

STATISTIKA 2

Statistika a statistické zpracování dat

Blok 1-1

Jiří Fišer

8. listopadu 2024

Vítejte ve světě statistiky

Předmět *Statistické zpracování dat* je určen pro studenty navazujícího studia.

- Naučíte se základní teoretické pojmy a koncepty.
- Zaměříme se na praktické aplikace nezbytné pro analýzu a zpracování dat.
- Předmět vám přinese užitečné dovednosti do praxe v oblastech jako ekonomie, management a marketing.

Struktura skript

Kapitoly jsou navrženy s logickou návazností:

- **Základní pojmy statistiky** – Přehled klíčových pojmů jako náhodný jev, náhodná veličina a pravděpodobnost.
- **Zpracování dat** – Postupy pro analýzu dat z reálného světa, využití popisné statistiky.
- **Matematická statistika** – Odhady parametrů, intervalové odhady a závěry o populaci.
- **Testování hypotéz** – Parametrické a neparametrické testy jako základ každého výzkumu.
- **Regresní a korelační analýza** – Predikce a modelování vztahů mezi proměnnými.

Co vás v kapitolách čeká

Každá kapitola obsahuje:

- **Teoretický výklad** – Základní pojmy a metody.
- **Řešené příklady** – Praktické příklady pro procvičení.
- **Rámečky** – Důležité body k zapamatování.
- **Shrnutí** – Připomenutí hlavních bodů.
- **Kontrolní otázky** – Pomocí otázek si ověříte porozumění látce.

Praktická aplikace a software

- Statistika je klíčová pro analýzu dat v ekonomii a managementu.
- Důraz na praktické úlohy a výpočty.
- Statistický software jako Excel vám usnadní řešení úloh a získá klíčové dovednosti pro praxi.

Motivace a podpora

- Statistika jako výzva a nástroj pro lepší porozumění datům.
- Příklady připraví na reálné situace.
- Chceme vás podpořit, aby statistika nebyla překážkou, ale užitečným nástrojem.

Základní statistické pojmy - Cíle kapitoly

Cíle:

- Představit základní principy a historii statistiky
- Rozlišit mezi deskriptivní a inferenční statistikou
- Definovat základní pojmy: populace, výběr, typy proměnných
- Porozumět měřítkům proměnných, centrální tendenci, variabilitě
- Rozlišit korelaci a kauzalitu

Úvod do statistiky

Statistika: věda o sběru, analýze a interpretaci dat.

Použití v rozhodování, analýze trendů, predikcích.

Dva hlavní typy statistiky:

- **Popisná statistika:** sumarizace dat, grafy, výpočty (např. průměr, medián)
- **Inferenční statistika:** závěry o populaci na základě vzorků

Populace a výběr

Populace: celkový soubor jednotek, o kterém zjišťujeme informace (např. obyvatelé, výrobky).

Výběr (vzorek): podmnožina populace pro analýzu.

- **Náhodný výběr:** stejná šance pro všechny členy
- **Výběrová chyba:** rozdíl mezi výběrovými výsledky a celkovou populací

Typy proměnných

Kvalitativní (kategorické): nelze číselně měřit, mají kategorie.

- **Nominální:** bez přirozeného pořadí (např. barvy)
- **Ordinální:** s přirozeným pořadím (např. spokojenost: nízká, střední, vysoká)

Kvantitativní (numerické): měřitelné číselně.

- **Diskrétní:** určité hodnoty, obvykle celá čísla (např. počet dětí)
- **Spojitý:** libovolné hodnoty v intervalu (např. výška)

Měřítko měření proměnných

- **Nominální škála:** kategorie bez pořadí (např. barvy)
- **Ordinální škála:** kategorie s pořadím (např. spokojenost: nízká, střední, vysoká)
- **Intervalová škála:** hodnoty s pořadím, bez absolutní nuly (např. teplota v °C)
- **Poměrová škála:** hodnoty s absolutní nulou (např. délka, váha)

Příklad: Typy proměnných

Tabulka dat o firmách:

Firma	Počet zaměstnanců	Roční obrat (mil.)	Obor
A	120	45.6	IT
B	300	120.8	Stavebnictví
C	50	15.2	Obchod
D	450	220.5	IT
E	90	30.1	Zdravotnictví

Interpretace:

- **Firma:** nominální proměnná
- **Počet zaměstnanců:** kvantitativní diskrétní
- **Roční obrat:** kvantitativní spojitá
- **Obor:** nominální proměnná

Popisná statistika

- Popisná statistika se zaměřuje
 - ▶ na sumarizaci
 - ▶ a popis základních charakteristik dat.
- Pomáhá nám porozumět tomu,
 - ▶ jaká data máme k dispozici,
 - ▶ a nabízí jednoduché nástroje pro jejich prezentaci.

Míry centrální tendence

Míry centrální tendence popisují střední hodnotu datového souboru:

- **Průměr:** Součet všech hodnot dělený počtem hodnot, citlivý na extrémní hodnoty.
- **Medián:** Prostřední hodnota v seřazeném datovém souboru, méně citlivý na extrémy.
- **Modus:** Hodnota, která se vyskytuje nejčastěji, použitelný i pro kvalitativní proměnné.

Míry variability

Míry variability udávají, jak se hodnoty v datovém souboru od sebe liší:

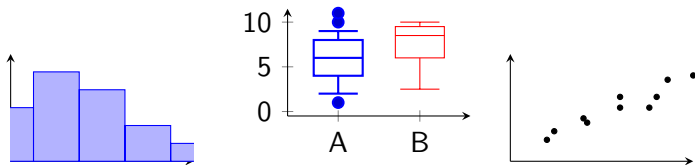
- **Rozptyl:** Průměrná čtvercová odchylka hodnot od průměru.
- **Směrodatná odchylka:** Odmocnina z rozptylu, měří průměrnou odchylku od průměru.
- **Variační koeficient:** Poměr směrodatné odchylky k průměru, umožňuje porovnání variability mezi datovými soubory.

Distribuce dat

- Distribuce dat popisuje
 - ▶ četnost výskytu jednotlivých hodnot v datovém souboru.
- Grafické a numerické popisy distribuce nám pomáhají pochopit
 - ▶ tvar,
 - ▶ centrální tendenci
 - ▶ a rozptyl dat.

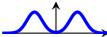
Grafické znázornění distribuce

- **Histogram:** Ukazuje četnost hodnot v intervalech, vhodný pro kvantitativní data.
- **Krabicový graf (boxplot):** Ukazuje rozložení dat pomocí pěti čísel, identifikuje odlehlé hodnoty.
- **Bodový diagram (scatter plot):** Zobrazuje vztah mezi dvěma proměnnými, užitečný pro analýzu korelace.



Obrázek: Histogram, krabicový diagram (boxplot) a bodový graf (scatterplot)

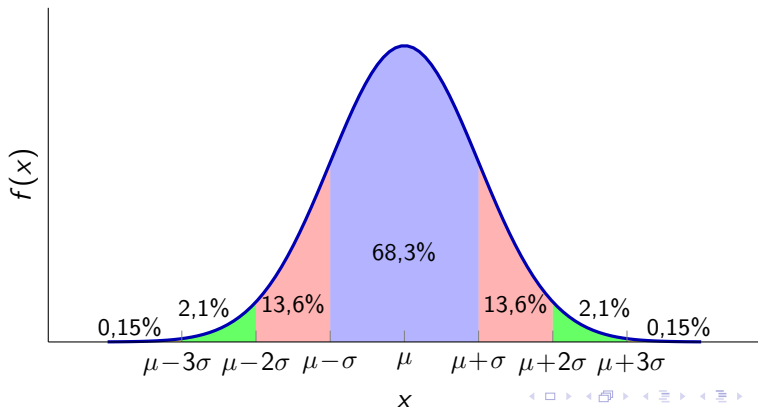
Tvar distribuce

- **Symetrická distribuce:**  Data jsou symetrická kolem centrální hodnoty (např. normální rozdělení).
- **Pravostranně šikmá:**  Dlouhý pravý ocas, většina dat vlevo.
- **Levostranně šikmá:**  Dlouhý levý ocas, většina dat vpravo.
- **Bimodální:**  Dva vrcholy, data mohou pocházet ze dvou skupin.
- **Špičatost (kurtosis):** Míra „ostrého“ nebo „plochého“ vrcholu.

Normální rozdělení

Normální rozdělení (Gaussovo) má charakteristický zvonovitý tvar:

- Symetrie kolem průměru.
- Průměr, medián a modus jsou stejné.
- Přibližně 68% hodnot je v intervalu jedné směrodatné odchylky od průměru, 95% v intervalu dvou a 99,7% ve třech směrodatných odchylkách.



Základy pravděpodobnosti

- Pravděpodobnost je nástroj, který nám
 - ▶ pomáhá kvantifikovat nejistotu.
- V rámci statistiky je klíčová
 - ▶ pro odhadování výsledků a rozhodování na základě dostupných dat.

Definice pravděpodobnosti

Pravděpodobnost události vyjadřuje, jak často bychom očekávali, že tato událost nastane při opakovaném experimentu za stejných podmínek.

$$P(A) \in [0, 1]$$

- $P(A) = 1$: událost A nastane jistě
- $P(A) = 0$: událost A nenastane nikdy

Základní pravidla pravděpodobnosti

- **Pravděpodobnost komplementu:** $P(\neg A) = 1 - P(A)$
- **Pravděpodobnost sjednocení:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Podmíněná pravděpodobnost:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Zákon velkých čísel

Zákon velkých čísel vysvětluje, že

- průměr výsledků z mnoha pokusů se blíží očekávané hodnotě.
- Čím více pokusů, tím bližší průměr skutečné hodnotě.

Příklad:

- Při házení mincí je pravděpodobnost panny 0,5.
- Pokud hodíme mincí jen desetkrát, může podíl pannen vybočovat,
- ale při tisíci hodech se podíl přiblíží 50 %.
- Tento zákon umožňuje v inferenční statistice odhadovat parametry populace na základě velkého vzorku.

Centrální limitní věta

Centrální limitní věta říká, že

- průměry vzorků z libovolné populace budou mít přibližně normální rozdělení,
- pokud je vzorek dostatečně velký.

Příklad:

- I když populace (např. věk lidí)
 - ▶ není normálně rozdělena,
- průměry dostatečně velkých vzorků z této populace
 - ▶ budou mít normální rozdělení.
- Centrální limitní věta umožňuje používat normální rozdělení při analýze dat, což je základ pro *testování hypotéz* a výpočet *intervalů spolehlivosti*.

Testování hypotéz (úvod)

- Testování hypotéz je
 - ▶ klíčovou metodou inferenční statistiky,
- která umožňuje rozhodnout, zda existuje dostatek důkazů
 - ▶ k zamítnutí nebo přijetí určitého tvrzení o populaci
 - ▶ na základě vzorku dat.

Základní pojmy v testování hypotéz

- **Nulová hypotéza (H_0):** Mezi proměnnými nebo skupinami neexistuje žádný vztah nebo rozdíl. Testujeme tuto hypotézu a snažíme se ji zamítnout.
- **Alternativní hypotéza (H_A):** Opačná hypotéza, tvrdí, že mezi proměnnými existuje vztah nebo rozdíl mezi skupinami.

Kroky při testování hypotéz

Testování hypotéz obvykle zahrnuje následující kroky:

- ➊ **Stanovení hypotéz:** Formulace nulové a alternativní hypotézy.
- ➋ **Výběr testu:** Volba vhodného statistického testu (např. t-test, chi-kvadrát test).
- ➌ **Stanovení hladiny významnosti (α):** Typicky 0,05 (5% šance na chybu I. druhu).
- ➍ **Výpočet testovací statistiky:** Vypočítání hodnoty statistiky ze vzorku dat.
- ➎ **Rozhodnutí:** Porovnání testovací statistiky s kritickou hodnotou nebo použití p-hodnoty k rozhodnutí o zamítnutí či přijetí H_0 .

Chyby při testování hypotéz

Existují dva druhy chyb:

- **Chyba I. druhu:** Zamítnutí pravdivé nulové hypotézy (falešně pozitivní).
- **Chyba II. druhu:** Nepřijetí nepravdivé nulové hypotézy (falešně negativní).

Rozhodnutí je vždy činěno na základě dat a existuje riziko chyb.

Příklad: Chyby při testování hypotéz

Testujeme účinnost nového léku:

- H_0 : Lék nemá účinek.
- H_A : Lék má účinek.
- **Chyba I. druhu:** Zamítnutí H_0 , tvrzení o účinnosti léku, přestože účinný není.
- **Chyba II. druhu:** Nepřijetí H_A , tvrzení o neúčinnosti léku, přestože účinný je.

Cílem je minimalizovat pravděpodobnost obou chyb, ale riziko chyb vždy existuje.

Intervaly spolehlivosti

- Interval spolehlivosti (confidence interval) je interval, který
 - ▶ s určitou pravděpodobností obsahuje skutečnou hodnotu parametru populace.
- Poskytuje informace
 - ▶ o bodovém odhadu parametru
 - ▶ a přesnosti tohoto odhadu.

Definice intervalu spolehlivosti

Interval spolehlivosti je založen na bodovém odhadu parametru:

$$CI = [\hat{\theta} - \Delta, \hat{\theta} + \Delta]$$

kde:

- $\hat{\theta}$ je bodový odhad (např. průměr),
- Δ je poloměr intervalu pro zvolenou úroveň spolehlivosti (např. pro 95%).

Úroveň spolehlivosti:

Nejběžnější úrovně spolehlivosti jsou 90%, 95% a 99%.

Interpretace intervalu spolehlivosti

Interval spolehlivosti udává možné hodnoty skutečného parametru.

- Příklad: Interval pro průměr $[45, 55]$ znamená, že s 95% jistotou leží skutečný průměr mezi 45 a 55.
- Vyšší spolehlivost vede k širším intervalům, nižší spolehlivost k užším intervalům.

Korelace a kauzalita

Korelace a kauzalita se týkají vztahu mezi dvěma proměnnými:

- **Korelace:** Měří sílu a směr lineárního vztahu.
- **Kauzalita:** Příčinný vztah, kdy změna jedné proměnné způsobuje změnu druhé.

Pearsonův korelační koeficient měří lineární vztah mezi dvěma proměnnými, nabývá hodnot od -1 do 1:

- Hodnota blízká 1: silná pozitivní korelace (obě proměnné rostou).
- Hodnota blízká -1: silná negativní korelace (jedna roste, druhá klesá).
- Hodnota blízká 0: žádný lineární vztah.

Kauzalita

Kauzalita znamená, že změna jedné proměnné způsobuje změnu druhé.

- Na rozdíl od korelace vyžaduje kauzalita další důkazy, aby bylo možné potvrdit příčinný vztah.

Korelace neznamená kauzalitu:

- Silná korelace mezi dvěma proměnnými neznamená, že jedna způsobuje změnu druhé.

Tento koncept je důležitý pro regresní analýzu a analýzu příčinných vztahů.

Historie a význam statistiky

Starověk:

- První formy statistiky se objevily před více než 4 000 lety v Číně a Egyptě, kde byly vedeny soupisy obyvatelstva a majetku.
- Tyto záznamy sloužily k efektivnějšímu výběru daní a organizaci vojenských sil.
- V antickém Řecku a Římě se statistika používala při správě měst a říší.

Středověk:

- Statistika byla zaměřena na administrativní a fiskální potřeby.
- V 11. století byl v Anglii zaveden „Domesday Book“, což byl katastrální soupis pro účely zdanění.

Historie a význam statistiky

18. století:

- Statistika se začala rozvíjet jako vědecký obor zaměřený na popis stavu státu.
- **Gottfried Achenwall** (1719–1772): zakladatel moderní statistiky, začal používat pojem „statistika“ v moderním smyslu jako věda o státu.

19. století:

- Statistika se rozvíjí jako matematická disciplína s teorií pravděpodobnosti.
- **Pierre-Simon Laplace** (1749–1827): rozvinul teorii pravděpodobnosti.
- **Carl Friedrich Gauss** (1777–1855): vyvinul metodu nejmenších čtverců a přispěl k normálnímu rozdělení.

20. století:

- **Sir Ronald A. Fisher** (1890–1962): přinesl inovace v návrhu experimentů, analýze rozptylu a teorii hypotéz.
- **Jerzy Neyman** (1894–1981) a **Egon Pearson** (1895–1980): vyvinuli Neyman-Pearsonovu teorii testování hypotéz.
- Rozvoj výpočetní techniky umožnil složité statistické analýzy.

Současnost:

- Statistika je klíčová v mnoha oblastech života – ve vědě, podnikání, správě, medicíně.
- S nárůstem big data je statistika nezbytná pro nacházení vzorců a trendů.
- Významní moderní statistici:
 - ▶ **John Tukey** (1915–2000): průzkumná analýza dat.
 - ▶ **William Feller** (1906–1970): teorie pravděpodobnosti a její aplikace.

Závěr - Historie a význam statistiky

- Historický vývoj ukazuje, jak statistika přerostla z nástroje pro správu státu na nezbytný vědecký obor pro rozhodování a analýzu.
- Díky práci mnoha významných osobností se statistika stala základním nástrojem ve vědě i praxi.