Intervalový odhad parametrů

Intervalový odhad představuje interval, ve kterém se bude se zadanou pravděpodobností či statistickou jistotou $(1 - \alpha)$ nacházet skutečná hodnota ("pravda") daného parametru Θ.

Neznámý parametr Θ odhadujeme dvěma číselnými hodnotami L_1 a L_2, které tvoří **meze** tzv. ***intervalu spolehlivosti*** (čili konfidenčního intervalu).

Interval spolehlivosti pokryje parametr Θ s předem zvolenou, statistickou jistotou čili dostatečně veli-kou pravděpodobností $P = (1 - \alpha)$

$$P(L_1 < \Theta < L_2) = 1 - \alpha$$

nazvanou ***koeficient spolehlivosti*** (čili konfidenční koeficient, statistická jistota). Je obyčejně roven 0.95 nebo 0.99.

Parametr α se nazývá ***hladina významnosti***.

Interval spolehlivosti vyjadřuje tvrzení:

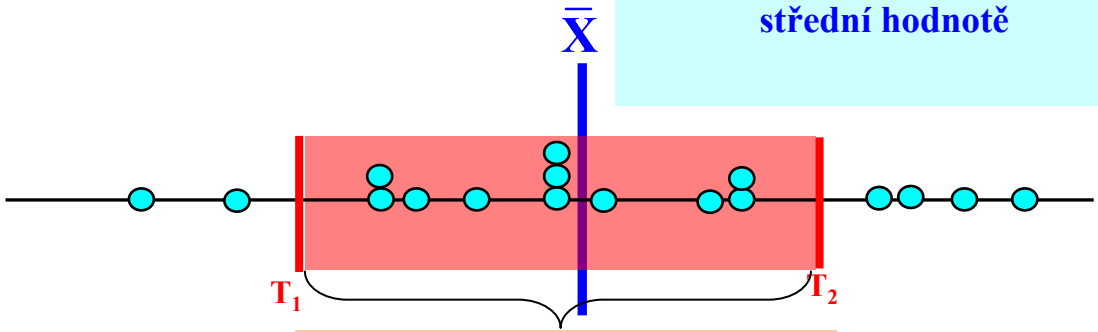
Statistická jistota, s jakou bude "pravda" Θ ležet v náhodných mezích L_1, L_2 je rovna právě $1 - \alpha$.

INTERVALOVÉ ODHADY PARAMETRŮ ZÁKLADNÍHO SOUBORU

Interval spolehlivosti pro parametr τ při **hladině významnosti** $\alpha \in (0,1)$ je určen statistikami T_1 a T_2:.

$$P\left(T_1 \le \tau \le T_2\right) = 1 - \alpha$$

toto je bodový odhad neznámé střední hodnoty μ vypočítaný z prvků výběru – nevíme nic o jeho vztahu ke skutečné střední hodnotě



toto je intervalový odhad neznámé střední hodnoty - předpokládáme, že s pravděpodobností $P = 1-\alpha$ leží μ kdekoli v tomto úseku číselné osy

Konstrukce intervalových odhadů

Postup konstrukce intervalu spolehlivosti střední hodnoty μ normálního rozdělení $N(\mu, \sigma^2)$:

1. Nejlepším bodovým odhadem střední hodnoty μ je výběrový průměr $\bar{x}$ s rozdělením $N(\mu, \sigma^2/n)$, pak v intervalu $\bar{x} \pm 1.96\sigma/\sqrt{n}$ leží přibližně 95% hodnot náhodných veličin výběru o rozsahu n,

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Hodnota 1.96 je totiž $100(1 - 0.05/2) = 97.5\%$ní kvantil normovaného Gaussova normálního rozdělení $u_{0.975}$.

2. **V praxi neznáme** směrodatnou odchylku σ. Jelikož má $\frac{\bar{x} - \mu}{s}\sqrt{n}$ Studentovo t-rozdělení, platí

$$P(-t_{1-\alpha/2}(\nu) \leq \frac{\bar{x} - \mu}{s}\sqrt{n} \leq t_{1-\alpha/2}(\nu)) = 1 - \alpha$$

kde $t_{1-\alpha/2}(\nu)$ je $100(1 - \alpha/2)$%ní kvantil Studentova rozdělení s $\nu = n - 1$ stupni volnosti.

100(1 - α)%ní int. spolehlivosti střední hodnoty μ

bude $$\bar{x} - t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}}$$

Meze int. spol. závisí vedle chyby s i na rozsahu výběru n. Pro větší rozsahy výběru ($n > 30$) lze použít místo kvantilu $t_{1-\alpha/2}$ kvantilu normovaného normálního rozdělení $u_{1-\alpha/2}$.

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

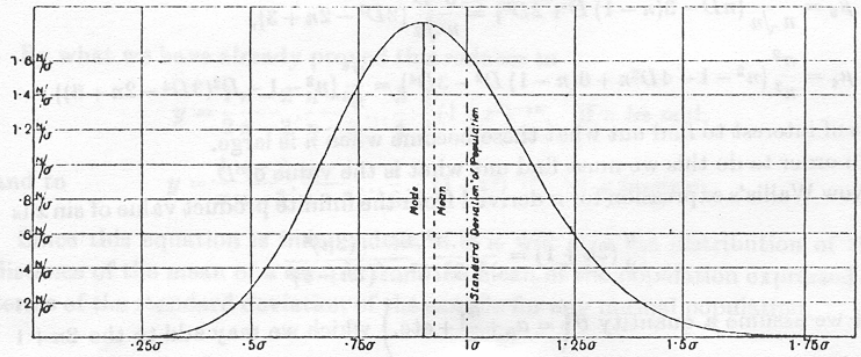
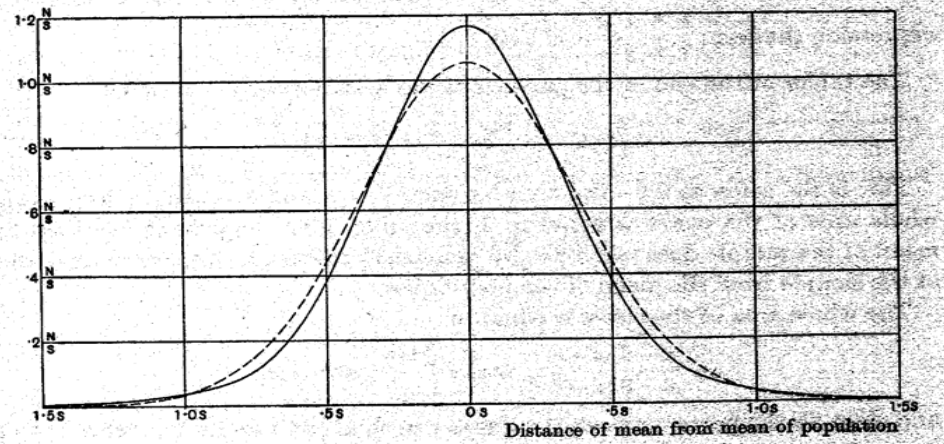
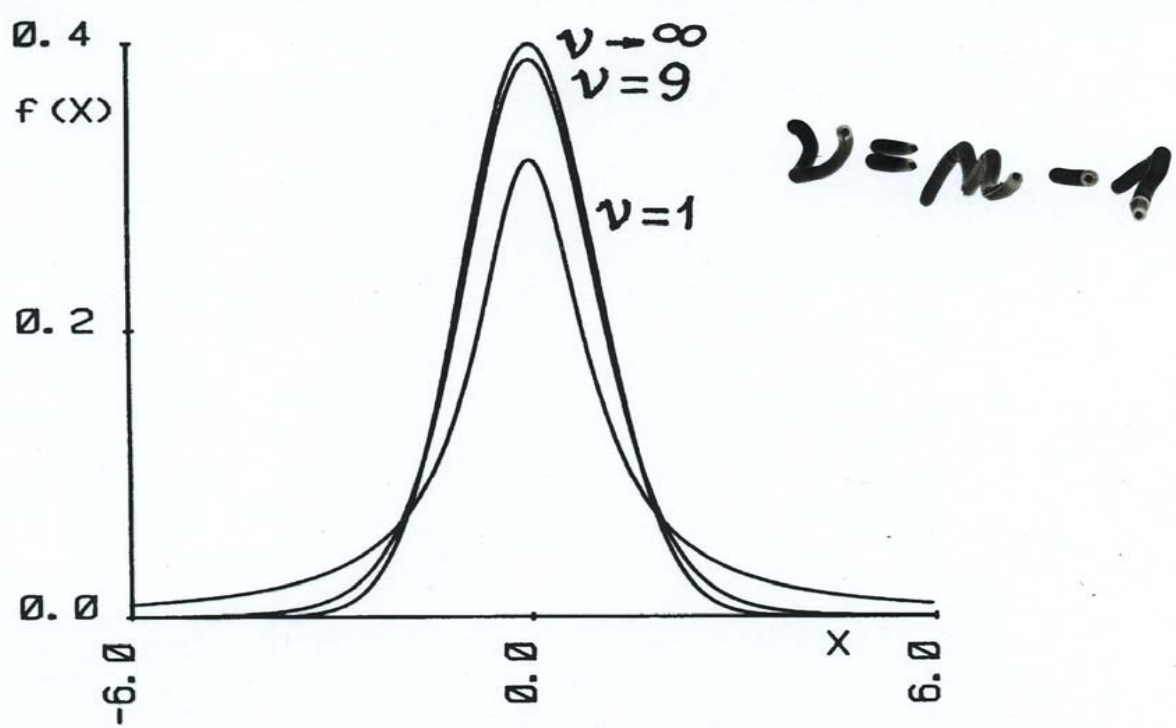


DIAGRAM II. Solid curve $y = \frac{N}{s} \times \frac{8}{7}\cdot\frac{6}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^{10}\theta, \quad x/s = \tan\theta$

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





Studentovo rozdělení pro stupně volnosti ν = 1, ν = 9 a normální rozdělení

Pro výběry pocházející z ***normálního rozdělení*** platí, že náhodná veličina

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

má **Studentovo rozdělení** s (n - 1) stupni volnosti a že náhodná veličina

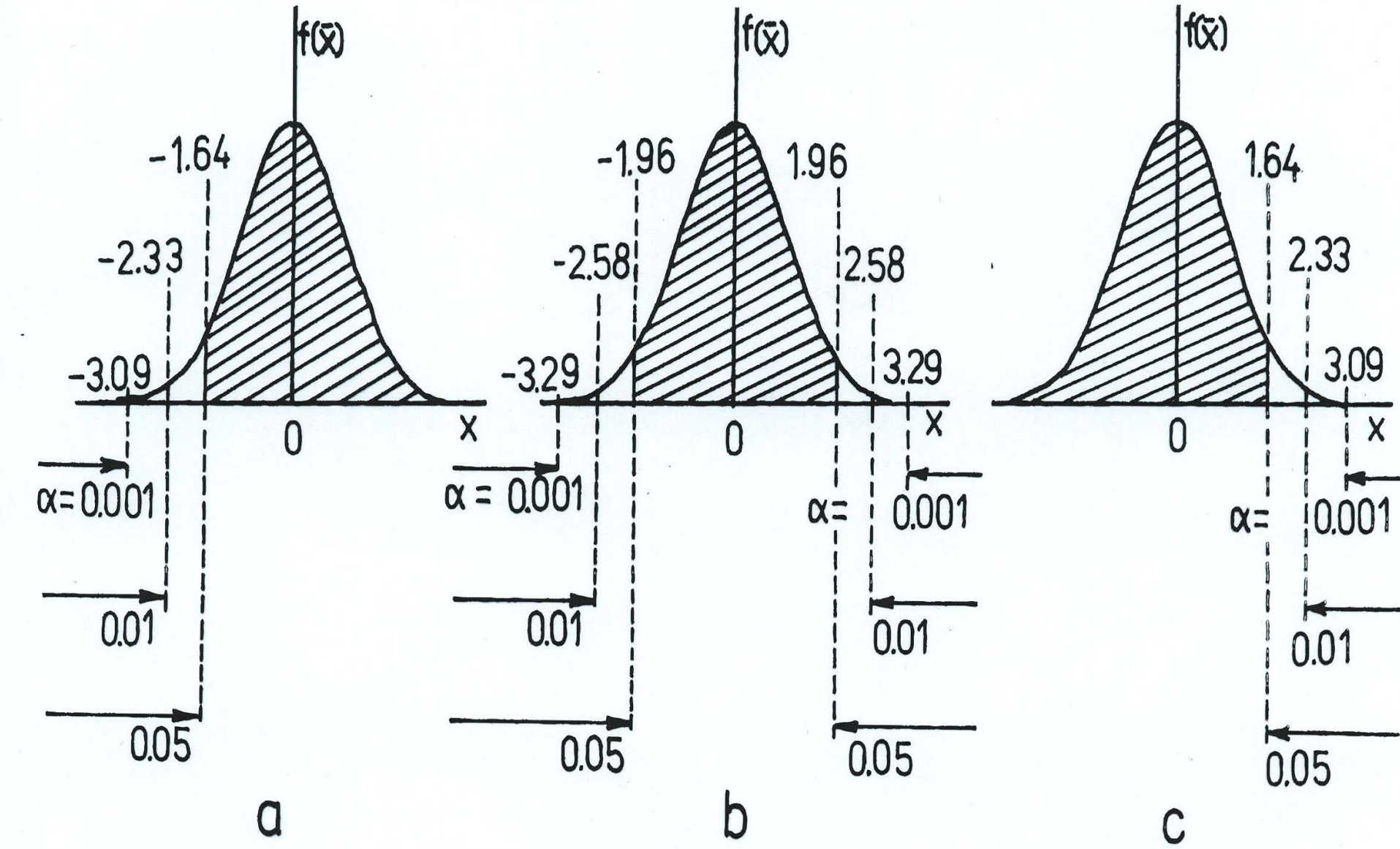
$$\chi^2 = \frac{(n - 1)\, s^2}{\sigma^2}$$

má χ^2 ***- rozdělení*** s (n - 1) stupni volnosti.

Pro dostatečně velký rozsah výběru ($n \geq 40$) lze pro normálně rozdělená původní data $\{x_i\}$, $i = 1, ..., n$, a libovolný odhad $\hat{\Theta}$ veličiny Θ považovat veličinu

$$U = \frac{(\hat{\Theta} - \Theta)}{\sqrt{D(\hat{\Theta})}}$$

za přibližně normálně rozdělenou.



Jednostranný ((a) levostranný a (c) pravostranný) a (b) oboustranný interval spolehlivosti průměru rozdělení N(0, 1) pro hodnoty $\alpha = 0.001$, 0.01 a 0.05

Vlastnosti intervalu spolehlivosti:

1. Čím je rozsah výběru n větší, tím je interval spolehlivosti užší.

2. Čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší,

3. Čím je vyšší statistická jistota $(1 - \alpha)$, tím je interval spolehlivosti širší.

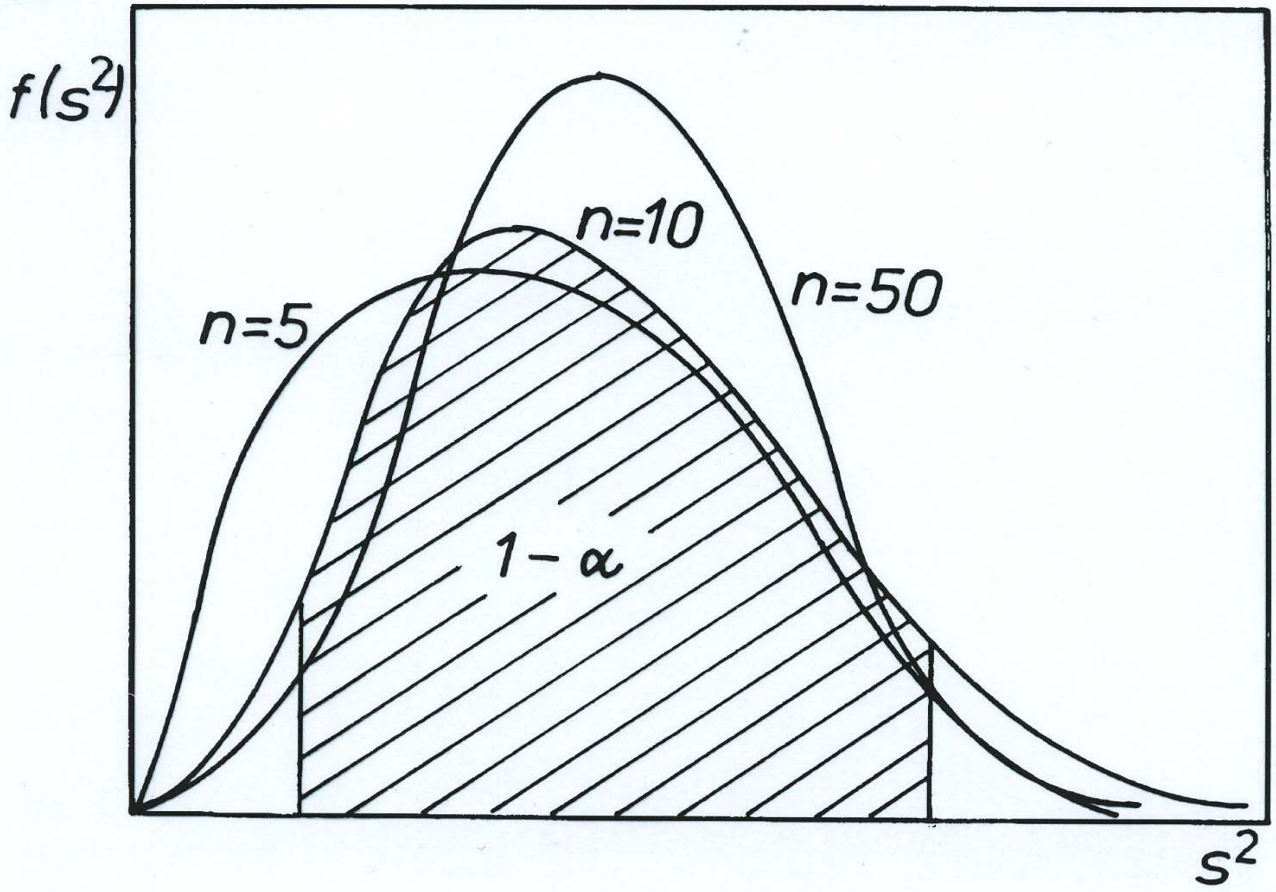
Obecně: **100(1 - α)%ní int. spolehlivosti parametru Θ** se vypočte dle asymptotického vztahu

$$\hat{\Theta} - u_{1-\alpha/2}\sqrt{D(\hat{\Theta})} \leq \Theta \leq \hat{\Theta} + u_{1-\alpha/2}\sqrt{D(\hat{\Theta})}$$

100(1 - α)%ní oboustranný interval spolehlivosti rozptylu σ^2 se vypočte dle

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

kde $\chi^2_{1-\alpha/2}(n-1)$ je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil rozdělení χ^2.



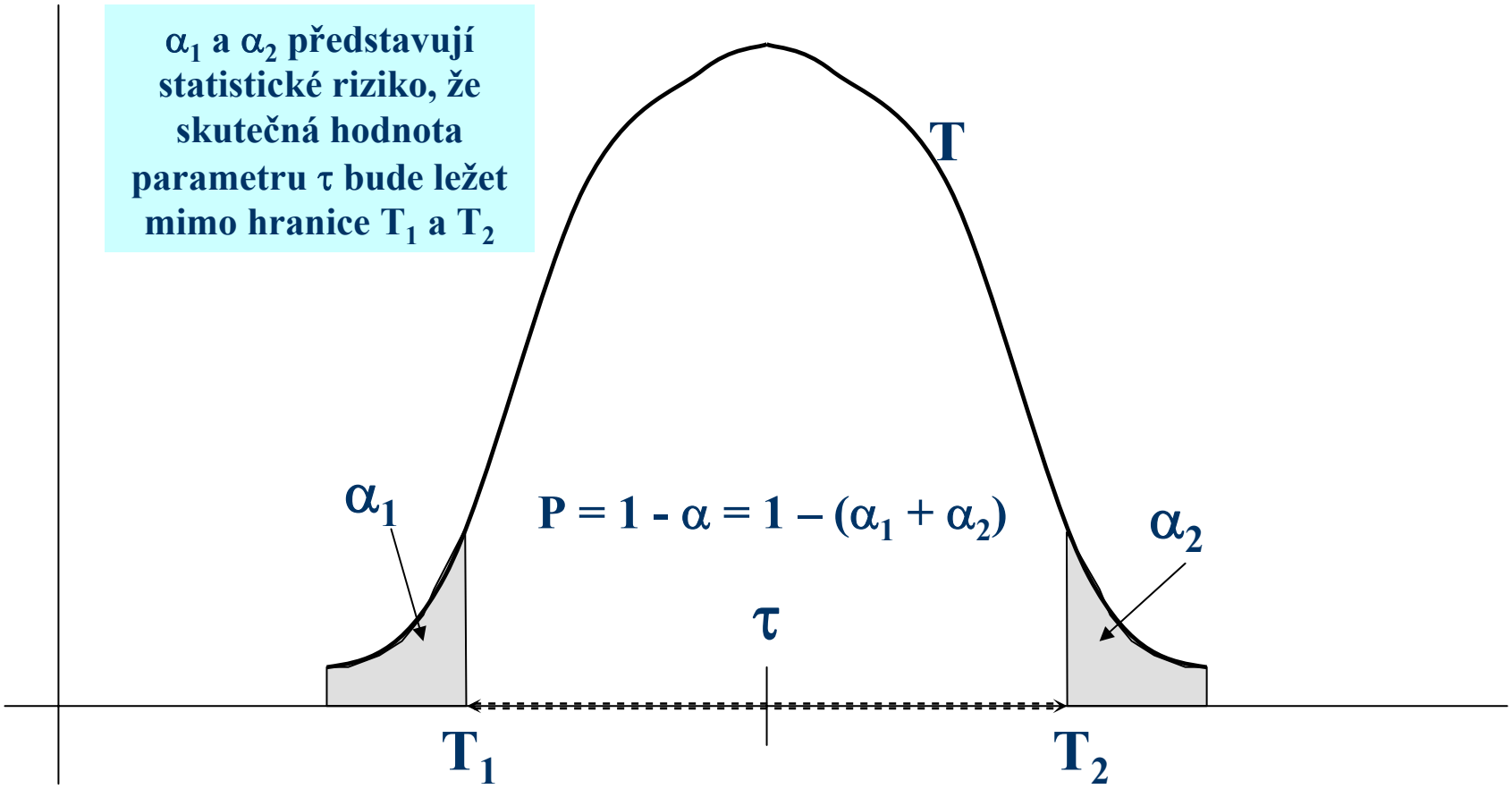
Závislost intervalu spolehlivosti a výběrového rozptylu na velikosti výběru

Interval spolehlivosti mediánu se přibližně vyčíslí

$$\tilde{x}_{0.5} - u_{1-\alpha/2}\frac{0.707\ s}{\sqrt{n}} \leq \mathbf{med} \leq \tilde{x}_{0.5} + u_{1-\alpha/2}\frac{0.707\ s}{\sqrt{n}}$$

INTERVALOVÉ ODHADY PARAMETRŮ ZÁKLADNÍHO SOUBORU

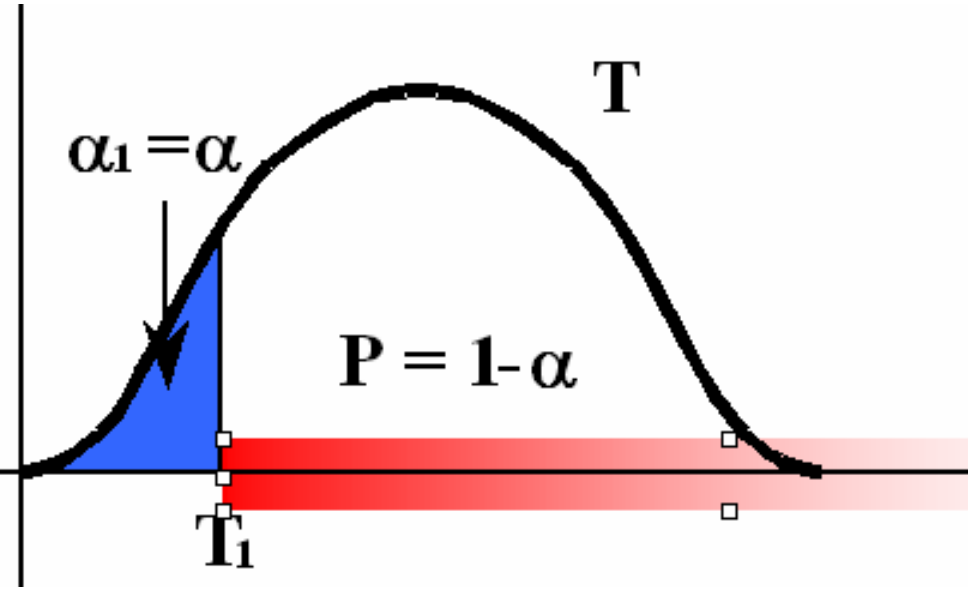
α_1 a α_2 představují statistické riziko, že skutečná hodnota parametru τ bude ležet mimo hranice T_1 a T_2



JEDNOSTRANNÉ INTERVALOVÉ ODHADY

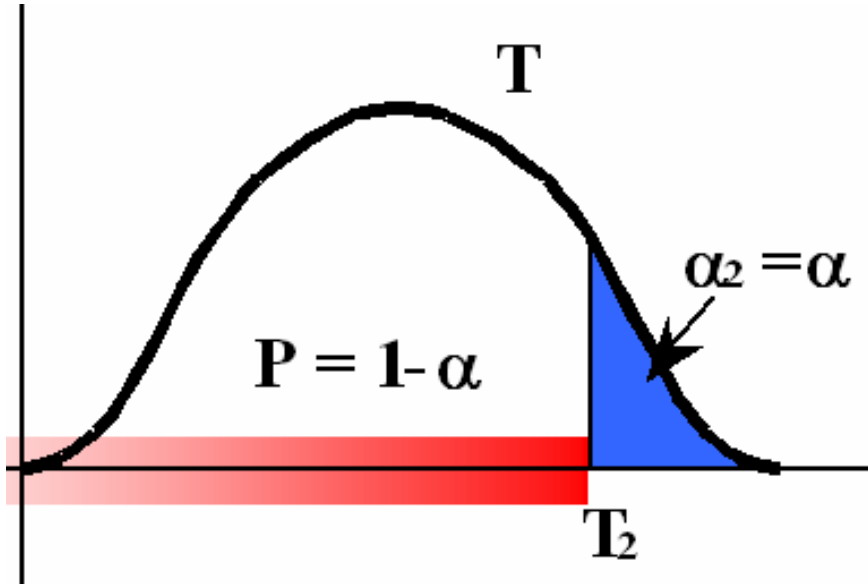
Levostranný odhad

$$P(\tau > T_1) = 1 - \alpha$$

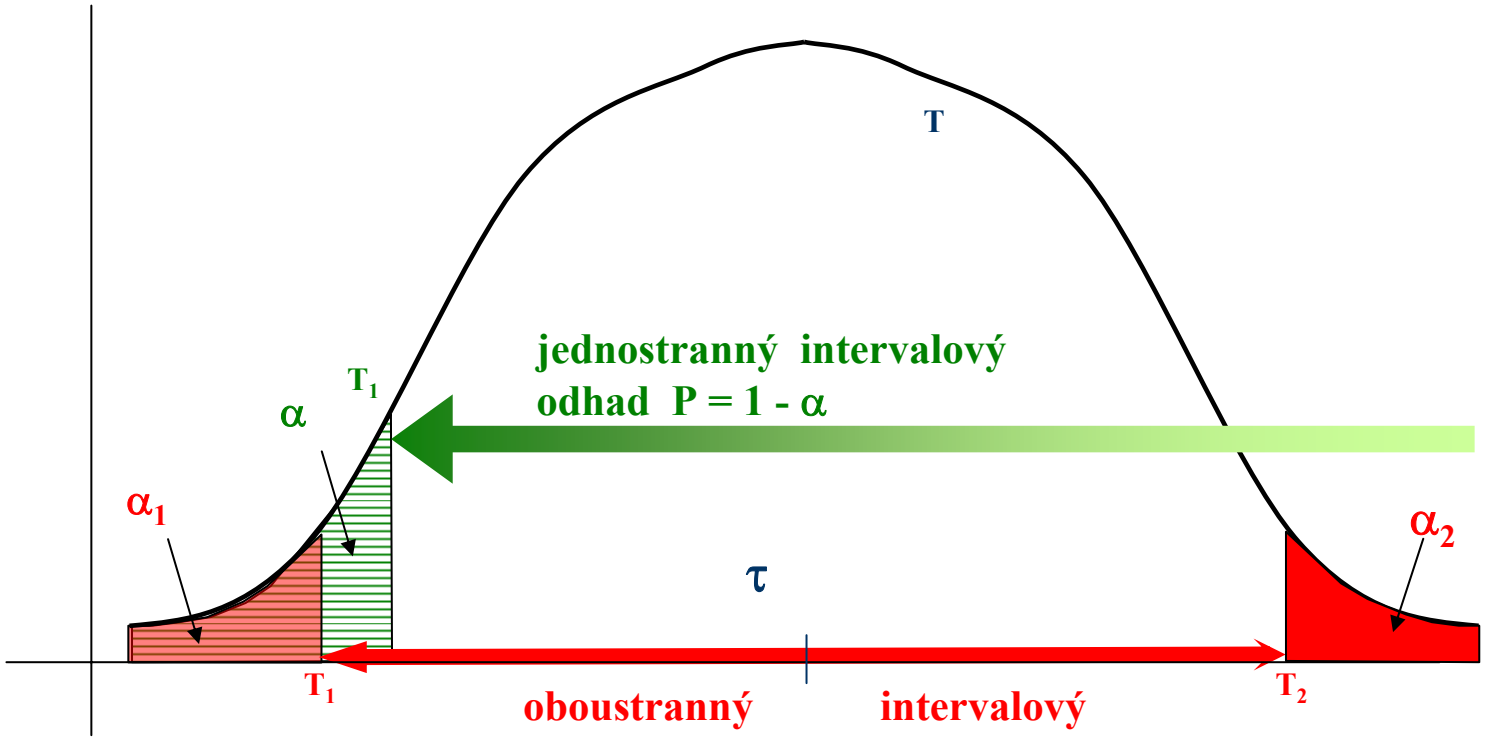


Pravostranný odhad

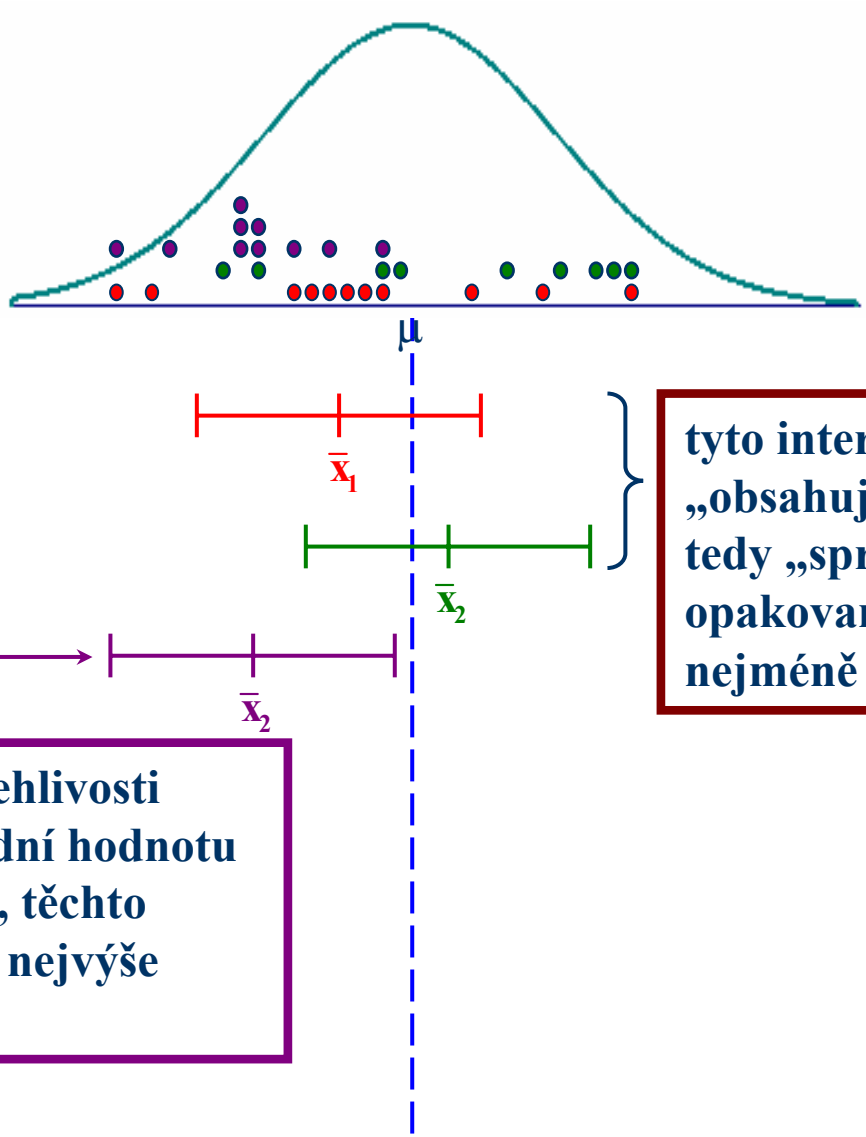
$$P(\tau < T_2) = 1 - \alpha$$



POROVNÁNÍ
JEDNOSTRANNÉHO A OBOUSTRANNÉHO ODHADU



HLADINA VÝZNAMNOSTI α V INTERVALOVÝCH ODHADECH



tyto intervaly spolehlivosti „obsahují“ střední hodnotu (jsou tedy „správné“), těch (při opakovaných výběrech) bude nejméně (1- α).100 %

tento interval spolehlivosti „neobsahuje“ střední hodnotu (je tedy „chybný“), těchto intervalů se objeví nejvýše (100α) %

INTERVAL SPOLEHLIVOSTI STŘEDNÍ HODNOTY μ

❶ je známa směrodatná odchylka σ základního souboru nebo je používán velký výběr (nad 30 prvků)

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

v případě velkého výběru lze použít místo σ výběrovou směrodatnou odchylku S

dolní hranice **horní hranice**

$z_{\alpha/2}$ je kvantil normovaného normálního rozdělení pro hladinu významnosti $\alpha/2$

INTERVAL SPOLEHLIVOSTI STŘEDNÍ HODNOTY μ

❷ není známa směrodatná odchylka σ základního souboru a je používán pouze malý výběr (do 30 prvků)

Platí, že veličina $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ má t-rozdělení s $k = (n - 1)$ stupni volnosti

$$\bar{x} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}$$

$t_{\alpha/2,n-1}$ je kvantil Studentova t-rozdělení pro hladinu významnosti $\alpha/2$ a $(n - 1)$ stupňů volnosti

INTERVAL SPOLEHLIVOSTI STŘEDNÍ HODNOTY μ

❸ velikost základního souboru je známa (N) a výběrový soubor je relativně velký (n > 5 % N)

Používá se **korekce na konečný základní soubor**:

$$\bar{x} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \leq \mu \leq \bar{x} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

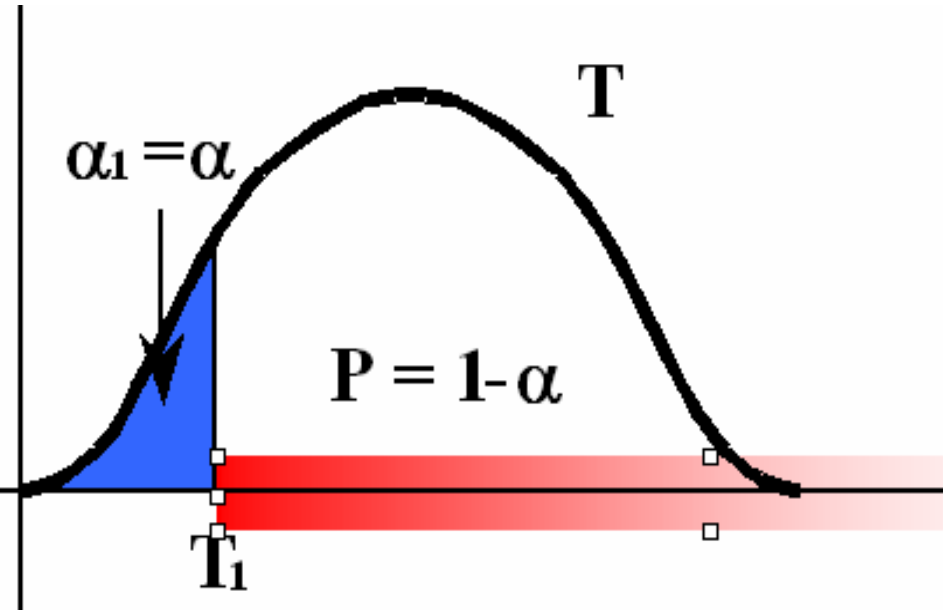
Účelem korekce je zmenšit standardní chybu $\bar{x}$.

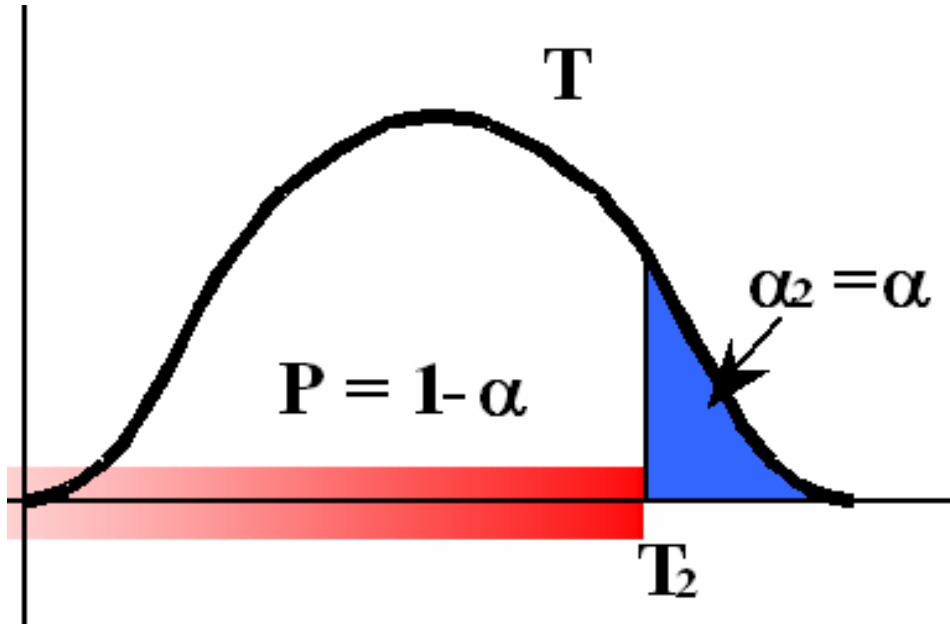
INTERVAL SPOLEHLIVOSTI STŘEDNÍ HODNOTY μ

❹ jednostranné intervaly

Jednostranné intervaly se počítají podle stejných vztahů jako oboustranné, pouze **hladina významnosti je α místo α/2**,

(veškeré statistické riziko „chybného“ intervalu je na jedné straně)





FAKTORY OVLIVŇUJÍCÍ VELIKOST INTERVALU SPOLEHLIVOSTI (IS)

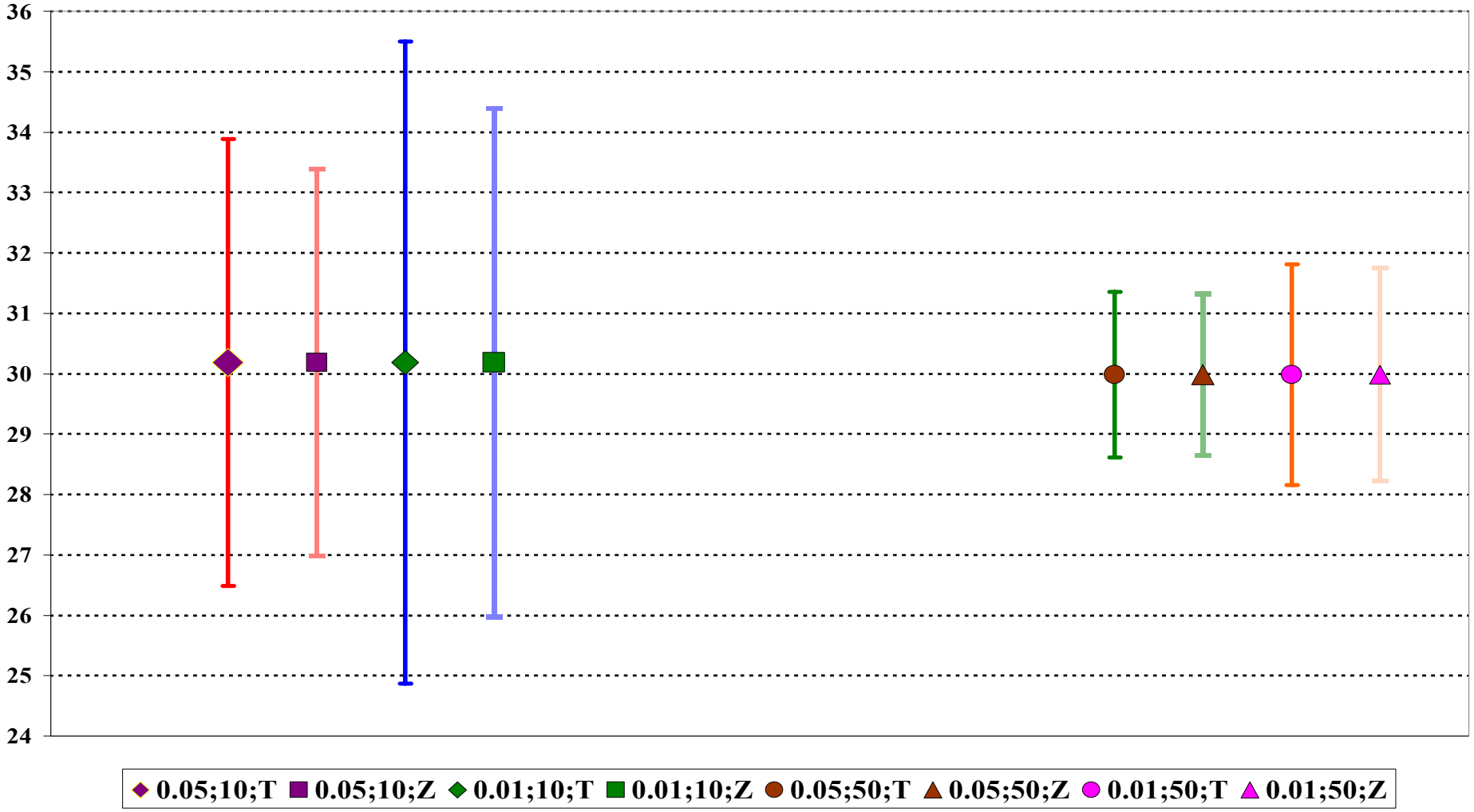
- **Velikost výběru** (čím větší výběr, tím užší IS).

- **Hladina významnosti α** (čím vyšší hodnota α, tím užší interval – **nižší hladina významnosti** (např. 0,01 místo 0,05) **znamená požadavek vyšší spolehlivosti určení IS** . Pokud určíme $\alpha = 0.01$, požadujeme spolehlivost IS $P = 99\%$. Pokud určíme $\alpha = 0.05$, požadujeme spolehlivost IS $P = 95\%$, IS musí být širší pro $P = 99\%$ než pro $P = 95\%$, protože musíme zaručit vyšší spolehlivost.

- **Variabilita** (čím vyšší hodnota směrodatné odchylky, tím širší IS).

- **Použitý vzorec** (pokud používáme *t*-rozdělení, je IS širší než při použití $N(0,1)$, rozdíl je markantnější u malých výběrů).

FAKTORY OVLIVŇUJÍCÍ VELIKOST INTERVALU SPOLEHLIVOSTI (IS)



INTERVAL SPOLEHLIVOSTI SMĚRODATNÉ ODCHYLKY σ

❶ pro malé výběry

Výpočet intervalu spolehlivosti směrodatné odchylky využívá χ^2-rozdělení a je nesouměrný – nesouměrnost je vyšší u odhadů vycházejících z malých výběrů.

$$\sqrt{\frac{n \cdot S^2}{\chi^2_{\frac{\alpha}{2}}}} \leq \sigma \leq \sqrt{\frac{n \cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}}}$$

INTERVAL SPOLEHLIVOSTI SMĚRODATNÉ ODCHYLKY σ

❷ pro velké výběry (nad 30 prvků)

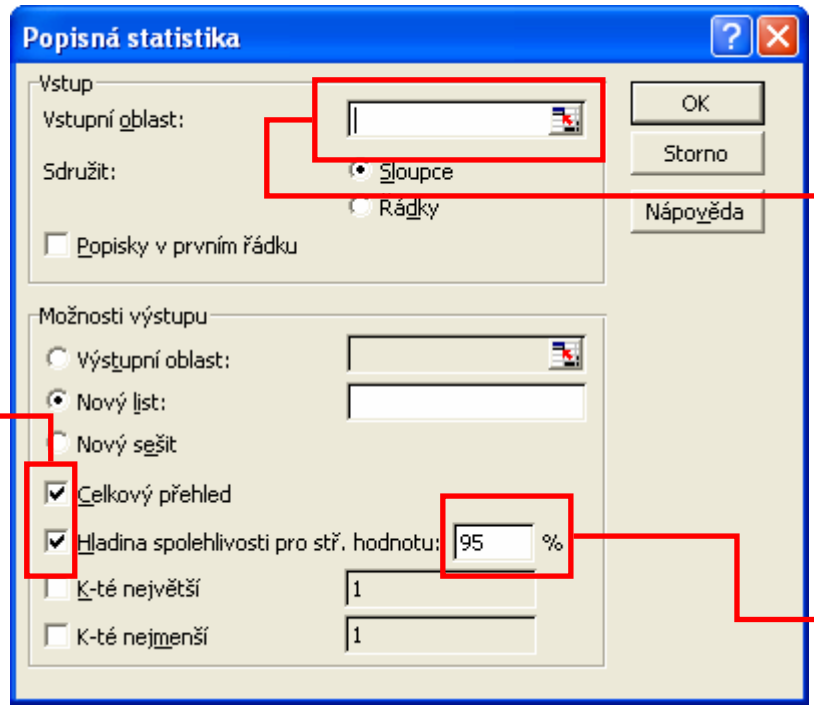
Výpočet intervalu spolehlivosti směrodatné odchylky pro velké výběry využívá normovaného normálního rozdělení a je souměrný.

$$\sigma = S \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{2n}}$$

INTERVALY SPOLEHLIVOSTI – PROVEDENÍ V EXCELU

❶ interval spolehlivosti střední hodnoty

a) pomocí doplňku Analýza dat



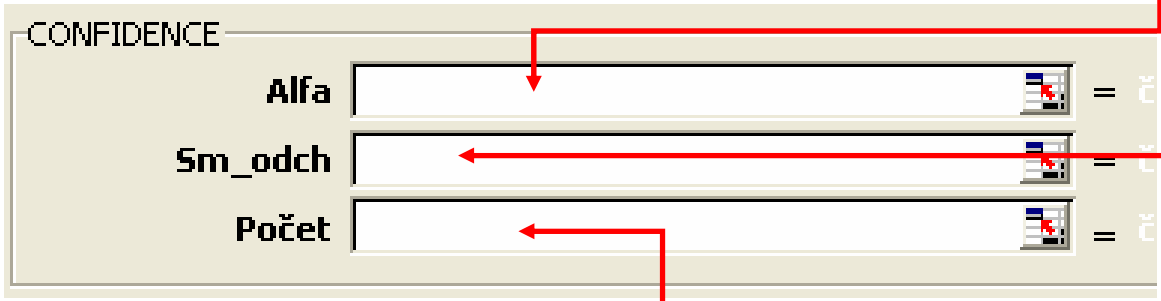
rozsah dat výběru

hodnota 100.(1-α)%

musí být zatrženo !!

INTERVALY SPOLEHLIVOSTI – PROVEDENÍ V EXCELU

❷ pomocí funkce CONFIDENCE



hodnota α

směrodatná odchylka (např. vypočítaná pomocí modulu „Popisná statistika“

velikost výběru

Způsob ❶ počítá interval spolehlivosti podle vzorce $t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}$

Způsob ❷ počítá interval spolehlivosti podle vzorce $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

VELIKOST VÝBĚRU

Obecně platí – čím větší – tím lepší.
Nejlepší je žádný výběr a použít základní soubor.

Obvyklá otázka:

Jakou **minimální** velikost výběru potřebuji vzhledem k **účelu analýzy** a k **požadované vypovídací schopnosti** o základním souboru?

Které kritérium použít?

Jedním ze základních kritérií velikosti výběru je **požadovaná přesnost a spolehlivost určení stanoveného parametru** (obvykle střední hodnoty).

VELIKOST VÝBĚRU

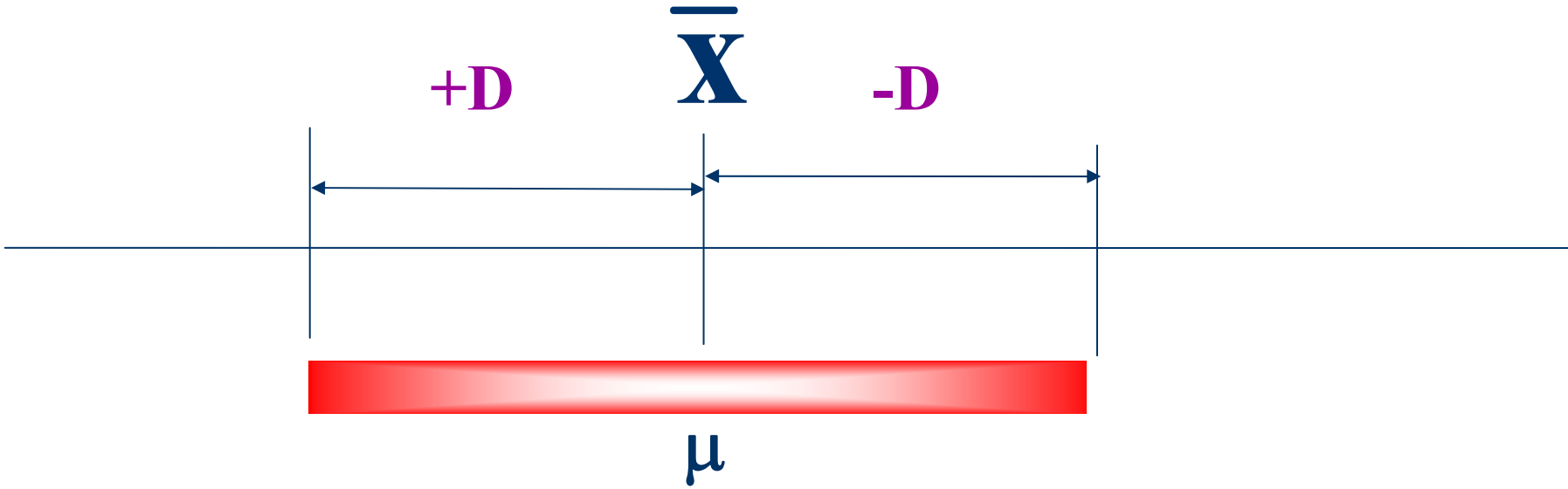
❶ Jaká bude **maximální povolená „vzdálenost“ mezi odhadem parametru a vlastním parametrem**? Odpovědí je **přesnost odhadu,** značená D.

❷ Jakou požadujeme **spolehlivost, že skutečná vzdálenost mezi odhadem a skutečným parametrem bude menší nebo nejvýše rovna D**? Odpovědí je **spolehlivost odhadu** daná hodnotou $100.(1-\alpha)$. Musíme tedy určit **hladinu významnosti α.**

❸ Jaká je **variabilita základního souboru** (většinou ji neznáme, je nutné použít co nejpřesnější odhad). Určíme ji pomocí **rozptylu (S^2)** nebo **variačního koeficientu (S%)**.

VELIKOST VÝBĚRU ZALOŽENÁ NA INTERVALU SPOLEHLIVOSTI μ

Jakou **velikost výběru n** ze základního souboru s **variabilitou** danou **rozptylem S^2 minimálně** potřebuji, abych se **spolehlivostí 100.(1-α) %** zabezpečil, že **střední hodnota μ** se bude pohybovat v **intervalu $\bar{x}$ ±D** (výběrový průměr ± přesnost odhadu)?



VELIKOST VÝBĚRU ZALOŽENÁ NA INTERVALU SPOLEHLIVOSTI μ

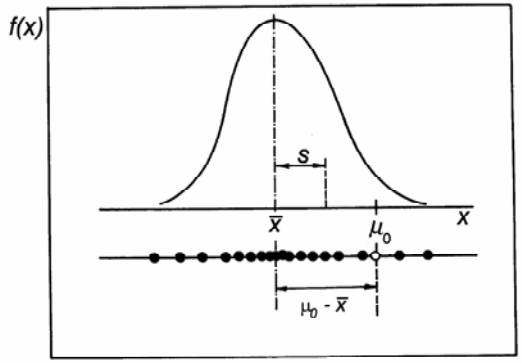
$$n = t_{\alpha/2}^2 \cdot \frac{S^2}{D^2} \qquad n = t_{\alpha/2}^2 \cdot \frac{(S\%)^2}{D^2\,[\%]}$$

$t_{\alpha/2}$ kvantil *t*-rozdělení pro hladinu významnosti $\alpha/2$ a pro n - 1 stupňů volnosti. Pro 1. aproximaci n odhadneme, pro 2. počítáme s výsledným (n - 1) z 1. aproximace a pokračujeme tak dlouho, dokud se n mění. Pro velké výběry můžeme použít přímo $z_{\alpha/2}$.

1) Vzorec vlevo se používá, pokud variabilitu (S^2) i přesnost odhadu (D^2) určujeme absolutně, tj.v jednotkách měřené veličiny.

2) Vzorec vpravo se používá, pokud obé určujeme relativně (v %).

Asymetrické rozdělení (pozitivně zešikmené) u dat
z oblasti stopové analýzy
při monitorování úrovně škodlivin v životním prostředí.



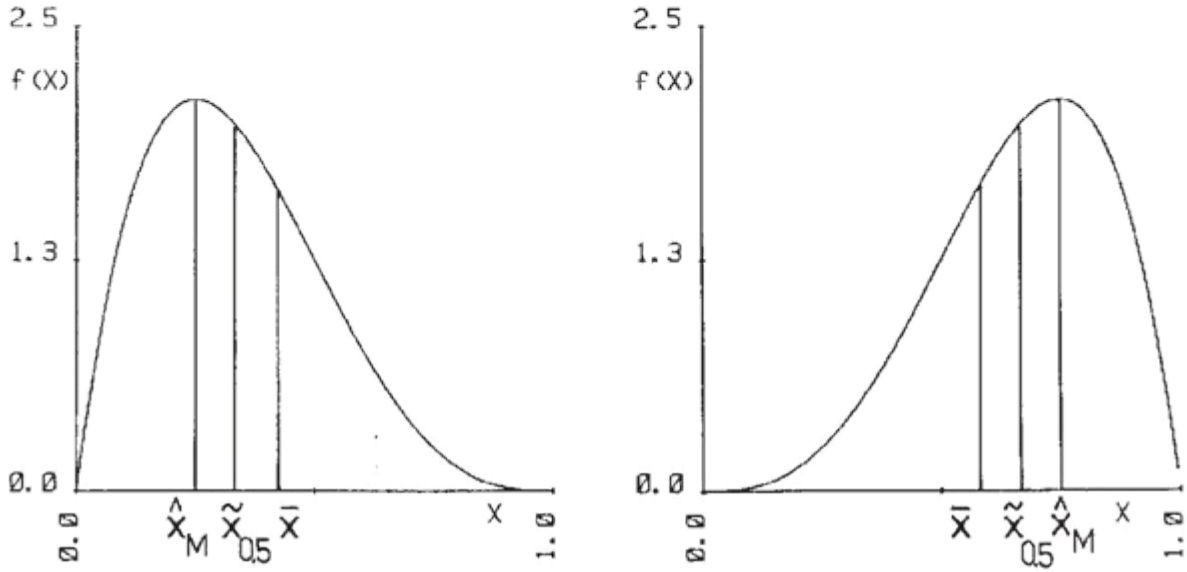
Úkol: zda odhad střední hodnoty výběru o velikosti *n* nepřekračuje úroveň μ_0,
zda μ_0 padlo do intervalu spolehlivosti *CI* parametru μ.

Specifické zvláštnosti dat:

a) Obsahují extrémně velké hodnoty, které nejsou důsledkem chyb měření.
b) Mohou být cenzurována zdola s ohledem na limitu detekce přístrojů.
c) Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám.
d) Jejich počet je omezen vzhledem k drahému vzorkování a složitému vyhodnocení.
e) Nelze opakované vzorkování za stejných podmínek, protože se stopové koncentrace škodlivin mění jak v času, tak i v prostoru.

Robustní techniky selhávají, protože eliminují extrémy, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

Bodový odhad korigovaného průměru asymetrického rozdělení



1. Korigovaný průměr Johnsonovou transformací

$$\bar{x}_R = \bar{x} + \frac{s\ g_1}{6\ n} \ .$$

velikost korekce souvisí se šikmostí a počtem měření.

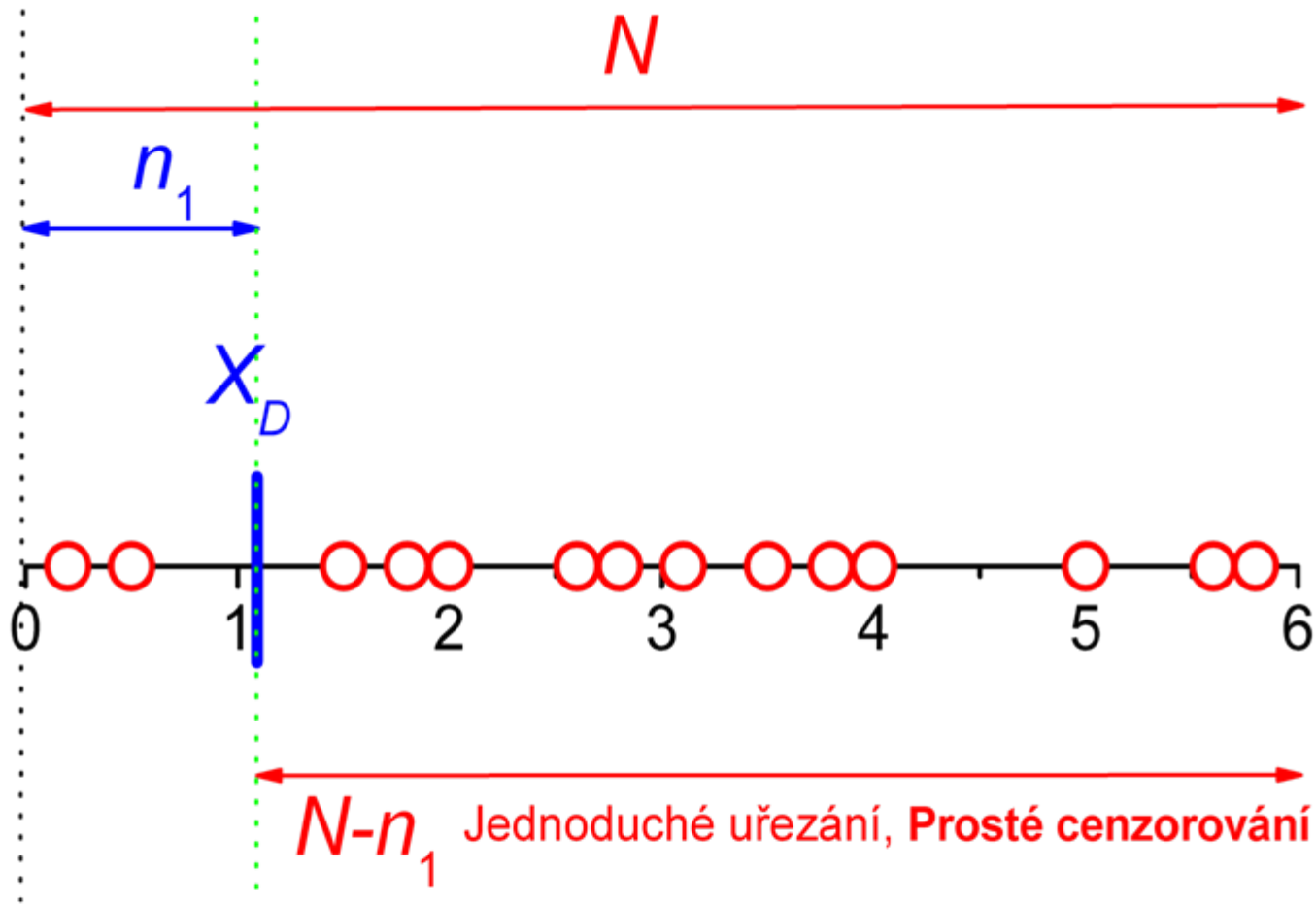
2. MCE odhad dle Chenové $\bar{x}_{R,\,MCE} = \bar{x} + ds$, kde d se vyčíslí dle

$$0.5\left[b - \frac{2\sqrt{n}}{g_1(x)} + \sqrt{4 - \frac{b^2}{3} + \frac{4n}{g_1(x)} + \frac{8\log(a}{b\,g_1(}\right.$$

a volba a a b se doporučuje $a = 1$ a $b = 2$, nebo $a = 10$ a $b = 3$.

3. Penalizovaný průměr se vypočte dle $\bar{x}_P = \bar{x} + \frac{4.5\,s^2}{\sqrt{n}}\,f(\bar{x})\,[1 - F(\bar{x})]$,

kde $f(\bar{x})$, resp. $F(\bar{x})$ jsou hustoty pravděpodobnosti a distribuční funkce $f(x) = \frac{\text{int}\,(\sqrt{n})}{2n\,A(\bar{x})}$, a $A(\bar{x})$ je k-tá nejmenší hodnota rozdílu $w_i = |x_i - \bar{x}|$, kde $k = \text{int}(n^{0.5})$ čili jde o k-tou pořádkovou statistiku. Hodnota distribuční funkce je počet prvků výběru ležících pod $\bar{x}$ a dělený n.


N
n1
XD
0
1
2
3
4
5
6
N-n1
Jednoduché uřezání, **Prosté cenzorování zdola**

PŘÍKLAD 1. *Určení koncentrace nečistot v surovině*

Byla sledována koncentrace nečistot v μg/g:

DL, DL, 1.24, 1.49, 1.50, 1.56, 1.61, 1.78,

kde DL značí hodnoty pod limitou detekce $x_D = 1$ μg/g .

Cíl: odhad střední hodnoty rozptylu a intervalu spolehlivosti za předpokladu normálního rozdělení.

Náhodný výběr reprezentovaný N-ticí dat $x_1, x_2, \ldots, x_N$.

Střední hodnotu, rozptyl a interval spolehlivosti střední hodnoty lze spolehlivě vypočítat jen při znalosti typu rozdělení pravděpodobnostního modelu měření.

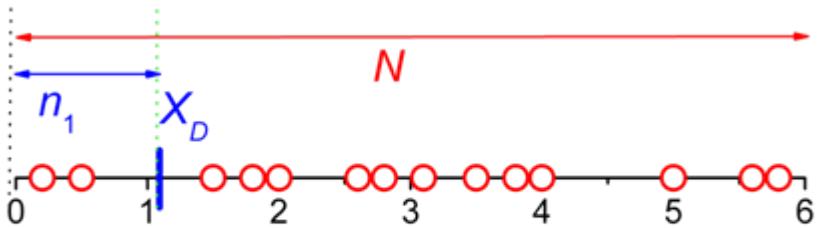
Existují dvě mezní situace dle rozmezí dat:

1) Rozmezí dat je v rámci jednoho řádu: užijeme standardní statistické metody za konstantního rozptylu a aditivního modelu měření.

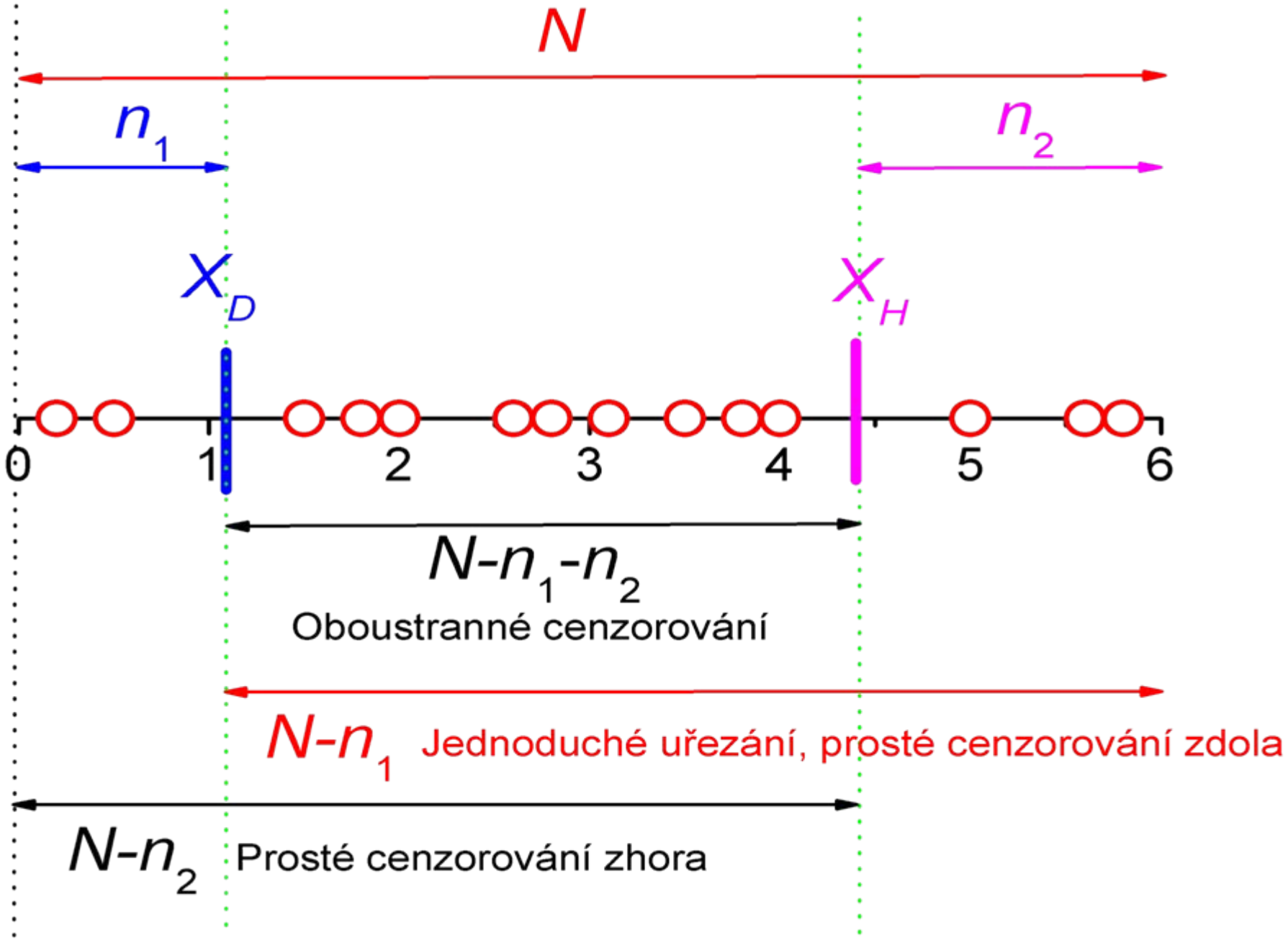
2) Rozmezí dat je v rozmezí několika řádů: použijeme logaritmickou transformaci dat nebo multiplikativní model měření.

Někdy nelze získat všechny výsledky měření,

protože některé jsou pod mezí detekce $DL = x_D$.



Náhodný výběr reprezentovaný N-ticí dat $x_1, x_2, \ldots, x_N$


N
n1
n2
XD
XH
0
1
2
3
4
5
6
N-n1-n2
Oboustranné cenzorování
N-n1
Jednoduché uřezání, prosté cenzorování zdola
N-n2
Prosté cenzorování zhora

1) Z počtu N je N - n_1 nad limitou detekce a zbytek n_1 je pod limitou detekce, **prosté cenzorování zdola**.
2) Při omezení shora je maximálně měřitelná hodnota $UL = x_U$ a n_2 měření je nad limitou intervalu stanovení, **prosté cenzorování shora**.
3) U **oboustranného cenzorování** jsou známy hodnoty pouze pro N - n_1 - n_2 prvků výběru.

Cohen definuje u prostého cenzorování zdola tři základní úlohy:

1) Data pod limitou detekce n_1 a ani celkový počet N nejsou známa. Je k dispozici pouze N - n_1 hodnot, **jednoduché uřezání**.

2) Je znám počet hodnot n_1 pod mezí detekce x_D a dále hodnoty N - n_1 prvků výběru celkové velikosti N, **cenzorování typu I.**

3) Nejmenších n_1 prvků pod mezí detekce nemá hodnotu. Pouze je známa hodnota x_D, která je menší než než nejmenší hodnota prvků nad limitou detekce $x_{(n1+1)}$, kde $x_{(n1+1)}$ značí pořádkovou statistiku, **cenzorování typu II**.

Standardně se řeší pouze cenzorování typu I.

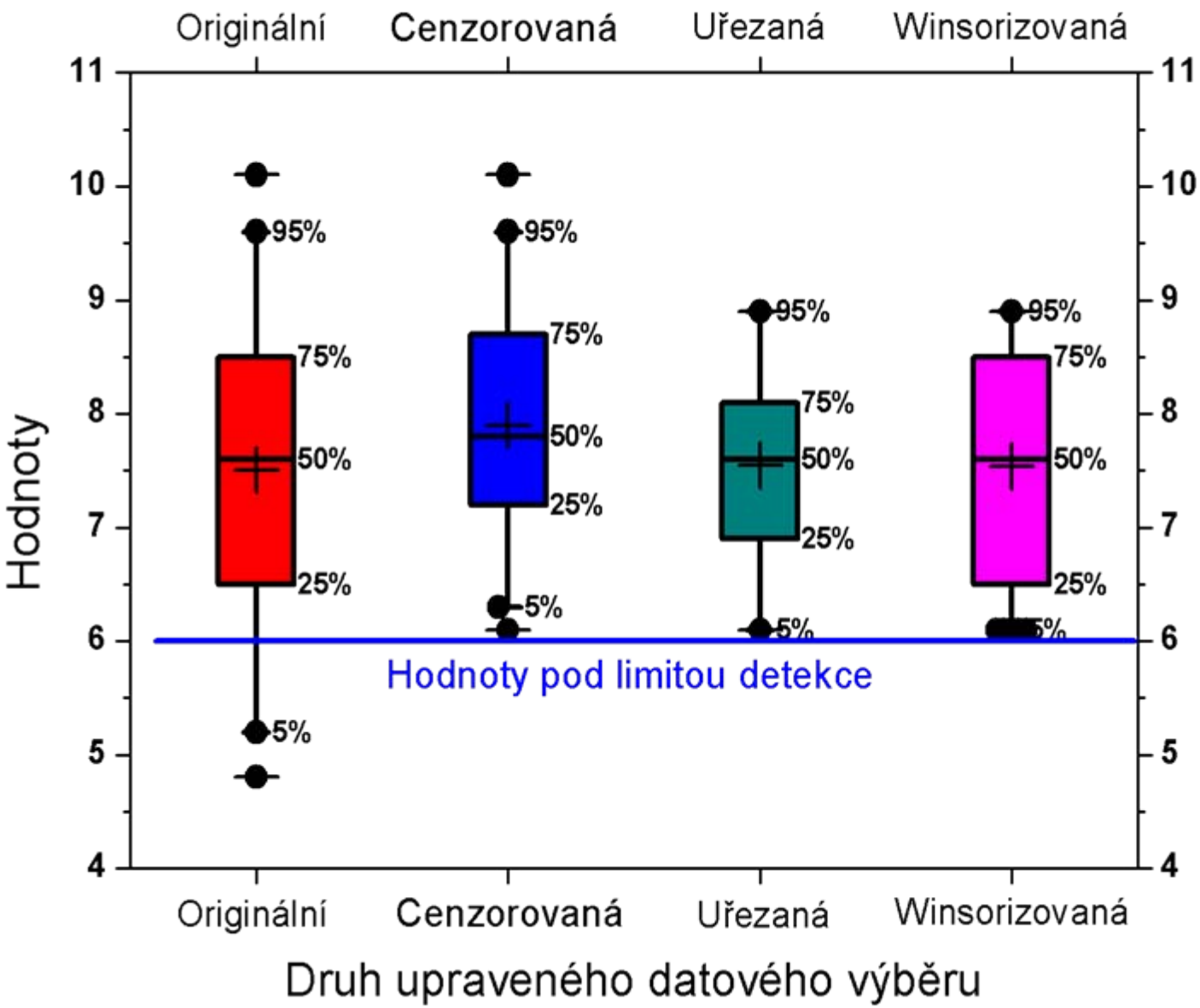
PŘÍKLAD 2. *Určení koncentrace ve stopové analýze*

Byla sledována koncentrace nečistot v μg/g, kdy řada hodnot je pod limitou detekce $x_D = 6$ μg/g .

Cíl: odhad střední hodnoty a intervalu spolehlivosti

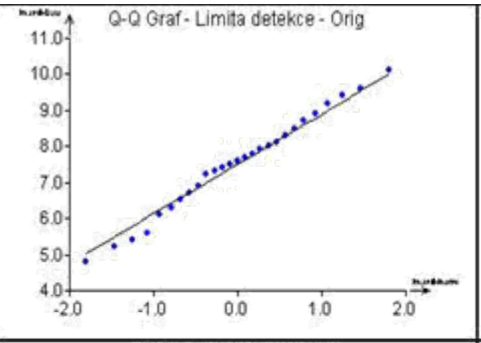
a) metodou původních nezměněných dat
b) metodou cenzorovaného výběru dat,
c) metodou uřezaného výběru dat, a
d) metodou winsorizovaných dat.

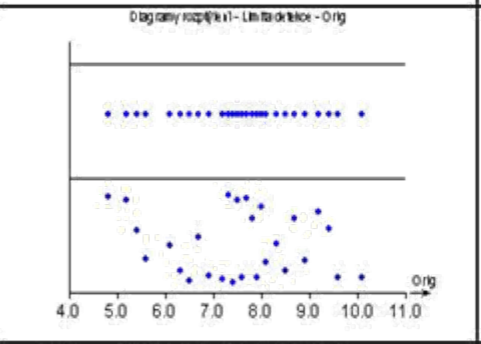
Data jsou	originální	cenzor.	uřezaná	winsoriz.
	4.8			6.1
	5.2			6.1
	5.4			6.1
Detekční limit = 6.0	5.6			6.1
	6.1	6.1	6.1	6.1
	6.3	6.3	6.3	6.3
	6.5	6.5	6.5	6.5
	6.7	6.7	6.7	6.7
	6.9	6.9	6.9	6.9
	7.2	7.2	7.2	7.2
	7.3	7.3	7.3	7.3
	7.4	7.4	7.4	7.4
	7.5	7.5	7.5	7.5
	7.6	7.6	7.6	7.6
	7.7	7.7	7.7	7.7
	7.8	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

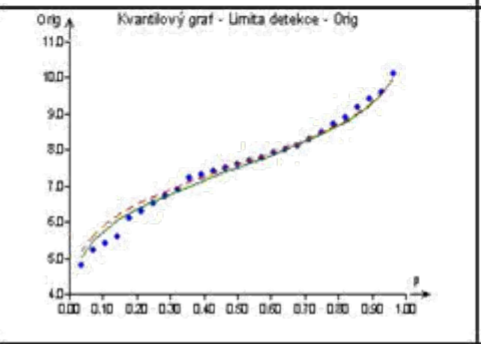

Originální
Cenzorovaná
Uřezaná
Winsorizovaná
11
10
9
8
7
6
5
4
95%
75%
50%
25%
5%
Hodnoty
Hodnoty pod limitou detekce
Druh upraveného datového výběru

Analýza dat s hodnotami pod LOD

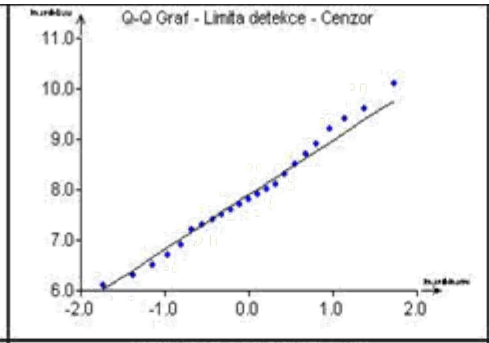
Originální data

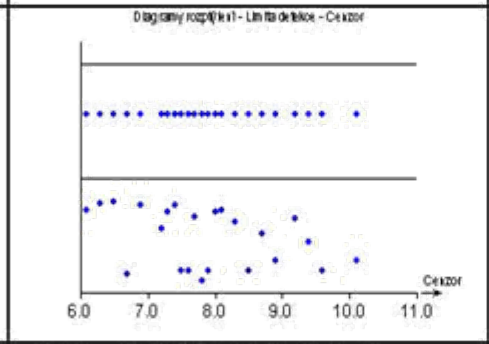


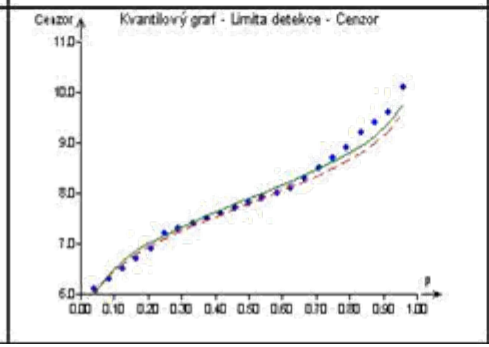




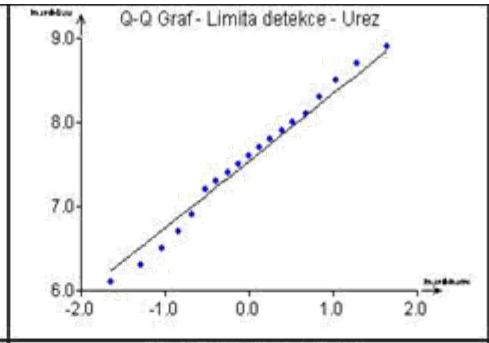
Cenzorovaná data

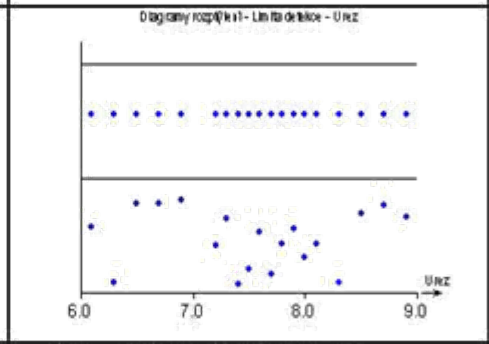


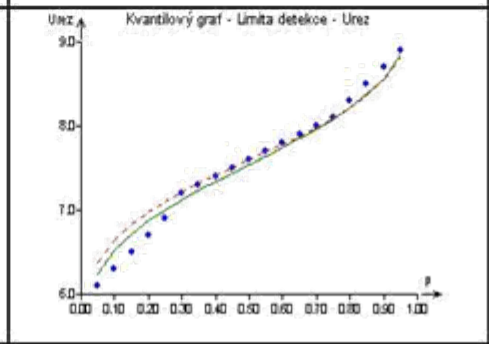




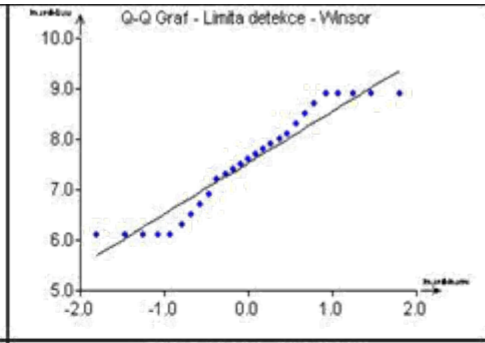
Uřezaná data

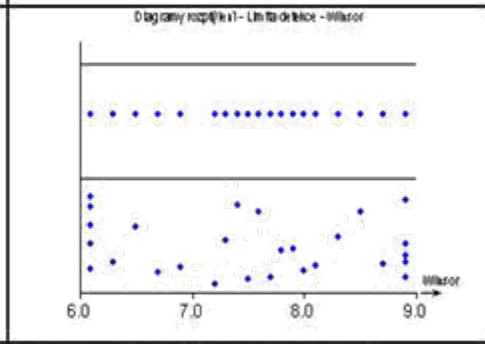


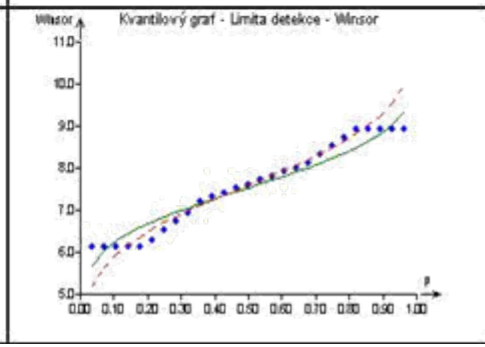




Winsorizovaná data







	Originální data	Cenzorovaná data	Uřezaná data	Winsorizovaná data
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

1. Povaha intervalového odhadu

Bodový odhad: malý význam, neříká nic kde leží skutečná hodnota parametru.

Intervalový odhad: se zadanou pravděpodobností $1 - \alpha$ se v něm nachází skutečná hodnota parametru Θ, L_D a L_H jsou meze ***intervalu spolehlivosti*** čili ***konfidenčního intervalu CI***.

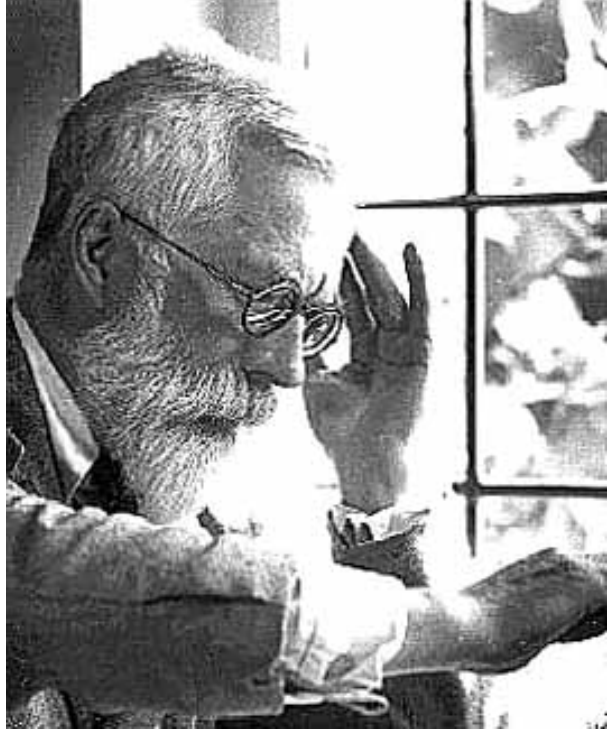
$$\text{Pravděpodobnost} \quad P(L_D < \Theta < L_H) = 1 - \alpha$$

s jakou bude skutečná hodnota Θ v náhodných mezích L_D, L_H, je rovna právě $1 - \alpha$. Nazývá se *koeficient spolehlivosti* čili *konfidenční koeficient, statistická jistota*, obyčejně 0.95 nebo 0.99. Parametr α se nazývá *hladina významnosti*.

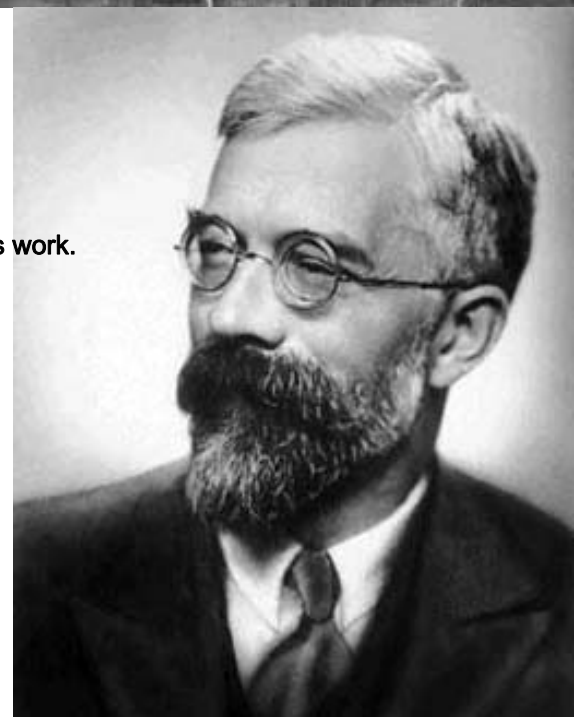
Vlastnosti *CI*:

a) čím je větší rozsah výběru n, tím je interval spolehlivosti užší,
b) čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší,
c) čím je vyšší statistická jistota $(1 - \alpha)$, tím je interval spolehlivosti širší.

Intervaly spolehlivosti L_D a L_H jsou o b o u s t r a n n é .
V praxi se užívá i jednostranný interval, buď l e v o s t r a n n ý $[L_D, \infty)$,
nebo p r a v o s t r a n n ý $(-\infty, L_H]$.



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

2. Konstrukce intervalových odhadů

100(1 − α)% interval spolehlivosti střední hodnoty se vypočte dle

1) Normální rozdělení $N(\mu, \sigma^2)$, $n > 30$: 95% konfidenční interval *CI* bude

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2\frac{\sigma}{\sqrt{n}}.$$

2 přesněji 1.96, je $100(1 - 0.05/2) = 97.5\%$ kvantil normovaného normálního rozdělení $u_{0.975}$.

2) V praxi, $n < 30$: známe odhad s a 95% konfidenční interval *CI* bude

$$\bar{x} - t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(\nu)\frac{s}{\sqrt{n}}.$$

Pro větší rozsahy výběru ($n > 30$) lze použít místo $t_{1-\alpha/2}$ kvantilu $u_{1-\alpha/2}$.

3) Relativní variabilita variačním koeficientem δ a $100(1-\alpha)\%$ interval spolehlivosti δ bude

$$D_M \leq \delta \leq H_M$$

kde

$$D_M = \frac{\hat{\delta}}{\sqrt{\left(\frac{C_1 + 2}{n} - 1\right)\hat{\delta}^2 + \frac{C_1}{n-1}}},$$

a

$$H_M = \frac{\hat{\delta}}{\sqrt{\left(\frac{C_2 + 2}{n} - 1\right)\hat{\delta}^2 + \frac{C_2}{n-1}}},$$

a kde $C_1 = \chi^2_{1-\alpha/2}(n-1)$ a $C_2 = \chi^2_{\alpha/2}(n-1)$ jsou kvantily χ^2-rozdělení.

4) Obecný 100(1 – α)% interval spolehlivosti libovolného parametru Θ bude

$$\hat{\Theta} - u_{1-\alpha/2}\sqrt{D(\hat{\Theta})} \leq \Theta \leq \hat{\Theta} + u_{1-\alpha/2}\sqrt{D(\hat{\Theta})}.$$

5) 100(1 – α)% oboustranný interval spolehlivosti rozptylu σ^2 bude

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

kde $\chi^2_{1-\alpha/2}(n-1)$ je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil χ^2-rozdělení.

6) 100(1 – α)% interval spolehlivosti mediánu bude

$$\tilde{x}_{0.5} - u_{1-\alpha/2}\frac{0.707\,s}{\sqrt{n}} \leq med \leq \tilde{x}_{0.5} + u_{1-\alpha/2}\frac{0.707\,s}{\sqrt{n}},$$

kde *med* označuje medián.

PŘÍKLAD 1 *Analýza pěti výběrů za nesprávného předpokladu normality*

Analýza výběrů velikosti $n = 51$ z rovnoměrného (R), normálního (N), exponenciálního (N), Laplaceova (L) a logaritmicko-normálního rozdělení (LN) za nesprávného předpokladu, že každý analyzovaný výběr pochází z normálního rozdělení.

Řešení: Předpoklad normálního rozdělení platí ve skutečnosti pouze u výběru $N(0, 1)$.

Rozdělení	$\bar{x}$	s	L_1	L_2
Rovnoměrné $R(0.5; 0.083)$	0.488	0.294	0.404	0.571
Normální $N(0; 1)$	−0.0574	1.089	−0.354	0.239
Exponenciální $E(1; -1)$	1.0059	1.167	0.674	1.338
Laplaceovo $L(0; 2)$	−0.0246	1.559	−0.468	0.419
logaritmicko-normální $LN(2.71; 47.21)$	4.077	8.636	1.623	6.532

Závěr: I když je konstrukce *CI* u čtyř výběrů založena na nesprávném předpokladu, pokrývají vyčíslené meze intervalu spolehlivosti ve všech případech správnou střední hodnotu.

Pro Laplaceovo a logaritmicko-normální rozdělení jsou však intervaly spolehlivosti širší.

3. Intervaly spolehlivosti pro zešikmená rozdělení

Standardní statistická analýza zde vede k *přehnaně optimistickým závěrům*.

Pro pozitivně zešikmená data: aritmetický průměr je menší než skutečná hodnota μ.

Výběr velikosti n je *nenormálního rozdělení* se střední hodnotou μ a rozptylem σ^2. Pak

má náhodná veličina z $$z = \sqrt{n}\,\frac{\bar{x} - \mu}{\sigma}$$

asymptoticky normální rozdělení.

Když σ^2 není známo, nahrazuje se výběrovou směrodatnou odchylkou s. Pak má

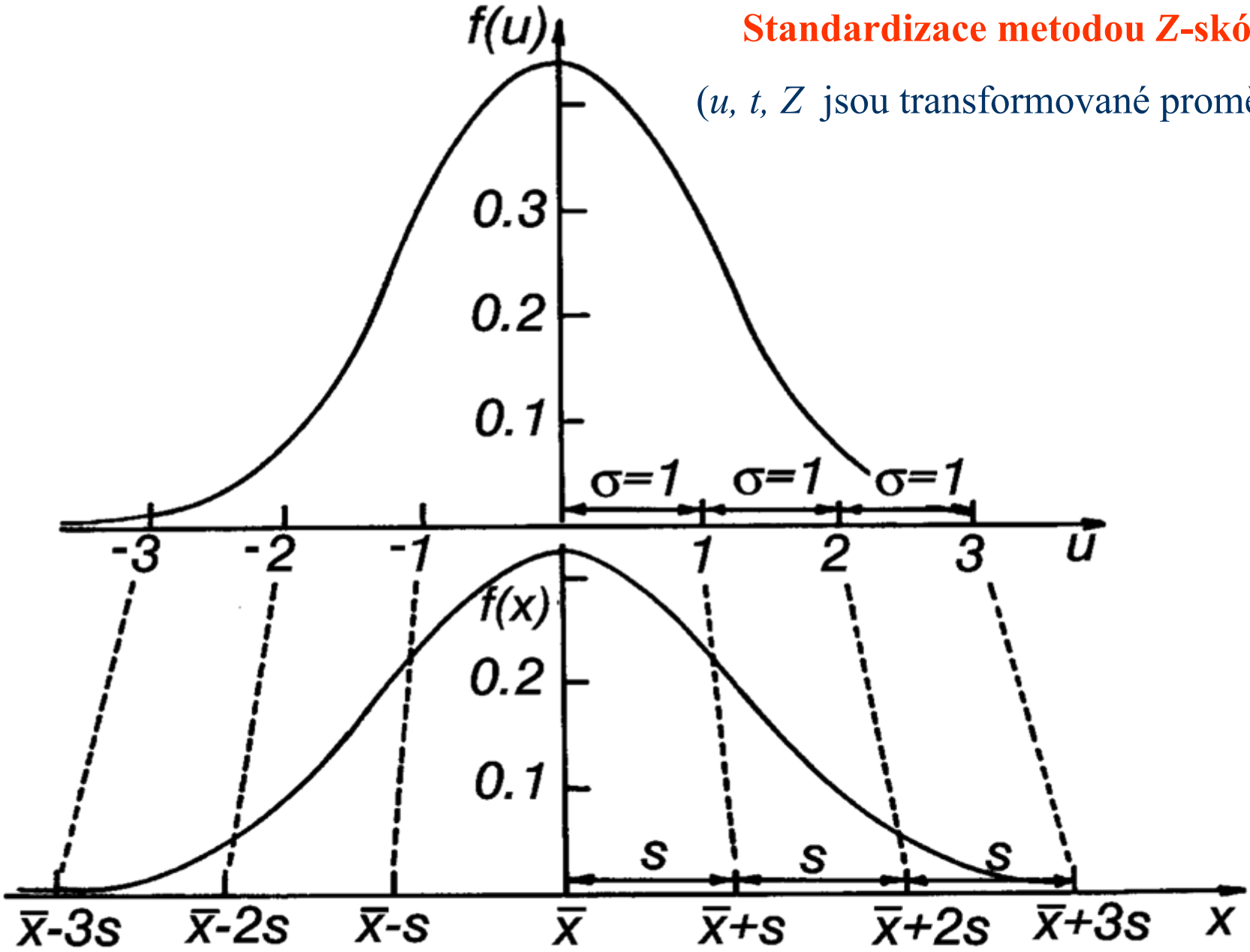
Studentova náhodná veličina t $$t = \sqrt{n}\,\frac{\bar{x} - \mu}{s}$$

Studentovo rozdělení s $n - 1$ stupni volnosti.

Asymptotická normalita veličiny z, resp. Studentovo rozdělení veličiny t, umožňuje:
konstrukci intervalu spolehlivosti střední hodnoty μ.

Standardizace metodou Z-skóre

(*u, t, Z* jsou transformované proměnné)



1) Při znalosti rozptylu σ^2: interval spolehlivosti *CI* = (*CID*, *CIH*) vyjádřit ve tvaru

$$\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}},$$

kde $z_{1-\alpha/2} = -t_{\alpha/2}$ jsou kvantily normovaného normálního rozdělení.

2) Pokud není σ^2 známo: interval spolehlivosti *CI* = (*CID*, *CIH*) lze vyjádřit ve tvaru

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},$$

kde $t_{1-\alpha/2}(n-1) = -z_{\alpha/2}(n-1)$ jsou kvantily Studentova rozdělení s $n-1$ stupni volnosti.

- Pro normální rozdělení mají oba intervaly přesně $100(1-\alpha)\%$ *pokrytí střední hodnoty*:
  a) jen u $100\alpha/2$ % případů je *střední hodnota* menší než *CID* (**nejistota *NP* zprava**),
  b) jen u $100\alpha/2$ % případů je větší než *CIH* (**nejistota *NL* zleva**).
- Pro nenormální rozdělení platí tyto intervaly pouze pro dostatečně vysoká *n*.
- Velikost *n* závisí na šikmosti $g_1(x)$ rozdělení.

Vliv šikmosti na rozdělení náhodné veličiny z vyjádříme prvním členem Edgeworthova rozvoje dle

$$P(z \leq x) = \Phi(x) - \frac{g_1(x)\ (x^2 - 1)}{6\sqrt{n}}\ \phi(x)\ ,$$

kde $\Phi(x)$ je distribuční funkce normovaného normálního rozdělení,
$\phi(x)$ je odpovídající hustota pravděpodobnosti.

Šikmost náhodné veličiny z je dána vztahem $g_1(z) = g_1(x)/\sqrt{n}$.

čím je $g_1(z)$ blíže k nule, tím je rozdělení veličiny z bližší normálnímu.

Pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ je kladné),
je také rozdělení veličiny z zešikmené k vyšším hodnotám (tj. $g_1(z)$ je kladné).

Interval spolehlivosti pak má: vyšší horní mez *CIH* a
vyšší dolní mez *CID*
než odpovídá reálnému rozdělení statistiky z.

Například: pro výběr rozsahu $n = 10$ z standard. exponenciálního rozdělení ($g_1(x) = 2$)

97.5% kvantil rozdělení veličiny z je roven 2.24 a
odpovídající kvantil normovaného normálního rozdělení je pouze 1.96.

2.5% kvantil rozdělení veličiny z je roven pouze -1.65 a
odpovídající kvantil normovaného normálního rozdělení je -1.96.

Závěr: Interval spolehlivosti je tedy celý posunut doprava oproti skutečnému.

Vliv šikmosti na rozdělení náhodné veličiny t lze užít prvního členu Edgeworthova rozvoje

$$P(t \leq x) = \Phi(x) - \frac{g_1(x)\ (2x^2 + 1)}{6\sqrt{n}}\ \phi_n(x)\ .$$

Porovnáním $P(z \leq x)$ a $P(t \leq x)$ je patrné opačné znaménko korekčního členu:

pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ je kladné)

je rozdělení náhodné veličiny t zešikmené k nižším hodnotám (tj. $g_1(x)$ je záporné).

Interval spolehlivosti pak má: nižší horní mez *CIH* a

nižší dolní mez *CID*,

než odpovídá reálnému rozdělení statistiky t.

Závěr: Interval spolehlivosti je tedy celý posunut doleva oproti skutečnému. To je zvláště nepříjemné u dat silně zešikmených vpravo a vede to k přehnaně optimistickým závěrům o úrovni kontaminace.

Postup nevyčísluje horní mez 95% intervalu spolehlivosti, ale jinou mez závislou na šikmosti dat a velikosti výběru.

Závěry: problémy výpočtu intervalů spolehlivosti střední hodnoty.

1) Pokud je rozdělení dat nenormální (zešikmené vpravo).

2) Pokud je velikost výběru malá.

3) Posun intervalu spolehlivosti směrem k nižším hodnotám.

4) Pro pozitivně zešikmená data je odhad $\bar{x}$ menší než μ.

5) Interval spolehlivosti je poměrně robustní.

6) **Základní techniky omezení vlivu zešikmení dat:**
   a) snížení asymetrie rozdělení náhodné veličiny *t*,
   b) výpočet korigovaného průměru.

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

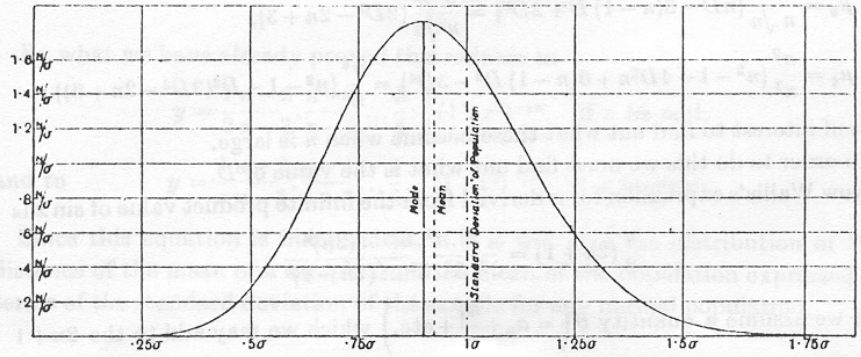
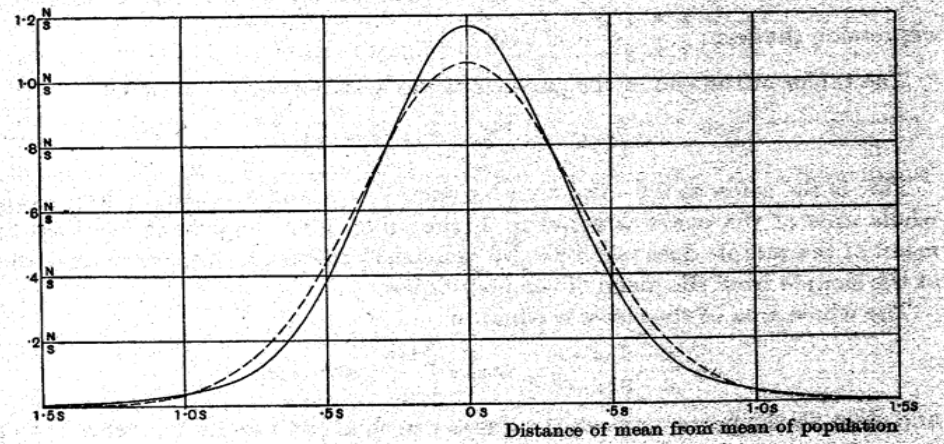


DIAGRAM II. Solid curve $y = \frac{N}{s} \times \frac{8}{7}\cdot\frac{6}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^{10}\theta, \quad x/s = \tan\theta$

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



4. Omezení asymetrie rozdělení Studentovy statistiky

Výhodné je použít Cornishův–Fisherův rozvoj, který umožňuje náhradu náhodné veličiny x, (mající μ, σ^2 a momenty m_r), náhodnou veličinou z (mající normované normální rozdělení):

$$x \approx \mu + \sigma z + \left(\frac{m_3}{6\sigma^2} \right) (z^2 - 1) + \dots ,$$

kde Cornishův–Fisherův rozvoj aritmetického průměru má tvar

$$\bar{x} = \mu + \frac{\sigma z}{\sqrt{n}} + \left(\frac{m_3}{6n\ \sigma^2} \right) (z^2 - 1) + \dots ,$$

a Cornishův–Fisherův rozvoj výběrového rozptylu je dán

$$s^2 = \sigma^2 + z \sqrt{\frac{m_4 - \sigma^4}{n}} .$$

Modifikovaná náhodná veličina t_1 Studentova rozdělení i pro nenormálně rozdělená data

je dána $$t_1 = \left[(\bar{x} - \mu) + \frac{m_3}{6n\,\sigma^2} + \frac{m_3}{3\,\sigma^4} (\bar{x} - \mu)^2 \right] \frac{s}{\sqrt{n}} .$$

Po úpravě vyjde **Johnsonův interval spolehlivosti**

$$\left(\bar{x} + \frac{\hat{m}_3}{6ns} \right) - t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} < \mu < \left(\bar{x} + \frac{\hat{m}_3}{6ns} \right) + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} .$$

Jisté její nedokonalosti odstraňuje **Hallova transformace**

$$t_H = K + \frac{g_1(x)\,K^2}{3} + \frac{g_1(x)^2\,K^3}{27} + \frac{g_1(x)}{6\,n} ,$$

resp. $$t_{H1} = \frac{g_1(x)}{6\,n} + \frac{3\sqrt{n}\,\exp\left(\frac{2\,K\,g_1(x)}{3\sqrt{n}} - 1 \right)}{2\,g_1(x)} , \quad \text{kde } K = (\bar{x} - \mu)/s .$$

Johnsonova a Hallova transformace násobené faktorem $n^{0.5}$ splňují podmínku

$$P(t_J \leq x) \approx \Phi(x) , \qquad (37)$$

tj. vedou k přibližné normalitě, redukci šikmosti a jsou invertovatelné.

Inverzní forma statistiky t_H se zahrnutou násobivou konstantou má tvar

$$t_H^{-1}(y) = \frac{3\sqrt{n}}{g_1(x)} \left[\left(1 + g_1(x) \left[\frac{y}{\sqrt{n}} - \frac{g_1(x)}{6n} \right] \right)^{1/3} - 1 \right].$$

Při sledování úrovně stopových prvků a škodlivin je zajímavý pouze pravostranný interval spolehlivosti (jednostranný interval zprava, tj. horní hranice střední hodnoty).

Tento interval se u rozdělení zešikmených vpravo používá často k určení povolené horní hranice, například znečištění.

Pro horní mez pravostranného intervalu spolehlivosti pak platí, že

$$\mu \leq \bar{x} + t_H^{-1}(z_{1-\alpha}) \frac{s}{\sqrt{n}} .$$

Místo transformace dle vztahu pro t_H^{-1} lze použít zjednodušenou verzi

$$t_a^{-1}(y) = y - \frac{g_1(x)\ (y^2/3 + 1/6)}{\sqrt{n}} .$$

Tato transformace se dosadí do uvedené nerovnosti pro μ.

Nejistota pokrytí zprava NP vyjadřuje pravděpodobnost, že skutečná střední hodnota je nižší než meze intervalu spolehlivosti.

Nejistota pokrytí zleva NL určuje pravděpodobnost, že skutečná střední hodnota je vyšší než meze intervalu spolehlivosti.

Nejistota pokrytí z obou stran NC je pak sjednocení obou chyb pokrytí, tj. $NC = NP + NL$.

Pro širokou třídu rozdělení bylo nalezeno, že

$$NP = \alpha/2 + [-0.73 + 0.71 \exp(-\alpha/2)] \, g_1/\sqrt{n}$$

a

$$NL = \alpha/2 + [0.19 + 0.026 \ln(\alpha/2)] \, g_1/\sqrt{n} \, .$$

Potřebná velikost výběru, aby byla zachována nejistota pokrytí: je rozdíl mezi požadovanou pravděpodobností pokrytí (například 0.95) a dosaženou pravděpodobností pokrytí (například 0.94).

Postup: fixuje se nejistota pokrytí na zvolené hodnotě a pro známé n i $g_1(x)$ se nalezne pravděpodobnost α^* pro výpočet kvantilu Studentova rozdělení a takto opravené kvantily se dosadí do vztahů pro určení *CI*.

Klasický pravostranný interval spolehlivosti má tvar

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

	Originální data	Cenzorovaná data	Uřezaná data	Winsorizovaná data
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

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

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

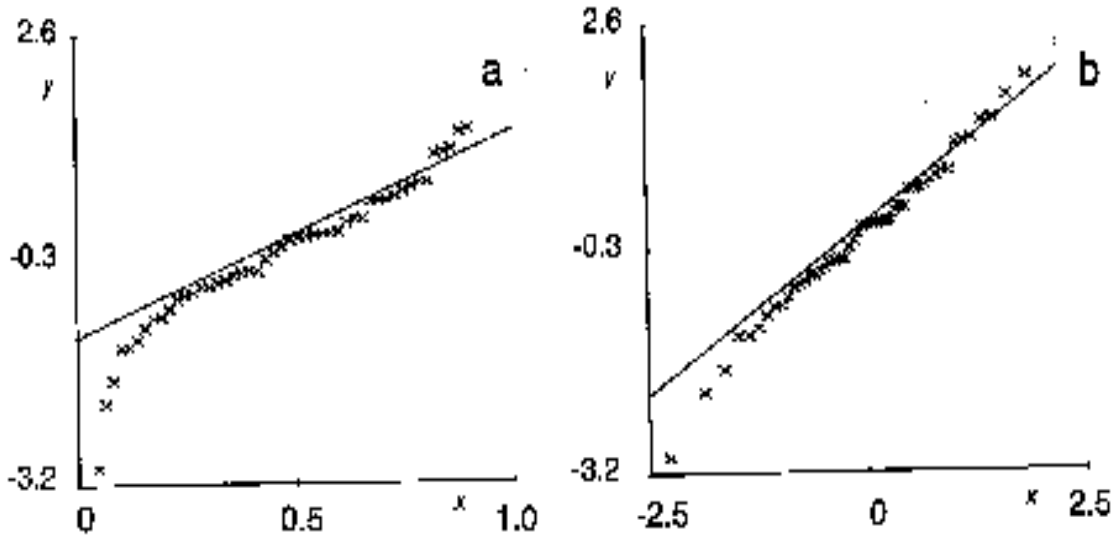
2) **Odhady parametrů polohy a rozptýlení:** odhad střední hodnoty odpovídá absolutnímu členu a odhad směrodatné odchylky směrnicí b regresní přímky.

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.

	Originální data	Cenzorovaná data	Uřezaná data	Winsorizovaná data
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.

PŘÍKLAD 1. *Určení koncentrace nečistot v surovině*

Koncentrace nečistot v μg/g: **DL, DL, 1.24, 1.49, 1.50, 1.56, 1.61, 1.78,** kde **DL** značí pod limitou detekce $x_D = 1$ μg/g. **Cílem:** odhad střední hodnoty, rozptylu a intervalu spolehlivosti pro normální rozdělení.

Řešení:

1. metoda: Postup s vynecháním hodnot pod mezí detekce (nevhodné, chybné!)

Průměr = 1.53 a výběrová směrodatná odchylka $s = 0.18$. Kvantil t rozdělení $t_{0.975}(5) = 2.571$ a 95% interval spolehlivosti ***UC* = 1.72, *LC* = 1.34.**

2. metoda: Maximalizace věrohodnostní funkce

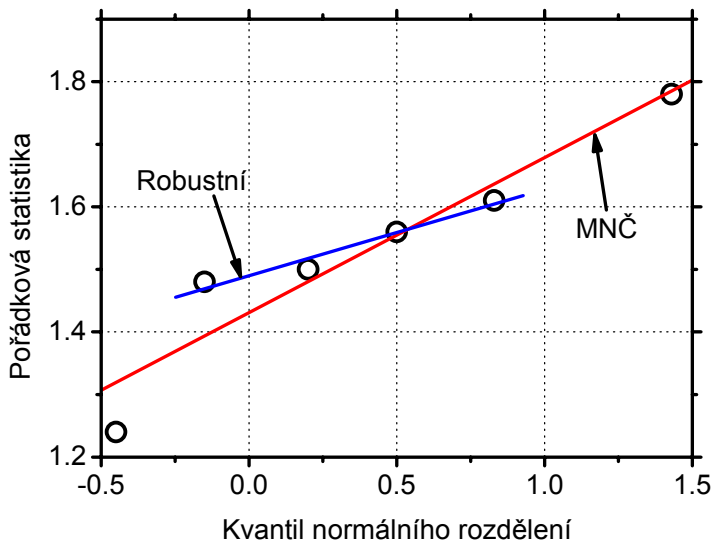
Parametr h = 0.25, parametr g = 0.11 a tabelovaná hodnota $\lambda = 0.3387$.
Průměr = 1.35 a výběrová směrodatná odchylka $s = 0.36$ a 95% interval spolehlivosti ***UC* = 1.72, *LC* = 0.98.**

3. metoda: Jednokroková aproximace maximalizace věrohodnostní funkce

Průměr = 1.46 a výběrová směr. odchylka $s = 0.20$ a 95% interval spolehlivosti ***UC* = 1.67, *LC* = 1.25.**

4. metoda: Pořádkové statistiky Ze směrnice a úseku určených klasickou MNČ vyšlo: **Průměr = 1.43** a výběrová směrodatná odchylka $s = 0.24$ a 95% interval spolehlivosti ***UC* = 1.68, *LC* = 1.18.**

Rankitový graf spolu s regresními přímkami je pro **klasickou** a **robustní MNČ**.



Závěr: Je patrné, že **postupy beroucí v úvahu limitu detekce** vedou k výrazně nižší dolní mezi intervalu spolehlivosti.

Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy **bez hlubšího rozboru** zde může vést ke katastrofálním závěrům.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Environmental Pollution ■■ (2005) ■■■–■■■
ENVIRONMENTAL POLLUTION
www.elsevier.com/locate/envpol

The analysis of soil cores polluted with certain metals using the Box–Cox transformation

Milan Meloun [a,*], Milan Sáňka [b], Pavel Němec [b], Soňa Křítková [a], Karel Kupka [c]

[a] *Department of Analytical Chemistry, University of Pardubice, CZ532 10 Pardubice, Czech Republic*
[b] *Central Institute for Supervisiting and Testing in Agriculture Division of Agrochemistry, Soil and Plant Nutrition, Hroznová 2, CZ656 06 Brno - Pisárky, Czech Republic*
[c] *Trilobyte Statistical Software Ltd., CZ530 02 Pardubice, Czech Republic*

Received 21 July 2004; accepted 28 January 2005

A new procedure of statistical analysis, with exploratory data diagnostics and Box–Cox transformation was used.

Abstract

To define the soil properties for a given area or country including the level of pollution, soil survey and inventory programs are essential tools. Soil data transformations enable the expression of the original data on a new scale, more suitable for data analysis. In the computer-aided interactive analysis of large data files of soil characteristics containing outliers, the diagnostic plots of the exploratory data analysis (EDA) often find that the sample distribution is systematically skewed or reject sample homogeneity. Under such circumstances the original data should be transformed. The Box–Cox transformation improves sample symmetry and stabilizes spread. The logarithmic plot of a profile likelihood function enables the optimum transformation parameter to be found. Here, a proposed procedure for data transformation in univariate data analysis is illustrated on a determination of cadmium content in the plough zone of agricultural soils. A typical soil pollution survey concerns the determination of the elements Be (16 544 values available), Cd (40 317 values), Co (22 176 values), Cr (40 318 values), Hg (32 344 values), Ni (34 989 values), Pb (40 344 values), V (20 373 values) and Zn (36 123 values) in large samples.

Keywords: Data transformation; Exploratory analysis; Soil pollution; Risk element contents

Table 1
Survey of summary statistics for the elements Be, Cd, Co, Cr, Hg, Ni, Pb, V and Zn including classical and robust measures of central tendency, measures of variability, and measures of shape

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16 544	40 317	22 176	40 318	32 344	34 989	40 344	20 373	36 123
Minimum x_1	0	0	0.2	0.1	0	0.1	0.17	0.37	0.7
Maximum x_n	9.33	28.1	110.5	1577.4	69.086	662.0	1121.0	86.0	2070.0
Lower quartile F_L	0.32	0.14	3.9	3.2	0.06	3.0	11.7	7.0	12.0
Upper quartile F_U	0.57	0.27	6.7	6.9	0.11	7.3	19.4	13.0	22.0
Interquartile range $F_U - F_L$	0.25	0.13	2.8	3.7	0.05	4.3	7.7	6.0	10.0
Classical estimates of location, scale and shape									
Sample mean $\bar{x}$	0.470 ± 0.004	0.238 ± 0.003	5.593 ± 0.039	7.104 ± 0.170	0.105 ± 0.006	6.033 ± 0.081	18.637 ± 0.299	10.878 ± 0.083	19.354 ± 0.234
Standard deviation s	0.264	0.300	2.930	17.35	0.534	7.728	30.594	6.015	22.73
Skewness $\hat{g}_1$	5.99	30.74	4.19	40.09	107.88	34.49	19.77	2.16	34.20
Kurtosis $\hat{g}_2$	119	2123.1	89.85	2608.52	12963.7	2298.8	528.2	12.41	2265.0
Robust estimates of location									
Median $\tilde{x}_{0.5}$	0.43 ± 0.01	0.19 ± 0.00	5.0 ± 0.0	4.60 ± 0.05	0.08 ± 0.00	4.70 ± 0.05	14.90 ± 0.05	9.60 ± 0.10	16.0 ± 0.05
Trimmed mean $\bar{x}(10\%)$	0.449 ± 0.003	0.210 ± 0.001	5.356 ± 0.033	5.361 ± 0.040	0.086 ± 0.001	5.320 ± 0.039	15.860 ± 0.067	10.320 ± 0.074	17.446 ± 0.089
Trimmed mean $\bar{x}(20\%)$	0.443 ± 0.003	0.203 ± 0.001	5.264 ± 0.032	5.072 ± 0.033	0.083 ± 0.001	5.109 ± 0.037	15.548 ± 0.063	10.050 ± 0.072	17.020 ± 0.085
Trimmed mean $\bar{x}(40\%)$	0.438 ± 0.003	0.194 ± 0.001	5.150 ± 0.030	4.795 ± 0.030	0.081 ± 0.001	4.883 ± 0.036	15.214 ± 0.061	9.752 ± 0.067	16.531 ± 0.084
Jarque–Berra normality test, critical value for $\alpha = 0.05$ is $\chi^2_{0.95}(2) = 5.99$									
Testing criterion C_1	157.1	291.1	145.9	311.1	382.0	294.3	259.3	111.4	294.9
Normality is	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected
Homogeneity test									
Number of outliers	265	2095	496	2285	1180	1128	1359	486	961
Box–Cox transformation									
Re-transformed mean $\bar{x}_R$	0.427 ± 0.003	0.187 ± 0.001	5.078 ± 0.030	4.922 ± 0.023	0.082 ± 0.001	4.797 ± 0.020	15.172 ± 0.050	9.611 ± 0.050	16.360 ± 0.050

Of particular interest here are sample size, minimum and maximum values within the sample, and both quartiles. The most rigorous estimates of location are re-transformed means after Box–Cox transformation.

Clin Chem Lab Med 2001; 39(3):244–255

Exploratory Biochemical Data Analysis: a Comparison of Two Sample Means and Diagnostic Displays

Milan Meloun[1], Martin Hill[2] and David Cibula[3]

[1]Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
[2]Institute of Endocrinology, Prague, Czech Republic
[3]Department of Obstetrics and Gynecology, General Teaching Hospital in Prague, Charles University Institute of Endocrinology, Prague, Czech Republic

1. Introduction

Statistics, when correctly used, can be a useful and constructive tool in the analysis of biochemical and clinical data; in careless or unscrupulous hands, however, it can be a dangerous weapon. Used properly, statistics will allow an investigator to quantify concepts and conclusions, and help both to take into account sources of systematic variation and to minimize the ef-

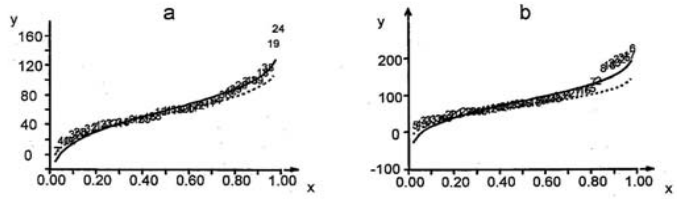


Fig. 2 The quantile plot (*x* axis: the order statistic $x_{(i)}$, *y* axis: the rank probability P_i) for (a) SHBG 0 data, (b) SHBG 1 data.

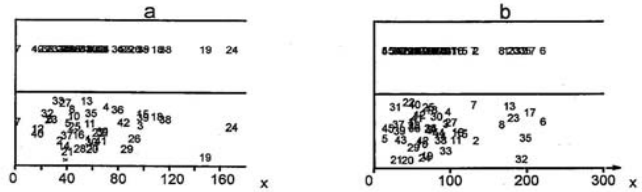


Fig. 3 The dot and jitter dot diagrams (*x* axis: the order statistic $x_{(i)}$, *y* axis: random variable) for (a) SHBG 0 data, (b) SHBG 1 data.

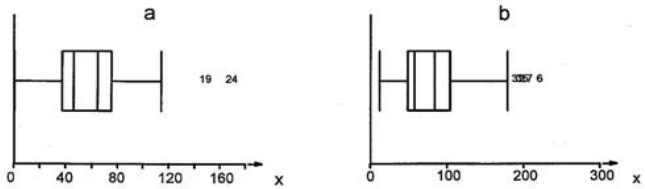


Fig. 4 The box-and-whisker plot (*x* axis: the order statistic $x_{(i)}$, *y* axis: no variable, diagram) for (a) SHBG 0 data, (b) SHBG 1 data.

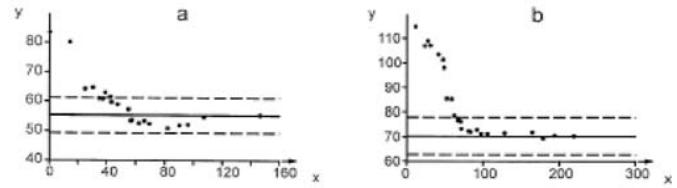


Fig. 5 The halfsum plot (*x* axis: the order statistic $x_{(i)}$, *y* axis: the halfsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$) for (a) SHBG 0 data, (b) SHBG 1 data.

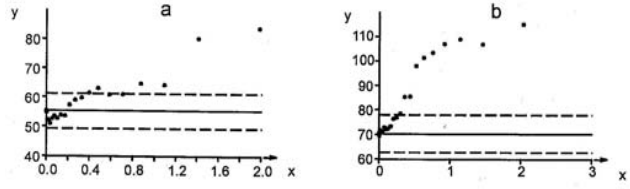


Fig. 6 The symmetry plot (*x* axis: the quantile of normalized Gaussian distribution u_{P_i} for $P_i = i/(n+1)$, *y* axis: the halfsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$) for (a) SHBG 0 data, (b) SHBG 1 data.

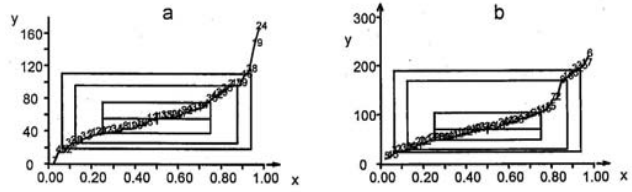


Fig. 7 The quantile-box plot (*x* axis: $P_i = (i - 1/3)/(n + 1/3)$, *y* axis: $x_{(i)}$) for (a) SHBG 0 data, (b) SHBG 1 data.

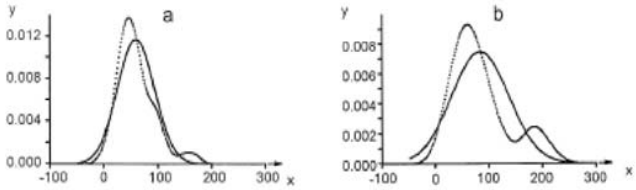


Fig. 8 The kernel estimate of the probability density plot (*x* axis: x_i, *y* axis: the kernel estimate $\hat{f}(x_i)$) for (a) SHBG 0 data, (b) SHBG 1 data.

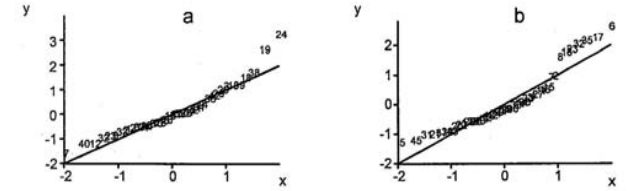


Fig. 9 The quantile-quantile plot (*x* axis: the theoretical quantile function $Q_1(P_i)$, *y* axis: the order statistic $x_{(i)}$) for (a) SHBG 0 data, (b) SHBG 1 data.

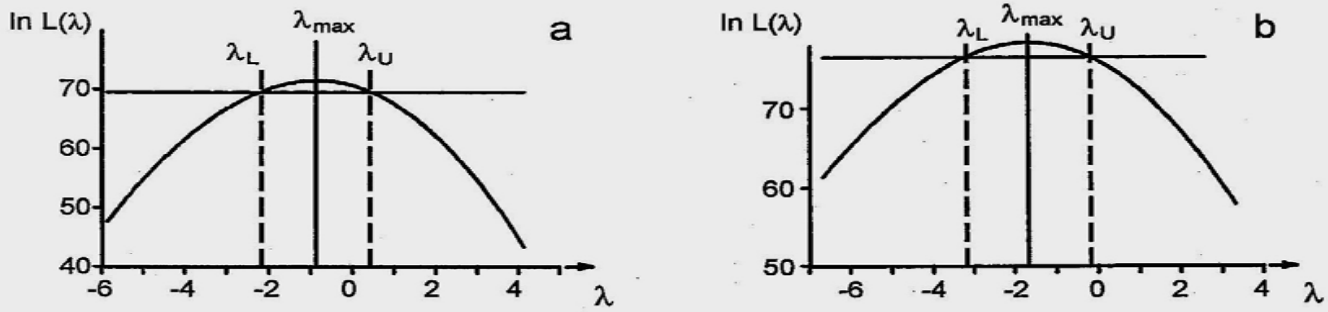


Fig. 11 The plot of the logarithms of the likelihood function ln $L(\lambda)$ in dependence on the power λ and estimation of the optimal power λ_{max} with its lower λ_L and upper λ_U limits of the confidence interval for the statistical probability 95% (x axis: ln $L(\lambda)$, y axis: λ) for (a) SHBG 0 data, (b) SHBG 1 data.

Tab. 7 A comparison of two sample means for selected steroids with the use of Student t-test of eq. [4].

Steroid	n	Mean (lower; upper limits)	SD	Median	Re-transf. mean (lower; upper limits)	Re-transf. stan. dev.	Skew-ness	Kurto-sis	Normality	Test of H_0: equal variances	Test of H_0: equal means
TESTO-0	43	2.10 (1.78; 2.42)	1.05	2.00 (1.66; 2.34)	1.98 (1.67; 2.31)	1.04	0.57	2.85	Accepted	H_0 is accepted	H_0 is accepted
TESTO-1	46	1.91 (1.61; 2.21)	1.01	1.65 (1.37; 1.93)	1.71 (1.46; 1.99)	0.89	**1.12**	4.02	Rejected		
SHBG-0	42	59.4 (48.7; 70.0)	34.19	55.4 (44.1; 66.6)	54.7 (44.7; 65.6)	33.4	**1.08**	**4.38**	Rejected	H_0 is rejected	H_0 is rejected
SHBG-1	45	84.3 (68.2; 100.4)	53.4	70.4 (55.8; 85.0)	72.9 (59.5; 88.3)	47.9	**1.00**	3.12	Accepted		
ADION-0	42	9.22 (7.66; 10.79)	5.03	8.23 (6.98; 9.47)	8.04 (6.91; 9.38)	1.71	**1.71**	**6.52**	Rejected	H_0 is accepted	H_0 is accepted
ADION-1	46	9.66 (8.45; 10.86)	4.06	9.71 (8.25; 11.16)	9.25 (8.08; 10.49)	4.06	0.37	2.48	Accepted		
DHEAS-0	43	6.30 (5.45; 7.15)	2.76	5.70 (4.35; 7.05)	5.88 (5.11; 6.73)	2.62	0.67	2.69	Accepted	H_0 is accepted	H_0 is accepted
DHEAS-1	47	6.55 (5.64; 7.46)	3.09	5.75 (4.97; 6.53)	5.83 (5.10; 6.68)	1.67	**0.93**	3.04	Rejected		
DHEA-0	43	7.99 (5.89; 10.10)	6.83	6.70 (5.84; 7.56)	6.45 (5.58; 7.51)	0.71	**4.59**	**26.87**	Rejected	H_0 is accepted	H_0 is accepted
DHEA-1	46	7.34 (6.12; 8.56)	4.09	6.63 (5.67; 7.58)	6.43 (5.42; 7.60)	3.66	**0.93**	3.01	Rejected		

Clin Chem Lab Med 2001; 39(1):53–61

Analysis of Large and Small Samples of Biochemical and Clinical Data

Milan Meloun[1], Martin Hill[2], Jiří Militký[3] and Karel Kupka[4]

[1] Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
[2] Institute of Endocrinology, Prague, Czech Republic
[3] Department of Textile Materials, Technical University, Liberec, Czech Republic
[4] Trilobyte Statistical Software Ltd., Pardubice, Czech Republic

With the advent of computers and sophisticated analytical instruments, the evaluation and interpretation of results seems to be the main problem. Due to the well-known fact that much experimental data in biochemistry exhibits a non-normal asymmetric distribution, classical analyses based on the assumption of normality cannot be employed; moreover, measurements are often corrupted by outliers. Tukey (1) has claimed that the techniques allowing the isolation of certain basic statistical features and patterns of data can be collec-

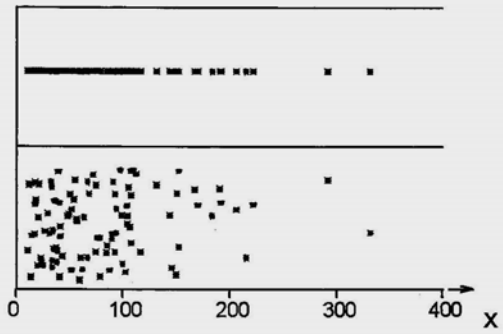


Fig. 2 Dot and jitter dot diagram of pregnenolone data.

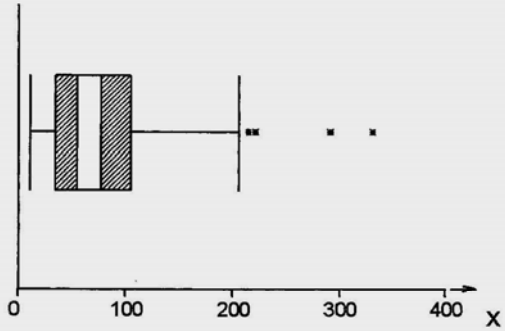


Fig. 3 Box-and-whisker plot of pregnenolone data.

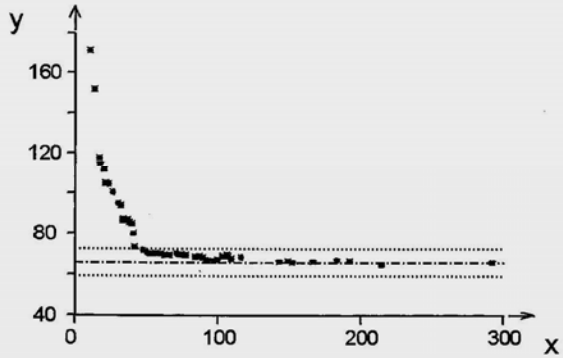


Fig. 4 Midsum plot of pregnenolone data.

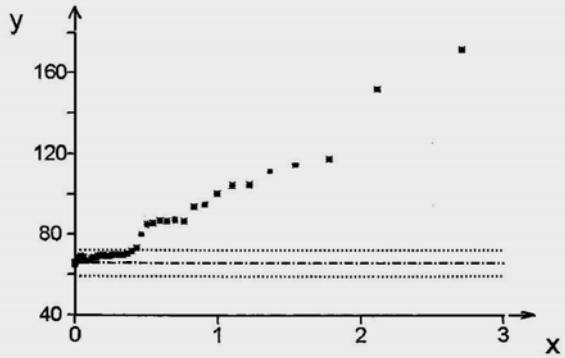


Fig. 5 Symmetry plot of pregnenolone data.

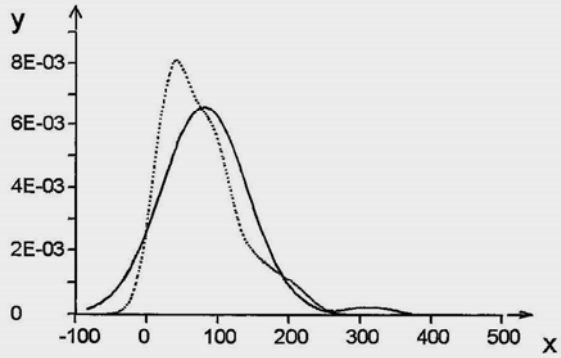


Fig. 6 The kernel estimator of the probability density plot of pregnenolone data, showing the empirical curve (dot curve) and the normal distribution approximation (full curve).

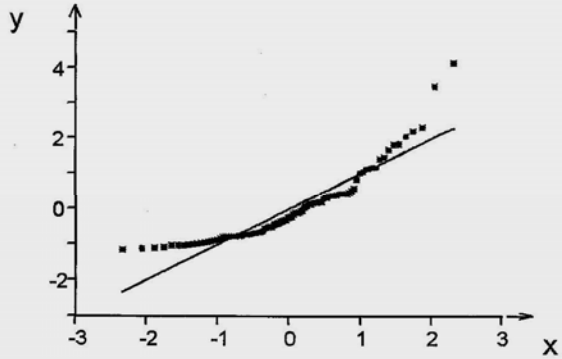


Fig. 7 Quantile-quantile plot (for normal distribution called a rankit plot) of pregnenolone data.