POKUD TO NEUMÍŠ VYSVĚTLIT JEDNODUŠE, NEROZUMÍŠ TOMU

ADRIANA ŘEHÁČKOVÁ

(NE)UČEBNICE STATISTIKY

TEORIE VYSOKOŠKOLSKÉ STATISTIKY JEDNODUŠE A ZÁBAVNĚ

Neučebnice statistiky
Adriana Řeháčková

Přebal knihy: Jiří Ondráček, Adriana Řeháčková
Úpravy obrázků a ilustrace: Adriana Řeháčková
Sazba: Adriana Řeháčková
Jazyková korektura: Michaela Kullová

Chyby a připomínky: info@statistickyneklasicky.cz
Pochvaly a recenze: info@statistickyneklasicky.cz

www.statistickyneklasicky.cz

OBSAH

STATISTIKA JINAK…

Tuto (ne)učebnici jsem napsala, protože přesně takovou knihu jsem při svém učení hledala….

Věřím, že všechno se dá vysvětlit jednoduše a se špetkou zábavy a jelikož je statistika jedním z nejobávanějších předmětů na mnoha vysokých školách, rozhodla jsem se to už před nějakou dobou změnit a dát ji trochu **lidského přístupu**.

Začalo to doučováním, organizováním desítek kurzů, vytvořením vlastního webu s online kurzy a poslední metou je zatím tato (ne)učebnice. Poznala jsem díky této práci více než tisíc studentů a některým, jak říkají, zachránila prdel, tak doufám, že i vám pomůžu tento předmět v pohodě zvládnout a hlavně si z něj i něco odnést.

PROČ (NE)UČEBNICE?

Tak především proto, že jsem si chtěla o statistice psát tak, jak já sama chci. Pokud si tedy potrpíte na přesné definice a složité zápisy, asi to nebude úplně váš šálek kávy. Cílem je, aby si z této knihy i běžný smrtelník odnesl, co možná nejvíce. Udělal si ve statistice pořádek, všechno si hezky spojil a pochopil nutný teoretický základ. V této knize **nenajdete řešení příkladů** a vysvětlování postupů, k tomuto účelu jsou vytvořeny moje **online kurzy**, které by měly být nedílnou součástí učení. Na webových stránkách si vyberete kurz přesně podle vaší školy (případně mi můžete napsat o doporučení vhodného kurzu) a dosáhnete tak ideálního spojení teorie a výpočtů. Klíčové je, aby vám do sebe všechno zapadalo, chápali jste proč děláte, co děláte a uměli to aplikovat na příkladech.

Tato kniha obsahuje, řekněme, **kompletní základní statistiku**, v některých částech jsem nešla tolik do detailů, jak bych chtěla, jelikož jedním z mých požadavků bylo, aby to bylo krátké a abyste se **učili to, co opravdu potřebujete**, bez omáčky kolem. Je tedy dost možné, že pro vaše účely se vám budou hodit jen některé z kapitol.

Nebojte se vesele přeskakovat a vybírat si, co sami potřebujete, je to na vás.

CO JE TO STATISTIKA?

"Statistika nuda je, jsou to samé výdaje"
Tak přesně toto si většina české populace představí, když se řekne **statistika.** A upřímně, i u mě to tak nějakou dobu bylo.

Jenže, co když je to vlastně celkem zábava a navíc, věděli jste, že statistika je všude kolem nás a má přesah do mnoha oborů? Může se jednat například o finance, psychologii, medicínu, demografii a spoustu dalších odvětví, pro které je statistika jejich nedílnou součástí. Dokážeme si spočítat pravděpodobnosti, nalézt v datech závislosti a také například modelovat a odhadovat budoucí vývoj dat.

Může nám velmi pomoc a je na nás, zda ji dokážeme využít k našemu dobru.

Například si díky ní dokážeme velmi jednoduše spočítat, jaká je pravděpodobnost, že vyhrajeme ve sportce. Ta je sice mizerná, ale i to vás statistika naučí, trochu toho kritického myšlení a obezřetnějšího pohledu na svět. A nyní tomu pojďme dát aspoň trochu té odbornosti.

Definice: Statistika je činnost nebo také věda, zabývající se sběrem, zpracováním a vyhodnocováním dat. Slouží například k odhadům, ke zjištění závislostí, k testování hypotéz a k výpočtům pravděpodobností. Jejím cílem je **získat co nejlepší (nejrelevantnější) informace z dostupných dat.**

Zároveň buďte vždy obezřetní a dávejte pozor na data, se kterými pracujete, jelikož, jak kdysi údajně řekl Aaron Levenstein:

Statistika je jako bikini. Co odhaluje je zajímavé, co skrývá je podstatné.

A ani všechny bikiny nejsou stejné, proto se statistika obecně dělí na: **popisnou statistiku, matematickou statistiku, inferenční statistiku,...** My se společně ponoříme do některých částí a vysvětlíme si základní, až lehce pokročilou teorii, společně s metody, které se k daným tématům vážou a které vám poslouží na většině vysokých škol a budou se hodit i v praxi.

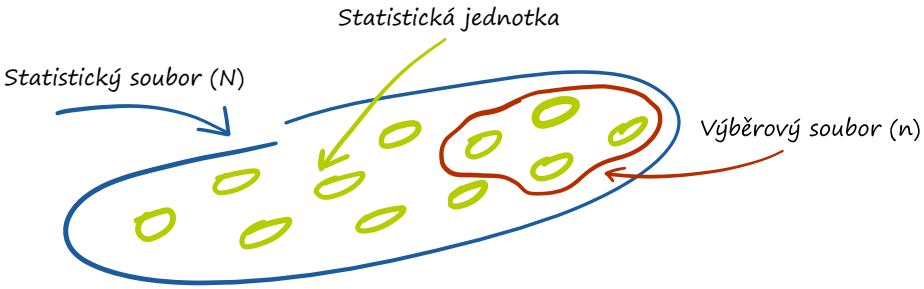
Tak pojďme na to!

ZÁKLADNÍ POJMY

Statistický soubor - Množina (skupina) prvků, které mají alespoň jednu společnou vlastnost, jenž je spojuje - **identifikační znak**.

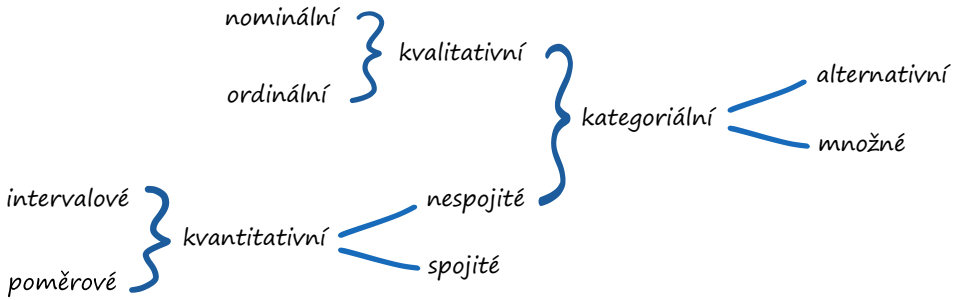
Statistická jednotka - Konkrétní objekt (prvek), který je součástí stat. souboru, předmětem pozorování.

Výběrový soubor - Jedná se o část (výběr) z našeho základního, statistického souboru. Tyto jednotky vybereme buď náhodně a nebo dle jistých pravidel.



Identifikační znak - Společná vlastnost pro všechny jednotky v souboru.

Statistická proměnná - Co literatura, či vysoká škola, to jiné členění. My si zde společně ukážem základní rozdělení, ve kterém se pokusím pokrýt, co nejvíce možných typů členění.



Kvalitativní proměnné - Slovně zadané proměnné, která se dělí na **nominální a ordinální** proměnné.

Nominální proměnná - Je proměnná vyjádřená slovy, kterou nejde seřadit a lze říci jen: Jsou stejné, nejsou stejné? *Například: škola, obor studia, povolání, pohlaví,...*

Ordinální proměnná - Nebo-li pořadová proměnná, je taková slovní proměnná, u které lze určit i pořadí. *Například: ukončený stupeň vzdělání, ročník, známka ve škole,...*

Kvantitativní proměnné - Číselně zadané proměnné, které se dělí na **intervalové a poměrové** proměnné a dle jiného dělení také na **nespojité a spojité.**

Intervalová proměnná - hodnotami jsou čísla a lze počítat o kolik je jedna hodnota větší než druhá. *Například: Teplota v Celsiích, počet dětí v rodině.*

Poměrová proměnná - k hodnotám můžeme navíc (oproti intervalovým), vypočítat i to, kolikrát je hodnota větší, tedy se jedná jen o **kladné hodnoty.** *Například: rychlost, hmotnost, délka,..*

Nespojitá proměnná - Je proměnná, která nabývá celočíselných hodnot, můžeme ji taky nazývat proměnnou **diskrétní**. *Například: počet dětí v rodině, počet karet v balíčku, počet knih v šuplíku,..*

Spojitá proměnná - Proměnná, která nabývá hodnot z konečného nebo nekonečného intervalu. *Například: výška, váha, roční příjem domácnosti,..*

Kategoriální proměnné - **Nominální, ordinální a nespojité** proměnné, lze označit jako kategoriální. A ty se dále dělí na alternativní a množné.

Alternativní proměnná - Jinak se jí říká **dichotomická** a nabývá pouze 2 obměn. *Například: muž/žena, kuřák/nekuřák,..*

Množná proměnná - Jinak se nazývá **vícekategoriální** a nabývá více než 2 kategorií. *Například: rodinný stav, počet dětí,...*

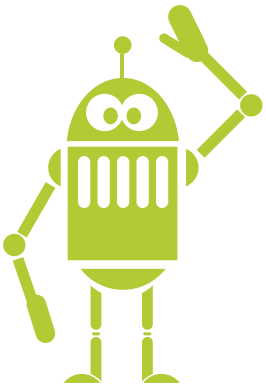
Než se pustíme do další kapitoly, ráda bych zde ještě vysvětlila rozdíl mezi pojmy **procenty a procentními body**, se kterými se ve statistice můžete setkat.

Podle webu www.focus-age.cz, Facebook v roce 2015 využívalo 71 % dotázaných **teenegerů**, v roce 2019 se jednalo o 51 % dotázaných teenegrů. Jaký je tedy pokles? To lze interpretovat dvěma způsoby:

Pokud se jedná o **procenta**, tak jde o pokles 28,17 %, nicméně mnohem více se zde hodí používat **procentní body**, kde hned na první pohled vidíme pokles o 20 procentních bodů (71-51).

Pokud by se jednalo například o růst mezi 20 % a 35 %, tak tento růst je o 75 %, ale v procentních bodech je to 15 procentních bodů.

Už vidíte ten obrovský rozdíl, jaký může (ne)správná interpretace udělat?



Bonus pro zvídavé hlavičky:
Slyšeli jste někdy o Českém statistickém úřadu? ČSÚ je hlavní orgán státní statistické služby, který koordinuje sběr a zpracování statistických údajů na území ČR, zabezpečuje získávání a zpracování údajů pro statistické účely a poskytuje statistické informace a co je důležité, je nezávislý na vládě a politických stranách.

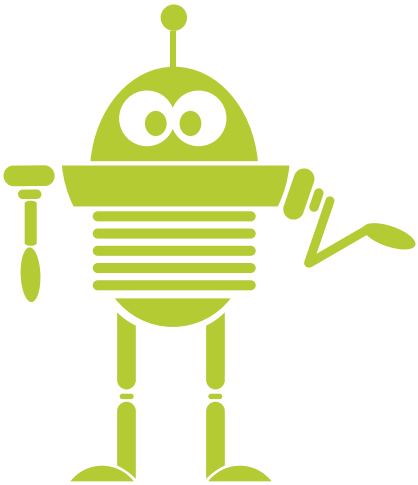
POLOHA A VARIABILITA

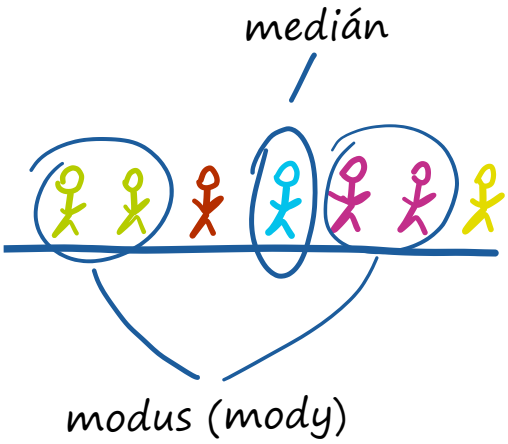
Úvod máme za sebou a čeká nás určování polohy a variability statistických znaků. Charakteristikou polohy jsou **průměr, medián, modus** a **kvantily.** Všechny tyto charakteristiky nám dávají informaci o umístění, poloze proměnné.

CHARAKTERISTIKY POLOHY

Medián - **Prostřední hodnota seřazeného souboru**. Medián dělí soubor na 50 % hodnot, které jsou menší než daná hodnota (medián) a 50 % hodnot, které jsou větší než medián. Zároveň patří do kvantilů a jedná se o druhý kvartil.

Modus - **Nejčetnější hodnota**. Jedná se o hodnotu, které je v souboru nejvíce, má největší četnost. Pokud v souboru žádná taková hodnota není, pak modus není. Pokud je takových hodnot více, například dvě hodnoty jsou v souboru 9x (kde uvažujme, že jde o největší četnost) potom jsou modem obě tyto hodnoty a jen se vypíšou.





Kvantily - Kvantily zkrátka dělí soubor na určitou část, kterou si sami zvolíme. Nejčastějším dělením jsou buď kvartily, decily nebo percentily. (Pozn. kvartily a kvantily jsou něco jiného, kvartil je podmnožinou kvantilů).

Kvartily - Kvartily dělí soubor na čtvrtiny, využili jsme je v našem krabicovém grafu, kde **Q1 = první kvartil** a rozděluje soubor na dolních 25 % hodnot a horních 75 % hodnot, potom máme **medián**, ten dělí soubor na dvě stejné poloviny a jako poslední máme **Q3 = třetí kvartil**, který dělí soubor na dolních 75 % hodnot a horních 25 % hodnot. Rozdíl mezi třetím a prvním kvartilem se nazývá **mezikvartilové rozpětí**, které rovněž bylo znázorněno v **krabicovém grafu** a právě ta „krabice" je naše mezikvartilové rozpětí.

Decily - Decily nám dělí soubor na desetiny. Tedy na deset částí a my vždy spočítáme takovou část, která nás zajímá, například dolních 40 % hodnot a horních 60 % hodnot.

Percentily - Percentily dělí soubor na setiny. Jistě si každý pamatuje, že ve výsledcích maturit, či jiných celostátních zkouškách, se velmi často objevuje percentil, ten vám řekne, kolik lidí dopadlo hůře a lépe, než právě vy. Například percentil 93 %, znamená, že 93 % lidí dopadlo hůře a 7 % dopadlo lépe, než vy.

Výpočty těchto charakteristik řeším v některém ze svých online kurzů. Obecně jde ale o velmi jednoduché výpočty, akorát dávejte pozor, abyste vždy měli hodnoty seřazené podle velikosti!

PRŮMĚRY

Společně si ukážeme **3 základní typy** průměrů.

Aritmetický průměr

To je takový ten klasický průměr, který zná určitě každý. Prostý aritmetický průměr se počítá tak, že sečteme hodnoty v souboru a vydělíme je jejich počtem. **Vážený aritmetický průměr** je v podstatě to samé, akorát zde využijeme četnosti. Když je nějaká hodnota v souboru třeba 10x, tak ji přece nebudu 10x sčítat, ale dám 10x daná hodnota a přesně to dělá vážený aritmetický průměr. Případně je schopen zaznamenat rozdílnou váhu, například kdybychom měli známky ze zkoušení a z domácího úkolu, **váha pak představuje četnost**.

prostá forma: $\overline{x} = \frac{\sum x_i}{n}$ vážená forma: $\overline{x} = \frac{\sum x_i \cdot n_i}{\sum n_i}$

Harmonický průměr

Harmonický průměr se používá při výpočtů průměrné rychlosti, průměrné hustoty nebo průměrné práce. Pracuje na principu převrácených hodnot. Zpravidla ho užijeme tehdy, když se jedná o již průměrné hodnoty. (Například chceme zjistit průměrnou rychlost z průměrných rychlostí.). Váženou variantu využijeme pokud se například liší délka trasy na které byla rychlost měřena.

prostá forma: $\bar{x} = \frac{n}{\sum \frac{1}{x_i}}$ vážená forma: $\bar{x} = \frac{\sum n_i}{\sum \frac{n_i}{x_i}}$

Geometrický průměr

Geometrický průměr se používá primárně k výpočtu tempa růstu. Je podobný aritmetickému, ale místo sčítání a dělení, používá násobení a odmocninu. Zpravidla ho využijí makroekonomové, pokud chcete spočítat tempo růstu, inflaci, HDP, růst cen, sáhněte po geometrickém průměru!

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

CHARAKTERISTIKY VARIABILITY

Je na čase zjistit něco o variabilitě, nebo-li rozptýlenosti hodnot. Nejvíc populární charakteristika je v této „branži" **rozptyl**. Je hodně populární, ale až jeho kamarádka **směrodatná odchylka**, nám o datech opravdu něco řekne a dokážeme si pod ní něco představit. Dalšími absolutními charakteristikami variability jsou například **rozpětí (variační, kvartilové,...)**, **průměrná absolutní odchylka**. Relativní charakteristikou variability je **variační koeficient**.

Rozptyl

Co to teda je? Říká se mu také druhý centrální moment, *„hmm, to asi moc nepomohlo"*. Dobře, tak si ukážeme, jak se počítá. Vezmete každou hodnotu v souboru, odečtete od ní průměr a každou tuto odchylku umocníte (kdybyste umocnění vynechali, potom platí, že suma těchto odchylek je rovna nule, což je nám naprosto k ničemu.) a následně všechny odchylky sečteme. Na závěr provedeme standardizace tím, že tyto odchylky vydělíme počtem hodnot a máme hotovo.

Populační vs výběrový rozptyl. Pozor na tento rozdíl. Pokud uvažujeme, že pracujeme s populací a chceme spočítat rozptyl, potom volíme klasický populační rozptyl. Pokud ale pracujeme s výběrem, tak sumu dělíme počtem hodnot zmenšených o jednotku.

populační rozptyl

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

výběrový rozptyl

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Směrodatná odchylka

Jak už jsme si řekli, je to kamarádka rozptylu a to proto, že je jeho odmocninou. Rozptyl vyjde kvůli mocnění v kvadrátu, zatímco směrodatná odchylka vychází již v původních jednotkách a můžeme tak variabilitu snadno interpretovat. Říká nám, jak moc jsou hodnoty rozptýleny od průměru hodnot. Samozřejmě, abychom měli směr. odchylku, musíme prvně spočítat rozptyl.

směrodatná odchylka: $s = \sqrt{s^2}$ $\sigma = \sqrt{\sigma^2}$

*(Poznámka: malým **s** značíme výběrovou hodnotu a **sigmou** značíme populační hodnotu)*

Variační koeficient

Variační koeficient udává relativní míru variability. Počítá se tak, že směrodatnou odchylku vydělíme průměrem, pokud toto číslo vynásobíme stem, máme var. koeficient vyjádřený v procentech. Relativní míra variability je vhodná hlavně k tomu, že **můžeme snadno porovnat variabilitu mezi více soubory**.

variační koeficient: $v_x = \frac{s_x}{\bar{x}}$

Rozpětí

Variační rozpětí - počítá se jako maximální hodnota souboru mínus minimální hodnota souboru. Udává nám rozpětí hodnot mezi kterými se soubor pohybuje.

Kvartilové rozpětí - Počítá se jako třetí kvartil mínus první kvartil.

variační rozpětí

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

mezikvartilové rozpětí

$$R = x_{0,75} - x_{0,25}$$

A CO VLASTNOSTI?

Pojďme si ukázat co platí pro průměr a rozptyl. Právě tato znalost vám usnadní mnoho výpočtů a hlavně se vyhnete nechtěným chybám.

Aritmetický průměr
U aritmetického průměru platí, že pokud u každé hodnoty v souboru přičteme nebo odečteme nějakou konstantu (libovolné číslo), tak se o tuto konstantu zvýší, případně sníží i průměr. To samé platí pro násobení a dělení. Všechny hodnoty vynásobím pěti, tak i průměr vynásobím pěti, toť vše. A na závěr ještě jedna vlastnost a to ta, že průměr jednoho čísla je to samé číslo. To dává smysl, ne?

Rozptyl a směrodatná odchylka
U rozptylu a směrodatné odchylky platí, že pokud ke každé hodnotě v souboru přičteme konstantu (opět nějaké libovolné číslo), potom se rozptyl a ani směrodatná odchylka **nezmění**.

Rozptyl a směr. odchylka jednoho čísla je **nula**, jedno číslo se zkrátka nemá od čeho odchýlit.

A na závěr, pokud všechny hodnoty vynásobíme nějakým číslem, konstantou, potom rozptyl musíme **vynásobit tímto číslem na druhou**. Tedy, pokud se všechny hodnoty zvýší 4 krát, rozptyl se zvýší 16 krát, zde je to ovlivněno tím, že rozptyl vychází v mocnině, proto se i násobené číslo umocňuje. Směrodat-

nou odchylku potom stačí jen vynásobit daným číslem, stejně jako v případě průměru.

vlastnosti arit. průměru	vlastnosti rozptylu
$\overline{x+k} = \bar{x}+k$	$s^2_{x+k} = s^2_x$
$\bar{k} = k$	$s^2_k = 0$
$\overline{x \cdot k} = \bar{x} \cdot k$	$s^2_{x \cdot k} = k^2 \cdot s^2_x$

Variační koeficient

Pro variační koeficient platí trochu jiná pravidla. Pokud všechny hodnot zvýšíme/snížíme o nějakou konstantu, potom se v případě **zvýšení hodnot**, variační koeficient **sníží**, v případě **snížení hodnot** o nějakou konstantu se variační koeficient **zvýší**. Je to dané tím, že směr. odchylka, která je v čitateli, se nemění a průměr, který je ve jmenovateli, se například zvýší, potom tedy dělíme větším číslem a var. koeficient se musí snížit. V případě násobení nebo dělení hodnot nějakou konstantou se variační koeficient **nemění**, protože se to ve zlomku hezky vykrátí a tedy to nezpůsobí žádnou změnu, rel. variabilita bude stejná.

Rozklad rozptylu

Může se nám stát, že budeme chtít spočítat rozptyl, směr. odchylku nebo variační koeficient za **více skupin dohromady.** To lze udělat poměrně snadno přes vzorec na rozklad rozptylu. Pozor, že nejde udělat žádné průměrování nebo sčítání, dělení, atd., je nutné použít speciální vzorec.

vnitroskupinový rozptyl — meziskupinový rozptyl

$$s^2_x = \overline{s^2} + s^2_{\bar{x}} = \frac{\sum s^2_i \cdot n_i}{\sum n_i} + \frac{\sum (\bar{x}_i - \bar{x})^2 \cdot n_i}{\sum n_i}$$