Statistická analýza dat

Alena Černíková

alena.cernikova@ujep.cz

22. listopadu 2023

Podmínky zápočtu

Průběžné odevzdávání domácích úkolů

- data do 18.10.2023
- popisné statistiky včetně intervalu spolehlivosti do 8.11.2023
- vztah dvou kategorických veličin
- vztah kategorické a číselné proměnné
- vztah dvou číselných proměnných

Obsah kurzu

- Co je statistika
- Konstrukce výběru a dotazníku
- Typy proměnných a jejich popisné statistiky
- Bodový vs intervalový odhad
- Parametrické a neparametrické testy o střední hodnotě
- Vztah dvou nominálních proměnných
- Vztah dvou ordinálních proměnných
- Závislost číselné proměnné na kategorické
- Korelace a 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 názor lidí na zřízení národního parku Křivoklátsko. Osloveno bylo 200 lidí, kterým byla položena otázka "Souhlasíte se zřízením NP Křivoklátsko?"se čtyřmi možnými odpověďmi: 1) Rozhodně ano, 2) Spíše ano, 3) Spiše ne, 4) Rozhodně ne. Co můžu z výsledků zjistit?*

Co je statistika

- **Nahodná veličina** – jakákoliv veličina, kterou měříme,
- **Populace** – soubor, pro nejž chceme udělat nějaký závěr, např. všichni dospělí obyvatelé České republiky, všichni starší 15 let žijící do 10 od hradu Křivoklát, atd.
- **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 podíl odpovědí
- **Výberová charakteristika** – charakteristika spočítaná na výběru pomocí níž odhadujeme populační ekvivalent, relativní četnost.

Konstrukce výběru

Před provedením výzkumu je třeba si rozmyslet

- Kdo je cílová populace a jak tedy realizovat náhodný výběr
- Jakým způsobem se budou data sbírat (pozorováním/osobním dotazníkem/on-line/experimentem/nalezením dat na internetu)
- Co chci dotazníkem zjistit, tedy jak navrhnout dotazník a jednotlivé otázky, abychom došli cíle
- Kolik dat chci získat (bude výstup kvalitativní nebo kvantitativní)

Konstrukce výběru

Jak zajistit, aby byl výběr **reprezentativní**?

- **náhodný výběr** ze seznamu celé populace – dělá se někde?
- **systematický výběr** ze seznamu (beru každého desátého)
- **kvótní výběr** – vybrat několik základních znaků a u nich určit **kvóty**, jak často mjí být zahrnuty ve výběru, aby výběr odpovídal populaci
  - Kvótní výběr pro dospělé obyvatele ČR – dodržují se kvóty pro pohlaví, věk (5 věkových skupin), vzdělání, kraj a příjem (5 příjmových skupin)
- **stratifikovaný výběr** – výběr i vyhodnocení realizuji ve stratech
- **vícestupňový výběr** – vybírám z několika velkých oblsatní, v rámci každé vybrané pak z několika podoblastí, atd.

A co anketa?

Konstrukce výběru

U pravděpodobnostních výběrů (náhodný, systematický) je možné dopředu určit minimální požadovaný rozsah výběru, aby výsledné zjištění mělo požadovanou přesnost.

- Chci aby **interval spolehlivosti** pro populační průměr měl délku maximálně 2Δ. Délka poloviny intervalu spolehlivosti je $z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$, kde α je hladina významnosti, n je počet pozorování a σ je směrodatná odchylka proměnné. Pak

$$n \geq \left(z(1-\alpha/2)\frac{\sigma}{\Delta}\right)$$

- Chci aby **dvouvýběrový t-test** měl spolehlivost α a sílu testu $1-\beta$. Při sdružené směrodatné odchylce obou výběrů s a požadovaném rozdílu mezi průměry Δ (při tomto rozdílu má být požadovaná síla testu) by rozsah výběru měl být

$$n \geq ((z(1-\alpha)+z(1-\beta))/\Delta)^2 s^2$$

Konstrukce dotazníku

Nejprve je třeba si stanovit výzkumné cíle

- Chci zjistit nějaká východiska – **explorační studie**
  – např. co vede lidi k návštěvě vybrané cukrárny
- Chci popsat nějakou situaci – **popisná studie**
  – např. jak se změnila spokojenost studentů FSE se studiem za posledních 10 let,
- Chci rozhodnout o pravdivosti nějakého tvrzení – **explanační studie**
  – např. Je návštěvnost cukrárny závislá na vybavenosti blízkého dětského hřiště?

Konstrukce dotazníku

Dotazník by měl zjistit vše potřebné a přitom nebýt příliš dlouhý

- příliš **málo otázek** – na některé výzkumné cíle nenajdu odpověď
- příliš **mnoho otázek** – respondent bude unavený a už nebude odpovídat podle pravdy

Otázky by měly být

- jednoznačné,
- složitostí přizpůsobené respondentům,
- zodpověditelné, tj. měly by dotazovat to, co respondent ví
- přehledné (pokud respondent vyplňuje dotazník sám)

Vlastnímu šetření by měla předcházet pilotáž.

Konstrukce dotazníku

Typy otázek

- **Uzavřená otázka** – klasická otázka v dotazníku s několika položkami
- **Otevřená otázka** – těžké na zpracování, ale mnohdy přinese zajímavé podněty, dávají se jen výpisy
- **Polootevřená otázka** – kompromis
- **Podstatné (meritorní) otázky** – k těmto se občas dávají i kontrolní otázky, abychom měli jistotu, že respondent odpovídá "správně"
- **Filtrační otázky** – podle odpovědi se pak dotazujeme dál jistým směrem (Př. máte auto? Jaké značky?)
- **Identifikační** – měly by být pokud možno v každém dotazníku, umožňují identifikovat skupiny (např. demografie)

Pozor na citlivé otázky.

Konstrukce dotazníku

Každá otázka v dotazníku reprezentuje náhodnou veličinu/proměnnou.
Při zpracování dat je důležité rozlišovat následující typy proměnných.

- **Číselné proměnné**, kardinální/intervalová škála – př. výška, váha, věk, atd.
- **Kategorické proměnné** – př. barva, kraj, povolání, nebo taky známka ve škole, číslo, které padne na kostce, atd.
- Kategorické proměnné se dále dělí na
  - **Nominální**, nominální škála – neuspořádané, př. barva, kraj
  - **Ordinální**, ordinální škála – uspořádané, př. známka, míra spokojenosti na stupnici

Otázky vedoucí na ordinální veličiny (**škálové otázky**) bývají v dotaznících nejoblíbenější.

Konstrukce dotazníku

Příklady vybraných škál

- **Nominální škála** – seznam položek, př. barva
- **Ordinální škála** – seznam položek, př. míra souhlasu (rozhodně souhlasím, spíše souhlasím, spíše nesouhlasím, nesouhlasím)
  – u ordinální škály nelze zaručit, že všichni respondenti vnímají stejně velké intervaly mezi jednotlivými body škály
- **Kardinální škála** – snaha o zajištění stejné vzdálenosti mezi body, pr.
  - vyberte hodnotu na stupnici od 1 do 10, kde 1 je velmi nespokojen a 10 velmi spokojen
  - zakreslete svou míru spokojenosti do úsečky, kde vlevo je velmi nespokojen a vpravo velmi spokojen
  - podle obrázků ohodnoťte svou míru spokojenosti (5 smajlíků)

Konstrukce dotazníku

Likertova metoda. Jednu věc můžeme hodnotit i vícerozměrně. Několik charakteristik sledované jednotky hodnotíme na škále a následně použijeme jedno souhrnné číslo, např. průměr, medián, součet, ...
Výsledná veličina je již spojitá (kromě mediánu).

Příklad. *Hodnotíme kvalitu výuky z několika pohledů: srozumitelnost vyjadřování, ochota odpovídat na dotazy, zajímavost vybraných příkladů, náročnost látky, náročnost u zkoušky, atd.*

Likertův koeficient diferenciace - málo diferencující otázky se vynechají (buďto přímo z šestření po pilotáži, nebo z vyhodnocení)

$$L_D = \frac{\text{součet čtvrtiny nejvyšších skóre} - \text{součet čtvrtiny nejnižších skóre}}{n/2}$$

kde n je počet pozorování. Pro pětibodovou škálu nabývá hodnot od 0 do 2.

Konstrukce dotazníku

Výběr ordinální škály

- neměla by být příliš **detailní** – př. 1) nikdy 2) vzdáleně 3) příležitostně 4) poměrně často 5) často 6) velmi často 7) téměř vždy 8) vždy
- zamyslet se nad možností/vhodností **sloučit položky** – př. 1) zřídka či nikdy 2) občas 3) často
- umožnit **střed** a v jaké podobě – př. 1) souhlasím 2) spíše souhlasím 3) nejsem si jistý 4) spíše nesouhlasím 5) nesouhlasím
- co s variantou **nevím** – často je možné zcela vynechat (ti, co neví, prostě neodpoví)
- škála by měla zahrnovat **"celý prostor"** –
  1) souhlasím 2)spíše souhlasím 3) spíše nesouhlasím 4) nesouhlasím
  versus
  1) silně souhlasím 2) souhlasím 3) nesouhlasím 4) silně nesouhlasím

Pozor na **objektivitu** odpovědí, odpověď respondenta vždy popisuje subjektivní postoj, ne nutně objektivní závěr.

Popisné statistiky

- **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
  - modus – nejčastěji uváděná hodnota
  - grafické charakteristiky – sloupcový a koláčový graf
- **Ordinální proměnné**
  - vhodné jsou absolutní i relativní četnosti, k obojímu též kumulativní četnosti
  - lze použít také průměr, medián atd.

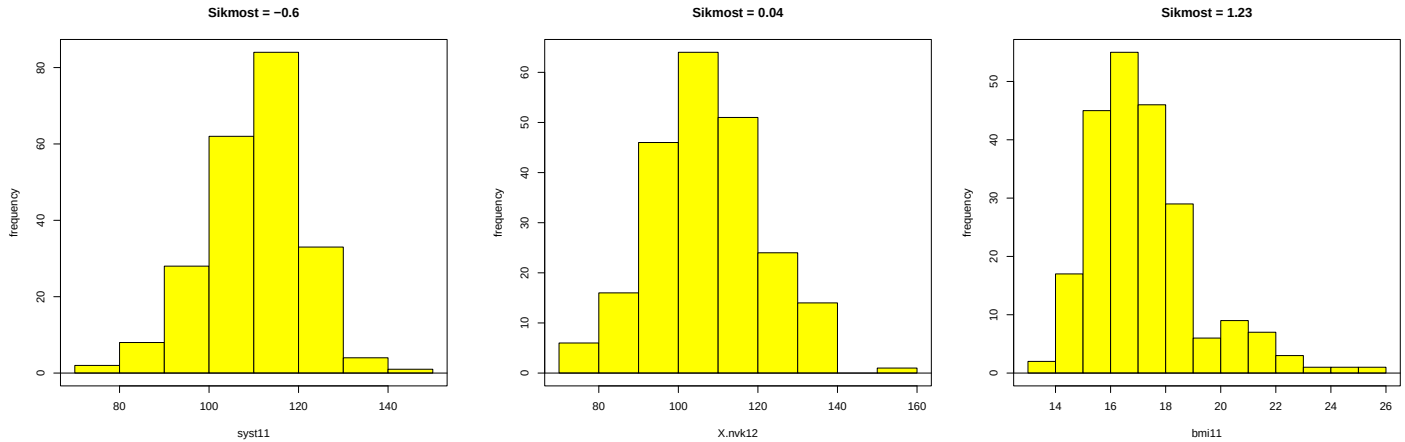
Opakování – popisné statistiky

Popisné statistiky pro číselnou proměnnou

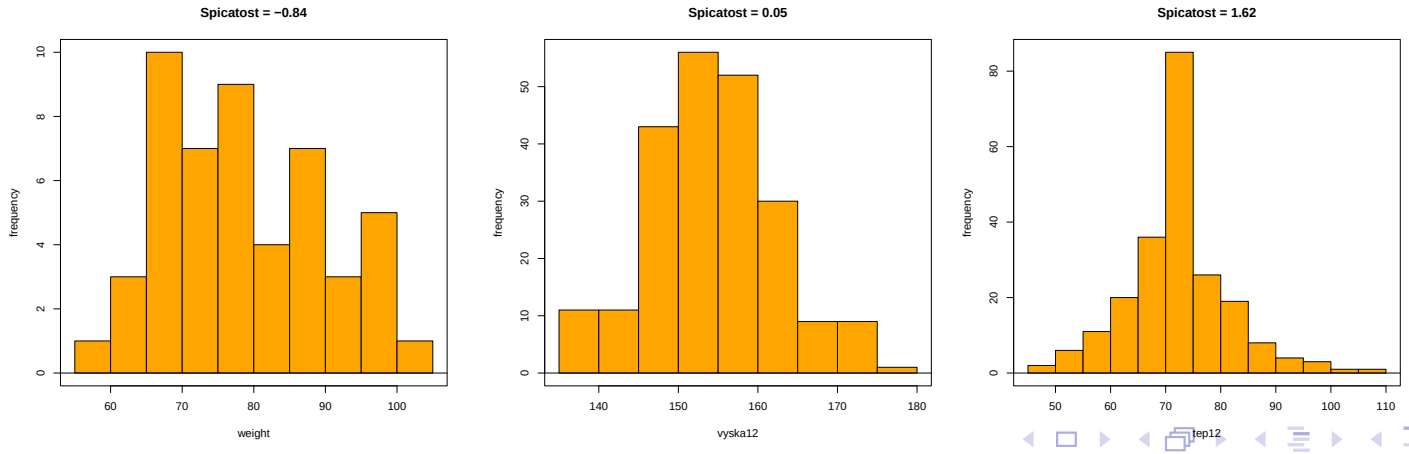
- **průměr** – $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, kde n je počet pozorování a $X_1, X_2, X_3, \ldots, X_n$ jsou jednotlivá měření
- **medián** – hodnota prostřední podle velikosti, nebo průměr prostředních dvou
- vybrané percentily, především **extrémy** a **kvartily** – hodnoty v jedné a ve třech čtvrtinách podle velikosti
- **Směrodatná odchylka** – $\mathrm{sd(X)} = \sqrt{\mathrm{VarX}} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2}{\mathrm{n}-1}}$
- **Mezikvartilové rozpětí** – $IQR(X) = Q_3 - Q_1$, kde Q_3 je třetí kvartil a Q_1 je první kvartil

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



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



Odhady střední hodnoty

U číselné proměnné nejčastěji odhadujeme populační průměr

- nejlepším bodovým odhadem střední hodnoty je **výběrový průměr** $\overline{X} = \sum_{i=1}^{n} X_i/n$
- nestranný odhad
- platí Centrální limitní věta – pro rostoucí počet pozorování konverguje rozdělení výběrového průměru k normálnímu pro $n \to \infty$
- střední chyba průměru je $\mathrm{SEM} = \mathrm{sd(X)}/\sqrt{\mathrm{n}}$
- **intervalový odhad** pro průměr je

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

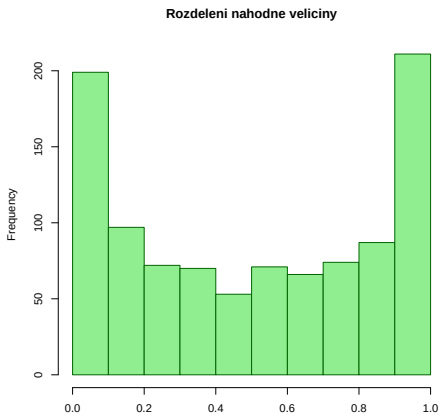
Centrální limitní věta

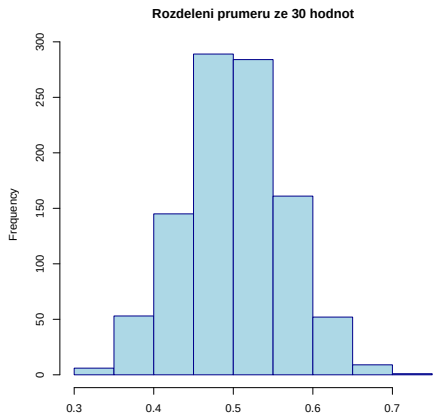
Věta

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.





Základy testování hypotéz

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme H_0
  – je v ní vždy pouze jedna varianta
- Alternativní hypotézu – značíme H_A
  – obsahuje více možností (např. interval)

Na základě testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu – platí alternativa
- Nezamítneme nulovou hypotézu

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

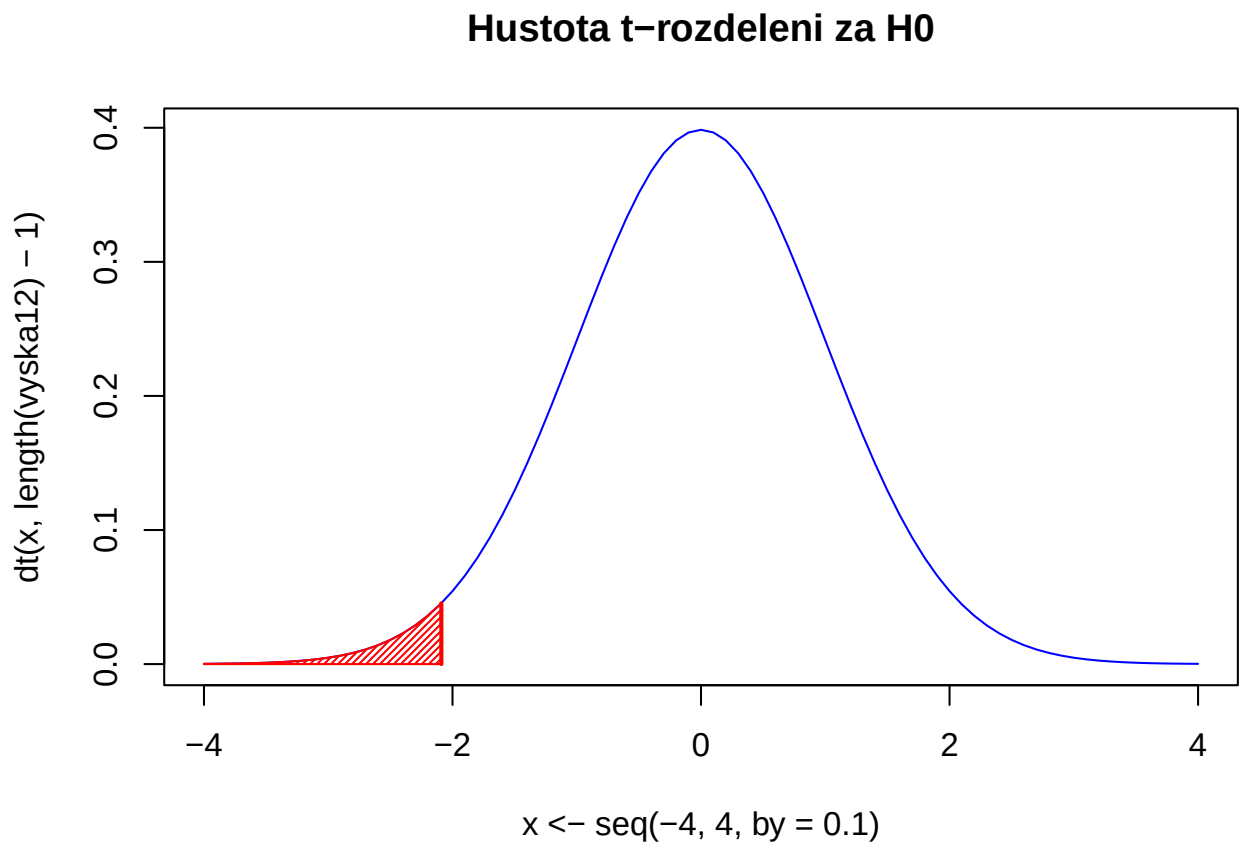
Základy testování hypotéz

Testovat mohu buď přes porovnání **testové statistiky** a **kritické hodnoty** (kvantil vybraného teoretického rozdělení), nebo přes porovnání *p*-**hodnoty** a **hladiny významnosti**.

Výsledkem testu v počítači je *p*-**hodnota**

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti H_0 nastal výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- *p*-hodnota $\leq \alpha$ potom ZAMÍTÁME H_0
- *p*-hodnota $> \alpha$ potom NEZAMÍTÁME H_0

Co je **p-hodnota**?

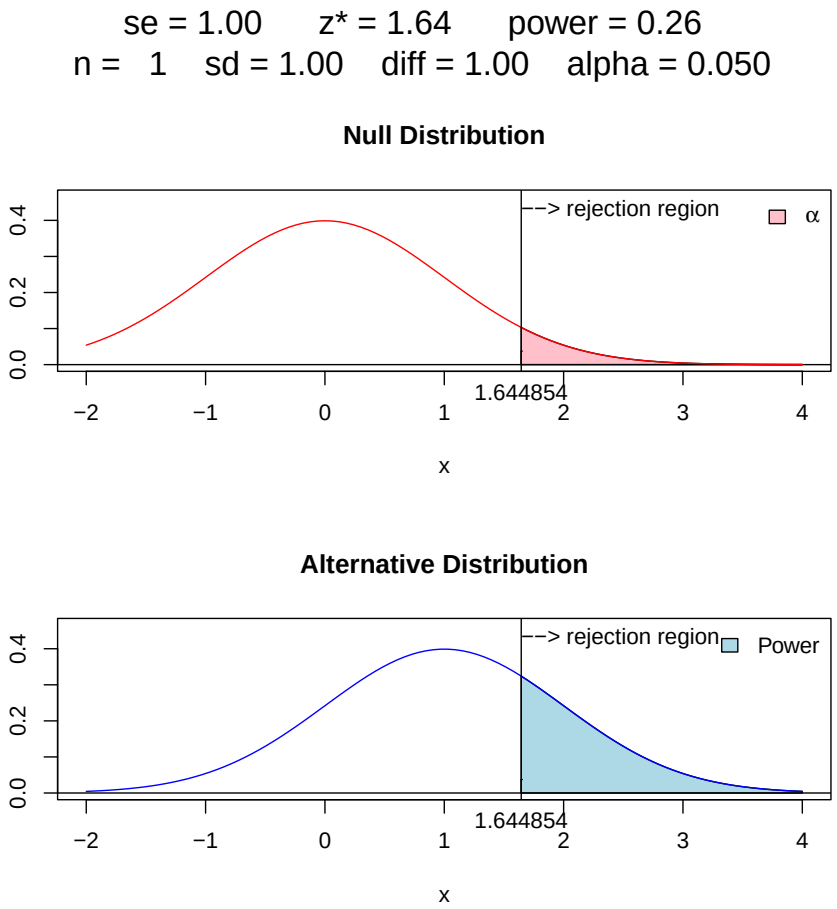


Základy testování hypotéz

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

Co je **síla testu**?

Na určité hladině významnosti chceme test proti vybrané alternativě co nejsilnější.



Jednovýběrový t-test

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

- H_0 střední hodnota $= \mu_0$
- H_1 střední hodnota $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$. **Testová statistika** jednovýběrového t-testu je

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

a za platnosti nulové hypotézy má tato statistika t-rozdělení o $n - 1$ stupních volnosti.
Jak souvisí jednovýběrový t-test s intervalem spolehlivosti?

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

Odhady pro pravděpodobnost

Podíl jedné konkrétní odpovědi na otázku v dotazníku

- nejlepším bodovým odhadem pravděpodobnosti je **relativní četnost** $p_i = n_i/n$
- nestranný odhad
- náhodná veličina $p = (p_i - \pi_i)/\sqrt{\pi_i(1-\pi_i)/n}$ konverguje k normálnímu rozdělení $N(0,1)$ pro $n \to \infty$
- **intervalový odhad** pro pravděpodobnost je

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

- pro použití tohoto intervalu musíme mít dostatečně velké n a p_i, má platit $np_i(1-p_i) > 9$

Test o pravděpodobnosti

Při testování pravděpodobnosti je možné využít buďto přesný binomický test (s využitím kritických hodnot binomického rozdělení), nebo aproximativní **"proportion test"** (s využitím kritických hodnot normálního rozdělení). Častěji využíváme ten druhý. Testujeme

- H_0 pravděpodobnost daného jevu $= \pi_0$
- H_1 pravděpodobnost daného jevu $\neq \pi_0$, nebo $< \pi_0$, nebo $> \pi_0$

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$ **Testová statistika** aproximativního testu o pravděpodobnosti je

$$Z = \frac{p_i - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

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

Párový t-test

V případě, že porovnáváme dva závislé výběry, tedy taková data, která tvoří přirozené páry, používá se **párový test**. Testované hypotézy v něm jsou

- H_0 střední hodnota rozdílu párů $= \mu_0$
- H_1 střední hodnota rozdílu $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Postup testu je takový, že v prvním kroku spočítám rozdíly mezi všemi páry

$$R_i = X_i - Y_i$$

kde X_i a Y_i jsou párová měření, a ve druhém kroku se testuje střední hodnota/ průměr tohoto rozdílu běžným **jednovýběrovým testem**.

Příklad. *Porovnávám věk otce a matky, srovnávám sílu pravé a levé ruky, srovnávám měření před a po podání nějakého léku, atd.*

Dvouvýběrový t-test

Pokud porovnávám dva nezávislé výběry (pozorování nemohu napárovat), pak je potřeba použít **dvouvýběrový test**.
Testujeme

- H_0 rozdíl středních hodnot $= \mu_0$
- H_1 rozdíl středních hodnot $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Testová statistika dvouvýběrového t-testu pro shodné rozptyly je

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

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.

Dvouvýběrový t-test

V případě, že výběry shodné rozdělení nemají, používá se **Welchova varianta dvouvýběrového t-testu**. Její **testová statistika** má tvar

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

Tato statistika má za platnosti nulové hypotézy 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.

Test shody rozptylů ve dvou výběrech

Chceme-li rozhodnout, kterou variantu dvouvýběrového t-testu máme použít, je nutné zjistit, zda jsou v obou výběrech stejné rozptyly.
Testujeme

- H_0 rozptyly jsou shodné
- H_1 rozptyly se liší

Testová statistika **F-testu pro dva rozptyly** má tvar

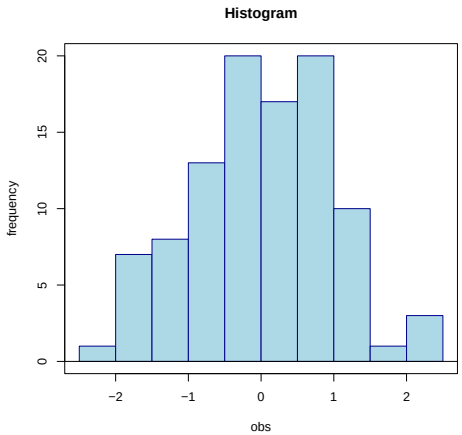
$$F = \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

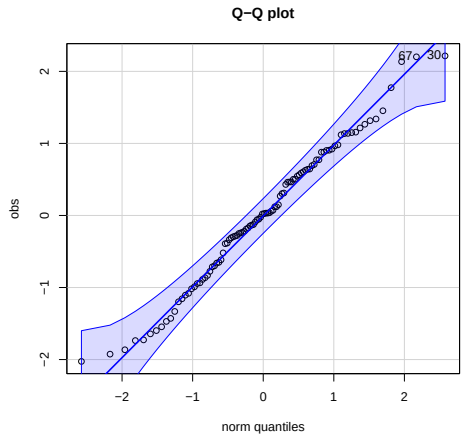
a za platnosti nulové hypotézy má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti.

Testování normality

Většina statistických postupů, odhadů a testů je odvozena 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

Testování normality

Nejčastěji používané číselné testy normality

- **Shapiro-Wilkův** – test odpovídající pravděpodobnostnímu grafu
  porovnává, jak si odpovídají teoretické percentily pro normální rozdělení a percentily naměřené pro sledovanou proměnnou
- **Kolmogorovův-Smirnovův** – test je založen na maximálním rozdílu empirické distribuční funkce a distribuční funkce normálního rozdělení
- **Andersonův-Darlingův** – test je založen na váženém průměru druhé mocniny rozdílu empirické distribuční funkce a distribuční funkce normálního rozdělení

Neparametrické testy

V případě, že náhodná veličina normální rozdělení nemá, respektive, žeodchylky od normálního rozdělení jsou takového typu, že nelze použít zvolený výše uvedený test, je potřeba zvolit odpovídají **neparametrickou metodu**.
Neparametrické testy bývají většinou založeny na pořadí naměřených hodnot v uspořádané řadě.

Příklad. *Uvažujme naměřené věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50. Data věků rodičů bývají sešikmena a často obsahují odlehlé hodnoty. Přiřadíme-li hodnotám pořadí podle velikosti, získáme řadu 6, 3.5, 7, 9, 3.5, 2, 5, 8, 1, 10. Takto získaná řada není sešikmená a nemá odlehlé hodnoty.*

Znaménkový test

Test o hodnotě mediánu jednoho výběru. Testujeme

- H_0 : medián $= m_0$
- H_1 : medián $\neq m_0, > m_0, < m_0$

Pro každé pozorování spočteme rozdíl $X_i - m_0$ a spočítáme, kolik těchto rozdílů je kladných. Tento součet označme jako Z.
Za platnosti nulové hypotézy má testová statistika Z binomické rozdělení $Bi(n, 1/2)$, kde n je počet pozorování.
Pro velká n je možné použít i transformaci

$$U = \frac{2Z - n}{\sqrt{n}}$$

Která má za platnosti H_0 $N(0, 1)$ rozdělení.

Znaménkový test

Příklad. *Pokračujme v příkladu s věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

- H_0 : *medián věku otců je 33 let*
- H_1 : *medián věku otců není 33 let*

Spočtěme rozdíly $X_i - m_0$*: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17. Kladných hodnot je mezi nimi* $Z = 4$*. P-hodnota testu vychází 0.75, což je hodnota* $> \alpha (= 0.05)$ *a* H_0 *tedy nezamítáme. Použitím U-transformace dostaneme* $U = -0.632$ *a p-hodnotu 0.527.*

Wilcoxonův jednovýběrový test

Znaménkový test porovnává pouze počet hodnot ležících pod mediánem a těch, co leží nad ním. Nezohledňuje však vzdálenost od mediánu. To dělá Wilcoxonův test, neboli **Mann-Whitneyův** test. Ten už je založen na pořadích. Testované hypotézy zůstávají stejné.

Postup testu

- spočítají se rozdíly od testované hodnoty $X_i - m_0$
- určí se jejich znaménko
- určí se pořadí absolutních hodnot rozdílů
- spočítá se součet těchto pořadí patřících kladným rozdílům
- označme tento součet S^+ a obdobně označme S^- součet pořadí pro záporné rozdíly, musí platit $S^+ + S^- = n(n+1)/2$.

Pro větší n lze užít transformaci

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$$

která má za platnosti H_0 $N(0,1)$ rozdělení.

Wilcoxonův jednovýběrový test

Příklad. *Pokračujme v příkladu s věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a opět testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

- H_0 : *medián věku otců je 33 let*
- H_1 : *medián věku otců není 33 let*

Spočtěme rozdíly $X_i - m_0$*: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17 a jejich absolutním hodnotám přiřaďme pořadí 1.5, 6, 1.5, 6, 6, 8, 3.5, 3.5, 9, 10. Sečtěme kladné (modré) pořadí* $S^+ = 21$ *a záporné (červené) pořadí* $S^- = 34$*. Testová statistika vychází* $U = -0.66$ *a p-hodnota 0,51* $> \alpha (= 0.05)$ *a* H_0 *tedy nezamítáme.*

Wilcoxonův párový test

V případě, že chceme porovnat dva **závislé výběry**, které nesplňují předpoklad normality, používá se **párový Wilcoxonův test.**
I zde zůstávají testované hypotézy stejné jako u párového t-testu.
V prvním kroku se spočítají **rozdíly** v rámci párů, tj. pro každé $X_i, Y_i, i = 1, \ldots, n$

$$R_i = X_i - Y_i$$

Pokud tyto rozdíly nemají normální rozdělení, použije se pro ně jednovýběrový Wilcoxonův test.

Wilcoxonův dvouvýběrový test

Pro porovnání dvou **nezávislých výběrů**, které nesplňují předpoklad normality, se používá **dvouvýběrový Wilcoxonův test**. Testujeme

- H_0 : střední hodnota X – střední hodnota $Y = 0$
- H_0 : střední hodnota X – střední hodnota $Y \neq 0, < 0$ nebo > 0

Test je založen na pořadích hodnot sdruženého výběru.

Postup

- oba výběry spojí do jednoho sdruženého
- sdružený výběr se uspořádá podle velikosti a každé pozorování dostane své pořadí
- pro oba výběry se vypočte součet pořadí a následně i průměrné pořadí
- pokud jsou si průměrná pořadí podobná, výběry se mezi sebou významně neliší

Wilcoxonův dvouvýběrový test

Technický výpočet: označme T_1, T_2 součet pořadí v prvním, respektive druhém výběru. Dále vypočteme

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2,$$

kde n_1, n_2 jsou rozsahy jednotlivých výběrů. Přesný test porovnává hodnotu $min(U_1, U_2)$ s kritickou hodnotou. Asymptoticky platí, že

$$U_0 = \frac{U_1 - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}$$

má za platnosti H_0 $N(0,1)$ rozdělení.

Wilcoxonův dvouvýběrový test

Příklad. *Chceme porovnat výsledky testů studentů v Ústí nad Labem a v Liberci. Studenti v Ústí dostali bodová ohodnocení 45, 79, 81, 56, 53, 77. Studenti v Liberci získali ohodnocení 76, 62, 84, 80, 41, 79, 66. Testujeme*

- H_0 : *Studenti v Ústí a v Liberci jsou stejní*
- H_1 : *Studenti v Ústí a v Liberci se liší.*
- *V prvním kroku srovnám všechny hodnoty do řady* 41, 45, 53, 56, 62, 66, 76, 77, 79, 79, 80, 81, 84
- *následně jim přiřadím pořadí* 1, 2, 3, 4, 5, 6, 7, 8, 9.5, 9.5, 11, 12, 13
- *pak vypočtu* $T_1 = 38.5$, $T_2 = 52.5$, $U_1 = 24.5$, $U_2 = 17.5$, $U_0 = 0.5$, $p = 0.6678$

P-hodnota $> \alpha$ *a tedy nezamítám nulovou hypotézu, neprokázal se rozdíl mezi studenty v Ústí a v Liberci.*

Test pro porovnání dvou pravděpodobností

Pokud chceme porovnat pravděpodobnost výskytu nějakého jevu ve dvou nezávislých výběrech, používá se test pro porovnání dvou pravděpodobnoostí.
Testujeme

- H_0 rozdíl pravděporobností $p_1 - p_2 = \pi_0$
- H_1 rozdíl pravděpodobností $p_1 - p_2 \neq \pi_0$, nebo $< \pi_0$, nebo $> \pi_0$

Testová staistika má tvar

$$Z = \frac{p_1 - p_2 - \pi_0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ kde } p = \frac{x_1 + x_2}{n_1 + n_2}$$

n_1, n_2 jsou počty pozorování v prvním a druhém výběru, x_1, x_2 jsou počty výskytů sledovaného jevu v prvním a druhém výběru, p_1, p_2 jsou relativní četnosti v prvním a druhém výběru.
Testová statistika má za platnosti H_0 $N(0, 1)$ rozdělení.

Test pro porovnání dvou pravděpodobností

Z testové statistiky lze odvodit i intervaol spolehlivosti pro rozdíl dvou pravděpodobností. Tento má tvar

$$p_1 - p_2 \pm z(1-\alpha/2)\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Jakou má tento interval interpretaci?

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.

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.

McNemarův test

V případě, že testujeme vývoj nějaké nominální charakteristiky v čase, můžeme tento vývoj popsat čtvercovou kontingenční tabulkou, kde v řádcích máme stav veličiny v čase t_0 a ve sloupcích stav veličiny v čase t_1. Pokud chceme testovat, že se **situace v čase nezměnila**, testujeme tím vlastně **symetrii** tabulky, tj.

- $H_0 : n_{i,j} = n_{j,i}$ pro všechna $i \neq j$
- H_1 : existuje alespoň jedna dvojice $j \neq j$ že $n_{i,j} \neq n_{j,i}$

Testová statistika testu má tvar

$$X = \sum_{i<j} \frac{(n_{i,j} - n_{j,i})^2}{n_{i,j} + n_{j,i}}$$

za platnosti nulové hypotézy má tato statistika χ^2-rozdělení o $k(k-1)/2$ stupních volnosti, kde k je počet kategorií veličiny.

Poměr šancí

Uvažujme dvouhodnotovou veličinu ve dvou populacích. Např. sledujeme výskyt chřipky ve městě a na venkově. Výsledky je možné zapsat do čtyřpolní tabulky

		město	venkov	
chřipka	ano	n_{11}	n_{12}	$n_{1.}$
	ne	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.

Cochran–Armitage test

Tímto testem se zjišťuje lineární nárůst sledovaného jevu v závislosti na rostoucí hodnotě uspořádané veličiny. Data v tomto případě tvoří kontingenční tabulka se dvěma řádky

	Y_1	$\dots$	Y_J	
$X = 1$	$n_{1,1}$	$\dots$	$n_{1,J}$	$n_{1.}$
$X = 0$	$n_{2,1}$	$\dots$	$n_{2,J}$	$n_{2.}$
	$n_{.1}$	$\dots$	$n_{.J}$	n

Příklad. *Zjišťujeme, zda nákup výrobku závisí na spokojenosti zákazníka při nákupu v obchodě. Proměnnou Y tvoří hodnocení spokojenosti na vícebodové škále, proměnná X pak má hodnotu 1 pokud zákazník výrobek koupil a hodnotu 0, pokud nekoupil.*

Cochran–Armitage test

Testují se hypotézy

- H_0 : procento daného jevu ($X = 1$) je stejné pro všechny úrovně ordinální proměnné
- H_1 : procento daného jevu se s rostoucí úrovní ordinální proměnné lineárně mění (roste nebo klesá)

Testová statistika má tvar

$$Z = \sum_{i=1}^{k} \frac{(n_{1,i}(R_i - \overline{R}))}{\sqrt{p_{1.}(1 - p_{1.}s^2)}}$$

kde

$$\overline{R} = \sum_{i=1}^{k} R_i n_{.i}/n, \qquad p_{1.} = n_{1.}/n, \qquad s^2 = \sum_{i=1}^{k} n_{1.}(R_i - \overline{R})^2$$

a R_i jsou skóry, většinou hodnoty $1, \ldots, k$. Za platnosti nulové hypotézy má testová statistika $N(0, 1)$ rozdělení.

Kendallovo τ

Pokud chceme zjistit, zda je lineární vztah mezi dvěma uspořádanými kategorickými proměnnými, je na místě uvažovat obdobu korelačního koeficientu. Pearsonův korelační koeficient, který se používá pro číselné proměnné, zde není zcela vhodný. Pro ordinální data se používá **Kendallovo** τ.

Uvažujme dvě porovnávané proměnné a označme je X a Y. Pro každou jednotku tak máme naměřenu dvojici hodnot (X_i, Y_i). Nyní uvažujme všechny dvojice jednotek a pokud pro danou dvojici platí, že $X_i < X_j \& Y_i < Y_j$ nebo $X_i > X_j \& Y_i > Z_j$, pak označme tuto dvojici jakou souhlasnou, pokud platí $X_i < X_j \& Y_i > Y_j$ nebo $X_i > X_j \& Y_i < Y_j$, označme ji za nesouhlasnou.

Kendallovo τ je založeno na rozdílu počtu souhlasných (n_s) a počtu nesouhlasných (n_n) dvojic.

Kendallovo τ

Konkrétně je **Kendallovo** τ definováno jako

$$\tau = \frac{n_s - n_n}{n} = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)$$

Rozptyl tohoto koeficientu je

$$\text{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

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

Kendallovo τ

Výše uvedený koeficient funguje dobře, pokud v datech nejsou stejné hodnoty. Pokud se stejné hodnoty vyskytnou, používají se následující obdoby tohoto koeficientu.
Pro proměnné se **stejným počtem hodnot**

$$\tau_B = \frac{n_s - n_n}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

kde $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$ a t_i jsou počty shodných hodnot u proměnné X, $n_1 = \sum_i u_i(u_i - 1)/2$ a u_i jsou počty shodných hodnot u proměnné Y.
Pro proměnné s **různým počtem hodnot**

$$\tau_C = \frac{2(n_s - n_n)}{n^2 \frac{m-1}{m}},$$

kde m je minimální počet hodnot u obou proměnných.
Výpočet rozptylů a následných testových statistik pro τ_B a τ_C je složitý. Přenechme ho tedy softwarům.

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	$F = MSA/MSe$	p
Chyba e	SSe	$dfe = n - k$	MSe		
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

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

ta 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.

Kruskal-Wallisův test

V případě, že není splněn předpoklad normality při porovnání více než dvou nezávislých výběrů, používá se **Kruskal-Wallisova ANOVA**. Kruskal-Wallisova ANOVA je přímým zobecněním Wilcoxonova dvouvýběrového testu. Testujeme

- H_0 : Střední hodnoty výběrů se neliší
- H_1 : Střední hodnoty výběrů se liší

Stejně jako u dvouvýběrového Wilcoxonova testu srovnáme všechny naměřené hodnoty do řady, určíme jejich pořadí a spočteme statistiky $T_1, \ldots, T_k$, kde k je počet výběrů. Pak platí, že testová statistika

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i}{n_i} - 3(n+1)$$

má za platnosti H_0 χ^2-rozdělení.

Friedmanův test

Friedmanův test umožňuje porovnat mezi sebou několik (více než 2) závislých výběrů. Jedná se o neparametrický test a hodí se tedy v případě, že data nemají normální rozdělení, nebo pokud porovnáváte hodnoty ordinální proměnné.
Testují se hypotézy

- H_0 : Střední hodnoty výběrů se neliší
- H_1 : Střední hodnoty výběrů se liší

Příklad. *Příkladem může být ochutnávka sušenek, jogurtů, atd., kdy každý jednotlivec dostane púřed sebou stejné vzorky, ochutná je a přiřadí jim bodové ohodnocení.*
Test probíhá tak, že se v rámci každého bloku (tj. v rámci jednotky, přes níž jsou výběry závislé) stanoví pořadí výběrů (výrobků) a pak se pro každý výběr tato pořadí sečtou. Jsou-li součty přibližně stejné, není mezi výběry rozdíl, pokud se součty lkiší, rozdíly existují.

Friedmanův test

Předpokládejme, že porovnáváme k výběrů a pro porovnání máme k dispozici l bloků. V každé kombinaci i-tý výběr, j-tý blok je naměřena pouze jedna hodnota.
Testová statistika má tvar

$$Q = \frac{12n}{k(k+1)} \sum_{i=1}^{k} \left(\bar{r}_{.i} - \frac{k+1}{2} \right)^2$$

kde

$$\bar{r}_{.i} = \frac{1}{l} \sum_{j=1}^{l} r_{ji}$$

a r_{ji} jsou pořadí výsledků v rámci j-tého bloku. Hodnota $\bar{r}_{.i}$ je tedy průměrné pořadí v i-tém výběru.
Za platnosti nulové hypotézy má testové statistika χ^2-rozdělení o $k-1$ stupních volnosti.

Kuchařka

- **Popis**
  - číselná proměnná – popisné statistiky polohy, variability, (tvaru rozdělení), krabicový graf, histogram
    – interval spolehlivosti pro průměr, jednovýběrový t-test nebo jednovýběrový Wilcoxonův test
  - kategorická proměnná – absolutní a relativní četnosti, kumulativní absolutní i relativní četnosti, sloupcový graf, koláčový graf,
    – interval spolehlivosti pro pravděpodobnost, test o hodnotě pravděpodobnosti

- **Vztah dvou proměnných**
  - Číselná vs. kategorická – graficky pomocí krabicových grafů
    – kategorická se dvěmi kategoriemi – dvouvýběrový t-test nebo dvouvýběrový Wilcoxonův test
    – kategorická s více kategoriemi – klasická ANOVA nebo Kruskal-Wallisova ANOVA
  - Dvě kategorické – obě nominální – χ^2-test, Fisherův exaktní test, poměr šancí
  - Dvě kategorické – jedna ordinální, druhá nominální – Kruskal-Wallisova ANOVA, Friedmanův test, Cochran-Armitage test
  - Dvě kategorické – obě ordinální – Kendallův korelační koeficient
  - Dvě číselné – graficky pomocí bodového grafu, Pearsonův nebo Spearmanův korelační koeficient