rozhodování v nejistotě?*
2. *Vysvětlete rozdíl mezi chybou prvního a druhého druhu.*
3. *Proč je zamítnutí nulové hypotézy pro praktické rozhodování užitečnější výsledek než nezamítnutí nulové hypotézy?*
4. *Kdy můžeme formulovat jednostrannou alternativu? Jakou nám to pak přináší výhodu?*
5. *Čím se liší párový t-test od jednovýběrového t-testu?*

Pojmy k zapamatování:

- *statistické testování hypotéz*
- *nulová hypotéza, alternativa*
- *chyby prvního a druhého druhu*
- *hladina významnosti*
- *síla testu*
- *testová statistika (kriterium)*
- *kritický obor*
- *jednovýběrový t-test*
- *dvouvýběrový t-test*
- *párové testy, párový t-test*
- *hodnota testové statistiky a odpovídající p-value*

Korespondenční úlohy č. 1 a 2

Budou zadány na začátku semestru.

3 Analýza rozptylu - jednoduché třídění

Jako analýza rozptylu (ANOVA) je označován soubor postupů induktivní statistiky užívaných při testování hypotéz o středních hodnotách při různém, často i velmi komplikovaném uspořádání experimentu. Analýzou rozptylu se podrobně zabývají specializované statistické monografie. Zde si ukážeme jen základní myšlenky analýzy rozptylu na úloze, která se nazývá analýza rozptylu s jednoduchým tříděním (one-way ANOVA). K prostudování této kapitoly by mělo stačit asi 2 až 3 hodiny.

Na analýzu rozptylu s jednoduchým tříděním můžeme pohlížet jako na zobecnění dvouvýběrového *t*-testu pro situaci, kdy máme testovat shodu středních hodnot ve více než dvou populacích. V takových úlohách nemůžeme použít opakovaně dvouvýběrový *t*-test pro všechny dvojice výběru, pokud chceme, aby pravděpodobnost chyby prvního druhu byla rovna zvolené hladině významnosti.

Předpokládejme, že máme I $(I \geq 2)$ nezávislých výběrů (tj. pozorovaná data jsou z I různých skupin). Náhodné veličiny (i jejich pozorované hodnoty) v i-tém výběru označíme $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, $n_i > 1$, $i = 1, 2, \ldots, I$. Výběry jsou z populací, které mají rozdělení $N(\mu_i, \sigma^2)$, tedy rozptyly ve všech populacích jsou shodné.

Celkem tedy máme k dispozici $n = \sum_{i=1}^{I} n_i$ nezávislých náhodných veličin.

Nulovou hypotézu, kterou chceme testovat, můžeme zapsat jako

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_I \tag{1}$$

Každou tuto náhodnou veličinu můžeme tedy vyjádřit jako součet

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, 2, \ldots, n_i; \qquad i = 1, 2, \ldots, I, \tag{2}$$

kde náhodné veličiny e_{ij} jsou *nezávislé* a mají stejné rozdělení $N(0, \sigma^2)$, $\sigma^2 > 0$. Tím jsme formulovali statistický model: Každou pozorovanou hodnotu Y_{ij} považujeme za součet hodnoty μ společné pro všechny skupiny, hodnoty α_i vyjadřující vliv i-té skupiny a normálně rozdělené náhodné složky ε_{ij} s nulovou střední hodnotou. Hodnoty $\mu, \sigma^2, \alpha_1, \alpha_2, \ldots, \alpha_I$ jsou neznámé parametry modelu. Pokud přidáme tzv. reparametrizační podmínku

$$\sum_{i=1}^{I} n_i \alpha_i = 0, \tag{3}$$

jsou hodnoty parametrů $\mu, \alpha_1, \alpha_2, \ldots, \alpha_I$ určeny jednoznačně a nulovou hypotézu (1) můžeme zapsat jako

$$H_0: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0 \tag{4}$$

Tato formulace je ekvivalentní formulaci (1). Parametr α_i pak můžeme chápat jako výsledek (efekt) charakterizující i-tou skupinu, v analýze rozptylu se někdy říká efekt i-tého ošetření (treatment). Testovaná hypotéza vyjadřuje, že skupiny se neliší, vliv ošetření je nulový.

Úkolem analýzy rozptylu je vlastně vysvětlit variabilitu všech vyšetřovaných náhodných veličin, čili vysvětlit variabilitu jejich pozorovaných hodnot.

Pro zkrácení dalšího zápisu zavedeme označení

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij} \quad \text{(skupinové součty)},$$

$$\overline{Y}_{i\bullet} = \frac{Y_{i\bullet}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad \text{(skupinové průměry)}$$

$$Y_{\bullet\bullet} = \sum_{i=1}^{I} Y_{i\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{(celkový součet)},$$

$$\overline{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{n} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{(celkový průměr)} \qquad (5)$$

V těchto zkratkách je vždy index, přes který se sčítá, vyznačen tečkou. Vidíme, že $\overline{Y}_{i\bullet}$ je výběrový průměr *i*-tého výběru (skupinový průměr), $\overline{Y}_{\bullet\bullet}$ je výběrový průměr ze všech pozorování (celkový průměr, grand mean).

Celkovou variabilitu pozorovaných hodnot charakterizuje součet čtverců odchylek od celkového průměru

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \qquad (6)$$

Tento tzv. *celkový* součet čtverců můžeme rozložit

$$\begin{aligned} S_T &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(Y_{ij} - \overline{Y}_{i\bullet}) + (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})\right]^2 = \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})\right] + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 = \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 + 2\sum_{i=1}^{I}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet}) + \sum_{i=1}^{I} n_i\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 = \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 + \sum_{i=1}^{I} n_i\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \end{aligned} \qquad (7)$$

Prostřední člen v součtu, $2\sum_{i=1}^{I}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet}) = 0$,

neboť $\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet}) = 0, \quad i = 1, 2, \ldots, I$ (součet odchylek od průměru je vždy roven nule).

Dva členy v posledním řádku (7) jsou charakteristikami variability

- *uvnitř* skupin $$S_e = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\bullet} \right)^2 \qquad (8)$$

  (součet čtverců odchylek pozorovaných hodnot od skupinových průměrů),

- *mezi* skupinami $$S_A = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} \right)^2 \qquad (9)$$

  (vážený součet čtverců odchylek skupinových průměrů od celkového průměru).

Vztah (7) tedy můžeme přepsat jako

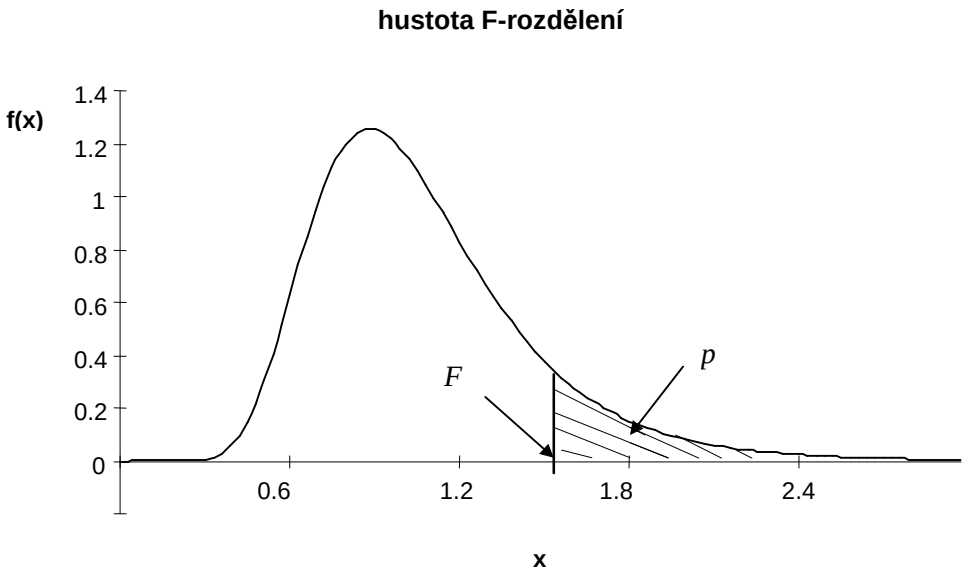
$$S_T = S_e + S_A \qquad (10)$$

Jak víme, celkový součet čtverců S_T má *(n - 1)* stupňů volnosti. Meziskupinový součet čtverců S_A má $(I-1)$ stupňů volnosti a součet čtverců uvnitř skupin (také se říká *residuální* nebo *chybový*, Error Sum of Squares) S_e má zbylé stupně volnosti, tj. *(n - I).* Pokud platí nulová hypotéza (4), je jak statistika $S_A / (I-1)$, tak statistika $S_e / (n-I)$ nestranným odhadem téhož rozptylu σ^2 a jejich podíl má tedy za platnosti nulové hypotézy *F*-rozdělení

$$F = \frac{S_A / (I-1)}{S_e / (n-I)} \sim F_{I-1,n-I} \qquad (11)$$

Pokud nulová hypotéza neplatí, je statistika $S_A / (I-1)$ výrazně větší. Kritickým oborem pro zamítnutí nulové hypotézy (4) je $W = \left[F_{I-1,n-I}(1-\alpha), +\infty \right)$.

Výsledky analýzy rozptylu jsou obvykle prezentovány v tabulkové formě, v počítačových výstupech i se sloupcem s hodnotou dosažené úrovně významnosti *p*, což je pravděpodobnost, že náhodná veličina mající rozdělení $F_{I-1,n-I}$ je větší nebo rovna hodnotě statistiky *F*. Význam hodnoty p vysvětluje následující obrázek. Je zřejmé, že pokud platí, $p \le \alpha$, nulovou hypotézu zamítáme, jinak nezamítáme.



Tabulka výsledků analýzy rozptylu s jednoduchým tříděním má následující tvar:

zdroj variability	suma čtverců	stupně volnosti	střední čtverec (mean square)	F	p
mezi skupinami	S_A	$I - 1$	$S_A / (I - 1)$	$\frac{S_A/(I-1)}{S_e/(n-I)}$	hodnota p
uvnitř skupin	S_e	$n - I$	$S_e / (n - I)$		
celkový	S_T	n – 1	$S_T / (n - 1)$		

U složitějších návrhů experimentu má tabulka výsledků analýzy rozptylu více řádků.

Zamítneme-li nulovou hypotézu o shodě všech středních hodnot

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$,

obvykle nás zajímá, která dvojice středních hodnot se liší. K tomu slouží testy nazývané *mnohonásobné porovnání* (*multiple comparison*). Těch je několik druhů, popis a základní informace k jejich užití nalezeneme v online manuálu NCSS, zájemce o podrobnější informace odkazujeme na literaturu, např. Anděl 1978, 1993, Havránek 1993 atd., podobně jako zájemce o složitější modely analýzy rozptylu.

Poznámka: Pokud bychom užili analýzu rozptylu s jednoduchým tříděním na data pocházející jen ze dvou výběrů, bude mít statistika F z rov. (11) tvar

$$F = \frac{S_A / 2}{S_e /(n-2)} \sim F_{1,n-2}$$

a hodnota statistiky *F* bude rovna druhé mocnině statistiky *t* ze dvouvýběrového oboustranného *t*-testu pro shodné rozptyly. Tyto dva testy jsou tedy ekvivalentní.

Rozkladu celkového rozptylu (10) můžeme užít pro výpočet směrodatné odchylky, máme-li k dispozici pouze skupinové charakteristiky - průměry $\overline{x}_i$, počty pozorování n_i a směrodatné odchylky s_i, $i = 1,2,\ldots,I$.

Směrodatná odchylka je odmocnina z celkového rozptylu, tj.

$$s = \sqrt{\frac{S_T}{n-1}} = \sqrt{\frac{S_e + S_A}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{I} s_i^2 \left(n_i - 1\right) + \sum_{i=1}^{I} n_i (\overline{x}_i - \overline{\overline{x}})^2\right]}, \qquad (12)$$

kde celkový průměr spočítáme jako vážený průměr skupinových průměrů,

$\overline{\overline{x}} = \frac{1}{n}\sum_{i=1}^{I} n_i \overline{x}_i$.

Aplikaci analýzy rozptylu s jednoduchým tříděním ukážeme na následujícím příkladu.

Příklad: Máme posoudit, zda střední hodnota veličiny *Delka* (data BI97) jsou stejné ve všech čtyřech lokalitách.

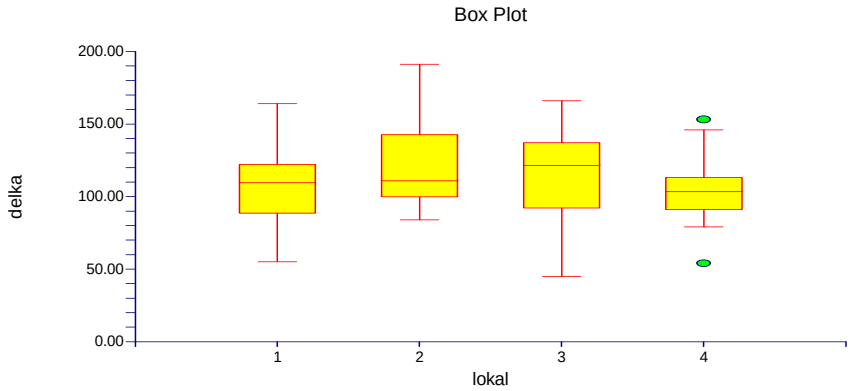
Pro test hypotézy o shodě středních hodnot

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

užijeme analýzu rozptylu s jednoduchým tříděním. Výpočet provedeme s pomocí programu NCSS. V něm z menu *Analysis* vybereme *ANOVA*, dále *One-way ANOVA*. Zadáme veličinu *Delka* jako *Dependent variable* a veličinu *Lokatita* jako *Factor variable* (tato veličina rozděluje pozorování do čtyřech skupin) a dostaneme výstup, který zda uvedeme ve zkrácené podobě:

Analysis of Variance Report

Response delka



Analysis of Variance Table

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
A (lokal)	3	3737.32	1245.773	1.68	0.176777
S(A)	87	64438.07	740.6674		
Total (Adjusted)	90	68175.38			

Z tabulky analýzy rozptylu vidíme, že $p = 0{,}177$. Tedy nulovou hypotézu nemůžeme zamítnout na žádné rozumně zvolené hladině významnosti. Rozdíly v poloze pozorovaných hodnot veličiny *Delka* v jednotlivých skupinách (viz krabicové diagramy na obrázku) nemůžeme přičítat nějakým systematickým rozdílům mezi skupinami, ale pouze důsledku nahodilého kolísání.

Kontrolní otázky:

1. *Jaká hypotéza se testuje v analýze rozptylu s jednoduchým tříděním?*
2. *Jaké jsou předpoklady pro užití analýzy rozptylu s jednoduchým tříděním?*
3. *Co je celkový průměr a skupinové průměry?*
4. *Čemu se říká celkový součet čtverců a jak jej lze rozložit?*
5. *Co je v analýze rozptylu s jednoduchým tříděním testovou statistikou, jaké má rozdělení za platnosti nulové hypotézy?*
6. *Kdy zamítáme nulovou hypotézu?*

Pojmy k zapamatování:

- *skupinové průměry a celkový průměr*
- *celkový součet čtverců a jeho rozklad*
- *import a export dat*
- *variabilita uvnitř skupin a mezi skupinami*
- *tabulka výsledků analýzy rozptylu*

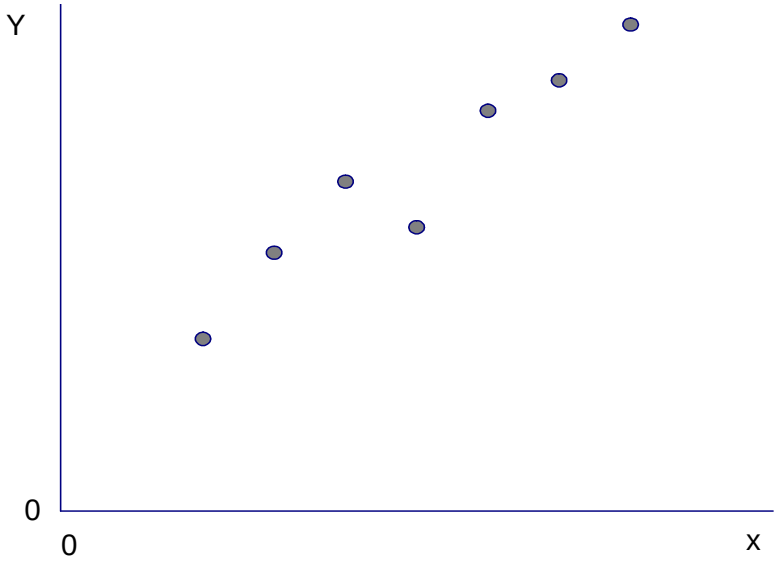
4 Základy lineární regrese

Regrese je snad nejčastěji užívaná statistická metoda. Odhaduje se, že 80 až 90 % aplikací statistiky je nějakou z variant regresní analýzy. Principy regresní analýzy se pokusíme vysvětlit na nejjednodušším tzv. klasickém lineárním regresním modelu. K prostudování této kapitoly si vyhraďte asi 4 hodiny.

Lineární regrese se zabývá problémem vysvětlení změn hodnot jedné veličiny lineární závislostí na jedné nebo více jiných veličinách. Uvažujme nejjednodušší případ, kdy vysvětlujeme veličinu Y linární závislostí na jedné vysvětlující veličině x. Data mají tvar, který je uveden v následující tabulce:

i	x_i	Y_i
1	x_1	Y_1
2	x_2	Y_2
⋮		
n	x_n	Y_n

Předpokládáme, že hodnoty veličiny x umíme nastavit přesně (např. teplotu v termostatu), hodnoty Y_i jsou zatíženy náhodným kolísáním, způsobeným třeba nepřesnostmi měřící metody (např. objem plynu). K dispozici tedy máme n dvojic pozorovaných hodnot. Grafické znázornění takových dat ukazuje následující obrázek.



Na obrázku vidíme, že s rostoucí hodnotou veličiny x se zhruba lineárně mění i hodnota Y, body na obrázku kolísají kolem myšlené přímky, kterou bychom mohli naměřenými body proložit.

Hodnoty veličiny Y_i můžeme vyjádřit jako součet dvou složek:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\ldots,n \tag{1}$$

kde β_0, β_1 jsou neznámé koeficienty určující lineární závislost a ε_i náhodná kolísání.

Pokud střední hodnoty náhodného kolísání jsou nulové, $E(\varepsilon_i) = 0,\ i = 1,2,\ldots,n$, rov. (1) můžeme přepsat

$$E(Y \mid x = x_i) = E(Y_i) = \beta_0 + \beta_1 x_i \tag{2}$$

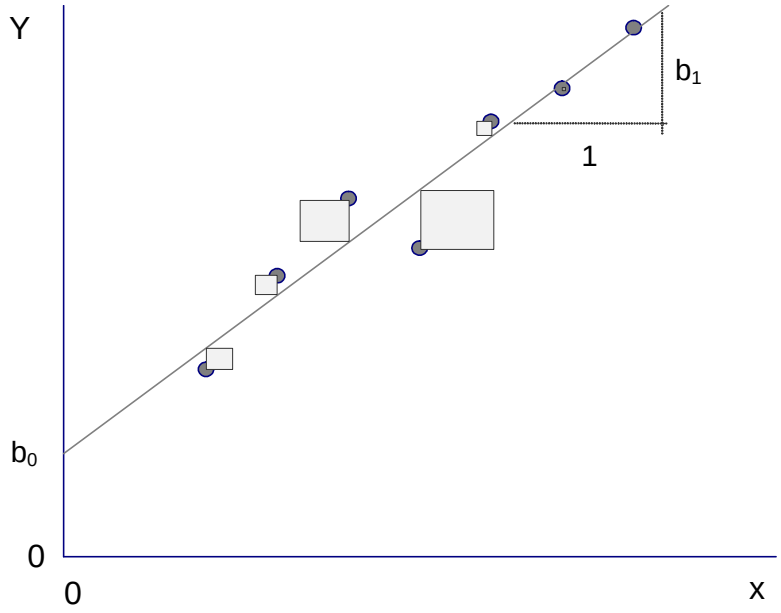
čili střední hodnoty náhodných veličin Y_i za podmínky, že veličina *x* má hodnotu x_i, leží na přímce dané rov. (2).

Rovnice (1) a (2) formulují regresní model, v tomto případě *lineární regresní model* s jednou vysvětlující proměnnou (regresorem) *x* a vysvětlovanou proměnnou *Y*. Neznámé koeficienty β_0, β_1 jsou *parametry regresního modelu*, také se jim říká regresní koeficienty. Regresní model je vlastně vyjádřením naší představy o závislosti veličiny *Y* na veličině *x*.

Jednou ze základních úloh regresní analýzy je odhad parametrů regresního modelu z pozorovaných dat. V případě našeho lineárního modelu je potřeba odhadnout regresní koeficienty β_0, β_1 z dat, tzn. nalézt takové hodnoty b_0, b_1, které by určovaly přímku $\hat{Y}_i = b_0 + b_1 x_i$ co nejlépe prokládající naměřená data. Hodnoty b_0, b_1, jsou pak odhady regresních koeficientů β_0, β_1 , $\hat{Y}_i$ je odhadem $E(Y|x = x_i)$. Co nejlepší proložení může být formulováno různými způsoby, nejčastěji se užívá *metoda nejmenších čtverců* (MNČ), tj. hledáme takové hodnoty b_0 (úsek, který vytíná přímka na ose *Y*), b_1 (směrnice přímky), aby součet čtverců odchylek pozorovaných hodnot Y_i od hodnot $\hat{Y}_i$ byl co nejmenší:

$$S_e = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 x_i\right)^2 \to \min \tag{3}$$

Metodu nejmenších čtverců vysvětluje následující obrázek. Řešíme úlohu, jak volit hodnoty b_0 a b_1, aby součet ploch vyznačených čtverců byl co nejmenší.



Hodnoty b_0, b_1 minimalizující S_e nalezneme tak, že parciální derivace S_e podle b_0, b_1 položíme rovny nule:

$$\frac{\partial S_e}{\partial b_0} = 0\,, \qquad \frac{\partial S_e}{\partial b_1} = 0\,.$$

Tím dostaneme soustavu tzv. *normálních rovnic* (v tomto případě dvou rovnic), v obecném případě, kdy regresní model má více parametrů než model s jedním regresorem, je počet normálních rovnic roven počtu parametrů. Jsou-li normální rovnice lineární jako v tomto regresním modelu, říkáme, že regresní model je *lineární v parametrech*.

Snadno nalezneme, že parciální derivace jsou rovny následujícím výrazům

$$\frac{\partial S_e}{\partial b_0} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 x_i\right) = -2\left(\sum_{i=1}^{n} Y_i - nb_0 - b_1\sum_{i=1}^{n} x_i\right),$$

$$\frac{\partial S_e}{\partial b_1} = -2\sum_{i=1}^{n}\left[\left(Y_i - b_0 - b_1 x_i\right)x_i\right] = -2\left(\sum_{i=1}^{n} x_i Y_i - b_0\sum_{i=1}^{n} x_i - b_1\sum_{i=1}^{n} x_i^2\right). \qquad (4)$$

V minimu jsou parciální derivace rovny nule, takže po jednoduchých úpravách dostaneme soustavu dvou normálních rovnic

$$nb_0 + b_1\sum x_i = \sum Y_i$$

$$b_0\sum x_i + b_1\sum x_i^2 = \sum x_i Y_i \qquad (5)$$

Řešení této soustavy rovnic můžeme vyjádřit explicitně takto:

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum x_i\right) = \bar{Y} - b_1 \bar{x} \tag{6}$$

$$b_1 = \frac{\sum x_i Y_i - \frac{\left(\sum x_i\right)\left(\sum Y_i\right)}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{n\sum x_i Y_i - \left(\sum x_i\right)\left(\sum Y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} \ . \tag{7}$$

Z rov. (6) vidíme, že přímka proložená metodou nejmenších čtverců, tj. splňující podmínku (3), prochází bodem $\left[\bar{x}, \bar{Y}\right]$.

Dosadíme-li z rov. (7) do (6), dostaneme

$$\begin{aligned} b_0 &= \frac{1}{n}\left[\sum Y_i - \frac{n\left(\sum x_i Y_i\right) - \left(\sum x_i\right)\left(\sum Y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}\sum x_i\right] = \\ &= \frac{\left(\sum Y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i Y_i\right)\left(\sum x_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} \end{aligned} \tag{8}$$

Nyní připomeneme některé rovnosti, které využijeme při dalším výkladu o statistických vlastnostech odhadů b_0 , b_1.

$$\begin{aligned} \sum\left(x_i - \bar{x}\right)^2 &= \sum\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \\ &= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \end{aligned} \tag{9}$$

$$\sum\left(x_i - \bar{x}\right)x_i = \sum\left(x_i^2 - \bar{x}x_i\right) = \sum x_i^2 - \bar{x}\sum x_i = \sum\left(x_i - \bar{x}\right)^2 \tag{10}$$

$$\begin{aligned} \sum\left(x_i - \bar{x}\right)\left(Y_i - \bar{Y}\right) &= \sum\left(x_i Y_i - \bar{Y}x_i - \bar{x}Y_i + \bar{x}\bar{Y}\right) = \\ &= \sum x_i Y_i - \bar{x}\sum Y_i - \bar{Y}\sum x_i + n\bar{x}\bar{Y} = \\ &= \sum x_i Y_i - n\bar{x}\bar{Y} - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} = \\ &= \sum x_i Y_i - n\bar{x}\bar{Y} = \sum x_i Y_i - \frac{\left(\sum x_i\right)\left(\sum Y_i\right)}{n} \end{aligned} \tag{11}$$

$$\begin{aligned} \sum\left(x_i - \bar{x}\right)Y_i &= \sum x_i Y_i - \bar{x}\sum Y_i = \\ &= \sum x_i Y_i - \frac{\sum x_i \sum Y_i}{n} = \sum\left(x_i - \bar{x}\right)\left(Y_i - \bar{Y}\right) \end{aligned} \tag{12}$$

Z rov. (7), (9) a (12) pak dostaneme

$$b_1 = \frac{\sum x_i Y_i - \frac{\left(\sum x_i\right)\left(\sum Y_i\right)}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{(n-1)\left[\sum (x_i - \bar{x})(Y_i - \bar{Y})\right]}{(n-1)\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2},$$

kde s_x^2 je výběrový rozptyl veličiny x a s_{xy} je výběrová kovariance.

Jelikož $r_{xy} = \frac{s_{xy}}{s_x s_y}$, vidíme, že $b_1 = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$.

Tzn., že směrnici regresní přímky můžeme vypočítat z hodnoty korelačního koeficientu. Jak vidíme, směrnice i korelační koeficient musí mít stejné znaménko.

S využitím (11) a (12) můžeme rov. (7) přepsat

$$b_1 = \frac{\sum \left(x_i - \bar{x}\right)\ Y_i}{\sum \left(x_i - \bar{x}\right)^2} \qquad (13)$$

Odtud

$$b_1 \sum \left(x_i - \bar{x}\right)^2 = \sum \left(x_i - \bar{x}\right) Y_i$$

Pak pro střední hodnoty náhodných veličin v předchozí rovnici platí

$$E(b_1)\sum \left(x_i - \bar{x}\right)^2 = \sum \left(x_i - \bar{x}\right) E(Y_i) = \sum \left(x_i - \bar{x}\right)(\beta_0 + \beta_1 x_i) =$$

$$= \beta_1 \sum \left(x_i - \bar{x}\right) x_i = \beta_1 \sum \left(x_i - \bar{x}\right)^2$$

Když tuto rovnost dělíme výrazem $\sum \left(x_i - \bar{x}\right)^2$, dostaneme $E(b_1) = \beta_1$, takže b_1 je *nestranným* odhadem parametru β_1.

Podobně pro b_0 můžeme dosadit do (6)

$$b_0 = \bar{Y} - b_1 \bar{x} = \sum \frac{Y_i}{n} - \frac{\sum \left(x_i - \bar{x}\right) Y_i}{\sum \left(x_i - \bar{x}\right)^2} \bar{x} = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum \left(x_i - \bar{x}\right)^2}\right] Y_i = \sum c_i Y_i .$$

Můžeme ukázat, že

$$\sum c_i = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum \left(x_i - \bar{x}\right)^2}\right] = \frac{n}{n} - \frac{\bar{x}\sum (x_i - \bar{x})}{\sum \left(x_i - \bar{x}\right)^2} = \frac{n}{n} - 0 = 1$$

a také, že

$$\sum c_i x_i = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum \left(x_i - \bar{x}\right)^2}\right] x_i = \frac{1}{n}\sum x_i - \frac{\bar{x}\sum\ (x_i - \bar{x}) x_i}{\sum \left(x_i - \bar{x}\right)^2} = \bar{x} - \bar{x} = 0$$

Pak pro střední hodnotu b_0 platí

$$E(b_0) = \sum c_i E(Y_i) = \sum c_i (\beta_0 + \beta_1) x_i = \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_0 .$$

Tedy i b_0 je *nestranným* odhadem parametru β_0.

Chceme-li určit rozptyly odhadů b_0 , b_1, potřebujeme ještě další předpoklady o náhodné složce e_i v rov. (1):

a) $E(\varepsilon_i) = 0, \qquad i = 1, 2, \ldots, n$

(tento předpoklad už byl vysloven dříve);

b) $\mathrm{var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2, \qquad i = 1, 2, \ldots, n$

(rozptyl e_i je konstantní, tzv. homoskedascita);

c) $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \ldots, n$

($\varepsilon_i, \varepsilon_j$ jsou nekorelované).

Z rov. (1) vidíme, že $\mathrm{var}(Y_i) = \mathrm{var}(e_i) = \sigma^2$. Pak z rov. (13) dostaneme

$$\mathrm{var}(b_1) = \frac{1}{\left[\sum \left(x_i - \bar{x}\right)^2\right]^2} \sum \left(x_i - \bar{x}\right)^2 \mathrm{var}(Y_i) = \frac{\sigma^2}{\sum \left(x_i - \bar{x}\right)^2} \ . \qquad (14)$$

Z rov. (14) vidíme, že rozptyl odhadu směrnice regresní přímky můžeme snížit vhodnou volbou hodnot regresoru tak, aby $\sum \left(x_i - \bar{x}\right)^2$ byla co největší.

Z rov. (6) dostaneme

$$\mathrm{var}(b_0) = \mathrm{var}(\bar{Y}) + \bar{x}^2 \ \mathrm{var}(b_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum \left(x_i - \bar{x}\right)^2} \right) \qquad (15)$$

Podobně tedy i rozptyl odhadu úseku regresní přímky můžeme snížit zvětšením rozsahu výběru a volbou hodnot regresoru tak, aby $\sum \left(x_i - \bar{x}\right)^2$ byla co největší.

Přidáme-li k předpokladům (a), (b), (c) ještě předpoklad (d)

d) $\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, 2, \ldots, n$

(odchylky hodnot Y_i od lineární závislosti mají normální rozdělení),

$$\text{pak} \quad \frac{b_j - \beta_j}{\sqrt{\mathrm{var}(b_j)}} \sim N(0,1), \quad j = 0, 1 \qquad (16)$$

Pokud bychom znali $\mathrm{var}(b_j)$, mohla by statistika definovaná rov. (16) sloužit jako testové kritérium pro testy hypotéz o parametrech regresního modelu.

Obyčejně však $\mathrm{var}(b_j)$ neznáme, neboť neznáme σ^2 - viz rov. (14) a (15). Hodnotu σ^2 (tzv. reziduální rozptyl) však můžeme odhadnout:

$$\hat{\sigma}^2 = s^2 = \frac{S_e^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 x_i\right)^2}{n-2} \ . \qquad (17)$$

Charakteristika s^2 definovaná rov. (17) - *výběrový residuální rozptyl* - je nestranným odhadem hodnoty σ^2. Dosadíme-li tento odhad do rov. (14) a (15) místo σ^2, získáme odhady rozptylů regresních parametrů. Označme odmocniny z těchto odhadů rozptylů $s(b_j),\ j = 0, 1$ (směrodatná odchylka nebo také standardní chyba odhadu regresního parametru). Pak náhodná veličina

$$\frac{b_j - \beta_j}{s(b_j)} \sim t_{n-2}, \quad j = 0,1, \tag{18}$$

a pro testování hypotéz $\beta_j = 0$ můžeme užít statistiku $\frac{b_j}{s(b_j)} \sim t_{n-2}$.

Poznámka:
Lineární regresní model (1) můžeme celkem snadno zobecnit, může obsahovat více než jeden regresor. Máme-li k regresorů, $k > 1$, lineární regresní model má tvar:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i, \quad i = 1, 2, \ldots, n$$

Pak residuální rozptyl se odhaduje jako

$$\hat{\sigma}^2 = s^2 = \frac{S_e}{n-k-1} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-k-1}$$

tj. součet residuálních čtverců se dělí rozsahem výběru zmenšeným o počet parametrů regresního modelu, což je *k+1* .

Pak platí $\frac{b_j - \beta_j}{s(b_j)} \sim t_{n-k-1}, \quad j = 0,1,\ldots,k$,

tedy tyto náhodné veličiny mají Studentovo *t*-rozdělení s *n-k-1* stupni volnosti.

Příklad:
Uvažujme data ze souboru BI97. Naším úkolem je odhad regresních parametrů lineárního modelu závislosti veličiny *VAHA* na veličině *DELKA*.

V řešení využijeme statistický program NCSS. Volbou *File/Open* otevřeme soubor BI97.S0 (tzv. s*avefile* vytvořený dříve programem NCSS) a v menu *Analysis* vybereme *Multiple Regression.*. V šabloně regrese zvolíme jako vysvětlovanou veličinu (*Dependent variable*) VAHA, jako regresory (*Independent variables*) zvolíme jedinou veličinu, a to DELKA. Po spuštění výpočtu dostaneme následující výstup (zde je uveden v trochu zkrácené podobě):

Multiple Regression Report

Dependent vaha

Regression Equation Section

Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level	Decision (5%)
Intercept	1.272396	4.163085	0.3056	0.760594	Accept Ho
delka	0.8864501	3.650991E-02	24.2797	0.000000	Reject Ho
R-Squared	0.868829				

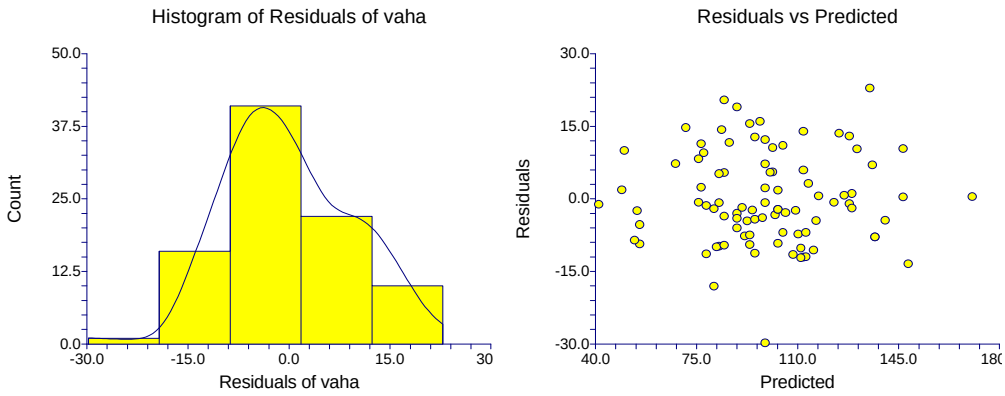
Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Stand. Coeff.
Intercept	1.272396	4.163085	-6.9995	9.5443	0.0000
delka	0.8864501	3.650991E-02	0.8139	0.9589	0.9321
T-Critical	1.986979				

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	899033.3	899033.3		
Model	1	53571.79	53571.79	589.5043	0.000000
Error	89	8087.964	90.87601		
Total(Adjusted)	90	61659.76	685.1084		

Root Mean Square Error	9.53289	R-Squared	0.8688
Mean of Dependent	99.39561	Adj R-Squared	0.8674
Coefficient of Variation	9.590857E-02	Press Value	8416.884
Sum |Press Residuals|	703.5859	Press R-Squared	0.8635



Možná je délka výstupu této naší jednoduché úlohy poněkud překvapivá, ale naučíme se v tomto výstupu číst.

Odhady parametrů lineárního regresního modelu jsou v části *Regression Equation Section*. Na řádku *Intercept* je odhad úseku regresní přímky - viz rov. (8) - a další charakteristiky týkající se tohoto parametru, na řádku *delka* pak je odhad směrnice - viz rov. (7) - a další charakteristiky týkající se tohoto parametru. Odhady parametrů b_0 , b_1, jsou tedy ve sloupci *Regression Coefficient***.** Ve sloupci *Standard Error* jsou pak $s(b_j), \quad j = 0,1$ - viz rov (14), (15) a následující text.

Ve sloupci *T-Value* jsou hodnoty testového kritéria $\frac{b_j}{s(b_j)}$ pro test hypotézy $\beta_j = 0$ - viz rov. (18) - a ve sloupci *Prob Level* jsou významnosti p pro oboustranný test.

Výsledkem naší úlohy jsou odhady b_0 (úsek) = 1,27 a b_1 (směrnice) = 0,886. Kromě toho vidíme, naše data nás opravňují zamítnout hypotézu $\beta_1 = 0$, (v tabulce výsledků má hodnota *p*-value 6 nul, tzn. $p < 0{,}0000005$), takže nulovou hypotézu můžeme zamítnout na jakékoli rozumně zvolené hladině významnosti. Zřejmě váha se s rostoucí délkou významně mění. Naproti tomu hypotézu $\beta_0 = 0$ zamítnout nemůžeme (p = 0,76) a tudíž je oprávněné předpokládat, že regresní přímka prochází počátkem. Takový regresní model jen s jedním parametrem, a to směrnicí, bychom měli prozkoumat v dalším kroku. Význam důležité charakteristiky *R-Squared* vysvětlíme později.

V části *Regression Coefficient Section* se opakují odhady regresních koeficientů a jejich směrodatných odchylek a dále jsou zde uvedeny 100(1-α)-procentní intervalové odhady regresních parametrů (ve sloupcích *Lower 95% C.L* a *Upper 95% C.L*.), hodnota α může být zvolena při zadání výpočtu.

Část Analysis of Variance Section vysvětlíme později. Z dalších charakteristik je užitečná Root Mean Square Error, což je odmocnina z Error Mean Square a je to směrodatná odchylka odhadu, odmocnina z výrazu daného rov. (17), tedy výběrová residuální směrodatná odchylka s.

Grafy ve výstupu - histogram residuí $Y_i - \hat{Y}_i$ a závislost residuí $Y_i - \hat{Y}_i$ na hodnotách $\hat{Y}_i$ predikovaných regresním modelem jsou užitečným nástrojem pro vizuální přibližné ověření předpokladů (a), (b), (c) a (d) užitých při odvozování vztahů pro odhad regresních parametrů a rozdělení statistik, zejména pro ověření konstantního rozptylu, nekorelovanosti residuí a jejich normálního rozdělení.
Konec příkladu.

Nyní se vrátíme k vysvětlení charakteristik, které jsme v předchozím příkladu přeskočili. Z odstavce o analýze rozptylu víme, že celkový součet čtverců odchylek naměřených hodnot veličiny Y od jejich průměru můžeme rozložit na dva sčítance:

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2 \qquad (19)$$

Označme jednotlivé sumy čtverců podle jejich významu

- celková suma čtverců (total sum of squares):

  $$TSS=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

- residuální suma čtverců (residual sum of squares):

  $$RSS=S_e=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

- modelová suma čtverců (model sum of squares):

  $$MSS=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$$

Rov. (19) tedy můžeme číst takto: Celkovou variabilitu vysvětlované veličiny rozložíme na část, která odpovídá variabilitě vysvětlené regresním modelem a na část, kterou model nevysvětluje, která zbývá, tedy je residuální. To můžeme zapsat:

$$TSS = MSS + RSS\,. \qquad (20)$$

Pak můžeme zavést *index determinace* R^2 (*R-squared*).

$$R^2=\frac{MSS}{TSS}=\frac{TSS-RSS}{TSS}=1-\frac{RSS}{TSS} \qquad (21)$$

Vidíme, že index (koeficient) determinace je vlastně podíl variability vysvětlený regresním modelem k celkové variabilitě závislé veličiny. Je zřejmé, že

$$0\le R^2\le 1 \qquad (22)$$

Hodnotu 1 dosahuje R^2 tehdy, když $RSS = 0$ (viz rov. 21), tj, že závislost Y na x je přesně lineární (model vysvětluje vše). Hodnotu 0 dosahuje index determinace tehdy, když model nevysvětluje z variability Y nic, tzn. $RSS = TSS$,
tedy regresní přímka je rovnoběžná s osou x v úrovni $b_0 = \bar{Y}$.

Lze také ukázat, že pro lineární regresní model s jedním regresorem - rov. (1) nebo (2) - je koeficient determinace roven druhé mocnině výběrového korelačního koeficientu, tedy

$$R^2 = r_{xy}{}^2 . \qquad (23)$$

Při používání tohoto vztahu nezapomeňte, že $-1 \le r_{xy} \le 1$ a znaménko korelačního koeficientu je shodné se znaménkem směrnice přímky.

Tabulka analýzy rozptylu je obvyklou součástí počítačových výstupů regresních programů. Její strukturu pro výběr o rozsahu n a regresní model s k parametry (počet regresorů je k -1) můžeme vyjádřit

zdroj variability	suma čtverců	stupně volnosti	střední čtverec (mean square)	F
model	*MSS*	*k*-1	*MSS / (k*-1)	$\frac{MSS/(k-1)}{RSS/(n-k)}$
error	*RSS*	*n-k*	*RSS / (n-k)*	
total	*TSS*	*n*-1		

Jsou-li splněny předpoklady (a) až (d), statistika F v předposledním sloupci tabulky má F rozdělení s $(k-1)$ a $(n-k)$ stupni volnosti. V případě modelu jen s jedním regresorem je tento test ekvivalentní s *t*-testem hypotézy, že $\beta_1 = 0$ (směrnice je nulová, tedy Y není na x lineárně závislé), dosažená úroveň významnosti p je u obou testů shodná, viz poznámka v závěru kapitoly o analýze rozptylu s jednoduchým tříděním. Statistiku F využijeme jen v úlohách s více než jedním regresorem. Je-li hodnota statistiky F v kritickém oboru, znamená to, že významná část variability veličiny Y je vysvětlena lineární závislostí na jednom nebo více regresorech.

Kontrolní otázky:

1. *Co vyjadřuje lineární regresní model, jaký má tvar?*
2. *Co jsou parametry lineárního modelu? Jak se odhadují z dat?*
3. *Co se minimalizuje v metodě nejmenších čtverců?*
4. *Jaké jsou předpoklady v klasickém lineárním modelu? Jak jejich platnost lze ověřit?*
5. *Jaké hypotézy o parametrech lze testovat? Co je testovou statistikou?*
6. *Jakých hodnot může nabývat koeficient determinace? Jak lze jeho hodnotu interpretovat?*

7. *Spočítejte úlohu řešenou v příkladu v této kapitole pomocí Excelu, zorientujte se ve výstupech a porovnejte výsledky.*

Pojmy k zapamatování:

- *lineární regresní model*
- *odhad parametrů regresního modelu, metoda nejmenších čtverců*
- *residuální rozptyl, rozptyly odhadů parametrů*
- *celkový a residuální součet čtverců, koeficient determinace*

Korespondenční úloha č. 3

Bude zadána na začátku semestru.

5 Neparametrické metody

V této rozsáhlé kapitole se seznámíme se základy tzv. neparametrických metod. Jsou to metody, kdy předmětem testu hypotézy není tvrzení o hodnotě parametru nějakého konkrétního rozdělení, ale nulová hypotéza je formulována obecněji, např. jako shoda rozdělení nebo nezávislost veličin. Tuto kapitolu doporučujeme studovat po jednotlivých podkapitolách a podle potřeby se v textu vracet a vzájemně porovnávat výhody a nevýhody jednotlivých testů. Postupy a algoritmy užívané v neparametrických metodách, zejména operace s pořadím hodnot, mohou být i inspirativní pro aplikaci v mnoha oborech informatiky.

Dosud jsme se setkávali jen s testy hypotéz o parametrech normálního rozdělení (*t*-testy, ANOVA, testy o parametrech lineárního regresního modelu). Všechny tyto testy vycházejí z předpokladu, že máme jeden nebo více výběrů z normálního rozdělení. Tak silný předpoklad při praktických aplikacích nebývá často splněn. Pak je na místě otázka, jakou statistickou metodu volit, abychom dostali spolehlivé výsledky a aby naše rozhodnutí při testu hypotézy nebylo ovlivněno právě jen nesplněním předpokladů pro použití těchto tzv. parametrických metod. Jedním z dlouhá léta osvědčených alternativních postupů je použití tzv. neparametrických metod. Nebudeme se podrobněji zabývat společnými vlastnostmi neparametrických metod, jen se spokojíme s tím, že neparametrické metody nevyžadují, aby výběry byly z normálního rozdělení. Většinou stačí, když jde o výběry ze spojitých rozdělení, u neparametrických metod se nulová hypotéza často týká mediánu rozdělení. Neparametrické metody často vycházejí z pořadí pozorovaných hodnot v jejich vzestupném uspořádání. Předpoklady pro aplikaci neparametrických metod jsou oproti parametrickým metodám daleko slabší, tzn. že při aplikacích jsou splněny častěji. Obecně však platí, že tato výhoda neparametrických testů je vyvážena nevýhodou – ve srovnání s testy parametrickými jsou neparametrické testy slabší, tzn. že pravděpodobnost zamítnutí nulové hypotézy v situaci, kdy zamítnuta být má, je menší. Proto by neparametrické testy měly být užívány jen tehdy, kdy předpoklady pro parametrické testy splněny nejsou.

5.1 Testy dobré shody

Testy dobré shody (angl. goodness-of-fit tests) se užívají k ověřování shody empirického rozdělení s nějakým teoretickým rozdělením. Ilustruje to následující příklad.

Příklad: Chceme ověřit, zda hrací kostka je „fair“, tzn. že všech 6 možných výsledků má stejnou pravděpodobnost. Uděláme tedy experiment, kdy kostkou hodíme opakovaně a zaznamenáme četnosti dosažených výsledků:

výsledek	1	2	3	4	5	6	*n*
četnost n_i	14	24	15	25	26	16	120

Testujeme nulovou hypotézu, že pravděpodobnosti p_i = 1/6. Můžeme tedy spočítat četnosti e_i, které bychom očekávali za platnosti nulové hypotézy ze 120 hodů za platnosti nulové hypotézy (n = 120), $e_i = n \cdot p_i = 120 \cdot (1/6) = 20$.

Nulovou hypotézu zamítneme, když se pozorované četnosti n_i budou hodně lišit od očekávaných četností e_i. Testovým kritériem je statistika

$$X = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}, \qquad (1)$$

kde k je počet možných výsledků, v našem příkladu k = 6. Tato statistika má při dostatečně velkém n (takovém, aby všechny $e_i \geq 5$) rozdělení chí-kvadrát s k-1 stupni volnosti,

$$X = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{k-1}. \qquad (2)$$

Nulovou hypotézu zamítneme, pokud odchylky od očekávaných četností jsou velké, tj. když hodnota testového kritéria X je v kritickém oboru W,
$W \equiv \left[\chi^2_{k-1}(1-\alpha), +\infty\right)$.

Pro náš příklad je výpočet ukázán v následující tabulce.

i	n_i	p_i	e_i	chi_kv
1	14	1/6	20	1.80
2	24	1/6	20	0.80
3	15	1/6	20	1.25
4	25	1/6	20	1.25
5	26	1/6	20	1.80
6	16	1/6	20	0.80
	120	1	120	7.70

Zvolíme-li $\alpha = 0,05$, je kritický obor $W \equiv [11.07, +\infty)$. Hodnota testové statistiky je 7,70, tedy neleží v kritickém oboru a nulovou hypotézu nemůžeme zamítnout. Na základě našeho experimentu jsme neprokázali, že kostka není „fair“.

Pro spojité veličiny a spojitá rozdělení je test dobré shody podobný, jen postup o trochu pracnější. Testujeme shodu rozdělení našich pozorovaných hodnot s nějakým spojitým teoretickým rozdělením, známe tedy distribuční funkci *F(x)* tohoto rozdělení. Potřebujeme tedy zjistit empirické četnosti n_i a očekávané četnosti e_i, tzn. předtím musíme obor hodnot empirických dat rozdělit na intervaly, v nich zjistit četnosti, spočítat očekávané četnosti a vyhodnotit testové kriterium (1). Současně potřebujeme, aby všechny očekávané četnosti byly $e_i \geq 5$. Je výhodné zvolit takové dělení na takových k intervalů, aby očekávané četnosti byly konstantní,

$$e_i = n \cdot p_i = \frac{n}{k} \geq 5\,, \tag{3}$$

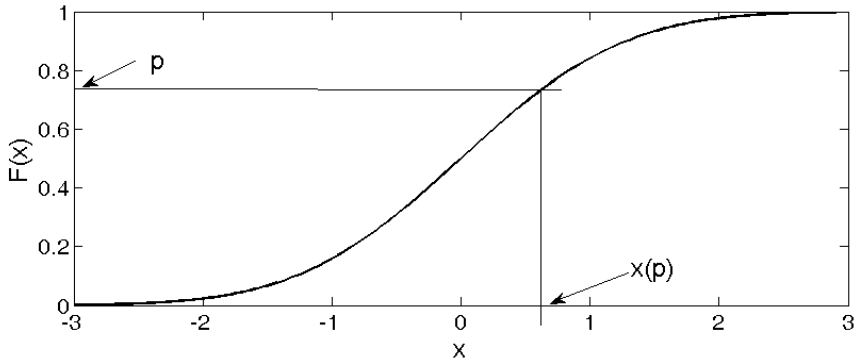
tedy *k* volíme tak, aby $k \leq n/5$.

Hranice intervalů jsou pak následující kvantily teoretického rozdělení,

$$x(i \cdot p_i) = x(i/k)\,, \qquad i = 0,1,\ldots,k\,. \tag{4}$$

Pak už se jen spočítají četnosti n_i, $i = 0,1,\ldots,k$, tj. počty hodnoty v jednotlivých intervalech a vyhodnotí testové kriterium (1).

Význam pojmu *p*-kvantil, tj hodnoty *x(p)* ilustruje obrázek.



Uvědomme si, že podmínka (3), znamená, že dělení na svislé ose hodnot *F(x)* je ekvidistantní, zatímco intervaly (jejich hranice dané vztahem (4) odečítáme na vodorovné ose) stejně široké většinou nejsou, záleží na tvaru distribuční funkce, čili na teoretickém rozdělení, s nímž testujeme shodu. Nejčastěji se testuje shoda s normálním rozdělením.

5.2 Kontingenční tabulky - test nezávislosti

Máme-li dvě nominální veličiny X, Y, kde X může nabývat hodnot $x_1, x_2, ..., x_C$ a veličina Y může nabývat hodnot $y_1, y_2, ..., y_R$, pak rozdělení četností pozorovaných hodnot můžeme vyjádřit kontingenční tabulkou, jak už známe z popisné statistiky.

		X						
Y		x_1	x_2	...	x_j	...	x_C	$n_{i.}$
	y_1	n_{11}	n_{12}		n_{1j}		n_{1C}	$n_{1.}$
	y_2	n_{21}	n_{22}				n_{2C}	$n_{2.}$
	:	:	:		:			:
	y_i	n_{i1}			n_{ij}		n_{iC}	$n_{i.}$
	:	:	:		:			:
	y_R	n_{R1}	n_{R2}		n_{Rj}		n_{RC}	$n_{R.}$
		$n.1$	$n.2$		$n.j$		$n.C$	$n.. = n$

Hodnoty n_{ij} jsou absolutní četnosti, tzn. počty sledovaných objektů, kdy veličina Y má hodnotu y_i a současně veličina X má hodnotu x_j. *Marginální četnosti* $n_{i\bullet}$ a $n_{\bullet j}$ jsou definovány jako řádkové, resp. sloupcové součty.

$$n_{i\bullet} = \sum_{j=1}^{C} n_{ij} \qquad n_{\bullet j} = \sum_{i=1}^{R} n_{ij} \tag{1}$$

Celkový počet objektů n je samozřejmě součet přes všechna políčka tabulky:

$$n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij} = \sum_{i=1}^{R} n_{i\bullet} = \sum_{j=1}^{C} n_{\bullet j} \tag{2}$$

Obvyklou úlohou statistické analýzy je rozhodnout, zda náhodné veličiny jsou nezávislé či mezi nimi existuje nějaký vtah a také nějakou vhodnou charakteristikou případnou závislost kvantifikovat.

Test nezávislosti dvou nominálních náhodných veličin X, Y je založen na tom, že můžeme odhadnout četnosti, které bychom pozorovali, kdyby opravdu veličiny X, Y nezávislé byly. Jsou-li X, Y nezávislé, pak pravděpodobnost jevu, že současně nastane jev $Y = y_i$ a jev $X = x_j$ vyjádřit jako součin pravděpodobností

$$P\left[(Y = y_i) \cap (X = x_j)\right] = P(Y = y_i).P(X = x_j)$$
$$i = 1, 2, \ldots, R, \quad j = 1, 2, \ldots, C \tag{3}$$

Pro zkrácení zápisu zavedeme označení

$$p_{ij} = P\left[(Y = y_i) \cap (X = x_j)\right], \quad p_{i\bullet} = P(Y = y_i), \quad p_{\bullet j} = P(X = x_j).$$

Pak rov.(3) můžeme přepsat

$$p_{ij} = p_{i\bullet}.p_{\bullet j} \quad i = 1,2,\ldots,R \quad j = 1,2,\ldots,C \tag{4}$$

Marginální pravděpodobnosti $p_{i\bullet}, p_{\bullet j}$ můžeme odhadnout jako relativní marginální četnosti (odhady jsou vyznačeny stříškou nad symbolem):

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}, \qquad \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}, \tag{5}$$

a četnost, kterou bychom očekávali v našich datech, pokud by veličiny *X, Y* byly nezávislé (tzv. *očekávaná* četnost, *expected* frequency) můžeme odhadnout pro každé políčko kontingenční tabulky jako

$$e_{ij} = n\ \hat{p}_{ij} = n\ \frac{n_{i\bullet}}{n}\ \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}\ n_{\bullet j}}{n}\ . \tag{6}$$

Nulovou hypotézu

H_0 : veličiny *X, Y* jsou nezávislé (7)

zamítneme tehdy, když pozorované četnosti n_{ij} budou podstatně odlišné od očekávaných četností e_{ij}, tj. hodnot, které bychom pozorovali v našich datech, pokud by nulová hypotéza platila. Testovou statistikou pro test nulové hypotézy (7) je

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}\ , \tag{8}$$

která má asymptoticky (tj. pro dostatečně velké četnosti) rozdělení χ^2 s $(R-1)\,(C-1)$ stupni volnosti, *přibližně* tedy platí

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \ \sim\ \chi^2_{(R-1)\,(C-1)}\ . \tag{9}$$

Jelikož (9) platí pouze přibližně, je při užití tohoto testu nutno posoudit, zda je splněna podmínka, že četnosti v tabulce jsou dostatečně velké. Obvykle se pro užití tohoto testu požaduje podmínka, aby všechny očekávané četnosti $e_{ij} \geq 1$ a naprostá většina (alespoň 80%) očekávaných četností byla $e_{ij} \geq 5$.

Kritickým oborem proto tento test nezávislosti je

$$W = \left[\chi^2_{(R-1)\,(C-1)}(1-\alpha), +\infty\right).$$

Zamítneme-li hypotézu o nezávislosti veličin X a Y, pak nás obvykle zajímá, které pozorované četnosti (která políčka kontingenční tabulky) se od četností očekávaných při nezávislosti veličin významně odchylují. Říkáme, že vyhledáváme zdroje závislosti.

Jedna z nejjednodušších metod posouzení těchto zdrojů závislosti je posouzení příspěvků jednotlivých políček tabulky k hodnotě testové statistiky (9). Velikost tohoto příspěvku je významná, když rozdíl pozorované a očekávané četnosti nelze považovat za náhodný, tj. tehdy, když

$$\frac{\left(n_{ij}-e_{ij}\right)^2}{e_{ij}} \geq \chi_1^2(1-\alpha), \qquad (10)$$

pro obvykle užívanou hodnotu $\alpha = 0,05$ $\chi_1^2(0,95) = 3,84$ (viz tabulky).

Pohodlnější je užít tzv. standardizovaná residua $\left(n_{ij}-e_{ij}\right)/\sqrt{e_{ij}}$, která mají přibližně normované normální rozdělení, tzn. významná jsou políčka s absolutní hodnotou standardizovaných residuí větší než 2. Užijeme-li standardizovaná residua, podle jejich znaménka vidíme, zda pozorovaná četnost je větší či menší než očekávaná.

Užití testu nezávislosti dvou nominálních veličin ukážeme na následujícím příkladu.

Příklad:

Máme posoudit, zda veličiny *Lokalita* a *Odruda* (data BI97) jsou nezávislé. Jinými slovy, zda zastoupeni obou odrůd všech čtyřech lokalitách můžeme považovat za shodné.

H_0 : *Lokalita* a *Odruda* jsou nezávislé veličiny

Výpočet provedeme s pomocí programu NCSS. V něm z menu *Analysis* vybereme *Descriptive Statistics*, dále *Cross Tabulation.* Zadáme veličinu *Lokalita a Odruda* jako *Table Columns* a *Table Row* . Pořadí ovlivňuje pouze tvar tabulek ve výstupu, nikoliv hodnotu spočtené testové statistiky. V šabloně *Report* vyznačíme, které výstupy požadujeme, v tomto příkladu *Counts* (pozorované četnosti), *Expected values* (očekávané četnosti), *Chi-square* (příspěvky políček do testové statistiky) a *Chi-square Stats* (testovou statistiku definovanou rov.(8)). Po provedení výpočtu dostaneme následující výstup, zde je uveden mírně zkrácen.

Cross Tabulation Report

Counts Section

		lokal			
odruda	**1**	**2**	**3**	**4**	**Total**
1	20	13	17	14	64
2	1	7	10	9	27
Total	21	20	27	23	91

Expected Counts Assuming Independence Section

		lokal			
odruda	**1**	**2**	**3**	**4**	**Total**
1	14.8	14.1	19.0	16.2	64.0
2	6.2	5.9	8.0	6.8	27.0
Total	21.0	20.0	27.0	23.0	91.0

Chi-Square Contribution Section

	lokal				
odruda	**1**	**2**	**3**	**4**	**Total**
1	1.85	0.08	0.21	0.29	2.43
2	4.39	0.19	0.49	0.69	5.76
Total	6.24	0.27	0.70	0.98	8.19

Chi-Square Statistics Section

Chi-Square	8.204673	
Degrees of Freedom	3.000000	
Probability Level	0.041966	Reject Ho

WARNING: At less one cell had a value less than 5.

V řádku Chi-Square vidíme, že hodnota testové statistiky je 8,20, odpovídající p = 0,042, tedy je menší než obvykle volená hladina významnosti $\alpha = 0{,}05$ a hypotézu o nezávislosti veličin *Lokalita* a *Odruda* můžeme na této hladině významnosti zamítnout, k čemuž nás ostatně nabádá i vysvětlující text ve výstupu, *Reject Ho.*

Varování, že některé pozorované četnosti v tabulce jsou malé, není příliš závažné, všechny očekávané četnosti jsou větší než 5, jak vidíme v části *Expected Counts Assuming Independence Section*

Podíváme-li se na zdroje závislosti (*Chi-Square Contribution Section*), vidíme, že pouze v jednom políčku (odruda = 2, lokalita = 1) je hodnota příspěvku políčka větší, než 3,84.

Celkově můžeme shrnout, že hypotézu o nezávislosti veličin *Lokalita* a *Odruda* jsme sice zamítnuli na hladině významnosti $\alpha = 0{,}05$, ale jen „s odřenýma ušima“ (hodnota p = 0,042 je jen o málo menší, než hladina významnosti) a navíc pouze jedno políčko tabulky přispívá významně k celkové hodnotě testové statistiky, takže zjištěnou závislost veličin *Lokalita* a *Odruda* můžeme přičítat jen malé četnosti odrůdy 2 v lokalitě 1. Jelikož víme, že test je asymptotický, tedy pouze přibližný, je nutno se závěrem, že sledované veličiny nejsou nezávislé, zacházet velmi opatrně.

Statistiku (8) lze užít pro test nezávislosti veličin, ale není vhodnou charakteristikou intenzity (těsnosti) závislosti, neboť její hodnota závisí na rozsahu výběru n. Zvětší-li se rozsah výběru k-krát při stejném proporcionálním obsazení políček tabulky, zvětší se i hodnota testové statistiky χ^2 k-krát. Pro spojité náhodné veličiny je mírou intenzity závislosti výběrový korelační koeficient nebo koeficient determinace. Podobné vlastnosti v případě dvou nominálních veličin, totiž nulovou hodnotu pro ideální nezávislost a hodnotu 1 pro dokonalou závislost mají některé z následujících charakteristik užívaných pro vyjádření těsnosti závislosti.

- Koeficient Φ $\Phi = \sqrt{\frac{\chi^2}{n}}$
- Cramerovo *V*, $V = \sqrt{\frac{\Phi^2}{\min(R, C)}}$
- Pearsonův koeficient kontingence $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$
- Čuprovův koeficient kontingence $T = \sqrt{\frac{\Phi^2}{(R-1)(C-1)}}$

Pro veličiny *Lokalita* a *Odruda* z uvedeného příkladu hodnoty těchto koeficientů získáme volbou *All Stats* v šabloně *Report*:

Phi	0.300269
Cramer's V	0.300269
Pearson's Contingency Coefficient	0.287584
Tschuprow's T	0.228155

Vidíme tedy, že vztah mezi veličinami opravdu není příliš těsný.

Poznámka:
Uvedený test nezávislosti můžeme užít nejen pro dvojici nominálních veličin, ale také pro veličiny ordinální. Je dokonce použitelný i pro spojité veličiny, pokud jejich hodnoty seskupíme do vhodných intervalů, ale v takové situaci je většinou pro posouzení vztahu veličin vhodnější korelační koeficient.

5.3 Znaménkový test

Obvyklá formulace jednovýběrového znaménkového testu je následující: Uvažujeme výběr ze spojitého rozdělení (nemusí být symetrické) a chceme testovat nulovou hypotézu, že medián tohoto rozdělení $\tilde{x}$ je roven jisté hodnotě x_0 proti jednostranné alternativě, např. že medián tohoto rozdělení je větší než x_0, tedy

H_0: $\tilde{x} = x_0$

H_1: $\tilde{x} > x_0$

Testovou statistikou je počet hodnot x_i ve výběru větší než x_0. Za platnosti nulové hypotézy má testová statistika Z binomické rozdělení, $Z \sim Bi(n, p)$, kde hodnota parametru $p = 0{,}5$ (z definice mediánu), n je rozsah výběru. Je-li hodnota testové statistiky rovna z, pak nulovou hypotézu zamítáme ve prospěch alternativy tehdy, když $P(Z \geq z) \leq \alpha$, kde α je zvolená hladina významnosti. Pravděpodobnost $P(Z \geq z) \leq \alpha$ lze snadno spočítat jako

$$P(Z \geq z\ ) = \sum_{k=z}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \frac{1}{2^n} \sum_{k=z}^{n} \binom{n}{k} = \frac{1}{2^n} \sum_{k=0}^{z} \binom{n}{k} .$$

Z vlastností binomického rozdělení můžeme určit střední hodnotu a rozptyl testové statistiky za platnosti nulové hypotézy

$$E(Z) = n\ p = \frac{n}{2} \qquad \text{a} \qquad \text{var}(Z) = n\ p\ (1-p) = \frac{n}{4} .$$

Pro větší rozsahy výběru lze aplikovat centrální limitní větu, pak normovaná náhodná veličina

$$U = \frac{Z - n/2}{\sqrt{n/4}} = \frac{2Z - n}{\sqrt{n}} \qquad (1)$$

má přibližně normované normální rozdělení $N(0, 1)$, což pak lze užít pro přibližné určení hodnoty $P(Z \geq z)$ u výběrů větších rozsahů.

Znaménkový test bývá velmi často užíván jako test párový, „přísná“ formulace tohoto párového testu je následující: Mějme dva závislé výběry ze spojitých rozdělení $(X_1, X_2, \ldots, X_n)$ a $(Y_1, Y_2, \ldots, Y_n)$ (tzn. dvě pozorování pro každý objekt) a testujeme hypotézu, že mediány obou veličin jsou shodné, většinou proti jednostranné alternativě, např.

H_0: $\tilde{X} = \tilde{Y}$

H_1: $\tilde{X} < \tilde{Y}$

Testovou statistikou je pak počet pozorování, kdy $Y_i > X_i$, další postup je stejný jako u jednovýběrového znaménkového testu.

Při volnější formulaci párového znaménkového testu se můžeme spokojit jen s kvalitativním porovnáním. Např. zjišťujeme, zda jistý léčebný postup přináší pacientům subjektivní pocit zlepšení zdravotního stavu. Léčebný postup je aplikován na n pacientů, dotazem na každého pacienta zjistíme, že u z pacientů nastalo zlepšení, u n-z zhoršení. Testujeme tedy hypotézu, že pravděpodobnost zlepšení je rovna 0,5 proti jednostranné alternativě, že tato pravděpodobnost je větší, tedy

H_0: $p = 0{,}5$
H_1: $p > 0{,}5$

Příklad:

Politická strana ABC si chtěla rychlým průzkumem ověřit, zda předvolební beseda přispěla ke zvýšení její důvěryhodnosti. V průzkumu bylo 16 náhodně vybraným účastníkům po besedě položena otázka, zda je jejich důvěra ve stranu ABC větší než před besedou. Odpovědí ANO bylo 10, NE odpovědělo 6 dotázaných. Lze se domnívat, že předvolební beseda přispěla ke zvýšení její důvěryhodnosti?

Odpověď na tuto otázku dá test hypotézy

H_0: $p = 0{,}5$ (beseda neměla vliv)

proti alternativě

H_1: $p > 0{,}5$ (beseda zvýšila důvěru)

Za platnosti H_0 má počet kladných odpovědí Z binomické rozdělení, $Z \sim Bi(16, 0{,}5)$.

$$P(Z \geq 10) = \frac{1}{2^{16}} \sum_{k=10}^{16} \binom{16}{k} = \frac{1}{2^{16}} \sum_{k=10}^{16} \binom{16}{16-k} =$$

$$= \frac{1}{2^{16}} \left[\binom{16}{6} + \binom{16}{5} + \cdots + \binom{16}{0} \right] \cong 0{,}22725$$

a tedy nulovou hypotézu zamítnout nemůžeme, tzn. není důvod věřit, že beseda zvýšila důvěryhodnost strany ABC.

Pokud bychom užili asymptotickou statistiku (1), dostaneme

$$u = \frac{2z - n}{\sqrt{n}} = \frac{2 \cdot 10 - 16}{\sqrt{16}} = 1.$$

Pravděpodobnost $P(U \geq 1) \cong 0{,}1587$, je o dost menší než přesná hodnota spočítaná z binomického rozdělení $Bi(16, 0{,}5)$, ale opět ani v tomto případě nemůžeme zamítnout nulovou hypotézu na jakékoliv rozumně zvolené hladině významnosti α. Dosti vysoký rozdíl mezi $P(Z \geq 10) \cong 0{,}22725$ a $P(U \geq 1) \cong 0{,}1587$, tj. přibližně 0,07 je způsoben malým rozsahem výběru (n = 16). Při větších hodnotách n se rozdíly snižují, jak ukazuje následující tabulka.

n	z	z/n	$P(Z \geq 10)$	u	$P(U \geq u)$
16	10	5/8	0,22725	1	0,15866
32	20	5/8	0,10766	$\sqrt{2}$	0,07868
64	40	5/8	0,02997	2	0,02275

V tabulce také vidíme, jak s rostoucím rozsahem výběru roste síla testu. Při stejné relativní četnosti kladných odpovědí 5/8 pro n = 16 a n = 32 nulovou hypotézu nezamítáme, pro n = 64 už bychom na hladině významnosti $\alpha = 0{,}05$ nulovou hypotézu zamítli.

5.4 Jednovýběrový Wilcoxonův test

Jednovýběrový Wilcoxonův test se podobně jako jednovýběrový znaménkový test užívá k testu hypotézy, že medián nějakého spojitého rozdělení je roven dané hodnotě. Oproti znaménkovému testu předpokládáme, že rozdělení, z něhož máme výběr $X_1, X_2, \ldots, X_n$, je nejen spojité, ale i *symetrické* kolem bodu *a*, tj. pro jeho hustotu f platí

$$f(a+x) = f(a-x)$$

a hodnota $a = \tilde{X}$ je hodnotou mediánu tohoto rozdělení. Jednovýběrovým Wilcoxonovým testem testujeme hypotézu

H_0: $\tilde{X} = x_0$

H_1: $\tilde{X} \neq x_0$

Předpokládejme, že žádná z hodnot X_i ve výběru není rovna x_0. Veličiny $Y_i = X_i - x_0$ (odchylky od předpokládané hodnoty x_0) seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty $|Y_{(1)}| \leq |Y_{(2)}| \leq \ldots \leq |Y_{(n)}|$. Nechť R_i^+ je pořadí hodnoty $|Y_{(i)}|$ v této posloupnosti. Je zřejmé, že za platnosti nulové hypotézy jsou $Y_1, Y_2, \ldots, Y_n$ nezávislé náhodné veličiny, jejichž rozdělení je symetrické kolem nuly. Proto by měly být součty pořadí nezáporných odchylek $S^+ = \sum_{i:Y_i \geq 0} R_i^+$ i záporných odchylek $S^- = \sum_{i:Y_i < 0} R_i^+$ zhruba stejné.

Samozřejmě platí, že součet pořadí je $S = S^+ + S^- = 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$ a nulovou hypotézu zamítneme, jestliže se hodnoty S^+, S^- podstatně liší, tzn. je-li $\min(S^+, S^-)$ menší nebo rovno kritické hodnotě $w_n(\alpha)$. Ta je pro menší hodnoty n tabelována (viz část Statistické tabulky nebo např. Anděl, 1993). Tabelované kritické hodnoty jsou spočítány kombinatoricky s využitím klasické pravděpodobnosti.

Pro větší rozsahy výběru lze užít asymptotickou aproximaci. Za platnosti nulové hypotézy je

$$E(S^+) = \frac{n(n+1)}{4} \qquad \text{a} \qquad \text{var}(S^+) = \frac{1}{24} n(n+1)(2n+1)$$

a bylo také dokázáno, že s rostoucím n se rozdělení statistiky S^+ blíží normálnímu rozdělení. Pak můžeme k testu nulové hypotézy užít statistiku

$$U = \frac{S^+ - E(S^+)}{\sqrt{\text{var}(S^+)}},$$

která má přibližně normované normální rozdělení $N(0,1)$. H_0 zamítneme, je-li absolutní hodnota této statistiky $|U| \geq u(1-\alpha/2)$, kde $u(1-\alpha/2)$ je $(1-\alpha/2)$ - kvantil rozdělení $N(0,1)$.

Příklad:

10 pokusných osob mělo bez předchozího výcviku nezávisle na sobě odhadnout, kdy od daného signálu uplyne jedna minuta. Byly získány následující výsledky (v sekundách): 53, 48, 45, 55, 63, 51, 66, 56, 50, 58.
Naším úkolem je testovat hypotézu H_0: $\tilde{X} = 60s$ proti alternativě H_1: $\tilde{X} \neq 60s$, tedy rozhodnout, zda naše pozorování nám poskytuje důvod odmítnout představu, že polovina osob v populaci délku jedné minuty podhodnocuje a polovina nadhodnocuje.

X_i	53	48	45	55	63	51	66	56	50	58
$Y_i = X_i - 60$	-7	-12	-15	-5	3	-9	6	-4	-10	-2

Hodnoty Y_i uspořádáme do neklesající posloupnosti podle $|Y_{(i)}|$:

pořadí	1	**2**	3	4	**5**	6	7	8	9	10
$Y_i = X_i - 60$	-2	**3**	-4	-5	**6**	-7	-9	-10	-12	-15

Kladné hodnoty Y_i jsou zvýrazněny.

Pak

$S^+ = 2 + 5 = 7$,

$$S^- = S - S^+ = \frac{10(10+1)}{2} - 7 = 55 - 7 = 48,$$

$\min(S^+, S^-) = 7$.

Kritická hodnota v tabulce je $w_{10}(0{,}05) = 8$, tzn. že H_0: $\tilde{X} = 60s$ můžeme zamítnout.

Pokud bychom i pro tak malý rozsah výběru užili asymptotický postup (je však doporučován pro rozsah výběru $n > 20$), dostaneme

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

$$\mathrm{var}(S^+) = \frac{n(n+1)(2n+1)}{24} = \frac{10 \cdot 11 \cdot 21}{24} = \frac{385}{24} = 96{,}25$$

$$U = \frac{S^+ - E(S^+)}{\sqrt{\mathrm{var}(S^+)}} = \frac{7 - 27{,}5}{\sqrt{96{,}25}} \cong -2{,}09$$

Protože $|U| \geq 1{,}96$, ($u(0{,}975) = 1{,}96$, viz tabulka normovaného normálního rozdělení, zamítli bychom nulovou hypotézu na hladině významnosti $\alpha = 0{,}05$ i tímto asymptotickým postupem.

Kdybychom v tomto příkladu užili znaménkový test, nulovou hypotézu bychom zamítnout nemohli. Při oboustranné alternativě H_1: $\tilde{X} \neq x_0$ můžeme zamítnout, když hodnota testové statistiky Z (počet kladných znamének) je buď příliš malá ($Z \leq k_1$) nebo příliš velká ($Z \geq k_2$). Hodnoty k_1, k_2, jsou nejmenší, resp. největší z čísel, pro která platí

$$P(Z \leq k_1) \leq \frac{\alpha}{2}, \qquad P(Z \geq k_2) \leq \frac{\alpha}{2}.$$

Za platnosti nulové hypotézy má $Z \sim Bi(n; 0{,}5)$, tzn. rozdělení je symetrické a $k_2 = n - k_1$. Hodnotu k_1 pro n = 10 a $\alpha = 0{,}05$ určíme takto:

k	$P(Z = k)$	$P(Z \leq k)$
0	$\frac{1}{2^{10}}\binom{10}{0} = \frac{1}{1024}$	0,0010
1	$\frac{1}{2^{10}}\binom{10}{1} = \frac{10}{1024}$	0,0108
2	$\frac{1}{2^{10}}\binom{10}{2} = \frac{45}{1024}$	0,0547

Hodnota $k_1 = 1$, počet kladných odchylek je roven 2, tedy větší než k_1 a nulovou hypotézu bychom zamítnout nemohli.

Uvedený příklad ilustruje, že Wilcoxonův jednovýběrový test je silnější než test znaménkový. Všimněme si, že $P(Z \leq 2) = 0{,}0547$, tzn. větší než $\alpha = 0{,}05$. Tedy znaménkový test by na této hladině významnosti nezamítnul H_0: $\tilde{X} = 60s$ ani proti jednostranné alternativě H_1: $\tilde{X} < 60s$.

Poznámka:

Používáme-li statistický software pro vyhodnocení neparametrických testů, je na místě obezřetnost při interpretaci výstupu z programu. Zejména při interpretaci tzv. *p-value*, Některé statistické programy uvádějí jako *p-value* jen hodnotu z asymptotického testu, neboť určení přesné hodnoty pro neparametrický test bývá výpočetně náročné. Proto zejména při zpracování výběrů menších rozsahů pečlivě pročtěte manuál nebo help programu a pokud je hodnota ve výstupu programu jen asymptotická, použijte kritické hodnoty ze statistických tabulek.

5.5 Dvouvýběrový Wilcoxonův test

Dvouvýběrový Wilcoxonův test je neparametrickou obdobou dvouvýběrového *t*-testu. V případě dvouvýběrového *t*-testu se testuje hypotéza o rovnosti středních hodnot dvou normálních rozdělení, ze kterých jsou dva nezávislé výběry. Wilcoxonův test je založen na pořadí a lze ho použít i pro výběry, které nejsou z normálních rozdělení.

Uvažujme dva nezávislé výběry ze dvou spojitých rozdělení:

- $X_1, X_2, \ldots, X_m$ náhodný výběr z rozdělení s distribuční funkcí F
- $Y_1, Y_2, \ldots, Y_n$ náhodný výběr z rozdělení s distribuční funkcí G

Wilcoxonův dvouvýběrový test je obecně zformulován jako test hypotézy o shodě distribučních funkcí

H_0: $F = G$
H_1: $F \neq G$

Ale většinou alternativu chápeme jako posunutí, tj. H_1: $G(x) = F(x - \Delta)$, $\Delta \neq 0$, pro kterou je tento test citlivý (má přijatelnou sílu). Pokud se distribuční funkce liší spíše jen rozptylem nebo tvarem, není užití dvouvýběrového Wilcoxonova testu vhodné.

Wilcoxonův dvouvýběrový test je založen na pořadí pozorovaných hodnot v tzv. sdruženém výběru. Všech $m+n$ hodnot $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ uspořádáme vzestupně, za platnosti nulové hypotézy jsou oba výběry z téhož rozdělení. Pořadí R_i ve sdruženém výběru má tedy hodnoty $1, 2, \ldots, m+n$. Pokud se ve sdruženém výběru vyskytují shodné hodnoty, přiřadíme jim odpovídající průměrné pořadí. Součet pořadí hodnot $X_1, X_2, \ldots, X_m$ označíme T_1, součet pořadí hodnot $Y_1, Y_2, \ldots, Y_n$ označíme T_2.
Je zřejmé, že

$$T_1 + T_2 = \sum_{i=1}^{m+n} R_i = \frac{1}{2}(m+n)(m+n+1)$$

a dále, že střední hodnoty ET_1 a ET_2 jsou za platnosti H_0 rovny násobku průměrného pořadí a rozsahu výběru, tj.

$$ET_1 = \frac{1}{2}m(m+n+1) \text{ a } ET_2 = \frac{1}{2}n(m+n+1) \ .$$

Lze dokázat, že

$$\operatorname{var} T_1 = \operatorname{var} T_2 = \frac{1}{12}m\,n\,(m+n+1) \ .$$

Nulovou hypotézu pak můžeme zamítnout, když statistika T_1 (nebo T_2) se příliš odlišuje od střední hodnoty očekávané za platnosti H_0. Pro větší rozsahy výběrů (m >10, n >10) lze k testu užít statistiku

$\frac{T_1 - ET_1}{\sqrt{\text{var}\, T_1}}$, která má přibližně rozdělení $N(0,1)$.

Místo veličiny T_1 (nebo T_2) můžeme užít statistiky

$$U_1 = mn + \frac{1}{2}m(m+1) - T_1$$

a

$$U_2 = mn + \frac{1}{2}n(n+1) - T_2$$

Snadno lze ukázat, že $U_1 + U_2 = mn$. Testu založeném na této statistice se říká Mannův-Whitneyův test a je ekvivalentní Wilcoxonovu testu. Nulovou hypotézu zamítneme, když $\min(U_1, U_2)$ je menší nebo rovno tabelované kritické hodnotě, viz část Statistické tabulky.

Pro větší rozsahy výběrů (m >10, n >10) lze k testu užít statistiku

$\frac{U_1 - EU_1}{\sqrt{\text{var}\, U_1}}$,

kde $E(U_1) = \frac{1}{2}m\ n$ a $\text{var}(U_1) = \frac{1}{12}mn\ (m+n+1)$, která má přibližně normované normální rozdělení $N(0,1)$.

Příklad:

Bylo vybráno 13 polí stejné kvality. Na 8 z nich se zkoušel nový způsob hnojení, zbývajících 5 bylo ošetřeno běžným způsobem. Výnosy pšenice v tunách na hektar jsou označeny X_i u nového a Y_i u běžného způsobu hnojení.

X_i	5,7	5,5	4,3	5,9	5,2	5,6	5,8	5,1
Y_i	5,0	4,5	4,2	5,4	4,4			

Máme zjistit, zda způsob hnojení má vliv na výnos pšenice.

Seřadíme hodnoty sdruženého výběru (X_i a Y_i) vzestupně:

Pořadí	X_i a Y_i	Způsob hnojení	Pořadí(X_i)
1	4.2	běžný	
2	4.3	nový	2
3	4.4	běžný	
4	4.5	běžný	
5	5.0	běžný	
6	5.1	nový	6
7	5.2	nový	7
8	5.4	běžný	
9	5.5	nový	9
10	5.6	nový	10
11	5.7	nový	11
12	5.8	nový	12
13	5.9	nový	13
			$T_1 = 70$

$$U_1 = m\,n + \frac{1}{2}m(m+1) - T_1 = 8 \cdot 5 + \frac{1}{2}8 \cdot 9 - 70 = 6,$$

$$U_2 = m\,n - U_1 = 40 - 6 = 34,$$

$$\min(U_1, U_2) = 6.$$

Jelikož kritická hodnota pro $\alpha = 0{,}05$ je 6, znamená to, $\min(U_1, U_2) = 6$ je v kritickém oboru, a proto zamítáme na hladině významnosti $\alpha = 0{,}05$ nulovou hypotézu, že způsob hnojení nemá vliv na výnos pšenice.

Povšimněme si, že hodnotu statistiky U_1 můžeme určit rychleji a jednodušeji, neboť U_1 znamená počet hodnot z druhého výběru, které následují ve sdruženém výběru za hodnotami z výběru prvního. Názorně to ukážeme na řešeném příkladu. Každý z výběrů uspořádáme vzestupně:

X_i	4,3	5,1	5,2	5,5	5,6	5,7	5,8	5,9
Y_i	4,2	4,4	4,5	5,0	5,4			

Pak už jen zjistíme počet hodnot ve druhém výběru, které jsou větší než hodnoty v prvním výběru:

počet hodnot $Y_i > 4{,}3$	4
počet hodnot $Y_i > 5{,}1$	1
počet hodnot $Y_i > 5{,}2$	1
počet hodnot $Y_i > 5{,}5$	0
$\vdots$	$\vdots$
počet hodnot $Y_i > 5{,}9$	0
	$U_1 = 6$

$U_2 = mn - U_1 = 40 - 6 = 34$, $\min(U_1, U_2) = 6$ a výpočet testové statistiky je hotov.

5.6 Kruskalův-Wallisův test

Kruskalův-Wallisův test je neparametrickou obdobou analýzy rozptylu s jednoduchým tříděním (one-way ANOVA). Je to zobecnění dvouvýběrového Wilcoxonova testu na situaci, kdy počet výběrů je větší než dva.

Nechť $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ je výběr z rozdělení se spojitou distribuční funkcí F_i. Uvažujme I takových výběrů, tj. $i = 1, 2, \ldots, I$. Chceme testovat hypotézu, že všechny distribuční funkce rozdělení, z nichž jsou výběry, jsou shodné

$$H_0: \qquad F_1 = F_2 = \ldots = F_I$$

proti alternativě, že aspoň v jedné dvojici se distribuční funkce liší. Všechny hodnoty Y_{ij} dohromady tvoří sdružený výběr o rozsahu $n_1 + n_2 + \ldots + n_I = n$. Hodnoty Y_{ij} ve sdruženém výběru se uspořádají vzestupně, určí se jejich pořadí R_{ij} a součty pořadí ve výběrech:

Výběr	Pořadí	Součet pořadí
1	$R_{11}, R_{12}, \ldots, R_{1n_i}$	T_1
2	$R_{21}, R_{22}, \ldots, R_{2n_i}$	T_2
$\vdots$	$\vdots$	$\vdots$
I	$R_{I1}, R_{I2}, \ldots, R_{In_i}$	T_I

Celkový součet všech pořadí je

$$T_1 + T_2 + \ldots + T_I = \frac{1}{2} n(n+1)$$

Střední hodnoty součtů pořadí jsou

$$ET_i = \frac{1}{2} n_i (n+1), \qquad i = 1, 2, \ldots, I$$

a testová statistika Q pro test nulové hypotézy je založena na součtu čtverců odchylek pozorovaných hodnot součtů pořadí od jejich středních hodnot

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{n_i} \left[T_i - \frac{1}{2} n_i (n+1) \right]^2 = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{T_i^2}{n_i} - 3(n+1)$$

Pro větší rozsahy výběrů má tato statistika přibližně rozdělení χ^2_{I-1}, takže H_0 zamítneme, je-li $Q \geq x_{I-1}(1-\alpha)$, kde $x_{I-1}(1-\alpha)$ je kvantil tohoto rozdělení. Pro malé rozsahy výběrů je možno použít některý ze statistických programů, např. StatXact, které počítají buď kombinatoricky nebo metodou Monte Carlo hodnotu *p-value* odpovídající zjištěné hodnotě statistiky *Q*.

Příklad:
Domy ve třech obcích se prodávají za následující ceny (tisíce EUR):

Obec	ceny			
A	39	45	71	
B	51	63	88	97
C	99	150	260	

Testujte, zda ceny domů jsou ze stejného rozdělení.

Nejdříve spočítáme součty pořadí v jednotlivých výběrech.

Obec	n_i	Pořadí				T_i
A	3	1	2	5		8
B	4	3	4	6	7	20
C	3	8	9	10		27
	10					

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{n_i} \left[T_i - \frac{1}{2} n_i (n+1) \right]^2 =$$

$$= \frac{12}{10 \cdot 11} \left(\frac{8^2}{3} + \frac{20^2}{4} + \frac{27^2}{3} \right) - 3 \cdot 11 = 6{,}745$$

Hodnota $x_2(0{,}95) = 5{,}9915$, tedy $Q = 6{,}745$ je v kritickém oboru a nulovou hypotézu zamítáme.

P-value odpovídající hodnotě statistiky $Q = 6{,}745$, tj. $P(X \geq 6{,}745)$, když $X \sim \chi^2_2$, je $p = 0{,}0343$. Přesná hodnota p spočítaná pomocí specializovaného

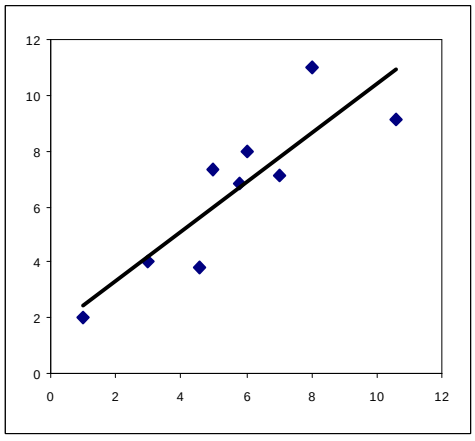
programu StatXact je $p = 0{,}010$. Vidíme tedy, že pro tak malé rozsahy výběrů se dosti liší od hodnoty p, získané z asymptotického rozdělení statistiky Q. Nicméně v tomto případě oba výsledky vedou k zamítnutí nulové hypotézy na hladině významnosti $\alpha = 0{,}05$.

5.7 Spearmanův koeficient pořadové korelace

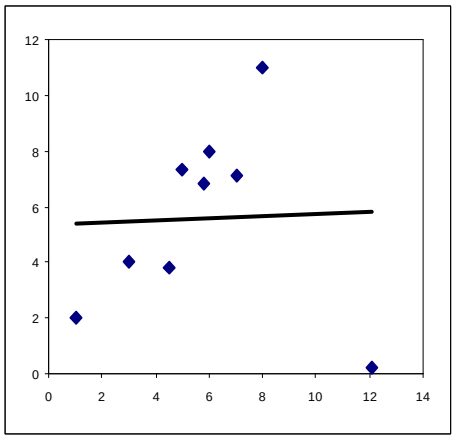
Jak víme, koeficient korelace vyjadřuje těsnost lineárního vztahu dvojice veličin. Korelační koeficient nabývá hodnot z intervalu $\langle -1,1 \rangle$. Výběrový korelační koeficient r_{xy} (tzv. Pearsonův) lze vyjádřit jako

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}} \qquad (1)$$

Víme už, že dobře „funguje“ pro posuzování vztahu dvou náhodných veličin majících dvourozměrné normální rozdělení. Pokud je rozdělení jiné než normální nebo výběr obsahuje odlehlé hodnoty, korelační koeficient r_{xy} o těsnosti vztahu veličin nemusí poskytovat dobrý obraz, viz následující obrázek, kdy jeden odlehlý bod velmi podstatně změnil hodnotu korelačního koeficientu.



$r_{xy} = 0{,}87$



$r_{xy} = 0{,}04$

Spearmanův koeficient korelace dostaneme tak, že místo původních hodnot X_i, Y_i dosadíme do vztahu (1) jejich pořadí.

Nechť $(X_1,Y_1)^T, (X_2,Y_2)^T, \ldots, (X_n,Y_n)^T$ je výběr ze spojitého dvourozměrného rozdělení,
$R_1, R_2, \ldots, R_n$ je pořadí hodnot $X_1, X_2, \ldots, X_n$,
$Q_1, Q_2, \ldots, Q_n$ je pořadí hodnot $Y_1, Y_2, \ldots, Y_n$.

Dvojice $(X_1,Y_1)^T, (X_2,Y_2)^T, \ldots, (X_n,Y_n)^T$ můžeme uspořádat vzestupně podle hodnot $X_1, X_2, \ldots, X_n$, pak $R_i = i$, $i = 1,2,\ldots,n$. Dosadíme-li do (1) za hodnoty X_i, Y_i jejich pořadí R_i a Q_i, dostaneme Spearmanův koeficient pořadové korelace r_S:

$$r_S = \frac{\sum_{i=1}^{n} R_i Q_i - n\overline{R}\,\overline{Q}}{\sum_{i=1}^{n} R_i^2 - n\overline{R}^2} \qquad (2)$$

Jelikož

$$\overline{R} = \overline{Q} = \frac{\sum_{i=1}^{n} R_i}{n} = \frac{n+1}{2},$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} Q_i^2 = = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{i=1}^{n} R_i\, Q_i = \frac{1}{2}\left(\sum_{i=1}^{n} R_i^2 + \sum_{i=1}^{n} Q_i^2\right) - \frac{1}{2}\sum_{i=1}^{n}(R_i - Q_i)^2 = \sum_{i=1}^{n} R_i^2 - \frac{1}{2}\sum_{i=1}^{n}(R_i - Q_i)^2,$$

můžeme vztah (2) upravit na

$$r_S = \frac{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} - \frac{1}{2}\sum_{i=1}^{n}(R_i - Q_i)^2}{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}} =$$

$$= 1 - \frac{\frac{1}{2}\sum_{i=1}^{n}(R_i - Q_i)^2}{\frac{2n(n+1)(2n+1) - 3n(n+1)^2}{12}} = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n(n^2-1)}$$

Označíme-li rozdíl v pořadí i-tého pozorování $d_i = R_i - Q_i$, Spearmanův korelační koeficient je

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (3)$$

- Jsou-li obě veličiny uspořádány shodně, tzn. $R_i = Q_i$, pak $\left(\sum_{i=1}^{n} d_i^2\right)_{\min} = 0$ a Spearmanův korelační koeficient $r_S = 1$.

- Jsou-li obě veličiny uspořádány opačně, tzn. $d_i = i - (n+1-i)$, $i = 1, 2, \ldots, n$, je pak součet čtverců rozdílu pořadí roven své maximální hodnotě $\left(\sum_{i=1}^{n} d_i^2\right)_{\max} = \frac{n(n^2-1)}{3}$ a Spearmanův korelační koeficient $r_S = -1$.

- Při náhodném uspořádání je součet čtverců rozdílu pořadí roven průměrné hodnotě $\frac{1}{2}\left[\left(\sum_{i=1}^{n} d_i^2\right)_{\min} + \left(\sum_{i=1}^{n} d_i^2\right)_{\max}\right] = \frac{n(n^2-1)}{6}$ a Spearmanův korelační koeficient $r_S = 0$.

Pomocí Spearmanova korelačního koeficientu lze testovat hypotézu o nekorelovanosti veličin X a Y. Pro malé rozsahy výběru jsou kritické hodnoty Spearmanova korelačního koeficientu tabelovány, viz např. část Statistické tabulky na konci tohoto textu. Pro $n > 30$ lze užít asymptotickou normalitu a nulovou hypotézu o nekorelovanosti veličin X a Y zamítnout při

$$|r_S| \geq \frac{u\left(1-\frac{\alpha}{2}\right)}{\sqrt{n-1}},$$

kde $u(1-\alpha/2)$ je kvantil normovaného normálního rozdělení $N(0,1)$.

Spearmanův korelační koeficient můžeme užít i pro hodnocení vztahu dvou veličin, i když jedna či obě jsou měřeny v ordinální škále.

Příklad:
Dva degustátoři hodnotili 7 vzorků vína. Vzorky jsou označeny A, B, C, D, E, F, G. Degustátoři ohodnotili pořadí vzorků vín takto

Degustátor	Uspořádání						
1	B	C	F	G	D	A	E
2	B	F	G	C	A	D	E

Ohodnoťte shodu degustátorů.

Určíme hodnoty pořadí R_i, Q_i:

vzorek	R_i	Q_i	d_i	d_i^2
B	1	1	0	0
C	2	4	-2	4
F	3	2	1	1
G	4	3	1	1
D	5	6	-1	1
A	6	5	1	1
E	7	7	0	0
				8

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8}{7 \cdot (7^2-1)} \cong 0{,}857$$

V tabulce 7 nalezneme, že kritická hodnota pro $\alpha = 0{,}05$ je 0,745. Zamítneme tedy na této hladině významnosti hypotézu, že hodnocení degustátorů nejsou korelované. Jinými slovy zamítáme hypotézu, že degustátoři vínu nerozumějí a vzorky uspořádali náhodně.

Kontrolní otázky:

1. *Proč se používají neparametrické metody? Jaké mají výhody a nevýhody v porovnání se svými parametrickými protějšky?*
2. *Zkuste zdůvodnit, proč jednovýběrový Wilcoxonův test je silnější než test znaménkový.*
3. *Které z testů uvedených v této kapitole jsou založeny na pořadí pozorovaných hodnot?*
4. *Proč je Spearmanův koeficient korelace méně citlivý na odlehlé hodnoty než Pearsonův korelační koeficient?*
5. *Jaká nulová hypotéza se testuje testem Chí-kvadrát popsaným v kapitole 5.6?*
6. *Příklad řešený v kapitole 5.6 (Chí-kvadrát test nezávislosti) spočtěte v Excelu (pro úsporu práce vhodně využijte absolutní a relativní adresy buněk při zápisu výrazů pro výpočet očekávaných četností a dalších veličin potřebných pro výpočet, abyste aritmetické výrazy mohli kopírovat).*

Pojmy k zapamatování:

- *neparametrické metody,*
- *statistiky založené na pořadí hodnot,*
- *znaménkový test, Mannův-Whitneyův test, Spearmanův koeficient korelace,*
- *kontingenční tabulka, test nezávislosti dvou nominálních veličin.*

Korespondenční úloha č. 4

Bude zadána na začátku semestru.

6 Programové prostředky pro statistické výpočty

Tato kapitola by vám měla pomoci v orientaci v programových prostředcích užívaných ve statistických výpočtech a analýze dat. Jsou zde uvedeny společné rysy těchto softwarových produktů. Podrobněji jsou zmíněny tabulkový procesor Excel a statistický paket NCSS, neboť s těmito produkty se nejpravděpodobněji setkáte při řešení vašich úloh při studiu na Ostravské universitě. Při prvním čtení této kapitoly, na které by mělo stačit 2 až 3 hodiny, postačí, když získáte orientaci v základních problémech a obtížích, se kterými se můžete ve výpočtech a interpretaci výsledků setkat. Spíše počítejte s tím, že při řešení konkrétního problému se budete k této kapitole vracet.

Podpora statistického zpracování dat je součástí mnoha obecných programových systémů orientovaných na práci s databázemi, na grafické zpracování dat, matematických programových prostředků (Matlab, Mathematica) a kromě toho existuje několik desítek specializovaných statistických programových paketů. Společným rysem těchto programových prostředků jsou operace s datovou maticí, tj. dvojrozměrnou tabulkou, ve které sloupce jsou veličiny a řádky pozorované objekty. Pro práci s tabulkami jsou určeny i tabulkové procesory (např. Excel), které jsou vybaveny celou řadou statistických funkcí a grafických prostředků. Tyto programové prostředky značně usnadňují statistické výpočty a dovolují uživateli soustředit se na správné použití statistických metod, nikoliv na výpočetní námahu.

6.1 Tabulkový procesor Excel

Excel je typickým představitelem tabulkových procesorů, některá jeho verze je dostupná prakticky na každém počítači. Standardní součástí Excelu je několik desítek statistických funkcí, které mohou být užity při statistických výpočtech. Je vybaven i poměrně kvalitní grafikou, která dovoluje pohodlné kreslení statistických grafů (prozatím s výjimkou např. krabicových diagramů a pár některých dalších ve statistice užívaných typů grafů).

Kromě toho lze Excel rozšířit o standardně dodávaný doplněk *Analýza dat*, který pokrývá prakticky všechny metody vysvětlované v základních kursech statistické analýzy dat. Vzhledem k tomu, že Excel je tzv. lokalizován, to znamená, že podrobná nápověda ke všem funkcím je k dispozici v češtině, a práce s tabulkovými procesory je součástí výuky předcházejících předmětů, nebudeme se jím nyní podrobněji zabývat. Pouze připojujeme upozornění na některé nedostatky zjištěné ve statistických funkcích a doplňku Analýza dat.

Dosti obecně lze říci, že zejména v české verzi Excelu se opakovaně vyskytují zmatení pojmů. Zaměňují se pojmy „průměr“ a „střední hodnota“, vysvětlení parametrů funkcí je zmatečné, výstupy z modulů doplňku *Analýza dat* jsou často redundantní (součet i průměr, směrodatná odchylka, směrodatná odchylka průměru i rozptyl, atd.), zbytečně vysoký počet významných číslic v .číselných hodnotách apod. Některé takové nedostatky ukazuje následující tabulka výstupu z modulu *Popisná statistika* doplňku *Analýza dat:*

Sloupec1	
Stř. hodnota	99.3956
Chyba stř. hodnoty	2.743841
Medián	99
Modus	101
Směr. odchylka	26.17458
Rozptyl výběru	685.1084
Špičatost	0.194895
Šikmost	0.164807
Rozdíl max-min	131
Minimum	40
Maximum	171
Součet	9045
Počet	91

Stř. hodnota je užita místo slova *Průměr*, Chyba stř. hodnoty místo *Směrodatná odchylka průměru*. Rozptyl výběru místo *Výběrový rozptyl*. Počet desetinných míst je nadbytečný.

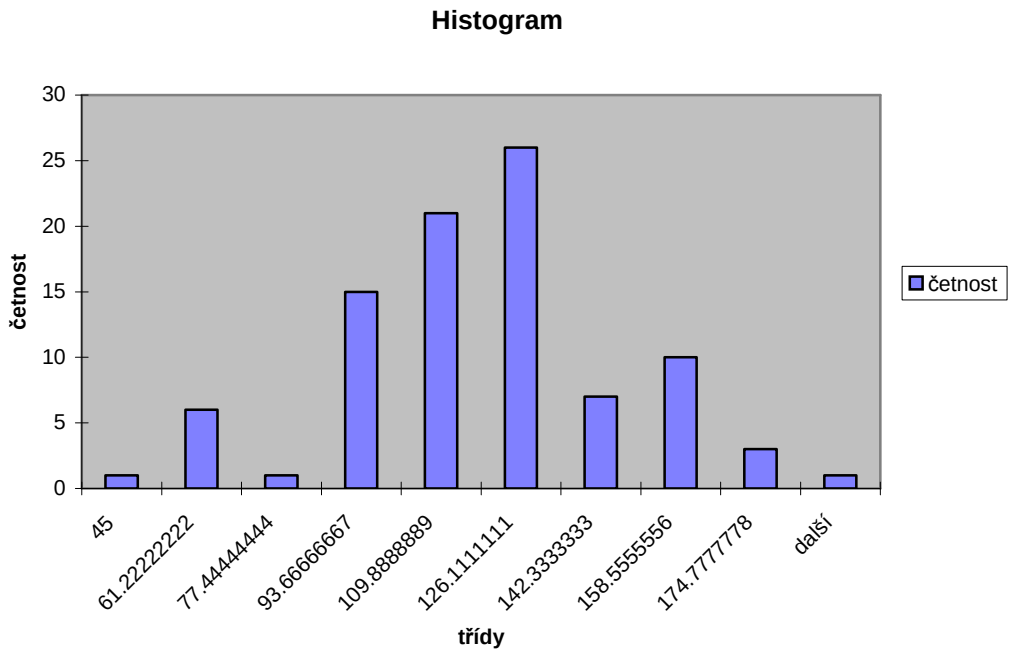
Chyby nalezneme i v jiných modulech doplňku *Analýza dat* pro běžné statistické testy. Např. dvouvýběrový *t*-test poskytne následující výstup:

Dvouvýběrový t-test s rovností rozptylů

	Soubor 1	Soubor 2
stř. hodnota	111.9219	107.7778
rozptyl	734.0097	831.0256
pozorování	64	27
společný rozptyl	762.3514	
hyp. rozdíl st. hodnot	0	
rozdíl	89	
t stat	0.654039	
P(T<=t) (1)	0.257387	
t krit (1)	1.662156	
P(T<=t) (2)	0.514773	
t krit (2)	1.986978	

Opět Stř. hodnota je užita místo *Průměr*. Pro uživatele rozlišujícího mezi jednostranným a oboustranným testem je výstup redundantní, uživateli mezi těmito variantami nerozlišujícímu tato redundance stejně nepomůže. Zájem může vzbudit statistika označená jako „*rozdíl*“. Skutečnost, že platí $rozdíl = n_1 + n_2 - 2$ (tedy je roven počtu *stupňů volnosti*) svádí k domněnce, že zkratku *df* interpretoval překladatel jako anglické *difference* a přeložil do češtiny. Tato chyba se vyskytuje ve většině testů implementovaných v doplňku *Analýza dat*.

Často užívaným modulem doplňku Analýzy dat je *Histogram*. S využitím implicitního nastavení vstupních parametrů můžete dostat následující obrázek:



Legenda a nadpis „Histogram“jsou zbytečné, jen zabírají místo, popis vodorovné osy neříká nic. Sloupce nejsou nad celou šířkou intervalů, počet významných číslic v popisu pod sloupci je nesmyslně velký. To lze napravit vhodnější volbou vstupních parametrů nebo dodatečnou úpravou grafu. Závažnějším nedostatkem je, že hodnoty popisující středy sloupců (středy jednotlivých intervalů) nejsou hodnoty odpovídající středu, ale pravému okraji intervalu.

Mezi statistickými funkcemi jsou i funkce pro výpočet hodnot distribučních funkcí a kvantilů často užívaných rozdělení. U nich je nápověda matoucí a místy zcela nesmyslná. Ukážeme to na příkladu funkce NORMDIST a z jejího helpu se dočteme následující:

nápověda:
NORMDIST
Vrací **kumulativní normální rozdělení** se zadanou střední hodnotou a směrodatnou odchylkou. Tato funkce má ve statistice velmi široké použití, včetně testování hypotéz.
Syntaxe
NORMDIST(x; průměr; směrod_odch; kumulativní)
X je **hodnota, pro niž počítáme rozdělení**.
Průměr je **aritmetický průměr rozdělení**.
Směrod_odch je směrodatná odchylka rozdělení.
Kumulativní je logická hodnota, která určuje tvar funkce. Pokud kumulativní je PRAVDA, NORMDIST vrací kumulativní distribuční funkci; je-li NEPRAVDA, vrací **pravděpodobnostní míru**.
.....
konec nápovědy.

Funkce NORMDIST jen stěží může vracet „*kumulativní normální rozdělení*“, ale z popisu lze vytušit, že tím je míněna hodnota *distribuční funkce* nebo *hustoty* (nikoli „*pravděpodobnostní míra*“) normálního rozdělení podle toho, jakou zadáme hodnotu posledního vstupního parametru „*kumulativní*“. Druhý parametr je vysvětlen jako „*aritmetický průměr rozdělení*“, což patrně vzniklo chybným překladem anglického termínu *mean*, který měl být přeložen jako *střední hodnota*.

Pozor při užívání funkcí navracející hodnoty kvantilů běžných rodělení. Funkce NORMINV s parametry *p*, μ, σ vrátí hodnotu příslušného kvantilu $x(p) = \sigma\, u(p) + \mu$, tedy na př. NORMINV(0,238; 175; 7]) vrátí hodnotu 170,01.

U jiných rozdělení je to však trochu odlišné. Pro určení kvantilů rozdělení χ^2 můžeme užít funkci CHIINV, která má dva vstupní parametry. Chceme-li, aby funkce vrátila hodnotu *p*-kvantilu, musíme její parametry zadat jako (*1-p)* a počet stupňů volnosti, takže např. zadáním CHIINV(0,05; 1) dostaneme hodnotu 0,95-kvantilu rozdělení χ^2_1, $x(0,95) =$ 3,84145. Ačkoliv v nápovědě k funkci CHIINV je, že to je inverzní funkce k distribuční funkci, není to úplně pravdivé. Funkce je navržena tak, aby vracela tzv. kritickou hodnotu (hranici kritického oboru) pro zadanou hodnotu významnosti α jako první parametr.

Podobně se chová i funkce FINV, p-kvantil dostaneme při zadání parametrů 1-*p*, *m*, *n*, např. FINV(0,05; 10; 20) vrátí hodnotu 2,347875, což je 0,95-kvantil.

Ještě o trochu komplikovanější to je u funkce funkce TINV pro výpočet kvantilů *t* - rozdělení. Pokud chceme, aby funkce TINV spočítala *p*-kvantil, musíme vstupní parametry zadat jako (1-2*p*, počet stupňů volnosti), např. vrací hodnotu *p*-kvantilu, např. TINV(0,05; 25) vrátí hodnotu 2,0595, což je hodnota 0,95 kvantilu *t* – rozdělení s 25 stupni volnosti. Podobně jako předchozí dvě funkce, i TINV vrací kritickou hodnotu, ale pro dvoustranný *t*-test.

Užíváte-li pro statistické výpočty Excel, vždy velmi pečlivě zkoumejte, co vlastně vám ve výsledcích Excel poskytuje a výstupy z Excelu, zejména z jeho české lokalizované verse, nepřenášejte bez rozmyslu do svých prezentací a dokumentů. Berte je jako polotovar, jehož editací a většinou i zkrácení lze vytvořit opravdu kvalitní a přehledný výstup.

6.2 Statistické programové systémy

Statistických programů komerčně šířených existuje veliké množství. Jako nejpopulárnější příklady můžeme zmínit SPSS, SAS, S-Plus, Statistica, Stata, Minitab, Unistat nebo NCSS. To jsou tzv. obecné, tj. pokrývají celou škálu statistických metod, jiné jsou specializované na analýzu některých dat (časové řady, kategoriální data apod.). Všechny statistické programy však mají tyto základní funkce:

- import dat (vstup datové tabulky připravené v jiném programovém prostředku, třeba v Excelu nebo v nějakém databázovém prostředku),
- manipulace s daty (transformace, uspořádávaní dat, výběry podmnožin datové matice, spojování datových matic),
- základní deskriptivní statistiky,
- grafické prostředky,
- ukládání dat k snadnému využití pro další zpracování (tzv. savefile),
- export dat (ve formátech vhodných pro jiné programové prostředky),
- presentace výsledků ve formě souborů pro další zpracování textovými procesory,
- řadu statistických metod, jako např. t-testy, analýzu rozptylu, několik regresních metod, neparametrické testy atd.

Ovládání statistických programů je v současné době možné většinou přes menu a ikony podobně jako u ostatních programových produktů pracujících pod Windows, dříve převažovalo ovládání pomocí příkazového jazyka, které bylo poněkud náročnější pro nepravidelného uživatele nebo začátečníka.

Vzhledem k tomu, že Ostravská universita je vybavena statistickým paketem NCSS, zaměříme se na práci s tímto produktem.

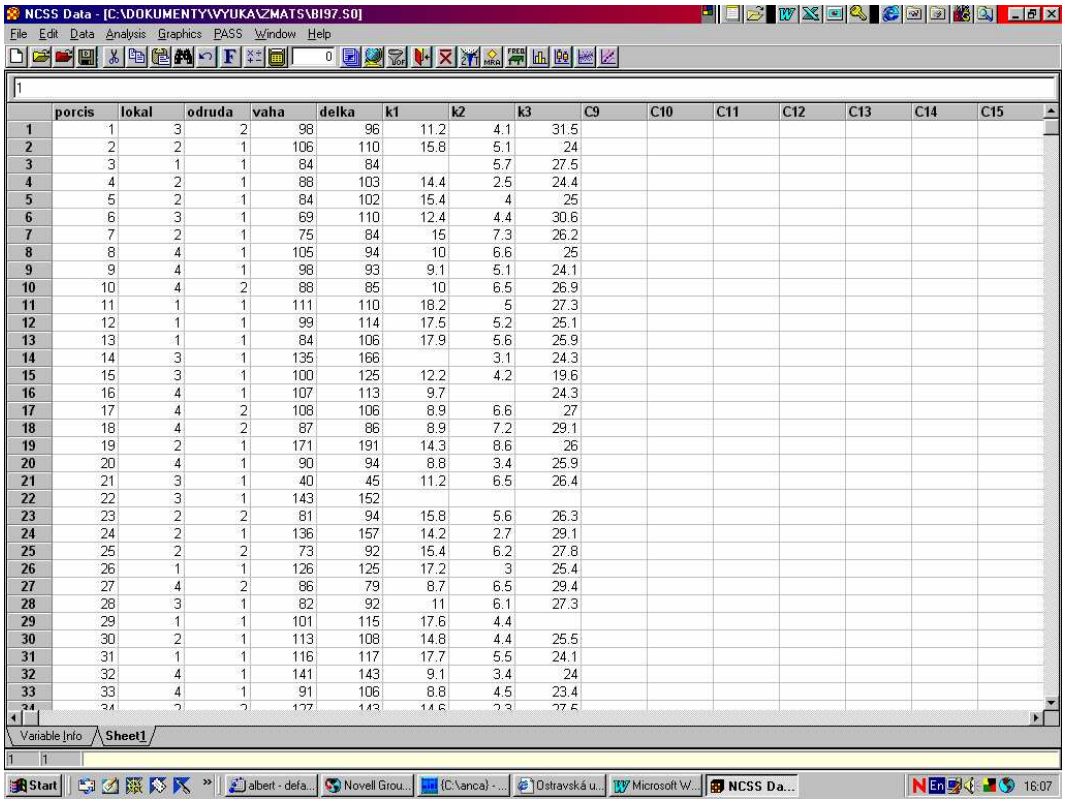
6.3 Programový paket NCSS

NCSS je universální statistický paket, doporučovaný zejména uživatelům-nestatistikům. Pokrývá však naprostou většinu požadavků i velmi sofistikované statistické analýzy dat. Ovládá se pomocí výběru z menu. NCSS komunikuje stylem „nabízím, co pravděpodobně můžete nebo máte v dané situaci požadovat, pokud vám to nevyhovuje, musíte to vyjádřit“. Výsledky (textový i grafický výstup společně) jsou ve formátu RTF (Rich Text Format), a tedy snadno importovatelné do běžných textových procesorů.

Základy ovládání NCSS ilustrují následující obrázky. Výběrem z menu přepínáme mezi pracovními okny se zpracovávanými daty, oknem tzv. šablon (templates), ve kterém specifikujeme vstupní parametry zvolené analytické procedury, oknem aktuálních výsledků a oknem tzv. LOG souboru s výsledky pro trvalé uložení po ukončení sezení. Hlavní způsob ovládání je výběr z menu a vyplňování formulářů pomocí myši, v mnohém podobné práci s tabulkovými procesory. Vyplněné šablony lze uložit pro opakované použití. Do LOG souboru se ukládají pouze ty

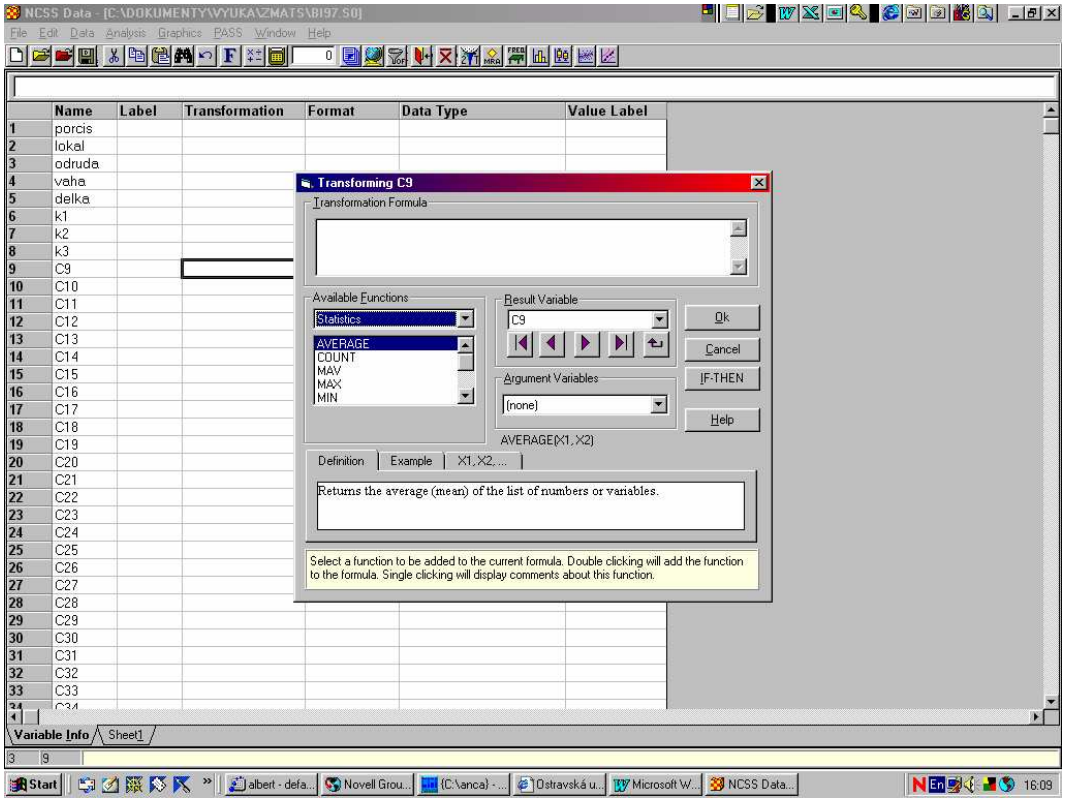
výsledky, které uživatel uloží explicitně, jinak jsou ztraceny a okno aktuálních výsledků je přepisováno následující spuštěnou procedurou. Zadávání transformací veličin a sdružování kategorií je jednoduché, spuštění výpočtu jen pro podmnožinu případů je možné, ale poměrně komplikované, je potřeba definovat logickou podmínku vybíráni podmnožiny pomocí funkce FILTER a při všech výpočtech tento filtr pak aktivovat ve vstupních parametrech výpočtu. Pokud úloha vyžaduje komplikovanější předzpracování dat, je většinou výhodné toto předzpracování udělat jiným programovým prostředkem např. Excelem, pokud data nejsou příliš rozsáhlá, a data pak do NCSS importovat. Import a export mnoha běžných formátů dat je součástí NCSS.

Tabulka s datovou maticí se liší od Excelu v tom, že názvy veličin jsou v názvech sloupců a na veličiny např. při zadávání vstupních parametrů výpočtu do šablony se odkazujeme pomocí jejich jmen.

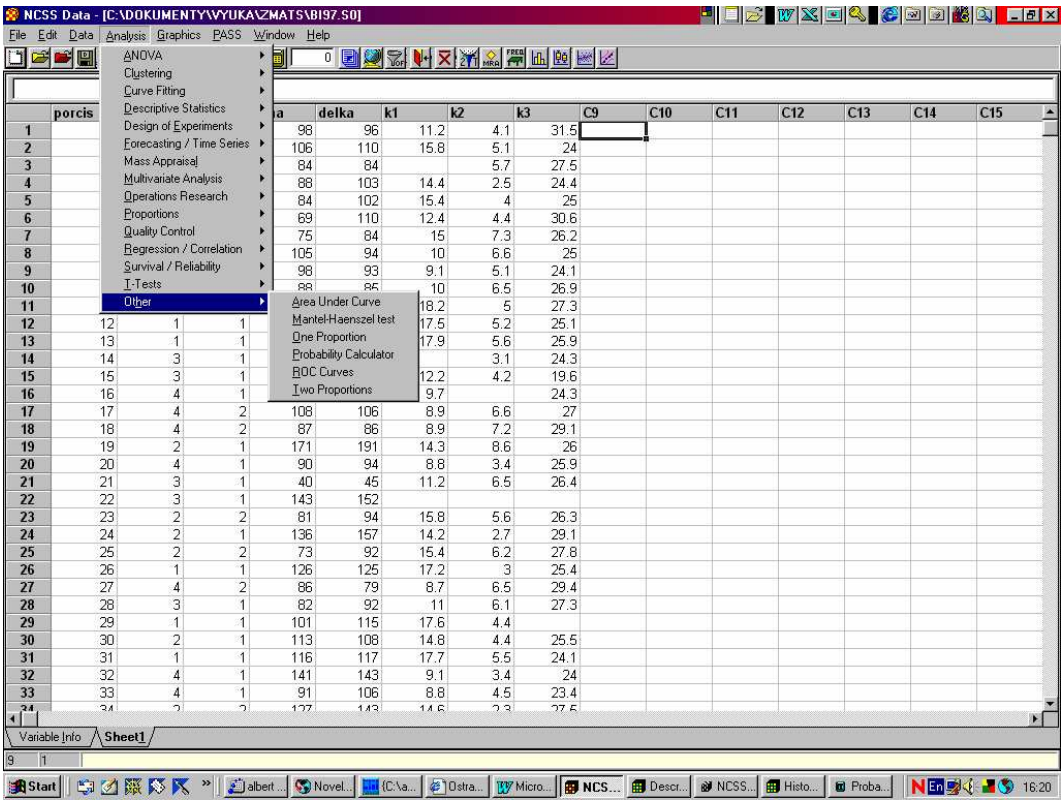


	porcis	lokal	odruda	vaha	delka	k1	k2	k3	C9	C10	C11	C12	C13	C14	C15
1	1	3	2	98	96	11.2	4.1	31.5							
2	2	2	1	106	110	15.8	5.1	24							
3	3	1	1	84	84		5.7	27.5							
4	4	2	1	88	103	14.4	2.5	24.4							
5	5	2	1	84	102	15.4	4	25							
6	6	3	1	69	110	12.4	4.4	30.6							
7	7	2	1	75	84	15	7.3	26.2							
8	8	4	1	105	94	10	6.6	25							
9	9	4	1	98	93	9.1	5.1	24.1							
10	10	4	2	88	85	10	6.5	26.9							
11	11	1	1	111	110	18.2	5	27.3							
12	12	1	1	99	114	17.5	5.2	25.1							
13	13	1	1	84	106	17.9	5.6	25.9							
14	14	3	1	135	166		3.1	24.3							
15	15	3	1	100	125	12.2	4.2	19.6							
16	16	4	1	107	113	9.7		24.3							
17	17	4	2	108	106	8.9	6.6	27							
18	18	4	2	87	86	8.9	7.2	29.1							
19	19	2	1	171	191	14.3	8.6	26							
20	20	4	1	90	94	8.8	3.4	25.9							
21	21	3	1	40	45	11.2	6.5	26.4							
22	22	3	1	143	152										
23	23	2	2	81	94	15.8	5.6	26.3							
24	24	2	1	136	157	14.2	2.7	29.1							
25	25	2	2	73	92	15.4	6.2	27.8							
26	26	1	1	126	125	17.2	3	25.4							
27	27	4	2	86	79	8.7	6.5	29.4							
28	28	3	1	82	92	11	6.1	27.3							
29	29	1	1	101	115	17.6	4.4								
30	30	2	1	113	108	14.8	4.4	25.5							
31	31	1	1	116	117	17.7	5.5	24.1							
32	32	4	1	141	143	9.1	3.4	24							
33	33	4	1	91	106	8.8	4.5	23.4							

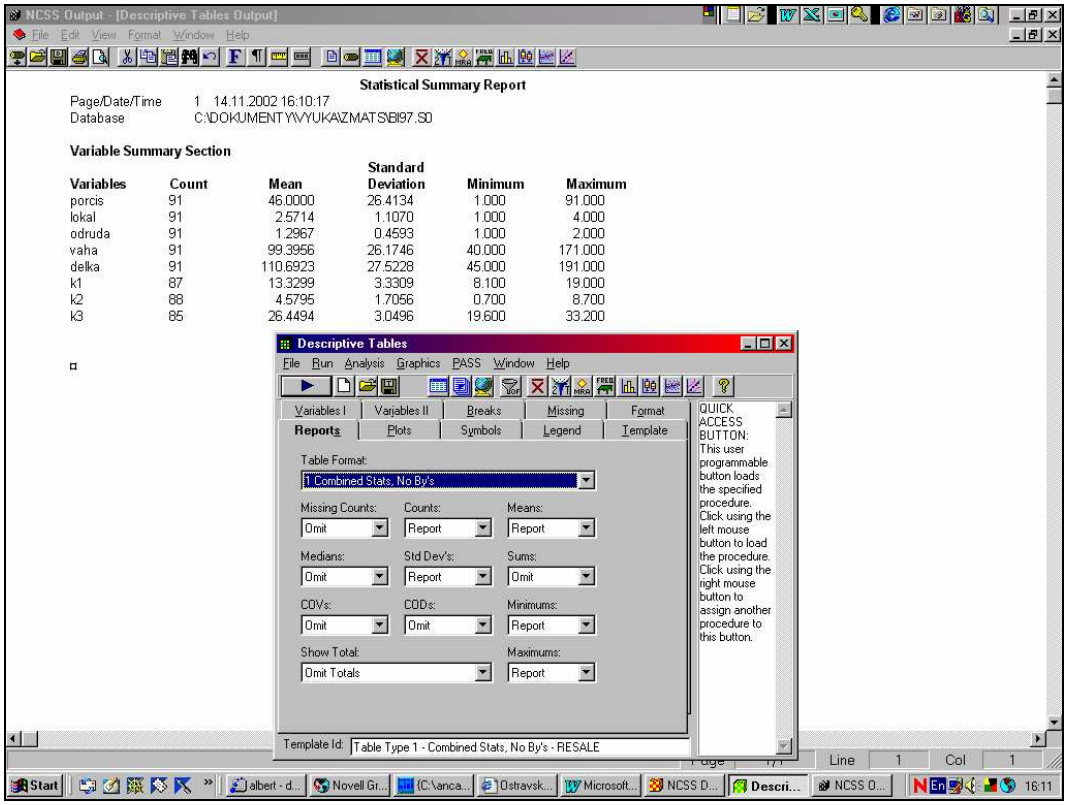
Kromě datové matice máme k dispozici i list s názvy veličin, ve kterém můžeme názvy veličin upravovat a také zadávat aritmetické výrazy pro výpočet odvozených veličin (transformace). Šablonu pro zadávání transformací otevřeme z položky Data v hlavním menu, odkud lze otevřít i šablonu pro nastavení a aktivaci filtru:



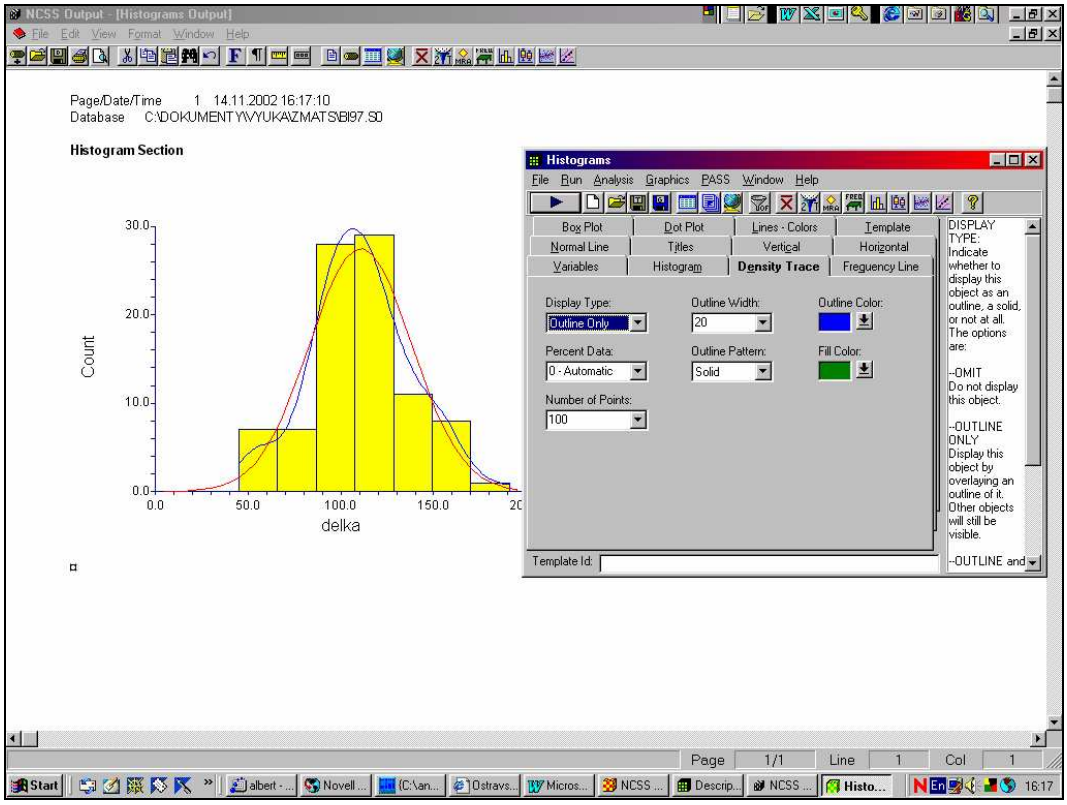
Požadované výpočty se zadávají volbou z menu, např. zde z položky Analysis hlavního menu rozbalíme skupiny implementovaných statistických metod:



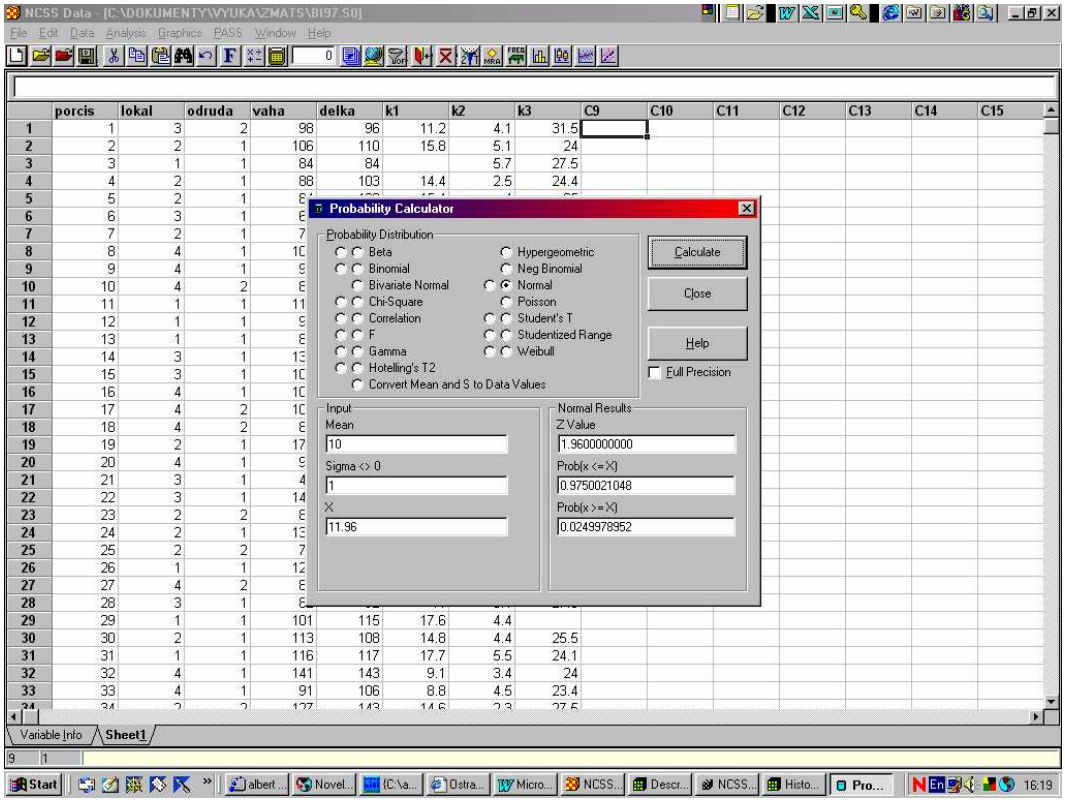
Vyplněním šablony se vstupními parametry výpočtu je možné specifikovat i úroveň podrobnosti a formát výstupu. Výstup je pak ve formátu RTF v okně aktuálního výstupu:



Podobně volbou Graphics v hlavním menu otevřeme nabídku grafických procedur. U všech těchto procedur je možné specifikovat obsah i vzhled grafických výstupů:



Součástí funkcí NCSS je i tzv. pravděpodobnostní kalkulátor, který nahrazuje obsáhlé statistické tabulky:



Výhodou NCSS je snadné ovládání pomocí menu, pohodlná práce s méně rozsáhlými daty, vysoká grafická kvalita výstupů i jejich snadný import do textových procesorů. K dispozici je i podrobná nápověda ve formě kompletního manuálu v angličtině. Pomocí NCSS byly zpracovány některé výsledky a grafy v těchto skriptech.

Přestože NCSS je kvalitní nástroj pro statistickou analýzu dat a dovolí vám velmi rychlou a efektivní práci, ale není, ostatně jako žádný jiný statistický program, pojistkou proti chybám v aplikacích statistiky.

Při užívání statistických programových prostředků věnujte pozornost i převodům zpracovávaných dat mezi různými programovými prostředky. Častým zdrojem obtíží při tomto převodu (bývá označován také jako import a export dat) mohou být zejména chybějící hodnoty v datech, které nemusí být předvedeny správně. Pokud data obsahují desetinná čísla, můžou vniknout potíže při neshodách oddělovače desetinných míst (čárka nebo tečka). Proto při operacích exportu a importu dat byste vždy měli zkontrolovat první a poslední řádek datové matice a základní popisné charakteristiky převáděného souboru, abyste tak s vysokou pravděpodobností mohli vyloučit nechtěnou změnu v datech způsobenou nesprávným převodem. Ze špatných dat nelze získat dobré výsledky.

Statistická analýza dat i s dobrým programovým vybavením je v naprosté většině případů duševně náročná činnost vyžadující soustředění a obezřetnost. Dovednost ovládání statistického software představuje jen menší část požadavků kladených na řešitele úlohy.

Kontrolní otázky:

1. *Jaká je obvyklá struktura dat zpracovávaná statistickými programy?*
2. *Co je to import dat a jaká jsou jeho úskalí?*
3. *Jaké jsou výhody a nevýhody Excelu ve srovnání se specializovanými statistickými pakety?*
4. *Na datech ze souboru BI97 si vyzkoušejte základní statistické funkce a doplněk Analýza dat.*

Pojmy k zapamatování:

- *statistická data, jejich struktura,*
- *obvyklé funkce ve statistických paketech,*
- *import a export dat,*
- *statistické funkce v Excelu a jejich nedostatky,*
- *doplněk Excelu Analýza dat.*

7 Prezentace výsledků analýzy dat

V této kapitole bude ukázány některá doporučení, jak prezentovat výsledky statistické analýzy. Část těchto doporučení vychází z knihy van Belle (2002). Části příkladů převzaté odtamtud jsou ponechány v angličtině. Následující příklad tří způsobů prezentace téhož jednoduchého výsledku ukazuje, že na formě prezentace výsledků záleží:

- The blood type in the population of the United States is approximately 40 %, 11 %, 4 % and 45 % for A, B, AB, and O, respectively.

- The blood type in the population of the United States is approximately 40% A, 11% B, 4% AB and 45% O.

- The blood type in the population of the United States is approximately,

O	45%
A	40%
B	11%
AB	4%.

Rozdíly ve snadnosti či obtížnosti vnímání tohoto jednoduchého výsledku nepotřebují žádné další vysvětlování a snad jsou dostatečným argumentem pro to, že na způsobu prezentace výsledků záleží a že bychom se nad tím měli důkladně zamýšlet.

7.1 Prezentace tabulek a užití vhodných grafů

Některé chyby ukazuje tabulka 1, ve které jsou uvedeny počty pracovníků v různých zdravotnických profesích v USA roku 1988, názvy kategorií jsou ponechány v angličtině. Tabulka je nedokonalá nejméně ve dvou ohledech:

- Číselné údaje jsou téměř jistě zatíženy různou nepřesností. Zatímco u lékařů, sester, dentistů a optiků to jsou hodnoty získané z příslušných registrů, u některých jiných kategorií jako řečových, fyzických a pracovních terapeutů nebo pedikérů (podiatrists) jde jen o odhad v tisících. Hodnoty v tabulce však vyvolávají dojem, že všechna čísla jsou přesná,
- van Belle jako chybu uvádí i to, že řádky tabulky jsou seřazeny podle abecedního pořadí názvů profesí, ne podle číselných hodnot. Možná se nám tato výhrada zdá neoprávněná, jsme asi zkaženi návyky jak z místních publikací, tak i většinou statistického softwaru, kde je četnostní tabulka seřazena podle názvů kategorií nebo jejich číselných kódů. Ale argument, že pořadí řádků by nemělo záviset na tom, v jakém jazyku publikujeme, nelze jen tak vyvrátit.

Tabulka 1**:** Počet aktivních zdravotníků v USA v roce 1980 (ze zprávy National Center for Health Statistics, 2000)

Occupation	**1980**
Chiropractors	25 600
Dentists	121 240
Nutritionists/Dieticians	32 000
Nurses, registered	1 272 900
Occupational Therapists	25 000
Optometrists	22 330
Pharmacists	142 780
Physical Therapists	50 000
Physicians	427 122
Podiatrists	7 000
Speech Therapists	50 000

Podle van Belleho by tabulka měla mít formu uvedenou v tabulce. 2, tj. číselné údaje zaokrouhlené na tisíce a řádky seřazeny sestupně podle číselných hodnot.

Tabulka 2**:** Údaje z tabulky 1 seřazené podle počtu, zaokrouhleno na tisíce.

Occupation in 1000's	**1980**
Nurses, registered	1273
Physicians	427
Pharmacists	143
Dentists	121
Physical Therapists	50
Speech Therapists	50
Nutritionists/Dieticians	32
Chiropractors	26
Occupational Therapists	25
Optometrists	22
Podiatrists	7

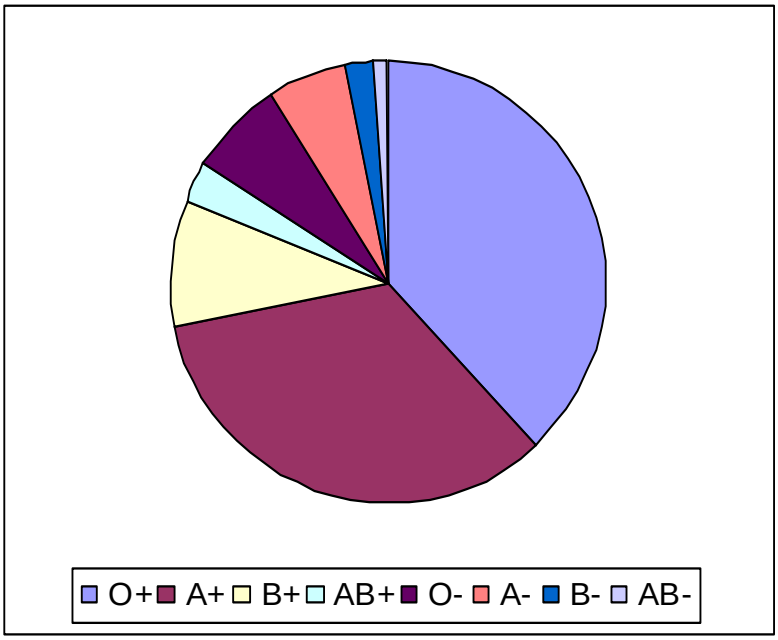
Dále se doporučuje užívat rozumný počet významných číslic. Pokud číselná hodnota je větší než 100, většinou stačí ji uvést jako celé číslo, tj. bez desetinných míst. Hodnoty ve sloupci mají být vhodně zarovnány, celá čísle vpravo, desetinná na desetinnou čárku (nebo tečku). Zejména v tabulkách je nutné brát ohled na tzv. „efektivní číslice“. To jsou ty číslice, jejichž hodnoty nejsou konstantní, ale mění se. Např. šestimístná čísla 354 691, 357 234, 356 991 mají jen čtyři efektivní číslice. Pokud bychom chtěli je prezentovat přijatelněji, pak bychom měli odečíst od těchto hodnot 350000 a uvádět tento výsledný rozdíl. V tabulkách ovšem mají být pokud možno nejvýše dvě až tři efektivní číslice, neboť více efektivních číslic člověk obtížně vnímá.

Všeobecně hlásaná zásada, že grafy místo číselných údajů jsou lepší, není vždy správná. Někdy je tabulka vhodnější než graf, zejména když zvolený typ grafu neodpovídá struktuře dat a tabulka ano. Jedním z doporučení je ***neužívat výsečové grafy.*** Van Belle uvádí citát: *„Jediná věc je horší než výsečový graf – několik nebo dokonce mnoho výsečových grafů“*.

Výsečové (koláčové) grafy ignorují strukturu dat, čtenář si musí propojovat legendu s výsečemi. Další van Bellův argument proti výsečovým grafům působí na první pohled úsměvně – při tisku výsečových grafů se spotřebuje moc inkoustu. Ale pokud se nad tím zamyslíme, je oprávněný. Porovnáme-li spotřebu inkoustu na bodový graf závislosti hodnot dvou veličin, kdy při malé spotřebě inkoustu získáme náhled na tuto závislost se spotřebou na výsečové grafy, kdy při velké spotřebě nezískáme nic (viz příklad, obr. 1), pak závažnost argumentu musíme uznat.

Tabulka 3: Relativní četnosti (v %) krevních skupin a Rh faktoru v populaci USA

Blood Type	Rh+	Rh-	Total
O	38	7	45
A	34	6	40
B	9	2	11
AB	3	1	4
Total	84	16	100



Obrázek 1: Relativní četnosti (v %) krevních skupin a Rh faktoru v populaci USA

Z výsečového grafu na obr. 1 se opravdu mnoho nedozvíme, struktura grafu neodpovídá struktuře dat, propojování legendy a výsečí je zbytečně namáhavé a

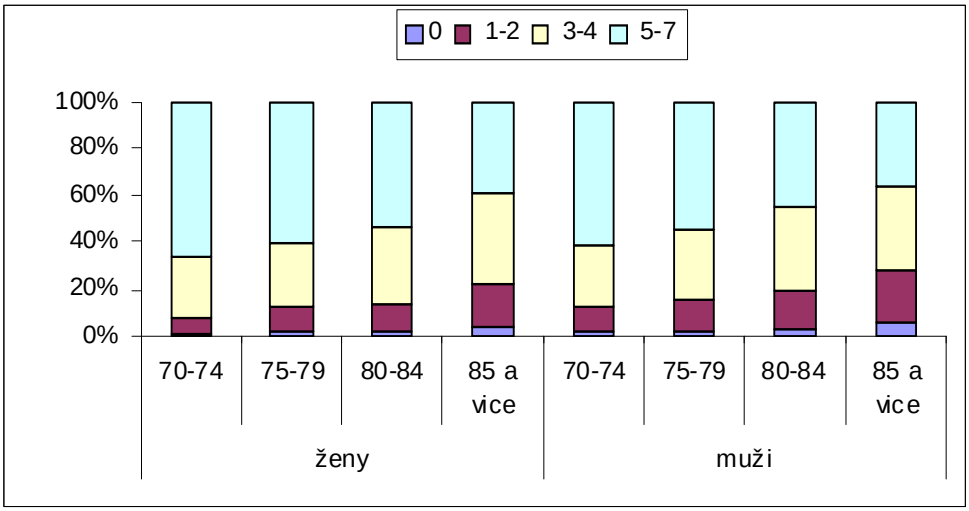
spotřeba inkoustu velká. Tabulka 3 prezentuje stejný výsledek daleko přehledněji a srozumitelněji.

Další van Belleho doporučení je ***neužívat spojované sloupcové grafy***. Spojované (kumulované, stackbar) sloupcové grafy jsou hůře čitelné než jednoduché sloupcové grafy a často lze najít efektivnější možnost, jak nahlédnout do struktury dat. To ilustrujeme na následujícím příkladu.

Souhrnná zdrojová data z průzkumu počtu aktivit provozovaných seniory v průběhu dvou týdnů jsou uvedena v tabulce 5. Ve zprávě Státního centra pro zdravotní statistiku byly tyto údaje prezentovány formou skládaného sloupcového grafu (obr. 2), což ke vnímání jejich obsahu nijak nepřispělo, spíše naopak. Prezentace by měla usnadňovat odpovědi na následující jednoduché a přirozené otázky: Mají více aktivit muži nebo ženy? Jak mění počet aktivit s věkem? Liší se tyto změny u mužů a žen? To ovšem spojovaný sloupcový graf na obrázku 2 rozhodně neusnadňuje.

Tabulka 5: Počet aktivit seniorů v průběhu dvou týdnů - četnosti v %

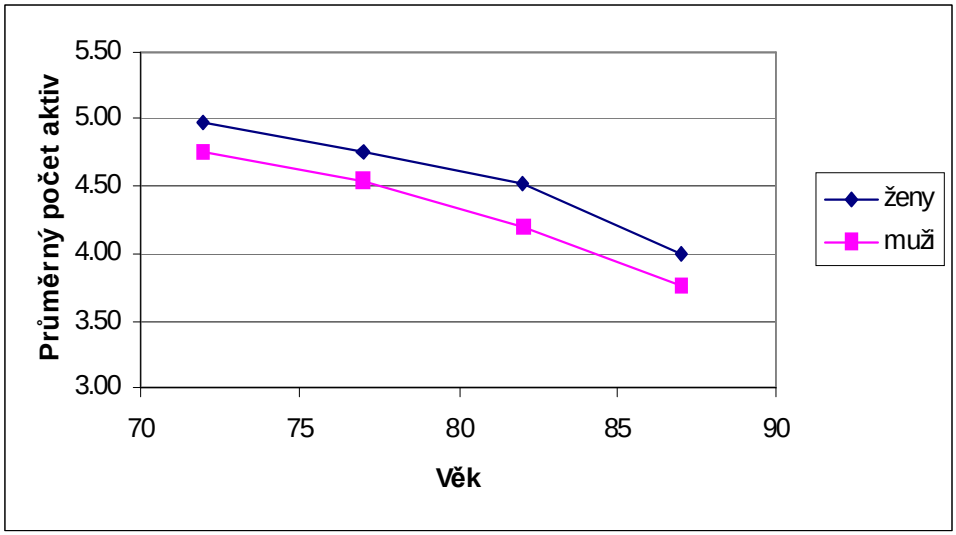
	Počet aktivit	70-74	75-79	80-84	85 a více
Ženy	0	1	1.3	2.1	3.1
	1-2	6.8	10.5	11.9	19.2
	3-4	26.8	27.5	32.5	38.3
	5-7	65.4	60.7	53.5	39.4
Muži	0	1.9	1.7	2.9	5.3
	1-2	10.5	13.3	15.9	23
	3-4	26.3	30.3	36.7	35.9
	5-7	61.2	54.7	44.5	35.9



Obrázek 2: Počet aktivit v průběhu dvou týdnů - četnosti v % (Kramarov et al., zpráva National Center for Health Statistics, 1999).

Přitom docela jednouchý přepočet a grafické zobrazení průměrných hodnot aktivit pro muže a ženy podle věkových kategorií na obrázku 3 vypovídá jasně, že ženy

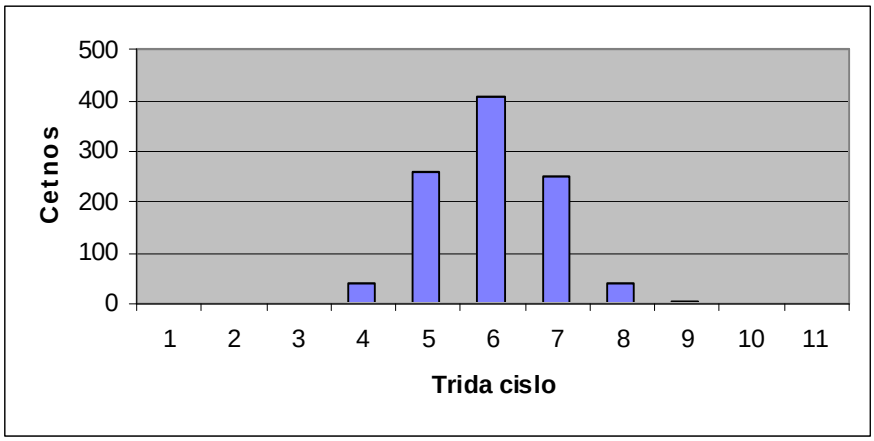
jsou o trochu aktivnější, počet aktivit s věkem klesá a rychlost tohoto poklesu je u obou pohlaví zhruba stejná.



Obrázek 3: Průměrný počet aktivit podle věku a pohlaví

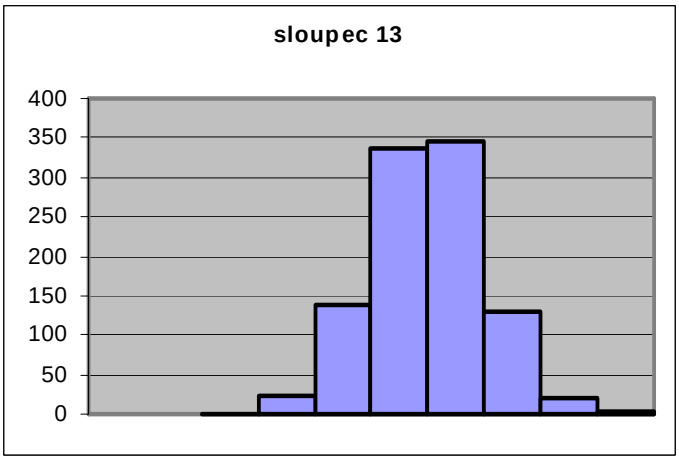
7. 2 Některé chyby prezentace ve studentských pracích

V tomto odstavci jsou komentovány chyby z korespondenčních úloh a semestrálních prací studentů v předmětu Analýza dat. Komentáře k chybám jsou psány kurzivou.



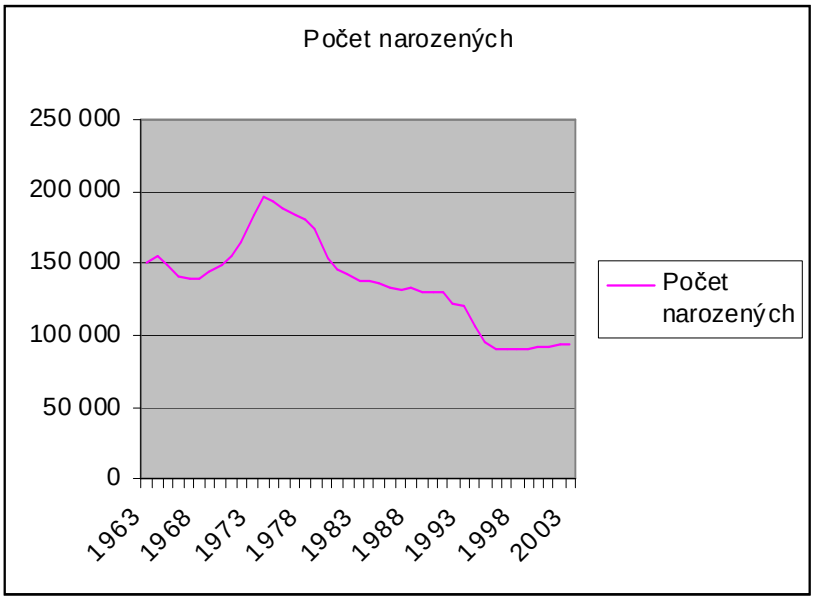
Obrázek 4**:** Histogram – častá chyba z naprosté nedbalosti

Histogram na obr. 4 je prezentován tak, jak ho nabízí Excel, zdravý rozum si vybral dovolenou, ohled na čtenáře žádný. Ponechány mezery mezi sloupci, nevhodně zvolené měřítko vodorovné osy (pět tříd s nulovou četností), nic nevypovídající popis vodorovné osy.



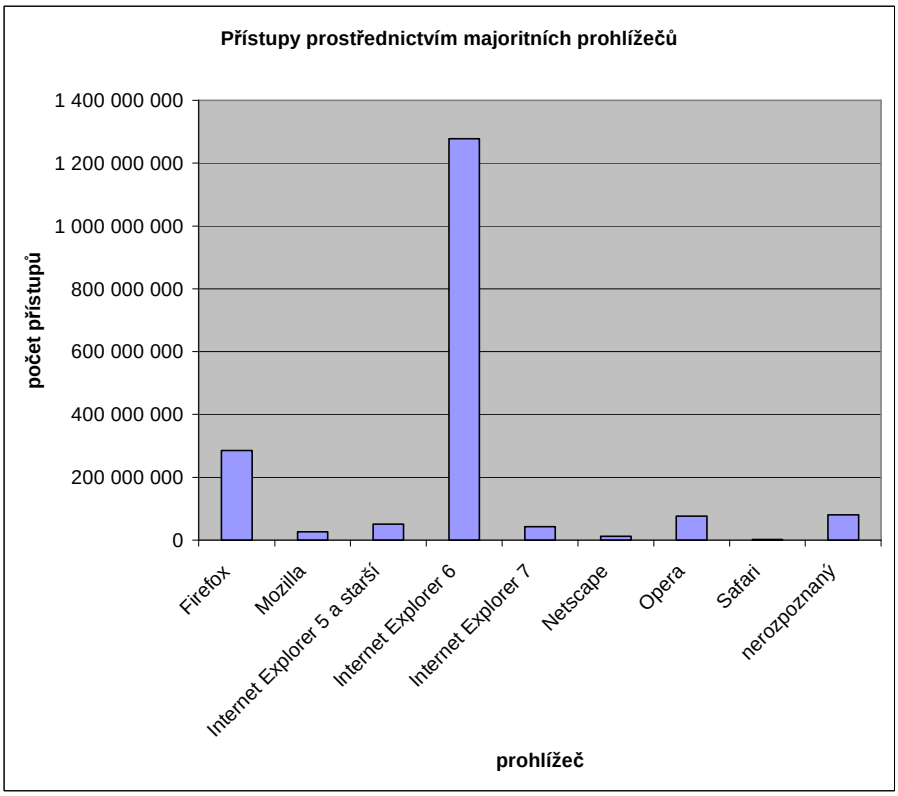
Obrázek 5: Histogram – další častá chyba způsobená nedbalostí

V histogramu na obr. 5 chybí popis os, zbytečný je nic neříkající nadpis histogramu ‚opět nevhodně zvolené měřítko vodorovné osy.



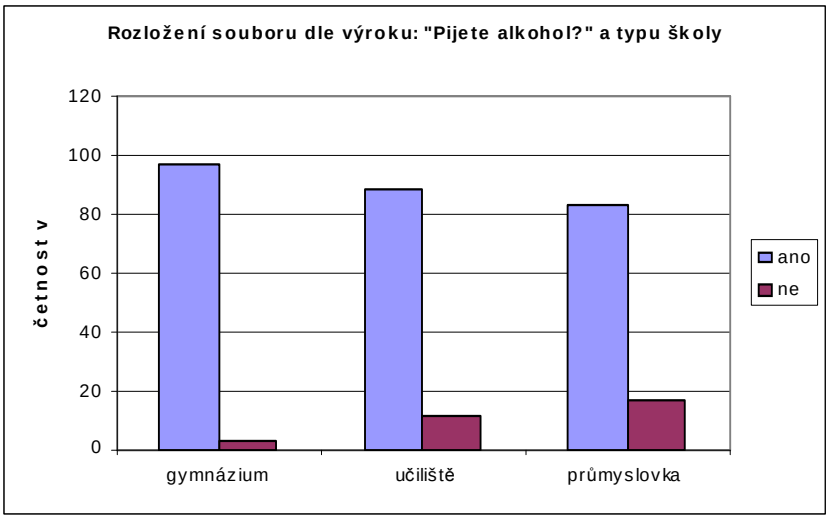
Obrázek 6: Časový průběh počtu narozených

Na obr. 6 chybí popis os grafu, nevhodné jednotky na svislé ose (tři neefektivní nuly, počet narozených měl být v tisících), legenda je nadbytečná a zbytečně zabírá značnou část kreslící plochy, význam čáry nejasný (bylo užito nějaké vyhlazování?), časová řada by měla být nakreslena jako body, případně se spojnicemi.



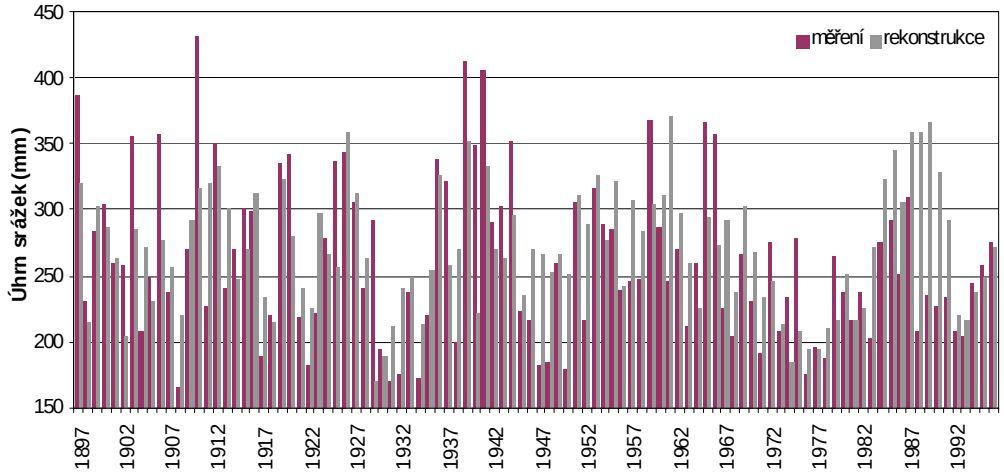
Obrázek 7: Nevhodný sloupcový graf

Na obr. 7 jsou užity nevhodné jednotky na svislé ose sloupcového grafu (8 neefektivních číslic), vhodnější by bylo uvádět počet přístupů v milionech nebo lépe ve stovkách milionů. Zobrazení devíti značně odlišných četností formou sloupcového grafu není nejvhodnější způsob prezentace tohoto výsledku, tabulka by vypovídala o struktuře a obsahu dat lépe.



Obrázek 8: Další nesprávný sloupcový graf

Na první pohled (pomineme-li neobratnou formulaci nadpisu) sloupcový graf na obr. 8 vypadá uspokojivě. Ale jaký je význam druhých sloupečků? Jsou to doplňky do 100%, takže jsou nadbytečné stejně jako legenda. Tři zjištěné relativní četnosti stačilo uvést jako tabulku, zabralo by to méně místa a vypovídalo jasně.



Obrázek 9: Nevhodně užitý typ grafu

Na obr. 9 je nevhodně zvolený typ grafu pro zobrazení dvou časových řad do jednoho obrázku, takže výsledek je nepoužitelný pro naprostou nečitelnost. Pro takové závislosti jsou vhodné bodové grafy, případně se spojnicemi bodů.

A ještě chyby v prezentaci číselných údajů:

H0: μ = 6
průměr x = 5,959409417
s = 0,99046792
hodnota testového kritéria: -1,29593994
Typická ukázka nesprávného a nepřehledného prezentování číselných výsledků s nadbytečným počtem platných číslic.

b1 = 0,90711042
b0 = 17,0189542
Se = Σ(Yi - b0 - b1x1)2 = 423,839904
s2 = Se / (n-2) = 26,489994
Podobné chyby jako v předchozí ukázce, tady navíc i neobratný a nepřesný zápis symbolů a vzorců.

Uvedené příklady chyb snad přispějí k tomu, že se v prezentacích podobné chyby nebudou opakovat. Van Belle požaduje, aby se v prezentaci výsledků statistických analýz věda spojovala s uměním. Možná je to požadavek příliš náročný, ale rozhodně bychom měli dbát alespoň na dobrou řemeslnou úroveň, využívat základní prezentační dovednosti, při prezentaci výsledků statistických analýz užívat zdravý rozum, přihlížet k možnostem vnímání čtenáře, mít ke čtenáři respekt a snažit se o co největší přehlednost a srozumitelnost výsledků.

Literatura - komentovaný seznam

Seznam je zlomkem rozsáhlé statistické literatury týkající se tohoto tématu. Zařazeny jsou především knihy a skripta českých autorů nebo české překlady z posledního období. Při výběru byl brán zřetel na dostupnost pro studenty Ostravské university a také na přístupnost textu začátečníkům ve statistice.

Anděl, J.: *Matematická statistika*, SNTL Praha, 1978
Nyní již klasická učebnice matematické statistiky. Úplné sledování vyžaduje hlubší znalosti matematické analýzy a lineární algebry, ale kniha obsahuje řadu příkladů, které jsou srozumitelné i bez těchto matematických znalostí a pomohou čtenáři orientovat se v aplikaci statistických metod.

Anděl, J.: Statistické metody, Matfyzpress Praha, 1993
Příručka pokrývající širokou paletu běžně užívaných metod statistické analýzy dat. Vysvětluje přístupným způsobem jejich matematicko-statistické základy. Velká pozornost je věnována i neparametrickým metodám.

Cyhelský, L., Kahounová, J. , Hindls, R.: Elementární statistická analýza, Management Press, Praha, 1996
Kniha přístupným způsobem vysvětluje základy deskriptivní statistiky a počtu pravděpodobnosti nutné pro aplikace statistiky. Zabývá se základy teorie odhadu a testování hypotéz. Neobsahuje analýzu rozptylu a regresi. Knihu je možno doporučit čtenáři se středoškolskými znalostmi matematiky jako první učebnici pro seznámení s problémy statistické analýzy dat. Dostupná v knihovně OU.

Havránek, T.: Statistika pro biologické a lékařské vědy, Academia, 1993
Kniha vynikajícího, bohužel předčasně zesnulého českého statistika, která vyšla až dva roky po jeho smrti. Kniha poměrně přístupným způsobem vykládá i obtížné partie statistické analýzy dat. Aplikace matematicko statistických metod je ilustrována na řadě netriviálních příkladů z autorovy praxe v analýze biomedicínských dat.

Hebák, P., Hustopecký, J.: Průvodce moderními statistickými metodami, SNTL Praha, 1990
Na více než třiceti příkladech inspirovaných praktickými úlohami je důkladně ilustrována aplikace různých metod induktivní statistiky, včetně formulace úlohy, zdůvodnění různých alternativ řešení a interpretace výsledků

Komenda, S.: Biometrie, skriptum PřF UP Olomouc, 1994
Autor do učebního textu promítá dlouholetou zkušenost z oblasti aplikací statistiky v biomedicínském výzkumu. Přístupnou formou jsou vysvětleny základy pravděpodobnosti, statistiky i mnohé metodologické otázky. Čtenářskou zajímavost textu zvyšuje řada původních aforismů. Vhodný úvodní text pro čtenáře nejen z okruhu biologů. Skriptum je dostupné ve více výtiscích v knihovně OU.

Křivý, I. : Základy matematické statistiky, skriptum PF Ostrava, 1985
Učební text pro studenty učitelství matematiky. Pokývá základní aplikační oblasti matematické statistiky. K úplnému sledování je potřeba vyšší než středoškolská úroveň matematiky. Skriptum je dostupné ve více výtiscích v knihovně OU.

Laga, J., Likeš, J.: Základní statistické tabulky, SNTL, 1978
Obsáhlé „klasické“ statistické tabulky českých autorů, obsahují i důkladné vysvětlení pojmů důležitých pro správné užití tabulek v aplikacích metod matematické statistiky.

Lepš, J.: Biostatistika, skriptum, Jihočeská universita, Čes. Budějovice, 1996
Netradičně napsaný učební text (autor je biolog), ve kterém je čtenář na příkladech veden od základních pojmů až ke shlukové analýze a dalším mnohorozměrným metodám analýzy dat.

Likeš, J., Machek, J.: Matematická statistika, SNTL, Praha, 1983
Učebnice statistiky pro vysoké školy technické, ale pokrývá i metody užívané v netechnických oborech. Předpokládá znalost základů matematické analýzy v rozsahu vyučovaném na technických školách.

Meloun, M., Militký, J.: Statistické zpracování experimentálních dat, PLUS, 1994
Rozsáhlá kniha aplikačně orientovaná, zejména na metody regresní analýzy. Je užitečná především pro chemické a technické obory, ale poslouží i pro jiné aplikace, zvláště s využitím statistického software.

NCSS 6.0 Statistical System for Windows – User ‘s Guide, NCSS Kaysville, 1995
Obsáhlý manuál k systému NCSS. Popisuje nejen ovládání programového systému, ale také základy implementovaných metod a doporučení pro interpretaci výsledků. K dispozici je on-line jako součást instalace NCSS.

Sprent, P., Smeeton, N.,C.: Applied Nonparametric Statistical Methods, Third Edition, Chapman & Hall/CRC, 2001
Obsáhlá monografie zaměřená i na výpočetní aspekty neparametrických metod a využití moderních algoritmů pro výpočet přesné pravděpodobnosti. Aplikace jsou ukázány na řadě příkladů.

Tvrdík J.: Základy statistické analýzy dat, Přírodovědecká fakulta Ostravské university, Ostrava 1998
Přístupně napsaný učební text zaměřený na pochopení důležitých pojmů nutných pro aplikaci statistických metod. Některé jeho části jsou v upravené formě převzaty i do opor k předmětům Základy matematické statistiky a Analýza dat.

Tvrdík J.: Základy matematické statistiky, 2. upravené vydání, Přírodovědecká fakulta Ostravské university, Ostrava, 2008
Opora ke stejnojmennému kursu, který předchází kursu Analýza dat.

van Belle G.: Statistical Rules of Thumb, John Wiley & Sons, 2002
Kniha autora s bohatou zkušeností z výuky i aplikací statistiky poskytuje řadu užitečných doporučení pro aplikace statistiky. Prezentací výsledků se zabývá v obsáhlé kapitole „Words, Tables, and Graphs“.

Wonnacot, T.H., Wonnacot, R.J.: Statistika pro obchod a hospodářství, Victoria Publishing, Praha, 1993
Rozsáhlá učebnice základů statistiky. Pokrývá mnoho statistických metod včetně těch, které se užívají v analýze ekonomických dat (časové řady atd.). Výklad je veden velmi přístupnou formou, problematika je ilustrována mnoha příklady.

Zvára, K.: Biostatistika, Karolinum, Praha, 1998
Velmi zdařilá učebnice statistiky, určená především studentům biologie. Je napsána přístupnou formou, důraz je kladen na aplikaci statistických metod, která je ilustrována řadou řešených příkladů z biologického výzkumu.

Zvára K., Štěpán J.,: Pravděpodobnost a matematická statistika, Matfyzpress, Praha, 2001
Vynikající učebnice původně napsaná pro studenty matematiky na pedagogických fakultách. Vhodná doplňující literatura, prohlubující znalosti matematické statistiky.

Interaktivní učebnice pro základní kurs statistiky:

Härdle W. et al., MM*Stat - Základy statistiky, http://www.quantlet.com/mdstat/scripts/mmcze/java/start.html, 2005

Řezanková, H., Marek, L., Vrabec, M., Kalenský, L., Řezanka, P., IASTAT - Interaktivní učebnice statistiky, http://badame.vse.cz/iastat/, 2005

Dear, K. et al., Surf-Stat, http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html, 2005

Statistické tabulky

Statistické tabulky byly pořízeny s využitím statistických funkcí NORMSDIST, CHIINV, TINV, FINV programu Microsoft Excel pro Windows 95, verse 7.0. Pokud jste u počítače, na kterém je nainstalován Excel nebo některý ze statistických programů (NCSS atd.) statistické tabulky nepotřebujete, neboť potřebné hodnoty distribučních funkcí či kvantilů snadno zjistíte pomocí těchto programových prostředků.

Distribuční funkce normovaného normálního rozdělení

$$X \sim N(0,1), \quad \Phi(x) = P(X < x)$$

$\Phi(x)$

x	*+0*	*+0,02*	*+0,04*	*+0,06*	*+0,08*
0,0	0,5000	0,5080	0,5160	0,5239	0,5319
0,1	0,5398	0,5478	0,5557	0,5636	0,5714
0,2	0,5793	0,5871	0,5948	0,6026	0,6103
0,3	0,6179	0,6255	0,6331	0,6406	0,6480
0,4	0,6554	0,6628	0,6700	0,6772	0,6844
0,5	0,6915	0,6985	0,7054	0,7123	0,7190
0,6	0,7257	0,7324	0,7389	0,7454	0,7517
0,7	0,7580	0,7642	0,7704	0,7764	0,7823
0,8	0,7881	0,7939	0,7995	0,8051	0,8106
0,9	0,8159	0,8212	0,8264	0,8315	0,8365
1,0	0,8413	0,8461	0,8508	0,8554	0,8599
1,1	0,8643	0,8686	0,8729	0,8770	0,8810
1,2	0,8849	0,8888	0,8925	0,8962	0,8997
1,3	0,9032	0,9066	0,9099	0,9131	0,9162
1,4	0,9192	0,9222	0,9251	0,9279	0,9306
1,5	0,9332	0,9357	0,9382	0,9406	0,9429
1,6	0,9452	0,9474	0,9495	0,9515	0,9535
1,7	0,9554	0,9573	0,9591	0,9608	0,9625
1,8	0,9641	0,9656	0,9671	0,9686	0,9699
1,9	0,9713	0,9726	0,9738	0,9750	0,9761
2,0	0,9772	0,9783	0,9793	0,9803	0,9812
2,1	0,9821	0,9830	0,9838	0,9846	0,9854
2,2	0,9861	0,9868	0,9875	0,9881	0,9887
2,3	0,9893	0,9898	0,9904	0,9909	0,9913
2,4	0,9918	0,9922	0,9927	0,9931	0,9934
2,5	0,9938	0,9941	0,9945	0,9948	0,9951

Vybrané kvantily rozdělení Chí-kvadrát

$$X \sim \chi^2_n, \quad P\left[X < x(p)\right] = p$$

	x(p)			
n	*p=0,025*	*p=0,95*	*p=0,975*	*p=0,99*
1	0,00	3,84	5,02	6,63
2	0,05	5,99	7,38	9,21
3	0,22	7,81	9,35	11,34
4	0,48	9,49	11,14	13,28
5	0,83	11,07	12,83	15,09
6	1,24	12,59	14,45	16,81
7	1,69	14,07	16,01	18,48
8	2,18	15,51	17,53	20,09
9	2,70	16,92	19,02	21,67
10	3,25	18,31	20,48	23,21
11	3,82	19,68	21,92	24,73
12	4,40	21,03	23,34	26,22
13	5,01	22,36	24,74	27,69
14	5,63	23,68	26,12	29,14
15	6,26	25,00	27,49	30,58
16	6,91	26,30	28,85	32,00
17	7,56	27,59	30,19	33,41
18	8,23	28,87	31,53	34,81
19	8,91	30,14	32,85	36,19
20	9,59	31,41	34,17	37,57
25	13,12	37,65	40,65	44,31
30	16,79	43,77	46,98	50,89
40	24,43	55,76	59,34	63,69
50	32,36	67,50	71,42	76,15
100	74,22	124,34	129,56	135,81

Vybrané kvantily Studentova t-rozdělení

$$X \sim t_n, \quad P\left[X < x(p)\right] = p$$

	x(p)				
n	*p=0,9*	*p=0,95*	*p=0,975*	*p=0,99*	*p=0,995*
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,36	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,41	1,89	2,36	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,11
12	1,36	1,78	2,18	2,68	3,05
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,14	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,72	2,09	2,53	2,85
25	1,32	1,71	2,06	2,49	2,79
30	1,31	1,70	2,04	2,46	2,75
40	1,30	1,68	2,02	2,42	2,70
50	1,30	1,68	2,01	2,40	2,68
70	1,29	1,67	1,99	2,38	2,65
100	1,29	1,66	1,98	2,36	2,63
500	1,28	1,65	1,96	2,33	2,59

Vybrané kvantily Fisherova Snedecorova F-rozdělení

$$X \sim F_{m,n}, \quad P\left[X < x(0{,}95)\right] = 0{,}95$$

x(0,95)

	m							
n	1	2	3	4	5	10	20	40
1	161,45	199,50	215,71	224,58	230,16	241,88	248,02	251,14
2	18,51	19,00	19,16	19,25	19,30	19,40	19,45	19,47
3	10,13	9,55	9,28	9,12	9,01	8,79	8,66	8,59
4	7,71	6,94	6,59	6,39	6,26	5,96	5,80	5,72
5	6,61	5,79	5,41	5,19	5,05	4,74	4,56	4,46
6	5,99	5,14	4,76	4,53	4,39	4,06	3,87	3,77
7	5,59	4,74	4,35	4,12	3,97	3,64	3,44	3,34
8	5,32	4,46	4,07	3,84	3,69	3,35	3,15	3,04
9	5,12	4,26	3,86	3,63	3,48	3,14	2,94	2,83
10	4,96	4,10	3,71	3,48	3,33	2,98	2,77	2,66
11	4,84	3,98	3,59	3,36	3,20	2,85	2,65	2,53
12	4,75	3,89	3,49	3,26	3,11	2,75	2,54	2,43
13	4,67	3,81	3,41	3,18	3,03	2,67	2,46	2,34
14	4,60	3,74	3,34	3,11	2,96	2,60	2,39	2,27
15	4,54	3,68	3,29	3,06	2,90	2,54	2,33	2,20
20	4,35	3,49	3,10	2,87	2,71	2,35	2,12	1,99
30	4,17	3,32	2,92	2,69	2,53	2,16	1,93	1,79
40	4,08	3,23	2,84	2,61	2,45	2,08	1,84	1,69
60	4,00	3,15	2,76	2,53	2,37	1,99	1,75	1,59
120	3,92	3,07	2,68	2,45	2,29	1,91	1,66	1,50
500	3,86	3,01	2,62	2,39	2,23	1,85	1,59	1,42

Kritické hodnoty pro jednovýběrový Wilcoxonův test

Nulová hypotéza se zamítá, je-li hodnota statistiky $\min(S^+, S^-)$ menší nebo rovna kritické hodnotě.

kritické hodnoty

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
6	0	
7	2	
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68

Kritické hodnoty pro dvouvýběrový Wilcoxonův (Mannův-Whitneyův) test

Nulová hypotéza se zamítá na hladině významnosti $\alpha = 0{,}05$, je-li hodnota statistiky $\min(U^+, U^-)$ menší nebo rovna kritické hodnotě.

	n											
m	4	5	6	7	8	9	10	11	12	13	14	15
4	0											
5	1	2										
6	2	3	5									
7	3	5	6	8								
8	4	6	8	10	13							
9	4	7	10	12	15	17						
10	5	8	11	14	17	20	23					
11	6	9	13	16	19	23	26	30				
12	7	11	14	18	22	26	29	33	37			
13	8	12	16	20	24	28	33	37	41	45		
14	9	13	17	22	26	31	36	40	45	50	55	
15	10	14	19	24	29	34	39	44	49	54	59	64

Kritické hodnoty Spearmanova korelačního koeficientu

Nulová hypotéza se zamítá na hladině významnosti α, je-li hodnota statistiky r_S větší nebo rovna kritické hodnotě.

kritické hodnoty

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	0.9000	
6	0.8286	0.9429
7	0.7450	0.8929
8	0.6905	0.8571
9	0.6833	0.8167
10	0.6364	0.7818
11	0.6091	0.7545
12	0.5804	0.7273
13	0.5549	0.6978
14	0.5341	0.6747
15	0.5179	0.6536
16	0.5000	0.6324
17	0.4853	0.6152
18	0.4716	0.5975
19	0.4579	0.5825
20	0.4451	0.5684