Základy popisné statistiky

V této kapitole se seznámíme se základy popisné statistiky, představíme si základní pojmy a budeme si je ilustrovat na praktických příkladech. Kapitola je psána formou volného textu, přesnou matematickou formulaci je možno nalézt např. v [1].

Zadání následujícího příkladu bude sloužit ilustraci představených pojmů a budeme se na něj v dalším odkazovat.

Ilustrační příklad. V chemické laboratoři byl zjišťován obsah alkoholu ve 30 různých vzorcích vín dodaných různými producenty vína. Výsledky obsahu alkoholu v procentech byly následující

$13,20;\ 13,16;\ 14,37;\ 13,24;\ 14,20;\ 14,39;\ 14,06;\ 14,83;\ 13,86;\ 14,10;$
$14,12;\ 13,75;\ 14,75;\ 14,38;\ 13,63;\ 14,30;\ 13,83;\ 14,19;\ 13,64;\ 14,06;$
$12,93;\ 13,71;\ 12,85;\ 13,50;\ 13,05;\ 13,39;\ 13,30;\ 13,87;\ 14,02;\ 13,73.$

Co je to statistika? Statistika je věda o získávání, zpracování a interpretaci informace obsažené v empirických pozorováních skutečného světa (např. v naměřených datech, průzkumech...) Jinak lze říci, že statistika je věda o zkoumání reality na základě napozorovaných dat.

Statistiku dělíme na popisnou a induktivní. Popisná neboli deskriptivní statistika se zabývá popisem konkrétních dat, kdy několika čísly a obrázky stručně vystihneme to důležité. Závěry můžeme vyvozovat pouze o daných

datech, nelze je zobecňovat. Induktivní neboli konfirmatorní statistika umožňuje na základě dat odpovídat na obecné otázky o populaci a získané závěry lze zobecnit.

Statistika má aplikace v přírodních vědách v biologii, chemii, fyzice, meteorologii, medicíně, genetice, farmakologii atd. Uplatňuje se také v ekonomii v makro a mikroekonomii, v bankovnictví a pojišťovnictví. Dále v technických vědách jako je telekomunikace, doprava, počítače, strojírenství, kontrola jakosti, řízení a organizace výroby a dalších. V neposlední řadě také ve společenských vědách, především v sociologii, behaviorálních vědách, archeologii, lingvistice, antropologii...

Statistika v chemii. Experiment je důležitým nástrojem výzkumu. V průběhu výzkumu bývají tvořeny složité fyzikálně-chemické modely a experiment slouží k jejich ověření. Statistické zpracování výsledků je pak součástí prakticky veškerého výzkumu. Statistické úlohy bývají nejčastěji ve formě plánování experimentu, detekci systematických chyb a tvorbě kalibračních přímek. Statistika se uplatňuje také v analytické chemii, optimalizaci a kontrole kvality v průmyslových výrobách. Dále bývají statisticky porovnány různé laboratoře, přístroje či podmínky.
V dalším textu se budeme zabývat pouze pojmy z popisné statistiky.

Co je popisná statistika. Z experimentálních měření získáváme data a ta chceme stručně a výstižně popsat. K tomuto účelu slouží popisná statistika. Popis konkrétního datového souboru je nedílnou součástí každé analýzy.

Data. Data jsou výsledkem pozorování nebo měření, které provádíme na nezávislých subjektech. Měříme nebo zjišťujeme hodnoty znaku, veličin, vlastností, například koncentrace určité látky, hmotnost, teplota, zabarvení, atd. Na jednom subjektu můžeme měřit více znaků. Výsledky zapisujeme do datové tabulky. Pozorování na jednotlivých subjektech jsou vetšinou v řádcích, jednotlivé měřené veličiny ve sloupcích. Statistickou analýzu provádíme většinou pomocí specializovaných statistických softwarů, pro příklad uveďme programy R, Statistica, SPSS, SAS atd.

Příkladem datového souboru jsou naměřené hodnoty obsahu alkoholu ve vzorcích vína z našeho lustračního příkladu.

Měřítka znaků. Měřítka můžeme dělit více způsoby. Prvním dělením je na

- nominální – jejich hodnoty jsou pouze označením různých kategorií (pohlaví, politický názor, barva, odrůda, ...),
- ordinální – jsou to uspořádané nominální hodnoty (vzdělání, spokojenost v práci (stupnice 1 až 5), stupeň bolesti, ...),
- intervalové – u nich lze uvažovat jejich rozdíly, ale nelze se ptát „kolikrát“ (rok narození, teplota ve stupních Celsia, ...),
- poměrové – většina veličin, které měříme (hmotnost, koncentrace, velikost, čas, ...).

Jiné dělení měřítek může být na

- kvalitativní neboli kategoriální faktory – existuje jen několik možných hodnot (kategorií) a zajímají nás četnosti jednotlivých hodnot, přičemž uvažovat charakteristiky jako průměr nemá smysl,
- kvantitativní neboli spojité – jejich hodnoty jsou čísla, zajímají nás charakteristiky polohy (průměr), variability atd.

Kvalitativní veličiny

Míry polohy

Průměr. Při výpočtu průměru $\bar{x}$ pozorujeme hodnoty $x_1, \ldots, x_n$. Průměr vypočteme jako

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1}$$

Někdy také bývá užitečné určit maximum a minimum zadaných hodnot.

Varianční řada. Při tvorbě variační řady postupujeme tak, že původní hodnoty $x_1, \ldots, x_n$ uspořádáme podle velikosti. Varianční řada

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \tag{2}$$

je neklesající posloupnost vytvořená z naměřených hodnot, přičemž x_1 je minimum, x_n je maximum. Je důležité uvědomit si rozdíl mezi x_1 a $x_{(1)}$.

Medián. Medián $\tilde{x}$ dělí data na dvě poloviny tak, že polovina je menší (nebo rovna) než $\tilde{x}$ a polovina vetší (nebo rovna) než $\tilde{x}$. Medián je tedy prostřední hodnota. Výpočet mediánu provádíme podle následujícího vzorce

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{je-li } n \text{ liché,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{je-li } n \text{ sudé.} \end{cases}$$

Kvantily. Kvantily neboli percentily můžeme charakterizovat následujícím způsobem.

- $\alpha - 100\,\%$ kvantil je hodnota taková, že $\alpha - 100\,\%$ hodnot v datech je menší nebo rovno a zbytek je vetší nebo rovno. Například $50\,\%$ kvantil je medián.
- Dolní kvartil $Q_1 = 25\,\%$ kvantil, je hodnota taková, že čtvrtina hodnot je menších (nebo rovných) a tři čtvrtiny jsou vetší (nebo stejné).
- Horní kvartil $Q_3 = 75\,\%$ kvantil, je hodnota taková že tři čtvrtiny hodnot jsou menší (nebo rovné) a čtvrtina je vetší (nebo stejná).

Úlohy z praxe, které využívají kvantilů mohou být například

- jaký obsah vápníku v krevním séru se považuje za nízký, tedy takový, jehož výskyt je u maximálně $5\,\%$ zdravých lidí,
- růstové křivky u dětí, jimiž zjišťujeme, zda není dítě extrémně malé nebo extrémně velké.

Příklad 1. V Motivačním příkladu určete pro hodnoty obsahu alkoholu průměr, variační řadu hodnot, minimum, maximum a medián.

Řešení. Připomeňme, že hodnoty byly

$$x_1 = 13,20;\ x_2 = 13,16;\ x_3 = 14,37;\ x_4 = 13,24;\ x_5 = 14,20;$$
$$x_6 = 14,39;\ x_7 = 14,06;\ x_8 = 14,83;\ x_9 = 13,86;\ x_{10} = 14,10;$$
$$x_{11} = 14,12;\ x_{12} = 13,75;\ x_{13} = 14,75;\ x_{14} = 14,38;\ x_{15} = 13,63;$$
$$x_{16} = 14,30;\ x_{17} = 13,83;\ x_{18} = 14,19;\ x_{19} = 13,64;\ x_{20} = 14,06;$$
$$x_{21} = 12,93;\ x_{22} = 13,71;\ x_{23} = 12,85;\ x_{24} = 13,50;\ x_{25} = 13,05;$$
$$x_{26} = 13,39;\ x_{27} = 13,30;\ x_{28} = 13,87;\ x_{29} = 14,02;\ x_{30} = 13,73.$$

- Průměr $\bar{x}$ vypočteme podle vztahu (1) jako

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}(13,20 + \ldots + 13,73) = 13,814.$$

- Varianční řada je tvaru

$$12,85;\ 12,93;\ 13,05;\ 13,16;\ 13,20;\ 13,24;\ 13,30;\ 13,39;\ 13,50;$$
$$13,63;\ 13,64;\ 13,71;\ 13,73;\ 13,75;\ 13,83;\ 13,86;\ 13,87;\ 14,02;$$
$$14,06;\ 14,06;\ 14,10;\ 14,12;\ 14,19;\ 14,20;\ 14,30;\ 14,37;\ 14,38;$$
$$14,39;\ 14,75;\ 14,83.$$

- Minimum je hodnota 12,85, maximum pak 14,83.

- Medián najdeme podle části vzorce pro n sudé, tedy

$$\tilde{x} = \frac{1}{2}(x_{(15)} + x_{(16)}) = \frac{1}{2}(13,83 + 13,86) = 13,845.$$

Míry variabily.

Míry variability měří rozptýlení neboli variabilitu či nestejnost.

Rozptyl. Rozptyl můžeme charakterizovat jako průměrný čtverec vzdálenosti od průměru. Spočteme jej podle vzorce

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \qquad (3)$$

Rozměrem je druhá mocnina původních jednotek.

Směrodatná odchylka Směrodatná odchylka je charakterizována jako odmocnina z rozptylu. Spočteme ji podle vzorce

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{4}$$

Směrodatná odchylka má stejný fyzikální rozměr jako původní data.
Existuje řada dalších popisných charakteristik (šikmost, špičatost, specializované popisné statistiky, ...). Ve statistické indukci slouží popisné statistiky jako odhady neznámých parametrů.

Příklad 2. V Ilustračním příkladu spočtěte pro zadaná data směrodatnou odchylku a rozptyl.

Řešení. Využijeme výsledků vypočtených v Příkladu 1. Dostáváme

$$\sum_{i=1}^{30} x_i^2 = 5732,319 \quad \text{a} \quad \bar{x}^2 = 190,817.$$

Odtud

$$s^2 = \frac{1}{29}(5732,31930 \cdot 190,817) = 0,269.$$

Směrodatnou odchylku potom vypočteme jako

$$s = \sqrt{0,269} = 0,519.$$

Grafické nástroje popisné statistiky.

Zmíníme se zde o dvou grafických nástrojích popisné statistiky a to o histogramu a krabicovém diagramu neboli boxplotu.

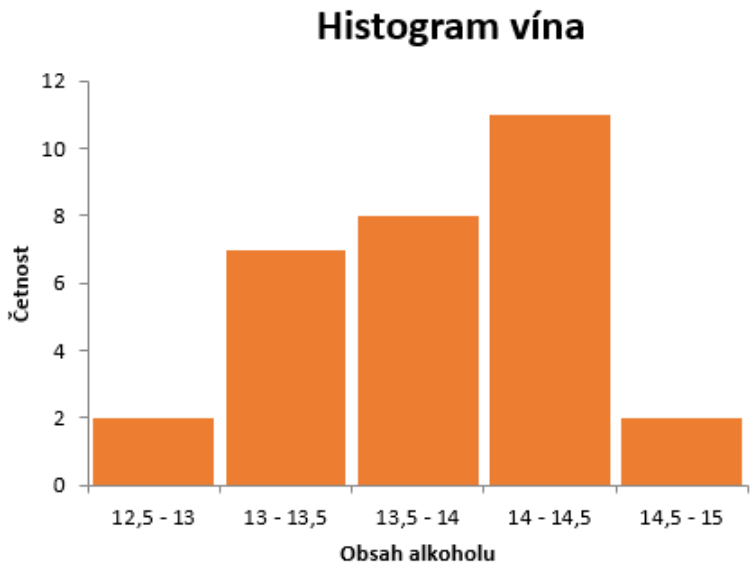
Histogram. Histogram dává nahlédnout, jak jsou jednotlivé hodnoty znaku v našich datech rozloženy, tedy které hodnoty se objevují často a které ojediněle. Histogram vytvoříme tak, že interval $I = [a; b]$, jenž pokrývá celé rozmezí dat, rozdělíme na K navazujících stejně velkých podintervalů A_k, kde $k = 1, \ldots, K$ a všechny budou délky $h = \frac{b-a}{K}$. S výjimkou prvního je bereme je například zprava uzavřené. Označíme n_k počet pozorování, které padly do A_K. Histogram je pak grafické znázornění intervalových četností n_K, neboli každému A_K odpovídá obdélník, jehož výška je rovna n_K.

Krabicový diagram. Krabicový diagram nemá úplně závaznou definici. Obvykle je v něm zakreslen výběrový medián a kvartily. Krabice je tvořena tak, že horní a dolní okraj určují výběrové kvartily Q_1 a Q_3, uprostřed se nachází čára určující výběrový medián. „Vousy“ ukazují rozmezí dat od kvartilu k minimu či maximu, není-li odlehlé. Odlehlé pozorování je takové, které je dál než $\frac{3}{2}(Q_3 - Q_1)$ od bližšího kvartilu.

Příklad 3. Pro hodnoty obsahu alkoholu z Motivačního příkladu nakreslete histogram.

Řešení. Postupujeme tak, že zvolíme $a = 12,5$, $b = 15$, $K = 5 \rightarrow h = 0,5$.

k	interval A_k	četnost n_k
1	[12, 5,13]	2
2	[13, 13,5]	7
3	[13,5, 14]	8
4	[14, 14,5]	11
5	[14,5, 15]	2



Literatura

[1] BUDÍKOVÁ, Marie, MIKOLÁŠ Štěpán , LERCH Tomáš : *Základní statistické metody.*, Vydání první. Brno: Masarykova univerzita, 2005. ISBN 80-210-3886.

[2] BUDÍKOVÁ Marie : Studijní materiály předmětu PřF:MAS01 [online]. [cit. 2014-01-09]. Dostupné z: `https://is.muni.cz/auth/el/1431/podzim2014/MAS01/um/50490616/`

[3] HUDECOVÁ Šárka: Matematická statistika [online]. [cit. 2014-01-09]. Dostupné z: `http://www.karlin.mff.cuni.cz/~hudecova/education/download/chem_predn/popisna_tisk.pdf`