

Popisné statistiky: Míry centrální tendence a variability

Mgr. Zuzana Szabó Lenhartová

Popisné statistiky

Popisná statistika

- popis statistického souboru
- popis charakteristických čísel
- popis rozptylu
- popis tvaru rozdělení

Popisná statistika

- popis statistického souboru
- popis charakteristických čísel
- popis rozptylu
- popis tvaru rozdělení

Centrální tendence

Sřední hodnoty, umístění

- nevhodná slova: střední, středně, průměrně, rovnoměrnost (hodně, málo) - špatná orientace
- normálně proměřená populační data

Jak sdělovat dat zveřejňovat?

- srovnání populací rozdílných ukazatelů centrální tendence a ukazatele variability
- nejvíce hodnota, která by všechny změřené hodnoty dobře reprezentovala

Ukazatel centrální tendence

- = ukazatel středních hodnot, ukazatel míry polohy
- charakteristika typické hodnoty dat
- ukazuje, kde se na měřítku skáče (špatně ošetřena data mážež)
- popisuje rozložení číselů jedné proměnné

Ukazatel variability

- ukazuje, jak moc či málo jsou data rozložena

Ukazatele centrální tendence

- popisná statistika (špatná charakteristická proměnná)
- ukazatel středních hodnot
- udávají průměrnou, typickou, reprezentativní, očekávanou hodnotu - jeden údaj
- jedno číslo - krásné a zřádné

- modus
- medián
- aritmetický průměr

Modus

Modus

- nejvíce se vyskytující hodnota
- nejvíce výskyt
- nejvíce výskyt
- nejvíce výskyt
- nejvíce výskyt

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Medián

Medián

- polovina dat je menší než medián
- polovina dat je větší než medián
- polovina dat je rovna mediánu
- polovina dat je menší než medián
- polovina dat je větší než medián

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Aritmetický průměr

Aritmetický průměr

- střední hodnota
- střední hodnota
- střední hodnota
- střední hodnota
- střední hodnota

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Míry variability (rozptýlenosti)

Míry variability (rozptýlenosti)

- rozptýlení středních hodnot - udávají, jak široce jsou hodnoty rozloženy
- ukazují, jak široce jsou hodnoty rozloženy
- ukazují, jak široce jsou hodnoty rozloženy
- ukazují, jak široce jsou hodnoty rozloženy
- ukazují, jak široce jsou hodnoty rozloženy

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Příklad

Příklad:

Příklad

Příklad:

Popisné statistiky: Míry centrální tendence a variability

Mgr. Zuzana Szabó Lenhartová

Popisné statistiky

Popisné statistiky se zabývají popisem základních charakteristik dat. Jejich cílem je poskytnout stručný a srozumitelný přehled o datové sadě.

Popisné statistiky zahrnují:

- Ukazatele centrální tendence
- Míry variability
- Ukazatele tvaru rozdělení

Centrální tendence

Sřední hodnoty, umístění

- nevhodná slovní popis - grafické zobrazení -
- rozsáhlá množina dat - širší orientace -
- normálně proměřená data -

Jak zobrazit datové množství?

- např. graficky -
- tabulkou -
- tabulkou -
- tabulkou -

Ukazatele centrální tendence

- ukazatel středních hodnot, ukazatel míry polohy
- charakteristika typické hodnoty dat
- ukazuje, kde se na měřítku střed (střední osa) data nacházejí
- popisuje rozložení číselové jedné proměnné

Ukazatel variability

- ukazuje, jak moc či málo jsou data rozložena

Ukazatele centrální tendence

- popisná statistika (široká charakteristická proměnná)
- ukazatel středních hodnot
- ukazuje průměrnou, typickou, reprezentativní, očekávanou hodnotu - jeden údaj
- jedno číslo - krásné a zřetelné

- modus
- medián
- aritmetický průměr

Modus

nejvíce se vyskytující hodnota

Průběh modu

Průběh modu je hodnota, která se v datové sadě vyskytuje nejčastěji. Je to hodnota, která se opakuje nejvícekrát.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Modus je 10.

Medián

ukazuje, kde se nachází polovina dat

Průběh mediánu

Medián je hodnota, která dělí datovou sadu na dvě poloviny. Je to hodnota, která se nachází uprostřed datové sady.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Medián je 4.

Aritmetický průměr

ukazuje, jak moc se data liší od průměru

Průběh aritmetického průměru

Aritmetický průměr je hodnota, která je součtem všech hodnot datové sady dělenou počtem hodnot.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Aritmetický průměr je 6.

Míry variability (rozptýlenosti)

ukazuje, jak moc se data liší od průměru

Průběh míry variability

Míry variability ukazují, jak moc se data liší od průměru. Jsou to ukazatele, které měří rozptýlenost datové sady.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Míry variability jsou 10, 2, 10, 4, 10, 4, 4, 6, 10, 9.



Příklad:

10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Průměr: 6

Medián: 4

Modus: 10

Popisné statistiky

- jednorozměrný popis a analýza proměnných

Kategoriální deskriptivy

- entropie
- modus

Pořadové deskriptivy

- medián
- kvartily
- percentily
- grafické zobrazení pomocí pořadových deskriptiv - boxplot

Odchylkové, momentové deskriptivy

- aritmetický průměr
- rozptyl, směrodatná odchylka
- zešíkmení
- špičatost

Centrální tendence

Střední hodnoty, umístění

- nevýhoda tabulky četností i grafického zobrazení - neúspornost (hodně čísel -> špatná orientace)
- nemůžeme proměnnou popsat rychle

Jak zobrazení dat zredukovat?

- úsporně popsat rozložení proměnných skrze ukazatele centrální tendence a ukazatele variability
- najít hodnotu, která by všechny naměřené hodnoty dobře reprezentovala

Ukazatel centrální tendence

- = ukazatel středních hodnot; ukazatel míry polohy
- charakteristika typické hodnoty dat
- ukazuje, kde se na měřené škále (číselné ose) data nalézají
- popisuje rozložení četností jedné proměnné

Ukazatel variability

- udává, jak moc či málo jsou data na škále rozptýlená

Ukazatele centrální tendence

- popisná statistika (číselná charakteristika proměnné)
- ukazatel středních hodnot
- udávají průměrnou, typickou, reprezentativní, očekávanou hodnotu - jeden údaj
- jedno číslo - krásné a zrádné
 - **modus**
 - **medián**
 - **aritmetický průměr**

Modus \hat{X}, Mo

- kategoriální typická hodnota
- **nejčastější hodnota**
(**hodnota s nejvyšší četností v datech**)
- jediná možnost u nominálních dat, u vyšších úrovní často užitečnou volbou
- když známe všechny naměřené hodnoty, stanovíme modus tak, že zjistíme, která hodnota se v daném souboru vyskytuje nejčastěji

Příklad: 14, 3, 18, 4, 8, 18, 4, 6, 8, 10, 8

- v případě tabulky četnosti s intervaly lze modus určit přibližně jako střed intervalu s největší četností
- *nezávislý na extrémních hodnotách naměřené veličiny*
 - modus nemusí být určen jednoznačně - se stejnou nejvyšší frekvencí se může vyskytovat více hodnot
 - rozdělení s jedním modem (vrcholem) - unimodální
 - rozdělení pravděpodobnosti s dvěma vrcholy - dvouvrcholová (bimodální).

Medián

\tilde{X}, Md

- pořadová střední hodnota
- prostřední hodnota z řady hodnot seřazených podle velikosti
- 50. percentil - rozděluje soubor dat na dvě stejné části
- při sudém počtu prvků je mediánem průměr ze dvou prostředních hodnot/ kterékoli číslo z intervalu mezi nejbližší vyšší a nejbližší nižší hodnotou (konsensuálně střed intervalu)
- používáme pro (ordinální) pořadová data a výše
- nezávislý na extrémních hodnotách měřené veličiny

Příklad: Měření vědomostí žáků didaktickým testem, výsledky:

14, 3, 18, 4, 8, 18, 4, 6, 8, 10, 8

Aritmetický průměr

- deviační, odchylková, momentová střední hodnota
- jak ho znáte ze školy:

součet všech naměřených údajů vydělený jejich počtem

- používáme pouze pro intervalová a poměrová data
- nevýhoda: velmi citlivý na extrémní hodnoty

Příklad: 1,3,6,8,9,9,10,10,10

Příklad:



	s 15	bez 15
Průměr	2,48	2,05
Medián	1,00	1,00
Modus	0	0

Příklad:

Určete průměr, medián a modus u těchto čtyř rozložení (sad dat):

- a. 3, 3, 4, 5, 6, 8, 8, 8, 9
- b. 2, 4, 4, 4, 6, 7, 7
- c. 7, 7, 8, 9, 10, 10, 10
- d. 1, 1, 3, 4, 5, 9

Míry variability (rozptýlenosti)

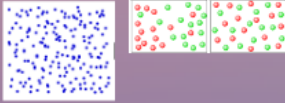
- omezenost středních hodnot - udávají pouze to, kolem jaké hodnoty se data "centrují" - které jsou nejčastější
- data se stejnou střední hodnotou mohou mít různou rozptýlenost

Variabilita - jak moc či málo jsou data na škále rozptýlená

- malá variabilita - většina hodnot v souboru je stejných nebo velmi blízkých
- vysoká variabilita - hodnoty jsou velmi rozmanité
- tři ukazatelé variability (podle škál)

Na nominální škále: Entropie

- veličina udávající "míru neuspořádanosti" zkoumaného systému
- míra neurčitosti systému



- v sociálních vědách se moc nepoužívá
- Pokud ano, tak:
- **variální poměr** či **nominální variance**

Na pořadové škále:

Variační rozpětí

- pokud můžeme seřadit hodnoty od nejmenší po největší a můžeme říct, co je minimum a co je maximum, máme rozpětí

$$R = X_{\max} - X_{\min}$$

- extrémně roste s velikostí vzorku - čím větší soubor, tím větší hodnota rozpětí
- nevýhoda: Vysoká citlivost vůči outlierům

Příklad: 2, 8, 9, 10, 1, 0, 5

Interkvartilové rozpětí

- vzdálenost mezi dvěma body na škále, které jsou na nějakém místě, které můžeme snadno definovat - používá se 25. a 75. percentil

$$- Q = Q_3 - Q_1 \text{ (75. percentil minus 25. percentil)}$$

- používáme spíše než jednoduché variační rozpětí

Na intervalové, poměrové škále

- charakteristiky založené na odchylkách od průměru
- měří rozptýlenost dat kolem aritmetického průměru

Rozptyl

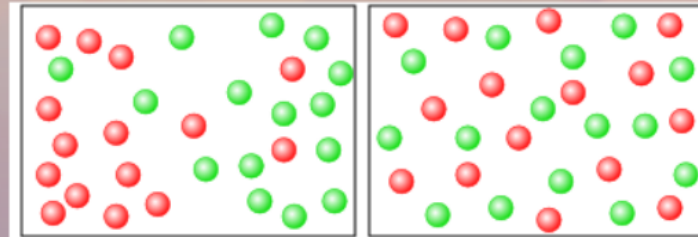
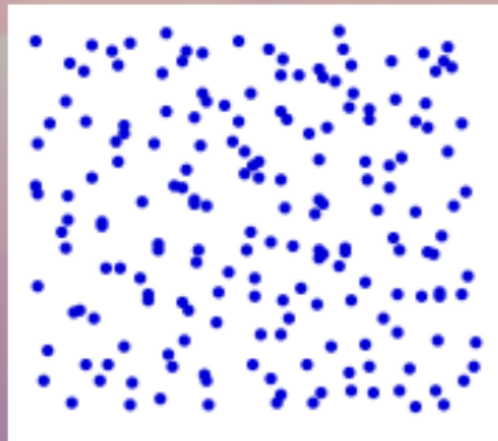
- aritmetický průměr čtverců odchylek od aritmetického průměru (průměrná kvadratická odchylka měření od aritmetického průměru, přičemž při průměrování této odchylky dělíme číslem $(n-1)$) = průměrná odchylka na druhou
- populační rozptyl: $(\sum x^2 / n)$
- výběrový rozptyl - vhodnější: $(\sum x^2 / (n-1))$ - při počítání pro všechny prvky populace součet odchylek na druhou = suma čtverců (sečtu odchylky od průměru na druhou)
- používá se v inferenční statistice

Směrodatná odchylka

- standardní odchylka
- odmocnina rozptylu - návrat k původní jednotce, ve které měříme

Na nominální škále: Entropie

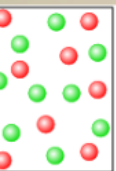
- veličina udávající "míru neuspořádanosti" zkoumaného systému
- míra neurčitosti systému



- v sociálních vědách se moc nepoužívá

Pokud ano, tak:

- **variační poměr** či **nominální variance**



Na pořadové škále:

Variační rozpětí

- pokud můžeme seřadit hodnoty od nejmenší po největší a můžeme říct, co je minimum a co je maximum, máme rozpětí

- $R = X_{\max} - X_{\min}$

- extrémně roste s velikostí vzorku - čím větší soubor, tím větší hodnota rozpětí

- nevýhoda: Vysoká citlivost vůči outlierům

Příklad: 2, 8, 9, 10, 1, 0, 5

Interkvartilové rozpětí

- vzdálenost mezi dvěma body na škále, které jsou na nějakém místě, které můžeme snadno definovat - používá se 25. a 75. percentil

- $Q = Q_3 - Q_1$ (75. percentil minus 25. percentil)

- používáme spíše než jednoduché variační rozpětí

- a
(p
při
= p
- p
- v
prv
so
na
- p

Na intervalové, poměrové škále

- charakteristiky založené na odchylkách od průměru
- měří rozptýlenost dat kolem aritmetického průměru

Rozptyl

- aritmetický průměr čtverců odchylek od aritmetického průměru (průměrná kvadratická odchylka měření od aritmetického průměru, přičemž při průměrování této odchylky dělíme číslem $(n-1)$)
= průměrná odchylka na druhou
- populační rozptyl: $(\sum x^2 / n)$
- výběrový rozptyl - vhodnější: $(\sum x^2 / (n - 1))$ - při počítání pro všechny prvky populace
součet odchylek na druhou = **suma čtverců** (sečtu odchylky od průměru na druhou)
- používá se v inferenční statistice

Směrodatná odchylka

- standardní odchylka
- odmocnina rozptylu - návrat k původní jednotce, ve které měříme

Popisné statistiky: Míry centrální tendence a variability

Mgr. Zuzana Szabó Lenhartová

Popisné statistiky

Popisné statistiky jsou soubor metod a postupů, které umožňují získat přehled o základních vlastnostech dat. Jejich cílem je poskytnout stručnou a srozumitelnou charakteristiku datového souboru.

Popisné statistiky zahrnují:

- Ukazuje, jak data vypadají (graficky)
- Ukazuje, jak data se chová (statisticky)
- Ukazuje, jak data se mění (dynamicky)
- Ukazuje, jak data se liší (variabilita)
- Ukazuje, jak data se souvisejí (korelace)

Centrální tendence

Sřední hodnoty, umístění

- nevhodná slova (číslo) - aritmetická průměr
- rozsáhlá (druhá část) - špatná orientace
- normálně proměnná (první část)

Jak odhadnout dat zjednodušen?

- například popisná statistika (popisná statistika)
- tendence ukazatele variability
- nejí hodnoty, která by všechny rozložení hodnoty dobře reprezentovala

Ukazuje centrální tendence

- ukazuje středních hodnot, ukazatel míry polohy
- charakteristika typické hodnoty dat
- ukazuje, kde se má střední šikma (špatná osa) data nachází
- popisuje rozložení číselní jedné proměnné

Ukazuje variabilitu

- ukazuje, jak moc či málo jsou data rozložena

Ukazuje centrální tendence

- popisná statistika (špatná charakteristická proměnná)
- ukazuje středních hodnot
- ukazuje průměrnou, typickou, reprezentativní, očekávanou hodnotu - jeden údaj
- jedno číslo - krásné a zřetelné

- modus
- medián
- aritmetický průměr

Modus

Modus je hodnota, která se v datovém souboru vyskytuje nejčastěji. Je to nejvíce výskytová hodnota. Modus je vhodný pro popisné statistiky, protože ukazuje, jak často se určitá hodnota vyskytuje.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Modus je 10, protože se vyskytuje nejčastěji (5krát).

Medián

Medián je hodnota, která dělí datový soubor na dvě poloviny. Je to střední hodnota. Medián je vhodný pro popisné statistiky, protože ukazuje, kde se nachází polovina dat.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Medián je 4, protože se nachází uprostřed datového souboru.

Aritmetický průměr

Aritmetický průměr je součet všech hodnot v datovém souboru dělený počtem hodnot. Je to nejčastěji používaný ukazatel centrální tendence. Aritmetický průměr je vhodný pro popisné statistiky, protože ukazuje, jak se hodnoty v datovém souboru chová.

Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9

Aritmetický průměr je 6, protože součet všech hodnot (60) dělený počtem hodnot (10) je 6.

Míry variability (rozptýlenosti)

Míry variability ukazují, jak rozptýlená jsou data. Jsou to ukazatele, které pomáhají pochopit, jak moc se hodnoty v datovém souboru liší od sebe. Míry variability jsou důležité pro popisné statistiky, protože pomáhají pochopit, jak moc data jsou rozptýlená.

Ukazuje, jak moc či málo jsou data rozložena

- ukazuje, jak moc či málo jsou data rozložena
- ukazuje, jak moc či málo jsou data rozložena
- ukazuje, jak moc či málo jsou data rozložena



Příklad:

Ukazuje, jak moc či málo jsou data rozložena. Příklad: 10, 2, 10, 4, 10, 4, 4, 6, 10, 9