- Výběrové šetření a výběrový soubor
- Bodový odhad
- Intervalový odhad

5. Bodový a intervalový odhad

Cíl kapitoly

Statistické soubory používané v hospodářské praxi jsou charakteristické svou rozsáhlostí. Informaci obsaženou v těchto souborech je tedy nutno shrnout. V předchozím textu jsme naznačili možnost použití popisných statistických veličin, jako jsou průměr či rozptyl. V případě rozsáhlých souborů je však nemyslitelné, že bychom byli schopni tyto veličiny vypočítat. Je nutno hodnotu popisných veličin nějakým způsobem odhadnout. K tomuto účelu je možno sestrojit jejich bodové či intervalové odhady pořízené na základě tzv. výběrového souboru, který představuje pouze vzorek celého (základního) statistického souboru. Metody vytváření bodových a intervalových odhadů jsou náplní následující kapitoly.

Časová zátěž

6 hodin (2. a 3. týden v listopadu)

Jak jsme již uvedli v první kapitole, výsledkem statistického zkoumání je statistický soubor. Statistické soubory používané v hospodářské praxi jsou charakteristické svou rozsáhlostí (z hlediska počtu statistických jednotek obsažených v tomto souboru). Je tedy nutno informaci obsaženou v takovém statistickém souboru nějakým způsobem shrnout či zhustit. První a druhá kapitola prezentovaly některé používané možnosti, ať už se jednalo o třídění, tabulky četností, či nejrůznější míry (průměry, rozptyly apod.). Všechny uvedené postupy je možno užít především na statistické soubory vzniklé při tzv. **vyčerpávajícím statistickém zjišťování (šetření)**. Tedy při šetření, kdy se prověřují veškeré jednotky, které tvoří statistický soubor. Tento typ šetření je v případě hospodářské statistiky poměrně neobvyklý. Prověřování všech statistických jednotek (např. v souboru obyvatel dané země), s sebou přináší obrovské organizační, časové i finanční problémy. Z těchto důvodů jsou vyčerpávající šetření obvyklou doménou státní statistické služby (např. populační census). Přesnost výsledků získaných při tomto typu šetření je ovšem často převážena zmíněnými náklady. Mnohem obvyklejší proto je využívání jiného typu šetření – výběrového šetření.

5.1 Výběrové šetření a výběrový soubor

Výběrové (nevyčerpávající) šetření je charakteristické tím, že jsou při něm zjišťovány pouze vybrané jednotky statistického souboru. Zjištěné vlastnosti a charakteristiky získané z těchto vybraných jednotek jsou poté zobecňovány a jsou považovány za charakteristiky a vlastnosti celého (základního) statistického souboru.

Příkladů využití výběrových šetření je v současné statistické praxi celá řada. Na makroekonomické úrovni je to například sledování cenového vývoje pomocí výběrových cenových indexů (zmíníme se o nich v druhém dílu učebnice). Na růst (či pokles) cenové hladiny se usuzuje na základě výběrového

šetření cen prováděného jen ve vybraných obchodech a u vybraných výrobků a služeb.

Mezi typická výběrová šetření užívaná spíše v sociologické či politologické oblasti patří běžné průzkumy veřejného mínění, kdy se názor, či nálady obyvatelstva odvozují z názoru malého vzorku obyvatel.

Výběrový soubor

Lze tedy konstatovat, že **výběrový soubor** vzniká jako výsledek **výběrového šetření**. Výběrové šetření je takovým statistickým zjišťováním, kdy ze základního souboru o N prvcích (např. všichni obyvatelé ČR) vybereme jeho část – výběrový soubor – o rozsahu n (tisíc oslovených obyvatel ČR). Takto získaný výběrový soubor zpracujeme (tříděním, výpočtem průměru, kvantilu, atd.) a z výsledků usuzujeme na vlastnosti základního souboru.

Konstrukce výběrového souboru není zpravidla zcela náhodná, či libovolná. Naopak se snažíme o to, aby výběrový soubor skutečně co možná nejlépe odrážel vlastnosti základního souboru. Je proto vhodné dodržovat některé základní postupy a techniky při pořizování těchto souborů.

Metody vytváření výběrového souboru:

- anketa,
- metoda základního masívu,
- záměrný výběr,
- prostý náhodný výběr.

Anketa

Anketa je metodou, kdy výběrový soubor získáváme oslovením vybrané části statistických jednotek. Obvykle bývá ve formě dotazníku, či formuláře, který dotčené osoby (obvykle dobrovolně) vyplňují. Anketa je charakteristická malou návratností stejně jako nízkou reprezentativností pořízených údajů. Obvykle proto není příliš vhodná pro získávání obecně platných (validních) informací.

Základní masív

Metoda základního masívu je využívána u statistických zjišťování, kde základní soubor obsahuje několik významných jednotek a celé řady menších (méně významných). Lze tedy tyto malé jednotky pominout, či zahrnout do šetření v mnohem menší míře než jednotky velké. Proces vytváření výběru se tak výrazně zjednoduší, nicméně i zde musíme mít na vědomí riziko nezáměrného zkreslení (například z důvodu přecenění některých či podcenění jiných jednotek).

Záměrný výběr

Záměrný výběr je obvykle užíván u již zmíněného zjišťování růstu cenové hladiny. Statistický úřad, či skupina odborníků, na základě svého úsudku určí okruh těch jednotek, u kterých bude zjišťování probíhat. Výběr je tedy velmi závislý na subjektivním názoru jeho tvůrců. Je obvykle konstruován dvěma způsoby – jako **typický** a jako **kvótní**. Typický výběr vybírá pokud možno nejcharakterističtější reprezentanty a u nich provádí zjišťování (výběr typických výrobků a služeb určujících spotřební chování obyvatel). Kvótní výběr usiluje o strukturální shodu výběrového a základního souboru.

Náhodný výběr

Náhodný výběr, je výběrem, kdy všechny jednotky základního souboru mají stejnou pravděpodobnost, že do něj budou zahrnuty. Stejně tak je

stejná pravděpodobnost zahrnutí jakékoli kombinace těchto jednotek. Náhodný výběr také musí mít tu vlastnost, že výběr jedné jednotky nemá vliv na výběr jiné jednotky (jedná se tedy o nezávislé náhodné jevy).

Technik pořízení náhodného výběru je celá řada (např. generováním náhodných čísel, mechanickým výběrem atd.). V každém případě patří náhodný výběr k nejužívanějším technikám v hospodářské statistice.

Některé z metod pořizování uvedených typů výběrových souborů naleznete například v učebnici Seger, Hindls: *Statistika v hospodářství* na stranách 131–138, případně v učebnici Hindls, Hronová, Novák: *Analýza dat v manažerském rozhodování* na stranách 58–63.

I výběrový soubor získaný pomocí výše uvedených technik však nemusí být dostatečně přehledný. Stejně jako tomu bylo u popisu základních souborů, i informaci ve výběrovém souboru je nutno nějakým způsobem zhustit, či utřídit. K charakteristikám (mírám), které jsou používány pro základní soubor (zabývali jsme se jimi v kapitole 2), se pro výběrové soubory konstruuje analogický balík statistických veličin. Jsou nazývány **výběrové charakteristiky**, někdy také statistiky.

Zatímco charakteristiky základního statistického souboru jsou přesné a pevné hodnoty dané pouze strukturou a vlastnostmi tohoto souboru, výběrové charakteristiky jsou závislé na konkrétním výběru. Lze říci, že se pro stejný základní soubor mění od jednoho náhodného výběru k druhému a mají tudíž podobu náhodných veličin.

Výběrové rozdělení

S měnícím se náhodným výběrem se tedy mění i příslušné výběrové charakteristiky. Jelikož se jedná, jak jsme již zmínili, o náhodnou veličinu, budeme hledat pravidlo podle nějž se chová – označovali jsme jej jako zákon rozdělení náhodné veličiny. Získáme tak tzv. **výběrové rozdělení**, které je základním podkladem pro zpracování výběrových souborů. Pomocí výběrového rozdělení můžeme posoudit, do jaké míry odráží výběrová charakteristika skutečnost – tedy příslušnou charakteristiku základního souboru.

Výběrové charakteristiky jsou tak určitým typem odhadu charakteristik základního souboru (výběrový průměr je odhadem průměru základního souboru). Tento odhad můžeme vytvořit dvěma způsoby, a to jako

- bodový odhad,
- intervalový odhad.

5.2 Bodový odhad

Pomocí bodového odhadu provádíme odhad neznámé charakteristiky základního souboru jednou hodnotou (bodem). Z hodnot výběrového souboru vypočítáme hodnotu příslušné charakteristiky, kterou poté prohlásíme za odhad odpovídající charakteristiky základního souboru.

Bodový odhad G pomocí hodnoty g zapisujeme jako $\operatorname{odh} G \cong g$ nebo $\operatorname{est} G = g$ a čteme: „odhadem (estimátorem) G je g“.

Například uvažujeme šetření, kdy provádíme průzkum průměrného hrubého příjmu v určité organizaci o 1500 zaměstnancích. Z hodnot, které uvedlo náhodně oslovených 150 zaměstnanců, jsme vypočítali průměrný příjem 14 550 Kč. V tomto případě lze konstatovat, že hodnota 14 550 Kč je bodovým odhadem průměrného hrubého příjmu všech zaměstnanců dané organizace.

Ne každá vypočítaná charakteristika se však stává bodovým odhadem. Bodový odhad by měl splňovat některé vlastnosti, které zaručí jeho vypovídací hodnotu a spolehlivost.

Požadavky na bodový odhad:

- **Nezkreslenost (nestrannost)**
  Aby byla příslušná výběrová charakteristika nezkresleným bodovým odhadem nesmí vést k systematickému zkreslení (nadhodnocení, či podhodnocení) odhadované charakteristiky. Ze statistického hlediska tedy musí být střední hodnota odhadu a odhadované charakteristiky totožná. Tedy platí
  $$E(g) = G.$$
- **Konzistence**
  Pokud nelze splnit podmínku nezkreslenosti, požadujeme, aby byl odhad alespoň konzistentní. Pro konzistentní odhad platí, že s rostoucím rozsahem výběrového souboru roste i pravděpodobnost, že se daný odhad shoduje s odhadovanou charakteristikou. tedy s rostoucím rozsahem výběru se zkreslení snižuje. Tedy platí pro libovolné kladné číslo ε
  $$\lim_{n\to\infty} P(|g - G| < \varepsilon) = 1.$$
  Tedy s rostoucím rozsahem výběru ($n \to \infty$) je pravděpodobnost, že zkreslení $|g - G|$ bude menší než pevně zvolené ε rovna jedné. Jedná se tedy o jev jistý.
- **Vydatnost**
  Splňuje-li více výběrových charakteristik vlastnost nezkreslenosti a konzistence, za bodový odhad prohlásíme tu z nich, která má nejmenší rozptyl. O takové charakteristice pak řekneme, že je vydatným (nejlepším nezkresleným) odhadem charakteristiky základního souboru. Tedy platí
  $$E(g) = G.$$

5.3 Intervalový odhad

Častěji než pomocí jednoho údaje odhadujeme hledanou charakteristiku základního souboru pomocí intervalu hodnot. V tomto případě tedy určíme pouze obor hodnot, ve kterých se hledaná charakteristika nejspíše pohybuje. Informace o tomto intervalu je také z těchto důvodů doplněna o příslušnou pravděpodobnost (spolehlivost), se kterou v něm můžeme očekávat výskyt skutečné charakteristiky.

Konstruujeme tedy **interval spolehlivosti** $\langle G_d, G_h \rangle$, při dané spolehlivosti odhadu $1 - \alpha$. Pro interval spolehlivosti (konfidenční interval) potom platí

$$P(G_d < G < G_h) = 1 - \alpha.$$

Spolehlivost odhadu vždy vychází ze zvolené pravděpodobnosti. Je zřejmé, že s rostoucí spolehlivostí se interval spolehlivosti rozšiřuje a je tedy méně přesný. Při konstrukci intervalu spolehlivosti je tedy nutno vyřešit trade-off mezi přesností a spolehlivostí.

V praktických příkladech je obvyklé stanovit hodnotu pravděpodobnostní hladiny poměrně vysoké. Zpravidla se používá spolehlivosti 0,90, 0,95, resp. 0,99. Hovoříme pak o 90% (devadesátiprocentním), 95% a 99% intervalu spolehlivosti.

Stanovíme-li nejobvyklejší 95% interval spolehlivosti na základě výběrových dat, máme 95% pravděpodobnost, že tento interval pokryje hledanou charakteristiku základního souboru. Jinými slovy na 95% nalezneme skutečnou charakteristiku v tomto intervalu.

Interval spolehlivosti lze konstruovat dvěma způsoby, a to jako:

- **jednostranný**, kdy je dána jen horní nebo dolní mez. Jednostranné intervaly dále můžeme rozlišit na
  - **pravostranné** (dána jen horní mez)
    $P(G \geq G_h) = \alpha$
    $P(G < G_h) = 1 - \alpha$
  - **levostranné** (dána jen dolní mez)
    $P(G \leq G_d) = \alpha$
    $P(G > G_d) = 1 - \alpha$
- **oboustranný**, kdy jsou dány obě meze
  $P(G \leq G_d) = P(G \geq G_h) = \frac{\alpha}{2}$
  $P(G_d < G < G_h) = 1 - \alpha$

Například v případě výše uvedeného šetření, kdy provádíme průzkum průměrného hrubého příjmu v určité organizaci o 1500 zaměstnancích bude levostranný 95% interval spolehlivosti možno formulovat jako „průměrný příjem v organizaci dosahuje s pravděpodobností 95% nejméně 14 550 Kč“. Pravostranný 95% interval spolehlivosti by měl tvar „průměrný příjem je s 95% pravděpodobností nejvýše 14 550 Kč“. Oboustranný 95% interval spolehlivosti by pak mohl mít například tvar $\langle 13500, 15000 \rangle$. *Tedy „průměrný hrubý příjem v dané organizaci se na 95% pohybuje mezi 13 500 Kč a 15 000 Kč“.*

Intervaly spolehlivosti je možno konstruovat pro většinu charakteristik základního souboru. Z hlediska využitelnosti dále naznačíme pouze výpočet intervalů spolehlivosti pro průměr základního souboru. Konstrukce dalších typů intervalů spolehlivosti (např. pro rozptyl nebo pro relativní četnost souboru) naleznete například v učebnici Seger, Hindls, Hronová: *Statistika v hospodářství* na stranách 152–159.

5.3.1 Výpočet intervalu spolehlivosti pro aritmetický průměr

Na základě vlastností výběrových průměrů lze dokázat, že při dostatečně velkém rozsahu výběru je rozdělení výběrových průměrů přibližně normální se střední hodnotou μ a rozptylem σ^2/n. S využitím definice a vlastností náhodné veličiny tedy 95% interval spolehlivosti bude nejmenší interval pod rozdělením výběrové charakteristiky, který odpovídá 95% pravděpodobnosti (uvědomte si, že pravděpodobnost je dána plochou pod křivkou rozdělení náhodné veličiny příslušnou danému intervalu).

Pro dostatečně velký výběr, kde lze předpokládat normální rozdělení nalezneme interval spolehlivosti pro aritmetický průměr jako interval daný vztahem

$$\mu \in \left(\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right),$$

kde

μ	průměr základního souboru
$\overline{x}$	výběrový průměr
$u_{1-\alpha/2}$	$1 - \frac{\alpha}{2}$ procentní kvantil normárního rozdělení
σ	směrodatná odchylka základního souboru (v některých případech ji lze nahradit tzv. výběrovou směrodatnou odchylkou s'_x počítanou z výběrového souboru)
n	počet prvků výběrového souboru
s'_x	výběrová směrodatná odchylka počítaná jako druhá odmocnina výběrového rozptylu pro nějž platí vztah

$$s'^{\,2}_x = \sum_{x=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} = \frac{n}{n-1} s_x^2.$$

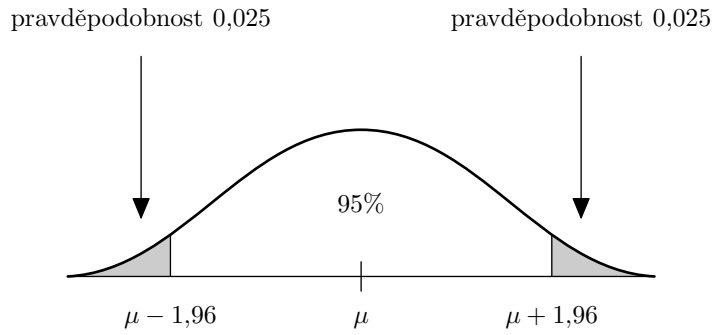
Pro 95% interval spolehlivosti poté dostáváme

$$\left(\overline{x} - u_{1-0,05/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-0,05/2}\frac{\sigma}{\sqrt{n}} \right) = \\ = \left(\overline{x} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}; \overline{x} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \right) = \overline{u} \pm 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

Výraz $\Delta = u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ je označován jako přípustná chyba odhadu, která udává rozpětí intervalu spolehlivosti. Jak je patrné z definičního vztahu, je její hodnota nepřímo úměrná rozsahu výběru n (braného v jeho druhé odmocnině).

Příklad 5.1

Při výzkumu průměrné spotřeby rychlého občerstvení za týden na jednu osobu byla u 180 respondentů zjištěna průměrná hodnota 0,82. Směrodatná odchylka byla zjištěna jako 0,48. Určete 95% a 99% interval spolehlivosti pro odhad průměrné spotřeby rychlého občerstvení za týden v celé populaci.



Obrázek 5.1: Princip konstrukce 95% intervalu spolehlivosti.

Řešení

Zadané hodnoty je $\overline{x} = 0{,}82$, $n = 180$, $s_x = 0{,}48$.

$$\sigma = s'_x = s_x \cdot \sqrt{\frac{n}{n-1}} = 0{,}48 \cdot \sqrt{\frac{180}{179}} = 0{,}483.$$

(pro velký rozsah výběru lze ztotožnit s výběrovou směrodatnou odchylkou)

Intervalový odhad průměru základního souboru vypočítáme podle výše uvedeného vztahu

$$\mu \in \left(\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right),$$

95% interval spolehlivosti má následující podobu

$$\mu = \overline{x} \pm 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} = 0{,}82 \pm 1{,}96 \cdot \frac{0{,}483}{\sqrt{180}} = 0{,}82 \pm 0{,}07.$$

99% interval spolehlivosti má následující podobu

$$\mu = \overline{x} \pm 2{,}58 \cdot \frac{\sigma}{\sqrt{n}} = 0{,}82 \pm 2{,}58 \cdot \frac{0{,}483}{\sqrt{180}} = 0{,}82 \pm 0{,}09.$$

Lze tedy konstatovat, že průměrná týdenní spotřeba rychlého občerstvení se s 95% pravděpodobností pohybuje mezi hodnotami 0,75 a 0,89 a s 99% pravděpodobností leží hodnota v intervalu $\langle 0{,}73; 0{,}91 \rangle$.

Podobně jako oboustranný interval spolehlivosti je možno zkonstruovat i jednostranné varianty. V tomto případě hledáme $(1-\alpha)\%$ kvantil normálního rozdělení, tedy kvantil $u_{1-\alpha}$. Podoba těchto intervalů je tedy následující

a) levostranný interval $\mu < \overline{x} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$
b) pravostranný interval $\mu > \overline{x} - u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

Příklad 5.2

Při výzkumu průměrného věku zaměstnanců vybrané organizace byl u 20 náhodně oslovených pracovníků zjištěn průměrný věk 45,3 roku. Směrodatná odchylka tohoto byla vypočítána na 5,4.

Určete 95% levo- a pravostranný interval spolehlivosti pro odhad průměrného věku všech zaměstnanců uvedeného podniku.

Řešení

Zadané hodnoty: $\overline{x} = 45{,}3$, $n = 120$, $s_x = 15{,}4$

$$\sigma = s'_x = s_x \cdot \sqrt{\frac{n}{n-1}} = 5{,}4 \cdot \sqrt{\frac{120}{119}} = 15{,}5.$$

Přípustnou chybu odhadu pro jednostranný interval lze vypočítat podle vztahu

$$\Delta = u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = 1{,}64 \cdot \frac{15{,}5}{\sqrt{120}} = 2{,}3.$$

- 95% levostranný interval

$$\mu > \overline{x} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = \overline{x} + \Delta = 45{,}3 + 2{,}3 = 47{,}6.$$

- pravostranný interval

$$\mu < \overline{x} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = \overline{x} - \Delta = 45{,}3 - 2{,}3 = 43{,}0.$$

Lze tedy konstatovat, že – na základě provedeného šetření – s pravděpodobností 95% nepřesahuje v dané organizaci průměrný věk zaměstnanců hodnotu 47,6. Nebo lze konstatovat, že (se stejnou pravděpodobností) průměrný věk ve sledované organizaci není nižší než 43 let.

5.3.2 Výpočet intervalu spolehlivosti pro průměr v případě malého rozsahu výběru

Předpoklad normálního rozdělení je v mnoha praktických statistických úlohách není příliš silný. Vlastnosti normálního rozdělení je proto možno užít i pro výběry o menší rozsahu (obvykle menším než 100).

Pro malý rozsah výběru existuje výrazně větší nejistota ohledně struktury souboru. V případě menšího rozsahu výběru proto již není možno předpokládat, že variabilita zjištěná výběrovým šetřením ve výběrovém souboru bude odpovídat variabilitě základního souboru. Z těchto důvodů je vhodnější pro odhady použít místo normálního rozdělení Studentovo t-rozdělení, které (na rozdíl od normálního) není závislé na parametru σ^2 (rozptylu).

Vztahy pro výpočet intervalu spolehlivosti pro malý rozsah výběru jsou jen mírně odlišné od výše uvedených vztahů. Je nutno pouze nahradit kvantily normálního rozdělení příslušnými kvantily t-rozdělení.

Pro oboustranný interval pak dostáváme

$$\mu \in \left(\overline{x} - t_{1-\alpha/2} \frac{s'_x}{\sqrt{n}}; \overline{x} + t_{1-\alpha/2} \frac{s'_x}{\sqrt{n}} \right),$$

Jednostranné varianty pak mají následující podobu

- levostranný interval $\mu > \overline{x} + t_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$
- pravostranný interval $\mu < \overline{x} - t_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

Opět je možno definovat přípustnou chybu odhadu, jakožto míru rozpětí intervalu spolehlivosti. Její výpočet je možný podle vztahu

$$\Delta = t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Pro výpočet hodnoty kvantilu t-rozdělení potřebuje znát tzv. **počet stupňů volnosti**. Jejich hodnota je vždy dána hodnotou $n - 1$, kde n je rozsah výběru.

Příklad 5.3

Určete 95% interval spolehlivosti pro odhad průměrného věku zaměstnanců v podniku uvedeném v příkladu 5.2 vyjděte ze stejných výsledků průzkumu, pouze předpokládejte, že počet dotázaných zaměstnanců byl 20.

Řešení

Zadané hodnoty: $\mu = 45{,}3$, $n = 20$, $s_x = 15{,}4$. Pro **oboustranný interval** spolehlivosti platí

$$\mu \in \left(\overline{x} - t_{1-\alpha/2} \frac{s'_x}{\sqrt{n}}; \overline{x} + t_{1-\alpha/2} \frac{s'_x}{\sqrt{n}} \right),$$

Odsud dostáváme pro 95% interval spolehlivosti a rozsahu výběru 20

$$\mu = \overline{x} \pm t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 45{,}3 \pm 2{,}09 \cdot \frac{15{,}0}{\sqrt{20}} = 45{,}3 \pm 7{,}0.$$

Průměrný věk zaměstnanců se v daném podniku s 95% pravděpodobností pohybuje mezi 38,3 a 52,3 roky.

Pro **jednostranné intervaly** spolehlivosti pak analogicky dostáváme

- levostranný interval

$$\mu > \overline{x} + t_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = 45{,}3 + 1{,}73 \cdot \frac{15{,}0}{\sqrt{20}} = 45{,}3 + 5{,}8 = 51{,}1.$$

- pravostranný interval

$$\mu < \overline{x} - t_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = 45{,}3 - 1{,}73 \cdot \frac{15{,}0}{\sqrt{20}} = 45{,}3 - 5{,}8 = 39{,}5.$$

Průměrný věk zaměstnanců daného podniku tedy na 95% nepřesáhuje 51,1 let, resp. na 95% není nižší než 39,5 let.

Všimněte si, že interval spolehlivosti pro menší rozsah výběr vychází výrazně širší než je tomu u dostatečně velkého rozsahu výběru. Je to způsobeno tím, že u dostatečně velkého rozsahu můžeme vycházet z předpokladu známé variability základního a naše nejistota je o tento fakt snížena.

Je proto vhodné, pokud je to možné využívat prvního vztahu pro výpočet intervalu spolehlivosti. V praxi to znamená tvořit taková výběrová šetření, která pokryjí alespoň 100 respondentů.

5.3.3 Určení rozsahu výběru

Jak jsme již výše uvedli, přesnost (spolehlivost) odhadu je velmi závislá na rozsahu výběru, na základě něhož odhad vytváříme. Před provedením vlastního šetření je proto vhodné nejprve stanovit rozsah výběru n, který povede k dostatečně zobecňujícím výsledkům. Jelikož jsme se dosud zabývali intervaly spolehlivosti pro aritmetický průměr, naznačíme odvození vztahu pro stanovení rozsahu výběru pro tento interval.

Vyjdeme z výrazu pro přípustnou chybu, přičemž budeme předpokládat, že rozsah výběru je dostatečný (tedy lze použít vlastnosti normálního rozdělení).

$$\Delta = u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$
$$\sqrt{n} \cdot \Delta = u_{1-\alpha/2} \cdot \sigma.$$

A odsud dostáváme vztah pro rozsah výběru při stanovené přípustné chybě Δ ve tvaru

$$n = \frac{u_{1-\alpha/2}^2 \cdot \sigma^2}{\Delta^2}.$$

5.3.4 Postup při hledání rozsahu výběru

Pro určení rozsahu výběru tedy vyjdeme z výše uvedeného vztahu. Jelikož obvykle není znám rozptyl základního souboru, je nutno jej odhadnout pomocí výběrového rozptylu. Pokud jej neznáme, či nemůžeme využít z minulého šetření, musíme nejprve provést tzv. **předvýběr**.

Předvýběr je náhodným výběrem o malém rozsahu n_1 prvků. Z něj vypočítáme výběrový rozptyl. Lze-li předpokládat, že posuzované znaky budou mít normální rozdělení, pak stačí poměrně malý rozsah předvýběru n_1. V tom případě je možno využít pro výpočet rozsahu výběru pozměněného vztahu pro minimální rozsah výběru, kdy nahradíme kvantily normálního rozdělení příslušnými kvantily rozdělení t.

$$n = \frac{t_{1-\alpha/2}^2 \cdot {s'}_x^2}{\Delta^2}.$$

Nelze-li předpokládat normalitu rozdělení, je nutno provést předvýběr většího rozsahu ($n_1 > 30$) a použít původního vztahu pro minimální rozsah výběru (obsahujícího kvantily normálního rozdělení).

V praktických úlohách postupujeme tak, že provedeme předvýběr o rozsahu n_1, na základě něj určíme výběrový rozptyl a rozsah výběru n Poté doplníme předvýběr (je-li $n_1 < n$) o n_2 prvků. Tak, aby platilo $n_1 + n_2 = n$. Z takto

vzniklého výběrové souboru poté vypočítáme interval spolehlivosti pro aritmetický průměr základního souboru.

Příklad 5.4

Chceme odhadnout průměrnou mzdu pracovníků vybraného odvětví. Z minulého šetření vyplývá, že směrodatná odchylka mezd je 750 Kč. Odhad chceme vytvořit s 95% spolehlivostí a jsme ochotni připustit chybu maximálně 50 Kč.

Jak velký musíme provést výběr, abychom zajistili požadovanou přesnost a spolehlivost?

Řešení

Zadané hodnoty: $\sigma = 750$ Kč, $\Delta = 50$ Kč, $n =?$

Jelikož známe směrodatnou odchylku z minulého šetření, není nutno provádět předvýběr. Proto můžeme přímo vypočítat minimální rozsah podle výše uvedeného vztahu.

$$n = \frac{u_{1-\alpha/2}^2 \cdot \sigma^2}{\Delta^2} = \frac{196^2 \cdot 750^2}{50^2} = \frac{3{,}84 \cdot 562500}{2500} = 864{,}3.$$

K zajištění požadované přesnosti tedy musíme oslovit alespoň 865 pracovníků. V praktických úlohách bývá minimální rozsah výběru zaokrouhlován. V tomto případě bychom tedy stanovili rozsah výběru na 900 pracovníků.

Příklad 5.5

Jistá firma by ráda zjistila průměrné roční výdaje obyvatel za elektroniku. Požaduje chybu maximálně 1000 Kč. Zajímá ji, jak velký počet občanů má oslovit, aby dosáhla spolehlivosti 95%.

Firma provedla pilotní průzkum, ve kterém náhodně oslovila 20 osob. Jejich odpovědi jsou v následující tabulce (v tisících Kč).

10	2	66	13	7	0	3	11	13	22
9	3	4	14	40	15	1	9	0	6

Řešení

Z předvýběru nejprve vypočítáme výběrový rozptyl. Lze užít vztahu pro výběrový rozptyl, který má podobný tvar jako vztah pro rozptyl. Dostáváme:

$$s_x'^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{4630}{19} = 243{,}7.$$

Interval spolehlivosti tedy určíme z následujících hodnot: $\sigma^2 = s_x'^2 = 243{,}7$, $\Delta = 1$, $n_1 = 20$. Jelikož je rozsah předvýběru poměrně malý ($n_1 < 30$), užijeme k výpočtu rozsahu výběru kvantilů t-rozdělení. Kvantily t-rozdělení budou mít $\nu = n_1 - 1$ stupňů volnosti.

$$n = \frac{t_{1-\alpha/2}^2 \cdot s_x'^2}{\Delta^2} = \frac{2{,}09^2 \cdot 243{,}7^2}{1^2} = 1067{,}7.$$

Je tedy nutno oslovit alespoň 1068 osob. Firma tedy musí oslovit 1100 občanů. Na základě jejich odpovědí bude vypočítaná průměrná hodnota spolehlivá na 95%.

Shrnutí kapitoly

Vytváření odhadu průměru či rozptylu základního statistického souboru je klíčovou úlohou při vytváření celé řady sociologických či ekonomických průzkumů. Jedná se především o úlohy, kdy se snažíme nalézt průměr, či rozptyl rozložení některé vlastnosti (charakteristiky) v populaci. Praktická statistika (z technických, či finančních důvodů) není schopna pro tyto účely oslovit všechny občany. Využívá pouze jejich vzorek – výběrový soubor.

Pátá kapitola naznačila některé možnosti, jak na základě výběrového souboru odhadnout hodnotu průměru základního souboru. Jsou uvedeny některé příklady jak zkonstruovat tzv. interval spolehlivosti, který označuje nejpravděpodobnější oblast hodnot, ve které hledaný průměr leží. Při hledání intervalu spolehlivosti je možno využít především průměru výběrového souboru a také některých vlastností standardních rozdělení náhodné veličiny, které jsme uvedli v předchozí kapitole.

Otázky k zamyšlení

1. Posuďte jak se změní výsledky příkladu 5.5 v případě, že třetí dotázaný uvede místo hodnoty 66 hodnotu 15 tisíc Kč.

2. Vyočítejte 95% interval spolehlivosti pro průměrnou hrubou mzdu u zaměstnanců malého podniku. U deseti zaměstnanců byly zjištěny následující hrubé platy (v tis. Kč.)

	1	2	3	4	5	6	7	8	9	10
plat	17	18	19	11	13	16	14	17	22	23

3. Statistický úřad připravuje statistické šetření průměrné roční návštěvnosti kulturních zařízení obyvateli v dané zemi. Před započetím hlavního šetření byl proveden pilotní průzkum u dvaceti náhodně oslovených občanů. Jejich odpovědi (počet návštěv kina, divadla, či koncertu za kalendářní rok) přináší následující tabulka:

Číslo odpovědi	1	2	3	4	5	6	7	8	9	10
Počet návštěv	30	45	12	50	16	40	23	64	13	5
Číslo odpovědi	11	12	13	14	15	16	17	18	19	20
Počet návštěv	7	42	76	53	26	65	48	29	63	19

Výsledkem šetření má být průměrný počet návštěv kulturního zařízení občany dané země za rok, přičemž statistický úřad by chtěl dosáhnout spolehlivosti této hodnoty 95%. Průměrná hodnota by měla ležet v intervalu, který určí tuto hodnotu s tolerancí ± 5 návštěv.

Určete kolik osob musí statistický úřad oslovit, aby splnil zadání uvedeného šetření.

Příloha kapitoly 5

Využití programu MS EXCEL pro výpočet intervalů spolehlivosti

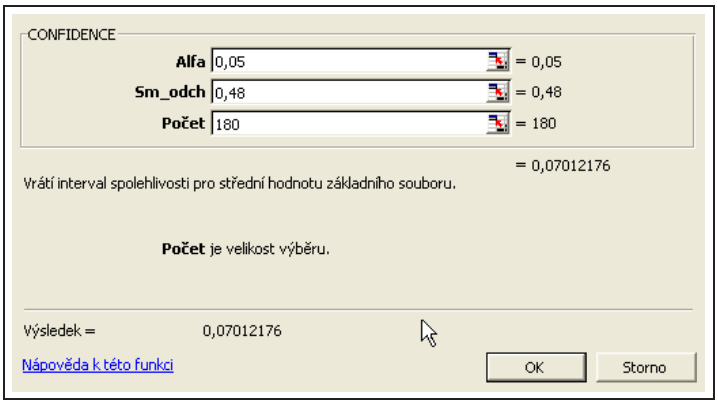
Pro výpočet intervalů spolehlivosti, resp. některých vlastností náhodné veličiny je možno využít programu MS EXCEL. Program EXCEL je možno využít jednak pro výpočet pomocných veličin – kvantilů rozdělení náhodné veličiny, výběrových rozptylů, průměrů a směrodatných odchylek, jednak i pro výpočet konkrétní podoby intervalu spolehlivosti.

Výpočet kvantilů rozdělení náhodné veličiny v prostředí programu EXCEL jsme se již zmínili v příloze kapitoly 4. Program lze však využít také pro výpočet dalších veličin potřebných pro stanovení intervalu spolehlivosti.

Výpočet výběrového rozptylu a výběrové směrodatné odchylky
Výběrový rozptyl statistického souboru jsme definovali v kapitole 5 pomocí vztahu mezi tímto rozptylem a rozptylem základního souboru. Mimo tohoto přepočtového vztahu je možno užít také přímého výpočtu v programu EXCEL pomocí funkce VAR.VÝBĚR, případně VARA.
Syntaxe: VAR.VÝBĚR(číslo1;číslo2;...)

Stejně tak lze vypočítat i výběrovou směrodatnou odchylku. Můžeme ji vypočítat buď jako odmocninu výběrového rozptylu (v EXCELu pomocí funkce ODMOCNINA(číslo)) nebo přímo pomocí funkce SMODCH.VÝBĚR:
Syntaxe: SMODCH.VÝBĚR(číslo1;číslo2;...)

Určení intervalu spolehlivosti
Program EXCEL umožňuje přímé určení intervalu spolehlivosti bez nutnosti výpočtu výběrového rozptylu a příslušných kvantilů rozdělení. Pomocí funkce CONFIDENCE můžeme spočítat přímo hodnotu přípustné chyby (označovali jsme ji symbolem Δ).
Syntaxe: CONFIDENCE(alfa;sm_odch;počet)



Obrázek 5.2: Zadání hodnot do funkce CONFIDENCE pro hodnoty z příkladu 5.1 uvedeného v této kapitole.

Alfa je hladina významnosti, pomocí které je vypočítána hladina spolehlivosti. Hladina spolehlivosti se rovná $100 \cdot (1 - \text{alfa})\%$, tzn. je-li argument alfa roven 0,05, hladina spolehlivosti je 95%. **Sm_odch** je směrodatná odchylka základního souboru pro danou oblast dat a pokládá se za známou. **Počet** je velikost výběru.