0.1 Základy statistického zpracování dat

Statistika se zabývá shromažďováním, tříděním a popisem velkých souborů dat. Někdy se pod pojmem statistika myslí přímo nashromážděná data, jindy spíše činnost spojená s jejich získáváním a zpracováním. Předmětem statistiky je také hledání zákonitostí v těchto datech a předpověď budoucího vývoje. V závěru našeho kurzu se seznámíte s tzv. testováním statistických hypotéz, zde, na začátku, si však pouze definujeme základní pojmy a budeme se věnovat různým popisným charakteristikám souboru dat.

Ve statistickém šetření zkoumáme vlastnosti určité skupiny objektů. Tyto objekty mohou být různého druhu: zaměstnanci podniku, u kterých sledujeme např. jejich výkonnost, vzdělání a plat; pokusné myši, u kterých sledujeme reakci na podanou látku; výrobky, u kterých sledujeme jejich kvalitu, apod.

Zkoumané objekty nazýváme **statistickými jednotkami**. Množinu všech statistických jednotek nazveme **statistickým souborem**. Vlastnosti statistických jednotek vyjadřují **statistické znaky**.
Zjišťujeme-li u každé statistické jednotky pouze jeden statistický znak, získáváme tak soubor **jednorozměrný**. Zjišťujeme-li dva nebo více znaků a zkoumáme-li jejich vzájemné vztahy, hovoříme o souborech **dvourozměrných**, resp. **vícerozměrných**.

Při statistickém zkoumání se snažíme udělat nějaký závěr ohledně vlastností celého statistického souboru (např. všech občanů ČR, všech výrobků určitého závodu, apod.). Často je však nemožné použít opravdu všechny statistické jednotky a musíme se omezit pouze na vybranou podmnožinou statistického souboru.

Podle rozsahu můžeme zkoumané soubory rozdělit na dva typy:

- **Základní soubor** (populace) – obsahuje všechny jednotky.
- **Výběrový soubor** (výběr) – obsahuje pouze některé jednotky.

Z vlastností výběrového souboru se pak snažíme dělat závěry pro celý základní soubor. Proto si při výběru prvků musíme počínat opatrně, výběrový soubor by měl být reprezentativní.

Příklad 0.1. Jestliže zvolíme za statistickou jednotku studenta VUT, lze tuto jednotku charakterizovat např. pomocí znaků udávajících ročník, fakultu, na které studuje, vážený studijní průměr, atd. Vidíme, že znaky mohou být několika různých typů. Některé lze popsat číselnou hodnotou, pro náš příklad by to byl ročník a průměr. Vyjádřit fakultu číslem však v podstatě nelze. Mohli bychom si sice zavést označení např. ale takto zvolená čísla by pak sloužila pouze jako indexy. Nemělo by význam počítat např. „průměrnou fakultu“ všech studentů.

Statistické znaky rozlišujeme:

- **Kvantitativní** – lze je vyjádřit číselnou hodnotou. Tyto znaky můžeme dále rozdělit na
  - spojité – mohou nabývat kterékoli hodnoty z určitého intervalu (např. spotřeba elektřiny),
  - nespojité (diskrétní) – mohou nabývat pouze hodnot z určité konečné nebo spočetné množiny, často se jedná o celočíselné hodnoty (např. počet dětí v rodině).
- **Kvalitativní** – jsou popsány slovně.

My se budeme zabývat převážně znaky kvantitativními.

0.2 Rozdělení četností

Budeme zkoumat jednorozměrný statistický soubor o celkovém rozsahu n statistických jednotek. Cílem je zjistit, jak často se v souboru vyskytují jednotlivé hodnoty sledovaného kvantitativního znaku x. Soubor seřadíme podle velikosti x. Další postup se však trochu liší pro znaky spojité a nespojité (diskrétní).

0.2.1 Diskrétní znaky

Předpokládejme, že v souboru o rozsahu n může sledovaný znak x nabývat k různých hodnot (variant) $x_1, x_2, \ldots, x_k$. **Četnost** varianty x_i je počet výskytů této hodnoty ve sledovaném souboru a označíme ji n_i, $i = 1, \ldots, k$. Pak platí

$$n_1 + n_2 + \cdots + n_k = n.$$

Často je přehlednější pracovat spíše s relativními četnostmi. Můžeme pak např. porovnávat rozdělení četností znaku u dvou souborů o různém rozsahu.

Relativní četnost varianty x_i zavedeme jako

$$f_i = \frac{n_i}{n}\,.$$

Pro relativní četnosti platí

$$f_1 + \cdots + f_k = \frac{n_1}{n} + \cdots \frac{n_k}{n} = \frac{n_1 + \cdots + n_k}{n} = 1.$$

Relativní četnost se často vyjadřuje i v procentech.

Užitečné jsou také tzv. **kumulativní četnosti** (opět absolutní nebo relativní). Ty udávají, kolik jednotek má hodnotu znaku menší nebo rovnou vybrané variantě x_i.

Pro zobrazení četností se u diskrétních kvantitativních znaků používá **spojnicový graf** (zvaný též polygon četností) nebo **sloupcový graf**, viz obrázky 1 a 2.

Příklad 0.2. Zkoumáme věk studentů nastupujících do 1. ročníku vysoké školy. Máme k dispozici tabulku, v níž jsou pořadová čísla studentů a jejich věky:

Varianta znaku	Četnost		Kumulativní četnost	
	absolutní	relativní	absolutní	relativní
x_1	n_1	f_1	n_1	f_1
x_2	n_2	f_2	$n_1 + n_2$	$f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	$n_1 + \cdots + n_k = n$	$f_1 + \cdots + f_k = 1$

Tab. 1: Tabulka četností a kumulativních četností

ID	Věk	ID	Věk	ID	Věk	ID	Věk	ID	Věk	ID	Věk	ID	Věk
1	19	11	19	21	19	31	19	41	20	51	19	61	19
2	19	12	19	22	19	32	19	42	19	52	22	62	19
3	19	13	20	23	20	33	23	43	20	53	19	63	20
4	19	14	19	24	21	34	20	44	20	54	19	64	20
5	19	15	19	25	20	35	18	45	19	55	19	65	19
6	19	16	19	26	19	36	19	46	19	56	19	66	20
7	20	17	21	27	19	37	20	47	20	57	19	67	20
8	20	18	19	28	20	38	19	48	19	58	20	68	21
9	19	19	19	29	19	39	19	49	19	59	19	69	19
10	22	20	19	30	19	40	20	50	19	60	20	70	19

Najděte rozdělení četností věku studentů.

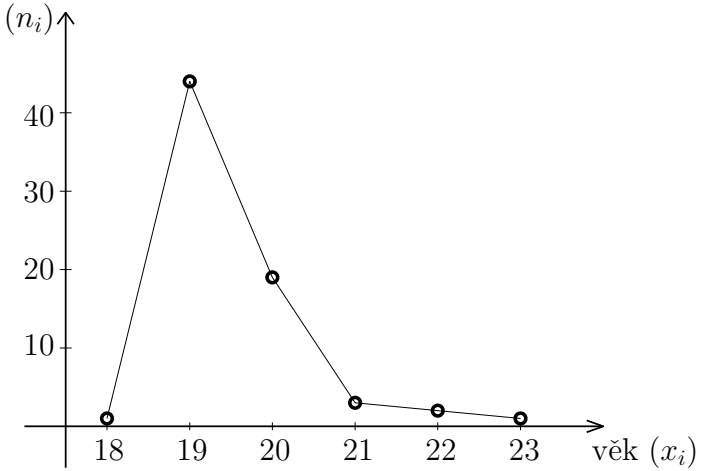
Řešení. Vidíme, že studentů je celkem $n = 70$ a že věk nabývá hodnot z množiny $\{18, 19, 20, 21, 22, 23\}$. Osmnáctiletý student je jeden, devetenáctiletých je 44, atd. Tabulka četností proto bude vypadat takto (relativní četnosti jsou zaokrouhleny na 3 desetinná místa):

Věk studenta x_i	Počet studentů n_i	Relativní četnost f_i	Kumulativní absolutní četnost	Kumulativní relativní četnost
18	1	0,014	1	0,014
19	44	0,629	45	0,643
20	19	0,271	64	0,914
21	3	0,043	67	0,957
22	2	0,029	69	0,986
23	1	0,014	70	1,000

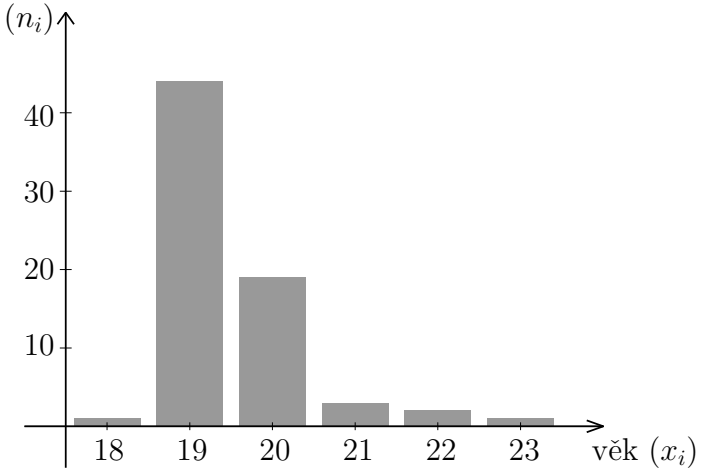
Graficky jsou absolutní četnosti znázorněny na obrázcích 1 a 2. Podobně by vypadal obrázek pro relativní četnosti.

□

Výše popsané způsoby zpracování četností jsou vhodné pro diskrétní znaky, které mohou nabývat pouze malého počtu hodnot. Zkoumáme-li diskrétní znak, který může nabývat mnoha různých hodnot, je lepší



Obr. 1: Spojnicový graf



Obr. 2: Sloupcový graf

hodnoty seskupit do intervalů a pracovat s těmito intervaly. Je to stejný postup, jaký se používá pro spojité znaky.

0.2.2 Spojité znaky

Spojité znaky mohou nabývat jakékoli hodnoty z určitého intervalu (záleží na povaze zkoumaného znaku). Tabulku četností popsanou v předchozí kapitole nemůžeme proto dost dobře sestavit. Může se stát, že máme soubor velkého rozsahu, ale žádná hodnota se v něm neopakuje. Proto pro spojité znaky nebo pro znaky sice diskrétní, ale s velkým počtem možných variant, konstruujeme intervalové rozdělení četností. Zde je důležitá otázka, do kolika intervalů máme hodnoty roztřídit. Příliš malý počet intervalů vede k velmi hrubému pohledu na rozdělení četností. Příliš velký počet intervalů vede k tomu, že graf je „střapatý“ a nevyniknou zákonitosti charakteristické pro daný soubor. Pro orientační odhad vhodného počtu intervalů se používají různá pravidla, z nichž nejpoužívanější je Sturgesovo (viz rámeček).

Při konstrukci **intervalového rozdělení četností** stanovujeme počty výskytů hodnot znaku, které náleží do předem vymezených intervalů. Pro stanovení počtu intervalů se často používá tzv. Sturgesovo pravidlo

$$k \doteq 1 + \log_2 n \doteq 1 + 3{,}3 \log n.$$

Pro grafické zobrazení intervalového rozdělení četností se používá **histogram**. Jsou-li všechny intervaly stejné šířky, pak je histogram sloupcový graf, kde nad každým intervalem sestrojíme obdélník, jehož výška je rovna příslušné četnosti. Histogram se někdy také normuje, aby součet obsahů všech obdélníků dal jedničku.

Jestliže jsou intervaly z nějakého důvodů různě široké, musíme při sestrojení histogramu tyto šířky vzít v potaz. DOPLNIT

Příklad 0.3. Zkoumáme průměrnou spotřebu benzinu u automobilů určité značky. Testováním 80 automobilů jsme získali následující hodnoty (v litrech na 100 km):

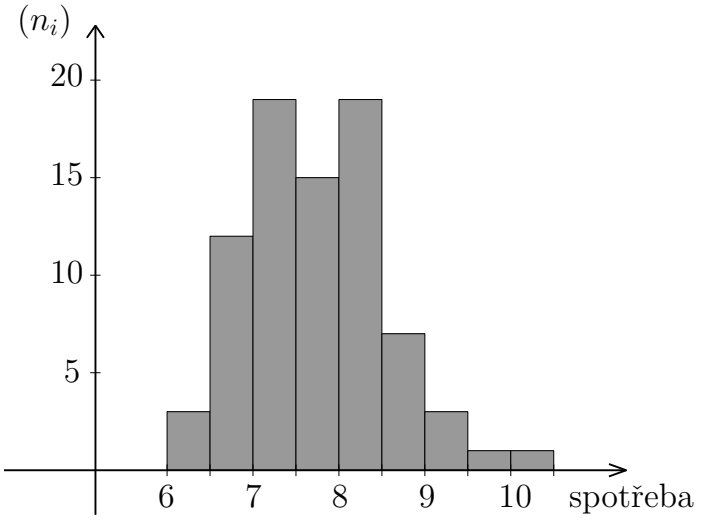
6,23	6,86	6,98	7,12	7,31	7,60	7,80	8,60	8,33	8,41	8,57	9,35
6,38	6,91	7,00	7,12	7,37	7,68	7,82	8,12	8,35	8,41	8,66	9,66
6,48	6,94	7,00	7,14	7,40	7,69	7,82	8,13	8,35	8,45	8,88	10,49
6,76	6,95	7,50	7,14	7,42	7,69	7,83	8,14	8,35	8,48	8,92	
6,79	6,95	7,80	7,23	7,46	7,71	7,88	8,22	8,35	8,48	8,95	
6,80	6,96	7,11	7,24	7,47	7,72	7,90	8,24	8,37	8,54	9,20	
6,82	6,98	7,11	7,29	7,53	7,76	7,98	8,28	8,40	8,55	9,25	

Najděte intervalové rozdělení četností a znázorněte je pomocí histogramu.

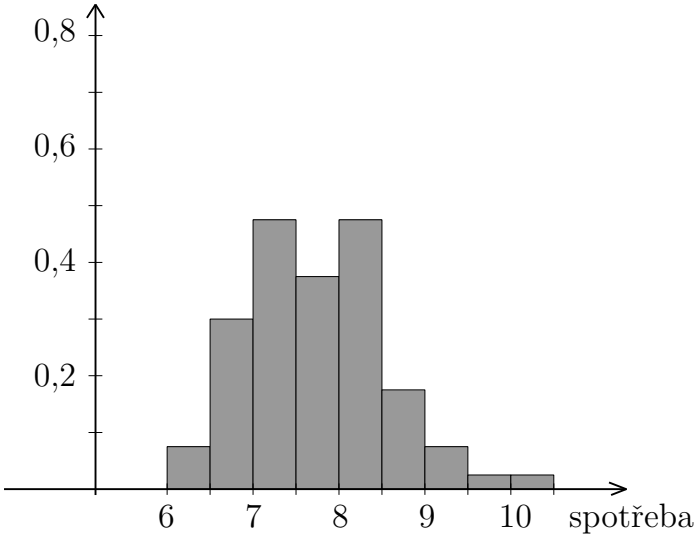
Řešení. Máme $n = 80$ hodnot v rozmezí 6,23 až 10,49. Můžeme je rozdělit např. do intervalů $\langle 6; 6{,}5)$, $\langle 6; 6{,}5)$ až $\langle 10; 10{,}5)$ (podle Sturgesova pravidla by intervalů mělo být zhruba $1 + 3{,}3 \log 80 \doteq 7$, my jich máme 9). V prvním intervalu leží 3 hodnoty, ve druhém 12, celkem tabulka intervalových četností dopadne takto:

Interval	Počet aut n_i	Relativní četnost f_i	Kumulativní absolutní četnost	Kumulativní relativní četnost
$\langle 6; 6{,}5)$	3	0,0375	3	0,0375
$\langle 6{,}5; 7)$	12	0,1500	15	0,1875
$\langle 7; 7{,}5)$	19	0,2375	34	0,4250
$\langle 7{,}5; 8)$	15	0,1875	49	0,6125
$\langle 8; 8{,}5)$	19	0,2375	68	0,8500
$\langle 8{,}5; 9)$	7	0,0875	75	0,9375
$\langle 9; 9{,}5)$	3	0,0375	78	0,9750
$\langle 9{,}5; 10)$	1	0,0125	79	0,9875
$\langle 10; 10{,}5)$	1	0,0125	80	1,0000

Na obrázku 3 vidíme příslušný histogram. Histogram na obrázku 4 vznikl normováním: vzali jsme relativní četnosti a vydělili je délkou dílčího intervalu, tj. 0,5. Výška prvního sloupce je tedy $2 \cdot 0{,}0375$ atd. □



Obr. 3: Histogram četností



Obr. 4: Normovaný histogram

0.3 Charakteristiky polohy

Charakteristiky polohy (nebo též úrovně) popisují, kolem jakých hodnot se zkoumaný znak zhruba pohybuje.

0.3.1 Aritmetický průměr

Aritmetický průměr patří mezi nejznámější a nejdůležitější charakteristiky statistického souboru.

Máme-li soubor rozsahu n a zjištěné hodnoty znaku jsou $x_1, \ldots, x_n$, pak jejich aritmetický průměr je

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Jestliže sledovaný znak x může nabývat k různých hodnot $x_1, x_2, \ldots, x_k$ a pro každou hodnotu x_i, $i = 1, \ldots, k$, známe její četnost n_i, resp. relativní četnost f_i, pak pro zjištění aritmetického průměru nemusíme všechny hodnoty sečítat. Platí totiž

$$\overline{x} = \frac{\overbrace{x_1 + \cdots + x_1}^{n_1\text{-krát}} + \overbrace{x_2 + \cdots + x_2}^{n_2\text{-krát}} + \cdots + \overbrace{x_k + \cdots + x_k}^{n_k\text{-krát}}}{n} = x_1 \cdot \frac{n_1}{n} + \cdots + x_k \cdot \frac{n_k}{n}.$$

Aritmetický průměr znaku, který nabývá hodnot $x_1, x_2, \ldots, x_k$ s četnostmi n_i a relativními četnostmi f_i, $i = 1, \ldots, k$, lze vypočítat jako

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_i = \sum_{i=1}^{k} x_i \cdot f_i. \tag{1}$$

Jestliže zkoumáme spojitý znak a známe rozložení intervalových četností, můžeme je pro výpočet aritmetického průměru využít podobně jako v případě (1). Za hodnoty znaku bereme středy intervalů. Aritmetický průměr však tímto způsobem nedostaneme úplně přesně.

Příklad 0.4. Celkem $n = 200$ studentů psalo písemku, na kterou bylo možno získat maximálně 15 bodů. V níže uvedené tabulce je úspěšnost studentů – četnosti n_i a relativní četnosti f_i jednotlivých počtů bodů. Vypočtěte průměrný počet bodů z písemky.

body	n_i	f_i	body	n_i	f_i	body	n_i	f_i	body	n_i	f_i
0	3	0,015	4	6	0,030	8	24	0,120	12	17	0,085
1	5	0,025	5	13	0,065	9	16	0,080	13	20	0,100
2	2	0,010	6	11	0,055	10	18	0,090	14	12	0,060
3	3	0,015	7	14	0,070	11	21	0,105	15	15	0,075

Řešení. Průměrný počet bodů je

$$\begin{aligned}\overline{x} &= \frac{1}{200}(0\cdot 3 + 1\cdot 5 + 2\cdot 2 + \cdots + 13\cdot 20 + 14\cdot 12 + 15\cdot 15) = \\ &= 0\cdot 0{,}015 + 1\cdot 0{,}025 + 2\cdot 0{,}010 + \cdots + 13\cdot 0{,}100 + 14\cdot 0{,}060 + 15\cdot 0{,}075 = 9{,}375\end{aligned}$$

□

Příklad 0.5. Vypočtěte průměrnou spotřebu benzínu pro hodnoty z příkladu 0.3.

Řešení. Využijeme-li intervalové rozložení četností a jako reprezentanta každého intervalu vezmeme jeho střed, dostaneme

$$\overline{x} \doteq \frac{1}{80}(6{,}25\cdot 3 + 6{,}75\cdot 12 + \cdots + 9{,}75\cdot 1 + 10{,}25\cdot 1) \doteq 7{,}74.$$

Jestliže však použijeme všechny hodnoty z tabulky a spočítáme průměr klasicky, vyjde hodnota lehce odlišná:

$$\overline{x} = \frac{1}{80}(6{,}23 + 6{,}38 + 6{,}48 + \cdots) \doteq 7{,}78.$$

□

Důležité vlastnosti aritmetického průměru

1. Jestliže ke všem hodnotám znaku přičteme konstantu $a \in \mathbb{R}$, přičte se a i k průměrné hodnotě:
$$\overline{x+a} = \overline{x} + a. \tag{2}$$
2. Jestliže každou hodnotu znaku vynásobíme konstantou $a \in \mathbb{R}$, výsledný průměr bude a-násobkem původního průměru:
$$\overline{a \cdot x} = a \cdot \overline{x}. \tag{3}$$
3. Jestliže v tomtéž statistickém souboru sledujeme dva znaky x a y, pak průměr z jejich součtu je součet průměrů:
$$\overline{x+y} = \overline{x} + \overline{y} \tag{4}$$

Právě uvedené vztahy není těžké dokázat. Jsou-li zjištěné hodnoty znaku $x_1, \ldots, x_n$, resp. $y_1, \ldots, y_n$, pak

$$\begin{aligned} \overline{x+a} &= \frac{1}{n}\left((x_1+a)+(x_2+a)+\cdots+(x_n+a)\right) = \frac{1}{n}(x_1+\cdots+x_n)+\frac{1}{n}(n\cdot a) = \overline{x}+a, \\ \overline{a\cdot x} &= \frac{1}{n}\left(a\cdot x_1+\cdots+a\cdot x_n\right) = \frac{a}{n}(x_1+\cdots+x_n) = a\cdot\overline{x}, \end{aligned}$$

$$\overline{x+y} \;=\; \frac{1}{n}\left((x_1+y_1)+(x_2+y_2)+\cdots+(x_n+y_n)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{x}+\overline{y}.$$

Ohledně (4) zdůrazněme, že musíme mít stejný počet x-ových a y-ových hodnot. Pro dva soubory čísel o různém rozsahu vztah (4) samozřejmě neplatí!

Příklad 0.6. a) Průměrný plat v určitém oddělení podniku byl 24 000 Kč. Pak dostali všichni 1 000 Kč přidáno. Nyní je průměrná mzda 25 000 Kč.

b) Studenti u profesora A. dostali na písemku průměrně 5 bodů. Profesor A. pak zjistil, že byl při hodnocení mnohem přísnější než profesor B., a rozhodl se, že každému studentovi body zvýší 1,2-krát. Nyní mají studenti průměrně 6 bodů.

c) Bylo provedeno statistické šetření mezi 3 000 domácností. Bylo zjištěno, že průměrné výdaje za bydlení jsou 5 000 Kč na měsíc a průměrné výdaje za jídlo jsou 4 500 Kč na měsíc. Kdybychom u každé domácnosti brali výdaje za bydlení a za jídlo jako jednu položku, dostali bychom průměrnou hodnotu 9 500 Kč na domácnost a měsíc.

V některých případech nám aritmetický průměr nemusí dát dobrou představu o typické úrovni hodnot souboru. Jestliže např. máme soubor, třeba i velkého rozsahu, který obsahuje několik extrémně velkých čísel, může tím být průměr značně vychýlen oproti obvyklým hodnotám.

Příklad 0.7. V jisté firmě pracuje 10 řadových pracovníků s platem 15 000 Kč, zatímco ředitel má 100 000 Kč. Průměrný plat je pak přibližně 22 727 Kč, ale zkuste to říct těm „dole“...

Proto se kromě aritmetického průměru užívají i další charakteristiky úrovně, které někdy mohou být i výstižnější. V dalších odstavcích popíšeme modus a medián.

0.3.2 Modus

> **Modus** statistického znaku označíme $\hat{x}$ a je to hodnota, která se v souboru vyskytuje **nejčastěji**.
> U spojitých znaků – známe-li intervalové rozdělení četností – stanovujeme tzv. modální (nejčetnější) interval. Za přibližnou hodnotu modu pak můžeme brát jeho střed.

Příklad 0.8. Modus statistického souboru z příkladu 0.2 (věk studentů) je $\hat{x} = 19$, modus souboru z příkladu 0.4 (výsledky písemky) je $\hat{x} = 8$ a modální intervaly souboru z příkladu 0.3 (spotřeba benzinu) jsou dva: $\langle 7; 7{,}5)$ a $\langle 8; 8{,}5)$.

0.3.3 Medián

Medián rozděluje statistický soubor na dvě stejně velké části. Občas může mít větší vypovídací hodnotu než průměr, viz příklad 0.7.

Medián statistického znaku označíme $\tilde{x}$ nebo též (v souladu s označením použitým v kapitole 0.4) $\tilde{x}_{0,5}$. Je to **prostřední hodnota** ze souboru uspořádaného podle velikosti:
Označíme-li prvky uspořádané podle velikosti jako $x_1, x_2, \ldots, x_n$ a počet prvků n je liché číslo, pak je medián přímo prostřední hodnota, tj.

$$\tilde{x} = x_{(n+1)/2}\,.$$

Je-li rozsah souboru n sudé číslo, je medián průměr ze dvou prostředních prvků, tj.

$$\tilde{x} = \frac{1}{2}\left(x_{n/2} + x_{(n/2)+1}\right).$$

Poznámka 0.9. Medián $\tilde{x}$ je tedy takové číslo, že alespoň 50 % hodnot souboru je menších nebo rovných $\tilde{x}$ a alespoň 50 % hodnot souboru je větších nebo rovných $\tilde{x}$, viz též příklad 0.10.

Příklad 0.10. Určete medián, jestliže zjištěné hodnoty zkoumaného znaku jsou

$$4, 7, 3, 5, 2, 4, 8, 6, 3, 4, 7, 2, 4, 5, 5.$$

Řešení. Setříděním podle velikosti dostaneme

$$2, 2, 3, 3, 4, 4, 4, \underline{4}, 5, 5, 5, 6, 7, 7, 8.$$

Hodnot je celkem 15, medián tedy bude osmá (prostřední) z nich, tj. $\tilde{x} = 4$.

Jestli někoho zarazilo slovo „alespoň“ v poznámce 0.9 a očekával by, že pod i nad mediánem leží přesně 50 % hodnot, pak zde můžeme na ukázku uvést, že v našem příkladu máme 8 hodnot menších nebo rovných mediánu, což je zhruba 53 %. Větších nebo rovných mediánu je dokonce 11 hodnot neboli přibližně 73 %. □

Příklad 0.11. Určete medián statistického souboru z příkladu 0.4 (výsledky písemky).

Řešení. Víme, že soubor má $n = 200$ prvků, medián tedy bude průměr ze 100. a 101. prvku. Zatím ale nevíme, jakou hodnotu 100. a 101. prvek mají. Ze zadání příkladu máme k dispozici tabulku četností. Pomocí ní nyní budeme počítat kumulativní četnosti, dokud nenarazíme na stovku:

body	četnost	kumul. četnost	body	četnost	kumul. četnost	body	četnost	kumul. četnost
0	3	3	4	6	19	8	24	81
1	5	8	5	13	32	9	16	97
2	2	10	6	11	43	10	18	115
3	3	13	7	14	57	...	...	...

Vidíme, že seřadíme-li soubor podle velikosti, pak prvních 97 prvků nabývá hodnoty menší nebo

rovné 9, zatímco prvních 115 prvků je menších nebo rovných 10. To znamená, že 100. i 101. prvek souboru je roven 10, a medián je proto

$$\tilde{x} = \frac{x_{100} + x_{101}}{2} = \frac{10 + 10}{2} = 10.$$

Příklad jsme mohli také vyřešit pomocí kumulativních relativních četností. V tomto případě bychom zkoumali, kdy bude dosaženo hodnoty 0,5. □

0.4 Kvantily

V předchozím odstavci jsme se seznámili s mediánem, který rozděluje soubor na dvě stejně početné části. Podobně můžeme zkoumat, jaké hranice soubor rozdělí na čtyři stejně početné části – pak mluvíme o tzv. kvartilech, apod. Obecně hledáme hranici, pod kterou leží určité vybrané procento hodnot celého souboru. V kapitole **??** se seznámíme s pojmem kvantil náhodné veličiny, který bude jednoznačně definován. Pro statistické soubory se však spokojíme s poněkud neurčitým popisem kvantilu:

> Pro $p \in (0, 1)$ je **kvantil** $\tilde{x}_p$ neboli p-kvantil takové číslo, které odděluje nejmenších $p \cdot 100$ % hodnot statistického znaku od největších $(1 - p) \cdot 100$ % hodnot.

Speciální případy kvantilů:

- **Medián** $\tilde{x}_{0,5}$ – dělí soubor seřazený podle velikosti zkoumaného znaku na poloviny.
- **Kvartily** $\tilde{x}_{0,25}, \tilde{x}_{0,5}, \tilde{x}_{0,75}$ – dělí soubor na čtvrtiny. Hodnotu $\tilde{x}_{0,25}$ nazýváme první kvartil, druhý kvartil splývá s mediánem a hodnotu $\tilde{x}_{0,75}$ nazýváme třetí kvartil.
- **Decily** $\tilde{x}_{0,1}, \ldots, \tilde{x}_{0,9}$ – dělí soubor na desetiny. Mluvíme o prvním, druhém, až devátém decilu.
- **Percentily** $\tilde{x}_{0,01}, \ldots, \tilde{x}_{0,99}$ – dělí soubor na setiny.

Zbývá ještě popsat, jak se kvantil pro konkrétní data najde. Musíme si uvědomit, že definicí v rámečku kvantil bohužel není dán jednoznačně. Různé statistické softwary také pro nalezení kvantilů používají různé algoritmy, které dávají rozdílné výsledky. Popíšeme zde jeden z možných postupů. Předpokládejme, že soubor už je seřazený podle velikosti zkoumaného znaku, jehož hodnoty jsou $x_1 \leq x_2 \leq \cdots \leq x_n$. Nejprve vypočítáme pořadové číslo prvku, který odděluje nejmenších $p \cdot 100\%$ hodnot. To můžeme udělat např. jako

$$k = (n+1)p \quad \text{nebo} \quad k = 1 + (n-1)p.$$

Všimněte, že pro medián, tj. 0,5-kvantil, vyjde v obou případech $k = (n+1)/2$. Je-li takto nalezené k celé číslo, je $\tilde{x}_p = x_k$ a kvantil jsme našli – v případě mediánu se tohle stane pro n liché. Často však k celé číslo

není. Např. pro medián a $n = 4$ nám vyjde $k = 2{,}5$. Příslušný „dva-a-půltý“ prvek určíme jako průměr prvku druhého a třetího, což můžeme též zapsat jako $\tilde{x}_{0,5} = x_2 + 0{,}5 \cdot (x_3 - x_2)$. Obecně, jestliže k leží v intervalu $\langle m, m+1)$, kde m je celé číslo, pak za hodnotu p-kvantilu můžeme brát

$$\tilde{x}_p = x_m + (k - m)(x_{m+1} - x_m).$$

Právě popsaný postup používají některé specializované statistické softwary nebo také MS Excel. Např. v Matlabu se kvantily hledají ještě jiným způsobem, zájemci si jej mohou najít v příslušném hesle nápovědy.

Příklad 0.12. Najděte první a třetí kvartil, jestliže zjištěné hodnoty zkoumaného znaku jsou

$$3, 4, 5, 6, 6, 7, 7, 8.$$

Řešení. Máme $n = 8$ hodnot. Pořadové číslo prvního kvartilu bude $k = 9 \cdot 0{,}25 = 2{,}25$. První kvartil je proto

$$\tilde{x}_{0,25} = x_2 + 0{,}25 \cdot (x_3 - x_2) = 4 + 0{,}25 \cdot (5 - 4) = 4{,}25.$$

Použijeme-li druhý uvedený způsob výpočtu k, dostaneme $\tilde{x}_{0,25} = 4{,}75$, zatímco Matlab dává výsledek $\tilde{x}_{0,25} = 4{,}5$.
Pokud jde o třetí kvartil, tak zde v každém případě vyjde $\tilde{x}_{0,75} = 7$. □

0.5 Charakteristiky variability

Charakteristiky variability popisují rozptýlenost hodnot. Zajímá nás, jestli se znak pohybuje nejčastěji jen v určitém nevelkém intervalu, nebo zda je jeho rozpětí široké. Nejčastěji zkoumáme, jak moc jsou hodnoty znaku rozptýlené kolem aritmetického průměru, existují však i jiné charakteristiky variability.

0.5.1 Nejjednodušší míry variability

Nejjednodušší, ale i nejhrubší mírou variability je variační rozpětí.

Variační rozpětí je rozdíl největší a nejmenší hodnoty znaku:

$$R = x_{\max} - x_{\min}.$$

Variační rozpětí vypočítáme velmi snadno, ovšem jeho nevýhodou je to, že extrémní hodnoty mohou být nahodilé a je možné, že naprostá většina hodnot znaku leží v intervalu daleko užším.

Příklad 0.13. Variační rozpětí hodnot z příkladu 0.3 (spotřeba benzínu) je $10{,}49 - 6{,}23 = 4{,}26$.

Další charakteristikou variability je tzv. mezikvartilové rozpětí. To udává, v jak širokém intervalu leží „prostřední“ polovina všech hodnot.

Mezikvartilové rozpětí je rozdíl třetího a prvního kvartilu:

$$\tilde{x}_{0,75} - \tilde{x}_{0,25}.$$

0.5.2 Rozptyl a směrodatná odchylka

Ve většině případů však dává statistická teorie i praxe přednost takovým mírám variability, jejichž velikost je závislá na všech hodnotách statistického souboru. Zajímavé také je, jak moc jsou hodnoty „nahuštěné“ kolem aritmetického průměru.

Třeba vás napadá, že by nebylo marné zkoumat průměr z hodnot $(x_i - \overline{x})$, $i = 1, \ldots, n$. Tudy však cesta bohužel nevede, protože výsledek je vždy roven nule:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\overline{x} = \overline{x} - \frac{1}{n}\cdot n \cdot \overline{x} = 0.$$

Dalším kandidátem na rozptyl je průměrná absolutní odchylka $\frac{1}{n}\sum |x_i - \overline{x}|$. Ta už je nenulová (samozřejmě kromě případu, že všechny hodnoty x_i jsou stejné) a jakousi informaci o variabilitě sděluje. Problém je však v tom,

že součet absolutních hodnot je obtížně matematicky zpracovatelný (např. obtížně se derivuje, apod.). Proto se nejčastěji používá průměrná kvadratická odchylka, tzv. rozptyl.

Rozptyl statistického znaku označíme σ^2 a definujeme jej jako

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{5}$$

Často užívanou veličinou je též tzv. **výběrový rozptyl** s^2, který používáme v případě, že máme k dispozici pouze výběrový soubor. Je definován jako

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{6}$$

Rozptyl nám tedy udává, jak moc se hodnoty statistického znaku průměrně liší od průměrné hodnoty, ovšem ve druhé mocnině. Výsledek je proto ve čtvercích použité měrné jednotky, což ztěžuje jeho interpretaci. Abychom se dostali zpátky na původní jednotky, rozptyl odmocníme, čímž získáme tzv. směrodatnou odchylku:

Směrodatná odchylka σ je odmocnina z rozptylu:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Není třeba se obávat, že bychom dostali odmocninu ze záporného čísla, protože rozptyl jako průměrná hodnota z druhých mocnin záporně vyjít nemůže.

Prakticky se pro výpočet rozptylu používá o něco jednodušší vzorec, než je (5).
Jeho odvození není složité. Nejprve roznásobíme $(x_i - \overline{x})^2$a sumu rozdělíme na dílčí tři sumy:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}\cdot x_i + \overline{x}^2) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\sum_{i=1}^{n}\overline{x}\cdot x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2$$

Z druhé sumy vytkneme průměr $\overline{x}$. Poslední suma je rovna $n\overline{x}$ (sčítáme n-krát tutéž hodnotu $\overline{x}$. Celkem máme

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \overline{x}^2\frac{n}{n} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 + \overline{x}^2.$$

Tím se dostáváme k finálnímu vzorci:

Rozptyl lze vypočítat jako rozdíl průměru z druhých mocnin x_i a druhé mocniny průměru,

$$\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2. \tag{7}$$

Podobně jako u průměru můžeme při výpočtu rozptylu použít četnosti jednotlivých variant znaku

Rozptyl znaku, který nabývá hodnot $x_1, x_2, \ldots, x_k$ s četnostmi n_i a relativními četnostmi f_i, $i = 1, \ldots, k$, lze vypočítat jako

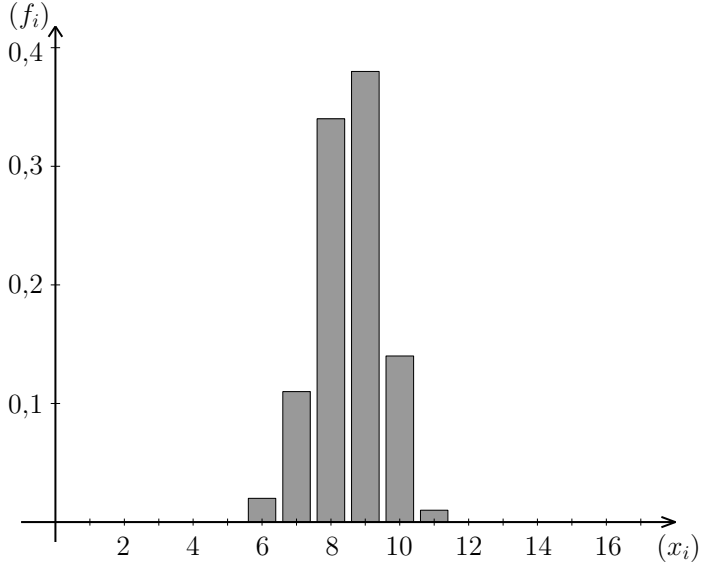
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i = \left(\frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_i\right) - \overline{x}^2,$$

případně jako

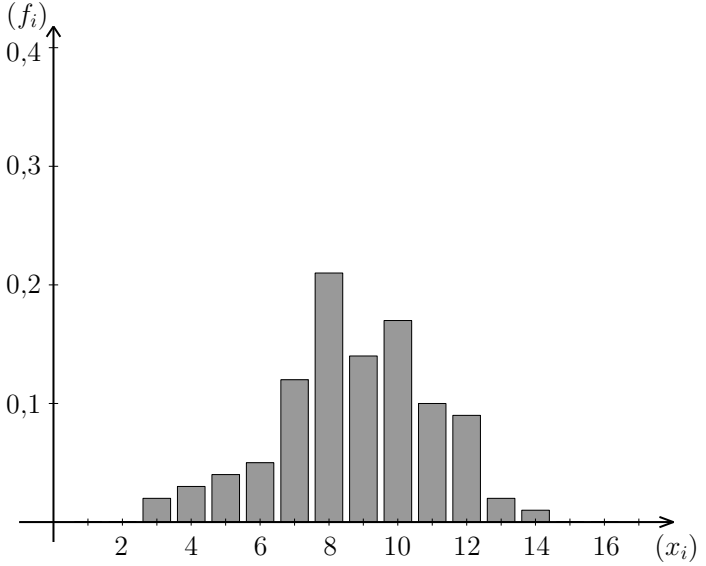
$$\sigma^2 = \sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot f_i = \left(\sum_{i=1}^{k} x_i^2 \cdot f_i\right) - \overline{x}^2. \tag{8}$$

Obrázky 5 a 6 ilustrují význam rozptylu. Zkoumali jsme dva statistické znaky se stejným průměrem,

ovšem druhý z nich má větší rozptyl než první. Vidíme, že rozdělení relativních četností druhého znaku je širší a plošší než u znaku prvního.



Obr. 5: Relativní četnosti pro znak s průměrem $\overline{x} \doteq 9$ a rozptylem $\sigma^2 \doteq 1$



Obr. 6: Relativní četnosti pro znak s průměrem $\overline{x} \doteq 9$ a rozptylem $\sigma^2 \doteq 7$

Příklad 0.14. Určete průměr, rozptyl a směrodatnou odchylku znaku, který nabývá hodnot

$$3, 4, 5, 6, 6, 7, 7, 8.$$

Řešení. Rozsah souboru je $n = 8$ a průměr je $\overline{x} = 5{,}75$. Rozptyl můžeme spočítat např. podle (7). Průměr z druhých mocnin je

$$\frac{1}{8}\sum_{i=1}^{8} x_i^2 = \frac{1}{8}\left(3^2 + 4^2 + \cdots + 8^2\right) = 35{,}5,$$

takže rozptyl je

$$\sigma^2 = 35{,}5 - 5{,}75^2 = 2{,}4375.$$

Směrodatná odchylka je pak $\sigma = \sqrt{2{,}4375} \doteq 1{,}5612$.

Rozptyl jsme samozřejmě mohli počítat i přímo z definice (5):

$$\sigma^2 = \frac{1}{8}\left((3 - 5{,}75)^2 + (4 - 5{,}75)^2 + \cdots + (8 - 5{,}75)^2\right) = 2{,}4375.$$

Výpočet by však byl o něco zdlouhavější. Vzorec (7) používáme právě kvůli tomu, že je méně náročný, i když při pohledu na něj nevyniká podstata věci – že vlastně zjišťujeme, jak moc se hodnoty liší od průměru. □

Důležité vlastnosti rozptylu

1. Jestliže ke všem hodnotám znaku přičteme konstantu $a \in \mathbb{R}$, rozptyl se nezmění:

$$\sigma^2_{x+a} = \sigma^2_x. \tag{9}$$

2. Jestliže každou hodnotu znaku vynásobíme konstantou $a \in \mathbb{R}$, výsledný rozptyl se změní a^2-krát:

$$\sigma^2_{a\cdot x} = a^2 \cdot \sigma^2_x. \tag{10}$$

Důkaz vztahu (9) nebudeme rozepisovat. Spíše si uvědomme podstatu věci: jestliže ke všem x_i přičteme konstantu, hodnoty se posunou po číselné ose, posune se i jejich průměr, ale rozptýlenost kolem průměru zůstane stejná.

Vztah (10) lze dokázat podobně jako (3) s tím, že už víme, že $\overline{a \cdot x} = a \cdot \overline{x}$:

$$\sigma^2_{a\cdot x} = \frac{1}{n}\left((ax_1 - a\overline{x})^2 + \cdots + (ax_n - a\overline{x})^2\right) = \frac{a^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2 \cdot \sigma^2_x.$$

0.6 Statistické testy

Občas někde čteme nebo slyšíme formulaci „Je statisticky dokázáno, že . . . “ Nyní se dozvíme, jak se to dělá – jakým způsobem se např. ověřuje, jestli inovace nějakého výrobního procesu skutečně přináší zlepšení, jestli procento lidí s určitou vlastností je v jedné populaci větší než ve druhé, apod.

0.6.1 Základní principy statistického testu

Příklad 0.15. **Soudní proces jako příklad rozhodovacího procesu.** Uvažujme jednoduchý soudní proces, ve kterém existuje pouze jediný možný trest a soud rozhodne, zda se tomuto trestu obžalovaný podrobí nebo ne. A navíc proti rozhodnutí soudu neexistuje žádné odvolání. Jedná se o jakýsi rozhodovací proces, u kterého mohou nastat čtyři možné výsledky:

1. Obžalovaný je vinen a soud jej odsoudí.
2. Obžalovaný je nevinen a soud jej osvobodí.
3. Obžalovaný je nevinen a soud jej odsoudí. Jedná se o chybné rozhodnutí - tuto chybu budeme označovat jako chybu prvního druhu.
4. Obžalovaný je vinen a soud jej osvobodí. Toto rozhodnutí je rovněž chybné - budeme tuto chybu označovat chybou druhého druhu.

V každém soudním procesu se musí hledat jistá rovnováha mezi tvrdostí a mírností. Jedním extrémem je benevolentní soudce, který k usvědčení obžalovaného vyžaduje velké množství důkazů. Takový soudce jen zřídka odsoudí nevinného (zřídka se dopustí chyby prvního druhu), ale dosti často osvobodí viníka (chyba druhého druhu). Druhým extrémem je přísný soudce, kterému k usvědčení stačí jen několik důkazů. Takový soudce posílá do vězení i jen při stínu podezření, čili častěji odsoudí nevinného (chyba prvního druhu), ale zřídka osvobodí darebáka (= zřídka se dopustí chyby druhého druhu).
Je otázkou, která z chyb je závažnější - zda chyba prvního druhu, nebo chyba druhého druhu. Všeobecně se má za to, že závažnější je uvěznit nevinného, než osvobodit darebáka. A proto se chybě odsouzení nevinného přisuzuje druh číslo 1 a věnuje se jí větší pozornost. Ale někde musí být stanovena jistá hranice, po jejímž překročení už soud přistoupí k rozhodnutí „vinen“ a bez skrupulí člověka potrestá.

Všimněme si jedné věci, která platí jako obecný princip. Pokud se soudce snaží být mírný a odsoudí člověka až po nahromadění velkého množství důkazů (snižuje tím možnost výskytu chyby prvního druhu), současně narůstá nebezpečí, že i když je obžalovaný vinen, potřebné množství důkazů se nenajde a soud jej osvobodí (roste možnost výskytu chyby druhého druhu). Tj. snižováním možnosti výskytu chyby prvního druhu roste možnost výskytu chyby druhého druhu – a naopak: pokud zvyšujeme možnost výskytu chyby prvního druhu, snižuje se možnost výskytu chyby druhého druhu. Je vidět, že žádnou z chyb není možné naprosto vyrušit: pokud totiž snižujeme možnost výskytu chyby prvního druhu až téměř na nulu, roste tím možnost výskytu chyby druhého druhu do obludných rozměrů a rozhodnutí učiněná tímto stylem jsou nerozumná, až nemoudrá. Strategií v rozhodovacích procesech tohoto typu je tedy zvolit

pravděpodobnost výskytu chyby prvního druhu malou, ale ne příliš malou.

Přejděme nyní ke konkrétnějšímu, i když možná méně vzletnému příkladu:

Příklad 0.16. Dva bratři, Vašek a Ondra, se pořád hádali, který z nich vynese odpadky, až jim otec nařídil, aby si vždycky hodili korunou. Dokonce jim na to vyhradil jednu starou, už neplatnou minci. Když padne líc, je na řadě Vašek. Když rub, tak Ondra. Vaškovi se zdá, že líc padá podezřele často a že ta mince je nějaká divná. Chtěl by to dokázat.

Řešení. Řešení se celkem nabízí: Vašek mincí mnohokrát hodí a bude pozorovat, jak často na ní padá líc a jak často rub. Dejme tomu, že hodil padesátkrát a líc padl třicetkrát. Přesvědčí nás to, že na minci padá líc častěji? Necháme teď Vaška házet a podíváme se na matematickou stránku věci.

Označíme X náhodnou veličinu udávající počet líců při 50 hodech mincí. Tato náhodná veličina má binomické rozdělení s parametry $n = 50$ a $p = 0{,}5$ – to ovšem v případě, že mince je vyvážená a líc na ní padá stejně často jako rub. Budeme pro tuto chvíli předpokládat, že mince vyvážená je, tj. že opravdu $p = 0{,}5$. Za tohoto předpokladu najdeme hranici, nad kterou se počet líců dostane jen s velmi malou pravděpodobností. Jestliže bude experimentem získaných 30 líců nad touto hranicí, stalo se něco nečekaného a učiníme závěr, že na minci líc padá opravdu podezřele často a že parametr p bude větší než 0,5. Bude-li 30 pod nalezenou hranicí, řekneme, že výsledek není průkazný a že 30 líců z 50 hodů je u vyvážené mince ještě v očekávaných mezích. Hraniční pravděpodobnost není nijak „shůry dána“, tu si volíme a v praxi se většinou volí 5%. Budeme tedy hledat k, pro které je

$$P(X > k) = 0{,}05 \quad \text{neboli} \quad P(X \leq k) = 0{,}95.$$

Připomeňme, že této hodnotě se říká 0,95-kvantil. (Protože X má diskrétní rozdělení, hledáme spíš nejmenší hodnotu k, pro kterou bude pravděpodobnost 0,95 překročena, protože přesně 0,95 nám pro žádné k vyjít nemusí.)

Kdybychom pracovali přímo s binomickým rozdělením, bylo by nalezení kvantilu při ručním výpočtu velmi pracné až nemožné. Naštěstí si můžeme práci zjednodušit pomocí normálního rozdělení. Střední hodnota a rozptyl náhodné veličiny X jsou

$$EX = \mu = 50 \cdot 0{,}5 = 25, \quad DX = \sigma^2 = 50 \cdot 0{,}5 \cdot (1 - 0{,}5) = 12{,}5.$$

Náhodná veličina X má proto přibližně normální rozdělení $No(25; 12{,}5)$, a tedy

$$U \doteq \frac{X - 25}{\sqrt{12{,}5}}\,.$$

Při hledání meze k budeme postupovat podobně jako v příkladu **??** d). Najdeme 0,95-kvantil náhodné veličiny U a pak jej zpětně transformujeme:

$$P(X \leq k) \doteq P(U \leq u) = \Phi(u) = 0{,}95 \quad \Rightarrow \quad u \doteq 1{,}65$$

$$\frac{k - 25}{\sqrt{1{,}25}} \doteq 1{,}65 \quad \Rightarrow \quad k \doteq 31.$$

Hledaná hranice, která bude překročena pouze s pravděpodobností 0,05, je tedy 31 líců (z 50 hodů). Protože Vaškovi padl líc 30-krát a 30 leží pod nalezenou mezí, nevyváženost mince se těsně, ale přece nepotvrdila.

Kdyby Vaškovi padl líc třeba 35-krát, už by to bylo opravdu podezřelé a závěr by byl, že mince vyvážená není. □

Nyní postup předvedený v předchozím příkladu ještě jednou projdeme s použitím odborné terminologie. Testovali jsme tzv. nulovou hypotézu $H_0\colon p = 0{,}5$ (líc padá stejně často jako rub), proti alternativní hypotéze $H_1\colon p > 0{,}5$ (líc padá častěji než rub). Rozhodovali jsme podle hodnoty testového kritéria – počtu líců při 50 hodech. Předpokládali jsme, že platí H_0, a za tohoto předpokladu jsme našli kritický obor, do kterého testové kritérium padne jen s velmi malou pravděpodobností, za tuto pravděpodobnost jsme zvolili $\alpha = 0{,}05$. Pro náš příklad vyšel kritický obor $\langle 31, 50\rangle$. Protože hodnota testového kritéria získaná experimentem, tj. 30, do kritického oboru nepatřila, hypotéza H_1 testem nebyla prokázána.

Obecně se testování provádí v těchto krocích:

1. Vyslovíme **nulovou hypotézu** H_0 a **alternativní hypotézu** H_1.
2. Stanovíme **testové kritérium** – náhodnou veličinu T, podle které chceme o platnosti nulové hypotézy H_0 rozhodnout.
3. Předpokládáme, že platí H_0, a najdeme **kritický obor** W, do kterého testové kritérium T padne jen se zvolenou malou pravděpodobností α. Hodnotu α nazýváme **hladinou významnosti testu**. Kritický obor W je tedy stanoven tak, aby

$$P(T \in W | \text{platí } H_0) = \alpha,$$

   a jeho hranici (hranice) tvoří odpovídající kvantil (kvantily) náhodné veličiny T.
4. Zjistíme hodnotu testového kritéria (zpracujeme výsledek konkrétního pokusu či měření).
5. Jestliže empirická (tj. pokusem získaná) hodnota kritéria leží v kritickém oboru, zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_1 – hypotéza H_1 byla prokázána. Pokud naměřená hodnota v kritickém oboru neleží, hypotézu H_0 nezamítáme a hypotéza H_1 se neprokázala.

Nulová hypotéza by měla jednoznačně určovat rozdělení zkoumaného znaku. Zpravidla bývá tvaru

$\theta = \theta_0$, kde θ je určitý parametr, např. $\mu, p, \sigma^2, \ldots$, a θ_0 je konkrétní hodnota. H_0 tedy může být např. $p = 0{,}5$, $\mu = 4$, apod. Nemůže být např. tvaru $p < 0{,}5$, protože pak bychom nemohli přesně najít kritický obor. Naproti tomu alternativní hypotéza často popisuje to, co se snažíme testem prokázat, a bývá ve tvaru nerovnosti, případně tvaru $\theta \neq \theta_0$, např. $p > 0{,}5$, $\mu \neq 4$, apod. Podle toho, jaké hodnoty T svědčí ve prospěch alternativy H_1 (nízké, vysoké, případně obojí), rozlišujeme testy jednostranné a oboustranné.

Jestliže testujeme nulovou hypotézu $H_0\colon \theta = \theta_0$ proti alternativní hypotéze

$$H_1\colon \theta \neq \theta_0,$$

provádíme **oboustranný test**. Kritický obor je pak tvaru (viz obrázek 7)

$$W = \left(T_{\min}, T_{\alpha/2}\right\rangle \cup \left\langle T_{1-\alpha/2}, T_{\max}\right) .$$

Jestliže je alternativní hypotéza tvaru

$$H_1\colon \theta > \theta_0,$$

provádíme **jednostranný, a to pravostranný test**. Kritický obor je pak tvaru (viz obrázek 8)

$$W = \left\langle T_{1-\alpha}, T_{\max}\right) .$$

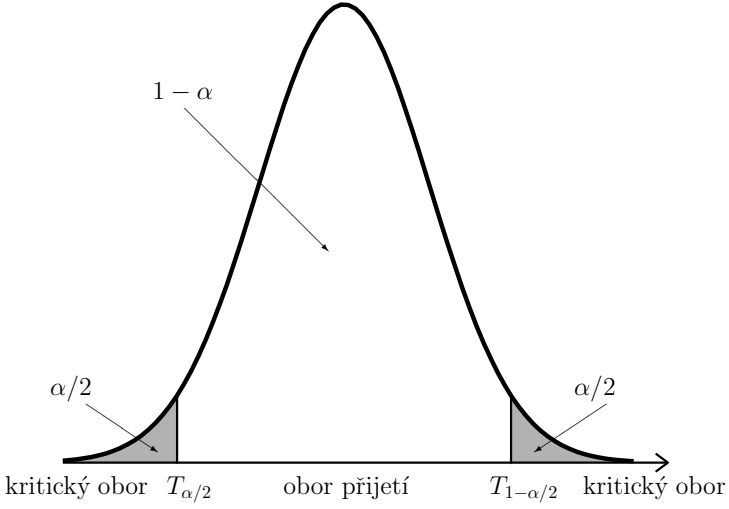
Jestliže je alternativní hypotéza tvaru

$$H_1\colon \theta < \theta_0,$$

provádíme **jednostranný, a to levostranný test**. Kritický obor je pak tvaru (viz obrázek 9)

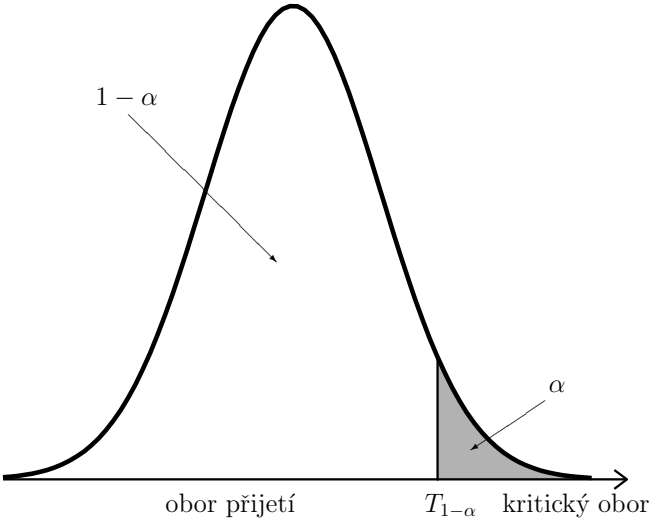
$$W = \left(T_{\min}, T_{\alpha}\right\rangle .$$

V příkladu 0.16 jsme tedy prováděli pravostranný test.

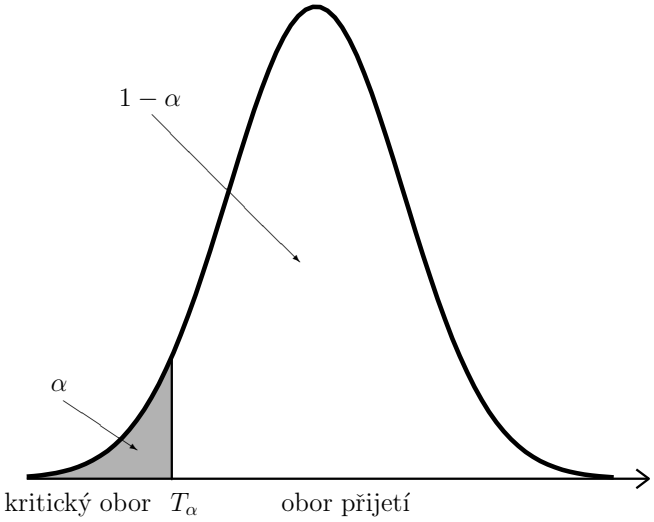


Obr. 7: Oboustranný test

V příkladu 0.15 jsme popsali možná špatná rozhodnutí při testování. Řeč byla o chybě 1. a 2. druhu. Jak tedy může testování dopadnout? Máme čtyři možnosti:



Obr. 8: Pravostranný test



Obr. 9: Levostranný test

	skutečnost: H_0 platí	skutečnost: H_1 platí
rozhodnutí: H_0 nezamítáme	správně	chyba 2.druhu
rozhodnutí: H_0 zamítáme	chyba 1.druhu	správně

Chyba 1. druhu nastane, jestliže nulová hypotéza H_0 platí, ale my ji zamítneme. Pravděpodobnost chyby 1. druhu je rovna hladině významnosti testu α,

$$P(H_0 \text{ zamítneme} | H_0 \text{ platí}) = \alpha.$$

Chyba 2. druhu nastane, jestliže nulová hypotéza H_0 neplatí (čili platí H_1), a přitom H_0 není zamítnuta. Pravděpodobnost chyby 2. druhu označíme β. S chybou 2. druhu souvisí tzv. **síla testu**. Je to pravděpodobnost, že správně zamítneme H_0, když platí alternativní hypotéze H_1,

$$\text{síla jednostranného testu} = P(H_0 \text{ zamítneme} | H_0 \text{ neplatí}) = 1 - \beta.$$

Síla testu je pozitivní pojem – čím je síla testu větší, tím je tento test vhodnější k nalezení závislosti mezi danými proměnnými. Ovšem sílu testu většinou neznáme, protože pravděpodobnost β často nedokážeme určit – k tomu bychom museli znát rozdělení testového kritéria za předpokladu, že platí alternativní hypotéza H_1. Se silou testu souvisí i následující věc: pokud naměřená hodnota kritéria nepřekročí teoretické kritické hodnoty, říkáme, že „hypotézu H_0 nezamítáme“, nikoliv „hypotézu H_0 přijímáme“. Pokud totiž náš použitý statistický test měl malou sílu, mohlo se stát, že ačkoliv závislost mezi veličinami nenalezl, ona ve skutečnosti existuje a H_0 neplatí. Z tohoto důvodu se používá tato „opatrná“ terminologie.

Další obrat jsme už také použili: pokud zamítáme H_0, někdy se říká, že výsledek testu je statisticky významný (resp. závislost mezi studovanými veličinami je statisticky významná, nebo vliv jedné veličiny na druhou je významný).

0.6.2 Statistický test střední hodnoty průměru měření při známém rozptylu

Máme-li n nezávislých náhodných veličin X_i se stejným rozdělením (tím pádem se stejnou střední hodnotou μ a rozptylem σ^2), pak průměr z těchto náhodných veličin $\overline{X}$ má buď přímo normální rozdělení (v případě, že $X_i \sim No(\mu, \sigma^2)$), nebo (pro velké n) se jeho rozdělení k normálnímu blíží. Nyní tento fakt využijeme k statistickým testům.

Test „μ =konst“

Příklad 0.17. V rámci testování výsledků našeho školství se před nějakou dobou ustanovilo, že všichni žáci posledního ročníku základních škol v České republice píší srovnávací test z matematiky. Je známo (z výsledků v předchozích letech), že ohodnocení testu má normální rozdělení se střední hodnotou $\mu = 500$ bodů a směrodatnou odchylkou $\sigma = 100$ bodů (jedná se o teoretické rozdělení celé populace žáků).

Jako součást projektu dotovaného Evropskou unií vyvinuli akademičtí pracovníci program INTEL, jehož cílem je zlepšit znalosti matematiky žactva, zejména pak zlepšit výsledky souhrnného testu.

Chtějí svůj program INTEL otestovat, a proto náhodně vybrali 25 žáků z ČR a program zaslali každému z nich. Po provedení testu z matematiky se ukázalo, že průměr ohodnocení daných 25 žáků je $\overline{x} = 540$.

Otázka zní: lze nyní říct, že program INTEL zlepšuje výkon v testu, nebo se jen náhodou vybralo 25 studentů s vyšším výkonnostním průměrem v matematice? Jedná se o „skutečný“ výsledek (= lze jej zobecnit pro celou populaci), nebo bylo vyššího průměru dosaženo jen díky náhodným faktorům? Tyto otázky nás přivádějí ke statistickému testu, který rozhodne. Jako hladinu významnosti testu zvolíme opět $\alpha = 0{,}05$.

Řešení. 1. H_0: $\mu = 500$ (Program INTEL nemá vliv na zlepšení matematických schopností, tj. střední hodnota bodového ohodnocení testu celé populace studentů i po rozšíření programu všem (celé populaci) zůstane stejná.)

H_1: $\mu > 500$ (Jednostranný test - můžeme předpokládat, že program znalosti matematiky nezhoršuje.)

2. Kritériem volíme právě veličinu $\overline{X}$, která popisuje průměr hodnot 25 náhodně vybraných žáků.

3. Za předpokladu platnosti H_0 má veličina $\overline{X}$ parametry

$$\overline{X} \sim No(\mu_{\overline{X}}, \sigma^2_{\overline{X}}), \quad \mu_{\overline{X}} = 500, \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = 400 \Longrightarrow \sigma_{\overline{X}} = 20.$$

Stanovená kritická U-hodnota je pro $\alpha = 0{,}05$ stejná jako u testu z příkladu 0.16 $u_{0{,}95} = 1{,}65$. Odtud kritická hodnota v rozměru veličiny $\overline{X}$ je

$$\overline{X}_k = \mu_{\overline{X}} + \sigma_{\overline{X}} \cdot 1{,}65 = 533,$$

kritický obor je tedy

$$W = \langle 533, \text{plný počet bodů z testu} \rangle$$

4. Hodnota získaná pokusem je 540 bodů, což v kritickém oboru leží.
5. Rozhodnutí testu: Protože $540 > 533$, zamítáme H_0 a uzavíráme, že program „skutečně“ (resp. na hladině významnosti $\alpha = 0{,}05$) zlepšuje matematické schopnosti studentů.

□

Test „$\mu_1 = \mu_2$“

Příklad 0.18. Vraťme se k situaci z příkladu 0.17. Komerční softwarová firma rovněž vyvinula program pro výuku matematiky (s názvem KILL) a chce jej školám prodávat. Ředitel školy zvažuje, zda program koupit, nebo zda využívat program INTEL. Chtěl by proto zjistit, který z obou konkurenčních programů INTEL a KILL je lepší, tj. který více zvyšuje úroveň matematických znalostí.

Získal testovací verzi programu KILL a předal ji 32 náhodně vybraným studentům. Jiných 32 náhodně vybraných studentů mělo pracovat s programem INTEL. Po provedení testu z matematiky získal od těchto 64 studentů výsledky jejich ohodnocení a spočetl průměry příslušných hodnot. U programu INTEL $\overline{x}_1 = 600$, u programu KILL $\overline{x}_2 = 533$ (v obou případech velikost vzorku $n = 32$).

Aby zjistil, do jaké míry je jeho měření reprezentativní a zda rozdíl průměrů není pouze náhodný (tj. způsobený např. tím, že program INTEL byl rozdán mezi studenty, kteří byli náhodou chytřejší, ale ne tím, že by INTEL byl lepší než KILL), sáhne ke statistickému testu, opět na hladině významnosti $\alpha = 0{,}05$.

Řešení. 1. H_0: $\mu_1 = \mu_2$ (kdyby se oba programy distribuovaly celé populaci, výsledná střední hodnota ohodnocení by byla u obou stejná).

H_1: $\mu_1 \neq \mu_2$ (musíme použít oboustranný test, protože nevíme, který z programů je lepší).

2. Testovým kritériem bude rozdíl náhodných veličin $\overline{X}_1 - \overline{X}_2$ s konkrétní naměřenou hodnotou $x_1 - x_2 = = 600 - 533 = 67$.

3. Za předpokladu platnosti H_0 je rozdělení kritéria $\overline{X_1} - \overline{X_2}$ normální (je to důsledek věty **??**) se střední hodnotou a rozptylem

$$E(\overline{X_1} - \overline{X_2}) = E\overline{X_1} - E\overline{X_2} = \mu_1 - \mu_2 = 0,$$

$$D(\overline{X_1} - \overline{X_2}) = D\overline{X_1} + D\overline{X_2} = \frac{10000}{32} + \frac{10000}{32} = 625$$

Pokud $\sigma^2_{\overline{X_1}-\overline{X_2}} = 625$, tak $\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{625} = 25$. Pro náš příklad není nutné, aby obě vyšetřované skupiny měly stejný počet studentů - jiný počet studentů v každé skupině by se projevil pouze na tom, že v posledním řádku odvození by v obou jmenovatelích nebylo číslo 32, ale číslo vyjadřující velikost dané skupiny.

Protože tentokrát provádíme oboustranný test, musíme najít 0,025-kvantil a 0,975-kvantil. Pro U-rozdělení jsou příslušné hodnoty $u \doteq \pm 1{,}96$, pro náhodnou veličinu $\overline{X}_1 - \overline{X}_2$ to bude

$$x_1 \doteq -1{,}96 \cdot 25 + 0 = -49, \quad x_2 \doteq 1{,}96 \cdot 25 + 0 = 49.$$

4. Pokusem zjištěná hodnota rozdílu průměrů je 67, což leží v kritickém oboru.

5. V našem případě tedy hypotézu H_0 zamítáme, program INTEL je lepší než program KILL.

□

Testy uvedené v této kapitole jsou příkladem prvních „praktických“ statistických testů, které jsou užívány. Naměříme hodnotu jedné veličiny u jedné skupiny pozorování, popřípadě u dvou, vypočteme průměr měření v každé ze skupin a tento průměr podrobíme jednostrannému nebo oboustrannému statistickému testu.

Ovšem přitom v těchto testech tiše předpokládáme, že rozptyl σ^2 celé populace je známý. To ale většinou není pravda a my jej musíme odhadnout (přibližně určit) z naměřených hodnot. Díky větší míře nejasnosti pak kritérium analogického statistického testu, který nepoužívá přímo σ^2, ale jeho odhad s^2 , nelze popsat normálním rozdělením, ale tzv. t**-rozdělením** - příslušný statistický test je v literatuře nazýván t**-test**.