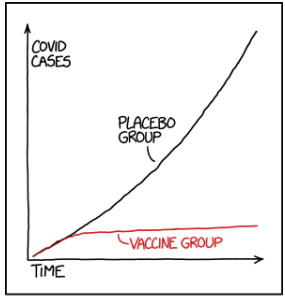
Výpočetní technika a lékařská informatika

Základy biostatistiky a analýzy dat

Mgr. Markéta Trnečková, Ph.D.

Data

- data – pojem známe z minula
- co s nimi dál?
- **Statistika**
  - popisná statistika (deskriptivní) – poskytuje pouze výčet pozorovaných jedinců a jejich vlastností, nepokouší se vyslovovat k vlastnostem jedinců, kteří nebyli sledováni
  - zobecňující statistika (induktivní) – umožňuje zkoumat pouze část celé populace a získat pokud možno co nejpřesnější společný odhad sledovaných (obecných) charakteristik celé populace



Statistika

- Zkoumáme **statistický soubor** (skupinu, množinu, ...) na základě nějakého znaku
- **Znaky**:
  - Kvalitativní – popisujeme různými slovy (omezenou skupinou slov) – např. muž/žena, jméno, ...
  - Kvantitativní – hodnoty jsou uspořádány a mohou vyjadřovat dokonce i určitou míru – věk, váha, ...
- **Data s neúplnou informací** – jevy, které nejsme schopni vždy přesně pozorovat (můžeme jen stanovit interval, ve kterém se sledovaná hodnota nalézá)

Příklad

V souboru `pacienti_dataset.xlsx` určete, co jsou sledované znaky. U každého určete, zda je kvalitativní nebo kvantitativní.

Statistika

- statistický soubor se skládá z **jednotek**, jejich počet je n
- x_i hodnota znaku i-té jednotky
- x_j^* jedinečné hodnoty
- n_j četnost x_j^* (kolikrát je x_j^* v souboru zastoupena)
- relativní četnost p_j četnost n_j/n (pravděpodobnost)
- kumulativní četnost c_j četnost $\sum_{k=1}^{j} n_k$ (případně kumulativní relativní četnost) – má význam u ordinálních dat (např. vzdělání – kdo má SŠ tak má určitě i ZŠ)

Příklad

Jaké jsou jednotky sledovaného znaku Pohlaví? Jaké jsou jedinečné hodnoty, jejich četnost a relativní četnost?

Statistika

Příklad (Jak to udělat v excelu?)

Jaké jsou jedinečné hodnoty?

`=UNIQUE(C2:C51)` → vložte do N2

Příklad (Jak to udělat v excelu?)

Jaká je jejich četnost?

`=COUNTIF(C2:C51;N2)` → vložte do O2

rozšiřte tento vzorec i na další řádek.

Příklad (Jak to udělat v excelu?)

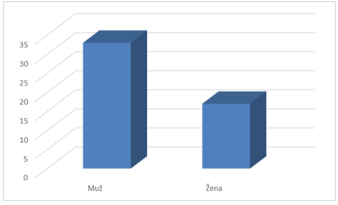
Jaká je jejich relativní četnost?

`=O2/POČET2(C2:C51)` → vložte do P2

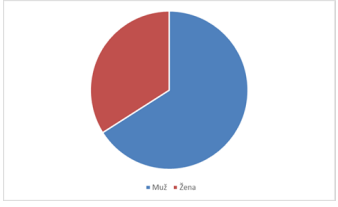
Vše lze udělat i pomocí kontingenčních tabulek.

Znázornění

- **sloupcový graf**



- **sloupcový graf**

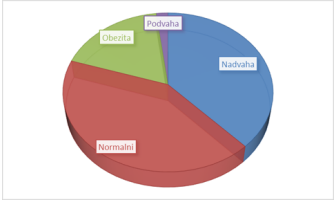


Příklad

Do tabulky v excelu přidejte sloupcový graf pro Pohlaví a koláčový graf pro Kouření.

Znázornění

- Některá data obsahují velké množství různých hodnot
- je vhodné je rozdělit na intervaly a dle toho znázornit (sloupcový graf, histogram)



Příklad

Do tabulky v excelu přidejte libovolný graf pro BMI. Data vhodně rozdělte do intervalů.

Znázornění

Příklad

Do tabulky v excelu přidejte libovolný graf pro BMI. Data vhodně rozdělte do intervalů.

- Pomocí příkazu `IFS` určíme interval
  `=IFS(I2<18,5;"Podvaha";I2<25;"Normalni";I2<30;"Nadvaha";PRAVDA;"Obezita")`
- spočítáme četnosti jednotlivých hodnot, jako dříve
- případně histogram (excel automaticky rozdělí hodnoty do intervalů)

Základní statistické charakteristiky

- **Rozsah hodnot** – interval od **minimální** do **maximální hodnoty**

Příklad

Zjistěte rozsah hodnot Věku. (minimální a maximální hodnotu)

- **Aritmetický průměr** – průměrná hodnota
- **Medián** – prostřední hodnota
- **Modus** – hodnota, která se vyskytuje nejčastěji

Příklad

Zjistěte průměrný Věk a medián a modus tohoto sloupce.

Základní statistické charakteristiky

- **Rozptyl** (variance) – rozptyl „průměrný čtverec vzdáleností“ naměřených hodnot od průměru
  $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- **Směrodatná odchylka** – odmocnina rozptylu
  $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
  Je mírou „různorodosti“ pozorovaných hodnot

Příklad

Zjistěte rozptyl a směrodatnou odchylku hodnot ve sloupci Věk.

- **Střední chyba** – Vyjadřuje míru variability (přesnost stanovení) průměru
  počítá se jako směrodatná odchylka vydělená odmocninou počtu měření
  $SE = \frac{s}{\sqrt{n}}$

Příklad

Vypočítejte střední chybu průměru věku.

Testy charakteristik

- Jednovýběrové testy
  - slouží k ověření, zda se průměr (nebo jiný parametr) jedné populace statisticky významně liší od známé hodnoty
  - např. ověření, zda průměrný cholesterol v souboru je vyšší než doporučená hodnota (např. $<$ 5,0 mmol/L)
  - Z-test, t-test,
- Dvouvýběrové testy
  - používají se k porovnání, zda se průměry (nebo jiné parametry) dvou nezávislých populací statisticky významně liší
  - např. srovnání průměrného BMI mužů a žen v souboru pacientů

Základní pojmy testování hypotéz

Nulová hypotéza (H_0)

- Předpoklad, který testujeme.
- Obvykle znamená žádný efekt nebo rozdíl.
- Příklad (medicína): Průměrná hladina cholesterolu pacientů s diabetem je stejná jako u zdravé populace.

Alternativní hypotéza (H_1)

- Předpoklad, který přijmeme, pokud data odporují H_0.
- Obvykle znamená existenci efektu nebo rozdílu.
- Příklad: Průměrná hladina cholesterolu pacientů s diabetem se liší od zdravé populace.

Konfidenční interval a p-hodnota

Konfidenční interval (Confidence Interval, CI)

- Interval, který s určitou pravděpodobností obsahuje skutečný parametr populace.
- Např. 95% CI: pokud bychom test opakovali mnohokrát, 95% intervalů by obsahovalo skutečný průměr populace.

p-hodnota (p-value)

- P-hodnota je pravděpodobnost, že bychom pozorovali data stejně extrémní nebo ještě extrémnější než ta naměřená, pokud by nulová hypotéza (H_0) byla pravdivá.
- Malá p-hodnota (např. $< 0,05$) $\rightarrow$ data jsou nepravděpodobná za předpokladu $H_0 \rightarrow$ odmítáme nulovou hypotézu.

Obecný postup výpočtu p-hodnoty:

1. Spočítáme testovou statistiku podle druhu testu
2. Určíme typ testu: pravostranný, levostranný, oboustranný.
3. Najdeme pravděpodobnost extrémnějších hodnot statistiky podle příslušného rozdělení
4. Porovnáme p-hodnotu s hladinou významnosti α:
   - $p \leq \alpha \rightarrow$ zamítáme H_0
   - $p > \alpha \rightarrow$ nezamítáme H_0

Statistická významnost

Statistická významnost

- Výsledek je **statisticky významný**, pokud $p < \alpha$ (typicky $\alpha = 0,05$).
- Znamená, že je výsledek dostatečně nepravděpodobný za předpokladu H_0.

Jednovýběrové a vícevýběrové testy

Jednovýběrové testy

- Zkoumají **vlastnosti jednoho souboru dat**.
- Typická otázka: „Liší se průměrná hodnota od známé (doporučené) hodnoty?“
- Příklad: Ověření, zda průměrný cholesterol pacientů je vyšší než 5,0 mmol/L.

Vícevýběrové testy (dvouvýběrové a více)

- Porovnávají **dvě nebo více skupin**.
- Typická otázka: „Liší se skupiny mezi sebou?“
- Dvouvýběrové testy – dvě skupiny (např. ženy vs. muži).
- ANOVA – více než dvě skupiny (např. různé věkové kategorie pacientů).

Statistické testy a metody

Jednovýběrové testy

- **Jednovýběrový t-test** – porovnání průměru se známou hodnotou.
- **Shapiro-Wilk test** – ověření normality dat.
- **Chi-kvadrát test dobré shody** – zda data odpovídají očekávanému rozdělení.

Dvouvýběrové testy

- **Dvouvýběrový t-test** – porovnání průměrů dvou skupin.
- **Mann-Whitney U test** – neparametrická alternativa k t-testu.
- **Chi-kvadrát test nezávislosti** – vztah mezi kategoriálními proměnnými.

Další metody (medicína a biostatistika)

- **ANOVA** – porovnání více než dvou skupin.
- **Korelace a regrese** – vztah mezi dvěma (či více) proměnnými.
- **Kaplan-Meier analýza** – přežití pacientů v čase.

Jednovýběrový t-test

- Jednovýběrový t-test je statistický test, který porovnává průměr vzorků s referenční hodnotou (např. průměrem populace).
- **Hypotézy:**
  - Nulová hypotéza H_0: $\mu = \mu_0$ (průměr vzorku se rovná referenční hodnotě)
  - Alternativní hypotéza H_1: $\mu \neq \mu_0$ (dvoustranný test) nebo $\mu > \mu_0$ / $\mu < \mu_0$ (jednostranný test)
- **Testovací statistika:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

kde:

- $\bar{x}$ = průměr vzorku
- μ_0 = referenční hodnota
- s = směrodatná odchylka vzorku
- n = velikost vzorku

Příklad: Jednovýběrový t-test

Příklad

Cíl: Ověřit, zda je průměrná hodnota cholesterolu v souboru pacientů vyšší než doporučená hranice 5,0 mmol/L.

Statistická formulace:

- **Nulová hypotéza (H_0):** $\mu = 5,0$
- **Alternativní hypotéza (H_1):** $\mu > 5,0$
- **Signifikance:** $\alpha = 0,05$
- **Test:** Jednovýběrový t-test (one-sample t-test)
- **Data:** Sloupec `Cholesterol` v tabulce pacientů

Úkol: Spočítejte testovou statistiku (t-test), p-hodnotu a rozhodněte, zda zamítneme H_0.

Příklad: Jednovýběrový t-test – cholesterol

Postup řešení:

1. **Vytvoření dat:** Sloupec *Cholesterol* obsahuje naměřené hodnoty, $F2 : F51$.
2. **Definice hypotéz:**
   - H_0: průměrný cholesterol = 5,0 mmol/L
   - H_1: průměrný cholesterol > 5,0 mmol/L (jednostranný test)
3. **Spočítejte základní charakteristiky:**
   - Průměr: `=PRŮMĚR(F2:F51)`
   - Směrodatná odchylka: `=STDEVA(F2:F51)`
   - Počet hodnot: `=POČET(F2:F51)`
4. **Testová statistika**:
   $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
   `=(PRŮMĚR(F2:F51)-5)/(STDEVA(F2:F51)/ODMOCNINA(POČET(F2:F51)))`
5. **Určení p-hodnoty**:
   - jednostranný test ($\bar{x} > \mu_0$):
     `=T.DIST.RT(t; POČET(F2:F51)-1)`
6. **Interpretace:**
   - $p < 0,05 \Rightarrow$ zamítáme H_0, průměrný cholesterol v souboru je statisticky významně vyšší než 5,0 mmol/L
   - $p \geq 0,05 \Rightarrow$ nemáme dostatek důkazů, že průměr je vyšší než 5,0 mmol/L

Dvouvýběrový t-test

Definice:

- Statistický test porovnávající průměry dvou skupin
- Nulová hypotéza H_0: $\mu_1 = \mu_2$
- Alternativní hypotéza H_1: $\mu_1 \neq \mu_2$ (oboustranný), nebo $\mu_1 > \mu_2$, $\mu_1 < \mu_2$ (jednostranný).

Použití v medicíně:

- Porovnání průměrného BMI mužů a žen
- Účinnost léčby: skupina pacientů s novým lékem vs. kontrolní skupina
- Průměrná hladina cholesterolu u kuřáků a nekuřáků

Typy:

- *Nezávislé výběry* (dvě různé skupiny pacientů)
- *Závislé výběry* (opakovaná měření u stejných pacientů)

Příklad: Dvouvýběrový t-test

Příklad

Otázka: Mají muži a ženy v našem souboru odlišný průměrný BMI?

Hypotézy:

- H_0: Průměrný BMI mužů $=$ průměrný BMI žen
- H_1: Průměrný BMI mužů $\neq$ průměrný BMI žen

Typ testu: Dvouvýběrový t-test pro nezávislé výběry, dvoustranný.

Řešení v Excelu

Postup:

1. Rozdělíme data na dvě oblasti:
   - BMI mužů (`=FILTER(I2:I51;C2:C51="Muž")`)
   - BMI žen (`=FILTER(I2:I51;C2:C51="Žena")`)
2. Použijeme funkci:
   $= T.TEST(FILTER(I2:I51; C2:C51 = "Muž"); FILTER(I2:I51; C2:C51 = "Žena"); 2; 2)$
   - $2 =$ dvoustranný test
   - $2 =$ dvouvýběrový test pro nezávislé výběry
3. Porovnáme p-hodnotu (výsledek funkce) s hladinou významnosti $\alpha = 0,05$.

Interpretace:

- p $< 0{,}05 \rightarrow$ zamítáme H_0 (BMI se liší mezi pohlavími).
- p $\geq 0{,}05 \rightarrow$ nezamítáme H_0 (BMI se statisticky významně neliší).

Řešení v Excelu

Vykreslení výsledku:

1. Vytvořte sloupcový graf pro průměrné hodnoty BMI mužů a žen
2. Přidejte směrodatné odchylky:
   - Klikněte na sloupce grafu → Přidat prvek grafu → Chybové úsečky → Další možnosti
   - Vyberte Vlastní a zadejte hodnoty standardních odchylek
   - Graf ukáže, jak moc se průměry liší vzhledem k rozptylu dat

Chi-kvadrát test nezávislosti

Definice: *Chi-kvadrát test nezávislosti* slouží k ověření, zda jsou dvě kategoriální proměnné statisticky nezávislé.

Použití:

- Analýza asociace mezi dvěma kategoriálními proměnnými
- Příklady:
  - Kouření vs. diagnóza pacienta
  - Pohlaví vs. volba produktu
  - Účast v programu vs. výsledek testu
- Výsledek testu: p-hodnota $\rightarrow$ rozhodnutí o nezávislosti (pravděpodobně existuje rozdíl mezi skupinami ve sledované kategorii)

Poznámka:

- Data musí být uspořádána v kontingenční tabulce
- Četnosti v jednotlivých buňkách by měly být dostatečně velké

Kontingenční tabulka se ve statistice užívá k přehledné vizualizaci vzájemného vztahu dvou statistických znaků.

Příklad: Chi-kvadrát test nezávislosti

Příklad

Zjistěte, zda je kouření nezávislé na pohlaví pacienta.

Postup v Excelu:

1. Vytvořte kontingenční tabulku Muž/Žena vs. Kouří/Nekouří (př. počet žen, které kouří `=COUNTIFS(C1:C51;"Žena";H1:H51;"Ano")`

	Nekouří	Kouří
Muži	Počet mužů, kteří nekouří	Počet mužů, kteří kouří
Ženy	Počet žen, které nekouří	Počet žen, které kouří

2. Vytvořte tabulku očekávaných hodnot

	Nekouří	Kouří
Muž	$E_{11} = \frac{(\text{suma řádku Muž})(\text{suma sloupce Nekouří})}{\text{celkem}}$	$E_{12} = \frac{(\text{suma řádku Muž})(\text{suma sloupce Kouří})}{\text{celkem}}$
Žena	$E_{21} = \frac{(\text{suma řádku Žena})(\text{suma sloupce Nekouří})}{\text{celkem}}$	$E_{22} = \frac{(\text{suma řádku Žena})(\text{suma sloupce Kouří})}{\text{celkem}}$

3. Použijte funkci `CHISQ.TEST` (`=CHISQ.TEST(kontingenční tabulka;tabulka očekávaných hodnot)`)
4. Zkontrolujte p-hodnotu a rozhodněte, zda zamítáte nulovou hypotézu nezávislosti.

Korelace – definice a výpočet

Definice:

Pearsonův korelační koeficient r měří sílu lineární závislosti dvou proměnných.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Interpretace:

- $r = 1$ ... dokonalá pozitivní lineární závislost
- $r = -1$ ... dokonalá negativní lineární závislost
- $r \approx 0$ ... žádná lineární závislost

Příklad: Korelace

Příklad

Rozhodněte, zda existuje lineární vztah mezi BMI a tlakem. Spočítejte Pearsonův korelační koeficient r.

Postup v Excelu

1. Použijte funkci:
   `=CORREL(I2:I51;G2:G51)`
2. Výsledek = hodnota r

Interpretace:

- $r > 0$ ... vyšší BMI je spojeno s vyšším tlakem
- $r < 0$ ... vyšší BMI je spojeno s nižším tlakem
- $r \approx 0$ ... lineární vztah není prokazatelný

ANOVA

Definice:

Analýza rozptylu (ANOVA) testuje, zda se průměry více než dvou skupin liší.

$$F = \frac{MS_{\text{mezi}}}{MS_{\text{uvnitř}}}$$

MS – průměrný čtverec odchylek.

Hypotézy:

- H_0: všechny průměry jsou stejné
- H_A: alespoň jeden průměr se liší

Typy ANOVY:

- Jednofaktorová ANOVA – vliv jednoho faktoru (např. diagnóza)
- Dvoufaktorová ANOVA – vliv dvou faktorů (např. diagnóza a pohlaví)
- ANOVA s opakovanými měřeními – měření u stejných jedinců v čase

Použití v medicíně:

- porovnání účinnosti léčby mezi více skupinami pacientů
- sledování vlivu různých faktorů na zdraví

ANOVA – výpočet krok za krokem

Kroky výpočtu:

1. Celkový průměr: $\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$
2. Součet čtverců mezi skupinami: $SS_{\text{mezi}} = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$
3. Součet čtverců uvnitř skupin: $SS_{\text{uvnitř}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$
4. Stupně volnosti: $df_{\text{mezi}} = k - 1, \qquad df_{\text{uvnitř}} = n - k$
5. Průměrné čtverce (Mean Squares): $MS_{\text{mezi}} = \frac{SS_{\text{mezi}}}{df_{\text{mezi}}}, \qquad MS_{\text{uvnitř}} = \frac{SS_{\text{uvnitř}}}{df_{\text{uvnitř}}}$
6. Testová statistika: $F = \frac{MS_{\text{mezi}}}{MS_{\text{uvnitř}}}$

- n ... celkový počet pozorování
- k ... počet skupin
- n_i ... počet pozorování v i-té skupině

Příklad: jednofaktorová ANOVA

Příklad

Ověřte, zda se BMI liší mezi skupinami pacientů podle diagnózy.

Data:

- Sloupce představují skupiny pacientů podle diagnózy
- Každý řádek obsahuje hodnotu BMI pacienta

ANOVA v Excelu

Předpoklad: čtyři skupiny v rozsazích `U2:U9`, `V2:V15`, `W2:W11`, `X2:X19`
průměry skupin v `U21`, `V21`, `W21`, `X21`.

1. **Celkový průměr** $\bar{x}$ (`Y21`)
   `=PRŮMĚR(U2:U9;V2:V15;W2:W11;X2:X19)`
2. **SS mezi skupinami** $SS_{\text{mezi}} = \sum_{i=1}^{4} n_i(\bar{x}_i - \bar{x})^2$
   `=POČET(U2:U9)*(U21-Y21)^2+POČET(V2:V15)*(V21-Y21)^2`
   `+POČET(W2:W11)*(W21-Y21)^2+POČET(X2:X19)*(X21-Y21)^2`
3. **SS uvnitř skupin**
   1. Pro každou skupinu vytvořte pomocný sloupec: $(x_{ij} - \bar{x}_i)^2$
      `=(U2-U21)^2`
   2. hodnoty sečtěte $SS_{\text{uvnitř}}$
      `=SUMA(U24:U31)+SUMA(V24:V37)+SUMA(W24:W33)+SUMA(X24:X41)`
4. **df, MS, F a p-hodnota**
   k = 4, n = `=POČET(U24:X41)`, df_{mezi} = k-1, $df_{\text{uvnitř}}$ = n-k
   $MS_{\text{mezi}} = SS_{\text{mezi}}/df_{\text{mezi}}$
   $MS_{\text{uvnitř}} = SS_{\text{uvnitř}}/df_{\text{uvnitř}}$
   $F = MS_{\text{mezi}}/MS_{\text{uvnitř}}$
   p = `F.DIST.RT(F;` $df_{\texttt{mezi}}$`;` $df_{\texttt{uvnitř}}$`)`

ANOVA v Excelu

Testované hypotézy:

- H_0: Všechny skupinové průměry BMI jsou stejné (nezávisí na diagnóze).
- H_A: Alespoň jeden průměr se liší.

Testovací statistika **F** je poměr průměrného čtverce mezi skupinami k průměrnému čtverci uvnitř skupin:
$F = \frac{MS_{\text{mezi}}}{MS_{\text{uvnitř}}}$

- MS_{mezi} – variabilita mezi průměry skupin
- $MS_{\text{uvnitř}}$ – průměrná variabilita uvnitř skupin
- $F \approx 1$ – skupiny se statisticky neliší (variabilita mezi skupinami je podobná jako uvnitř skupin)
- $F \gg 1$ – mezi skupinami je větší rozdíl než uvnitř $\rightarrow$ možné statisticky významné rozdíly

Poznámka: Skutečný význam F hodnoty určuje *p-hodnota*, kterou porovnáváme s hladinou významnosti (např. 0,05).

- Pro $p \geq 0{,}05$, nulovou hypotézu **nezamítáme**.
- Statisticky tedy **není prokázán rozdíl** v BMI mezi skupinami pacientů podle diagnózy.