ANALÝZA JEDNOROZMĚRNÝCH DAT

Soubor zahrnuje množství sledovaných případů, resp. předmětů, které mají z daného hlediska stejné vlastnosti.

a) **Konečný** soubor: množství může být omezené,

b) **Nekonečný** soubor: množství může být neomezené,

Náhodný výběr:

1. Všechny prvky výběru $\{x_i\}$, $i = 1, ..., n$, se chápou jako náhodné veličiny, které se řídí ***stejným zákonem rozdělení*** t. zn., že výběr je **homogenní.**
2. Hodnoty x_i výběru jsou vybrány nezávisle na sobě.
3. Je charakterizován:
   a) informací o ***střední hodnotě***,
   b) informací o ***rozptýlení*** kolem střední hodnoty.
   c) informací o ***tvaru*** výběrového rozdělení.

ZÁKLADNÍ POJMY a analýza výběru

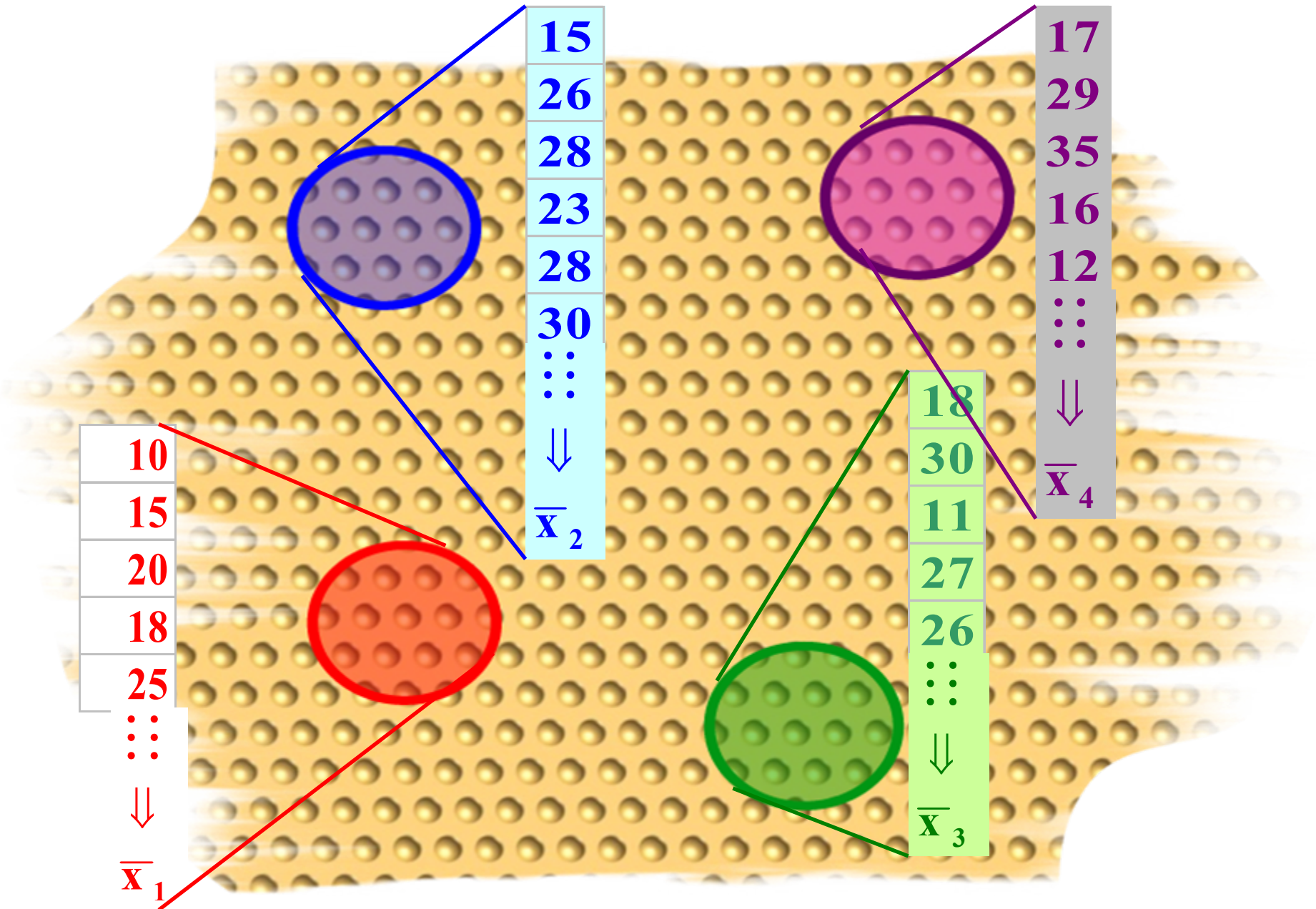
PARAMETR je statistická charakteristika **základního souboru** (značí se řeckými písmeny, např. střední hodnota μ).

STATISTIKA je statistická charakteristika **výběrového souboru** (značí se latinkou, např. výběrový průměr $\bar{x}$).

ODHAD PARAMETRU ZÁKLADNÍHO SOUBORU je určitá hodnota (hodnoty) vhodné statistiky získaná postupem zvaným **bodový** nebo **intervalový odhad**.

VÝBĚROVÉ ROZDĚLENÍ je rozdělení pravděpodobnosti všech možných hodnot statistiky, které mohou být zjištěny z náhodných výběrů stejné velikosti vybraných z určitého základního souboru.

VÝBĚROVÝ PRŮMĚR a jeho rozdělení



Bodové odhady parametrů

Odhad parametru (polohy, rozptýlení a tvaru) rozdělení náhodné veličiny výběrovou charakteristikou nazýváme ***bodový odhad***.

Vlastnosti:

1. ***Konzistence odhadu***: odhad $\hat{\Theta}$ je konzistentní, když s rostoucím počtem pozorování se odhad blíží k teoretické hodnotě s pravděpodobností 1.
2. ***Nestrannost odhadu***: odhad je nestranný, když pro daný rozsah výběru n je jeho střední hodnota rovna parametru souboru Θ čili kolísá symetricky kolem Θ na obě strany.
3. ***Vydatnost odhadu***: odhad je vydatný, když je jeho rozptyl okolo skutečné hodnoty Θ minimální vzhledem ke všem možným odhadům tohoto parametru.
4. ***Robustnost odhadu***: odlehlé hodnoty nemají velký vliv na hodnotu odhadu.

Za nejlepší nestranný odhad se považuje ten,
který je zároveň nestranný, vydatný a postačující.

ZÁKLADNÍ VLASTNOSTI BODOVÝCH ODHADŮ

- **Nespornost (konzistence):** se vzrůstající velikostí výběru vzrůstá i pravděpodobnost, že odhad bude blízký skutečnému parametru.

- **Nevychýlenost (nestrannost):** střední hodnota výběrové statistiky se rovná odhadovanému parametru základního souboru

$$E(T) = P$$

- **Vydatnost:** znamená, že daný nevychýlený odhad má se srovnání se všemi ostatními nevychýlenými odhady nejmenší variabilitu (rozptyl)

Metoda maximální věrohodnosti

Pravděpodobnost současného výskytu všech prvků výběru

$$L(\Theta) = \prod_{i=1}^{n} p(x_i, \Theta)$$

Pro ***spojité náhodné veličiny*** je věrohodnostní funkce $L(\Theta)$ sdružená hustota pravděpodobnosti $f(\Theta, x_1, \ldots, x_n)$

$$L(\Theta) = \prod_{i=1}^{n} f(x_i, \Theta)$$

místo $L(\Theta)$ se používá i jejího logaritmu.

K určení **maximálně věrohodných odhadů** je soustava rovnic

$$\frac{\delta \ln L(\Theta)}{\delta \Theta_j} = \sum_{i=1}^{n} \frac{\delta \ln f(x_i, \Theta)}{\delta \Theta_j} = 0$$

pro $j = 1, \ldots, m$, kde m je počet odhadovaných parametrů.

Platí: pro dostatečně veliké rozsahy výběru (asymptoticky pro $n \to \infty$) jsou tyto odhady nejlepší, nestranné a mají normální rozdělení.

Konstrukce bodových odhadů

- **Metoda maximální věrohodnosti (MLE):** Konstruují se odhady maximalizující věrohodnostní funkci L (resp její logaritmus lnL**)**.

- Věrohodnostní funkce je sdružená hustota pravděpodobnosti všech prvků výběru. Pro nezávislá měření (nezávislé chyby ε_i) platí

$$\ln L(\mu, \sigma^2, a) = \sum \ln f(x_i \mu, \sigma^2, a)$$

- Podmínky maxima představují soustavu rovnic

$$\frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \mu} = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \sigma^2} = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2, a)}{\partial a} = 0$$

Pro **rozptyl maximálně věrohodného odhadu** bude platit

$$D(\hat{\Theta}) = -\frac{1}{E\left[\frac{d^2 \ln L}{d\Theta^2}\right]}$$

kde D(.) je operátor rozptylu a E(.) operátor střední hodnoty.

Pro variabilitu se vyčísluje kovarianční matice $\mathbf{C} = \mathbf{A}^{-1}$ obsahující na diagonále rozptyly $D(\hat{\Theta}_j)$ maximálně věrohodných odhadů. Prvky A_{ij} matice **A** jsou dány vztahem $A_{ij} = -E\left[\frac{\delta^2 \ln L}{\delta\Theta_j \, \delta\Theta_i}\right]$.

Závěr: ***rozdělení maximálně věrohodných odhadů*** je asymptoticky normální $N(\Theta, \mathbf{A}^{-1})$.

MLE pro normální rozdělení

- **Věrohodnostní funkce pro náhodný nezávislý výběr z normálního rozdělení**

$$\ln(L) = (-N/2) * [\ln(2\pi) + \ln(\sigma^2)] - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$$

- **Derivace**

$$\frac{\partial \ln(L)}{\partial \mu} = 2\sum_i x_i - 2 * N * \mu$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = \frac{\sum_i (x_i - \mu)^2}{[\sigma^2 \sqrt{2}]^2} - \frac{N}{2\sigma^2}$$

- **Výběrový průměr a rozptyl jsou tedy MLE odhady.**

Derivací logaritmu věrohodnostní funkce

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

podle parametru μ a po úpravě vyjde odhad střední hodnoty, tj. **odhad výběrového průměru**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

a druhou derivací ln L podle μ a dosazením dostaneme **rozptyl** tohoto odhadu

$$D(\hat{\mu}) = \frac{\sigma^2}{n}$$

Odhad rozptylu bude

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

s rozptylem tohoto odhadu $D(\hat{\sigma}^2) = \dfrac{2\,\sigma^4}{n}$

V praxi: parametr μ není znám a nahrazuje se $\hat{\mu} = \bar{x}$.

Rozptyl $\hat{\sigma}^2$ představuje vychýlený odhad, protože

$$E(\hat{\sigma}^2) = K\ \sigma^2, \quad \text{kde } K = (n - 1)/n.$$

Za nevychýlený odhad rozptylu se užívá **odhad výběrového rozptylu**

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

VÝBĚROVÝ PRŮMĚR a jeho rozdělení

Pro rozdělení výběrového průměru platí:

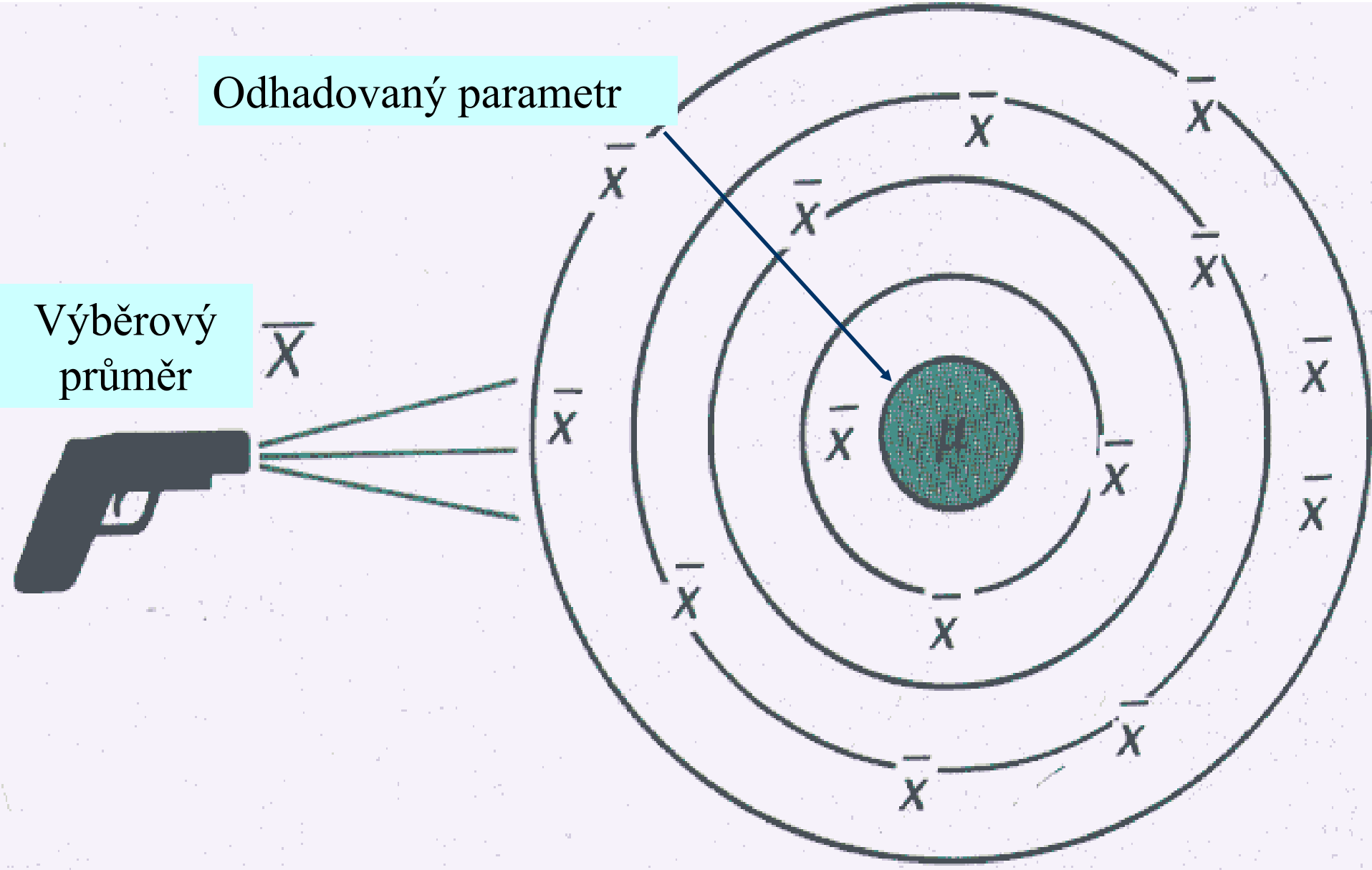
$$E(\bar{X}) = \mu \qquad \sigma_{\bar{X}} = SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Jestliže výběr pochází z normálního rozdělení se střední hodnotou μ a směrodatnou odchylkou σ, potom výběrový průměr $\bar{x}$ má **výběrové normální rozdělení $N(\mu,\sigma^2/n)$**

Z toho vyplývá:

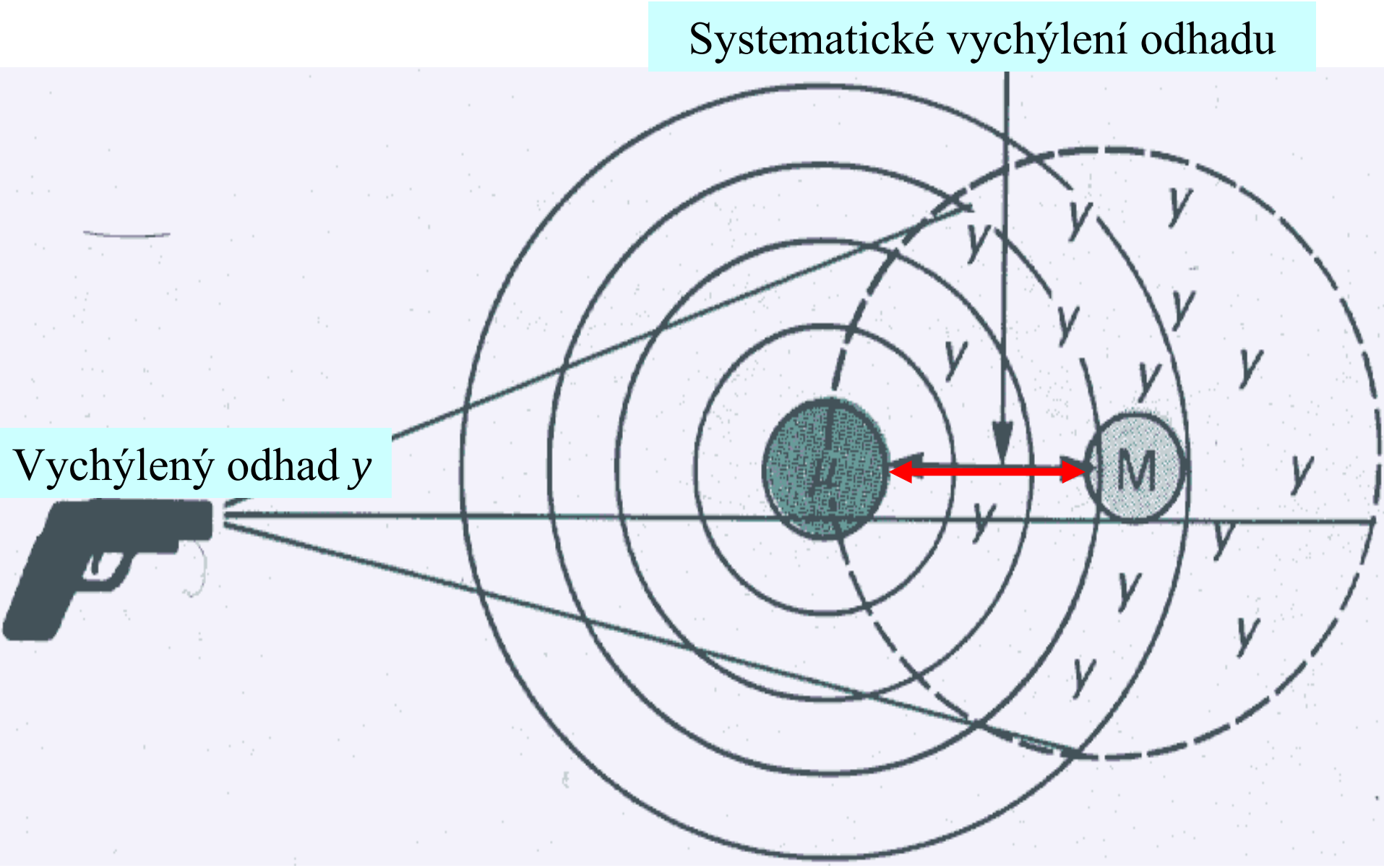
- výběrový průměr je dobrým odhadem střední hodnoty μ
- se vzrůstající velikostí výběru n se snižuje standardní chyba průměru SE a určení střední hodnoty je tím spolehlivější

ZÁKLADNÍ VLASTNOSTI BODOVÝCH ODHADŮ

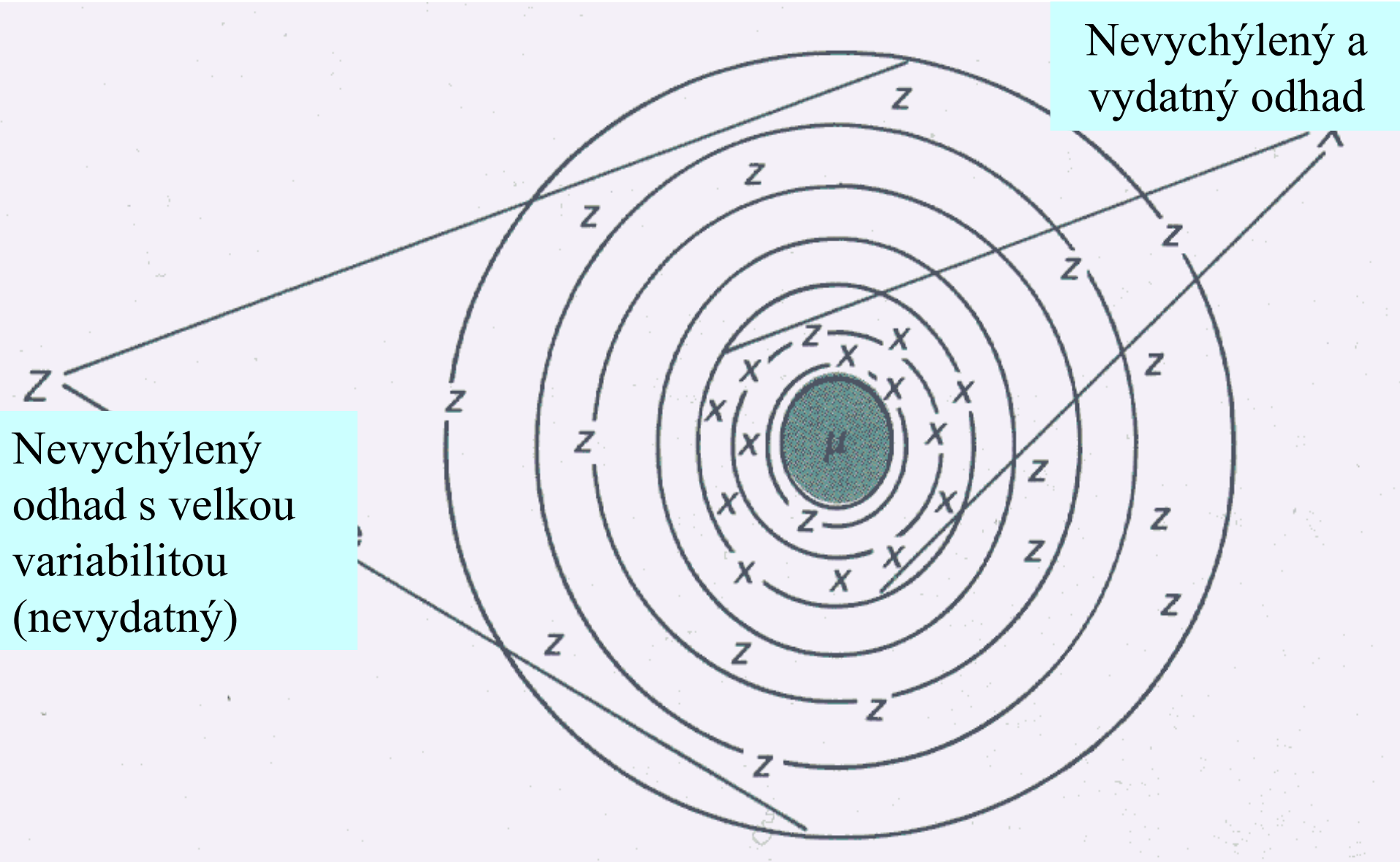


Grafické znázornění nevychýleného odhadu

ZÁKLADNÍ VLASTNOSTI BODOVÝCH ODHADŮ



ZÁKLADNÍ VLASTNOSTI BODOVÝCH ODHADŮ



Bodové míry polohy, rozptýlení a tvaru

1. Mírou polohy je **výběrový průměr** $\bar{x}$, který je také ***prvním centrálním momentem***,

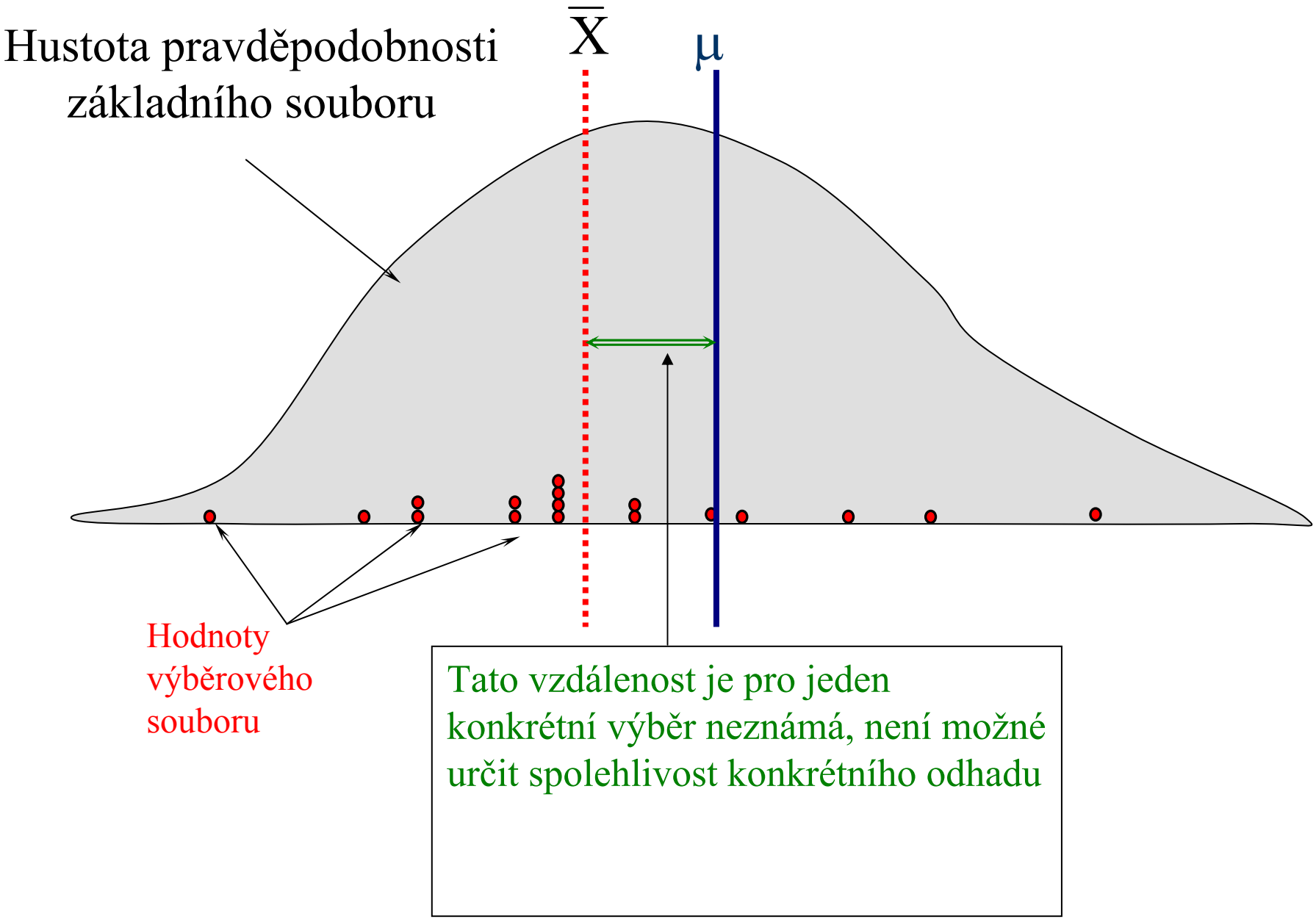
$$E(\hat{x}) = \mu \qquad D(\hat{x}) = \frac{\sigma^2}{n}$$

2. Mírou variability je **výběrový rozptyl** s^2, který je také ***druhým centrálním momentem.***

$$E(s^2) = \sigma^2 \qquad D(s^2) = \frac{\sigma^4}{n}\left[g_2 - \frac{n-3}{n-1}\right]$$

3. **Módus**, $\hat{x}_M$ je lokální maximum na grafu hustoty pravděpodobnosti čili nejčastější prvek výběru.

BODOVÉ ODHADY ZÁKLADNÍCH PARAMETRŮ

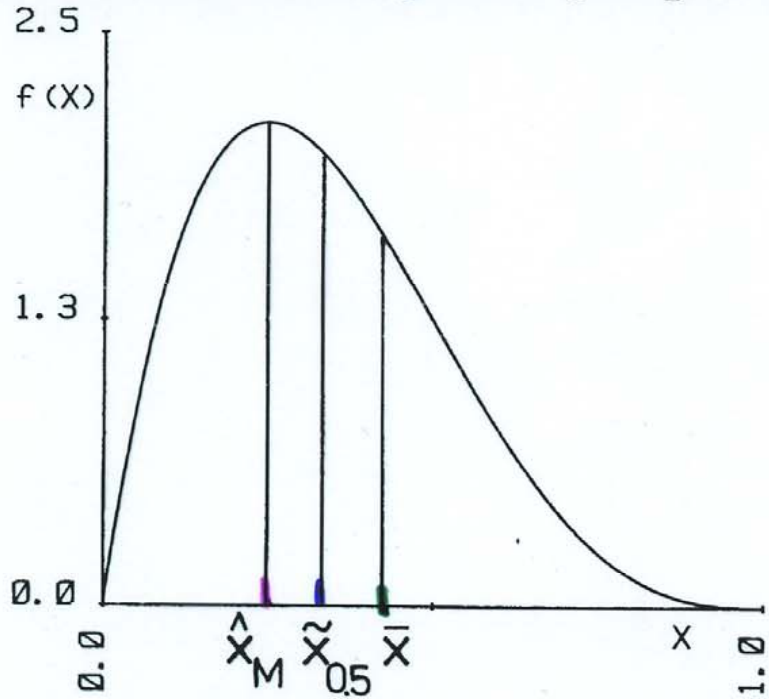


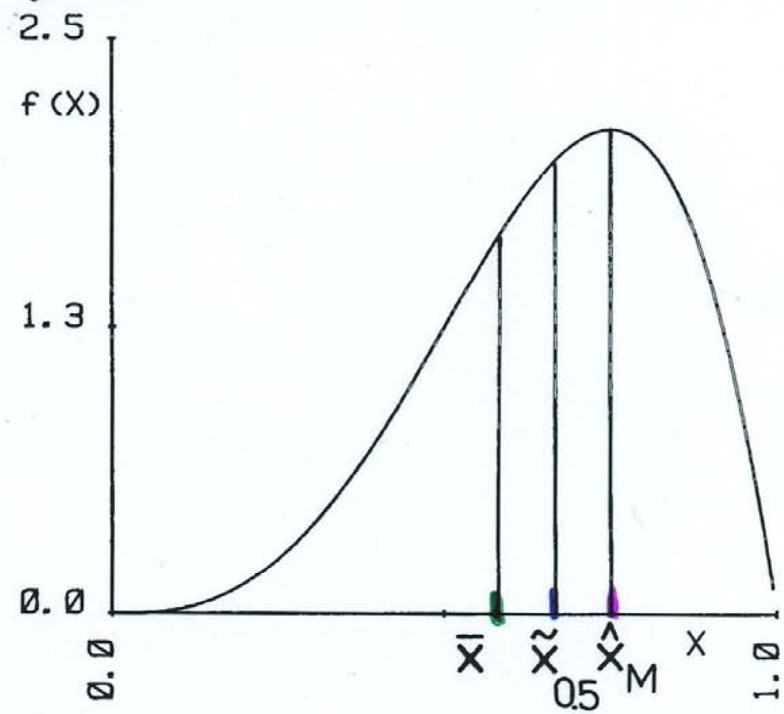
Medián je maximálně věrohodným odhadem polohy u Laplaceova rozdělení a má pro toto rozdělení rozptyl

$$D_L(\tilde{x}_{0.5}) = \frac{\sigma^2}{2\ n}$$

Parametr	Odhad	Rozptyl odhadu	Efektivnost	Rozdělení odhadu
Střední hodnota μ	$\bar{x}$	σ^2/n	1	$N(\mu, \sigma^2)$
	$\tilde{x}_{0.5}$	$\sigma^2\ \pi/(2\ n)$	0.63	$N(\mu, \sigma^2)$
	$\hat{x}_p$	$\sigma^2\ \pi^2/(24\ \ln(n))$	$24\ \ln(n)/(\pi^2\ n)$	$N(\mu, \sigma^2)$
Rozptyl σ^2	$\hat{\sigma}^2$	$2\ \sigma^4/n$	1*)	$N(\sigma^2, D(\sigma^2))$
	s^2	$2\ \sigma^4/(n-1)$	1	
Směrodatná odchylka σ	$\hat{\sigma}$	$\sigma^2/(2\ n)$	~ 1*)	$N(\sigma, D(\sigma))$
	s	$\sigma^2/[2\ (n-1)]$	1	
	R	$\sim 1.36\ \sigma^2/n$	~ 0.368	
	d	$\sigma^2/[(\pi-2)\ n]$	~ 0.876	

*) jde o vychýlené odhady





Vztah mezi módem $\hat{x}_M$, průměrem $\bar{x}$ a mediánem $\tilde{x}_{0.5}$ pro sešikmená beta rozdělení (a) k nižším, (b) k vyšším hodnotám

4. **Výběrový medián**, $\tilde{x}_{0.5}$, dělí výběr na dvě části, každá obsahuje 50% prvků. Pro prvky setříděné vzestupně $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ (pořádkové statistiky), je

pro n liché $\tilde{x}_{0.5} = x_{(k)}$, kde $k = (n + 1)/2$,

a **pro n sudé** $\tilde{x}_{0.5} = [x_{(k)} + x_{(k+1)}]/2$, kde $k = n/2$.

5. **Kvantilové odhady** jsou robustnější, méně citlivé na vybočující hodnoty než je tomu u momentových odhadů.

6. Pro rovnoměrné rozdělení dat je vydatným odhadem polohy **výběrová polosuma** $\hat{x}_P$ definovaná vztahem

$$\hat{x}_P = \frac{x_{(1)} + x_{(n)}}{2}$$

kde $x_{(1)}$ je minimální a $x_{(n)}$ maximální prvek ve výběru.

Rozptyl odhadu polosumy je $D_R(\hat{x}_P) = \dfrac{6\,\sigma^2}{(n-1)\,(n-2)}$

7. U řady měřicích přístrojů se měří za podmínek konstantní **relativní chyby měření** .Pro rozptyl měření platí, že

$$\sigma_x^2 = x_i^2\,\delta^2 .$$

8. Má-li každé měření x_i normální rozdělení s hustotou pravděpodobnosti a rozptylem $\sigma^2(x_i)$, lze pro odhad střední hodnoty odvodit vztah tzv. **vážený aritmetický průměr** s vahami $1/\sigma^2(x_i)$,

$$\hat{x}_w = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma^2(x_i)}}{\sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}}$$

Rozptyl tohoto odhadu má tvar

$$D(\bar{x}_w) = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}}$$

9. Když se provádí měření za podmínek konstantní relativní chyby δ, má **vážený odhad střední hodnoty** tvar

$$\bar{x}_w = \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{\sum_{i=1}^{n} \frac{1}{x_i^2}}$$

a pro rozptyl tohoto odhadu platí $D(\bar{x}_w) = \dfrac{\delta^2}{\sum_{i=1}^{n} \frac{1}{x_i}}$

10. Pro odhad variability je možné užít kvantilové odhady, např. **interkvartilové rozpětí**

$$R = 0.7413\ (\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

kde $\tilde{x}_{0.75}$ je odhad horního a $\tilde{x}_{0.25}$ odhad dolního kvartilu.

11. **Průměrná absolutní odchylka d**

$$d = \sqrt{\frac{\pi}{2}} \; \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|$$

faktor $\sqrt{\pi/2}$ zajišťuje pro normální rozdělení asymptoticky přiblížení k směrodatné odchylce σ.

12. Pro relativní rozptýlení dat se užívá relativní směrodatná odchylka nebo-li **variační koeficient** $\delta = \sigma/\mu$ a odhad je

$$\hat{\delta} = \frac{s}{\bar{x}}$$

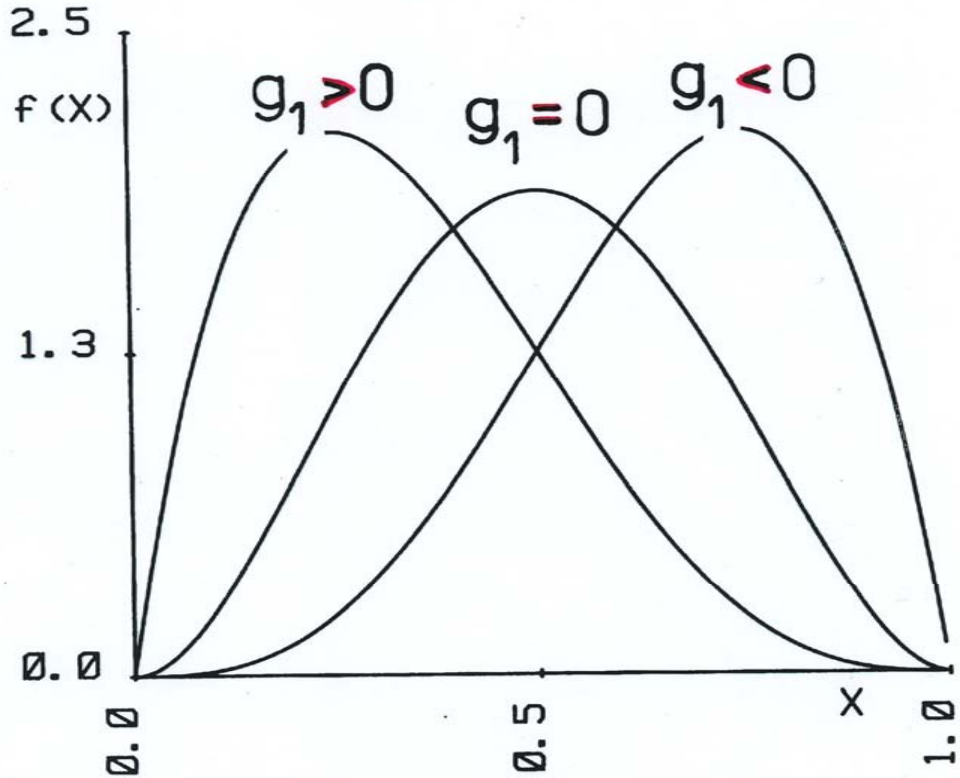
Pro rozptyl tohoto odhadu platí $D(\delta) \approx \delta^2 \dfrac{n + \delta^2 (2n + 1)}{2n(n-1)}$.

13. Momentový odhad ***šikmosti*** g_1 je prvním parametrem tvaru rozdělení

$$\hat{g}_1 = \frac{\sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^3\right]^{3/2}}$$

a střední hodnota pro výběry z normálního rozdělení je rovna nule, $E(\hat{g}_1) = 0$, a pro asymptotický rozptyl odhadu platí

$$D(\hat{g}_1) = \frac{(n - 2)}{(n + 1)(n + 3)}$$



14. Momentový odhad ***špičatosti*** g_2 je druhým parametrem tvaru rozdělení

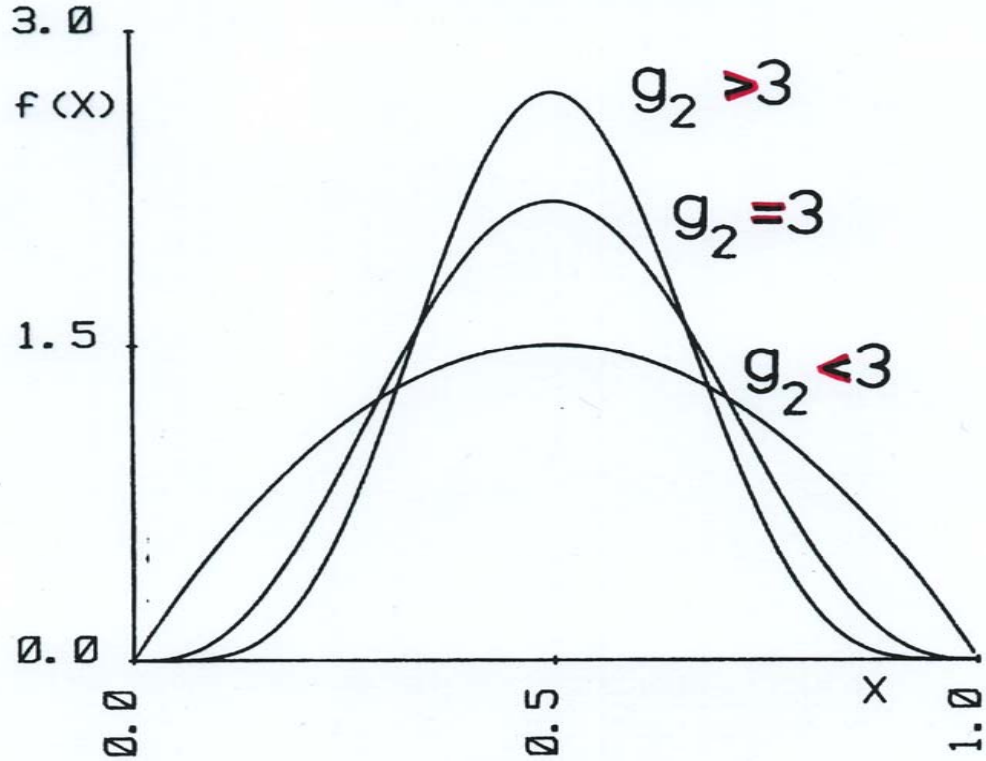
$$\hat{g}_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}$$

a střední hodnota pro výběry z normálního rozdělení je

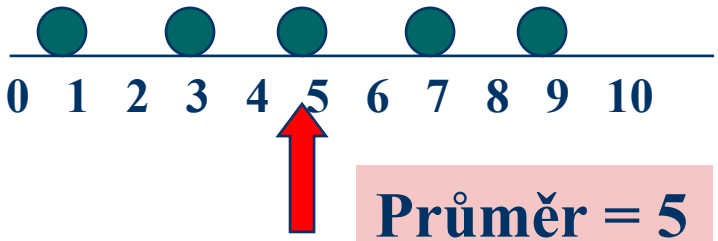
$$E(\hat{g}_2) = 3 - \frac{6}{n+1}$$

a pro asymptotický rozptyl tohoto odhadu platí

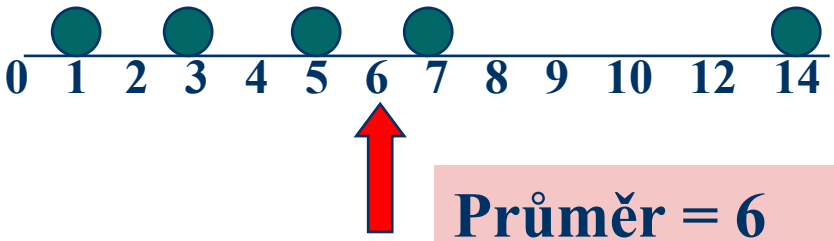
$$D(\hat{g}_2) \approx \frac{24\, n\, (n-2)\, (n-3)}{(n+1)^2\, (n+3)\, (n+5)}$$



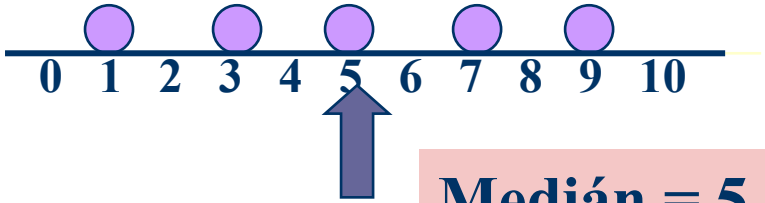
Ukázky příkladů určení centrální tendence



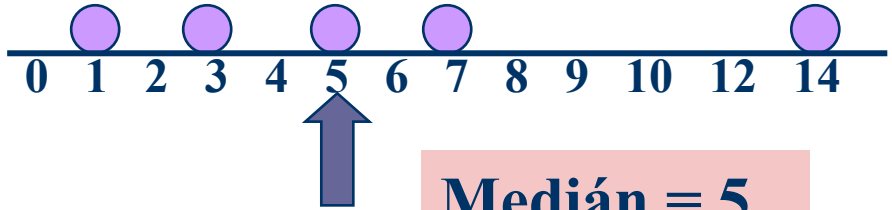
Průměr = 5



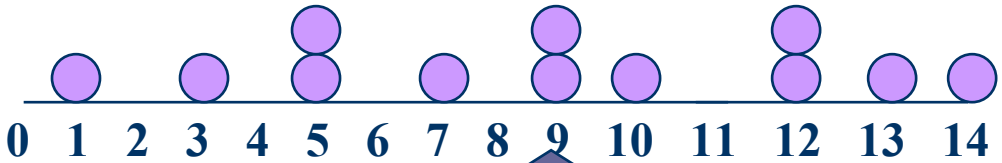
Průměr = 6



Medián = 5



Medián = 5

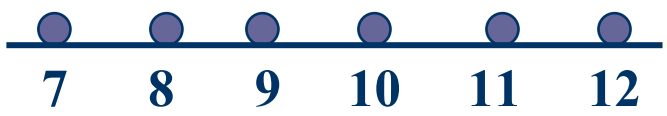


Módus = 9

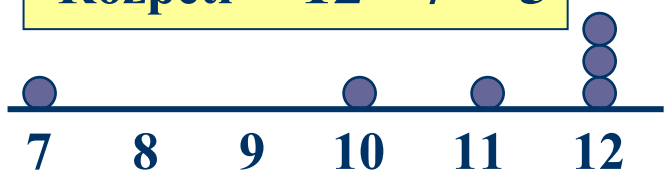
Není módus

Ukázky příkladů rozličné variability (rozptýlení)

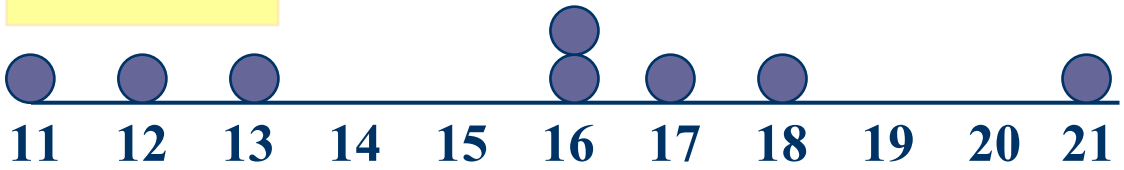
Rozpětí = 12 - 7 = 5



Rozpětí = 12 - 7 = 5



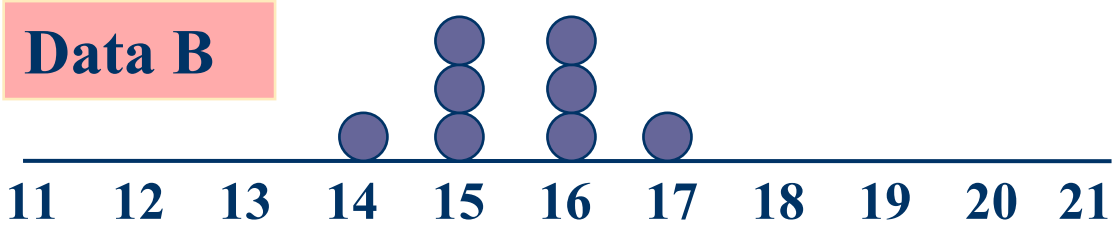
Data A



Průměr = 15.5

s = 3.338

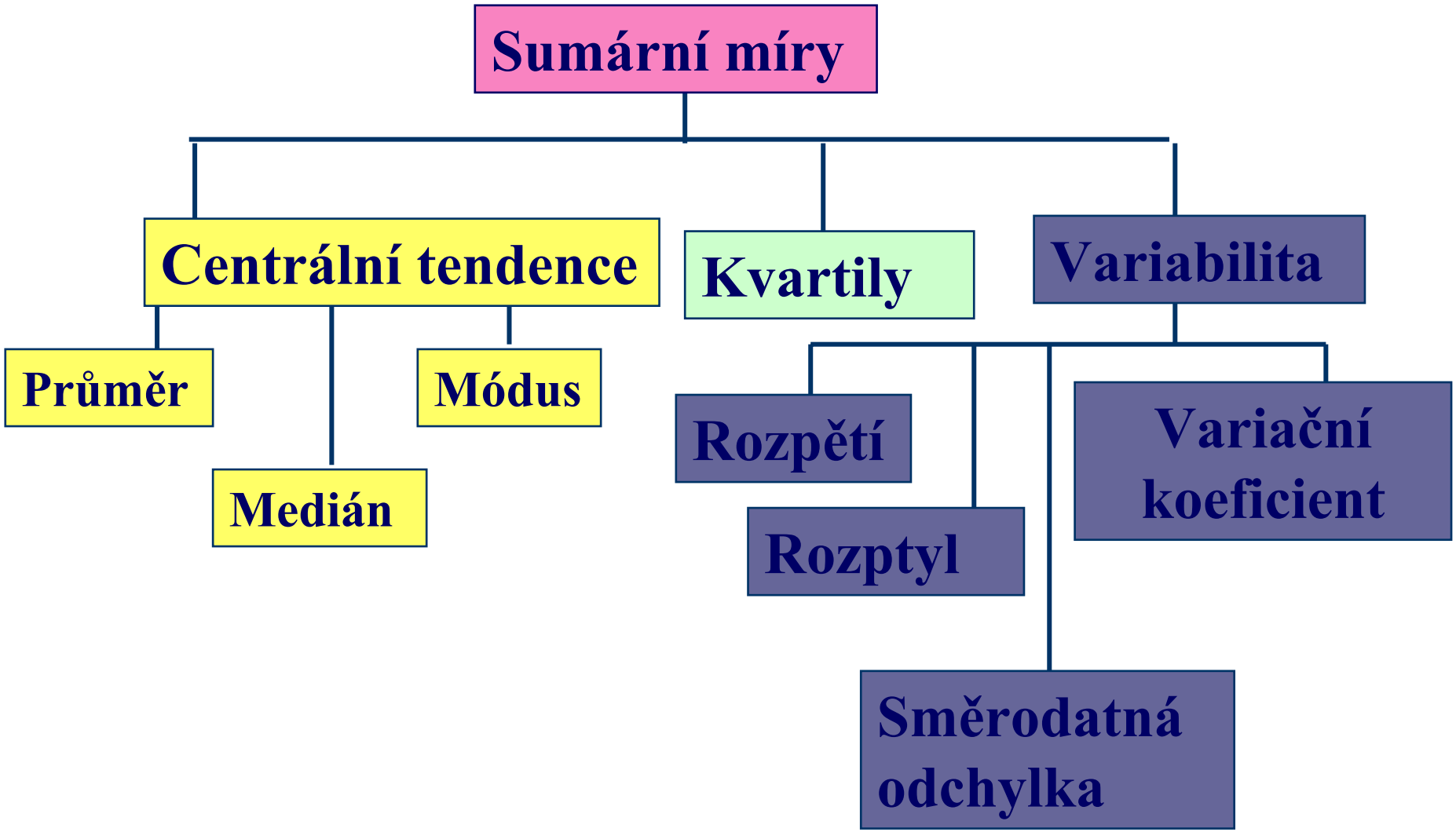
Data B



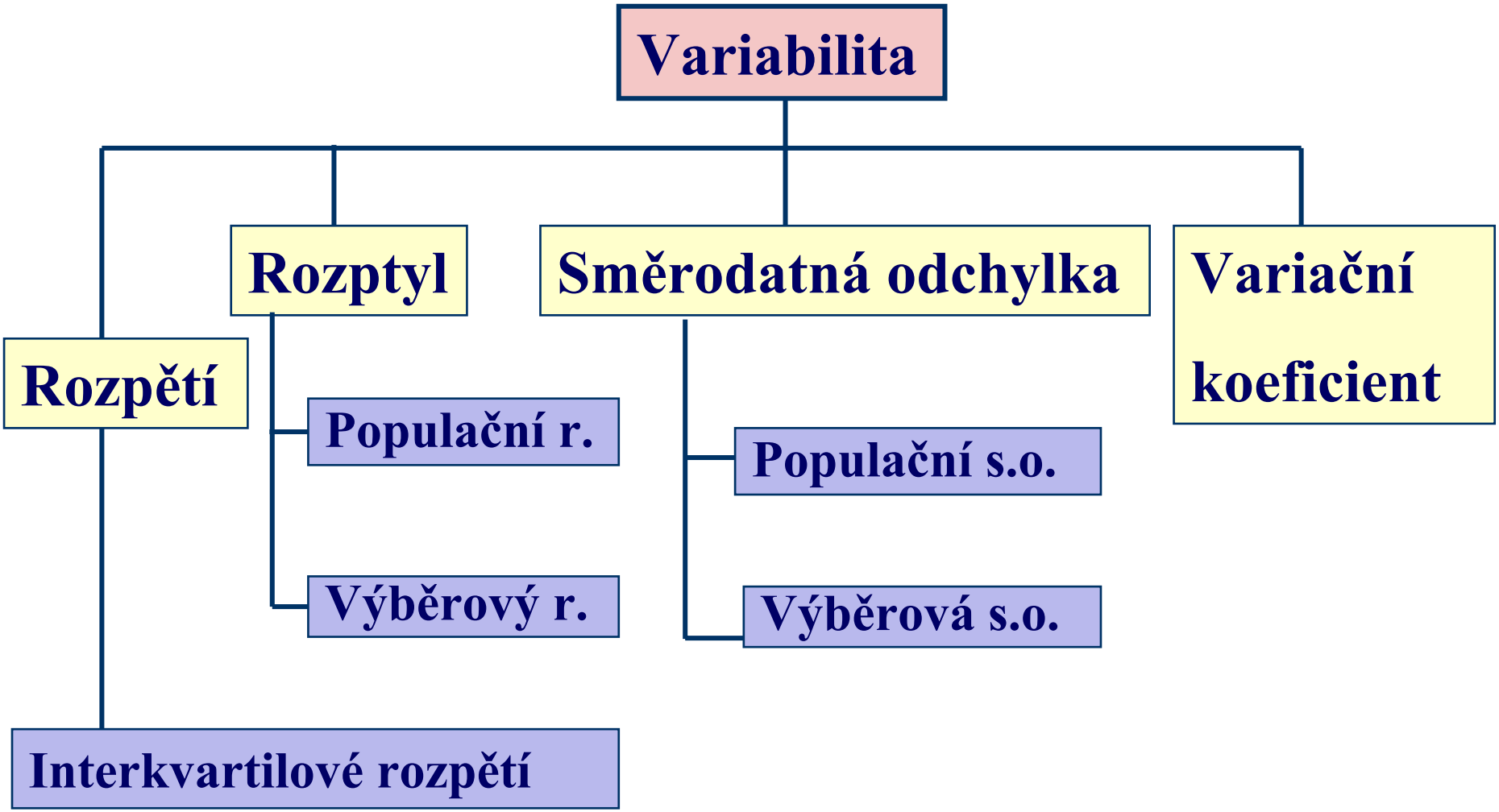
Průměr = 15.5

s = .9258

Přehled sumarizace dat



Přehled měr variability (rozptýlení)



Popisné statistiky polohy, rozptýlení a tvaru

Parametry polohy:

Centrální tendence

- Průměr aritmetický: ($\Sigma x/n$)
- Medián: Centrální bod (dělicí bod)

Parametry rozptýlení:

Variace

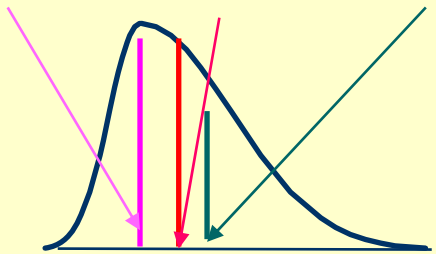
- směrodatná odchylka

Parametry tvaru:

- šikmost – náklon rozdělení
- špičatost – délka konců

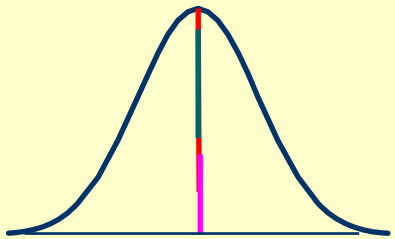
Zešikmené vpravo

Mode < Median < Mean



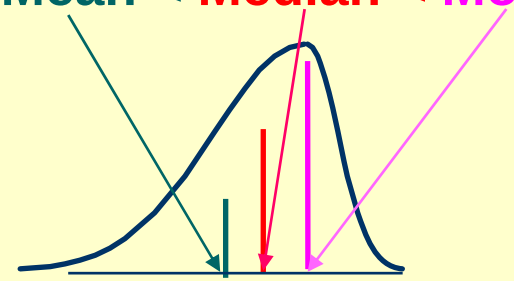
Symetrické

Mean = Median =Mode



Zešikmené vlevo

Mean < Median < Mode



Pokud x_i nemají normální rozdělení $D(s) \approx \frac{\sigma^2 (g_2 - 1)}{4N}$

Parametry rozptýlení - směrodatná odchylka

- **Směrodatná odchylka s je vychýleným odhadem veličiny σ.**
- **Platí, že E(s) < σ. Pro nevychýlený odhad lze odvodit**

$$\hat{\sigma} = K_u * s \approx \sqrt{\frac{s^2 (N-1)}{N - 1.45}}$$

kde $$K_u = \Gamma\left(\frac{N-1}{2}\right)\sqrt{\frac{N-1}{2}} \Big/ \Gamma\left(\frac{N}{2}\right) \approx \sqrt{\frac{N-1}{N}}\left(1 - \frac{3}{4N} - \frac{37}{32N^2} \cdots\right)$$

$$D(\hat{\sigma}) = \sigma^2 \big/ (2N)$$

$$D(s) \approx \frac{\sigma^2}{2(N-1)}$$

Parametry rozptýlení - variační koeficient

- Výběrový odhad V a odpovídající rozptyl D(V) lze vyjádřit ve tvaru

$$V = \frac{\bar{x}}{s}$$

$$\mathrm{D(V)} \approx \delta \left[\frac{N + (2N-1)\delta^2}{2N(n-1)} \right] \propto \frac{\delta^2}{2N}(1+2\delta^2)$$

Parametry tvaru - šikmost a špičatost

Odhady jsou pouze asymptotické (platí pro velké N).

- Pro odhad šikmosti g_1 se používá výběrová šikmost $\hat{g}_1$ s rozptylem $D(\hat{g}_1)$, kde

$$\hat{g}_1 = \frac{\sqrt{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{3/2}} \qquad D(\hat{g}_1) \approx \frac{6N(N-1)}{(N-2)(N+1)(N+3)}$$

- Pro odhad špičatosti g_2 platí

$$\hat{g}_2 = \frac{N\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{2}} \qquad D(\hat{g}_2) \approx \frac{24N(N-1)^2}{(N-3)(N-2)(N+5)(N+3)}$$

Kdy použít mediánu?

Aritmetický průměr $\overline{X}$ je efektivní (má minimální rozptyl) pouze pro normální rozdělení.

Pro případ Laplaceova rozdělení je efektivní výběrový **medián** $\tilde{x}_{0.5}$ pro jehož rozptyl platí

$$D(\tilde{x}_{0.5}) = \frac{1}{4Nf^2(M)}$$

kde $f^2(M)$ je hodnota hustoty pravděpodobnosti v místě teoretického mediánu M.

Pro případ Laplaceova rozdělení je $f(M) = 1/(\sigma\sqrt{2})$ a tedy

$$D(\tilde{x}_{0.5}) = \sigma^2/(2N)$$

- Pro Laplaceovo rozdělení je poměr

$$E_M = D(\bar{x}) \big/ D(\tilde{x}_{0.5}) = 2$$

tj. medián je 2x efektivnější než aritmetický průměr.

- Pro třídu exponenciálních rozdělení s parametrem α lze obecně určit E_M ze vztahu

$$E_M = \alpha^2 \Gamma(3/\alpha) \big/ \Gamma^3(1/\alpha)$$

Z této rovnice plyne, že pro rozdělení $\alpha > 2$ (dlouhé konce) je medián efektivnější než aritmetický průměr

Kdy použít polosumy?

Polosuma je efektivnější než $\bar{x}$ pro $g_2 > 2.2$.

- Pro rozdělení s plochými vrcholy se doporučuje použití **kvartilové polosumy**

$$P_F = (\tilde{x}_{0.75} - \tilde{x}_{0.25})/2$$

kde $\tilde{x}_{0.75}$ resp. $\tilde{x}_{0.25}$ je horní, resp. dolní kvartil

- V případě ohraničených rozdělení (arkussínové a lichoběžníkové třídy) je efektivní tzv. **polosuma**

$$\hat{x}_P = (x_{max} - x_{min})/2$$

kde x_{max} je maximální a x_{min} je minimální prvek výběru

Rozptyl odhadu polosumy je pro **normální rozdělení** roven

$$D(\hat{x}_P) = \frac{(\pi^2\sigma^2)}{(24\ln(N))}$$

Pro **rovnoměrné rozdělení** je

$$D(\hat{x}_P) = \frac{(6\sigma^2)}{[(N-1)(N-2)]}$$

Arkussínové rozdělení je definováno v intervalu (-A, A) a pro jeho hustotu pravděpodobnosti platí $f(x) = \frac{1}{\pi\sqrt{A^2 - x^2}}$

Pro **arkussínové rozdělení**

$$D(\hat{x}_P) = \frac{(5\pi^4\sigma^2)}{N^4}$$

Uřezaný průměr

- α-uřezaný průměr $\bar{x}(\alpha)$ je definován vztahem

$$\bar{x}(\alpha) = \frac{1}{n - 2M} \sum_{i=M+1}^{N-M} x_{(i)}$$

kde M = int(αN/100) je celá část výrazu α N/100 a $x_{(i)}$ jsou pořádkové statistiky (vzestupně setříděné prvky výběru).

Kombinovaný odhad centrální hodnoty

- Pro symetrická rozdělení s vybočujícími hodnotami je doporučen za odhad středu symetrie čili centrální hodnoty použít **medián** dle vzorce

$$\tilde{x}_C = \mathbf{med}\{\bar{x}\ ,\tilde{x}_{0.5}\ ,\ \hat{x}_P\ ,\ \mathbf{PF},\ \bar{x}(0.25)\}$$

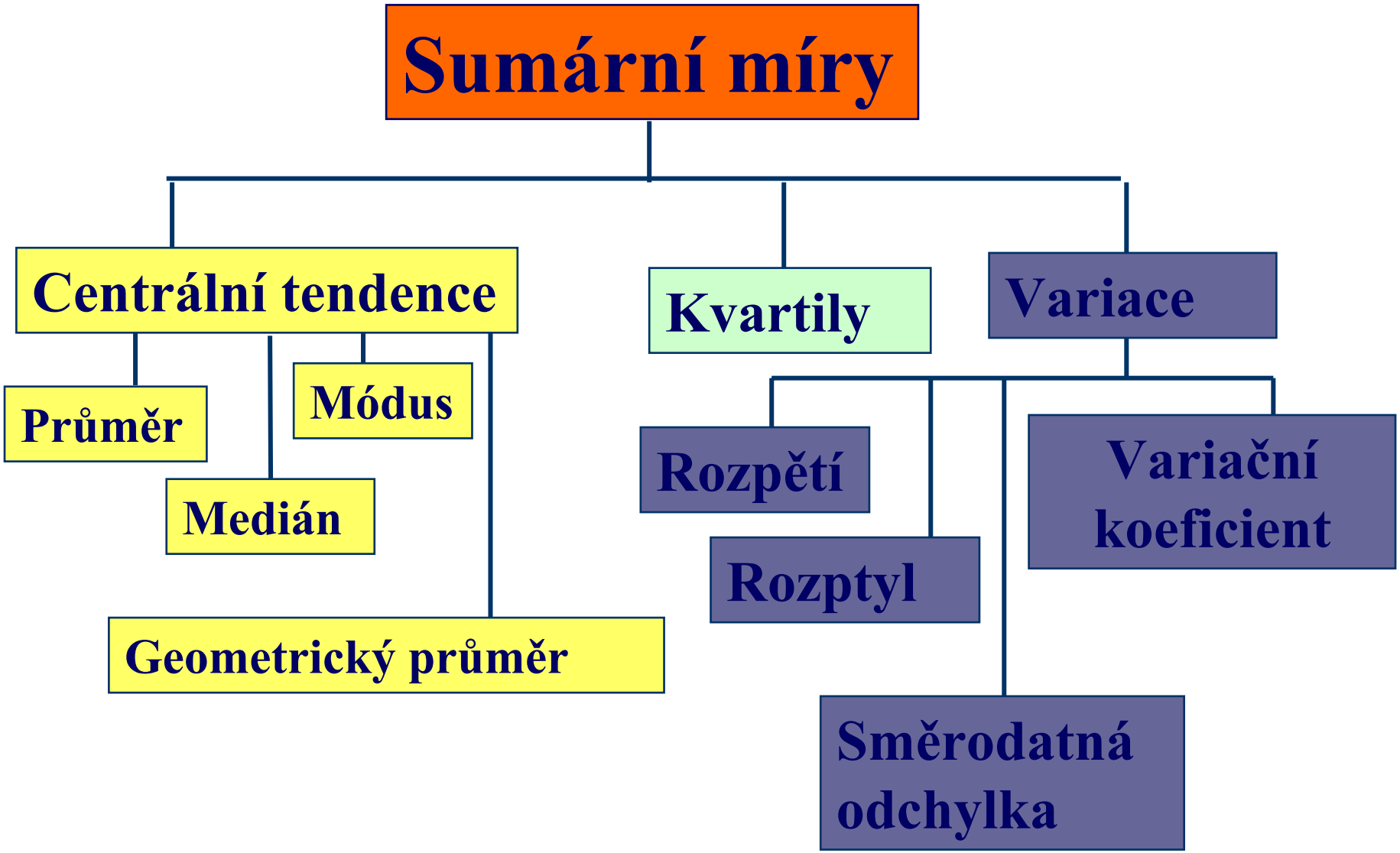
kde med{.} označuje medián z prvků v závorce.

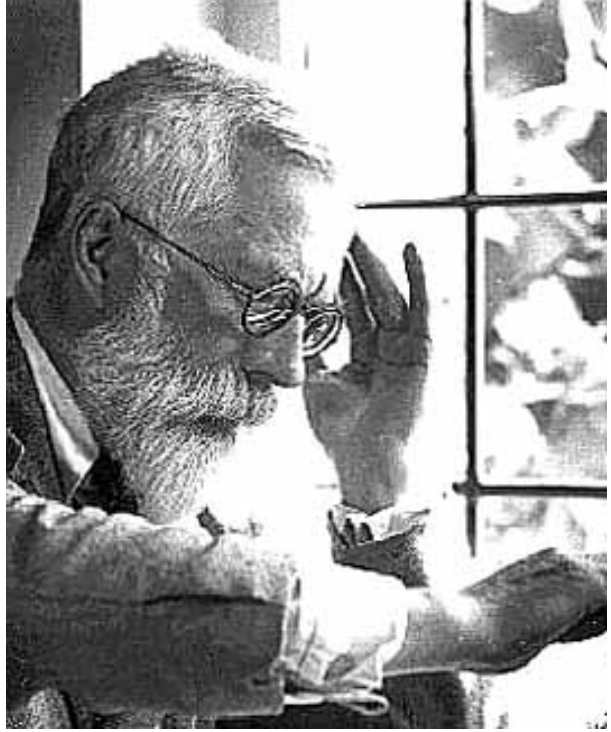
Pro odhad rozptylu odhadu $\tilde{x}_C$ je možno použít interkvantilové délky

$$k_{0.9} = \left(\tilde{x}_{0.95} - \tilde{x}_{0.05}\right)/2$$

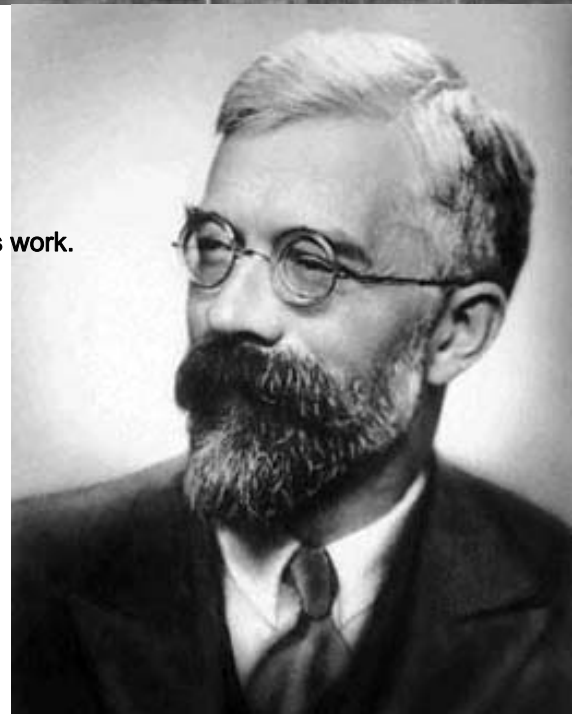
$$D(\tilde{x}_C) = k_{0.9}^2 \big/ (2.72N)$$

SUMARIZACE DAT





Fisher, Sir Ronald Aylmer, 1890-1962



Sir Ronald Fisher F.R.S. (1890-1962) was one of the leading scientists of the 20th century; making major contributions to Statistics, Evolutionary Biology and Genetics. This website has information about him and his work.

"perhaps the most original mathematical scientist of the [twentieth] century"
Bradley Efron *Annals of Statistics* (1976)

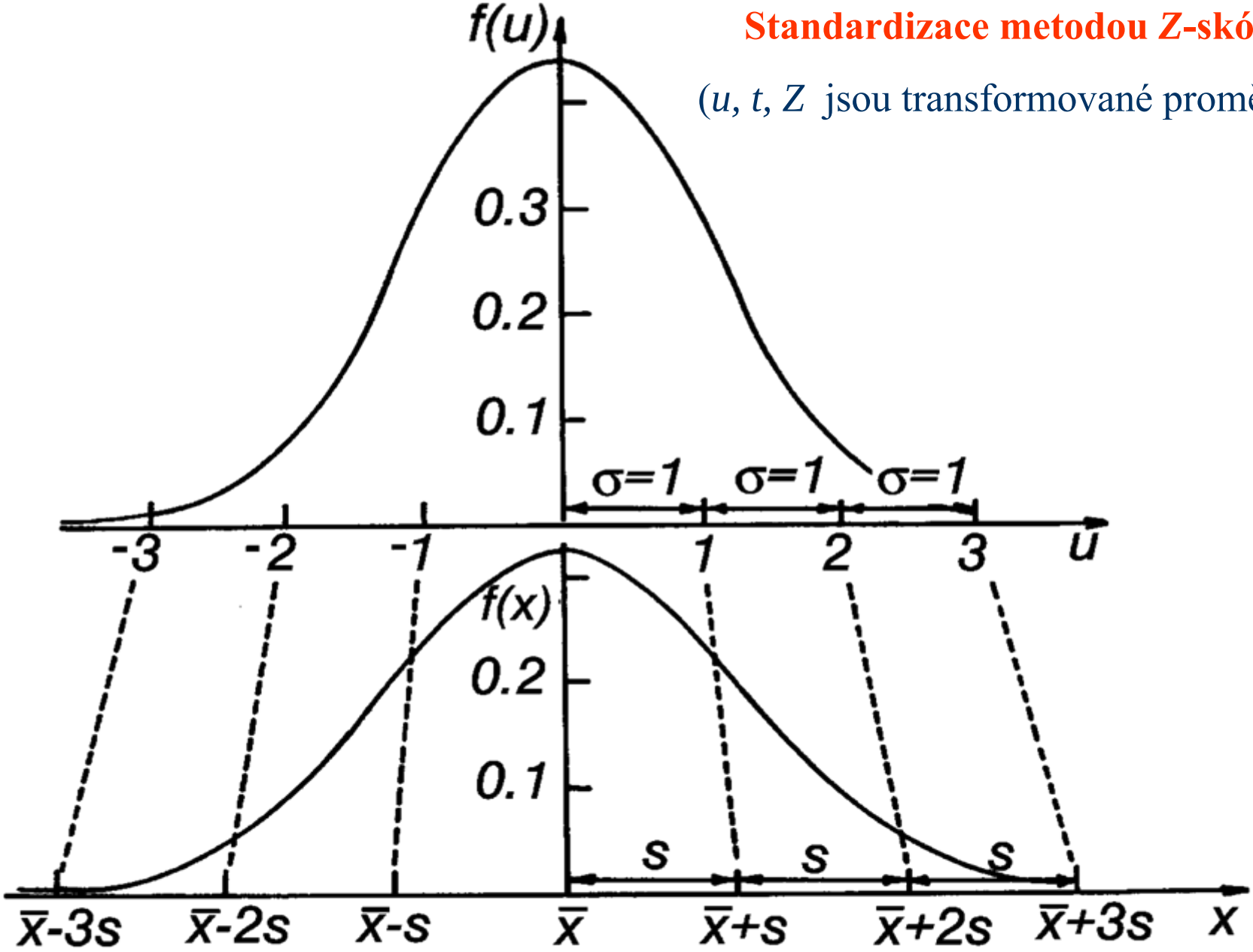
"Fisher was a genius who almost single-handedly created the foundations for modern statistical science …."
Anders Hald *A History of Mathematical Statistics* (1998)

"Sir Ronald Fisher … could be regarded as Darwin's greatest twentieth-century successor."
Richard Dawkins *River out of Eden* (1995)

http://www.library.adelaide.edu.au/ual/special/fisher.html

Standardizace metodou Z-skóre

(u, t, Z jsou transformované proměnné)



'Student' in 1908

Gosset, William Sealy ("Student"), 1876-1937

The probable error of a mean [Paper on the t-test], *Biometrika* **6** (1908), pp. 1-25

DIAGRAM I. Frequency Curve giving the Distribution of Standard Deviations of samples of 10 taken from a Normal Population

Equation $y = \frac{N}{7.5.3} \frac{10^{\frac{9}{2}}}{\sigma^9} \sqrt{\left(\frac{2}{\pi}\right)} x^8 e^{-\frac{10x^2}{2\sigma^2}}$

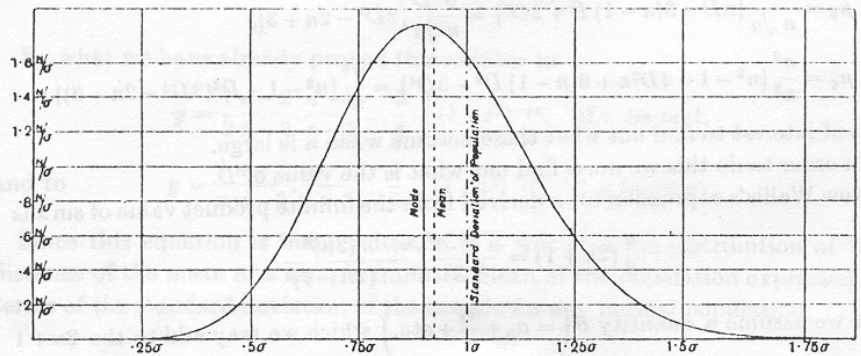
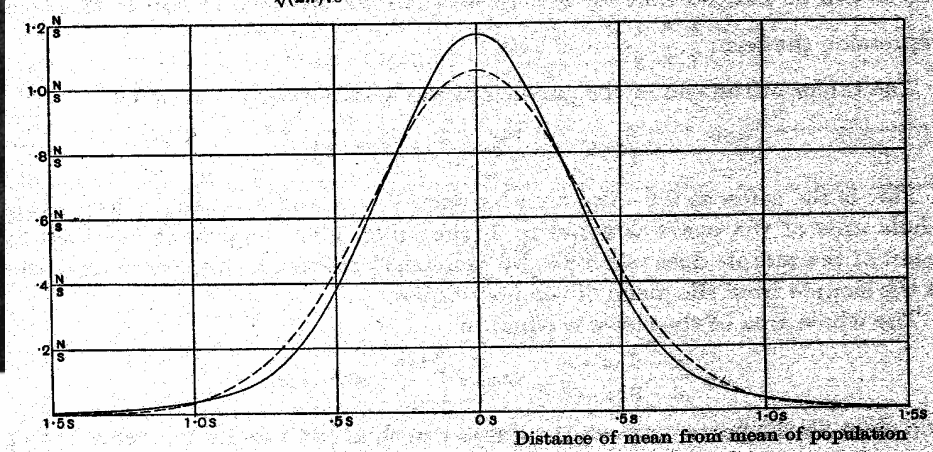


DIAGRAM II. Solid curve $y = \frac{N}{s} \times \frac{8}{7} \cdot \frac{6}{5} \cdot \frac{4}{3} \cdot \frac{2}{\pi} \cos^{10} \theta, \quad x/s = \tan \theta$

Broken line curve $y = \frac{\sqrt{7}.N}{\sqrt{(2\pi)}.s} e^{-\frac{7x^2}{2s^2}}$, the normal curve with the same standard deviation



Odhady parametrů

Rozdělení měření pro oba modely obsahuje střední hodnotu μ a rozptyl σ^2 a odhady získáme:

1) Momentová metoda: pro normální rozdělení: odhadem je aritmetický průměr x_A a výběrový rozptyl s^2.

2) Metoda maximální věrohodnosti získává odhad maximalizující logaritmus věrohodnostní funkce L, tj. $\ln(L) = (-N/2)[\ln(2\pi) + \ln(\sigma^2)] - \dfrac{\sum_i (x_i - \mu)^2}{2\sigma^2}$.

Pro první derivace logaritmu věrohodnostní funkce pak platí

$$\frac{\partial \ln(L)}{\partial \mu} = 2\sum_i x_i - 2N\mu ,$$

$$\frac{\partial \ln(L)}{\partial \sigma^2} = \frac{\sum_i (x_i - \mu)^2}{[\sigma^2 \sqrt{2}]^2} - \frac{N}{2\sigma^2}$$

maximálně věrohodné odhady střední hodnoty a rozptylu jsou totožné s průměrem a výběrovým rozptylem.

1) **Stanovení typu rozdělení:** pro výpočet $F_t^{-1}(P_i)$ je třeba znát obecně parametry teoretické rozdělení. V řadě případů je však možná standardizace $s = (x - Q)/R$, kde R je parame rozptýlení.

Standardizované kvantilové funkce $Q_s(P_i) = Fs_t^{-1}(P_i)$ obsahují jen tvarové faktory.

V případě shody obou rozdělení pak resultuje přímková závislost

$$x_{(i)} = Q + R \,.\, Q_S(P_i) = a + b \,.\, Q_S(P_i) \,.$$

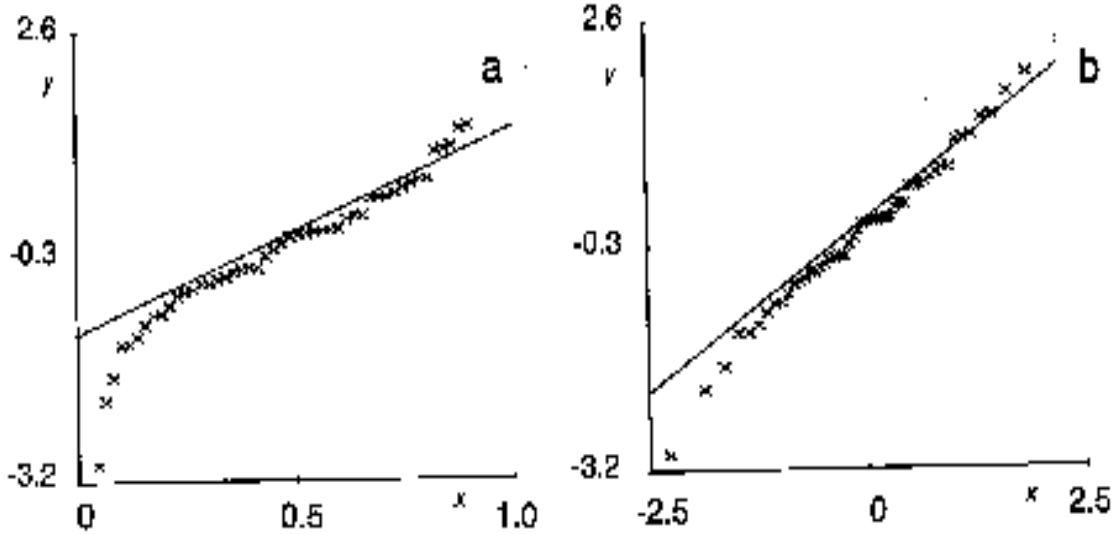
2) **Odhady parametrů polohy a rozptýlení:** odhad střední hodnoty odpovídá absolutnímu členu a odhad směrodatné odchylky směrnicí b regresní přímky.

Pro odhad parametrů z Q-Q grafů je možno použít bud MNČ.

Transformací $z = \ln[(1+\sigma * Y)^{1/\sigma}]$ má veličina z normované normální rozdělení N(0,1).

Pořádková statistika $x_{(i)}$ pak souvisí s pořádkovou statistikou normovaného normálního rozdělení $z_{(i)}$ dle

$$x_{(i)} = \mu + \tau \left[\frac{\exp(\sigma z_{(i)}) - 1}{\sigma} \right] \approx \mu + \tau g_i(\sigma)$$



V řadě případů je však možná standardizace

$$s = (x - Q) / R \,,$$ kde R je parametr rozptýlení.

V případě shody obou rozdělení pak resultuje přímková závislost

$$x_{(i)} = Q + R \,.\, Q_S(P_i) = a + b \,.\, Q_S(P_i) \,.$$

Odhad střední hodnoty odpovídá absolutnímu členu a
odhad směrodatné odchylky směrnicí b regresní přímky.

Pro odhad parametrů z Q-Q grafů je možno použít bud MNČ.

Cenzorované výběry

Pro odhady parametrů v cenzorovaných výběrech lze použít jak metodu maximální věrohodnosti, tak i metody založené a pořádkových statistikách.

Cenzorování typu I: známe limitu detekce x_L (mez pod kterou se zaznamenává pouze, Přítomnost měření) a předpokládejme, známe rozdělení dat charakterizované hustotou pravděpodobnosti $f(x)$, resp. distribuční funkcí $F(x)$.

Pro cenzorovaná měření lze při znalosti distribuční funkce měření určit pouze pravděpodobnost s jakou leží pod mezí detekce, která je rovna $F(x_L)$.

Všechny možné kombinace n_1 prvků, které ve výběru velikosti N leží pod limitou detekce jsou dány binomickým koeficientem $N!/(n_1!\ (N - n_1)!)$.

Věrohodností funkce má pro tento případ tvar

$$\ln(L) = \frac{N!}{n_1!*(N-n_1)!} F(X_L)^{n_1} * \prod_{i=n_1+1}^{N} f(x_{(i)}) .$$

Data jsou	originální	cenzor.	uřezaná	winsoriz.
	4.8			6.1
	5.2			6.1
	5.4			6.1
Detekční limit = 6.0	5.6			6.1
	6.1	6.1	6.1	6.1
	6.3	6.3	6.3	6.3
	6.5	6.5	6.5	6.5
	6.7	6.7	6.7	6.7
	6.9	6.9	6.9	6.9
	7.2	7.2	7.2	7.2
	7.3	7.3	7.3	7.3
	7.4	7.4	7.4	7.4
	7.5	7.5	7.5	7.5
	7.6	7.6	7.6	7.6
	7.7	7.7	7.7	7.7
	7.8	7.8	7.8	7.8
	7.9	7.9	7.9	7.9
	8.0	8.0	8.0	8.0
	8.1	8.1	8.1	8.1
	8.3	8.3	8.3	8.3
	8.5	8.5	8.5	8.5
	8.7	8.7	8.7	8.7
	8.9	8.9	8.9	8.9
	9.2	9.2		8.9
	9.4	9.4		8.9
	9.6	9.6		8.9
	10.1	10.1		8.9

Pro případ normálního rozdělení dat nalezl Cohen vztahy pro odhad střední hodnoty $\overline{x}_C$ a rozptylu odpovídající maximalizaci věrohodnostní funkce s využitím odhadů z necenzorované části dat

$$\overline{x}_N = \frac{1}{N - n_1} \sum_{i=n_1+1}^{N} x_{(i)} ,$$

$$s_N^2 = \frac{1}{N - n_1 - 1} \sum_{i=n_1+1}^{N} (x_{(i)} - \overline{x}_N)^2 .$$

Platí, že

$$\overline{x}_C = \overline{x}_N - \lambda * (\overline{x}_N - x_L) ,$$

$$s_C^2 = s_N^2 + \lambda * (\overline{x}_N - x_L)^2$$

Parametr λ závisí:

1) na odhadnutém podílu cenzorovaných dat $h = n_1 / N$ a

2) na parametru $g = s_N^2 / (\overline{x}_N - x_L)^2$.

Hodnoty λ jsou tabelovány a existují také empirické vztahy.

Postačuje **jednokrokový odhad** založený na předpokladu, že počet hodnot pod limitou detekce má binomické rozdělení a pro **odhad střední hodnoty** $\overline{x}_{CJ}$ a **rozptylu** s^2_{CJ} pak platí

$$\overline{x}_{CJ} = \overline{x}_N - q * s_N$$

$$s^2_{CJ} = \frac{\sum_{i=n_1+1}^{N} x^2_{(i)}}{N - n_1} - (\overline{x}_N)^2 - s^2_N * (q * \Phi^{-1}(h) - q^2)$$

Korekční faktor q má tvar

$$q = \frac{N}{(N - n_1)\sqrt{2\pi}} \exp(-0.5 * [\Phi^{-1}(h)]^2) \quad .$$

Odhady x_{CJ} a s^2_{CJ} lze tedy určit relativně snadno bez nutnosti použití speciálních tabulek.

Pro dvou-parametrové logaritmicko-normální rozdělení stačí místo hodnot $x_{(i)}$ použít jejich logaritmů $ln\ (x_{(i)})$ a logaritmovat i limitu detekce.

Počet =	27	23	19	27
Průměr :	7.51	7.90	7.55	7.53
L_d	6.96	7.43	7.16	7.13
L_h	8.05	8.37	7.93	7.94
Směr. odchylka :	1.383	1.076	0.797	1.022
Medián :	7.60	7.80	7.60	7.60
L_d	6.76	7.01	6.74	6.76
L_h	8.44	8.59	8.46	8.44
Medianová směr. odchylka :	0.408	0.383	0.408	0.408
10% Průměr :	7.52	7.87	7.55	7.54
L_d	6.92	7.36	7.14	7.06
L_h	8.12	8.39	7.97	8.02
10% Směr. odchylka :	1.000	0.737	0.646	0.861
40% Průměr :	7.56	7.85	7.58	7.56
L_d	6.96	7.34	7.14	6.96
L_h	8.16	8.35	8.02	8.16
40% Směr. odchylka :	0.423	0.344	0.274	0.423
Mocn. transformace, průměr:	7.57	7.84	7.58	7.58
Box-Cox transf., průměr:	7.56	7.84	7.58	7.58
Normalita, vypočt. hlad. význam.:	Přijata, 0.88	Přijata, 0.78	Přijata, 0.90	Přijata, 0.94

Praktické doplňky

Při zpracování experimentálních dat záleží na množství informací, které jsou:

I. **Víme vše:** známe pravděpodobnostní model a stačí pouze ověření předpokladů před konfirmativní statistickou analýzou

II. **Nevíme nic:** postavíme datově závislý pravděpodobnostní model a provede se komplex analýza dat (1. EDA průzkumová, 2. Ověření předpokladů, 3. Transformace, 4. Porovnání výběrovéh rozdělení s teoretickými).

III. **Něco víme:** postavíme empirický model se známými tak i datově závislými informacen pak se provede 1., 2., 3. a 4. analýza dat)

Doporučené další postupy:

1) Robustní metody,
2) Využití zešikmených rozdělení,
3) Počítačově intenzivní metody,
4) Generalizovaná lineární regrese.