Jednoduchá lineární závislost

Regresní funkce: $y' = f(x, b_0, ..., b_m)$

Předpoklad: Funkce je lineární v parametrech:

$y' = b_0 f_0(x) + ... + b_m f_m(x)$

$f_0(x)$ ... $f_m(x)$ = regresory

b_0 ... b_m = regresní parametry – určujeme **METODOU NEJMENŠÍCH ČTVERCŮ**

Regresní funkce je tedy funkcí m+1 neznámých parametrů b_0, b_1, ..., b_m, jejíž hodnoty musíme nalézt tak, aby bylo splněno kritérium nejmenších čtverců: $\sum_{i=1}^{n} (y_i - y_i')^2 = \min$.

Extrém této funkce najdeme tak, že najdeme první parciální derivace postupně podle všech m+1 neznámých parametrů, položíme je rovny nule a vzniklou soustavu lineární normálních rovnic řešíme.

Pro vyhnutí se derivování využijeme pravidla, definujícího j-tou normální rovnici jako

$$\sum_{i=1}^{n} y_i f_j(x_i) - \sum_{i=1}^{n} \sum_{j=0}^{m} b_j f_j(x_i) f_j(x_i) = 0$$

$$y' = b_0 + b_1 \cdot \frac{1}{x_i} \rightarrow f_0(x) = 1,\ f_1(x) = \frac{1}{x_i}$$

$$\sum y_i \cdot 1 - \sum \left(b_0 + b_1 \cdot \frac{1}{x_i} \right) \cdot 1 = \sum y_i - n b_0 - b_1 \sum \frac{1}{x_i} = 0$$

$$\sum y_i \cdot \frac{1}{x_i} - \sum \left(b_0 + b_1 \cdot \frac{1}{x_i} \right) \cdot \frac{1}{x_i} = \sum \frac{y_i}{x_i} - b_0 \sum \frac{1}{x_i} - b_1 \sum \frac{1}{x_i^2} = 0$$

$$y' = b_0 + b_1 \cdot x_i \rightarrow f_0(x) = 1,\ f_1(x) = x_i$$

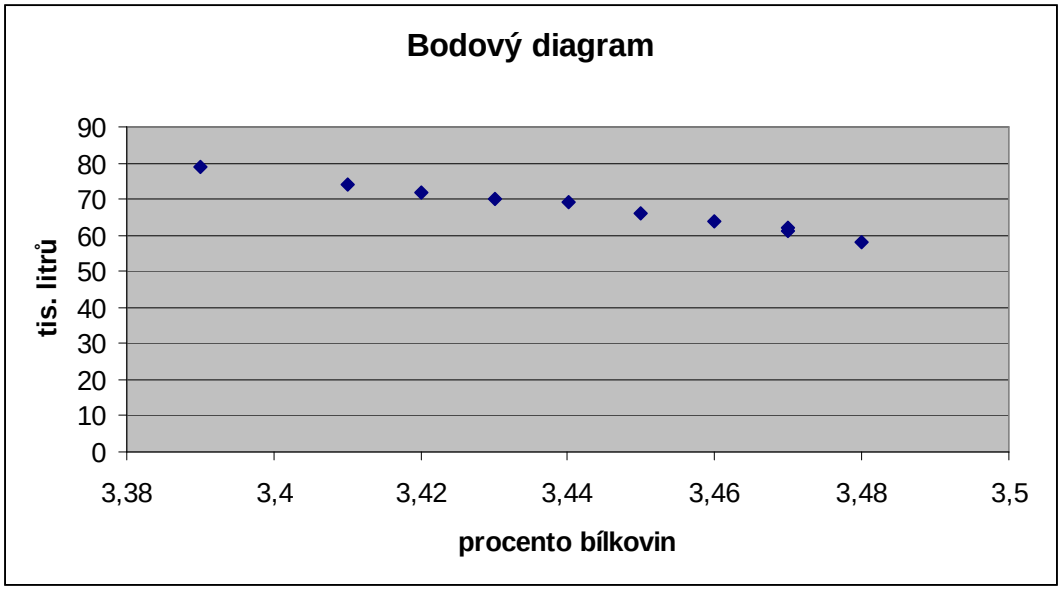
$$\sum y_i \cdot 1 - \sum (b_0 + b_1 \cdot x_i) \cdot 1 = \sum y_i - n b_0 - b_1 \sum x_i = 0$$

$$\sum y_i \cdot x_i - \sum (b_0 + b_1 \cdot x_i) \cdot x_i = \sum y_i \cdot x_i - b_0 \sum x_i - b_1 \sum x_i^2 = 0$$

Příklad: Při sledování závislosti obsahu bílkovin v mléce (v relativním vyjádření) (y) na objemu produkce v 1000 l (x) byly zjištěny následující údaje, které jsou uvedeny v tabulce:

3,39	3,41	3,42	3,43	3,44	3,45	3,46	3,47	3,47	3,48
79	74	72	70	69	66	64	62	61	58

1. Sestrojte bodový diagram (EXCEL, UNISTAT)
2. Zvolte vhodný typ funkce, určete její rovnici na základě MNČ



Lineární regrese

Výsledky regrese

Platný počet pozorování: 10, 0 Vynechán
Závisle proměnná: bílkoviny

	Koeficient	Směrodatná chyba	t-statistika	Významn.	Dolní 95%	Horní 95%
Konstanta	3,7454	0,0106	351,7640	0,0000	3,7208	3,7699
tis. litrů	-0,0045	0,0002	-28,6095	0,0000	-0,0049	-0,0041

Reziduální součet čtverců =	0,0001
Směrodatná chyba =	0,0031
Průměr Y =	3,4420
Směrodatná odchylka Y =	0,0294
Korelační koeficient =	0,9951
Čtverec R =	0,9903
Upravené R-kvadrát =	0,9891
F(1,8) =	818,5024
Významnost F =	0,0000
Durbin-Watsonova statistika =	1,1311
Log fce věrohodnosti =	44,6905
Potlačená statistika =	0,0001

Index determinace:

$$I_{yx}^2 = \frac{s_{y'}^2}{s_y^2} = \frac{\sum_{i=1}^{n}(y_i' - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum y_i \cdot y_i' - n \cdot \bar{y}^2}{\sum y_i^2 - n \cdot \bar{y}^2} = 0{,}9903$$

Index korelace:

$$I_{yx} = \sqrt{I^2{}_{yx}} = 0{,}9951$$

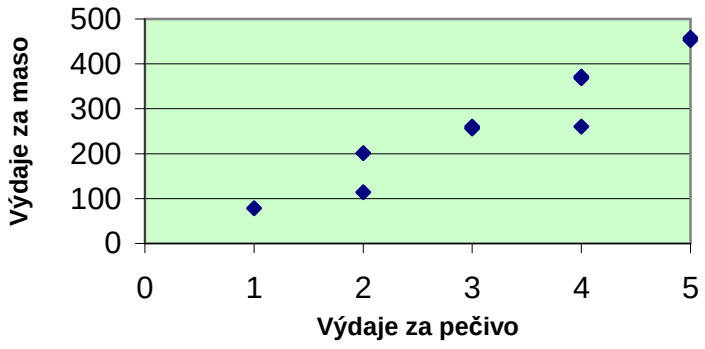
Sdružené regresní přímky

Sestrojte bodový graf, vypočtěte sílu závislosti a rovnice sdružených regresních přímek pro lineární vztah mezi výdaji za maso a masné výrobky (y) a výdaji za pečivo (x) v souboru vybraných domácností.

x_i	y_i	x_i*y_i	x_i^2	y_i^2			
4	372	1488	16	138384	n = 10	**průměr x**	3,3
5	458	2290	25	209764		**průměr y**	282
3	256	768	9	65536		**průměr x^2**	10,89
1	78	78	1	6084		**průměr y^2**	79524
4	260	1040	16	67600			
2	201	402	4	40401			
4	368	1472	16	135424			
3	260	780	9	67600			
5	453	2265	25	205209			
2	114	228	4	12996			
33	**2820**	**10811**	**125**	**948998**			

Sestrojení bodového grafu pro závisle a nezávisle proměnnou.

Bodový graf



Regresní koeficienty

$$b_{yx} = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^{\,2} - (\sum x_i)^2} = \frac{\sum y_i x_i - n\overline{xy}}{\sum x_i^{\,2} - n\overline{x}^2} = \frac{s_{xy}}{s^2{}_x}$$

$$b_{xy} = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum y_i^{\,2} - (\sum y_i)^2} = \frac{\sum y_i x_i - n\overline{xy}}{\sum y_i^{\,2} - n\overline{y}^2} = \frac{s_{xy}}{s^2{}_y}$$

Absolutní členy

$$a_{yx} = \overline{y} - b_{yx} \cdot \overline{x}$$

$$a_{xy} = \overline{x} - b_{xy} \cdot \overline{y}$$

$$b_{yx} = \frac{10811 - 10 \cdot 3{,}3 \cdot 282}{125 - 10 \cdot 10{,}89} = \frac{1505}{16{,}1} = \underline{\underline{93{,}478}}$$

$$a_{yx} = 282 - 93{,}478 \cdot 3{,}3 = \underline{\underline{-26{,}478}}$$

$$b_{xy} = \frac{10811 - 10 \cdot 3{,}3 \cdot 282}{948998 - 10 \cdot 79524} = \frac{1505}{153758} = \underline{\underline{0{,}009788}}$$

$$a_{xy} = 3{,}3 - 0{,}009788 \cdot 282 = \underline{\underline{0{,}53976}}$$

Sdružené regresní přímky:

$$\boxed{y' = a_{yx} + b_{yx} \cdot x} \quad \Rightarrow \quad y' = -26{,}478 + 93{,}478 \cdot x$$

$$\boxed{x' = a_{xy} + b_{xy} \cdot y} \quad \Rightarrow \quad x' = 0{,}53976 + 0{,}009788 \cdot y$$

Posunem počátku souřadnicové soustavy do bodu, kde se sdružené regresní přímky protínají (je to v průměrech $\bar{x}$ a $\bar{y}$) dostaneme regresní přímky v **transformovaném tvaru:**

$$\boxed{y' = \bar{y} + b_{yx}(x - \bar{x})} \quad \Rightarrow \quad y' = 282 + 93{,}478(x - 3{,}3)$$

$$\boxed{x' = \bar{x} + b_{xy}(y - \bar{y})} \quad \Rightarrow \quad x' = 3{,}3 + 0{,}009788(y - 282)$$

Jelikož jsou regresní koeficienty nesouměřitelné, provádí se - k dosažení srovnatelnosti sklonu různých regresních přímek – normování regresních koeficientů jejich násobením podílem směrodatných odchylek:

$$\boxed{\beta = \frac{s_{xy}}{s^2{}_x} \cdot \frac{s_x}{s_y} = \frac{s_{xy}}{s_x \cdot s_y}}$$

Vypočetli jsme tzv. normovaný **Beta-koeficient**, který je pro obě přímky stejný a nezávisí na zvolených měrných jednotkách.

Dospějeme k němu i normováním obou veličin:

$$\boxed{Z = \frac{Y - \bar{y}}{s_y}} \qquad \boxed{U = \frac{X - \bar{x}}{s_x}}$$

Obdržíme sdružené regresní přímky v **normovaném tvaru**:

$$\boxed{z' = \beta \cdot u} \quad \Rightarrow \quad z' = 0{,}957 \cdot u$$

$$\boxed{u' = \beta \cdot z} \quad \Rightarrow \quad u' = 0{,}957 \cdot z$$

Z rozkladu rozptylu pro metodu nejmenších čtverců vyplyne **koeficient determinace**, který je zvláštním případem indexu determinace pro **přímočarou závislost.** (Zase může být vyjádřen i v procentech.)

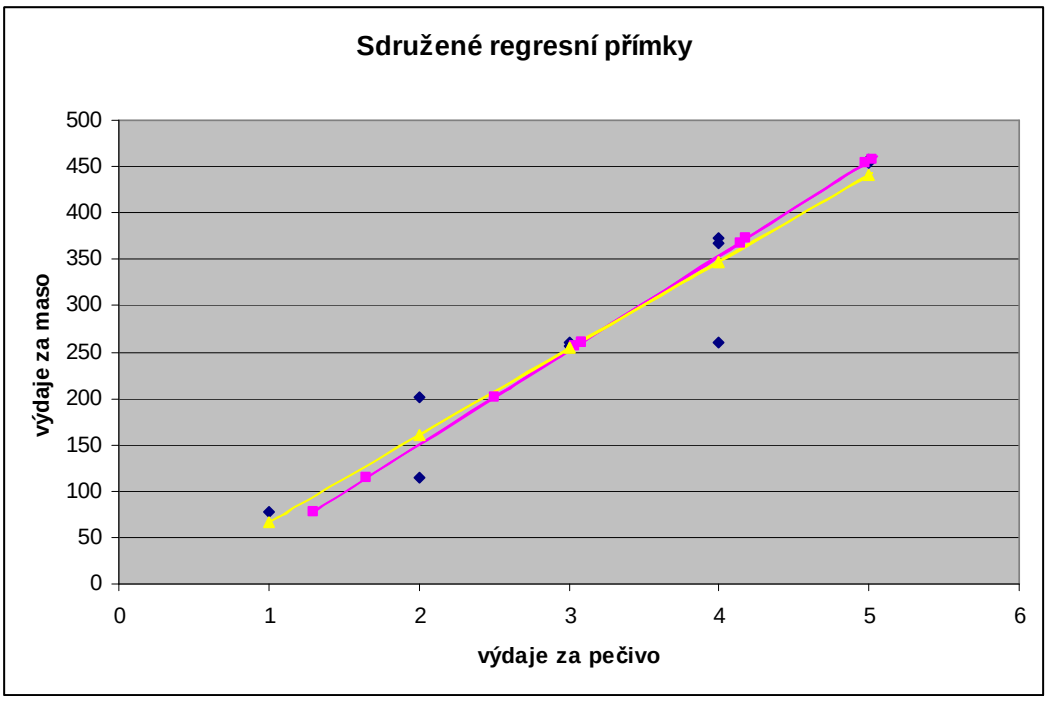
$$\boxed{I^2_{yx} = \frac{b^2_{yx} \cdot s^2_x}{s^2_y} = \frac{s^2_{xy}}{s^2_x \cdot s^2_y} = \beta^2 = r^2}$$

$$I^2_{yx} = 0{,}957^2 = \underline{\underline{0{,}9158}}$$

K vyjádření síly přímočaré závislosti slouží druhá odmocnina koeficientu korelace a tou je **koeficient korelace**:

$$\boxed{r = \frac{\sum y_i x_i - n\bar{x}\bar{y}}{\sqrt{\left[\sum x_i^2 - n\bar{x}^2\right]\left[\sum y_i^2 - n\bar{y}^2\right]}} = \frac{s_{xy}}{s_x \cdot s_y} = \pm\sqrt{b_{yx} \cdot b_{xy}}}$$

$$r = \frac{10811 - 10 \cdot 3{,}3 \cdot 282}{\sqrt{[125 - 10 \cdot 10{,}89][948998 - 10 \cdot 79524]}} = \frac{1505}{1573{,}37} = \underline{\underline{0{,}95654}}$$



Závislost slovních (kvalitativních) znaků

Naším úkolem je zjistit, zda existuje závislost (popř. jak je silná) mezi dvěma otázkami z marketingového průzkumu „Uplatnění absolventů ekonomické fakulty v praxi“.

A. Kde v současné době pracujete ?
1. ve státním podniku
2. v české soukromé firmě
3. v zahraniční či nadnárodní firmě v pracovním poměru
4. v družstvu
5. soukromě podnikám, nikoho nezaměstnávám
6. soukromě podnikám a zaměstnávám další osoby
7. jiná forma

B. Odpovídáte ve své funkci za práci jiných ?
1. ano
2. ne

Na základě odpovědí respondentů (absolventů naší fakulty) byla sestavena **kontingenční tabulka**.

Odpovědnost PRACOVIŠTĚ	ANO	NE	SOUČET ŘÁDKU
1	2	18	20
2	16	25	41
3	14	12	26
5	1	2	3
6	2	0	2
7	1	1	2
Součet sloupce	36	58	94

Kde znak *A* (otázka č. 1) nabývá obměn a_1 až a_7 a můžeme jej považovat např.za nezávisle proměnný znak, a znak *B* (otázka č. 2) nabývá obměn b_1 až b_2 a půjde o závisle proměnný znak.

K výpočtu ukazatele potřebujeme znát kromě **skutečných četností** (zjištěných průzkumem) i **četnosti teoretické** (vypočítané za předpokladu nezávislosti obou znaků), u kterých platí, že čím více se budou lišit od těch skutečných tím silnější bude závislost obou znaků.

$$n'_{ij} = \frac{n_i \cdot n_j}{n}$$

, kde n_i, n_j jsou příslušné okrajové četnosti a n je rozsah souboru.

Na základě tohoto vztahu vypočítáme teoretické četnosti pro všechny četnosti skutečné.

Očekávané četnosti	ano	ne
1	7,6596	12,3404
2	15,7021	25,2979
3	9,9574	16,0426
5	1,1489	1,8511
6	0,7660	1,2340
7	0,7660	1,2340

Míru intenzity vzájemné závislosti dvou slovních znaků v kontingenční tabulce měří **Čtvercová kontingence χ^2.**

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

Čtvercová kontingence může nabývat libovolných nezáporných hodnot, nejsme schopni určit pomocí této míry sílu závislosti, proto konstruujeme různé míry kontingence, které z ní vycházejí:

Průměrná čtvercová kontingence Φ^2: $$\Phi^2 = \frac{\chi^2}{n}$$

Maximální možná hodnota je opět různá.

Pearsonův koeficient kontingence P: $$P = \sqrt{\frac{\Phi^2}{1+\Phi^2}} = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Nabývá hodnot z intervalu <0, 1), hodnoty jedna nemůže nikdy dosáhnout. Hodnota je závislá na rozměrech tabulky.

Čuprovův koeficient kontingence T: $$T = \sqrt{\frac{\Phi^2}{(r-1)(s-1)}}$$

Je z intervalu <0, 1> pouze pro čtvercové kontingenční tabulky (r = s).

Cramérův koeficient kontingence C: $$C = \sqrt{\frac{\Phi^2}{\min\{r-1; s-1\}}}$$

Je z intervalu $0 \le C \le 1$ bez ohledu na velikost tabulky.

Vypočítejte uvedené míry kontingence pro naši tabulku a vyjádřete se o síle závislosti mezi otázkami. (Unistat)

Statistika Chí-kvadrát =	12,8159
Stupně volnosti =	5,0000
Pravostranná pravděpodobnost =	0,0252
Průměrná čtvercová kontingence $Fí^2$ =	0,1363
Fí =	0,3692
Cramerovo V =	0,3692
Pearsonův koeficient kontingence =	0,3464
Somerova delta (sl) =	-0,3597
Somerova delta (řád) =	-0,2478
Goodman-Kruskalova Gama =	-0,5071
Kendallovo tau b =	-0,2985
Kendallovo tau c =	-0,3400

Měření asociace

- zvláštní případ kontingenční závislosti pro $r = s = 2$,
- zvláštní případ korelační závislosti dvou znaků, z nichž každý nabývá pouze dvou hodnot – nula a jedna.

Příklad: V ovocném sadě byl proveden postřik ovocných stromů proti červivosti ovoce. Ze 450 stromů jich bylo postřikem ošetřeno 335, neošetřeno zůstalo 115. V asociační tabulce jsou uvedeny výsledky ošetření stromů vzhledem k červivosti ovoce.

Červivost / Postřik		ANO y = 1	NE y = 0	Součet
ANO	x = 1	$n_{11} = 12$	$n_{10} = 323$	$n_{1*} = 335$
NE	x = 0	$n_{01} = 53$	$n_{00} = 62$	$n_{0*} = 115$
Součet		$N_{*1} = 65$	$n_{*0} = 385$	n = 450

Kde * v indexu říká, že četnosti jsou sčítány přes index znaku, který je nahrazen hvězdičkou.

K měření intenzity asociační závislosti se používá **koeficient asociace**, který je koeficientem korelace v případě nula-jedničkových veličin (se stejnými vlastnostmi):

$$V = \frac{n \cdot n_{11} - n_{1*} \cdot n_{*1}}{\sqrt{n_{1*} \cdot n_{*1} \cdot n_{0*} \cdot n_{*0}}}$$

$$V = \frac{450 \cdot 12 - 335 \cdot 65}{\sqrt{335 \cdot 65 \cdot 115 \cdot 385}} = \underline{\underline{-0{,}527}}$$

Na základě výsledku můžeme mluvit o negativní střední závislosti mezi postřikem a červivostí ovoce.

Stanovení velikosti výběrového souboru

Klasická úvaha o velikosti souboru je, že čím je výběrový soubor větší, tím přesnější výsledky lze získat. Tato představa je správná jen za podmínek, které se v praxi málokdy podaří splnit:

1. Podíl skutečně prošetřených výběrových jednotek by nesměl záviset na velikosti výběrového souboru.

Nesměla by existovat žádná nevýběrová, systematická chyba.Testy homogenity rozptylů

Směrodatná chyba výběru

je to směrodatná odchylka výběrové charakteristiky

$$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x}_i - \mu\right)^2}{k}} = \xrightarrow[\text{úprava}]{\text{matematická}} = \frac{\sigma_x}{\sqrt{n}}$$

pro výběr bez opakování násobíme opravným koeficientem

$$s_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Směrodatnou odchylku základního souboru σ_x pouze odhadujeme:

- na základě pravidla 6 sigma $\sigma_x = \frac{x_{\max} - x_{\min}}{6}$
- nebo pomocí směrodatné odchylky výběrového souboru počítané ze stupňů volnosti

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} \quad \text{potom} \quad s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n(n-1)}} = \frac{s_x}{\sqrt{n}}$$

Pokud máme směrodatnou odchylku počítanou z *n* hodnot, použijeme opravný koeficient $\sqrt{\frac{n}{n-1}}$

Přípustná chyba výběru (Δ)

součin směrodatné chyby a koeficientu spolehlivosti (normované veličiny standardizovaného nebo Studentova rozdělení)

pro n<30 $\Delta = s_{\bar{x}} \cdot t_{1-\frac{\alpha}{2}}$

pro n>30 $\Delta = s_{\bar{x}} \cdot u_{1-\frac{\alpha}{2}}$

Δ nám říká, s jakou pravděpodobností se bude vyskytovat směrodatná chyba

Stanovení rozsahu výběrového souboru

výběr s opakováním: $n = \frac{u^2_{1-\frac{\alpha}{2}} \cdot \sigma_x^2}{\Delta^2} = \frac{t^2_{1-\frac{\alpha}{2}} \cdot s_x^2}{\Delta^2}$

platí, chceme-li odhadovat průměr

Stupně volnosti

- měří prostor (volnost) výsledků výběrů
- jednotky informací

Abychom pochopili název “stupně volnosti”, uvažujme výběr rozsahu *n* = 2 pozorování, např. 21 a 15. Průměr pak bude $\bar{x} = 18$ a odchylky 3 a -3. Druhá odchylka je záporným ekvivalentem první. Zatímco první odchylka je “volná”, druhá je přísně determinována. Je zde tedy 1 stupeň volnosti pro odchylky.
Obecně pro výběr velikosti *n* je prvních *n* - 1 odchylek volných, zatímco poslední je přísně determinována požadavkem, že součet všech odchylek je roven 0; $\left(x - \bar{x}\right) = 0$.

Určete minimální rozsah výběrového souboru pro odhad aritmetického průměru základního souboru, jestliže znáte:

$s_x = 9$, $\Delta_{0,975} = 3$, $\bar{x} = 40$

$u_{0,975} = 1,960$

$$n = \frac{1{,}96^2 \cdot 9^2}{3^2} = 34{,}57 \rightarrow 35 \text{ vzorků}$$

Tab.III - Kvantily u_p normovaného normálního rozdělení.

Určete počet vzorků, které musíte vybrat, jestliže chcete odhadnout průměrnou hmotnost vzorku s přesností p = 0,95 a s přesností

a) 1,5 g
b) 1 g
c) 0,2 g

Předvýběr 25 vzoků poskytl tyto výsledky:
s_x = 6 g, $\bar{x}$ =120

$t_{0,975}$ = 2,064 tabulková hodnota pro 24 st. volnosti
Tab.V - Kvantily t_p Studentova rozdělení
Δ - přesnosti

a) $n = \frac{2{,}064^2 \cdot 6^2}{1{,}5^2} = 68{,}16 \rightarrow 69$ vzorků

b) $n = \frac{2{,}064^2 \cdot 6^2}{1^2} = 153{,}36 \rightarrow 154$ vzorků

c) $n = \frac{2{,}064^2 \cdot 6^2}{0{,}2^2} = 3834{,}00 \rightarrow 3834$ vzorků

U rozsáhlého souboru vajec má být odhadnuta průměrná hmotnost s přesností na
a) 1 g
b) 0,5 g
c) 0,1 g
Jak rozsáhlý má být výběr vajec, aby byla dosažena požadovaná přesnost s pravděpodobností *p* = 0,99? Předvýběr 25 vajec poskytl tyto výsledky:
s_x^2 = 10 g^2 , $\bar{x}$ = 58

$t_{0,995}$ = 2,797 tabulková hodnota pro 24 st. volnosti

a) $n = \frac{2{,}797^2 \cdot 10}{1^2} = 78{,}23 \rightarrow 79$ vzorků

b) $n = \frac{2{,}797^2 \cdot 10}{0{,}5^2} = 312{,}92 \rightarrow 313$ vzorků

c) $n = \frac{2{,}797^2 \cdot 10}{0{,}1^2} = 7823{,}00 \rightarrow 7823$ vzorků

Bodový odhad

- odhadujeme základní charakteristiku (*T*) pomocí výběrové charakteristiky (*t*) jako jediné číslo

Pravděpodobnost bezchybného odhadu je rovna 0. Chyby se dopouštíme s pravděpodobností 1.

Intervalový odhad

- odhad příslušné charakteristiky (*T*) základního souboru pomocí intervalu
- odhad je reprezentován tzv. **intervalem spolehlivosti** (**konfidenčním intervalem**), který s danou pravděpodobností bude obsahovat skutečnou hodnotu odhadované charakteristiky základního souboru. Tato pravděpodobnost se nazývá **spolehlivostí odhadu** a značí se 1 - α. Čím větší pravděpodobnost, tím je odhad spolehlivější. Pravděpodobnost opačného jevu, tj 1 - (1 - α) = α se nazývá **riziko odhadu**.

Interval spolehlivosti pro střední hodnotu

vycházíme z normálního ($n>30$) nebo Studentova ($n \leq 30$] rozdělení

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

kde: $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

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

Interval spolehlivosti pro rozptyl

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

Interval spolehlivosti pro směrodatnou odchylku

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

Odhadněte s pravděpodobností 0,95 pomocí oboustranného intervalu spolehlivosti průměrnou hmotnost živě narozených selat, když u 100 náhodně vybraných jedinců byly zjištěny tyto hmotnosti:

hmotnost (kg)	1,7	1,8	1,9	2,0	2,1

počet selat	7	20	45	18	10

$n = 100$, $\bar{x} = 1{,}094$

$$\sum_{i=1}^{n} x_i \cdot n_i = 190{,}4, \qquad \sum_{i=1}^{n} x_i^2 \cdot n_i = 363{,}58$$

$$s_x^2 = \frac{363{,}58}{100} - 1{,}904^2 = 3{,}6358 - 3{,}625 = 0{,}01, \quad s_x = 0{,}103$$

Tabulky: kvantily u_p normovaného normálního rozložení: $u_{0{,}975} = 1{,}960$

$$P[1{,}904 - 1{,}96 \cdot \frac{0{,}103}{\sqrt{100}} \le \mu \le 1{,}904 + 1{,}96 \cdot \frac{0{,}103}{\sqrt{100}}] = 0{,}95$$

$$P(1{,}884 \le \mu \le 1{,}91) = 0{,}95$$

Odhadněte s pravděpodobností 0,95 pomocí oboustranného intervalu spolehlivosti průměrnou hmotnost všech jablek určité odrůdy, když u 100 vzorků náhodně vybraných bylo zjištěno:

hmotnost (g)	140	145	150	155	160
počet jablek	10	18	45	21	6

n = 100, $\bar{x} = 149{,}75$

$$\sum_{i=1}^{n} x_i \cdot n_i = 14975 \qquad \sum_{i=1}^{n} x_i^2 \cdot n_i = 2245075$$

$$s_x^2 = \frac{224075}{100} - 149{,}75^2 = 22450{,}75 - 22425{,}0625 = 25{,}6875$$

$s_x = 5{,}07$, $u_{0{,}975} = 1{,}960$

$$P[149{,}75 - 1{,}96 \cdot \frac{5{,}07}{\sqrt{100}} \le \mu \le 149{,}75 + 1{,}96 \cdot \frac{5{,}07}{\sqrt{100}}] = 0{,}95$$

$$P(148{,}76 \le \mu \le 150{,}74) = 0{,}95$$

Odhadněte variabilitu hmotnosti jablek s pravděpodobností 0,95.

v=99, $\chi^2_{0{,}025}=73{,}34$ $\chi^2_{0{,}975}=128{,}45$
v=100, $\chi^2_{0{,}025}=74{,}20$ $\chi^2_{0{,}975}=129{,}60$

$$P\left[\frac{99 \cdot 25{,}6875}{128{,}45} \le \sigma_x^2 \le \frac{99 \cdot 25{,}6875}{73{,}34}\right] = 0{,}95$$

$$P\left[19{,}8 \le \sigma_x^2 \le 34{,}67\right] = 0{,}95$$

$$P\left[4{,}45 \le \sigma_x \le 5{,}89\right] = 0{,}95$$

Určete oboustranný interval spolehlivosti aritmetického průměru základního souboru, jestliže znáte:

$n = 25$, $\bar{x} = 50$, $s_x^2 = 12$, $\alpha = 0{,}05$

$v = 24$ $t_{0{,}975} = 2{,}064$

$$P[50 - 2{,}064 \cdot \frac{3{,}46}{\sqrt{25}} \le \mu \le 50 + 2{,}064 \cdot \frac{3{,}46}{\sqrt{25}}] = 0{,}95$$

$$P(48{,}57 \le \mu \le 51{,}43) = 0{,}95$$

Testování statistických hypotéz

Statistická hypotéza - určitý předpoklad o statistických datech vyslovený dřív, než došlo ke zkoumání dat.
Testování - procedura vedoucí k zamítnutí nebo nezamítnutí hypotézy v podmínkách nejistoty.
Test významnosti - smyslem testování je ověřit, zda rozdíl mezi skutečnou (naměřenou) a předpokládanou hodnotou je statisticky významný.

Postup

1. Formulace hypotézy → stanovení nulové hypotézy H_0
např. $H_0 \equiv \mu_1 = \mu_2$ → $H_0 \equiv \mu_1 - \mu_2 = 0$
aby byla ověřitelná, musí být zformulována v negativním smyslu
H_1 alternativní hypotéza → přijmeme ji, jestliže nepřijmeme H_0
už ji netestujeme
oboustranné x jednostranné

2. Volba hladiny významnosti α
hladina významnosti - pravděpodobnost chybného zamítnutí pravdivé hypotézy $\alpha \to 0$ (α = 0,05; α = 0,01)

3. Provedení náhodného výběru, výpočet testového kritéria T a stanovení jeho rozdělení

4. Vyhodnocení testu
$T_{vyp} < T_{tab}$ → H_0 se nezamítá (rozdíl je statisticky nevýznamný)
$T_{vyp} > T_{tab}$ → H_0 se zamítá (rozdíl je statisticky významný nebo vysoce významný)

Chyby při testování

1. Chyba prvního druhu (pravděpodobnost → α) - chybné zamítnutí H_0
2. Chyba druhého druhu (pravděpodobnost → β) - chybné nezamítnutí nesprávné H_0

Testy

- parametrické - veličiny v normálním rozložení, odhady para-metrů
- neparametrické - neznáme zákon rozložení veličiny, vychází z velikostního třídění jednotek podle zkoumaných znaků

Testování homogenity rozptylu

$H_0 \equiv \sigma_1^2 = \sigma_1^2 = \ldots \sigma_k^2 = \sigma^2$

- u dvou rozptylů: $H_0 \equiv \frac{\sigma_1^2}{\sigma_2^2} = 1$

testové kritérium $F = \frac{s^2_{\max}}{s^2_{\min}}$ o n_1-1 a n_2-1 stupních volnosti

n_1 - výběr s větším rozptylem
n_2 - výběr s menším rozptylem

- u více rozptylů:

a) výběry mají stejný rozsah:

Davidův test $V_{(k,n-1)} = \frac{s^2_{\max}}{s^2_{\min}}$

Cochranův test $q_{(k,n-1)} = \frac{s^2_{\max}}{\sum_{i=1}^{n} s_i^2}$

b) výběry mají různý rozsah:

Bartletův test $\chi^2_{(k-1)} = \frac{2{,}30259}{C}\left[(n-k)\log s^2 - \sum_{i=1}^{k}(n_i-1)\log s_i^2\right]$

event. $B = \frac{1}{C}\left[(n-k)\ln s^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2\right]$,

kde s_i^2 (i=1, ..., k) je nestranný výběrový rozptyl

$$s^2 = \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{n-k}, \quad n = \sum_{i=1}^{k} n_i, \quad C = 1 + \frac{1}{3(k-1)}\cdot\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right)$$

Testování průkaznosti rozdílu mezi průměry

$H_0 \equiv \mu_1 = \mu_2 = \ldots \mu_k = \mu$
Předpokladem použití testu je potvrzení normality rozdělení a homogenity rozptylů.

1. Testujeme průměr základního souboru (μ) a výběrového ($\bar{x}$):

$$\mu,\ \bar{x}: \quad \underline{\underline{t_{(n-1)} = \frac{|\bar{x}-\mu|}{s_{\bar{x}}}}} = |\bar{x}_1 - \mu| \cdot \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

2. Testujeme průměry výběrových souborů ($\bar{x}_1$, $\bar{x}_2$):

a) stejné rozsahy ($n_1 = n_2 = n$)

$$\underline{\underline{t_{(2n-2)} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}}}} = |\bar{x}_1 - \bar{x}_2| \cdot \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2 + \sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}}$$

b) různé rozsahy ($n_1 \neq n_2$)

$$t_{(n_1+n_2-2)} = \left|\bar{x}_1 - \bar{x}_2\right| \cdot \sqrt{\frac{(n_1+n_2-2)\, n_1 \cdot n_2}{(n_1+n_2)\cdot\left[\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2 + \sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2\right]}}$$

Výpočtový tvar pro čtverec odchylek:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

3. Párový t-test

(testování průkaznosti rozdílu mezi dvěma průměry závislých souborů)
hodnocení na základě rozdílů mezi jednotlivými páry, takže se ze dvou výběrových souborů původních hodnot dostane jediný soubor rozdílů.

$$t_{(n-1)} = \frac{\left|\bar{d}-\mu_d\right|}{s_{\bar{d}}} = \frac{\bar{d}-0}{s_{\bar{d}}} = \frac{\bar{d}}{s_{\bar{d}}}$$

kde $\bar{d} = \dfrac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_{2i})}{n} = \dfrac{\sum_{i=1}^{n}d_i}{n}$

$$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n(n-1)}} = \sqrt{\frac{\sum_{i=1}^{n}d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}d_i\right)^2}{n(n-1)}}$$

$H_0 \equiv E(D) = 0$ Náhodná veličina D má normální rozložení se střední hodnotou $E(D)$ a disperzí $D^2(D)$.

Máme rozhodnout na (hladině významnosti α = 0,05), zda dvě váhy pracují se stejnou náhodnou chybou. Máme k dispozici vždy 7 měření od každé váhy, přičemž s_1 = 0,198 a s_2 = 0,098.

$$H_0 \equiv \frac{\sigma_1^2}{\sigma_2^2} = 1$$

F-test $\quad F = \dfrac{0{,}198^2}{0{,}098^2} = 4{,}08$

$\upsilon_1 = 6,\ \upsilon_2 = 6,\ \alpha = 0{,}05 \rightarrow F_{tab} = 4{,}28$ (při $\alpha = 0{,}01 \rightarrow F_{tab} = 8{,}47$)

$F_{vyp} < F_{tab} \rightarrow$ nezamítáme H_0

Rozptyly jsou homogenní.

Předchozí příklad doplníme o další váhu se stejným počtem měření a zjištěnou $s_3 = 0{,}206$.

K ověření hypotézy $H_0 \equiv \sigma_1^2 = \sigma_2^2 = \sigma_1^2$ použijeme test kritéria $Q \rightarrow$ Cochranův test

$$q_{(3,6)} = \frac{0{,}206^2}{0{,}198^2 + 0{,}098^2 + 0{,}206^2} = 0{,}465$$

$n = 3$ - počet rozptylů, $\upsilon = 6$ - stupně volnosti
$\alpha = 0{,}05 \rightarrow q_{tab(3,6)} = 0{,}6770$

$F_{vyp} < F_{tab} \rightarrow$ nezamítáme H_0. Rozptyly jsou homogenní (náhodná chyba měření není závislá na použité váze).

Automat má dávkovat krmnou směs po 100 g. Technická kontrola vybrala náhodně 50 vzorků, u kterých byla zjištěna přesná hmotnost. Rozhodněte, zda se hmotnost směsi statisticky průkazně neliší od požadované normy.

Hmotnost (g)	96	98	100	102	104	Σ
Počet vzorků	7	29	9	3	2	50
$x_i \,.\, n_i$	672	2842	900	306	208	4928
$x_i^2 \,.\, n_i$	64512	278516	90000	31212	21636	485872

$$\bar{x} = \frac{4928}{50} = 98{,}56$$

$$s_x^2 = \frac{485872}{50} - 98{,}56^2 = 3{,}37$$ upraveno opravným koeficientem $\rightarrow$

$$s_{x(n-1)}^2 = 3{,}37 \cdot \frac{50}{49} = 3{,}44$$

$$s_x = 1{,}85$$

$$s_{\bar{x}} = \frac{s_{x(n-1)}}{\sqrt{n}} = \frac{1{,}85}{\sqrt{50}} = 0{,}262$$

$$t_{(49)} = \frac{|98{,}5 - 100|}{0{,}262} = 5{,}496\ **$$

$t_{tab(0{,}975)} = 2{,}010$
$t_{tab(0{,}995)} = 2{,}682$

$t_{vyp} > t_{tab} \rightarrow$ zamítáme H_0.

Rozhodněte, zda se průkazně liší délka klasů 2 odrůd pšenice obecné, pěstované ve stejných podmínkách, když u 100 vzorků každé odrůdy bylo zjištěno:

$\bar{x}_1 = 69{,}5mm,\ s_{x_1} = 4{,}18mm$

$\bar{x}_1 = 66{,}1mm,\ s_{x_2} = 3{,}90mm$

$$s_{\bar{x}_1} = \frac{4{,}18}{\sqrt{100}} = 0{,}418 \quad s_{\bar{x}_2} = \frac{3{,}90}{\sqrt{100}} = 0{,}390$$

$$t_{2n-2(198)} = \frac{|69{,}5 - 66{,}1|}{\sqrt{0{,}418^2 + 0{,}39^2}} = \frac{3{,}4}{0{,}572} = 5{,}94\,**$$

$t_{tab(0{,}975)} = 1{,}960$
$t_{tab(0{,}995)} = 2{,}576$

$t_{vyp} > t_{tab} \rightarrow$ zamítáme H_0. Délka klasů se vysoce průkazně liší.

Zjistěte, zda existuje průkazný rozdíl v hmotnosti kokosových ořechů vypěstovaných na různých místech ostrova. Z každého místa je oznámen jiný počet měření.

$$n_1 = 10 \quad \sum x_1 = 3{,}500 \quad \sum x_1^2 = 1{,}230 \quad (\bar{x}_1 = 0{,}350 kg)$$
$$n_2 = 8 \quad \sum x_2 = 2{,}400 \quad \sum x_2^2 = 0{,}800 \quad (\bar{x}_2 = 0{,}300 kg)$$

$$t_{(10+8-2)} = |0{,}350 - 0{,}3| \cdot \sqrt{\frac{(10+8-2)\cdot 10 \cdot 8}{(10+8)\left[\left(1{,}23 - \frac{3{,}5^2}{10}\right) + \left(0{,}8 - \frac{3{,}4^2}{8}\right)\right]}}$$

$$t_{(16)} = 0{,}05 \cdot \sqrt{\frac{1280}{18[0{,}005 + 0{,}08]}} = 1{,}446$$

$t_{tab16(0{,}975)} = 2{,}12$

$t_{vyp} < t_{tab} \rightarrow$ nezamítáme H_0. Mezi hmotností kokosových ořechů z různých míst nebyl prokázán rozdíl.

Příklad na párový t-test

Je třeba porovnat 2 metody určování obsahu cukru (%) v bulvách cukrovky. Bylo náhodně vybráno 15 bulev a pro každou z nich bylo oběmi metodami stanoveno % cukru.
Rozdíly (diference) mezi oběmi metodami byly:

Číslo vzorku	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Diference	0,2	0,0	0,1	0,5	-0,2	0,4	0,1	-0,3	-0,1	0,2	0,3	-0,2	0,1	0,1	-0,1

Zjistěte, zda existuje průkazný rozdíl v určování % cukru mezi oběmi metodami.

$$\sum d_i = 1{,}1 \quad \sum d_i^2 = 0{,}81$$

$$\bar{d} = \frac{1{,}1}{15} = 0{,}073$$

$$s_{\bar{d}} = \sqrt{\frac{0{,}8 - \frac{1}{15}(1{,}1)^2}{15 \cdot 14}} = 0{,}059$$

anebo $$s_d^2 = \left(\frac{\sum_{i=1}^{n} d_i^2}{n} - \bar{d}^{\,2} \right) \cdot \frac{n}{n-1} = \left(\frac{0{,}81}{15} - 0{,}073^2 \right) \cdot \frac{15}{14} = 0{,}052$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{\sqrt{0{,}052}}{\sqrt{15}} = \frac{0{,}228}{3{,}873} = 0{,}059$$

$$t_{(14)} = \frac{|0{,}073|}{0{,}059} = 1{,}24$$

$t_{tab14(0{,}975)} = 2{,}145$

$t_{vyp} < t_{tab} \rightarrow$ nezamítáme H_0. Není průkazný rozdíl mezi metodami.

Analýza rozptylu

Modely slouží k tomu, aby se jich používalo, nikoli k tomu, aby se jim věřilo.

Henri Theil

- metoda testování průkaznosti rozdílu mezi průměry několika souborů na sobě nezávislých (porovnáváme dva a více výběrů a chceme zjistit, zda tyto výběry mohou vycházet ze společného základního souboru, zda zjištěné odchylky lze vysvětlit jako náhodné)
- hodnocení biopokusů – polní pokusnictví
- správná volba uspořádání pokusu:
  1. slouží k ověření účinnosti ověřovaných zásahů, tj. faktorů na sledovaný pokusný materiál
  2. slouží k podchycení nekontrolovatelných zdrojů proměnlivosti (půdní rozdíly)
  3. slouží ke snížení vlivu náhodných zdrojů proměnlivosti vzniklých nekontrolovatelnými vlivy (počasí, poškození, chyba).
- vhodné matematicko – statistické zhodnocení
- úkolem je rozčlenit celkovou variabilitu na dílčí složky (podle vlivu jednotlivých sledovaných faktorů) a na složku reziduální (nelze vysvětlit – neznámé, náhodné faktory)

Jednofaktorová analýza rozptylu

A. Tabulka uspořádání dat (jednofaktorová)

Pozorování (jedinci)	**Faktor A**					**Celkem**
	$\mathbf{a_1}$	$\mathbf{a_2}$	**…**	$\mathbf{a_i}$	$\mathbf{a_a}$	
1	$\mathbf{y_{11}}$	$\mathbf{y_{21}}$				
2	$\mathbf{y_{12}}$					
…	**…**					
j	$\mathbf{y_{1j}}$					
…	**…**					
$\mathbf{n_i}$	$\mathbf{y_{1ni}}$					
Součty	$\mathbf{Y_{1.}}$	$\mathbf{Y_{2.}}$	**…**	$\mathbf{Y_{i.}}$		$\mathbf{Y_{..}}$
Průměry	$\mathbf{y_{1.}}$	$\mathbf{y_{2.}}$	**…**	$\mathbf{y_{i.}}$		$\mathbf{y_{..}}$

Rozptyl $$s_i^2 = \left(\frac{\sum_{i=1}^{n} y_{ij}^2}{n_i} - \bar{y}^2\right)\frac{n_i}{n_i - 1}$$

Faktor **A** má počet úrovní $\mathbf{a_1, a_2, \ldots, a_a}$.
Faktor **B** má počet úrovní $\mathbf{b_1, b_2, \ldots, b_b}$.
Faktor **R** má počet úrovní $\mathbf{r_1, r_2, \ldots, r_r}$.
Naměřené hodnoty se značí **y**, např. $\mathbf{y_{1,2,3}}$ → obecně $y_{i,j,k}$
Součty se značí **Y**
Tečková symbolika - zajišťuje přesnost a výstižnost

součty pro úroveň faktoru **A**: $Y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$, pro úroveň faktoru **B**: $Y_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{r} y_{ijk}$,

pro opakování **R**: $Y_{..k} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ijk}$, součet všech naměřených hodnot: $Y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$

Obdobně (ale malými písmeny) se značí průměry pro jednotlivá kritéria, např. $\mathbf{y_{i..}}$, $\mathbf{y_{.j.}}$,$\mathbf{y_{..k}}$, $\mathbf{y_{...}}$
Nemůže dojít k záměně, protože naměřená hodnota má vždy všechny indexy vyplněné (nemá tečku) - $\mathbf{y_{i,j,k}}$

B. Testování homogenity rozptylu

1. Cochranův test (pro stejný rozsah výběrových souborů)
$H_0 \equiv \sigma_1^2 = \sigma_1^2 = \ldots \sigma_k^2 = \sigma^2$

Testové kritérium $Q_{(k,n-1)} = \dfrac{s_{\max}^2}{\sum_{i=1}^{n} s_i^2}$, porovnáváme s tabulkovou hodnotou $q_{0,05\ (0,01)}$

pro počet výběrů **k** a **n -1** stupňů volnosti

2. Bartlettův test (pro různé rozsahy souborů) $n_i > 6$
$H_0 \equiv \sigma_1^2 = \sigma_1^2 = \ldots \sigma_k^2 = \sigma^2$

Testové kritérium $\chi^2_{(k-1)} = \dfrac{\ln 10}{C}\left[(N-k)\log s^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2\right]$

Tabulka: **k** - počet výběrů
n_1, n_2 … rozsahy
χ^2 - Pearsonovo rozdělení (k - 1 stupňů volnosti)
s_i^2 (i=1, ..., k) je nestranný výběrový rozptyl

kde $s^2 = \dfrac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k}$

$n = \sum_{i=1}^{k} n_i$, $C = 1 + \dfrac{1}{3(k-1)} \cdot \left(\sum_{i=1}^{k} \dfrac{1}{n_i - 1} - \dfrac{1}{n-k}\right)$

C. Rozklad rozptylu a stupňů volnosti

Značení: **n** - počet pozorování celkem
n_i - počet pozorování ve skupině
a - počet skupin
y_{ij} - hodnota jednoho pozorování (v i-té skupině j-tý jedinec)

Celkový	=	Skupin	+	Reziduální
$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$	=	$\sum_{i=1}^{a} n_i(\bar{y}_i-\bar{y})^2$	+	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$
S_T	=	S_A	+	S_e
n - 1	=	a - 1	+	n - a
υ_T	=	υ_A	+	υ_e

Průměrná čvercová odchylka MS (Mean Square) = průměrný čtverec (= dílčí rozptyl)

D. Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců	Stupně volnosti	Průměrný čtverec	Testové kritérium
Skupiny Faktor A	S_A	υ_A	$MS_A = \frac{S_A}{\upsilon_A}$	$F = \frac{MS_A}{MS_e}$
Jedinci Reziduum e	S_e	υ_e	$MS_e = \frac{S_e}{\upsilon_e}$	x
Celkem	S_T	υ_T	x	

Vyhledáme tabulkovou hodnotu **F** pro **α = 0,05** nebo **α = 0,01**
pro stupně volnosti čitatele (tj. skupin) = **a - 1**
a stupně volnosti jmenovatele (tj. rezidua) = **n - a**
$F_{vyp} > F_{tab}$ … H_0 se zamítá

E. Výpočtový tvar

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - K$$

$$S_A = \frac{1}{n_i}\sum_{i=1}^{a} Y_{i.}^2 - K = \frac{1}{n_i}\left(Y_{1\bullet}^{\ 2} + Y_{2\bullet}^2 + \ldots + Y_{a\bullet}^2\right) - K$$

$$S_e = S_T - S_A$$

kde $K = \frac{1}{n} Y_{\bullet\bullet}^2$ (korekce)

Značení: významný rozdíl + (α = 0,05)
vysoce významný rozdíl ++ (α = 0,01)

F. Metody následného testování

1. <u>Metoda minimální průkazné diference</u>

Střední chyba diference $s_{\bar{d}} = \sqrt{\frac{2MS_e}{n_i}}$

$$(d) = t_{1-\alpha} \cdot s_{\bar{d}}$$

t_{tab} pro 1 - α a stupeň volnosti rezidua
Výpočet minimálního rozdílu, který můžeme označit za průkazný

2. Tukeyův test

$D = Q \cdot s_{y_{i\bullet.}} \quad (\equiv s_{\bar{x}})$

kde $s_{yi..}$ - střední chyba $\sqrt{\frac{MS_e}{n_i}}$

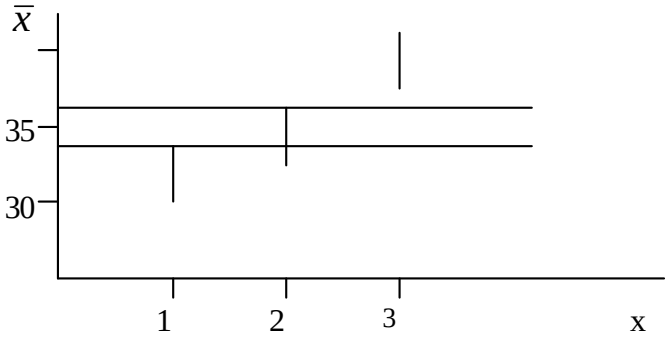
Q - kritické hodnoty q studentizovaného rozpětí podle počtu úrovní faktoru (a) a stupňů volnosti rezidua (n - a)

Výživa	a_1	a_2	a_3	a_4
a_5	+	+		
a_4	++			
a_3		+		
a_2	++			

Vyhodnotí se v tabulce rozdílů průměrů

3. Grafická metoda

Pomocí konfidenčních intervalů kolem průměru (průkazný rozdíl mezi těmi, které se nepřekrývají)



Rozdíl mezi 2 a 3 malý,
mezi 1 a 3 průkazný

4. Scheffeho metoda kontrastů

nejpřesnější metoda na odhalení průkazného rozdílu

kontrast: ψ(psí) = $k_1\mu_1 + k_2\mu_2 + \ldots + k_p\mu_p$, kde $k_1, k_2, \ldots, k_p$ jsou konstanty

a platí $\sum_{i=1}^{p} k_i = 0$

Při porovnání 2 středních hodnot volíme $k_1 = 1$, $k_2 = -1$

Bodový odhad kontrastu $\hat{\psi}$ je: $\psi = k_1 y_{1.} + k_2 y_{2.} + \ldots + k_p y_p$

Směrodatná chyba kontrastu $\hat{\psi}$ je: $s_{\hat{\psi}} = \sqrt{\frac{MS_e}{n_i} \sum_{i=1}^{p} k_i^2}$

Testová charakteristika:

$|t| = \frac{\hat{\psi}}{s_{\hat{\psi}}}$ > S → kontrast je průkazný

$S = \sqrt{\upsilon_A \cdot F_{\alpha(\upsilon_A,\upsilon_e)}}$, kde υ_A - … stupně volnosti skupin

υ_e - … stupně volnosti rezidua

$F\alpha_{(\upsilon A,\ \upsilon e)}$ - tabulková hodnota F-rozdělení

Dvoufaktorová analýza rozptylu

Každá pokusná jednotka je podrobena dvěma způsobům třídění současně - sloupcové a řádkové třídění.
V průsečíku řádku a sloupce:

- je vždy jedno pozorování
- je různý počet pozorování
- je stejný počet pozorování

A.Tabulka uspořádání vstupních dat

Faktor $A_{(i)}$	Faktor $B_{(j)}$	b_1	b_2	…	b_j	Celkem
a_1		y_{111} y_{112} … y_{11ni}	y_{121} y_{122} … y_{12ni}		y_{1j1} y_{1j2} … y_{1jni}	$Y_{1..}$
…		…	…	…	…	
a_i		y_{i11} y_{i12} … y_{i1ni}	y_{i21} y_{i22} … y_{i2ni}		y_{ij1} y_{ij2} … y_{ijni}	
Součet průměr		$Y_{.1.}$ $y_{.1.}$	$Y_{.2.}$ $y_{.2.}$		$Y_{.j.}$ $y_{.j.}$	$Y_{...}$ $y_{...}$

y_{ijk} i = 1, …, a
j = 1, …, b
k = 1, …, n_i

n = a.b.n_i (počet pozorování celkem)
n_i (počet pozorování ve skupině)

Dílčí proměnlivost je dána úrovněmi faktoru A, faktoru B, (kombinacemi obou faktorů = interakcemi) a reziduálními vlivy.
Počítáme ANOVU bez interakcí.
B. Rozklad součtu čtverců a stupňů volnosti

Celkem	=	Faktor A	+	Faktor B	+	Reziduum
S_T	=	S_A	+	S_B	+	S_e
n - 1	=	a - 1	+	b - 1	+	n-1-(a-1)-(b-1)
υ_T	=	υ_A	+	υ_B	+	υ_e

C. Tabulka analýzy rozptylu dvoufaktorové

Zdroje variability	Součty čtverců	Stupně volnosti	Průměrný čtverec	Testové kritérium
FaktorA	S_A	υ_A	$MS_A = \frac{S_A}{\upsilon_A}$	$F = \frac{MS_A}{MS_e}$
FaktorB	S_B	υ_B	$MS_B = \frac{S_B}{\upsilon_B}$	$F = \frac{MS_B}{MS_e}$
Reziduum	S_e	υ_e	$MS_e = \frac{S_e}{\upsilon_e}$	x
Celkem	S_T	υ_T	x	

D. Výpočtový tvar

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_i} y_{ijk}^2 - K$$

$$S_A = \frac{1}{b \cdot n_i}\sum_{i=1}^{a} Y_{i\bullet\bullet}^2 - K$$

$$S_B = \frac{1}{a \cdot n_i}\sum_{j=1}^{b} Y_{\bullet j\bullet}^2 - K$$

$$S_e = S_T - S_A - S_B$$

$$K = \frac{1}{a \cdot b \cdot n_i} \cdot Y_{\bullet\bullet\bullet}^2$$ korekce

Vícefaktorový pokus zachycuje i interakci faktorů, tzn. jejich vzájemné spolupůsobení.

Např. 2 faktory + 2 úrovně (a) a 3 úrovně (b) = 6 kombinací
a_1b_1 a_2b_1
a_1b_2 a_2b_2
a_1b_3 a_2b_3

Úkolem analýzy rozptylu:
Rozčlenit celkovou variabilitu na dílčí složky (podle vlivu jednotlivých faktorů) a na složku reziduální (nelze vysvětlit).

Postup při rozkladu rozptylu

1. vypočítáme celkový průměr, tj průměr všech výsledků pokusu a určíme odchylky jednotlivých hodnot pozorování od tohoto průměru, které umocníme na druhou a sečteme

= celkový součet čtverců - S_T $\quad S_T = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}\right)^2$

2. vypočítáme průměry jednotlivých skupin podle faktorů. Určíme odchylky těchto průměrů od celkového a jejich čtverce, pro každý faktor dostaneme tzv. kvadratickou složku - S_A, S_B, …

$$S_A = \sum_{i=1}^{k} n_i \left(y_i' - \bar{y} \right)^2$$

3. Odečtením všech kvadratických složek od celkového součtu čtverců zůstane složka reziduální - S_e (nevysvětlená, náhodná)

Rozklad součtu čtverců

$$\sum_{i=1}^{k}\sum_{j=1}^{n} \left(y_{ij} - \bar{y} \right)^2 = \sum_{i=1}^{k} n_i \left(y_i' - \bar{y} \right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n} \left(y_{ij} - y_i' \right)^2$$

$S_{T\ celkový}$ = $S_{A\ skupin}$ + $S_{e\ reziduální}$

n - celkový počet prvků
n_i - počet pozorování ve skupině
k - počet skupin
y_{ij} - hodnota naměřená v i-té skupině u j-tého jedince

- čitatelová složka rozptylu
- jmenovatel → stupně volnosti (υ)

$\upsilon = (n-1)$ → odpovídá celkovému počtu pozorování
$\upsilon_A, \upsilon_B, \ldots$ → $(n_i - 1)$ → počty stupňů volnosti jednotlivých skupin
υ_e → stupně volnosti rezidua

Rozklad stupňů volnosti

n - 1	=	k - 1	+ n - k
St. $v._{celkem}$	=	St. $v._{skupin}$	+ St. $v._{rezidua}$

- Lze vypočítat průměrné čtvercové odchylky - MS
- Testovým kritériem je hodnota Fisher-Snedecorova rozdělení $F = \frac{MS_A}{MS_e}$

Tabulka analýzy rozptylu:

Zdroj variability	Součet čtverců S	Stupně volnosti υ	Průměrný čtverec MS	Testové kritérium F
Faktor A	$S_A = \sum_{i=1}^{k} n_i \left(y_i' - \bar{y} \right)^2$	k - 1	$MS_A = \frac{S_A}{k-1}$	$F = \frac{MS_A}{MS_e}$
Reziduum (e)	$S_e = \sum_{i=1}^{k}\sum_{j=1}^{n} \left(y_{ij} - y_i' \right)^2$	n - k	$MS_e = \frac{S_e}{n-k}$	x
Celkem	$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n} \left(y_{ij} - \bar{y} \right)^2$	n - 1	x	

H_0 se zamítá ← $F_{vyp} > F_{tab}$

Praktické poznámky:

- součet čtverců S nemůže být záporný
- korekční člen k slouží ke zjednodušení výpočtu (mocnina celkového součtu dělená počtem všech měření)
- Tečková symbolika - zajišťuje přesnost a výstižnost

Máme srovnat výkonnost 4 odrůd kukuřice. Abychom mohli použít analýzu rozptylu, musíme ověřit homogenitu rozptylu.

Výsledky pokusu:

	a_1	a_2	a_3	a_4
	45	35	33	41
	46	33	34	41
	49		35	43
	44		34	41
			34	44
				42
				44
				41
				41
$\overline{x}$	46	34	34	42
n_i	4	2	5	9
Σx_i^2	8478	2314	5782	15890
$_{(n)}s^2_{xi}$	3,5	1,0	0,4	1,55
$_{(n-1)}s^2_{xi}$	4,67	2	0,5	1,75

$$_{(n-1)}s^2_{x_i} = {}_{(n)}s^2_{x_i} \cdot \frac{n}{n-1}$$

Bartlettův test

$$C = 1 + \frac{1}{3(4-1)} \cdot \left(\frac{1}{4-1} + \frac{1}{2-1} + \frac{1}{5-1} + \frac{1}{9-1} - \frac{1}{20-4} \right) = 1{,}8287$$

$$s^2 = \frac{4{,}67(4-1) + 2(2-1) + 0{,}5(5-1) + 1{,}75(9-1)}{20-4} = 1{,}9987$$

$$\chi^2_{(k-1)} = \frac{\ln 10}{1{,}8287}\left[(20-4)\log 1{,}9987 - \left[(4-1)\log 4{,}6 + (2-1)\log 2 + (5-1)\log 0{,}5 + (9-1)\log 1{,}75\right]\right] = 2{,}22$$

$\chi^2_{tab(3)} = 7{,}81$ ($\alpha = 0{,}05$)

H_0 se nezamítá → rozptyly jsou homogenní

Jsou sledovány 2 odrůdy ječmene při 3 úrovních výživy. Srovnejte počet zrn na rostlině.

Faktor A	Faktor B			Celkem skup. A
	b_1	b_2	b_3	
a_1	99 100 107 103	103 102 101 105	104 98 108 105	1235
a_2	113 107 107 106	111 105 112 112	104 104 106 106	1293
Celkem skup. B	842	851	835	2528
	105,25	106,375	104,375	105,33

a = 2, b = 3, n_i = 4
n = (a.b.n_i) = 24

$$K = \frac{1}{24} \cdot 2528^2 = 266282{,}66$$

$$S_T = 266648 - 266282{,}66 = \underline{\underline{365{,}34}}$$

$$S_A = \frac{1}{3 \cdot 4}\left(1235^2 + 1293^2\right) - 266282{,}66\overline{6} = \frac{1}{12}\left(1525225 + 1671849\right) - 266282{,}66\overline{6} =$$

$$= \frac{1}{12}\ 3197074 - 266282{,}66\overline{6} = 266422{,}8\overline{3} - 266282{,}66\overline{6} = \ \underline{\underline{140{,}17}}$$

$$S_B = \frac{1}{2 \cdot 4}\left(842^2 + 851^2 + 835^2\right) - 266282{,}66\overline{6} =$$

$$= \frac{1}{8}\left(708964 + 724201 + 697225\right) - 266282{,}66\overline{6} =$$

$$= \frac{1}{8}\ 2130390 - 266282{,}66\overline{6} = 266298{,}75 - 266282{,}66\overline{6} = \ \underline{\underline{16{,}09}}$$

$$S_e = 365{,}34 - 140{,}17 - 16{,}08 = \underline{\underline{209{,}08}}$$

1. Jednofaktorová analýza

Zdroj variability	Součet čtverců S	Stupně volnosti υ	Průměrný čtverec MS	Testové kritérium F
Skupiny (B) výživa	16,09	2	$\frac{16,09}{2} = 8,045$	$\frac{8,045}{16,63} = 0,48$
Jedinci (e) reziduum	349,25	21	$\frac{349,25}{21} = 16,63$	x
Celkem	365,34	23	x	

$F_{tab(2,21)} = 3,49/5,85$

Není průkazný rozdíl v úrovni výživy ječmene.

2. Dvoufaktorová analýza

Zdroj variability	Součet čtverců S	Stupně volnosti υ	Průměrný čtverec MS	Testové kritérium F
Skupiny (A) odrůda	140,17	1	$\frac{140,17}{1} = 140,17$	$\frac{140,17}{10,454} = 13,41**$
Skupiny (B) výživa	16,09	2	$\frac{16,09}{2} = 8,045$	$\frac{8,045}{10,454} = 0,77$
Jedinci (e) reziduum	209,08	20	$\frac{209,08}{20} = 10,454$	x
Celkem	365,34	23	x	

$F_{A\text{-}tab(1,20)} = 4,35/8,10$
$F_{B\text{-}tab(2,20)} = 3,49/5,85$

Vysoce průkazný rozdíl mezi odrůdami ječmene.

<u>Metody následného testování:</u>

3. Scheffeho metoda kontrastů

tabulka kontrastů $s_{\psi} = \sqrt{\frac{MS_e}{n_i}\sum_{i=1}^{p} k_i^2} = \sqrt{\frac{16063}{8} \cdot 2}$

Úroveň výživy	$y_{i.}$	Kontrast $\hat{\psi}$ (rozdíl průměrů)	$\lvert t \rvert = \frac{\hat{\psi}}{s_{\hat{\psi}}}$	Významnost kontrastu
1	105,25	-	-	-
2	106,375	1,125	0,55	t < S
3	104,375	0,875	0,43	t < S
2	106,375	-	-	-
3	104,375	2,000	0,98	t < S

Nejsou průkazné rozdíly (což nám už řekla tabulka analýzy rozptylu)

Posuďte, zda se 5 plemen hodnocených v pokuse odlišuje v mléčné užitkovosti (při zachování stejných podmínek chovu). Z každého plemene bylo vybráno 10 krav.

A. Tabulka vstupních dat

Jedinci	Skupiny (plemeno)					$i = 1, 2, \ldots, a$
	1	2	3	4	5	$j = 1, 2, \ldots, n_i$
1	10	8	18	10	6	
2	12	10	13	12	12	
3	8	6	10	10	8	
4	13	7	12	8	10	
5	7	8	14	10	10	252
6	10	7	12	14	10	
7	11	7	12	10	7	
8	8	8	10	12	9	
9	10	6	11	11	8	
10	11	9	14	9	10	246
$Y_{i.}$	100	76	126	106	90	$Y.. = 498$
$y_{i.}$	10,0	7,6	12,6	10,6	9,0	$y.. = 9,96$
$\sum_{i=1}^{a} y_i^2$	1032	592	1638	1150	838	$\sum_{i=1}^{a} y_{ij}^2 = 5250$

B. Cochranův test homogenity rozptylů:

$$s_{y_1}^2 = \frac{1032}{10} - 10^2 = 3{,}20 \qquad 3{,}20 \cdot \left(\frac{10}{9}\right) = 3{,}56$$

$$s_{y_2}^2 = \frac{592}{10} - 7{,}6^2 = 1{,}44 \qquad 1{,}44 \cdot \left(\frac{10}{9}\right) = 1{,}6$$

$$s_{y_3}^2 = \frac{1638}{10} - 12{,}6^2 = 5{,}04 \qquad 5{,}04 \cdot \left(\frac{10}{9}\right) = 5{,}6$$

$$s_{y_4}^2 = \frac{1150}{10} - 10{,}6^2 = 2{,}64 \qquad 2{,}64 \cdot \left(\frac{10}{9}\right) = 2{,}93$$

$$s_{y_5}^2 = \frac{838}{10} - 9^2 = 2{,}80 \qquad 2{,}80 \cdot \left(\frac{10}{9}\right) = 3{,}11$$

$$Q_{(k,n-1)} = \frac{s_{\max}^2}{\sum_{i=1}^{a} s_i^2} \qquad q_{(5,9)} = \frac{5{,}6}{16{,}8} = 0{,}3333$$

Tabulka Cochranova statistika $q_{tab(5,9)} = 0{,}4241$

H_0 nezamítáme, rozptyly jsou homogenní.

E. Výpočtový tvar

$$K = \frac{1}{n} \cdot Y_{..}^2 = \frac{1}{50} \cdot 498^2 = 4960{,}08$$

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - K = 520 - 4960{,}08 = \underline{\underline{289{,}92}}$$

$$S_A = \frac{1}{n_i} \sum_{i=1}^{a} Y_{i.}^2 - K = \frac{1}{10}\left(100^2 + 76^2 + 126^2 + 106^2 + 90^2\right) - 4960{,}08 = \underline{\underline{138{,}72}}$$

$$S_e = S_T - S_A = 289{,}92 - 138{,}72 = \underline{\underline{151{,}2}}$$

D.Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců S	Stupně volnosti υ	Průměrný čtverec MS	Testové kritérium F
Skupiny (plemeno)	138,72	4	34,68	10,32**
Jedinci (e)	151,2	45	3,36	x
Celkem	289,92	49	x	

$F_{tab(4,45)}$ = 2,6 / 3,8

**Vysoce průkazný rozdíl.

F. Metody následného testování

a) minimální průkazná diference (Least Square Difference - LSD)

$$s_{\bar{d}} = \sqrt{\frac{2\,MS_e}{n_i}} = \sqrt{\frac{2 \cdot 3{,}36}{10}} = 0{,}8198 \qquad (d) = t_{1-\frac{\alpha}{2}} \cdot s_{\bar{d}}$$

$t_{0,975(45)}$ = 2,01 (d) = 2,01 . 0,8198 = **1,65** porovnáme s tabulkou
$t_{0,995(45)}$ = 2,68 (d) = 2,68 . 0,8198 = **2,20** porovnáme s tabulkou

Tabulka rozdílů průměrů

	a_1	a_2	a_3	a_4
a_2	1,00	1,4	3,6++	1,6
a_3	0,6	3,0++	2,0+	
a_4	2,6++	5++		
a_5	2,4++			

b) Tukeyův test

$$s_{y_{i.}\bullet} = \sqrt{\frac{MS_e}{n_i}} = \sqrt{\frac{3{,}36}{10}} = 0{,}5797 \qquad D = Q \cdot s_{y_{i.}\bullet}$$

$Q_{5,45(0,05)}$ = 4,02 $D_{(0,05)}$ =4,02 . 0,5797 = **2,33** porovnáme s tabulkou
$Q_{5,45(0,01)}$ = 4,90 $D_{(0,05)}$ =4,90 . 0,5797 = **2,84** porovnáme s tabulkou

Q - hodnoty studentizovaného rozpětí
podle počtu úrovní faktoru

podle stupňů volnosti rezidua
Tabulka rozdílů průměrů

a_i	a_1	a_2	a_3	a_4
a_2	1,00	1,4	3,6++	1,6
a_3	0,6	3,0++	2,0+	
a_4	2,6+	5++		
a_5	2,4+			

Test je přísnější!

c) Scheffeho metoda kontrastů

$$s_{\hat{\psi}} = \sqrt{\frac{MS_e}{n_i}\sum_{i=1}^{p} k_i^2} = \sqrt{\frac{3{,}36}{10}2} = 0{,}8198$$

$$S = \sqrt{\upsilon_A \cdot F_{\alpha(\upsilon_A,\upsilon_e)}}$$, kde υ_A - … stupně volnosti skupin

$F_{0,05(4,45)} = 2{,}6$ $\quad S = \sqrt{4 \cdot 2{,}6} = 3{,}22$ porovnáme s tabulkou

$F_{0,01(4,45)} = 3{,}8$ $\quad S = \sqrt{4 \cdot 3{,}8} = 3{,}90$ porovnáme s tabulkou

$\left|t\right| = \frac{\hat{\psi}}{s_{\hat{\psi}}} > S$ $\quad \rightarrow \left|t\right| > S_{0,05} \rightarrow$ kontrast je významný

$\rightarrow \left|t\right| > S_{0,01} \rightarrow$ kontrast je vysoce významný

$k_1 = 1$, $k_2 = -1$, $\hat{\psi}$ = rozdíl průměrů!

Tabulka kontrastů

$\left\|t\right\|$	a_1	a_2	a_3	a_4
a_5	1,22	1,71	4,39++	1,95
a_4	0,73	3,66+	2,44+	
a_3	3,17	6,1++		
a_2	2,93			

Je nejpřísnější ze všech metod!

SHRNUTÍ

Rozdíl skupin	Hodnocení průkaznosti rozdílu		
	LSD	Tukey	Scheffe
1-2	++	+	
1-3	++	+	
1-4			
1-5			
2-3	++	++	++
2-4	++	++	+
2-5			
3-4	+		
3-5	++	++	++
4-5			

d 9.4 (FYTO)

Ověřte, zda mezi 5 odrůdami révy vinné existuje průkazný rozdíl v cukernatosti.
(10 vzorků od každé odrůdy)

A. Tabulka vstupních dat

Číslo vzorku	Faktor A (odrůdy)					i = 1, 2, …, a
Jedinci	1	2	3	4	5	j = 1, 2, …, n_i
1	16	22	21	23	19	
2	15	20	19	22	16	
3	17	23	18	24	17	
4	18	21	21	23	16	
5	15	23	18	22	18	487
6	16	20	20	21	17	
7	17	21	21	22	16	
8	15	20	20	23	16	
9	17	19	19	24	17	
10	19	21	18	21	18	478
Součet $(\Sigma x)Y_{i.}$	165	210	195	225	170	**Y..** = 965
Průměr $y_{i.}$	16,5	21	19,5	22,5	17,0	19,3
$\sum_{i=1}^{a} x_i^2$	2739	4426	3817	5073	2900	$\sum_{i=1}^{a} x_{ij}^2 = 18955$

B. Cochranův test homogenity rozptylu

$$s_{x_1}^2 = \frac{2739}{10} - 16{,}5^2 = 1{,}65 \qquad 1{,}65 \cdot \left(\frac{10}{9}\right) = 1{,}83$$

$$s_{x_2}^2 = \frac{4426}{10} - 21^2 = 1{,}6 \qquad 1{,}6 \cdot \left(\frac{10}{9}\right) = 1{,}78$$

$$s_{x_3}^2 = \frac{3817}{10} - 19{,}5^2 = 1{,}45 \qquad 1{,}45 \cdot \left(\frac{10}{9}\right) = 1{,}61$$

$$s_{x_4}^2 = \frac{5073}{10} - 22{,}5^2 = 1{,}05 \qquad 1{,}05 \cdot \left(\frac{10}{9}\right) = 1{,}17$$

$$s_{x_5}^2 = \frac{2900}{10} - 17{,}0^2 = 1{,}0 \qquad 1{,}0 \cdot \left(\frac{10}{9}\right) = 1{,}11$$

$$Q_{(k,n-1)} = \frac{s_{\max}^2}{\sum_{i=1}^{a} s_i^2} \qquad q_{(5,9)} = \frac{1{,}83}{7{,}5} = 0{,}244$$

Tabulka Cochranova statistika $q_{tab(5,9)} = 0{,}4241$

H_0 nezamítáme, rozptyly jsou homogenní.

E. Výpočtový tvar

$$K = \frac{1}{n} \cdot Y_{\bullet\bullet}^2 = \frac{1}{50} \cdot 965^2 = 18624{,}5 \qquad (n = a \,.\, n_i)$$

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - K = 18955 - 18624{,}5 = \underline{\underline{330{,}5}}$$

$$S_A = \frac{1}{n_i} \sum_{i=1}^{a} Y_{i.}^2 - K = \frac{1}{10}\left(165^2 + 210^2 + 195^2 + 225^2 + 170^2\right) - 18624{,}5 = \underline{\underline{263}}$$

$$S_e = S_T - S_A = 330{,}5 - 263 = \underline{\underline{67{,}5}}$$

D.Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců S	Stupně volnosti υ	Průměrný čtverec MS	Testové kritérium F
Odrůdy (A)	263	4	65,75	43,83**
Jedinci (e)	67,5	45	1,5	x
Celkem	330,5	49	x	

$F_{tab(4,45)}$ = 2,6 / 3,8

**Vysoce průkazný rozdíl.

F. Metody následného testování

a) minimální průkazná diference (Least Square Difference - LSD)

$$s_{\bar{d}} = \sqrt{\frac{2\,MS_e}{n_i}} = \sqrt{\frac{2 \cdot 1{,}5}{10}} = 0{,}55 \qquad (d) = t_{1-\frac{\alpha}{2}} \cdot s_{\bar{d}}$$

$t_{0,975(45)}$ = 2,016 (d) = 2,016 . 0,55 = **1,1088** porovnáme s tabulkou
$t_{0,995(45)}$ = 2,693 (d) = 2,693 . 0,55 = **2,481** porovnáme s tabulkou

Tabulka rozdílů průměrů

d	a_1	a_2	a_3	a_4
a_5	0,5	4++	2,5++	5,5++
a_4	6++	1,5++	3+	
a_3	3++	1,5++		
a_2	4,5++			

b) Tukeyův test

$$s_{y_i.} = \sqrt{\frac{MS_e}{n_i}} = \sqrt{\frac{1{,}5}{10}} = 0{,}39 \qquad D = Q \,.\, s_{y_i.}$$

$Q_{5,45(0,05)}$ = 4,03 $D_{(0,05)}$ =4,03 . 0,39 = **1,573** porovnáme s tabulkou

$Q_{5,45(0,01)}$ = 4,90 $D_{(0,05)}$ =4,90 . 0,39 = **1,911** porovnáme s tabulkou

Q - hodnoty studentizovaného rozpětí (Tab. 8,9)
podle počtu úrovní faktoru

podle stupňů volnosti rezidua
Tabulka rozdílů průměrů

a_i	a_1	a_2	a_3	a_4
a_5		++	++	++
a_4	++		++	
a_3	++			
a_2	++			

Test je přísnější!

c) Scheffeho metoda kontrastů

$$s_{\hat{\psi}} = \sqrt{\frac{MS_e}{n_i}\sum_{i=1}^{p} k_i^2} = \sqrt{\frac{1{,}5}{10}\cdot 2} = 0{,}55 \qquad S = \sqrt{\upsilon_A \cdot F_{\alpha(\upsilon_A,\upsilon_e)}}$$

kde υ_A - … stupně volnosti skupin

$F_{0,05(4,45)} = 2{,}6$ $\quad S = \sqrt{4\cdot 2{,}6} = 3{,}22$ porovnáme s tabulkou

$F_{0,01(4,45)} = 3{,}8$ $\quad S = \sqrt{4\cdot 3{,}8} = 3{,}90$ porovnáme s tabulkou

$|t| = \frac{\hat{\psi}}{s_{\hat{\psi}}} > S$ $\quad \rightarrow |t| > S_{0,05} \rightarrow$ kontrast je významný

$\rightarrow |t| > S_{0,01} \rightarrow$ kontrast je vysoce významný

$k_1 = 1$, $k_2 = -1$, $\hat{\psi}$ = rozdíl průměrů!

Tabulka kontrastů

$\lvert t\rvert$	a_1	a_2	a_3	a_4
a_5	0,91	7,27++	4,55++	10++
a_4	10,91++	2,73	5,45++	
a_3	5,45++	2,73		
a_2	8,18++			

Je nejpřísnější ze všech metod!

SHRNUTÍ

Rozdíl skupin	Hodnocení průkaznosti rozdílu		
	LSD	Tukey	Scheffe
1-2	++	++	++
1-3	++	++	++
1-4	++	++	++
1-5			
2-3	++		
2-4	++		
2-5	++	++	++
3-4	++	++	++
3-5	++	++	++
4-5	++	++	++

d) konfidenční intervaly kolem průměrů skupin

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot s_{\bar{x}} \qquad \left(s_{\bar{x}} = \sqrt{\frac{MS_e}{n_i}} \right) \qquad s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \sqrt{\frac{s_x^2}{n}}$$

$$s_{\bar{x}_1} = \sqrt{0{,}183} = 0{,}428$$

$$s_{\bar{x}_2} = \sqrt{0{,}178} = 0{,}422$$

$$s_{\bar{x}_3} = \sqrt{0{,}161} = 0{,}401$$

$$s_{\bar{x}_4} = \sqrt{0{,}117} = 0{,}342$$

$$s_{\bar{x}_5} = \sqrt{0{,}111} = 0{,}333$$

95%	99%
$15{,}54 < \mu_1 < 17{,}47$	$15{,}11 < \mu_1 < 17{,}80$
$20{,}65 < \mu_2 < 21{,}95$	$19{,}63 < \mu_2 < 22{,}37$
$18{,}59 < \mu_3 < 20{,}41$	$18{,}20 < \mu_3 < 20{,}80$
$21{,}73 < \mu_4 < 23{,}27$	$21{,}39 < \mu_4 < 23{,}61$
$16{,}25 < \mu_5 < 17{,}75$	$15{,}92 < \mu_5 < 18{,}08$

$t_{0{,}975(9)} = 2{,}262$
$t_{0{,}995(9)} = 3{,}250$

Grafické znázornění konfidenčních intervalů

Výsledky z programu UNISTAT ver. 5. 6

F-test

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
H	30	39,5333	2,9447	0,5376
CS	20	30,2500	1,9967	0,4465
Celkem	50	35,8200	2,6109	0,3692

F(29,19) =	2,1750
Pravostranná pravděpodobnost =	0,0405
95% Konfidenční interval =	0,9055 <> 4,8530

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
H	30	39,5333	2,9447	0,5376
J	12	18,1667	1,6967	0,4898
Celkem	42	33,4286	2,6605	0,4105

F(29,11) =	3,0121
Pravostranná pravděpodobnost =	0,0287
95% Konfidenční interval =	0,9638 <> 7,4556

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
CS	20	30,2500	1,9967	0,4465
J	12	18,1667	1,6967	0,4898
Celkem	32	25,7188	1,8922	0,3345

F(19,11) =	1,3849
Pravostranná pravděpodobnost =	0,2947
95% Konfidenční interval =	0,4271 <> 3,8286

t-test (spol.rozptyl)

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
CS	20	30,2500	1,9967	0,4465
J	12	18,1667	1,6967	0,4898
Celkem	32	25,7188	1,8922	0,3345

t-statistika =	17,4881
Stupně volnosti =	30,0000
dvoustranná pravděpodobnost =	0,0000
Rozdíl mezi průměry =	12,0833
95% Konfidenční interval =	10,6722 <> 13,4944

t-test (různé rozptyly)

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
H	30	39,5333	2,9447	0,5376
CS	20	30,2500	1,9967	0,4465
Celkem	50	35,8200	2,6109	0,3692

t-statistika =	13,2838
Stupně volnosti =	47,9695
dvoustranná pravděpodobnost =	0,0000
Rozdíl mezi průměry =	9,2833
95% Konfidenční interval =	7,7679 <> 10,7988

Datová proměnná: dojivost
Dílčí výběr vybrán: plemeno

plemeno	Příp.	Průměr	Směrodatná odchylka	Směrodatná chyba
H	30	39,5333	2,9447	0,5376
J	12	18,1667	1,6967	0,4898
Celkem	42	33,4286	2,6605	0,4105

t-statistika =	29,3787
Stupně volnosti =	34,4860
dvoustranná pravděpodobnost =	0,0000
Rozdíl mezi průměry =	21,3667
95% Konfidenční interval =	19,5208 <> 23,2125

Testy homogenity rozptylů

Pro dojivost

	Testovací statistika	Významn.
tříděno podle plemeno		
Bartlettův test chí-kvadrát	5,8175	0,0545
Bartlett-Boxův F test	2,9261	0,0537
Cochranovo C (max var/sum var)	0,5581	0,0192
Hartleyovo F (max var/min var)	3,0121	
Levenův F test	2,5281	0,0884

Analýza rozptylu

Přístup: Klasický experiment
Závisle proměnná: dojivost

~~Zdroj variability~~	Součet čtverců	St. vol.	Průměrný čtverec	Stat F	Významn.
Hlavní efekty	4050,036	2	2025,018	332,911	0,0000
plemeno	4050,036	2	2025,018	332,911	0,0000
Vysvětleno	4050,036	2	2025,018	332,911	0,0000
Chyba	358,883	59	6,083		
Celkem	4408,919	61	72,277		

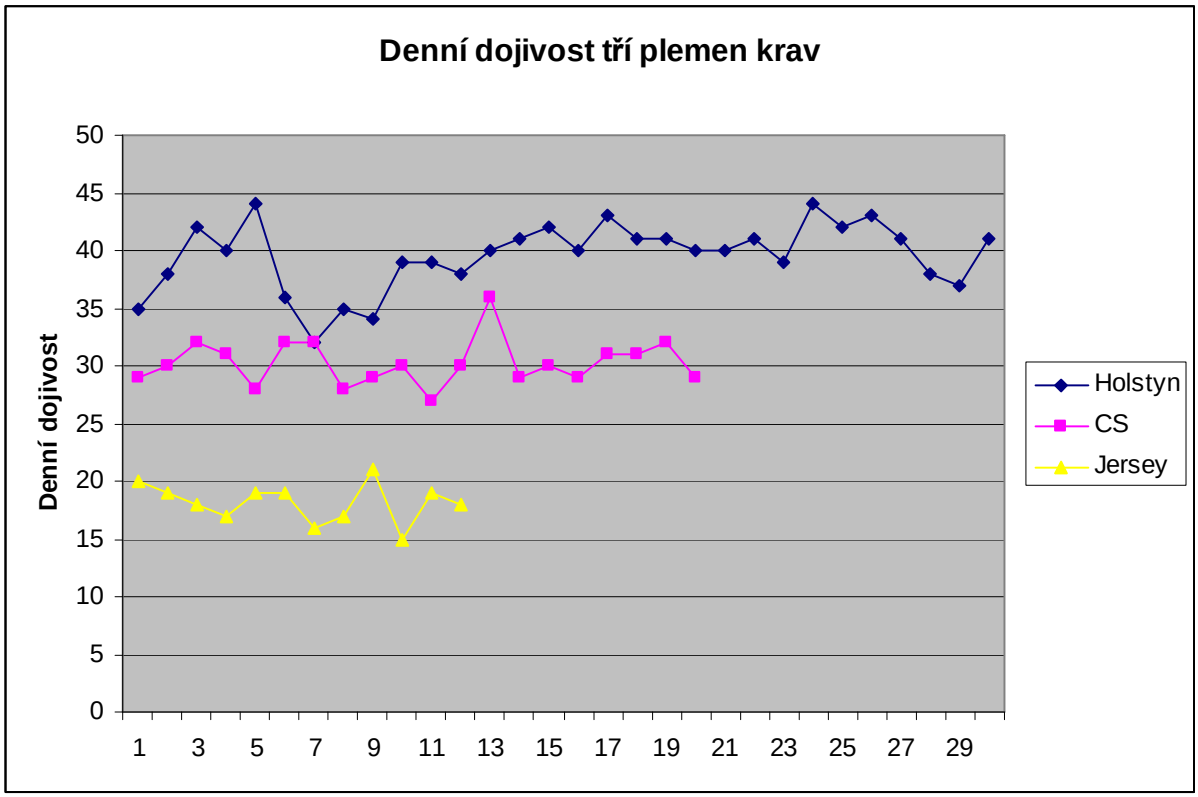
Mnohonásobná porovnávání

Tukey-HSD

Pro dojivost, tříděno podle plemeno
Střední kvadratická chyba: 6,08276836158195, Stupně volnosti: 59
** označuje významně odlišné páry.
Párový test je významný, pokud q hodnota je větší než tabulková hodnota q.

Skupina	Příp.	Průměr	J	CS	H	
J	12	18,1667		**	**	|
CS	20	30,2500	**		**	|
H	30	39,5333	**	**		|

Srovnání	Rozdíl	Směrodatná chyba	q Stat	Tabulka q	Významn.	Dolní 95%	Horní 95%	Výsledek
H - J	21,3667	0,8424	35,8697	3,4001	0,0000	19,3413	23,3920	**
CS - J	12,0833	0,9006	18,9750	3,4001	0,0000	9,9181	14,2485	**
H - CS	9,2833	0,7120	18,4399	3,4001	0,0000	7,5716	10,9951	**



JEDNODUCHÁ NELINEÁRNÍ KORELAČNÍ ZÁVISLOST

Aditivní typy:

$y' = b_0 + b_1 x$ lineární (přímka)

$y' = b_0 + b_1 x + b_2 x^2$ kvadratický (parabola 2.st.)

$y' = b_0 + b_1 x + b_2 x^2 + b_3 x^3$ kubický (parabola 3.st.)

$y' = b_0 + \frac{b_1}{x}$ lomený 1.st. (hyperbola 1.st.)

$y' = b_0 + \frac{b_1}{x} + \frac{b_2}{x^2}$ lomený 2.st. (hyperbola 2.st.)

$y' = b_0 + b_1 x + b_2 \sqrt{x}$ odmocninný

$y' = b_0 + b_1 \log x$ logaritmický

Multiplikativní typy:

$y' = b_0 \cdot b_1^{\,x}$

$(\log y' = \log b_0 + x \log b_1)$ | exponenciální

$y' = b_0 \cdot x^{b_1}$

mocninný

$(\log y' = \log b_0 + b_1 \log x)$ |

ROZKLAD EMPIRICKÉHO ROZPTYLU

Empirický rozptyl lze rozložit na součet rozptylu teoretického a rozptylu reziduálního:

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

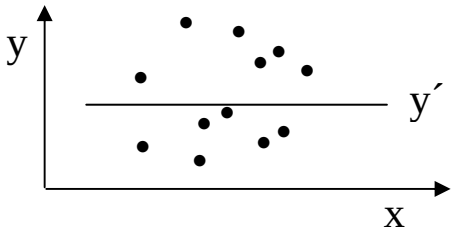
V symbolické formě je rozklad rozptylu vyjádřen jako

$$var\ y = var\ y' + var\ (y-y')$$, resp. $$s_y^2 = s_{y'}^2 + s_{y-y'}^2$$

Při výpočtu indexu determinace s ohledem na podíl složek na empirickém rozptylu mohou nastat tři možnosti:

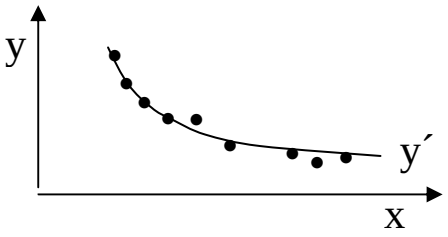
a) *var y´ = 0*, takže *var y = var (y-y´)*

Jde o limitní případ, kdy je y_i' nezávislé na x_i, takže regresní čarou je přímka rovnoběžná s osou *x*. Jde o ***nezávislost***.



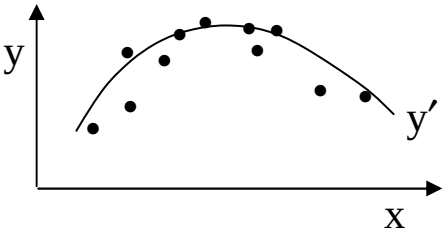
b) *var(y-y´) = 0*, takže *var y = var y´*

Jde o druhý limitní případ, kdy je každé y_i' stejné s y_i. Všechny body leží přímo na regresní křivce a jde tedy o ***pevnou závislost***.



c) *var y´≠0, var (y-y´) ≠0*, takže *var y = var y´+ var (y-y´)*

V daném případě jde o ***volnou závislost***, která je předmětem statistického zkoumání.



VÍCENÁSOBNÁ A DÍLČÍ

KORELAČNÍ ZÁVISLOST

- **lineární regrese**

$$y'_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}$$

$$y'_i = a + b_{yx_1 \cdot x_2 x_3 \ldots x_k} x_{1i} + b_{yx_2 \cdot x_1 x_3 \ldots x_k} x_{2i} + \ldots + b_{yx_k \cdot x_1 x_2 \ldots x_{k-1}} x_{ki}$$

$$y'_i = \bar{y} + b_1 (x_{1i} - \bar{x}_1) + b_2 (x_{2i} - \bar{x}_2) + \ldots + b_k (x_{ki} - \bar{x}_k)$$

- **nelineární regrese**

kvadratická

$$y'_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + c_1 x_{1i}^2 + c_2 x_{2i}^2 + \ldots + c_k x_{ki}^2$$
$$(+d_{1,2} x_{1i} x_{2i} + d_{1,3} x_{1i} x_{3i} + \ldots + d_{k-1,k} x_{(k-1)i} x_{ki})$$

lomená

$$y'_i = a + \frac{b_1}{x_{1i}} + \frac{b_2}{x_{2i}} + \ldots + \frac{b_k}{x_{ki}}$$

exponenciální

$$y'_i = a \cdot b_1^{x_{1i}} \cdot b_2^{x_{2i}} \cdot \ldots \cdot b_k^{x_{ki}}$$

mocninná

$$y'_i = a \cdot x_{1i}^{b_1} \cdot x_{2i}^{b_2} \cdot \ldots \cdot x_{ki}^{b_k}$$

Index vícenásobné korelace

$$I_{y \cdot x_1 x_2 \ldots x_k} = \sqrt{\frac{\text{var } y'}{\text{var } y}}$$

Př.:

NORMÁLNÍ ROVNICE

Vícefaktorový model vyjádřený mocninnou funkcí

(tzv. Cobb-Douglasovou funkcí)

$$y' = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \ldots \cdot x_k^{b_k}$$

se převede logaritmováním na aditivní tvar

$$\log y' = \log a + b_1 \log x_1 + b_2 \log x_2 + \ldots + b_k \log x_k$$

a pak se vyvodí soustava normálních rovnic.

Např.

dvoufaktorová mocninná funkce

$$y' = a \cdot x_1^{b_1} \cdot x_2^{b_2}$$

$$\log y' = \log a + b_1 \log x_1 + b_2 \log x_2$$

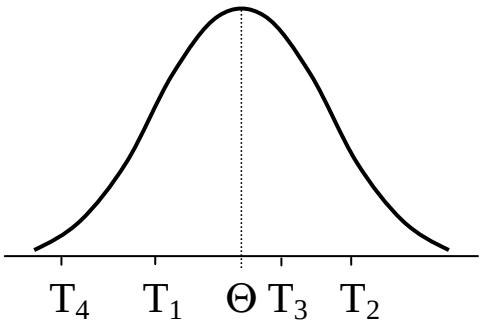
soustava normálních rovnic :

$$\sum \log y = n \log a + b_1 \sum \log x_1 + b_2 \sum \log x_2$$

$$\sum (\log y)(\log x_1) = \log a \sum \log x_1 + b_1 \sum (\log x_1)^2 + b_2 \sum (\log x_1)(\log x_2)$$

$$\sum (\log y)(\log x_2) = \log a \sum \log x_2 + b_1 \sum (\log x_1)(\log x_2) + b_2 \sum (\log x_2)^2$$

SMĚRODATNÁ CHYBA
(dříve střední chyba)



T_i … výběrová charakteristika (s normálním rozdělením) kde: i = 1, 2, …, k

Θ … charakt. zákl. souboru

<u>výběrové chyby:</u> $T_i - \Theta$ (+, -, velké, malé)

Směrodatná chyba průměru

opravný koeficient při výběru bez opakování

$$s_{\bar{x}} = \sqrt{\frac{\sum(\bar{x}-\mu)^2}{k}} \quad \Rightarrow \quad \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

směrodatná odchylka základního souboru

$$\sigma_x \approx s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$\sigma_x \approx \frac{R}{6}$$

přičemž

$$s_{\bar{x}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n(n-1)}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}{n(n-1)}}$$

ROZSAH VÝBĚRU

Jediným kritériem je přesnost odhadu.

Výpočetní postup:

(Platí pro náhodný výběr s opakováním, ale lze jej použít i při praktičtějším výběru bez opakování, neboť je přísnější.)

$$\Delta = u_{1-\frac{\alpha}{2}} \cdot s_{\bar{x}}$$

$$\Delta = u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma_x}{\sqrt{n}}$$

$$\Delta^2 = u^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma_x^2}{n}$$

kde: Δ - přípustná chyba
u - kvantil norm. rozdělení
σ_x^2 - rozptyl zákl. souboru
σ_x - směrodatná odchylka
$s_{\bar{x}}$ - směrodatná chyba průměru
n - rozsah souboru

Rozsah výběru je tím větší, čím je větší stanovená pravděpodobnost výpočtu a variabilita základního souboru a čím je menší přípustná chyba.

$$n = \frac{u^2_{1-\frac{\alpha}{2}} \cdot \sigma_x^2}{\Delta^2}$$

Při praktickém výpočtu se obvykle vychází z předvýběru, takže vzorec má menší úpravu:

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

kde: t - kvantil Student. rozdělení
s_x^2 - rozptyl výběru

Výpočet je značně ovlivněn rozptylem stanoveným z před- výběru, je proto vhodný spíše pro jednorázové použití.

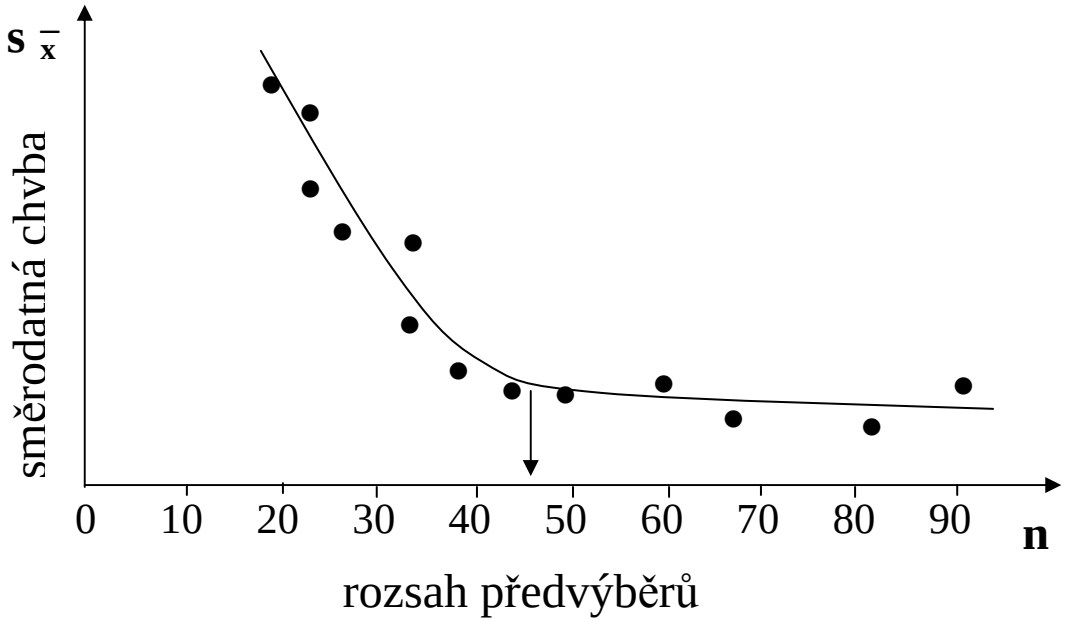
ROZSAH VÝBĚRU

Jediným kritériem je přesnost odhadu.

Grafický postup:

Využívá se tehdy, kdy se výběrové šetření často opakuje (např. každodenní odebírání vzorků) a kdy záleží na tom, aby rozsah výběru byl co nejmenší a přitom reprezentativní.

- ze základního souboru se odebere více předvýběrů o různém rozsahu
- z každého předvýběru se vypočte směrodatná chyba
- sestrojí se bodový graf, na vodorovné ose se vynáší *rozsah předvýběrů* a na svislé ose jejich *směrodatné chyby*
- body se položí křivka
- zlom na křivce představuje vhodný rozsah výběru



NEPARAMETRICKÉ TESTY

U **parametrických testů** je známé rozdělení a parametry (úplné určení) nebo alespoň rozdělení (částečné určení).

U **neparametrických testů** není známé rozdělení, jsou však formulovány různé předpoklady jako např. spojitost distribuční funkce, minimální či maximální rozsah souboru apod.

~~TESTY SHODY ROZDĚLENÍ~~

Mann – Whitneův test

shoda dvou empirických rozdělení

Kolmogorův test pro dva výběry

shoda dvou empirických rozdělení

Kolmogorův test pro dva výběry

shoda empirického rozdělení s rozdělením teoretickým

TESTY PRŮKAZNOSTI ROZDÍLU STŘED. HODNOT

Znaménkový test

dva závislé soubory (nahrazuje párový test)

Wilcoxonův test

dva závislé soubory (nahrazuje párový test) – přísnější

Kruskal – Wallisův test

více nezávislých souborů (nahrazuje jednofakt. analýzu variance)

Friedmanův test

více závislých souborů (nahrazuje dvoufakt. analýzu variance)

TESTY PRŮKAZNOSTI ODCHYLEK

Dixonův test

test extrémních odchylek (za Grubbsův test)

Test náhodnosti uspořádání

kolísání hodnot vlivem náhody nebo vlivem nenáhodných faktorů.