Úvod do teorie měření

Alena Černíková

alena.cernikova@ujep.cz

30. dubna 2024

Podmínky zápočtu

- **tři domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat na univerzitní OneDrive – bude upřesněno později
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  zpracování závislosti dvou proměnných
  ucelený text od výzkumné otázky až po interpretaci výsledku

Obsah kurzu

- Data
- 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 průměrná výška dospělých lidí v České republice. Změříme 200 lidí a popíšeme získané hodnoty. Na základě těchto údajů chceme něco říci i o celé populaci.*

Data

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

- byly získány objektivně
- tvořily reprezentativní výběr
  pokud chceme informace za celou ČR, nemůžeme získávat data jen v Ústí nad Labem
- jednotlivé hodnoty byly vzájemně nezávislé
  není dobré, aby mezi 200 vybranými lidmi bylo 50 z jedné sportovní školy

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.

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.

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.
  - a pro malé počty kategorií i absolutní a relativní četnosti

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

Popisné statistiky polohy

Příklad. *Mějme náhodný výběr 18-ti dospělých lidí a předpokládejme, že jsme u nich naměřili výšky* 176, 184, 167, 193, 174, 182, 181, 179, 187, 165, 168, 172, 184, 178, 160, 168, 171, 159. *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, nebo průměr prostředních dvou

Jak vypočítat **kvartily**

- z uspořádané řady – hodnoty v jedné a ve třech čtvrtinách

Jak vypočítat **extrémy**

- minimum a maximum

Popisné statistiky polohy – výpočet kvartilů podle **R**

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
- $X_{(k)}$ – hodnoty z uspořádané řady, k-tý nejmenší prvek
- 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 je určen vybraným pravidlem
  nejčastěji se používá *Sturgesovo pravidlo*

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

kde n je počet pozorování

V praxi je často používaný **vážený průměr**

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

kde

- X_i jsou hodnoty
- w_i jsou váhy

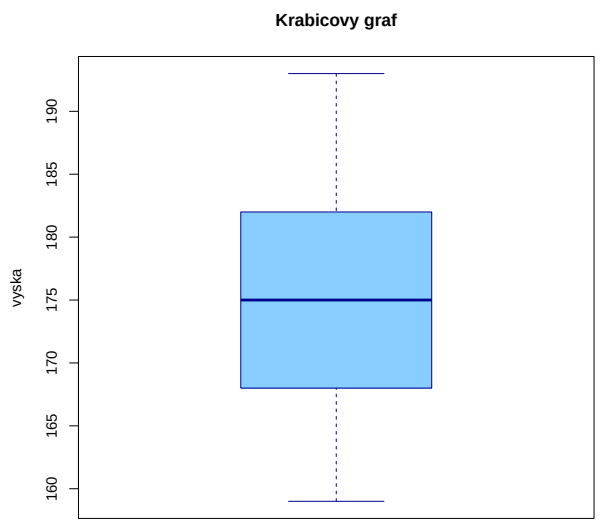
Příklad. *Spočtete průměrnou známku u termínu zkoušky z matematiky, když víte, že 5 studentů dostalo 1, 7 studentů dostalo 2 a 13 studentů dostalo 3.*

Popisné statistiky – číselné proměnné

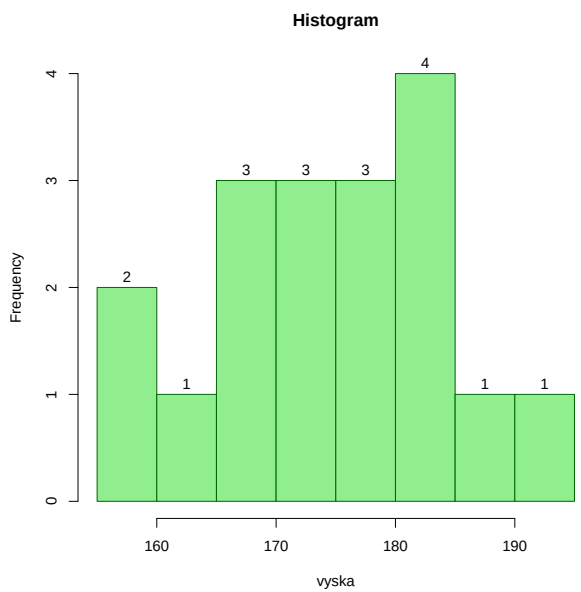
Popisné statistiky polohy – výsledky

- průměr – 174.89
- medián – 175
- kvartily – 168, 181.75
- extrémy – 159, 193

Grafy



1. Krabicový graf



2. 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 tvaru rozdělení

Pro obě statistiky (šikmost i špičatost) je třeba nejprve spočítat standardizované proměnné, tak zvané **Z-skóry**

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

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

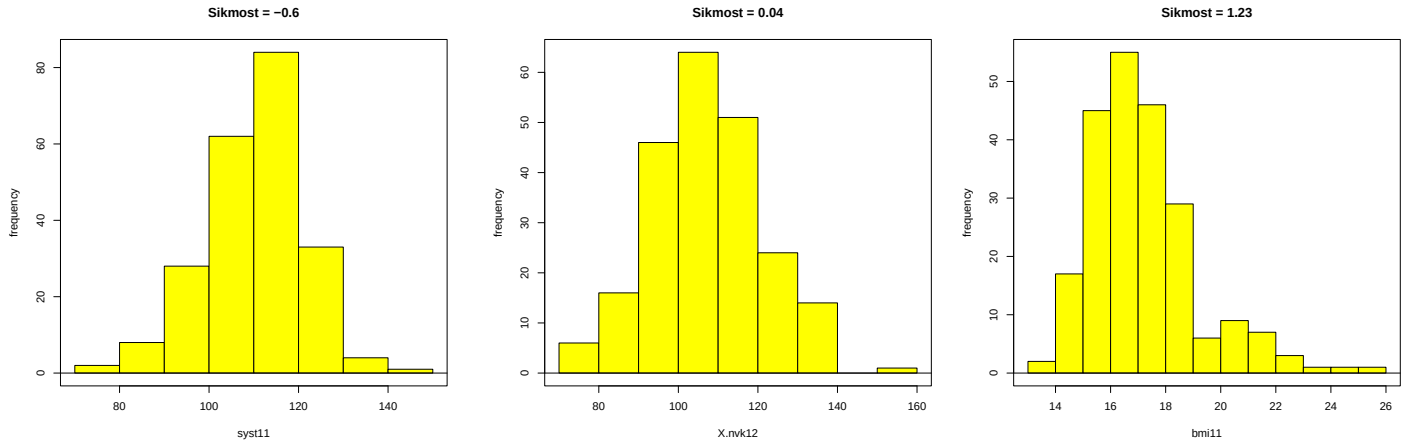
$$\mathrm{Skew}(X) = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{\mathrm{sd}(X)} = \frac{\sum_{i=1}^{n} Z_i^3}{n}$$

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

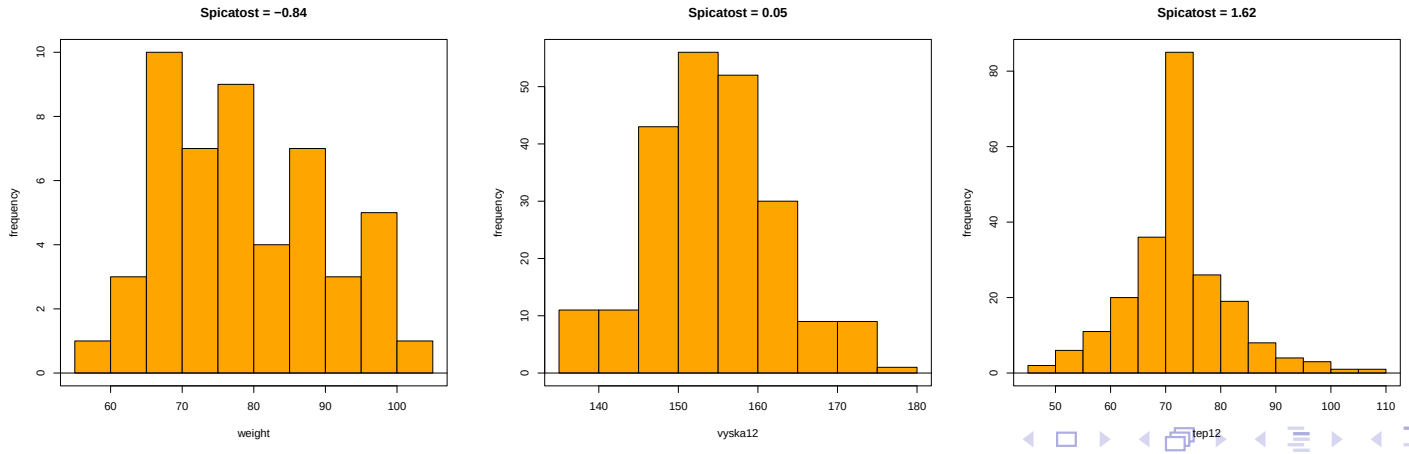
$$\mathrm{Kurt}(X) = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{\mathrm{sd}(X)} - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



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



Popisné statistiky – číselné proměnné

Popisné statistiky variability – výsledky

- rozptyl – 88.81
- směrodatná odchylka – 9.42
- mezikvartilové rozpětí – 13.75
- variační koeficient – 0.054

Popisné statistiky tvaru rozdělení – výsledky

- šikmost – 0.027
- špičatost – -1.04

Otázky na promyšlení

- Kdy kterou charakteristiku použít a proč
- Jaké mají jednotlivé statistiky rozměry
- Jak se jednotlivé statistiky mění v závislosti na posunutí a změně měřítka u původní veličiny

Číselné popisné statistiky

Příklad. *Mějme náhodný výběr 10-ti dospělých lidí a předpokládejme, že jsme u nich 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.

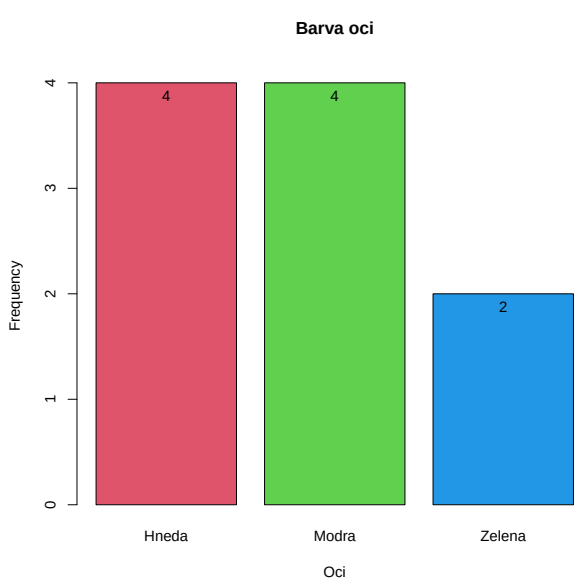
Barva	Absolutní	Relativní %
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

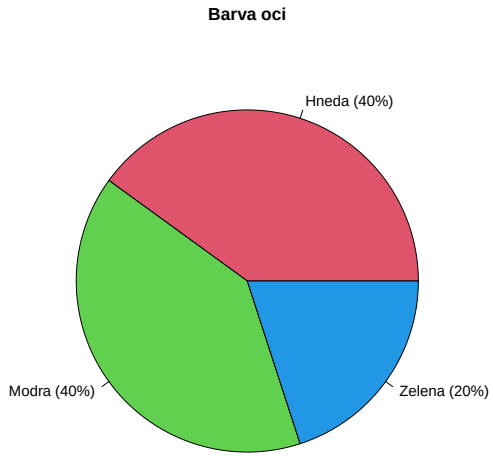
Jak vypočítat **relativní četnost?**
Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

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

Grafické popisné statistiky

Sloupcový a koláčový graf – je možné je popisovat v absolutních počtech, nebo v procentech





Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení popisuje pravděpodobnosti možných výsledků náhodného pokusu.

- **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 kosce 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

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*

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

Náhodné jevy

- **Vzorec pro celkovou pravděpodobnost** – chceme spočítat pst jevu A, když známe pouze podmíněné psti $P(A|H_i)$, kde H_i jsou neslučitelné jevy, jejichž sjednocení je jev jistý, tj. $H_1 \cup H_2 \cup \ldots \cup H_k = \Omega$ a $H_i \cap H_j = \emptyset$ pro všechna i, j

$$P(A) = \sum_{i=1}^{k} P(A|H_i)P(H_i)$$

- **Bayesův vzorec** – jak vypočítat podmíněnou pravděpodobnost $P(A|B)$ ze znalosti $P(B|A)$

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

neboli vzorec v obecné podobě

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_{j=1}^{k} P(A|H_j)P(H_j)}$$

pravděpodobnosti $P(H_i)$ se nazývají *apriorní* a pravděpodobnosti $P(H_i|A)$ *aposteriorní*

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné na

- **Spojitá** – pro číselné proměnné ,
  př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní** – pro kategorické proměnné (mohou být jak nominální, tak ordinální)
  př. binomické, poissonovo, alternativní, . . .

Funkce určující rozdělení

- **Distribuční funkce** – $F(t) = P(X \leq t), t \in \mathbb{R}$
  – neklesající, zprava spojitá, obor hodnot je mezi 0 a 1
- **Pravděpodobnostní funkce** – $p(t) = P(X = t), t \in \mathbb{R}$
  – definovaná pouze pro diskrétní rozdělení
  – nespojitá, nenulová jen v hodnotách, kterých může náhodná veličina nabývat
- **Hustota** – $f(t) = \frac{d}{dt}F(t)$
  – definovaná pouze pro spojitá rozdělení – obdoba pravděpodobnostní funkce, ale nedefinuje konkrétní pravděpodobnosti
  – pravděpodobnost jedné konkrétní hodnoty u spojitého rozdělení je 0
  – derivace funkce distribuční

Další charakteristiky pro diskrétní i spojitá rozdělení

- Střední hodnota

$$E(X) = \sum_{i=1}^{n} X_i p_i, \qquad EX = \int_{-\infty}^{\infty} xf(x)dx$$

- Rozptyl

$$Var(X) = \sum_{i=1}^{n} (X_i - E(X))^2 p_i, \; Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

Binomické rozdělení – zástupce diskrétních 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

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

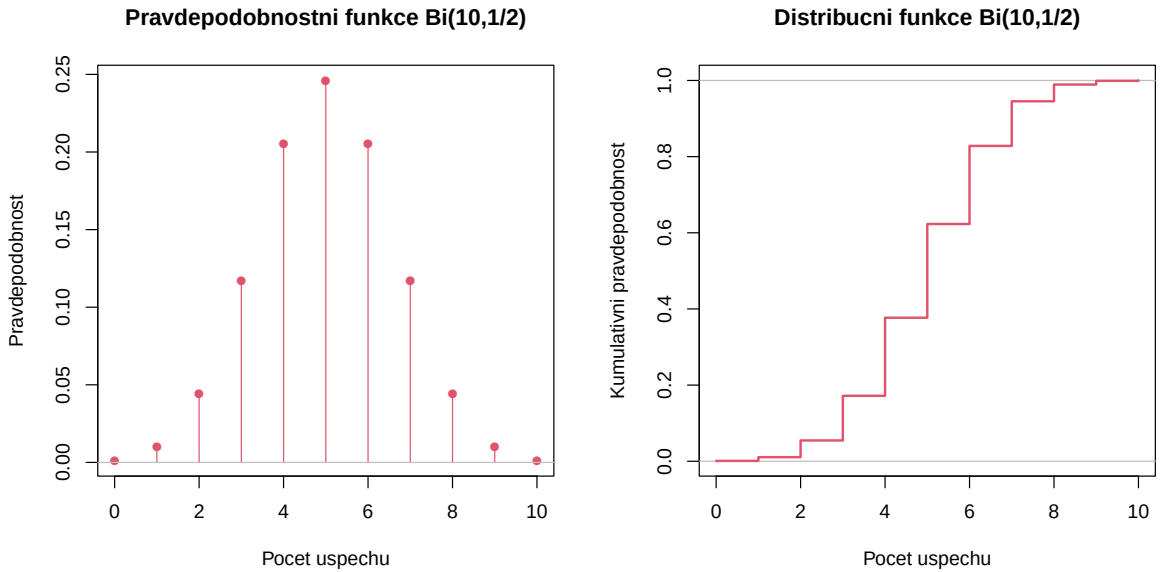
Střední hodnota a rozptyl

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

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

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

Pravděpodobnostní rozdělení

Normální rozdělení – zástupce spojitých 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

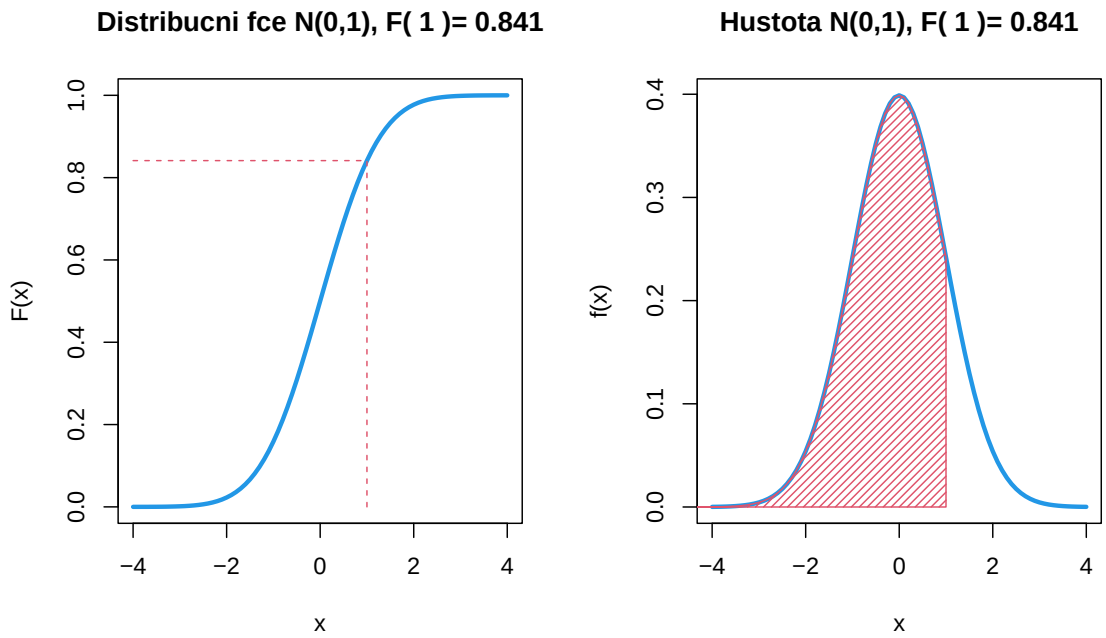
Hustota normálního rozdělení má tvar

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

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

Normální rozdělení

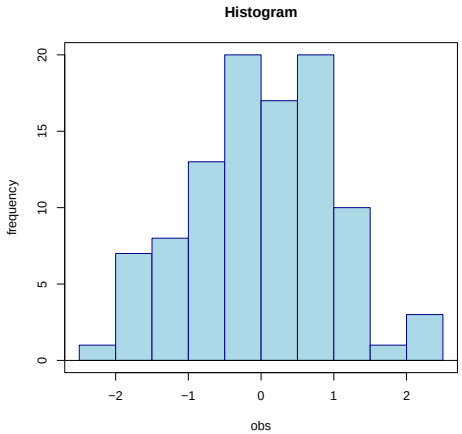
Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota.

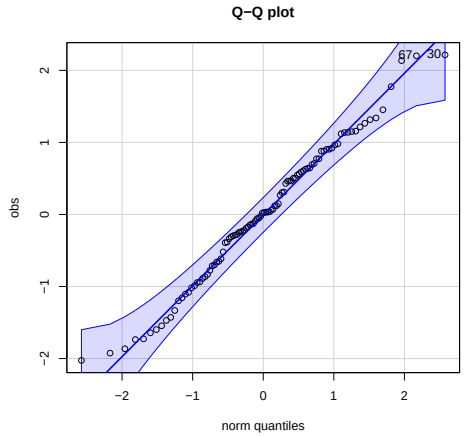


Pravděpodobnostní rozdělení

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

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





- **Číselné testy** – nejčastěji Shapiro-Wilkův test

Bodový odhad střední hodnoty

Příklad. *Mějme situaci, kdy potřebujeme odhadnout průměrnou výšku dospělých lidí v celé České republice. Náhodně jsme vybrali a změřili 500 lidí. Výběrový průměr vyšel 178.12 cm a výběrová směrodatná odchylka 7.9 cm. Odhadněte populační průměr výšky dospělých lidí.*

- nejlepší bodový odhad je výběrový průměr $\overline{X} = 178.12$
- jaká je pravděpodobnost, že se populační průměr bude rovnat přesně tomuto číslu?
- jaká je chyba tohoto odhadu
- střední chyba odhadu průměru

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

Intervalový odhad střední hodnoty

Chceme interval, ve kterém se s vysokou pravděpodobností bude nacházet skutečný populační průměr/ 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 dostanu

Intervalový odhad střední hodnoty

Výpočet intervalu spolehlivosti vychází z faktu, že výběrový průměr má normální rozdělení

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n}),$$

kde μ je odhadovaná teoretická střední hodnota, σ je teoretická směrodatná odchylka a n je počet pozorování.
Když znám skutečný rozptyl dat, pak interval spolehlivosti pro střední hodnotu má tvar

$$\left(\overline{X} - z(1-\alpha/2)\sigma/\sqrt{n}, \overline{X} + z(1-\alpha/2)\sigma/\sqrt{n}\right)$$

kde $z(1-\alpha/2)$ je kvantil standardního normálního rozdělení.

Častější je případ, že rozptyl neznám, pak

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

a meze intervalu spolehlivosti pak jsou

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\mathrm{sd(X)}/\sqrt{\mathrm{n}}, \overline{\mathrm{X}} + \mathrm{t_{n-1}}(1-\alpha/2)\mathrm{sd(X)}/\sqrt{\mathrm{n}}\right)$$

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

Základy testování hypotéz

Je možné říci, že platí následující tvrzení?

- Nový lék je lepší než ten stávající.
- Průměrná výška lidí se za posledních 50 let zvýšila.
- Výnosy z jednotlivých druhů jabloní se liší.
- Krevní tlak závisí na hmotnosti.

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, výnosy druhů jabloní jsou stejné, průměrná výška lidí je 175 cm
- Alternativní hypotézu – značíme H_1
  – obsahuje více možností (např. interval)
  – př. nový lék je lepší než ten stávající, výnosy druhů jabloní se liší, lidé jsou v průměru vyšší než 175 cm
  – není přesně řečeno, jak moc je nový lék lepší, o kolik se liší výnosy jabloní, nebo o kolik jsou lidé vyšší než 175 cm

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

Základy testování hypotéz

S testovou statistikou se většinou pracuje při ručním výpočtu.
Statistické softwary vrací jako výsledek testu *p*-hodnotu.

***p*-hodnota** je

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti H_0
  nastal výsledek, jaký nastal,
  nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- definice *p*-hodnoty se týká testové statistiky

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

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

za platnosti nulové hypotézy má tato statistika t-rozdělení o $n - 1$ stupních volnosti.
Testovou statistiku T porovnáváme s kritickými hodnotami t-rozdělení (tzv. kvantily), na základě čehož buď můžeme přímo rozhodnot o zamítutí nebo nezamítnutí nulové hypotézy, nebo můžeme spočítat p-hodnotu a test vyhodnocovat na základě ní.

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

Jednovýběrový t-test

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 $\alpha = 0.05$.

Jednovýběrový t-test

Pokračování příkladu.
Testová statistika vyšla

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

Tuto hodnotu porovnám s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$. Jelikož testová statistika je v absolutní hodnotě větší než kritická hodnota, zamítám nulovou hypotézu. P-hodnota vyšla $p = 0.005 < 0.05$, což také vede na zamítnutí nulové hypotézy.
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.

- *Jsou otcové v průměru o 10 cm vyšší než matky?*
- *Mají praváci silnější pravou ruku než levou?*
- *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 Vaší pravou ruku se sousedovou levou).
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 222 dětí v jedenáctém a dvanáctém roce věku. Průměrná výška jedenáctiletých vyšla 148.8 cm, u dvanáctiletých pak 154.9 cm. Směrodatná odchylka u jedenáctiletých vyšla 7.1 cm, u dvanáctiletých pak 7.9 cm. Průměrná hodnota rozdílu výšek vyšla 6.1 cm a směrodatná odchylka 2.8 cm. Vyrostly děti mezi jedenáctým a dvanáctým rokem v průměru alespoň o 5 cm?*

Do testové statistiky vkládáme charakteristiky rozdílu (tedy nikoliv rozdíl průměrů, ale průměr rozdílů).

$$T = \frac{\overline{X} - \mu_0}{\text{sd(X)}}\sqrt{n} = \frac{6.1 - 5}{2.8}\sqrt{222} = 5.9$$

Tuto testovou statistiku porovnáváme s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$. Jelikož testová statistika je větší než příslušný kvantil, zamítám nulovou hypotézu. P-hodnota pro tento případ vychází $p = 7.26 * 10^{-9}$, což je menší než $\alpha = 0.05$.

Závěr: Prokázali jsme, že mezi jedenáctým a dvanáctým rokem děti vyrostly v průměru o více než o 5 cm.

Dvouvýběrový t-test

Porovnáváme-li střední hodnotu dvou **nezávislých** výběrů, používá se **dvouvýběrový test**.
Testová statistika má tvar

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

kde S je střední chyba rozdílu průměrů. Tato střední chyba se počítá jinak, když oba výběry mají stejné rozptyly, a když je mají různé. V případě stejných rozptylů je S následující

$$S = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \right) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

n_1, n_2 je rozsah výběru X, respektive Y.
Za platnosti nulové hypotézy má tato statistika t-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.
Předpokladem použití dvouvýběrového testu je normalita dat v obou výběrech.

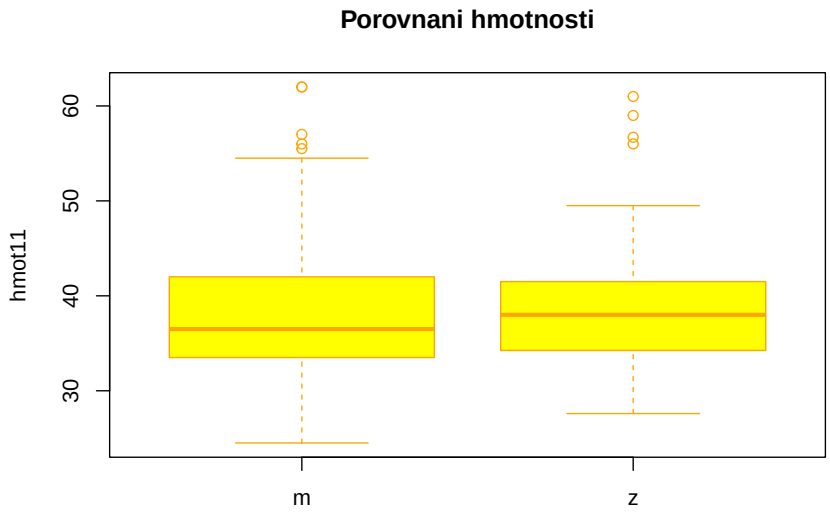
Dvouvýběrový t-test

Příklad. *Ve výběru mám 222 jedenáctiletých dětí, z toho 159 hochů a 63 dívek. Průměrná hmotnost hochů vyšla 38.1 kg a u dívek 39.1. Směrodatná odchylka pro hochy vyšla 6.7 kg a pro dívky 7.1. Je hmotnost jedenáctiletých dětí u obou pohlaví stejná?*

Testované hypotézy

- H_0 : hmotnost hochů a hmotnost dívek se neliší
- H_1 : hmotnost hochů a dívek se liší

Grafické porovnání



Dvouvýběrový t-test

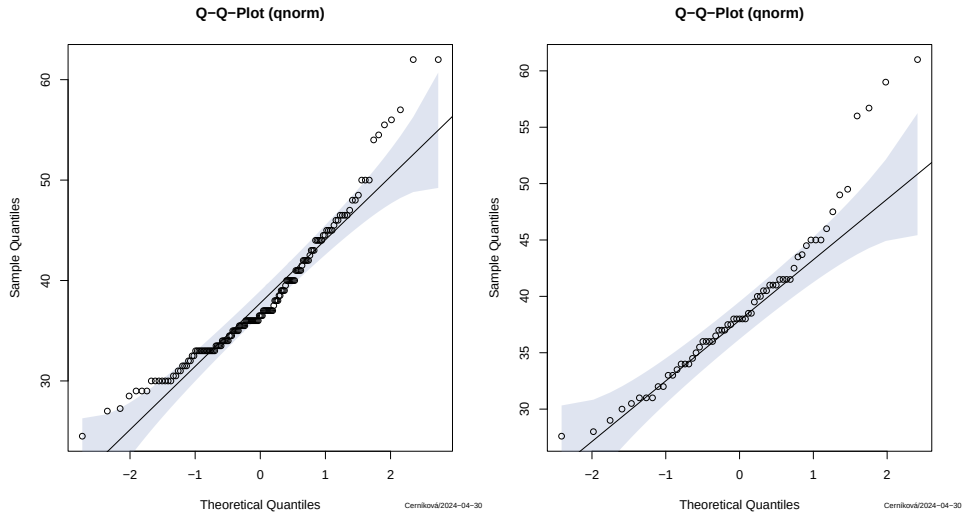
Testová statistika testu vychází

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S} = \frac{38.1 - 39.1}{1.0168} = -1.001$$

Tomu odpovída p-hodnota 0.3151. P-hodnota je větší než $\alpha = 0.05$, nulovou hypotézu nezamítám

Závěr: Na hladině významnosti 5% jsem neprokázala, že by se hmotnost jedenáctiletých hochů a dívek lišila.

Kontrola normality.



Analýza rozptylu – ANOVA

Porovnáváme-li střední hodnotu ve více než ve dvou nezávislých výběrech, používá se **analýza rozptylu**. Vždy se testují následující hypotézy

- H_0 : všechny střední hodnoty jsou stejné
- H_1 : alespoň jedna střední hodnota se liší

Myšlenka spočívá v porovnání variability mezi výběry s variabilitou v rámci výběrů.

Stejně jako u dvouvýběrového testu budeme brát pouze analýzu rozptylu pro normálně rozdělená data

Analýza rozptylu – ANOVA

Označme X_{ij} i-té pozorování z j-tého výběru, $\overline{X}_{i.}$ průměr i-tého výběru, $\overline{X}_{..}$ celkový průměr všech pozorování, n_i rozsah i-tého výběru a k počet výběrů.

Analýza rozptylu rozkládá celkovou variabilitu

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

na variabilitu vysvětlenou výběry (mezi výběry) SS_A a variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SS_e. Platí

$$\begin{aligned} SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{..})^2 = \\ &= \sum_{i=1}^{k} n_i (\overline{X}_{i.} - \overline{X}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2 = \\ &= SSA + SSe \end{aligned}$$

Analýza rozptylu – ANOVA

Výstupem z analýzy rozptylu je tzv. **tabulka analýzy rozptylu**

	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	p-hodnota
Faktor A	SSA	$dfA = k - 1$	$MSA = \frac{SSA}{dfA}$	$F = MSA/MSe$	p
Chyba e	SSe	$dfe = n - k$	$MSe = \frac{SSe}{dfe}$		
Celkem	SST	$dft = n - 1$			

Za platnosti nulové hypotézy má testová statistika F-rozdělení o $k - 1$ a $n - k$ stupních volnosti.

Párové srovnání

Zajímá-li nás, které konkrétní dvojice výběrů se od sebe významně liší, nelze toto zjistit větším počtem běžných dvouvýběrových testů, neboť by tím příliš vzrostla chyba prvního druhu (tj. neudržela by se celková hladina významnosti). Je nutné použít párové srovnání, např. **Tukeyův test**.
Testuje se

- H_0 : střední hodnoty μ_i a μ_j jsou stejné
- H_1 : střední hodnoty μ_i a μ_j se liší

pro všechny dvojice i a j.
Testová statistika má tvar

$$Q = \frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{s^*}, \text{ kde } s^* = \sqrt{\frac{SSe}{n(n-k)}}$$

Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní kritické hodnoty.

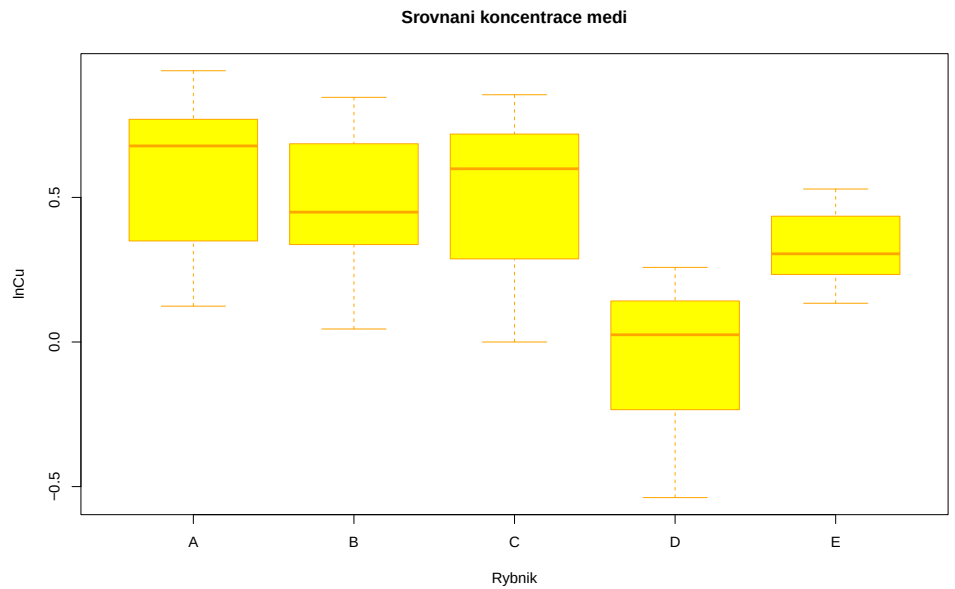
Analýza rozptylu – ANOVA

Příklad. *Byla měřena koncentrace mědi v těle ryb. Porovnáváno bylo 5 rybníků, kde z každého byl vyloven vzorek 7-mi ryb. Výběrové rozptyly 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

Grafické porovnání



Analýza rozptylu – ANOVA

Abychom mohli vybrat správnou verzi analýzy rozptylu, otestujme nejprve shodu rozptylů ve všech výběrech. Tyto rozptyly vyšly postupně 0.10, 0.08, 0.10, 0.08 a 0.02. Testujeme

- H_0 rozptyly jsou shodné
- H_1 rozptyly se liší

Testová statistika Bartlettova testu vyšla 3.67 při čtyřech stupních volnosti, což dává p-hodnotu 0.45. Jelikož je p-hodnota větší než $\alpha = 0.05$, nulovou hypotézu nezamítáme a můžeme použít klasickou ANOVU pro shodné rozptyly.

Analýza rozptylu – ANOVA

Tabulka analýzy rozptylu vyšla

	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	*p*-hodnota
Rybník	1.796	4	0.4491	5.896	0.00127
Chyba	2.285	30	0.0762		
Celkem	4.081	34			

P-hodnota vyšla menší než $\alpha = 0.05$, což znamená, že nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší.

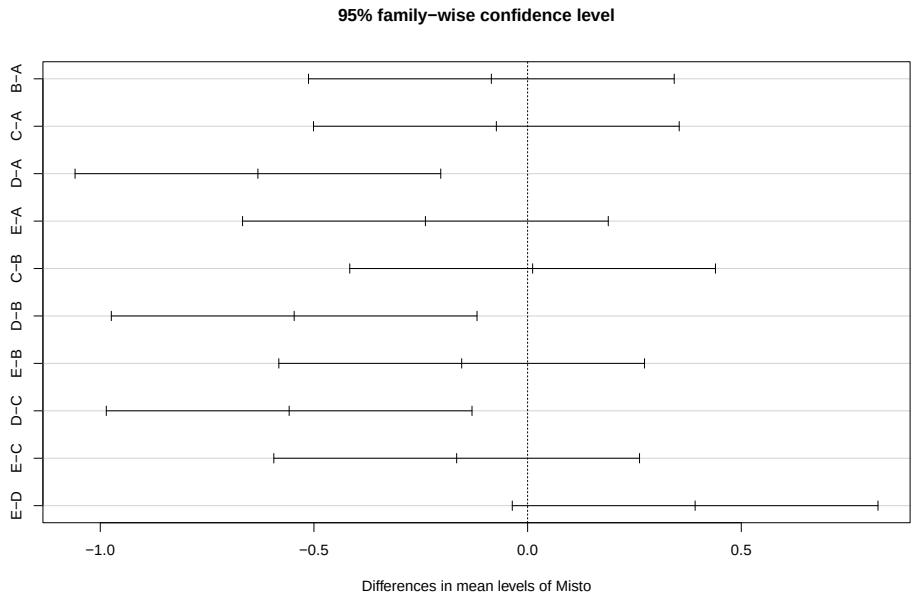
Analýza rozptylu – ANOVA

Párové srovnání vrátí následující tabulku

	rozdíl	dolní mez	horní mez	p-hodnota
B-A	-0.08485714	-0.51274077	0.3430265	0.9777112
C-A	-0.07314286	-0.50102648	0.3547408	0.9871500
D-A	-0.63114286	-1.05902648	-0.2032592	**0.0015454**
E-A	-0.23914286	-0.66702648	0.1887408	0.4960690
C-B	0.01171429	-0.41616934	0.4395979	0.9999904
D-B	-0.54628571	-0.97416934	-0.1184021	**0.0070956**
E-B	-0.15428571	-0.58216934	0.2735979	0.8319549
D-C	-0.55800000	-0.98588362	-0.1301164	**0.0057762**
E-C	-0.16600000	-0.59388362	0.2618836	0.7920009
E-D	0.39200000	-0.03588362	0.8198836	0.0850175

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.