MASARYKOVA UNIVERZITA

Z1069 Statistické metody a zpracování dat

IV. Odhady parametrů

K čemu to je dobré?

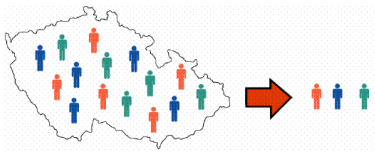
Obvyklým případem při zpracování hromadných jevů je, že máme poměrně malý počet pozorování nějaké veličiny a chceme učinit závěry o tom, co bychom obdrželi, kdybychom měli pozorování mnohokrát více.

Cílem je ukázat:

1) Jaké vlastnosti má mít (náhodný) výběr
2) Jaké vlastnosti (rozdělení) mají výběrové statistiky
3) **Jak lze odhadnout parametry základního souboru ze souboru výběrového**

Výběrové metody zkoumání

- **Základní soubor** (populace) a jeho parametry
- **Výběrový soubor** a jeho statistiky



Z **výběru** provádíme úsudky (**odhady)** skutečných hodnot parametrů základního souboru

To si můžeme dovolit pokud má výběrový soubor stejné vlastnosti jako soubor základní (je reprezentativní)

Toho dosáhneme **náhodným výběrem**

Jaké jsou **důvody**, proč pracujeme s výběrovými soubory?

Odhadování jako základ statistického usuzování

Používáme statistickou indukci - usuzujeme z části (výběr) na celek (základní soubor).

Odhad neznámých parametrů základního souboru provádíme:

1) na základě statistických charakteristik výběru.
2) na základě jistých předpokladů o jejich rozdělení

Vztahy mezi základním souborem a výběry

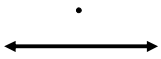
Základní pojmy a symboly

tohle neznáme a proto to odhadujeme

	Základní soubor	Výběrový soubor
• rozsah	N	n
• i-tý prvek	a_i	x_i
• aritmetický průměr	μ	$\bar{x}$
• směrodatná odchylka (rozptyl)	σ (σ^2)	s (s^2)

Odhady parametrů základního souboru: $\hat{\mu}$ $\hat{\sigma}$

Dva způsoby odhadu parametrů základního souboru



- bodový odhad
- intervalový odhad

Nejčastěji odhadujeme střední hodnotu a rozptyl základního souboru

K odhadování potřebujeme poznatky o teoretických rozděleních

Odhady parametrů základního souboru z výběru provádíme s určitou **pravděpodobností (přesností, spolehlivostí)**

Bodový odhad parametrů základního souboru

Je to odhad parametru základního souboru (střední hodnoty, rozptylu) z výběrového souboru pomocí jedné hodnoty.

Bodový odhad aritmetického průměru základního souboru

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

Průměr výběrového souboru **je nezkresleným odhadem** střední hodnoty (průměru) základního souboru

Bodový odhad parametrů základního souboru

Bodový odhad směrodatné odchylky základního souboru

Určuje se z odchylek jednotlivých prvků od výběrového průměru. Pro n-1 **stupňů volnosti** platí:

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

Směrodatná odchylka (rozptyl) výběrového souboru **není nezkresleným odhadem** směrodatné odchylky (rozptylu) základního souboru

(poznámka)

Stupně volnosti

Máme odhad aritmetického průměru a platí následující výraz:

$$\sum_{i=1}^{n} x_i = n \cdot \bar{x}$$

K určení hodnoty $\hat{\sigma}$ lze tedy využít pouze *(n-1)* nezávislých členů tzv. **stupňů volnosti**

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

Odhadem průměru „ztrácíme“ jeden nezávislý „pokus“

Příklad:

- průměr vypočtený ze tří měření je 5
- dvě náhodná (nezávislá) měření budou 4 a 5
- zbývající třetí měření musí být 6, aby byl průměr roven 5, tedy není nezávislé

Bodový odhad parametrů základního souboru

Směrodatná odchylka základního souboru

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

Směrodatná odchylka výběrového souboru

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

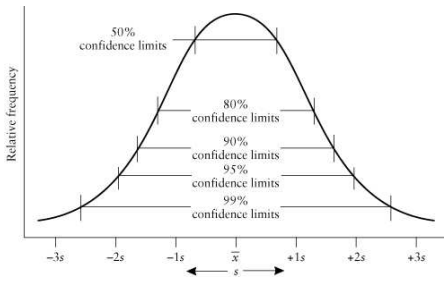
Pro malá n (n<30) bychom měli výběrovou směrodatnou odchylku *s* počítat podle výše uvedeného vzorce.

Bodový odhad je odhad jedním číslem. Nevíme jak je toto číslo „**přesné**“ či „**spolehlivé**“.

Výhodnější je odhad pomocí **intervalu**, který bude s vysokou **pravděpodobností** obsahovat neznámý parametr.

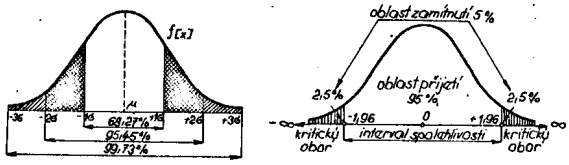
Pomocí výběrového průměru konstruujeme interval uvnitř kterého se skutečná hodnota populačního průměru nachází s vysokou pravděpodobností

Intervalový odhad, intervaly spolehlivosti



Z vlastností normálního rozdělení lze pomocí hodnoty aritmetického průměru a násobků směrodatné odchylky určit **meze, které vyjadřují pravděpodobnosti, s nimiž dané hodnoty leží v určitém intervalu**

Intervaly spolehlivosti



Vnitřní interval vymezený jistým násobkem směrodatné odchylky se označuje jako **interval spolehlivosti**.

Odchylky od průměru, které se nacházejí uvnitř tohoto intervalu označujeme jako **odchylky přípustné**, nevýznamné.

Analogicky jsou definovány **odchylky významné**.

Meze spolehlivosti dále vymezují tzv. **kritický obor (**oblast zamítnutí) a **oblast přijetí**.

Intervaly spolehlivosti

Šířku intervalu spolehlivosti volíme podle povahy problému a závisí také na rozsahu náhodného výběru. Nejčastěji používané intervaly:

Násobky s	Oblast přijetí	Oblast zamítnutí
1,960	95 %	5 %
2,576	99 %	1 %
3,291	99,9%	0,1 %

Interpretace intervalů spolehlivosti: 95 % interval spolehlivosti stanovený na základě náhodného výběru zahrne s pravděpodobností 95 % skutečnou hodnotu odhadovaného parametru.

Intervalový odhad parametrů základního souboru

Na rozdíl od bodového odhadu zde určujeme interval, v němž se zadanou pravděpodobností leží odhadovaný neznámý parametr.

Intervalový odhad se liší podle rozsahu souboru a také podle toho, jaké parametry známe.

Dále budeme značit:

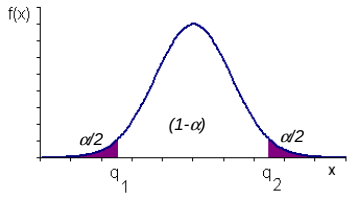
q_1, q_2 - krajní hodnoty intervalu spolehlivosti – meze spolehlivosti

α – **hladina významnosti** - pravděpodobnost, že skutečný parametr základního souboru není z intervalu spolehlivosti.

(1-α) – **hladina spolehlivosti** (spolehlivost odhadu) – představuje pravděpodobnost, že skutečný parametr základního souboru se nachází uvnitř intervalu spolehlivosti.

Intervalový odhad dvoustranný

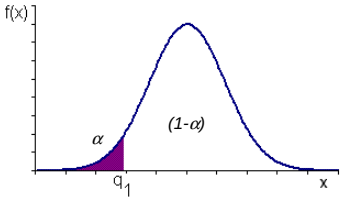
$$P(q_1 \le \mu \le q_2) = 1-\alpha$$



Interpretace: Pravděpodobnost, že parametr μ základního souboru se nachází mezi hodnotami q_1, q_2 je *(1-α)*

Intervalový odhad jednostranný

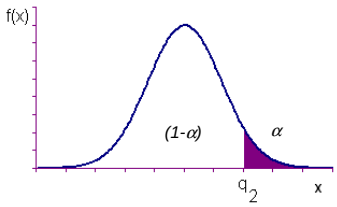
zdola ohraničený $P(q_1 \le \mu) = 1-\alpha$



Interpretace: Pravděpodobnost, že parametr μ základního souboru má větší hodnotu než q_1, je *(1-α)*

Intervalový odhad jednostranný

shora ohraničený $P(\mu \le q_2) = 1-\alpha$



Interpretace: Pravděpodobnost, že parametr μ základního souboru má menší hodnotu než q_2, je *(1-α)*

Intervalový odhad parametru μ pro velké rozsahy výběru (n > 30)

Intervalový odhad lze obecně zapsat: $P(q_1 \le \mu \le q_2) = 1-\alpha$

Pokud známe hodnotu σ hodnoty q_1, q_2 lze určit takto:

$$q_1 = \overline{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \qquad q_2 = \overline{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$z_{1-\frac{\alpha}{2}}$ je příslušný **kvantil normovaného normálního rozdělení** (lze ho najít v tabulkách či vypočítat)

$\frac{\sigma}{\sqrt{n}}$ je tzv. směrodatná chyba odhadu průměru (viz poznámky)

σ – směrodatná odchylka základního souboru
n – rozsah výběru
α – hladina významnosti

Intervalový odhad parametru μ pro velké rozsahy výběru (n > 30)

Pokud neznáme hodnotu σ hodnoty q_1, q_2 lze určit takto:

$$q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} \qquad q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}$$

s – směrodatná odchylka výběrového souboru

Intervalový odhad parametru μ lze potom zapsat:

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

Intervalový odhad parametru μ pro velké rozsahy výběru (n > 30)

Výše uvedená nerovnice je splněna s pravděpodobnosti *(1-α):*

$$P[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

analogicky při neznámém σ

$$P[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}] = 1-\alpha$$

Výraz (delta) $\Delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ se označuje jako **přípustná chyba**

Intervalový odhad parametru *μ* lze jednoduše zapsat jako

$$\mu = \bar{x} \pm \Delta$$

Příklad (1/1): Určete 95% interval spolehlivosti pro průměrnou návštěvnost rekreačního střediska, když pro náhodný výběr 100 návštěvníků je průměrná délka pobytu 2,2 dne a rozptyl délky pobytu **všech** návštěvníků je 0,36

$n = 100$

$\bar{x} = 2{,}2$

$\sigma = \sqrt{0{,}36} = 0{,}6$

$\alpha = 0{,}05$

Z tabulek kvantilů normovaného normálního rozdělení určíme hodnotu z pro α=0,05: $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0{,}05}{2}} = z_{0{,}975} = 1{,}96$

p	z_p	p	z_p	p	z_p	p	z_p
0,50	0,000	0,75	0,674	0,950	1,645	0,975	1,96
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,977
0,52	0,050	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034

dále vypočítáme hranice intervalu spolehlivosti …

Příklad (1/2):

Vypočítáme hranice intervalu spolehlivosti:

$$q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2{,}2 - 1{,}96 \cdot \frac{0{,}6}{\sqrt{100}} = 2{,}0824$$

$$q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2{,}2 + 1{,}96 \cdot \frac{0{,}6}{\sqrt{100}} = 2{,}3176$$

Výsledný intervalový odhad lze zapsat:

$$P(2{,}0824 \le \mu \le 2{,}3176) = 0{,}95$$

Můžeme tvrdit, že s pravděpodobností 95% (na hladině významnosti α=0,05) se průměrná délka pobytu všech návštěvníků rekreačního střediska pohybuje v intervalu <2,0824;2,3176>

Často užívané intervalové odhady parametru *μ*

α=0,1 $P[\bar{x} - 1{,}645\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}645\frac{\sigma}{\sqrt{n}}] = 90\%$

α=0,05 $P[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}] = 95\%$

α=0,01 $P[\bar{x} - 2{,}576\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2{,}576\frac{\sigma}{\sqrt{n}}] = 99\%$

Intervalový odhad parametru μ pro malé rozsahy výběru (n < 30)

V případě výběrů malého rozsahu je nutné nahradit hodnotu jistého kvantilu normovaného normálního rozdělení (z) kritickou hodnotou **t-rozdělení** pro $\nu = n-1$ stupňů volnosti.

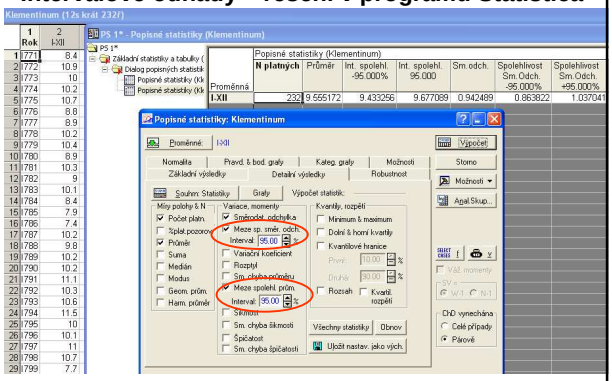
Pokud tedy známe hodnotu rozptylu σ^2 potom pro krajní hodnoty intervalu spolehlivosti *q1, q2* dostáváme:

$$q_1 = \bar{x} - t_{1-\frac{\alpha}{2};(n-1)} \frac{\sigma}{\sqrt{n}} \qquad q_2 = \bar{x} + t_{1-\frac{\alpha}{2};(n-1)} \frac{\sigma}{\sqrt{n}}$$

Pokud neznáme hodnotu rozptylu σ^2 potom použijeme k jeho odhadu výběrové hodnoty *s*:

$$q_1 = \bar{x} - t_{1-\frac{\alpha}{2};(n-1)} \frac{s}{\sqrt{n-1}} \qquad q_2 = \bar{x} + t_{1-\frac{\alpha}{2};(n-1)} \frac{s}{\sqrt{n-1}}$$

Intervalové odhady – řešení v programu Statistica



Intervalový odhad parametru σ² základního souboru

Předpokládáme, že základní soubor má normální rozdělení. Intervalový odhad bude mít obecný tvar:

$$P(q_1 \le \sigma^2 \le q_2) = 1-\alpha$$

Intervalový odhad se opírá o poznatek rozdělení výběrového rozptylu, že totiž náhodná veličina $(n-1)s^2/\sigma^2$ má χ^2 rozdělení s $\nu = n-1$ stupni volnosti.

Hodnoty q_1, q_2 určujeme pomocí odhadnuté hodnoty s z výběrového souboru:

$$q_1 = \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};(n-1)}} \qquad q_2 = \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};(n-1)}}$$

Ze statistických tabulek či s využitím vhodného statistického programu potřebujeme určit kritické hodnoty χ2 rozdělení pro *(n-1)* stupňů volnosti

Intervalový odhad parametru σ² základního souboru

Intervalový odhad parametru σ² lze potom zapsat:

$$P[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};(n-1)}}] = 1-\alpha$$

Odmocněním získáme výraz pro intervalový odhad směrodatné odchylky základního souboru.

Řešení v programu Statistica:

Příklad: Pro výběrový soubor 12 měření výšky vodní hladiny byla zjištěna hodnota rozptylu $s^2 = 0,64$. Určete intervalový odhad rozptylu pro hladinu spolehlivosti 0,90

$$P[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};(n-1)}}] = 1-\alpha$$

Určení rozsahu *n* náhodného výběru

Potřebujeme ho k tomu, abychom z výběru odhadli neznámý průměr s předem zvolenou přesností – tedy aby měl interval spolehlivosti požadovanou šířku.

Rozsah vypočteme ze vztahu pro výpočet tzv. **přípustné chyby** (delta), která je polovinou požadované šířky intervalu spolehlivosti.

$$\mu = \bar{x} \pm \Delta$$

$$\Delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$ z čehož pro n platí: $$n = \left(\frac{\sigma \cdot z_{1-\frac{\alpha}{2}}}{\Delta} \right)^2$$

Určení rozsahu *n* náhodného výběru

Příklad: Z náhodného výběru 60-ti zákazníků hypermarketu jsme zjistili jejich průměrný věk 28 roků. Za předpokladu, že známe směrodatnou odchylku **všech** zákazníků (9 roků) určete:

$n = 60$
$\bar{x} = 28$
$\sigma = 9$
$\alpha = 0,05$

a) 95 % interval pro průměrný věk všech zákazníků

$$(28 - 1,96\frac{9}{\sqrt{60}} \le \mu \le 28 + 1,96\frac{9}{\sqrt{60}})$$

$$(25,7 \le \mu \le 30,3)$$

b) potřebujeme, aby 95 % interval byl pouze plus minus 2 roky. Jak velký výběr je zapotřebí?

Předpokládáme, že přípustná chyba Δ je 2

$$n = \left(\frac{\sigma \cdot z_{1-\frac{\alpha}{2}}}{\Delta} \right)^2 = \left(\frac{9 \cdot 1,96}{2} \right)^2 = 8,82^2 = 78$$

Výběr by musel obsahovat 78 zákazníků

Poznámky

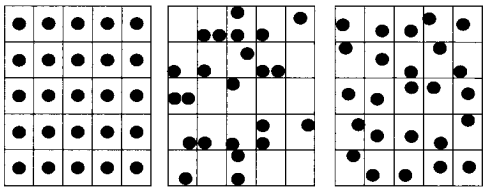
Základní dělení způsobů výběru

Je-li pravděpodobnost každého členu základního souboru, že bude zařazen do výběru, stejná, potom hovoříme o **náhodném** výběru

- prostý náhodný výběr
- výběr s opakováním resp. bez opakování
- výběr oblastní (typický, stratifikovaný)
- výběr systematický (mechanický)
- výběr vícestupňový
- výběr záměrný (subjektivní – ne náhodný)

Techniky losování a generování náhodných čísel k zajištění požadavku náhodnosti výběru

Základní dělení způsobů výběru

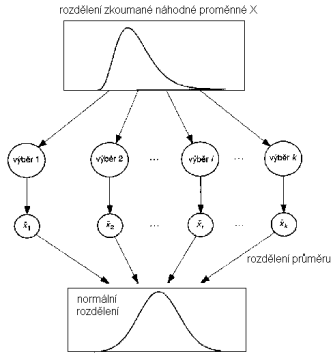


Příklad systematického, náhodného a stratifikovaného náhodného výběru

Výběrové metody souvisí teorií odhadu …

Výběrová rozdělení

Z jistého základního souboru můžeme učinit několik náhodných výběrů – jejich statistické charakteristiky budou odlišné – jsou náhodnými proměnnými.



Průměr výběrových průměrů

$$\mu_{\bar{x}} = \left(\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_{r-1} + \bar{x}_r\right)/r = \frac{1}{r}\sum_{i=1}^{r}\bar{x}_i$$

Směrodatná odchylka výběrových průměrů

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{r}\left(\bar{x}_i - \mu_{\bar{x}}\right)^2}{r}}$$

kde *r* je počet výběrů.

Výběrový průměr a výběrové rozdělení průměrů

V případě velkého rozsahu základního souboru s normálním rozdělením a s parametry μ, σ platí, že **rozdělení výběrových průměrů** je také normální s parametry:

průměr $\mu_{\bar{x}} = \mu$

směrodatná odchylka $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Směrodatná odchylka rozdělení výběrových průměrů je menší než směrodatná odchylka základního souboru a to tím menší, čím větší je rozsah výběru.

(poznámka)

Rozptyl výběrových průměrů

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \left(\sigma^2 + \sigma^2 + \sigma^2 + ... + \sigma^2\right) = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

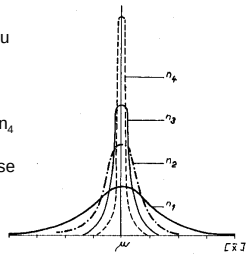
a tedy směrodatná odchylka výběrového průměru:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Vlastnosti parametrů výběrového rozdělení průměrů

Závislost tvaru rozdělení (a také hodnot rozptylu a směrodatné odchylky) na rozsahu výběru

$n_1 < n_2 < n_3 < n_4$



• Bez ohledu na tvar původního rozdělení se rozdělení výběrového průměru blíží k normálnímu rozdělení pro rozsah výběru jdoucí do nekonečna.

• Rozdělení velkého počtu takových výběrových průměrů bude tedy užší než původní rozdělení a bude mít stejný střed.

• Směrodatná odchylka výběrového rozdělení průměrů se nazývá **směrodatná chyba odhadu průměru** (nebo též střední chyba průměru).

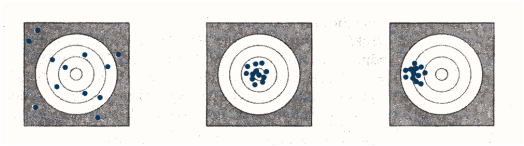
Vlastnosti odhadů ve statistice

• Odhad musí být **konzistentní** – rozdíl mezi odhadnutou a skutečnou hodnotou se zmenšuje s růstem n (rozsah výběru).

• Odhad má být **nezkreslený** (nevychýlený) - všechny odchylky odhadu od skutečné hodnoty se kompenzují (naopak – odhad vychýlený).

• Odhad má být **vydatný** – vydatnou je charakteristika, jejíž rozptyl je ze všech možných výběrů nejmenší

• Odhad neznámých parametrů základního souboru provádíme s jistou **přesností** a **spolehlivostí**.



Přesnost a spolehlivost odhadu

• **Přesnost odhadu** – je dána násobkem směrodatné chyby odhadu (tj. šířkou intervalu spolehlivosti)

• **Spolehlivost odhadu** – je určena pravděpodobností, se kterou je možné určitý odhad považovat za správný.

• Pro určení přesnosti a spolehlivosti je nutná **znalost rozdělení** výběrových charakteristik.

• Pro n > 30 se výběrové rozdělení obvykle považuje za normální. Jiná teoretická rozdělení se používají u malých výběrů.