PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA

doc. RNDr. Tomáš Mrkvička, Ph.D.

December 13, 2016

Chapter 1

Zpracování statistického materiálu

1.1 Rozložení četností a jejich znázornění

Definice 1.1 *Mějme soubor dat o rozsahu n:* $x_1, \ldots, x_n$. *Nechť a je minimální hodnota (popř. infimum) , b je maximální hodnota (popř. suprémum), tj.* $x_{min} = a$, $x_{max} = b$.

1. *Interval* $< a, b >$ *nazýváme* **variačním oborem**

2. *Rozdíl* $x = b - a$ *nazýváme* **variačním rozpětím**.

3. *Variační obor* $< a, b >$ *rozkládáme na menší části, nazývané* **třídy** *(popř. třídní intervaly).*

4. Šířkou (délkou) h *třídy příslušného třídního intervalu* $< a, b >$ *nazýváme číslo* $h = b_k - a_k$. *Číslo* $\frac{1}{2}\ (a_k + b_k)$ *nazýváme středem třídy, číslo* a_k *dolní hranicí uvažované třídy, číslo* b_k *horní hranicí uvažované třídy.*

5. *Hodnotu* x_k *argumentu X, která je zpravidla dána středem k-té třídy a zastupuje všechny hodnoty patřící do této třídy, nazýváme* **třídním znakem k-té třídy.**

Při rozkladu variačního oboru $< a, b >$ v třídy budeme dbát zpravidla těchto zásad:

1. Obsahuje-li soubor jen malý počet různých hodnot, volíme každou hodnotu x_k za samostatnou třídu. Pokud statistický soubor má značně velký počet různých

hodnot x_k (popř. je jich nekonečně mnoho), sdružujeme hodnoty argumentu v třídy. Přitom šířky tříd volíme obvykle stejně velké. Pro výpočet šířky h lze použít přibližného vzorce $h \approx \frac{8}{100} \cdot (b - a)$.
Při volbě počtu třídních intervalů se doporučuje, aby jich bylo 8 až 20. Záleží na rozsahu souboru a účelu statistické tabulky. Počet k třídních intervalů volíme např. $k \approx 3,3log(n)$ nebo $k \approx \sqrt{n}$. Dvě pozorování považujeme za ekvivalentní, jakmile padnou do téhož třídního intervalu.

2. Jestliže na hranici dvou sousedních tříd padne více hodnot argumentu, zařazujeme polovinu z nich do nižší třídy a druhou polovinu do třídy vyšší. Zbyla-li ještě jedna hodnota (toto odpovídá lichému počtu hodnot ležících na hranic), rozhodneme o její příslušnosti k dané třídě losem. Není vhodné zařazovat stereotypně takové hraniční hodnoty vždy do vyšší, popř. nižší třídy, neboť by se tím mohl zkreslit celkový obraz rozložení uvažovaného souboru ve prospěch vyšších, popř. nižších tříd.

4. Vyskytuje-li se v hraničních třídách velmi málo hodnot , je vhodné tyto třídy spojit se sousední třídou v třídu jedinou.

Definice 1.2 *Druhy četností:*

1. *Počet prvků souboru patřících do k-té třídy nazýváme* **absolutní četností** *argumentu v k-té třídě nebo absolutní třídní četností (stručně četností) k-té třídy. Značíme* f_k.

2. *Jeli* f_k *absolutní třídní četnost k-té třídy,* n *rozsah uvažovaného souboru, potom*

   a) $\frac{f_k}{n}$ *nazýváme* **relativní četností** *k-té třídy,*

   b) $100 * \frac{f_k}{n}$ *nazýváme* **procentní relativní četností** *k-té třídy.*

3. **Kumulativní (součtovou) absolutní četností** f_k *k-té třídy nazýváme součet všech četností* f_k *až do k-té třídy včetně, tj.*

$$F_k = \sum_{j=1}^{k} f_j.$$

4. Kumulativní relativní četností R_k *k-té třídy nazýváme součet*

$$R_k = \sum_{j=1}^{k} \frac{f_j}{n} = \frac{F_k}{n}.$$

Poznámka 1.1 *Pro četnosti platí některé vlastnosti (uvažujeme statistický soubor rozsahu n, který je rozdělen do r tříd)*

1.

$$\sum_{k=1}^{r} f_k = n$$

2.

$$F_r = n$$

3.

$$\sum_{k=1}^{r} \frac{f_k}{n} = 1$$

Definice 1.3 Tabulkou rozložení četností *daného statistického souboru nazýváme tabulku, v níž jsou uvedeny hodnoty s příslušnými absolutními, popř. relativními četnostmi.*

Příklad 1.1 *Na telefonní stanici zaznamenávali počet telefonních výzev za dobu 1 min. Během jedné hodiny bylo v určité denní době dosaženo těchto výsledků (v každém řádku jsou hodnoty získané během 10 minut):*

3,2,2,3,1,1,0,4,2,1
1,4,0,1,2,3,1,2,5,2
3,0,2,4,1,2,3,0,1,2
1,3,1,2,0,7,3,2,1,1
4,0,0,1,4,2,3,2,1,3
2,2,3,1,4,0,2,1,1,5.

Sestavte tabulku rozložení daného statistického souboru.

Počet telefonních výzev za 1 min	Absolutní četnost	Relativní četnost
0	8	0.133
1	17	0.283
2	16	0.266
3	10	0.166
4	6	0,1
5	2	0,033
7	1	0,016
Celkem	60	1

Table 1.1: **Tabulka rozložení četností**

Argument statistického souboru představuje náhodnou veličinu X. Ze zákona velkých čísel plyne, že relativní četnost $\frac{f_k}{n}$ udává (přibližně) pravděpodobnost, že X padne do k-té třídy, takže platí $p_k = P(a_k \leq X \leq b_k) \approx \frac{f_k}{n}$, přičemž interval $\langle a_k, b_k \rangle$ je k-tou třídou.

Definice 1.4 *Typy znázornění absolutních či relativních četností:*

1. Histogram *rozložení absolutních (relativních) četností sestavíme tak, že na*

osu x vyneseme středy jednotlivých tříd a nad každou úsečkou zobrazující určitou třídu (šířky h) sestrojíme pravoúhelník s výškou rovnou příslušné absolutní četnosti f_k, popř. relativní četnosti $\frac{f_k}{n}$. Horní obraz pravoúhelníka představuje histogram rozložení četností. Histogram relativních četností aproximuje hustotu rozdělení náhodné veličiny X.

2. Úsečkový diagram*(nebo graf) rozložení absolutních (relativních) četností dostaneme, jestliže na ose x zobrazíme středy jednotlivých tříd a v každém z nich sestrojíme ve směru osy y úsečku o délce rovné příslušné absolutní četnosti f_k, popř. relativní četnosti $\frac{f_k}{n}$.*

3. Polygon *rozložení četností (spojnicový diagram) dostaneme, jestliže koncové body úsečkového diagramu rozložení četnosti spojíme úsečkami a vytvoříme tak lomenou čáru, která pak představuje hledaný polygon neboli spojnicový diagram.*

4. *Graf, polygon nebo histogram kumulativních četností dostaneme analogicky jako v bodech 1,2 a 3.*

5. Ogivní křivku *(stručně* **ogivu***) dostaneme, sestrojíme-li polygon kumulativních relativních četností. Ogiva aproximuje graf distribuční funkce uvažované náhodné veličiny X.*

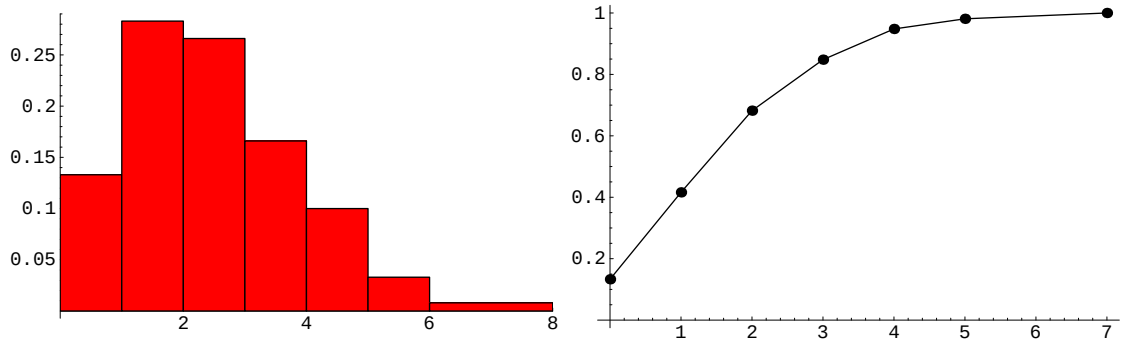

Figure 1.1: Histogram a ogiva dat z Příkladu 1.1

1.2 Charakteristiky polohy

Definice 1.5 *Nechť je dán statistický soubor o hodnotách $x_1, x_2, ..., x_n$, které jsou popř. roztříděny do r tříd, přičemž f_k značí absolutní četnost k-té třídy.*

1. Aritmetický průměr $\bar{X}$ *je definován vztahy*

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\sum_{i=1}^{r} f_i x_i. \tag{1.1}$$

2. Geometrický průměr $\bar{X}_g$ *je definován vztahem*

$$\bar{X}_g = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_1} \tag{1.2}$$

3. Harmonický průměr $\bar{X}_h$ *je definován vztahy*

$$\bar{X}_h = \frac{1}{A}, kdeA = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{x_k} = \frac{1}{n}\sum_{i=1}^{r} \frac{f_i}{x_i}. \tag{1.3}$$

Ve vztazích 1.1, 1.3 jsou uvedeny dva tvary. První tvar odpovídá souboru neroztříděné a druhy tvar roztříděnému.

Věta 1.1

$$\bar{X}_h \leq \bar{X}_g \leq \bar{X}.$$

Nechť je dán statistický soubor o hodnotách $x_1, x_2, ..., x_n$. Setřídíme-li hodnoty podle velikosti dostaneme tzv. setříděný statistický soubor

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)},$$

kde $X_{(1)}$ označuje nejmenší hodnotu, $X_{(2)}$ označuje druhou nejmenší hodnotu, ... Obecně $X_{(i)}$ označuje i-tou pořadovou hodnotu.

Definice 1.6 Medián *je určen dvěma způsoby, v závislosti na počtu prvků statistického souboru. V případě lichého počtu hodnot vezmeme za medián* $\tilde{x}$ *prostřední hodnotu*

$$\tilde{x} = X_{\left(\left[\frac{n}{2}\right]+1\right)}.$$

Pokud X *má sudý počet hodnot, vezmeme za medián* $\tilde{x}$ *aritmetický průměr prostředních dvou hodnot*

$$\tilde{x} = \frac{X_{\left(\left[\frac{n}{2}\right]\right)} + X_{\left(\left[\frac{n}{2}\right]+1\right)}}{2}.$$

Medián je speciálním případem *výběrového kvantilu*. Výběrovým kvantilem nazývám hodnotu zvolenou tak, že pozorování, která jsou menší než tato hodnota, tvoří předepsaný díl výběru (např. 25% výběrový kvantil se nazývá dolní výběrový kvartil, 50% výběrový kvantil je medián a 75% výběrový kvantil se nazývá horní výběrový kvartil).

Definice 1.7 *Nechť argument statistického souboru může nabývat pouze konečně mnoha hodnot. Pak* **Modus** *je hodnota argumentu s největší absolutní četností. Modus nemusí být určen jednoznačně.*

Příklad 1.2 *Uvažujme následující hypotetický příklad. Ve firmě F existují 4 platové třídy, s platy uvedenými v následující tabulce. Počet zaměstnanců udává, kolik zaměstnanců je v dané platové třídě.*

Spočtěme některé charakteristiky polohy. Aritmetický průměr $\overline{X} = 13500$, *geometrický průměr* $\overline{X}_g = 12381.3$, *harmonický průměr* $\overline{X}_h = 11726.6$. *Jelikož máme 44 hodnot, bude medián průměr 22. a 23. pořadové hodnoty, tedy*

třída	zařazení	plat v Kč	počet zaměstnanců
1.	výkonná síla	10.000	30
2.	mistr	16.000	10
3.	náměstek	28.000	3
4.	ředitel	50.000	1

Table 1.2: Tabulka četností příjmu zaměstnanců ve firmě F.

$\widetilde{x} = 10.000$. *Dolní výběrový kvartil bude průměr 11. a 12. pořadové hodnoty, tj. 10.000 a horní výběrový kvartil je 16.000.*

Každá charakteristika polohy nám dává jen parciální informaci o statistickém souboru, zatímco grafy rozložení četností nám dávají úplnou informaci o statistickém souboru.

1.3 Charakteristiky variability

Definice 1.8 *Charakteristiky variability:*

1. Rozptylem (disperzí) *statického souboru s rozsahem n nazýváme aritmetický průměr kvadratických odchylek $(x_k - \bar{X})^2$ hodnot argumentu X od aritmetického průměru $\bar{X}$*

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{r} f_i(x_i - \bar{X})^2. \tag{1.4}$$

Rozptyl uvedený ve vzorci 1.4 rozptyl náhodné veličiny podhodnucuje, proto se k výpočtu rozptylu častěji používá vzorce:

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{r} f_i\left(x_i - \bar{X}\right)^2, \tag{1.5}$$

2. Směrodatnou odchylkou *nazýváme*

$$\sqrt{s^2} = s \geq 0. \tag{1.6}$$

3. Průměrnou odchylkou $\bar{d}$ *nazýváme aritmetický průměr absolutních hodnot odchylek od aritmetického průměru* $\bar{X}$*, tj.*

$$\bar{d} = \frac{1}{n}\sum_{k=1}^{n}|x_k - \bar{X}| = \frac{1}{n}\sum_{i=1}^{r} f_i|x_i - \bar{X}|. \tag{1.7}$$

4. Variační koeficient v *statistického souboru je definován jako*

$$v = \frac{s}{\bar{X}}. \tag{1.8}$$

Poznámka 1.2 *Rozptyl je definován vzorcem 1.4, pro jeho výpočet se však častěji používá vzorce*

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k^2) - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{r} f_i x_i^2 - \bar{X}^2. \tag{1.9}$$

nebo

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k^2) - \frac{n}{n-1}\bar{X}^2 = \frac{1}{n-1}\sum_{i=1}^{r} f_i x_i^2 - \frac{n}{n-1}\bar{X}^2. \tag{1.10}$$

Poznámka 1.3 *Hodnoty statistického souboru jsou realizace nějaké náhodné veličiny. Např. Počet telefonních hovorů na ústředně za 1 minutu (viz. Příklad 1.1) je náhodná veličina, která má Poissonovo rozdělení* $X \sim \mathrm{Po}(\lambda)$. *Všechny charakteristiky polohy aproximují střední hodnotu náhodné veličiny* $\mathbb{E}X = \lambda$. *Podobně rozptyl statistického souboru aproximuje rozptyl náhodné veličiny* $\mathrm{Var}X = \lambda$.

Chapter 2

Náhodný výběr

Rozdělení náhodných výběrů podle způsobu provedení

a) *Prostý náhodný výběr s vrácením* (neboli *s opakováním*) je takový výběr, při němž se každý prvek základního souboru vrátí po vybrání zpět do souboru a další prvek se vybírá opět z celého základního souboru.

b) *Prostý náhodný výběr bez vrácení* je takový výběr, při němž se vybraný prvek

nevrací zpět do základního souboru.

c) *Oblastní (stratifikovaný) výběr* spočívá v tom, že základní výběr rozdělíme na stejnorodé disjunktní části a v každé z nich pak provedeme náhodný výběr. Musíme ovšem předpokládat, že o základním souboru máme dostatečné informace umožňující správnou volbu jednotlivých oblastí.

d) *Systematický (mechanický) náhodný výběr* spočívá v tom, že prvky základního statistického souboru seřadíme do určitého pořadí, z prvních k prvků souboru ($N \geq kn$, kde N je rozsah základního, n je rozsah výběru) vybereme náhodně jeden prvek a od něho počínaje vybereme každý *k-tý*, *2k-tý*...prvek.

Rozdělení náhodných výběrů podle rozsahu

a) *Malý* náhodný výběr - rozsah výběru $n < 30$.

b) *Velký* náhodný výběr - rozsah výběru $n \geq 30$.

Budeme uvažovat pouze prostý náhodný výběr s vrácením. Ve spojitosti s teorií

pravděpodobnosti, budeme o prostém náhodném výběru uvažovat následovně.

Definice 2.1 *Nechť* **Z** *je statistický soubor, jehož argument představuje náhodnou veličinu* X. **Náhodným výběrem** *z rozdělení náhodné veličiny* X, *budeme nazývat posloupnost* n *nezávislých realizací pokusu, danou náhodnými veličinami* $X_1, X_2, \ldots, X_n$, *které mají totéž rozdělení jako náhodná veličina* X *a jsou sdruženě nezávislé. (Nebo-li náhodným výběrem nazýváme takový výběr, který poskytuje každému prvku základního statistického souboru stejnou a nezávislou pravděpodobnost , že bude zahrnut do výběru.)*

Definice 2.2 *Charakteristiky základního souboru* **Z** *(náhodné veličiny* X*) budem nazývat* **teoretickými**. *Charakteristiky získané z empirického výběru budeme nazývat* **empirickými (výběrovými)**.

Příklad 2.1 *Statistický soubor představují všichni muži České republiky. Argumentem je jejich věk. Náhodná veličina* X *určuje věk náhodného muže z České republiky. Pro určení charakteristik náhodné veličiny* X *provedeme*

náhodný výběr o rozsahu n. Věk každého vybraného muže je jednou realizací náhodné veličiny X. Výsledné empirické charakteristiky pak odhadují teoretické charakteristiky.

Příklad 2.2 *(viz. Příklad 1.1) X je náhodná veličina udávající počet telefonních výzev za dobu 1 minuty. Byl proveden náhodný výběr z rozdělení X, jehož výsledky jsou zaznamenány v Příkladu 1.1. Předpokládejme, že $X \sim \mathrm{Po}(\lambda)$. Určíme empirickou střední hodnotu např. aritmetickým průměrem $\bar{X} = 2$. Určíme empirický rozptyl např. podle vzorce 1.5 $S^2 = 2.1356$. Z teorie pravděpodobnosti víme, že $\mathbb{E}X = \lambda = \mathrm{VarX}$ pro Poissonova rozdělení. Položme si otázku, zda empirická data prokazují úvodní hypotézu ($X \sim \mathrm{Po}(\lambda)$). Tyto otázky a mnohé další řeší matematická statistika, kterou se budeme zabývat v následujících kapitolách. Zatím pouze položme teoretickou střední hodnotu $\mathbb{E}X = 2$, neboli $\lambda = 2$ a napišme si příslušné pravděpodobnosti $P[X = k]$ pro $k = 0, 1, 2, ...7$ a porovnejme je s příslušnými relativními četnostmi. Z tabulky je vidět, že teoretické pravděpodobnosti se chovají podobně jako relativní četnosti, ale jestli stačí tato podobnost na prohlášení, že $X \sim \mathrm{Po}(2)$ zatím říct*

k	$P[X = k]$	Relativní četnost pro k výzev za jednu minutu
0	0.135	0.133
1	0.271	0.283
2	0.271	0.266
3	0.180	0.166
4	0.090	0,1
5	0.36	0,033
6	0.012	0
7	0.003	0,016

Table 2.1: **Porovnání teoretických pravděpodobností s relativními četnostmi.**

nemůžeme.

Definice 2.3 *Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení, které má střední hodnotu μ a konečný rozptyl σ^2. Zaveďme veličiny*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

kde $\bar{X}$ je **výběrový průměr** *a S^2 je* **výběrový rozptyl**.

Věta 2.1

$$\mathbb{E}\bar{X} = \mu, \qquad \mathrm{Var}\bar{X} = \frac{\sigma^2}{n}, \qquad \mathbb{E}S^2 = \sigma^2.$$

Věta 2.2 Silný zákon velkých čísel

$$\bar{X} \to \mu \qquad \text{skoro jist.}$$

Konvergence skoro jistě znamená, že existuje pouze množina ($A \subset \Omega$) pravděpodobnosti 0 (P(A)=0), pro kterou výraz nekonverguje.

Věta 2.3 Náhodný výběr z normálního rozdělení *Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde $\sigma^2 > 0$. Pak platí následující tvrzení:*

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- *Je-li $n \geq 2$, pak $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.*
- *Je-li $n \geq 2$, pak $\bar{X}$ a S^2 jsou nezávislé.*
- *Je-li $n \geq 2$, pak $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$.*

Chapter 3

Odhady parametrů

Rozlišujeme dva druhy odhadů

- Bodové odhady
- Intervalové odhady neboli intervaly spolehlivosti

Bodové odhady střední hodnoty a rozptylu: Věta 2.1 nám říká, že $\bar{X}$ je nestranný odhad střední hodnoty μ ($\mathbb{E}\bar{X} = \mu$),

S^2 je nestranný odhad σ^2.

Odhad intervalový

(koeficient spolehlivosti) $q = 1 - \alpha$. α se nejčastěji volí 0.05, 0.01 nebo ve vyjímečných případech, kdy potřebujeme mít zaručenou velkou jistotu, 0.001.

Definice 3.1 *Jsou-li B_1, B_2 takové statistiky příslušné parametru β základního souboru, že pro číslo $\alpha \in (0, 1)$ platí*

$$P(B_1 \leq \beta \leq B_2) = 1 - \alpha,$$

pak interval $[B_1, B_2]$ nazýváme **konfidenčním intervalem** *pro parametr β o spolehlivosti $1 - \alpha$. Používá se také názvu interval $100(1 - \alpha)$ - procentní spolehlivosti pro parametr β nebo názvu konfidenční interval pro parametr β se $100(1 - \alpha)$ - procentní spolehlivostí.*

3.1 Intervalové odhady pro parametry normálního rozdělení

Mějme $X_1, \ldots, X_n$ náhodný výběr z $\mathrm{N}(\mu, \sigma^2)$, parametr $\sigma^2 > 0$ není znám. Potom podle Věty 2.3 platí

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1},$$

tudíž podle definice kritické hodnoty studentova rozdělení je

$$P\left[t_{n-1}(\alpha/2) \leq \frac{\bar{X} - \mu}{S}\sqrt{n} \leq t_{n-1}(1-\alpha/2)\right] = 1 - \alpha,$$

přeuspořádáním dostaneme oboustranný intervalový odhad pro střední hodnotu μ normálního rozdělení o spolehlivosti $1 - \alpha$

$$\left[\bar{X} - t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}\right]. \tag{3.1}$$

Intervalový odhad pro rozptyl σ^2 dostaneme obdobně.

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

$$P\left[\chi^2_{n-1}\left(\frac{\alpha}{2}\right) \leq (n-1)S^2/\sigma^2 \leq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\right] = 1-\alpha,$$

přeuspořádáním dostaneme oboustranný intervalový odhad pro rozptyl σ^2 normálního rozdělení o spolehlivosti $1-\alpha$

$$\left[\frac{S^2(n-1)}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}, \frac{S^2(n-1)}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}\right]. \tag{3.2}$$

3.2 Intervalový odhad střední hodnoty pomocí CLV

V případě, že náhodné veličiny nemají normální rozdělení, nemůžeme použít předchozí odhady. Je-li však náhodných veličin větší počet, můžeme pak využít Centrální limitní věty, která říká, že součet většího počtu náhodných veličin se chová jako normální rozdělení. Pro použití aproximace pomocí CLV se obvykle doporučuje rozsah náhodného výběru $n \geq 20$.

Mějme $X_1, \ldots, X_n$ náhodný výběr z rozdělení s konečnou střední hodnotou μ a

konečným rozptylem σ^2. Potom podle Centrální limitní věty má

$$\frac{\bar{X}-\mu}{S}\sqrt{n} \to_{n\to\infty} \Phi \sim N(0,1)$$

asymptoticky normované normální rozdělení. Podle definice kritické hodnoty normovaného normálního rozdělení je

$$P\left[-u(1-\frac{\alpha}{2}) \leq \frac{\bar{X}-\mu}{S}\sqrt{n} \leq u(1-\frac{\alpha}{2})\right] = 1-\alpha,$$

přeuspořádáním dostaneme oboustranný intervalový odhad pro střední hodnotu μ o spolehlivosti $1-\alpha$

$$\left[\bar{X} - u(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}}, \bar{X} + u(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}}\right]. \tag{3.3}$$

Chapter 4

Parametrické testy

Hypotézy: H_0 (nulová) proti alternativní H_1.
Předpokládejme, že rozdělení náhodné veličiny závisí na parametru θ. O parametru θ se domníváme, že by mohl být roven danému číslu θ_0. V tomto případě nulovou hypotézu zapisujeme ve tvaru $H_0 : \theta = \theta_0$. Alternativní hypotéza H1 může být buď ve tvaru $H_1 : \theta \neq \theta_0$, nebo $H_1 : \theta > \theta_0$, popř. $H_1 : \theta \neq \theta_0$. V prvním případě se jedná o oboustrannou hypotézu, ve druhém o jednostrannou (přesněji

pravostrannou, popř. levostrannou).

Při svém rozhodnutí o platnosti H_1 či H_0 se můžeme dopustit jedné ze dvou chyb. Stane-li se, že zamítneme H_0, ačkoli je správná, uděláme tzv. *chybu prvního druhu*. Stane-li se, že nezamítneme H_0, ačkoli správná není, uděláme tzv. *chybu druhého druhu*. Při testování samozřejmě požadujeme, aby pravděpodobnosti obou chyb byly co možná nejmenší.

Při rozhodování o správnosti té či oné hypotézy se opíráme o tak zvanou testovací statistiku T. Testovací statistika je předem daný funkční předpis závisející na nějakém náhodném výběru $X_1, X_2, ..., X_n$, z určitého rozdělení. Hodnoty statistiky T mohou ležet v jedné ze dvou disjunktních množin, a to buď v kritickém oboru W (obor zamítnutí hypotézy H_0) nebo v oboru přijetí V (obor nezamítnutí hypotézy H_0). Jak už bylo řečeno můžeme se při testování dopustit jedné ze dvou chyb, přičemž se obvykle trvá jen na požadavku, aby pravděpodobnost chyby prvního druhu byla rovna α, kde α je nějaké dané číslo z intervalu (0,1). V praxi se nejčastěji volí $\alpha = 0.05$ nebo $\alpha = 0.01$ a číslu α se říká hladina testu.

Poznámka 4.1 *V současné době udává běžný statistický software (Statistica, S+, SAS, ale i Excel) dosaženou hladinu (v anglicky psané literatuře udávané pod názvem P-value, significance value). Je to nejmenší hladina testu, při které bychom ještě hypotézu H_0 zamítli. Tudíž zvolíme-li $\alpha = 0.05$ a P-value vyjde menší než 0.05 (nebo rovna), pak zamítáme hypotézu H_0 na hladině $\alpha = 0.05$. Pokud P-value vyjde větší než 0.05, pak nezamítáme hypotézu H_0 na hladině $\alpha = 0.05$.*

4.1 Jednovýběrový t test

Nechť $X_1, \ldots, X_n$, je náhodný výběr z $\mathrm{N}(\mu, \sigma^2)$, kde $n > 1$. Parametr $\sigma^2 > 0$ není znám. Je třeba testovat hypotézu $H_0 : \mu = \mu_0$, kde μ_0 je dané číslo, proti alternativě $H_1 : \mu \neq \mu_0$. Hypotézu H_0 zamítneme, bude-li $\bar{X}$ hodně vzdáleno od čísla μ_0. Z Věty 2.3 víme, že za platnosti hypotézy H_0 má statistika

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

studentovo rozdělení o n-1 stupních volnosti. Podle definice kritické hodnoty studentova rozdělení, dostaneme, že

$$P\left[|T| \geq t_{n-1}(1 - \alpha/2)\right] = \alpha.$$

Tedy hypotézu H_0 zamítneme na hladině α, jestliže platí

$$|T| \geq t_{n-1}(1 - \alpha/2).$$

p-hodnotu vypočteme z distribuční funkce t rozdělení:

$$p = 2(1 - F_{t_{n-1}}(|T|))$$

V případě jednostranné alternativy $H_1 : \mu > \mu_0$, resp. $H_1 : \mu < \mu_0$ hypotézu H_0 zamítneme, jestliže

$$T \geq t_{n-1}(1-\alpha), \quad \text{resp.} \quad T \leq -t_{n-1}(1-\alpha).$$

p-hodnotu jednostranného testu vypočteme jako:

$$p = 1 - F_{t_{n-1}}(T), \quad \text{resp.} \quad p = F_{t_{n-1}}(T)$$

4.2 Test o rozptylu normálního rozdělení.

Nechť $X_1, \ldots, X_n$, je náhodný výběr z $\mathrm{N}(\mu, \sigma^2)$, kde $n > 1$. Je třeba testovat hypotézu $H_0 : \sigma^2 = \sigma_0^2$, kde σ_0^2 je dané číslo, proti alternativě $H_1 : \sigma^2 \neq \sigma_0^2$. Hypotézu H_0 zamítneme, bude-li S^2 hodně vzdáleno od čísla σ_0^2. Z Věty 2.3 víme, že za platnosti hypotézy H_0 má statistika

$$T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

χ^2 rozdělení o n-1 stupních volnosti. Podle definice kritické hodnoty studentova rozdělení, dostaneme, že

$$P\left[\chi^2_{n-1}\left(\frac{\alpha}{2}\right) \leq (n-1)S^2/\sigma_0^2 \leq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\right] = 1-\alpha,$$

Tedy hypotézu H_0 zamítneme na hladině α, jestliže platí

$$T \leq \chi^2_{n-1}\left(\frac{\alpha}{2}\right) \quad \text{nebo} \quad T \geq \chi^2_{n-1}\left(1-\frac{\alpha}{2}\right).$$

p-hodnotu vypočteme z distribuční funkce χ^2 rozdělení:

$$p = \min\left(2(1 - F_{\chi^2_{n-1}}(T)), 2(F_{\chi^2_{n-1}}(T))\right)$$

V případě jednostranné alternativy $H_1 : \sigma^2 > \sigma_0^2$, resp. $H_1 : \sigma^2 < \sigma_0^2$ hypotézu H_0 zamítneme, jestliže

$$T \geq \chi^2_{n-1}(1-\alpha), \quad \text{resp.} \quad T \leq \chi^2_{n-1}(\alpha).$$

p-hodnotu jednostranného testu vypočteme jako:

$$p = 1 - F_{\chi^2_{n-1}}(T), \quad \text{resp.} \quad p = F_{\chi^2_{n-1}}(T)$$

4.3 Párový t test

Mějme náhodný výběr $(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$ z nějakého dvourozměrného rozdělení jehož vektor středních hodnot je (μ_1, μ_2). Chceme testovat hypotézu $H_0 : \mu_1 - \mu_2 = \Delta$ proti alternativě $H_1 : \mu_1 - \mu_2 \neq \Delta$, kde Δ je nějaké dané číslo (nejčastěji $\Delta = 0$). Položíme

$$X_1 = Y_1 - Z_1, X_2 = Y_2 - Z_2, \ldots, X_n = Y_n - Z_n.$$

Veličiny $X_1, X_2, ..., X_n$ jsou nezávislé. Předpokládejme, že $X_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$. Zřejmě $\mu = \mu_1 - \mu_2$. Jsou-li tyto předpoklady splněny, pak je úloha převedena na jednovýběrový t test. Z veličin $X_1, X_2, ..., X_n$ vypočteme $\bar{X} a S^2$. Hypotézu H_0 zamítneme na hladině α, platí-li

$$|T| = \left| \frac{(\bar{X} - \Delta)\sqrt{n}}{S} \right| \geq t_{n-1}(1 - \alpha/2).$$

p-hodnotu vypočteme opět z distribuční funkce t rozdělení:

$$p = 2(1 - F_{t_{n-1}}(|T|))$$

Párový t test se používá v situacích, kdy máme na každém z n objektů měřeny dvě veličiny. Jednotlivé objekty lze zpravidla pokládat za nezávislé, ale měření na témž objektu nikoli. Párový t test použijeme, když např. testujeme účinnost nějakého léku na n pacientech, přičemž Y_i jsou hodnoty naměřené před podáním léku a Z_i jsou hodnoty naměřené po podání léku.

4.4 Dvouvýběrový t test

Nechť $X_1, X_2, \ldots, X_n$ je výběr z $N(\mu_1, \sigma^2)$ a $Y_1, Y_2, \ldots, Y_m$ výběr z $N(\mu_2, \sigma^2)$. Nechť tyto dva výběry jsou na sobě nezávislé. Předpokládejme, že $n \geq 2, m \geq 2, \sigma^2 > 0$ a σ^2 neznáme. Chceme testovat hypotézu $H_0 : \mu_1 - \mu_2 = \Delta$ proti $H_1 : \mu_1 - \mu_2 \neq \Delta$, kde Δ je nějaké dané číslo (nejčastěji $\Delta = 0$). Označme $\bar{X}, S_X^2$ a $\bar{Y}, S_Y^2$ charakteristiky těchto výběrů. Hypotézu H_0 zamítneme na hladině α platí-li

$$| T |= \left| \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \right| \geq t_{n+m-2}(1 - \alpha/2).$$

p-hodnotu vypočteme opět z distribuční funkce t rozdělení:

$$p = 2(1 - F_{t_{n+m-2}}(|T|))$$

Dvouvýběrový t test používáme v případech, kdy se např. na n pacientech zkouší působení léku A a na jiných m pacientech působení léku B. Účelem pokusu je zjistit, zda působení obou léků je stejné.

Často dochází k záměně párového a dvouvýběrového t testu,což je hrubá chyba. Dvouvýběrový t test můžeme použít pouze v případě, když máme zajištěnu nezávislost všech veličin $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_n$. V případě záměny těchto testů, dojdeme zpravidla k nerozumným a nesmyslným výsledkům.

Předpoklady: Pro výše uvedené testy platí určité předpoklady. Jedním z nich je nezávislost jednotlivých veličin. Tento předpoklad je nejdůležitější, neboť jeho porušení má závažné důsledky a činí závěry založené na předchozích testech chybnými. Dalším předpokladem je normalita rozdělení. Vzhledem k centrální limitní větě a zákonu velkých čísel její porušení při větším rozsahu náhodného výběru není závažné. Navíc v Odstavci 4.7 je uveden test za pomoci CLV, který normalitu nepředpokládá. Při závažném porušení normality a malém rozsahu náhodného výběru dáváme přednost použití některého neparametrického testu. U dvouvýběrového t testu je další požadavek, a to shodnost rozptylů obou rozdělení. V případě, že rozdíl ve velikosti rozptylů není příliš veliký, porušení tohoto požadavku neovlivní podstatným způsobem celkový výsledek. O shodnosti rozptylů rozhodneme na základě následujícího testu.

4.5 Test shodnosti dvou rozptylů

Nechť $X_1, X_2, \ldots, X_n$ je výběr z $N(\mu_1, \sigma_1^2)$ a $Y_1, Y_2, \ldots, Y_m$ výběr z $N(\mu_2, \sigma_2^2)$. Nechť tyto dva výběry jsou na sobě nezávislé. Předpokládejme, že $n \geq 2, m \geq 2, \sigma_1^2 > 0, \sigma_2^2 > 0$. Testujeme hypotézu $H_0 : \sigma_1^2 = \sigma_2^2$ proti $H_1 : \sigma_1^2 \neq \sigma_2^2$ Protože S_X^2 je nestranný odhad parametru σ_1^2 a S_Y^2parametru σ_2^2, lze očekávat, že za platnosti hypotézy H_0 bude podíl $\frac{S_X^2}{S_Y^2}$ blízký jedné. Proti H_0 budou tedy svědčit buď hodnoty blízké nule, nebo hodnoty velké. Hypotézu H_0 zamítneme, jestliže

$$\frac{S_X^2}{S_Y^2} \leq k_1 \quad \text{nebo} \quad \frac{S_X^2}{S_Y^2} \geq k_2,$$

přičemž

$$k_1 = F_{n-1,m-1}(\frac{\alpha}{2}) = \frac{1}{F_{m-1,n-1}(1-\alpha/2)}, \quad k_2 = F_{n-1,m-1}(1-\frac{\alpha}{2}),$$

kde $F_{n-1,m-1}(\alpha/2)$ je kritická hodnota Fisherova-Snedecorova rozdělení o n-1 a m-1 stupních volnosti. p-hodnotu vypočteme z distribuční funkce F rozdělení:

$$p = \min\left(2(1 - F_{F_{n-1,m-1}}(\frac{S_X^2}{S_Y^2})), 2(F_{F_{n-1,m-1}}(\frac{S_X^2}{S_Y^2}))\right)$$

4.6 Porovnávání středních hodnot při nestejných rozptylech

Nechť $X_1, X_2, ..., X_n$ je výběr z $\mathrm{N}(\mu_1, \sigma_1^2)$ a $Y_1, Y_2, ..., Y_n$ je výběr z $\mathrm{N}(\mu_2, \sigma_2^2)$ nezávislý na prvním výběru. Víme-li, že $\sigma_1^2 \neq \sigma_2^2$, můžeme střední hodnoty porovnat následovně. Je-li $m \geq n$ utvoříme rozdíly $X_1 - Y_1, X_2 - Y_2, ...X_n - Y_n$. Na ně lze aplikovat jednovýběrový t test, neboť jednotlivé rozdíly jsou na sobě nezávislé a každý z nich má rozdělení $\mathrm{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. Nevýhodou tohoto postupu je nejen ztráta m-n veličin Y-ových, ale i neefektivní využití zbývajících veličin.

Místo předcházející metody se v praxi dává přednost tomuto přibližnému testu: Nejprve se vypočte

$$S_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right), \quad S_Y^2 = \frac{1}{m-1}\left(\sum_{j=1}^{m} Y_j^2 - m\bar{Y}^2\right)$$

$$S = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}, \quad v_x = \frac{S_X^2}{n}, \quad v_Y = \frac{S_Y^2}{m}.$$

Testujeme-li $H_0 : \mu_1 - \mu_2 = 0$ proti $H_1 : \mu_1 - \mu_2 \neq 0$, pak H_0 zamítá v případě, že platí nerovnost

$$\frac{|\bar{X} - \bar{Y}|}{S} \geq \frac{v_x t_{n-1}(\alpha) + v_y t_{m-1}(\alpha)}{v_x + v_y}.$$

Tento test má přibližně hodnotu α.

4.7 Test o střední hodnotě pomocí CLV

V případě, že náhodné veličiny výrazně nesplňují normalitu, nemůžeme použít předchozí testy. Je-li však náhodných veličin větší počet, můžeme pak využít Centrální limitní věty, která říká, že součet většího počtu náhodných veličin se chová jako normální rozdělení. Pro použití aproximace pomocí CLV se obvykle doporučuje rozsah náhodného výběru $n \geq 20$.

Mějme $X_1, \ldots, X_n$ náhodný výběr z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2. Je třeba testovat hypotézu $H_0 : \mu = \mu_0$, kde μ_0 je dané číslo, proti alternativě $H_1 : \mu \neq \mu_0$. Hypotézu H_0 zamítneme, bude-li $\bar{X}$ hodně vzdáleno od čísla μ_0. Podle Centrální limitní věty má statistika

$$T = \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} \rightarrow_{n\to\infty} \Phi \sim N(0,1)$$

za platnosti H_0 asymptoticky normované normální rozdělení.

Podle definice kritické hodnoty normovaného normálního rozdělení je asymptoticky

$$P\left[|T| \leq u(1-\frac{\alpha}{2})\right] = 1-\alpha.$$

Tedy hypotézu H_0 zamítneme na hladině α, jestliže platí

$$|T| \geq u(1-\frac{\alpha}{2}).$$

p-hodnotu vypočteme z distribuční funkce normovaného normálního rozdělení:

$$p = 2(1-\Phi(|T|))$$

V případě jednostranné alternativy $H_1 : \mu > \mu_0$, resp. $H_1 : \mu < \mu_0$ hypotézu H_0 zamítneme, jestliže

$$T \geq u(1-\alpha), \quad \text{resp.} \quad T \leq -u(1-\alpha).$$

p-hodnotu jednostranného testu vypočteme jako:

$$p = 1-\Phi(T), \quad \text{resp.} \quad p = \Phi(T)$$

V případě, že σ_0^2 není známo, použijeme místo něj ve výpočtu statistiky T jeho nestranný odhad S^2.

Chapter 5

Porovnání více výběrů

5.1 Analýza rozptylu jednoduchého třídění

Tento test je zobecněním dvouvýběrového t testu, který rozšíříme na případ ($I \geq 3$) výběru. Uvažujme tedy I nezávislých výběrů,

$$Y_{11}, ..., Y_{1n_1} \quad \text{je výběr z } N(\mu_1, \sigma^2)$$

atd. až

$$Y_{I1}, ..., Y_{In_I} \quad \text{je výběr z } N(\mu_I, \sigma^2).$$

Chceme testovat hypotézu $H_0 : \mu_1 = \ldots = \mu_I$ proti alternativě, že existují alespoň dvě střední hodnoty, které si rovny nejsou.

Někdy se uvedená situace zapisuje modelem:

$$Y_{ij} = \mu + \alpha_i + e_{ij},$$

kde $\mu + \alpha_i = \mu_i$ a $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ je chyba vyplývající z nepřesnosti měření nebo ze systematické odchylky od průměru. Hypotézu H_0 přepíšeme na jednodušší model, který je splněn, pokud platí hypotéza H_0:

$$Y_{ij} = \mu + e_{ij}.$$

Test provedeme následovně. Nejprve si označme průměry jednotlivých výběrů

$$\overline{Y}_i = \frac{Y_{i1} + ... + Y_{in_i}}{n_i} \quad \text{pro} \quad i = 1, ..., I$$

a průměr všech hodnot

$$\overline{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

kde $n = n_1 + \ldots + n_I$. Nyní spočtěme celkový součet čtverců S_T (tj. celková kvadratická chyba modelu za platnosti H_0, tedy v případě že $\mu_1 = \ldots = \mu_I = \mu$.) Za odhad μ se bere $\overline{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \overline{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\overline{Y}^2.$$

Reziduální součet čtverců S_e je celková kvadratická chyba modelu za předpokladu, že hypotéza H_0 neplatí, tedy v případě že $\mu_1 \neq \ldots \neq \mu_I$. Za odhad μ_i se bere $\overline{Y}_i$.

$$S_e = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2 = \sum_i \sum_j Y_{ij}^2 - \sum_i n_i \overline{Y}_i^2.$$

Veličina $S_A = S_T - S_e$ se interpretuje jako součet čtverců připadající na rozdíly v ošetřeních. Tato veličina je vždy kladná, protože chyba obecnějšího modelu S_e je vždy menší než chyba jednoduššího modelu S_T. Je-li S_A malé, pak oba modely jsou si podobné a tudíž nebudeme zamítat hypotézu H_0. Je-li S_A velké, pak obecnější model vysvětluje velkou část celkové chyby S_T a tudíž zamítneme H_0.

Za platnosti hypotézy H_0 má statistika

$$F_A = \frac{(n-I)S_A}{(I-1)S_e} \sim F_{I-1,n-I}$$

F rozdělení o $I-1$ a $n-I$ stupních volnosti. Tedy hypotézu H_0 zamítneme na hladině α v případě že

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

Výsledky celého testu se stručně zapisují do tabulky (viz. Tabulka 5.1).

Variabilita	součet čtverců S	počet stupňů volnosti f	podíl S/f	F
ošetření	S_A	$f_A = I-1$	S_A/f_A	F_A
reziduální	S_e	$f_e = n-I$	S_e/f_e	-
celková	S_T	$f_t = n-1$	-	-

Table 5.1: Tabulka analýzy rozptylu

Přepoklady tohoto testu jsou obdobné předpokladům dvouvýběrového t testu (viz. strana 40). Nejdůležitější je opět nezávislost jednotlivých výběrů, normalita může být porušena, pokud rozsahy výběrů umožňují použití CLV. Není-li tomu

tak je vhodnější provést neparametrickou obdobu tohoto testu, která se nazývá Kruskalův-Wallisův test. Posledním předpokladem je shodnost rozptylů všech výběrů. Pokud by odhad některého rozptylu vycházel velmi odlišně od ostatních, měli bychom provést test shody rozptylů (viz. např. [**?**]).

Veličina $s^2 = S_e/(n - I)$ se nazývá reziduální rozptyl a je nestranným odhadem rozptylu σ^2.

Poznámka 5.1 *Mohlo by se zdát, že výše uvedený test by se dal provádět sadou dvouvýběrových t testů, provedených na každou dvojici výběrů. Ovšem takových testů bychom museli udělat* $I(I-1)/2$. *Kdyby každý z nich byl proveden na hladině* α, *byla by výsledná hladina výrazně větší než* α. *Pokud bychom hladinu každého testu snížili na* $2\alpha/I(I-1)$, *byla by celková hladina naopak podstatně menší než* α. *Ukazuje se, že takovýto postup nevede k dobrým výsledkům.*

V případě, že hypotézu H_0 zamítneme, je často třeba rozhodnout, pro které dvojice

indexů platí $\mu_i \neq \mu_j$. Tento problém řeší **Tukeyova metoda mnohonásobného porovnání**.

Protože $\overline{Y}_i$ je odhadem pro μ_i, vytvoří se nejprve tabulka rozdílů $\overline{Y}_i - \overline{Y}_j$ (viz. Tabulka 5.2).

	j			
i	2	3	...	I
1	$\overline{Y}_1 - \overline{Y}_2$	$\overline{Y}_1 - \overline{Y}_3$	...	$\overline{Y}_1 - \overline{Y}_I$
2		$\overline{Y}_2 - \overline{Y}_3$	...	$\overline{Y}_2 - \overline{Y}_I$
⋮	⋮	⋮	⋮	⋮
$I-1$				$\overline{Y}_{I-1} - \overline{Y}_I$

Table 5.2: Rozdíly průměrů

Statistika

$$\frac{|\overline{Y}_i - \overline{Y}_j|}{s\sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j})}} \sim q_{I,n-I}$$

má rozdělení nazývající se studentizované rozpětí. Kritická hodnota $q_{I,f}(\alpha)$ studentizovaného rozpětí je takové číslo, pro něž platí $P[Q \geq q_{I,f}(\alpha)] = \alpha$. Tyto

kritické hodnoty jsou tabelovány. Tudíž platí-li

$$|\overline{Y}_i - \overline{Y}_j| \geq sq_{I,n-I}(\alpha)\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

zamítáme hypotézu o rovnosti $\mu_i = \mu_j$. Provedeme-li tento postup pro všechny dvojice, pak hladina testu je menší nebo rovna α. Rovnost nastává v případě, že všechny výběry mají stejný rozsah.

Abychom se lépe orientovali ve výsledcích Tukeyovy metody, připisuje se do Tabulky 5.2 ke každému rozdílu hvězdička, pokud je rozdíl významný (signifikantní) na hladině 0.05. Dvě hvězdičky pro významnost na hladině 0.01 a tři pro 0.001.

5.2 Analýza rozptylu dvojné třídění

Uvažujme model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad \text{kde} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{5.1}$$

kde μ, α_i pro $i = 1, \ldots, I$ a β_j pro $j = 1, \ldots, J$ jsou neznámé parametry a $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ je chyba vyplývající z nepřesnosti měření nebo ze systematické odchylky od průměru. To znamená, že naměřené veličiny Y_{ij} závisí jak na sloupci, tak na řádku, v kterém se vyskytují. Navíc v každém řádku máme stejný počet prvků. Představme si např. situaci, kdy měříme na J pacientech tlak v I okamžicích (např. ráno, v poledne a večer). Každý pacient má jinou průměrnou hodnotu tlaku $\mu + \beta_j$. Výchylky během dne jsou určeny parametry α_i. Je vidět, že ve výše uvedeném modelu jsou dva parametry nadbytečné. Abychom tomuto předešli, klademe na parametry dvě dodatečné podmínky:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

Nyní chceme testovat hypotézu $H_0 : \alpha_1 = \ldots = \alpha_I = 0$ (tj. že nezáleží na řádkovém třídění), kterou přepíšeme na jednodušší model odpovídající jednoduchému třídění:

$$Y_{ij} = \mu + \beta_j + e_{ij}.$$

Test provedeme následovně. Nejprve si označme průměry jednotlivých výběrů

$$\overline{Y}_{i.} = \frac{Y_{i1} + ... + Y_{iJ}}{J} \quad \text{pro} \quad i = 1, ..., I,$$

$$\overline{Y}_{.j} = \frac{Y_{1j} + ... + Y_{Ij}}{I} \quad \text{pro} \quad j = 1, ..., J$$

a průměr všech hodnot

$$\overline{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

kde $n = IJ$. Nyní spočtěme celkový součet čtverců S_T (tj. celková kvadratická chyba, v případě že $\alpha_1 = \ldots = \alpha_I = \beta_1 = \ldots = \beta_J = 0$.) Za odhad μ se bere $\overline{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \overline{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\overline{Y}^2.$$

Součet čtverců chyb v řádcích označíme

$$S_A = J \sum_i \overline{Y}_{i.}^2 - n\overline{Y}^2.$$

Součet čtverců chyb ve sloupcích označíme

$$S_B = I \sum_j \overline{Y}_{.j}^2 - n\overline{Y}^2.$$

Reziduální součet čtverců $S_e = S_T - S_A - S_B$ je celková kvadratická chyba modelu 5.1.

Stupně volnosti jednotlivých součtů čtverců jsou:

$$F_T = n - 1, \quad F_A = I - 1, \quad F_B = J - 1, \quad F_e = n - I - J - 1.$$

Hypotézu H_0 zamítneme na hladině α v případě že

$$F_A = \frac{S_A/f_A}{S_e/f_e} \geq F_{f_A,f_e}(1-\alpha).$$

Podobně budeme postupovat v případě testování hypotézy $H_0' : \beta_1 = \ldots = \beta_J = 0$ (tj. že nezáleží na sloupcovém třídění), kterou přepíšeme na jednodušší model odpovídající jednoduchému třídění:

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$

Hypotézu H_0' zamítneme na hladině α v případě že

$$F_B = \frac{S_B/f_B}{S_e/f_e} \geq F_{f_B,f_e}(1-\alpha).$$

Výsledky celého testu se stručně zapisují do tabulky (viz. Tabulka 5.3).

Variabilita	součet čtverců S	počet stupňů volnosti f	podíl S/f	F
řádková	S_A	$f_A = I-1$	S_A/f_A	F_A
sloupcová	S_B	$f_B = J-1$	S_B/f_B	F_B
reziduální	S_e	$f_e = n-I-J-1$	S_e/f_e	-
celková	S_T	$f_t = n-1$	-	-

Table 5.3: Tabulka analýzy rozptylu dvojného třídění

Reziduální rozptyl $s^2 = S_e/f_e$ je nestranným odhadem rozptylu σ^2.

V případě, že hypotézu H_0 zamítneme, je často třeba rozhodnout, pro které dvojice indexů neplatí rovnost. Tento problém řeší, stejně jako u jednoduchého třídění, **Tukeyova metoda mnohonásobného porovnání**.

Rovnost $\alpha_i = \alpha_l$ zamítneme, platí-li

$$|\overline{Y}_{i.} - \overline{Y}_{l.}| \geq sq_{I,n-I-J+1}(\alpha)\sqrt{\frac{1}{J}},$$

Rovnost $\beta_j = \beta_l$ zamítneme, platí-li

$$|\overline{Y}_{.j} - \overline{Y}_{.l}| \geq sq_{I,n-I-J+1}(\alpha)\sqrt{\frac{1}{I}},$$

Poznámka 5.2 *Model 5.1 se dá dále zobecňovat. Např. pro každé i, j můžeme mít P dat.*

$$Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}, \quad kde \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad p = 1, \ldots, P.$$

Je možné sledovat interakce v modelu mezi řádky a sloupci

$$Y_{ijp} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijp}.$$

*Dále je také možné sledovat závislost veličin na třech typech parametrů tzv. trojné třídění. Tyto modely jsou řešeny např. v [**?**], [**?**]. Neparamitrickou obdobou výše popsaného dvojného třídění je Friedmanův test.*

Chapter 6

Korelační analýza

V kapitole 1 jsme uvedli, že z nezávislosti náhodných veličin plyne nekorelovanost, neboli že korelační koeficient $\rho = 0$. Tudíž zamítneme-li hypotézu $H_0 : \rho = 0$, pak můžeme i zamítnout hypotézu nezávislosti. Zabývejme se tedy nyní hypotézou $H_0 : \rho = 0$.

6.1 Výběrový korelační koeficient

Mějme náhodný výběr $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ z nějakého dvourozměrného rozdělení. Korelační koeficient je definován jako

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}X \ \mathrm{Var}Y}}.$$

Pro odhad Var X a Var Y použijeme výběrový rozptyl

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Z věty 2.1 víme, že $\mathbb{E}S_X^2 = \mathrm{Var}X$ a $\mathbb{E}S_Y^2 = \mathrm{Var}Y$. Podobně definujme výběrovou kovarianci vztahem

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

pro kterou platí $\mathbb{E}S_{XY} = \mathrm{Cov}(X, Y)$. Tudíž pokud $S_X^2 > 0$ a $S_Y^2 > 0$, definujeme výběrový korelační koeficient r jako

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

Po drobné úpravě dostaneme vzorec vhodný pro výpočet:

$$r = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)(\sum Y_i^2 - n\bar{Y}^2)}}.$$

Ze Schwarzovy nerovnosti dostaneme, že $-1 \leq r \leq 1$.

Výběrový korelační koeficient není nestranný odhad ρ, jako tomu je u výběrového rozptylu a kovariance.

Předpokládejme nyní, že $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ je náhodný výběr z nějakého dvourozměrného normálního rozdělení a $\mathrm{Var}\, X > 0$, $\mathrm{Var}\, Y > 0$, $|\rho| < 1$. Za těchto předpokladů je

$$\mathbb{E}r = \rho - \frac{1 - \rho^2}{n} + o(n^{-1}),$$

kde $o(n^{-1})$ značíme funkci $f(n)$, pro kterou platí $\lim_{n\to\infty} \frac{f(n)}{n} = 0$.

Testujme nyní hypotézu $H_0 : \rho = 0$ proti alternativě $H_1 : \rho \neq 0$. Za platnosti hypotézy H_0 a za výše uvedených předpokladů má statistika

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$$

Studentovo rozdělení o $n-2$ stupních volnosti. Tudíž hypotézu H_0 zamítneme na hladině α, v případě, že

$$|T| \geq t_{n-2}(1-\alpha/2).$$

U tohoto testu je normalita náhodného výběru podstatný předpoklad. Nejsme-li si jisti tímto předpokladem, použijeme pro test nezávislosti raději Spearmanův korelační koeficient.

Chapter 7

Lineární regrese

7.1 Lineární regrese s jednou vysvětlující proměnnou

Regresní model

$$Y = f(x)$$

vysvětluje závislost veličiny Y na hodnotách x skrze regresní funkci f. Cílem regrese je najít regresní funkci f, známe-li n pozorovaných dvojic

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

kde x_i jsou hodnoty nezávislé , hodnoty vysvětlující proměnné x a y_i jsou hodnoty závislé, hodnoty vysvětlované veličiny Y. Předpokládejme, že hodnoty y_i jsou naměřeny s určitou chybou e_i. Pro odvození všech testů a intervalových odhadů v průběhu celé této kapitoly, klademe na chyby e_i předpoklad, že mají normální rozdělení N(0, σ^2). Pro odvození bodových odhadů tento předpoklad není nutný. Jinak řečeno, máme n pozorování vysvětlované veličiny Y v n známých hodnotách vysvětlující proměnné x, tudíž máme n rovnic

$$Y_i = f(x_i) + e_i, \quad i = 1, 2, \ldots, n.$$

Lineární regresí budeme uvažovat regresi, jejíž regresní funkce je lineární

$$f(x) = \beta_0 + \beta_1 \cdot x.$$

Cílem lineární regrese je nalezení parametrů β_0 a β_1. Tento úkol provedeme metodou nejmenších čtverců. Tato metoda spočívá v tom, že hledáme parametry β_0 a β_1 pro něž je součet čtverců chyb modelu minimální. Tedy hledáme minimum funkce

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \cdot x_i))^2.$$

Tudíž řešíme soustavu rovnice

$$\frac{\delta g(\beta_0, \beta_1)}{\delta \beta_0} = 0, \quad \frac{\delta g(\beta_0, \beta_1)}{\delta \beta_1} = 0.$$

Po úpravách obdržíme tyto odhady

$$b_1 = \frac{\sum (x_i - \overline{x}) \cdot Y_i}{(x_i - \overline{x})^2} = \frac{\sum x_i Y_i - n\overline{x}\,\overline{Y}}{\sum x_i^2 - n\overline{x}^2}, \quad b_0 = \overline{Y} - b_1 \cdot \overline{x}, \tag{7.1}$$

kde $\overline{x} = \frac{1}{n} \sum x_i$ a $\overline{Y} = \frac{1}{n} \sum Y_i$. Odhady b_0, b_1 jsou nejlepší nestranné odhady, tzn. že odhady b_0, b_1 jsou nestranné ($\mathbb{E}b_0 = \beta_0, \quad \mathbb{E}b_1 = \beta_1$) a mají nejmenší rozptyl ze všech nestranných odhadů.

Minimum funkce g

$$S_e = g(b_0, b_1) = \sum (Y_i - (b_0 + b_1 \cdot x_i))^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i$$

se nazývá **reziduální součet čtverců**. Odhad rozptylu chyb σ^2 je

$$s^2 = \frac{S_e}{n-2}.$$

Totální součet čtverců

$$S_T = \sum (Y_i - \overline{Y})^2$$

vyjadřuje celkovou kvadratickou chybu regresního modelu.

Vhodnost modelu posuzujeme koeficientem determinace

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

který vyjadřuje jaká část celkové chyby S_T je vysvětlena regresním modelem. (Chyba S_e obsahuje to, co regresní model nedokázal vysvětlit). Koeficient determinace

můžeme také počítat podle vzorce

$$R^2 = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2},$$

kde $\widehat{Y}_i = \widehat{f}(x_i) = b_0 + b_1 \cdot x_i$ je regresní odhad hodnoty regresní funkce v bodě x_i. Je zřejmé, že čím blíže je R^2 jedné, tím lépe regresní model vystihuje naměřená data. Někdy se uvádí: je-li koeficient determinace větší než 0.85, můžeme říci, že model je vhodně zvolen.

Nejčastěji se zabýváme otázkou, zda je možné model zjednodušit tak, že hodnoty Y_i vůbec nezávisí na x_i. Tudíž testujeme hypotézu

$$H_0 : \beta_1 = 0 \quad \text{proti} \quad H_1 : \beta_1 \neq 0$$

Za platnosti H_0 má testová statistika

$$T = \frac{b_1}{s} \cdot \sqrt{\sum x_i^2 - n\overline{x}^2} \sim t_{n-2}$$

studentovo rozdělení o $n-2$ stupních volnosti. Tudíž pokud $|T| \geq t_{n-2}(1-\alpha/2)$ zamítneme hypotézu H_0 na hladině spolehlivosti α. Nezamítneme-li hypotézu H_0

tohoto testu pak jsme vlastně potvrdili lineární závislost Y_i na x_i zamaskovanou náhodnými chybami e_i.

Intervaly spolehlivosti Standartním způsobem můžeme vytvořit intervalový odhad pro parametr β_1 o spolehlivosti $1-\alpha$:

$$\left(b_1 - \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\overline{x}^2}}, b_1 + \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\overline{x}^2}}\right).$$

Častěji ovšem hledáme intervalový odhad pro $\beta_0 + \beta_1 x$:

$$\left(b_0 + b_1 x - t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2}}, b_0 + b_1 x + t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{}{\sum}}\right.$$

Tento interval překrývá hodnotu $\beta_0 + \beta_1 x$ s pravděpodobností $1-\alpha$. Sestrojíme-li takovéto intervaly pro všechna $x \in [\min x_i, \max x_i]$, vytvoříme tzv. **pás spolehlivos kolem regresní přímky**. Hranice pásu jsou tvořeny dvěma větvemi hyperboly.

Příklad 7.1 *Za prvních sedm měsíců roku má firma záznamy o počtu hodin provozu výrobní linky* (x_i) *a o nákladech na její údržbu* (Y_i) *v tisících Kč.*

x_i	275	350	250	325	375	400	300
Y_i	149	170	140	164	192	200	165

Najděme nejprve regresní přímku $Y = b_0 + b_1 x$. *Dosadíme-li do vzorců 7.1, dostaneme:* $b_0 = 42.75$ *a* $b_1 = 0.387$. *Nyní spočtěme reziduální součet čtverců* $S_e = 2622.89$, *tudíž odhad rozptylu chyb* e_i *je* $s^2 = 524.579$. *Regresní součet čtverců můžeme snadno spočítat jako* $S_T = (n-1)S_Y^2$, *kde* S_Y^2 *je výběrový rozptyl* Y. $S_T = 2771.71$. *Tudíž koeficient determinace* $R^2 = 0.9463$.

Nyní se zabývejme hypotézou $H_0 : \beta_1 = 0$. *Spočteme statistiku* $T = 9.38$ *a porovnáme jí s hodnotou kvantilu* $t_5(0.975) = 2.57$. *Tudíž zamítáme hypotézu* H_0 *na 5% hladině. Jak koeficient determinace tak tento test nám potvrdil vhodnost tohoto lineární modelu.*

Podívejme se ještě na intervalové odhady. Intervalový odhad o spolehlivosti 95% pro parametr β_1 *je* $[0.2811, 0.4931]$. *Odtud je také vidět, že zamítáme hypotézu* H_0. *Pás spolehlivosti kolem regresní přímky je ukázán na obrázku 7.1.*

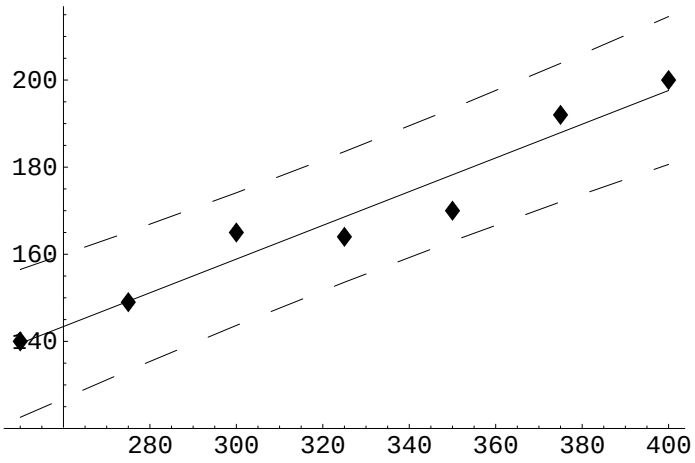

Figure 7.1: Závislost provozních nákladů na době provozu. Body zobrazují naměřené hodnoty, plná čára představuje odhadnutou regresní přímku $Y = b_0 + b_1 x$ a čárkovaně jsou vyznačeny hranice pásu spolehlivosti kolem regresní přímky.

Interpretace modelu: Absolutní člen b_0 odhaduje fixní měsíční náklady, nezávislé na délce provozu linky. Lineární člen $b_1 x$ odhaduje variabilní náklady přímo úměrné délce provozu.

7.2 Lineární regrese s více vysvětlujícími proměnnými

Regrese patří k základním statickým metodám. Jejím cílem je najít regresní funkci, která se snaží vysvětlit vznik většího počtu pozorovaných náhodných veličin $Y_1, Y_2, ..., Y_n$ pomocí známých vlivů X_{ij} a pomocí poměrně malého počtu parametrů $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Budeme se zabývat lineární regresí, tzn. že závislost na parametrech $\beta_0, \beta_1, \beta_2, ..., \beta_k$ je lineární.

K dispozici tedy máme na jednu vysvětlovanou proměnou k vysvětlujících proměnných. Tedy v tomto odstavci budeme pracovat s modelem:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k. \tag{7.2}$$

Pokud máme n pozorování, dostaneme pak n rovnic o *k+1* neznámých ve tvaru

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + e_i, \quad \text{kde } i = 1, 2, ..., n, \tag{7.3}$$

Zde e_i jsou náhodné chyby. Pro odvození všech testů a intervalových odhadů v průběhu celé této kapitoly, klademe na chyby e_i předpoklad, že mají normální rozdělení N(0, σ^2). Pro odvození bodových odhadů tento předpoklad není nutný.

Maticový zápis tohoto modelu má tvar

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e},$$

kde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}. \tag{7.4}$$

Cílem lineární regrese je odhadnout parametry modelu $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Pro odhad těchto parametrů se nejčastěji používá *metoda nejmenších čtverců*. Tato metoda spočívá v minimalizaci funkce

$$g(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))^2. \tag{7.5}$$

Nutnou podmínkou pro existenci extrému je nulovost parciálních derivací. Vzhledem k tomu, že daná funkce je ve svém definičním oboru konvexní, je to i postačující

podmínka. Zderivujeme-li danou funkci podle všech proměnných a položíme-li parciální derivace rovné nule, dostaneme soustavu následujících rovnic

$$\frac{\partial g(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))(-1) = 0$$

a

$$\frac{\partial g(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_j} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}))(-X_{ij}) = 0,$$

kde $j = 1, 2, ..., k$.

Po menších úpravách obdržíme

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2} + ... + \beta_k \sum_{i=1}^{n} X_{ik} = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_{i1} + \beta_1 \sum_{i=1}^{n} X_{i1}^2 + \beta_2 \sum_{i=1}^{n} X_{i2} X_{i1} + ... + \beta_k \sum_{i=1}^{n} X_{ik} X_{i1} = \sum_{i=1}^{n} Y_i X_{i1}$$

$$\vdots$$

$$\beta_0 \sum_{i=1}^{n} X_{ik} + \beta_1 \sum_{i=1}^{n} X_{ik}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{ik}X_{i2} + ... + \beta_k \sum_{i=1}^{n} X_{ik}^2 = \sum_{i=1}^{n} Y_i X_{ik}. \quad (7.6)$$

Vyřešením této soustavy získáme odhady $b_0, b_1, ..., b_k$ parametrů $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Výše uvedená soustava se nazývá *soustava normálních rovnic.* Maticový zápis soustavy normálních rovnic je

$$(\mathbf{X}^T\mathbf{X}) \cdot \beta = \mathbf{X}^T\mathbf{Y}. \quad (7.7)$$

Je-li matice $(\mathbf{X}^T\mathbf{X})$ regulární (tzn. existuje k ní matice inverzní, označme ji $(\mathbf{X}^T\mathbf{X})^{-1}$), potom odhad parametrů $\beta = \beta_0, \beta_1, \beta_2, ..., \beta_k$ je

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (7.8)$$

Minimum funkce g nazýváme **reziduální součet čtverců** a vypočteme jej

$$S_e = g(\mathbf{b}) = \sum (Y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + + b_k x_{ik}))^2 = \sum (Y_i - \widehat{Y}_i)^2,$$

kde $\widehat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$ je regresní odhad hodnoty Y_i. Odhad rozptylu chyb σ^2 je $s^2 = \frac{S_e}{n-k-1}$. s^2 nazýváme reziduální rozptyl.

Regresní součet čtverců $S_T = \sum(Y_i - \overline{Y})^2$ vyjadřuje celkovou kvadratickou chybu regresního modelu.

Vhodnost modelu posuzujeme koeficientem determinace

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

který vyjadřuje jaká část celkové chyby S_T je vysvětlena regresním modelem. (Chyba S_e obsahuje to, co regresní model nedokázal vysvětlit). Koeficient determinace můžeme také počítat podle vzorce

$$R^2 = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2},$$

kde $\widehat{Y}_i = \widehat{f}(x_i) = b_0 + b_1 \cdot x_{i1} + \ldots + b_k \cdot x_{ik}$ je regresní odhad Y_i. Je zřejmé, že čím blíže je R^2 jedné, tím lépe regresní model vystihuje naměřená data. Někdy se

uvádí: je-li koeficient determinace větší než 0.85, můžeme říci, že model je vhodně zvolen.

Metodou nejmenších čtverců získáme bodové odhady parametrů $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Někdy nás však zajímají i intervalové odhady o spolehlivosti $1-\alpha$ konstruované pro parametry $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Intervalový odhad o spolehlivosti $1-\alpha$ pro parametr β_i je interval

$$\left(b_i - t_{n-k-1}(1-\alpha/2) \cdot s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}, b_i + t_{n-k-1}(1-\alpha/2) \cdot s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}} \right), \tag{7.9}$$

kde $(\mathbf{X}^T\mathbf{X})^{-1}_{ii}$ je prvek matice $(\mathbf{X}^T\mathbf{X})^{-1}$, nacházející se na i-tém řádku a i-tém sloupci.

Je zřejmé, že čím méně budeme mít vysvětlujících proměnných, tím bude model jednodušší. Proto se nejčastěji zabýváme otázkou, zda je možné model zjednodušit tak, že hodnoty Y_i vůbec nezávisí na x_{ij}. Tudíž testujeme hypotézu

$$H_0 : \beta_j = 0 \quad \text{proti} \quad H_1 : \beta_j \neq 0$$

Za platnosti H_0 má testová statistika

$$T = \frac{b_j}{s \cdot \sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-k-1} \tag{7.10}$$

studentovo rozdělení o $n-k-1$ stupních volnosti. Tudíž pokud $|T| \geq t_{n-k-1}(1-\alpha/2)$ zamítneme hypotézu H_0 na hladině spolehlivosti α. Nezamítneme-li hypotézu H_0 tohoto testu pak jsme vlastně potvrdili lineární závislost Y_i na i-té vysvětlující proměnné zamaskovanou náhodnými chybami e_i.

Někdy se ptáme, zda je možné model zjednodušit o více než jeden parametr, v takovém případě nepoužijeme dva předchozí testy, protože jejich společná hladina by nebyla α, ale použijeme následující test.

Testujeme hypotézu

$$H_0 : \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_l} = 0, \quad 1 \leq j_1, \ldots, j_l \leq k$$

proti alternativě, že zjednodušený model neplatí (tj. že alespoň jedno $\beta_{j_i} \neq 0$). Číslo l zde označuje počet parametrů, které se pokoušíme z modelu vypustit. Maticový

zápis zjednodušeného modelu má tvar

$$\mathbf{Y} = \widetilde{\mathbf{X}}\widetilde{\beta} + \widetilde{\mathbf{e}},$$

kde matice $\widetilde{X}$ vznikne z matice X vynecháním sloupců příslušejícím parametrům $\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_l}$. Vektor $\widetilde{\beta}$ vznikne z vektoru β vynecháním parametrů $\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_l}$ Podobně vznikne i $\widetilde{e}$.

Parametry zjednodušeného modelu $\widetilde{\beta}$ odhadneme pomocí

$$\widetilde{\mathbf{b}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{Y}. \tag{7.11}$$

Poté spočteme reziduální součet čtverců pro zjednodušený model

$$\widetilde{S}_e = \sum (Y_i - \widetilde{\widehat{Y}}_i)^2,$$

kde $\widetilde{\widehat{Y}}_i$ je regresní odhad Y_i ve zjednodušeném modelu. Je zřejmé, že $\widetilde{S}_e \geq S_e$, neboť S_e je minimum funkce $g(\beta)$ bez jakýchkoli omezení na vektor β, zatímco $\widetilde{S}_e$ je minimum funkce $g(\beta)$ za podmínky $\beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_l} = 0$.

Za platnosti H_0 má pak testová statistika

$$F = \frac{(n-k-1)(\widetilde{S}_e - S_e)}{lS_e} \sim F_{l,n-k-1}$$

rozdělení $F_{l,n-k-1}$. Tudíž pokud $F \geq F_{l,n-k-1}(1-\alpha)$ zamítneme hypotézu H_0 na hladině spolehlivosti α a tudíž model nemůžeme zjednodušit.

Příklad 7.2 *V 60-tých letech proběhla ve Velké Británii následující studie. Ve 30 hrabstvích byli naměřeny veličiny: A = změna populace za posledních 10 let, B = počet zaměstnanců v zemědělství, C = velikost daní z nemovitostí, D = procento obyvatel mající telefon, E = procento obyvatel žijící na vesnici, F = průměrný věk. Těmito veličinami měla být vysvětlena veličina Y = procento obyvatel žijících pod hranicí bídy. Tudíž dostáváme lineární regresní model*

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i + \beta_E E_i + \beta_F F_i + e_i.$$

Matice $\mathbf{X}$ *bude obsahovat 7 sloupců, kde v prvním budou samé jedničky, ve druhém budou hodnoty veličiny A, ve třetím B, atd. Nyní podle vzorce 7.8*

spočteme odhad jednotlivých parametrů

$\mathbf{b} = (b_0, b_A, b_B, b_C, b_D, b_E, b_F)^T = (31.26, -0.39, 0.0007, 1.23, -0.083, 0.16, -0,42)^T.$

Dále spočteme reziduální součet čtverců $S_e = 265.66$ *a regresní součet čtverců* $S_T = 1197.72$. *Odtud dostáváme, že* $R^2 = 0.78$.

Nyní nás zajímá, jestli některé proměnné můžeme z modelu vypustit ($H_0 : \beta_j = 0$*). Za tímto účelem, spočteme pro každou proměnnou hodnotu statistiky* T *a to podle vzorce 7.10.*

$\mathbf{T} = (T_0, T_A, T_B, T_C, T_D, T_E, T_F)^T = (2.35, -4.87, 1.69, 0.38, -0.63, 2.67, -1.64)^T.$

Tyto hodnoty porovnáme s kvantilem $t_{23}(0.975) = 2.068$ *a vidíme, že hypotézu nulovosti parametrů zamítáme u* $\beta_0, \beta_A, \beta_E$. *Tyto testy nám říkají, že můžeme z modelu vypustit proměnnou B nebo C nebo D nebo F. Ovšem nevíme, zda můžeme vypustit všechny proměnné najednou. Na tuto otázku nám odpoví následující F-test.*

Uvažujme tedy zjednodušený model (podmodel)

$$Y_i = \beta_0 + \beta_A A_i + \beta_E E_i + e_i.$$

Pro podmodel spočteme odhad jednotlivých parametrů

$$\widetilde{\mathbf{b}} = (\widetilde{b}_0, \widetilde{b}_A, \widetilde{b}_E)^T = (16.67, -0.40, 0.13)^T.$$

Dále spočteme reziduální součet čtverců podmodelu $\widetilde{S}_e = 393.03$. *Odtud dostáván* *že* $R^2 = 0.67$. *Nyní sestrojíme statistiku*

$$F = \frac{(30 - 6 - 1)(\widetilde{S}_e - S_e)}{3S_e} = 3.67 > 3.03 = F_{3,23}(0.95).$$

Z toho plyne, že H_0 *zamítáme, neboli nemůžeme vypustit všechny čtyři proměnné* *zároveň.*

Musíme tedy některou proměnnou do podmodelu přidat. Přidejme proměnnou B, *protože* $T_B > T_C, T_D, T_F$ *(tj. proměnná* B *je v modelu významější než* C, D *a* F*). Uvažujme tedy podmodel*

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_E E_i + e_i.$$

Pro podmodel spočteme odhad jednotlivých parametrů

$$\widetilde{\mathbf{b}} = (\widetilde{b}_0, \widetilde{b}_A, \widetilde{b}_B, \widetilde{b}_E)^T = (10.99, -0.40, 0.001, 0.19)^T.$$

Dále spočteme reziduální součet čtverců podmodelu $\widetilde{S}_e = 318.83$. *Odtud dostáván že* $R^2 = 0.73$. *Nyní sestrojíme statistiku*

$$F = \frac{(30 - 6 - 1)(\widetilde{S}_e - S_e)}{4S_e} = 1.15 < 2.80 = F_{4,23}(0.95).$$

Z toho plyne, že H_0 *nezamítáme, neboli z původního modelu můžeme vypustit proměnné C, D a F. Popíšeme tedy veličinu Y proměnnými A, B a E.*

7.3 Polynomiální regrese

Kvadratická regrese: Pod pojmem kvadratická regrese míníme model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + e_i, \quad i = 1, 2, ..., n,$$

kde $e_i \sim N(0, \sigma^2), n \geq 4$. Zde náhodná veličina Y_i závisí kvadraticky na veličinnách X_i.

Položíme-li $Z_i = X_i^2, \quad i = 1, 2, ..., n$ dostáváme model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + e_i, \quad i = 1, 2, ..., n.$$

V tomto modelu závisí náhodná veličina Y_i lineárně na veličinách X_i a Z_i. Neboli úloha kvadratické regrese byla převedena na úlohu lineární regrese se dvěma vysvětlujícími proměnnými.

Podobně budeme postupovat i pro regrese vyšších stupňů.

Odhad stupně regresního polynomu: Uvažujme nyní model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \ldots + \beta_p \cdot X_i^p + e_i, \quad i = 1, 2, ..., n,$$

kde stupeň p regresního polynomu není znám. Počet parametrů tohoto modelu označme $k = p + 1$.

Uvažujme, že skutečný stupeň polynomu je p_0 a tedy skutečný počet parametrů modelu je $k_0 = p_0 + 1$. Označme s_k^2 reziduální rozptyl modelu s k parametry (reziduální rozptyl je definován na str. 75). Dá se ukázat, že

$$\mathbb{E}s_k^2 > \sigma^2 \quad \text{pro } k < k_0, \quad \mathbb{E}s_k^2 = \sigma^2 \quad \text{pro } k \geq k_0.$$

Tudíž, je třeba najít bod, kde posloupnost s_k^2 se mění z klesající posloupnosti na oscilující. Toto je obtížná úloha, proto ji převedeme na úlohu hledání minima posloupnosti. Vytvoříme posloupnost

$$A_k = s_k^2 \left(1 + \frac{k}{\sqrt[4]{n}}\right),$$

která větším k přidává větší váhu. Hodnotu k, pro kterou je A_k minimální, pak vezmeme jako odhad skutečného počtu parametrů k_0.

7.4 Nelineární regrese

Uvažujme nelineární regresní model

$$Y_i = f(X_i, \beta) + e_i, \quad i = 1, 2, ..., n,$$

kde f je regresní funkce a β je vektor neznámých parametrů. Odhad parametrů β metodou nejmenších čtverců dostaneme minimalizací výrazu

$$S(\beta) = \sum_{i=1}^{n} (Y_i - f(X_i, \beta))^2.$$

Tuto úlohu iteračně řeší různé statistické a matematické programy. Tyto programy ovšem vyžadují počáteční aproximaci vektoru $\mathbf{b}$.

Počáteční aproximaci můžeme snadno získat u tzv. **linearizovatelných modelů** tj. modelů, které se dají převést na lineární model. Jako příklad si uveďme model, jehož regresní funkce je exponenciální:

$$Y_i = \beta_0 e^{\beta_1 X_i} + e_i, \quad i = 1, 2, ..., n.$$

Při počáteční aproximaci si můžeme dovolit zapomenout na chyby e_i a model zlogaritmovat

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i, \quad i = 1, 2, ..., n.$$

Zavedeme-li nové parametry $\alpha_0 = \ln \beta_0$ a $\alpha_1 = \beta_1$, dostaneme lineární regresní

model

$$\ln Y_i = \alpha_0 + \alpha_1 X_i, \quad i = 1, 2, ..., n,$$

který vyřešíme podle kapitoly **??** a dostaneme odhady a_0, a_1. Za odhad parametrů původního modelu pak vezmeme odhady

$$b_0 = e^{a_0}, \quad b_1 = a_1.$$

Na závěr uveďme příklady některých linearizovatelných modelů.

1. $Y = e^{\beta_0 + \beta_1 X}$
2. $Y = \beta_0 X^{\beta_1}$
3. $Y = \beta_0 + \beta_1 \ln x$
4. $Y = \ln(\beta_0 + \beta_1 X)$
5. $Y = \frac{1}{\beta_0 + \beta_1 X}$