Statistika v Biologii

Alena Černíková

alena.cernikova@ujep.cz

4. prosince 2023

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 tří proměnných
  od zadání *výzkumu* až po interpretaci

Obsah kurzu

- Co je statistika
- Typy proměnných
- Popisné statistiky
- Pravděpodobnostní rozdělení
- Bodový vs intervalový odhad
- Základy testování
- Jednovýběrový, párový a dvouvýběrový test
- Analýza rozptylu
- Korelace
- Jednoduchá lineární regrese

Co je statistika

Statistika je přesná věda o nepřesných číslech.
Zkoumáme náhodnou veličinu na nějaké populaci. Celou populaci změřit neumíme. Uděláme náhodný výběr, na kterém změříme sledovanou veličinu a na základě náhodného výběru 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.*

Co je statistika

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

$$k = 1 + 3.32 \log(n)$$

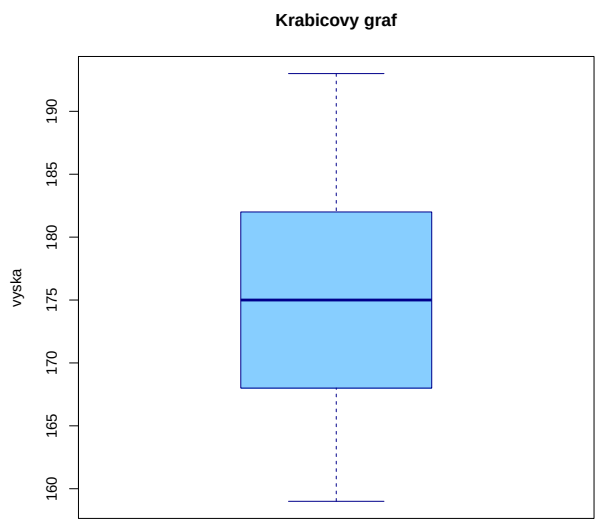
kde n je počet pozorování

Popisné statistiky – číselné proměnné

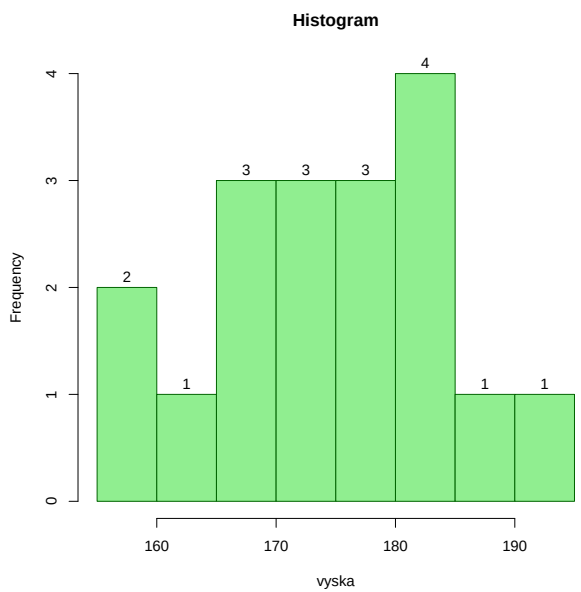
Popisné statistiky polohy – výsledky

- 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**

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

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

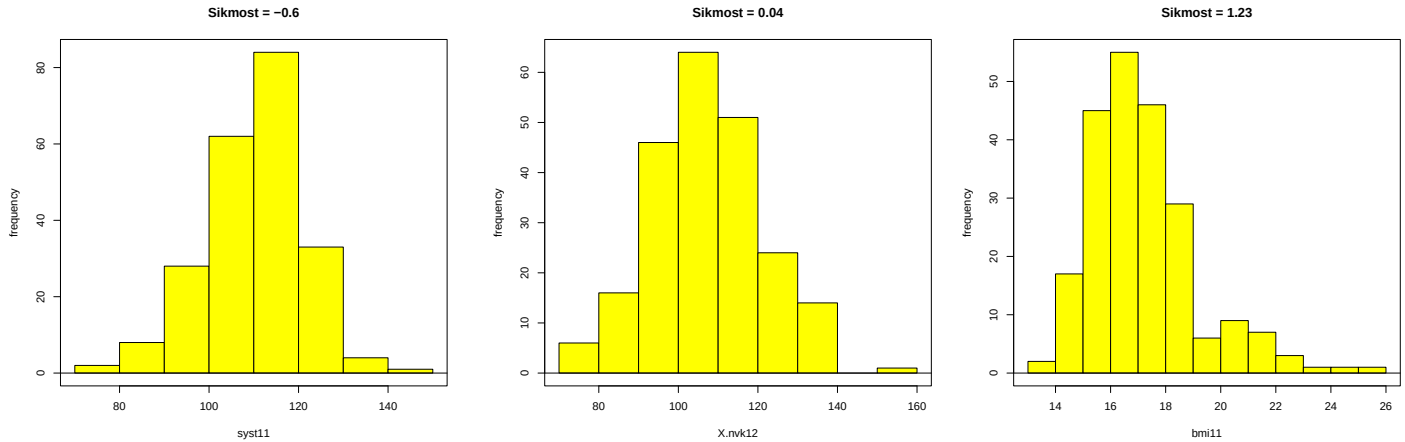
$$\mathrm{Skew(X)} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{\mathrm{sd(X)}} = \frac{\sum_{i=1}^{n} Z_i^3}{n}$$

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

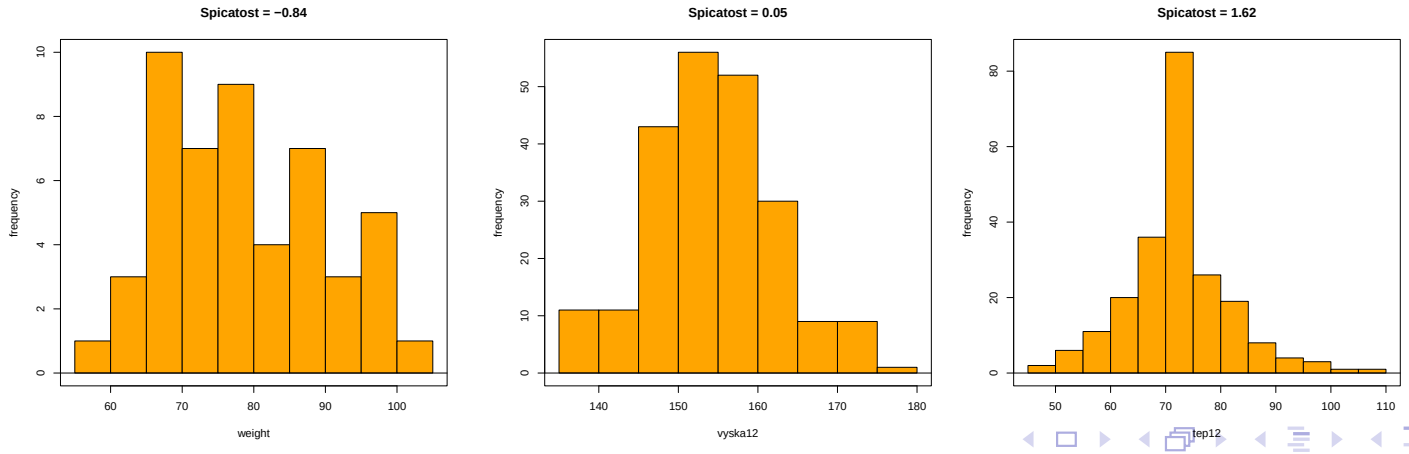
$$\mathrm{Kurt(X)} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{\mathrm{sd(X)}} - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

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.

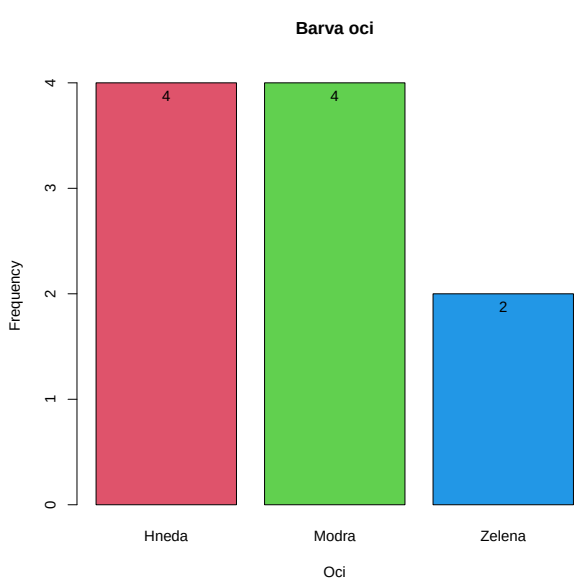
Barva	Absolutní	Relativní %
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

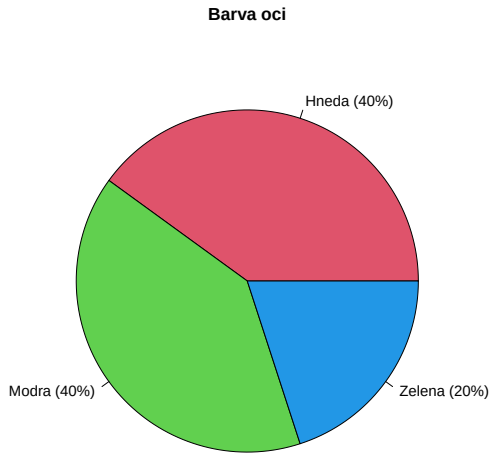
Jak vypočítat **relativní četnost?**
Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

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

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

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*

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

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*

Senzitivita a specificita testu

termíny používané v medicíně

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

Pravděpodobnostní rozdělení

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

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

- 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 – 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

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

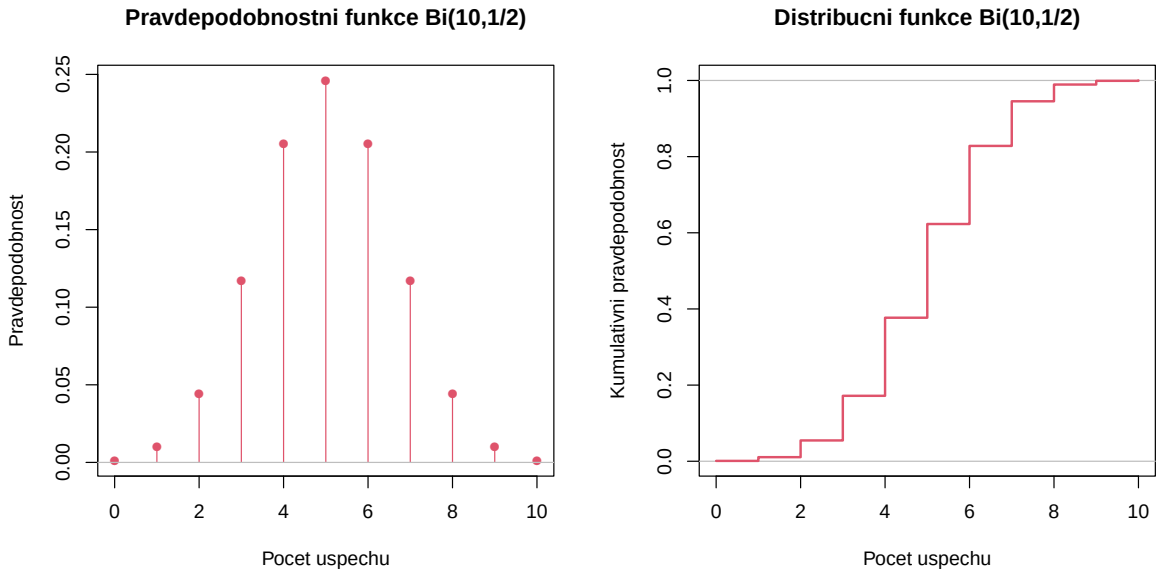
Střední hodnota a rozptyl

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

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

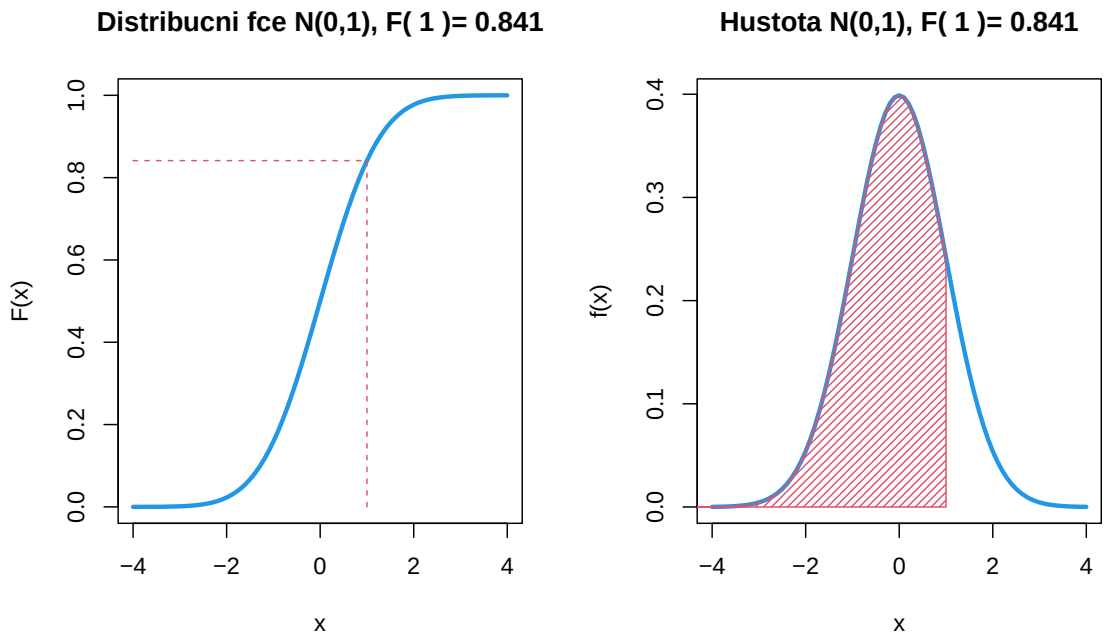
Hustota normálního rozdělení má tvar

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

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

Normální rozdělení

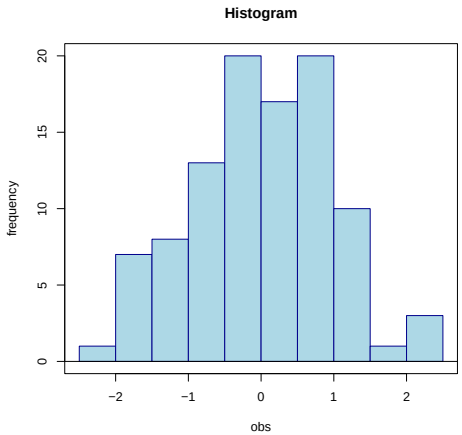
Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota.

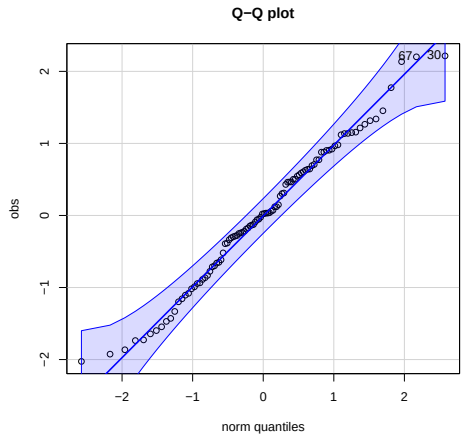


Pravděpodobnostní rozdělení

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

Centrální limitní věta

U testů pracujících s průměrem stačí, když má normální rozdělení průměr.

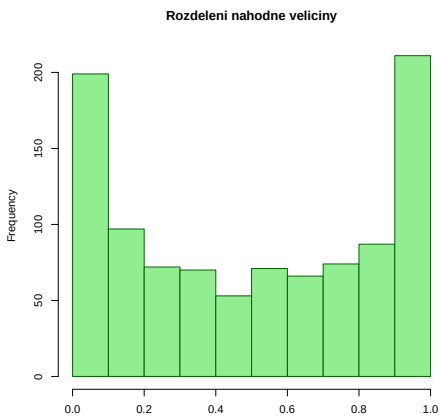
Většinu případů lze tedy vyřešit pomocí **Centrální limitní věty**.

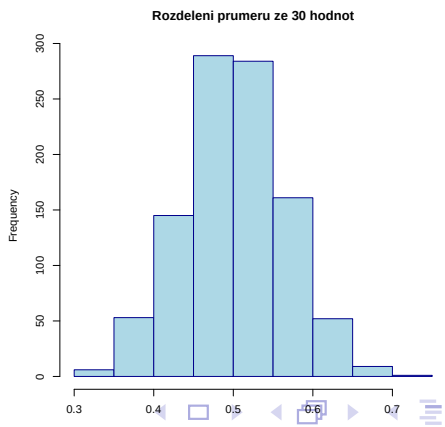
Věta

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

V praxi to znamená, že 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 dětí.*

- nejlepší bodový odhad je výběrový průměr $\overline{X} = 173.12$
- 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 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

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

Intervalový odhad pravděpodobnosti

Interval spolehlivosti mohu počítat prakticky pro libovolný odhad parametru. Předpokládejme binomické rozdělení s parametrem p, který chci ho odhadnout z dat.

Příklad. *Házíme 100 krát kostkou, z těchto 100 hodů mi šestka padla 20 krát a zajímá mne interval spolehlivosti pro pravděpodobnost, že padne 6.*

Za předpokladu dostatečně velkých hodnot n a p, konkrétně $n\hat{p}(1-\hat{p}) > 9$, kde $\hat{p}$ je odhad parametru p platí, že

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

jj. pro velká n má relativní četnost normální rozdělení.
Interval spolehlivosti pro pravděpodobnost se tedy dá vypočítat podle vzorce

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

Základy testování hypotéz

Statistické odhady jsou jen jednou částí statistiky, někdy je potřeba otestovat nějaké tvrzení. Např.

- 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.

Při statistickém rozhodování 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.
- Alternativní hypotézu – značíme H_1
  – obsahuje více možností (např. interval) – př. nový lék je lepší než ten stávající, výnosy druhů jabloní se liší

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

Jiný závěr udělat nemohu, to co mě zajímá (to, co chci prokázat), musí být v alternativě. 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

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, jsou tabelovány)
- 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**.

- 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

(Ne)zamítnout H_0 nestačí, tento výsledek je třeba interpretovat vzhledem k položené otázce.

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

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

a za platnosti nulové hypotézy má tato statistika t-rozdělení o $n - 1$ stupních volnosti.
Testovou statistiku T porovnáváme s kritickými hodnotami t-rozdělení (tj. 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 (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

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.

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**.
Budeme zde brát dva typy dvouvýběrového t-testu:

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

K tomu, abychom mohli vybrat správnou verzi testu, je potřeba otestovat shodu rozptylů v obou výběrech. Testuje se

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

Testová statistika má za platnosti H_0 F-rozdělení.

$$F = \frac{\mathrm{Var(X)}}{\mathrm{Var(Y)}} \sim F_{n_1-1,n_2-1}$$

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika tohoto testu má tvar

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

kde

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

a 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.

Welchův test

Testová statistika tohoto 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?
Nejprve otestujeme shodu rozptylů, testová statistika vychází

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

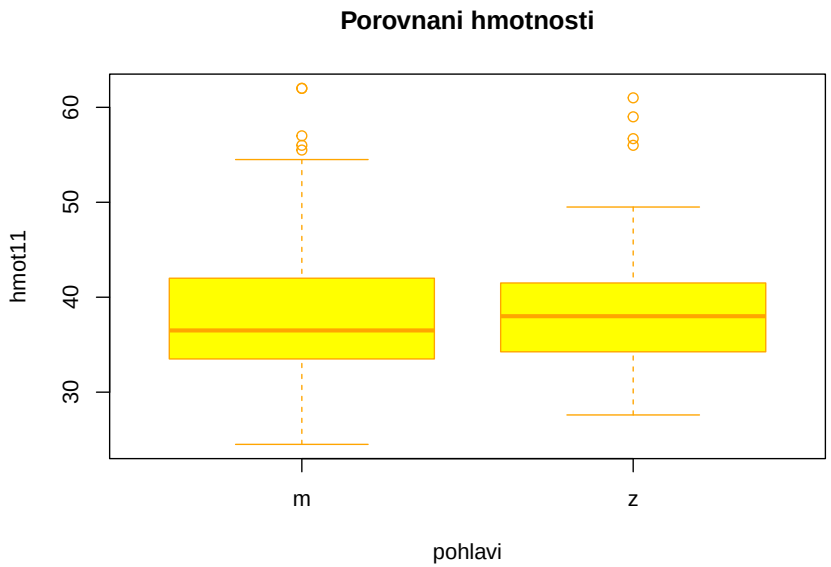
P-hodnota testu vyšla 0,56, což je více než $\alpha = 0.05$. Nulovou hypotézu tudíž nezamítáme, rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly.

Dvouvýběrový t-test

Testujeme

- 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

Grafické porovnání



Dvouvýběrový t-test

Testová statistika testu vychází

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

Tuto testovou statistiku porovnáváme s kvantilem t-rozdělení $t_{220}(1 - 0.025) = 1.97$ (kvantil pro oboustrannou alternativu). Jelikož testová statistika je v absolutní hodnotě menší než tento kvantil, tak nulovou hypotézu nezamítám.
P-hodnota testu vyšla 0.3151, tedy číslo větší než $\alpha = 0.05$

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.

Klasická (níže uvedená) ANOVA je určena pro normálně rozdělená data a výběry se shodnými rozptyly. Existuje i Welchova obdoba pro různé rozptyly ve skupinách a neparametrická verze pro data, ketrá nemají normální rozdělení.

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.

Bartlettův test

Předpokladem analýzy rozptylu je shoda rozptylů ve všech výběrech. Tento předpoklad můžeme zkontrolovat např. prostřednictvím **Bartlettova testu**.

Testujeme

- H_0 rozptyly jsou shodné
- H_1 rozptyly se liší

Testová statistika je založena na výběrových rozptylech v každém výběru zvlášť. Označme $\mathrm{Var(X)_i}$ výběrový rozptyl v i-tém výběru a

$$S^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\mathrm{Var(X)_i}}{n-k},$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right)$$

Testová statistika

$$B = \frac{1}{C}\left((n-k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln \mathrm{Var(X)_i}\right)$$

má za platnosti nulové hypotézy χ^2-rozdělení o $k-1$ 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**, případně **Tukey HSD test** pro různě velké výběry.
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 tabelované kritické hodnoty.

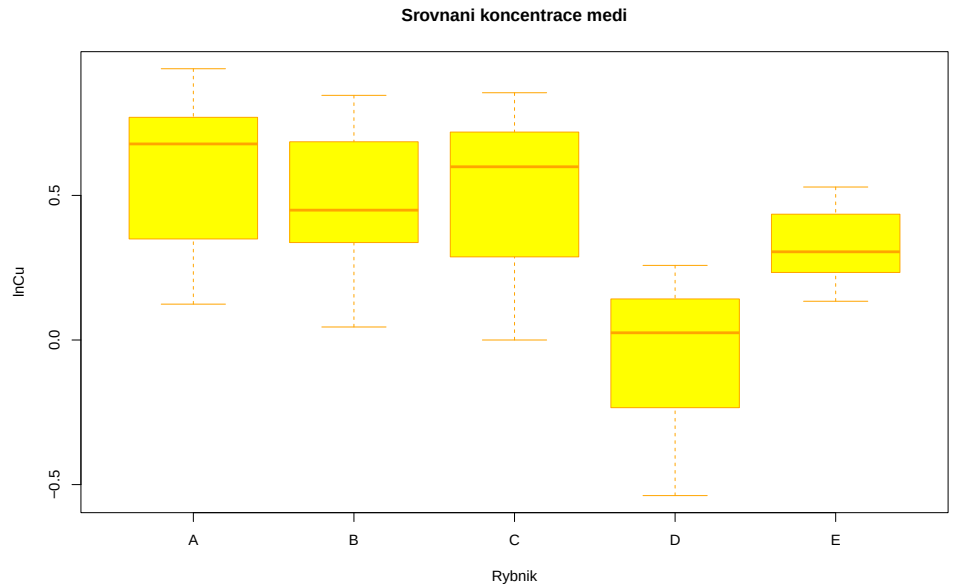
Analýza rozptylu – ANOVA

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			

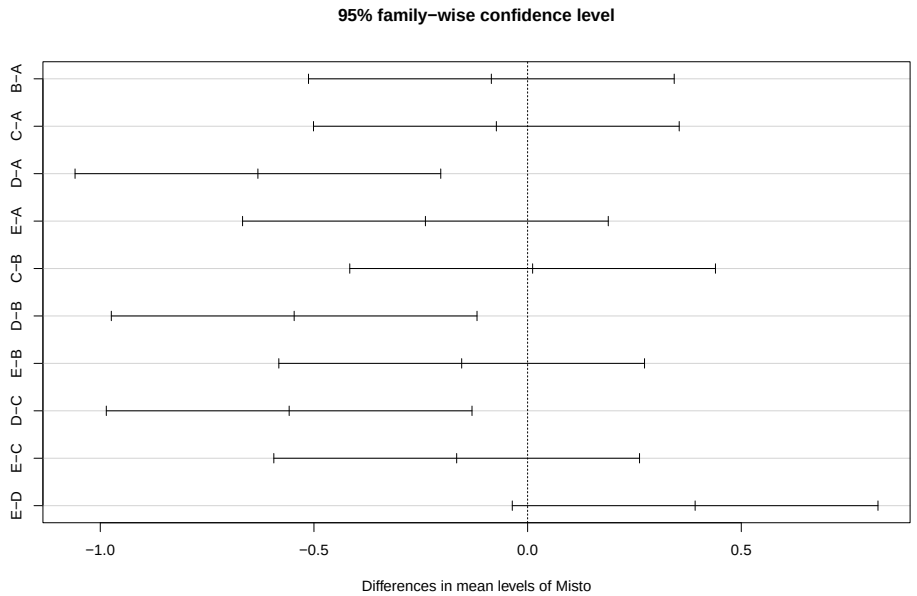
P-hodnota vyšla menší než $\alpha = 0.05$, což znamená, že nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší.

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

Je-li cílem výzkumu zjistit, zda spolu lineárně souvisí dvě číselné proměnné, používá se **korelační koeficient.**
Pearsonův korelační koeficient vypočteme jako

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

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

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

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

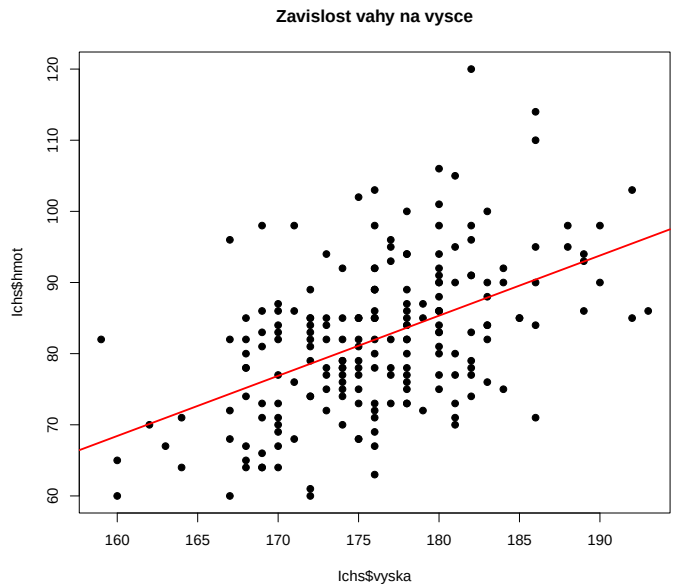
$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

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

Pearsonův korelační koeficient

Příklad. *Do výzkumu bylo zahrnuto 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční. U těchto mužů byly měřeny různé charakteristiky. Souvisí spolu výška a hmotnost těchto mužů?*

Nejprve grafické porovnání



Z grafu je patrná rostoucí závislost mezi oběma proměnnými.

Pearsonův korelační koeficient

Dále jsme testovali

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

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. Závislost je průkazná.

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

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

Hodnoty $\hat{Y}_i$ se nazývají odhady, nebo též predikce. Hodnoty b_0, b_1 jsou pak odehady regresních koeficientů. Pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné.

Koeficient determinace

Zajímavý ukazatel je koeficient determinace

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

Říká, kolik procent variability závisle proměnné se modelem vysvětlí.
Jinými slovy, z kolika procent závisle proměnná závisí na X a z kolika na něčem jiném. Na základě modelu lze též zkonstruovat test nezávislosti. Testujeme

- H_0 váha na výšce lineárně nezávisí, $\beta_1 = 0$
- H_1 váha na výšce lineárně závisí, $\beta_1 \neq 0$

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

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

Odhadli jsme model ve tvaru

$$Y_i = -66.85 + 0.85X_i$$

Střední chyba odhadu lineárního členu vyšla 0.1 a testová statistika tedy 8.19. Tu jsme porovnali s kvantilem t-rozdělení $t_{202}(1 - 0.975) = 1.97$. Jelikož je testová statistika větší, tak zamítáme nulovou hypotézu. P-hodnota testu vyšla $2.93 * 10^{-14}$, což je menší než $\alpha = 0.05$.
Závěr: Můžeme tedy říci, že u mužů s jedním rizikovým faktorem ischemické choroby srdeční hmotnost na výšce závisí. Závislost je přímá a vysvětlí se jí 25% variability závisle proměnné.

Test dobré shody

Uvažujme jednu kategorickou proměnnou, která může nabývat více než dvou kategorií. Takováto proměnná má tzv. **Multinomické rozdělení**. Jedná se o zobecnění binomického rozdělení.

Máme k kategorií, kterých může náhodná veličina nabývat. Opakujme n-krát pokus a počítejme, kolikrát byla každá z kategorií pozorována. Tyto počty označme jako proměnné $X_1, \ldots, X_k$. Multinomické rozdělení je pak dáno pravděpodobnostmi

$$P(X_1 = c_1, \ldots, X_k = c_k) = \frac{n!}{c_1! \cdot \ldots \cdot c_k!} p_1^{c_1} \cdot \ldots \cdot p_k^{c_k}$$

Dále platí, že

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

Test dobré shody

Chceme-li otesovat, že výše uvedené pravděpodobnosti nabývají nějakých konkrétních hodnot, tj. testujeme

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

použijeme χ^2 **test dobré shody**.

Testová statistika má tvar

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

a za platnosti nulové hypotézy má χ^2-rozdělení o $k - 1$ stupních volnosti. Předpokladem testu je, že všechny očekávané četnosti , tj. všechny hodnoty $n\pi_i$, jsou větší než 5. Tímto testem můžeme i testovat, zda náhodná veličina má nějaké konkrétní rozdělení.

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 dobré shody

Příklad. *Naměřili jsme hodnoty* $c_1 = 8, c_2 = 5, c_3 = 12, c_4 = 7, c_5 = 9, c_6 = 9$. *Teoretická hodnota* $n\pi_i = 50 * 1/6 = 8.3333$. *Dosadíme do vzorce a dostaneme*

$$\chi^2 = \frac{(8-8.3333)^2}{8.3333} + \frac{(5-8.3333)^2}{8.3333} + \frac{(12-8.3333)^2}{8.3333} + \\ + \frac{(7-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} = 3.28$$

Kritická hodnota χ^2*- rozdělení o* 5*-ti stupních volnosti je* $\chi^2_5 = 11.07$ *a p-hodnota vyšla* $p = 0.6569$. *Testová statistika je větší než kritická hodnota a p-hodnota menší než* α*, tedy nezamítáme nulovou hypotézu.*

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

Kontingenční tabulka absolutních četností pak má tvar

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

χ^2-test nezávislosti

Test nezávislosti je založen na porovnání pozorovaných četností v tabulce a četností očekávaných za platnosti nulové hypotézy. Testujeme

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

Testová statistika má tvar

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

Tato testová statistika má za platnosti nulové hypotézy χ^2-rozdělení o $(k-1)(l-1)$ stupních volnosti.
Očekávané četnosti se dopočítávají z definice nezávislosti $P(A \cap B) = P(A)P(B)$.

Fisherův exaktní test

Předpokladem χ^2-testu je, že všechny očekávané četnosti jsou větší než 5. Pokud předpoklad není splněn, používá se **Fisherův exaktní test**, známý též jako **Fisherův faktoriálový test**. Tento test 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.

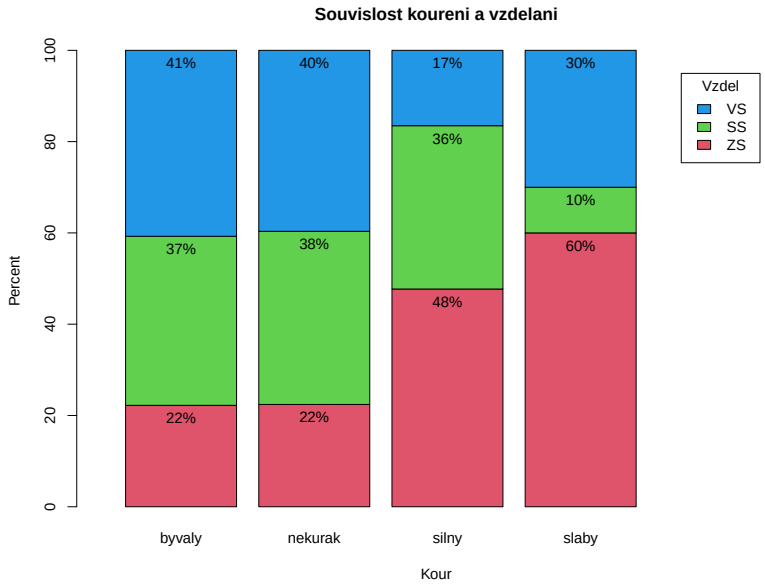
χ^2-test nezávislosti

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

Testová statistika vyšla 21.286. Porovnáváme ji s kvantilem χ^2-rozdělení $\chi^2_6 = 12.59$. Jelikož testová statistika vyšla vyšší, tak zamítáme nulovou hypotézu. P-hodnota testu vyšla 0.00163, což je menší než $\alpha = 0.05$.
Jelikož však nejsou splněny předpoklady testu, měli bychom vypočítat ještě p-hodnotu Fisherova exaktního testu. Ta vychází 0.00084.
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. Výsledky je možné zapsat do čtyřpolní tabulky

	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

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

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. Jelikož testová statistika je větší než kritická hodnota a p-hodnota je menší než* α, *zamítáme nulovou hypotézu. Ve městě je významně větší šance dostat chřipku než na venkově.*