# 9 NÁHODNÉ VÝBĚRY A JEJICH ZPRACOVÁNÍ

**Čas ke studiu kapitoly: 30 minut**

**Cíl:** Po prostudování tohoto odstavce budete rozumět pojmům

- Základní soubor, populace, výběr, výběrové šetření, výběrová statistika

a budete znát základní výběrové statistiky pro výběry z normálního rozdělení

## Výklad:

***Motto:***

*Chceme-li vědět, jak chutná víno v sudu, nemusíme vypít celý sud. Stačí jenom malý doušek a víme na čem jsme.*

Statistika – to je sběr a zpracování dat. V mnoha oborech se setkáme s průzkumy opírajícími se o relativně malý počet zkoumaných jednotek (**výběr**). Statistika pak používá postupy pomocí nichž můžeme, sice s určitým rizikem (předem stanoveným), na základě toho mála usuzovat na chování celku (**populace**). Tomuto zobecňování říkáme **statistická indukce.**

## 9.1 Statistické zjišťování

Pro většinu statistických souborů, s nimiž se v praxi setkáváme, je typický vysoký **rozsah** (počet zkoumaných jednotek). Jakmile jsme tedy postavení před úkol provést určité šetření a analyzovat údaje z něj zjištěné, musíme nejprve rozhodnout, zda budeme toto šetření realizovat jako vyčerpávající nebo výběrové.

**Vyčerpávající šetření** – to je prošetření všech jednotek statistického souboru (populace). Zpravidla se jedná o záležitost velmi nákladnou (personálně, finančně, časově), mnohdy dokonce prakticky nerealizovatelnou (destrukční zkoušky). Pokud však toto šetření proběhne, mezi jeho nesporné výhody patří přesnost zjištěných charakteristik a detailnost formací o každé zkoumané jednotce. Příkladem vyčerpávajícího šetření je například sčítání lidu.

**Výběrové šetření** – jde o prošetření vybraných jednotek statistického souboru (populace). Z takto pořízených charakteristik pak více či méně usuzujeme na vlastnosti celé populace. Výběrová šetření se používají například při zjišťování jaká je podpora politických stran, při ověřování pevnosti trubek vyráběných určitým podnikem, apod. Mírou objektivnosti informací, které z něho získáme, je kvalita provedení výběrového šetření.

## 9.2 Typy výběrových šetření

Základní soubory, z nichž vybíráme mohou být buď konečně nebo nekonečně velké. Příkladem konečně velkého základního souboru je dodávka výrobku (např. praček), příkladem nekonečně velkého základního souboru je nepřetržitá pásová výroba (např. praček). Při konstrukci výběrového souboru se snažíme o to, aby výběrový soubor měl stejné vlastnosti jako základní soubor, z něhož výběr pochází.

Mezi druhy výběrových šetření řadíme anketu, metodu základního masivu, záměrný výběr a náhodný výběr.

**Anketa** oslovuje pouze nesystematicky vybranou část populace (osob, podniků, institucí). Dotazník se k respondentům (dotazovaným) dostává prostřednictvím sdělovacích prostředků (anketa televizních diváků, anketa časopisu Mládí, …) nebo je zaslán adresně. Návratnost dotazníku je však malá (odhaduje se že 30%). Informace získané anketním šetřením nelze zobecňovat.

**Metoda základního masivu** se používá v případech, kdy se základní soubor skládá z několika velkých jednotek a z většího počtu jednotek malých. (např. při šetření v oblasti hutnictví se můžeme podle této metody zaměřit na několik „obřích“ společností, tam provést šetření a „malé“ podniky vynechat. Výhody: menší pracnost a menší časová náročnost šetření. Nevýhody: zobecnění poznatků má menší platnost (nevystihuje specifika menších jednotek).

**Záměrný výběr** spočívá v tom, že skupina odborníku na danou problematiku vybere podle svého nejlepšího uvážení ty jednotky, o nichž se lze domnívat, že ve svém souhrnu nejlépe umožní provést šetření. S tímto typem šetření se setkáme například při průzkumech trhu a při průzkumech veřejného mínění. Nevýhoda: subjektivní přístup k výběru zpochybňuje možnost zobecnění.

**Prostý náhodný výběr** je základním a v praxi nejpoužívanějším typem výběru. Jde o výběr, při němž mají všechny jednotky základního souboru stejnou pravděpodobnost, že do výběru budou zařazeny.

Nyní se náhodným výběrem budeme zabývat podrobněji (a formálněji).

## 9.3 Náhodný výběr

**Náhodný výběr ($\underline{X}$)** je speciální náhodný vektor, jehož složky jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti.

Pokusíme se uvedenou definici vysvětlit. Opakujeme-li $n$-krát nezávisle pokus (pozorování, měření), jehož výsledek je náhodná veličina $X$ s distribuční funkcí $F(x)$, pozorujeme vlastně náhodný vektor $\underline{X} = (X_1,...,X_n)^T$, $X_i \approx F(x)$, jehož složky jsou vzájemně nezávislé náhodné veličiny s touž distribuční funkcí $F(x)$. Tento vektor nazýváme **náhodný výběr** z rozdělení $F(x)$ nebo **náhodný výběr ze základního souboru** (nebo **populace**) s rozdělením $F(x)$.Číslo $n$ se nazývá **rozsah** náhodného výběru.

Podle rozsahu obvykle rozdělujeme náhodné výběry na **malé** ( $n \leq 30$) a **velké** ($n > 30$).

Náhodný výběr má zřejmě **sdruženou distribuční funkci $F(\underline{x})$**:

$$F(\underline{x}) = F(x_1,...,x_n) = P(X_1 < x_1;\ldots;X_n < x_n) = P(X_1 < x)\cdot\ldots\cdot P(X_n < x_n) =$$
$$= F(x_1)\cdot\ldots\cdot F(x_n) = \prod_{i=1}^{n} F(x_i)$$

a podobně i **sdruženou hustotu pravděpodobnosti**:

$$f(\underline{x}) = \prod_{i=1}^{n} f(x_i)$$

Číselný vektor $(x_1,..., x_n)'$, který získáme při realizaci náhodného výběru $(X_1, \ldots, X_n)'$, nazýváme **statistický soubor** nebo vzorek o rozsahu $n$. Množina všech těchto vektorů se nazývá **výběrový prostor**. Je to zřejmě podmnožina množiny $R^n$.

Řadu informací o posuzované náhodné veličině X poskytují její číselné charakteristiky, např. *EX*, *DX*, $\sigma_X$ atd. Při statistické indukci jsme při určování jejich hodnot odkázáni na realizace náhodných výběrů, tedy na statistické soubory. Užíváme přitom následující pojmy:

Funkci náhodného výběru $\underline{X} = (X_1,..., X_n)^T$, k jejímuž určení není třeba znát konkrétní hodnoty parametrů příslušného rozdělení, nazýváme **statistika** nebo **výběrová charakteristika** a značíme ji $T(\underline{X}) = T(X_1, \dots, X_n)$. Je to obecně náhodná veličina.

Její hodnotu $t = T(x_1,\dots,x_n)$, kterou nabývá na statistickém souboru $(x_1,\dots,x_n)^T$, nazýváme **pozorovaná hodnota statistiky *T*** nebo empirická charakteristika.

**Používáme zejména následující statistiky**:

1. $T_1(\underline{X}) = \overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ … **výběrový průměr**

   $E(T_1(\underline{X})) = E\overline{X} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} EX_i = \dfrac{EX_i}{n} \cdot \sum_{i=1}^{n} 1 = EX_i$

2. $T_2(\underline{X}) = S^2 = \dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$ … **výběrový rozptyl**

   Není těžké ukázat, že $ES^2 = DX_i$

3. $T_3(\underline{X}) = \sqrt{S^2} = S$ … **výběrová směrodatná odchylka**

Nechť v daném výběru je počet prvků se sledovanou vlastností $x_V$, pak:

4. $T_4(\underline{X}) = p = \dfrac{x_V}{n}$ … **výběrová relativní četnost (výběrový podíl)**

## 9.4 Výběrová rozdělení – rozdělení statistik či výběrových charakteristik

Předpokládejme, že daný náhodný výběr pochází z normálního rozdělení:

$$\underline{X} = (X_1,..., X_n)^T, \quad X_i \to N(\mu, \sigma^2)$$

1. $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} \to N\left(\mu, \dfrac{\sigma^2}{n}\right)$

   (plyne jednak z centrální limitní věty pro velká *n*, ale dá se také ukázat na základě odvození rozdělení součtu náhodných veličin)

2. $Z = \dfrac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$

   (plyne z transformace předešlého rozdělení)

3. $\chi = \frac{S^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$

(bylo vysvětleno při diskusi rozdělení $\chi^2$)

4. $T = \frac{\overline{X} - \mu}{S} \cdot \sqrt{n} \to t_{n-1}$

(odvozeno při diskusi o použití Studentova rozdělení)

5. $P_1 = \frac{p - \pi}{\sqrt{\pi(1-\pi)}} \cdot \sqrt{n} \to N(0;1)$

(odvozeno při diskuzi aplikaci centrální limitní věty – kap. 7.5.1)

Nyní předpokládáme dva výběry z normálních rozdělení:

$$\underline{X} = (X_1,...,X_{n_X})^T,\ X_i \to N(\mu_X, \sigma_X^2),\ \underline{Y} = (Y_1,...,Y_{n_Y})^T,\ Y_j \to N(\mu_Y, \sigma_Y^2).$$

Potom platí:

6. $Z_2 = \frac{(\overline{X} - \overline{Y}) - (\mu_{x1} - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \to N(0,1)$

$$\overline{X} \to N\left(\mu_x; \frac{\sigma_x^2}{n_x}\right);\quad \overline{Y} \to N\left(\mu_Y; \frac{\sigma_Y^2}{n_Y}\right);\quad \left(\overline{X} - \overline{Y}\right) \to N\left(\mu_X - \mu_Y; \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

7. $F = \frac{\frac{\frac{S_x^2}{\sigma_X^2} \cdot (n_X - 1)}{n_X - 1}}{\frac{\frac{S_y^2}{\sigma_Y^2} \cdot (n_Y - 1)}{n_Y - 1}} = \frac{\frac{S_x^2}{\sigma_X^2}}{\frac{S_y^2}{\sigma_Y^2}} \to F_{n_X-1, n_Y-1}$

(zdůvodněno v souvislosti s *F*-rozdělením)

Předpokládejme speciální případ, že **rozptyly jsou neznámé avšak stejné**: $\sigma_X^2 = \sigma_Y^2$. Potom se dá ukázat, že platí:

8. $T_2 = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{S_x^2(n_X - 1) + S_y^2(n_Y - 1)}} \cdot \sqrt{\frac{n_X \cdot n_Y \cdot (n_X + n_Y - 2)}{n_X + n_Y}} \to t_{n_X + n_Y - 2}$

Nechť mají dané výběrové soubory počty prvků se sledovanou vlastností $x_V$ a $y_V$. Pak výběrové relativní četnosti určíme jako:

$$p_X = \frac{x_V}{n_X};\quad p_Y = \frac{y_V}{n_Y}$$

Pak platí:

9. $P_2 = \dfrac{(p_X - p_Y) - (\pi_X - \pi_Y)}{\sqrt{p(1-p)\left(\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right)}} \to N(0;1)$, kde $p = \dfrac{x_V + y_V}{n_X + n_Y}$

## Shrnutí:

Statistika pak používá postupy pomocí nichž můžeme, sice s určitým rizikem (předem stanoveným), na základě toho mála usuzovat na chování celku (**populace**). Tomuto zobecňování říkáme **statistická indukce**.

Jakmile jsme postavení před úkol provést určité šetření a analyzovat údaje z něj zjištěné, musíme nejprve rozhodnout, zda budeme toto **šetření** realizovat jako **vyčerpávající nebo výběrové**.

**Vyčerpávající šetření** – to je prošetření všech jednotek statistického souboru (populace).

**Výběrové šetření** – jde o prošetření vybraných jednotek statistického souboru (populace).

Mezi druhy výběrových šetření řadíme **anketu**, **metodu základního masivu**, **záměrný výběr** a **náhodný výběr**.

**Náhodný výběr ($\underline{X}$)** je speciální náhodný vektor, jehož složky jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti.
Číselný vektor $(x_1,..., x_n)'$, který získáme při realizaci náhodného výběru $(X_1, ... , X_n)'$, nazýváme **statistický soubor** nebo vzorek o rozsahu $n$. Množina všech těchto vektorů se nazývá **výběrový prostor**.

Funkci náhodného výběru $\underline{X}=(X_1, ... , X_n)'$, k jejímuž určení není třeba znát konkrétní hodnoty parametrů příslušného rozdělení, nazýváme **statistika** nebo **výběrová charakteristika** a značíme ji $T(\underline{X})$. Její hodnotu $t=T(x_1,..., x_n)$, kterou nabývá na statistickém souboru $(x_1,..., x_n)'$, nazýváme **pozorovaná hodnota statistiky *T*** nebo empirická charakteristika.

Používáme zejména následující statistiky: **výběrový průměr $\bar{x}$, výběrový rozptyl $s^2$ a výběrovou směrodatnou odchylku s** a **výběrový podíl p**.

Za předpokladu, že náhodný výběr pochází z normálního rozdělení pravděpodobnosti, se dají z daného náhodného výběru odvodit další významné statistiky se známým rozdělením:

| Výběrová charakteristika | Rozdělení výběrové charakteristiky |
|---|---|
| $Z = \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n}$ | N(0;1) |
| $T_{n-1} = \frac{\overline{X} - \mu}{S} \cdot \sqrt{n}$ | $t_{n-1}$ |
| $\chi = \frac{(n-1)S^2}{\sigma^2}$ | $\chi^2_{n-1}$ |
| $P_1 = \frac{p - \pi}{\sqrt{\pi(1-\pi)}} \cdot \sqrt{n}$ | N(0;1) |

Máme-li k dispozici dva výběry z normálního rozdělení, setkáváme s následujícími výběrovými statistikami:

| Výběrová charakteristika | Rozdělení výběrové charakteristiky |
|---|---|
| $Z_2 = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$ | N(0;1) |
| $T_2 = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$ | $t_{n_X + n_Y - 2}$ |
| $F = \frac{\frac{S_x^2}{\sigma_X^2}}{\frac{S_y^2}{\sigma_Y^2}}$ | $F_{n_X - 1, n_Y - 1}$ |
| $P_2 = \frac{(p_X - p_Y) - (\pi_X - \pi_Y)}{\sqrt{p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}$ | N(0;1) |

## Otázky

1. Co je statistická indukce ?
2. Charakterizujte pojmy náhodný výběr a statistický soubor.
3. Co jsou výběrové charakteristiky a které z nich se nejčastěji používají ?
4. Vyjmenujte některé z dalších výběrový statistik, tj. statistik odvozených z náhodného výběru z normálního rozdělení a u některých z takto vyjmenovaných statistik se pokuste zdůvodnit rozdělení pravděpodobnosti.