Výběrové charakteristiky a jejich rozdělení

Jiří Neubauer

Katedra ekonometrie, FVL, UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Výběrové šetření

Statistické šetření

- úplné (vyčerpávající)
- neúplné (výběrové)

U výběrového šetření se snažíme o to, aby výběrový soubor dobře reprezentoval základní soubor. Za tohoto předpokladu je možné zobecnit poznatky z výběrového souboru na celý základní soubor.

Výběrové šetření

Statistické šetření

- úplné (vyčerpávající)
- neúplné (výběrové)

U výběrového šetření se snažíme o to, aby výběrový soubor dobře reprezentoval základní soubor. Za tohoto předpokladu je možné zobecnit poznatky z výběrového souboru na celý základní soubor.

Výběrové šetření

Statistické šetření

- úplné (vyčerpávající)
- neúplné (výběrové)

U výběrového šetření se snažíme o to, aby výběrový soubor dobře reprezentoval základní soubor. Za tohoto předpokladu je možné zobecnit poznatky z výběrového souboru na celý základní soubor.

Výběrové šetření

Výběrový soubor získáme buď **záměrným** nebo **náhodným výběrem**. Nejjednodušší náhodný výběr je tzv. **prostý náhodný výběr**, tj. přímý výběr, kde každá jednotka má stejnou pravděpodobnost výběru.

Problém výběru s vracením a bez vracení

Je-li $n/N \leq 0{,}05$ nebo základní soubor je nekonečný, hypotetický, považuje se požadavek na nezávislost za splněný, tedy mezi výběrem svracením a bez vracení nebudeme dělat rozdíl.

Výběrové šetření

Výběrový soubor získáme buď **záměrným** nebo **náhodným výběrem**. Nejjednodušší náhodný výběr je tzv. **prostý náhodný výběr**, tj. přímý výběr, kde každá jednotka má stejnou pravděpodobnost výběru.

Problém výběru s vracením a bez vracení
Je-li $n/N \leq 0{,}05$ nebo základní soubor je nekonečný, hypotetický, považuje se požadavek na nezávislost za splněný, tedy mezi výběrem svracením a bez vracení nebudeme dělat rozdíl.

Nezávislé náhodné veličiny

Náhodné veličiny $X_1, X_2, \ldots, X_n$ jsou nezávislé právě tehdy, když pro libovolná čísla $x_1, x_2, \ldots, x_n \in \mathbb{R}$ platí

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) \cdots P(X_n \leq x_n).$$

Nezávislé náhodné veličiny

Mějme náhodný vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)$, jehož složky $X_1, X_2, \dots, X_n$ jsou náhodné veličiny. Nechť

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

je **sdružená distribuční funkce** a $F(x_1), F(x_2), \dots, F(x_n)$ jsou distribuční funkce náhodných veličin $X_1, X_2, \dots, X_n$. Náhodné veličiny $X_1, X_2, \dots, X_n$ jsou nezávislé právě tehdy, když

$$F(x_1, x_2, \dots, x_n) = F(x_1) \cdot F(x_2) \cdots F(x_n).$$

Nezávislé náhodné veličiny

Je-li **X** náhodný vektor, jehož složky jsou diskrétní náhodné veličiny, funkce

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

je **sdružená pravděpodobnostní funkce** a $p(x_1), p(x_2), \dots, p(x_n)$ jsou pravděpodobnostní funkce náhodných veličin $X_1, X_2, \dots, X_n$, pak platí: Náhodné veličiny $X_1, X_2, \dots, X_n$ jsou nezávislé právě tehdy, když

$$p(x_1, x_2, \dots, x_n) = p(x_1) \cdot p(x_2) \cdots p(x_n).$$

Nezávislé náhodné veličiny

Je-li **X** náhodný vektor, jehož složky jsou spojité náhodné veličiny, funkce

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$$

je **sdružená funkce hustoty pravděpodobnosti**
a $f(x_1), f(x_2), \ldots, f(x_n)$ jsou funkce hustoty pravděpodobnosti náhodných veličin $X_1, X_2, \ldots, X_n$, pak platí:
Náhodné veličiny $X_1, X_2, \ldots, X_n$ jsou nezávislé právě tehdy, když

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n).$$

Náhodný výběr

- U každé jednotky, která se dostane do výběrového souboru, zjistíme hodnotu zkoumaného znaku x_i $(i = 1, 2, \dots, n)$.
- Tuto hodnotu můžeme chápat jako jednu z možných hodnot náhodné veličiny X_i. Každá z těchto n náhodných veličin má stejné rozdělení.

Náhodný výběr

- U každé jednotky, která se dostane do výběrového souboru, zjistíme hodnotu zkoumaného znaku x_i $(i = 1, 2, \ldots, n)$.
- Tuto hodnotu můžeme chápat jako jednu z možných hodnot náhodné veličiny X_i. Každá z těchto n náhodných veličin má stejné rozdělení.

Náhodný výběr

Definice

Náhodný výběr o rozsahu n je posloupnost nezávislých náhodných veličin $X_1, X_2, \ldots, X_n$ se stejným rozdělení.

Můžeme ho chápat jako vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Konkrétní realizaci budeme značit $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Naměřené hodnoty $x_1, x_2, \ldots, x_n$ nazýváme pozorování nebo také vstupní (empirická) data.

Náhodný výběr

Protože jsou náhodné veličiny $X_1, X_2, \ldots, X_n$ nezávislé a mají stejné rozdělení, platí pro distribuční funkci $F(\mathbf{x})$ náhodného výběru

$$F(\mathbf{x}) = F(x_1)F(x_2)\cdots F(x_n), \quad x_i \in \mathbb{R}.$$

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z rovnoměrného rozdělení na intervalu $(0, 1)$. Určete distribuční funkci $F(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim R(0, 1)$ tedy $F(x_i) = x_i$ pro $0 < x_i < 1$,

$$F(\mathbf{x}) = F(x_1)F(x_2)\cdots F(x_n) = x_1 \cdot x_2 \cdots x_n.$$

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z rovnoměrného rozdělení na intervalu $(0, 1)$. Určete distribuční funkci $F(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim R(0, 1)$ tedy $F(x_i) = x_i$ pro $0 < x_i < 1$,

$$F(\mathbf{x}) = F(x_1)F(x_2)\cdots F(x_n) = x_1 \cdot x_2 \cdots x_n.$$

Náhodný výběr

Pravděpodobnostní funkce $p(\mathbf{x})$ náhodného výběru v případě diskrétního rozdělení veličin $X_1, X_2, \ldots, X_n$ je

$$p(\mathbf{x}) = p(x_1)p(x_2)\cdots p(x_n), \quad x_i \in \mathbb{R}.$$

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z Poissonova rozdělení s parametrem λ. Určete pravděpodobnostní funkci $p(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim Po(\lambda)$ tedy $p(x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$ pro $x_i = 0, 1, 2, \ldots,\ i = 1, 2, \ldots, n$

$$p(\mathbf{x}) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \frac{1}{x_1! \cdot x_2! \cdots x_n!}.$$

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z Poissonova rozdělení s parametrem λ. Určete pravděpodobnostní funkci $p(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim Po(\lambda)$ tedy $p(x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$ pro $x_i = 0, 1, 2, \ldots,\ i = 1, 2, \ldots, n$

$$p(\mathbf{x}) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \frac{1}{x_1! \cdot x_2! \cdots x_n!}.$$

Náhodný výběr

Hustota rozdělení $f(\mathbf{x})$ náhodného výběru z rozdělení s hustotou $f(x)$ je

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n),$$

kde $x_i \in \mathbb{R}, i = 1, 2, \ldots, n$.

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$. Určete hustotu pravděpodobnosti $f(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim N(\mu, \sigma^2)$ tedy $f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$ pro $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$

$$f(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Příklad

Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$. Určete hustotu pravděpodobnosti $f(\mathbf{x})$ náhodného výběru.

Řešení:

$X_i \sim N(\mu, \sigma^2)$ tedy $f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$ pro $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$

$$f(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Výběrové charakteristiky

Definice

Funkce náhodných veličin $X_1, X_2, \dots, X_n$ se nazývá **statistika**

$$T = T(X_1, X_2, \dots, X_n) = T(\mathbf{X}).$$

Výběrové charakteristiky

- Výběrový úhrn

$$M = \sum_{i=1}^{n} X_i$$

- Výběrový průměr

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Výběrové charakteristiky

- Výběrový úhrn

$$M = \sum_{i=1}^{n} X_i$$

- Výběrový průměr

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Výběrové charakteristiky

- Výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

- Základní rozptyl výběru

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

Výběrové charakteristiky

- Výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

- Základní rozptyl výběru

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

Výběrové charakteristiky

- Výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

- Základní rozptyl výběru

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

Výběrové charakteristiky

- Výběrový r-tý obecný moment

$$M_r' = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

- Výběrový r-tý centrální moment

$$M_r = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^r$$

Výběrové charakteristiky

- Výběrový r-tý obecný moment

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

- Výběrový r-tý centrální moment

$$M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r$$

Výběrové charakteristiky

- Výběrový koeficient šikmosti

$$A_3 = \frac{M_3}{M_2^{3/2}}$$

- Výběrový koeficient špičatosti

$$A_4 = \frac{M_4}{M_2^2} - 3$$

Výběrové charakteristiky

- Výběrový koeficient šikmosti

$$A_3 = \frac{M_3}{M_2^{3/2}}$$

- Výběrový koeficient špičatosti

$$A_4 = \frac{M_4}{M_2^2} - 3$$

Rozdělení výběrového úhrnu

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení se střední hodnotou μ a rozptylem σ^2, tzn. $E(X_i) = \mu$, $D(X_i) = \sigma^2$, pro $i = 1, 2 \ldots, n$. Střední hodnota a rozptyl výběrového úhrnu jsou

$$E(M) = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E(X_i) = n\mu$$

$$D(M) = D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} D(X_i) = n\sigma^2$$

Rozdělení výběrového úhrnu

Věta

Je-li $X_1, X_2, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$, má také výběrový úhrn normální rozdělení

$$M \sim N(n\mu, n\sigma^2).$$

Rozdělení výběrového průměru

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení se střední hodnotou μ a rozptylem σ^2, tzn. $E(X_i) = \mu$, $D(X_i) = \sigma^2$, pro $i = 1, 2 \ldots, n$. Střední hodnota a rozptyl výběrového průměru jsou

$$E(\overline{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$D(\overline{X}) = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Rozdělení výběrového průměru

Věta

Je-li $X_1, X_2, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$, má také výběrový průměr normální rozdělení

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Normováním dostaneme statistiku

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n},$$

která má normované normální rozdělení $N(0, 1)$.

Rozdělení výběrového průměru

Je-li $X_1, X_2, \dots, X_n$ náhodný výběr z rozdělení se střední hodnotou μ a rozptylem σ^2, pak ná veličina

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

má pro $n \geq 30$ přibližně normované normální rozdělení $N(0, 1)$ – viz centrální limitní věta.

Rozdělení výběrového rozptylu

Mějme náhodný výběr $X_1, X_2, \dots, X_n$ z rozdělení se střední hodnotou μ a rozptylem σ^2, tzn. $E(X_i) = \mu$, $D(X_i) = \sigma^2$, pro $i = 1, 2 \dots, n$. Střední hodnoty výběrového a základního rozptylu jsou

$$E(S^2) = \sigma^2 \quad E(S_n^2) = \frac{n-1}{n}\sigma^2.$$

Rozdělení výběrového rozptylu

Při odvození střední hodnoty výběrového rozptylu použijeme vztahy:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$D(X_i) = E(X_i^2) - E(X_i)^2 \rightarrow E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$D(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \rightarrow E(\overline{X}^2) = D(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Rozdělení výběrového rozptylu

Při odvození střední hodnoty výběrového rozptylu použijeme vztahy:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$D(X_i) = E(X_i^2) - E(X_i)^2 \to E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$D(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \to E(\overline{X}^2) = D(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Rozdělení výběrového rozptylu

Při odvození střední hodnoty výběrového rozptylu použijeme vztahy:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$D(X_i) = E(X_i^2) - E(X_i)^2 \rightarrow E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$D(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \rightarrow E(\overline{X}^2) = D(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Rozdělení výběrového rozptylu

$$
\begin{aligned}
E(S_n^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - E(\overline{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \frac{1}{n} n(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

$$
E(S^2) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2
$$

Rozdělení výběrového rozptylu

Věta

Mějme náhodný výběr z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Náhodná veličina

$$\chi^2 = \frac{n-1}{\sigma^2} S^2$$

má χ^2-rozdělení s $n-1$ stupni volnosti.

Rozdělení výběrového rozptylu

Předpokládejme, že máme výběr z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Víme, že $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$ a $\chi^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$. Náhodná veličina

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \cdot \frac{\sqrt{n-1}}{\sqrt{\frac{n-1}{\sigma^2}S^2}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \cdot \frac{\sigma}{S} = \frac{\overline{X}-\mu}{S}\sqrt{n}$$

má Studentovo t-rozdělení s $n-1$ stupni volnosti.

Rozdělení výběrového rozptylu

Věta

Mějme náhodný výběr z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Náhodná veličina

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n}$$

má Studentovo t-rozdělení s $n - 1$ stupni volnosti.

Rozdělení výběrového podílu

Předpokládejme, že rozdělení v základním souboru je alternativní s parametrem π. Náhodný výběr mohou být buď jedničky nebo nuly. Náhodnou veličina $X = X_1 + X_2 + \cdots + X_n$ potom určuje počet jedniček ve výběru (tzv. **výběrová absolutní četnost**). Podíl

$$P = \frac{X}{n}$$

bývá označován jako **výběrová relativní četnost** nebo častěji **výběrový podíl**

Rozdělení výběrového podílu

Předpokládejme, že máme náhodný výběr velkého rozsahu n z alternativního rozdělení s parametrem π. Náhodná veličina $P = \frac{X}{n}$ má přibližně normální rozdělení se střední hodnotou π a směrodatnou odchylkou $\sqrt{\pi(1-\pi)/n}$ – viz centrální limitní věta. Z CLV plyne, že normovaná náhodná veličina

$$Z = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}}$$

má pro velká n přibližně normální rozdělení. Aproximace je vhodná, pokud $n\pi \geq 5$ a zároveň $n(1-\pi) \geq 5$.