Aktivita 1
Seminář základů statistiky a workshop (Prof. Ing. Milan Palát, CSc., Ing. Kristina Somerlíková, Ph.D.)

1 Statistické třídění

Základní metoda statistického zpracování.
Seskupování hodnot proměnné, které jsou z hlediska klasifikačního znaku stejné nebo podobné. Zároveň se uvádí četnosti.
Znaky rozlišujeme

- *tříděné* (univarietní nebo multivarietní)
- *třídící* (kvalitativní nebo kvantitativní)

Třídění:
Prosté – podle jednoho třídícího znaku
Vícenásobné – podle několika znaků

Třídící znaky:

- *Časové* (podle doby relevantní události)
- *Prostorové* (podle místa)
- *Věcné* (podle popisného stavu nebo typu experimentálního ošetření)

- *Dvojné* (podle pohlaví, vakcinace, březosti, zdravotního stavu)
- *Množné* (podle variety, druhu, plemene)

Spojité (kontinuální) - např. podle vykázaného zisku, tržeb, nákladů
Nespojité (diskrétní) - např. podle počtu členů v rodině

Variační řady - *rozdělení četností* (u nespojitých proměnných)
- *intervalové rozdělení četností* (u spojitých proměnných)

Význam třídění
- lepší organizace dat, poznání struktury
- výpočet aritmet. průměru, populačních parametrů
- metody GOF (goodness of fit)

Variační rozpětí (R) - rozdíl mezi maximální a minimální hodnotou.
Variační třídy - disjunktní intervaly na číselné ose, uvnitř intervalů nerozlišujeme hodnoty, ztrácíme část informací, ale získáme na přehlednosti. Většinou pracujeme s *6-15* třídami.
Třídy „musí“ být stejně široké.

Pravidlo pro počet intervalů:
$n < 100$ k = 5-9 intervalů
$100 < n < 500$ k = 10-15 intervalů
$n > 500$ $k = 1+3{,}3 * \log n$
Hranice a středy tříd by měla být vhodná čísla.
Každou třídu reprezentuje její fyzický střed – x_i (ne průměr hodnot!),
Úhrn třídy je pak roven $x_i \cdot n_i$ a nahrazuje přesnou hodnotu součtu všech hodnot třídy.

Příprava tabulky četností
Četnost - počet pozorování v souboru, třídě
Absolutní četnost (n_i) - fyzický počet pozorování výběrového souboru zařazených do třídy
Kumulativní (součtová) četnost (kn_i) - součet všech absolutních četností předcházejících dané absolutních četností.
Relativní četnost (p_i) - podíl absolutní četnosti k celkovému počtu hodnot souboru
Relativní četnosti vyjadřujeme v pravděpodobnostech nebo v procentech.
Kumulativní relativní četnost - součtová relativní četnost (kp_i)
Kumulativní četnosti jsou vyjádřitelné *ascendentním* nebo *descendentním* způsobem.

2 Základní variační charakteristiky statistického souboru.

1. Lokační míry (obecné polohy) -> střední hodnoty
2. Míry proměnlivosti (variability) -> variační míry
3. Míry šikmosti (symetrie) -> míry souměrnosti
4. Míry koncentrace (špičatosti) ->míry špičatosti

1. Měření obecné úrovně.

Střední hodnoty

a.) Průměry

Aritmetický	$\bar{x}$
Geometrický	$\bar{x}_G$
Harmonický	$\bar{x}_H$
Kvadratický	$\bar{x}_Q$
Chronologický	$\bar{x}_{CH}$

b.) Ostatní střední hodnoty

Medián	$\tilde{x}$
Modus	$\hat{x}$

Průměry jsou charakteristiky obecné polohy a jsou funkcemi všech hodnot v souboru.

Aritmetický průměr ($\bar{x}$)

Prostá výpočtová forma: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ Vážená forma: $\bar{x} = \frac{\sum_{i=1}^{k} x_i * n_i}{n}$,

Kde $n = \sum_{i=1}^{k} n_i$. Jsou-li absolutní četnosti nahrazeny relativními četnostmi, vážená forma se redukuje na $\bar{x} = \frac{\sum_{i=1}^{k} x_i * p_i}{\sum_{i=1}^{k} p_i}$

Vážená forma se aplikuje na tříděná data (rozdělení četností nebo intervalové rozdělení četností), u dat, kde jsou známy parciální průměry. Prostá forma se používá u menších netříděných souborů.

Vlastnosti aritmetického průměru:

1. Součet absolutních odchylek jednotlivých hodnot souboru je roven nule.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2. Součet čtverců odchylek je minimální.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = Min \text{ , tj. } \sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - c)^2, \forall c \neq \bar{x}$$

3. Aritmetický průměr konstanty je roven konstantě
4. Průměr součtů (rozdílů) dvou proměnných je roven součtu (rozdílu) jejich aritmet. průměrů.
5. U vážené formy, jsou-li všechny četnosti násobeny (děleny) stejnou konstantou, průměr se nemění.
6. Je-li ke každé hodnotě přičtena (odečtena) určitá konstanta, o tuto konstantu se zvýší (sníží) i aritmetický průměr.
7. Je-li každá hodnota souboru násobena (dělena) určitou konstantou *c*, bude aritmetický průměr *c*-krát větší (menší).

Harmonický průměr ($\bar{x}$)

Převrácená hodnota součtu převrácených hodnot zkoumaného znaku. Používá se při průměrování neprímo vyjádřených veličin jako rychlosti, výnosy, výkony atd.

Prostá forma: $\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ Vážená forma: $\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Geometrický průměr ($\bar{x}_G$)

n-tá odmocnina ze součinu n hodnot.

Prostá forma výpočtu: $\bar{x}_G = \sqrt[n]{x_1 * x_2 * \ldots * x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

V logaritmickém tvaru: $\log \bar{x}_G = \frac{1}{n}\sum_{i=1}^{n} \log x_i$

Vážená forma výpočtu: $\bar{x}_G = \sqrt[n]{x_1^{n_1} * x_2^{n_2} * \ldots * x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$

V logaritmickém tvaru: $\log \bar{x}_G = \frac{1}{n}\sum_{i=1}^{n} n_i * \log x_i$

Používá se při analýze bezrozměrných indexů zřetězených v čase.

Medián ($\tilde{x}$)

Prostřední hodnota setříděné řady hodnot souboru. Jedná se o x_{50} , tedy 50% kvantil. Představuje hodnotu, která rozdělí setříděný soubor na dvě stejné části, co do počtu hodnot. 50% hodnot je menších než medián a 50% je větších než medián.

Při lichém počtu hodnot je prostřední hodnota medián.
Při sudém počtu hodnot je mediánem průměr dvou prostředních hodnot setříděného souboru.

Modus ($\hat{x}$)

Je hodnota souboru s nejvyšší četností.

U symetrického normálního rozdělení je $\bar{x} = \tilde{x} = \hat{x}$

U levostranně nesouměrného rozdělení je $\hat{x} < \tilde{x} < \bar{x}$

U pravostranně nesouměrného rozdělení je $\bar{x} < \tilde{x} < \hat{x}$

Míry proměnlivosti

A. Variační rozpětí R = Y_{max} - Y_{min}
B. Kvantilové (kvartilové) odchylky

Mezi-kvartilové rozpětí(IQR): IQR = x_{75} - x_{25}
Kvartilová odchylka : Q = IQR / 2

C. Průměrné odchylky absolutní a relativní
Vypočítávají se průměrné odchylky buďto od průměru nebo od mediánu.
Průměrná absolutní odchylka:

Prostý tvar:

$$\bar{d}_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} / d_i / = \frac{1}{n}\sum_{i=1}^{n} / x_i - \bar{x} /$$

Vážený tvar:

$$\bar{d}_{\bar{x}} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} / d_i / * n_i = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} / x_i - \bar{x} / * n_i$$

Relativní průměrná odchylka:
Vyjádřitelná v % z aritmetického průměru.

$$\bar{d}'_{\bar{x}} = \frac{\bar{d}_{\bar{x}}}{\bar{x}} * 100$$

D. Rozptyl a směrodatná odchylka
Prostá forma (nevychýlená):

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \qquad s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}, \qquad s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 / n}{n-1}$$

Vážená forma:

$$s_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 * n_i}{n-1}$$

Vlastnosti rozptylu:

Je nezáporný.
Je nejmenší průměrnou čtvercovou odchylkou.
Změnou hodnot o konstantu se rozptyl nemění.
Násobením (dělením) všech hodnot konstantou k se rozptyl zvětší (zmenší) k-krát.

Rozptyl součtu (rozdílu) dvou proměnných je roven součtu (rozdílu) jejich rozptylů plus (minus) dvojnásobek jejich kovariance.

$$s_{(x \pm y)}^2 = s_x^2 + s_y^2 \pm 2 * s_{xy}$$

Celkový rozptyl z dílčích souborů je roven průměru dílčích rozptylů a rozptylu dílčích průměrů.

$$s_x^2 = \overline{s_x^2} + s_{\bar{x}}^2$$

Směrodatná odchylka: $s_x = \sqrt{s_x^2}$

Je uvedena ve stejných jednotkách jako naměřené hodnoty.

E. Variační koeficient

$$v_x = \frac{s_x}{\bar{x}} * 100[\%]$$

Používá se při porovnávání variability jednoho znaku v různých souborech nebo různých znaků v jednom souboru.
Míry nesouměrnosti (šikmosti)

1. Pearsonova míra šikmosti:

$$\tau = \frac{\bar{x} - \hat{x}}{s_x}, \text{ popř. } \tau = \frac{3(\bar{x} - \tilde{x})}{s_x},$$

záporné hodnoty indikují pravostrannou nesouměrnost.

2. Koeficient nesouměrnosti - asymetrie(α_3):

$$\alpha_3 = \frac{1}{s_x^3} \frac{1}{n} \sum (x_i - \bar{x})^3 n_i$$

Míry špičatosti (koncentrace, kartéze):
1. Koeficient špičatosti (α_4):

$$\alpha_4 = \frac{1}{s_x^4} \frac{1}{n} \sum (x_i - \bar{x})^4 n_i - 3$$

Kladná hodnota indikuje špičatější rozdělení oproti normálnímu rozdělení.
Záporná hodnota znamená podnormální špičatost (plochost) rozdělení.

3 Jednoduchá lineární regrese a korelace

Cílem je zkoumání příčinné závislosti mezi dvěma, či více proměnnými.

Regresní úloha: spočívá v nalezení rovnice regresní funkce, která vhodně popisuje typ a průběh závislosti y = f(x).

Podle typu funkce regresní závislost dělíme na lineární nebo nelineární.
Podle počtu proměnných na regresi jednoduchou nebo vícenásobnou.

Modelová rovnice jednoduché regresní úlohy je:

$$Y = a + b*x + e,$$

Kde

Y je závisle proměnná (odezva)
a je prostý člen (intercept)
b je regresní koeficient b_{yx}
X je nezávisle proměnná (regresor)
E je residuální odchylka

Při oboustranné závislosti jsou možné dvě regresní přímky:

$$\mathbf{Y' = a_{yx} + b_{yx} * x}$$
$$\mathbf{X' = a_{xy} + b_{xy} * y}$$

$\mathbf{a_{yx}}$, $\mathbf{b_{yx}}$, $\mathbf{a_{xy}}$, $\mathbf{b_{xy}}$ jsou neznámé koeficienty, jejichž hodnotu získáme řešením soustavy tzv. normálních rovnic.

x,y jsou empirické(skutečné hodnoty závisle proměnné.
x´,y´ jsou teoretické hodnoty závisle proměnné vypočtené z regresní rovnice.

Hodnoty potřebné pro výpočet regresních hodnot:
Součty čtverců odchylek od průměru:

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

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y}) * (x_i - \bar{x})$$

Základní forma regresního koeficientu je pak:

$$b_{yx} = \frac{S_{xy}}{S_{xx}} \qquad b_{xy} = \frac{S_{xy}}{S_{yy}}$$

Forma I.

$$b_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Forma II.

$$b_{yx} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad b_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}$$

Forma III.

$$b_{yx} = \frac{\sum_{i=1}^{n} x_i y_i - 1/n * \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - 1/n * (\sum_{i=1}^{n} x_i)^2} \qquad b_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - 1/n * \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i^2 - 1/n * (\sum_{i=1}^{n} y_i)^2}$$

Absolutní člen je pak:

$$a_{yx} = \bar{y} - b_{yx} * \bar{x} \qquad a_{xy} = \bar{x} - b_{xy} * \bar{y}$$

INTERPRETACE:

Regresní koeficient b_{yx} udává jednotkovou změnu závisle proměnné (y), když se nezávisle proměnná (x) změní o jednotku.
Absolutní člen (intercept) a_{yx} udává hodnotu teoretické proměnné y´, je-li hodnota regresoru x rovna nule.

Vlastnosti metody **LS** (nejmenší čtverce):

$\sum_{i=1}^{n}(y_i - y_i') = 0$, suma odchylek empirických a teoretických hodnot rovny nule

$\sum_{i=1}^{n}(y_i' - \bar{y}) = 0$, suma odchylek teoretických hodnot a průměru rovny nule

$\sum_{i=1}^{n}(y_i - \bar{y}) = 0$, suma odchylek empirických hodnot a průměru rovny nule

$\sum_{i=1}^{n}(y_i - y_i')^2 = \min$, suma čtverců odchylek empirických a teoretických hodnot je minimální

Koeficient korelace (r).

Je bezrozměrná veličina v intervalu –1 <= r => +1.
Znaménkem se musí shodovat s oběma regresními koeficienty.

Kladná hodnota znamená kladnou, pozitivní závislost.
Záporná hodnota znamená zápornou, negativní závislost.

r = 0 znamená lineární nezávislost.
/r/ = 1 znamená pevnou funkční závislost.

Absolutní hodnota r	Těsnost závislosti	Typ závislosti
0	Nulová	Nezávislost
0,0-0,3	Nízká	Volná závislost
0,3-0,5	Mírná	
0,5-0,7	Význačná	
0,7-0,9	Velká	
0,9-0,99	Velmi vysoká	
1,0	Pevná funkční	Pevná závislost

Výpočet:

$r = \pm\sqrt{b_{yx} * b_{xy}}$ geometrický průměr obou regresních koeficientů,
kde znaménko odpovídá znaménku regresního koeficientu

úpravou vztahu lze získat výrazy pro výpočet koeficientů regrese:

$$b_{yx} = r\frac{s_y}{s_x}, \qquad b_{yx} = \frac{r^2}{b_{xy}}, \qquad b_{xy} = r\frac{s_x}{s_y}, \qquad b_{xy} = \frac{r^2}{b_{yx}}$$

kde hodnoty směrodatných odchylek *s* se počítají vychýleným způsobem.

Obecně korelační koeficient dostaneme:

$$r = \frac{\mathrm{cov}_{xy}}{\sqrt{\mathrm{var}_x * \mathrm{var}_y}}$$

výpočtové tvary:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}}$$

nebo

$$r = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} * \frac{y_i - \bar{y}}{s_x} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

nebo

$r = \sqrt{\frac{\mathrm{var}(y')}{\mathrm{var}(y)}}$, kde var(y´) je variance teoretických hodnot a var(y) je variance empirických hodnot závisle proměnné.

4 Náhodná veličina, rozdělení pravděpodobnosti

Náhodná veličina

= libovolná kvantitativní charakteristika náhodného pokusu

- proměnná nabývající hodnot v závislosti na náhodě
- hodnota je tedy jednoznačně určena výsledkem náhodného pokusu, kterou je číselná hodnota - realizace x náhodné veličiny X)
- pro náhodnou veličinu se užívá označení X_1, X_2, X_3, Y, Z, pro hodnoty realizace pak x_1, x_2, x_3, y, z apod.

Základní druhy náhodné veličiny:

nespojitá (diskrétní)

- alternativní rozdělení, Binomické rozdělení, Poissonovo rozdělení, Hypergeometrické

spojitá

- normální (Gaussovo) rozdělení, rozdělení χ^2, t, F (Fisher- Snedecorovo)

Zákon rozdělení pravděpodobnosti

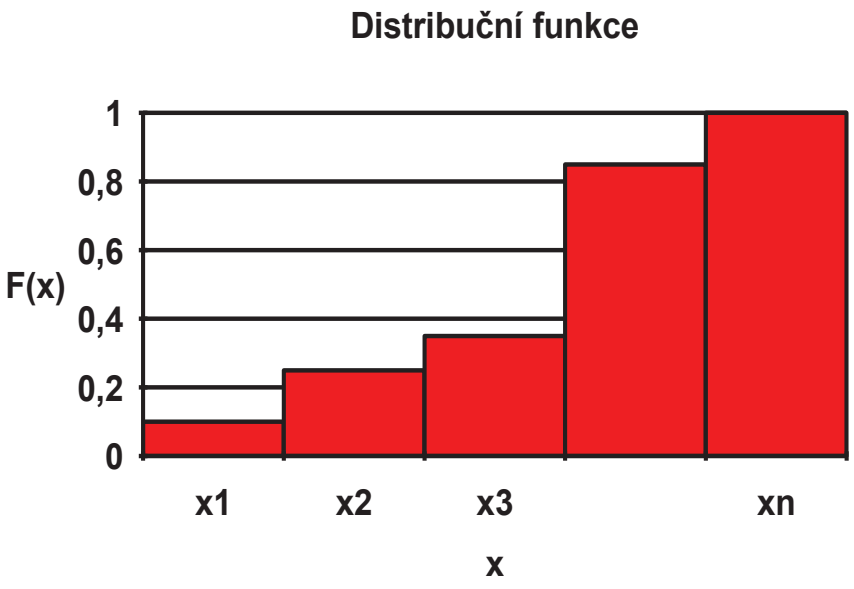
= pravidlo, podle kterého jsou jednotlivým možným hodnotám náhodné veličiny X přiřazeny jejich pravděpodobnosti.

způsoby vyjádření zákona rozdělení pravděpodobností - vzorcem, tabulkou, graficky

Základním prostředkem vyjádření zákona rozdělení náhodné veličiny X je distribuční funkce $F(x)=P(X \leq x)$

Vlastnosti distribuční funkce:

- $0 \leq F(x) \leq 1$
- $P(x_1 < X < x_2) = F(x_2) - F(x_1)$
- Distribuční funkce je neklesající, tj. pro všechna $x_1 < x_2$ platí, že $F(x_1) \leq F(x_2)$
- Distribuční funkce je spojitá zprava
- $F(-\infty) = 0$, $F(\infty) = 1$



Kvantily

100 α% kvantil x_α spojité náhodné veličiny X nazýváme hodnotu, pro kterou platí $F(x_\alpha) = \alpha$

je-li $\alpha=0{,}05 \rightarrow$ 5 % kvantil
$\alpha=0{,}95 \rightarrow$ 95 % kvantil

Kvantily umožňují konstruovat takové intervaly, do nichž spadá hodnota náhodné veličiny se zvolenou pravděpodobností.

např. $x_{0,05} = 1{,}18$
$x_{0,95} = 5{,}94$

pak $P(1{,}18 < X < 5.94) = 0{,}90$

POZN. Pro praktickou práci jsou kvantily důležitých pravděpodobnostních rozdělení tabelovány

Statistiky

Základní používané statistiky

- aritmetický průměr $\overline{X}$, jehož realizace je $\overline{x}_n$
- rozptyl resp. směrodatná odchylka - 2 tvary (výběrový a základního souboru)

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

5 Teorie odhadu

Bodový odhad

- je odhad na základě jednoho čísla
- odhadem charakteristiky či parametru základního souboru Θ je výběrová charakteristika či parametr T (obvykle je volen tzv. výběrový protějšek)

výběrová charakteristika charakteristika zákl. souboru

$$R \xrightarrow{odhad} \Theta$$

pak

$$\overline{x} \rightarrow \mu$$

$$s_x^2 \rightarrow \sigma_x^2$$

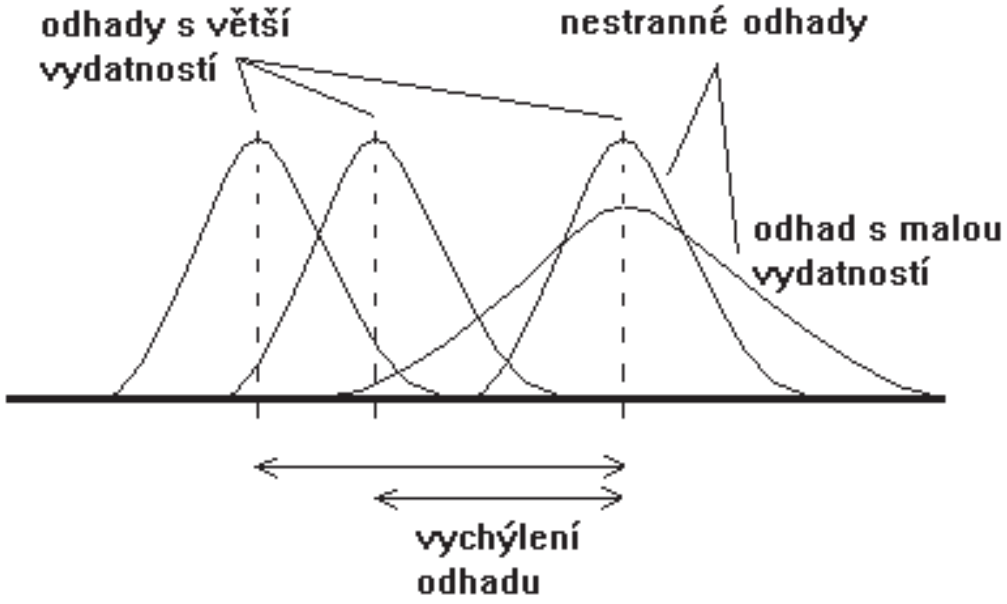
$$r \rightarrow \rho$$

$$b_{yx} \rightarrow \beta_{yx}$$

Bodový odhad má splňovat:

- *nestrannost* - tj. odhad střední hodnoty charakteristiky výběrového souboru je roven odhadované charakteristice základního souboru $E(T) = \Theta$
- *konzistence* - vzrůstající rozsah výběru snižuje výběrovou chybu $\lim_{n\to\infty} P(|T-\Theta| < \varepsilon) = 1$
- *vydatnost* - takový odhad, který má z charakteristik přicházejících v úvahu nejmenší rozptyl $D(T)<D(T^+)$

kde T - výběrová charakteristika splňující vydatnost odhadu
T^+ - jakákoli jiná výběrová charakteristika



Vydatnost lze měřit *mírou vydatnosti* $e(T^+)$:

$$e(T^+) = \frac{D(T)}{D(T^+)} \qquad 0<e(T^+)<1$$

Lze uvést:

$$\lim_{n\to\infty} e(T^+) = 1$$

Intervalový odhad

- odhadem charakteristiky či parametru základního souboru se rozumí stanovení intervalu, v němž se odhadovaná charakteristika či parametr nachází

Pro 100(1-α) procentní interval spolehlivosti charakteristiky Θ platí:

$$P(T^{/} \le \Theta \le T^{//}) = 1-\alpha$$

kde $T^{/}$ - dolní hranice intervalu
$T^{//}$ - horní hranice intervalu

- hodnoty α jsou rizika odhadu

- za α se obvykle volí $\alpha=0{,}05$ nebo $\alpha=0{,}01$ (95% resp. 99% interval spolehlivosti)
- *intervaly spolehlivosti* se označují též termínem konfidenční intervaly
- při stanovení intervalů spolehlivosti se často využívá normální aproximace. Vychází se z normované veličiny normálního rozdělení výběrové charakteristiky

$$U = \frac{T - E(T)}{\sqrt{D(T)}} = \frac{T - \Theta}{\sqrt{D(T)}} \qquad U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Distribuční funkce normovaného normálního rozdělení je tabelována pro různé hodnoty u

- Intervaly spolehlivosti mohou být *jednostranné* nebo *oboustranné*

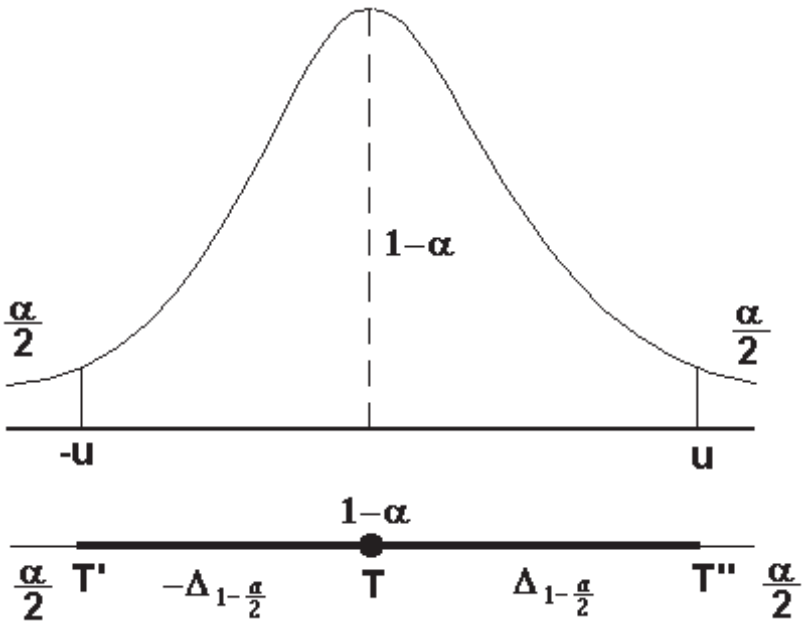
Oboustranný interval spolehlivosti Θ:

$$P(-u \le U \le u) = P(-u \le \frac{T - \Theta}{\sqrt{D(T)}} \le u) =$$

$$P\left[-u\sqrt{D(T)} \le T - \Theta \le u\sqrt{D(T)}\right] = P\left[T - u\sqrt{D(T)} \le \Theta \le T + u\sqrt{D(T)}\right]$$

takže platí:

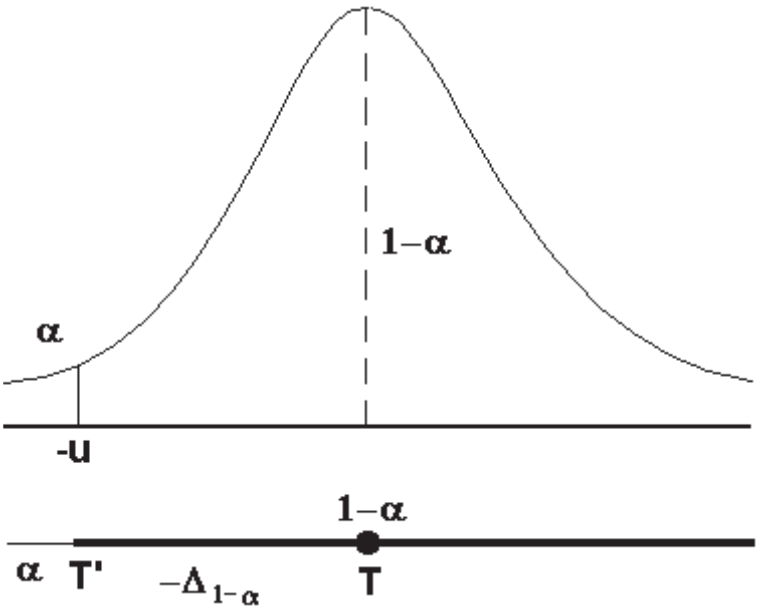
$$P\left[T - u_{1-\frac{\alpha}{2}}\sqrt{D(T)} \le \Theta \le T + u_{1-\frac{\alpha}{2}}\sqrt{D(T)}\right] = 1 - \alpha$$



Jednostranné intervaly spolehlivosti charakteristiky Θ pak:

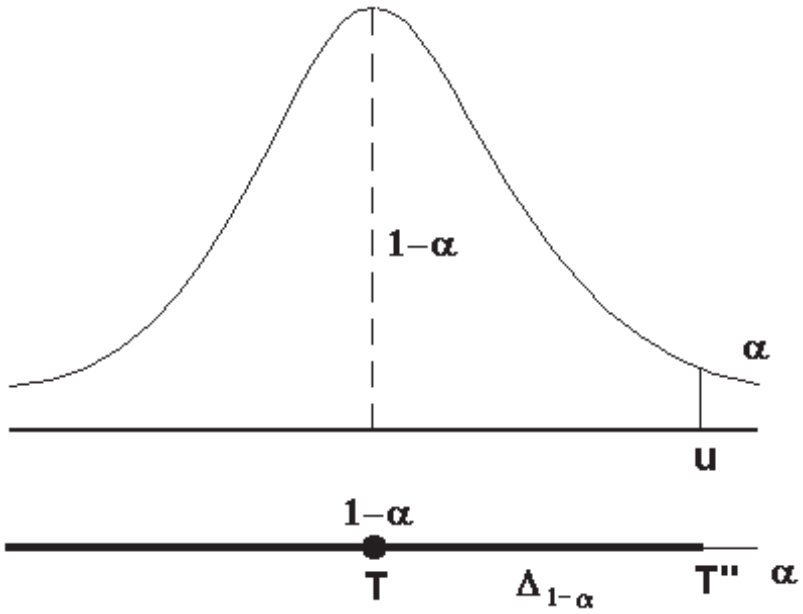
levostranný interval

$$P\left[T - u_{1-\alpha}\sqrt{D(T)} \le \Theta\right] = 1 - \alpha$$



pravostranný interval

$$P\left[\Theta \le T + u_{1-\alpha}\sqrt{D(T)}\right] = 1 - \alpha$$



Δ - *přípustná chyba* = násobek normované veličiny normálního či Studentova rozdělení a střední chyby

$$\Delta = u_{1-\frac{\alpha}{2}} \sqrt{D(T)} \qquad \Delta = u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Stanovení minimálního rozsahu výběru:

$$n \geq \frac{t^2_{1-\frac{\alpha}{2}} \cdot s_x^2}{\Delta^2}$$

při rozsahu výběru n>30 lze neznámý parametr σ bez problémů nahradit jeho bodovým odhadem - směrodatnou odchylkou S_{n-1} (nahrazení normálním rozdělením)

$$n \geq \frac{u^2_{1-\frac{\alpha}{2}} \cdot \sigma^2}{\Delta^2}$$

při rozsahu výběru n<30 je při neznámém parametru σ nutno použít vztah

$$P(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}) = 1 - \alpha$$

kde $t_{1-\alpha/2}$ je kvantil Studentova rozdělení pro n - 1 stupňů volnosti

- Grafické stanovení minimálního rozsahu výběru - je spolehlivější

Interval spolehlivosti aritmetického průměru

Oboustranný interval

$$P(\bar{x} - u_{1-\frac{\alpha}{2}} \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + u_{1-\frac{\alpha}{2}} \cdot s_{\bar{x}}) = 1 - \alpha$$

kde $S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, popř. $S_{\bar{x}} = \frac{S_{n-1}}{\sqrt{n}}$

Levostranný

$$P(\bar{x} - u_{1-\alpha} \cdot s_{\bar{x}} \leq \mu) = 1 - \alpha$$

Pravostranný

$$P(\mu \leq \bar{x} + u_{1-\alpha} \cdot s_{\bar{x}}) = 1 - \alpha$$

Interval spolehlivosti rozptylu

- s využitím χ^2 rozdělení

$$P\left[\frac{(n-1)s_x^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma_x^2 \leq \frac{(n-1)s_x^2}{\chi^2_{\frac{\alpha}{2}}}\right] = 1 - \alpha$$

Interval spolehlivosti relativních a absolutních četností

- relativní četnosti

$$P(p_i - t_{1-\frac{\alpha}{2}} \cdot s_{pi} \leq P_i \leq p_i + t_{1-\frac{\alpha}{2}} \cdot s_{pi}) = 1 - \alpha$$

kde: $s_{pi} = \sqrt{\frac{p_i(1-p_i)}{n}}$

- absolutní četnosti

$$P\left[N(p_i - t_{1-\frac{\alpha}{2}} \cdot s_{pi}) \leq N_i \leq N(p_i + t_{1-\frac{\alpha}{2}} \cdot s_{pi})\right] = 1 - \alpha$$

Intervalový odhad charakteristik korelace a regrese

Závislost
- podle stupně závislosti - pevná, volná
- podle druhu znaků - korelační, asociační, kontingenční

Druhy korelační závislosti
- podle počtu kvantitativních znaků - jednoduchá, vícenásobná
- podle typu regresní funkce - lineární, nelineární
- podle změn - pozitivní, negativní

- *korelační koeficient*

výběrový koeficient korelace r neodpovídá kritériím bodového odhadu, proto:

$$r \xrightarrow{Fisherova\_transformace} z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$$ (tabelováno)

$$P(z_r - u_{1-\frac{\alpha}{2}} s_{zr} \le \xi \le z_r + u_{1-\frac{\alpha}{2}} s_{zr}) = 1-\alpha$$

kde $s_{zr} = \frac{1}{\sqrt{n-3}}$

ale pro r < 0,5 a n > 100 platí:

$$P(r - u_{1-\frac{\alpha}{2}} s_r \le \rho \le r + u_{1-\frac{\alpha}{2}} s_r) = 1-\alpha$$

kde $s_r = \frac{1-r^2}{\sqrt{n-k-1}}$

- *regresní koeficient b_1, popř. b_{yx}*

Přímka může být zapsána buď ve tvaru: $y'_i = a_{yx} + b_{yx} x_i$
nebo $y'_i = b_0 + b_1 x_i$. Potom pro intervaly spolehlivosti platí:

$$P(b_1 - t_{1-\frac{\alpha}{2}} s_{b_1} \le \beta_1 \le b_1 + t_{1-\frac{\alpha}{2}} s_{b_1}) = 1-\alpha$$

kde $s_{b1} = s_e \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}$

popř.

$$s_{b_1} = \frac{s_y}{s_x} \cdot \sqrt{\frac{1-r^2}{n-k-1}}$$

- Absolutní člen b_o, popř. a_{yx}

$$P(b_0 - t_{1-\frac{\alpha}{2}} s_{b_0} \le \beta_0 \le b_0 + t_{1-\frac{\alpha}{2}} s_{b_0}) = 1-\alpha$$

kde

$$s_{b0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

s_e je reziduální směrodatná odchylka $\sqrt{\frac{\sum (y - y')^2}{n-2}}$

- *regresní přímka*

$y'_i = a_{yx} + b_{yx} x_i$ popř. $y'_i = b_0 + b_1 x_i$.

$$P(y'_i - t_{1-\frac{\alpha}{2}} s_{y'i} \le y'_j \le y'_i + t_{1-\frac{\alpha}{2}} s_{y'i}) = 1-\alpha$$

kde $s_{yi'} = s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

popř.

$$s_{y'i} = s_{\bar{y}} \sqrt{1 + \frac{(x_i - \bar{x})^2}{s_x^2}}$$

a $s_{\bar{y}} = \frac{s_y}{\sqrt{n}}$

Nejpřesnější je odhad v blízkosti aritmetického průměru, interval spolehlivosti je v tomto místě nejužší.

Poznámka: Pro n>30 lze t rozdělení aproximovat normálním

- *pás spolehlivosti kolem regresní funkce*

Hodnoty závisle proměnné konkrétního statistického znaku jsou rozptýleny kolem regresní funkce. Tento pás, ve kterém se tyto skutečné hodnoty nacházejí, lze stanovit se zvolenou pravděpodobností.

- Pás spolehlivosti kolem regresní přímky

$$P_{y_{i(H,D)}}(y_i' \pm t_{1-\frac{\alpha}{2}} \cdot s_{yx}) = 1-\alpha$$

kde s_{yx} je směrodatná (standardní) chyba

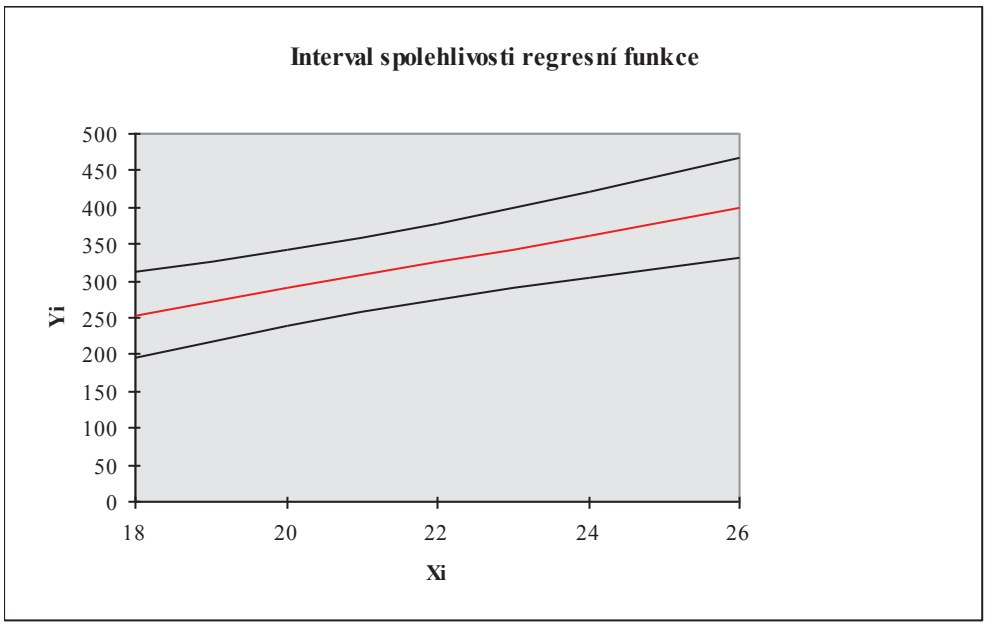
$$s_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i)^2 - \sum_{i=1}^{n} y_i \cdot y_i'}{n-k-1}}$$

k - počet parametrů regresní funkce mimo absolutní člen, popř. počet nezávisle proměnných (vysvětlujících proměnných)

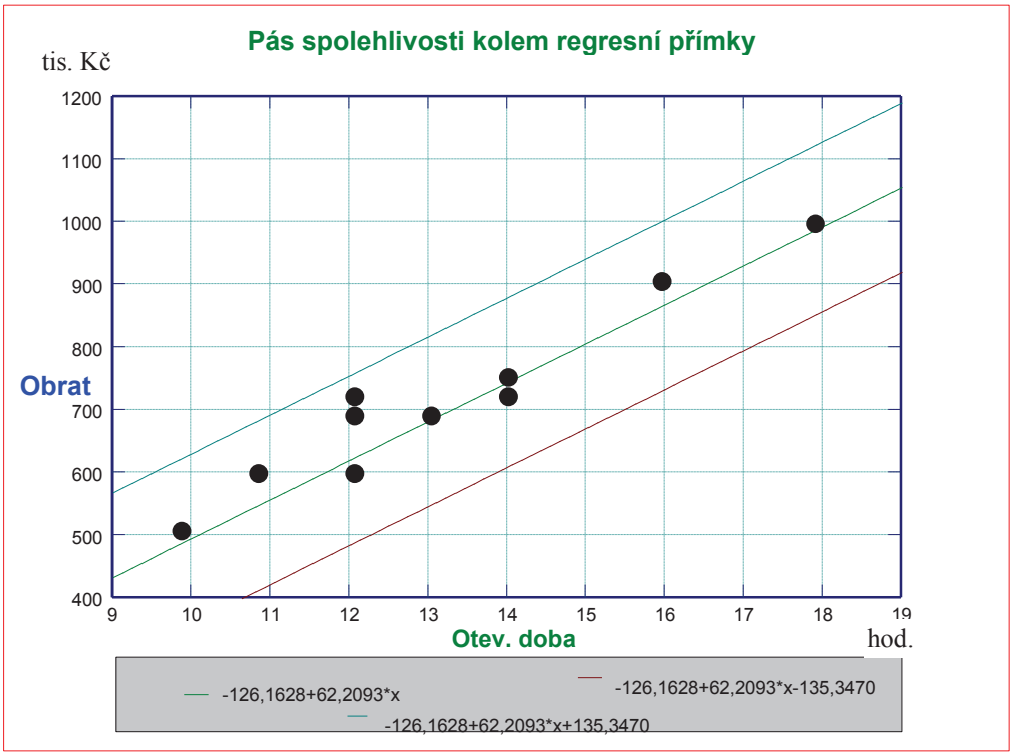
Vzorce pro s_x, r

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \sqrt{b_{yx} \cdot b_{xy}}$$



E. Pás spolehlivosti



6 Testování statistických hypotéz

- spjato se statistickými odhady
- Principem je vyslovení předpokladu o charakteristice základního souboru - nulová hypotéza H_o a její testování

$\mu=c$, - střední hodnota je rovna konstantě
$\rho=0$ - korelační koeficient je roven 0
$\beta=0$ - regresní koeficient je roven 0
$\mu_1=\mu_2$ - stř. hodnoty 2 výběrů se rovnají
apod.

Proti nulové hypotéze - alternativní hypotéza H_1

- u dvoustranného testu - $\mu \neq c$
- u jednostranného testu - $\mu > c$

Chyba 1. druhu - H_0 je pravdivá a zamítá se, pravděpodobnost chyby je α
Chyba 2. Druhu - H_0 je nepravdivá a nezamítáme ji - pravděpodobnost chyby je β

Hladina významnosti - pravděpodobnost chyby 1. druhu - α

Postup při testování hypotéz:

1. formulace hypotézy
2. volba testového kritéria

3. sestrojení kritického oboru
4. výpočet hodnoty testového kritéria
5. formulace výsledků testu

Platí-li, že hodnota testového kritéria je větší než tabulková hodnota při: $\alpha = 0,05$ - test je statisticky průkazný

$\alpha = 0,01$ - test je statisticky vysoce průkazný

Testy o střední hodnotě při velkém výběru (n>30) ze základního souboru, popř. při známém rozptylu (δ^2)

Testové kritérium: $U = \dfrac{\bar{X} - C}{\dfrac{s_x}{\sqrt{n}}}$

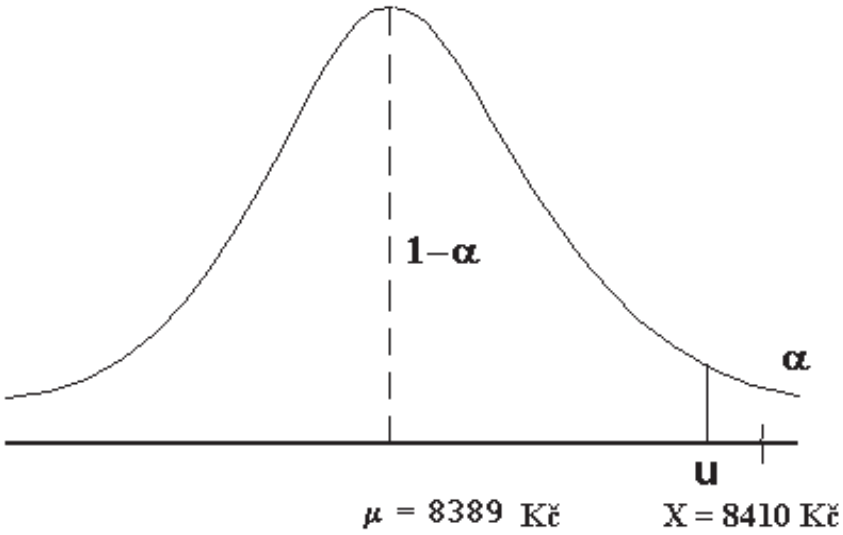
Př. Otestujte, zda-li průměrný plat pracovníků školství je vyšší než 8389 Kč.

Nulovou hypotézu lze formulovat jako:

H_0: $\mu = 8389$ Kč

Alternativní hypotézu jako:

H_1: $\mu > 8389$



Za tímto účelem byl proveden náhodný výběr 100 osob pracujících ve oboru. Byla zjištěna průměrná odměna 8410 Kč a směrodatná odchylka $s_x = 90$ Kč. Test provedeme na hladině významnosti $\alpha = 0,05$
Pro hodnotu testového kritéria platí:

$$U = \frac{8410 - 8389}{\frac{90}{\sqrt{100}}} = 2,33$$

Tabulková hodnota 95% kvantilu $u_{0,95}$ je 1,64
I při hladině významnosti $\alpha = 0,01$ je test statisticky významný ($u_{0,95} = 2,326$).

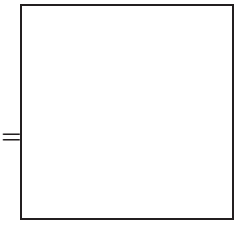
Podobně, pokud by byla průměrná odměna zjištěna jako 8368 Kč a testové kritérium U=-2,33 a alternativní hypotéza H_1 byla $\mu < 8389$, platilo by, že $u_\alpha = -u_{1-\alpha}$

Závěrem, lze říci, že zamítáme nulovou hypotézu, že průměrná odměna je 8389 Kč. Z toho tedy plyne, že průměrná odměna je vyšší.

Testy o střední hodnotě při malém výběru (n<30) ze základního souboru, popř. neznámém rozptylu zákl. souboru

Jedná se o podobný postup jako při testování výběrů větších jak 30 s tím rozdílem, že testovým kritériem je hodnota t.

Testové kritérium má tvar

$t =$ 

t má studentovo rozdělení o n-1 stupních volnosti.

Př. U náhodného výběru spotřebitelů o rozsahu n=15 byl zjištěn průměrný měsíční výdaj na osobu za potraviny 1850, směrodatná odchylka 80. Zjistěte, zda-li lze zamítnou hypotézu, že průměrný výdaj za potraviny na osobu a měsíc je v ČR 1828 Kč.

Nulovou hypotézu lze formulovat jako:

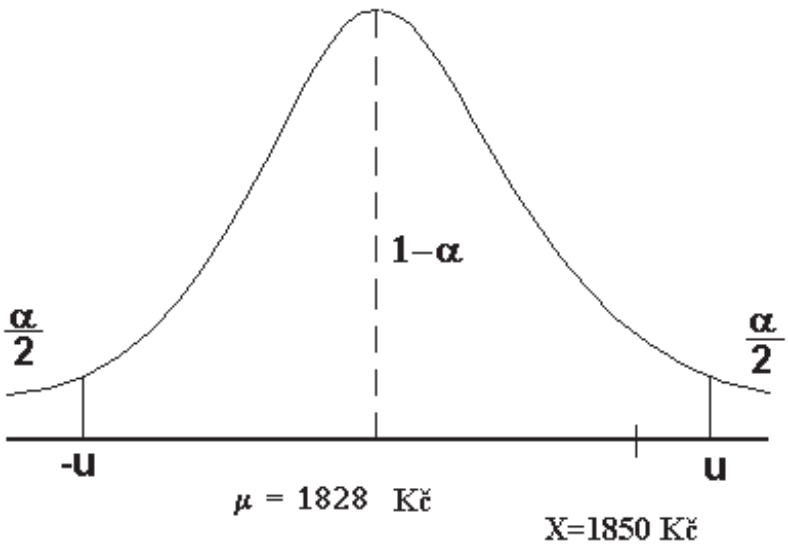
H_0: $\mu = 1828$ Kč

Alternativní hypotézu jako:

H_1: $\mu \neq 1828$

Test provedeme na hadině významnosti $\alpha = 0,05$
Testové kritérium t má tvar:

$$t = \frac{1850 - 1828}{\frac{80}{\sqrt{15}}} = 1,06$$



Tabulková hodnota t-rozdělení pro oboustrannou hypotézu pro 14 stupňů volnosti je $t_{0,975}=2,145$.
Nezamítáme nulovou hypotézu, že střední hodnota se rovná 1828 Kč.

Princip a postup při testování hypotéz pro regresi, regresní koeficienty a index korelace je podobný.

Test hypotézy o shodě 2 průměrů:

za předpokladu známých rozptylů v obou základních souborech
pro srovnávání 2 alternativ, posouzení významnosti změn apod.

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Příklad:
Na 5% hladině významnosti testu ověřte, zda výkon pracovníků v jednom závodě je významně vyšší než v jiném, zaměřeném na stejný typ výroby. Je znám rozptyl výkonů $\sigma_1^2 = 5$ a $\sigma_2^2 = 3$. K ověření testované hypotézy byl proveden náhodný výběr v prvním závodě $n_1 = 50$ pracovníků a $n_2 = 40$ pracovníků, průměrné výkony byly $\bar{x}_1 = 35$ a $\bar{x}_2 = 30$.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

$$U = \frac{35 - 30}{\sqrt{\frac{5}{50} + \frac{3}{40}}} = 11{,}95$$

$u_{0,95} = 1{,}645$

$11{,}95 > 1{,}645$

Nulová hypotéza se zamítá, na zvolené 5% hladině významnosti je výkon pracovníků v prvním závodě vyšší než ve druhém.

Testování průkaznosti regresního modelu - analýza rozptylu (variance)

Definovaný model testujeme pomocí analýzy rozptylu, kdy zjišťujeme variabilitu vysvětlenou regresí a ovlivněnou náhodnými vlivy. Testovým kritériem je F-test

Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců	Stupně volnosti	Rozptyl	F-hodnota
Regrese	S_R	$\nu_R=k$	$s^2_R=S_R/\nu_R$	s^2_R/ s^2_e
Reziduum	S_e	$\nu_e=n-k-1$	$s^2_e=S_e/\nu_e$	
Celkem	S_T	$\nu_T=n-1$		

$$S_R = \sum_{i-1}^{n} (y_i' - \bar{y})^2$$

$$S_e = \sum (y_i - y_i')^2$$

$$S_T = \sum (y_i - \bar{y})^2$$

Pro sumy čtverců a stupně volnosti platí:
$S_T = S_R + S_e$, tj. celková = způsobená regresí + reziduální
$\nu_T = \nu_R + \nu_e$

k ... počet parametrů regresního modelu kromě absolutního členu, popř. počet nezávisle proměnných

Při testování vycházíme z nulové hypotézy H_0: „model je statisticky neprůkazný“
Testovým kritériem je F-hodnota získaná jako podíl rozptylu teoretických hodnot (rozptyl vysvětlený regresí) k rozptylu kolem regrese (reziduální).

$$F_{(k,\, n-k-1)} = \frac{s_R^2}{s_e^2}$$

F má Fisher-Snedecorovo rozdělení s k a n-k-1 stupni volnosti.

Př. Ve 12 regionech byly sledovány 2 proměnné: cena za určitý výrobek a množství, které spotřebitelé za tuto cenu požadovali (poptávka). Zjistěte, jaký je vztah mezi cenou a množstvím. Proveďte testování regresního modelu.

Cena	Množství	*Vyrovnané hodnoty*
7	200	*181,42*
7.5	180	*176,62*
8	170	*171,81*
8.5	161	*167,01*
9	153	*162,21*
9.5	148	*157,40*
10	145	*152,60*
10.5	143	*147,79*
11	141	*142,99*
11.5	140	*138,19*
12	140	*133,38*
12.5	139	*128,58*

Řešení:
Metodou nejmenších čtverců bylo vypočítána rovnice přímky:
y’ = 248,68 - 9,61.x

Hodnota korelačního koeficientu byla 0,896

Regresní model lze testovat analýzou rozptylu. Bylo vypočteno:

$S_R = \sum_{i-1}^{n} (y_i' - \bar{y}_i)^2 = (181{,}4-155)^2 + \ldots\ldots + (128{,}58-155)^2 = 3300{,}5$

$S_T = \sum (y_i - \bar{y})^2 = (200-155)^2 + \ldots\ldots + (139-155)^2 = 4110$

$S_e = \sum (y_i - y_i')^2 = S_T - S_R = 809{,}5$

$s^2_R = S_R/\nu_R = 3300,5/1 = 3300,5$
$s^2_e = S_e/\nu_e = 809,5/10 = 80,95$

Pro testové kritérium F potom platí:

$$F_{(1,10)} = \frac{3300,5}{80,95} = 40,771$$

$F_{tab} = 4,965$

$F_{vyp} > F_{tab}$, platí proto, že zamítáme nulovou hypotézu H_0, že regresní model je neprůkazný.

Výsledná data pro analýzu rozptylu jsou uvedena v tabulce.

Analýza rozptylu

Vliv	Suma čtverců S	St.v. ν	Rozptyl s^2	F-hod.	St.význ.
Regrese R	3300.483	1	3300.483	40.771	0.0001
Chyba (e)	809.517	10	80.952		
Celkem (T)	4110.000	11	373.636		

Testování parametrů regresní funkce

Nulová hypotéza H_0 je ve tvaru: $\beta_j=0$, tj. že parametry regresní funkce jsou nevýznamné, rovny 0, neovlivňují závisle proměnnou. Alternativní hypotéza H_1 je $\beta_j \neq 0$. Pro testové kritérium t platí:

$t_{(n-k-1)} = \frac{b_j}{s_{bj}}$, b_j je parametr funkce, s_{bj} je směrodatná chyba odhadu

kde pro j=0 (absolutní člen) platí

$$s_{b0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

pro j=1 (regresní koeficient) platí

$$s_{b1} = s_e \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$, popř. $$s_{b_1} = \frac{s_y}{s_x} \cdot \sqrt{\frac{1-r^2}{n-k-1}}$$

s_e je reziduální směrodatná odchylka $\sqrt{\frac{\sum (y - y')^2}{n-2}}$

s_x a s_y jsou směrodatné odchylky proměnných x a y.

Hodnota t má Studentovo rozdělení t s n-k-1 stupni volnosti. Pro n>30 se kvantily nahrazují kvantily normálního rozdělení.

Byly zjištěny tyto hodnoty regresní přímky:

	Koeficient
Konstanta	248.68188
Směrnice	-9.60839192

Otestujte parametry regresní funkce na hladině významnosti $\alpha = 0,05$.

$$s_{b0} = 9\sqrt{\frac{1}{12} + \frac{95,06}{35,75}} = 14,9$$ pro t-hodnotu platí: $t = \frac{248,7}{14,9} = 16,69$

$$s_{b1} = 9\sqrt{\frac{1}{35,75}} = 1,50$$ pro t-hodnotu platí: $t = \frac{-9,61}{1,5} = -6,4$

$t_{tab} = 2,228$, resp. -2,228

Jelikož hodnota vypočtená je větší než tabulková, můžeme na hladině významnosti $\alpha=0,05$ zamítnout hypotézu o nulové hodnotě koeficientů regresní funkce.

Testování statistické významnosti korelačního koeficientu

Testovým kritériem je opět hodnota F, která má Fisher-Snedecorovo rozdělení s k a n-k-1 stupni volnosti.

$$F = \frac{r_{yx}^2 (n-k-1)}{(1-r_{yx}^2).k}$$

Pozn.: Jedná-li se o jednoduchou regresi, lze použít testové kritérium t s n-2 stupni volnosti. Potom platí

$$t = \frac{r_{yx}.\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$$

Z příkladu v kapitole 1.5.3 byl zjištěn korelační koeficient r = 0,896. Na hladině významnosti $\alpha = 0,05$ testujte hodnotu korelačního koeficientu.

$$F_{(1,10)} = \frac{0,803.10}{(1-0,803).1} = 40,77$$

ANALÝZA VARIANCE jednofaktorová

Model pozorování: $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Skupiny / Jedinci	1	2	…i…	k
1	x_{11}	x_{21}	x_{i1}	x_{k1}
2	x_{12}	x_{22}	x_{i2}	x_{k2}
…				
j	x_{1j}	x_{2j}	x_{ij}	x_{kj}
…				
n_i	x_{1n_1}	x_{2n_2}	x_{in_i}	x_{kn_k}
Součet	$X_{1.}$	$X_{2.}$	$X_{i.}$	$X_{k.}$
Průměr	$x_{1.}$	$x_{2.}$	$x_{i.}$	$x_{k.}$

Průměry skupin:

$$x_{i.} = \frac{X_{i.}}{n_i}$$

Celkový průměr:

$$x_{..} = \frac{\sum X_{i.}}{\sum n_i} = \frac{X_{..}}{n}$$

Nulová hypotéza: $H_0 \equiv \mu_1 = \mu_2 = \ldots = \mu_k = \mu$

„Rozdíl mezi průměry je statisticky neprůkazný“

Tabulka analýzy variance

Zdroj variability	Součet čtverců S	Stupně voln. ν	Prům.čtverec MS
Skup. (faktor)	$S_1 = \sum_{i=1}^{k} (x_{i.} - x_{..})^2 n_i$	$\nu_1 = k-1$	$MS_1 = \frac{S_1}{\nu_1}$
Jed. (reziduum)	$S_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2$	$\nu_e = n-k$	$MS_e = \frac{S_e}{\nu_e}$
C e l k e m	$S_T = \sum_{i=1}^{k} \sum_{j=1}^{n_1} (x_{ij} - x_{..})^2$	$\nu_T = n-1$	

Test. kritérium:

$$F_{(k-1,n-k)} = \frac{MS_1}{MS_e}$$

Vyhodnocení:

$F_{vyp} > F_{tab}$ … H_0 se zamítá

$F_{vyp} < F_{tab}$ … H_0 se nezamítá

Příklad:

ANALÝZA VARIANCE

JEDNOFAKTOROVÁ

Zadání: Úkolem je zjistit, zda má 5 různých krmných dávek rozdílný vliv na přírůstek živé hmotnosti skotu

Krmná dávka	Denní přírůstek živé hmotnosti (x)			
A	0,67	0,55	0,42	0,67
B	0,98	0,96	0,91	0,66
C	0,60	0,69	0,50	0,35
D	0,79	0,64	0,81	0,70
E	0,90	0,70	0,79	0,88

Řešení: H_0 → *„Dané krmné dávky nemají rozdílný vliv na přírůstek živé hmotnosti.“*

Jedinci	A	B	C	D	E	
1	67	98	60	79	90	
2	55	96	69	64	70	
3	42	91	50	81	79	
4	67	66	35	70	88	
Σ	231	351	214	294	327	1 417
$\bar{x}$	57,75	87,75	53,50	73,50	81,75	

$$K = \frac{\left(\sum x\right)^2}{n_i \cdot k} = \frac{1\,417^2}{4 \cdot 5} = 100\,394{,}45$$

$$S_C = \sum x^2 - K = 67^2 + 55^2 + \cdots + 88^2 - 100\,394{,}45 = 5\,698{,}55$$

$$S_1 = \frac{1}{n_i}\left[\left(\sum x_A\right)^2 + \left(\sum x_B\right)^2 + \cdots + \left(\sum x_E\right)^2\right] - K =$$

$$= \frac{1}{4}\left(231^2 + 351^2 + 214^2 + 294^2 + 327^2\right) - 100\,394{,}45 = 3\,536{,}30$$

$$S_e = S_C - S_1 = 5\,698{,}55 - 3\,536{,}30 = 2\,162{,}25$$

Tabulka analýzy variance:

Příčina proměnlivosti	Součet čtverců	Stupně volnosti	Průměrný čtverec
Faktor (krm.dávky)	$S_1 = 3336,30$	$k-1 = 4$	$MS_1 = 884,075$
Reziduum (jedinci)	$S_e = 2162,25$	$n-k = 15$	$MS_e = 144,150$
C e l k e m	$S_C = 5698,55$	$n-1 = 19$	

Testové kritérium:

$$F_{(k-1,n-k)} = \frac{\text{průměrný čtverec faktoru}}{\text{průměrný čtverec rezidua}} = \frac{MS_1}{MS_e}$$

$$F_{(4,15)} = \frac{884,075}{144,150} = 6,13$$

Vyhodnocení:

F_{vyp}		F_{tab}
6,13	>	3,06 (pro P=0,05)

Nulová hypotéza se na hladině významnosti P=0,05 zamítá, existuje tedy rozdílný vliv krmných dávek na přírůstek živé hmotnosti (alespoň v jedné dvojici).

NÁSLEDNÉ METODY
při zamítnutí H_0 z analýzy variance

Metoda minimální průkazné diference

$$(d) = t \cdot s_{d_{i.}}$$

t … pro stupně volnosti rezidua a danou hladinu významnosti

směrodatná chyba průměrného rozdílu:

$$s_{i.} = \sqrt{MS_e \left(\frac{1}{n_i} + \frac{1}{n_p} \right)}$$

skupiny o různém rozsahu

$$s_{d_{i.}} = \sqrt{\frac{2\,MS_e}{n_i}}$$

skupiny o stejném rozsahu

Scheffeho metoda kontrastů

$$t_{(n_i+n_p-2)} = \frac{/\,x_{i.} - x_{p.}\,/}{S_{x_{i.}-x_{p.}}}$$

směrodatná chyba rozdílu průměrů:

$$s_{x_{i.}-x_{p.}} = \sqrt{MS_e \frac{n_i + n_p}{n_i \cdot n_p}}$$

skupiny o různém rozsahu

$$s_{d_{i.}} = \sqrt{\frac{2\,MS_e}{n_i}}$$

skupiny o stejném rozsahu

Metoda minimální diference

$$s_{\bar{d}} = \left(s_{d_{i.}}\right) = \sqrt{\frac{2\,MS_e}{n_i}} = \sqrt{\frac{2 \cdot 144{,}15}{4}} = 8{,}49$$

$$(d) = t \cdot s_{\bar{d}} \begin{cases} = 2{,}13 \cdot 8{,}49 = 18{,}08 \ \text{(pro P=0,05)} \\ = 2{,}95 \cdot 8{,}49 = 25{,}03 \ \text{(pro P=0,01)} \end{cases}$$

Tabulka rozdílů mezi průměry skupin:

Průměr	Skupina	E	D	A	C
87,75	B	6,00	14,25	30,00 **	34,25 **
81,75	E	-	8,25	24,00 *	28,25 **
73,50	D		-	15,75	20,00 *
57,75	A			-	4,25
53,50	C				

Poznámka: * rozdíl statisticky průkazný (P=0,05)
** rozdíl statisticky vysoce průkazný (P=0,01)

ANALÝZA VARIANCE dvoufaktorová

bez interakce, s jedním pozorováním

Model pozorování: $x_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

kde: μ - průměrná hodnota zkoumaného znaku
α_i - efekt i-té skupiny podle faktoru A
β_j - efekt j-té skupiny podle faktoru B
ε_{ijk} - vliv náhodných veličin

Faktor A / Faktor B	1	2	…i …	k	Součet	Průměr
1	x_{11}	x_{21}	x_{i1}	x_{k1}	$X_{.1}$	$x_{.1}$
2	x_{12}	x_{22}	x_{i2}	x_{k2}	$X_{.2}$	$x_{.2}$
…						
j	x_{1j}	x_{2j}	x_{ij}	x_{kj}	$X_{.j}$	$x_{.j}$
…						
n	x_{1n}	x_{2n}	x_{in}	x_{kn}	$X_{.n}$	$x_{.n}$
Součet	$X_{1.}$	$X_{2.}$	$X_{i.}$	$X_{k.}$	$X_{..}$	
Průměr	$x_{1.}$	$x_{2.}$	$x_{i.}$	$x_{k.}$		$x_{..}$

Testové kritérium:

- pro sloupce: $F_{[(k-1),(k-1)(n-1)]} = \frac{MS_A}{MS_e}$

- pro řádky: $F_{[(n-1),(k-1)(n-1)]} = \frac{MS_B}{MS_e}$

Vyhodnocení:

Pro každý faktor zvlášť, tj. samostatně pro sloupce a pro řádky. Poté průkaznost mezi dvojicemi následnými testy.

Tabulka analýzy variance

Zdroj variability	Součet čtverců S	Stupně volnosti ν	Průměrný čtverec MS
Sloupce (faktor A)	$S_A = n\sum_{i=1}^{k}(x_{i.} - x..)^2$	$\nu_A = k-1$	$MS_A = \frac{S_A}{\nu_A}$
Řádky (faktor B)	$S_B = k\sum_{j=1}^{n}(x_{.j} - x..)^2$	$\nu_B = n-1$	$MS_B = \frac{S_B}{\nu_B}$
Reziduum	$S_e = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - x_{i.} - x_{.j} + x..)^2$	$\nu_e = (k-1)(n-1)$	$MS_e = \frac{S_e}{\nu_e}$
Celkem	$S_T = \sum\sum(x_{ij} - x..)^2$	$\nu_T = nk-1$	

Zdroje proměnlivosti	Součty čtverců S	Stupně volnosti ν	Průměrný čtverec MS	Testové kriterium F
odrůdy	22,69	1	22,69	23,39++
spon B	15,19	1	15,19	15,65++
hnojení C	43,56	3	14,52	14,96++
opakování R	5,54	2	2,77	2,85
interakce A.B	0,02	1	0,02	0,02
A.C	0,40	3	0,13	0,13
B.C	3,56	3	1,18	1,21
A.B.C	0,39	3	0,13	0,13
reziduum	29,13	30	0,97	-
celková proměnlivost	120,48	47	-	-

8 Ukázka testování regresního modelu a jeho parametrů ve statistickém systému UNISTAT

Závislost mezi cenou a požadovaným množstvím
Závisle proměnná: množství

	Koeficient	Směr. chyba	t-hodnota	Významnost
Konstanta	248.6818181818	14.8997998937	16.6902790612	0.0000
Směrnice	-9.608391608392	1.504787216347	-6.385216131563	0.0001

Reziduální suma čtverců =	809.5174825175
Směrodatná chyba =	8.997318947984
Průměr Y =	155
Směrodatná odch. y =	19.33
Index determinace =	0.80303711
F(1,10) =	40.77098504677
významnost F =	0.0001
Počet řádků =	12

Analýza rozptylu regrese

Vliv	Suma čtverců	St.v.	Rozptyl	F-hod.	význ.
Regrese	3300.483	1	3300.483	40.771	0.0001
Chyba	809.517	10	80.952		
Celkem	4110.000	11	373.636		

Rozklad sumy čtverců

Vliv	Suma čtverců	St.v.	Rozptyl	F-hod.	Význ.
cena	3300.483	1	3300.483	40.771	0.0001
Celkem	3300.483	1	3300.483	40.771	0.0001

95% interval spolehlivosti pro koeficienty regresní funkce

Koeficient	Hodnota	Směrodatná ch.	dolní mez	Horní mez
konstanta	248.6818181818	14.8997998937	215.4830	281.8806
směrnice	-9.608391608392	1.504787216347	-12.9613	-6.2555

95% interval spolehlivosti pro přímku a pás spolehlivosti

	dolní m.pás	dolní mez př.	Teoret. Y	horní mez př.	Horní m.pás
1	158.6108	170.5370	181.4231	192.3092	204.2354
2	154.4311	167.1109	176.6189	186.1269	198.8066
3	150.1395	163.5734	171.8147	180.0560	193.4898
4	145.7279	159.8651	167.0105	174.1559	188.2931
5	141.1894	155.8964	162.2063	168.5162	183.2231
6	136.5194	151.5546	157.4021	163.2496	178.2848
7	131.7152	146.7504	152.5979	158.4454	173.4806
8	126.7769	141.4838	147.7937	154.1036	168.8106
9	121.7069	135.8441	142.9895	150.1349	164.2721
10	116.5102	129.9440	138.1853	146.4266	159.8605
11	111.1934	123.8731	133.3811	142.8891	155.5689
12	105.7646	117.6908	128.5769	139.4630	151.3892

Příklad: Zjistěte statistickou průkaznost závislosti mezi počtem zaměstnanců a tržbami. Testování proveďte na hladině významnosti $\alpha = 0{,}05\%$. Úkol proveďte pro přímku, s pomocí výpočetní techniky i pro parabolu. Výsledek komentujte.

počet zaměst.	*tržby v mil. Kč.*
10	2
15	6
20	9
25	11
30	12
35	12.5
40	12.9
45	13
50	12
55	8

Po proložení přímkou byla zjištěny tyto výsledky:

	Koeficient
Abs.člen	5.026060606061
směrnice	0.14812121

Reziduální suma čtverců =	72.15296969697
Směrodatná chyba	3.003185177794
Průměr Y =	9.84
Směr. Odch. Y =	3.611770879899
Index determinace =	0.38543006

Výsledky:

	Coefficient	Standard Error	t-Statistic	Significance
Constant	5.026060606061	2.349637292995	2.139079346861	0.0649
poczam	0.14812121	0.066127961	2.239918022592	0.0554

Residual Sum of Squares =	72.15296969697
Standard Error =	3.003185177794
Mean of Y =	9.84
Stand Dev of y =	3.611770879899
R-squared =	0.38543006
Adjusted R-squared =	0.38543006
F(1,8) =	5.017232747933
significance of F =	0.0554
Number of Rows =	10

ANOVA of Regression

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Signif
Regression	45.251	1	45.251	5.017	0.0554
Error	72.153	8	9.019		
Total	117.404	9	13.045		

95% Confidence Intervals for Regression Coefficients

Constant	Coefficient	Standard Error	Lower Bound	Upper Bound
abs.člen	5.026060606061	2.349637292995	-0.3922	10.4443
směrnice	0.14812121	0.066127961	-0.0044	0.3006

95% Confidence Intervals for Mean and Actual Y Values

	lb Actual Y	lb Mean of Y	Fitted Y	ub Mean of Y	ub Actual Y
1	-1.5257	2.4369	6.5073	10.5777	14.5403
2	-0.4902	3.7957	7.2479	10.7000	14.9860
3	0.4792	5.0851	7.9885	10.8918	15.4978
4	1.3762	6.2584	8.7291	11.1997	16.0820
5	2.1963	7.2468	9.4697	11.6926	16.7431
6	2.9369	7.9874	10.2103	12.4332	17.4837
7	3.5980	8.4803	10.9509	13.4216	18.3038
8	4.1822	8.7882	11.6915	14.5949	19.2008
9	4.6940	8.9800	12.4321	15.8843	20.1702
10	5.1397	9.1023	13.1727	17.2431	21.2057

Závěr: Po proložení přímkou lze zjistit, že model není statisticky významný. Je proto třeba zvolit jiné, vhodnější proložení. V tomto případě odpovídá zjištěným datům parabola, kdy všechny testy vycházení průkazné.

Příklad: Otestujte model, koeficienty regresní funkce, korelační koeficient u závislosti mezi prodejem automobilů a spotřebou pohonných hmot.

Prodej	*PHM*
143	345
150	340
165	350
190	400
170	380
178	390
210	410
148	344
130	320
250	450

Vypočítané hodnoty:

	koeficient
konstanta	184.2049299178
směrnice	1.088206863219

Index determinace = 0.95169887

Výsledky:

Směr. Chyba	t-hodnota	Significance
15.31633696677	12.02669609043	0.0000
0.086675298	12.55498264334	0.0000

Analýza rozptylu regrese

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Signif
Regression	13720.547	1	13720.547	157.628	0.0000
Error	696.353	8	87.044		
Total	14416.900	9	1601.878		

Index determinace = 0.95169887
F(1,8) = 157.6275891746
significance of F = 0.0000

Literatura

STÁVKOVÁ, J., DUFEK, J. *Biometrika.* 1. vyd. Brno: Mendelova zemědělská a lesnická univerzita v Brně, 2000. 178 s. ISBN 80-7157-486-4.

ANDĚL, J. *Statistické metody.* 1. vyd. Praha: MATFYZPRESS, 1993. 246 s.

MELOUN, M., MILITKÝ, J. *Kompendium statistického zpracování dat : metody a řešené úlohy včetně CD.* 1. vyd. Praha: Academia, 2002. 764 s. ISBN 80-200-1008-4.

MENDENHALL, W., SINCICH, T. *Statistics for the Engineering and Computer Sciences.* 2. vyd. San Francisco: Dellen Publishing Company, 1988. 16 s. ISBN 0-02-380460-2.

NAVIDI, W. *Statistics for engineers and scientists.* Boston: McGraw-Hill, 2006. 869 s. ISBN 0-07-121492-5.

ROD, J., VONDRÁČEK, J. *Polní pokusnictví : Pokusnická technika se základy biometriky.* Brno: VŠZ, 1975. 230 s.

SEGER, J., HINDLS, R. *Statistické metody v tržním hospodářství.* 1. vyd. Praha: Victoria Publishing, 1995. 435 s. ISBN 80-7187-058-7.

PALÁT, M. Aplikace biometrických metod a modelování v lesnické ekologii. In FLAK, P. *Biometrické metódy a modely v pódohospodárskej vede, výskume a výučbe. XVI. letná škola biometriky, Račkova dolina, 21. - 25. júna 2004.* Nitra: VES SPU v Nitre, 2004, s. 265-277. ISBN 80-891620-6-1.

Seminář základů statistiky a workshop – Ing. Kristina Somerlíková

V teoretické části semináře jsou vysvětleny základní pojmy a charakteristiky a objasněny používané statistické metody. V následující praktické části, budou uvedené charakteristiky a metody prakticky ukázány na souvislém příkladě.

Soukromý zemědělec vlastní stádo mléčného skotu tří různých plemen různého stáří. Jeho hlavním produktem je mléko, vede si denní záznamy o produkci jednotlivých krav.

1. Navrhněte tabulku rozdělení četností z uvedených dat. Dopočítejte relativní četnost a kumulativní četnosti. Grafické zobrazení četností.
2. Nalezněte významné hodnoty variační řady. Analýza struktury. Sestrojení Lorenzovy koncentrační křivky.
3. Vypočítejte z uvedených dat charakteristiky obecné úrovně a charakteristiky variability. Pracujte s daty tříděnými i netříděnými.
4. Výpočet regresní úlohy. Výpočet indexu korelace. Grafické znázornění regresní funkce.
5. Výpočet sdružených regresních přímek a korelačního koeficientu. Grafické znázornění přímek.
6. Měření závislosti slovních znaků. Výpočet koeficientů kontingence a asociace.
7. Střední a přípustná chyba výběru, stanovení rozsahu výběrového souboru.
8. Výpočet konfidenčních intervalů pro střední hodnotu, rozptyl a směrodatnou odchylku, jejich grafické zobrazení.
9. Testování homogenity rozptylu, t – testy: testování významnosti rozdílu dvou střeních hodnot u nezávislých i závislých souborů.
10. Jednofaktorová a vícefaktorová analýza rozptylu.
11. Metody následného testování.