Statistické metody v ochraně kulturního dědictví

Lubomír Prokeš

II.

Náhodný výběr

= reprezentativní vzorek základního souboru.

1. Jednotlivá pozorování v náhodném výběru pocházejí z téhož rozdělení, tj. jsou realizována za stejných podmínek.
2. Hodnoty náhodné veličiny v náhodném výběru musí být vybrány nezávisle, tj. výběr kterékoli hodnoty nesmí ovlivnit výběr hodnoty následující.

Popisná statistika II

výběrové odhady parametrů používaných
k charakteristice náhodných výběrů

- Výběrový průměr $\bar{x}$
- Výběrový medián $\tilde{x}$
- Výběrové variační rozpětí: $R = x_{max} - x_{min}$
- Výběrové kvartily Q_{III} a Q_I
- Výběrový rozptyl (s^2) a výběrová směrodatná odchylka (s) a výběrový variační koeficient

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

- Výběrová šikmost
- Výběrová špičatost

- Výběrový modus

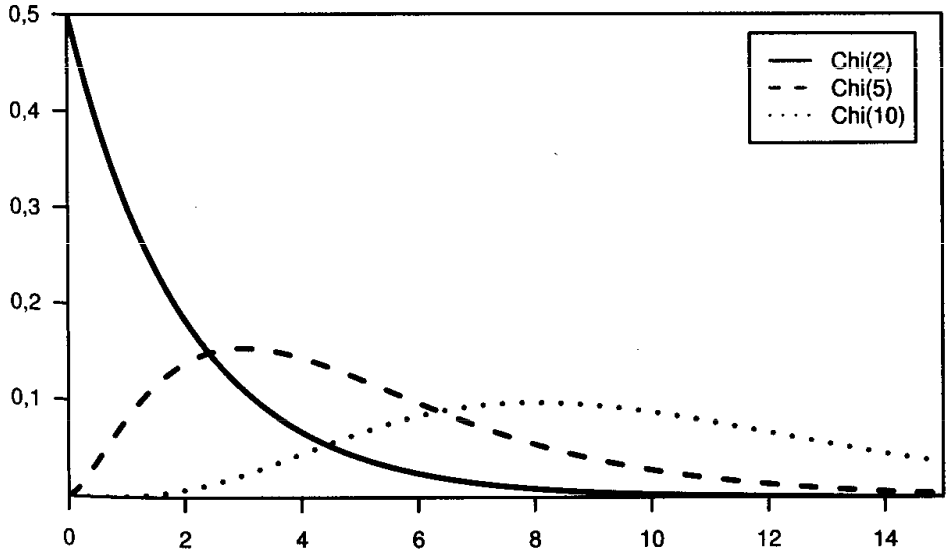
Statistická indukce

– zobecnění závěrů získaných zpracováním výběru na celý základní soubor.

Rozdělení χ2 (chí kvadrát)

Pro výběr n prvků z normovaného normálního rozdělení (z_1, z_2, ... , z_n) lze provést součet jeho čtverců χ2.

$$\chi^2 = \sum_{i=1}^{n} z_i^2$$

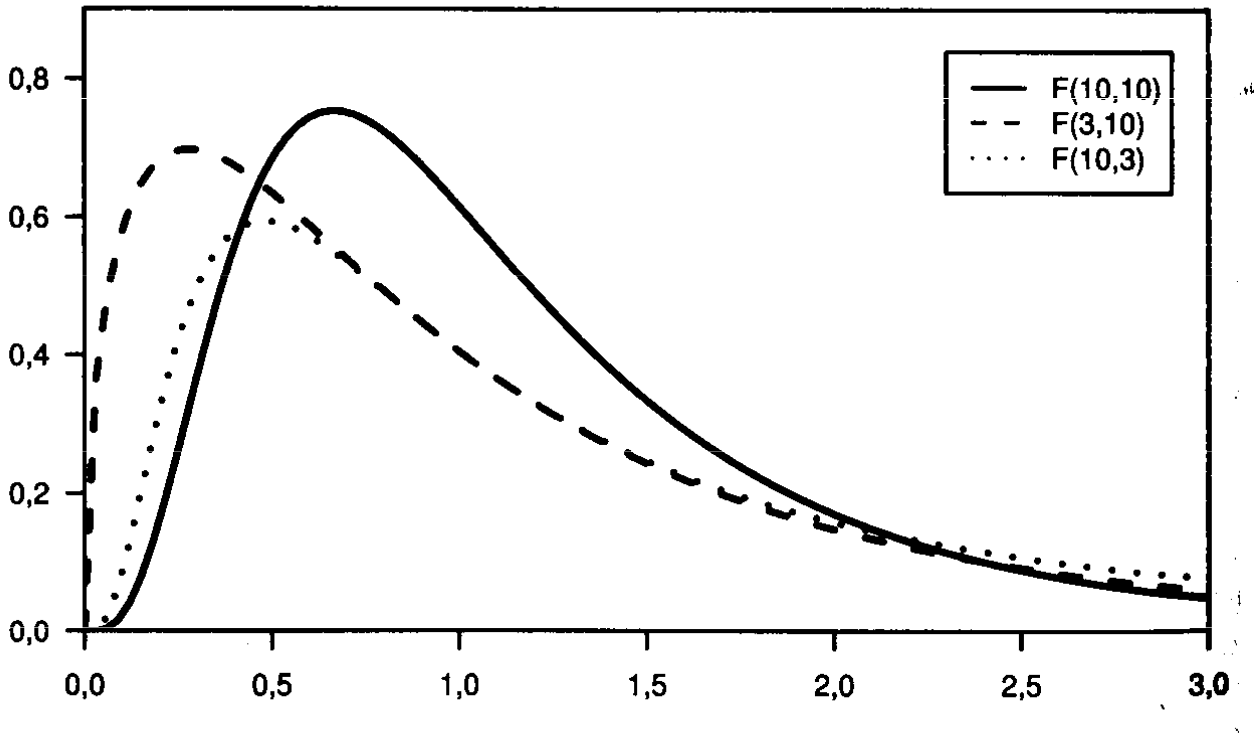


S rostoucí počtem stupňů volnosti (v) se blíží rozdělení normálnímu.

Fisher – Snedecorovo rozdělení
(F-rozdělení)

Náhodná veličina F je definována jako poměr dvou nezávislých náhodných veličin, které mají rozdělení χ2 s v1, resp. v2 stupni volnosti.

$$F = \frac{\frac{\chi_1^2}{\nu_1}}{\frac{\chi_2^2}{\nu_2}}$$



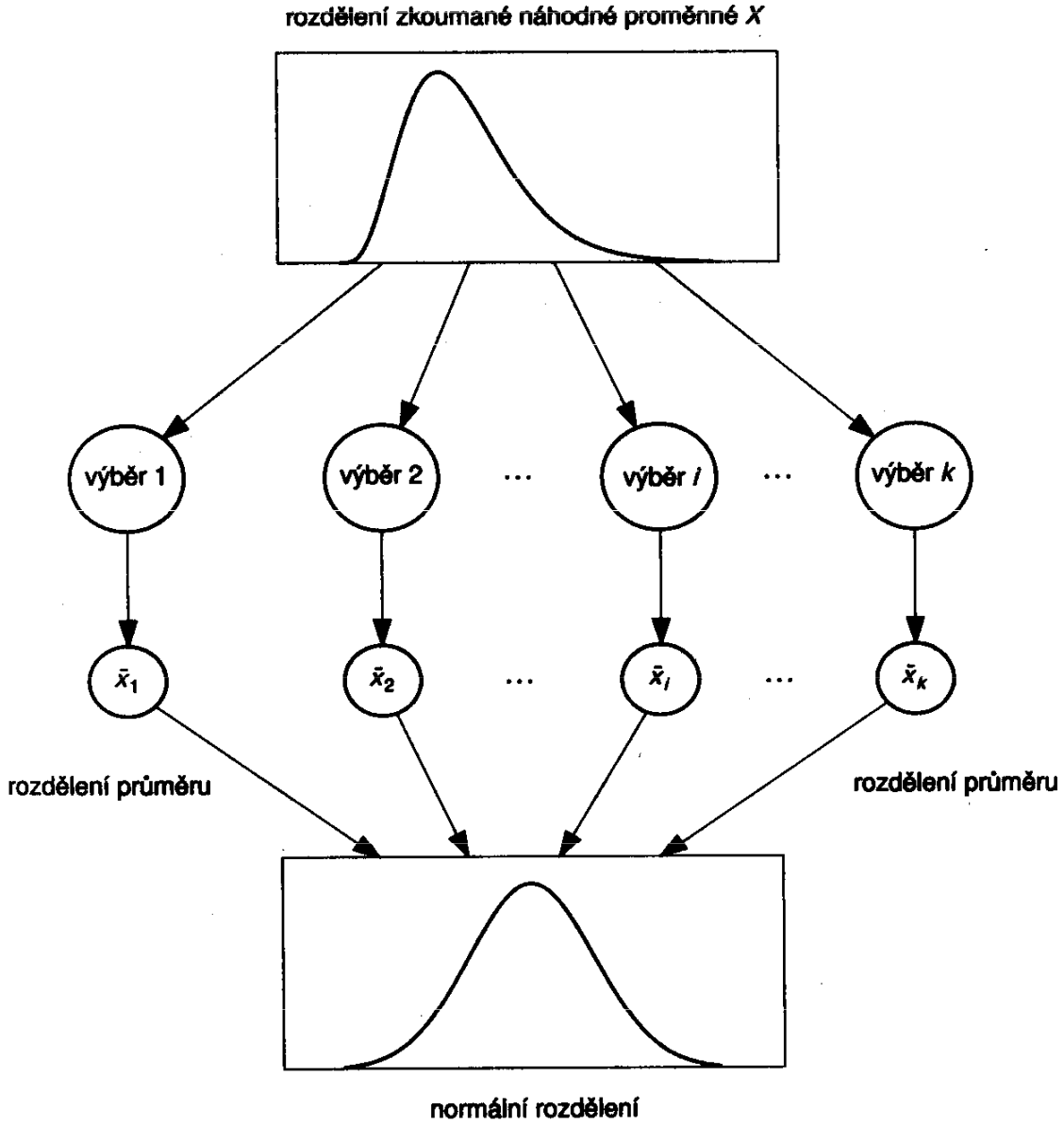
Aritmetický průměr jako náhodná veličina

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{\bar{x} - \mu}{\sigma_x}$$

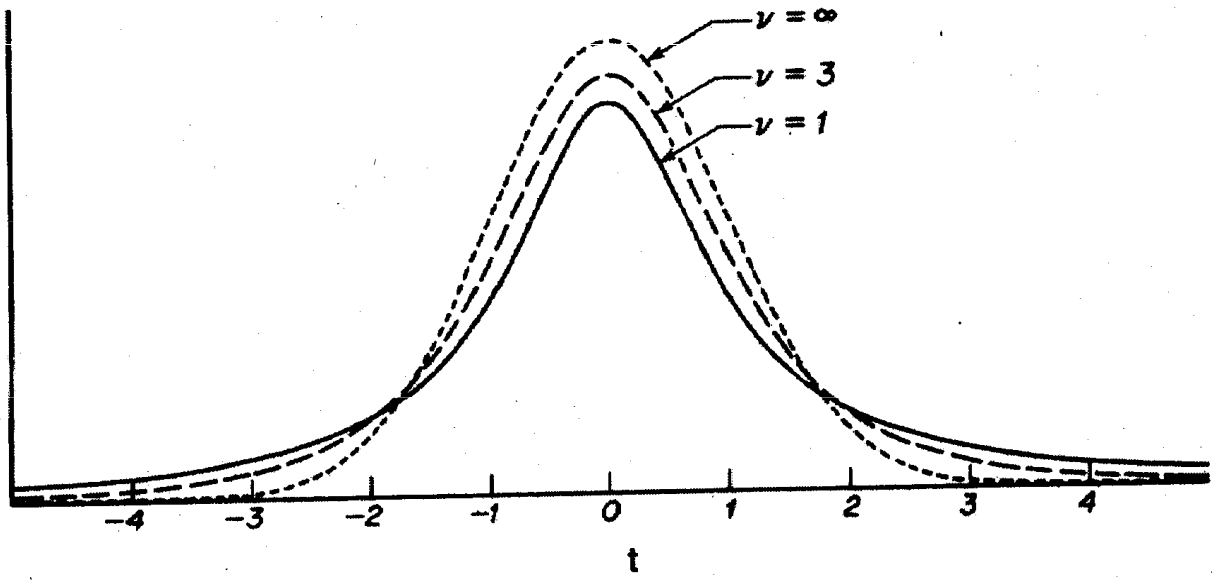
$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s_x}$$



Studentovo rozdělení
(t-rozdělení)

$$t = \frac{\bar{x} - \mu}{s_x}$$



Obr. 5-1 *t*-distribuce pro různé stupně volnosti, ν. Pro ν=∞, *t* distribuce je identická s normální distribucí.

Stratifikovaný výběr

pokud známe faktor, který by mohl sledovanou vlastnost ovlivňovat, můžeme populaci rozdělit do dílčích skupin (vrstev, strat) a provádět náhodný výběr odděleně v každé vrstvě. Zjištěné výsledky se pak slučují vhodnou metodou, respektující velikost vrstev.

Odhad střední hodnoty a rozptylu na základě znalosti odhadů z dílčích výběrů

Na základě dílčích průměrů

$$\bar{x}_1, \bar{x}_2, \ldots \bar{x}_k$$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k}$$

Na základě znalosti dílčích rozptylů s^2_1, s^2_2, …, s^2_k

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s^2{}_i + \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{k} n_i - 1}$$

Statistické odhady

- **Bodové** = 1 hodnota: vlastní odhad parametru základního souboru z výběrových charakteristik

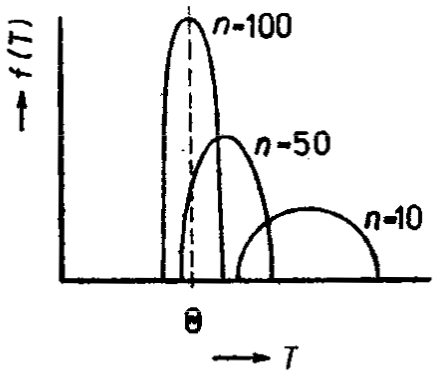
- **Intervalové** = bodové odhady + jejich přesnost (ta roste s rozsahem výběru)

  2 hodnoty: hranice intervalu spolehlivosti

Vlastnosti bodových odhadů

- Konzistence

Odhad je konzistentní, pokud se s rostoucím *n* zmenšuje rozdíl mezi odhadem a skutečnou hodnotou parametru.

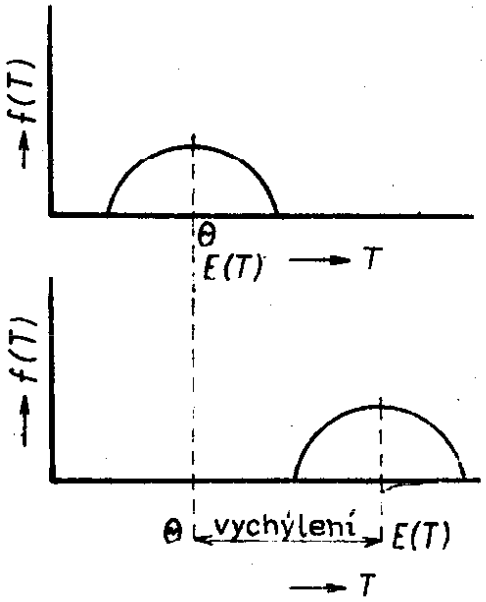


Obr. 5. Konzistentní odhad

Vlastnosti bodových odhadů

- Nestrannost (nevychýlenost)

Odhad je nestranný pokud je při malém n stejná pravděpodobnost že bude odhad podhodnocený stejně jako nadhodnocený.

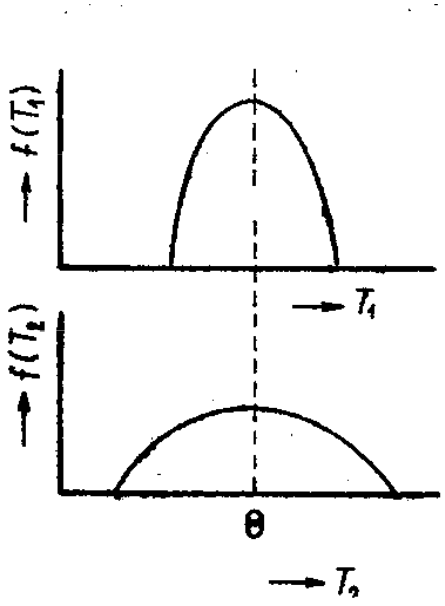


Vychýlený a nevychýlený odhad

Vlastnosti bodových odhadů

- Vydatnost (eficience)

Odhad je vydatný, když se jeho rozptyl okolo skutečné hodnoty s rostoucím n zmenšuje.



Obr. 6. Eficientní odhad

Intervalový odhad

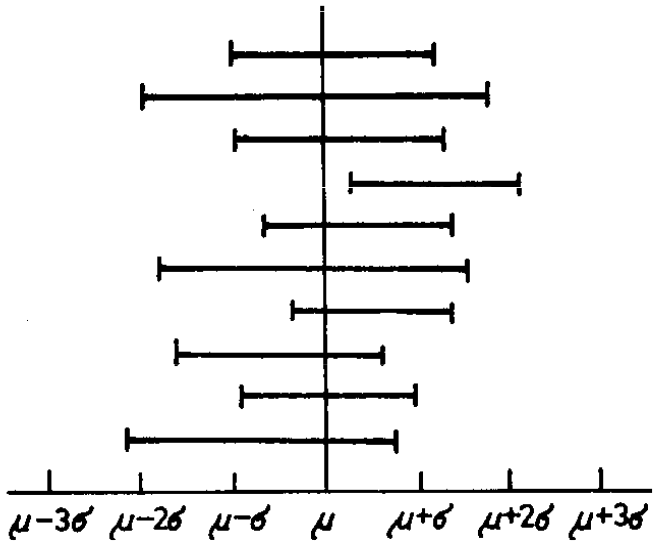
Střední hodnota výběrového aritmetického průměru leží s určitou pravděpodobností (1 - α) v intervalu

$$\mu \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

Pro výběrový aritmetický průměr platí

$$\mu - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$



$$\bar{x} - t_\alpha (n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_\alpha (n-1) \frac{s}{\sqrt{n}}$$

Interval spolehlivosti střední hodnoty

- S použitím kvantilů t-rozdělení

$$\bar{x} - t_\alpha(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_\alpha(n-1)\frac{s}{\sqrt{n}}$$

$t_\alpha(n-1)$ jsou tabelovány

- S použitím variačního rozpětí R (Dean a Dixon)

$$\bar{x} - K_n R \le \mu \le \bar{x} + K_n R$$

K_n jsou tabelovány.

Intervalový odhad

Pro výběrový rozptyl platí

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

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}$$

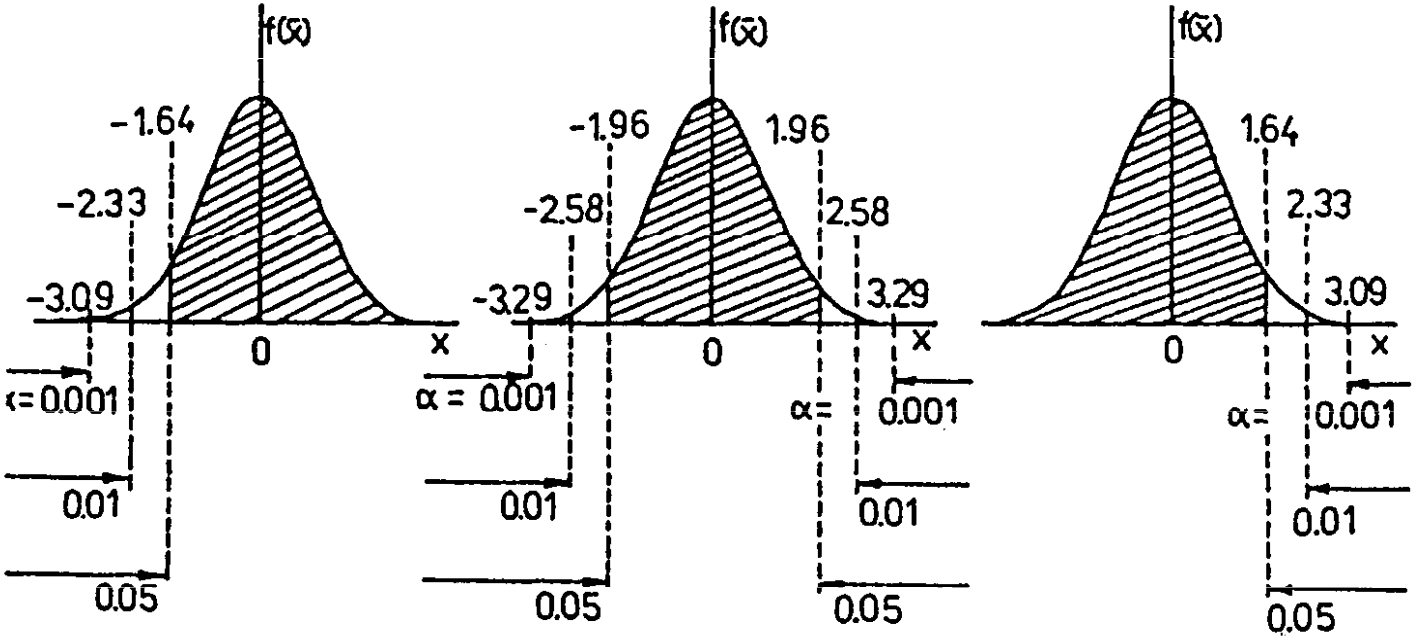
Interval spolehlivosti

- Jednostranný
- Oboustranný

$$H_0 : \mu \leq \mu_0$$

$$H_0 : \mu \geq \mu_0$$

$$H_0 : \mu = \mu_0$$



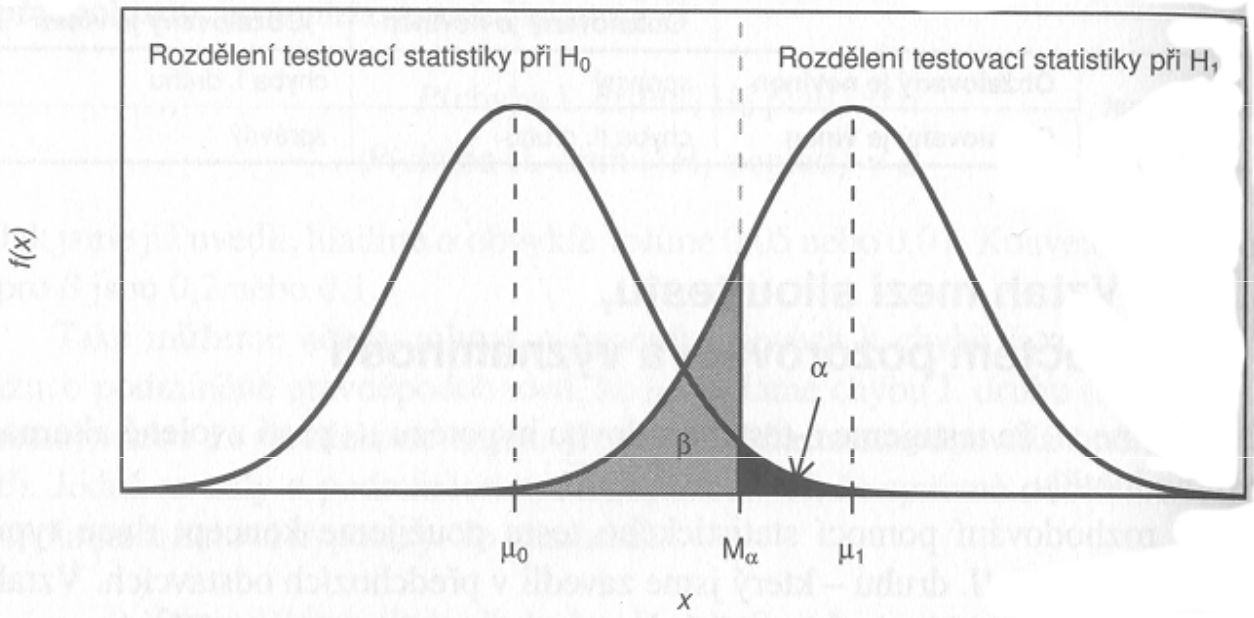
Testování hypotéz

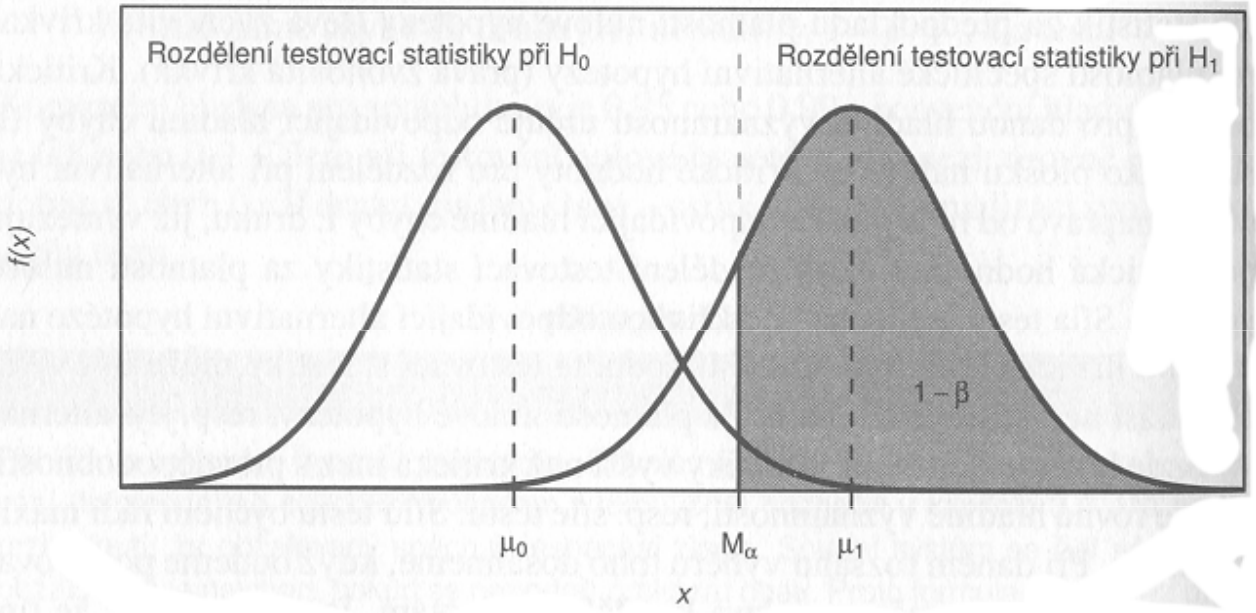
- Formulace hypotézy
  - » nulová hypotéza (H_0)
  - » alternativní hypotéza (H_1)
- Volba hladiny významnosti α
- Volba testu a výpočet testovacího kritéria.
- Interpretace výsledků (zamítnutí/nezamítnutí H_0)

Testování hypotéz

Závěr	Skutečnost	
	H_0 je pravdivá	H_0 je nepravdivá
Zamítáme H_0	riziko α chyba I. druhu	(1- β) správný závěr
Nezamítáme H_0	(1- α) správný závěr	riziko β chyba II. druhu

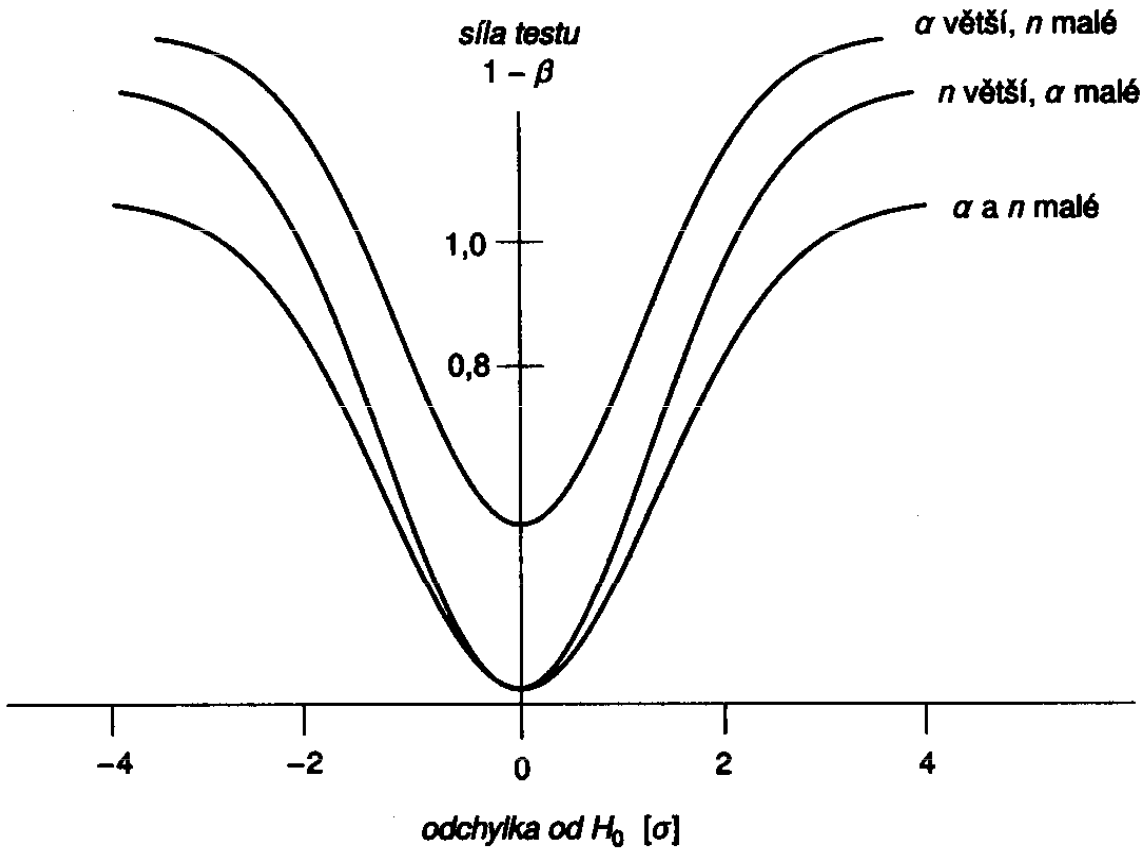
Testování hypotéz





α = hladina významnosti (nejčastěji 0,05)

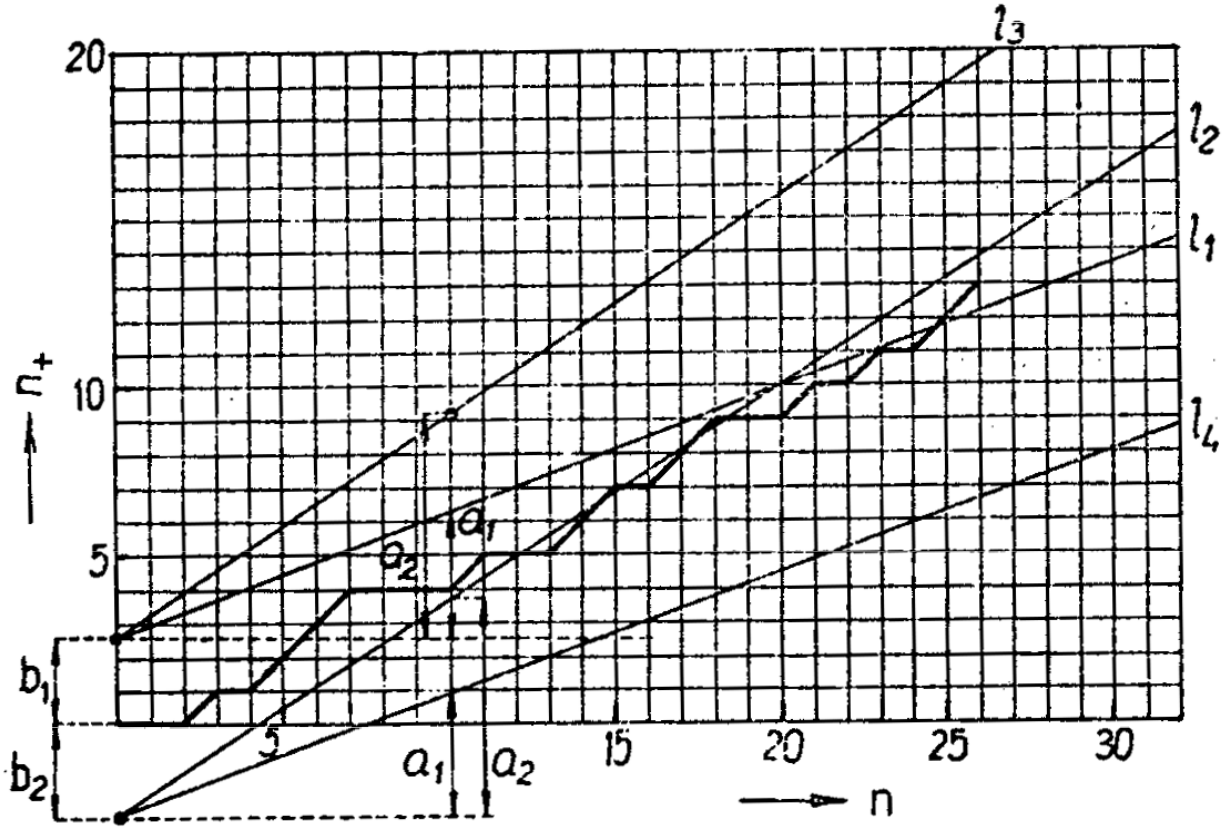
Síla testu = pravděpodobnost, že se vyhneme chybě II.druhu



Nezamítnutí hypotézy H_0 tedy může nastat nejen díky její platnosti, ale také, zejména pro malé rozsahy výběrů, i jako důsledek chyby II. druhu !!!!

Sekvenční testy

Spolehlivost statistických testů je do značné míry závislá na rozsahu zpracovávaného souboru (počtu stanovení), takže při malém počtu výsledků mohou být závěry nesprávné (důsledek chyby II. druhu).



Sekvenční analýza umožňuje rozhodnutí mezi třemi alternativami:

1. je zcela spolehlivě prokázána statistická nevýznamnost, přijímáme hypotézu H_0 na hladině významnosti α.

2. je zcela spolehlivě prokázána statistická významnost, přijímáme hypotézu H_1 na hladině významnosti β.

3. počet výsledků n je příliš malý k spolehlivému přijetí jedné z obou alternativních hypotéz (hladiny α a β volíme dle závažnosti rizika nesprávného rozhodnutí; nejčastěji se volí α = β = 0,05).

Základní předpoklady o datech

- Nezávislost (náhodnost výběru)
- Minimální velikost výběru
- Homogenita
- Odlehlé hodnoty
- Normalita

Nezávislost

- Test autokorelace

významnost autokorelačního koeficientu prvního řádu ρ_A podle testovacího kritéria

$$t = T_1 \frac{\sqrt{n+1}}{1-T_1}$$

$$T_1 = (1 - T/2)\sqrt{\frac{n^2-1}{n^2-4}}$$

a T je von Neumanův poměr

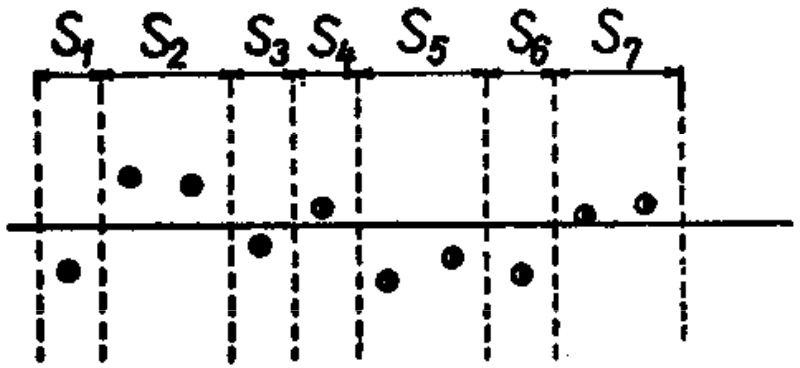
$$T = \frac{\sum_{k=2}^{n}(x_k - x_{k-1})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

Za předpokladu, že platí nulová hypotéza $\rho_A = 0$, má veličina T_1 Studentovo rozdělení s n+1 stupni volnosti s kritickým oborem $|t| > t_{1-\alpha/2}(n+1)$.

Nezávislost

Skupinový test.

Mediálou (přímka rovnoběžná s osou x) rozdělíme data, vzhledem k ose x na dvě poloviny), data pak rozdělíme do skupin podle toho, zda jsou nad, či pod mediánou. Počet takto získaných skupin z n hodnot porovnáme s tabulkou.



Trend výsledků

Nezávislost

- Spearmanův korelační koeficient

  (viz korelace)

- Znaménkový test

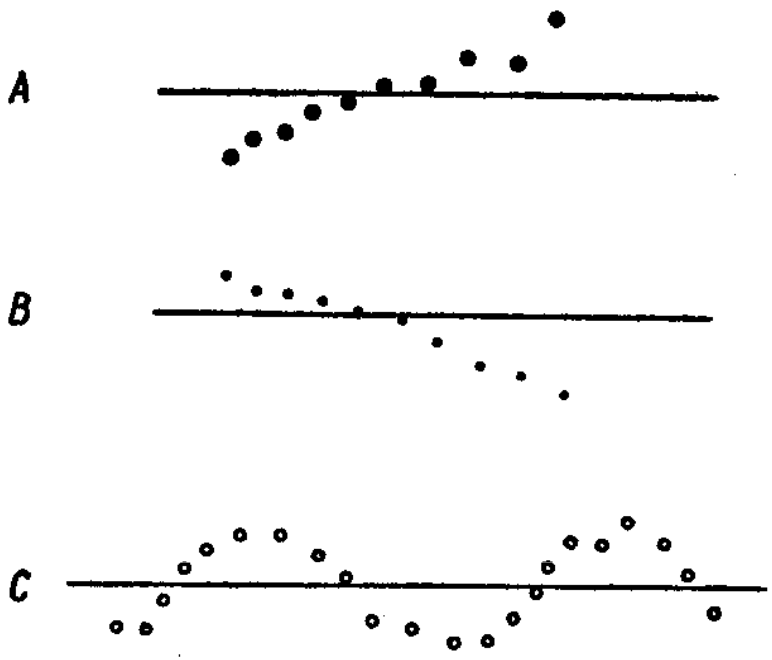
  vypočítají se odchylky testu a určí se poměr n_+/n_-, ten se testuje pomocí binomického rozdělení.

Nezávislost

Body zvratu

pro periodický trend **x**i-1 < **x**i > **x**i+1, resp. **x**i-1 > **x**i < **x**i+1

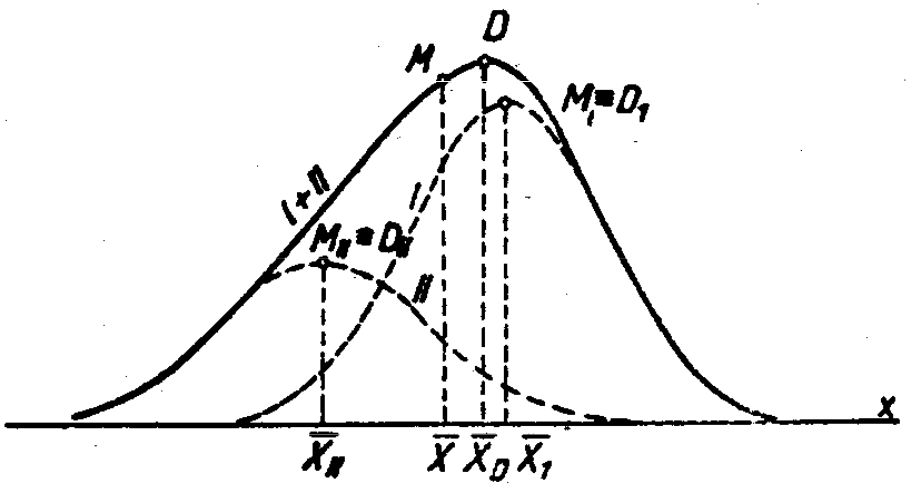
$$U = \frac{Z - \frac{2n-4}{3}}{\sqrt{\frac{16n-29}{90}}}$$



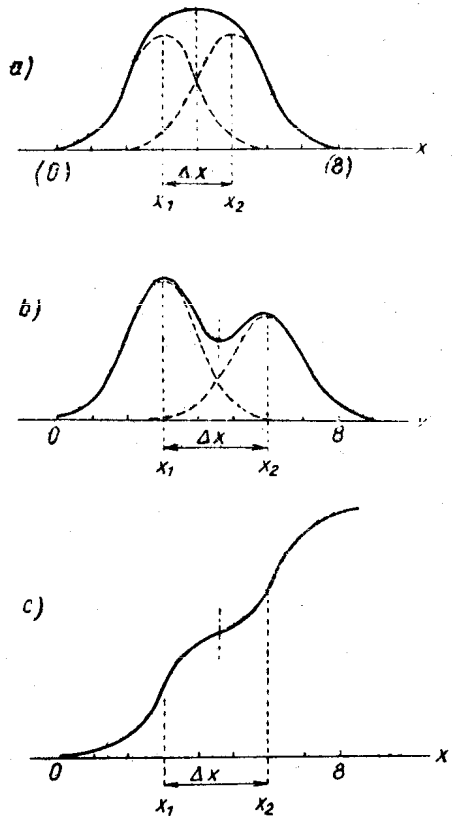
Z je celkový počet bodů zvratu ve sledované posloupnosti. Veličina U má při $n \to \infty$ asymptoticky normální rozdělení N(0, 1). Hypotézu zamítáme když $|U| \geq u(\alpha/2)$

Homogenita

$$(1-\varepsilon)f(x,\Theta)+\varepsilon.f_{\varepsilon}(x,\Theta_{\varepsilon})=(1-\varepsilon)\frac{1}{\sigma\sqrt{2\pi}}\exp-\left(\frac{x-\mu}{2\sigma}\right)^2+\varepsilon\frac{1}{\sigma_{\varepsilon}\sqrt{2\pi}}\exp-\left(\frac{x-\mu}{2\sigma_{\varepsilon}}\right)^2$$



Vznik asymetrie frekvenční křivky



Vznik ploché a dvouvrcholové frekvenční křivky (*x* značí střední hodnotu základního souboru)

Odlehlé hodnoty

vedou k vychýleným odhadům

- Grafické metody

  box and whisker plot

Grubbsův test

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

$$T_n = \frac{x_n - \bar{x}}{S_n} \qquad T_1 = \frac{\bar{x} - x_1}{S_n}$$

Deanův a Dixonův test

$$Q_n = \frac{x_n - x_{n-1}}{R} \qquad Q_1 = \frac{x_2 - x_1}{R}$$

- *Metoda modifikace vnitřních hradeb*

Modifikované vnitřní hradby jsou definovány

- dolní vnitřní hradba: $B_D^* = \tilde{x}_{0,25} - K(\tilde{x}_{0,75} - \tilde{x}_{0,25})$
- horní vnitřní hradba: $B_H^* = \tilde{x}_{0,75} - K(\tilde{x}_{0,75} - \tilde{x}_{0,25})$

Parametr K se volí tak, aby byla vysoká pravděpodobnost, že z výběru velikosti *n* z normálního rozdělení nebude žádný prvek mimo modifikované vnitřní hradby (obvykle se volí pravděpodobnost 0,95). Pro n v rozmezí $8 \leq n \leq 100$ lze použít aproximace

$$K = 2{,}25 - 3{,}6/n$$

Odlehlé hodnoty

- Vyloučení odlehlých hodnot ze souboru (nedoporučuje se, zejm. u malých výběrů)
- Použití robustních parametrů polohy

medián

uřezaný průměr $$\bar{x}_u = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} x_i$$

winsorizovaný průměr $$\bar{x}_w = \frac{1}{n}\sum_{i=m+2}^{n-m-1} x_i + (m+1)(x_{m+1} + x_{n-m})$$

$m = \text{int}(Un/100)$ U je procento uřezaných pořádkových statistik, nejlépe 10%

Minimální velikost výběru

- Pro zvolenou střední chybu průměru (x – μ):

$$n_{\min} > \left(\frac{\sigma}{\bar{x} - \mu} \right)^2 = \left(\frac{t_\alpha s}{\bar{x} - \mu} \right)^2$$

Nutná je znalost směrodatné odchylky nebo jejího odhadu. Pro $\alpha = 0{,}05$ je t_α přibližně rovno 2.

Normalita

- Grafické metody

  box and whisker plot

  histogram a jádrový odhad

Kvantil-kvantilový (QQ) graf

osa x: výběrové kvantily

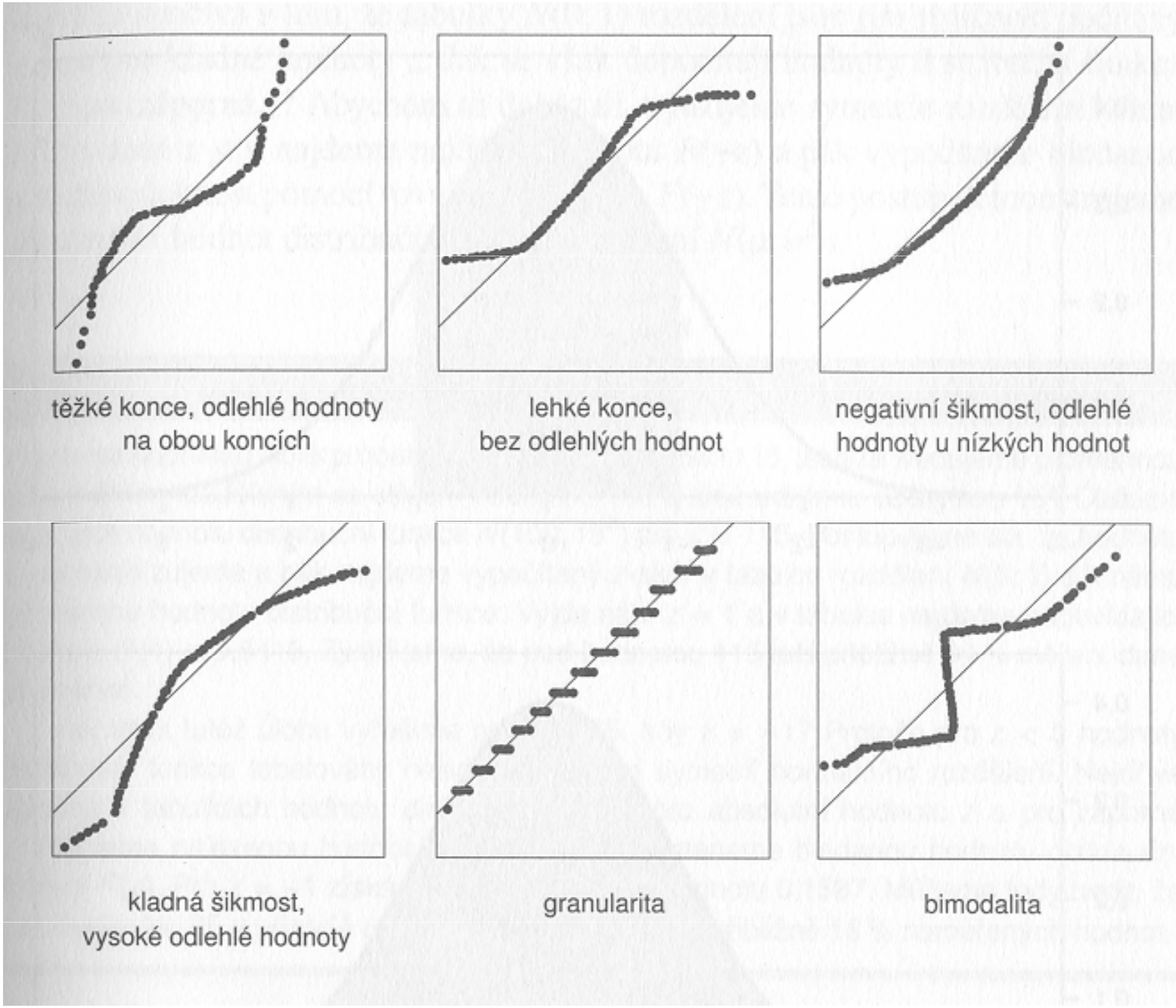
osa y: kvantily teoretického rozdělení (nejč. norm. normálního rozd.)

Pravděpodobnostní (PP) graf

osa x: standardizovaná proměnná $\frac{x_i - \bar{x}}{s}$ $\frac{x_i - \tilde{x}}{MD}$

osa y: standardizovaná distr. funkce teoretického rozdělení (nejč. norm. normálního rozd.)

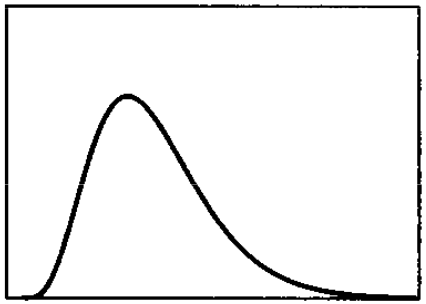
Kvantil – kvantilový graf



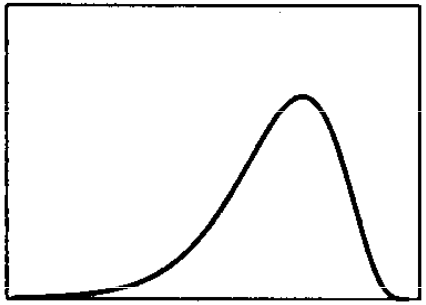
Normalita

- Anderson – Darlingův test
- Shapirův – Wilkův test
- Test šikmosti a špičatosti
- Test dobré shody
- Kolmogorovův a Lilieforsův test

Šikmost

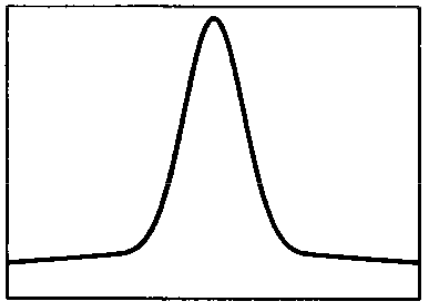


zešikmené zprava
(kladné zešikmení)

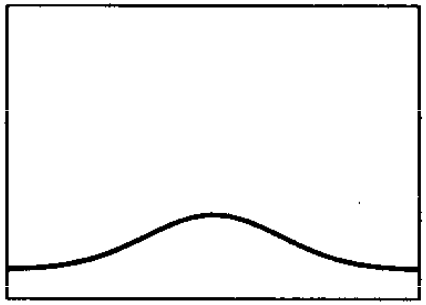


zešikmené zleva
(záporné zešikmení)

Špičatost

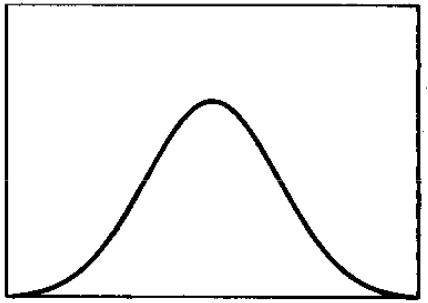


leptokurtické
(špičatější než normální)

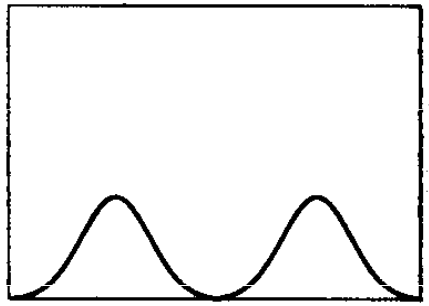


platykurtické
(méně špičaté než normální)

Počet vrcholů



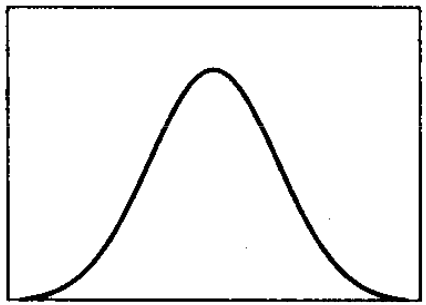
jeden vrchol
(unimodální)



dva vrcholy
(bimodální)

$$\hat{x} > \tilde{x} > \bar{x}$$

pravostranná asymetrie



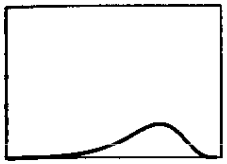
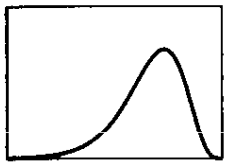
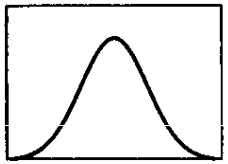
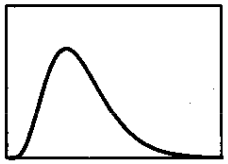
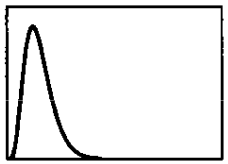
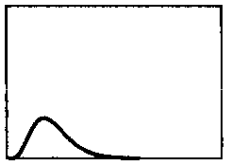
symetrické, jeden vrchol,
zvonovitý tvar

$$\bar{x} > \tilde{x} > \hat{x}$$

levostranná asymetrie

Transformace dat

- Logaritmická
- Mocninná
- Box-Coxova

Problém	Transformace	Efekt
	$X = X^3$	snižuje (extrémní) zešikmení zleva
	$X = X^2$	snižuje zešikmení zleva
	$X = X^1$	žádný účinek
	$X = X^{1/2} = \sqrt{X}$	snižuje zešikmení zprava
	$X = \ln X$	snižuje zešikmení zprava
	$X = -X^{-1/2} = -1/\sqrt{X}$	snižuje (extrémní) zešikmení zprava

Testy shody

- Středních hodnot (testy správnosti)
- Rozptylů (testy přesnosti)
- Rozdělení

s jedním výběrem
se dvěma výběry

Test shody středních hodnot s jedním výběrem (μ je známo)

- Studentův test

$$t = \frac{|\bar{x} - \mu|}{s}$$

- Lordův test

$$u_n = \frac{|\bar{x} - \mu|}{R}$$

Test shody středních hodnot se dvěma výběry

Pro $n_1 = n_2$

Studentův test

$$t = \frac{|\bar{x}_1 - \bar{x}_2|\sqrt{n-1}}{\sqrt{(s_1^2 + s_2^2)}}$$

Lordův test

$$u = \frac{|\bar{x}_1 - \bar{x}_2|}{R_1 + R_2}$$

Test shody středních hodnot se dvěma výběry

Pro $n_1 \neq n_2$

Studentův test

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\left[s_1^2/(n_1 - 1) + s_2^2/(n_2 - 1)\right]}}$$

Moorův test

$$U = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{R_1 + R_2}$$

t-testy výběrů s nestejnými rozptyly

Shoda s_1^2 a s_2^2 se testuje F-testem

$s_1^2 = s_2^2$

v tabulkách

$$t_\alpha = t(n_1 + n_2 - 2)$$

$s_1^2 \neq s_2^2$

$$t_\alpha = \frac{t_1 s_1^2 /(n_1 - 1) + t_2 s_2^2 /(n_2 - 1)}{s_1^2 /(n_1 - 1) + s_2^2 /(n_2 - 1)}$$

Neparametrické testy shody středních hodnot

test shody mediánů

Wilcoxonův test

Mann – Whitneyův test

znaménkový test

Závislé hodnoty (bloky)

- Párový t- test

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

- Znaménkový test
- Wilcoxonův test
- Permutační (Bootstrap) test

není nutný předpoklad náhodného výběru.

Párový t-test a ANOVA

- Párový t-test lze užít **pouze** pro srovnání **dvou** souborů!!!
- **Nelze** ho použít pro srovnání více souborů způsobem „každý s každým“ – výsledky nejsou nezávislé a je problém s odhadem α (chyby I. druhu).
- V případech více než dvou souborů lze použít pouze analýzu rozptylu (ANOVU)

Párový t-test a ANOVA

Počet průměrů (k)	Hladina signifikance užívaná v *t* testech					
	0.20	0.10	0.05	0.02	0.01	0.001
2	0.20	0.10	0.05	0.02	0.01	0.001
3	0.41	0.23	0.13	0.05	0.03	0.003
4	0.58	0.36	0.21	0.09	0.05	0.006
5	0.71	0.47	0.23	0.13	0.07	0.009
10	0.96	0.83	0.63	0.37	0.23	0.034
20	1.00	0.98	0.92	0.71	0.52	0.109
∞	1.00	1.00	1.00	1.00	1.00	1.00

Pravděpodobnost, že se dopustíme chyby I. druhu, budeme-li užívat více *t* testů při hledání rozdílů mezi všemi páry ve skupině *k* průměrů.

Test shody rozptylů

- F-test (Fisher-Snedecorův)

$$F = \frac{s_1^2}{s_2^2}$$

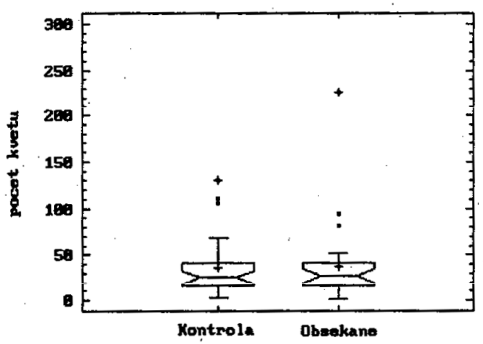
- Leveneův test
- Jacknife testy

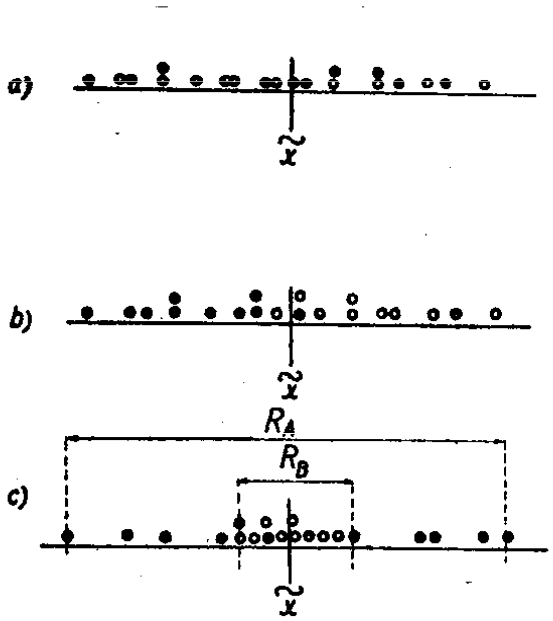
Test shody středních hodnot a rozptylů

Grafické metody

box and whisker plot

- histogramy
- stem and leaf plot
- Lewisův test





LEWISŮV test

● - *A*, ○ - *B*; *a* - metoda *A* i *B* dává stejné výsledky, *b* - metoda *B* dává vyšší výsledky, metoda *A* nižší, *c* - metoda *A* dává méně přesné výsledky než metoda *B*

Kolmogorovův a Smirnovův test

Srovnání empirické výběrové distribuční funkce s

distribuční funkcí (Kolmogorov)

jinou empirickou distribucí (Smirnov)

$D = \max /F(x) - F_n(x)/$