

STATISTIKA 2

Statistika a statistické zpracování dat

Blok 1-2

Jiří Fišer

8. listopadu 2024

Zpracování dat z výběrových zjišťování

Cíle kapitoly:

- Definovat pojem náhodného výběru a výběrového souboru.
- Představit základní výběrové charakteristiky a jejich rozdělení.
- Uvést základní metody výběrových šetření a jejich použití.

Úvodní příklad - Výběrové šetření

Představte si, že jste analytikem ve velké maloobchodní společnosti. Vedení chce zjistit průměrnou spokojenost zákazníků, ale místo dotazování všech provede výběrové šetření.

Úkol: Navrhnout spolehlivé šetření, kde každý zákazník má stejnou šanci být zahrnut (*prostý náhodný výběr*).

Výpočet průměrné spokojenosti

Příklad: Z 100 zákazníků získáte následující hodnocení spokojenosti na škále 1–10:

$$x = (8, 7, 9, 6, \dots)$$

Součet všech hodnocení je 750.

Výběrový průměr:

$$\bar{x} = \frac{1}{100} \cdot 750 = 7,5$$

Tento průměr použijeme jako odhad průměrné spokojenosti všech zákazníků.

Směrodatná odchylka

Směrodatná odchylka měří, jak moc se odpovědi liší od průměru.

Příklad: Výpočet směrodatné odchylky:

$$s = \sqrt{\frac{1}{99} \sum_{i=1}^{100} (x_i - \bar{x})^2} \approx 1,2$$

Interpretace: Většina odpovědí se pohybuje v rozmezí $7,5 \pm 1,2$, tedy mezi 6,3 a 8,7.

Význam směrodatné odchylky

Směrodatná odchylka ukazuje, jak jsou odpovědi konzistentní:

- Nižší hodnota: spokojenost zákazníků je podobná.
- Vyšší hodnota: větší rozdíly mezi zákazníky (někteří velmi spokojeni, jiní méně).

Pomocí tohoto šetření můžeme také odhadnout interval spokojenosti pro celou populaci zákazníků.

Závěr - Výběrové šetření

Tento příklad ukazuje důležitost správně provedeného výběrového šetření:

- Správný výběr zajišťuje reprezentativnost výsledků.
- Výběrové charakteristiky, jako průměr a směrodatná odchylka, pomáhají analyzovat a odhadovat parametry populace.

Pravděpodobnostní (náhodný) výběr zajišťuje, že každý prvek souboru má určitou pravděpodobnost, že bude zahrnut do výběrového souboru. Tento přístup umožňuje reprezentativnost výsledků na základě náhodně vybraných jednotek.

Tvorba výběrového souboru

Tvorba výběrového souboru zahrnuje dvě hlavní složky:

- **Pravděpodobnost vybrání:** Každá jednotka má přiřazenou pravděpodobnost být vybrána.
- **Náhodnost výběru:** Výběr je prováděn náhodně, rozhoduje pouze náhoda.

Spojení těchto dvou složek je klíčové pro zajištění reprezentativnosti.

Pravděpodobnostní a náhodný výběr

- Termín „**pravděpodobnostní výběr**“ převažuje nad „náhodný výběr“.
- Pravděpodobnosti vybrání mohou být různé, pokud jsou stejné, jedná se o **prostý náhodný výběr**.

Pochybnosti o náhodném výběru

Pochybnosti neodborníků: Náhodný výběr se může jevit jako náhodný proces bez kontroly.

Reakce: Náhodný výběr s předem známými pravděpodobnostmi využívá výhod náhody a lze matematicky kontrolovat jeho zákonitosti, což zajišťuje reprezentativnost.

Výhody pravděpodobnostního výběru

Pravděpodobnostní výběry umožňují spolehlivé odhady:

- **Konzistentní odhady:** Přesnost odhadů roste se zvyšováním počtu jednotek.
- **Nevychýlené odhady:** Odhady v průměru ani nepřehánějí, ani nebagatelizují skutečnou hodnotu.

Měření přesnosti odhadů

Přesnost odhadů lze měřit pomocí:

- **Střední výběrové chyby** (standardní chyba průměru) $\frac{s}{\sqrt{n}}$
- **Intervalových odhadů**: poskytují interval, ve kterém se skutečná hodnota téměř jistě nachází.

Další kapitoly se budou podrobněji věnovat těmto odhadům a jejich interpretaci.

Prostý náhodný výběr

Prostý náhodný výběr je základní forma **pravděpodobnostního výběru**. Každý prvek základního souboru (ZS) má stejnou pravděpodobnost, že bude do výběru zahrnut.

Definice: Prostý náhodný výběr je výběr o rozsahu n , kdy každá množina n prvků má stejnou pravděpodobnost, že bude vybrána.

Podmínka: Je nutné mít:

- očíslovaný seznam všech prvků základního souboru (**opora výběru**),
- generátor náhodných čísel pro výběr prvků.

Postup při prostém náhodném výběru

Postup při provádění prostého náhodného výběru:

- 1 Sestavte oporu výběru a přiřaďte každému prvku celé číslo od 1 do N .
- 2 Rozhodněte o rozsahu výběru n .
- 3 Vygenerujte n náhodných čísel mezi 1 a N .
- 4 Získejte data od prvků identifikovaných těmito čísly.

Výběrový poměr

Výběrový poměr je poměr mezi rozsahem výběru n a velikostí základního souboru N :

$$\frac{n}{N}$$

- Vyjadřuje pravděpodobnost, že prvek ZS bude zahrnut do výběru.
- Výběr může být **s vrácením** (prvek může být vybrán vícekrát) nebo **bez vrácení**.

Náhradní metody při neproveditelnosti prostého náhodného výběru

Pokud je prostý náhodný výběr neproveditelný nebo příliš nákladný, lze použít:

- **Stratifikovaný náhodný výběr:** Populace je rozdělena na skupiny (straty) a z každé straty se provede náhodný výběr.
- **Systematický výběr:** Po seřazení ZS je náhodně vybrán jeden prvek z prvních k a dále každý k -tý prvek.
- **Vícestupňový shlukový výběr:** Výběr probíhá ve více stupních (např. okresy, města, sídliště, domácnosti).

Stratifikovaný náhodný výběr

Stratifikovaný náhodný výběr: ZS je rozdělen na dílčí skupiny (straty) a z každé se provede náhodný výběr.

Vhodné pro: Populace, které lze stratifikovat podle určitého znaku (např. věk, pohlaví) pro zajištění reprezentace každé podskupiny.

Systematický výběr

Systematický výběr: Ze seřazeného ZS se náhodně vybere jeden prvek z prvních k , poté každý k -tý prvek.

Příklad: Při ZS o 100 prvcích a výběru 10 prvků ($k=10$) náhodně vybereme prvek, např. 5, a následně vybereme 5, 15, 25, ..., 95.

Vícestupňový shlukový výběr

Vícestupňový shlukový výběr: Metoda vhodná pro rozsáhlé základní soubory, např. při výzkumu veřejného mínění.

- 1 Náhodně vybereme vzorek okresů.
- 2 Z každého okresu vybereme vzorek měst.
- 3 Z vybraných měst vybereme vzorek sídlišť.
- 4 Z vybraných sídlišť vybereme domácnosti k dotazování.

Výhoda: Efektivní a méně nákladný než prostý náhodný výběr.

Výběrové charakteristiky a jejich rozdělení

- Výběrové charakteristiky jsou statistiky
 - ▶ získané z výběrového souboru,
 - ▶ poskytují informace o základním souboru.
- Jsou funkcemi náhodných veličin,
 - ▶ protože závisí na konkrétním výběru vzorku.

Základní pojmy

- **Základní soubor:**

- ▶ Skládá se z N jednotek,
- ▶ zajímá nás znak X .

- **Výběrový soubor:**

- ▶ Obsahuje n jednotek
- ▶ náhodně vybraných ze základního souboru.

- **Statistický model:**

- ▶ Pravděpodobnostní rozdělení náhodné veličiny X tvoří statistický model.

Výběrové charakteristiky

- Výběrové charakteristiky jsou funkce náhodných veličin

$$X_1, X_2, \dots, X_n,$$

- definované jako:

$$T = T(X_1, X_2, \dots, X_n),$$

kde

- ▶ X_1 je náhodná veličina náhodného výběru 1. prvku ve výběru,
- ▶ X_2 je náhodná veličina náhodného výběru 2. prvku ve výběru, ...

Výběrový obecný moment a výběrový průměr

Výběrový obecný moment: k -tý obecný moment je

$$m_k = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

Výběrový průměr:

$$\bar{x} = m_1 = \frac{1}{n} \sum_{i=1}^n x_i.$$

U prostého náhodného výběru platí $\mathbb{E}(\bar{X}) = \mu$, což z něj činí vhodný **odhad střední hodnoty ZS**:

$$\mu \approx \bar{x}.$$

Výběrový centrální moment a výběrový rozptyl

Výběrový centrální moment: k -tý centrální moment:

$$m'_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Výběrový rozptyl:

$$s^2 = m'_2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Výběrový rozptyl je vhodný pro **odhad rozptylu základního souboru**:

$$\sigma^2 \approx s^2.$$

Výběrová směrodatná odchylka a výběrová kovariance

Výběrová směrodatná odchylka:

$$s = \sqrt{s^2}.$$

Výběrová kovariance: Pokud sledujeme dva znaky X a Y :

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

- Kovariance nám ukazuje, zda a jakým směrem se dvě proměnné pohybují společně (kladná hodnota znamená pozitivní vztah, záporná hodnota negativní vztah).
- Její velikost však závisí na měřítku proměnných, což komplikuje interpretaci.

Výběrový lineární korelační koeficient

Výběrový lineární korelační koeficient:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{s_X s_Y},$$

kde s_X a s_Y jsou výběrové směrodatné odchylky znaků X a Y .

- **Korelační koeficient standardizuje kovarianci** do intervalu $[-1, 1]$, což usnadňuje interpretaci:
 - ▶ $r_{XY} \approx 1$: silná pozitivní korelace
 - ▶ $r_{XY} \approx -1$: silná negativní korelace
 - ▶ $r_{XY} \approx 0$: žádný lineární vztah

Úloha výběrového šetření

Úkolem výběrového šetření je odhadnout neznámé parametry základního souboru pomocí výběrových charakteristik:

- **Parametry základního souboru** (μ, σ^2)
- **Výběrové charakteristiky** ($\bar{x}, s_X, r_{XY}$)

Cílem je získat odhady, které jsou přesné a nevychýlené.

Řešené příklady

- Příklady ilustrují probranou látku.
- Nejsou příliš náročné na výpočty.
- V praxi by byly výběry rozsáhlejší.

Příklad: Stratifikovaný výběr a výběrové charakteristiky

Situace: Obchodní řetězec analyzuje průměrné nákupy zákazníků ve dvou regionech (A a B).

Celkový počet zákazníků: 20 000 (12 000 v regionu A, 8 000 v regionu B)
Stratifikovaný náhodný výběr: 5 zákazníků z každého regionu.

Údaje o útratách zákazníků ve výběru

Region A: $x_A = (800, 1500, 700, 1200, 900)$

Region B: $x_B = (1000, 1100, 950, 1300, 750)$

Cílem je vypočítat:

- 1 Průměrnou útratu zákazníků ve výběru v regionu A a regionu B.
- 2 Směrodatnou odchylku útrat zákazníků v regionu A a regionu B.

Výpočet průměrné útraty zákazníků

Region A:

$$\bar{x}_A = \frac{1}{5} \sum_{i=1}^5 x_{A,i} = \frac{800 + 1500 + 700 + 1200 + 900}{5} = 1020 \text{ Kč.}$$

Region B:

$$\bar{x}_B = \frac{1}{5} \sum_{i=1}^5 x_{B,i} = \frac{1000 + 1100 + 950 + 1300 + 750}{5} = 1020 \text{ Kč.}$$

Průměrná útrata je stejná v obou regionech: 1020 Kč.

Výpočet směrodatné odchylky

Region A:

$$s_A = \sqrt{\frac{1}{4} \sum_{i=1}^5 (x_{A,i} - \bar{x}_A)^2} = \sqrt{107\,000} \approx 327,11 \text{ Kč.}$$

Region B:

$$s_B = \sqrt{\frac{1}{4} \sum_{i=1}^5 (x_{B,i} - \bar{x}_B)^2} = \sqrt{40\,750} \approx 201,87 \text{ Kč.}$$

Směrodatná odchylka je vyšší v regionu A, což značí větší rozptyl kolem průměru.

Interpretace výsledků

Stratifikovaný náhodný výběr „ukázal“, že:

- Průměrná útrata je stejná v obou regionech (1020 Kč).
- Směrodatná odchylka je vyšší v regionu A (327,11 Kč) než v regionu B (201,87 Kč), což naznačuje větší variabilitu v regionu A.

Příklad: Kovariance a korelační koeficient

Situace: Jako analytik pro investiční společnost zkoumáte vztah mezi ročními výnosy dvou akcií (X a Y) za posledních 5 let.

Výnosy (v %) jsou následující:

- **Akcie X:** $X = (5, 10, 12, 6, 8)$
- **Akcie Y:** $Y = (3, 8, 9, 5, 6)$

Cílem je vypočítat průměrné výnosy, kovarianci a korelační koeficient.

Výpočet průměrných výnosů

Průměrný výnos pro akcii X:

$$\bar{X} = \frac{1}{5} \sum_{i=1}^5 X_i = \frac{5 + 10 + 12 + 6 + 8}{5} = 8,2 \text{ \%}.$$

Průměrný výnos pro akcii Y:

$$\bar{Y} = \frac{1}{5} \sum_{i=1}^5 Y_i = \frac{3 + 8 + 9 + 5 + 6}{5} = 6,2 \text{ \%}.$$

Průměrné roční výnosy: $X = 8,2 \text{ \%}$, $Y = 6,2 \text{ \%}$.

Výpočet kovariance

Kovariance mezi výnosy X a Y:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

$$\text{Cov}(X, Y) = \frac{26,8}{4} = 6,7.$$

Kovariance je kladná, což značí pozitivní vztah mezi výnosy X a Y.

Výpočet korelačního koeficientu

Korelační koeficient mezi X a Y:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{s_X s_Y}.$$

Po výpočtu směrodatných odchylek získáme:

$$r_{XY} \approx \frac{6,7}{2,86 \cdot 2,38} \approx 0,98.$$

Korelační koeficient 0,98 znamená silnou pozitivní korelaci.

Interpretace kovariance a korelačního koeficientu

- **Kovariance** $\text{Cov}(X, Y) = 6,7$: kladná hodnota znamená, že výnosy akcií X a Y mají pozitivní vztah.
- **Korelační koeficient** $r_{XY} = 0,98$: vysoká hodnota ukazuje na velmi silnou pozitivní korelaci.

Tento výsledek naznačuje, že výnosy akcií X a Y se pohybují podobně, což může ovlivnit rozhodnutí o diverzifikaci portfolia.