

Texty k přednáškám

# Matematická Statistika

Ivan Nagy, Jitka Kratochvílová

# Obsah

<b>1 Náhodný výběr</b>	<b>3</b>
1.1 Pojem náhodného výběru (Skripta str. 68) . . . . .	3
1.2 Charakteristiky výběru (Skripta str. 69) . . . . .	4
1.3 Výběrový průměr (Skripta str. 70) . . . . .	4
1.4 Charakteristiky výběrového průměru (Skripta str. 70) . . . . .	4
1.5 Rozdělení výběrového průměru (Skripta str. 66,70) . . . . .	5
1.6 Výběrový rozptyl (Skripta str. 71) . . . . .	5
1.7 Výběrový podíl (Skripta str. 72) . . . . .	5
<b>2 Výběrové charakteristiky</b>	<b>6</b>
2.1 Výběrový průměr při známém rozptylu souboru (Skripta str. 70) . . . . .	6
2.2 Výběrový průměr při neznámém rozptylu souboru (Skripta str. 72) . . . . .	6
2.3 Výběrový rozptyl (Skripta str. 71) . . . . .	6
2.4 Výběrový podíl (Skripta str. 72) . . . . .	7
2.5 Dva výběrové průměry 74 . . . . .	7
2.6 Dva výběrové rozptyly . . . . .	8
2.7 Dva výběrové podíly . . . . .	8
<b>3 Bodové odhady a jejich vlastnosti</b>	<b>9</b>
3.1 Statistika (Skripta str. 77) . . . . .	9
3.2 Bodový odhad parametru rozdělení (Skripta str. 77) . . . . .	9
3.3 Vlastnosti bodových odhadů (Skripta str. 78) . . . . .	10
3.4 Konstrukce bodových odhadů (Skripta str. 82) . . . . .	12
<b>4 Intervalové odhady</b>	<b>15</b>
4.1 Pojem intervalu spolehlivosti (Skripta str. 85) . . . . .	15
4.2 Druhy intervalů spolehlivosti pro jednu náhodnou veličinu (Skripta str. 86-94)	16
<b>5 Parametrické testy hypotéz</b>	<b>19</b>
5.1 Pojem parametrického testu (Skripta str. 95-96) . . . . .	19
5.2 Základní pojmy (Skripta str. 97) . . . . .	19
5.3 P-hodnota (Skripta str. 101-103) . . . . .	21
5.4 Obecné schéma testu hypotézy (Skripta str. 101) . . . . .	21
5.5 Vybrané parametrické testy (Skripta str. 103-115) . . . . .	23

<b>6 Chi<sup>2</sup> testy hypotéz</b>	<b>25</b>
6.1 Test dobré shody (Skripta str. 115-117)	25
6.2 Test nezávislosti (Skripta str. 118-119)	27
<b>7 Další neparametrické testy hypotéz</b>	<b>29</b>
7.1 Test mediánu	29
7.2 Test nezávislosti prvků výběru	30
7.3 Test nezávislosti výběrů	30
7.4 Test typu rozdělení	32
<b>8 Regresní analýza</b>	<b>33</b>
8.1 Lineární regrese (Skripta str. 121-126)	33
8.2 Nelineární regrese	35
<b>9 Korelační analýza (Skripta str. 127-141)</b>	<b>37</b>
9.1 Intervaly spolehlivosti (Skripta str. 136,138-139)	37
9.2 Testy hypotéz (Skripta str. 134-135,140)	38
<b>10 Analýza rozptylu (ANOVA)</b>	<b>40</b>
10.1 ANOVA při jednoduchém třídění	40
10.2 ANOVA při dvojném třídění	41
<b>11 Analýza hlavních komponent</b>	<b>44</b>
11.1 Rozklad kovarianční matice	45
11.2 Rozklad datové matice	47

# 1 Náhodný výběr

Hlavním úkolem statistiky je rozbor dat, která vykazují náhodné kolísání. Jedná se většinou o data měřená na určitém procesu za účelem (lepšího) poznání tohoto procesu. Například: měříme hustoty a intenzity dopravního proudu na několika místech určité dopravní oblasti, abychom byli schopni (lépe) řídit dopravu v této oblasti.

## 1.1 Pojem náhodného výběru (Skripta str. 68)

Zkoumaný proces chápeme jako náhodnou veličinu s určitým, nám neznámým (nebo neúplně známým), rozdelením a měřená data jako realizace této náhodné veličiny. Pro jednoduchost a lepší představu se při výkladu omezíme na nejběžnější druh procesu, který nazveme *základním statistickým pokusem* a který spočívá v dotazu vybrané jednotky z dané množiny jednotek na určitou věc. Množinu všech odpovědí nazveme *soubor* a podmnožinu získaných odpovědí nazveme *výběr*.

Například: Sledování spotřeby automobilů určitého typu, vyrobené v daném roce a při najetí 5000 km. Souborem jsou spotřeby automobilů vyrobených v daném roce. Výběrem jsou spotřeby sledovaných automobilů. Dotaz je symbolický název pro změření spotřeby automobilu. Jednotka je automobil, vyrobený v daném roce (prvek souboru). Množina spotřeb všech automobilů z daného roku je *soubor* a podmnožina spotřeb sledovaných automobilů je *výběr*.

Abychom získali objektivní informace o zkoumaném procesu, je třeba, aby výběr dotazovaných jednotek byl nezávislý.

Například: Zjišťujeme-li průměrné stáří automobilů, je nesmyslné omezit se na autobazary.

Jestliže výběr, tj. sérii dotazů, opakujeme, dostaneme jiné odpovědi. Je tím, že dotazy při dalším výběru zahrnují jiné jednotky. Abstraktně lze definovat *náhodný výběr* jako uspořádanou množinu (vektor) náhodných veličin, jejíž realizací dostaneme jednu konkrétní realizaci výběru – číselný vektor.

### Definice 1.1 (Náhodný výběr)

Náhodný výběr  $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n]$  je vektor nezávislých a stejně rozdělených náhodných veličin. Číslo  $n$  se nazývá *rozsah výběru*. Vektor realizací náhodných veličin  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  nazveme *realizací náhodného výběru*.

Jedna z typických úloh statistiky je odhad střední hodnoty souboru, přičemž předpokládáme, že typ rozdělení souboru je známý. Odhadujeme jeden z jeho parametrů – střední hodnotu. O té vypovídají všechna data náhodného výběru. Je proto užitečné znát rozdělení celého náhodného výběru.

### Tvrzení 1.1 (Distribuční funkce náhodného výběru)

$\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n]$  je náhodný výběr a  $F(x_i)$ ,  $i = 1, 2, \dots, n$  je distribuční funkce, určující rozdělení jeho složek. Pak distribuční funkce náhodného výběru  $H(\mathbf{x})$  je rovna součinu distribučních funkcí jeho složek

$$H(\mathbf{x}) = H(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \dots F(x_n).$$

*Ověření:* Plyně přímo ze skutečnosti, že složky náhodného výběru jsou nezávislé a distribuční funkce nezávislých náhodných veličin je rovna součinu jejich distribučních funkcí.

## 1.2 Charakteristiky výběru (Skripta str. 69)

Realizace náhodného výběru je datový soubor. Ten lze popsat pomocí charakteristik popisné statistiky. Těmito charakteristikami lze popsat i náhodný výběr, s jednou důležitou odlišností. Charakteristiky datového souboru jsou konstanty (např. pro střední hodnotu platí: sečtu-li daná čísla odpředu nebo odzadu, dostanu vždy totéž). Charakteristiky náhodného výběru jsou náhodné veličiny. Náhodnost zde vzniká vzhledem k opakováním výběrům. Provedeme-li první výběr a spočteme charakteristiku (např. průměr) této realizace, dostaneme číslo. Další výběr poskytne novou realizaci a protože je náhodný, budou v ní jiná čísla. Proto i charakteristika (průměr) bude mít jinou hodnotu. Tedy: každý výběr dá jinou realizaci a jinou hodnotu charakteristiky – realizaci charakteristiky náhodného výběru, která je náhodnou veličinou.

Různé charakteristiky náhodného výběru budou dále velmi důležité. Každá zkoumaná vlastnost bude mít svou charakteristiku, které budeme říkat *statistika*. Protože statistika je náhodná, má své rozdělení (hustotu pravděpodobnosti statistiky). Ta je základním nástrojem pro veškerá odvození klasické statistiky.

Nejprve se podrobněji podíváme na nejznámější charakteristiku výběrový průměr, v příští kapitole si pak uvedeme další základní charakteristiky a jejich vlastnosti.

## 1.3 Výběrový průměr (Skripta str. 70)

### Definice 1.2 (Výběrový průměr)

Pro výběr  $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n]$  ze spojité náhodné veličiny  $X$  se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$  je výběrový průměr definován vztahem

$$\bar{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^n \mathcal{X}_i. \quad (1)$$

**Komentář** k definici

1. Všimněme si, že ve vzorci pro výběrový průměr figurují velká písmena. Není to tedy průměr z čísel, ale formálně zapsaný průměr z náhodných veličin.

## 1.4 Charakteristiky výběrového průměru (Skripta str. 70)

Pro výběrový průměr, jako náhodnou veličinu, uvedeme jeho střední hodnotu a rozptyl. Jsou to vlastně charakteristiky definované na charakteristice výběrový průměr. Tyto charakteristiky se počítají pro všechny možné hodnoty výběrového průměru. Závisí proto na všech realizacích souboru. Jsou to tedy již konstanty (důkladně rozmyslet).

PŘÍKLAD: Uvažujme soubor s hodnotami  $\{1, 2, 3\}$  a výběr z tohoto souboru o rozsahu  $n = 2$ . Všechny možné výběry jsou uvedeny jako sloupce následující tabulky

možné výběry	1 1 2 3 1 2 2 3 1 2 3	1 2 3 1 2 2 3 1 2 3	1 1.5 2 1.5 2 2.5 2 3 1 2.5 3
výb. průměry	1 1.5 2 1.5 2 2.5 2 3 1 2.5 3		

Jestliže zprůměrujeme všechny výběrové průměry, dostaneme střední hodnotu výběrového průměru. Zde je to 2.

**Střední hodnota výběrového průměru  $\bar{X}$**  je rovna střední hodnotě souboru  $E[X] = \mu$

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu. \quad (2)$$

**Rozptyl výběrového průměru  $\bar{X}$**  je roven rozptylu souboru  $D[X] = \sigma^2$ , dělenému rozsahem výběru  $n$

$$s_{\bar{X}}^2 = D[\bar{X}] = D\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} D\left[\sum_{i=1}^n X_i\right] \underset{\text{nezávislost}}{\equiv} \frac{1}{n^2} \sum_{i=1}^n D[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}. \quad (3)$$

## 1.5 Rozdělení výběrového průměru (Skripta str. 66,70)

Pro výběr z náhodné veličiny s normálním rozdělením  $N(\mu, \sigma^2)$ , nebo pro dostatečně velký výběr (viz Centrální limitní věta) platí

$$\bar{X} \sim N(E[\bar{X}], D[\bar{X}]) = N(\mu, \sigma^2/n)$$

tj, výběrový průměr má normální rozdělení se střední hodnotou  $\mu$  a rozptylem  $\sigma^2/n$ .

## 1.6 Výběrový rozptyl (Skripta str. 71)

### Definice 1.3 (Výběrový rozptyl)

Pro výběr  $\mathbf{X} = [X_1, X_2, \dots, X_n]$  ze spojité náhodné veličiny  $X$  se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$  je výběrový rozptyl definován vztahem

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (4)$$

## 1.7 Výběrový podíl (Skripta str. 72)

### Definice 1.4 (Výběrový podíl)

Pro výběr  $\mathbf{X} = [X_1, X_2, \dots, X_n]$  z alternativní náhodné veličiny  $X$  se souborovým podílem  $\pi$  je výběrový podíl definován vztahem

$$p = \frac{n^+}{n}, \quad (5)$$

kde  $n^+$  je počet příznivých pokusů ve výběru a  $n$  je rozsah výběru.

## 2 Výběrové charakteristiky

Výběrový průměr je nejznámější, ale zdaleka ne jedinou charakteristikou výběru. Nyní uvedeme ty charakteristiky, které budeme dále nejčastěji používat. Mohou se týkat jednoho nebo dvou výběrů.

### JEDEN NÁHODNÝ VÝBĚR

Uvažujeme výběr z jednoho rozdělení, např. zjišťujeme stáří náhodně vybraných automobilů. Budeme sledovat tři základní charakteristiky náhodného výběru: výběrový *průměr*, *rozptyl* a *podíl*. Protože je výběr náhodný, jsou i tyto charakteristiky náhodné a každá z nich má své rozdělení.

#### 2.1 Výběrový průměr při známém rozptylu souboru (Skripta str. 70)

**Normovaný výběrový průměr** z normálního rozdělení  $N(\mu; \sigma^2)$ , nebo dostatečně velký výběr z libovolného rozdělení se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$  je

$$Z = \frac{\bar{x} - \mu}{\sigma} \sqrt{n} \sim N(0, 1) \quad (6)$$

a má normované normální rozdělení  $N(0; 1)$ .

#### 2.2 Výběrový průměr při neznámém rozptylu souboru (Skripta str. 72)

V případě neznámého rozptylu souboru, ze kterého je výběr pořízen, se neznámý rozptyl odhadne pomocí výběrového rozptylu (4).

**Normovaný výběrový průměr** je definován

$$t = \frac{\bar{x} - \mu}{S} \sqrt{n} \sim St(n - 1) \quad (7)$$

a má Studentovo rozdělení s  $n - 1$  stupni volnosti.

#### 2.3 Výběrový rozptyl (Skripta str. 71)

**Normovaný výběrový rozptyl** je definován vztahem

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim Chi2(n - 1) \quad (8)$$

a má rozdělení rozdělení  $\chi^2$  s  $n - 1$  stupni volnosti.

## 2.4 Výběrový podíl (Skripta str. 72)

Normovaný výběrový podíl je

$$Z = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim N(0; 1) \quad (9)$$

a má přibližně normální normované rozdělení  $N(0; 1)$ . To platí pro  $np > 5$  a zároveň  $n(1-p) > 5$ .

## DVA NÁHODNÉ VÝBĚRY

Uvažujeme dva výběry, tj. dvě náhodné veličiny, které chceme zkoumat. Například měříme nahuštění předních pneumatik u náhodně vybraných automobilů. Tlak v levé pneumatice je realizací první náhodné veličiny, tlak v pravé je realizací druhé náhodné veličiny. Charakteristikami budou opět *průměr*, *rozptyl* a *podíl* a vztahují se na rozdíl obou náhodných veličin. Předpokládáme, že rozptyl rozdělení není znám a je nahrazen výběrovým rozptylem.

Značíme:  $x_1, x_2$  výběry,  $n_1, n_2$  rozsahy výběrů,  $\mu_1, \mu_2$  střední hodnoty výběrů a  $\sigma_1^2, \sigma_2^2$  rozptyly výběrů.

## 2.5 Dva výběrové průměry 74

Normovaný rozdíl dvou výběrových průměrů při shodných rozptylech

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_P \sqrt{1/n_1 + 1/n_2}} \sim St(n-1), \quad S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}. \quad (10)$$

a má Studentovo rozdělení s  $n-1$  stupni volnosti.

Normovaný rozdíl dvou výběrových průměrů při různých rozptylech

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim St(n-1) \quad (11)$$

a má Studentovo rozdělení s  $\delta$  stupni volnosti, kde

$$\delta = \frac{(k_1 + k_2)^2}{\frac{k_1^2}{n_1-1} + \frac{k_2^2}{n_2-1}}, \quad k_1 = \frac{s_1^2}{n_1}, \quad k_2 = \frac{s_2^2}{n_2}$$

Normovaný rozdíl dvou výběrových průměrů při párových výběrech

$$t = \frac{\bar{D} - (\mu_1 - \mu_2)}{S_D \sqrt{n}} \sim St(n-1), \quad (12)$$

kde  $D = \mathcal{X}_1 - \mathcal{X}_2$ ,  $\bar{D}$  je výběrový průměr a  $S_D$  výběrový rozptyl náhodného vektoru  $D$ .

$$D_i = X_{1,i} - X_{2,i}; \quad \bar{D} = \frac{1}{n} \sum_{i=1}^n D_i; \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

Tato charakteristika má Studentovo rozdělení s  $n - 1$  stupni volnosti.

**PŘÍKLAD:** Sledujeme nahuštění předních pneumatik u osobních automobilů. U každého auta změříme tlak v levé pneumatici (prvek výběru 1) a v pravé pneumatici (prvek výběru 2). Tedy, pro každý objekt (automobil) měříme dvě vlastnosti (pravou a levou pneumatiku). Tak dostaneme párové výběry.

**POZNÁMKA:** Párový rozdíl výběrových průměrů dostaneme tak, že odečteme položky jednotlivých výběrů, tak dostaneme výběr  $D$  a z něho sestavíme obyčejný normovaný výběrový průměr (7).

## 2.6 Dva výběrové rozptyly

**Normovaný podíl dvou výběrových rozptylů**

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1). \quad (13)$$

Tato charakteristika má  $F$  rozdělení s  $n_1 - 1$  stupni volnosti pro čitatele a  $n_2 - 1$  stupni volnosti pro jmenovatele.

## 2.7 Dva výběrové podíly

**Normovaný rozdíl dvou výběrových podílů**  $p_1$  a  $p_2$  pro dvě alternativní rozdělení s podíly  $\pi_1$  a  $\pi_2$  je

$$Z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0; 1) \quad (14)$$

a má přibližně normální rozdělení  $N(0; 1)$ .

### 3 Bodové odhady a jejich vlastnosti

#### 3.1 Statistika (Skripta str. 77)

Výběr pořizujeme proto, abychom se (více) doveděli o souboru, ze kterého jsme výběr pořídili. Zde se soustředíme na situaci, kdy známe rozdělení souboru až na jeden nebo více parametrů. Např. víme, že rozdělení souboru je normální, ale neznáme jeho střední hodnotu, případně i rozptyl. Z výběru se snažíme hodnoty těchto neznámých parametrů odhadnout. Předpis, pomocí kterého z výběru vypočteme hodnotu neznámého parametru se nazývá *statistika*. V souladu s tím je i následující definice.

##### Definice 3.1 (Statistika)

Statistika  $T = T(\mathcal{X})$  je funkce výběru  $\mathcal{X}$ .

**Komentář** k definici

1. Statistika určená pro odhadování se nazývá *odhadová statistika*, pro testování *testová statistika*.
2. Definice neříká nic o tom, jak statistiku volit vzhledem k jejímu cílovému využití (odhad, test). Její vhodnost či nevhodnost budeme zkoumat později.

#### 3.2 Bodový odhad parametru rozdělení (Skripta str. 77)

##### Definice 3.2 (Bodový odhad)

Sledujeme rozdělení s hustotou pravděpodobnosti  $f(x; \theta)$  s neznámým parametrem  $\theta$ . Provedli jsme realizaci náhodného výběru  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  z tohoto rozdělení a definovali statistiku  $T(\mathcal{X})$ . *Bodový odhad*  $\hat{\theta}$  parametru  $\theta$  pro realizaci náhodného výběru  $\mathbf{x}$  je hodnota statistiky  $T$  s dosazenou realizací náhodného výběru  $\mathbf{x}$

$$\hat{\theta} = T(\mathbf{x}) \quad (15)$$

**Komentář** k definici

1. Pro každou novou realizaci výběru obdržíme jiný bodový odhad. Odtud je zřejmé, že bodový odhad nemůže dát úplně přesnou hodnotu parametru.
2. Vlastní volbu statistiky jsme zatím nechali stranou. Lze pro ni použít metodu momentů nebo maximální věrohodnosti, o které se zmíníme. Statistiku je také možno volit heuristicky, potom je však třeba ověřit její vlastnosti.

PŘÍKLAD: Odhadujeme parametr střední hodnotu  $\mu$ . Provedli jsme výběr

$$\mathbf{x} = \{x_1, x_2, \dots, x_n\} = \{3, 5, 2, 4, 5\}.$$

Protože víme, že střední hodnotu si lze přibližně představit jako průměr všech možných realizací, usoudíme, že jejím vhodným odhadem bude aritmetický průměr  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . Po dosazení realizace výběru dostaneme  $\bar{x} = 19/5$ . Vlastnosti zvolené statistiky je však třeba ověřit (viz dále).

### 3.3 Vlastnosti bodových odhadů (Skripta str. 78)

Vlastnosti bodového odhadu  $\hat{\theta}$  vypovídají o vhodnosti použité statistiky  $T$  k odhadu hodnoty parametru  $\theta$ . Uvedeme tři vlastnosti: *nestrannost*, *konzistence* a *vydatnost*. {{}}

#### NESTRANNOST<sup>1</sup>

##### Definice 3.3 (Nestrannost)

Statistika  $T$  poskytuje *nestranný bodový odhad* parametru  $\theta$ , jestliže její střední hodnota se rovná tomuto parametru

$$E[T] = \theta \quad (16)$$

**Komentář** k definici

1. Střední hodnotu  $E[T]$  je třeba chápat jako ”průměrování” přes všechny možné výběry. Jestliže bychom chtěli naznačit výpočet této střední hodnoty, bylo by třeba postupovat takto: provedeme první výběr, spočteme bodový odhad, provedeme druhý výběr a opět spočteme bodový odhad, atd. Po provedení všech možných výběrů uděláme průměr ze všech jednotlivých bodových odhadů. To je hledaná střední hodnota.
2. Nestrannost říká, že odhad je ”v průměru” (rozumíme přes všechny možné výběry) přesný.

**PŘÍKLAD:** Ověříme nestrannost výběrového průměru vzhledem ke střední hodnotě. Máme dokázat, že platí  $E[\bar{X}] = \mu$ . Dosadíme definici výběrového průměru a manipulujeme s operátorem střední hodnoty

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

**Vychýlení bodového odhadu  $B(\theta)$ <sup>2</sup>** je definováno jako rozdíl mezi střední hodnotou statistiky a odhadovaným parametrem

$$B(\theta) = E(T) - \theta \quad (17)$$

**POZNÁMKA:** Z definice nevychýlenosti přímo plyne, že je-li statistika  $T$  pro odhad parametru  $\theta$  nevychýlená, pak vychýlení  $B(\theta)$  je nulové.

#### KONZISTENCE<sup>3</sup>

##### Definice 3.4 (Konzistence)

---

<sup>1</sup>Skripta str. 78

<sup>2</sup>Skripta str. 79

<sup>3</sup>Skripta str. 79

Statistika  $T$  dává konzistentní bodový odhad parametru  $\theta$ , jestliže pro rostoucí rozsah výběru se hodnota statistiky (v pravděpodobnosti) neomezeně blíží skutečnému parametru

$$\lim_{n \rightarrow \infty} P(|T - \theta| < \epsilon) = 1, \quad \forall \epsilon > 0 \quad (18)$$

### Komentář k definici

Tato vlastnost je limitní. Lze ji sledovat jen při zvětšujícím se rozsahu výběru – např. změříme 10 hodnot, pak 100, atd.

### Tvrzení 3.1 (Kriterium konzistence<sup>4</sup>)

Odhad je konzistentní, jestliže je

- asymptoticky nestranný,
- jeho rozptyl jde k nule s rozsahem výběru jdoucím k nekonečnu.

*Ověření: Plyne z použití Čebyševovy nerovnosti pro obecný odhad.*

PŘÍKLAD: Ověříme konzistenci výběrového průměru vzhledem k odhadu střední hodnoty. Pro důkaz využijeme Čebyševovu nerovnost (??), kterou zapíšeme pro výběrová průměr a střední hodnotu

$$P(|\bar{X} - \mu| < \epsilon) > 1 - \frac{\sigma^2}{n\epsilon^2}$$

Protože  $\lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} = 0$ , je odhad konzistentní.

### VYDATNOST<sup>5</sup>

### Definice 3.5 (Vydatnost)

Pro dvě nestranné statistiky  $T$  a  $U$  definujeme jako vydatnější tu z nich, která má menší rozptyl

$$D[T] < D[U] \implies T \text{ je vydatnější než } U \quad (19)$$

PONÁMKA: Pro posouzení dvou statistik, které nejsou nestranné, je třeba zavést středně kvadratickou chybu  $MSE$ <sup>6</sup> definovanou vztahem

$$MSE = E[(T - \theta)^2] = D[T] + (B(\theta))^2 \quad (20)$$

Jako vydatnější definujeme statistiku s menší  $MSE$ . (Pro nestranné statistiky  $MSE$  přechází na rozptyl a obě definice jsou shodné.)

Ze vztahu pro  $MSE$  je patrné, že v obecném případě tato charakteristika posuzuje jak rozptyl odhadové statistiky, tak i její vychýlení.  $MSE$  bude minimální, jestliže bude minimální rozptyl i vychýlení statistiky.

---

<sup>5</sup>Skripta str. 80

<sup>6</sup>Skripta str. 80

### 3.4 Konstrukce bodových odhadů (Skripta str. 82)

Řekli jsme co je bodový odhad a jaké u něho sledujeme vlastnosti. Nyní se budeme věnovat otázce, jak lze takový odhad zkonstruovat. Ukážeme dvě základní metody pro konstrukci statistiky, vhodné pro odhad daného parametru.

#### METODA MOMENTŮ<sup>7</sup>

Tato metoda je velmi jednoduchá, obecně však nedává příliš kvalitní výsledky. Spočívá v porovnání obecných (nebo centrálních) momentů souboru a výběru. Podle toho, kolik parametrů odhadujeme, kolik momentů musíme porovnat. Momenty souboru počítáme s pomocí hustoty pravděpodobnosti souboru  $f(x, \theta)$ . Budou tedy obsahovat neznámý parametr  $\theta$ . Výběr je množina změrených hodnot. Moment výběru bude tedy číslo. Porovnáním momentů získáme rovnice kde neznámé budou odhadované parametry. Z nich odhad vypočteme.

**PŘÍKLAD:** Budeme odhadovat neznámý parametr  $\delta$  exponenciálního rozdělení s hustotou pravděpodobnosti  $f(x, \delta) = \delta \exp\{-\delta x\}$ ,  $\delta > 0$ ,  $x \in (0, \infty)$  z výběru  $\mathbf{x} = [x_1, x_2, \dots, x_n]$ .

Protože odhadujeme jediný parametr, stačí porovnat první momenty, tj. střední hodnotu souboru (exponenciálního rozdělení) a výběrový průměr změřeného výběru.

Střední hodnota souboru je

$$E[X] = \int_0^\infty x f(x, \delta) dx = \int_0^\infty x \delta \exp\{-\delta x\} dx = 1/\delta.$$

Výběrový průměr je

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n x_i = \text{číslo.}$$

Porovnáním dostaneme

$$\frac{1}{\delta} = \bar{\mathbf{x}} \Rightarrow \hat{\delta} = \frac{1}{\bar{\mathbf{x}}},$$

kde symbolem  $\hat{\delta}$  jsme označili bodový odhad parametru  $\delta$ .

#### METODA MAXIMÁLNÍ VĚROHODNOSTI<sup>8</sup>

##### ODHAD PRO OBECNÉ ROZDĚLENÍ

Tato metoda dává velmi kvalitní výsledky a je často používána. Pro normální rozdělení souboru je ekvivalentní s metodou nejmenších čtverců, o které budeme mluvit v regresní analýze. Metoda je založena na minimalizaci tzv. věrohodnostní funkce nebo jejím logaritmu (což je totéž – proč?).

---

<sup>7</sup>Skripta str. 82

<sup>8</sup>Skripta str. 82

### Definice 3.6 (Věrohodnostní funkce)

Pro rozdělení s hustotou pravděpodobnosti  $f(x, \theta)$  a realizaci náhodného výběru  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  definujeme věrohodnostní funkci  $\mathcal{L}_n(\theta)$  vztahem

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n f(x_i, \theta). \quad (21)$$

### Definice 3.7 (Maximálně věrohodný odhad)

Maximálně věrohodným odhadem parametru  $\theta$  rozdělení  $f(x, \theta)$  nazveme odhad  $\hat{\theta} \in \theta^*$ , který maximalizuje (logaritmus) věrohodnostní funkce, tj. platí

$$\log \mathcal{L}_n(\hat{\theta}) \geq \log \mathcal{L}_n(\theta), \quad \forall \theta \in \theta^* \quad (22)$$

kde  $\theta^*$  označuje množinu všech přípustných hodnot parametru  $\theta$ .

Obecný postup při určení maximálně věrohodného odhadu je následující:

1. Sestavíme věrohodnostní funkci tak, že násobíme hustoty pravděpodobnosti souboru a do každé dosadíme za  $x$  jeden prvek výběru  $x_i$ .
2. Je-li to výhodné, věrohodnostní funkci logaritmujeme. Protože řada rozdělení má hustotu pravděpodobnosti ve tvaru exponenciály, bývá logaritmus užitečný pro zjednodušení součinu. Není však nutný.
3. Hledáme maximum logaritmu věrohodnostní funkce. V jednoduchém případě např. pomocí derivace, ve složitějším případě numericky.

Bod, kde leží maximum je maximálně věrohodný odhad.

## ODHAD PRO EXPONENCIÁLNÍ TŘÍDU ROZDĚLENÍ

### Definice 3.8 (Exponenciální třída rozdělení)

Řekneme, že hustota pravděpodobnosti patří do exponenciální třídy, jestliže je možno ji zapsat ve tvaru

$$f(x, \theta) = \exp\{Q(\theta) U(x) + R(\theta) + V(x)\}, \quad (23)$$

kde  $Q, R$  jsou funkcemi jen  $\theta$  (ne  $x$ ) a  $U, V$  jsou funkcemi jen  $x$  (ne  $\theta$ ).

PŘÍKLAD: Alternativní rozdělení je z exponenciální třídy, neboť platí

$$\begin{aligned} f(x, \pi) &= \pi^x (1 - \pi)^{1-x} = \exp\{\log(\pi^x (1 - \pi)^{1-x})\} = \\ &= \exp\{x \log(\pi) + (1 - x) \log(1 - \pi)\} = \exp\{[\log(\pi) - \log(1 - \pi)] x + \log(1 - \pi)\}. \end{aligned}$$

Zde platí:  $Q = \log(\pi) - \log(1 - \pi)$ ,  $U = x$ ,  $R = \log(1 - \pi)$ ,  $V = 0$ .

Je-li rozdělení z exponenciální třídy, dostáváme následující jednoduché řešení úlohy maximálně věrohodného odhadu.

### Tvrzení 3.2 (Max. věr. odhad pro exponenciální třídu)

Pro rozdělení s hustotou pravděpodobnosti  $f(x, \theta) = \exp\{Q(\theta)U(x) + R(\theta) + V(x)\}$  a realizaci výběru  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  je extrém logaritmické věrohodnostní funkce dán řešením rovnice

$$Q'(\theta)S(x) + nR'(\theta) = 0,$$

kde  $Q', R'$  jsou derivace podle  $\theta$  a  $S(x) = \sum_{i=1}^n U(x_i)$ .

Aby nalezený extrém byl maximum, musí být ještě splněna nerovnost

$$Q''(\theta)S(x) + nR'' < 0.$$

Ověření: Věrohodnostní funkce je

$$\begin{aligned} \mathcal{L}_n(\theta) &= \prod_{i=1}^n \exp\{Q(\theta)U(x_i) + R(\theta) + V(x_i)\} = \exp\left\{\sum_{i=1}^n (Q(\theta)U(x_i) + R(\theta) + V(x_i))\right\} = \\ &= \exp\left\{Q(\theta)\sum_{i=1}^n U(x_i) + nR(\theta) + \sum_{i=1}^n V(x_i)\right\}. \end{aligned}$$

Tento výraz logaritmujeme a se zavedeným značením dostaneme

$$\log \mathcal{L}_n(\theta) = Q(\theta)S(x) + nR(\theta) + \sum_{i=1}^n V(x_i).$$

Po derivaci podle  $\theta$  poslední výraz zmizí. Anulováním derivace dostáváme podmínu pro extrém. Podmína pro maximum je záporná druhá derivace.

**PŘÍKLAD:** Metodou maximální věrohodnosti určete odhadovou statistiku pro parametr  $\pi$  alternativního rozdělení.

Uvažujeme alternativní rozdělení, jehož exponenciální tvar jsme již odvodili (viz příklad k (23)) a zjistili jsme, že platí  $Q = \log(\pi/(1-\pi))$ ,  $U = x$ ,  $R = \log(1-\pi)$ ,  $V = 0$ . Abychom mohli použít tvrzení 3.2, potřebujeme ještě spočítat funkci  $S$  příslušné derivace.

$$S = \sum_{i=1}^n x_i = n\bar{x}, \quad Q' = \frac{1}{\pi(1-\pi)}, \quad Q'' = -\frac{2\pi-1}{\pi^2(1-\pi)^2}, \quad R' = -\frac{1}{1-\pi}, \quad R'' = -\frac{1}{(1-\pi)^2}$$

Podmínka extrému

$$Q'S + nR' = \frac{1}{\pi(1-\pi)}n\bar{x} - n\frac{1}{1-\pi} = 0 \Rightarrow \hat{\pi} = \bar{x}$$

Podmínka maxima (s dosazeným odhadem  $\bar{x} = \hat{\pi}$ )

$$-\frac{2\hat{\pi}-1}{\hat{\pi}^2(1-\hat{\pi})^2}n\hat{\pi} - n\frac{1}{(1-\hat{\pi})^2} < 0 \Rightarrow \hat{\pi} < 1$$

což je vždy splněno, neboť  $\pi = 1$  je patologický případ.

## 4 Intervalové odhady

### 4.1 Pojem intervalu spolehlivosti (Skripta str. 85)

Bodové odhady poskytují hodnotu odhadu, ale neříkají nic o jeho přesnosti. Ta je zahrnuta v intervalu spolehlivosti - tj. intervalu, ve kterém leží neznámý parametr s danou pravděpodobností. Přesnost (nebo neurčitost) odhadu je dána šírkou intervalu.

Z definice pravděpodobnosti vyplývá, že pravděpodobnost parametru v daném intervalu je asymptoticky rovna relativní četnosti odhadů, které padnou do tohoto intervalu (při opakovaných výběrech). Proto lze interval spolehlivosti charakterizovat také jako interval, do kterého padne dané procento odhadů. Relativní četnost odhadů určuje hustota pravděpodobnosti statistiky pro bodový odhad. Ta je také základem pro určení intervalu spolehlivosti.

#### Definice 4.1 (Interval spolehlivosti)

Interval  $I_\alpha = (\theta_D, \theta_H)$  nazveme  $\alpha$ -interval spolehlivosti pro parametr  $\theta$ , jestliže platí

$$P(\theta \in I_\alpha) = 1 - \alpha. \quad (24)$$

#### Komentář k definici

1. Interval spolehlivosti budeme zkracovat IS.
2.  $I_\alpha$  se také nazývá  $100(1 - \alpha)$ -procentní interval spolehlivosti, tj. 0.05-IS je 95% IS.
3. Pokud je  $-\infty < \theta_D < \theta_H < \infty$ , tj.  $I_\alpha$  je zdola i shora omezený, hovoříme o oboustranném IS. Je-li  $\theta_D = -\infty$ , tj.  $I_\alpha = (-\infty, \theta_H)$ , jedná se o pravostranný IS, pro  $\theta_H = \infty$ , tj.  $I_\alpha = (\theta_D, \infty)$  jde o levostanný IS.

PŘÍKLAD: Určete 95% IS pro střední hodnotu  $\mu$  normálního rozdělení s rozptylem  $\sigma^2 = 1$  na základě výběru o rozsahu  $n = 100$  s průměrem  $\bar{x} = 3.7$ .

Podle požadavků na IS musí platit

$$P(\mu_D < \bar{x} < \mu_H) = 1 - \alpha.$$

Normujeme

$$P\left(\zeta_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\sigma} \sqrt{n} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

a v argumentu vyjádříme  $\mu$  a použijeme vztah  $\zeta_{\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$

$$-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Upravenou podmínu dosadíme zpět

$$P\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

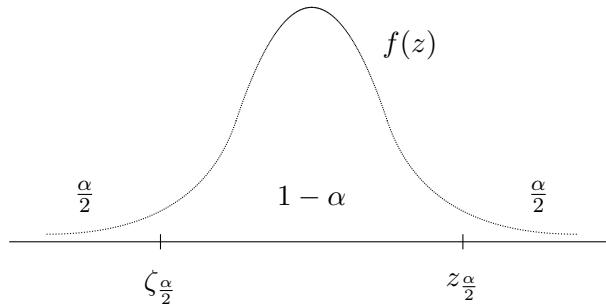
Podle definice je IS

$$I_\alpha = (\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

nebo píšeme

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Na obrázku je hustota pravděpodobnosti bodového odhadu, tj. výběrového průměru a na ní jsou vyznačeny příslušné pravděpodobnosti, kvantil a kritická hodnota.



Konkrétně:

$$\alpha = 0.05/2 = 0.025; \quad \bar{x} = 3.7; \quad \sigma = 1; \quad n = 100; \quad z_{\frac{\alpha}{2}} = 1.96 \text{ (z tabulek)}$$

$$I_{0.05} = (3.7 - \frac{1}{10}1.96; 3.7 + \frac{1}{10}1.96) = (3.504; 3.896)$$

## 4.2 Druhy intervalů spolehlivosti pro jednu náhodnou veličinu

(Skripta str. 86-94)

Budeme uvažovat následující IS.

**Střední hodnota (známé  $\sigma^2$ )<sup>9</sup>**

$$I_\alpha : \mu \in \bar{x} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \quad z \sim N(0; 1)$$

• V programu `Octave` lze pro tento interval použít funkci

```
is=z_int(x,v,alpha,alt),
```

kde `is` je interval spolehlivosti, `x` je výběr, `v` je rozptyl souboru, `alpha` je hladina významnosti, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ )

---

<sup>9</sup>Skripta str. 86

## Střední hodnota (neznámé $\sigma^2$ )<sup>10</sup>

$$I_\alpha : \mu \in \bar{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2}, \quad t \sim St(n-1)$$

• V programu `Octave` lze pro tyto intervaly použít funkci

```
is=t_int(x,alpha,alt),
is=t_int_2s(x1,x2,alpha,alt),
is=t_int_2n(x1,x2,alpha,alt),
is=t_int_2p(x1,x2,alpha,alt),
```

kde `is` je interval spolehlivosti, `x,x1,x2` výběr, `alpha` hladina významnosti, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ )

## Rozptyl<sup>11</sup>

$$I_\alpha : \left( \frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \right), \quad \chi^2 \sim \text{Chi}^2(n-1)$$

• V programu `Octave` lze pro tento interval použít funkci

```
is=var_int(x,alpha,alt),
```

kde `is` je interval spolehlivosti, `x` výběr, `alpha` hladina významnosti, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ )

## Podíl<sup>12</sup>

$$I_\alpha : \pi \in p \pm \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}, \quad z \sim N(0; 1)$$

• V programu `Octave` lze pro tento interval použít funkci

```
is=prop_int(x,n,alpha,alt),
is=prop_int_2(x1,n1,x2,n2,alpha,alt),
```

kde `is` je interval spolehlivosti, `x,x1,x2` je výběrový podíl nebo počet, `alpha` hladina významnosti, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ )

**Doplněk** k uvedeným intervalům:

---

<sup>10</sup>Skripta str. 89

<sup>11</sup>Skripta str. 90

<sup>12</sup>Skripta str. 92

výběrový průměr:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , výběrový rozptyl:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ,  
 výběrový podíl:  $p = \frac{n^+}{n}$ ,  $n^+$  je počet jedniček (úspěchů).

PŘÍKLAD: V jakém intervalu lze očekávat životnost zakoupené pneumatiky s pravděpodobností 0.95, jestliže pro 10 náhodně vybraných pneumatik byly zjištěny následující životnosti (v rocích)

2 4 5 6 10 8 6 5 6 7.

Přípravné výpočty:

$$\bar{x} = 5.9, \quad s = 2.183, \quad t_{0.025}(9) = 2.262$$

Interval:

$$I_{0.05} = (4.34; 7.46)$$

## 5 Parametrické testy hypotéz

### 5.1 Pojem parametrického testu (Skripta str. 95-96)

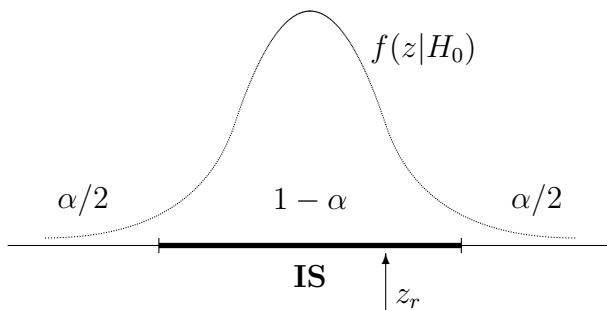
Na základě výběru srovnáváme dvě tvrzení o hodnotě určitého parametru  $\theta$  rozdelení  $f(x, \theta)$ . První tvrzení (které většinou obhajuje stávající stav věcí) se nazývá *nulová hypotéza* a značí se  $H_0$ , druhé tvrzení (které většinou prosazuje, že věci se změnily) je *alternativní hypotéza* označená  $H_A$ . Nulová hypotéza něco tvrdí: např., že střední hodnota  $\mu$  je rovna  $\mu_0$  a alternativní hypotéza ji odporuje. To může mít tři různé podoby:

hypotéza	příklad	(25)
I. parametr má podle $H_A$ <b>větší</b> hodnotu než podle $H_0$	$\mu > \mu_0$	
II. parametr má podle $H_A$ <b>menší</b> hodnotu než podle $H_0$	$\mu < \mu_0$	
III. parametr se podle $H_A$ <b>nerovná</b> hodnotě parametru podle $H_0$	$\mu \neq \mu_0$	

Tvrzení testujeme na základě *testové statistiky*, kterou je statistika pro bodový odhad parametru podle  $H_0$ . Pro parametrické testy lze podstatu testování vyložit v souvislosti s IS následujícím způsobem (pro jednoduchost budeme uvažovat test pro střední hodnotu a se známým rozptylem souboru).

Nulová hypotéza říká, že  $\mu = \mu_0$ . Jestliže je tato hypotéza pravdivá a kolem bodu  $\mu_0$  sestrojíme  $\alpha$  IS a tam by s také s pravděpodobností  $1 - \alpha$  měl padnout bodový odhad, pořízený z výběru. Pokud tam padne, hypotézu  $H_0$  nezamítáme – řekneme, že data neprokázala její neplatnost. Pokud bodový odhad padne mimo IS, hypotézu  $H_0$  zamítнемe. Jediný (formální) rozdíl testů a intervalů je v tom, že při intervalu používáme nenormovaný tvar statistiky, např. pro střední hodnotu se známým rozptylem je to výběrový průměr  $\bar{X}$ , zatímco pro test použijeme normovaný výběrový průměr  $z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$ . Jeho realizaci označíme  $z_r$ .

P O Z N Á M K A: *Všimněte si, že pro normování použijeme  $\mu_0$ , což je střední hodnota podle nulové hypotézy. Celý test probíhá za platnosti  $H_0$ , která bud' dále trvá, nebo je testem vyvrácena.*



### 5.2 Základní pojmy (Skripta str. 97)

V předchozím odstavci jsme dosti netradičně nastínili podstatu testování parametrických hypotéz. Nyní uvedeme základní pojmy a postupy pro testování tak, jak je lze běžně nalézt v učebnicích klasické statistiky. Pojmy budeme ihned demonstrovat na následujícím příkladě.

**PŘÍKLAD:** Závod vyrábí televizní obrazovky u kterých udává střední životnost 1200 h a rozptyl životnosti 900 h<sup>2</sup>. Vývojové oddělení provedlo některé technologické změny při výrobě a tvrdí, že životnost nově vyrobených obrazovek je větší. Své tvrzení dokládá výběrem 10 obrazovek z nové série pro něž byla zjištěna průměrná životnost 1216 h. Testujte  $H_0$  : "střední životnost obrazovek je 1200 hodin" na hladině významnosti 0.05.

**Nulová hypotéza**  $H_0$  je základní hypotéza, která většinou potvrzuje stávající stav.

[ $H_0$ : střední životnost obrazovek je 1200 h.]

**Alternativní hypotéza**  $H_A$  je nová hypotéza, která jedním ze způsobů (25) popírá  $H_0$ .

[ $H_A$ : střední životnost obrazovek je větší.]

**Testová statistika** je normovaná statistika pro bodový odhad testovaného parametru.

[Zde se jedná o střední hodnotu, jejíž odhadová statistika je výběrový průměr.]

Dále se jako parametr bude objevovat rozptyl se statistikou výběrový rozptyl a podíl se statistikou výběrový podíl.

**Hladina významnosti**  $\alpha$  je pravděpodobnost  $\alpha$  z IS. Je to tzv. pravděpodobnost I. druhu, tj., že  $H_0$  bude zamítnuta, zatímco je ve skutečnosti pravdivá.

[V příkladu je  $\alpha = 0.05$ , což bývá nejčastěji používaná hodnota.]

**Obor přijetí** je množina hodnot normované statistiky, která odpovídá IS. Pokud normovaná statistika padne do oboru přijetí, není  $H_0$  zamítnuta.

**Kritický obor**  $W$  je množina hodnot normované statistiky, která odpovídá doplňku IS. Pokud normovaná statistika padne do kritického oboru, je  $H_0$  zamítnuta.

**P O Z N Á M K A:** Podle alternativní hypotézy  $H_A$  poznáme směrování testu.

pro  $W = (\bullet; \infty)$  podle I. hovoříme o pravostranném testu;

pro  $W = (-\infty; \bullet)$  podle II. hovoříme o levostranném testu;

pro  $W = (-\infty; \bullet) \cup (\bullet; \infty)$  podle III. hovoříme o oboustranném testu.

**PŘÍKLAD: (pokračování)** Vypočteme příklad o televizních obrazovkách.

Statistika pro odhad:  $\bar{x} = 1216$  (zadáno).

Hladina významnosti:  $\alpha = 0.05$  (zadáno).

Normovaná statistika:  $z_r = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{1216 - 1200}{\sqrt{900}} \sqrt{10} = 1.687$ .

Obor přijetí: dostaneme normováním pravostranného IS, tj.  $(-\infty; z_\alpha) = (-\infty; 1.645)$ .

Kritický obor: je doplňkem oboru přijetí, tj.  $W = (1.645; \infty)$ .

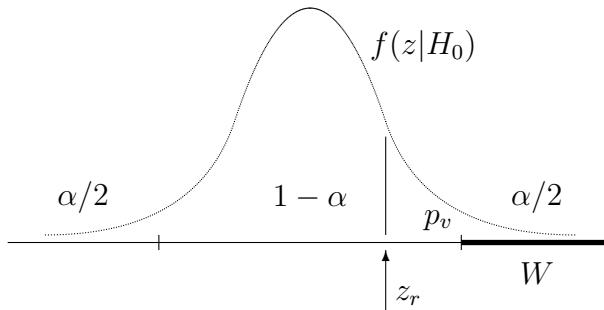
Závěr:  $z \in W \Rightarrow H_0$  zamítáme. Tedy, není pravda, že střední životnost obrazovek je 1200 hodin.

### 5.3 P-hodnota (Skripta str. 101-103)

Z předchozího příkladu je patrné, že z klasického testu nepoznáme, "jak moc"  $H_0$  zamítáme nebo "jak daleko" od zamítnutí se  $H_0$  nachází. Abychom tento nedostatek odstranili, a také abychom výsledek testu vyjádřili jediným číslem, zavádíme *p-hodnotu*  $p_v$ . Pro pravostranný test střední hodnoty je p-hodnota definována vztahem

$$p_v = P(Z > z_r | H_0), \quad (26)$$

což je plocha pod hustotou pravděpodobnosti normované statistiky, vpravo od realizované statistiky.



Z obrázku je patrné, že:

- bude-li  $p_v = \alpha$ , budeme na hranici zamítnutí,
- pro  $p_v < \alpha$  leží realizovaná statistika v kritickém oboru  $W$ , a tedy  $H_0$  zamítáme,
- pro  $p_v > \alpha$  leží realizovaná statistika mimo kritický obor  $W$ , a tedy  $H_0$  nezamítáme.

PŘÍKLAD: **(pokračování)** K předchozímu příkladu ještě dopočteme p-hodnotu:

$$p_v = P(Z > 1.687) = 0.046.$$

Protože  $p_v < \alpha$ , hypotézu  $H_0$  zamítáme.

### 5.4 Obecné schéma testu hypotézy (Skripta str. 101)

Pro jednotlivé případy z kapitoly 2 lze použít obecné schéma:

**Známe:** zadané a spočtené charakteristiky a konstanty.

**Testová statistika  $T$ :** použitá normovaná statistika a její rozdělení (vzorec - normovaný podle  $H_0$ ).

**Hodnota statistiky  $T_r$ :** statistika s dosazeným výběrem (vypočtené číslo).

**P-hodnota:** spočteme pravděpodobnost  $P_r = P(T < T_r)$   
 (kvantil pro hodnotu statistiky - funkce `xxx.cdf`)

$$\begin{aligned} p_v &= 1 - P_r && \text{pro pravostranný test, tj. } \theta_0 > \theta \\ p_v &= P_r && \text{pro levostranný test, tj. } \theta_0 < \theta \\ p_v &= 2 \min\{P_r, 1 - P_r\} && \text{pro oboustranný test, tj. } \theta_0 \neq \theta \end{aligned}$$

**Závěr:** slovní interpretace výsledku.

**PŘÍKLAD:** **(test pro dva podíly)** Na dvou pracovištích A a B byly sledovány pracovní prostory. Pracoviště A bylo sledováno v  $n_A = 800$  časových okamžicích a bylo zaznamenáno  $n_A^+ = 116$  prostojů, zatímco pracoviště B bylo sledováno  $n_B = 1200$  krát a zjištěno  $n_B^+ = 138$  prostojů. Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu o rovnosti středních podílů prostojů na obou pracovištích.

Řešení provedeme podle uvedeného schématu:

**Známe:**

Hypotézy:  $H_0 : p_A = p_B, H_A : p_A \neq p_B$

Směrování: oboustranný test.

Podíly:

$$p_A = \frac{116}{800} = 0.145, \quad p_B = \frac{138}{1200} = 0.115, \quad p_P = \frac{p_1(1-p_1)}{n_1-1} + \frac{p_2(1-p_2)}{n_2-1} = 2.4 \cdot 10^{-4}$$

Hladina významnosti:  $\alpha = 0.05$

**Testová statistika:**

$$z = \frac{p_1 - p_2}{\sqrt{p_P}} \sim N(0; 1)$$

**Hodnota statistiky:**  $z_r = 1.94$

**Kritický obor**  $W = (-1.96; 1.96)$

**P-hodnota:**  $p_v = 2 \times 0.0265 = 0.053$ .

**Závěr:** Na hladině 0.05 hypotézu  $H_0$  nezamítáme, podíly prostojů na pracovištích A a B nejsou stejné. P-hodnota navíc ukazuje, že zamítnutí je poměrně těsné.

**P O Z N Á M K A:**

1. Výběr charakteristiky (vzorce) pro test určuje  $H_0$  - o čem vypořídá tvrzení nulové hypotézy, to se testuje.
2. O směrování testu rozhoduje  $H_A$  - podle toho jak odporuje nulové hypotéze ( $<$ ,  $>$ ,  $\neq$ ) určíme směrování.

## 5.5 Vybrané parametrické testy (Skripta str. 103-115)

Budeme sledovat parametrické testy pro střední hodnotu, rozptyl a podíl, jako pro IS.

### Střední hodnota (známé $\sigma^2$ )<sup>13</sup>

- V programu `Octave` lze pro tyto testy použít funkci

```
[pval, z] = z_test (x, m, v, alt),  
[pval, z] = z_test_2(x, y, v_x, v_y, alt),
```

kde `pval` je p-hodnota, `z` je hodnota statistiky, `x,y` je výběr, `v,v_x,v_y` je rozptyl souboru, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ ).

### Střední hodnota (neznámé $\sigma^2$ )<sup>14</sup>

- V programu `Octave` lze pro tyto testy použít funkci

```
[pval, t, df] = t_test (x, m, alt),  
[pval, t, df] = t_test_2s (x, y, alt),  
[pval, t, df] = t_test_2n (x, y, alt),  
[pval, t, df] = t_test_2p (x, y, alt),
```

kde `pval` je p-hodnota, `t` je hodnota statistiky, `df` jsou stupně volnosti, `x,y` je výběr, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ ).

### Rozptyl<sup>15</sup>

- V programu `Octave` lze pro tyto testy použít funkci

```
[pval, ch2, df] = var_test (x, v0, alt),  
[pval, f, df_num, df_den] = var_test_2 (x, y, alt),
```

kde `pval` je p-hodnota, `ch2,t` je hodnota statistiky, `df,df_num,df_den` jsou stupně volnosti, `x,y` je výběr, `v0` je rozptyl podle nulové hypotézy, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ ).

### Podíl<sup>16</sup>

U tohoto testu se v literatuře obvykle uvádí trochu jiná statistika. V zájmu jednotnosti jsme ponechali stejnou statistiku, jako pro odhad (což je zvykem). Rozdíly jsou zanedbatelné.

- V programu `Octave` lze pro tento test použít funkci

---

<sup>13</sup>Skripta str. 103

<sup>14</sup>Skripta str. 105,108-113

<sup>15</sup>Skripta str. 106

<sup>16</sup>Skripta str. 107,113-114

```
[pval, z] = prop_test_2 (x1, n1, x2, n2, alt),
```

kde `pval` je p-hodnota, `z` je hodnota statistiky, `x1,x2` jsou výběrové podíly (počty), `n1,n2` jsou počty pokusů, `alt` typ testu ( $<$ ,  $>$ ,  $<>$ ).

## 6 Chi2 testy hypotéz

Chi2 testy jsou založeny na porovnání rozdílnosti mezi naměřenými četnostmi a četnostmi ideálními. Rozdíl je měřen pomocí normovaného kvadratického kritéria, které má ch2 rozdělení – odtud chi2 testy. Pro test se používají absolutní četnosti  $O$  výskytu sledovaného znaku, tzv. pozorované (observed) četnosti a absolutní četnosti  $E$ , které přesně odpovídají  $H_0$ , tzv. teoretické nebo očekávané četnosti (expected).

Pro naměřené četnosti  $O_i$ ,  $i = 1, 2, \dots, n$  a teoretické četnosti  $E_i$ ,  $i = 1, 2, \dots, n$  má statistika tvar

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim Chi2(n-1) \quad (27)$$

Statistika měří vzdálenost mezi pozorovanými a teoretickými četnostmi (je nezáporná). Jsou-li četnosti stejné, rovná se nule. Čím více jsou četnosti jiné, tím je hodnota statistiky větší. Nulová hypotéza testu je shoda četností, alternativní hypotéza je neshoda. Test je vždy pravostranný a kritickým oborem

$$W = (\chi_\alpha^2, \infty)$$

a p-hodnotou

$$p_v = P(\chi^2 > \chi_r^2).$$

Nejvýznamnější aplikace tohoto testu jsou pro testování typu rozdělení (test dobré shody) a testování nezávislosti dvou rozdělení (test nezávislosti). Oba testy vyložíme na příkladech.

### 6.1 Test dobré shody (Skripta str. 115-117)

Hodnoty testované náhodné veličiny rozdělíme na intervaly a měříme četnosti výskytu realizací náhodné veličiny na těchto intervalech. Tak získáme pozorované četnosti  $O$ . Teoretické četnosti  $E$  určíme bud' pomocí hodnot distribuční funkce nebo jinak (jako v následujícím příkladě).

PŘÍKLAD: **(Test rovnoměrnosti)** Sledujeme nehodovost na pražských silnicích během různých dní týdne. Po určité době jsme shromáždili následující údaje

den	počet nehod
Po - Pá	1879
So	421
Ne	406

Na hladině významnosti 0.05 testujte tvrzení (nulové hypotézy), že nehody se během týdne vyskytují rovnoměrně.

Pozorované četnosti jsou zadány.

$$O = [1879, 421, 406]$$

Teoretické četnosti určíme takto: máme 3 intervaly s délkami  $[5, 1, 1]$ . Celkový počet pozorování je  $1879 + 421 + 406 = 2706$ . Tento počet pozorování máme rozdělit na dané intervaly rovnoměrně. První bude mít  $5/7$  druhý  $1/7$  a třetí také  $1/7$ . Bude tedy

$$E = 2706 \times [5/7, 1/7, 1/7] = [1932.86, 386.57, 386.57]$$

Hodnota statistiky

$$\chi_r^2 = \frac{(1879 - 1932.86)^2}{1932.86} + \frac{(421 - 386.57)^2}{386.57} + \frac{(406 - 386.57)^2}{386.57} = 5.54$$

Kritický obor:  $W = (\chi_\alpha^2(3-1); \infty) = (5.99; \infty)$  [5.99=chisquare\_inv(0.95,2)]

P-hodnota:  $p_v = P(\chi^2 > 5.54) = 0.063$  [0.063=1-chisquare\_cdf(5.54,2)]

Závěr: Nulovou hypotézu nelze vyvrátit. Zůstává v platnosti tvrzení: "nehody se vyskytují rovnoměrně".

**PŘÍKLAD:** (**Test normality**) Testujeme, zda rychlosti osobních automobilů jedoucích po nuselském mostě mají normální rozdělení s  $\mu = 60$  a  $\sigma = 7.6$ . Výsledky naměřených rychlostí jsou v tabulce

Interval rychlosti [km/h]	(20-50)	(50-60)	(60-70)	(70-120)
Pozorovaná četnost	35	268	315	21

Pozorované četnosti:  $O = [35, 268, 315, 21]$

Teoretické četnosti určíme pomocí distribuční funkce normálního rozdělení. Normujeme hranice intervalů  $h_i = [20, 50, 60, 70, 120]$

$$x_i = \frac{h_i - \mu}{\sigma}, \Rightarrow x = [-5.263, -1.316, 0, 1.316, 7.895]$$

Odpovídající hodnoty distribuční funkce standardního normálního rozdělení jsou

[0.094=stdnormal\_cdf(-1.316)]

$$F(x) = [0, 0.094, 0.5, 0.906, 1].$$

Pravděpodobnosti  $p_i$  intervalů dostaneme jako rozdíly  $\zeta_{i+1} - \zeta_i$ ,  $i = 2, 3, 4, 5$

$$p_i = [0.094, 0.406, 0.406, 0.094].$$

Z celkového počtu pozorování  $35 + 268 + 315 + 21 = 639$  tedy jednotlivým intervalům přísluší teoretické četnosti

$$E = 639 \times [0.094, 0.406, 0.406, 0.094] = [60.06, 259.43, 259.43, 60.06]$$

Hodnota statistiky:  $\chi_r^2 = 48.05$

Kritický obor:  $W = (7.82; \infty)$

P-hodnota:  $p_v = P(\chi^2 > 48.05) = 2.10^{-10}$

Závěr: Zjištěné četnosti ani náhodou nepochází z normálního rozdělení se střední hodnotou 60 a směrodatnou odchylkou 7.6.

•• V programu Octave lze pro tento test použít funkci

```
[pval, ch2, df] = chisquare_test_homogeneity (x, y, c),
[pval, ch2] = chisquare_test(o,e),
```

kde pval je p-hodnota, ch2 je statistika, df počet stupňů volnosti, x,y výběry, c intervaly pro určení četností, o,e pozorované a teoretické četnosti.

## 6.2 Test nezávislosti (Skripta str. 118-119)

Používá kontingenční tabulkou absolutních četností dvou náhodných veličin, jejichž nezávislost testujeme. Podle definice nezávislosti  $f(x, y) = f(x)f(y)$  určuje tabulkou teoretických (nezávislých) četností takto:

- tabulkou normalizuje na pravděpodobnosti (dělením prvků celkovým součtem prvků),
- určí marginální četnosti (součty) v sloupcích i řádcích,
- vypočte tabulkou nezávislých pravděpodobností (prvek  $(i, j)$  je součinem  $i$ -té sloupové a  $j$ -té řádkové marginálky),
- tabulkou re-normalizuje na absolutní četnosti (násobením všech prvků celkovým součtem původních prvků). Test je pravostranný a má  $(n_x - 1)(n_y - 1)$  stupňů volnosti.

Pomocí statistiky (27) se porovnává původní tabulka s tabulkou absolutních četností nezávislých veličin. Statistiku počítáme pro všechny prvky tabulek (srovnáme obě tabulky do vektorů). Nulová hypotéza  $H_0$  je ”jsou nezávislé”.

**PŘÍKLAD:** Testujeme, zda u řidičů osobních automobilů souvisí věk a reakční doba (měřená časem, za který řidič přehlédne křížovatku). Zjištěné údaje jsou sestaveny do následující tabulky

věk $V$ (roky)	1. (18-30)	2. (30-50)	3. (50-70)
reakční doba $R$			
1. menší než 2 sec	56	42	23
2. větší než 2 sec	32	49	37

Nezávislost testujete na hladině významnosti  $\alpha = 0.05$ .

Pozorované četnosti jsou:  $o = [56, 32, 42, 49, 23, 37]$ .

Teoretické četnosti dostaneme podle uvedeného postupu:

- normalizovaná tabulka

$$\begin{array}{ccc} 0.234 & 0.176 & 0.096 \\ 0.134 & 0.205 & 0.155 \end{array}$$

- marginální pravděpodobnosti jsou

$$\begin{array}{ccc} 0.368 & 0.381 & 0.251 \\ & & 0.50628 \\ & & 0.49372 \end{array}$$

- tabulka nezávislých pravděpodobností

$$\begin{array}{ccc} 0.186 & 0.193 & 0.12710 \\ 0.182 & 0.188 & 0.124 \end{array}$$

- tabulka absolutních četností

$$\begin{array}{ccc} 44.552 & 46.071 & 30.377 \\ 43.448 & 44.929 & 29.623 \end{array}$$

Teoretické četnosti:  $e = [44.552, 43.448, 46.071, 44.929, 30.377, 29.623]$

Hodnota statistiky:  $\chi^2_r = 10.315$ .

Kritický obor:  $W = (\chi^2_\alpha((3-1)(2-1)); \infty) = (5.99; \infty)$ .

P-hodnota:  $p_v = 0.006$ .

Závěr: Nulová hypotéza je vyvrácena, reakční doby řidičů jsou závislé na věku.

- V programu Octave lze pro tento test použít funkci

```
[pval, ch2, df] = chisquare_test_independence (X),
```

kde `pval, ch2, df` jsou p-hodnota, statistika, stupně volnosti, `X` je kontingenční tabulka (viz table).

## 7 Další neparametrické testy hypotéz

V minulé kapitole jsme uvedli dva nejznámější Chi2 testy dobré shody a nezávislosti. Oba se opírají o kontingenční tabulky absolutních četností a mají poměrně široké použití.

Nyní se zmíníme o některých speciálních testech, které lze využít pro statistické zpracování dat. Zaměříme se při tom spíše na význam a použití testu, odvození nebudeme provádět.

### 7.1 Test mediánu

**Znaménkový test** Máme výběr  $\mathcal{X} = [X_1, X_2, \dots, X_n]$  ze spojitého rozdělení s neznámým mediánem  $x_{0.5}$ . Testujeme nulovou hypotézu  $H_0 : x_{0.5} = x_0$ , kde  $x_0$  je dané číslo.

Vypočteme

$$D_i = X_i - x_0, \quad i = 1, 2, \dots, n$$

a písmenem  $b$  označíme počet kladných  $D_i$ .  $b$  je statistika s binomickým rozdělením, kterou lze pro  $n \rightarrow \infty$  approximovat normálním rozdělením  $N(n/2; n/4)$ .

Normovaná statistika testu je

$$z = \frac{2b - n}{\sqrt{n}} \sim N(0; 1)$$

a lze ji testovat pomocí  $z$ -testu.

PŘÍKLAD: (Jen demonstrace postupu! Není  $n \rightarrow \infty$ ) Testujeme, zda výběr

$$x = [5.3, 4.2, 6.8, 5.7, 5.1, 3.1]$$

pochází z rozdělení s mediánem  $x_0 = 5$ .

Počet dat výběru:  $n = 6$ .

Rozdíly  $x_i - x_0$  jsou

$$0.3, -0.8, 1.8, 0.7, 0.1, -1.9$$

a z nich  $b = 4$  jsou kladné.

Normovaná statistika

$$z = \frac{2 \times 4 - 6}{\sqrt{6}} = 0.817$$

a p-hodnota pro oboustranný test je  $p_v = 0.207$ .

Závěr testu: Nulová hypotéza "data pochází z rozdělení s mediánem 5" se nepopírá.

• V programu Octave lze pro test použít funkci

```
[pval, b, n] = sign_test (x, y, alt),
```

kde  $pval$  je p-hodnota,  $b$  je statistika (binomického rozdělení),  $n$  je počet stupňů volnosti statistiky  $b$ ,  $x, y$  jsou realizace výběrů,  $alt$  je směrování testu.

## 7.2 Test nezávislosti prvků výběru

**Pořadový test nezávislosti** Uvažujeme výběr  $\mathcal{X}$  o rozsahu  $n$  a určíme jeho výběrový medián  $\hat{x}_{0.5}$ . Test vychází z rozdílů mezi prvky výběru  $X_i, i = 1, 2, \dots, n$  a výběrového mediánu  $\hat{x}_{0.5}$  (srovnej znaménkový test)

$$D_i = X_i - \hat{x}_{0.5} \quad i = 1, 2, \dots, n.$$

Na těchto rozdílech definuje sériu, tj. souvislé posloupnosti prvků se stejným znaménkem mezi dvěma změnami znaménka. Jako statistiku  $b$  definujeme počet sérií v posloupnosti rozdílů  $D$ . Tato statistika má přibližně normální rozdělení  $N(n/2 + 1; \sqrt{n-1}/2)$ .

Normovaná statistika je

$$z = \frac{2b - (n-2)}{\sqrt{n-1}} \sim N(0; 1).$$

Nulovou hypotézu  $H_0$  : ”prvky výběru jsou nezávislé” lze testovat pomocí levostranného  $z$ -testu.

PŘÍKLAD: Testujeme nezávislost prvků výběru

$$x = \{2.4, 2.2, 1.6, 1.8, 1.5, 1.8, 2.2, 2.3, 2.3, 2.5\}$$

o rozsahu  $n = 10$ . Medián výběru je  $\hat{x}_{0.5} = 2.2$  a diference

$$x - \hat{x}_{0.5} = \{0.2, 0.0, -0.6, -0.4, -0.7, -0.4, 0.0, 0.1, 0.1, 0.3\}.$$

V nich je možno nalézt  $b = 3$  série se stejnými znaménky. Statistika tedy je

$$z = \frac{2 \times 3 - (10-2)}{\sqrt{10-1}} = -0.667$$

a p-hodnota  $p_v = 0.25$ . Prvky výběru tedy lze považovat za nezávislé.

P O Z N Á M K A: *Tento test je velmi významný nejen pro testování, zda nás výběr je skutečně nezávislý (jak požaduje definice výběru), ale také například pro test reziduí po výpočtu regresní analýzy, pro ověření kvality regrese, kterou se budeme zabývat v následující kapitole.*

• V programu Octave lze pro test použít funkci

$$[pval, z] = wztest(x),$$

kde  $pval$  je p-hodnota,  $z$  je statistika,  $x$  je testovaný výběr.

## 7.3 Test nezávislosti výběrů

**Pearsonův test** Pro dva náhodné výběry  $\mathcal{X}$  a  $\mathcal{Y}$  o rozsahu  $n$  vypočteme výběrový korelační koeficient  $r = \text{cov}(x, y) / \sqrt{\text{var}(x)\text{var}(y)}$ .

Statistika je

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

a má Studentovo rozdělení s  $n - 2$  stupni volnosti. Pro test  $H_0$  : "výběry jsou nezávislé" lze použít oboustranný  $t$ -test.

P O Z N Á M K A: *Tento test budeme používat i pro ověření výsledku regresní analýzy.*

• V programu **Octave** lze pro test použít funkci

```
s=cor_test(x,y,alt,"p"),
```

kde **s** je struktura, obsahující výsledky testu, **x**, **y** jsou výběry, **alt** je směrování testu ("<", ">", "<>").

**Spearmanův test** Uvažujme dva náhodné výběry  $\mathcal{X}$  a  $\mathcal{Y}$ , oba o rozsahu  $n$ . Pro oba výběry definujeme pořadí  $P$ , resp.,  $Q$ , tj. např. pro  $x = [6.2, 2.8, 4.1]$  je  $p = [3, 1, 2]$ , protože 6.2 je na třetím místě uspořádaného výběru<sup>17</sup>  $x$ , 2.8 na prvním a 4.1 na druhém. Tato pořadí dosadíme do vzorce pro výběrový korelační koeficient a dostaneme statistiku

$$r_S = \frac{\text{cov}(p, q)}{\sqrt{\text{var}(p)\text{var}(q)}} = 1 - \frac{6}{n(n^2 - 1)}S,$$

kde  $S = \sum_{i=1}^n (p_i - q_i)^2$ .

Tato statistika se testuje podle speciálních tabelovaných hodnot Spearmanova testu. Nulová hypotéza  $H_0$  : "výběry jsou nezávislé".

P Ř Í K L A D: Testujte nezávislost rozdělení, z nichž pochází výběry

```
x = [2.5, 3.4, 1.3, 5.8, 3.6, 2.7, 4.3, 5.1, 2.9, 4.5]
```

• V programu **Octave** lze pro test použít funkci

```
struc=cor_test(x,y,alt,"s"),
```

kde **struc**: je struktura, obsahující výsledky testu, **x**, **y**: jsou výběry, **alt**: je typ testu.

**Kendalův test** Uvažujme dva náhodné výběry  $\mathcal{X}$  a  $\mathcal{Y}$  o rozsahu  $n$  a jejich pořadí  $P$  a  $Q$ . Z pořadí sestavíme dvourádkovou matici a její sloupce uspořádáme tak, aby v prvním řádku bylo  $1, 2, \dots, n$ . Druhý řádek uspořádané matice označíme  $R$  a jeho prvky  $r_1, r_2, \dots, r_n$ . Písmenem  $k_i$  označíme počet všech prvků  $r_{i+1}, r_{i+2}, \dots, r_n$ , které jsou větší než  $r_i$ . Dále označíme  $K = \sum_{i=1}^{n-1} k_i$ . Statistika pak je

$$r_K = \frac{4K}{n(n-1) - 1}$$

a testuje se opět podle speciálních hodnot Kendalova testu. Nulová hypotéza  $H_0$  : "výběry jsou nezávislé".

• V programu **Octave** lze pro test použít funkci

---

<sup>17</sup>Uspořádaný výběr dostaneme, jestliže prvky výběru uspořádáme podle velikosti. V našem příkladě je uspořádaný výběr  $\tilde{x} = [2.8, 4.1, 6.2]$ .

```
struc=cor_test(x,y,alt,"k") ,
```

kde `struc`: je struktura, obsahující výsledky testu, `x`, `y`: jsou výběry, `alt`: je typ testu.

## 7.4 Test typu rozdělení

**Kolmogorov-Smirnovův test** Tento test slouží k ověření, zda zkoumané rozdělení má dané rozdělení. Je založen na porovnání distribuční funkce  $F(x)$  daného rozdělení  $X$  a výběrové distribuční funkce  $F_n(x)$ <sup>18</sup>, určené z výběru  $X$  o rozsahu  $n$ .

Statistika testu je definována vztahem

$$k_s = \sup_{x_i \in X} |F_n(x_i) - F(x_i)|$$

a má různá rozdělení, podle typu testované distribuční funkce. Pro důležitá testovaná rozdělení jsou hodnoty rozdělení  $k_s$  tabelovány. Nulová hypotéza  $H_0$  : "rozdělení má předpokládaný typ".

• V programu `Octave` lze pro test použít funkci

```
[pval,ks]=kolmogorov_smirnov_test(x,dist,params,alt),
```

kde `pval` je p-hodnota testu, `ks` je hodnota statistiky, `x` je realizace výběru, `dist` je typ rozdělení ("binomial", "poisson", "uniform", "normal", "exponential", "lognormal" a další), `params` jsou parametry rozdělení, `alt` je typ testu.

PŘÍKLADE:

```
[pval,ks]=kolmogorov_smirnov_test(x,"normal",0,1),
```

kde `alt` je "`<>`" jako předvolba.

**P O Z N Á M K A:** *Velmi důležitý test, neboť řada jiných statistických procedur vyžaduje normalitu!, případně jiné rozdělení. Není radno dělat závěry, pokud nemáme ověřenu platnost předpokladů!!!*

---

<sup>18</sup>Je to schodovitá funkce: nulová do  $x_{(1)}$ , v každém bodě  $x_{(i)}$  má přírůstek  $1/n$ , a od  $x_{(n)}$  dále je rovna jedné.  $x_{(i)}$ ,  $i = 1, 2, \dots, n$  jsou prvky výběru, uspořádané podle velikosti.

## 8 Regresní analýza

Regresní analýza poskytuje nástroj k hledání stochastické závislosti mezi dvojicí náhodných veličin  $X$  – nezávisle proměnná a  $Y$  – závisle proměnná. V nejběžnější (lineární) podobě zkoumá, zda mezi oběma veličinami existuje *lineární vztah*. Velice jednoduše lze také zkoumat např. polynomický nebo exponenciální vztah. Dále se budeme věnovat především lineární regresi, o ostatních se stručně zmíníme později.

### 8.1 Lineární regrese (Skripta str. 121-126)

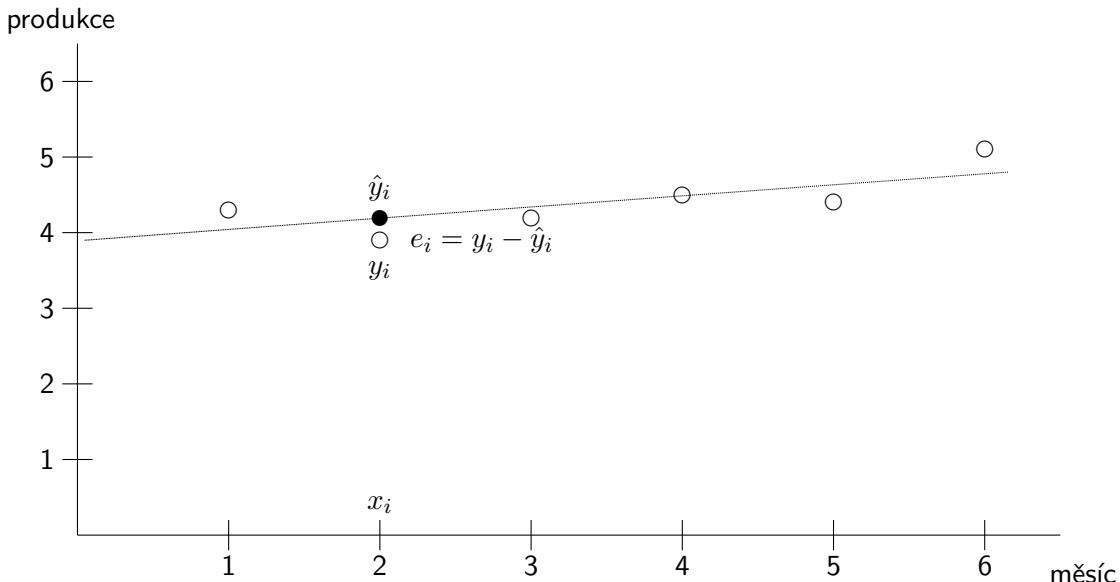
PŘÍKLAD: Sledujeme produkci automobilového závodu během půl roku. Produkce v jednotlivých měsících byla

měsíc	1	2	3	4	5	6
produkce (ks×100)	4.3	3.9	4.2	4.5	4.4	5.1

Odhadněte lineární trend těchto dat a určete, zda mají tendenci k růstu nebo poklesu.

Zadaná data jsou vykreslena na obrázku a je jimi "od oka" proložena přímka. Ihned nás napadnou otázky

- jsou tato data vhodná k approximaci přímkou?
- je nakreslená přímka tou nejlepší, která data approximuje?



Abychom na otázky z příkladu dokázali odpovědět, budeme datové dvojice  $[x_i, y_i]$ ,  $i = 1, 2, \dots, n$  reprezentovat geometricky, jako body v rovině. Přímku, kterou chceme body proložit, budeme uvažovat ve směrnicovém tvaru  $y = b_1 x + b_0$ , kde  $[x, y]$  je libovolný bod přímky,  $b_1$  je směrnice a  $b_0$  absolutní člen (úsek na ose  $y$ ). Optimální přímku nazveme *regresní přímka* a budeme požadovat, aby poloha této regresní přímky vůči datovým bodům minimalizovala určité kriterium vzdálenosti. Abychom mohli být konkrétní, zavedeme následující

pojmy a uvedeme vzorce pro výpočet regresních koeficientů (odvození je v [1] nebo podrobněji v [2]):

**Predikce**  $\hat{y}_i$  je bod, jehož  $x$ -ová souřadnice je  $x_i$  a  $y$ -ová  $b_1x_i + b_0$ , tj. leží na proložené přímce.

**Chyba predikce (reziduum)**  $e_i$  je rozdíl mezi datovým bodem a predikcí, tj. svislá vzdálenost bodu od přímky.

**Model datových bodů** lze pomocí proložené přímky vyjádřit takto: datový bod je predikce plus chyba predikce, tj.

$$y_i = b_1x_i + b_0 + e_i \quad (28)$$

**Kritérium optimality** pro "ideální regresní přímku" definujeme jako součet kvadrátů všech chyb predikce

$$J = \sum_{i=1}^n e_i^2, \quad (29)$$

a požadujeme, aby byl minimální. Chyby predikce odpovídající regresní přímce (tj. jejichž kriterium  $J$  je minimální) nazýváme *rezidua*.

**Koeficienty regresní přímky  $b_1$  a  $b_0$**  jsou

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1\bar{x}, \quad (30)$$

s označením

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

**Korelační koeficient  $r$**  je

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad (31)$$

kde  $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ .

**P O Z N Á M K A:** Uvedený vzorec není v rozporu s tím, který jsme uvedli v souvislosti s Pearsonovým testem. Po vydělení čitatele i jmenovatele výrazem  $n - 1$  dostaneme totéž.

**Význam:**

- Koeficient  $b_1$  vypovídá o trendu regresní přímky. Je-li  $b_1 > 0$  přímka roste, pro  $b_1 < 0$  klesá a v případě  $b_1 = 0$  je vodorovná.
- Korelační koeficient  $r$  nese informaci o tom, jak silná je vazba mezi daty  $x$  a  $y$ . Jeho rozsah je  $r \in (0, 1)$  a je-li nulový, veličiny  $x$  a  $y$  spolu nesouvisí – regrese nemá význam.

Je-li kladný přímka roste, je-li záporný, klesá. Čím více se blíží  $\pm 1$  tím je vazba silnější.  
Pro  $r = \pm 1$  leží datové body na regresní přímce.

PŘÍKLAD: (pokračování) Podle zadání našeho příkladu o produkci automobilů bude

$$\bar{x} = 3.5, \quad \bar{y} = 4.4, \quad S_{xx} = 17.5, \quad S_{yy} = 0.8, \quad S_{xy} = 2.9, \quad b_1 = 0.166, \quad b_0 = 3.82, \quad r = 0.775$$

Predikce v bodech měření (nebo i kdekoliv jinde) určíme tak, že do odhadnuté regresní přímky dosadíme příslušná  $x$

$$\begin{aligned}\hat{y}_i &= b_1 x_i + b_0 = 0.166 x_i + 3.82, \quad i = 1, 2, \dots, n \\ \text{tj.} \quad &3.99, 4.15, 4.32, 4.48, 4.65, 4.82\end{aligned}$$

Hodnota regresního koeficientu  $r = 0.775$ , stejně jako směrnice regresní přímky  $b_1 = 0.166$  potvrzuje, že data mají tendenci nárůstu. Regresní koeficient navíc ukazuje, že data lze dosti dobře approximovat přímou ( $0.775 \rightarrow 1$ ).

• V programu Octave lze pro výpočet regresní přímky použít funkci

```
p = lin_reg(x,y),
[b1,b0,r] = reg_desc(x,y),
ch = reg_info(x,y),
```

kde  $p=[b1,b0]$ ,  $b1,b0,r$  jsou parametry regresní přímky,  $ch$  je struktura s výsledky a  $x,y$  jsou výběry.

## 8.2 Nelineární regrese

Přímka je nejznámější, nikoliv však jedinou funkcí, kterou lze měřenými body prokládat. Jako další si uvedeme polynomickou a exponenciální regresi. Důležité při tom je, abychom dokázali vyjádřit závislost proměnnou  $y$  (nebo její funkci) jako skalární součin parametrů a vektoru, jehož složky jsou funkce  $x$ . Tak dostaneme obecnou regresní rovnici

$$\begin{aligned}\tilde{y}_i &= b_m f_m(x_i) + b_{m-1} f_{m-1}(x_i) + \dots + b_1 f_1(x_i) + b_0 + e_i = \\ &= [\tilde{x}_i^{(m)}, \tilde{x}_i^{(m-1)}, \dots, \tilde{x}_i^{(1)}, 1] \times [b_m, b_{m-1}, \dots, b_1, b_0]' + e_i,\end{aligned}\tag{32}$$

kde  $f_k(x_i) = \tilde{x}_i^{(k)}$ ,  $k = 1, 2, \dots, m$ ,  $i = 1, 2, \dots, n$  jsou příslušné (známé) funkce  $x$ ,  $\tilde{y}_i$ ,  $i = 1, 2, \dots, n$  je (známá) funkce  $y$  a symbol ' značí transpozici.

Koeficienty regresní přímky obdržíme takto:

- Sestavíme rozšířenou datovou matici

$$D = \begin{bmatrix} \tilde{y}_1, & \tilde{x}_1^{(m)}, & \tilde{x}_1^{(m-1)}, & \dots, & \tilde{x}_1^{(1)}, & 1 \\ \tilde{y}_2, & \tilde{x}_2^{(m)}, & \tilde{x}_2^{(m-1)}, & \dots, & \tilde{x}_2^{(1)}, & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \tilde{y}_n, & \tilde{x}_n^{(m)}, & \tilde{x}_n^{(m-1)}, & \dots, & \tilde{x}_n^{(1)}, & 1 \end{bmatrix}.$$

- Vypočteme matici  $V = D'D$ , kde  $D'$  značí matici transponovanou.

- Rozdělíme matici  $V$

$$V = \begin{bmatrix} V_y & V'_{xy} \\ V_{xy} & V_x \end{bmatrix}$$

tak, že  $V_y = V(1, 1)$  je číslo,  $V_{xy} = V(2 : n, 1)$  je sloupcový vektor a  $V_x = V(2 : n, 2 : n)$  je matici.

- Koeficienty regresní přímky jsou

$$b = [b_m, b_{m-1}, \dots, b_0] = V_x^{-1}V_{xy}. \quad (33)$$

P O Z N Á M K A: *Všimněte si, že pro skalární případ tento vzorec souhlasí s dříve uvedeným.*

P Ř Í K L A D: (polynomiální regrese) Regresní přímka má tvar

$$y = b_m x^m + b_{m-1} x^{m-1} + \dots + b_1 x + b_0, \quad (34)$$

který je vzhledem k parametrům lineární.

Datová matice bude

$$D = \begin{bmatrix} y_1, & x_1^m, & x_1^{m-1}, & \dots, & x_1, & 1 \\ y_2, & x_2^m, & x_2^{m-1}, & \dots, & x_2, & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_n, & x_n^m, & x_n^{m-1}, & \dots, & x_n, & 1 \end{bmatrix}.$$

Další postup podle obecného návodu výše.

- V programu Octave lze pro polynomiální regrese použít funkci

$$p = \text{pol\_reg}(x, y, k),$$

kde  $p$  jsou parametry regrese,  $x, y$  jsou výběry,  $k$  je stupeň regresního polynomu.

P Ř Í K L A D: (exponenciální regrese) Regresní přímka má tvar

$$y = a \exp\{b_1 x\}, \quad (35)$$

který je nelineární. Linearizujeme je logaritmováním

$$\ln y = \ln a + b_1 x \Rightarrow \tilde{y} = b_1 x + b_0, \quad (36)$$

kde  $\tilde{y} = \ln y$  a  $b_0 = \ln a$ .

Dále již postupujeme podle obecného návodu.

P O Z N Á M K A: *Pozor! Z linearizované regresní přímky dostaneme logaritmy predikcí  $\ln \hat{y}$ . Skutečné predikce jsou  $\hat{y} = \exp\{\ln \hat{y}\}$ .*

- V programu Octave lze pro polynomiální regrese použít funkci

$$p=\exp\_reg(x, y),$$

kde  $p$  jsou parametry regrese,  $x, y$  jsou výběry.

## 9 Korelační analýza (Skripta str. 127-141)

V předchozí kapitole jsme ukázali, jak k daným výběrům  $x$  a  $y$  sestrojit regresní přímku. Zároveň jsme konstatovali, že korelační koeficient  $r$  vypovídá o kvalitě regrese. Přesnější vyjádření, založené na intervalech spolehlivosti a testech hypotéz, poskytuje korelační analýza.

Pro další úvahy budeme předpokládat, že skutečně existuje *ideální regresní přímka*  $y = \beta_1 x + \beta_0$  a že naše regresní přímka  $y = b_1 x + b_0$  tuto ideální přímku odhaduje.

V tomto smyslu jsou  $b_1$  a  $b_0$  bodovými odhady  $\beta_1$  a  $\beta_0$  a predikce  $\hat{y}_i$  je odhadem hodnoty ideální regresní přímky v bodě  $x_i$ .

P O Z N Á M K A: *Pro lepší pochopení je možno si představit, že skutečný proces generuje data přesně na přímce (tj. je popsán ideální regresní přímkou) a my tato data měříme s chybami. Úkolem regresní analýzy je dobrat se přes zašuměná data ke skutečnému procesu.*

### 9.1 Intervaly spolehlivosti (Skripta str. 136,138-139)

Tyto intervaly lze využít pro ověření kvality regrese.

#### Interval pro $\beta_1$

Je to interval se středem v  $b_1$ , ve které bude ležet  $\beta_1$  s pravděpodobností  $1 - \alpha$ .

Interval spolehlivosti:

$$\beta_1 \in b_1 \pm \frac{s_e}{\sqrt{S_{xx}}} t_{\alpha/2},$$

kde

$$s_e = \sqrt{\frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}}, \quad \text{a } t_{\alpha/2} \text{ je krit. hod. } St(n-1)$$

VYUŽITÍ: obsahuje-li interval nulu, data nejsou příliš vhodná pro lineární regresi.

#### Interval pro hodnotu regresní přímky

Pro opakování výběry dostaneme různé regresní přímky. Tento interval pro hodnoty  $y$  v pevném bodě  $x_p$  je interalem, kterým bude procházet  $(1 - \alpha) \times 100\%$  všech možných regresních přímek.

Interval spolehlivosti:

$$y(x_p) \in \hat{y}_p \pm s_e \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$

VYUŽITÍ: Čím širší je tento interval, tím méně vhodná je lineární regrese.

P O Z N Á M K A: Často se uvádí celý pás spolehlivosti, tvořený intervaly na husté síti bodů  $x$ .

⊕ V programu Octave lze pro intervaly spolehlivosti v regresi použít funkci

```
[is_e,is_p,pval_a,pval_r]=reg_infe(x,y,xp,alpha,alt),
```

kde `is_e, is_p` jsou IS pro střední hodnotu a predikci, `pval_a, pval_r` jsou p-hodnoty pro směrnici a korelační koeficient, `x, y` výběry, `xp` pevný bod x, `alpha` hladina významnosti pro test, `alt` směrování testu.

## 9.2 Testy hypotéz (Skripta str. 134-135,140)

Testy ověřují vhodnost dat pro lineární regresi.

### t-test korelačního koeficientu (Pearsonův test)

Testujeme vhodnost lineární regrese. Statistikou je výběrový korelační koeficient  $r$ .

Normovaná statistika:

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

Test je oboustranný, s  $H_0$  : ”data nejsou vhodná pro regresi”.

P O Z N Á M K A: *Tento test jsme již uvedli mezi neparametrickými testy.*

### F-test poměru vysvětleného a nevysvětleného rozptylu

Je velmi kvalitním testem vhodnosti regrese, který lze použít i v případě více nezávislých proměnných. Jeho podstata je následující:

Definujeme

celkový součet čtverců odchylek dat od průměru

$$S_y = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{yy},$$

regresní součet čtverců odchylek predikcí od průměru

$$S_{\hat{y}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = S_{xy}^2 / S_{xx},$$

reziduální součet čtverců odchylek dat od predikcí (tj. od přímky)

$$S_{\hat{e}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = S_{yy} - S_{xy}^2 / S_{xx} = S_y - S_{\hat{y}}.$$

Celkový součet vypovídá o rozptýlenosti změrených dat bez ohledu na regresi. Regresní součet čtverců ukazuje, kolik původního rozptylu lze vysvětlit na základě regrese. Reziduální součet vypovídá o zbylému, tedy nevysvětleném, rozptylu.

V tomto testu se uvažuje statistika

$$F = \frac{(n - 2)S_{\hat{y}}}{S_{\hat{e}}} \sim F(1, n - 2),$$

což je podíl rozptylu vysvětleného regresí a nevysvětleného. Statistika má rozdělení  $F$  se stupni volnosti 1 a  $n - 2$ . Nulovou hypotézu  $H_0$  : "data nejsou vhodná pro regresi" testujeme pomocí pravostranného F-testu.

### **z-test nezávislosti reziduí**

Použijeme Pořadový test nezávislosti podle odstavce 7.2.

• V programu Octave lze pro testy v regresi použít funkci

```
[is_a,is_e,pval_a,pval_r] = reg_infe(x,y,xp,alpha,alt),  
[pval,t,df] = t_test_reg(x,y,alt),  
[pval, F, df_num, df_den] = f_test_reg(x,y),  
[pval,z] = wztest(x)
```

kde `is_a, is_e` jsou IS, `pval` je p-hodnota testů, `t, F, z` statistiky, `df` stupně volnosti, `x, y` výběry, `xp` pevný bod x, `alpha` hladina významnosti pro test, `alt` směrování testu.

# 10 Analýza rozptylu (ANOVA)

Podobně jako korelační analýza, která z rozboru celkového rozptylu dat dělá závěry o funkci regrese, pracuje i analýza rozptylu (ANalysis Of VAriance) z celkovým rozptylem dat, rozkládá jej do jednotlivých tříd a dělá závěry o těchto třídách.

## 10.1 ANOVA při jednoduchém trídění

Uvažujeme několik podobných zdrojů dat a chceme se přesvědčit, že všechny tyto zdroje fungují stejně, tj. průměry dat změrených na jednotlivých zdrojích jsou stejné. Postup testu uvedeme podle příkladu.

PŘÍKLAD: Testujeme tři automobily stejné značky pro závod do vrchu. Vlivem nestejných podmínek (různí řidiči, povrch vozovky, atd.) se naměřené časy liší. Naším úkolem je zjistit, zda rozdíly v různých průměrných časech při opakovaných jízdách jsou způsobeny rozdílnou kvalitou automobilů nebo je lze přičíst na vrub náhodným vlivům. Data, která jsme naměřili, jsou v tabulce

Časy při opakovaných jízdách automobilů (min)					
automobil 1	5.32	5.24	5.47	4.98	5.16
automobil 2	5.88	5.31	4.86	5.45	5.12
automobil 3	5.32	4.21	5.44	5.33	5.24

Označíme:

$a$  počet tříd (automobilů),  $n$  počet měření,

$x_i = [x_{1;i}, x_{2;i}, \dots, x_{n;i}]$  data od jednotlivých automobilů,

$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{j;i}$  průměry dat od jednotlivých automobilů,

$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{j;i} - \bar{x}_i)^2$  výběrové rozptyly dat jednotlivých automobilů,

$\bar{\bar{x}} = \frac{1}{a} \sum_{i=1}^a \bar{x}_i$  průměr z průměrů pro jednotlivé automobily,

pro  $i = 1, 2, \dots, a$ .

Definujeme:

**rozptyl mezi třídami**  $s_x^2$  je výběrový rozptyl jednotlivých průměrů

$$s_x^2 = \frac{1}{a-1} \sum_{i=1}^a (\bar{x}_i - \bar{\bar{x}})^2,$$

**rozptyl uvnitř tříd**  $s_P^2$  je průměr jednotlivých výběrových rozptylů

$$s_P^2 = \frac{1}{a} \sum_{i=1}^a s_i^2.$$

**Statistika pro test** je dána podílem

$$F = \frac{n s_x^2}{s_P^2} \sim F(a-1, a(n-1)),$$

což je podíl rozptylu vysvětleného třídami a rozptylu nevysvětleného. Statistika má rozdělení  $F$  se stupni volnosti  $a-1$  a  $a(n-1)$ .

Nulová hypotéza  $H_0$ : "střední hodnoty tříd jsou stejné". Test je pravostranný.

PŘÍKLAD: **(dokončení)** V našem příkladě o závodních automobilech je  $a = 3$ ;  $n = 5$ ; průměry  $\bar{x}_i : 5.23, 5.32, 5.11$ ; rozptyly  $s_i^2 : 0.033, 0.145, 0.257$ .

rozptyl mezi třídami  $s_x^2 = 0.0118$ ; rozptyl uvnitř tříd  $s_P^2 = 0.145$ .

Statistika:  $F_r = \frac{5 \times 0.0118}{0.145} = 0.405$

p-hodnota:  $p_v = P(F > F_r) = 0.676$   
při použití rozdělení  $F(a-1, a(n-1)) = F(2, 12)$ .

Závěr testu: automobily jsou stejné.

⊕ V programu Octave lze pro analýzu rozptylu s jednoduchým tříděním použít funkci

```
pval = anova(x),
```

kde pval je p-hodnota, x data (matice se skupinami ve sloupcích).

## 10.2 ANOVA při dvojném třídění

Uvažujeme podobně jako v předchozím několik zdrojů dat. Rozdíl je v tom, že na tyto zdroje mohou mít vliv ještě další faktory. Cílem je určit, zda rozdíly v datech je možno vysvětlit vlivem těchto dalších faktorů, nebo zda zdroje jsou skutečně různé.

PŘÍKLAD: Testujeme tři automobily stejné značky pro závod do vrchu a máme k dispozici pět řidičů. Každého z řidičů necháme vyjet závodní dráhu se všemi automobilemi a zaznamenáme jejich časy. Ty jsou uvedeny v tabulce

Časy při jízdách automobilů (min)					
	řidič 1	řidič 2	řidič 3	řidič 4	řidič 5
automobil 1	5.32	5.24	5.47	4.98	5.16
automobil 2	5.88	5.31	4.86	5.45	5.12
automobil 3	5.32	4.21	5.44	5.33	5.24

Naším úkolem je zjistit, zda rozdíly v různých průměrných časech při jízdách automobilem jsou způsobeny rozdílnou kvalitou automobilem nebo je lze přičíst na vrub rozdílům mezi řidiči.

**P O Z N Á M K A:** V příkladě pro jednoduché třídění jsme se snažili rozdílnosti v časech vysvětlit pouze rozdílností automobilů. Nyní rozdíly v časech vysvětlujeme jednak rozdílností automobilů, ale také rozdílností řidičů.

Označíme:

$a$  počet automobilů,  $b$  počet řidičů,

$x_{i,j}$  je čas  $i$ -tého automobilu s  $j$ -tým řidičem

$\bar{x}_{\bullet j} = \frac{1}{a} \sum_{i=1}^a x_{i,j}$  je průměrný čas  $j$ -tého řidiče (přes všechny automobily),

$\bar{x}_{i\bullet} = \frac{1}{b} \sum_{j=1}^b x_{i,j}$  je průměrný čas  $i$ -tého automobilu (se všemi řidiči),

$\bar{\bar{x}}$  je celkový průměrný čas z celé tabulky.

Definujeme:

### rozptyl mezi průměry automobilů

$$s_A^2 = \frac{b}{a-1} \sum_{i=1}^a (\bar{x}_{i\bullet} - \bar{\bar{x}})^2,$$

### rozptyl mezi průměry řidičů

$$s_B^2 = \frac{a}{b-1} \sum_{j=1}^b (\bar{x}_{\bullet j} - \bar{\bar{x}})^2,$$

### reziduální rozptyl (uvnitř tříd)

$$s_R^2 = \frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b (x_{i,j} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{\bar{x}})^2,$$

který je vypočten z reziduí – rozdílů mezi daty  $x_{i,j}$  a jejich předpověďmi  $\hat{x}_{i,j}$ , kde

$$\hat{x}_{i,j} = \underbrace{\bar{\bar{x}}}_{\text{celkový průměr}} + \underbrace{(\bar{x}_{i\bullet} - \bar{\bar{x}})}_{\text{efekt auta}} + \underbrace{(\bar{x}_{\bullet j} - \bar{\bar{x}})}_{\text{efekt řidiče}}$$

### Statistika pro test automobilů

$$F_A = \frac{s_A^2}{s_R^2} \sim F(a-1, (a-1)(b-1)),$$

což je podíl rozptylu vysvětleného rozdíly v automobilech a rozptylu nevysvětleného. Statistika má rozdělení  $F$  se stupni volnosti  $a-1$  a  $(a-1)(b-1)$ .

Nulová hypotéza  $H_0$ : "střední hodnoty pro automobily jsou stejné". Test je pravostranný.

**Statistika pro test řidičů** je dána podílem

$$F_B = \frac{s_B^2}{s_R^2} \sim F(b-1, (a-1)(b-1)),$$

což je podíl rozptylu vysvětleného rozdíly v řidičích a rozptylu nevysvětleného. Statistika má rozdělení  $F$  se stupni volnosti  $b-1$  a  $(a-1)(b-1)$ .

Nulová hypotéza  $H_0$ : "střední hodnoty pro řidiče jsou stejné". Test je pravostranný.

**P O Z N Á M K A:** *Na první pohled by se mohlo zdát, že ANOVA s dvojným tříděním provádí pouze dva paralelní testy pro dvě veličiny, které mají vliv na sledovaná data. Není to však pravda. V každém z obou testů se berou v úvahu vlivy obou veličin. To co jsme dříve museli prohlásit za nevysvětlený rozptyl (který byl třeba proti vysvětlenému příliš veliký), lze nyní vysvětlit pomocí druhé veličiny. Tím se nevysvětlený rozptyl zmenší a test může dopadnout zcela jinak.*

**P Ř Í K L A D:** **(dokončení)** V našem příkladě o závodních automobilech je  $a = 3$ ,  $b = 5$ ,

průměry pro automobily  $\bar{x}_{i\bullet}$  : 5.23 5.32 5.11,

průměry pro řidiče  $\bar{x}_{\bullet j}$  : 5.51 4.92 5.26 5.25 5.17.

$$\begin{aligned} \text{rozptyl mezi automobily } s_A^2 &= 0.059, & \text{rozptyl mezi řidiči } s_B^2 &= 0.132, \\ \text{reziduální rozptyl } s_R^2 &= 0.152. \end{aligned}$$

Statistika:  $F_{aut} = 0.388$ , p-hodnota<sub>aut</sub> = 0.69

Statistika:  $F_{rid} = 0.874$ , p-hodnota<sub>rid</sub> = 0.52

Závěr testu: ani automobily ani řidiči nejsou odlišní.

• V programu **Octave** lze pro analýzu rozptylu s dvojným tříděním použít funkci

```
[pv_col, pv_row] = anova_2(x),
```

kde **pv\_col**, **pv\_row** jsou p-hodnoty, **x** data (matice s prvním faktorem ve sloupcích a druhým faktorem v řádcích).

## 11 Analýza hlavních komponet

Tato úloha provádí transformaci měřených dat na menší počet tzv. fiktivních dat tak, aby většina informace obsažená v původních datech zůstala zachována. Jedná se tedy o úlohu redukce dat, potřebných pro popis sledovaného znaku.

PŘÍKLADEM: Stav dopravní oblasti lze určovat na základě detektorů, umístěných v blízkosti křižovatek oblasti. Tyto detektory jsou většinou umístěny v každém dopravním pruhu a často v několika vzdálenostech od křižovatky. Je zřejmé, že takových měřících míst je v oblasti velké množství. Přitom ale, informace kterou nesou musí být závislá – např. u dvou detektorů postavených za sebou jsou signály jen zpožděné. Naším cílem je redukovat "skutečné detektory" oblasti na menší počet "fiktivních detektorů", které obdržíme "kombinací skutečných detektorů". Přitom požadujeme, aby tyto fiktivní detektory nesly téměř stejnou informaci o stavu oblasti, jako detektory skutečné.

V následujících odstavcích uvedeme dva možné způsoby jak zmíněnou redukci provést. Budeme se při nich opírat o rozklad matice pomocí vlastních a singulárních čísel. Proto nejprve stručně uvedeme ty výsledky, které budeme dále potřebovat.

### Definice 11.1 (Rozklad pomocí vlastních čísel)

Pro pozitivně definitní, symetrickou matici  $R$  existuje rozklad

$$R = V L V', \quad (37)$$

kde

$L$  je diagonální matice s diagonálou tvořenou vlastními čísly matice  $R$ ,

$V$  je čtvercová matice jejích sloupce jsou tvořeny vlastními vektory matice  $R$ , odpovídajícími vlastním číslům matice  $L$ .

Matice  $V$  je ortogonální, tj. platí pro ni  $V V' = V' V = I$  s jednotkovou maticí  $I$ . Z vlastnosti ortogonalita plyne, že platí  $V^{-1} = V'$

• V programu Octave lze pro tento rozklad použít funkci

$$[V, L] = \text{eig}(R).$$

### Definice 11.2 (Rozklad pomocí singulárních čísel)

Pro matici  $D$  typu  $N \times n$  existuje rozklad

$$D = U S V', \quad (38)$$

kde

$U$ , resp.  $S$ , jsou ortogonální matice typu  $N \times N$ , resp.  $n \times n$ ,

$S$  je matice typu  $N \times n$ , která má na hlavní diagonále tzv. singulární čísla matice  $D$  a jinak samé nuly.

Pokud v matici  $S$  vynecháme nulové řádky a sloupce matic  $U$  a  $V$ , které by se v součinu "potkali" s nulami matice  $S$ , budou mít matice  $U$ ,  $S$  a  $V$  rozměry  $N \times n$ ,  $n \times n$  a  $n \times n$ . Ty nazýváme zkrácené vyjádření rozkladu.

• V programu `Octave` lze pro tento rozklad použít funkci

$$[U, S, V] = \text{svd}(D).$$

## 11.1 Rozklad kovarianční matice

První způsob, který uvedeme je založen na rozkladu výběrové kovarianční matice dat podle (37). V časových okamžicích  $t = 1, 2, \dots, N$  měříme datový vektor

$$d_t = [d_{1;t}, d_{2;t}, \dots, d_{n;t}],$$

např. údaje z detektorů podle zmíněného příkladu.

Měřené datové vektory sestavíme do datové matice  $D$ , resp., centované datové matice  $D_c$

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} d_{1;1} & d_{2;1} & \dots & d_{n;1} \\ d_{1;2} & d_{2;2} & \dots & d_{n;2} \\ \vdots & \vdots & \dots & \vdots \\ d_{1;N} & d_{2;N} & \dots & d_{n;N} \end{bmatrix}, \quad \text{resp.,} \quad D_c = \begin{bmatrix} d_1 - \bar{d} \\ d_2 - \bar{d} \\ \vdots \\ d_N - \bar{d} \end{bmatrix},$$

kde  $\bar{d}$  je vektor výběrových průměrů  $\bar{d} = \frac{1}{N} \sum_{t=1}^N d_t$ .

Dále určíme výběrovou kovarianční matici dat

$$R = \frac{1}{n-1} \sum_{t=1}^N (d_t - \bar{d})'(d_t - \bar{d}) = \frac{1}{n-1} D_c' D_c,$$

kterou rozložíme podle (37) takto

$$R = V L V',$$

kde

$V$  je matice typu  $n \times n$ , která má ve sloupcích vlastní vektory matice  $R$ ,

$L$  je diagonální matice stupně  $n$  s vlastními čísly matice  $R$  na diagonále.

Zavedeme transformaci dat

$$g_t = (d_t - \bar{d}) V \quad \text{nebo maticově} \quad G = D_c V,$$

kde matice  $G$  má za řádky transformovaná data  $g_t$ ,  $t = 1, 2, \dots, N$ .

Snadno lze ukázat, že kovarianční matice transformovaných dat  $G$  je diagonální a na diagonále má vlastní čísla kovarianční matice  $R$

$$\frac{1}{n-1} G' G = \frac{1}{n-1} V' D_c' D_c V = V' R V = V' (V L V') V = L,$$

protože matice  $V$  je ortogonální, tj. platí  $V' V = I$ , kde  $I$  je jednotková matice.

Nová (transformovaná) data jsou tedy navzájem nezávislá a mají výběrové rozptyly rovny vlastním číslům matice  $R$ .

Kovarianční matici dat (podobně jako rozptyl u regresní analýzy nebo analýzy rozptylu) chceme nyní vysvětlit. Protože transformované datové veličiny jsou nezávislé, vysvětlují kovarianční matici samostatně, každá zvlášť svým vlastním rozptylem. Pokud jsou datové veličiny závislé, projeví se to tím, že některé transformované veličiny budou mít rozptyly vzhledem k ostatním velmi malé. Tyto veličiny pak můžeme vynechat a pracovat jen s těmi, které měly rozptyl větší. Přitom většinou požadujeme, aby bylo vysvětleno např. 99% celé kovarianční matice.

Prakticky postupujeme např. takto:

1. Vlastní čísla (prvky na diagonále matice  $L$ ) uspořádáme podle velikosti od největšího k nejmenšímu.
2. Stejným způsobem, jako jsme přehazovali vlastní čísla, přeházíme i vlastní vektory, tj. sloupce matice  $V$ .
3. Vlastní čísla normujeme tak, aby jejich součet byl roven jedné.
4. Postupně sčítáme normovaná vlastní čísla od největšího a hlídáme, kdy jejich součet překročí stanovenou hranici (např. 0.99).
5. Jako nové proměnné vezmeme ty transformované veličiny, jejichž rozptyly jsme v předchozím kroku posčítali.

**PŘÍKLAD:** Změřili jsme 1000 vzorků dopravních dat (intenzity a hustoty dopravního proudu ve strahovském tunelu). Každé měření zahrnovalo 5 hodnot. Určili jsme kovarianční matici dat

$$R = \begin{bmatrix} 87.7 & 41.9 & 22.4 & 202.4 & 23.5 \\ 41.9 & 350.1 & 6.1 & 95.9 & 8.5 \\ 22.4 & 6.1 & 9.0 & 68.3 & 6.9 \\ 202.4 & 95.9 & 68.3 & 2194.0 & 53.7 \\ 23.5 & 8.5 & 6.9 & 53.7 & 10.6 \end{bmatrix}$$

Její rozklad  $R = V L V'$  je

$$L = \begin{bmatrix} 2222.2 & 0 & 0 & 0 & 0 \\ 0 & 348.8 & 0 & 0 & 0 \\ 0 & 0 & 73.6 & 0 & 0 \\ 0 & 0 & 0 & 4.2 & 0 \\ 0 & 0 & 0 & 0 & 2.6 \end{bmatrix} \quad V = \begin{bmatrix} 0.096 & 0.111 & 0.921 & -0.349 & 0.093 \\ 0.053 & 0.991 & -0.119 & 0.015 & -0.009 \\ 0.032 & 0.013 & 0.235 & 0.396 & -0.887 \\ 0.993 & -0.065 & -0.097 & -0.002 & 0.008 \\ 0.025 & 0.023 & 0.271 & 0.849 & 0.452 \end{bmatrix}$$

Normovaná vlastní čísla a jejich kumulativní součet jsou

$$L_{norm} = [0.975 \ 0.024 \ 0.001 \ 3.5e-6 \ 1.3e-6] \quad \text{a} \quad L_{cum} = [0.975 \ 0.999 \ 1 \ 1 \ 1].$$

Z vektoru  $L_{cum}$  je patrné, že pro pokrytí 99% celé kovarianční matice postačí uvažovat první dvě transformované proměnné. Označíme-li  $D_c$  matici  $1000 \times 5$  s původními centroványmi proměnnými  $d_1, \dots, d_5$  ve sloupcích,  $G_r$  matici  $1000 \times 2$  vybraných transformovaných proměnných  $g_{r,1}, g_{r,2}$  a matici

$V_r$  jako submatice matice  $V$  obsahující prvé dva sloupce (tj. vlastní vektory odpovídající vybraným vlastním čísly) bude platit

$$G_r = D_c V_r = [d_{c,1}, d_{c,2}, d_{c,3}, d_{c,4}, d_{c,5}] \begin{bmatrix} 0.096 & 0.111 \\ 0.053 & 0.991 \\ 0.032 & 0.013 \\ 0.993 & -0.065 \\ 0.025 & 0.023 \end{bmatrix}$$

Odtud je také patrno, jak jsou původní proměnné kombinovány.  $g_{r,1}$  je tvořena převážně ze čtvrté a  $g_{r,2}$  z druhé původní veličiny. V tomto případě lze tedy místo s dvěma transformovanými veličinami pracovat s druhou a čtvrtou původní veličinou. Ostatní již mnoho nové informace nenesou.

## 11.2 Rozklad datové matice

Jiný způsob, jak provést redukci dat přímo ve spojení s regresní analýzou těchto dat, je založen na rozkladu datové matice pomocí tzv. SVD (Singular Value Decomposition) rozkladu (38). Uvažujme regresi pro  $t = 1, 2, \dots, N$

$$y_t = d_t \theta + e_t \quad \text{nebo maticově} \quad Y = D \theta + E, \quad (39)$$

kde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad D = \begin{bmatrix} d_{1;1} & \dots & d_{n,1} \\ \vdots & \dots & \vdots \\ d_{1;N} & \dots & d_{n;N} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}.$$

Datovou matici  $D$  typu  $N \times n$  rozložíme podle (38)

$$D = U S V, \quad (40)$$

kde

$U$ , resp.  $V$ , jsou ortogonální matice typu  $N \times N$ , resp.  $n \times n$ ,

$S$  je matice typu  $N \times n$ , která má na hlavní diagonále singulární čísla matice  $D$  a jinak nuly.

Dosadíme-li tento rozklad do regresní rovnice (39), dostaneme

$$Y = (U S) (V \theta) + E,$$

kde  $US = DV'$  (protože  $V^{-1} = V'$  - ortogonalita) a  $V\theta = \tilde{\theta}$  jsou modifikované parametry.

Nyní je situace obdobná jako pro vlastní čísla kovarianční matice dat. Jsou-li datové veličiny závislé, budou některá singulární čísla malá a lze je nahradit nulami. Počet nenulových singulárních čísel označíme  $r$ . Jestliže v regresní rovnici vynecháme sloupce a řádky jednotlivých matic, které by po vynásobení byly stejně nulové, bude (ponecháme stejně názvy)

$U$  ortogonální matice typu  $N \times r$ ,

$S$  diagonální matice typu  $r \times r$  a

$V$  ortogonální matice typu  $r \times n$ .

Vektor  $\tilde{\theta} = V\theta$  dimenze  $r$  představuje transformované parametry regrese a matice  $G = US$  typu  $N \times r$  transformovaná data. S využitím rovnice (40) pro rozklad datové matice a skutečnosti, že matice  $V$  je ortogonální, a tedy platí  $V^{-1} = V'$ , lze psát  $US = DV'$ . Matici transformovaných dat  $G$  lze tedy vyjádřit také jako transformaci původních dat  $D$  s transformační maticí  $V'$ , tj.

$$G = DV' \quad \text{typu } N \times r \quad \text{tedy s } r < n \text{ proměnnými.}$$

Tento model lze odhadovat podle následujícího algoritmu:

1. Z naměřených dat sestavíme vektor  $Y$  a datovou matici  $D$ .
2. Matici  $D$  rozložíme:  $D = USV$ .
3. Rozhodneme, která singulární čísla ponecháme (počet označíme  $r$ ) a která zanedbáme.
4. Provedeme redukci matice  $V$  tak, že v ní ponecháme je  $r$  sloupců, které odpovídají nenulovým singulárním číslům.
5. Pro odhad modelu použijeme vektor  $Y$  a transformovanou datovou matici  $G$ .

Pro odhad nebo predikci z dalších dat (pro  $t = N+1, N+2, \dots$ ) lze předpokládat, že transformační matice  $V$  se ”příliš nemění” a používat ji i dále. Občas je možno ji pro nová data přepočítat opět podle uvedeného algoritmu.

## Reference

- [1] I. Nagy, *Pravděpodobnost a matematická statistika - cvičení*, skriptum FD ČVUT, Praha, 2002.
- [2] I. Nagy, *Základy bayesovského odhadování a řízení*, Vydavatelství ČVUT, Praha, 2003.
- [3] J. Novovičová, *Pravděpodobnost a matematická statistika*, skriptum FD ČVUT, Praha, 1999.
- [4] J. Hátle, J. Likeš, *Základy počtu pravděpodobnosti a matematické statistiky*, SNTL, Praha, 1974.
- [5] J. Anděl, *Matematická statistika*, SNTL, Praha, 1978.
- [6] R. C. Rao, *Lineární metody statistické indukce a jejich aplikace*, ACADEMIA, Praha, 1978.