Statistika v Biologii

Alena Černíková

alena.cernikova@ujep.cz

8. listopadu 2024

Podmínky zápočtu

- **tři domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat přes moje internetové stránky – kapitola Statistika v Biologii, odkaz Úkoly
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  zpracování vzájemného vztahu tří proměnných
  souvislý text obsahující všechny podstatné statistické informace

Obsah kurzu

Co je statistika

Statistika je přesná věda o nepřesných číslech.

Cílem je zjistit chování náhodné veličiny v určité populaci. Celou populaci změřit neumíme. Uděláme náhodný výběr, na kterém změříme sledovanou veličinu, a na základě zjištěných informací děláme závěry pro celou populaci.

Příklad. *Zajímá nás průměrná výška dospělých lidí v celé České republice. Všechny dospělé lidi změřit nemumíme, uděláme náhodný výběr o cca 200 lidech a na základě získaných výsledků se snažíme celkovou průměrnou výšku odhadnout. Průměrná výška pro těchto 200 lidí vyšla 175 cm.*

Základní pojmy

- **Nahodná veličina** – jakákoliv veličina, kterou můžeme měřit opakovaně, např. výška
- **Populace** – soubor, pro nejž chceme udělat nějaký závěr, např. všichni dospělí obyvatelé České republiky
- **Náhodný výběr** – v porovnání s populací malý soubor pozorování získaný náhodně, jde o nezávislé, stejně rozdělené náhodné veličiny, např. výběr 200 lidí
- **Populační charakteristika** – charakteristika popisující populaci, např. populační průměr
- **Výberová charakteristika** – charakteristika spočítaná na výběru, pomocí níž odhadujeme populační ekvivalent, např. výběrový průměr.

Typy proměnných

Náhodná veličina se často nazývá proměnná.

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

- **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 charakteristiky – krabicový graf, histogram
- **Nominální proměnné**
  - číselné charakteristiky – 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

Čištění dat

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 hodnot

Jak vypočítat **kvartily**

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

Jak vypočítat **extrémy**

- minimum a maximum

Popisné statistiky polohy

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

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

- p – díl dat, které chcete p-tým percentilem oddělit
  číslo mezi 0 a 1
- $X_{(k)}$ – 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 (odlehlá 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*

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

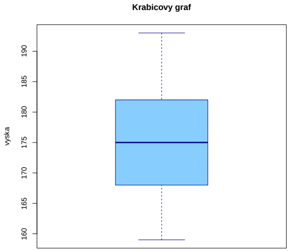
kde n je počet pozorování

Příklad

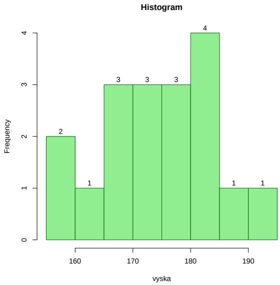
Popisné statistiky polohy – výsledky

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

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

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

Popisné statistiky tvaru rozdělení

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

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

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

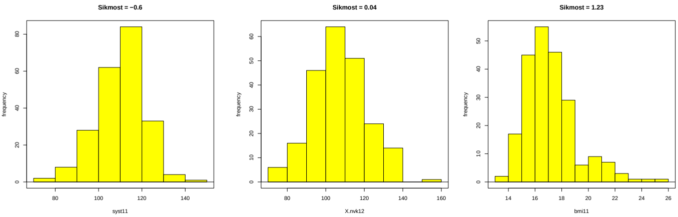
$$\mathrm{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\mathrm{sd}(X)}\right)^3 = \frac{\sum_{i=1}^{n} Z_i^3}{n}$$

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

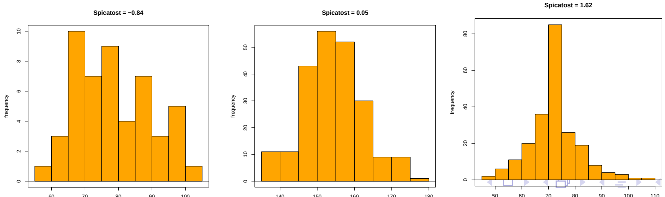
$$\mathrm{Kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\mathrm{sd}(X)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



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



Pokračování příkladu

Popisné statistiky variability – výsledky

- rozptyl – 88.81 cm^2
- směrodatná odchylka – 9.42 cm
- mezikvartilové rozpětí – 13.75 cm
- variační koeficient – 0.054

Popisné statistiky tvaru rozdělení – výsledky

- šikmost – 0.027
  hodnota blízká nule, téměř symetrické rozdělení
- špičatost – -1.04
  záporná hodnota, pložší rozdělení, než je rozdělení normální

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.

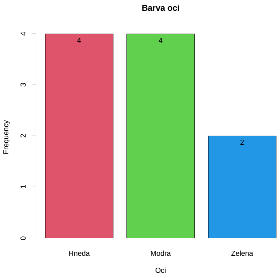
Barva	Absolutní	Relativní (%)
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

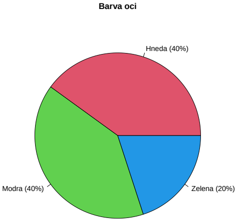
Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

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

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 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

Příklad

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*

Jak se vypočte **pravděpodobnost náhodného jevu** A?

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

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, $P(\overline{A}) = 1 - P(A)$
  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

Vztah dvou jevů

- **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*

Bayesův vzorec

- **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*

Senzitivita a specificita testu

Charakteristiky medicínských testů označují podmíněné pravděpodobnosti

- **Senzitivita testu** – pravděpodobnost, že test vyjde pozitivně, pokud je osoba nemocná
  $P(\text{test je pozitivní}|\text{osoba je nemocná})$
- **Specificita testu** – pravděpodobnost, že test vyjde negativně, pokud je osoba zdravá
  $P(\text{test je negativní}|\text{osoba je zdravá})$

Senzitivita a specificita testu

Příklad. *Výzkumu se zúčastnilo 2000 pacientů, z nichž 50 bylo HIV pozitivních. Všichni podstoupili test na HIV. Test vyšel pozitivní pro 45 pozitivních pacientů a pro 200 negativních. Spočtěte senzitivitu a specificitu testu a také pravděpodobnost, že člověk bude skutečně HIV pozitivní, pokud mu vyjde pozitivní test.*

		Skutečnost Pozitivní	Negativní	Celkem
Test	Pozitivní	45	200	245
	Negativní	5	1750	1755
	Celkem	50	1950	2000

- **Senzitivita testu** – $P(\text{test je pozitivní}|\text{osoba je nemocná}) = 45/50 = 0.9$
- **Specificita testu** – $P(\text{test je negativní}|\text{osoba je zdravá}) = 1750/1950 = 0.897$
- **Jsem nemocný, když mám pozitivní test?** – $P(\text{osoba je nemocná}|\text{test je pozitivní}) = 45/245 = 0.184$
  Pomocí Bayesovy věty

$$P(ON|TP) = \frac{P(ON \cap TP)}{P(TP)} = \frac{P(TP|ON)P(ON)}{P(TP|ON)P(ON) + P(TP|OZ)P(OZ)} =$$

$$= \frac{\text{Senzitivita} * \text{podíl nemocných}}{\text{Senzitivita} * \text{podíl nemocných} + (1 - \text{Specificita}) * \text{podíl zdravých}} = \frac{0.9 * 0.025}{0.9 * 0.025 + 0.102 * 0.975} = 0.184$$

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné na

- **Spojitá** – pro číselné proměnné nabývající všech reálných hodnot v určitém intervalu,
  př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní** – pro číselné proměnné které nabývají jasně oddělitelných hodnot
  mohou nabývat i nekonečně hodnot, najčastěji počty
  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
  – definovaná pro všechna rozdělení
- **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

Střední hodnota a rozptyl

Další charakteristiky pro diskrétní i spojitá rozdělení

- Střední hodnota

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

- 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 (diskrétní)

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**

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

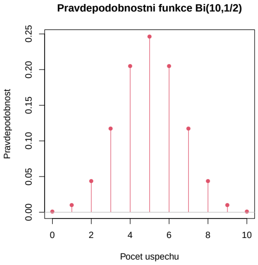
Střední hodnota a **rozptyl**

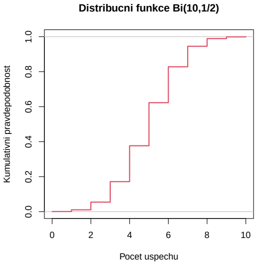
$$\mathrm{E(X) = np,} \qquad \mathrm{Var(X) = np(1 - p)}$$

Binomické rozdělení (diskrétní)

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 (spojité)

Známé rozdělení s hezkými vlastnostmi. Toto rozdělení má výška lidí určitého věku, IQ, ...

Značení $N(\mu, \sigma^2)$, kde

- μ – střední hodnota
- σ^2 – rozptyl

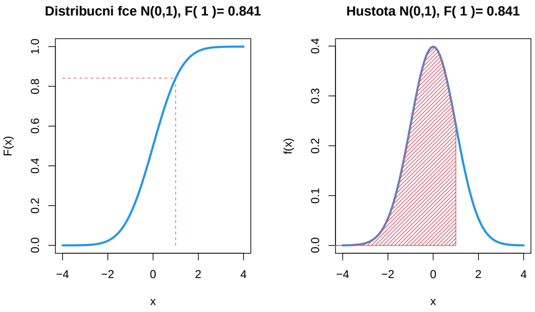
Hustota normálního rozdělení má tvar

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

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

Normální rozdělení (spojité)

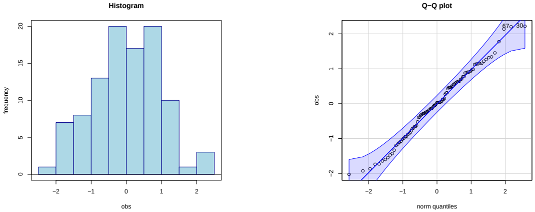
Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota.



Testování normality

Existují statistické postupy určené pro normální rozdělení a takové, které normalitu nepožadují.
Je tedy potřeba umět normalitu otestovat

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



- **Číselné testy** – např. Shapiro-Wilkův, Andersonův-Darlingův, Kolmogorovův-Smirnovův, Lillieforsův a další

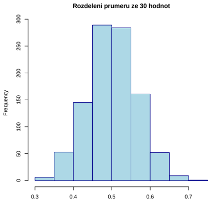
Centrální limitní věta

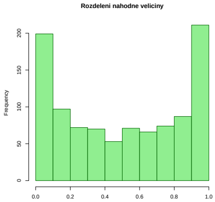
Věta

Rozdělení součtu nezávislých, stejně rozdělených náhodných veličin konverguje k normálnímu pro počet těchto náhodných veličin rostoucí nade všechny meze.

Tedy: čím více hodnot sčítáte/průměrujete, tím spíše bude mít průměr normální rozdělení.

Ukázka, jak vypadá rozdělení průměru 30-ti hodnot z beta rozdělení v porovnání s rozdělením samotným.





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 173.12 cm a výběrová směrodatná odchylka 8.9 cm. Odhadněte populační průměr výšky dospělých lidí.*

- **nejlepší bodový odhad** je výběrový průměr $\overline{X} = 173.12$ cm
- jaká je pravděpodobnost, že se populační průměr bude rovnat přesně tomuto číslu?
- chyba odhadu, tzv. střední chyba odhadu průměru

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

Intervalový odhad střední hodnoty

Chceme interval, ve kterém se s vysokou pravděpodobností bude nacházet populační průměr/ skutečná střední hodnota.

Z čeho se interval spolehlivosti počítá

- 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 mám a tím užší bude 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

Výběrový průměr má normální rozdělení

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- μ je teoretická střední hodnota
- σ je teoretická směrodatná odchylka
- n je počet pozorování
- $N(0, 1)$ je standartní normální rozdělení

Interval spolehlivosti pro střední hodnotu, když **znám** skutečný **rozptyl** dat

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

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

Intervalový odhad střední hodnoty

Běžnější je případ, že **rozptyl neznám** a musím ho nahradit výběrovou směrodatnou odchylkou. Pak platí

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

- t_{n-1} značí t-rozdělení o $n-1$ stupních volnosti

meze intervalu spolehlivosti jsou

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

- $t_{n-1}(1-\alpha/2)$ je kvantil t-rozdělení o $n-1$ stupních volnosti
- kvanlity t-rozdělené jsou větší než kvantily normálního rozdělení
- interval spolehlivosti je širší než v předchozím případě

Intervalový odhad pravděpodobnosti

Chceme odhadnout parametr binomického rozdělení (p).

Příklad. *Házíme 100 krát kostkou, z těchto 100 hodů mi šestka padla 20 krát. Jaká je pravděpodobnost, že padne 6.*

- **nejlepším bodovým odhadem** je relativní četnost $\hat{p} = n_u/n$
  - n_u je počet úspěchů
  - n je počet pokusů
- **intervalový odhad** vychází z předpokladu, že

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

  pro $n\hat{p}(1-\hat{p}) > 9$
- meze intervalu spolehlivosti jsou

$$\left(\hat{p} - z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Základy testování hypotéz

Testování hypotéz se využívá, když potřebujeme ověřit platnost nějakého tvrzení, např.

- Nový lék je lepší než ten stávající.
- Průměrná výška lidí je 175 cm.
- Průměrná výška lidí se za posledních 50 let zvýšila.
- Výnosy z jednotlivých druhů jabloní se liší.
- Krevní tlak závisí na hmotnosti.

Testované hypotézy

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme H_0
  – je v ní vždy pouze jedna varianta, např.
  - nový lék je stejný jako ten stávající (rozdíl je roven nule)
  - výška se rovná 175 cm
  - výnosy všech druhů jabloní jsou stejné (rozdíl je roven nule)
- Alternativní hypotézu – značíme H_1
  – obsahuje více možností (interval hodnot), např.
  - nový lék je lepší než ten stávající (rozdíl je větší než nula)
  - výška se liší od 175 cm
  - výnosy druhů jablonÍ se liší (rozdíl se nerovná nule)

Závěr testu

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

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

Jiný závěr udělat nemohu, z čehož plyne, že

- to co mě zajímá (to, co chci prokázat), **musí být v alternativě**
- platnost nulové hypotézy nelze prokázat
- na interpretaci závěru testu je kladen speciální důraz

.

Chyby při 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
- **chyba druhého druhu** – nezamítneme H_0, přestože neplatí
  – značí se β a hodnota $1 - \beta$ se nazývá síla testu
  – za dané hladiny významnosti chceme test co nejsilnější

	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

Postup testování

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

Testované hypotézy

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

Testujeme nulovou hypotézu

- 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 $\alpha = 0.05$

Testová statistika

Testová statistika jednovýběrového t-testu je

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

- za platnosti nulové hypotézy má t-rozdělení o $n - 1$ stupních volnosti
- její hodnotu porovnáváme s kvantily t-rozdělení (kritické hodnoty)
- je-li rozdíl mezi tím, co jsme naměřili ($\overline{X}$) a nulovou hypotézou μ_0 příliš velký, nulovou hypotézu ZAMÍTÁME

Předpokladem jednovýběrového t-testu je, že průměr testované veličiny má normální rozdělení (díky CLV většinou splněno).

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

Vyhodnocení testu

- testová statistika vyšla

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

- porovnám ji s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$
- jelikož $|T| = 2.5618 > 1.65$, tak nulovou hypotézu zamítáme
- p-hodnota vyšla $p = 0.005 < 0.05 \Rightarrow$ zamítáme H_0
- nakonec je potřeba ověřit normalitu

Závěr: Prokázali jsme, že průměrná výška jedenáctiletých dětí je menší než 150 cm.

Párový 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).

Postup testování

- V prvním kroku jsou pro všechny páry vypočteny **rozdíly**:

$$R_i = X_i - Y_i$$

- střední hodnota těchto rozdílů je testována jednovýběrovým testem

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 rozdílů výšek vyšla 6.1 cm a směrodatná odchylka 2.8 cm. Vyrostly děti mezi jedenáctým a dvanáctým rokem v průměru alespoň o 5 cm?*

Testované hypotézy

- H_0 : výška ve 12 letech – výška v 11 letech $=$ 5 cm
- H_1 : výška ve 12 letech – výška v 11 letech $>$ 5 cm

Párový t-test

Vyhodnocení testu

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

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

- porovnáváme ji s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$
- jelikož $|T| = 5.9 > 1.65$ tak, nulovou hypotézu zamítáme
- p-hodnota vyšla $p = 7.26 \times 10^{-9} < 0.05 \Rightarrow$ zamítáme H0
- nakonec je potřeba ověřit normalitu rozdílu

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

Dvouvýběrový test

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

Příklady

- Je nový lék lepší než ten stávající?
- Jsou muži vyšší než ženy?
- Liší se od sebe dva druhy hnojiva?

Testované hypotézy

- H_0 : průměr 1. skupiny – průměr 2. skupiny $= \mu_0$
- H_1 : průměr 1. skupiny – průměr 2. skupiny $\neq \mu_0$, $> \mu_0, < \mu_0$

Dvouvýběrový test testuje rozdíl průměrů.

Dvouvýběrový *t*-test

Budeme zde brát dva typy dvouvýběrového testu:

- Dvouvýběrový t-test pro shodné rozptyly
- Welchův dvouvýběrový test pro různé rozptyly

Oba tyto testy předpokládají normální rozdělení obou průměrů.

Test shody rozptylů

K tomu, abychom mohli vybrat správnou verzi testu, je potřeba otestovat **shodu rozptylů** v obou výběrech.
Testované hypotézy

- H_0 : rozptyly se ve výběrech neliší
- H_1 : rozptyly se ve výběrech liší

Testová statistika je

$$F = \frac{\mathrm{Var(X)}}{\mathrm{Var(Y)}}$$

za platnosti H_0 má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti, kde n_1, n_2 jsou rozsahy prvního, respektive druhého výběru.

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika testu má tvar

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

kde

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

za platnosti nulové hypotézy má tato statistika t-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.

Welchův t-test

Testová statistika testu má tvar

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{\mathrm{Var(X)}}{n_1} + \frac{\mathrm{Var(Y)}}{n_2}}}$$

a za platnosti nulové hypotézy má t-rozdělení o ν stupních volnosti, kde

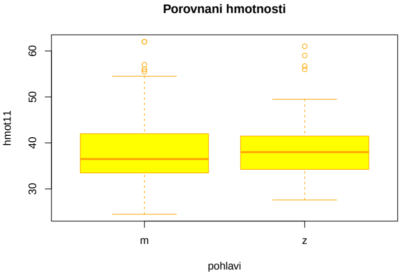
$$\nu = \frac{(\mathrm{Var(X)}/\mathrm{n}_1 + \mathrm{Var(Y)}/\mathrm{n}_2)^2}{\frac{(\mathrm{Var(X)}/\mathrm{n}_1)^2}{n_1 - 1} + \frac{(\mathrm{Var(Y)}/\mathrm{n}_2)^2}{n_2 - 1}}.$$

kritické hodnoty je možno odvodit, přestože ν není celé číslo.

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 v průměru stejná pro hochy jako pro dívky?

Grafické porovnání



Test shody rozptylů

Testované hypotézy

- H_0 : rozptyly jsou stejné
- H_1 : rozptyly se liší

Vyhodnocení testu

- test shody rozptylů má testovou statistiku

$$F = \frac{\mathrm{Var(X)}}{\mathrm{Var(Y)}} = \frac{45.1}{50.6} = 0.89$$

- p-hodnota vyšla $p = 0.56 > 0.05$ a nulovou hypotézu nezamítáme
- rozptyly ve skupinách jsou přibližně stejné
- můžeme použít dvouvýběrový t-test pro shodné rozptyly

Dvouvýběrový t-test

Testované hypotézy

- H_0 : hmotnost hochů a hmotnost dívek se neliší
  hmotnost hochů – hmotnost dívek = 0
- H_1 : hmotnost hochů a dívek se liší
  hmotnost hochů – hmotnost dívek $\neq$ 0

Dvouvýběrový t-test

Vyhodnocení testu

- Testová statistika testu vychází

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{38.1 - 39.1}{6.83}\sqrt{\frac{159 * 63}{159 + 63}} = -1.001$$

- porovnáme ji s kvantilem t-rozdělení $t_{220}(1 - 0.025) = 1.97$
- jelikož $|T| = 1.001 < 1.97$ tak nulovou hypotézu nezamítáme
- p-hodnota vyšla $p = 0.3151 > 0.05 \Rightarrow$ nezamítáme H_0
- nakonec je potřeba ověřit normalitu v obou výběrech

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

Analýza rozptylu – ANOVA

Porovnáváme-li střední hodnotu ve více než ve dvou nezávislých výběrech, používá se **analýza rozptylu.** Testujeme

- 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.

Analýza rozptylu – ANOVA

Stejně jako u dvouvýběrového t-testu, budeme i zde brát dvě varianty analýzy rozptylu

- klasická analýza rozptylu (ANOVA) – pro shodné rozptyly ve všech výběrech
- Welchova verze analýzy rozptylu – pro různé rozptyly ve výběrech

Předpokladem obou těchto verzí ANOVY je normalita dat ve všech výběrech či ekvivalentně normalita residuí odpovídajícího lineárního modelu.

Klasická 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
- k počet výběrů

Analýza rozptylu rozkládá celkovou variabilitu (čitatel rozptylu)

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

Klasická ANOVA

Rozklad variability na

- variabilitu vysvětlenou výběry (mezi výběry) SS_A
- variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SS_e

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

Klasická 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.

Welchova ANOVA

Welchova ANOVA je založena na vážené variabilitě mezi výběry

$$F \frac{\frac{1}{k-1}\sum_{j=1}^{k} w_j(\overline{X}_j - \overline{X}_w)^2}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left(\frac{1}{n_j-1}\right)\left(1 - \frac{w_j}{w}\right)^2}$$

kde

$$w_j = \frac{n_j}{s_j^2}, \quad w = \sum_{j=1}^{k} w_j, \quad \overline{X}_w = \frac{\sum_{j=1}^{k} w_j \overline{X}_j}{w}$$

Testová statistika má F-rozdělení o $k-1$ a ν stupních volnosti

$$\nu = \frac{k^2-1}{3\sum_{j=1}^{k}\left(\frac{1}{n_j-1}\right)\left(1 - \frac{w_j}{w}\right)^2}$$

Bartlettův test

Test shody rozptylů pro více než 2 rozptyly.
Testované hypotézy

- H_0 : rozptyly jsou shodné
- H_1 : rozptyly se liší

Označme $\mathrm{Var}(\mathrm{X_j})$ výběrový rozptyl v j-tém výběru a

$$
\begin{aligned}
S^2 &= \frac{\sum_{j=1}^{k}(n_j-1)\mathrm{Var}(\mathrm{X_j})}{n-k},\\
C &= 1+\frac{1}{3(k-1)}\left(\sum_{j=1}^{k}\frac{1}{n_j-1}-\frac{1}{n-k}\right)
\end{aligned}
$$

Testová statistika Bartlettova testu

$$
B=\frac{1}{C}\left((n-k)\ln S^2-\sum_{j=1}^{k}(n_j-1)\ln \mathrm{Var}(\mathrm{X_j})\right)
$$

má za platnosti nulové hypotézy χ^2-rozdělení o $k-1$ stupních volnosti.

Párové srovnání

Porovnání dvojic výběrů

- nelze pomocí dvouvýběrových testů – příliš roste chyba prvního druhu
- prostřednictvím tzv. párového srovnání
- nejčastěji **Tukeyho test**, respektive **Tukey HSD test** pro různě velké výběry

Testované hypotézy

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

Párové srovnání

Testová statistika má tvar

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

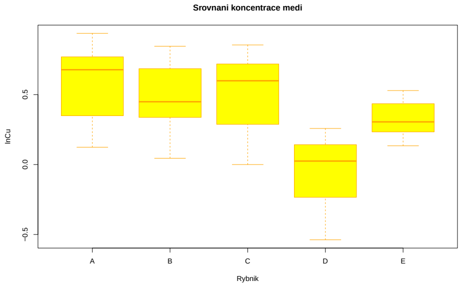
Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty. Vyhodnocení

- na základě p-hodnot nebo intervalú spolehlivosti pro rozdíly všech dvojic
- přestože ANOVA ukáže významný rozdíl mezi skupinami, nemusí se tento projevit v párovém srovnání

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 průměry logaritmizované koncentrace mědi pro jednotlivé rybníky vyšly 0.57, 0.48, 0.50, -0.06 a 0.33. Liší se od sebe tyto rybníky?*

Grafické porovnání středních hodnot



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

Testované hypotézy

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

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.
Nakonec je potřeba ověřit normalitu residuí nebo jednotlivých výběrů

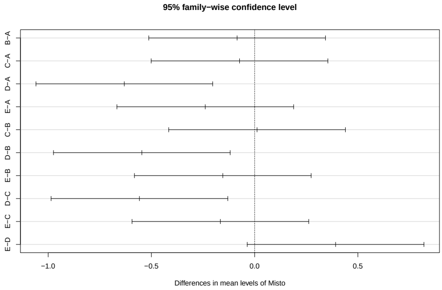
Analýza rozptylu – ANOVA

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.

Pearsonův korelační koeficient

Lineární vztah dvou číselných proměnných zkoumá **korelační koeficient. Pearsonův korelační koeficient** vypočteme jako

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

Libovolný korealční koeficient nabývá hodnot mezi -1 a 1 a platí, že

- absolutní nepřímá závislost má $\mathrm{Cor}(X,Y)=-1$
- lineární nezávislost/ nekorelovanost má $\mathrm{Cor}(X,Y)=0$
- absolutní přímá závislost má $\mathrm{Cor}(X,Y)=1$

Pearsonův korelační koeficient

O statistické významnosti závislosti rozhodujeme testem

- H_0 : korelační koefcient $= 0$
- H_1 : korelační koefcient $\neq 0$,
  H_1 : korelační koefcient > 0,
  H_1 : korelační koefcient < 0

Za platnosti nulové hypotézy platí, že testová statistika

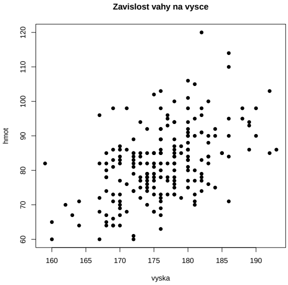
$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

má t-rozdělení o $n - 2$ stupních volnosti.
Předpokladem použití Pearsonova korelačního koeficientu je normalita obou prpoměnných

Pearsonův korelační koeficient

Příklad. *Výzkumu se účastnilo 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční, u nichž byly měřeny fyzické údaje. Souvisí spolu výška a hmotnost těchto mužů?*

Nejprve grafické porovnání



Z grafu je patrná rostoucí závislost mezi oběma proměnnými.

Pearsonův korelační koeficient

Testované hypotézy

- H_0 : váha a výška spolu nesouvisí, korelační koeficient $= 0$
- H_1 : váha a výška spolu souvisí, korelační koeficient $\neq 0$

Korelační koeficient vyšel 0,5 a testová statistika

$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}}\sqrt{(n-2)} = \frac{0.5}{\sqrt{1-0.25}}\sqrt{202} = 8.19.$$

Testová statistika je větší než kvantil t-rozdělení $t_{202}(1 - 0.975) = 1.97$. P-hodnota testu vyšla 2.926×10^{-14}, což je menší než $\alpha = 0.05$. Nulovou hypotézu tedy zamítáme.
Závěr: Závislost mezi váhou a výškou je průkazná a přímá.
Nakonec je potřeba ověřit normalitu obou proměnných.

Lineární regrese

Vztah mezi dvěma spojitými proměnnými lze hodnotit i z pohledu **lineární regrese**, která zkoumá příčinnou závislost. V tomto případě máme

- **nezávisle proměnnou** X – příčinu
- **závisle proměnnou** Y – důsledek

Výsledkem je odhad lineárního modelu ve tvaru

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

kde

- 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

Graficky popisujeme pomocí bodového grafu, ale není jedno, která proměnná je na které ose

- na x-ovou osu se kreslí nezávisle proměnná
- na y-ovou osu se kreslí závisle proměnná

Odhad probíhá **metodou nejmenších čtverců**, která minimalizuje součet druhých mocnin residuí

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

- $\hat{Y}_i$ se nazývají odhady, nebo též predikce
- b_0, b_1 jsou odhady regresních koeficientů
- pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné

Koeficient determinace

Často nazývaný *R-squared*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \mathrm{cor}(X, Y)^2$$

- ukazatel kvality modelu
- procento variability závisle proměnné vysvětlené modelem tj. z kolika procent závisí Y na X a z kolika na něčem jiném

Regresní koeficienty

Interpretace regresních koeficientů

- b_0 – kde protíná regresní přímka x-ovou osu
  – kolik by vyšlo Y, kdyby X bylo 0
- b_1 – o kolik se v průměru změní závisle proměnná Y, když se nezávisle proměnná X zvýší o 1

Test nezávislosti

- H_0 : Y na X lineárně nezávisí, $\beta_1 = 0$
- H_1 : Y na X lineárně závisí, $\beta_1 \neq 0$

Test je založen na faktu, že $b_1/\mathrm{se(b_1)} \sim \mathrm{N}(0, 1)$, kde b_1 je odhad lineárního členu β_1 a $\mathrm{se(b_1)}$ je jeho střední chyba.

- test vychází stejně jako u korelačního koeficientu

Lineární regrese

Příklad. *Pokračujme příkladem závislosti hmotnosti na výšce u mužů s jedním rizikovým faktorem ischemické choroby srdeční.*

Odhadli jsme model ve tvaru

$$Y_i = -66.85 + 0.85X_i$$

- se(b_1) vyšla 0.1 , testová statistika 8.19, kvantil t-rozdělení $t_{202}(1 - 0.975) = 1.97$
- koeficient determinace je $R^2 = 0.25$
- p-hodnota testu vyšla $2.93 \times 10^{-14} < \alpha = 0.05$
- tedy zamítáme nulovou hypotézu

Závěr: U mužů s jedním rizikovým faktorem ischemické choroby srdeční závisí hmotnost na výšce. Závislost je přímá a vysvětlí se jí 25% variability závisle proměnné.

Test dobré shody

Kategorická proměnná s více než dvěma kategoriemi má tzv. **Multinomické rozdělení**. Jedná se o zobecnění binomického rozdělení.

Označme

- k počet kategorií, kterých může náhodná veličina nabývat
- n počet pokusů pokus/ pozorování
- $X_1, \ldots, X_k$ počty, kolikrát nastala která kategorie v n pokusech

Pravděpodobnosti multinomického rozdělení jsou

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k}$$

Střední hodnota a rozptyl

$$\mathrm{E}(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i)$$

Test dobré shody

Test o hodnotách jednotlivých pravděpodobností.
Testované hypotézy

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

Testová statistika má tvar

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

- za platnosti nulové hypotézy má χ^2-rozdělení o $k-1$ stupních volnosti
- předpokladem je, že všechny očekávané četnosti , tj. hodnoty $n\pi_i$, jsou větší než 5.

Tímto testem je možné testovat i konkrétní rozdělení veličiny.

Test dobré shody

Příklad. *Házíme 50 krát šestistěnnou kostkou a počítáme, kolikrát padla která hodnota. Jednička padla 8 krát, dvojka 5 krát, trojka 12 krát, čtyřka 7 krát, pětka 9 krát a šestka také 9 krát. Můžeme o kostce říci, že je spravedlivá?*

Testujeme hypotézy

- $H_0 : p_1 = p_2 = \ldots = p_6 = 1/6$
- H_1 : *alespoň jedna z pravděpodobností* $p_1, \ldots, p_6$ *se nerovná* $1/6$.

Test porovnává

- *pozorované četnosti* $n_1 = 8, n_2 = 5, n_3 = 12, n_4 = 7, n_5 = 9, n_6 = 9$
- *očekávané četnosti* $n\pi_i = 50 \times 1/6 = 8.3333$.

Test dobré shody

Příklad. *Testová statistika*

$$\begin{aligned} X^2 &= \frac{(8-8.3333)^2}{8.3333} + \frac{(5-8.3333)^2}{8.3333} + \frac{(12-8.3333)^2}{8.3333} + \\ &+ \frac{(7-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} = 3.28 \end{aligned}$$

- *kritická hodnota χ^2- rozdělení o 5-ti st. volnosti je* $\chi^2_5 = 11.07$
- *p-hodnota vyšla $p = 0.6569$*
- *testová statistika $X^2 < \chi^2_5$ a p-hodnota $< \alpha$*
- *tedy nezamítáme nulovou hypotézu*

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

χ^2-test nezávislosti

Vztah dvou kategorických proměnných popisujeme **tabulkou absolutních četností.** 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

Obecná kontingenční tabulka absolutních četností má tvar

$$
\begin{array}{c|ccc|c}
 & Y_1 & \dots & Y_c & \\
\hline
X_1 & n_{1,1} & \dots & n_{1,c} & n_{1.} \\
\vdots & & \ddots & & \vdots \\
X_r & n_{r,1} & \dots & n_{r,c} & n_{r.} \\
\hline
 & n_{.1} & \dots & n_{.c} & n
\end{array}
$$

Hodnoty uvedené v tabulce jsou tzv. *pozorované četnosti*.

χ^2-test nezávislosti

Test je založen na porovnání

- pozorovaných četností v tabulce
- očekávaných četností za platnosti nulové hypotézy

Testované hypotézy

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

Testová statistika má tvar

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

- za platnosti nulové hypotézy má χ^2-rozdělení o $(k-1)(l-1)$ stupních volnosti
- očekávané četnosti odpovídají definici nezávislosti $P(A \cap B) = P(A)P(B)$

Fisherův exaktní test

Pro malé četnosti se používá **Fisherův exaktní test**

- není-li splněn předpoklad χ^2-testu, tj. některé očekávané četnosti jsou menší než 5
- počítá přímo p-hodnotu, tj. pravděpodobnost, že za platnosti H_0 bude pozorována právě naše tabulka četností

Pro čtyřpolní tabulku

	Y_1	Y_2	
X_1	n_{11}	n_{12}	$n_{1.}$
X_2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

se p-hodnota vypočítá následujícím způsobem

$$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

Pro větší tabulky je test složitější.

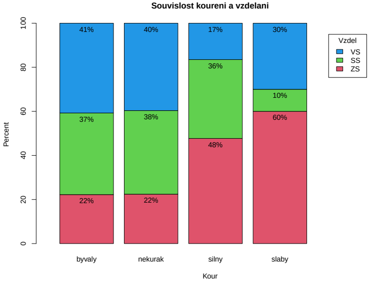
χ^2-test nezávislosti

Příklad. *U 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční bylo zjišťováno i vzdělání a kategorie kouření. Výsledky jsou shrnuty v následující tabulce absolutních četností. Souvisí spolu tyto dvě veličiny?*

	ZŠ	SŠ	VŠ
bývalý kuřák	6	10	11
nekuřák	13	22	23
slabý kuřák	52	39	18
silný kuřák	6	1	3

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 : kouření se vzděláním nesouvisí
- H_1 : kouření se vzděláním souvisí

Výsledky testu

- testová statistika χ^2-testu vyšla 21.286
- kvantil χ^2-rozdělení $\chi^2_6 = 12.59$
- p-hodnota testu vyšla 0.00163
- **ale nejsou splněny předpoklady χ^2-testu**
- p-hodnota Fisherova exaktního testu 0.00084
- p-hodnota $< \alpha$ tedy zamítáme nulovou hypotézu

Závěr: Prokázali jsme, že kouření se vzděláním souvisí.

Poměr šancí

Uvažujme dvouhodnotovou veličinu ve dvou populacích. Např. sledujeme výskyt chřipky ve městě a na venkově.

	Chřipku má	Chřipku nemá	
Město	n_{11}	n_{12}	$n_{1.}$
Venkov	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Rozdíl mezi populacemi je možné popsat poměrem šancí. Nejprve definujme **šanci** "mít chřipku proti nemít chřipku"jako

$$Odds = \frac{P(\text{má chřipku})}{P(\text{nemá chřipku})}$$

Poměr šancí je pak podíl této šance v jedné populaci ku šanci v druhé populaci.

Poměr šancí

Pro naši tabulku je pak **poměr šancí** definovaný jako

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Interpretace tohoto poměru říká, kolikrát je větší šance na chřipku ve městě než na venkově.

Pokud chceme otestovat, že šance na chřipku jsou stejné ve městě jako na venkově, testujeme

- H_0 : poměr šancí $OR = 1$
- H_1 : poměr šancí $OR \neq 1$

Testová statistika tohoto testu je rovna

$$Z = \frac{\ln(OR)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

a za platnosti nulové hypotézy má $N(0,1)$ rozdělení.

Poměr šancí

Pro poměr šancí je možné spočítat i **interval spolehlivosti**

$$\ln(OR) \pm \left(\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\right) z(\alpha/2).$$

Co je možné tímto intervalem zjistit?
Např. můžeme vyhodnocovat, zda se tento poměr může rovnat nějaké konkrétní hodnotě.

Poměr šancí

Příklad. *Uvažujme následující čtyřpolní tabulku*

	Chřipku má	*Chřipku nemá*	
Město	58	17	75
Venkov	32	30	62
	90	47	137

Šance mít chřipku ve městě vychází $58/17 = 3.41$*, šance mít chřipku na venkově vychází* $32/30 = 1.07$*. Poměr šancí ve městě vs. na venkově vychází* $3.41/1.07 = 3.2$*. Ve městě je více než třikrát větší šance mít chřipku než na venkově. Testová statistika vychází 3.27, kritická hodnota 1.96 a p-hodnota 0.001. Testová statistika je větší než kritická hodnota a p-hodnota je menší než* α*, zamítáme nulovou hypotézu.*
Závěr: *Ve městě je významně větší šance dostat chřipku než na venkově.*