

Proč se zabývat statistikou?

Předmět „Statistika“ je součástí osnov všech lékařských fakult. Vede k tomu několik důvodů:

V praxi odborní zdravotničtí pracovníci často potřebují získat informace ze shromážděných dat. Měli by tedy být schopni vybrat pro zpracování dat adekvátní statistické metody, použít je ve vhodném statistickém programovém systému a získané výsledky správně interpretovat.

Pokud se zdravotnický pracovník zabývá studiem odborné literatury, rovněž se neobejde bez znalosti statistických pojmů, aby byl schopen text pochopit a kriticky zhodnotit.

Při publikování článků v biomedicínských časopisech je používání statistických metod pravidlem.

Čtyři etapy statistického zkoumání

- 1) **plánování statistického šetření** (důležité je stanovení cíle statistického šetření, výběr vhodných statistických metod, ověření předpokladů jejich použitelnosti, stanovení rozsahu výběrového souboru)
- 2) **sběr dat** (orientace na vhodné objekty a jejich podstatné vlastnosti, realizace měření či příprava dotazníků, proškolení týmu, který bude data sbírat)
- 3) **průzkum získaných dat** (kontrola dat z hlediska formálního, logického i početního, roztrídění dat, tvorba tabulek, konstrukce grafů, práce s chybějícími a odlehlými hodnotami, výpočet číselných charakteristik dat)
- 4) **analýza dat** (získání bodových a intervalových odhadů důležitých parametrů dat, testování statistických hypotéz).

Jednotlivé etapy na sebe navazují a vzájemně se ovlivňují. **Opomeneme-li některé podstatné okolnosti při plánování statistického šetření nebo se dopustíme hrubých chyb při sběru dat, pak ani sebesložitější analýza nemůže poskytnout věrohodné výsledky!**

Při průzkumu a analýze dat se využívají různé statistické programové systémy. Umožňují každému uživateli získat velmi snadno i výsledky náročných statistických analýz. Samozřejmě však neupozorní, že je prováděna analýza, která pro daná konkrétní data nemá smysl. Proto je důležité, aby uživatel rozuměl principům metod, znal a ověřoval jejich předpoklady. Na základě statistického zkoumání lze získat adekvátní představu o věcném problému, který uživatel řeší.

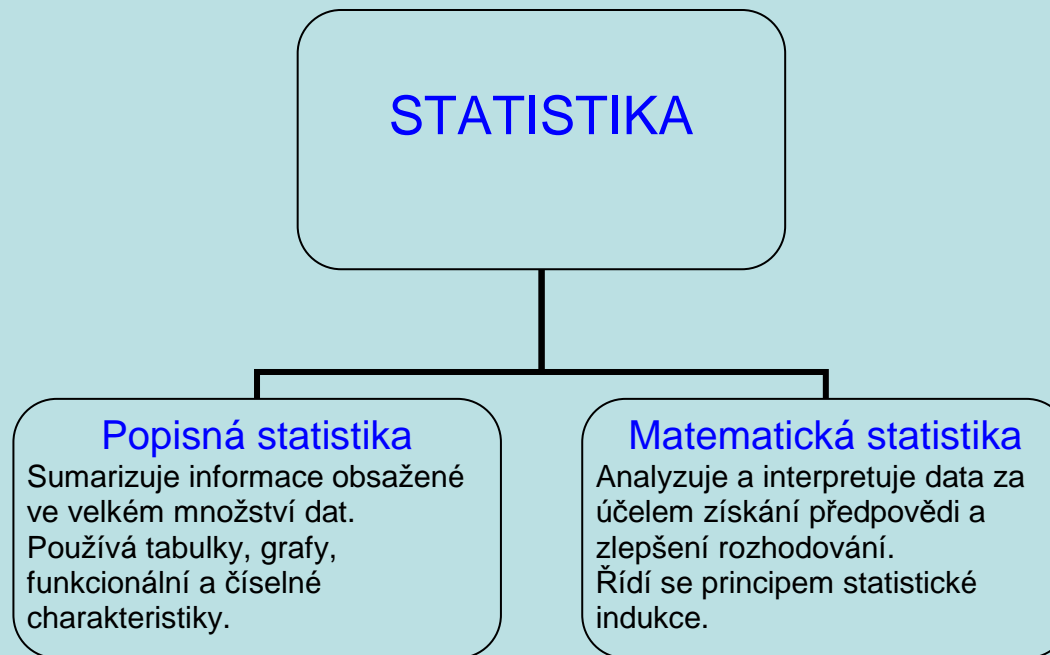
Statistika – různé definice

Statistikou rozumíme soubor číselných údajů o hromadných jevech.

Statistikou rozumíme činnost, která spočívá ve sběru dat a jejich analýze.

Statistikou rozumíme vědeckou disciplínu, která se zabývá získáváním informací z numerických údajů – dat. (My se budeme zabývat statistikou v tomto pojetí).

Rozdělení statistiky



Základem matematické statistiky je počet pravděpodobnosti, který se zabývá studiem zákonitostí v náhodných pokusech. Matematickými prostředky modeluje situace, v nichž hraje roli náhoda. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění při různých provedeních téhož pokusu a nepodléhají naší kontrole.

Popisná statistika

Popisná statistika je disciplína, která popisuje a sumarizuje informace obsažené ve velkém množství dat pomocí tabulek, grafů, funkcionálních a číselných charakteristik. Činí tak pomocí základních matematických operací. Cílem popisné statistiky je zpřehlednit informace „ukryté“ v datových souborech.

Popisná statistika je velmi důležitá minimálně ze dvou důvodů:

- v praxi se často používá (všichni znají takové pojmy, jako je průměr, směrodatná odchylka, tabulka rozložení četností, výsečový graf apod.)
- motivuje pojmy, se kterými pak pracuje počet pravděpodobnosti (např. relativní četnost motivuje pravděpodobnost, hustota četnosti motivuje hustotu pravděpodobnosti, průměr motivuje střední hodnotu apod.)

Dobré pochopení pojmů popisné statistiky tedy velmi usnadní studium počtu pravděpodobnosti.

Základní, výběrový a datový soubor

Základním souborem (population) rozumíme libovolnou neprázdnou množinu E .

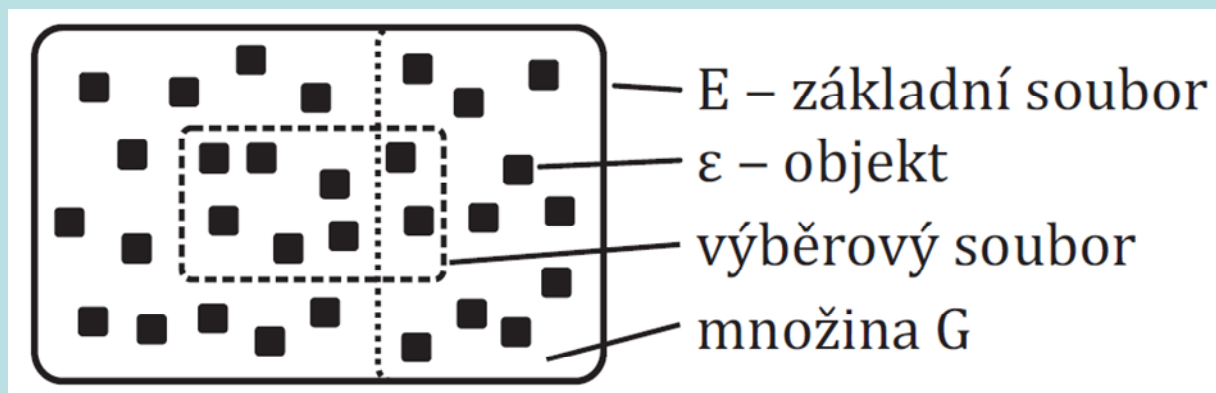
Prvky množiny E značíme ε a nazýváme je **objekty (units)**.

Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, \dots, \varepsilon_n\}$ základního souboru E nazýváme **výběrový soubor rozsahu n (sample size n)**.

Je-li množina $G \subseteq E$, pak symbolem $N(G)$ rozumíme **absolutní četnost (absolute frequency)** množiny G ve výběrovém souboru, tj. počet těch objektů množiny G , které patří do výběrového souboru.

Relativní četnost (absolute frequency) množiny G ve výběrovém souboru zavedeme vztahem $p(G) = \frac{N(G)}{n}$.

Ilustrace



Způsoby získání výběrového souboru

1. **Prostý náhodný výběr (simple random sample)** – objekty výběrového souboru získáme losováním z objektů základního souboru.
2. **Systematický náhodný výběr (systematic random sample)** – objekty výběrového souboru získáme pomocí pořadových čísel nebo podle nějaké vlastnosti, která nesouvisí se sledovanou vlastností (např. podle data narození, podle počátečního písmena příjmení apod. Nevýhoda – při nevhodné volbě výběrového kritéria může dojít k nežádoucí selekci.
3. **Stratifikovaný náhodný výběr (stratified random sample)** – základní soubor rozdělíme do několika skupin a dále z každé této skupiny vybíráme metodou prostého nebo systematického náhodného výběru.
4. **Párový výběr (random pair)** – užívá se zejména v klinické praxi. K osobám s určitou vlastností (např. s určitou nemocí) se vyberou osoby, které tuto nemoc nemají, ale s původními osobami se shodují ve všech vlastnostech, které by mohly ovlivnit výsledek výzkumu, např. věk, pohlaví, zaměstnání apod.

Příklad: Základním souborem E je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, \dots, \varepsilon_{20}\}$. Z těchto 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

Řešení:

$$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20,$$

$$p(G_1) = \frac{12}{20} = 0,6,$$

$$p(G_2) = \frac{15}{20} = 0,75,$$

$$p(G_1 \cap G_2) = \frac{11}{20} = 0,55$$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

Vlastnosti relativní četnosti: Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

- $p(\emptyset) = 0$
- $p(G) \geq 0$ (nezápornost)
- $p(G) \leq 1$
- $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$
- $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$
- $p(G_1 \cup G_2) + 0 \leq p(G_1) + p(G_2)$ (subaditivita)
- $G_1 \cap G_2 = \emptyset \Rightarrow p(G_1 \cup G_2) = p(G_1) + p(G_2)$ (aditivita)
- $p(G_2 \setminus G_1) = p(G_2) - p(G_1 \cap G_2)$
- $G_1 \subseteq G_2 \Rightarrow p(G_2 \setminus G_1) = p(G_2) - p(G_1)$ (subtraktivita)
- $G_1 \subseteq G_2 \Rightarrow p(G_1) \leq p(G_2)$ (monotonie)
- $p(E) = 1$ (normovanost)
- $p(G) + p(\bar{G}) = 1$ (komplementarita)

Pojem podmíněné relativní četnosti (conditional relative frequency): Pokud se v daném základním souboru zajímáme o dvě podmnožiny, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé podmnožiny.

Nechť E je základní soubor, G_1, G_2 jeho podmnožiny, $\{\varepsilon_1, \dots, \varepsilon_n\}$ výběrový soubor. Definujeme:

podmíněnou relativní četnost množiny G_1 ve výběrovém souboru za předpokladu G_2 :

$$p(G_1 / G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)},$$

podmíněnou relativní četnost G_2 ve výběrovém souboru za předpokladu G_1 :

$$p(G_2 / G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}.$$

Příklad: Pro údaje z příkladu o studentech vypočtete podmíněnou relativní četnost úspěšných matematiků mezi úspěšnými angličtináři a podmíněnou relativní četnost úspěšných angličtinářů mezi úspěšnými matematiky.

(Připomínáme, že z 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech.)

Řešení:

$$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20,$$

$$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{11}{15} = 0,73 \text{ (tzn., že 73\% těch studentů, kteří}$$

byli úspěšní v angličtině, uspělo i v matematice)

$$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{11}{12} = 0,92 \text{ (tzn., že 92\% těch studentů, kteří byli}$$

úspěšní v matematice, uspělo i v angličtině)

Pojem četnostní nezávislosti dvou množin (independence of two sets): O četnostní nezávislosti dvou množin v daném výběrovém souboru hovoříme tehdy, když informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ i z druhé množiny.

Řekneme že množiny G_1, G_2 jsou **četnostně nezávislé** v daném výběrovém souboru, jestliže

$$p(G_1 \cap G_2) = p(G_1)p(G_2).$$

(V praxi jen zřídka dojde k tomu, že uvedený vztah platí přesně. Většinou je jen naznačena určitá tendence četnostní nezávislosti.)

Příklad: Pro údaje z příkladu o studentech zjistěte, zda úspěchy v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé. (připomínáme, že oboustranně úspěšných studentů bylo 55 %, úspěšných matematiků 60 % a úspěšných angličtinářů 45 %.)

Řešení:

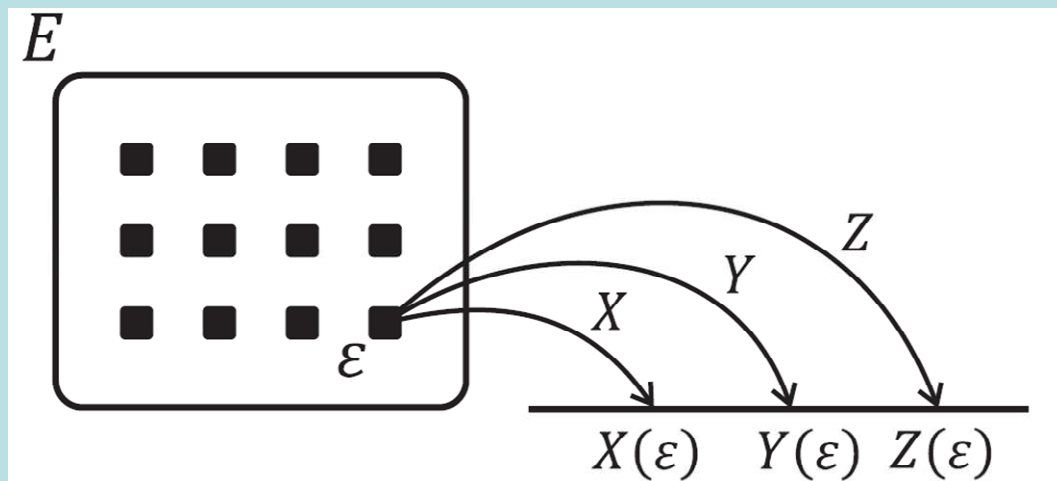
$$p(G_1 \cap G_2) = 0,55, p(G_1)p(G_2) = 0,6 \times 0,75 = 0,45,$$

tedy skutečná relativní četnost oboustranně úspěšných studentů je větší než by odpovídalo četnostní nezávislosti množin G_1, G_2 v daném výběrovém souboru. Znamená to, že úspěch v matematice se zpravidla sdružuje s úspěchem v angličtině a naopak.

Pojem skalárního a vektorového znaku (scalar and vector variable): Vlastnosti objektů vyjadřujeme číselně pomocí znaků.

Nechť E je základní soubor. Funkce $X: E \rightarrow \mathbb{R}$, $Y: E \rightarrow \mathbb{R}$, ..., $Z: E \rightarrow \mathbb{R}$, které každému objektu přiřazují číslo, se nazývají **(skalární) znaky**. Uspořádaná p -tice (X, Y, \dots, Z) se nazývá **vektorový znak**.

Ilustrace



Označení: Nechť je dán výběrový soubor $\{\varepsilon_1, \dots, \varepsilon_n\} \subseteq E$. Hodnoty znaků X, Y, \dots, Z pro i -tý objekt označíme $x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, ..., $z_i = Z(\varepsilon_i)$, $i = 1, \dots, n$.

Pojem datového souboru (data set):

Matrice $\begin{pmatrix} X_1 & Y_1 & \cdots & Z_1 \\ X_2 & Y_2 & \cdots & Z_2 \\ \cdots & \cdots & \cdots & \cdots \\ X_n & Y_n & \cdots & Z_n \end{pmatrix}$ typu $n \times p$ se nazývá **datový soubor**.

Její řádky odpovídají jednotlivým objektům, sloupce znakům.

Libovolný sloupec této matice nazýváme **jednorozměrným datovým souborem (one-dimensional data set)**.

Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle

velikosti, dostaneme **uspořádaný datový soubor (orderly data set)** $\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}$, kde $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

Vektor $\begin{pmatrix} X_{[1]} \\ \vdots \\ X_{[r]} \end{pmatrix}$, kde $x_{[1]} < \dots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá **vektor variant (vector of variants)**.

Příklad:

Pro studenty z výběrového souboru uvedeného výše byly zjišťovány hodnoty těchto znaků:
X – známka z matematiky v 1. zkušebním termínu, Y – známka z angličtiny v 1.
zkušebním termínu, Z – pohlaví studenta (0 ... žena, 1 ... muž). Byl získán datový soubor:

```
1 2 0
2 3 1
3 4 0
4 5 1
5 6 0
6 7 1
7 8 0
8 9 1
9 10 0
10 11 1
```

Utvořte jednorozměrný neuspořádaný a uspořádaný datový soubor pro známky z matematiky a vektor variant pro známky z matematiky.

Řešení:

```
1 2 3 4 5 6 7 8 9 10
0 1 0 1 0 1 0 1 0 1
0 1 0 1 0 1 0 1 0 1
```

Pojem jevu (event):

Nechť $\{\varepsilon_1, \dots, \varepsilon_n\}$ je výběrový soubor, X, Y, \dots, Z jsou znaky, B, B_1, B_p jsou číselné množiny.

Zápis $\{X \in B\}$ znamená jev „znak X nabyl hodnoty z množiny B “.

Zápis $\{X \in B_1 \wedge Y \in B_2 \wedge \dots \wedge Z \in B_p\}$ znamená jev „znak X nabyl hodnoty z množiny B_1 a současně znak Y nabyl hodnoty z množiny B_2 atd. až znak Z nabyl hodnoty z množiny B_p “.

Symbol $N(X \in B)$ značí **absolutní četnost** jevu $\{X \in B\}$ ve výběrovém souboru, tj. počet těch objektů ve výběrovém souboru, pro něž $x_i \in B$.

Symbol $p(X \in B)$ znamená **relativní četnost** jevu $\{X \in B\}$ ve výběrovém souboru, tj. $p(X \in B) = \frac{N(X \in B)}{n}$.

Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge \dots \wedge Z \in B_p)$ resp.

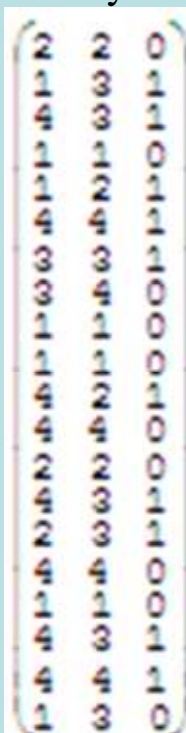
$p(X \in B_1 \wedge Y \in B_2 \wedge \dots \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge \dots \wedge Z \in B_p\}$ ve výběrovém souboru.

Příklad:

Pro datový soubor s údaji o známkách a pohlaví studentů najděte relativní četnost:

- matematických jedničkářů
- úspěšných matematiků
- oboustranně neúspěšných studentů.

Datový soubor má tvar:



1	7
úspěšný	12
0	4

Řešení:

$$\text{Ad a)} p(X = 1) = \frac{7}{20} = 0,35$$

Ad b)

$$p(X \leq 3) = \frac{12}{20} = 0,6$$

Ad c)

$$p(X = 4 \wedge Y = 4) = \frac{4}{20} = 0,2$$

Zjistili jsme, že jedničku z matematiky má 35 % studentů, zkoušku z matematiky úspěšně složilo 60 % studentů a oboustranně neúspěšných bylo 20 % studentů.

Jednorozměrné bodové rozložení četností

Jestliže počet variant znaku X v jednorozměrném datovém souboru není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o **bodovém rozložení četností**.

Nechť je dán jednorozměrný datový soubor $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, v němž znak X nabývá r variant.

Pro $j = 1, \dots, r$ definujeme:

$n_j = N(X = x_{[j]})$ – **absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru**

$p_j = \frac{n_j}{n}$ – **relativní četnost varianty $x_{[j]}$ ve výběrovém souboru**

$N_j = N(X \leq x_{[j]}) = n_1 + \dots + n_j$ – **absolutní kumulativní četnost prvních j variant ve výběrovém souboru**

$F_j = \frac{N_j}{n} = p_1 + \dots + p_j$ – **relativní kumulativní četnost prvních j variant ve výběrovém souboru**

Tabulka typu

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
\vdots	\vdots	\vdots	\vdots	\vdots
$x_{[r]}$	n_r	p_r	N_r	F_r

se nazývá **variační řada** (nebo též **tabulka rozložení četností – frequency table**).

Příklad: Máme jednorozměrný datový soubor, který obsahuje údaje o známkách z matematiky (znak X) u 20 studentů.

(
2
1
4
1
1
4
3
3
1
1
4
4
2
4
2
4
1
4
4
1
)

Sestavte tabulku rozložení četností.

Řešení:

$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	$7/20=0,35$	7	$7/20=0,35$
2	3	$3/20=0,15$	10	$10/20=0,50$
3	2	$2/20=0,10$	12	$12/20=0,60$
4	8	$8/20=0,40$	20	$20/20=1,00$
Σ	20	1,00	-	-

Četnostní funkce

Pomocí relativních četností zavedeme **četnostní funkci (frequency function)**.

Funkce $p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]}, j = 1, \dots, r \\ 0 & \text{jinak} \end{cases}$ se nazývá četnostní funkce.

(Hodnota četnostní funkce v každé variantě znaku X je rovna relativní četnosti této varianty a jinde je nulová.)

Četnostní funkce je

nezáporná ($\forall x \in R: p(x) \geq 0$)

a normovaná (součet všech jejích hodnot je 1, tj. $\sum_{x=-\infty}^{\infty} p(x) = 1$).

Empirická distribuční funkce

Pomocí kumulativních relativních četností zavedeme **empirickou distribuční funkci (empirical distribution function)**.

$$\text{Funkce } F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \leq x < x_{[j+1]}, j = 1, \dots, r-1 \\ 1 & \text{pro } x \geq x_{[r]} \end{cases} \text{ se nazývá empirická distribuční funkce.}$$

(Vysvětlení: Funkce $F(x)$ je nulová až do první varianty $x_{[1]}$ znaku X . Tam má skok na hodnotu $F_1 = p_1$. Na hodnotě F_1 setrvává až do druhé varianty $x_{[2]}$, kde skočí na hodnotu F_2 , tedy skok má velikost p_2 . Tak se pokračuje dál, až v poslední variantě $x_{[r]}$ skočí na 1 a tam už setrvává.)

Empirická distribuční funkce je

neklesající ($\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2: F(x_1) \leq F(x_2)$),

zprava spojitá ($\forall x_0 \in \mathbb{R}$ libovolné, ale pevně dané: $\lim_{x \rightarrow x_0+} F(x) = F(x_0)$)

a normovaná ($\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$).

Příklad: Pro známky z matematiky nakreslete graf četnostní funkce a empirické distribuční funkce.

Řešení:

Variační řada

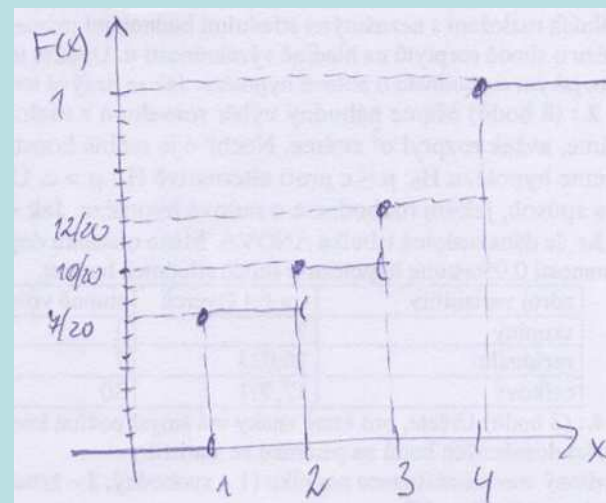
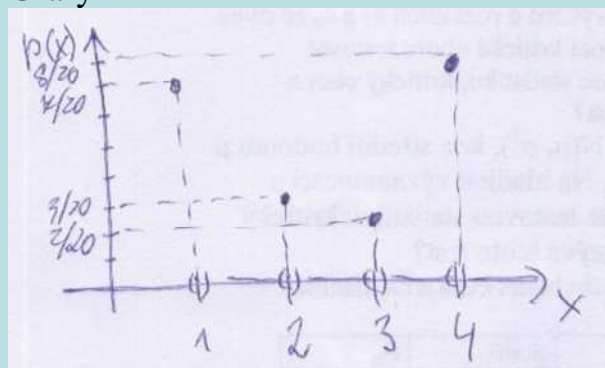
$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	$7/20=0,35$	7	$7/20=0,35$
2	3	$3/20=0,15$	10	$10/20=0,50$
3	2	$2/20=0,10$	12	$12/20=0,60$
4	8	$8/20=0,40$	20	$20/20=1,00$
Σ	20	1,00	-	-

Vzorce

$$p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]}, j=1, \dots, r \\ 0 & \text{jinak} \end{cases}$$

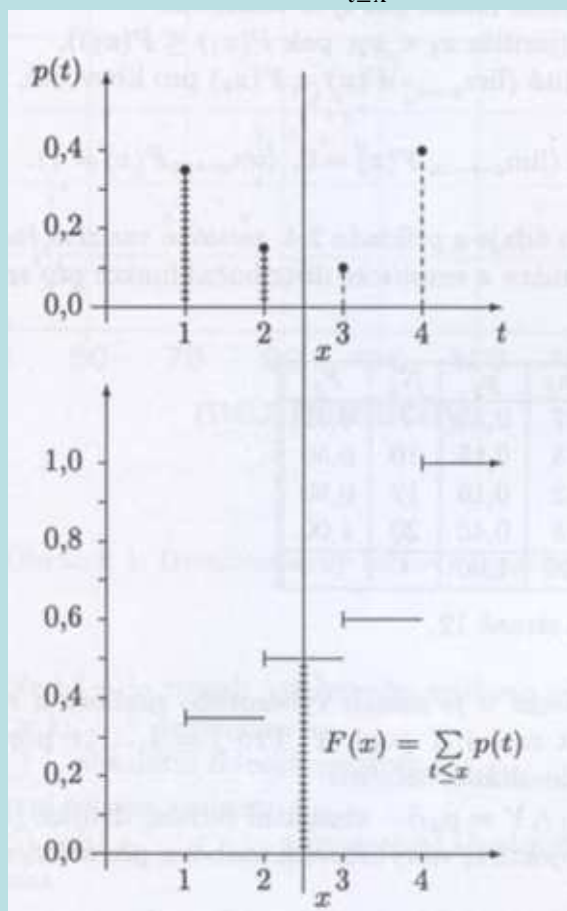
$$F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \leq x < x_{[j+1]}, j=1, \dots, r-1 \\ 1 & \text{pro } x \geq x_{[r]} \end{cases}$$

Grafy



Vztah mezi četnostní funkcí a empirickou distribuční funkcí

$$\forall x \in \mathbb{R} : F(x) = \sum_{t \leq x} p(t)$$



Grafické znázornění bodového rozložení četností

Tečkový diagram (dot diagram): na číselné ose vyznačíme jednotlivé varianty znaku X a nad každou variantu nakreslíme tolik teček, jaká je její absolutní četnost.

Polygon četnosti (frequency polygon): je lomená čára spojující body, jejichž x-ová souřadnice je varianta znaku X a y-ová souřadnice je absolutní či relativní četnost této varianty.

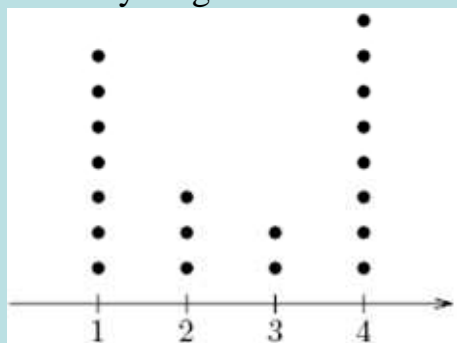
Sloupkový diagram (bar chart): je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní či relativní četnost této varianty.

Výsečový graf (pie chart): je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním četnostem variant znaku X.

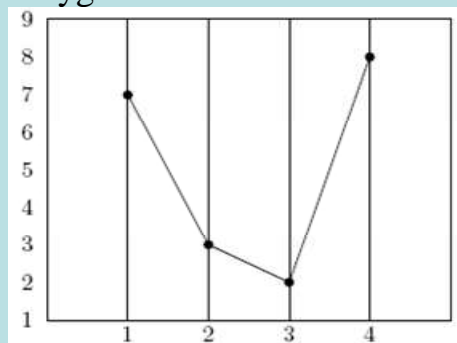
Příklad: Pro jednorozměrný datový soubor známek z matematiky sestrojte tečkový diagram, polygon četností, sloupkový diagram a výsečový graf.

Řešení:

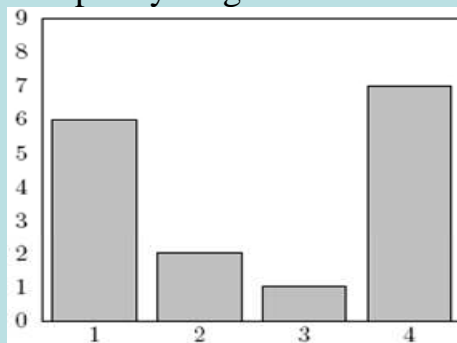
Tečkový diagram



Polygon četností



Sloupkový diagram



Výsečový graf



Dvourozměrné bodové rozložení četností

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} \mathbf{X}_1 & \mathbf{y}_1 \\ \dots & \dots \\ \mathbf{X}_n & \mathbf{y}_n \end{pmatrix}$, kde znak X má r variant a znak Y má s variant. Pak definujeme:

$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]})$ – **simultánní absolutní četnost dvojice $(x_{[j]}, y_{[k]})$** ve výběrovém souboru

$p_{jk} = \frac{n_{jk}}{n}$ – **simultánní relativní četnost dvojice $(x_{[j]}, y_{[k]})$** ve výběrovém souboru

$n_{.j} = N(X = x_{[j]}) = n_{j1} + \dots + n_{js}$ – **marginální absolutní četnost varianty $x_{[j]}$**

$p_{.j} = \frac{n_{.j}}{n} = p_{j1} + \dots + p_{js}$ – **marginální relativní četnost varianty $x_{[j]}$**

$n_{.k} = N(Y = y_{[k]}) = n_{1k} + \dots + n_{rk}$ – **marginální absolutní četnost varianty $y_{[k]}$**

$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \dots + p_{rk}$ – **marginální relativní četnost varianty $y_{[k]}$**

Simultánní četností zapisujeme do **kontingenční tabulky (contingency table)**.

Kontingenční tabulka simultánních absolutních četností má tvar:

	y	$y_{[1]}$	\dots	$y_{[s]}$	$n_{.j}$
x	n_{jk}				
$X_{[1]}$		n_{11}	\dots	n_{1s}	$n_{1.}$
\vdots		\dots	\dots	\dots	\dots
$X_{[r]}$		n_{r1}	\dots	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	\dots	$n_{.s}$	n

Příklad: Máme datový soubor, který obsahuje údaje o známkách z matematiky (znak X), z angličtiny (znak Y) a pohlaví studenta (znak Z, 0 – žena, 1 – muž) u 20 studentů:

X	2	1	4	1	1	4	3	3	1	1	4	4	2	4	2	4	1	4	4	1
Y	2	3	3	1	2	4	3	4	1	1	2	4	2	3	3	4	1	3	4	3
Z	0	1	1	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	0

Vytvořte kontingenční tabulku simultánních absolutních a relativních četností pro známky z matematiky a angličtiny.

Řešení:

Kontingenční tabulka simultánních absolutních četností

		<i>y</i>				<i>n_{j.}</i>
		1	2	3	4	
<i>x</i>	<i>n_{jk}</i>					
	1		4	1	2	0
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
<i>n_{.k}</i>		4	4	7	5	<i>n</i> = 20

Kontingenční tabulka simultánních relativních četností

		<i>y</i>				<i>p_{j.}</i>
		1	2	3	4	
<i>x</i>	<i>p_{jk}</i>					
	1		0,20	0,05	0,10	0,00
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
<i>p_{.k}</i>		0,20	0,20	0,35	0,25	1,00

Simultánní a marginální četnostní funkce

Pomocí simultánních relativních četností zavedeme **simultánní četnostní funkci**:

$$\text{Funkce } p(x, y) = p(x, y) = \begin{cases} p_{jk} \text{ pro } x = x_{[j]}, y = y_{[k]}, j = 1, \dots, r, k = 1, \dots, s \\ 0 \text{ jinak} \end{cases}$$

se nazývá simultánní četnostní funkce.

Pomocí marginálních relativních četností zavedeme **marginální četnostní funkce pro znaky X a Y**.

Odlišíme je indexem takto:

$$p_1(x) = \begin{cases} p_{.j} \text{ pro } x = x_{[j]}, j = 1, \dots, r \\ 0 \text{ jinak} \end{cases}, p_2(y) = \begin{cases} p_{.k} \text{ pro } y = y_{[k]}, k = 1, \dots, s \\ 0 \text{ jinak} \end{cases}.$$

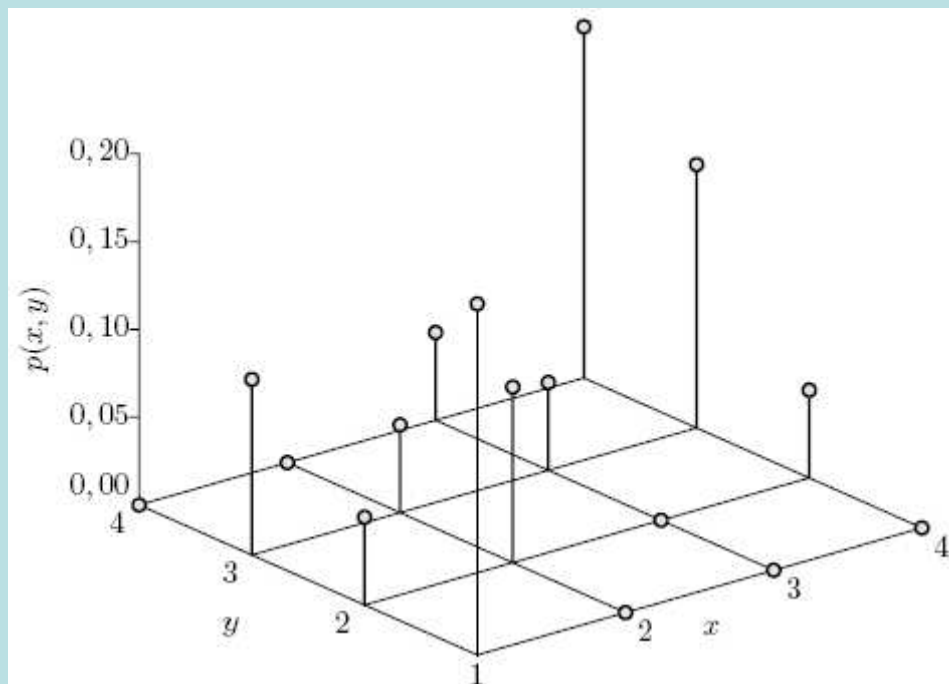
Mezi simultánní četnostní funkcí a marginálními četnostními funkcemi platí vztahy:

$$p_1(x) = \sum_{y=-\infty}^{\infty} p(x, y), p_2(y) = \sum_{x=-\infty}^{\infty} p(x, y).$$

Příklad: Sestrojte graf simultánní četnostní funkce pro známky z matematiky a angličtiny.

Řešení: Vyjdeme z kontingenční tabulky simultánních relativních četností.

	y	1	2	3	4	$P_{j.}$
x	P_{jk}					
1		0,20	0,05	0,10	0,00	0,35
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
$P_{.k}$		0,20	0,20	0,35	0,25	1,00



Četnostní nezávislost znaků v daném výběrovém souboru

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé, právě když

pro všechna $j = 1, \dots, r$ a všechna $k = 1, \dots, s$ platí multiplikativní vztah: $p_{jk} = p_j \cdot p_k$

neboli pro $\forall (x, y) \in \mathbb{R}^2 : p(x, y) = p_1(x)p_2(y)$.

Příklad: Ověřte, zda v našem datovém souboru jsou známky z matematiky a angličtiny četnostně nezávislé.

Řešení: Vyjdeme z kontingenční tabulky simultánních relativních četností:

	<i>y</i>	1	2	3	4	p_j
<i>x</i>	p_{jk}					
1		0,20	0,05	0,10	0,00	0,35
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
p_k		0,20	0,20	0,35	0,25	1,00

Známky z matematiky a angličtiny nejsou četnostně nezávislé, protože už pro $j = 1, k = 1$ je multiplikativní vztah porušen:

$p_{11} = 0,20, p_{.1} = 0,35, p_{1.} = 0,20$, tudíž $0,20 \neq 0,35 \cdot 0,20$

Řádkově a sloupcově podmíněné relativní četnosti

$$p_{j(k)} = \frac{n_{jk}}{n_{.k}} - \text{sloupcově podmíněná relativní četnost varianty } x_{[j]} \text{ za předpokladu } y_{[k]}$$

$$p_{(j)k} = \frac{n_{jk}}{n_{j.}} - \text{řádkově podmíněná relativní četnost varianty } y_{[k]} \text{ za předpokladu } x_{[j]}$$

Podmíněné relativní četnosti zapisujeme do kontingenční tabulky. Často je vyjadřujeme v procentech.

Příklad: Pro datový soubor známek z matematiky a angličtiny sestavte kontingenční tabulku sloupcově a poté řádkově podmíněných relativních četností.

Řešení:

Nejprve vypočítáme sloupcově podmíněné relativní četnosti. Použijeme kontingenční tabulku simultánních absolutních četností.

	y	1	2	3	4	$n_{j\cdot}$
x	n_{jk}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
$n_{\cdot k}$		4	4	7	5	$n = 20$

	y	1	2	3	4
x	$P_{j(k)}$				
1		1,00	0,25	0,29	0,00
2		0,00	0,50	0,14	0,00
3		0,00	0,00	0,14	0,20
4		0,00	0,25	0,43	0,80
Σ		1,00	1,00	1,00	1,00

Interpretujeme např. třetí sloupec: z těch studentů, kteří měli trojku z angličtiny, mělo $2/7 = 29\%$ jedničku z matematiky, $1/7 = 14\%$ dvojku z matematiky, $1/7 = 14\%$ trojku z matematiky a $3/7 = 43\%$ čtyřku z matematiky.

Nyní vypočítáme řádkově podmíněné relativní četnosti. Opět použijeme kontingenční tabulku simultánních absolutních četností.

	y	1	2	3	4	$n_{j\cdot}$
x	n_{jk}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
$n_{\cdot k}$		4	4	7	5	$n = 20$

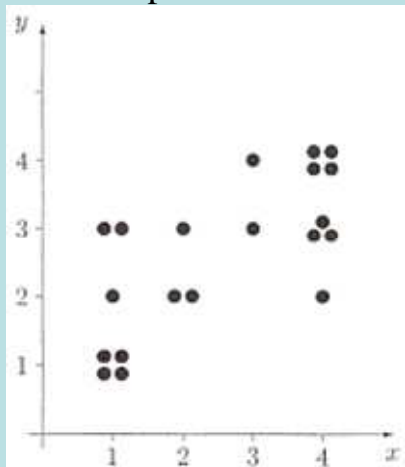
	y	1	2	3	4	Σ
x	$P_{(j)k}$					
1		0,57	0,14	0,29	0,00	1,00
2		0,00	0,67	0,33	0,00	1,00
3		0,00	0,00	0,50	0,50	1,00
4		0,00	0,12	0,38	0,50	1,00

Interpretujeme např. první řádek: z těch studentů, kteří měli jedničku z matematiky, mělo $4/7 = 57\%$ jedničku z angličtiny, $1/7 = 14\%$ dvojku z angličtiny a $2/7 = 29\%$ trojku z angličtiny.

Dvourozměrný tečkový diagram (scatter plot)

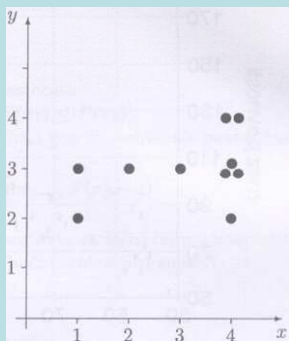
Dvourozměrné rozložení četností lze znázornit pomocí **dvourozměrného tečkového diagramu**. Na vodorovnou osu vyneseme varianty znaku X, na svislou varianty znaku Y a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dané dvojice.

V našem příkladě se studenty dostaneme tento diagram:



Dvourozměrný tečkový diagram svědčí o nepříliš výrazné tendenci k podobné klasifikaci v obou předmětech. Zcela odlišný vzhled však mají diagramy pro muže a pro ženy:

Pro muže



Pro ženy

