UČEBNÍ TEXTY OSTRAVSKÉ UNIVERZITY

Přírodovědecká fakulta

ANALÝZA DAT

Petr Bujok

Ostravská univerzita 2019

ANALÝZA DAT
KIP/ANDAT

texty pro distanční studium

Autor: Petr Bujok

Ostravská univerzita, Přírodovědecká fakulta
Katedra Informatiky a počítačů

Jazyková korektura nebyla provedena, za jazykovou stránku odpovídá autor.

Obsah

Předmluva

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.

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.txt. Pokud si chcete uvedená řešení sami ověřit a zopakovat, tato data si můžete stáhnout z webových stránek autora textu, http://www1.osu.cz/∼bujok/.

Každá kapitola začíná pokyny pro její studium. Tato část je vždy označena jako **Průvodce studiem** s ikonou na okraji stránky.

Pojmy a důležité souvislosti k zapamatování jsou vyznačeny na okraji stránky textu ikonou.

V rozsahu celého textu jsou umístěny **Příklady**, jejichž podrobné řešení umožňuje porozumět probírané problematice do větší hloubky a tak si snáze osvojit praktiky pro další aplikace.

V závěru každé kapitoly je rekapitulace nejdůležitějších pojmů. Tato rekapitulace je označena textem **Shrnutí** a ikonou na okraji.

Oddíl **Kontrolní otázky** označený ikonou by vám měl pomoci zjistit, zda jste prostudovanou kapitolu pochopili a snad vyprovokuje i vaše další otázky, na které budete hledat odpověď.

U některých kapitol je připomenuta **Korespondeční úloha**. Pro kombinované a distanční studium jsou korespondenční úlohy zadávány v rámci kurzu daného semestru. Úspěšné vyřešení korespondenčních úloh je součástí podmínek pro celkové hodnocení předmětu.

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. 6 o prezentaci výsledků.

1 Parametrické testy o shodě středních hodnot

Průvodce studiem:

Tato kapitola shrnuje tři statistické parametrické testy. Text je rozdělen do několika logicky ucelených částí. K prostudování celé této kapitoly budete potřebovat asi 12 hodin. Studium vám ulehčí četné ilustrativní příklady. V případě nejasností je možné se obrátit na oporu předchozího kursu Základy pravděpodobnosti a statistiky.

Cíl: Po prostudování této části kapitoly byste měli:

- znát detaily testování statistických hypotéz,
- určit a interpretovat testové kritérium jednotlivých testů,
- chápat rozdíly mezi základními parametrickými testy,
- rozlišovat mezi jednostrannou a oboustrannou alternativní hypotézou.

1.1 Jednovýběrový t-test

Jednovýběrový oboustranný t-test byl podrobně vysvětlen v učebním textu Základy pravděpodobnosti a 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 $T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$ a při oboustranné alternativě $\mu \neq \mu_0$ je kritický obor $W \equiv (-\infty, t_{n-1}(\alpha/2)] \cup [t_{n-1}(1-\alpha/2), +\infty)$. Pokud vypočtená hodnota T leží v kritickém oboru, tak nulovou hypotézu $\mu = \mu_0$ pro dané α zamítáme (kvantily t-rozdělení jsou tabelovány 8.3).

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*:

$H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0$ (tzv. *pravostranná* alternativa)

Další postup testu bude zcela analogický jako u oboustranného testu, pouze kritický obor bude jiný, totiž $W \equiv [t_{n-1}(1-\alpha), +\infty)$. Nulovou hypotézu můžeme zamítnout

ve prospěch této alternativy tehdy, když výběrový průměr $\overline{X}$ je o hodně větší než μ_0. Přesněji vyjádřeno, když pro hodnotu testového kritéria platí

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

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 $W \equiv (-\infty, t_{n-1}(\alpha)]$.

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

Příklad 1.1 Na vybranou optimalizační úlohu byl aplikován deterministický algoritmus, který dosáhl jejího řešení za 2400 výpočetních kroků. Dlouhodobý výzkum ukazuje, že nedeterministický (stochastický) přístup může být efektivnější. Proto byl na stejnou úlohu nasazen rovněž stochastický algoritmus. Z důvodu stochastičnosti byl experiment opakován 60 krát, a výsledné počty kroků k dosažení řešení téže úlohy jsou v tabulce:

2622	2906	3816	1371	2812	1058	1749	2262	3992	1841	1200	2895
1675	4512	2530	2182	3044	2510	3087	3852	1220	1285	3984	2588
2232	2727	1741	2742	1932	3790	1548	1085	899	2269	2804	1336
3457	2525	1933	2307	2821	2333	2523	1122	1405	2661	1828	788
2018	1400	2130	635	1419	3275	2767	957	2594	1413	3124	3923

Zjistěte, zda je stochastický algoritmus efektivnější než deterministický.

Pokud máme zjistit, zda jsou výsledky opakování stejného pokusu stochastického algoritmu lepší než výsledek deterministického „etalonu“, použijeme jednovýběrový t-test. Nulovou hypotézu můžeme formulovat $H_0 : \mu_{\text{stoch}} = 2400$. Dosadíme-li do testového kritéria, obdržíme $T = -0.909$, a pokud výsledek porovnáme s tabulkou 8.3 ($\alpha = 0.05$, proto T_{krit}=2), zjistíme, že $|T| < T_{krit}$. Na základě toho nemůžeme zamítnout H_0, tudíž nově aplikovaný stochastický algoritmus dosahuje obdobných výsledků jako deterministický přístup (a to i přes to, že je průměrný počet kroků stochastického algoritmu roven 2291). Ukázka výstupu programu JASP:

					95% CI for Mean Difference	
	t	df	p	Mean Difference	Lower	Upper
kroky	-0.909	59	0.367	-109.067	-349.254	131.121

	N	Mean	SD	SE
kroky	60.000	2290.933	929.780	120.034

1.2 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 textu Základy pravděpodobnosti a statistiky víme, 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:

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

Pokud chceme testovat hypotézu, že střední hodnoty v obou populacích jsou shodné, tj.

$H_0 : \mu_1 = \mu_2$

proti některé z alternativ

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

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

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

užijeme testovou statistiku

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

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

$$T_{noneq} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}, \tag{2}$$

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 opora Základy pravděpodobnosti a statistiky)
$\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ a $\frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$
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 $n_1 - 1,\ n_2 - 1$,

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

Lze se dohodnout, že indexování výběrů zvolíme tak, aby platilo $s_1^2 \geq s_2^2$. Prakticky to znamená, že ve jmenovateli (3) bude menší z obou výběrových rozptylů. Pak kritickým oborem bude

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

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 MS Excel 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 komerčním statistickém software je ve výsledcích vyhodnocena zpravidla jak testová statistika (1) pro rovnost rozptylů, tak kritérium (2) pro neshodu rozptylů. Je na uživateli, aby si vybral správnou část výsledku pro interpretaci. Postup si ukážeme na příkladu.

Příklad 1.2 Máme posoudit, zda střední hodnoty *platu* (data *lide*) jsou stejné v populaci žijící ve Středomoří (1) i v populaci žijící ve Skandinávii (-1). Použijeme pro-

gram JASP, z menu *T-tests* vybereme *Independent Samples T-tests*. Zadáme *Prijem* jako *Variables* a veličinu *zeme* jako *Split* (tato veličina rozděluje pozorování do dvou skupin) a dostaneme výstup, který zde uvedeme ve zkrácené podobě.

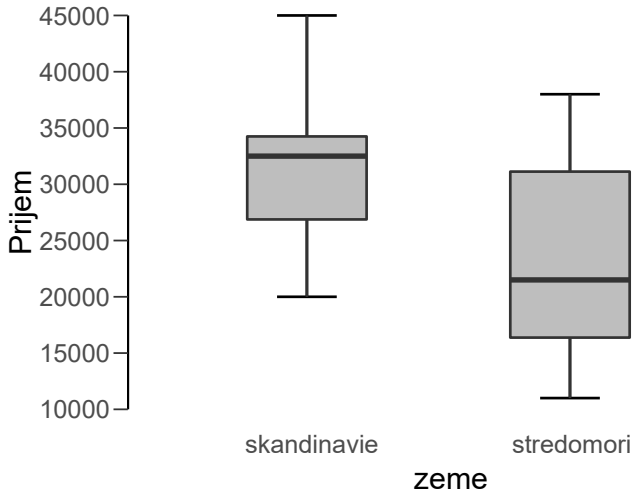
	Group	N	Mean	SD	SE
Prijem	-1	16	31406.250	6983.836	1745.959
	1	16	23468.750	9078.305	2269.576

Independent samples T-test

	Test	Statistic	df	p	Location Parameter	SE Difference
Prijem	**Student**	2.772	30.000	0.009	7937.500	2863.451

Test of equality of variances (Levene's)

	F	df	p
Prijem	3.239	1	0.082



Obrázek 1: Krabicový graf (tento obrázek je rovněž v příloze).

I zkrácený výstup je dosti obsáhlý a napoprvé nám dá trochu práce se v něm orientovat a správně interpretovat výsledky. Výhodou tohoto software je automatizovaná volba správného t-testu. Naším úkolem je testovat nulovou hypotézu o shodě středních hodnot proti oboustranné alternativě, tj. $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

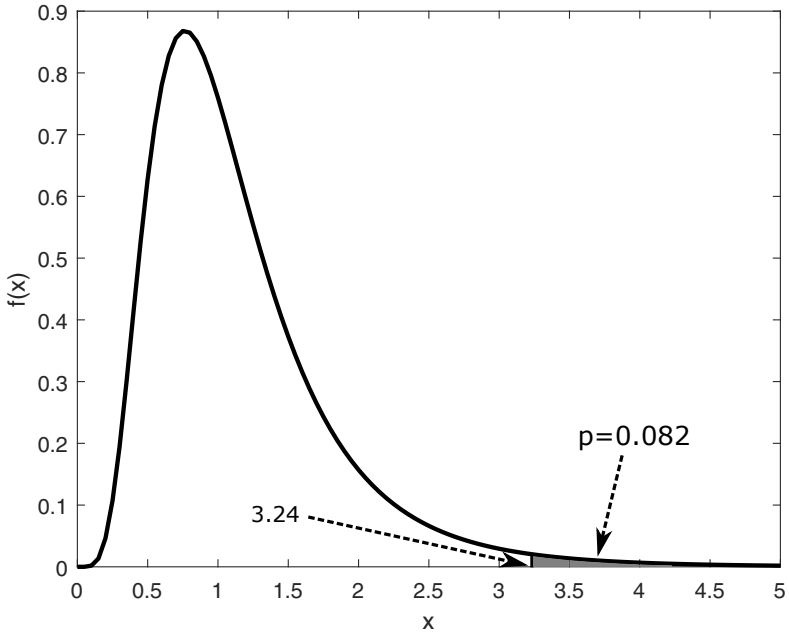
Stejnou nulovou i alternativní hypotézu můžeme formulovat i takto: $H_0 : \mu_1 - \mu_2 = 0$

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

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). 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 (*Test of equality of variances (Levene's)*). Tam nalezneme hodnotu testové statistiky spočtené podle vztahu (3) a kromě toho také tzv. dosaženou úroveň významnosti této hodnoty, která je uvedena ve sloupci p. Tato významnost (p, někdy označovaná také *p-value*, *prob-level*, *significance*) je často užívanou charakteristikou, která usnadňuje interpretaci výsledků. V případě *jednostranného* testu, což je tento test, 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 $p = P(X \geq 3,239) \cong 0,082$.

Smysl p v tomto příkladu i v jiných jednostranných testech vysvětluje následující obrázek.

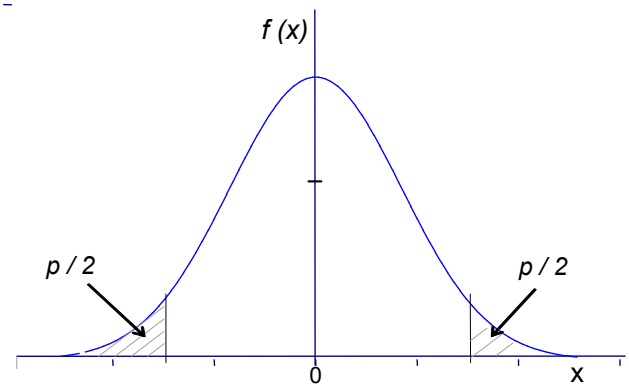


Obrázek 2: Princip p-hodnoty (tento obrázek je rovněž v příloze).

Je zřejmé, že pokud platí $p \leq \alpha$, nulovou hypotézu zamítáme, jinak nezamítáme. Jelikož v našem příkladu vyšlo $p \cong 0,082$, tedy nepatrně 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 *Independent samples T-test*. Její hodnota je 2,772 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 $p = P(|X| \geq 2,772) \cong 0,009$. Jednoduše řečeno, u oboustran-

ných 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 $p = 0,009$, 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 $\alpha = 0,001$, 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 $p \leq \alpha$, víme, že hodnota testového kriteria je v kritickém oboru, pokud $p > \alpha$, hodnota testového kriteria v kritickém oboru není.

Poznámka 1.1

Na hodnotu p lze nahlížet v určitém ohledu také jako na *pravděpodobnost důvěry v nulovou hypotézu.*

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 dostatečně velké. Jak víme z odstavce o centrální limitní větě, při dostatečně velkém počtu pozorování má testové kriterium

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5}$$

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.

1.3 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, hmotnost před a po dietě, nástupní a současná mzda 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 $\overline{D}$ 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

$$T_p = \frac{\overline{D}}{s_D/\sqrt{n}}, \tag{6}$$

která má rozdělení t_{n-1}. Podobně jako u jednovýběrového t-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, $H_0 : \mu_1 - \mu_2 = a$ (například hmotnost po dietě je alespoň o 10 kg nižší). Pak testovou statistikou je

$$T_p = \frac{\overline{D} - a}{s_D/\sqrt{n}}, \tag{7}$$

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

Příklad 1.3 Máme zjistit, zda ve firmě dochází ke zvyšování platů zaměstnanců (data *employs*). Pracuje zde 474 osob, a každá dospěla od nástupního platu (*nplat*) k platu, který má nyní (*plat*). V první tabulce výstupu JASP vidíme, že v průměru se plat zaměstnanců od nástupu liší zhruba o 50 %. Provedeme párový t-test (nabídka *T-tests* a *Paired Samples T-tests*), kde obě veličiny zadáme do kolonky *Variables*. Dosažená T statistika je $T = 35,036$, což velkoryse umožňuje zamítnout H_0 (také proto, že p je podstatně menší než 0,05).

	N	Mean	SD	SE
plat	474	34419.568	17075.661	784.311
nplat	474	17016.086	7870.638	361.510

Paired Samples T-Test

	t	df	p	Mean Difference	SE Difference
plat - nplat	35.036	473	1.610e-133	17403.481	496.732

Mzdy zaměstnanců firmy byly již letmým pohledem evidentně zásadně vyšší, v porovnání s nástupním platem. Samotný test pak tento verdikt jasně potvrdil.

Shrnutí:

- 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 hodnota p

Korespondenční úkol:

Korespondenční úlohy budou zadávány vždy na začátku semestru.

2 Analýza rozptylu - jednoduché třídění

Průvodce studiem:

Jako analýza rozptylu (angl. ANalysis Of VAriance – 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 4 až 5 hodin.

Cíl: Po prostudování této části kapitoly byste měli:

- rozlišovat dvouvýběrový t-test a jeho zobecnění,
- pochopit mechanismus testování nulové hypotézy ANOVA,
- dokázat interpretovat tzv. post-hoc testy.

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 $Y_{i1}, Y_{i2}, \ldots, Y_{in_i},\ n_i > 1, \quad i = 1,\ 2,\ \ldots, I$. 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

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \tag{8}$$

Každou tuto náhodnou veličinu můžeme tedy vyjádřit jako součet

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, 2, \ldots, n_i; \quad i = 1, 2, \ldots, I, \tag{9}$$

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*

$$\sum_{i=1}^{I} n_i \alpha_i = 0, \tag{10}$$

jsou hodnoty parametrů $\mu, \alpha_1, \alpha_2, \ldots, \alpha_I$ určeny jednoznačně a nulovou hypotézu (8) můžeme zapsat jako

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0. \tag{11}$$

Tato formulace je ekvivalentní formulaci (8). 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

$$\begin{aligned}
Y_{i\bullet} &= \sum_{j=1}^{n_i} Y_{ij} && \textit{(skupinové součty)},\\
\overline{Y}_{i\bullet} &= \frac{Y_{i\bullet}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} && \textit{(skupinové průměry)}\\
Y_{\bullet\bullet} &= \sum_{i=1}^{I} Y_{i\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} && \textit{(celkový součet)},\\
\overline{Y}_{\bullet\bullet} &= \frac{Y_{\bullet\bullet}}{n} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} && \textit{(celkový průměr)}
\end{aligned} \tag{12}$$

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

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{13}$$

Tento tzv. *celkový* součet čtverců můžeme rozložit

$$
\begin{aligned}
S_T &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(Y_{ij}-\overline{Y}_{i\bullet})+(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})\right]^2 = \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(Y_{ij}-\overline{Y}_{i\bullet})(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})\right] + \\
&+ \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2 + \qquad (14)\\
&+2\sum_{i=1}^{I}(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet}) + \sum_{i=1}^{I} n_i\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2 = \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2 + \sum_{i=1}^{I} n_i\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2
\end{aligned}
$$

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

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

Poznámka 2.1

Dva členy v posledním řádku (14) jsou charakteristikami variability

- **uvnitř** skupin

$$S_e = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2 \qquad (15)$$

  (součet čtverců odchylek pozorovaných hodnot od skupinových průměrů),

- **mezi** skupinami

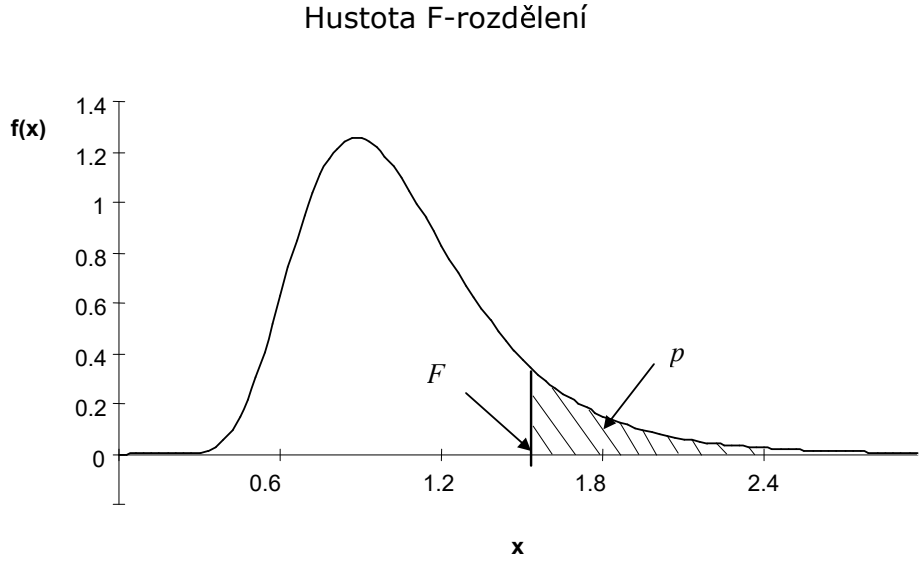
$$S_A = \sum_{i=1}^{I} n_i\left(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}\right)^2 \qquad (16)$$

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

Vztah (14) tedy můžeme přepsat jako

$$S_T = S_e + S_A \qquad (17)$$

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 (11), 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

$$F = \frac{S_A/(I-1)}{S_e/(n-I)} \sim F_{I-1,n-I}. \tag{18}$$

Pokud nulová hypotéza neplatí, je statistika $S_A/(I-1)$ výrazně větší. Kritickým oborem pro zamítnutí nulové hypotézy (11) je $W = [F_{I-1,n-I}(1-\alpha), +\infty)$.

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, $p \leq \alpha$, 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*, nebo tzv. *post-hoc* testy). Těch je několik druhů, popis a základní informace k jejich užití nalezneme v nápo-

vědě statistického software, zájemce o podrobnější informace odkazujeme na literaturu, např. Anděl 1978, 1993, Havránek 1993 atd., podobně jako zájemce o složitější modely analýzy rozptylu.

Poznámka 2.2

Pokud bychom užili analýzu rozptylu s jednoduchým tříděním na data pocházející jen ze dvou výběrů, bude mít statistika F z rov. (18) 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 (17) 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, $\quad i = 1, 2, \ldots, I$.

Směrodatná odchylka je odmocnina z celkového rozptylu, tj.

$$s = \sqrt{\frac{S_T}{n-1}} = \sqrt{\frac{S_e + S_A}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{I} {s_i}^2\,(n_i - 1) + \sum_{i=1}^{I} n_i(\overline{x}_i - \overline{x})^2\right]}, \qquad (19)$$

kde celkový průměr spočítáme jako vážený průměr skupinových průměrů,

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

Aplikaci analýzy rozptylu s jednoduchým tříděním ukážeme na následujícím příkladu.

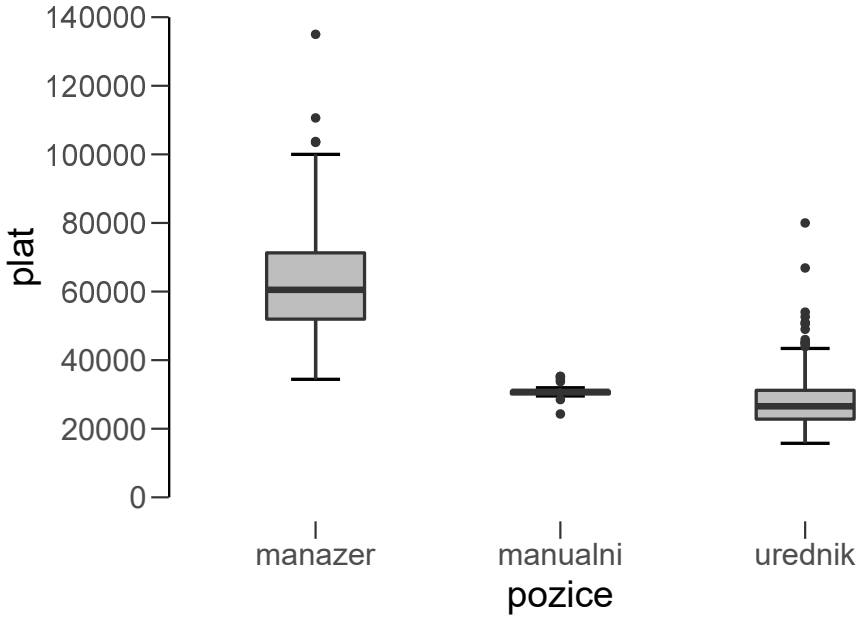
Příklad 2.1 Máme posoudit, zda střední hodnota veličiny *plat* (data *employs*) jsou stejné ve všech třech profesích (*manualni*, *urednik*, *manazer*).

Pro test hypotézy o shodě středních hodnot

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

užijeme analýzu rozptylu s jednoduchým tříděním (třídicí faktor je *pozice*). Výpočet provedeme s pomocí statistického software JASP. V něm z menu *ANOVA* vybereme *ANOVA*. Zadáme veličinu *plat* jako *Dependent variable* a veličinu *pozice* jako *Fixed Factors* (tato veličina rozděluje pozorování do třech skupin) a dostaneme výstup, který zda uvedeme ve zkrácené podobě:

pozice	Mean	SD	N
manazer	63977.798	18244.776	84
manualni	30938.889	2114.616	27
urednik	27838.540	7567.995	363



ANOVA – plat

Cases	Sum of Squares	df	Mean Square	F	p
pozice	8.944e+10	2.000	4.472e+10	434.481	1.165e-107
Residual	4.848e+10	471.000	1.029e+8		

		Mean Difference	SE	t	$\mathbf{p}_{bonf}$
manazer	manualni	33038.909	2244.409	14.721	3.810e-40
	urednik	36139.258	1228.352	29.421	2.883e-108
manualni	urednik	3100.349	2023.760	1.532	0.379

Z tabulky analýzy rozptylu vidíme, že p <0,05, tedy nulovou hypotézu můžeme zamítnout. Rozdíly v poloze pozorovaných hodnot veličiny *plat* v jednotlivých skupinách (viz krabicové diagramy na obrázku výše) jsou způsobeny systematickými rozdíly mezi skupinami zaměstnanců (*pozice*), nikoliv pouze v důsledku nahodilého kolísání.

V takovém případě je vhodné, dále zkoumat, ve kterých skupinách zaměstnanců se plat liší, a ve kterých nikoliv. K tomu poslouží tzv. *post-hoc* testy (někdy také *multiple comparison tests*), které vzájemně porovnají rozdíly mezi platy jednotlivých dvojic-skupin zaměstnanců. Existuje několik versí těchto testů, v tomto případě jsme zvolili Bonferroniho test (dosažená významnost je ve sloupci $\mathbf{p}_{bonf}$) s verdiktem, že manuální zaměstnanci a úředníci nemají významně rozdílné příjmy. Naopak

je významný rozdíl ve střední hodnotě platu manažerů a zbylých dvou skupin zaměstnanců (což lze odečítat i z krabicového grafu).

Shrnutí:

- 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. 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

Korespondenční úkol:

Korespondenční úlohy budou zadávány vždy na začátku semestru.

3 Základy lineární regrese

Průvodce studiem:

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 10 hodin.

Cíl: Po prostudování této části kapitoly byste měli:

- chápat základní princip lineární regrese,
- rozpoznat rozdíl mezi regresí a korelací,
- umět nalézt odhady parametrů s pomocí metody nejmenších čtverců,
- umět identifikovat nevhodnou nezávisle proměnnou,
- zvládnout vyčíslit a interpretovat vhodnost modelu.

Lineární regrese se zabývá problémem vysvětlení změn hodnot jedné veličiny lineární závislostí na jedné nebo více jiných veličinách. Uvažujme nejjednodušší případ, kdy vysvětlujeme veličinu $\boldsymbol{Y}$ lineární závislostí na jedné vysvětlující veličině $\boldsymbol{x}$.

Příklad 3.1 Data mají tvar, který je uveden v následující tabulce:

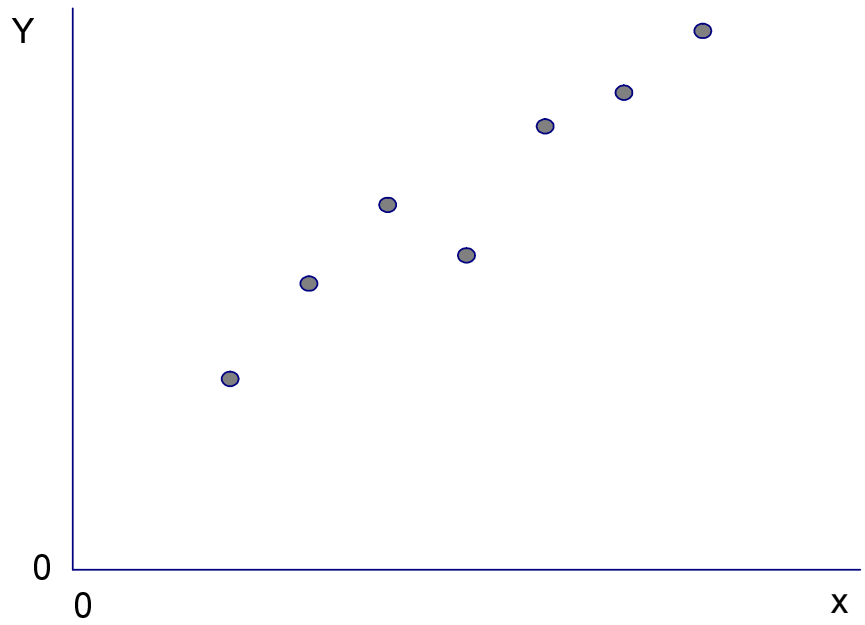
i	x_i	Y_i
1	x_1	Y_1
2	x_2	Y_2
$\vdots$		
n	x_n	Y_n

Předpokládáme, že hodnoty veličiny $\boldsymbol{x}$ 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 rostoucími hodnotami veličiny $\boldsymbol{x}$ se zhruba lineárně mění i hodnoty $\boldsymbol{Y}$. Body na obrázku kolísají kolem myšlené přímky, kterou bychom mohli naměřenými body proložit.

Hodnoty veličiny Y_i můžeme vyjádřit jako součet dvou složek:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{20}$$



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

Pokud střední hodnoty náhodného kolísání jsou nulové, $E(\varepsilon_i) = 0, i = 1, 2, \ldots, n$, rov. (20) můžeme přepsat

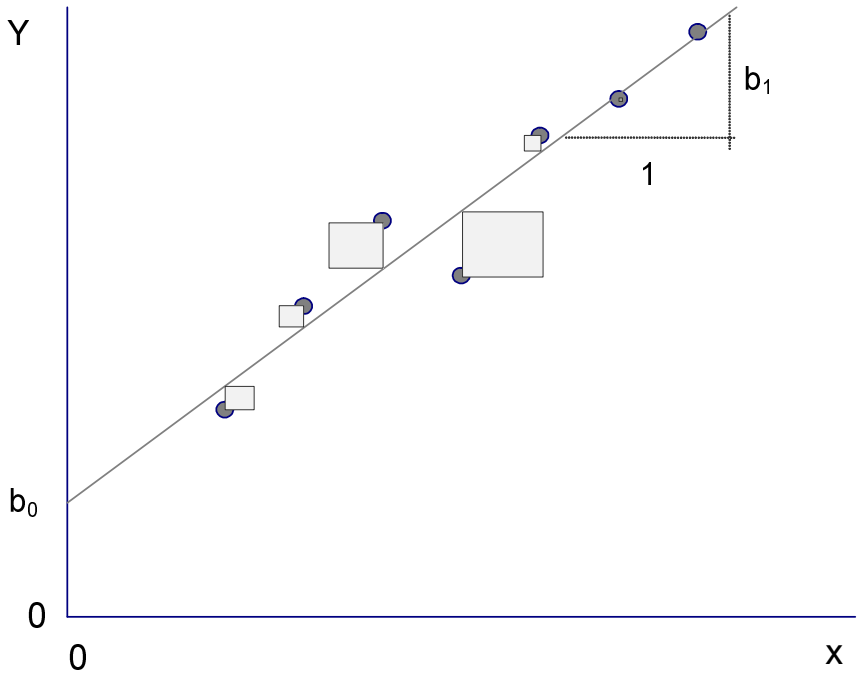
$$E(\boldsymbol{Y}|\boldsymbol{x} = x_i) = E(Y_i) = \beta_0 + \beta_1 x_i \tag{21}$$

čili střední hodnoty náhodných veličin Y_i za podmínky, že veličina $\boldsymbol{x}$ má hodnotu x_i, leží na přímce dané rov. (21).

Rovnice (20) a (21) formulují regresní model, v tomto případě *lineární regresní model* s jednou vysvětlující proměnnou (regresorem) $\boldsymbol{x}$ a vysvětlovanou proměnnou $\boldsymbol{Y}$. 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 $\boldsymbol{Y}$ na veličině $\boldsymbol{x}$.

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 $E(\boldsymbol{Y}|\boldsymbol{x} = x_i)$. 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 $\boldsymbol{Y}$) b_1 (směrnice přímky), aby součet čtverců odchylek pozorovaných hodnot Y_i od hodnot $\hat{Y}_i$ byl co nejmenší:

$$S_e = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 \to \min \tag{22}$$



Obrázek 3: Metoda nejmenších čtverců (tento obrázek je rovněž v příloze).

Příklad 3.2 Metodu nejmenších čtverců vysvětluje následující obrázek. Řešíme úlohu, jak volit hodnoty b_0 a b_1, aby součet obsahů 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

$$\begin{aligned} \frac{\partial S_e}{\partial b_0} &= -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 x_i\right) = -2\left(\sum_{i=1}^{n} Y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i\right), \\ \frac{\partial S_e}{\partial b_1} &= -2\sum_{i=1}^{n}\left[\left(Y_i - b_0 - b_1 x_i\right) x_i\right] = -2\left(\sum_{i=1}^{n} x_i Y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2\right). \end{aligned} \tag{23}$$

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

$$\begin{aligned} nb_0 + b_1 \sum x_i &= \sum Y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 &= \sum x_i Y_i \end{aligned} \tag{24}$$

Řešení této soustavy rovnic můžeme vyjádřit explicitně takto:

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum x_i\right) = \overline{Y} - b_1 \overline{x} \tag{25}$$

$$b_1 = \frac{\sum x_i Y_i - \frac{(\sum x_i)(\sum Y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{n \sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - (\sum x_i)^2}. \tag{26}$$

Z rov. (25) vidíme, že přímka proložená metodou nejmenších čtverců, tj. splňující podmínku (22), prochází bodem $[\overline{x}, \overline{Y}]$.

Dosadíme-li z rov. (26) do (25), dostaneme

$$\begin{aligned} b_0 &= \frac{1}{n}\left[\sum Y_i - \frac{n(\sum x_i Y_i) - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - (\sum x_i)^2} \sum x_i\right] = \\ &= \frac{(\sum Y_i)(\sum x_i^2) - (\sum x_i Y_i)(\sum x_i)}{n \sum x_i^2 - (\sum x_i)^2} \end{aligned} \tag{27}$$

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

$$\begin{aligned} \sum (x_i - \overline{x})^2 &= \sum \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right) = \sum x_i^2 - 2\overline{x} \sum x_i + n\overline{x}^2 = \\ &= \sum x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 = \sum x_i^2 - n\overline{x}^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \end{aligned} \tag{28}$$

$$\sum (x_i - \overline{x})\, x_i = \sum \left(x_i^2 - \overline{x}x_i\right) = \sum x_i^2 - \overline{x} \sum x_i = \sum (x_i - \overline{x})^2 \tag{29}$$

$$\begin{aligned} \sum (x_i - \overline{x})\left(Y_i - \overline{Y}\right) &= \sum \left(x_i Y_i - \overline{Y}x_i - \overline{x}Y_i + \overline{x}\overline{Y}\right) = \\ &= \sum x_i Y_i - \overline{x} \sum Y_i - \overline{Y} \sum x_i + n\overline{x}\overline{Y} = \\ &= \sum x_i Y_i - n\overline{x}\overline{Y} - n\overline{x}\overline{Y} + n\overline{x}\overline{Y} = \\ &= \sum x_i Y_i - n\overline{x}\overline{Y} = \sum x_i Y_i - \frac{(\sum x_i)(\sum Y_i)}{n} \end{aligned} \tag{30}$$

$$\begin{aligned} \sum (x_i - \overline{x})\, Y_i &= \sum x_i Y_i - \overline{x} \sum Y_i = \\ &= \sum x_i Y_i - \frac{\sum x_i \sum Y_i}{n} = \sum (x_i - \overline{x})\left(Y_i - \overline{Y}\right) \end{aligned} \tag{31}$$

Z rov. (26), (28) a (31) pak dostaneme

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

kde s_x^2 je výběrový rozptyl veličiny $\boldsymbol{x}$ 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 (30) a (31) můžeme rov. (26) přepsat

$$b_1 = \frac{\sum (x_i - \overline{x})\, Y_i}{\sum (x_i - \overline{x})^2} \tag{32}$$

Odtud

$$b_1 \sum (x_i - \overline{x})^2 = \sum (x_i - \overline{x})\, Y_i$$

Pak pro střední hodnoty náhodných veličin v předchozí rovnici platí

$$E(b_1) \sum (x_i - \overline{x})^2 = \sum (x_i - \overline{x})\, E(Y_i) = \sum (x_i - \overline{x})\,(\beta_0 + \beta_1 x_i) =$$

$$= \beta_1 \sum (x_i - \overline{x})\, x_i = \beta_1 \sum (x_i - \overline{x})^2$$

.

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

Podobně pro b_0 můžeme dosadit do (25)

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

Můžeme ukázat, že

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

a také, že

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

.

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. (20):

a) $E(\varepsilon_i) = 0, \qquad i = 1, 2, \ldots, n$ (tento předpoklad už byl vysloven dříve);

b) $var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2, \qquad i = 1, 2, \ldots, n$
(rozptyl e_i je konstantní, tzv. homoskedasticita);

c) $cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \ldots, n$
($\varepsilon_i, \varepsilon_j$ jsou nekorelované).

Z rov. (20) vidíme, že $var(Y_i) = var(e_i) = \sigma^2$. Pak z rov. (32) dostaneme

$$var(b_1) = \frac{1}{\left[\sum (x_i - \overline{x})^2\right]^2} \sum (x_i - \overline{x})^2 \, var(Y_i) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2}. \tag{33}$$

Z rov. (33) vidíme, že rozptyl odhadu směrnice regresní přímky můžeme snížit vhodnou volbou hodnot regresoru tak, aby $\sum (x_i - \overline{x})^2$ byla co největší.

Z rov. (25) dostaneme

$$var(b_0) = var(\overline{Y}) + \overline{x}^2 var(b_1) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2} \right) \tag{34}$$

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

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

d) $\varepsilon_i \sim N(0, \sigma^2) \quad i = 1, 2, \ldots, n$
(odchylky hodnot Y_i od lineární závislosti mají normální rozdělení), pak

$$\frac{b_j - \beta_j}{\sqrt{var(b_j)}} \sim N(0, 1), \qquad j = 0, 1 \tag{35}$$

Pokud bychom znali $var(b_j)$, mohla by statistika definovaná rov. (35) sloužit jako testové kritérium pro testy hypotéz o parametrech regresního modelu.

Obyčejně však $var(b_j)$ neznáme, neboť neznáme σ^2 - viz rov. (33) a (34). Hodnotu σ^2 (tzv. reziduální rozptyl) však můžeme odhadnout:

$$\hat{\sigma}^2 = s^2 = \frac{S_e^2}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2}{n-2}. \tag{36}$$

Charakteristika s^2 definovaná rov. (36) - *výběrový residuální rozptyl* - je nestranným odhadem hodnoty σ^2. Dosadíme-li tento odhad do rov. (33) a (34) 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

$$\frac{b_j - \beta_j}{s(b_j)} \sim t_{n-2}, \quad j = 0, 1, \tag{37}$$

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

Poznámka 3.1

Lineární regresní model (20) 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 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 3.3 Uvažujme data ze souboru *lide*. Naším úkolem je odhad regresních parametrů lineárního modelu závislosti veličiny *Hmotnost* na veličině *Vyska*. V řešení využijeme například statistický program JASP. Volbou *File/Open* otevřeme soubor lide.jasp (vytvořený dříve programem JASP) a v menu *Regression* vybereme *Linear Regression*. V šabloně regrese zvolíme jako vysvětlovanou veličinu (*Dependent vari-*

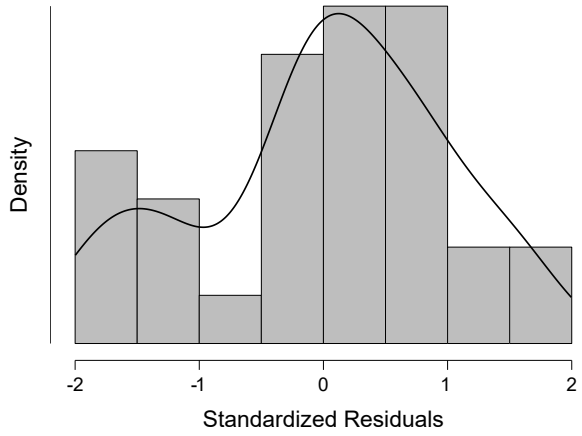
able) Hmotnost, jako regresor (*Covariates*) zvolíme jedinou veličinu, a to Vyska. Po spuštění výpočtu dostaneme následující výstup:

Coefficients

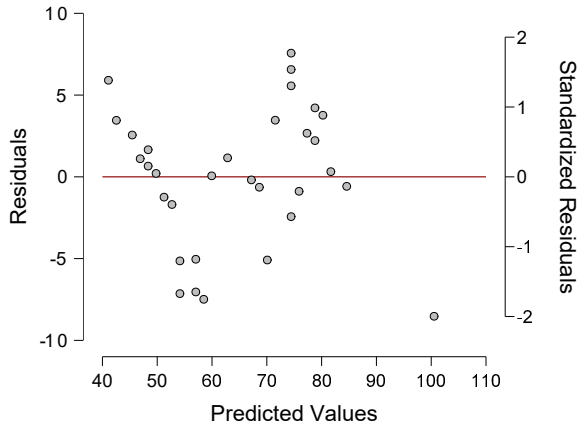
					95% CI	
	Unstandar.	**Std. Error**	**t**	**p**	**Lower**	**Upper**
(Intercept)	-186.488	13.445	-13.870	1.379e-14	-213.947	-159.029
Vyska	1.450	0.078	18.696	4.413e-18	1.291	1.608

ANOVA

R	**R^2**	**Adjusted R^2**	**RMSE**	**R^2 Change**	**F Change**
0.960	0.921	0.918	4.342	0.921	349.525



Obrázek 4: Histogram residuí (tento obrázek je rovněž v příloze).



Obrázek 5: Rozdělení residuí (tento obrázek je rovněž v příloze).

Odhady parametrů lineárního regresního modelu jsou v části *Coefficients*. Na řádku *Intercept* je odhad úseku regresní přímky - viz rov. (27) - a další charakteristiky týkající se tohoto parametru, na řádku *Vyska* pak je odhad směrnice - viz rov. (26) - a další charakteristiky týkající se tohoto parametru. Odhady parametrů b_0, b_1, jsou tedy ve sloupci *Unstandardized*. Ve sloupci *Std. Error* jsou pak $s(b_j)$, $j = 0, 1$ -

viz rov. (33), (34) a následující text. Ve sloupci t jsou hodnoty testového kritéria $\frac{b_j}{s(b_j)}$ pro test hypotézy $\beta_j = 0$ - viz rov. (37) - a ve sloupci p jsou významnosti p pro oboustranný test.

Výsledkem naší úlohy jsou odhady b_0 (úsek) = -186,5 a b_1 (směrnice) = 1,45. Kromě toho vidíme, naše data nás opravňují zamítnout hypotézu $\beta_1 = 0$, (v tabulce výsledků má hodnota p-value 17 nul, tzn. $p < 5 \times 10{-}18$), takže nulovou hypotézu můžeme zamítnout na jakékoli rozumně zvolené hladině významnosti. Zřejmě se váha s rostoucí délkou významně mění. Stejně tak můžeme zamítnout hypotézu $\beta_0 = 0$ ($p < 5 \times 10^{-13}$), tudíž regresní přímka neprochází počátkem. Takový regresní model jen s jedním parametrem, a to směrnicí, bychom měli prozkoumat v dalším kroku. Význam důležité charakteristiky $\mathbf{R}^2$ vysvětlíme později.

V části *Coefficients* jsou rovněž uvedeny $100(1-\alpha)$-procentní intervalové odhady regresních parametrů (ve sloupcích *Lower* a *Upper* 95 % C.I.), hodnota α může být zvolena při zadání výpočtu.

Část *ANOVA* vysvětlíme později. Z dalších charakteristik je užitečná *RMSE*, což je směrodatná odchylka odhadu, odmocnina z výrazu daného rov. (36), 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.

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 $\boldsymbol{Y}$ od jejich průměru můžeme rozložit na dva sčítance:

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \tag{38}$$

Označme jednotlivé sumy čtverců podle jejich významu

- celková suma čtverců (total sum of squares):

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

- 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):

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

Rov. (38) 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:

$$TSS = MSS + RSS. \tag{39}$$

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

$$R^2 = \frac{MSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{40}$$

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

$$0 \leq R^2 \leq 1 \tag{41}$$

Hodnotu 1 dosahuje R^2 tehdy, když $RSS = 0$ (viz rov. (40)), tj, že závislost $\boldsymbol{Y}$ na $\boldsymbol{x}$ je přesně lineární (model vysvětluje vše). Hodnotu 0 dosahuje index determinace tehdy, když model nevysvětluje z variability $\boldsymbol{Y}$ nic, tzn. $RSS = TSS$, tedy regresní přímka je rovnoběžná s osou $\boldsymbol{x}$ v úrovni $b_0 = \overline{Y}$.

Lze také ukázat, že pro lineární regresní model s jedním regresorem - rov. (20) nebo (21) - je koeficient determinace roven druhé mocnině výběrového korelačního koeficientu, tedy

$$R^2 = r_{xy}^2. \tag{42}$$

Poznámka 3.2

Při používání tohoto vztahu nezapomeňte, že $-1 \leq r_{xy} \leq 1$ 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 $\boldsymbol{Y}$ není na $\boldsymbol{x}$ 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 $\boldsymbol{Y}$ je vysvětlena lineární závislostí na jednom nebo více regresorech.

Shrnutí:

- Regrese je jedna z nejužívanějších metod statistiky.
- Nejjednodušší a nejznámější forma regrese je lineární model.
- Cílem regrese je modelovat závislost mezi závisle proměnnou a nezávisle proměnnou(-ými).
- Hledání koeficientů modelu zajišťuje nejčastěji metoda nejmenších čtverců.
- Vhodnost modelu lze odhadovat s pomocí koeficientu determinace.
- Pro každý z regresorů je potřeba testovat, zda má v modelu uplatnění.
- Na lineární regresi jsou kladeny předpoklady, které je potřeba ověřit.

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 úkol:

Korespondenční úlohy budou zadávány vždy na začátku semestru.

4 Neparametrické metody

Průvodce studiem:

V této rozsáhlé kapitole se seznámíme se základy tzv. neparametrických metod. Metod, 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.

Cíl: Po prostudování této části kapitoly byste měli:

- porozumět obecnému principu použití neparametrických metod,
- být schopni pracovat s výpočtu testu dobré shody a testu nezávislosti,
- podrobněji pochopit princip Wilcoxonova a Kruskal-Wallisova testu.

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.

4.1 Test dobré shody

Test dobré shody (angl. goodness-of-fit test) se užívá k ověřování shody empirického rozdělení s nějakým teoretickým rozdělením. Ilustruje to následující příklad.

Příklad 4.1 Chceme ověřit, zda je počet útoků do počítačové sítě v rámci vybrané společnosti stejný v rámci týdne. Jinými slovy, zda všech sedm možných výsledků má stejnou pravděpodobnost. Firma zaznamenala následující četnosti v úhrnu jednoho kalendářního měsíce:

výsledek	Po	Út	St	Čt	Pá	So	Ne	n
četnost n_i	119	94	93	82	126	139	117	770

Testujeme nulovou hypotézu, že pravděpodobnosti $p_i = 1/7$. Můžeme tedy spočítat četnosti e_i, které bychom očekávali za platnosti nulové hypotézy ze 770 útoků za platnosti nulové hypotézy ($n = 770$), $e_i = n \cdot p_i = 770 \cdot (1/7) = 110$.

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

$$X = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}, \tag{43}$$

kde k je počet možných výsledků, v našem příkladu $k = 7$. Tato statistika má při dostatečně velkém n (takovém, aby všechny $e_i \geq 5$) rozdělení chí-kvadrát s $k - 1$ stupni volnosti,

$$X = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{k-1}. \tag{44}$$

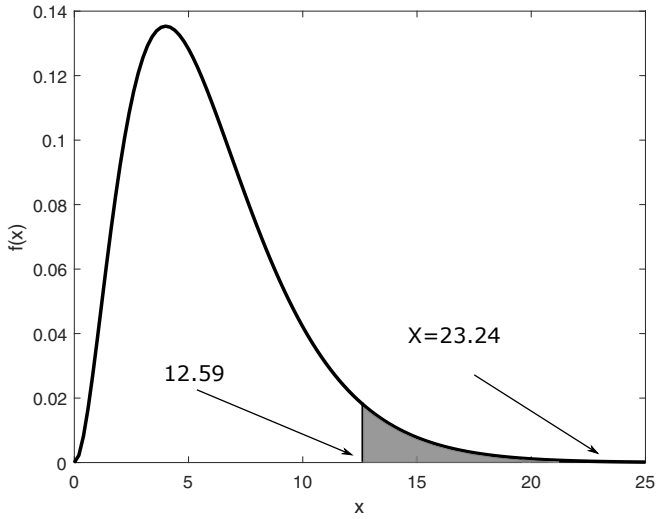
Nulovou hypotézu zamítneme, pokud odchylky od očekávaných četností jsou velké, tj. když hodnota testového kritéria X je v kritickém oboru W,

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

Pro náš příklad je výpočet ukázán v následující tabulce.

i	n_i	p_i	e_i	χ^2	ε_i
Po	119	1/7	110	0.74	0.86
Út	94	1/7	110	0.80	-1.53
St	93	1/7	110	1.25	-1.62
Čt	82	1/7	110	1.25	-2.67
Pá	126	1/7	110	1.80	1.53
So	139	1/7	110	0.80	2.77
Ne	117	1/7	110	0.80	0.67
Σ	770	1	770	23.24	0

Zvolíme-li $\alpha = 0,05$, je kritický obor $W \equiv [12.59, +\infty)$. Hodnota testové statistiky je 23,24, leží tedy v kritickém oboru, a proto nulovou hypotézu zamítáme. Je tedy zřejmé, že počet útoků do počítačové sítě společnosti se v rámci dnů týdne liší, což znázorňuje i obrázek:



V takovém případě nás dále zajímá, zda některý den dochází k podstatně vyšším, či podstatně nižším počtům útoků, oproti zbytku týdne. K tomu poslouží tzv. standardizovaná residua – ε_i, která mají zhruba normované normální rozdělení:

$$\varepsilon_i = \frac{(n_i - e_i)}{\sqrt{e_i}}.$$

Je evidentní, že residua mohou nabývat kladných i záporných hodnot, což v případě překročení mezní hodnoty normovaného normálního rozdělení ($\approx \pm 1.96$) odhalí zvýšený (+) či naopak snížený (−) počet výskytů útoků. Z posledního sloupce naší tabulky vidíme, že v sobotu dochází k významně vyšším počtům útoků, a naopak ve čtvrtek je počítačová síť společnosti ohrožována významně nejméně.

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 (43). 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,

$$e_i = n \cdot p_i = \frac{n}{k} \geq 5, \tag{45}$$

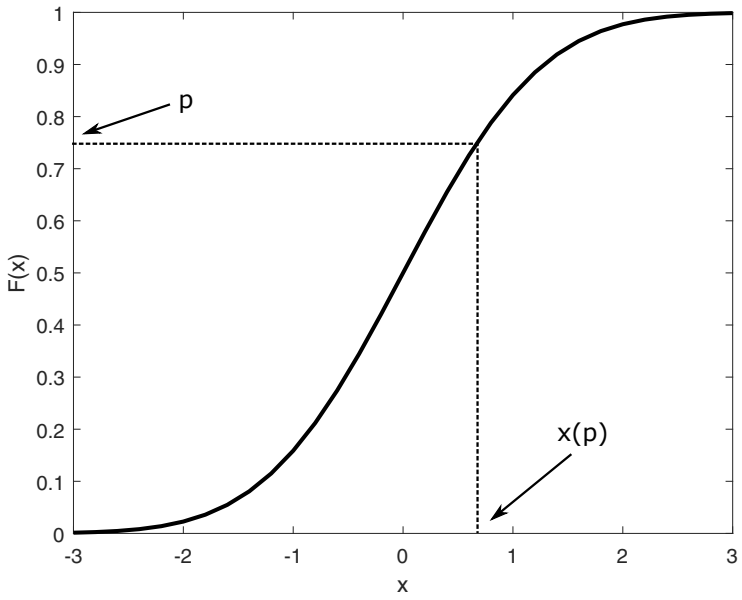
tedy k volíme tak, aby $k \leq n/5$.

Hranice intervalů jsou pak následující kvantily teoretického rozdělení,

$$x(i \cdot p_i) = x(i/k), \quad i = 0, 1, \ldots, k. \tag{46}$$

Pak už se jen spočítají četnosti n_i, $i = 0, 1, \ldots, k$, tj. počty hodnoty v jednotlivých intervalech a vyhodnotí testové kriterium (43).

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



Obrázek 6: Kvantil a distribuční funkce (tento obrázek je rovněž v příloze).

Uvědomme si, že podmínka (45), znamená, že dělení na svislé ose hodnot $F(x)$ je ekvidistantní, zatímco intervaly (jejich hranice dané vztahem (46) 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.

4.2 Kontingenční tabulky - test nezávislosti

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

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

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

$$n_{i\bullet} = \sum_{j=1}^{C} n_{ij} n_{\bullet j} = \sum_{i=1}^{R} n_{ij} \tag{47}$$

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

$$n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij} = \sum_{i=1}^{R} n_{i\bullet} = \sum_{j=1}^{C} n_{\bullet j} \tag{48}$$

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 $\boldsymbol{X}, \boldsymbol{Y}$ je založen na tom, že můžeme odhadnout četnosti, které bychom pozorovali, kdyby opravdu veličiny $\boldsymbol{X}, \boldsymbol{Y}$ nezávislé byly. Jsou-li $\boldsymbol{X}, \boldsymbol{Y}$ nezávislé, pak pravděpodobnost jevu, že současně nastane jev $Y = y_i$ a jev $\boldsymbol{X} = x_j$ vyjádřit jako součin pravděpodobností

$$\begin{gathered} P\left[(Y = y_i) \bigcap (X = x_j)\right] = P(Y = y_i) \cdot P(X = x_j), \\ i = 1, 2, \dots, R, j = 1, 2, \dots, C. \end{gathered} \tag{49}$$

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.(49) můžeme přepsat

$$p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \quad i = 1, 2, \dots, R, \quad j = 1, 2, \dots, C \tag{50}$$

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

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}, \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}, \tag{51}$$

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

$$e_{ij} = n\hat{p}_{ij} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}. \tag{52}$$

Nulovou hypotézu

$$H_0 : \textit{veličiny } X, Y \textit{ jsou nezávislé} \tag{53}$$

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 (53) je

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{54}$$

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

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(R-1)(C-1)}. \tag{55}$$

Jelikož (55) 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 $\boldsymbol{X}$ a $\boldsymbol{Y}$, 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 (55). Velikost tohoto příspěvku je významná, když rozdíl pozorované a očekávané četnosti nelze považovat za náhodný, tj. tehdy, když

$$\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \geq \chi_1^2 (1 - \alpha) , \tag{56}$$

pro obvykle užívanou hodnotu $\alpha = 0,05$ je $\chi_1^2(0,95) = 3,84$ (viz tabulky 8.1).

Pohodlnější je užít standardizovaná residua $\varepsilon_{ij} = (n_{ij} - e_{ij}) / \sqrt{e_{ij}}$. 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 4.2 Máme posoudit, zda veličiny *vzdel* a *pozice* (data employs) jsou nezávislé. Jinými slovy, zda zastoupeni všech tří stupňů vzdělání je ve třech profesích shodné.

H_0: *vzdel* a *pozice* jsou nezávislé veličiny

Výpočet provedeme s pomocí programu JASP. V něm z menu *Frequencies* vybereme *Contingency Tables*. Zadáme veličinu *vzdel* a *pozice* jako *Rows* a *Columns*.

Pořadí ovlivňuje pouze tvar tabulek ve výstupu, nikoliv hodnotu spočtené testové statistiky. Ovšem doporučujeme do sloupců (columns) uvádět veličinu, která by měla být závislá na druhé veličině v řádcích (rows). V našem příkladu tedy do sloupců zadáme *pozice* a do řádků *vzdel*, protože logicky úroveň vzdělání může ovlivnit pracovní pozici, a ne naopak.

Contingency Tables

		pozice			
vzdel		manazer	manualni	urednik	Total
SS	Count	1.0	13.0	182.0	196.0
	Expected count	34.7	11.2	150.1	196.0
VS	Count	83.0	1.0	141.0	225.00
	Expected count	39.9	12.8	172.3	225.0
ZS	Count	0.0	13.0	40.0	53.0
	Expected count	9.392	3.0	40.6	53.0
Total	Count	84.0	27.0	363.0	474.0
	Expected count	84.0	27.0	363.0	474.00

V šabloně *Results* vyznačíme, které výstupy požadujeme, v tomto příkladu *Count* (pozorované četnosti), *Expected count* (očekávané četnosti) a *Chi-square Tests* (tes-

tovou statistiku definovanou rov. (54)). Po provedení výpočtu dostaneme následující výstup, zde je uveden mírně zkrácen.

Chi-Square Tests

	Value	df	p
X^2	145.472	4	1.901e-30
N	474		

Standardized Residuals

vzdel	manazer	manualni	urednik
SS	-5.72	0.55	2.60
VS	6.83	-3.30	-2.39
ZS	0	5.74	-0.09

V řádku X^2 vidíme, že hodnota testové statistiky je 145,5, hodnota významnosti p je menší než obvykle volená hladina významnosti $\alpha = 0,05$ a hypotézu o nezávislosti veličin *vzdel* a *pozice* můžeme na této hladině významnosti zamítnout.

Vidíme, že všechny očekávané četnosti jsou větší než 5, jak vidíme v řádcích *Expected count*. Program JASP v aktuální variantě nepočítá standardizovaná residua, tyto však lze snadno spočítat např. v MS Excel, jak ukazuje poslední tabulka.

Celkově můžeme shrnout, že hypotézu o nezávislosti veličin *vzdel* a *pozice* jsme zamítli na hladině významnosti $\alpha = 0,05$, s přehledem bychom ji však zamítli i na jakékoli jiné hladině významnosti. Podíváme-li se na zdroje závislosti (*Standardized Residuals*), vidíme, že je značně mnoho vysokoškolsky vzdělaných manažerů, středoškolsky vzdělaných úředníků a manuálně pracujících se základním vzděláním, což lze považovat, za logický fakt.

Statistiku (54) 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.

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

Pro veličiny *vzdel* a *pozice* z uvedeného příkladu hodnoty těchto koeficientů získáme navolením příslušných políček:

Cramer's V	0.392
Contingency Coefficient	0.485

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

Poznámka 4.1

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.

4.3 Znaménkový test

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

$H_0 : \tilde{x} = x_0$

$H_1 : \tilde{x} > x_0$

Testovou statistikou je počet hodnot x_i ve výběru větších 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

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

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

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

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

$$U = \frac{Z - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2Z - n}{\sqrt{n}} \tag{57}$$

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

Poznámka 4.2

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.

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

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

Testovou statistikou je počet pozorování, pro která platí $Y_i > X_i$, přičemž 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 určitý „nový“ přístup v oblasti informatiky přináší uživatelům subjektivní pocit zlepšení práce či zábavy. Nový přístup je aplikován na n počítačích (či u n uživatelů), dotazem na každého uživatele zjistíme, že u z uživatelů došlo ke zlepšení, u $n - z$ ke zhoršení (nebo můžeme říci, že nedošlo ke zlepšení). 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 4.3 Nejmenovaná úspěšná počítačová společnost si chtěla malým průzkumem ověřit, zda představení nového produktu přispělo ke zvýšení její prestižnosti. V průzkumu bylo osloveno náhodně 50 účastníků, kteří měli po představení nového výrobku odpovědět na otázku, zda *jejich důvěra v tuto společnost je větší než před představením*. Kladných odpovědí (ANO) bylo 32, NE odpovědělo zbylých 18 dotázaných. Lze se domnívat, že představený výrobek přispívá ke zvýšení důvěryhodnosti společnosti?

Odpověď na tuto otázku dá test hypotézy
$H_0 : p = 0.5$ (představení nemělo vliv)

proti alternativě

$H_1 : p > 0.5$ (představení zvýšilo důvěru)

Za platnosti H_0 má počet kladných odpovědí Z binomické rozdělení, $Z \sim Bi(50, 0.5)$.

$$P(Z \geq 32) = \frac{1}{2^{50}} \sum_{k=32}^{50} \binom{50}{k} = \frac{1}{2^{50}} \sum_{k=32}^{50} \binom{50}{50-k} =$$
$$= \frac{1}{2^{50}} \left[\binom{50}{18} + \binom{50}{17} + \cdots + \binom{50}{0} \right] \cong 0.0325$$

a tedy nulovou hypotézu zamítáme, tzn. je důvod věřit, že představení nového produktu zvýšilo důvěryhodnost počítačové společnosti.

Pokud bychom užili asymptotickou statistiku (57), dostaneme

$$u = \frac{2z - n}{\sqrt{n}} = \frac{2 \cdot 32 - 50}{\sqrt{50}} = 1.98.$$

Pravděpodobnost $P(U \geq 1.98) \cong 0.0239$, je o něco menší než přesná hodnota spočítaná z binomického rozdělení $Bi(50,\ 0.5)$, ale opět i v tomto případě zamítáme nulovou hypotézu na hladině významnosti $\alpha = 0.05$. Rozdíl mezi $P(Z \geq 32) \cong 0.0325$ a $P(U \geq 1.98) \cong 0.0239$, tj. přibližně 8.6×10^{-3} je způsoben malým rozsahem výběru ($n = 50$). Při větších hodnotách n se rozdíly snižují, , při menších naopak zvyšují, jak ukazuje následující tabulka.

n	z	z/n	$P(Z \geq z)$	u	$P(U \geq u)$
25	16	16/25	0.1148	1.4	0.0808
50	32	16/25	0.0325	1.98	0.0239
100	64	16/25	0.0033	2.8	0.0026

V tabulce také vidíme, jak s rostoucím rozsahem výběru roste síla testu, a naopak. Při stejné relativní četnosti kladných odpovědí 16/25 pro $n = 25$ nulovou hypotézu nezamítáme, pro $n = 50$ a $n = 100$ už na hladině významnosti $\alpha = 0.05$ nulovou hypotézu zamítneme.

4.4 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 $\left|Y_{(1)}\right| \leq \left|Y_{(2)}\right| \leq \ldots \leq \left|Y_{(n)}\right|$. Nechť R_i^+ je pořadí hodnoty $\left|Y_{(i)}\right|$ 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 kladných a záporných pořadí je $S = S^+ + S^- = 1 + 2 + \ldots + n = \dfrac{n(n+1)}{2}$ a nulovou hypotézu zamítneme, jestliže se hodnoty S^+, S^- podstatně liší, tzn. pokud je $\min(S^+, S^-)$ menší nebo rovno kritické hodnotě $w_n(\alpha)$. Tyto hodnoty jsou pro menší výběry n tabelovány (například v tabulce 8.5), přičemž jsou spočítány kombinatoricky s využitím klasické pravděpodobnosti.

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

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

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

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

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

Příklad 4.4 V rámci pokusu zdravotní organizace bylo testováno, zda je srdeční rytmus populace ČR v rámci doporučených mezí (pro tento příklad uvažujeme ideální tep 75 úderů za minutu). Bylo náhodně osloveno deset osob, kterým byl za patřičných podmínek změřen srdeční tep. Na základě měření byly získány následující výsledky: 68, 72, 94, 80, 71, 67, 97, 89, 81, 66.

Naším úkolem je testovat hypotézu $H_0 : \tilde{X} = 75$ (úderů) proti alternativě $H_1 : \tilde{X} \neq 75$, tedy rozhodnout, zda nám naše pozorování poskytuje důvod odmítnout představu, že polovina osob v populaci má tep nižší a polovina vyšší než doporučených 75 úderů za minutu.

X_i	68	72	94	80	71	67	97	89	81	66
$Y_i = X_i - 75$	-7	-3	19	5	-4	-8	22	14	6	-9

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

pořadí	1	2	**3**	**4**	5	6	7	**8**	**9**	**10**
$Y_i = X_i - 75$	-3	-4	**5**	**6**	-7	-8	-9	**14**	**19**	**22**

Kladné hodnoty Y_i jsou zvýrazněny.

Potom
$S^+ = 3 + 4 + 8 + 9 + 10 = 34,$
$S^- = S - S^+ = 1 + 2 + 5 + 6 + 7 = 21,$
$\min(S^+, S^-) = 21.$

Kritická hodnota v tabulce je $w_{10}(0.05) = 8$, tzn. že $H_0 : \tilde{X} = 75$ úderů nemůžeme zamítnout.

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

$$E(S^+) = \frac{n(n+1)}{4} = \frac{10 \cdot 11}{4} = 27,5$$
$$var(S^+) = \frac{n(n+1)(2n+1)}{24} = \frac{10 \cdot 11 \cdot 21}{24} = \frac{385}{24} = 96,25$$
$$U = \frac{S^+ - E(S^+)}{\sqrt{var(S^+)}} = \frac{34 - 27,5}{\sqrt{96,25}} \cong 0.66$$

Protože $|U| < 1.96,\ (u(0,975) = 1.96)$, viz tabulka normovaného normálního rozdělení 8.1, nemohli bychom zamítnout nulovou hypotézu na hladině významnosti $\alpha = 0.05$ ani tímto asymptotickým postupem.

Kdybychom v tomto příkladu užili znaménkový test, nulovou hypotézu bychom zamítnout rovněž 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
3	$\frac{1}{2^{10}}\binom{10}{3} = \frac{120}{1024}$	0,1719
4	$\frac{1}{2^{10}}\binom{10}{4} = \frac{210}{1024}$	0,3770
5	$\frac{1}{2^{10}}\binom{10}{5} = \frac{252}{1024}$	0,6231

Hodnota $k_1 = 1$, počet kladných odchylek je roven 5, tedy větší než k_1 a nulovou hypotézu zamítnout nemůžeme.

Všimněme si, že $P(Z \leq 5) = 0.6231$, tzn. větší než $\alpha = 0.05$. Znaménkový test by na této hladině významnosti nezamítl $H_0 : \tilde{X} = 75$ úderů ani proti jednostranné alternativě $H_1 : \tilde{X} < 75$.

Poznámka 4.3

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 nápovědu programu a pokud je hodnota ve výstupu programu jen asymptotická, použijte kritické hodnoty ze statistických tabulek.

4.5 Dvouvýběrový Wilcoxonův test

Poznámka 4.4

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 máme k dispozici dva nezávislé výběry. V případě, kdy je normalita dat porušena, je potřeba použít neparametrickou alternativu. Jendou z možností je Wilcoxonův test, který je založen na pořadí pozorování.

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 (tvaru) distribučních funkcí:
$H_0 : F = G$
$H_1 : F \neq G$

Ale většinou alternativu chápeme jako posunutí Δ, tj. $H_1 : G(x) = F(x - \Delta), \quad \Delta \neq 0$, pro které je tento test citlivý (má přijatelnou sílu). Pokud se distribuční funkce

F a G liší spíše jen rozptylem nebo tvarem, není užití dvouvýběrového Wilcoxonova testu vhodné.

Obrázek 7: Tvary rozdělení populací (tento obrázek je rovněž v příloze).

Wilcoxonův dvouvýběrový test je založen na pořadí pozorovaných hodnot v tzv. sdruženém výběru. Všech $m + n$ hodnot z obou výběrů $X_1, X_2, \ldots, X_m$ a $Y_1, Y_2, \ldots, Y_n$ 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.

$$E(T_1) = \frac{1}{2}m(m+n+1) \quad \text{a} \quad E(T_2) = \frac{1}{2}n(m+n+1).$$

Lze dokázat, že

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

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

$$\frac{T_1 - E(T_1)}{\sqrt{var(T_1)}},$$

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 8.6.

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

$$\frac{U_1 - E(U_1)}{\sqrt{var(U_1)}},$$

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

Příklad 4.5 Na 29 počítačích se stejnou konfigurací a nastavením bylo testováno, kolik chybných paketů bylo přeneseno počítačovou sítí. Na 14 počítačích byl použitý opravný mechanismus $(24, 35, 41, 45, 46, 65, 67, 73, 88, 100, 127, 141, 171, 199)$, zbylých 15 počítačů bylo bez opatření $(57, 66, 68, 87, 96, 113, 128, 133, 143, 146, 149, 156, 158, 159, 176)$. Počet chybných přenesených paketů je označen X_i pro PC bez opatření a Y_i pro PC s opravným mechanismem.

Chceme zjistit, zda má opravné opatření vliv na zvýšení kvality (snížení počtu chyb) přenosu dat sítě. Seřadíme hodnoty sdruženého výběru (X_i a Y_i) vzestupně:

pořadí	oprava	chyby	pořadí_opr	pořadí	oprava	chyby	pořadí_opr
1	s	24	1	16	bez	113	
2	s	35	2	17	s	127	17
3	s	41	3	18	bez	128	
4	s	45	4	19	bez	133	
5	s	46	5	20	s	141	20
6	bez	57		21	bez	143	
7	s	65	7	22	bez	146	
8	bez	66		23	bez	149	
9	s	67	9	24	bez	156	
10	bez	68		25	bez	158	
11	s	73	11	26	bez	159	
12	bez	87		27	s	171	27
13	s	88	13	28	bez	176	
14	bez	96		29	s	199	29
15	s	100	15	pořadí T1=			163

$$U_1 = mn + \frac{1}{2}m(m+1) - T_1 = 14 \cdot 15 + \frac{1}{2}14 \cdot 15 - 163 = 152,$$
$$U_2 = mn - U_1 = 210 - 152 = 58,$$
$$\min(U_1, U_2) = 58.$$

Jelikož kritická hodnota pro $\alpha = 0.05$ je 59, znamená to, $\min(U_1, U_2) = 58$ je v kritickém oboru, a proto zamítáme na hladině významnosti $\alpha = 0.05$ nulovou hypotézu. Způsob opravy chybných paketů má vliv na kvalitu přenosu dat.

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	57	66	68	87	96	113	128	133	143	146	149	156	158	159	176
Y_i	24	35	41	45	46	65	67	73	88	100	127	141	171	199	

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

počet hodnot $X_i > 24$	15
počet hodnot $X_i > 35$	15
počet hodnot $X_i > 41$	15
počet hodnot $X_i > 45$	15
počet hodnot $X_i > 46$	15
počet hodnot $X_i > 65$	14
počet hodnot $X_i > 67$	13
počet hodnot $X_i > 73$	12
počet hodnot $X_i > 88$	11
počet hodnot $X_i > 100$	10
počet hodnot $X_i > 127$	9
počet hodnot $X_i > 141$	7
počet hodnot $X_i > 171$	1
počet hodnot $Y_i > 199$	0
	U_1 =152

$U_2 = m \cdot n - U_1 = 210 - 152 = 58$, $\min(U_1, U_2) = 58$ a výpočet testové statistiky je hotov.

4.6 Kruskalův-Wallisův test

Další z metod, které jsou určené pro data nemající normální rozložení, je test s názvem Kruskal-Wallis.

Poznámka 4.5

Kruskalův-Wallisův test je neparametrickou obdobou analýzy rozptylu s jednoduchým tříděním (*one-way ANOVA*). Přičemž jednoduché třídění znamená, že sledujeme závislost spojité veličiny pouze na jedné diskrétní veličině. Jedná se o 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. $i = 1,\ 2, \ldots, I$. Cílem Kruskal-Wallisova testu je testovat hypotézu, že všechny distribuční funkce rozdělení, z nichž jsou výběry, jsou shodné:

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

proti alternativě, že alespoň jedna dvojice distribučních funkcí se 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 spočtou se 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

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

a testová statistika Q pro test nulové hypotézy je založena na součtu čtverců odchylek pozorovaných hodnot součtů pořadí (T_i) od jejich středních hodnot $(E(T_i))$:

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

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

Příklad 4.6 Rychlost tří databázových jazyků, při spouštění dotazu nad stejnou databází je naměřená v sekundách:

Jazyk	čas (s)					
A	37	22	17	14	18	
B	48	23	23	45		
C	66	63	31	44	30	53

Chceme testovat, zda je rychlost dotazovacích jazyků ze stejného rozdělení. Nejdříve spočítáme součty pořadí v jednotlivých výběrech.

Jazyk	n_i	Pořadí						T_i
A	5	9	4	2	1	3		19
B	4	12	5	6	11			34
C	6	15	14	8	10	7	13	67
	15							

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{n_i} \left[T_i - \frac{1}{2} n_i (n+1) \right]^2 = \frac{12}{15 \cdot 16} \left(\frac{441}{5} + \frac{4}{4} + \frac{361}{6} \right) = 7.4683$$

Hodnota $\chi_2(0.95) = 5.9915$, tedy $Q = 7.4683$ je v kritickém oboru a nulovou hypotézu zamítáme.

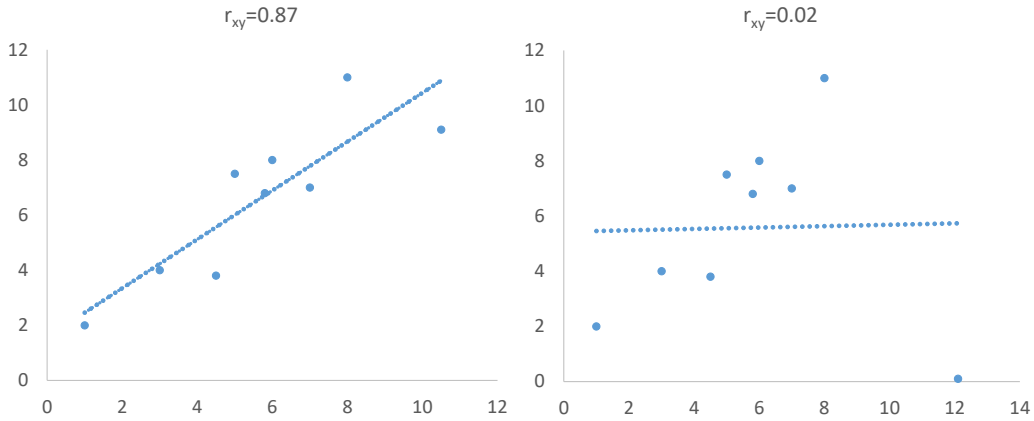
P-value odpovídající hodnotě statistiky $Q = 7.4683$, tj. $P(X \geq 7.4683)$, když $X \sim \chi^2_2$, je $p = 0.02389$. Vidíme tedy, že pro tak malé rozsahy výběrů se dosti liší od hodnoty p, získané z asymptotického rozdělení statistiky Q. Nicméně v tomto případě oba výsledky vedou k zamítnutí nulové hypotézy na hladině významnosti $\alpha = 0.05$.

4.7 Spearmanův koeficient pořadové korelace

Korelace je jednou z možností ohodnocení těsnosti lineárního vztahu mezi dvojicí veličin. Korelační koeficient ρ_{xy} nabývá hodnot z intervalu $\langle -1, 1 \rangle$. Pearsonův výběrový korelační koeficient r_{xy} lze vyjádřit jako

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} = \\ = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}} \tag{58}$$

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



Obrázek 8: Korelační koeficient (tento obrázek je rovněž v příloze).

Poznámka 4.6

Spearmanův koeficient korelace dostaneme tak, že místo původních hodnot X_i, Y_i dosadíme do vztahu (58) 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 (58) za hodnoty X_i, Y_i jejich pořadí R_i a Q_i, dostaneme Spearmanův koeficient pořadové korelace r_S:

$$r_S = \frac{\sum_{i=1}^{n} R_i Q_i - n\bar{R}\bar{Q}}{\sum_{i=1}^{n} R_i^2 - n\bar{R}^2} \tag{59}$$

Jelikož

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

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

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

můžeme vztah (59) 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

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \tag{60}$$

Poznámka 4.7

- 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 $\boldsymbol{X}$ a $\boldsymbol{Y}$. 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 $\boldsymbol{X}$ a $\boldsymbol{Y}$ 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 4.7 Dva uživatelé hodnotili 10 programovacích jazyků podle oblíbenosti. Jazyky pro jednoduchost označíme písmeny abecedy A, B, C, D, E, F, G, H, I, J. Hodnocení uživatelů shrnuje následující tabulka:

Uživatel	Uspořádání									
U_1	B	E	G	A	D	F	H	J	C	I
U_2	F	J	H	I	C	B	D	A	E	G

Ohodnoťte shodu degustátorů. Určíme hodnoty pořadí R_i, Q_i:

jazyk	R_i	Q_i	d_i	d_i^2
B	1	6	-5	25
E	2	8	-6	36
G	3	7	-4	16
A	4	10	-6	36
D	5	9	-4	16
F	6	1	5	25
H	7	5	2	4
J	8	4	4	16
C	9	2	7	49
I	10	3	7	49
$\sum$				272

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 272}{10 \cdot (10^2-1)} \cong -0.6485$$

V tabulce kritických hodnot Spearmanova koeficientu korelace nalezneme, že kritická hodnota pro $\alpha = 0.05$ je 0.6364. Na této hladině významnosti tedy zamítneme nulovou hypotézu, že hodnocení programovacích jazyků dvěma uživateli nejsou korelované. Jinými slovy zamítáme hypotézu, že jsou jazyky u vybraných uživatelů stejně oblíbené, naopak záporný koeficient naznačuje opačné uspořádání.

Shrnutí:

- Neparametrické metody jsou vhodné pro data nesplňující normální rozdělení.
- Test dobré shody zjišťuje shodu empirického rozdělení s „tabulkovým“.
- Test nezávislosti umožňuje testovat závislost dvou kategorických veličin.
- Znaménkový jednovýběrový test je založen na binomickém rozdělení.
- Wilcoxonův a Kruskal-Wallisův test užívají k testování hypotézy uspořádání hodnot.
- Spearmanův korelační koeficient je vhodný pro porovnání dvou veličin s odlehlými hodnotami.

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 úkol:

Korespondenční úlohy budou zadávány vždy na začátku semestru.

5 Programové prostředky pro statistické výpočty

Průvodce studiem:

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 MS Excel a statistický paket JASP, 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.

Cíl: Po prostudování této části kapitoly byste měli umět:

- orientovat ve statistických funkcích MS Excel a programu JASP,
- vybírat vhodné části výstupů analýzy,
- interpretovat výsledné analýzy do jednoznačných závěrů.

Podpora statistického zpracování dat je součástí mnoha obecných programových systémů orientovaných na práci s databázemi, na grafické zpracování dat, matematických programových prostředků (Matlab, Mathematica, aj.) a kromě toho existuje celá řada specializovaných statistických programových paketů. Společným rysem těchto statistických programových prostředků jsou operace s datovou maticí, tj. dvojrozměrnou tabulkou, ve které sloupce jsou veličiny a řádky představují pozorované objekty. Pro práci s tabulkami jsou určeny i tabulkové procesory (např. MS 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.

5.1 Tabulkový procesor MS Excel

MS Excel je typickým představitelem tabulkových procesorů, některá jeho verze je dostupná prakticky na každém počítači. Standardní součástí MS 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 MS Excel rozšířit o standardně dodávaný doplněk *Analýza dat*, který pokrývá prakticky všechny metody vysvětlované v základních kursech statistické analýzy dat, vyjímaje neparametrické verse testů. Vzhledem k tomu, že MS 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.

Poznámka 5.1

Dosti obecně lze říci, že zejména v české verzi MS Excel 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:*

Příklad 5.1

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

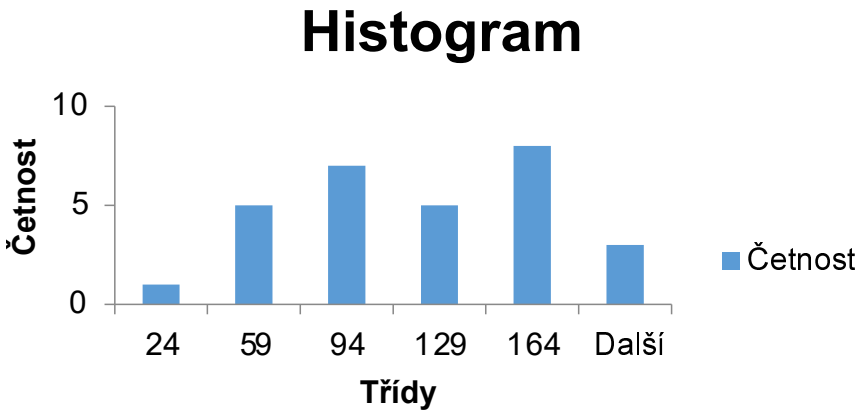
Stř. hodnota je užita místo slova *Průměr*, Chyba stř. hodnoty místo *Směrodatná odchylka průměru*. Rozptyl výběru místo *Výběrový rozptyl*. Počet desetinných míst je nadbytečný.

Příklad 5.2 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*.

Příklad 5.3 Často užívaným modulem doplňku Analýzy dat je *Histogram*. S využitím implicitního nastavení vstupních parametrů můžete dostat následující obrázek:

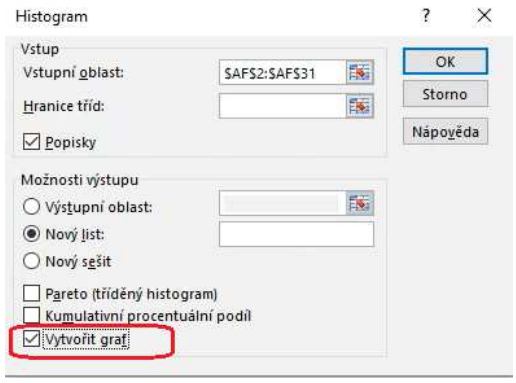


Legenda a nadpis „Histogram“ jsou zbytečné, jen zabírají místo, popis vodorovné osy neříká nic. Sloupce nejsou nad celou šířkou intervalů. To lze napravit vhodnější volbou vstupních parametrů nebo dodatečnou úpravou grafu. Závažnějším nedostat-

kem je, že hodnoty popisující středy sloupců (středy jednotlivých intervalů) nejsou hodnoty odpovídající středu, ale pravému okraji intervalu.

Poznámka 5.2

Nutno podotknout, že v nabídce tvorby histogramu je nutné zaškrtnout políčko „Vykreslit graf“:



Příklad 5.4 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 popisu v nápovědě se dočteme následující:

nápověda:

NORMDIST (funkce)

Vrátí normální rozdělení se zadanou střední hodnotou a směrodatnou odchylkou. Tato funkce má ve statistice velmi široké použití, včetně testování hypotéz.

Syntaxe

NORMDIST(x, střed_hodn, sm_odch, kumulativní)

Syntaxe funkce NORMDIST obsahuje následující argumenty:

X: Povinný argument. Jedná se o hodnotu, pro kterou chcete zjistit hodnotu rozdělení.

Střed_hodn: Povinný argument. Jedná se o aritmetickou střední hodnotu.

Sm_odch: Povinný argument. Jedná se o směrodatnou odchylku rozdělení.

Kumulativní: Povinný argument. Logická hodnota, která určuje typ použité funkce. Pokud je argument kumulativní nastaven na hodnotu TRUE, vrátí funkce

NORMDIST kumulativní distribuční funkci. V případě hodnoty FALSE vrátí hromadnou pravděpodobnostní funkci.

konec nápovědy.

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

Příklad 5.5 Pozor při užívání funkcí navracející hodnoty kvantilů běžných rozdělení. Funkce NORMINV s parametry p, μ, σ vrátí hodnotu příslušného kvantilu $x(p) = \sigma u(p) + \mu$, tedy např. NORMINV(0,238; 175; 7]) vrátí hodnotu 170,01.

Příklad 5.6

Poznámka 5.3

U jiných rozdělení je to však trochu odlišné. Pro určení kvantilů rozdělení χ^2 můžeme užít funkci CHIINV (nebo novější CHISQ.INV.RT), která má dva vstupní parametry. Chceme-li, aby funkce vrátila hodnotu p-kvantilu, musíme její parametry zadat jako $(1-p)$ a požadovaný 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 nesprávný 0,95-kvantil.

Příklad 5.7 Ještě o trochu komplikovanější to je u funkce TINV pro výpočet kvantilů studentova t-rozdělení. Pokud chceme, aby funkce TINV spočítala p-kvantil, musíme vstupní parametry zadat jako $(1 - 2 \cdot p$ 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.

Poznámka 5.4

Pokud užíváte pro statistické výpočty MS Excel, vždy velmi pečlivě zkoumejte, co vlastně vám ve výsledcích MS Excel poskytuje a výstupy z MS 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. Protože se však MS Excel stále postupně vyvíjí, doporučujeme v novějších versích tohoto balíku kontrolovat statistické výpočty se statistickými tabulkami, či ověřeným statistickým software.

5.2 Statistické programové systémy

Statistických programů šířených komerčně existuje značné 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.). Vedle těchto „placených“ aplikací ecistuje celá řada volně dostupných statistických software, které se mnohdy dokážou těm komerčním přinejmenším rovnat. Zmíníme některé oblíbené produkty jako MYSTAT, GRETL, JASP, či velmi universální a rozšířený jazyk R. Všechny statistické programy však mají zpravidla tyto základní funkce:

Poznámka 5.5

- import dat (vstup datové tabulky připravené v jiném programovém prostředku, třeba v MS 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 v prostředí dané aplikace,
- 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. Zástupcem příkazového ovládání je velmi oblíbený jazyk R, který ovšem obsahuje obrovskou podporu s nápovědou a řešeními, které i nezkušeným uživatelům umožňuje poměrně snadný průnik do ovládání této aplikace.

5.3 Volně dostupný program JASP

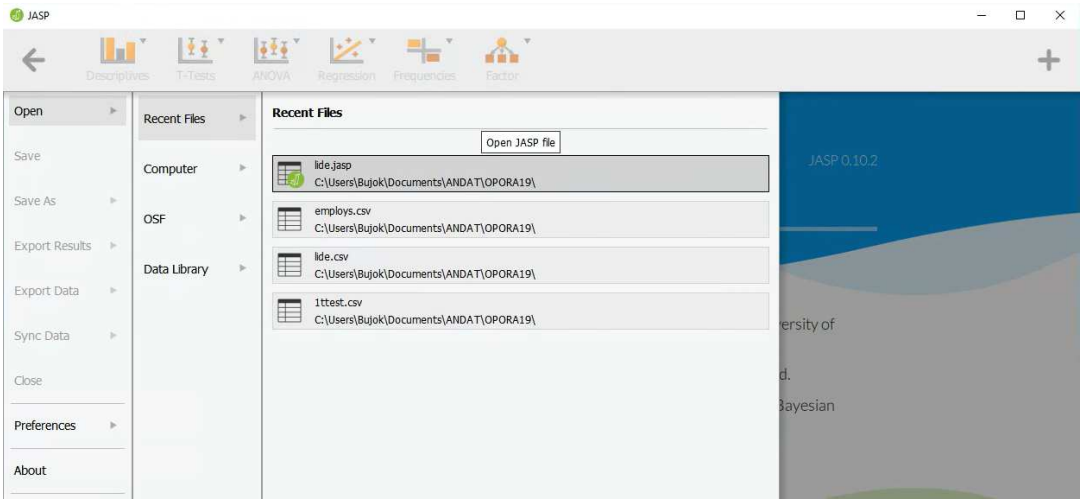
V rámci tohoto textu si ukážeme několik ukázek z programu JASP [1], který je volně k dispozici. Hlavním důvodem volby této aplikace je v prvé řadě nezávislost na licenčním software, a zejména pak poměrně přesná nabídka statistických funkcí pro potřeby tohoto kursu. JASP je poměrně universální statistický balík, doporučovaný zejména méně zkušeným uživatelům. Pokrývá však podstatnou část požadavků i popisné i inferenční statistické analýzy dat. Ovládá se pomocí výběru z jednoduchého menu. JASP oproti jiným podobným aplikacím sice vůbec nenabízí prvky transformace dat (je však propojen se zdrojovým souborem), oproti tomu však nabízí paralelně i několik statistických metod zpracovaných pomocí tzv. Bayesovské statistiky, což ocení zejména uživatelé se smyslem pro experimentování. Výsledky (textový i grafický výstup společně) jsou editovatelné a lze je ukládat přímo do formy jazyka LaTeX (text), což je vyspělé sazební prostředí pro tvorbu textů a EPS (grafika). Úvodní okno programu JASP je zobrazeno na obrázku 9. Na počátku lze pouze data načíst, přičemž máme na výběr ze čtyř možností, z nichž budeme nejčastěji užívat *Computer* (výběr souboru z PC).

JASP, stejně jako mnohé jiné statistické balíky, spolupracuje pouze s tabulkami ve formátu CSV. Po importu doporučujeme datovou matici uložit do formátu *.jasp*.

V nabídce nastavení aplikace (*Preferences*) části *Results* lze zvolit výpočet p-hodnot namísto úrovně významnosti - což lze pro začátek doporučit, a počet desetinných míst výsledných analýz. Základy ovládání jednoduché nabídky JASP ilustrují následující obrázky. Ačkoli JASP nenabízí příliš propracovanou nabídku transformace dat, tlačítkem filtru (vlevo nahoře matice dat) obdržíme možnost filtrování dat, což oceníme hlavně při analýze rozměrných dat (obrázek 10).

JASP umožňuje rovněž vytvářet nové proměnné (tlačítko „+“ vpravo nad datovou maticí), odvozené z existujících proměnných, či založené na některé z vestavěných

[1] https://jasp-stats.org/



Obrázek 9: JASP - úvodní okno

Obrázek 10: JASP - filtr dat

funkcí. I přesto je však pohodlnější proměnné přichystat v některém z tabulkových procesorů.

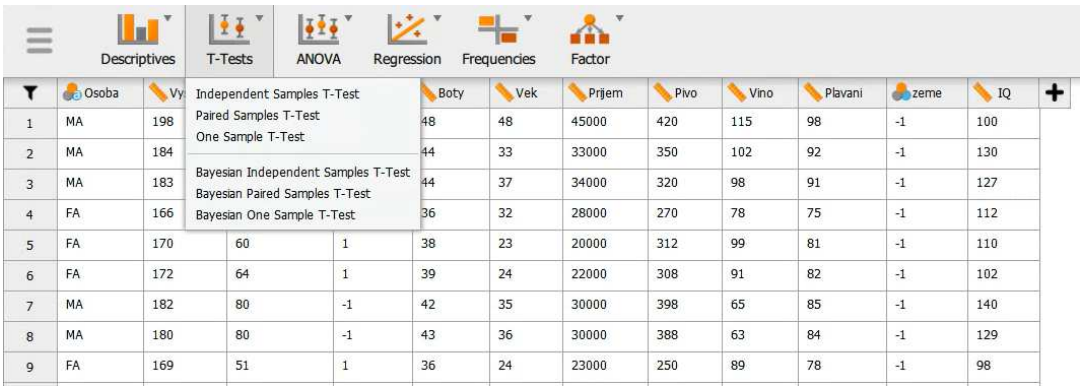
Příklad 5.8 Tabulka s datovou maticí se liší od MS Excelu v tom, že názvy veličin jsou v názvech sloupců a na veličiny např. při zadávání vstupních parametrů výpočtu do šablony se odkazujeme pomocí jejich jmen. Proměnné jsou navíc, podobně jako např. v SPSS, označeny symbolem nominální, ordinální nebo spojité škály, pro větší přehled (obrázek 11).

Příklad 5.9 Požadované výpočty se zadávají volbou z přehledného menu (obrázek 12), např. zde z položky *T-Tests* rozbalíme podskupiny implementovaných statistických metod (současně navíc v provedení Bayesovské statistiky):

Příklad 5.10 Vyplněním nabídky konkrétní metody se vstupními parametry výpočtu je možné specifikovat i úroveň podrobnosti. Nastavení metod aplikace JASP je

	Osoba	Vyska	Hmotnost	Vlasy	Boty	Vek	Prijem	Pivo	Vino	Plavani	zeme	IQ
1	MA	198	92	-1	48	48	45000	420	115	98	-1	100
2	MA	184	84	-1	44	33	33000	350	102	92	-1	130
3	MA	183	83	-1	44	37	34000	320	98	91	-1	127
4	FA	166	47	-1	36	32	28000	270	78	75	-1	
5	FA	170	60	1	38	23	20000	312	99	81	-1	110
6	FA	172	64	1	39	24	22000	308	91	82	-1	102
7	MA	182	80	-1	42	35	30000	398	65	85	-1	140
8	MA	180	80	-1	43	36	30000	388	63	84	-1	129
9	FA	169	51	1	36	24	23000	250	89	78	-1	98
10	FA	168	52	1	37	27	23500	260	86	78	-1	100
11	MA	183	81	-1	42	37	35000	345	45	90	-1	105

Obrázek 11: JASP - matice dat



	Osoba	Vy			Boty	Vek	Prijem	Pivo	Vino	Plavani	zeme	IQ
1	MA	198			48	48	45000	420	115	98	-1	100
2	MA	184			44	33	33000	350	102	92	-1	130
3	MA	183			44	37	34000	320	98	91	-1	127
4	FA	166			36	32	28000	270	78	75	-1	112
5	FA	170	60	1	38	23	20000	312	99	81	-1	110
6	FA	172	64	1	39	24	22000	308	91	82	-1	102
7	MA	182	80	-1	42	35	30000	398	65	85	-1	140
8	MA	180	80	-1	43	36	30000	388	63	84	-1	129
9	FA	169	51	1	36	24	23000	250	89	78	-1	98

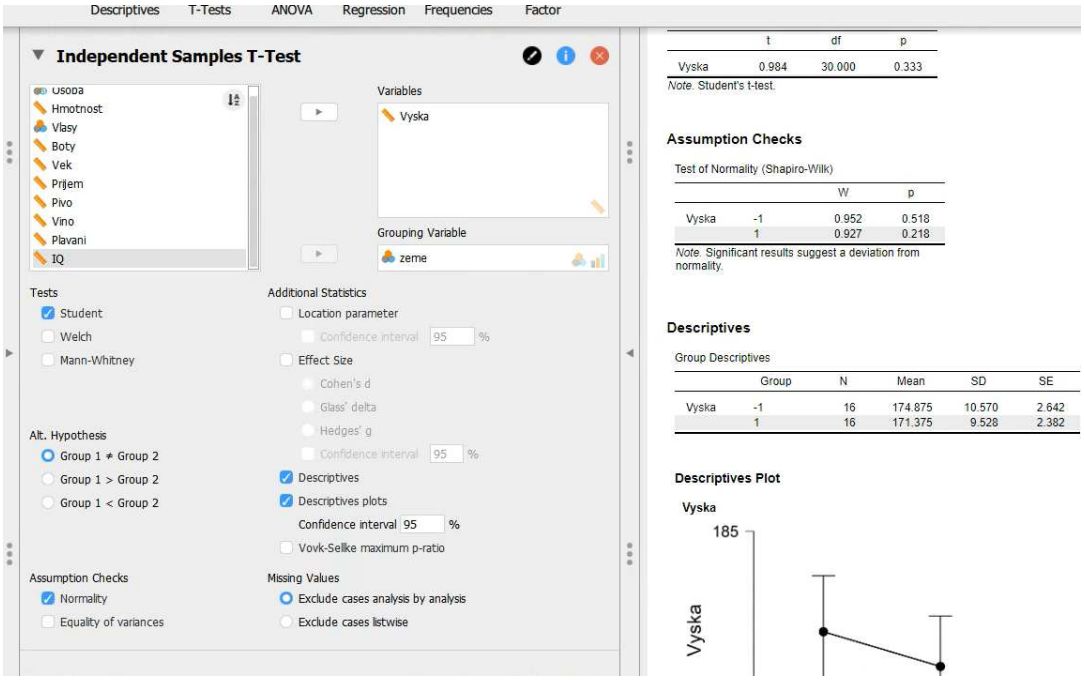
Obrázek 12: JASP - jednoduché menu

intuitivní, což ocení začínající statistici (obrázek 13). Výsledky analýzu jsou pak v okně aktuálního výstupu.

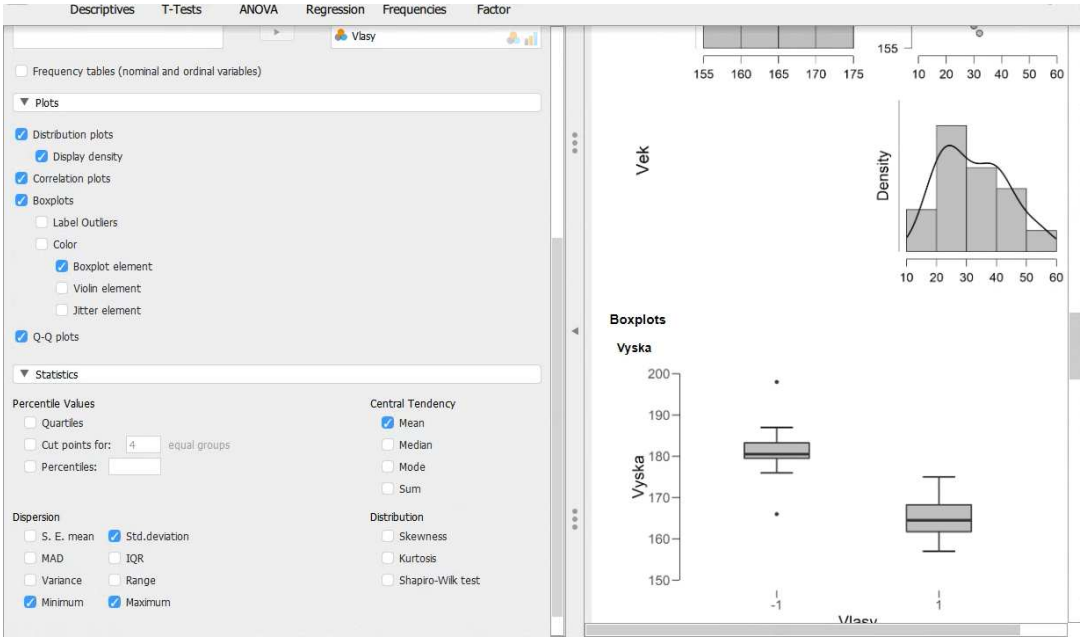
Příklad 5.11 Oproti některým obsáhlejším komerčním produktům JASP nenabízí samostatnou nabídku pro tvorbu grafů. Ovšem v menu *Descriptives* - *Descriptive Statistics* lze kromě základních popisných statistik rovněž docílit vykreslení běžných statistických grafů (oobrázek 14):

Příklad 5.12 JASP, podobně jako například SPSS, má výstupy analýzy pouze pro prohlížení. Pro uložení a editaci výstupů je možné data (či obrázky) přenést buď jako obyčejný text do „klasických editorů“ nebo jako upravený text do formátu LaTeX. Do výsledků v prostředí JASP lze psát alespoň drobné poznámky (obrázek 15).

Poznámka 5.6

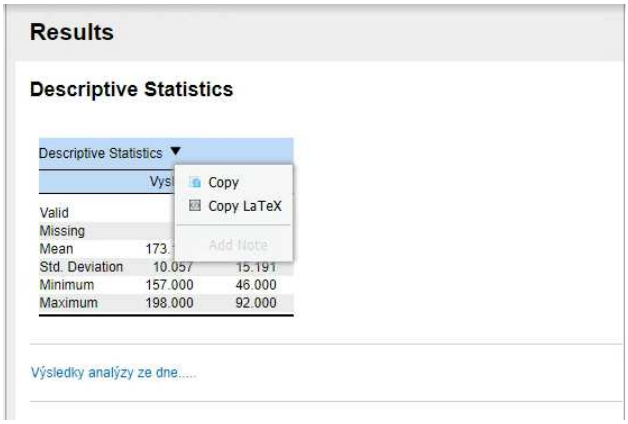


Obrázek 13: JASP - nastavení testu



Obrázek 14: JASP - deskriptivní statistika a grafy

Přestože JASP 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.



Obrázek 15: JASP - kopírování výsledné analýzy

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). Zejména v JASP si dejte při importu pozor na oddělovače tisíců. 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.

Shrnutí:

- Programové prostředky pro analýzu dat.
- Import a export dat.
- Data ve formátu statistické aplikace.
- Filtrování a transformace dat.
- Nastavení (výstupu) statistického balíku.
- Převod výstupů do patřičného textového editoru.

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 MS 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 MS Excelu a jejich nedostatky,
- doplněk MS Excelu Analýza dat.

6 Prezentace výsledků analýzy dat

Průvodce studiem:

V této kapitole budou ukázána některá doporučení, jak prezentovat výsledky statistické analýzy. Část těchto doporučení vychází z knihy van Belle (2002). Proto jsou části převzatých příkladů ponechány v angličtině. Příklad tří způsobů prezentace téhož jednoduchého výsledku ukazuje, že na formě prezentace výsledků záleží.

Cíl: Po prostudování této části kapitoly byste měli umět:

- rozlišovat požadavky na přesnost numerických výsledků,
- umět vhodně a věcně vytvářet popis analýzy,
- oprostit se mechanického postupování při zpracování dat.

Příklad 6.1

- 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.

6.1 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:

1. 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.

2. Autor 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 software, kde je tabulka četností 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.

Příklad 6.2

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 1 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

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

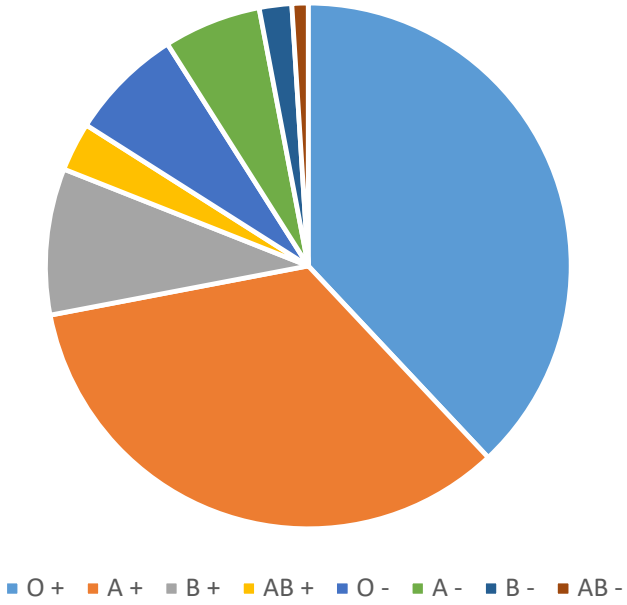
Neméně důležité je zaměřit se na rozumný počet významných číslic. Pokud číselná hodnota je větší než 100, většinou stačí ji uvést jako celé číslo, tj. bez desetinných míst. Hodnoty ve sloupci mají být vhodně zarovnány (celá číslice vpravo, desetinná na desetinnou čárku nebo tečku). Zejména v tabulkách je nutné brát ohled na tzv. „efektivní číslice“. To jsou 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 350 000 a uvádět výsledný rozdíl (tedy 4 691, 7 234 a 6 991). V tabulkách ovšem mají být pokud možno nejvýše dvě až tři efektivní číslice, neboť více efektivních číslic člověk obtížně vnímá.

Všeobecná zásada, že grafy jsou lepší než číselné údaje, není vždy správná. Někdy je totiž 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, a navíc si čtenář 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. 16), pak závažnost argumentu musíme uznat.

Příklad 6.3

Z výsečového grafu na obr. 16 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.



Obrázek 16: Relativní četnosti (v %) krevních skupin a Rh faktoru v populaci USA

Další van Belleho doporučení v kontextu grafické prezentace výsledků je ***neužívat skládané sloupcové grafy.*** Skládané (kumulované, stackbar) sloupcové grafy jsou hůře čitelné než jednoduché sloupcové grafy a často lze najít efektivnější možnost, jak nahlédnout do struktury dat, jak to ilustruje následující příklad.

Příklad 6.4 Souhrnná zdrojová data z průzkumu počtu aktivit provozovaných seniory v průběhu dvou týdnů jsou uvedena v tabulce 4. Ve zprávě Státního centra pro zdravotní statistiku byly tyto údaje prezentovány formou skládaného sloupcového grafu (obr. 17), což ke vnímání jejich obsahu nijak nepřispělo, spíše naopak. Prezentace by měla usnadňovat odpovědi na následující jednoduché a přirozené otázky:

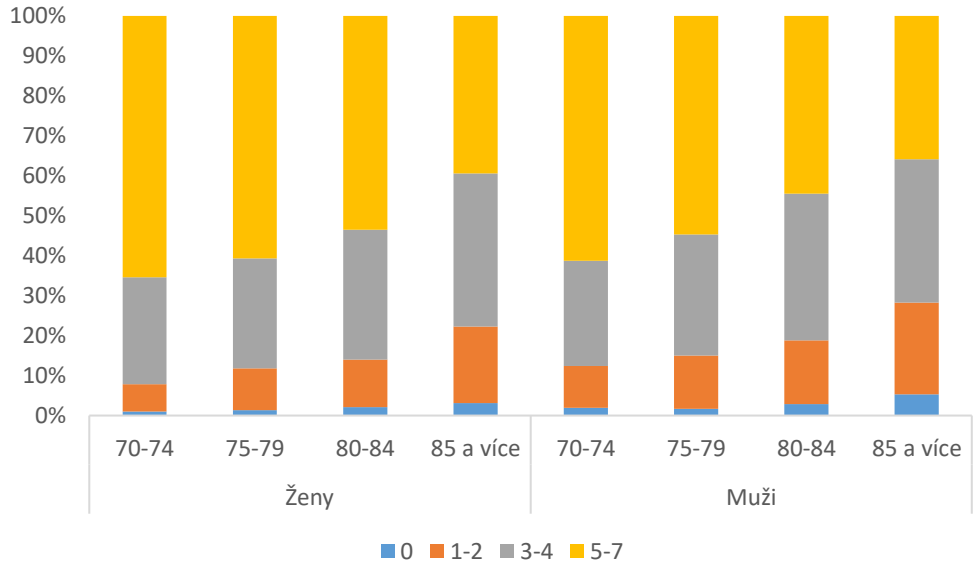
- Mají více aktivit muži nebo ženy?
- Jak mění počet aktivit s věkem?
- Liší se tyto změny u mužů a žen?

To ovšem spojovaný sloupcový graf na obrázku 17 rozhodně neusnadňuje.

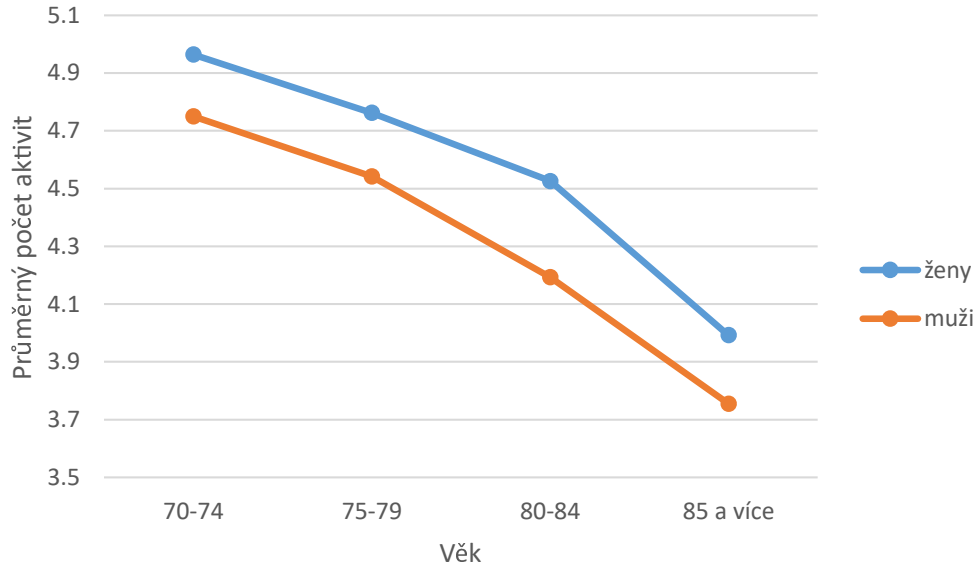
Přitom docela jednoduchý přepočet a grafické zobrazení průměrných hodnot aktivit pro muže a ženy podle věkových kategorií na obrázku 18 vypovídá jasně, že ženy jsou o trochu aktivnější, počet aktivit s věkem klesá a rychlost tohoto poklesu je u obou pohlaví zhruba stejná.

Tabulka 4: 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



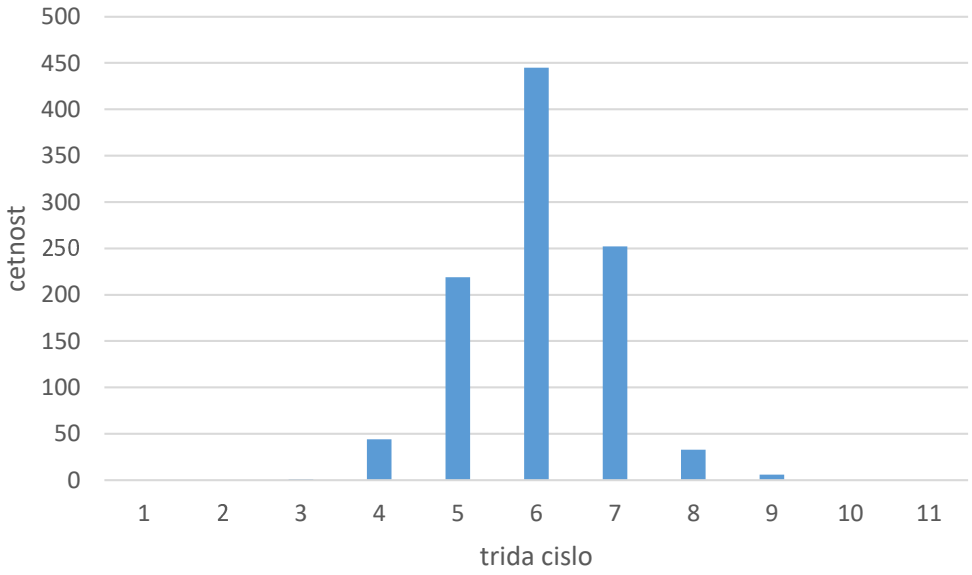
Obrázek 17: Počet aktivit v průběhu dvou týdnů - četnosti v % (Kramarov et al., zpráva National Center for Health Statistics, 1999).



Obrázek 18: Průměrný počet aktivit podle věku a pohlaví (tento obrázek je rovněž v příloze).

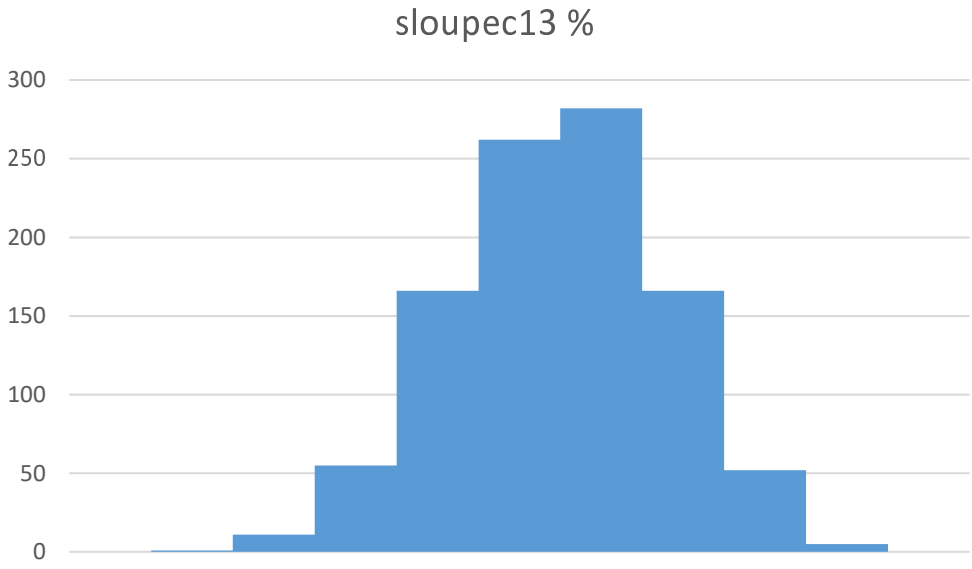
6.2 Jakým chybám se vyhnout?

V tomto odstavci je uvedeno několik typických chyb z korespondenčních úloh a semestrálních prací studentů v předmětu Analýza dat. Komentáře k chybám jsou pro větší přehled psány kurzivou.



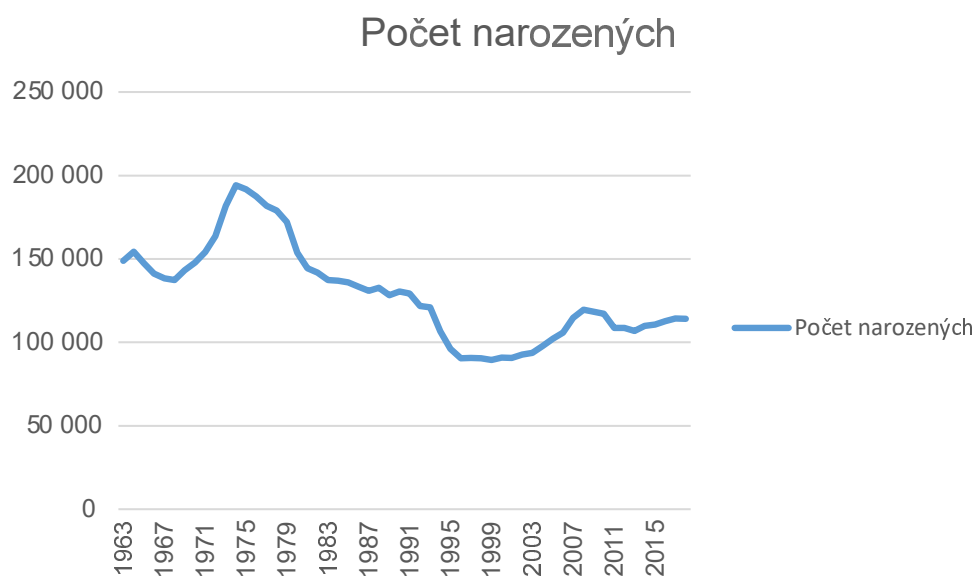
Obrázek 19: Histogram – častá chyba z naprosté nedbalosti.

Histogram na obr. 19 je prezentován tak, jak ho nabízí MS Excel, zdravý rozum si zřejmě 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.



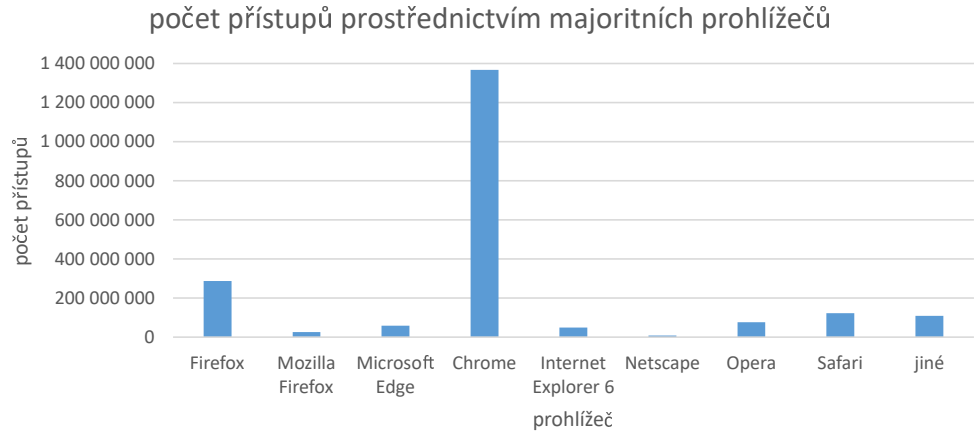
Obrázek 20: Histogram – další častá chyba způsobená nedbalostí.

V histogramu na obr. 20 chybí popis os, zbytečný je nic neříkající nadpis histogramu, opět nevhodně zvolené měřítko vodorovné osy.



Obrázek 21: Časový průběh počtu narozených.

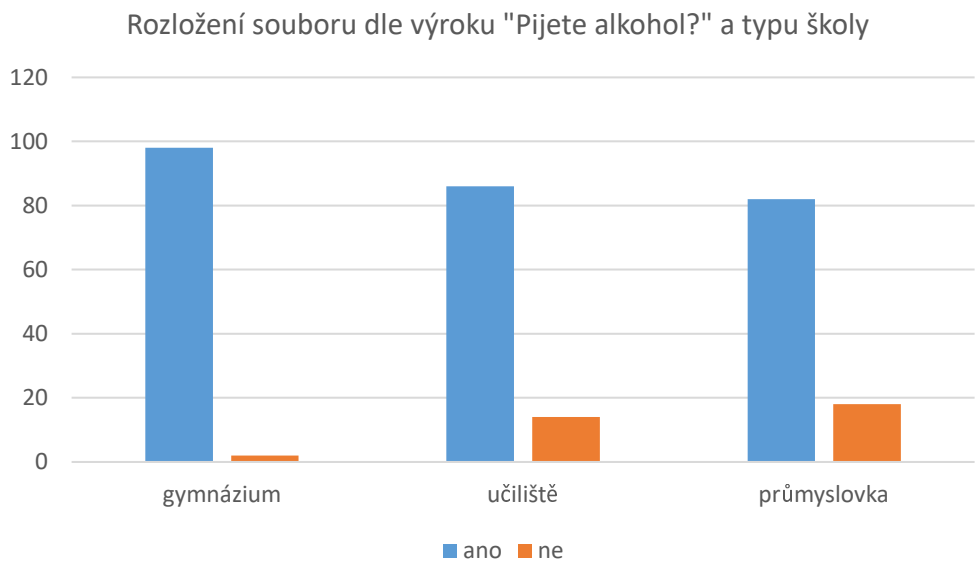
Na obr. 21 chybí popis os grafu, nevhodné jednotky na svislé ose (tři neefektivní nuly, počet narozených měl být v tisících), legenda je nadbytečná a zbytečně zabírá značnou část kreslící plochy, význam čáry nejasný (bylo užito nějaké vyhlazování?), časová řada by měla být nakreslena jako body, případně se spojnicemi.



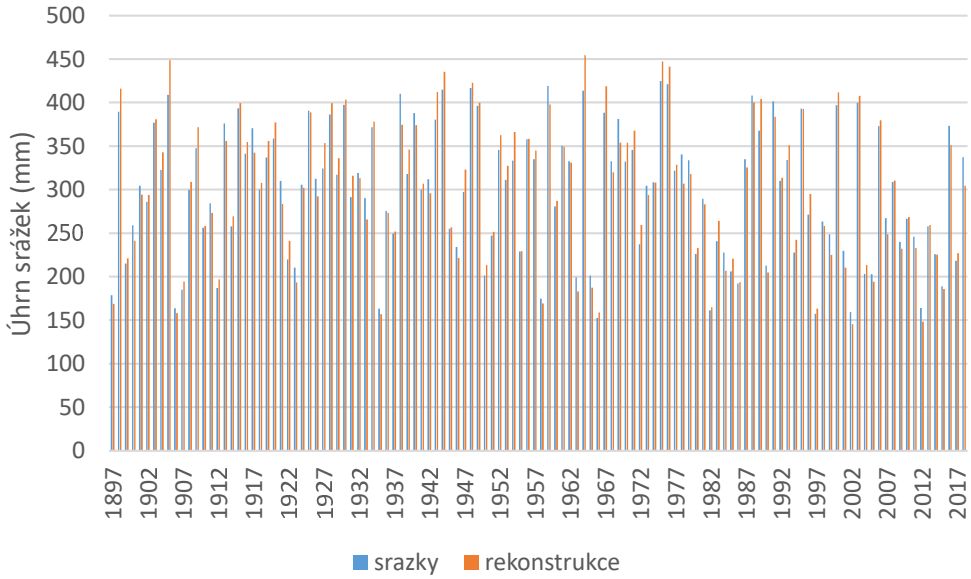
Obrázek 22: Nevhodný sloupcový graf.

Na obr. 22 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.

Na první pohled (pomineme-li neobratnou formulaci nadpisu) sloupcový graf na obr. 23 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 23: Další nesprávný sloupcový graf.



Obrázek 24: Nevhodně užitý typ grafu.

Na obr. 24 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:

$H_0 : \mu = 6$

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 - b1 x1)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.

7 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.

Bujok, P., Tvrdík J., Poláková R.: Základy pravděpodobnosti a statistiky, Ostravská universita, Ostrava, 2015

Opora ke stejnojmennému kursu, který předchází kursu Analýza dat.

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.

Goss-Sampson, M.: Statistical analysis in JASP: A guide for students, 2018

Přehledný manuál balíku JASP orientovaný zejména pro studenty.

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 PřF 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.

Pekár, S., Brabec, M.: Moderní analýza biologických dat 1, Scientia, 2009

Prakticky zaměřená publikace aplikované statistiky v prostředí jazyka R.

Pekár, S., Brabec, M.: Moderní analýza biologických dat 1, Scientia, 2012

Pokračování prakticky zaměřené publikace aplikované statistiky v prostředí jazyka R.

Řezanková H.: Analýza dat z dotazníkových šetření, 4. přepracované vydání, Professional publishing, 2017

Publikace zaměřená na analýzu dat z dotazníkového šetření, tedy zejména kategorických veličin, v prostředí SPSS.

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.

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.

8 Statistické tabulky

Statistické tabulky byly pořízeny s využitím statistických funkcí NORMSDIST, CHIINV, TINV, FINV programu Microsoft Excel 2016 pro Windows 10. Pokud jste u počítače, na kterém je nainstalován Excel nebo některý ze statistických programů statistické tabulky nepotřebujete, nebot' potřebné hodnoty distribučních funkcí či kvantilů snadno zjistíte pomocí těchto programových prostředků.

8.1 Distribuční funkce normovaného normálního rozdělení

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

	$\Phi(x)$				
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

8.2 Vybrané kvantily rozdělení Chí-kvadrát

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

	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

8.3 Vybrané kvantily Studentova t-rozdělení

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

	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

8.4 Vybrané kvantily Fisherova Snedecorova F-rozdělení

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

	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

8.5 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

8.6 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

8.7 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