# Základní pojmy statistiky

Statistika je věda, která zkoumá, zpracovává a vyhodnocuje data.

## Populace #

Cílem statistiky je provést experiment, jehož účelem je zjistit něco zajímavého o dané populaci. Populací se myslí obecně jakýkoliv soubor prvků, které chceme zrovna zkoumat. Pokud chceme zjistit, jaký je průměrný věk obyvatel České republiky, tak naší populací budou všichni obyvatelé České republiky.

Pokud ale budeme chtít zjistit průměrnou spotřebu benzinu osobních aut na sto kilometrů, bude naše populace rovna množině všech osobních aut (na daném území).

## Výběr, výběrový soubor #

Často není možné pracovat se všemi prvky populace. Představme si, že chceme zjistit, co si lidé v České republice myslí o povinné maturitě z matematiky. Abychom opravdu zjistili, co si lidé o povinné maturitě myslí, museli bychom chodit dům od domu, most od mostu a ptát se každého občana, co si o povinné maturitě z matematiky myslí. Něco takového není v praxi možné. Některé důvody:

- Je to příliš drahé. Dotázat se všech přibližně deseti a půl milionu obyvatel není levná záležitost. Například první přímá volba prezidenta stála 625 milionů korun.
- Trvá to příliš dlouho. Volby se jistě připravovaly několik měsíců – pokud potřebujete výsledek statistiky za týden, tak je to příliš dlouhá doba.
- Ne všichni budou chtít odpovědět. Někdo vám už z principu nebude chtít odpovědět vaše otázky. Pokud jsou naší populací nějaké stroje, tak se zase mohou rozbít. Pokud byste u aut sledovali počet ujetých kilometrů, tak se může stát, že se tachometr rozbije nebo ho někdo záměrně přetočí.
- Experiment může být příliš nebezpečný. Z dotazu na povinnou maturitu asi nikdo infarkt nedostane, ale můžeme si vzít jiný příklad – testování nového léku „hojito“. Co by se stalo, kdybychom „hojito“ testovali na celé populaci České republiky a během testování by se zjistilo, že 20 % testovaných lidí okamžitě dostane ukrutný průjem? No, asi bude lepší, když ten lék nejprve otestujeme na menší skupině lidí, že?

Abychom se vyhnuli těmto nevýhodám, volíme z dané populace pouze nějaký výběr (nebo též výběrový soubor). Pokud máme populaci P, tak výběrovým souborem V je každá podmnožina P, tedy $V \subseteq P$ .Náš experiment poté provedeme pouze na tomto výběru V a výsledky zevšeobecníme na celou populaci. Tyto výsledky budou samozřejmě nepřesné – jak moc nepřesné budou záleží především na tom, že jak velký je výběr V a jakou metodu jsme zvolili pro výběr prvků do V.

Typické chyby tak mohou být:

- Příliš malý počet prvků ve V. Pokud se na povinnou maturitu zeptáte prvních sedmi lidí, které uvidíte, tak nemůžete dostat smysluplné výsledky.
- Nereprezentativní výběr prvků z populace. Pokud se na povinnou maturitu z matematiky zeptáte tisícovky absolventů Matematicko-Fyzikální fakulty, tak získáte jiné odpovědi, než kdybyste se zeptali tisícovky studentů třetích ročníků středních škol.

# Proměnné #

Během experimentu zkoumáme prvky výběrového souboru. Údaje, které sledujeme, nazýváme proměnné a hodnoty proměnných nazýváme varianty. Existují základní typy proměnných:

- Kvalitativní proměnná: tuto proměnnou typicky nemá smysl měřit, jedná se o nějaké slovní ohodnocení. Typickým příkladem může být dotaz na národnost. Variantami takové proměnné bude např. hodnoty „česká národnost“, „slovenská národnost“ apod. Nemá přitom smysl měřit nebo porovnávat českou a slovenskou národnost. Můžeme porovnávat počty Čechů a Slováků, ale samotnou národnost porovnávat nemůžeme.

Do této kategorie spadá i otázka na povinné maturity, kde se očekávají odpovědi „ano, chci povinnou maturitu z matematiky“ nebo „ne, nechci povinnou maturitu z matematiky“, což jsou varianty této proměnné. Opět můžeme porovnávat počty odpovědí, ale nemá smysl porovnávat samotné „ano“ a „ne“.

- Kvantitativní proměnné: tuto proměnnou změříme. Jedná se tak o délky, hmotnosti, časy, počty a podobně. Kvantitativní proměnné dále dělíme na diskrétní a spojité proměnné:

# Diskrétní proměnná #

Diskrétní proměnná obsahuje konečný počet variant nebo obsahuje spočetný počet variant (viz dále). Poměrně často se jedná o celá čísla. Například počet žáků ve třídě – v běžné třídě bude řekněme něco mezi patnácti a čtyřiceti dětmi.

Diskrétní proměnná se vyznačuje tím, že jsme vždy schopni říci, jaké jsou další a předchozí varianty. Pokud je ve třídě 3B 28 dětí, tak předchozí varianta je 27 dětí a následující 29 dětí. U kvalitativní proměnné to většinou nejsme schopni udělat – jaká je následující varianta za českou národností?

Diskrétní proměnná může být i nekonečná, ale musí být spočetná – to znamená, že stále musíme být schopni určit předchozí a následující variantu. Například bychom mohli zavést proměnnou „vzdálenost dvou objektů s přesností na jeden kilometr“. Pokud změříme, že vzdálenost dvou objektů, například auta a stodoly, je 12 kilometrů, tak opět platí, že další a předcházející varianta je 13, respektive 11 kilometrů. Přitom vzdálenost není nejspíš nijak omezena. Pokud máme dva objekty od sebe vzdálené 1 500 000 kilometrů, jistě najdeme i objekty, které jsou od sebe vzdálené 1 500 001 kilometrů.

Proměnná by zůstala diskrétní, i kdybychom změnili přesnost na desetiny kilometru (tj. na stovky metrů). Pak bychom mohli naměřit vzdálenost 15,7 km a následující a předchozí hodnoty by byly 15,8 a 15,6.

Pokud předchozí nebo následující varianta neexistuje, tak to není v rozporu s tím, že je proměnná diskrétní. Například pro vzdálenost nula kilometrů neexistuje předchozí varianta – vzdálenost minus jeden kilometr nedefinujeme. Přesto je vzdálenost s přesností na jeden kilometr diskrétní proměnná.

# Spojitá proměnná #

Spojitá proměnná vždy obsahuje nekonečný počet variant. Hodnotami jsou typicky reálná čísla, takže se jedná například o vzdálenost (bez dodatku o přesnosti). U spojitých proměnných nedokážeme určit předchozí ani následující variantu. Pokud změříme, že vzdálenost něčeho je 3,58745 metrů, tak nedokážeme najít číslo, které je přesně za tímto číslem.

V množině reálných čísel jsou i iracionální čísla s nekonečným desetinným rozvojem. My samozřejmě nemáme přístroje, které by dokázaly změřit vzdálenost na takovou přenost, takže v realitě je každá taková proměnná stejně diskrétní – právě proto, že každý přístroj má nějakou přesnost. Pokud něco měříte pravítkem, tak tam máte přesnost na jeden milimetr. Můžete tak změřit, že knížka má šířku 167 mm nebo 168 mm, ale nic mezi tím; samozřejmě, pokud to nějak neodhadnete atp.

Pokud máte nějaký vědečtější přístroj, můžete mít přesnost na jeden mikrometr. Na úplně přesně změření objektu to ale ani tak nejspíš stačit nebude.

Přesto všechno obyčejně mluvíme o vzdálenost nebo o hmotnosti jako o spojitých proměnných. V praxi je podobné zjednodušení nutné a obyčejně ničemu nevadí.

# Náhodná proměnná #

Náhodná proměnná je diskrétní nebo spojitá proměnná, pro kterou nedokážeme před provedením experimentu určit její výslednou hodnotu. Náhodnou proměnnou tak může být výsledek hodu šestistěnnou kostkou. Dokud touto kostkou nehodíme, tak nemůžeme vědět, jaké číslo nám na kostce padne.

Můžeme být schopni předpovědět, že nějaké hodnoty budou pravděpodobnější než ostatní, to nám nevadí, jen si nesmíme být úplně jistě, že získáme nějakou konkrétní hodnotu. Například pokud bychom náhodně vylosovali jednoho obyvatele ČR a zeptali bychom se ho, v jakém městě žije, je pravděpodobnější, že bude žít v Praze než někde v Kravařích. V Praze zkrátka žije více lidí.

Pokud bychom měli kostku, která by měla na pěti stranách šest puntíků a na zbylé šesté straně dva puntíky, je daleko pravděpodobnější, že nám při hodu padne šest puntíků. Pořád je to ale náhodná proměnná, protože není jisté, že padne šest puntíků.

Kdybychom tuto kostku upravili tak, aby na všech šesti stěnách bylo šest puntíků, nebyl by hod kostkou náhodnou proměnnou, protože by nám vždy padlo šest puntíků.

# Rozložení četnosti

Rozložení nám popisuje množinu hodnot, kterých může nabýt naše náhodná proměnná.

## Jednorozměrné rozložení četnosti #

Rozložení typicky ukazujeme ve formě tabulky nebo grafu. Začneme tím, že provedeme nějaký experiment. Paní učitelka Logaritmová učí v osmé cé matematiku, můžeme se podívat, jaké známky dostali její žáci na vysvědčení. Populací tohoto experimentu budou všichni žáci osmé cé. Je jich celkem 30. Výsledky experimentu zapíšeme do tabulky:

| Známka | Počet žáků |
|---|---|
| Výborně | 7 |
| Chvalitebně | 13 |
| Dobře | 6 |
| Dostatečně | 3 |
| Nedostatečně | 1 |

Tato tabulka nám tak zobrazuje rozložení četnosti diskrétní proměnné „výsledná známka z matematiky na vysvědčení“. Stejné rozložení bychom mohli zobrazit pomocí sloupcového grafu takto:


## Skupinové rozložení četnosti #

V předchozím příkladě jsme měli u každé známky jen jeden sloupeček, který nám říkal, kolik žáíků dostalo danou známku. My můžeme tyto četnosti ještě rozdělit na nějaké zajímavé podskupiny. Například by nás mohlo zajímat, jakou známku dostali kluci a jakou holky. Místo jednoho sloupce tak zavedeme dva, jeden pro kluky a druhý pro holky. Taková tabulka už by pak zobrazovala skupinové rozložení četnosti.

| Známka | Počet kluků | Počet holek |
|---|---|---|
| Výborně | 5 | 2 |
| Chvalitebně | 6 | 7 |
| Dobře | 4 | 2 |
| Dostatečně | 1 | 2 |
| Nedostatečně | 1 | 0 |

Stejná data ve sloupcovém grafu:

# Relativní četnost #

Někdy nepotřebujeme znát absolutní počet žáků, kteří dostali takovou nebo makovou známku. Můžeme jen chtít vědět, kolik procent žáků dostalo nedostatečnou a podobně. K tomu slouží relativní četnost.

Víme, že ve třídě osmé cé máme celkem 30 žáků. Označme tento počet N, tedy N = 30. Pokud chceme vypočítat relativní četnost, vezmeme absolutní četnost a vydělíme N. Takže relativní četnost výskytu známky dobře, pokud nebudeme rozlišovat kluky a holky, zjistíme tak, že vydělíme 6/30, kde 6 je počet žáků, kteří dostali známku dobře. Výsledkem je relativní četnost 0, 2. Pokud chceme výsledek v procentech, vynásobíme toto číslo stem: 0, 2 · 100 = 20 %. Relativní četnost můžeme opět zapsat do tabulky:

| Známka | Absolutní četnost | Relativní četnost |
|---|---|---|
| Výborně | 7 | 0,2333… |
| Chvalitebně | 13 | 0,4333… |
| Dobře | 6 | 0,2 |
| Dostatečně | 3 | 0,1 |
| Nedostatečně | 1 | 0,0333… |

I relativní četnost můžeme zobrazit v grafu:




# Kumulativní četnost #

Vedle sloupečku s absolutní četností výsledných známek můžeme zobrazit ještě jeden sloupeček s kumulativní četností. Prohlédněte si následující tabulku:

| Známka | Počet žáků | Kumulativní četnost |
|---|---|---|
| Výborně | 7 | 7 |
| Chvalitebně | 13 | 20 |
| Dobře | 6 | 26 |
| Dostatečně | 3 | 29 |
| Nedostatečně | 1 | 30 |

Ve druhém řádku máme ve sloupci kumulativní četnost hodnotu 20. Tu jsme získali tak, že jsme sečetli hodnoty počty žáků v prvním a ve druhém sloupci, tedy 7 + 13 = 20. Ve třetím řádku tak máme kumulativní četnost 26, protože 7 + 13 + 6 = 26. Jinými slovy jsou to součty četností všech řádků, které jsou výše než současný řádek plus hodnota z aktuálního řádku.

Na prvním řádku tak bude kumaltivní četnost shodná s absolutní četností, na posledním řádku bude kumulativní četnost rovna velikosti celé populace (ve třídě je třicet žáků).

Můžeme také spočítat kumulativní relativní četnosti:

| Známka | Relativní četnost | Kumulativní relativní četnost |
|---|---|---|
| Výborně | 0,2333 | 0,2333… |
| Chvalitebně | 0,4333 | 0,6666… |
| Dobře | 0,2 | 0,86666… |
| Dostatečně | 0,1 | 0,96666… |
| Nedostatečně | 0,0333 | 1 |

V prvním řádku opět máme u kumulativní relativní četnosti stejnou hodnotu jako ve sloupečku Relativní četnost. V posledním máme 1, což znamená 100 % populace.

# Rozložení četnosti spojité proměnné #

Pokud náhodná proměnná není diskrétní, ale spojitá, nezobrazuje se obvykle v tabulce ani ve sloupcovém grafu, ale v běžném grafu jako klasická funkce. Spojitá proměnná je například teplota. Na stránkách chmi.cz, Český hydrometerologický ústav, si můžeme vyhledat průměrnou teplotu v Praze v měsíci lednu za roky 1961–1990 a zobrazit je v grafu jako spojitou křivku:

Průměrná teplota v lednu v Praze29162330-3,0-1,50,01,53,0Den v lednu°C