Úvod do teorie měření

Alena Černíková

alena.cernikova@ujep.cz

14. dubna 2026

Podmínky zápočtu

- **DVA domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat na univerzitní OneDrive – odkaz u mě na stránce
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  zpracování závislosti dvou proměnných
  ucelený text od výzkumné otázky až po interpretaci výsledku

Obsah kurzu

- Cíle výzkumu
- Typy proměnných
- Popis dat
- Pravděpodobnostní rozdělení
- Bodový vs intervalový odhad
- Základy testování
- Jednovýběrový, párový a dvouvýběrový test
- Analýza rozptylu
- Korelace
- Jednoduchá lineární regrese
- Chyby měření

Výuka

Výuka bude probíhat ve statistickém software R, v prostředí R Commander.

- volně stažitelný software – návod na stažení a instalaci v podkladech pro praktickou část
- většinu používaných metod je možné volit přes menu – snadná obsluha
- vyžaduje základní znalosti angličtiny

Data

Realizujeme měření / výzkum / pokus, jehož výsledkem jsou čísla nebo nebo i textové charakteristiky. Tyto informace – **data** uložíme do databáze, kterou následně načteme do statistického softwaru a analyzujeme.
Co nás zajímá?

- popis získaných dat
- zobecnění získaných výsledků na celou populaci

Příklad. *Zajímá nás jak se změní koncentrace látky A při přidání látky B do roztoku. Provedeme pokus v laboratoři - každý z 56 studentů udělá svůj pokus. Na základě výsledků chceme něco říci o tom, jak se obecně roztok chová.*

Data

Data, která jsme naměřili, se nazývají **výběr**. Chceme po nich, aby

- byly získány objektivně
- tvořily reprezentativní výběr
  pokud chceme obecné informace, nemůžeme dělat pokus vždy jen při vysokých teplotách
- jednotlivé hodnoty byly vzájemně nezávislé
  není dobré, aby se opakovaně používal ten samý roztok (vždy mícháme nový)

Pokud máme nezávislá data, získaná "náhodně", která tvoří reprezentativní výběr z celé populace, říkáme, že pracujeme s **náhodným výběrem**. Na jeho základě je možné výsledky zobecnit na celou populaci.

Data

Terminologie

- **Nahodná veličina** – cokoliv, co měříme a můžeme to měřit opkovaně, např. výška, koncentrace, úroveň vzdělání
- **Populace** – úplný soubor, pro nejž chceme udělat nějaký závěr, např. všichni dospělí obyvatelé České republiky
- **Náhodný výběr** – v porovnání s populací malý soubor pozorování, který tvoří nezávislé, stejně rozdělené náhodné veličiny, např. výběr 200 lidí
- **Populační charakteristika** – charakteristika popisující populaci, např. populační průměr
- **Výběrová charakteristika** – charakteristika spočítaná na výběru pomocí níž odhadujeme populační ekvivalent, např. výběrový průměr.

Značení

- X – náhodná veličina (případně Y, Z)
- n – počet pozorování
- i – index, pořadový nebo sčítací
- X_i – naměřené hodnoty
- $\sum_{i=1}^{n}$ – znak pro součet přes i od 1 do n
- $X_{(i)}$ – uspořádaná řada naměřených hodnot

Popisné statistiky

Problémy v datech – aneb co dělat když

- **Chybějící pozorování**
  snažíme se, aby jich bylo co nejméně,
  když jich je málo, tak pracujeme bez nich – většina statistických metod implementovaných v různých softwarech si s tím poradí
  je možné je doplnit na základě nějakého modelu (*imputation*)
- **Odlehlé hodnoty**
  kontrola, zda nedošlo k chybě měření
  pokud ne, tak z popisných statistik se většinou nevynechávají, ale je dobré zmínit, že se jedná o odlehlé hodnoty
  pro popis proměnné je pak lépe zvolit ukazatele necitlivé na odlehlé pozorování
  ze složitějších analýz se často vynechávají

Typy proměnných

Abychom správně určili, které charakteristiky máme pro proměnnou počítat, je třeba nejprve určit typ proměnné.

- **Číselné proměnné** – pr. výška, váha, věk, atd.
- **Kategorické proměnné** – pr. barva, kraj, povolání, nebo taky známka ve škole, číslo, které padne na kostce, atd.
- Kategorické proměnné se dále dělí na
  - **Nominální** – neuspořádané, př. barva, kraj
  - **Ordinální** – uspořádané, př. známka, číslo na kostce

Popisné statistiky

Jak popisujeme jednotlivé typy proměnných

- **Číselné proměnné**
  - popisné statistiky polohy – průměr, medián, vybrané percentily (kvartily, extrémy)
  - popisné statistiky variability – rozptyl, směrodatná odchylka, mezikvartilové rozpětí, koeficient variace
  - popisné statistiky tvaru rozdělení – šikmost, špičatost
  - grafické charakteritiky – krabicový graf, histogram
- **Nominální proměnné**
  - číselné charakteritiky – absolutní a relativní četnosti
  - grafické charakteristiky – sloupcový a koláčový graf
- **Ordinální proměnné**
  - lze použít jak průměr, medián atd.
  - pro malé počty kategorií i absolutní a relativní četnosti, plus kumulativní četnosti

Popisné statistiky pro číselné proměnné

Popisné statistiky polohy

- nejprve nás zajímá "úroveň"
- kolem jakých hodnot se naměřená data pohybují

Popisné statistiky variability

- cílem je zjistit, jak jsou data rozprostřena kolem střední hodnoty
- v podstatě se měří šířka intervalu, na němž jsou naměřená data
- musí nabývat kladných hodnot

Popisné statistiky tvaru rozdělení

- zajímá nás tvar histogramu
- většinou porovnáváme s normalitou (Gaussova křivka)

Popisné statistiky polohy

Příklad. *Uvažujme výsledky experimentu, kde studenti naměřili koncentraci látky A v roztoku. Experiment provádělo 10 studentů a máme tedy 10 hodnot:* $37, 42, 35, 63, 44, 35, 40, 39, 43, 41$. *Spočtěme průměr, medián, kvartily a extrémy.*

Jak vypočítat **průměr** z n hodnot značených $X_1, X_2, X_3, \ldots, X_n$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Jak vypočítat **medián**

- z uspořádané řady – hodnota prostřední podle velikosti (při lichém počtu pozorování), nebo průměr prostředních dvou (při sudém počtu pozorování)

Jak vypočítat **kvartily**

- z uspořádané řady – hodnoty v jedné a ve třech čtvrtinách (výpočet viz. dále)

Jak vypočítat **extrémy**

- minimum a maximum

Popisné statistiky polohy

Výpočet pro obecný p-tý percentil – vážený průměr dvou sousedních uspořádaných hodnot.
Označme

- p – číslo mezi 0 a 1, díl dat, které chcete p-tým percentilem oddělit
- k – pořadí hodnoty v uspořádané řadě (hodnoty $X_{(k)}$ a $X_{(k+1)}$ budeme průměrovat)
- q – koeficient, kterým se násobí uspořádané hodnoty do váženého průměru

$$p - ty\ percentil = (1 - q)X_{(k)} + qX_{(k+1)}$$

$$k = \lfloor 1 + (n - 1)p \rfloor$$

$$q = 1 + (n - 1)p - k$$

Grafické popisné statistiky

Pro popis číselné proměnné se používají 2 typy grafů

- **Krabicový graf**
  jsou v něm zobrazeny vybrané percentily (medián a kvartily), tykadla dosahují k nejvzdálenějšímu neodlehlému pozorování (odlehlé pozorování se vyznačují zvlášť)
  odlehlé pozorování je takové, které je od bližšího kvartilu dále než jeden a půl násobek mezikvartilového rozpětí $1.5(Q_3 - Q_1)$
- **Histogram**
  počet sloupců závisí na počtu pozorování
  nejčastěji se používá *Sturgesovo pravidlo*

$$k = 1 + 3.32 \log_{10}(n)$$

kde n je počet pozorování

Vážený průměr

V praxi je často používaný **vážený průměr**

$$\overline{X}_w = \frac{\sum_{i=1}^{k} X_i w_i}{\sum_{i=1}^{k} w_i}$$

kde

- X_i jsou hodnoty
- w_i jsou váhy

Příklad. *Spočtete průměrnou známku u termínu zkoušky z matematiky, když víte, že 5 studentů dostalo 1, 7 studentů dostalo 2 a 13 studentů dostalo 3.*

Příklad

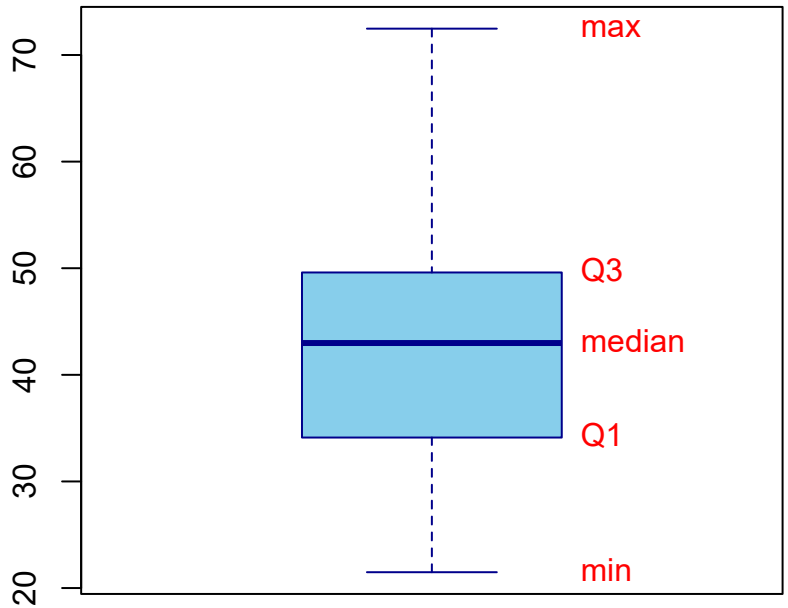
Spočtěte popisné statistiky polohy pro koncentraci látky *B* z databáze `ChemData.RData`.

```
> numSummary(ChemData[,"B", drop=FALSE], statistics=c("mean", "quantiles"),
      quantiles=c(0,.25,.5,.75,1))
mean       0%       25%      50%      75%     100%      n
43.88686 21.48576 34.13538 42.98286 49.43253 72.47179 56
```

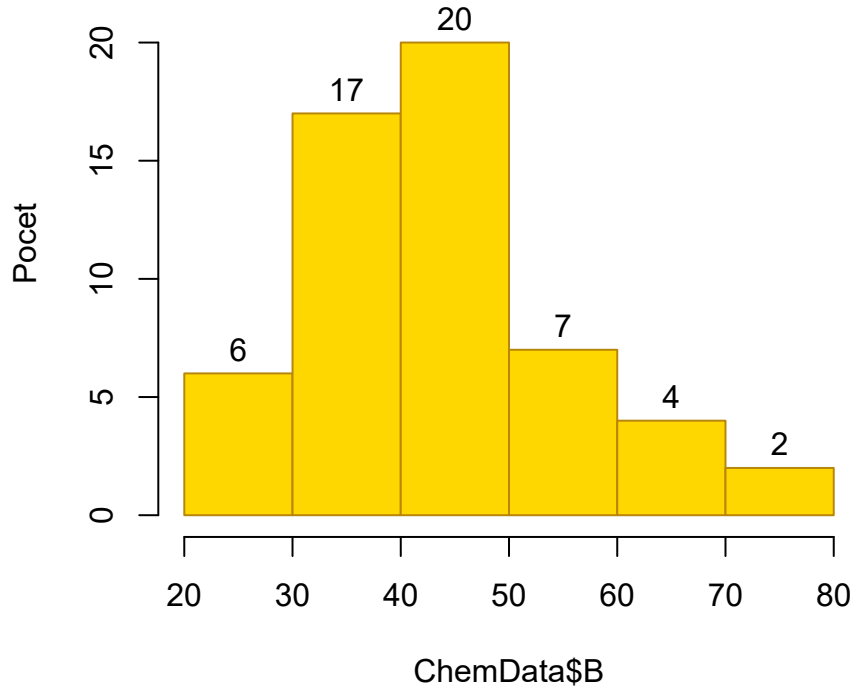
Výsledky

- počet pozorování (n): 56
- průměr (mean): 43.89
- medián (50%): 42.98
- kvartily (25% a 75%): 34.14, 49.43
- extrémy (0% a 100%): 21.49, 72.47

Grafické popisné statistiky



Krabicový graf



Histogram

Popisné statistiky variability

- Rozptyl a směrodatná odchylka

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}, \qquad \mathrm{sd}(X) = \sqrt{\mathrm{Var}X}$$

- Mezikvartilové rozpětí

$$IQR(X) = Q_3 - Q_1$$

kde Q_3 je třetí kvartil a Q_1 je první kvartil

- Variační koeficient

$$\mathrm{cv}(X) = \frac{\mathrm{sd}(X)}{\overline{X}}$$

Popisné statistiky variability

Rozptyl a **směrodatná odchylka**

- něco jako průměrná odchylka od průměru
- v intervalu průměr $\pm 3 \times$ směrodatná odchylka leží víc než 99% hodnot

Mezikvartilové rozpětí

- šířka krabice v boxplotu
- šířka prostřední poloviny dat

Variační koeficient

- kolik procent z průměru zabírá směrodatná odchylka
- používá se pro srovnání variability mezi různými proměnnými

Popisné statistiky tvaru rozdělení

- dvě charakteristiky popisující symetrii (**šikmost**) a relativní rozložení (**špičatost**) dat
- obě se počítají přes standardizované proměnné, tak zvané **Z-skóry**

$$Z_i = \frac{X_i - \overline{X}}{\mathrm{sd}(X)}$$

- standardizace vzhledem k průměru (0) a ke směrodatné odchylce (1)
- hodnoty kolem nuly většinou v intervalu $< -3; 3 >$

Popisné statistiky tvaru rozdělení

- **Šikmost** – průměr ze třetích mocnin z-skórů

$$\text{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\text{sd}(X)}\right)^3 = \frac{\sum_{i=1}^{n} Z_i^3}{n}$$

- **Špičatost** – průměr ze čtvrtých mocnin z-skórů mínus 3

$$\text{Kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\text{sd}(X)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

Popisné statistiky tvaru rozdělení

Šikmost

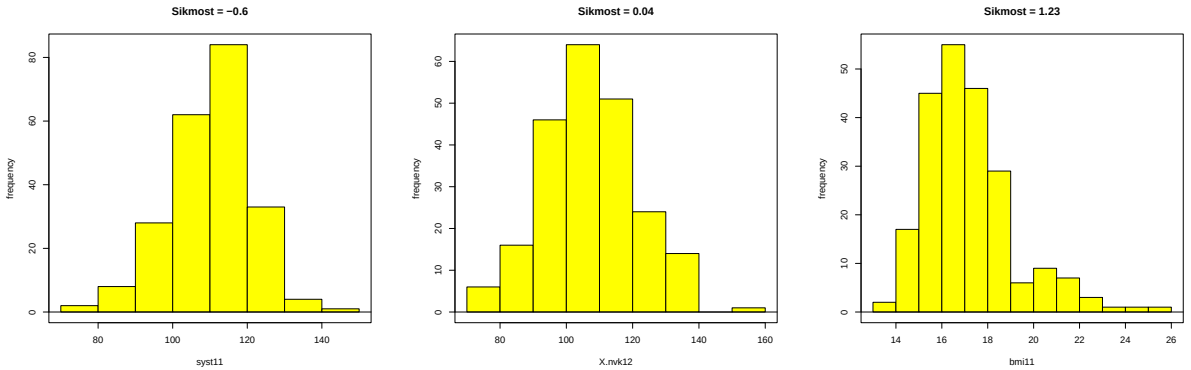
- měří symetrii rozdělení
- pro symetrické (např. normální) rozdělení je šikmost = 0
- histogram pro šikmost v intervalu $(-0.3; 0.3)$ se jeví symetricky

Špičatost

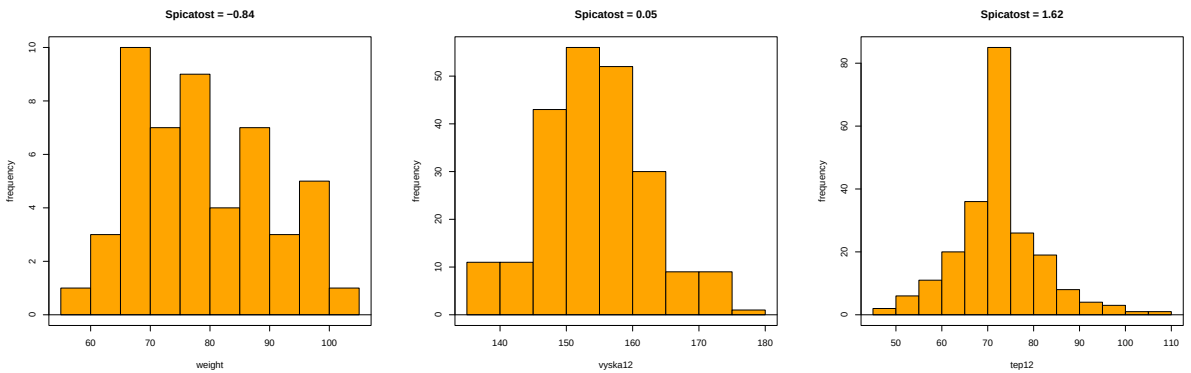
- měří špičatost v porovnání s normálním rozdělením
- normální rozdělení má špičatost rovnu 0
- odchylky od nuly bývají vidět jen v symetrických rozděleních pro hodnoty mimo interval $(-0.5; 0.5)$

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



Ukázka záporné, nulové (špičatost normálního rozdělení) a kladné špičatosti



Příklad

Spočtěte popisné statistiky variability a tvaru rozdělení pro koncentraci látky *B* z databáze `ChemData.RData`.

```
> numSummary(ChemData[,"B", drop=FALSE], statistics=c("mean", "sd", "se(mean)", "
      IQR", "cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="3")
mean       sd      se(mean)   IQR      cv         skewness   kurtosis  n
43.88686 12.0739 1.613443 15.29715 0.2751143 0.5399053 -0.302757 56
```

Výsledky

- počet pozorování (n): 56
- směrodatná odchylka (sd): 12.07
- střední chyba průměru (se(mean)): 1.61
- mezikvartilové rozpětí (IQR): 15.3
- koeficient variace (cv): 0.275
- šikmost (skewness): 0.54
- špičatost (kurtosis): -0.30

Příklad

Interpretace výsledků

- střední chyba průměru udává, do jaké míry je výběrový průměr dobrým odhadem průměru teoretického (je to velikost chyby)
- směrodatná odchylka vychází 12.07, což je 27.5% velikosti průměru (hodnota koeficientu variace)
- šikmost 0.54 říká, že máme mírně sešikmená data s protažením doprava
- špičatost -0.3 říká, že špičatost dat je srovnatelná s normálním rozdělením, respektive že jsou mírně složší než normální rozdělení

Číselné popisné statistiky

Příklad. *Mějme náhodný výběr 10-ti dospělých lidí, u kterých jsme zjišťovali barvu očí. Ve výběru jsme rozlišovali 3 barvy: modrá (M), hnědá (H) a zelená (Z). Zjistili jsme následující barvy M, M, Z, H, H, H, M, Z, M, H. Popišme zjištěné výsledky.*

Tabulka absolutních a relativních četností.

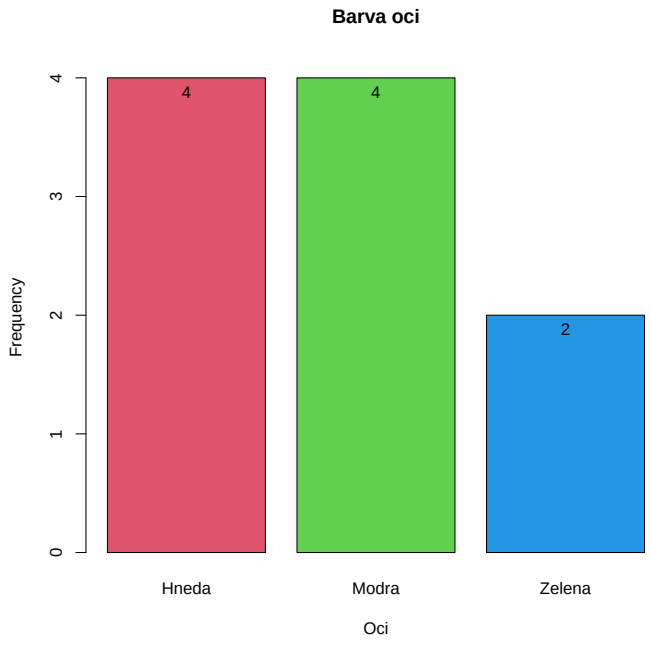
Barva	Absolutní	Relativní %
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

Jak vypočítat **relativní četnost?**
Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

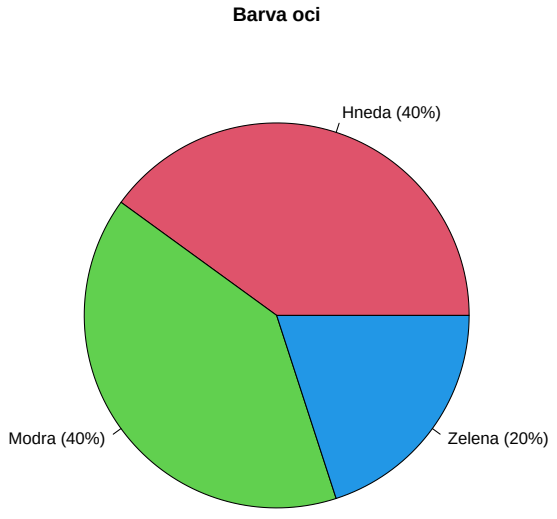
$$p_j = \frac{n_j}{n}$$

Grafické popisné statistiky

Graficky je možné znázornit jak absolutní počty, tak procenta



Sloupcový graf



Koláčový graf

Příklad

Spočtěte popisné statistiky pro barvu roztoku z databáze `ChemData.RData`

```
> local({
      +  .Table <- with(ChemData, table(barva))
      +  cat("\ncounts:\n")
      +  print(.Table)
      +  cat("\npercentages:\n")
      +  print(round(100*.Table/sum(.Table), 2))
      + })

counts:
barva
cervena ruzova fialova
   12      23      21

percentages:
barva
cervena ruzova fialova
  21.43  41.07   37.50
```

Příklad

Výsledky

- nejvíce vzorků (23) se zbarvilo do růžova, celkem 42%
- nejméně vzorků (12) se zbarvilo červeně, celkem 21.43%
- kumulativní četnosti R-ko nepočítá
  - červená: 12 – počet vzorků s červenou barvou
  - růžová: 35 – počet vzorků s červenou nebo růžovou barvou
  - fialová: 56 – počet vzorků s červenou nebo růžovou nebo fialovou barvou

Náhodné jevy

Jak může dopadnout náhodný pokus.

- **Náhodný pokus** – pokus konaný za přesně daných podmínek, o němž není dopředu známo jak dopadne
  Př. hod kostkou, měření výšky lidí, výsledek studenta u zkoušky
- **Náhodný jev** – možný výsledek náhodného pokusu
  Př. na kostce padne sudé číslo, výška člověka bude větší než 170 cm, student zkoušku udělá
- **Elementární jev** – nejmenší možné náhodné jevy, které nemohou nastat současně, ale musí nastat vždy alespoň jeden z nich
  Př. na kostce padne 1, 2, 3, 4, 5 nebo 6, výška člověka bude 160 cm, student zkoušku udělá nebo neudělá
- Součet všech elementárních jevů je prostor všech možných výsledků náhodného pokusu

Pravděpodobnostní rozdělení

Náhodné jevy

- **Jev jistý** Ω – soubor všech elementárních jevů, tj. celý prostor možných výsledků, $P(\Omega) = 1$
  Př. na kostce padne číslo od jedné do šesti
- **Jev nemožný** $\emptyset$ – jev, který neobsahuje ani jeden elementární jev, $P(\emptyset) = 0$
  Př. na kostce padne mínus jedna
- **Jev opačný** k jevu A, tj. $\overline{A}$ – soubor elementárních jevů, které nastanou právě když nenastane jev A,
  Př. na kostce padne sudé číslo, a na kostce padne liché číslo
- **Neslučitelné jevy** – jevy A a B jsou neslučitelné, když mají prázdný průnik
  Př. na kostce padne sudé číslo, a na kostce padne 1
- **Podjev** – jev A je podjevem jevu B, když je jeho částí
  Př. na kostce padne liché číslo a na kostce padne 3

Pravděpodobnostní rozdělení

Pravděpodobnost je funkce, která náhodnému jevu A přiřadí hodnotu mezi 0 a 1. Značíme ji $P(A)$.

- pravděpodobnost nemožného jevu $P(\emptyset) = 0$
- pravděpodobnost jistého jevu $P(\Omega) = 1$
- jsou-li A a B dva náhodné jevy, pro něž platí, že $A \subset B$, pak $P(A) \leq P(B)$
- pro každé dva náhodné jevy A a B platí $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- pro náhodný jev A a opačný jev $\overline{A}$ platí $P(\overline{A}) = 1 - P(A)$

V diskrétním případě se pravděpodobnost náhodného jevu A vypočte jako

$$P(A) = \frac{\text{počet příznivých možností}}{\text{počet všech možností}}$$

Pravděpodobnostní rozdělení

Příklad. *Házíme dvěma šestistěnnými kostkami, červenou a modrou. Elementární jevy jsou všechny možné dvojice hodnot (1,1), (1,2), (1,3), . . . , (6,5), (6,6). Celkem jich je 36. Nás zajímají pravděpodobnosti následujících náhodných jevů.*

- *Na červené kostce padne liché číslo*
- *Na modré kostce padne číslo dělitelné třemi*
- *Součet na obou kostkách bude větší nebo rovno 10*

Pravděpodobnostní rozdělení

Náhodné jevy

- **Podmíněná pravděpodobnost** – hledáme pravděpodobnost jevu A za podmínky že víme, že nastal jev B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Předpokládáme $P(B) > 0$.
Př. jaká je pst, že součet bodů na dvou kostkách je větší nebo rovno 10, když víme, že na modré kostce padlo sudé číslo.

- **Nezávislost jevů** – jevy A a B jsou nezávislé, když

$$P(A) = P(A|B)$$

nebo jinak zapsáno

$$P(A)P(B) = P(A \cap B)$$

Př.: Jsou jevy "na červené kostce padne liché číslo"a "na modré kostce padne číslo dělitelné třemi" nezávislé?

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné na

- **Spojitá** – pro spojité číselné proměnné
  může nastat libovolná reálná hodnota z nějakého intervalu
  př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní** – pro diskrétní číselné proměnné
  nastávají pouze oddělené hodnoty na číselné ose, např. počet
  př. binomické, poissonovo, alternativní, . . .

Funkce určující rozdělení

Distribuční funkce

- pravděpodobnost, že nastane hodnota menší nebo rovna danému číslu
- $F(t) = P(X \leq t), t \in \mathbb{R}$
- neklesající, zprava spojitá funkce
- definiční obor je celá reálná osa
- obor hodnot je mezi 0 a 1

Funkce určující rozdělení

Pravděpodobnostní funkce

- pravděpodobnost, že nastane konkrétní hodnota
- $p(t) = P(X = t), t \in \mathbb{R}$
- definovaná pouze pro diskrétní rozdělení
- nespojitá, nenulová jen v hodnotách, kterých může náhodná veličina nabývat
- součty pravděpodobností až do konkrétní hodnoty dávají distribuční funkci v daném bodě
- $F(t) = \sum_{-\infty}^{t} p(t), t \in \mathbb{R}$

Funkce určující rozdělení

Hustota

- hladká křivka definovaná na celé reálné ose
- derivace distribuční funkce
- $f(t) = \frac{d}{dt}F(t)$
- definovaná pouze pro spojitá rozdělení
- obdoba pravděpodobnostní funkce, ale nedefinuje konkrétní pravděpodobnosti
- pravděpodobnost jedné konkrétní hodnoty u spojitého rozdělení je 0
- distribuční funkce je integrálem z hustoty až do daného bodu (velikost plochy pod křivkou)

Základní charakteristiky rozdělení

Definice **Střední hodnoty**

- diskrétní rozdělení

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

- spojité rozdělení

$$EX = \int_{-\infty}^{\infty} xf(x)dx$$

- populační, nebo také teoretický průměr
- když budu pokus opakovat mnohokrát, tak jaký bude **dlouhodobý průměr**
- očekávaná hodnota

Základní charakteristiky rozdělení

Definice **Rozptylu**

- diskrétní rozdělení

$$Var(X) = \sum_{i=1}^{n} (X_i - E(X))^2 p_i$$

- spojité rozdělení

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

- populační, neboli teoretická variabilita
- mám-li k dispozici celou populaci, počítá se jako průměr druhých mocnin odchylek od průměru

Binomické rozdělení

Mějme náhodný pokus, který může skončit jedním ze dvou výsledků: úspěch – neúspěch. Opakujme tento pokus mnohokrát a počítejme počet úspěchů. Počet úspěchů má binomické rozdělení.

Značení $Bi(n, p)$, kde

- n – počet pokusů,
- p – pravděpodobnost úspěchu

Hodnoty pravděpodobnostní funkce

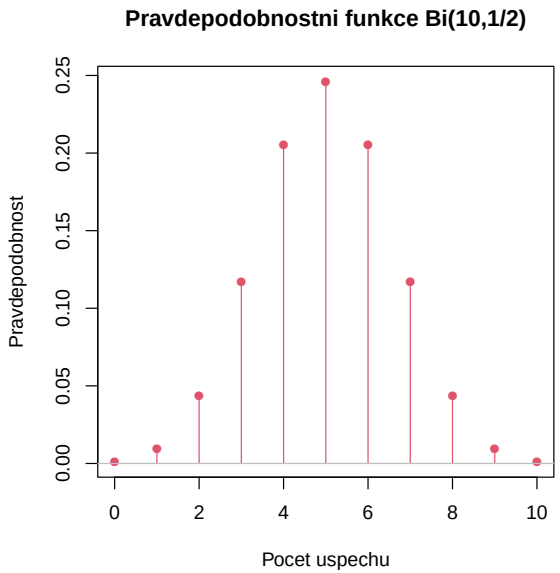
$$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

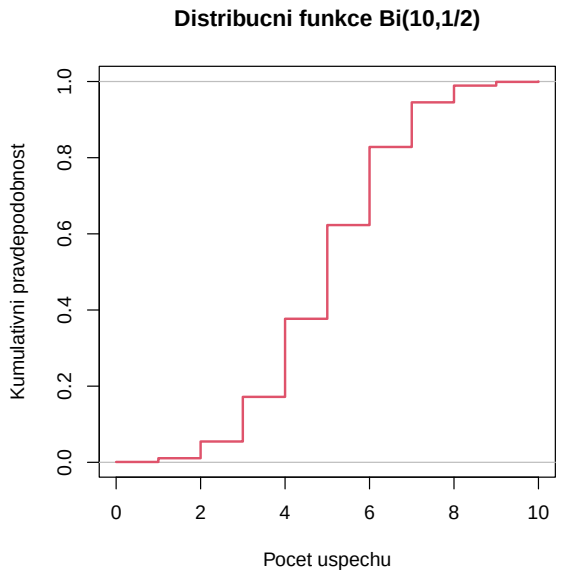
Střední hodnota a rozptyl

$$\mathrm{E(X) = np}, \qquad \mathrm{Var(X) = np(1 - p)}$$

Binomické rozdělení

Příklad. *Házíme 10x mincí a počítáme, kolikrát padla panna. Počet pokusů je* $n = 10$*, pravděpodobnost úspěchu* $p = 1/2$*. Máme tedy rozdělení* $Bi(10, 1/2)$.
Pravděpodobnostní a distribuční funkce.





Střední hodnota a rozptyl

$$E(X) = np = 10\frac{1}{2} = 5, \qquad Var(X) = np(1-p) = 10\frac{1}{2}\frac{1}{2} = 2.5$$

Normální rozdělení

Jedná se o "hezké" rozdělení, se kterým se dobře pracuje. Toto rozdělení má výška lidí určitého věku, IQ, ...

Značení $N(\mu, \sigma^2)$, kde

- μ – střední hodnota
- σ^2 – rozptyl

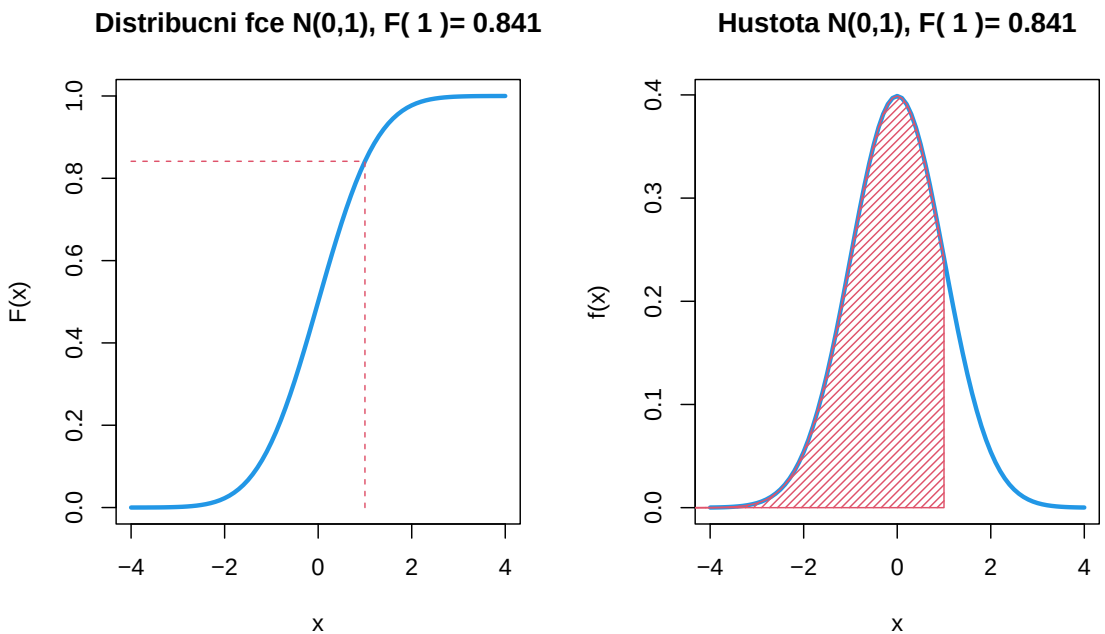
Hustota normálního rozdělení má tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Je to tak zvaná Gaussova křivka.
Ve statistice se nejčastěji používá standardní normální rozdělení $N(0, 1)$.

Normální rozdělení

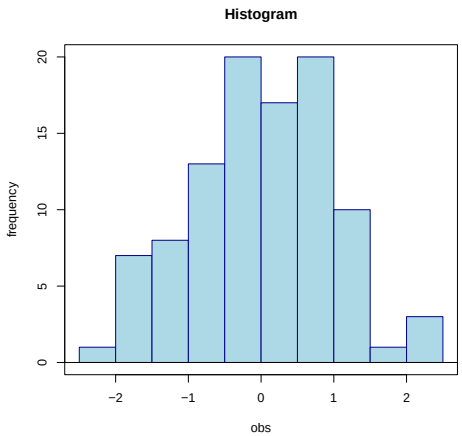
Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota.

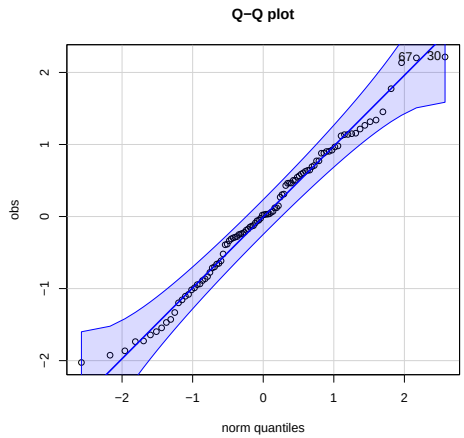


Testování normality

Většina statistických postupů, odhadů a testů je odvozena právě pro normální rozdělení. Je proto dobré zjistit, zda náhodná veličina normální rozdělení má či nemá.

- **Grafické testy** – histogram a pravděpodobnostní graf





- **Číselné testy** – nejčastěji Shapiro-Wilkův test
- **Popisné statistiky tvaru rozdělení** – šikmost i špičatost se mají rovnat 0

Cílem je odhadnout nějakou populační charakteristiku

- **náhodný výběr**: soubor nezávislých stejně rozdělených náhodných veličin (měříme stejným způsobem stejnou veličinu na různých jednotkách)
- na základě náhodného výběru děláme závěr o populaci
- na základě správně nastavených experimentů děláme obecné závěry o chování měřené veličiny
- výběr/nastavení experimentů musí být reprezentativní
- u číselných veličin nejčastěji odhadujeme **střední hodnotu**
- u kategorických veličin odhadujeme **pravděpodobnost jevu**
- teoreticky můžeme odhadovat i jiné veličiny (rozptyl, vybraný kvantil, ...)

Bodový odhad střední hodnoty

Příklad. *Zajímá nás průměrná koncentrace látky B v roztoku vzniklém určitým postupem. Za různých podmínek bylo připraveno 100 roztoků (za různých teplot, při různých tlakových podmínkách, atd.). Výběrový průměr koncentrace látky B vyšel 17.2 a výběrová směrodatná odchylka 2.9. Co můžeme říci o očekávané koncentraci (střední hodnotě) látky B?*

- **nejlepší bodový odhad** je výběrový průměr $\overline{X} = 17.2$
- jaká je pravděpodobnost, že se střední hodnota bude rovnat přesně tomuto číslu?
- je tento odhad dobrý?
- střední chyba odhadu průměru

$$\text{SEM} = \frac{\text{sd}(X)}{\sqrt{n}} = \frac{2.9}{\sqrt{100}} = 0.29$$

Intervalový odhad střední hodnoty

Chceme interval, ve kterém se s vysokou pravděpodobností bude nacházet skutečná střední hodnota.
Na čem tento interval závisí a jak?

- Výběrový průměr – leží ve středu intervalu spolehlivosti
- Výběrový rozptyl – čím větší variabilitu výběr má, tím širší bude interval spolehlivosti
- Počet pozorování – čím více pozorování, tím přesnější odhad a tím užší interval spolehlivosti
- Požadovaná spolehlivost – čím spolehlivější výsledek chci, tj. čím větší pravděpodobnost, že výběrový průměr bude ležet uvnitř intervalu spolehlivosti, tím širší interval

Intervalový odhad střední hodnoty

Lze odvodit, že standardizovaný výběrový průměr má t-rozdělení o $n-1$ stupních volnosti

$$\frac{\overline{X}-\mu}{\mathrm{sd(X)}/\sqrt{\mathrm{n}}} \sim t_{n-1},$$

z tohoto faktu a z požadavku

$$P(\mu \in \text{interval spolehlivosti}) = 1-\alpha$$

je možné dále odvodit meze intervalu spolehlivosti

$$\overline{X} \pm t_{n-1}(1-\alpha/2)\frac{\mathrm{sd(X)}}{\sqrt{n}}$$

kde $t_{n-1}(1-\alpha/2)$ je $1-\alpha/2$ procentní kvantil t-rozdělení o $n-1$ stupních volnosti

Intervalový odhad střední hodnoty

95%-ní interval spolehlivosti pro výše uvedený příklad vychází

$$(16.62; 17.78)$$

Interpretace

- Se spolehlivostí 95% skutečná střední hodnota koncentrace látky B leží v intervalu od 16.62 do 17.78.
- Kdybychom opakovali pokus víckrát, tak v 95% případů bude interval spolehlivosti obsahovat skutečnou střední hodnotu.
- Na základě tohoto intervalu nemůžeme říci nic o výsledcích jiných pokusů, pouze o teoretické střední hodnotě

Bodový odhad pravděpodobnosti

Příklad. *Máme určitý roztok, do kterého přidáme látku A. S jakou pravděpodobností se tento roztok zbarví do zelena? Pokus opakovalo 100 studentů a roztok se jim zbarvil do zelena v 78 případech.*
Zbarvení roztoku do zelena má alternativní rozdělení s parametrem (pravděpodobností) p.

- **nejlepším bodovým odhadem** psti je relativní četnost $\hat{p} = 78/100 = 0.78$
- jaká je pravděpodobnost, že skutečná pravděpodobnost se rovná přesně tomuto číslu?
- je tento odhad dobrý?
- střední chyba odhadu pravděpodobnosti

$$\mathrm{SEP} = \frac{\mathrm{sd}(X)}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Intervalový odhad pravděpodobnosti

Interval spolehlivosti vychází z aproximace odhadu pravděpodobnosti normálním rozdělením

$$p = (\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1)$$

pro $n\hat{p}(1-\hat{p}) > 9$, tedy pro větší n.
Z požadavku, že

$$P(p \in \text{interval spolehlivosti}) = 1 - \alpha$$

lze odvodit, že meze intervalu spolehlivosti pro pravděpodobnost jsou

$$\hat{p} \pm z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}$$

Existují i jiné vzorce pro interval spolehlivosti využívající buď přesné binomické rozdělení, nebo jiné aproximace.

Intervalový odhad pravděpodobnosti

95%-ní interval spolehlivosti pro pravděpodobnost zeleného zbarvení roztoku

$$(0.6988; 0.8612)$$

Interpretace

- Se spolehlivostí 95% skutečná pravděpodobnost, že se roztok zbarví do zelena, leží v intervalu od 69.88% do 86.12%.
- Kdybychom opakovali pokus víckrát, tak v 95% případů bude interval spolehlivosti obsahovat skutečnou pravděpodobnost.
- Na základě tohoto intervalu nemůžeme říci nic o výsledcích jiných pokusů, pouze o teoretické pravděpodobnosti

Základy testování hypotéz

Je možné říci, že platí následující tvrzení?

- Nový lék je lepší než ten stávající.
- Střední hoenodta koncentrace látky v roztoku je 50.
- Koncentrace látky v roztoku se liší podle světelných podmínek.
- Množství uvolněné látky z roztoku závisí na teplotě.

Platnost tvrzení je možné ověřit pomocí **statistických testů**.

Základy testování hypotéz

Při statistickém testu testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme H_0
  – je v ní vždy pouze jedna varianta
  – př. nový lék je stejný jako ten stávající, střední koncentrace látky v roztoku je 50
- Alternativní hypotézu – značíme H_1
  – obsahuje více možností (např. interval)
  – př. nový lék je lepší než ten stávající, střední koncentrace látky v roztoku je větší než 50
  – není přesně řečeno, jak moc je nový lék lepší, nebo o kolik je koncentrace větší než 50

Testované hypotézy

Nejčastěji testované dvojice hypotéz

- **Porovnání výběrů** / skupin
  - H_0: mezi skupinami není rozdíl
  - H_1: mezi skupinami je rozdíl
- **Test nezávislosti**
  - H_0: proměnné spolu nesouvisí
  - H_1: proměnné spolu souvisí
- k výše uvedeným nulovým hypotézám připadají v úvahu i jednotranné alternativy

Základy testování hypotéz

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu
  – tím jsme prokázali platnost alternativy
- Nezamítneme nulovou hypotézu
  – tím jsme neprokázali nic

Důležité je

- závěr je pomocí nulové hypotézy
- prokázat lze pouze platnost alternativy
- to, co mě zajímá, musí být v alternativě
- musíte vědět, co Vám test říká vzhledem k Vaší otázce

Základy testování hypotéz

Při rozhodování můžeme udělat chybu

- chyba prvního druhu – zamítneme H_0, přestože platí
  – značí se α, a jmenuje se hladina významnosti
  – závažnější z obou chyb
  – každý test má velikost této chyby předem omezenu
- chyba druhého druhu – nezamítneme H_0, přestože neplatí
  – značí se β
  – hodnota $1 - \beta$ se nazývá síla testu
  – při dané hladině významnosti chceme test co nejsilnější

Základy testování hypotéz

	Skutečně platí H_0	Skutečně platí H_1
Zamítáme H_0	Chyba I. druhu $\leq \alpha$	OK síla testu
Nezamítáme H_0	OK	Chyba II. druhu β

Základy testování hypotéz

Podle toho, co testujeme a podle typu dat vybereme vhodný statistický test, kterým budeme o platnosti testovaných hypotéz rozhodovat. Rozhodnutí můžeme udělat buď na základě

- porovnání **testové statistiky** (T) a kritické hodnoty (c)
- porovnání ***p*-hodnoty** a hladiny významnosti (α)

Platí, že

- absolutní hodnota testové statistiky $|T| \geq c$ nebo p-hodnota $\leq \alpha$ potom ZAMÍTÁME H_0
- absolutní hodnota testové statistiky $|T| < c$ nebo p-hodnota $> \alpha$ potom NEZAMÍTÁME H_0

P-hodnota

S testovou statistikou se většinou pracuje při ručním výpočtu

- testovou statistiku je možné ručně spočítat a kritické hodnoty jsou tabelovány

Statistické softwary vrací jako výsledek testu ***p*-hodnotu**

- *p*-hodnota se také nazývá *aktuální dosažená hladina testu*
- počítá se kombinací hodnoty testové statistiky a příslušné kritické hodnoty
- **definice** *p*-hodnoty
  - pravděpodobnost, že za platnosti H_0
    nastal výsledek, jaký nastal,
    nebo jakýkoliv jiný, který ještě více odpovídá alternativě

P-hodnota

P-hodnotu si také můžeme přestavit následovně

- Jak pravděpodobný je náš výsledek, když ve skutečnosti platí nulová hypotéza?
- Je pozorovaný odklon od nulové hypotézy (např. rozdíl mezi skupinami) dílem náhody?

Jednovýběrový t-test

Nejjednodušším testem je **jednovýběrový test o střední hodnotě**.

Testujeme

- H_0 : střední hodnota $= \mu_0$

Proti jedné ze tří alternativ

- H_1 : střední hodnota $\neq \mu_0$
- H_1 : střední hodnota $< \mu_0$
- H_1 : střední hodnota $> \mu_0$

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$

Jednovýběrový t-test

Testová statistika jednovýběrového t-testu je

$$T = \frac{\overline{X} - \mu_0}{\text{sd}(X)}\sqrt{n}$$

za platnosti nulové hypotézy má tato statistika t-rozdělení o $n - 1$ stupních volnosti.

- čím větší je rozdíl mezi $\overline{X}$ a μ_0, tím větší absolutní hodnota testové statistiky
- porovnáváme s kritickými hodnotami (kvantily) t-rozdělení (většinou cca 1.96)
- z testové statistiky a kritické hodnoty se počítá **p-hodnota**

Předpokladem jednovýběrového t-testu je, že průměr testované veličiny má normální rozdělení.

Příklad

Příklad. *Bylo změřeno 222 jedenáctiletých dětí. Průměrná výška tohoto výběru je 148.8 cm, a směrodatná odchylka výšky vyšla 7.1. Dá se předpokládat, že průměrná výška všech jedenátiletých dětí v republice je menší než 150 cm?*

Testované hypotézy

- H_0 : průměrná výška $=$ 150 cm
- H_1 : průměrná výška $<$ 150 cm

Testujeme na hladině významnosti $\alpha = 0.05$.

Příklad

Testová statistika vyšla

$$T = \frac{\overline{X} - \mu_0}{\text{sd(X)}}\sqrt{n} = \frac{148.8 - 150}{7.1/\sqrt{222}} = -2.5618$$

- kritická hodnota $t_{221}(1 - 0.05) = 1.65$
- jelikož $|T| = 2.56 > t_{221}(1 - 0.05) = 1.65$, zamítám nulovou hypotézu
- p-hodnota $p = 0.005 < 0.05$, tedy zamítám nulovou hypotézu

Závěr: Prokázala jsem, že průměrná výška jedenáctiletých dětí je menší než 150 cm.

Párový t-test

Párový test se používá v případě, že porovnáváme střední hodnotu ve dvou **závislých** výběrech.
Např.

- *Je koncentrace jedné látky v roztoku stejná jako koncentrace druhé látky?*
- *Snížila se koncentrace látky v roztoku po 15 minutách?*
- *Klesl pacientům po podání léku krevní tlak?*

Ať je otázka formulována jakkoliv, tak test porovnává průměrné hodnoty. Vyjde nám tedy odpověď, jak je to "v průměru".

Závislé výběry poznám tak, že data tvoří přirozené páry.

Párový t-test

Při aplikaci testu je důležité udržet párová data u sebe, (abyste neporovnávali koncentraci jedné látky ve Vašem roztoku s koncentrací druhé látky u souseda).
V prvním kroku jsou pro všechny páry vypočteny **rozdíly**:

$$R_i = X_i - Y_i$$

dále je testována střední hodnota těchto rozdílů, tedy je aplikován jednovýběrový t-test na hodnoty rozdílu.
Předpokladem testu je normalita **rozdílů** R_i.

Párový t-test

Příklad. *Bylo měřeno 56 vzorků roztoku a v každém byla zjišťována koncentrace látek B a C. Průměrná koncentrace látky B vyšla* 43.89*, průměrná koncentrace látky C* 48.38 *a průměr z rozdílů těchto koncentrací v každém roztoku vyšel* -4.5*. Směrodatné odchylky jsou pro koncentraci látky B* 12.1*, pro koncentraci látky C* 11.8 *a pro rozdíl koncentrací* 19.7*. Můžeme říci, že látka C má vyšší koncentraci než látka B?*

Do testové statistiky vkládáme charakteristiky rozdílu

$$T = \frac{\overline{X} - \mu_0}{\mathrm{sd}(X)}\sqrt{n} = \frac{-4.5 - 0}{19.7}\sqrt{56} = -1.7$$

- jelikož $|T| > t_{55}(1 - 0.05) = 1.67$, zamítám nulovou hypotézu
- p-hodnota $p = 0.0466 < \alpha = 0.05$, tedy zamítám nulovou hypotézu

Závěr: Prokázali jsme, že koncentrace látky C bývá vyšší než koncentrace látky B.

Dvouvýběrový t-test

Porovnáváme-li střední hodnotu dvou **nezávislých** výběrů, používá se **dvouvýběrový test**.

- testová statistika má tvar

$$T = \frac{\overline{X} - \overline{Y}(-\mu_0)}{S}$$

- S je střední chyba rozdílu průměrů
- S se počítá jinak, když oba výběry mají stejné rozptyly, a když je mají různé

Dvouvýběrový t-test

- pro různé rozptyly je

$$S = \sqrt{\frac{\mathrm{Var(X)}}{n_1} + \frac{\mathrm{Var(Y)}}{n_2}}$$

- n_1, n_2 je rozsah výběru X, respektive Y
- za platnosti H_0 má T t-rozdělení o počtu stupních volnosti, který se dá odvodit ze vztahu rozptylů
- předpokladem použití dvouvýběrového testu je normalita dat v obou výběrech

Dvouvýběrový t-test

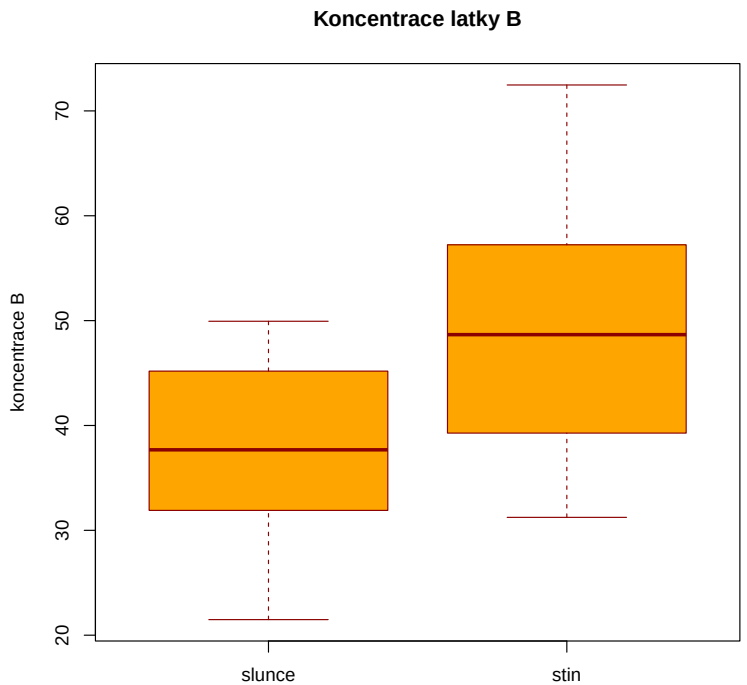
Příklad. *Ve výběru mám 56 roztoků, 28 měřených na přímém slunci a 28 ve stínu. Průměrná koncentrace látky B na slunci je 37.8 se směrodatnou odchylkou 7.9, a ve stínu je průměrná koncentrace 50 a směrodatná odchylka 12.5. Ovlivňuje sluneční světlo procesy v roztoku vzhledem ke koncentraci látky B?*

Testované hypotézy

- H_0 : koncentrace látky B je stejná na slunci i ve stínu
- H_1 : koncentrace látky B na slunci a ve stínu se liší

Dvouvýběrový t-test

Grafické porovnání



Dvouvýběrový t-test

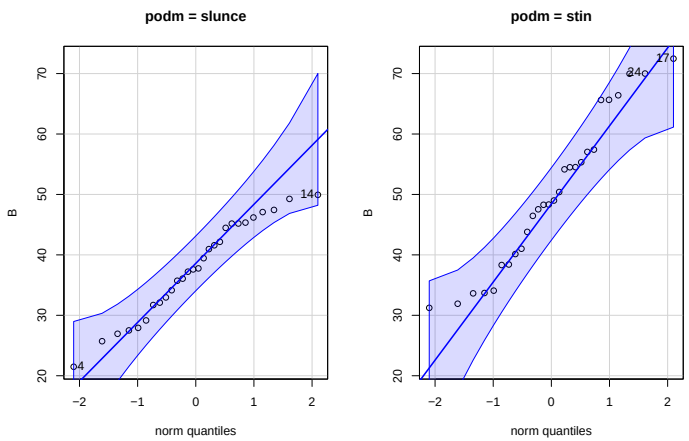
Testová statistika testu vychází

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S} = \frac{37.8 - 50}{2.8} = -4.35$$

Tomu odpovída p-hodnota 0.00008. P-hodnota je menší než $\alpha = 0.05$, nulovou hypotézu zamítám

Závěr: Na hladině významnosti 5% jsem prokázala rozdíl mezi roztoky na slunci a ve stínu v koncentraci látky B.

Kontrola normality.



Analýza rozptylu – ANOVA

Porovnávání střední hodnoty ve více než ve dvou nezávislých výběrech

- testované hypotézy
  - H_0 : všechny střední hodnoty jsou stejné
  - H_1 : alespoň jedna střední hodnota se liší
- porovnává se variabilita mezi výběry s variabilitou v rámci výběrů.
- předpoklad testu je normalita dat v každém výběru
- existuje varianta pro stejné rozptyly ve výběrech a pro rozptyly různé

Analýza rozptylu – ANOVA

Analýza rozptylu rozkládá celkovou variabilitu (rozptyl)

$$(n-1)\mathrm{VarX} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X_i} - \overline{\mathrm{X}})^2$$

- variabilita vysvětlená (mezi výběry):
  jak moc se liší skupinové průměry od celkového

$$\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2$$

- variabilita nevysvětlená (zbytková, v rámci výběrů):
  jak moc se liší jednotlivá pozorování od průměru ve "své" skupině

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$$

Analýza rozptylu – ANOVA

- výstupem z analýzy rozptylu je tzv. **tabulka analýzy rozptylu**
- v prvním řádku je variabilita vysvětlená
- ve druhém variabilita nevysvětlená
- testová statistika je dává do podílu
- za platnosti H_0 má testová statistika F-rozdělení o $k-1$ a $n-k$ stupních volnosti

Párové srovnání

Rozdíl mezi dvojicemi skupin se testuje pomocí **párového srovnání**

- rozdíl nelze testovat větším počtem dvouvýběrových testů
- používá se tzv. **Tukeyho test**
- testované hypotézy
  - H_0 : střední hodnoty μ_i a μ_j jsou stejné
  - H_1 : střední hodnoty μ_i a μ_j se liší
- testová statistika porovnává standardizovaný rozdíl průměrů $Q = \frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{s^*}$ s kritickou hodnotou

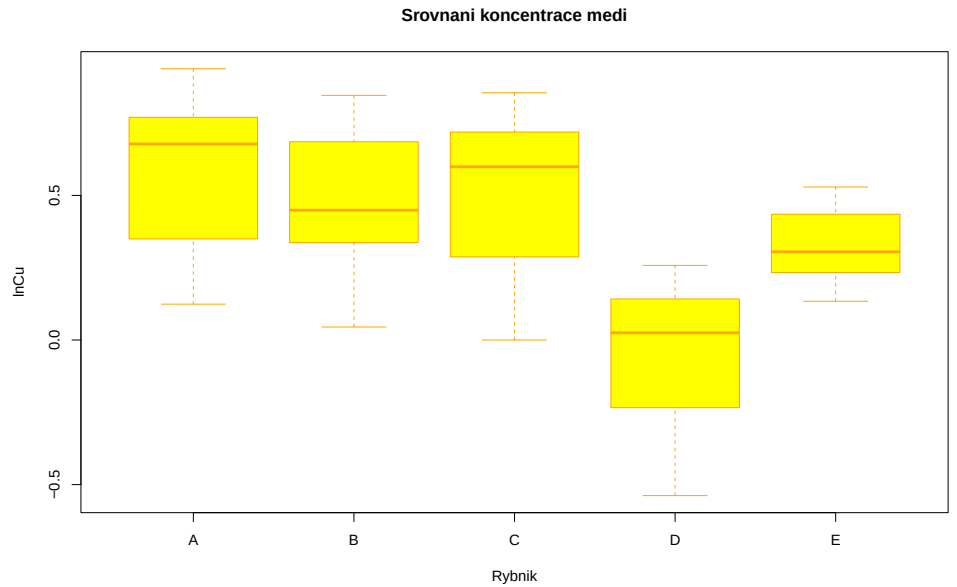
Analýza rozptylu – ANOVA

Příklad. *Byla měřena koncentrace mědi v těle ryb. Porovnáváno bylo 5 rybníků, kde z každého byl vyloven vzorek sedmi ryb. Výběrové průměry pro jednotlivé rybníky vyšly 0.57, 0.48, 0.50, -0.06 a 0.33. Liší se od sebe tyto rybníky?*
Testujeme

- H_0 : všechny rybníky jsou stejné
- H_1 : alespoň jeden rybník se liší

Analýza rozptylu – ANOVA

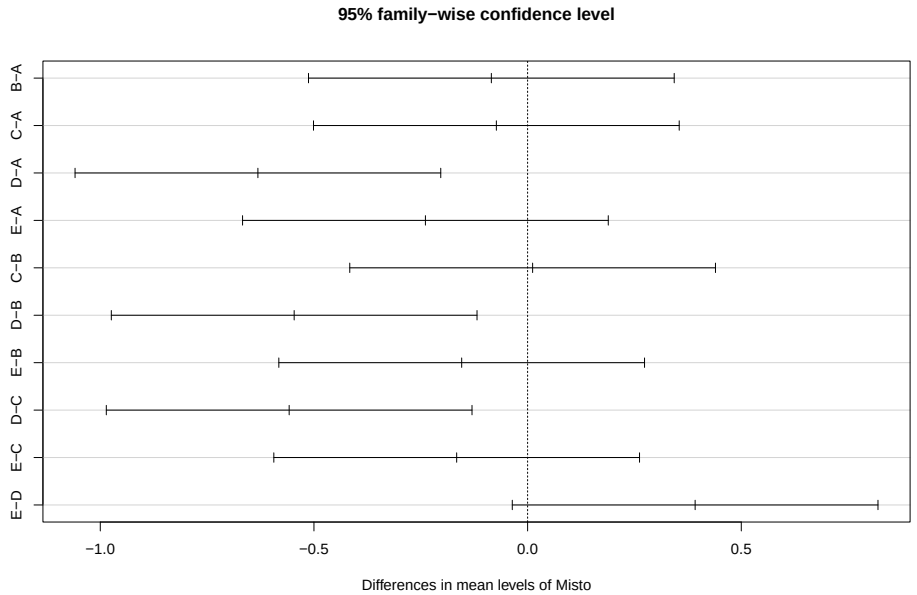
Grafické porovnání



- p-hodnota vyšla 0.00127, tedy menší než $\alpha = 0.05$
- $\Rightarrow$ nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší.

Analýza rozptylu – ANOVA

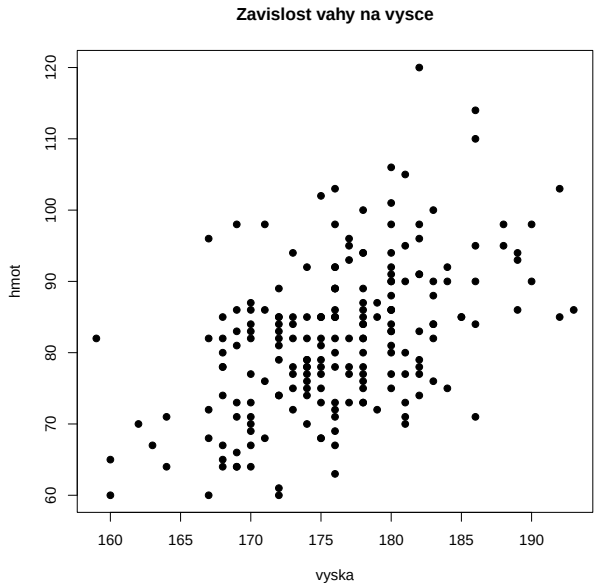
Graf pro párové srovnání. Pro kterou dvojici rybníků interval spolehlivosti neobsahuje svislou čárkovanou čáru (nulu), pak mezi ní je významný rozdíl.



Závěr: Rybníky se v koncentraci mědi v těle ryb významně liší, konkrétně se liší rybník D od rybníků A, B a C.

Bodový graf

Vztah dvou číselných proměnných se znázorňuje bodovým grafem



- korelační koeficient měří sílu závislosti
- lineární regrese prokládá grafem přímku

Pearsonův korelační koeficient

Měří lineární vztah dvou číselných proměnných

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- předpokladem je normální rozdělení obou proměnných
- nabývá hodnot mezi -1 a 1 a platí
- určuje směr a sílu závislosti

Pearsonův korelační koeficient

- absolutní nepřímá závislost má $\mathrm{Cor}(X, Y) = -1$
- lineární nezávislost/ nekorelovanost má $\mathrm{Cor}(X, Y) = 0$
- absolutní přímá závislost má $\mathrm{Cor}(X, Y) = 1$
- kladná hodnota značí přímou / pozitivní závislost
- záporná hodnota značí nepřímou / negativní závislost
- hodnoty $|\mathrm{Cor}(X, Y)| \leq 0.3$ značí slabou závislost
- hodnoty $0.3 < |\mathrm{Cor}(X, Y)| < 0.7$ značí střední závislost
- hodnoty $|\mathrm{Cor}(X, Y)| \geq 0.7$ značí silnou závislost

Pearsonův korelační koeficient

O statistické významnosti závislosti rozhodujeme testem

- H_0 : proměnné spolu nesouvisí, korelační koeficient $= 0$
- H_1 : proměnné spolu souvisí, korelační koeficient $\neq 0$,

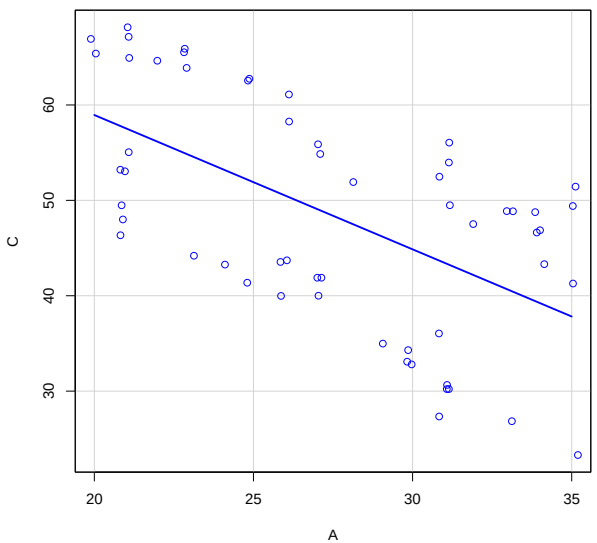
Za platnosti nulové hypotézy platí, že testová statistika

$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

má t-rozdělení o $n - 2$ stupních volnosti.

Pearsonův korelační koeficient

Příklad. *Bylo měřeno 56 roztoků a v každém se měřila koncentrace látek A a C. Můžeme tvrdit, že mezi těmito látkami existuje lineární vztah?*



- z grafu je patrná klesající / nepřímá závislost mezi oběma proměnnými

Pearsonův korelační koeficient

- korelační koeficient vyšel -0.57
- jedná se o střední negativní závislost
- testované hypotézy
  - H_0 : koncentrace látek A a C spolu nesouvisí, kor. koef. $= 0$
  - H_1 : koncentrace látek A a C spolu souvisí, kor. koef. $\neq 0$
- testová statistika $|T| = 5.16 > t_{54}(0.975) = 2$, tedy zamítáme nulovou hypotézu
- p-hodnota $0.000003615 < \alpha = 0.05$, tedy zamítáme : H_0

Závěr: Závislost je průkazná.

Lineární regrese

Lineární regrese zkoumá příčinnou závislost jedné číselné proměnné na druhé

- bodovým grafem prokládáme přímku a hledáme její rovnici
- na x-ovou osu kreslíme nezávisle proměnnou (příčinu) X
- na y-ovou osu kreslíme závisle proměnnou (důsledek) Y
- **odhadujeme model**

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

- Y_i jsou hodnoty závisle proměnné
- X_i jsou hodnoty nezávisle proměnné
- β_0 je absolutní člen
- β_1 je lineární člen
- e_i jsou náhodné chyby

Lineární regrese

Odhad probíhá **metodou nejmenších čtverců**

- minimalizuje součet druhých mocnin residuí

$$\min \sum_{i=1}^{n} R_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

- hodnoty $\hat{Y}_i$ se nazývají odhady, či predikce (body na přímce)
- b_0, b_1 jsou odhady regresních koeficientů
- residua jsou vzdálenosti budů od přímky po kolmici

Pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné.

Lineární regrese

Koeficient determinace

- kolik procent variability závisle proměnné se modelem vysvětlí
- tedy z kolika procent závisle proměnná závisí na X a z kolika na něčem jiném

Test nezávislosti obou proměnných

- H_0 : proměnná Y na X lineárně nezávisí, $\beta_1 = 0$
- H_1 : proměnná Y na X lineárně závisí, $\beta_1 \neq 0$
- testová statistika $b_1 / \mathrm{se}(\mathrm{b}_1) \sim \mathrm{N}(0, 1)$
  - b_1 je odhad lineárního členu β_1 a $\mathrm{se}(\mathrm{b}_1)$ je jeho střední chyba
- p-hodnota testu je stejná jako u korelačního koeficientu

Předpokladem lineární regrese je, že residua jsou nezávislá, s normálním rozdělením a konstantním rozptylem.

Lineární regrese

Příklad. *Pokračujme příkladem závislosti koncentrace látky C na koncentraci látky A.*

- odhadli jsme model ve tvaru $C = 87.12 - 1.41A$
- interpretace koeficientů
  - $b_0 = 87.12$ v tomto bodě regresní prímka protíná osu y
  - $b_1 = -1.4$ při nárůstu koncentrace látky A o jednu jednotku očekáváme pokles koncentrace látky C o 1.4 jednotek
- koeficient determinace (R-squared) vyšel 0.33
  - závislostí na koncentraci látky A se vysvětlí 33% variability koncentrace látky C
- p-hodnota testu nezávislosti vyšla $0.000003615 < \alpha = 0.05$, závislost je statisticky významná

Test dobré shody

Test používaný v situacích, kdy potřebujeme porovnat naměřené hodnoty s hodnotami očekávanými. Používá se

- pro testy parametrů v tzv. *multinomickém rozdělení*
  - řídí se jím kategorická náhodná veličina s k kategoriemi
  - testujeme konkrétní hodnoty pravděpodobností těchto kategorií
- testy o *konkrétním rozdělení*
  - porovnávají se naměřené kvantily s teoretickými kvantily z daného rozdělení

Test dobré shody

Test o pravděpodobnostech multinomického rozdělení

- $H_0 : p_1 = \pi_1, \ldots, p_k = \pi_k$
- H_1 : neplatí $p_1 = \pi_1, \ldots, p_k = \pi_k$

Testová statistika má tvar

$$\chi^2 = \sum_{i=1}^{k} \frac{(np_i - n\pi_i)^2}{n\pi_i}$$

- za platnosti H_0 má χ^2-rozdělení o $k - 1$ stupních volnosti
- porovnává pozorované četnosti (np_i) a očekávané četnosti ($n\pi_i$)
- předpoklad je, že všechny očekávané četnosti jsou větší než 5

Test dobré shody

Příklad. *Házíme 50 krát šestistěnnou kostkou a počítáme, kolikrát padla která hodnota: 1 – 8×, 2 – 5×, 3 – 12×, 4 – 7×, 5 – 9× a 6 – 9×. Můžeme o kostce říci, že je spravedlivá?*

Testujeme hypotézy

- $H_0 : p_1 = p_2 = \ldots = p_6 = 1/6$
- H_1 : *alespoň jedna* $p_i, i = 1, \ldots, 6$ *se nerovná* $1/6$.
- *očekávané četnosti* $n\pi_i = 50 \times 1/6 = 8.\overline{3}$ *jsou větší než 5.*
- *testová statistika* $\chi^2 = 3.28$
- *p-hodnota* $p = 0.6569$ *je větší než* $\alpha = 0.05$, *tedy nezamítáme nulovou hypotézu.*

Závěr: *Neprokázali jsme, že by kostka byla falešná.*

χ^2-test nezávislosti

Vztah dvou kategorických proměnných popisujeme **kontingenční tabulkou**. Označme

- $X_1, \ldots, X_k$ hodnoty jedné kategorické proměnné
- $Y_1, \ldots, Y_l$ hodnoty druhé kategorické proměnné
- $n_{i,j}$ četnost současného výskytu znaků X_i, Y_j
- $n_{i.}$ marginální četnost znaku X_i
- $n_{.j}$ marginální četnost znaku Y_j
- n celkový počet pozorování

χ^2-test nezávislosti

Kontingenční tabulka absolutních četností

	Y_1	$\dots$	Y_l	
X_1	$n_{1,1}$	$\dots$	$n_{1,l}$	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
X_k	$n_{k,1}$	$\dots$	$n_{k,l}$	$n_{k.}$
	$n_{.1}$	$\dots$	$n_{.l}$	n

Testované hypotézy testu **nezávislosti**

- H_0 : proměnné na sobě nezávisí
- H_1 : proměnné na sobě závisí

χ^2-test nezávislosti

Testová statistika má obdobný tvar jako u předešlého testu

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(pozorovane_{i,j} - ocekavane_{i,j})^2}{ocekavane_{i,j}} =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{i,j} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = \sum_{i=1,j=1}^{k,l} \frac{(np_{ij} - n\pi_{ij})^2}{n\pi_{ij}}$$

- za platnosti nulové hypotézy má χ^2-rozdělení o $(k-1)(l-1)$ stupních volnosti
- očekávané četnosti se dopočítávají z definice nezávislosti $P(A \cap B) = P(A)P(B)$
- předpokladem testu je, že všechny očekávané četnosti jsou větší než 5

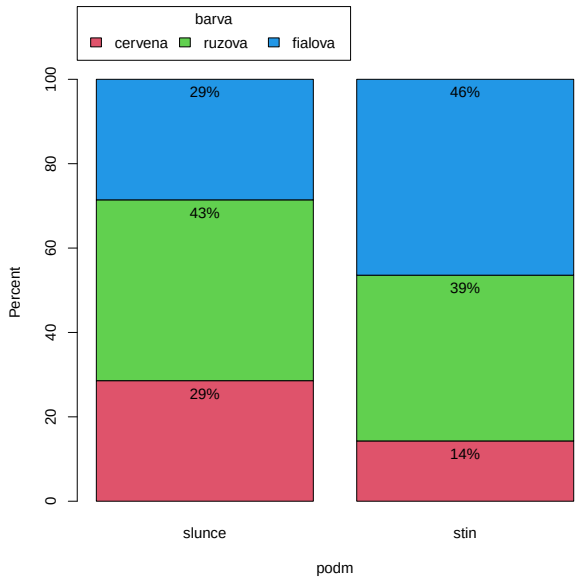
χ^2-test nezávislosti

Příklad. *Z 56 roztoků se 12 zbarvilo do červena, 23 do růžova a 21 do fialova. Polovina roztoků byla na slunci a polovina ve stínu. Přesné rozložení barev za různých slunečních podmínek znázorňuje tabulka. Je možné říci, berva roztoku nezávisí na slunečním svitu?*

	červená	růžová	fialová	Celkem
slunce	8	12	8	28
stín	4	11	13	28

χ^2-test nezávislosti

Vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

χ^2-test nezávislosti

Testem nezávislosti jsme zjišťovali

- H_0 : barva se slunečním svitem nesouvisí
- H_1 : barva se slunečním svitem souvisí

Výsledky testu

- testová statistika vyšla 2.57, což je větší než kritická hodnota 5.99
- p-hodnota testu vyšla 0.277, což je větší než $\alpha = 0.05$
- tedy nezamítáme nulovou hypotézu
- R-ko nevypsalo Warning, předpoklady testu jsou splněny

Závěr: Neprokázali jsme, že by barva roztoku závisela na slunečním svitu.

Chyby měření

Chyba měření je rozdíl mezi skutečnou hodnotou a naměřenou hodnotou. Rozlišujeme několik typů chyb

- Hrubá chyba – velká chyba daná nepozorností, volbou špatné metody či jinou nečekanou situací
  často je možné ji odhalit přes *odlehlé pozorování*
  náprava je možná jen přes nové měření/ vymazání
- Systematická chyba – je dána přesností měřícího přístroje
  většinou udávána výrobcem přístroje, není-li dána, uvažujeme ji jako polovinu nejmenší měřitelné jednotky
  je možné ji korigovat
- Náhodná chyba – vzniká náhodnými rušivými vlivy
  v přírodě může mít poměrně velkou hodnotu, v laboratorních podmínkách bývá malá
  není možné ji opravit, ale je možné ji měřit opakovaným pozorováním
  často má normální rozdělení

Chyby měření

Nejistota měření – souvisí s výsledkem měření a charakterizuje rozsah hodnot, které je možné přiřadit k měřené veličině.

Celková nejistota bývá součtem nejistoty statistické a nejistoty odhadnutelné.

- Statistická nejistota – jedná se o střední chybu odhadu pro výběrový průměr je rovna

$$u_A = \mathrm{se}(\overline{X}) = \frac{\mathrm{sd}(X)}{\sqrt{n}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}}{\sqrt{n}}$$

- Odhadnutelná nejistota – dána chybou měření udávanou výrobcem, změnou referenčních podmínek, výsledky minulých šetření, atd.

$$u_B = \sqrt{\sum_{j=1}^{m} A_j^2 u_{Bj}^2}$$

kde A_j jsou součinitelé citlivosti zdrojů a u_{Bj} jsou maximální odchylky zdrojů

Chyby měření

Jak **reportovat** chybu měření

- Absolutní chyba měření – meřena přímo v jednotkách měřené veličiny
  Absolutní chyba = ukazovaná hodnota – skutečná hodnota
  např. naměříme hodnotu 1 ± 0.05, což znamená, že skutečná hodnota bude v rozmezí od 0.95 do 1.05.
- Relativní chyba měření – nejčastěji měřena v procentech
  Relativní chyba = (absolutní chyba)/(měřená hodnota)
  např. relativní chyba při výše zmíněném příkladu je $0.05/1 \times 100\% = 5\%$
- Interval spolehlivosti – nejistota měření se často udává pomocí intervalu spolehlivosti, který v sobě zahrnuje všechny dostupné chyby

Chyby měření

Chyby **přístrojů**

- Základní chyba měření – dosahuje se jí při předepsaných referenčních podmínkách
  nejčastěji jsou dány podmínky na teplotu, tlak, stabilitu napájení, atd.
  bývá dáno velice úzké rozpětí hodnot
  této chyby lze většinou dosáhnout pouze v laboratorních podmínkách
- Pracovní chyba měření – chyba, kterou dosahuje přístroj při běžných pracovních podmínkách
  podmiňuje se na stejné veličiny, ale rozpětí povolených hodnot bývá širší
  tato chyba je větší než chyba základní, podle normy ČSN 61557 může být relativní chyba měřícího přístroje maximálně 30%

Chyby měření

Pravidla o chybách měření

- Měříme-li veličiny x_1, x_2, které ve výsledku sčítáme ($y = x_1 + x_2$), měly by mít obě přibližně stejnou *absolutní chybu*. Má-li jedna chybu větší, rozhoduje pak sama o chybě výsledku.
- Je-li výsledkem sčítání $y = x_1 + x_2$ malá hodnota, snažíme se ji měřit přímo, jinak je výsledek zatížen velkou relativní chybou.
- Je li výsledkem součin měřených veličin $y = x_1 x_2$, pak by obě veličiny měly mít stejnou *relativní chybu*