Statistické metody a zpracování dat

IV. Odhady parametrů

Petr Dobrovolný

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.

- výběrové metody
- bodový a intervalový odhad parametrů základního souboru

Cílem je ukázat, jak odhadnout např. průměr základního souboru ze souboru výběrového

Základní pojmy

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

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

(rozsáhlost, nekonečnost, nákladnost, efektivita, rychlost, …)

Výběrové metody zkoumání

- Souvisí teorií odhadu.
- Používáme statistickou indukci (usuzujeme z části (výběr) na celek (základní soubor).
- Odhad neznámých parametrů základního souboru provádíme na základě statistických charakteristik výběru.

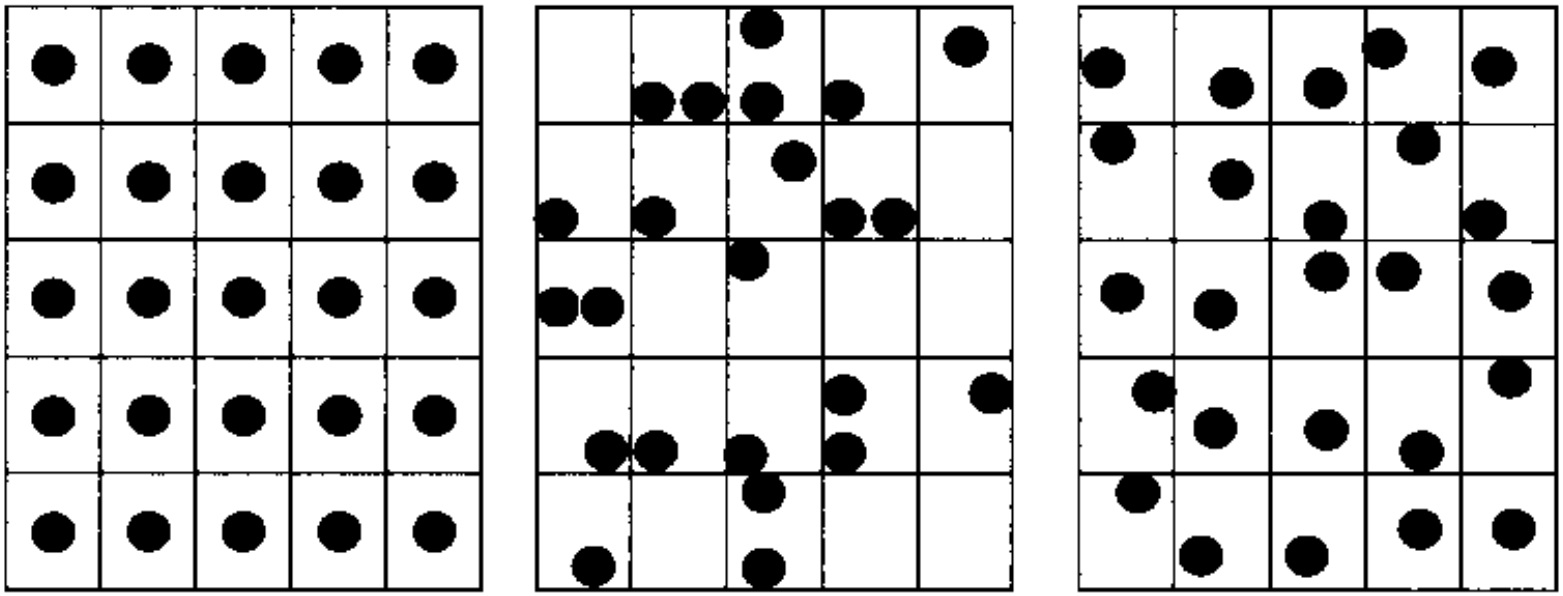
Je-li pravděpodobnost každého členu základního souboru , že bude zařazen do výběru, stejná, potom hovoříme o **náhodném** výběru

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

Základní dělení způsobů 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



Příklad systematického, náhodného a stratifikovaného náhodného výběru

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 střední výběrové chyby (je to směrodatná odchylka příslušné charakteristiky ze všech teoreticky možných výběrů).

• **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.

•Neznámé parametry základního souboru odhadujeme dvěma způsoby

- **bodový odhad**
- **intervalový odhad**

Vztahy mezi základním souborem a výběry

základní pojmy a symboly

	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: $\hat{x}$

$\hat{\sigma}$

Průměr výběrových průměrů

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

kde r je počet výběrů.

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, jeho parametry a rozdělení

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.

Výběrový průměr, jeho parametry a rozdělení

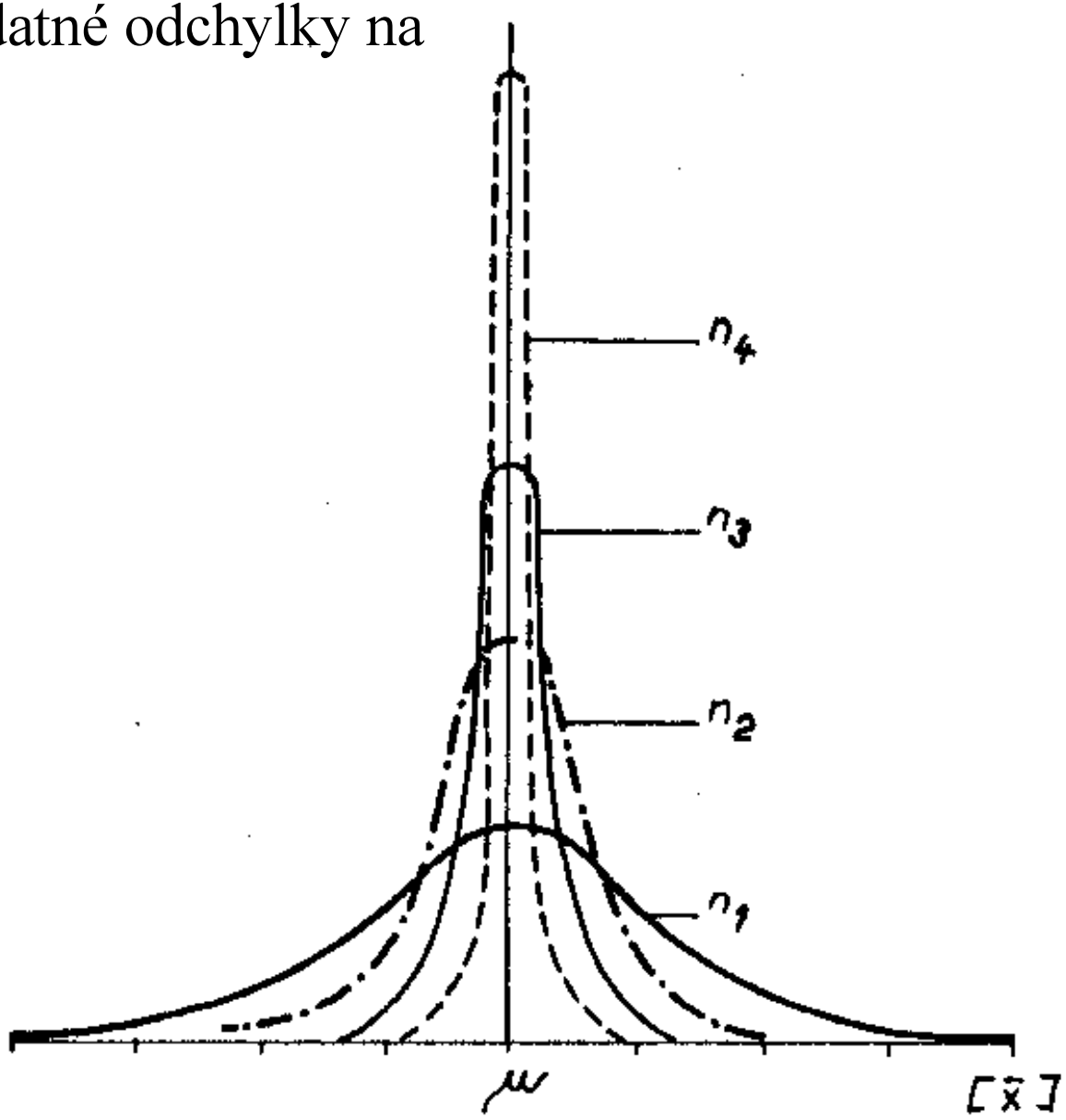
- Matematicky lze dokázat, že směrodatná odchylka rozdělení výběrových průměrů je rovna směrodatné odchylce původního rozdělení vydělené druhou odmocninou rozsahu výběru.
- Směrodatná odchylka výběrového rozdělení průměrů se nazývá **směrodatná chyba průměru** (nebo též střední chyba průměru).

Vztahy mezi výběry a základním souborem

- 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.
- Je rozumné očekávat, že čím větší bude rozsah výběru, tím více se bude průměr výsledného rozdělení blížit středu původního rozdělení a výsledné rozdělení bude užší.

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

$n_1 > n_2 > n_3 > n_4$



Bodový odhad parametrů základního souboru

Bodový odhad aritmetického průměru základního souboru

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

Bodový odhad směrodatné odchylky základního souboru

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

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

Bodový odhad parametrů základního souboru

Je-li výběrová směrodatná odchylka s rovna:

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

potom z toho plyne, že $\hat{\sigma} > s$

Další úpravou lze získat:

$$\frac{\hat{\sigma}}{s} = \sqrt{\frac{\frac{1}{n-1}}{\frac{1}{n}}} \qquad \text{a dále} \qquad \hat{\sigma} = s \cdot \sqrt{\frac{n}{n-1}}$$

Bodový odhad parametrů základního souboru

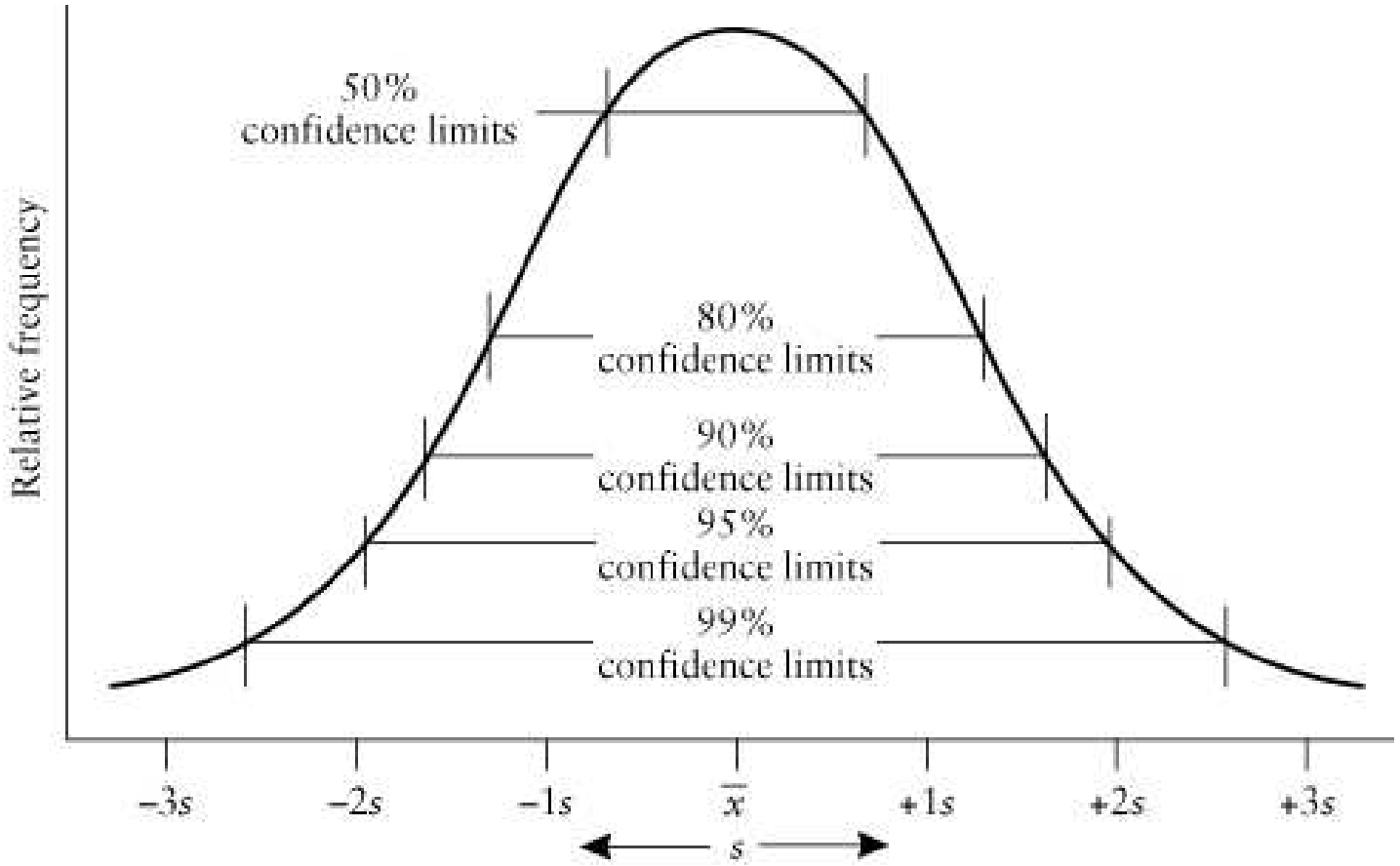
Pro odhad směrodatné odchylky výběrových průměrů:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{a dále} \qquad \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$$

Odhady parametrů základního souboru ($\hat{\mu}, \hat{\sigma}$) se výběr od výběru mění.

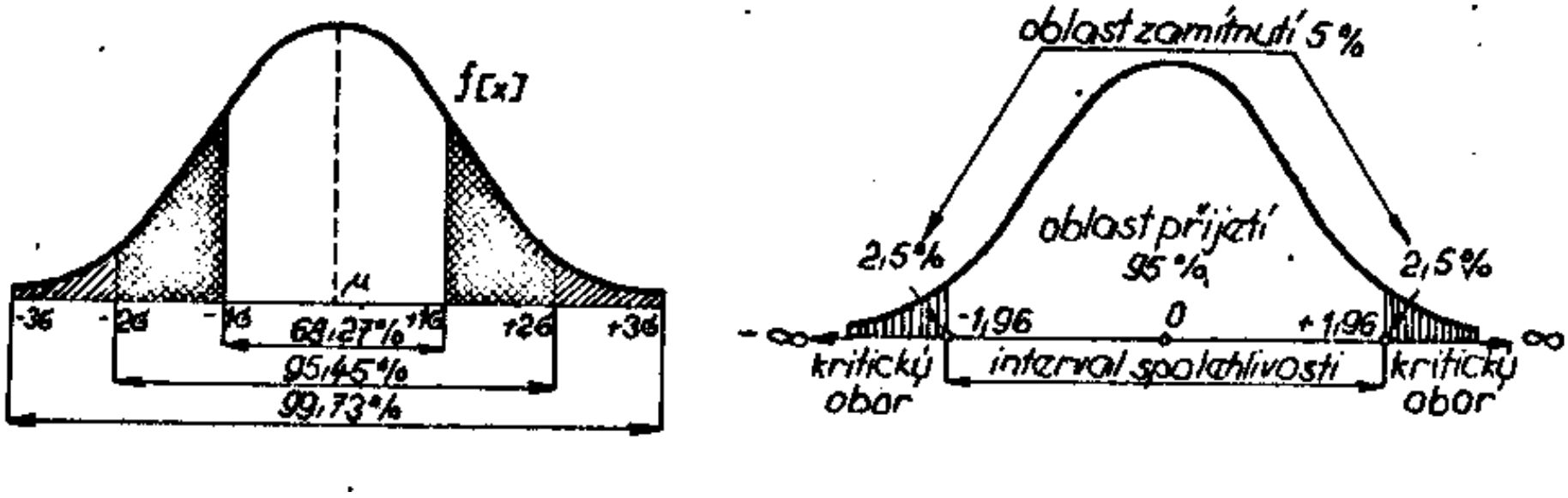
Musíme proto stanovit jejich odchylky od skutečných parametrů (μ, σ) a také určit jejich přesnost odhadu pomocí tzv. **intervalů spolehlivosti.**

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 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 sd	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.

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

Jak plyne z výše uvedeného, rozdělení výběrových průměrů lze považovat za normální s parametry:

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

Intervalový odhad parametrů základního souboru

Například: interval $\mu_{\bar{x}} \pm 2{,}576 \cdot \sigma_{\bar{x}}$ bude zahrnovat 99 % všech výběrových průměrů. Tedy téměř s jistotou bude výběrový parametr $\overline{x}$ součástí intervalu:

$$\mu_{\bar{x}} - 2{,}576 \cdot \sigma_{\bar{x}} \leq \bar{x} \leq \mu_{\bar{x}} + 2{,}576 \cdot \sigma_{\bar{x}}$$

Uvedený násobek směrodatné odchylky lze nahradit hodnotou u_p, kde index p vyjadřuje pravděpodobnost, s níž náhodná veličina překročí kritickou hodnotu (zde $u_{0,01} = \pm 2{,}576$).

Intervalový odhad

Předchozí vztah můžeme dále psát

$$\mu_{\bar{x}} - u_p \cdot \sigma_{\bar{x}} \leq \bar{x} \leq \mu_{\bar{x}} + u_p \cdot \sigma_{\bar{x}}$$

Po dosazení za $\mu_{\bar{x}}$ a $\sigma_{\bar{x}}$ dále:

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

Řešením pro intervalový odhad neznámého průměru μ dostaneme:

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

Intervalový odhad

Hodnotu směrodatné odchylky obvykle neznáme a proto pracujeme s jejím odhadem $\hat{\sigma}$ a s výběrovými charakteristikami:

$$\bar{x} - u_p \cdot \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + u_p \cdot \frac{s}{\sqrt{n-1}}$$

Uvedený výraz definuje **intervalový odhad parametru** μ **základního souboru.**

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 $$n = u_p^2 \cdot \frac{s^2}{\delta}$$

kde δ je polovina požadované šířky intervalu spolehlivosti.

Často definujeme také tzv. **směrodatnou chybu aritmetického průměru** $c_{\bar{x}}$

$$c_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Určení rozsahu *n* náhodného výběru

Potom pravděpodobná chyba výběrového průměru $pc_{\bar{x}}$

$$pc_{\bar{x}} = 0{,}6745 \cdot \frac{s}{\sqrt{n}}$$

Z této rovnice můžeme určit rozsah výběru nutného k odhadu průměru tak, aby jeho chyba měla předem zvolenou velikost:

$$\frac{N'}{n} = \frac{pc_{\bar{x}}^2}{PC_{\bar{x}}^2}$$

zde *n* je rozsah výběru, *N'* je hledaný rozsah výběru, $pc_{\bar{x}}$ vypočtená pravděpodobná chyba a $PC_{\bar{x}}$ zvolená pravděpodobná chyba.

Intervalový odhad parametru μ pro n < 30

V případě výběrů malého rozsahu je nutné nahradit hodnotu u_p kritickou hodnotou t-rozdělení t_p pro $\nu = n - 1$ stupňů volnosti.

Potom dostáváme:

$$\bar{x} - t_p \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_p \frac{s}{\sqrt{n-1}}$$

Intervalový odhad parametru σ pro n < 30

Za předpokladu normálního rozdělení základního souboru náhodná veličina ns^2/σ^2 má χ^2 rozdělení s $\nu = n-1$ stupni volnosti.

Pro určení hranic intervalu spolehlivosti pro parametr normálního rozdělení vycházíme ze vztahu:

$$\chi^2_{1-0,5p} \leq \frac{ns^2}{\sigma^2} \quad \text{a} \quad \chi^2_{0,5p} \geq \frac{ns^2}{\sigma^2}$$

Výrazy na levých stranách označují kritické hodnoty náhodné veličiny χ^2 s $\nu = n-1$ stupni volnosti (uvedeny v tabulkách).

Intervalový odhad σ pro n < 30

Řešením výrazů pro σ^2 dostaneme jeho interval spolehlivosti:

$$\frac{ns^2}{\chi^2_{0,5\,p}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{1-0,5\,p}}$$

Odmocněním získáme výraz pro **intervalový odhad směrodatné odchylky základního souboru.**

Průměrná chyba směrodatné odchylky $c_\sigma = \frac{s}{\sqrt{2n}}$