Úvod do statistické analýzy 2. část

Ing. Josef Chudoba, Ph.D.

Ústav nových technologií a aplikované informatiky
Fakulta mechatroniky, Technická univerzita v Liberci

Verze 1.1.1 – 16. 2. 2021

Učební text vychází především ze skript:

1) M. Litschmannová – Vybrané kapitoly z pravděpodobnosti, Ostrava 2011, VŠB-TU Ostrava
2) M. Litschmannová – Úvod do statistiky, Ostrava 2011, VŠB-TU Ostrava

Obsah

7 – Teorie odhadu

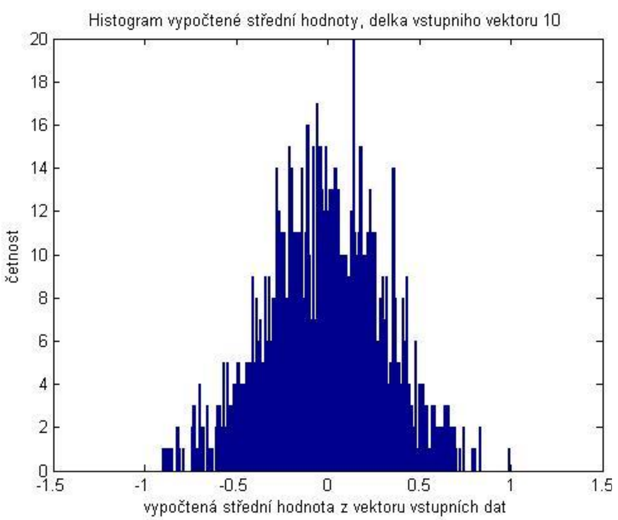
- 7.1 Úvod do teorie odhadu
- 7.2 Bodové odhady
- 7.3 Intervalové odhady

- 1 výběr
  - 7.4 Intervalový odhad střední hodnoty normálního rozdělení
  - 7.5 Intervalový odhad rozptylu normálního rozdělení
  - 7.6 Intervalový odhad směrodatné odchylky normálního rozdělení
  - 7.7 Intervalový odhad relativní četnosti
  - 7.8 Odhad rozsahu výběru
  - 7.9 Intervalový odhad mediánu
  - 7.10 Intervalový odhad parametrů spojitých rozdělení

- 2 výběry
  - 7.11 Intervalový odhad poměru rozptylů dvou výběrů s normálním rozdělením
  - 7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením
  - 7.13 Intervalový odhad pro rozdíl relativních četností dvou populací

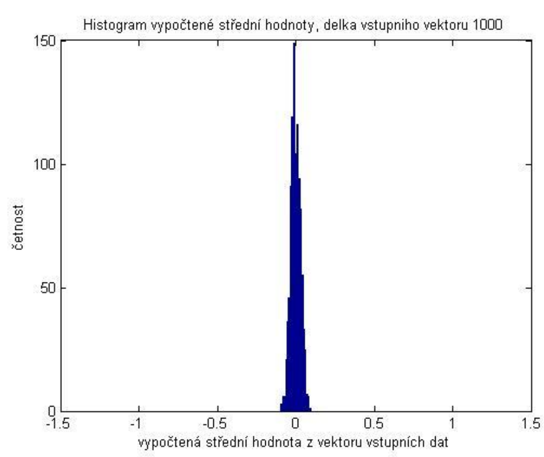
7.1 Úvod do teorie odhadu

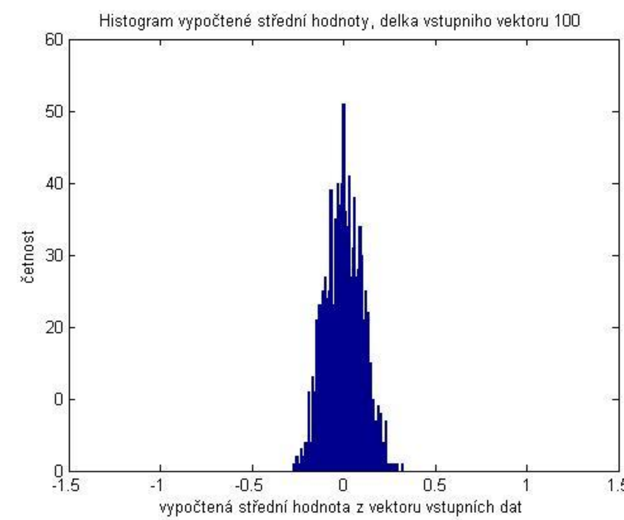
- Př. Vygenerujeme 1000 náhodných vektorů z normovaného normálního rozdělení ($\mu = 0, \sigma^2 = 1$), které mají délku n. Vytvoříme průměr z vektoru n a vyneseme do histogramu.
- Z vlastností výběrového průměru: $E(\bar{X}) = \frac{\sum_{i=1}^{n} E(X_i)}{n}$ a $\mathrm{D}(\bar{X}) = \frac{\sum_{i=1}^{n} D(X_i)}{n^2}$ lze odhadnout střední hodnotu a rozptyl v histogramech.
- V případě menšího množství vygenerovaných dat střední hodnota více kolísá.



$$E(X) = 0$$
$$D(X) = \frac{1}{10}$$
$$\sigma(X) = 0.316$$

$$E(X) = 0$$
$$D(X) = \frac{1}{1000}$$
$$\sigma(X) = 0.0316$$





$$E(X) = 0$$
$$D(X) = \frac{1}{100}$$
$$\sigma(X) = 0.1$$

7.1 Úvod do teorie odhadu

- U příkladu byla data vygenerována z normovaného normálního rozdělení se známou střední hodnotou a rozptylem.
- V praxi je situace odlišná – máme data o délce n, vypočteme výběrovou střední hodnotu. Chceme vědět v jakém intervalu se s určitou pravděpodobností bude nacházet skutečná střední hodnota získaná z „velkého množství dat“.

- Příklad výsledku:
  – Z 10 vstupních dat jsme zjistili výběrovou střední hodnotu 0,5. S pravděpodobností 95 % se bude skutečná střední hodnota nacházet v intervalu $\langle 0.15; 0.85\rangle$.

7.2 Bodové odhady

- Mějme náhodný výběr $X_1, X_2, \cdots, X_n$ z určitého rozdělení, které závisí na parametru Θ. Odhadem T parametru Θ je pak výběrová charakteristika $T(X_1, X_2, \cdots, X_n)$, která nabývá hodnot „blízkých“ k neznámému parametru Θ.
- Parametr Θ může být například střední hodnota, rozptyl, četnost, medián apod.
- Bodový odhad musí splňovat základní vlastnosti:
  - Nestrannost
  - Konzistentnost
  - Vydatnost
  - Robustnost

7.2 Bodové odhady

- Nestrannost
  - Výběrová charakteristika T je nestranným odhadem statistiky Θ, je-li $E(T) = \Theta$.
  - Platí-li, že $\lim_{n\to\infty} (E(T_n) - \Theta\ ) = 0$, pak je statistika T asymptoticky nestranným odhadem Θ.
- Konzistentnost
  - Za konzistentní odhad statistiky Θ označíme takovou statistiku T, která splňuje rovnost

$$\lim_{n\to\infty} P(|\Theta - T_n| < \varepsilon\ ) = 1$$

  - Jestliže je bodový odhad parametru Θ konzistentní, pak je malá pravděpodobnost, že se při zvyšujícím se rozsahu výběru dopustíme velké chyby při odhadu parametru Θ.

7.2 Bodové odhady

- Vydatnost
  - Za vydatný odhad statistiky Θ se označí taková statistika, která má ze všech nestranných odhadů nejmenší rozptyl.
  - $\lim_{n \to \infty} D(T_n) = 0$
- Robustnost
  - Za robustní odhad statistiky Θ se označí taková statistika, která není výrazně ovlivnitelná hodnotami způsobené například hrubou chybou.
  - Robustní odhad není – minimum a maximum
  - Robustní odhad je aritmetický průměr

7.3 Intervalové odhady

- Odhad parametru Θ určujeme pomocí intervalového odhadu.
  - Interval spolehlivosti (také konfidenční interval) pro parametr Θ je taková dvojice statistik (T_d, T_h), pro kterou s pravděpodobností 1-α ($\alpha \in \langle 0,1 \rangle$) platí:

$$P(T_d \leq \Theta \leq T_h) = 1 - \alpha$$

  - Parametr Θ může být střední hodnota, medián, rozptyl, četnost apod.
  - Často u intervalových odhadů uvádíme „pravděpodobnost 1-α“ jako „spolehlivost 1-α“

7.3 Intervalové odhady

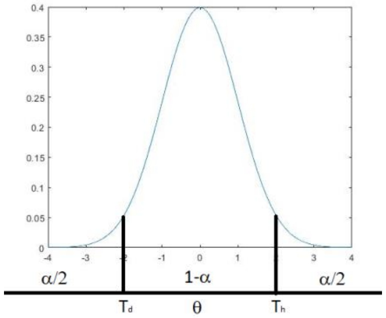
- Spolehlivost odhadu 1-α předpokládá, že při opakovaných výběrech s konstantním rozsahem n z dané populace:
  - $100 \cdot (1-\alpha)$ % intervalových odhadů obsahuje skutečnou hodnotu odhadovaného parametru Θ
  - $100 \cdot \alpha$ % intervalových odhadů neobsahuje skutečnou hodnotu odhadovaného parametru Θ

- Spolehlivost odhadu 1-α požadujeme blízkou 1.
  - α se obvykle uvažuje 0.05, vzácněji 0.1 nebo 0.01
  - Se snižujícím se α se rozšiřuje šířka intervalu.
- Pro zúžení intervalu je třeba mít více naměřených dat.
  - Pro dvojnásobné zúžení intervalu je třeba míti 4x více dat.

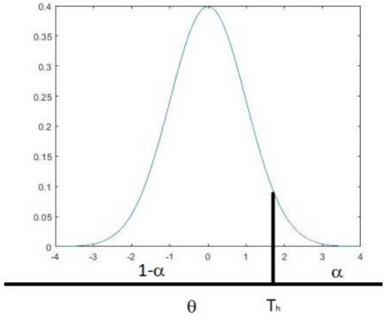
7.3 Intervalové odhady

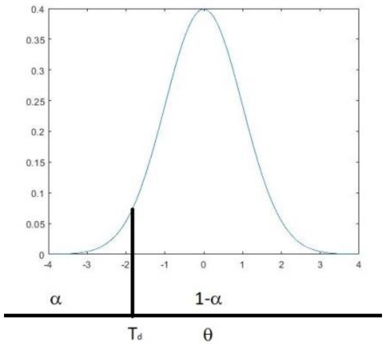
- Na obrázku je z dat získána střední hodnota 0. Intervalový odhad označuje interval, ve kterém se bude s pravděpodobností $1-\alpha$ nacházet skutečná střední hodnota.

- Oboustranný interval spolehlivosti
  - U oboustranných intervalů spolehlivosti hledáme interval $< T_d , T_h >$, ve kterém daný parametr leží se spolehlivostí 1-α.
  - Výsledek udává obě meze T_d i T_h.
  - $P(\Theta < T_d) = \frac{\alpha}{2}$ $\quad P(\Theta > T_h) = \frac{\alpha}{2}$ $\quad P(T_d \leq \Theta \leq T_h) = 1-\alpha$



- Jednostranné intervaly spolehlivosti
  - V matlabu označený „left“
    - U intervalu spolehlivosti se udává pouze horní mez T_h
    - Potom jednostranný interval nabývá $P(\Theta \leq T_h) = 1-\alpha$ interval $(-\infty, T_H)$.
  - V matlabu označený „right“
    - U intervalu spolehlivosti se udává pouze dolní mez T_d
    - Potom jednostranný interval nabývá $P(\Theta \geq T_d) = 1-\alpha$ interval (T_D, ∞).





- V matlabu je přednastaven oboustranný interval spolehlivosti. Jednostranné intervaly se definují pomocí „left“, „right“.
- Pozor v matlabu nekoresponduje „left“ s českým označením levostranný („right“ s pravostranný) interval spolehlivosti.

7.4 Intervalový odhad střední hodnoty normálního rozdělení

- Rozlišují se 2 případy
  - 1) směrodatná odchylka σ je předem známá a definována
  - 2) směrodatná odchylka σ je neznámá
- Nutný předpoklad – data pocházejí z normálního rozdělení

- 7.4.1 σ je předem známá – ztest (vzácný případ)
- 7.4.2 příklad výpočtu
- 7.4.3 odvození vzorců
- 7.4.4 σ je neznámá - ttest (obvyklý případ)
- 7.4.5 příklad

7.4.1 Intervalový odhad střední hodnoty σ je předem známá

- Náhodná veličina X je z normálního rozdělení s neznámou střední hodnotou μ a předem definovanou směrodatnou odchylkou σ. Vybereme vzorek z populace o rozsahu n, který má průměr $\bar{x}$.
- Potom intervalový odhad s pravděpodobností $1-\alpha$ se vypočte:
  - Oboustranný $\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle$
  - Jednostranný odhad left $(-\infty, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} >$
  - Jednostranný odhad right $< \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}, \infty)$
  - $z_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálního rozdělení.

7.4.1 Intervalový odhad střední hodnoty σ je předem známá

- Funkce v matlabu: ztest
  - Funkce ztest obsahuje více parametrů

- [h,p,ci]=ztest(x,m,sigma,alpha,tail)
  - x vektor vstupních dat
  - m střední hodnota se kterou je průměr porovnáván (používá se u testování hypotéz)
  - sigma směrodatná odchylka
  - alpha hladina významnosti testu, (1-alpha) představuje spolehlivost intervalového odhadu
  - Tail typ intervalového odhadu $P(1-\alpha)$
    - 'both' oboustranný interval $\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle$
    - ‘left' jednostranný interval $(-\infty, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} >$
    - ‘right' jednostranný interval $< \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}, \infty)$

  - h výsledek hypotézy (používá se u testování hypotéz)
  - p p-value (používá se u testování hypotéz)
  - ci konfidenční interval

7.4.1 Intervalový odhad střední hodnoty σ je předem známá

- [h,p,ci]=ztest(x,m,sigma,alpha,tail)
- Příklad: Máte naměřená data (vektor x), zjistěte 95% oboustranný intervalový odhad střední hodnoty, jestliže víte, že směrodatná odchylka je rovna 1.
- Řešení
  - >> x=[8.5,8.8,9.1,9.2,9.4,9.5,9.7,9.9,10.2];
  - >> [h,p,ci]=ztest(x,10,1,0.05,"both")

  - h = 0 používá se u testování hypotéz
  - p = 0.0574 používá se u testování hypotéz
  - ci = 8.7133 10.0200
  - 95% oboustranný intervalový interval je $< 8.71{,}10.02 >$

- Střední hodnota vektoru je 9.37.
- Inženýrsky lze výsledek interpretovat: S pravděpodobností 95 % bude skutečná střední hodnota v intervalu $< 8.71{,}10.02 >$.
- Všimněte si, že intervalový odhad střední hodnoty je symetrický kolem z dat zjištěné střední hodnoty.

7.4.2 Intervalový odhad střední hodnoty příklad výpočtu

- Vygenerujeme 10 dat z normálního rozdělení se střední hodnotou 5 a definovanou směrodatnou odchylkou 2.
  - x=normrnd(5,2,1,10)
  - Vstupní vektor: x=[6.6808, 3.2239, 5.2002, 3.9109, 5.6070, 3.7993, 5.9799, 6.4787, 8.4238, 4.6118]
  - Střední hodnota vektoru je 5.392.
- Chceme určit 95% intervalový odhad, spolehlivost 0.95 (proto se zadává 0.05)
  - [h,p,ci]=ztest(x,5,2,0.05,'both')
  - Konfidenční meze: ci = 4.1520 6.6312
- Změníme spolehlivost na 0.99
  - [h,p,ci]=ztest(x,5,2,0.01,'both')
  - Konfidenční meze: ci = 3.7625 7.0207

7.4.3 Intervalový odhad střední hodnoty odvození vzorců

- 1) Pro dostatečně velký rozsah lze rozdělení průměru aproximovat normálním rozdělením s parametry:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- 2) Náhodná veličina $\mathrm{X} \rightarrow N(\mu, \sigma^2)$ lze přetransformovat na náhodnou veličinu $\mathrm{Z} \rightarrow N(0{,}1)$ pomocí transformace

$$\mathrm{Z} = \frac{X - \mu}{\sigma}$$

- 3) Převedeme rovnici na normované normální rozdělení.

$$\mathrm{Z} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

- 4) Oboustranný intervalový odhad je tvořen kvantily $\frac{\alpha}{2}$ a $1 - \frac{\alpha}{2}$. Tím se dosáhne spolehlivost α. Parametr α lze zvolit v intervalu $(0{,}1)$. Pro kvantil $\frac{\alpha}{2}$ určíme z inverzní funkce k distribuční funkci hodnotu (příkaz norminv) hodnotu $z_{\frac{\alpha}{2}}$ a $z_{1-\frac{\alpha}{2}}$.
- 5) Rovnici v bodě 3 upravíme a Z nahradíme buď $z_{\frac{\alpha}{2}}$ nebo $z_{1-\frac{\alpha}{2}}$.

$$\mu_1 = \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} + \bar{X} \qquad \mu_2 = \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} + \bar{X}$$

7.4.3 Intervalový odhad střední hodnoty odvození vzorců

- 5) pokr. $\mu_1 = \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} + \bar{X}$ $\mu_2 = \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} + \bar{X}$
- 6) $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ $\mu_1 = \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} + \bar{X}$ $\mu_2 = -\frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} + \bar{X}$
- 7) Konfidenční meze jsou :

$$\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle$$

7.4.4 Intervalový odhad střední hodnoty σ je neznámá

- Náhodná veličina X je z normálního rozdělení s neznámou střední hodnotou μ a nedefinovanou směrodatnou odchylkou σ. Vybereme vzorek z populace o rozsahu n, který má průměr $\bar{x}$ a směrodatnou odchylku s.
- Potom intervalový odhad se vypočte:
  - Oboustranný $\left\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \right\rangle$
  - Jednostranný left $(-\infty, \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha} >$
  - Jednostranný right $< \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha}, \infty)$
  - $t_{1-\frac{\alpha}{2}}$ je kvantil Studentova rozdělení s $n-1$ stupni volnosti.

7.4.4 Intervalový odhad střední hodnoty σ je neznámá

- Funkce v matlabu: ttest
  - Funkce ttest obsahuje více parametrů

- [h,p,ci]=ttest(x,m,alpha,tail)
  - x vektor vstupních dat
  - m střední hodnota se kterou je průměr porovnáván (používá se u testování hypotéz)
  - alpha hladina významnosti testu, (1-alpha) představuje spolehlivost intervalového odhadu
  - Tail typ intervalového odhadu
    - 'both' oboustranný interval
    - 'left' jednostranný interval 95% interval je $(-\infty, T_H)$
    - 'right' jednostranný interval 95% interval je (T_D, ∞)

  - h výsledek hypotézy (používá se u testování hypotéz)
  - p p-value (používá se u testování hypotéz)
  - ci konfidenční interval
- Porovnej funkce ttest a ztest; rozdíl je pouze v přítomnosti směrodatné odchylky

7.4.5 Příklad

- Př. : Deset balíčků mouky pocházející z balícího stroje mělo hmotnost v gramech: 987, 1001, 993, 994, 993, 1005, 1007, 999, 995 a 1002. Sestrojte 95% interval spolehlivosti pro zjištění maximální hmotnosti balíčku mouky.

- 1) směrodatná odchylka není definována, použiji ttest
- 2) načtu data a vypočtu střední hodnotu
  - >> x=[987, 1001, 993, 994, 993, 1005, 1007, 999, 995, 1002];
  - >> mean(x)
  - ans = 997.6000
- 3) vypočtu interval spolehlivosti
  - >> [h,p,ci]=ttest(x,1000,0.05,'left')
  - h = 0 (bude vysvětleno v kapitole 8)
  - p = 0.1274 (bude vysvětleno v kapitole 8)
  - ci = -∞ 1001.2
- 4) Interval spolehlivosti je $\langle -\infty; 1001.2 \rangle$

- S pravděpodobností 95 % skutečná střední hodnota hmotnosti mouky v balíčku je nižší než 1001.2 g.

7.5 Intervalový odhad rozptylu normálního rozdělení

- Náhodná veličina X je z normálního rozdělení s neznámou střední hodnotou μ a neznámým rozptylem σ^2. Vybereme vzorek z populace o rozsahu n, který má výběrový rozptyl s^2.
- Potom intervalový odhad rozptylu se vypočte:
  - Oboustranný $\left\langle \frac{(n-1)\cdot s^2}{\chi^2{}_{1-\frac{\alpha}{2}}}, \frac{(n-1)\cdot s^2}{\chi^2{}_{\frac{\alpha}{2}}} \right\rangle$
  - Jednostranný left $(0, \frac{(n-1)\cdot s^2}{\chi^2{}_{\alpha}} >$
  - Jednostranný right $< \frac{(n-1)\cdot s^2}{\chi^2{}_{1-\alpha}}, \infty)$
  - $\chi^2{}_{1-\frac{\alpha}{2}}$ je kvantil χ^2 rozdělení s $n-1$ stupni volnosti.

7.5 Intervalový odhad rozptylu normálního rozdělení

- Funkce v matlabu: vartest
  - Funkce vartest obsahuje více parametrů

- [h,p,ci]=vartest(x,v,alpha,tail)
  - x vektor vstupních dat
  - v rozptyl se kterým je výběrový rozptyl porovnáván (používá se u testování hypotéz)
  - alpha hladina významnosti testu, (1-alpha) představuje spolehlivost intervalového odhadu
  - Tail typ intervalového odhadu
    - 'both' oboustranný intervalový odhad
    - 'left' horní intervalový odhad interval $(0, T_H)$
    - 'right' dolní intervalový odhad interval (T_D, ∞)
  - h výsledek hypotézy (používá se u testování hypotéz)
  - p p-value (používá se u testování hypotéz)
  - ci konfidenční interval

7.5 Intervalový odhad rozptylu normálního rozdělení

- Př. : Deset balíčků mouky pocházející z balícího stroje mělo hmotnost v gramech: 987, 1001, 993, 994, 993, 1005, 1007, 999, 995 a 1002. Sestrojte 95% interval spolehlivosti pro rozptyl hmotnosti.
- 1) použije se funkce vartest
- 2) načtu data a vypočtu střední hodnotu a rozptyl
  - >> x=[987, 1001, 993, 994, 993, 1005, 1007, 999, 995, 1002];
  - >> mean(x)
  - ans = 997.6000
  - >> var(x)
  - ans = 38.9333
- 3) vypočtu interval spolehlivosti
  - >> [h,p,ci]=vartest(x,100,0.05,'both')
  - h = 0
  - p = 0.1181
  - ci = 18.4200 129.7591
- 4) Interval spolehlivosti rozptylu je $\langle 18.42; 129.8 \rangle$

- Všimněte si, že interval spolehlivosti není symetrický kolem vypočteného rozptylu. Důvodem je, že rozptyl se vypočítává kvadrátem odchylek od střední hodnoty.

7.6 Intervalový odhad směrodatné odchylky normálního rozdělení

- Náhodná veličina X je z normálního rozdělení s neznámou střední hodnotou μ a neznámým rozptylem σ. Vybereme vzorek z populace o rozsahu n, který má výběrovou směrodatnou odchylku s.
- Potom intervalový odhad směrodatné odchylky se vypočte:
  - Oboustranný $\left\langle \sqrt{\frac{(n-1)\cdot s^2}{{\chi^2}_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)\cdot s^2}{{\chi^2}_{\frac{\alpha}{2}}}} \right\rangle$
  - Jednostranný left $(0, \sqrt{\frac{(n-1)\cdot s^2}{{\chi^2}_{\alpha}}} >$
  - Jednostranný right $< \sqrt{\frac{(n-1)\cdot s^2}{{\chi^2}_{1-\alpha}}}, \infty)$
  - ${\chi^2}_{1-\frac{\alpha}{2}}$ je kvantil χ^2 rozdělení s $n-1$ stupni volnosti.
- Vychází se z předpokladu, že směrodatná odchylka σ je odmocnina z rozptylu σ^2.
- V matlabu není intervalový odhad implementován, protože lze použít funkci pro výpočet intervalového odhadu rozptylu. Následným odmocněním se získá intervalový odhad směrodatné odchylky.

7.7 Intervalový odhad relativní četnosti

- Mějme výběrový soubor o rozsahu n. Nechť x prvků ze souboru má určitou vlastnost; pravděpodobnost této vlastnosti je $p = \frac{x}{n}$. Dále nechť je rozsah souboru:
  - Dostatečně velký ($n > 30$)
  - Menší než 5 % rozsahu základního souboru ($\frac{n}{N} < 0.05$)
  - Splňující podmínku: $n > \frac{9}{p(1-p)}$
- Pak lze relativní četnost π odhadnout pomocí intervalů:
  - Oboustranný $\left\langle p - \sqrt{\frac{p\cdot(1-p)}{n}}\, z_{1-\frac{\alpha}{2}}, p + \sqrt{\frac{p\cdot(1-p)}{n}}\, z_{1-\frac{\alpha}{2}} \right\rangle$
  - Jednostranný left $(0, p + \sqrt{\frac{p\cdot(1-p)}{n}}\, z_{1-\alpha} >$
  - Jednostranný right $< p - \sqrt{\frac{p\cdot(1-p)}{n}}\, z_{1-\alpha}, 1)$

7.7 Intervalový odhad relativní četnosti

- Intervalový odhad není v matlabu implementován.
  - Nutno počítat pomocí vzorců
  - Vstupem je: pravděpodobnost P , rozsah n a hladina významnosti α

  - Oboustranný Ci=[p-sqrt(p*(1-p)/n)*norminv(1-alfa/2,0,1), p+sqrt(p*(1-p)/n)*norminv(1-alfa/2,0,1)]
  - Jednostranný Ci=[0, p+sqrt(p*(1-p)/n)*norminv(1-alfa,0,1)]
  - Jednostranný Ci=[p-sqrt(p*(1-p)/n)*norminv(1-alfa,0,1), 1]

- Př. V předvolebním průzkumu se zjistilo, že 11 % lidí by volilo stranu HAF. Celkem bylo dotázáno 200 lidí. Zkuste zjistit oboustranný intervalový odhad se spolehlivostí 0.95 na povolební výsledky strany HAF.
  - n=200 p=0.11 alfa=0.95

  - >> n=200;
  - >> p=0.11;
  - >> alfa=0.05;
  - >> Ci=[p-sqrt(p*(1-p)/n)*norminv(1-alfa/2,0,1),p+sqrt(p*(1-p)/n)*norminv(1-alfa/2,0,1)]
  - Ci = 0.0666 0.1534
  - Stranu HAF bude u voleb s pravděpodobností 95 % volit mezi 6.66 až 15.34 % voličů.

7.8 Odhad rozsahu výběru

- Chceme znát rozsah výběru, jestliže intervalový odhad má mít určitou šířku.
- Velikost šířky Δ intervalového odhadu je závislá na velikosti vstupu n.
- Odhad rozsahu výběru odvodíme:
  - Odečtením mezí oboustranného intervalového odhadu a vyjádřením n.
  - Vzorec pro intervalový odhad střední hodnoty je:

$$\left\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\rangle$$

- Δ je občas definována jako polovina z maximální šířky intervalu!
- Odvození:
  - 1) odečteme oboustranný intervalový odhad

$$\Delta_{max} \geq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} - \left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right) = \frac{2 \cdot \sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

  - 2) vyjádříme n:

$$n \geq \left(\frac{2\sigma}{\Delta_{max}} \cdot z_{1-\frac{\alpha}{2}} \right)^2$$

7.8 Odhad rozsahu výběru

- 1) σ je předem definována

$$n \geq \left(\frac{2\sigma}{\Delta_{max}} \cdot z_{1-\frac{\alpha}{2}}\right)^2$$

  - z je kvantil normovaného normálního rozdělení
- 2) σ je neznámá

$$n \geq \left(\frac{2s}{\Delta_{max}} \cdot t_{1-\frac{\alpha}{2}}\right)^2$$

  - Studentovo rozdělení má $n - 1$ stupňů volnosti

- 3) intervalový odhad rozsahu výběru

$$n \geq \frac{4 \cdot p \cdot (1 - p)}{{\Delta_{max}}^2} \cdot {z_{1-\frac{\alpha}{2}}}^2$$

7.9 Intervalový odhad mediánu

- Jestliže data nepochází z normálního rozdělení, nelze stanovit intervalový odhad střední hodnoty.
- Využívá se například intervalový odhad mediánu s využitím interkvartilového rozpětí.
- Intervalový odhad se spolehlivostí 95 % se stanoví pomocí vzorce:

$$\left\langle \widehat{x_{0.5}} - 1.57\frac{(\widehat{x_{0.75}} - \widehat{x_{0.25}})}{\sqrt{n}}; \widehat{x_{0.5}} + 1.57\frac{(\widehat{x_{0.75}} - \widehat{x_{0.25}})}{\sqrt{n}} \right\rangle$$

- V matlabu není funkce implementována.
  - (median(x)-1.57*iqr(x)/sqrt(n), median(x)+1.57*iqr(x)/sqrt(n))

7.10 Intervalový odhad parametrů nenormálních rozdělení

- Intervalový odhad parametrů lze zjistit i pro náhodnou veličinu, která není z normálního rozdělení.
- Předpokládá se znalost statistického rozdělení, potom pomocí metody maximální věrohodnosti (není ve skriptech řešena) se vypočtou intervalové odhady parametrů.

- V matlabu je implementováno pro následující rozdělení:
  - Diskrétní
    - Binomické rozdělení [par, io] = binofit (x,n,alpha)
    - Poissonovo rozdělení [par, io] = poissfit(x,alpha)
  - Spojité
    - Normální rozdělení [par, io] = normfit(x, alpha, cens, freq)
    - Log normální rozdělení [par, io] = lognfit(x, alpha, cens, freq)
    - Exponenciální rozdělení [par, io] = expfit(x, alpha, cens, freq)
    - Weibullovo rozdělení [par, io] = wblfit(x, alpha, cens, freq)
    - Gamma rozdělení [par, io] = gamfit(x, alpha, cens, freq)

7.10 Intervalový odhad parametrů nenormálních rozdělení

- Alpha – obdržíme (1-alpha)% intervalový odhad, například alpha = 0.05 znamená 95% intervalový odhad

- Funkce v matlabu pro diskrétní náhodnou veličinu:
  - [par, io]=poissfit(x,alpha)
  - x – vektor naměřených dat
  - alpha – hladina významnosti testu, (1-alpha) představuje spolehlivost intervalového odhadu

- Funkce v matlabu pro spojitou náhodnou veličinu:
  - [par, io]=expfit(x, alpha, cens, freq)

  - x – vstupní vektor
  - alpha – hladina významnosti testu, (1-alpha) představuje intervalový odhad
  - cens – zkouška ukončena poruchou 0, zkouška ukončena časem 1
  - freq – počet výskytů

- par – odhad hodnoty parametrů
- io – konfidenční interval uvedený po sloupcích

7.10 Intervalový odhad parametrů nenormálních rozdělení

- Zjišťovali jsme počet nehod na dálnici D1 v jednotlivých dnech. Obdrželi jsme následující výsledky:
- Nehod=[0,0,1,2,1,2,0,1,3,1,0,0,1,0,2,1,3,1,1,1]
- Určete parametry Poissonova rozdělení a 95% intervalový odhad parametru lambda.

- Řešení:
  - Nehod=[0,0,1,2,1,2,0,1,3,1,0,0,1,0,2,1,3,1,1,1]
  - [par,io]=poissfit(Nehod,0.05)

  - par = 1.0500
  - io = 0.6500, 1.6050

- Parametr lambda poissonova rozdělení je roven 1.05 (aritmetický průměr). 95% intervalový odhad parametru poissonova rozdělení je: <0.65, 1.605>.
  - Střední hodnota Poissonova rozdělení je λt. Střední počet nehod za den na dálnici D1 je 1.05, 95% intervalový odhad je $\langle 0.65, 1.605 \rangle$.
  - Jestliže bychom použili aproximaci na normální rozdělení (funkce ttest), potom by byl intervalový odhad počtu nehod $\langle 0.608, 1.492 \rangle$.

    [h,p,ci,stat]=ttest(Nehod)   h = 1   p = 8.4707e-05   ci = 0.6080 1.4920

  - Pravděpodobnost, že se nestane nehoda na dálnici D1 je: poisspdf(0,1.05) = 0.3499
  - 95% intervalový odhad (musí se zjistit lokální extrémy funkce pro různá λ) , že se nestane nehoda na D1 je:
    - poisspdf(0,[0.65,1.605])
    - ans= 0.5220 0.2009
    - 95% intervalový odhad pravděpodobnosti, že se nestane dopravní nehoda je <0.2009,0.5220>

7.10 Intervalový odhad parametrů spojitých rozdělení

- Příklad z kapitoly 5.2
- Máte 10 výrobků a chcete zjistit střední dobu do poruchy a její intervalový odhad. Doba do poruchy je popsána exponenciálním rozdělením. Zkouška probíhá 1000 hodin. Za 1000 hodin se porouchalo 5 výrobků v časech 100, 200, 300, 500, 800 hodin. Po poruše nebyly nahrazeny. Zjistěte parametry exponenciálního rozdělení.
- $E(X) = \frac{\sum_{i=1}^{n} t_i}{r} = \frac{100+200+300+500+800+5\cdot 1000}{5} = \frac{6900}{5} = 1380\ h$
- $\lambda = \frac{1}{E(X)} = \frac{1}{1380} = 7.2 \cdot 10^{-4} h^{-1}$
- Matlab:
  - x=[100,200,300,500,800,1000];
  - cens=[0,0,0,0,0,1]
  - freq=[1,1,1,1,1,5];
  - [phat,pci]=expfit(x,0.05,cens,freq)

  - phat = 1380
  - pci = 673.7 4250.1
- Střední doba do poruchy je 1380 h. 95 % intervalový odhad střední doby do poruchy je <674,4250> h.

7.10 Intervalový odhad parametrů spojitých rozdělení

- Máte 10 výrobků a chcete zjistit střední dobu do poruchy a její intervalový odhad. Doba do poruchy je popsána Weibullovým rozdělením. Zkouška probíhá 1000 hodin. Za 1000 hodin se porouchalo 8 výrobků v časech 100, 200, 300, 500, 800, 900, 950, 980 hodin. Po poruše nebyly nahrazeny. Zjistěte parametry Weibullova rozdělení.

- Matlab:
  - x=[100,200,300,500,800,900,950,980,1000];
  - cens=[0,0,0,0,0,0,0,0,1]
  - freq=[1,1,1,1,1,1,1,1,2];
  - [phat,pci]=wblfit(x,0.05,cens,freq)

  - phat = 836.3 1.60
  - pci = 541.5 0.86
    1291 2.96
  - Parametr a Weibullova rozdělení je 836.3 a jeho 95% intervalový odhad je $\langle 541.5{,}1291 \rangle$. Parametr b Weibullova rozdělení je 1.60 a jeho 95 % intervalový odhad je $\langle 0.86, 2.96 \rangle$.
  - Pokud by intervalový odhad parametru b neobsahoval 1, lze říci, že komponenta degraduje ($b_{min} > 1$); komponenta je v období časných poruch ($b_{max} < 1$),

- Obtížně se pro zjištění parametrů používá Weibullův papír, protože některé časy nejsou do poruchy.

7.11 Intervalový odhad poměru rozptylů dvou výběrů s normálním rozdělením

- Mějme dva výběry X_1 a X_2 z normálního rozdělení. Vybereme vzorek z populace o rozsahu n_1 a n_2, který má výběrový rozptyl ${s_1}^2$ a ${s_2}^2$.
- Potom intervalový odhad podílů rozptylů $\frac{{s_1}^2}{{s_2}^2}$ se vypočte:
  - Oboustranný $\left\langle \frac{1}{F_{1-\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2}, \frac{1}{F_{\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2} \right\rangle$
  - Jednostranný left $(-\infty, \frac{1}{F_{\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2} >$
  - Jednostranný right $< \frac{1}{F_{1-\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2}, \infty)$
  - $F_{1-\frac{\alpha}{2}}$ je kvantil Fisher Snedecorova rozdělení s $n_1 - 1$ stupni volnosti v čitateli a $n_2 - 1$ stupni volnosti ve jmenovateli.

7.11 Intervalový odhad poměru rozptylů dvou výběrů s normálním rozdělením

- Funkce v matlabu: vartest2
  - Funkce vartest2 obsahuje více parametrů

- [h,p,ci]=vartest2(x,y,alpha,tail)
  - x 1. vektor vstupních dat
  - y 2. vektor vstupních dat
  - alpha hladina významnosti testu,
    (1-alpha) představuje spolehlivost intervalového odhadu
  - Tail typ intervalového odhadu
    - 'both' oboustranný interval
    - 'left' jednostranný interval interval $(0, T_H)$
    - 'right' jednostranný interval interval (T_D, ∞)

  - h výsledek hypotézy (používá se u testování hypotéz)
  - p p-value (používá se u testování hypotéz)
  - ci konfidenční interval

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

- Budeme řešit následující základní úlohy:
  - 1) Známe rozptyly ${\sigma_1}^2$ a ${\sigma_2}^2$ (jsou předem definované)
  - 2) neznáme rozptyly obou populací, ale lze předpokládat, že jsou shodné
  - 3) neznáme rozptyly obou populací, ale lze předpokládat, že nejsou shodné.
- Výpočtu musí předcházet test shody rozptylů dvou výběrů kap. 8.4.1 (nebo 7.11)

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

1) Známe rozptyly $\boldsymbol{\sigma_1}^2$ a $\boldsymbol{\sigma_2}^2$ (jsou předem definované)

- Mějme dvě náhodné veličiny s normálním rozdělením X_1 a X_2, s neznámou střední hodnotou μ_1 a μ_2, jejichž rozptyly $\sigma_1{}^2$ a $\sigma_2{}^2$ jsou předem definované. Vybereme vzorek z populace o rozsahu n_1 a n_2, který mají průměr $\bar{x}_1$ a $\bar{x}_2$.
- Potom intervalový odhad rozdílu středních hodnot se vypočte:
  - Oboustranný

$$\left\langle (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}\, z_{1-\frac{\alpha}{2}}, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}\, z_{1-\frac{\alpha}{2}} \right\rangle$$

  - Jednostranný left $(-\infty, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}\, z_{1-\alpha} >$
  - Jednostranný right $< (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}\, z_{1-\alpha}, \infty)$

  - $z_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálního rozdělení.

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

1) Známe rozptyly ${\sigma_1}^2$ a ${\sigma_2}^2$ (jsou předem definované)

- V matlabu není případ ad1) implementován.
- V praxi velmi vzácný případ, obvykle se nestává, že neznáme střední hodnotu a rozptyl je již předem definován.
- Při výpočtu se častěji využívají vzorce uvedené v ad 2) nebo v ad 3).

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

2) Neznáme rozptyly obou populací, ale lze předpokládat, že jsou shodné

- Mějme dvě náhodné veličiny s normálním rozdělením X_1 a X_2, s neznámou střední hodnotou μ_1 a μ_2. Také rozptyly ${\sigma_1}^2$ a ${\sigma_2}^2$ jsou neznámé, ale lze předpokládat, že jsou shodné. Vybereme vzorek z populace o rozsahu n_1 a n_2, které mají průměr $\bar{x}_1$ a $\bar{x}_2$, a výběrové směrodatné odchylky s_1 a s_2.
- Potom intervalový odhad rozdílu středních hodnot se vypočte:
  - Oboustranný
    $$\left\langle (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{(n_1-1)\cdot {s_1}^2 + (n_2-1)\cdot {s_2}^2}{n_1+n_2-2}}\, t_{1-\frac{\alpha}{2}}, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{(n_1-1)\cdot {s_1}^2 + (n_2-1)\cdot {s_2}^2}{n_1+n_2-2}}\, t_{1-\frac{\alpha}{2}} \right\rangle$$
  - Jednostranný left $(-\infty, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{(n_1-1)\cdot {s_1}^2 + (n_2-1)\cdot {s_2}^2}{n_1+n_2-2}}\, t_{1-\alpha} >$
  - Jednostranný right $< (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{(n_1-1)\cdot {s_1}^2 + (n_2-1)\cdot {s_2}^2}{n_1+n_2-2}}\, t_{1-\alpha}, \infty)$

  - $t_{1-\frac{\alpha}{2}}$ je kvantil Studentova rozdělení s $n_1 + n_2 - 2$ stupni volnosti.
  - Shodnost rozptylu dvou výběrů lze ověřit statistickým testem, který obvykle předchází výpočtu intervalového odhadu.

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

3) Neznáme rozptyly obou populací a lze předpokládat, že nejsou shodné

- Mějme dvě náhodné veličiny s normálním rozdělením X_1 a X_2, s neznámou střední hodnotou μ_1 a μ_2. Také rozptyly ${\sigma_1}^2$ a ${\sigma_2}^2$ jsou neznámé, ale lze předpokládat, že nejsou shodné. Vybereme vzorek z populace o rozsahu n_1 a n_2, který mají průměr $\bar{x}_1$ a $\bar{x}_2$, a výběrové směrodatné odchylky s_1 a s_2.
- Potom intervalový odhad rozdílu středních hodnot se vypočte:
  - Oboustranný $\left\langle (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\, t_{1-\frac{\alpha}{2}}, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\, t_{1-\frac{\alpha}{2}} \right\rangle$
  - Jednostranný left $(-\infty, (\bar{x}_1 - \bar{x}_2) + \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\, t_{1-\alpha} >$
  - Jednostranný right $< (\bar{x}_1 - \bar{x}_2) - \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\, t_{1-\alpha}, \infty)$
  - $t_{1-\frac{\alpha}{2}}$ je kvantil Studentova rozdělení s $\frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\left(\frac{{s_1}^2}{n_1}\right)^2 \cdot \frac{1}{n_1+1} + \left(\frac{{s_2}^2}{n_2}\right)^2 \cdot \frac{1}{n_2+1}} - 2$ stupni volnosti.
  - (Ne)shodnost rozptylu dvou výběrů lze ověřit statistickým testem, který obvykle předchází výpočtu intervalového odhadu.

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

- Funkce v matlabu: ttest2
  - Funkce ttest2 obsahuje více parametrů

- [h,p,ci]=ttest2(x,y,alpha,tail,vartype)
  - x 1. vektor vstupních dat
  - y 2. vektor vstupních dat
  - alpha hladina významnosti testu,
    (1-alpha) představuje spolehlivost intervalového odhadu
  - tail typ intervalového odhadu
    - 'both' oboustranný interval
    - 'left' jednostranný interval interval $(-\infty, T_H)$
    - 'right' jednostranný interval interval (T_D, ∞)
  - Vartype shodnost/neshodnost rozptylu
    - 'equal' rozptyly v obou výběrech jsou shodné
    - 'unequal' rozptyly nejsou shodné

  - h výsledek hypotézy (používá se u testování hypotéz)
  - p p-value (používá se u testování hypotéz)
  - ci konfidenční interval

7.12 Intervalový odhad rozdílu středních hodnot dvou výběrů s normálním rozdělením

- U výrobků A a B se zjišťovala jejich životnost. Zjistěte pomocí intervalového odhadu, zda životnost výrobku B je shodná jako u výrobku A. Parametr $\alpha = 0.05$ Byly zjištěny následující hodnoty: A=[24,26,27,28,29,31], B=[25,27,29,29,30].
- Uvažujte, že rozptyly jsou shodné.
  - >> A=[24,26,27,28,29,31];
  - >> B=[25,27,29,29,30];
  - >> [h,p,ci]=ttest2(A,B,0.05,'both','equal')
  - h = 0
  - p = 0.7219
  - ci = -3.5799 2.5799
- Protože konfidenční mez rozdílu obsahuje zároveň 0, lze říci, že životnost výrobku B je shodná s životností výrobku A.

- Výsledky v případě neshodnosti rozptylů:
  - [h,p,ci]=ttest2(A,B,0.05,'both','unequal')
  - ci = -3.5210 2.5210

7.13 Intervalový odhad pro rozdíl relativních četností dvou populací

- Mějme dvě populace, z nichž byly provedeny dva nezávislé náhodné výběry o rozsahu n_1 a n_2, kde počet prvků s určitou vlastností je x_1 a x_2 . Potom pravděpodobnost výskytu určité vlastnosti je $p_1 = \frac{x_1}{n_1}$ a $p_2 = \frac{x_2}{n_2}$.

- Nechť dále platí o rozsahu výběru, že:
  - Výběry jsou dostatečně velké ($n_1 > 30, n_2 > 30$)
  - Menší než 5 % rozsahu základního souboru ($\frac{n_1}{N_1} < 0.05, \frac{n_2}{N_2} < 0.05$)
  - Splňující podmínky: $n_1 > \frac{9}{p_1(1-p_1)}$ a $n_2 > \frac{9}{p_2(1-p_2)}$

- Nechť dále platí: $p = \frac{x_1+x_2}{n_1+n_2}$

7.13 Intervalový odhad pro rozdíl relativních četností dvou populací

- Nechť dále platí: $p = \frac{x_1+x_2}{n_1+n_2}$
- Potom intervalový odhad pro rozdíl relativních četností dvou populací se vypočte:
  - Oboustranný $\left\langle (p_1 - p_2) - \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\, z_{1-\frac{\alpha}{2}}, (p_1 - p_2) + \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\, z_{1-\frac{\alpha}{2}} \right\rangle$
  - Jednostranný $(-\infty, (p_1 - p_2) + \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\, z_{1-\alpha} >$
  - Jednostranný $< (p_1 - p_2) - \sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\, z_{1-\alpha}, \infty)$

- Příklad: 1. výběr je tvořen 100 prvky a pravděpodobnost úspěchu je 0.42; 2. výběr je tvořen 50 prvky a pravděpodobnost úspěchu je 0.26. Vypočtěte intervalový odhad pro rozdíl relativních četností $p_1 - p_2$. $\propto = 0.05$.
  - >> n1=100 >> n2=50 >> p1=0.42 >> p2=0.26
  - >> x1=n1*p1 x1 = 42 >> x2=n2*p2 x2 = 13
  - >> p=(x1+x2)/(n1+n2) p = 0.3667
  - >> Ci=[(p1-p2)-sqrt(p*(1-p)*(1/n1+1/n2))*norminv(0.975,0,1), (p1-p2)+sqrt(p*(1-p)*(1/n1+1/n2))*norminv(0.975,0,1)]
  - Ci = -0.0036 0.3236

7.14 Příkazy v matlabu /octave

- 1 výběr
  - Střední hodnota
    - Směrodatná odchylka definována ztest kap. 7.4.1
    - Směrodatná odchylka neznáma ttest kap. 7.4.4
  - Rozptyl vartest kap. 7.5
  - Směrodatná odchylka vartest kap. 7.6
  - Nenormální rozdělení expfit, …fit kap. 7.10
- 2 výběry
  - Shoda středních hodnot ttest2 kap. 7.11
  - Shoda rozptylů vartest2 kap. 7.12

8 Testy hypotéz

- 8.1 Princip testování hypotéz
- 8.2 Přístup k testování hypotéz
- 8.3 Jednovýběrové testy (a párový test)
- 8.4 Dvouvýběrové testy
- 8.5 Vícevýběrové testy

8.1 Princip testování hypotéz

- Pomocí statistického usuzování rozhodujeme na základě informací získaných z náhodných výběrů, zda přijmeme, nebo zamítneme určitou hypotézu týkající se základního souboru

- Statistickou hypotézou rozumíme jakékoliv tvrzení, které se může týkat:
  - neznámých parametrů výběru (například střední hodnoty, rozptylu, mediánu) kap. 8,
  - typu rozdělení (normální, exponenciální, shoda typu rozdělení u 2 výběrů, …) kap. 9
  - nezávislosti dat a dalších vlastností základního souboru (základních souborů). kap. 10
  - Proložení spojitých dat funkcí kap. 11

- Parametrická hypotéza – statistická hypotéza pojednává o parametrech rozdělení náhodné veličiny (střední hodnota, rozptyl,…)
  - Rovnost středních hodnot dvou výběrů
  - Testování rozptylu náhodného výběru

- Neparametrická hypotéza – statistická hypotéza nepojednává o parametrech rozdělení náhodné veličiny (např. typ rozdělení, nezávislost výběrů,…)
  - Pochází data z normálního rozdělení?

8.1 Princip testování hypotéz

- Rozdělení hypotéz podle počtu výběrů
  - Jednovýběrové
    - Střední hodnota náhodné veličiny je rovna 8.
    - Rozptyl náhodné veličiny je menší než 16.
  - Dvouvýběrové
    - Střední hodnota prvního výběru je větší než druhého.
    - Podíl rozptylů prvního a druhého výběru je shodný.
  - Vícevýběrové
    - Střední hodnoty všech výběrů jsou shodné; a následné zjištění výběrů, které se střední hodnotou odlišují
    - Rozptyly všech výběrů jsou shodné

8.1 Princip testování hypotéz

- Používají se dva druhy hypotéz:
  - Nulová hypotéza H_0 představuje tvrzení, že sledovaný efekt je nulový a bývá vyjádřena rovností mezi testovaným parametrem θ a jeho očekávanou hodnotou θ_0.
  - Alternativní hypotéza H_A, která popírá tvrzení dané nulovou hypotézou.
  - Mohou být následující tvrzení hypotéz:

		matlab (H_0)
$H_0: \theta = \theta_0$	$H_A: \theta \neq \theta_0$	„both“
$H_0: \theta \geq \theta_0$	$H_A: \theta < \theta_0$	„left“
$H_0: \theta \leq \theta_0$	$H_A: \theta > \theta_0$	„right“

Nulová hypotéza musí vždy obsahovat rovnost

8.1 Princip testování hypotéz

Příklady na nulovou a alternativní hypotézu:

- Ověřte, zda průměrná hmotnost výrobku je 1 kg?
  - Nulová hypotéza: $H_0: \mu = 1kg$
  - Alternativní hypotéza: $H_A: \mu \neq 1\ kg$
- Zvýšila se úpravou technologie životnost výrobku?
  - Nulová hypotéza: $H_0: \mu_{new} \leq \mu_{old}$
  - Alternativní hypotéza: $H_A: \mu_{new} > \mu_{old}$
  - Všimněte si, že nulová hypotéza vždy obsahuje rovnost

- Nulovou hypotézu považujeme za pravdivou, až do okamžiku, kdy nás výsledky potvrdí o opaku. Výsledkem je:
  - Zamítáme hypotézu H_0 ve prospěch hypotézy H_A
  - Nezamítáme H_0
  - Pozn. Pro rozlišení, zda (ne)zamítneme používáme testovací statistiky, založené na stejném principu jako u intervalů spolehlivosti.
  - Pozn. Všimněte si, že pouze zamítáme / nezamítáme H_0. Nelze říci, že H_0 přijímáme, protože rozšiřujícím se počtem dat se může stát, že bude zamítnuta.

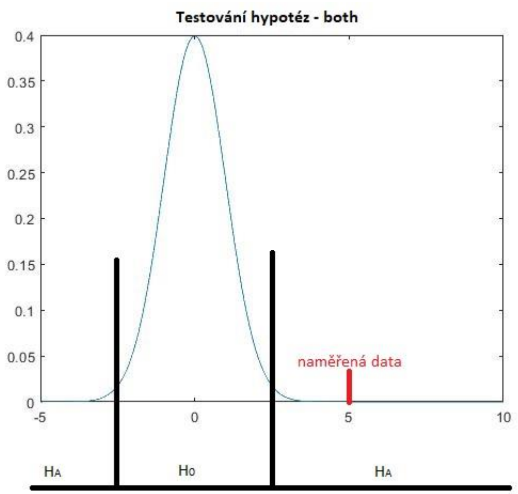
8.1 Princip testování hypotéz

- Příklad: U 10 lidí chtějící přibrat bylo zjištěno, že jejich roční váhový přírůstek je [1,3,4,4,5,5,6,7,7,8] kg. Průměrné zvýšení hmotnosti je 5 kg Otestujte, zda:

hmotnost je stejná

$H_0 : \theta = 0 \qquad H_A : \theta \neq 0$

„both“

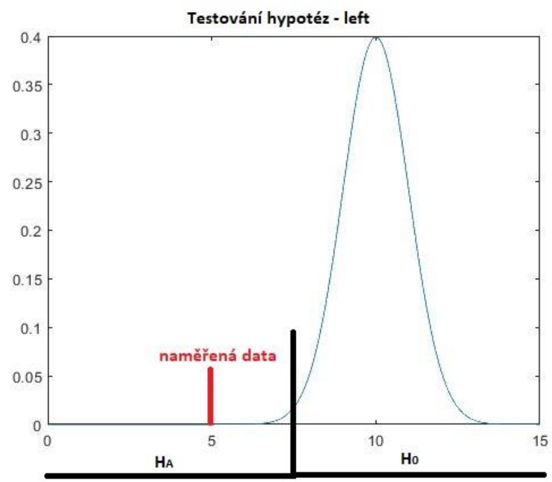


H0 – hmotnost je stejná
HA – hmotnost je nižší
HA – hmotnost je vyšší

došlo ke zvýšení hmotnosti max. o 10 kg

$H_0 : \theta \geq 10 \qquad H_A : \theta < 10$

„left“

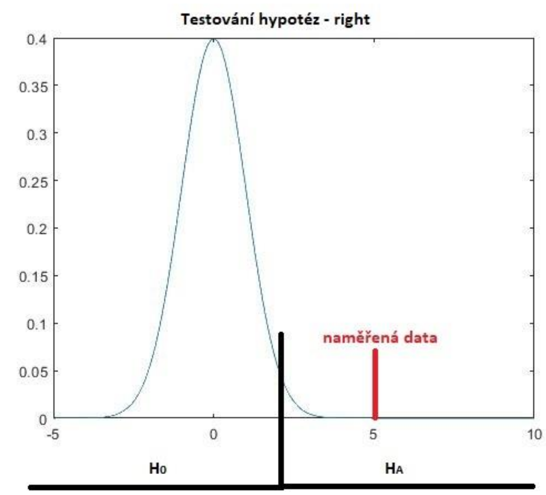


H0 – hmotnost je vyšší právě o 10 kg
H0 – hmotnost je vyšší než o 10 kg
HA – hmotnost je vyšší max. o 10 kg.

došlo ke zvýšení hmotnosti

$H_0 : \theta \leq 0 \qquad H_A : \theta > 0$

„right“



H0 – hmotnost je stejná
H0 – hmotnost je nižší
HA – hmotnost je vyšší

8.1 Princip testování hypotéz

- Chyba I. a II. druhu

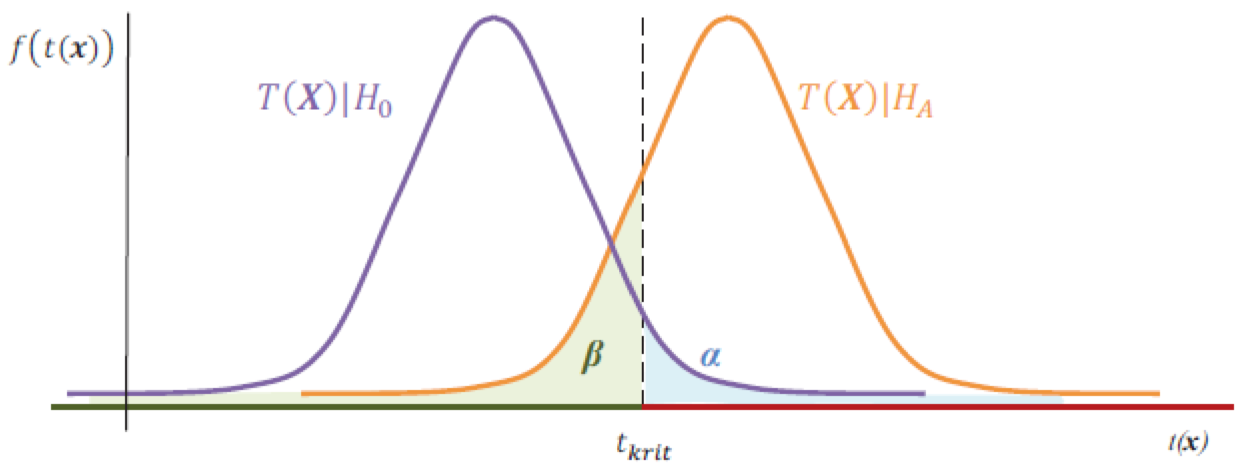
		Výsledek testu	
		Nezamítáme H_0	**Zamítáme H_0**
Skutečnost	**Platí H_0**	Správné rozhodnutí 1-α (spolehlivost testu)	Chyba I. druhu α (hladina významnosti)
	Platí H_A	Chyba II. druhu β	Správné rozhodnutí 1-β (síla testu)

  - Nulová hypotéza je platná a my ji zamítneme, dopouštíme se chyby I. druhu. Tuto pravděpodobnost nazýváme hladinu významnosti $\propto$.
  - Nulová hypotéza je platná a není zamítnuta. Tato pravděpodobnost je $1 - \propto$. Nazýváme ji spolehlivost testu.
  - Nulová hypotéza není platná a je zamítnuta. Vzniká s pravděpodobností $1 - \beta$ a označujeme ji sílou testu.
  - Nulová hypotéza není platná, ale je přijata. Dopouštíme se chyby II. druhu. Vzniká s pravděpodobností β.

8.1 Princip testování hypotéz

- Zmenšováním chyby $\propto$ se zvyšuje chyba β.
- Snaha o minimalizaci chyb.
- Obvykle volíme $\propto = 0.05$. Vzácněji i $\propto = 0.01$.

		Výsledek testu	
		Nezamítáme H_0	**Zamítáme H_0**
Skutečnost	**Platí H_0**	Správné rozhodnutí 1-α (spolehlivost testu)	Chyba I. druhu α (hladina významnosti)
	Platí H_A	Chyba II. druhu β	Správné rozhodnutí 1-β (síla testu)



8.2 Přístup k testování hypotéz

- Postup klasického testu hypotéz
  1. Formulace nulové a alternativní hypotézy
  2. Volba druhu testové statistiky. Na základě výsledku se rozhodne o zamítnutí/nezamítnutí nulové hypotézy – viz bod 5.
  3. Stanovení hladiny významnosti testu $\propto$.
  4. Zjištění hranice kdy zamítáme /nezamítáme H_0 (na základě vypočtené testové statistiky)
  5. Výpočet testové statistiky $T(X)$
  6. Formulace závěru
- Používáno historicky, když se úloha řešila pomocí papíru, kalkulačky a tabulek

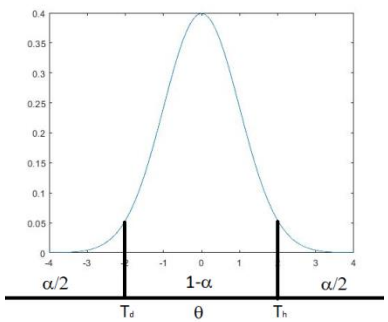
8.2 Přístup k testování hypotéz

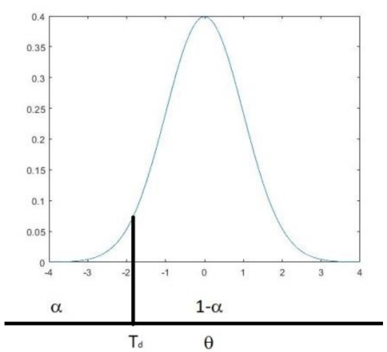
- Postup čistého testu významnosti
  1. Formulace nulové a alternativní hypotézy
  2. Volba druhu testové statistiky.
  3. Výpočet testové statistiky $T(X)$
  4. Výpočet p-value
     - Na základě p-value nezamítáme/zamítáme hypotézu H_0
     - Čím nižší vyjde p-value, tím více jsme přesvědčeni, že nulová hypotéza není správná a je třeba jí zamítnout.
     - $p_{value} < \propto \; H_0$ zamítáme
     - $p_{value} > \propto \; H_0$ nezamítáme
  5. Rozhodnutí na základě p-value
  6. Formulace závěru

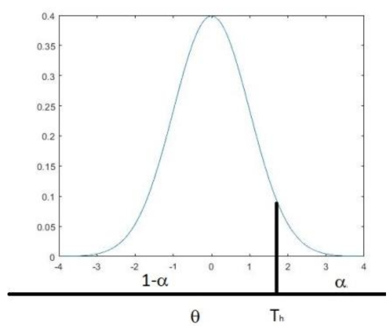
8.2 Přístup k testování hypotéz

- Co je to p-value a jak ji vypočítat?
  - Klasickým přístupem se zamítne/nezamítne hypotéza H_0. Čistý test významnosti poskytne p-value, která poskytuje obecnější informaci o výsledku – lze zjistit jak moc je výsledek testu významný.
  - Čím nižší vyjde p-value, tím více jsme přesvědčeni, že nulová hypotéza není správná a je třeba jí zamítnout.
  - Mějme výsledek testovací statistiky $T(X)$. Pro výsledek $T(X)$ určíme kvantil distribuční funkce (např. funkce normcdf, chi2cdf,tcdf apod).
    - $H_0\colon \theta = \theta_0, H_A\colon \theta \neq \theta_0$ $\quad p_{value} = 2 \cdot \min\big(F_0(T(X)), 1 - F_0(T(X))\big)$ 'both'
    - $H_0\colon \theta \geq \theta_0\,, H_A\colon \theta < \theta_0$ $\quad p_{value} = F_0(T(X))$ 'left'
    - $H_0\colon \theta \leq \theta_0\,, H_A\colon \theta > \theta_0$ $\quad p_{value} = 1 - F_0(T(X))$ 'right'

Matlab: both left right







8.2 Přístup k testování hypotéz

- Stanovení intervalů spolehlivosti a testování hypotéz probíhá obdobným způsobem.
- Například intervalový odhad střední hodnoty a testování zda střední hodnota naměřených dat je m.

  [h,p,ci,stat]=ttest(x,m,alpha,tail)

  - Pravá strana:
    - x vstupní data
    - m střední hodnota se kterou porovnáváme vstupní data
    - alpha hladina významnosti (u intervalů spolehlivosti – spolehlivost)
    - tail jedno nebo oboustranný interval spolehlivosti
- Levá strana
    - h výsledek hypotézy
    - p velikost p-value
    - ci konfidenční interval
    - stat výsledky testové statistiky T, počet stupňů volnosti

8.2 Přístup k testování hypotéz

- Příklad: Mějme naměřená data - životnost stroje. Výrobce tvrdí, že životnost výrobku je vyšší než 800 h. Mějme naměřená data: x=[700,750,780,820,860,900,940,950,980,1020]. Jejich průměrná hodnota je 870.
- 1) Formulace nulové a alternativní hypotézy
  - H0: $\mu \leq 800$ h
  - H1: $\mu > 800$ h
  - Podle H0 zjistíme, že budeme v matlabu zadávat „right“.
- 2) Volba druhu testové statistiky. V příslušné kapitole skript zjistíme, že úlohu řeší funkce „ttest“
- 3) Výpočet testové statistiky $T(X)$ a 4) výpočet p-value
  - Příkaz v matlabu: [h,p,ci,stat]=ttest(x,800,0.05,’right')
    - Příkaz stejný jako u intervalových odhadů
    - Hodnota 800, jako 2. parametr funkce odkazuje na hodnotu v nulové a alternativní hypotéze
- 5) Rozhodnutí na základě p-value
  - h = 1 — příkláníme se k alternativní hypotéze
    - 0 platí nulová hypotéza, 1 přikláníme se k alternativní hypotéze
  - p = 0.0330 — pvalue je menší než 0.05, zamítáme H0
  - ci = 808.6523 Inf — 95% interval spolehlivosti
  - stat =
    - tstat: -2.0917 — výsledek testové statistiky T — výpočet dle vzorce
    - df: 9 — počet stupňů volnosti — obvykle počet dat – počet odhadovaných parametrů
    - sd: 105.8301 — směrodatná odchylka vstupů

- Na hladině významnosti 5 % zamítáme hypotézu H0, že životnost stroje je menší nebo rovna 800 h.
- Laicky: Na hladině významnosti 5 % jsme prokázali, že životnost stroje je vyšší než 800 h

8.3 Jednovýběrové testy hypotéz

- 8.3.1 Test rozptylu normálního rozdělení
- 8.3.2 Test střední hodnoty normálního rozdělení
- 8.3.3 Párový test
- 8.3.4 Znaménkový test
- 8.3.5 Kvantilový test
- 8.3.6 Wilcoxonův test
- 8.3.7 Test o parametru π relativní četnosti
- 8.3.8 Testy hodnoty parametrů nenormálních rozdělení

8.3.1 Test rozptylu normálního rozdělení

- Mějme data z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, kde oba parametry neznáme. Na základě výběru $X_1, X_2, \dots, X_n$ chceme ověřit předpoklad, že rozptyl populace σ^2 se rovná hodnotě výběrového rozptylu z naměřených dat s^2.
- Potom testovací kritérium je následující:

$$T(X) = \frac{s^2}{\sigma^2} \cdot (n-1)$$

- Testovací kritérium má χ^2 rozdělení s $(n-1)$ stupni volnosti.

8.3.1 Test rozptylu normálního rozdělení

- Seznam hypotéz:
  - $H_0: s^2 = \sigma^2$ $\quad$ $H_A: s^2 \neq \sigma^2$ $\quad$ 'both'
    - H_0 nezamítám, když testovací statistika je:
      $\chi^2_{\frac{\alpha}{2}}(n-1\ st.v.) \leq T(X) \leq \chi^2_{1-\frac{\alpha}{2}}(n-1\ st.v.)$
  - $H_0: s^2 \geq \sigma^2$ $\quad$ $H_A: s^2 < \sigma^2$ $\quad$ 'left'
    - H_0 nezamítám, když testovací statistika je:
      $$T(X) \geq \chi^2_{\alpha}(n-1\ st.v.)$$
    - Např. $H_0: s^2 \geq 800$ $\quad$ $H_A: s^2 < 800$
  - $H_0 : s^2 \leq \sigma^2$ $\quad$ $H_A: s^2 > \sigma^2$ $\quad$ 'right'
    - H_0 nezamítám, když testovací statistika je:
      $$T(X) \leq \chi^2_{1-\alpha}(n-1\ st.v.)$$
    - Např. $H_0 : s^2 \leq 800$ $\quad$ $H_A: s^2 > 800$
- H_0 zamítám, když je p-value menší než $\propto$.

8.3.1 Test rozptylu normálního rozdělení

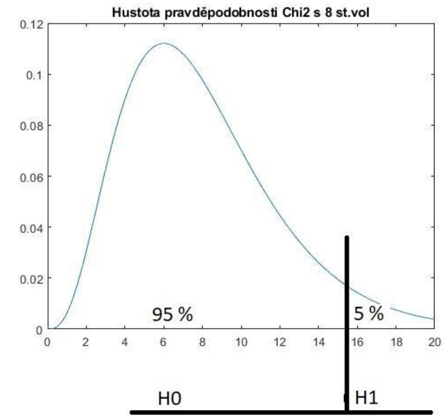
- Funkce v matlabu: vartest
  - Funkce vartest obsahuje více parametrů
  - Shodný test jako u intervalového odhadu

- [h,p,ci,stats]=vartest(x,v,alpha,tail)
  - x vektor vstupních dat
  - v rozptyl se kterým je výběrový rozptyl porovnáván

  - alpha hladina významnosti
  - tail typ intervalového odhadu
    - 'both' H_0: $s^2 = \sigma^2$ H_A: $s^2 \neq \sigma^2$
    - 'left' H_0: $s^2 \geq \sigma^2$ H_A: $s^2 < \sigma^2$
    - 'right' H_0: $s^2 \leq \sigma^2$ H_A: $s^2 > \sigma^2$
  - h výsledek hypotézy
  - p p-value
  - ci konfidenční interval
  - stats statistické výsledky
    - chisqstat výsledek testovací statistiky
    - df počet stupňů volnosti
- Všechny parametry se nemusí uvádět. Pokud je test na hladině významnosti 5 % a je oboustranný, tak stačí [h,p]=vartest(x,v)

8.3.1 Test rozptylu normálního rozdělení

- Stroj balí písek do 50 kg pytlů. Maximální povolený rozptyl je 0.0121 kg^2. Otestujte na hladině významnosti 5 %, zda rozptyl je menší nebo roven 0.0121 kg^2
- Data: 49.87,49.74,49.91,49.85,49.71,49.63,49.75,49.82,50.11
- Rozptyl dat je: var(x)=0.0194



- $H_0 : s^2 \leq 0.0121$ $H_A : s^2 > 0.0121$
  - H_0 nezamítám, když testovací statistika je:

$$T(X) \leq {\chi^2}_{1-\alpha}(n-1\, st.v.)$$

- Matlab
  - >> x=[49.87, 49.74, 49.91, 49.85, 49.71, 49.63, 49.75, 49.82, 50.11];
  - >> [h,p,ci,stats]=vartest(x,0.0121,0.05,'right')
  - h = 0
  - p = 0.1183
  - ci = 0.0100 Inf
  - stats : chisqstat: 12.8173 df: 8
  - Na hladině významnosti 5 % nezamítáme hypotézu H_0, že rozptyl je menší nebo roven 0.0121 kg^2 (pval= 0.1183, chistat=12.8173, 8 st. vol).

- 1) Formulace nulové a alternativní hypotézy
- 2) Volba druhu testové statistiky.
- 3) Výpočet testové statistiky $T(X)$
- 4) výpočet p-value
- 5) Rozhodnutí na základě p-value

8.3.1 Test rozptylu normálního rozdělení

- Stroj balí písek do 50 kg pytlů. Maximální povolený rozptyl je 0.0121 kg^2. Otestujte na hladině významnosti 5 %, zda rozptyl je menší nebo roven 0.0121 kg^2
- Data: 49.87,49.74,49.91,49.85,49.71,49.63,49.75,49.82,50.11
- Rozptyl dat je: var(x)=0.0194

- $H_0 : s^2 \leq 0.0121 \quad H_A : s^2 > 0.0121$
  - H_0 nezamítám, když testovací statistika je: $T(X) \leq \chi^2{}_{1-\alpha}(n-1\ st.v.)$

- Tužka, papír (historicky, nebo když nemáte vstupní data, ale pouze vypočtenou střední hodnotu a rozptyl)
  - Vypočtu rozptyl z dat var(x)=0.0194
  - Dosadím do vzorce testové statistiky $T(X) = \frac{s^2}{\sigma^2} \cdot (n-1) = \frac{0.0194}{0.0121} \cdot 8 = 12.8173$
  - Výsledek testové statistiky je 12.8173
  - Zjistím kvantily $\chi^2{}_{0.95}(8) = chi2inv(0.95{,}8) = 15.5073$ (hraniční bod zamítnutí H0)
  - Zjistím pvalue pom=chi2cdf(12.8173,8)=0.8817
  - Pvalue
    - Oboustranný interval $pval = 2 \cdot \min(pom, 1 - pom)$
    - Jednostranný left $pval = pom$
    - Jednostranný right $pval = 1 - pom$
    - $pval = 1 - pom = 1 - 0.8817 = 0.1283$
  - Závěr: Na hladině významnosti 5 % nezamítáme hypotézu H0, že rozptyl je menší nebo roven 0.0121.

8.3.2 Test střední hodnoty normálního rozdělení

- Mějme populaci z normálního rozdělení s neznámou střední hodnotou μ. Na základě výběru $X_1, X_2, \dots, X_n$ chceme ověřit předpoklad, že střední hodnota populace μ se rovná naměřeným hodnotám $\bar{X}$.
- Potom testovací kritérium je následující:

  1) Rozptyl je předem definovaný

  $$T(X) = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$$

  , kde testovací kritérium má normované normální rozdělení

  2) Rozptyl není definovaný

  $$T(X) = \frac{\bar{X} - \mu}{s} \cdot \sqrt{n}$$

  , kde testovací kritérium má Studentovo t rozdělení s $n - 1$ stupni volnosti.
- Případ ad1 je velmi vzácný.

8.3.2 Test střední hodnoty normálního rozdělení

- Seznam hypotéz:
  - $H_0: \bar{X} = \mu$ $\quad$ $H_A: \bar{X} \neq \mu$ $\quad$ 'both'
    - H_0 nezamítám, když testovací statistika je:
      $t_{\frac{\alpha}{2}}(n-1\ st.v.) \leq T(X) \leq t_{1-\frac{\alpha}{2}}(n-1\ st.v.)$
  - $H_0 : \bar{X} \geq \mu$ $\quad$ $H_A: \bar{X} < \mu$ $\quad$ 'left'
    - H_0 nezamítám, když testovací statistika je:
      $T(X) \geq t_{\alpha}(n-1\ st.v.)$
  - $H_0: \bar{X} \leq \mu$ $\quad$ $H_A: \bar{X} > \mu$ $\quad$ 'right'
    - H_0 nezamítám, když testovací statistika je:
      $T(X) \leq t_{1-\alpha}(n-1\ st.v.)$
- H_0 zamítám, když je p-value menší než $\propto$.

8.3.2 Test střední hodnoty normálního rozdělení

- Funkce v matlabu:
  - Rozptyl předem definovaný ztest Rozptyl neznámý ttest
  - Funkce ztest a ttest obsahují více parametrů

- [h,p,ci,stats]=ztest(x,m,sigma,alpha,tail)
- [h,p,ci,stats]=ttest(x,m,alpha,tail)
  - x vektor vstupních dat
  - m střední hodnota se kterou je průměr porovnáván
  - sigma rozptyl dat (používá se pouze u funkce ztest)
  - alpha hladina významnosti
  - tail typ intervalového odhadu
    - 'both' $H_0: \bar{X} = \mu$ $H_A: \bar{X} \neq \mu$
    - 'left' $H_0: \bar{X} \geq \mu$ $H_A: \bar{X} < \mu$
    - 'right' $H_0: \bar{X} \leq \mu$ $H_A: \bar{X} > \mu$

  - h výsledek hypotézy
  - p p-value
  - ci konfidenční interval
  - stats funkce ztest – pouze výsledek testovací statistiky

    funkce ttest:
    - tstat výsledek testovací statistiky df počet stupňů volnosti
    - sd vypočtená směrodatná odchylka
- Všechny parametry se nemusí uvádět. Často se uvádí např. [h,p]=ttest(x,m)

8.3.2 Test střední hodnoty normálního rozdělení

- Stroj balí písek do 50 kg pytlů. Maximální povolená hmotnostní odchylka je 0.05 kg. Otestujte na hladině významnosti 5 %, zda stroj na balení písku se nedopouští systematické chyby sypáním menšího/většího množství – tj. zda střední hodnota je rovna 50 kg.
- Data: 49.87,49.74,49.91,49.85,49.71,49.63,49.75,49.82,50.11
  - $H_0: \bar{X} = \mu$ $H_A: \bar{X} \neq \mu$ 'both'
    - H_0 nezamítám, když testovací statistika je:

      $$t_{\frac{\alpha}{2}}(n-1\ st.v.) \leq T(X) \leq t_{1-\frac{\alpha}{2}}(n-1\ st.v.)$$
  - Střední hodnota vektoru je 49.82. Ptáme se, zda je rozdíl již statisticky významný.
- Matlab
  - >> x=[49.87,49.74,49.91,49.85,49.71,49.63,49.75,49.82,50.11];
  - >> [h,p,ci,stats]=ttest(x,50,0.05,'both')
  - h = 1
  - p = 0.0048
  - ci = 49.7141 49.9281
  - stats = tstat: -3.8544 df: 8 sd: 0.1392
- Zamítáme na hladině významnosti 5 % hypotézu, že střední hodnota hmotnosti písku v pytli je 50 kg.
- Všimněte si, že z 9 vstupních dat je 8 menších než 50 kg. V případě, že očekáváme střední hodnotu 50 kg měl by být přibližně poloviční počet dat menších než 50 kg

8.3.3 Párový test

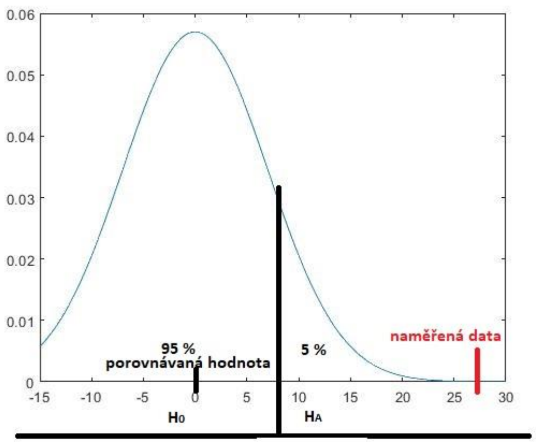
- Mějme populaci z normálního rozdělení s neznámou střední hodnotou μ_1 uskutečněnou před a střední hodnotou μ_2 uskutečněném po určité operaci. Na základě výběru $X_1, X_2, \dots, X_n$ chceme ověřit předpoklad, že střední hodnota populací μ_1 a μ_2 je shodná.
  - U každého měření jsme schopni říci, jaká hodnota byla naměřena před a jaká po dané operaci.
  - Rozdíl dvou normálních rozdělení $N(\mu_1, {\sigma_1}^2) - N(\mu_2, {\sigma_2}^2)$ má opět normální rozdělení s parametry: $N(\mu_1 - \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$.
  - Rozdílem výsledků po a před danou operací obdržíme změnu. Testujeme, zda vliv této změny je nulový či nikoliv.
- Testování probíhá pomocí funkce ttest, kde vstupem jsou rozdíly po a před měřením.
- Je možno provádět i jednostranné testy (např. došlo ke zlepšení), nebo s posunem definováním parametru m (např. došlo ke zlepšení o 5)

8.3.3 Párový test

- Př. Následující tabulka uvádí výsledky pevnosti oceli před kalením a po kalení. Určete na hladině významnosti 5 %, zda se kalením zvýšila pevnost oceli.

Před μ_1	450	480	510	490	510	500	440	490
Po μ_2	470	520	540	510	530	520	490	510
Rozdíl	20	40	30	20	20	20	50	20

  - U každého z měření známe hodnotu tvrdosti před a po kalení.
  - $H_0: \mu_2 - \mu_1 \leq 0$ $\qquad$ $H_A: \mu_2 - \mu_1 > 0$
  - >> x=[20, 40, 30, 20, 20, 20, 50, 20]
  - x = 20 40 30 20 20 20 50 20
  - >> [h,p,ci,stat]=ttest(x,0,0.05, 'right')
  - h = 1
  - p = 1.4172e-04
  - ci = 19.697 Inf
  - stat = tstat: 6.6767 df: 7 sd: 11.65



- Na hladině významnosti 5 % zamítám hypotézu H_0, že kalení nemá vliv na pevnost oceli.

8.3.3 Párový test

- Otestujte na stejném příkladě, zda se zvýšila pevnost oceli minimálně o 20.
  - $H_0: \mu_2 - \mu_1 \leq 20$ $H_A: \mu_2 - \mu_1 >$20
  - >> x=[20, 40, 30, 20, 20, 20, 50, 20]
  - >> [h,p,ci,stat]=ttest(x,20,0.05, 'right')

  - h = 0
  - p = 0.0557
  - ci = 19.6967 Inf
  - stat = tstat: 1.8209 df: 7 sd: 11.6496
- Na hladině významnosti 5 % přijímáme hypotézu H0, že pevnost oceli se zvýšila maximálně o 20.

8.3.4 Znaménkový test

- Znaménkový test umožňuje na základě výběru $X_1, X_2, \dots, X_n$ ověřit předpoklad, že se medián náhodného výběru $x_{0.5}$ rovná testované hodnotě $x_{test0.5}$.
- Sleduje se četnost Z_{-} naměřených hodnot menších než testovaná hodnota $x_{test0.5}$.
- Potom testovací kritérium je následující:
  - H_0: $x_{0.5} = x_{test0.5}$ H_A: $x_{0.5} \neq x_{test0.5}$ $T(X) = 2 \cdot \min\left(\sum_{i=0}^{Z_-} \binom{n}{i} 0.5^n, \sum_{i=0}^{Z_+} \binom{n}{i} 0.5^n\right)$
  - H_0: $x_{0.5} \geq x_{test0.5}$ H_A: $x_{0.5} < x_{test0.5}$ $T(X) = \sum_{i=0}^{Z_+} \binom{n}{i} 0.5^n$
  - H_0: $x_{0.5} \leq x_{test0.5}$ H_A: $x_{0.5} > x_{test0.5}$ $T(X) = \sum_{i=0}^{Z_-} \binom{n}{i} 0.5^n$
  - Testová statistika $T(X)$ zároveň představuje pvalue.
- Jedná se o neparametrický test, protože není nutný předpoklad o tvaru rozdělení.
- Jestliže některé z hodnot se rovnají testovanému mediánu, budou vynechány.

8.3.4 Znaménkový test

- Funkce v matlabu: signtest

- [p,h]=signtest(x,m,alfa,'tail','both')
  - x — vektor vstupních dat
  - m — testovaná hodnota mediánu
  - alfa — hladina významnosti
  - tail —
    H$_0$: $x_{0.5} = x_{test0.5}$ H$_A$: $x_{0.5} \neq x_{test0.5}$ 'both',
    H$_0$: $x_{0.5} \geq x_{test0.5}$ H$_A$: $x_{0.5} < x_{test0.5}$ 'left',
    H$_0$: $x_{0.5} \leq x_{test0.5}$ H$_A$: $x_{0.5} > x_{test0.5}$ 'right'
  - p — p-value
  - h — výsledek hypotézy

- Pro výpočet testovacího kritéria lze použít také funkce Binocdf
  - H$_0$: $x_{0.5} = x_{test0.5}$ H$_A$: $x_{0.5} \neq x_{test0.5}$ $\text{pvalue} = 2 \cdot binocdf(\min(Z_-, Z_+), n, 0.5)$
  - H$_0$: $x_{0.5} \geq x_{test0.5}$ H$_A$: $x_{0.5} < x_{test0.5}$ $\text{pvalue} = binocdf(Z_+, n, 0.5)$
  - H$_0$: $x_{0.5} \leq x_{test0.5}$ H$_A$: $x_{0.5} > x_{test0.5}$ $\text{pvalue} = binocdf(Z_-, n, 0.5)$

8.3.4 Znaménkový test

- Př. Mějte data: 4,5,5,6,6,7,7,8,8,8,9,10,12 a otestujte, zda medián je na hladině významnosti 0.05 roven 5.5.
  - Matlab
  - >> x=[4,5,5,6,6,7,7,8,8,8,9,10,12]
  - >> [p,h]=signtest(x,5.5,0.05)
  - p =   0.0923
  - h =   0
  - Nezamítáme hypotézu H0 na hladině významnosti 5 %, že medián je roven 5.5

- Jestliže se bude testovat medián roven 5, je nutno odstranit ze vstupu naměřené hodnoty 5 a výsledek počítat pouze ze zbývajících 11 naměřených hodnot.
  - pvalue výpočtem: $2 \cdot$Binocdf(1,11,0.5)=0.0117
  - matlab: [p,h]=signtest(x,5,0.05)
    p =   0.0117
  - Zamítáme hypotézu H0, že medián je roven 5

- Jestliže chceme otestovat H0 $x_{0.5} \leq 5$, oproti hypotéze H1 $x_{0.5} > 5$
  - Matlab: [p,h]=signtest(x,5,0.05,'tail','right')
    p =   0.00585
  - Zamítáme hypotézu H0, že medián je menší nebo roven 5.

8.3.5 Kvantilový test

- Principem vychází ze znaménkového testu.

- Kvantilový test umožňuje na základě výběru $X_1, X_2, \dots, X_n$ ověřit předpoklad, že se určitý kvantil $\propto$ náhodného výběru $x_\propto$ rovná testované hodnotě $x_{test\,\propto}$.
- Sleduje se četnost Z_- naměřených hodnot menších než testovaná hodnota $x_{test\,\propto}$.
- Potom testovací kritérium je následující:

$$T(X) = \sum_{i=0}^{Z_-} \binom{n}{i} \propto^i \cdot (1 - \propto)^{n-i}$$

- Z výsledné statistiky $T(X)$ lze zjistit hodnotu p-value pro oboustranný interval:

$$pvalue = 2 \cdot \min\{T(X), 1 - T(X)\}$$

- Jedná se o neparametrický test, protože není nutný předpoklad o tvaru rozdělení.
- Jestliže některé z hodnot se rovnají testované hodnotě kvantilu, budou vynechány.
- Výsledky pvalue se porovnají s hladinou významnosti $\propto$
- Při kvantilovém testu se vyžaduje větší množství naměřených dat.

8.3.5 Kvantilový test

- Pro výpočet testovacího kritéria lze použít funkci „binocdf" v matlabu

$$binocdf(Z_{-}, n, \propto)$$

- Př. Mějte data: 4,5,5,6,6,7,7,8,8,8,9,10,12 a otestujte, zda 90% kvantil je na hladině významnosti 0.05 roven 5.5.

- Matlab
  - H_0: $x_{0.9} = 5.5$; H_A: $x_{0.9} \neq 5.5$
  - >> x=[4,5,5,6,6,7,7,8,8,8,9,10,12]

  - >> Stat=binocdf(3,13,0.9) stat = 2.1E-8
  - >> pvalue=2*min(Stat,1-Stat)
  - pvalue = 4.2E-8
  - Hypotézu H_0 na hladině významnosti 5 % zamítáme.

- Jestliže se bude testovat 90% kvantil roven 8, je nutno odstranit ze vstupu naměřené hodnoty 8 a výsledek počítat pouze z 10 naměřených hodnot.
  - H_0: $x_{0.9} = 8$; H_A: $x_{0.9} \neq 8$
  - >> stat=binocdf(7,10,0.9) stat = 0.0702
  - >> pvalue=2*min(statistika,1-statistika)
  - pvalue = 0.1404
  - Hypotézu H_0 na hladině významnosti 5 % nezamítáme.

- Při testování malých / velkých kvantilů je nutno mít velké množství vstupních dat.

8.3.6 Wilcoxonův test

- Wilcoxonův test umožňuje na základě výběru $X_1, X_2, \dots, X_n$ ze spojitého rozdělení s hustotou $f(x)$, která je symetrická kolem mediánu $x_{0.5}$, zda se rovná testované hodnotě $x_{test0.5}$.
- Postup testování:
  1. Pro každou naměřenou hodnotu z výběru určeme $Y_i = X_i - x_{test0.5}$.
  2. Seřaďme veličiny Y_i vzestupně podle absolutní hodnoty a zaznamenejme jejich původní znaménko.
  3. Určeme pořadí veličiny $R_i = |Y_1| \leq |Y_2| \leq \dots \leq |Y_n|$. V případě shodných hodnot Y_i průměrujte pořadí
  4. Označme $R_i{}^+$ pořadí veličin s kladným znaménkem a $R_i{}^-$ se záporným
  5. Testová statistika je potom : $T(X) = min\left(\sum_{Y_i \geq 0} R_i{}^+, \sum_{Y_i \leq 0} R_i{}^-\right)$ a je nutno následně hledat v tabulkách.

- Pro stanovení zamítnutí / nezamítnutí hypotézy H_0 lze použít tabelovaných hodnot, nebo pro větší množství naměřených dat převést výsledek na normované normální rozdělení použitím vztahu:

$$z = \frac{\sum_{Y_i \geq 0} R_i{}^+ - \frac{n \cdot (n-1)}{4}}{\sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}}}$$

8.3.6 Wilcoxonův test

- Ze statistiky z lze příkazem normcdf(z,0,1) stanovit podle typu hypotézy pvalue.
  - $H_{0:}\, x_{0.5} = x_{test0.5}$ H$_A$: $x_{0.5} \neq x_{test0.5}$
    $pvalue = 2 \cdot \min\left(normcdf(z,0,1), 1 - normcdf(z,0,1)\right)$
  - $H_{0:}\, x_{0.5} \leq x_{test0.5}$ H$_A$: $x_{0.5} > x_{test0.5}$ $pvalue = 1 - normcdf(z,0,1)$
  - $H_{0:}\, x_{0.5} \geq x_{test0.5}$ H$_A$: $x_{0.5} < x_{test0.5}$ $pvalue = normcdf(z,0,1)$

- Jedná se o neparametrický test, protože není nutný předpoklad o tvaru rozdělení.
- Jestliže některé z hodnot se rovnají testovanému mediánu, budou vynechány.

8.3.6 Wilcoxonův test

- Funkce v matlabu signrank

- [p,h,stats]=signrank(x,m,alpha,method,tail)
  - x vektor vstupních dat
  - m porovnávaný medián
  - alpha hladina významnosti
  - method typ výpočtu pvalue.
    - Zadává se: 'method','exact‘ nebo 'method',‘approximate'
  - tail typ intervalového odhadu, zadává se: 'tail','both'
    - Nebo lze nahradit 'both' za 'right' nebo 'left'

  - p p-value
  - h výsledek hypotézy
  - stats $\sum_{Y_i \geq 0} R_i^{+}$ a hodnota z
- Parametry „alpha“, „method“, „tail“ a u výsledků „h“ a „stats“ lze vynechat

8.3.6 Wilcoxonův test

- Př. Mějme naměřená data: x =[4,5,5,6,6,7,7,8,8,8,9,10,12]. Ověřte na hladině významnosti 10 %, že medián je roven 5.5 pomocí přímého výpočtu a pomocí matlabu.
- Matlab:
  - >> [p,h,stats]=signrank(x,5.5,0.1,'method','exact','tail','both')
  - p = 0.0129
  - h = 1
  - stats = signedrank: 80
- Pro porovnání výsledky z matlabu pro method approximate
  - p = 0.0153
  - h = 1
  - stats = zval: 2.4244 signedrank: 80
- Na hladině významnosti 10 % zamítáme hypotézu H0, že medián je roven 5.5

8.3.6 Wilcoxonův test

- Přímý výpočet

X	4	5	5	6	6	7	7	8	8	8	9	10	12
Y	-1.5	-0.5	-0.5	0.5	0.5	1.5	1.5	2.5	2.5	2.5	3.5	4.5	6.5
Abs(Y)	1.5	0.5	0.5	0.5	0.5	1.5	1.5	2.5	2.5	2.5	3.5	4.5	6.5
pořadí	6	2.5	2.5	2.5	2.5	6	6	9	9	9	11	12	13
R+				2.5	2.5	6	6	9	9	9	11	12	13
suma(R+)	80												

$$z = \frac{\sum_{Y_i \geq 0} {R_i}^{+} - \frac{n \cdot (n-1)}{4}}{\sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}}} = \frac{80 - 39}{\sqrt{\frac{13 \cdot 14 \cdot 27}{24}}} = \frac{41}{14{,}30} = 2.8653$$

- pvalue= 0.0041
- Odlišný výsledek pvalue je způsoben odlišnou aproximací v matlabu a pomocí vzorce na normální rozdělení.

8.3.7 Test o parametru π relativní četnosti

- V sérii n nezávislých pokusů se náhodný jev A, vyskytl k krát. Pravděpodobnost náhodného jevu je $p = \frac{k}{n}$ a chceme ověřit, zda teoretická pravděpodobnost π se rovná p.
- Pro provedení testu je nutné mít alespoň $n > \frac{9}{p \cdot (1-p)}$ pokusů.
- Testovací statistika je:

$$T(X) = \frac{p - \pi}{\sqrt{\pi \cdot (1 - \pi)}} \sqrt{n}$$

- Testovací kritérium má normované normální rozdělení.

8.3.7 Test o parametru π relativní četnosti

- Seznam hypotéz:
  - $H_0: p = \pi$ $H_A: p \neq \pi$ H_0 nezamítám, když $z_{\frac{\propto}{2}} \leq T(X) \leq z_{1-\frac{\propto}{2}}$
  - $H_0: p \geq \pi$ $H_A: p < \pi$ H_0 nezamítám, když $T(X) \geq z_{\propto}$
  - $H_0: p \leq \pi$ $H_A: p > \pi$ H_0 nezamítám, když $T(X) \leq z_{1-\propto}$
  - V matlabu není funkce implementována.

- Př. Zeptali jsme 1000 voličů zda budou u voleb volit stranu LEŽ. 24 % odpovědělo, že ano. Strana LEŽ deklaruje, že u voleb získá 30 % hlasů. Ověřte na hladině významnosti 5% shodu středních hodnot.
  - $H_0: p = \pi = 30\ \%$ $H_A: p \neq 30\ \%$
  - $T(X) = \frac{p-\pi}{\sqrt{\pi \cdot (1-\pi)}}\sqrt{n} = \frac{0.24-0.30}{\sqrt{0.3 \cdot 0.7}}\sqrt{1000} = -4.1404$
  - pval= 2*normcdf(T,0,1)
  - $pvalue = 3.46 \cdot 10^{-5}$.
  - H_0, že strana LEŽ bude míst 30 %, na hladině významnosti 5 % zamítáme

8.4 Dvouvýběrové testy hypotéz

- 8.4.1 Test o shodě dvou rozptylů, výběrů z normálního rozdělení
- 8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení
- 8.4.3 Mann-Whitneyův test mediánů
- 8.4.4 Testování relativních četností

8.4.1 Test o shodě dvou rozptylů, výběrů z normálního rozdělení

- Mějme dva nezávislé výběry $X_1, X_2, \dots, X_{n_X}$ a $Y_1, Y_2, \dots, Y_{n_Y}$, které pocházejí z populací mající normální rozdělení $N(\mu_x, {\sigma_x}^2)$ a $N\left(\mu_y, {\sigma_y}^2\right)$. Parametry $\mu_x, {\sigma_x}^2, \mu_y, {\sigma_y}^2$ jsou neznámé.
- Chceme otestovat, zda ${\sigma_x}^2 = {\sigma_y}^2$.
- Vypočteme výběrový rozptyl z obou výběrů.
- Potom testovací kritérium je:

$$T(X,Y) = \frac{{s_X}^2}{{s_Y}^2}$$

- Testovací kritérium má F rozdělení s $n_X - 1, n_Y - 1$ stupni volnosti.
- Jestliže se testuje ${\sigma_y}^2 = k \cdot {\sigma_x}^2$ upravuje se testovací statistika na:

$$T(X,Y) = \frac{\frac{{s_X}^2}{{\sigma_X}^2}}{\frac{{s_Y}^2}{{\sigma_Y}^2}} = \frac{\frac{{s_X}^2}{{\sigma_X}^2}}{\frac{{s_Y}^2}{k \cdot {\sigma_X}^2}} = \frac{k \cdot {s_X}^2}{{s_Y}^2}$$

8.4.1 Test o shodě dvou rozptylů, výběrů z normálního rozdělení

- Seznam hypotéz:
  - $H_0: {s_X}^2 = {s_Y}^2$ $H_A: {s_X}^2 \neq {s_Y}^2$ 'both'
    - H_0 nezamítám, když testovací statistika je:
    $F_{\frac{\alpha}{2}}(n_X - 1, n_Y - 1\ st.v.) \leq T(X) \leq F_{1-\frac{\alpha}{2}}(n_X - 1, n_Y - 1\ st.v.)$
  - $H_0: {s_X}^2 \geq {s_Y}^2$ $H_A: {s_X}^2 < {s_Y}^2$ 'left'
    - H_0 nezamítám, když testovací statistika je:
    $T(X) \geq F_{\alpha}(n_X - 1, n_Y - 1\ st.v.)$
  - $H_0 : {s_X}^2 \leq {s_Y}^2$ $H_A: {s_X}^2 > {s_Y}^2$ 'right'
    - H_0 nezamítám, když testovací statistika je:
    $T(X) \leq F_{1-\alpha}(n_X - 1, n_Y - 1\ st.v.)$
- H_0 zamítám, když je p-value menší než $\propto$.

8.4.1 Test o shodě dvou rozptylů, výběrů z normálního rozdělení

- Funkce v matlabu: vartest2
  - Funkce vartest2 obsahuje více parametrů

- [h,p,ci,stats]=vartest2(x,y,alpha,tail)
  - x vektor X vstupních dat
  - y vektor Y vstupních dat
  - alpha hladina významnosti
  - Tail typ intervalového odhadu
    - 'both' oboustranný interval
    - 'left' $H_A: {s_X}^2 < {s_Y}^2$
    - 'right' $H_A: {s_X}^2 > {s_Y}^2$

  - h výsledek hypotézy
  - p p-value
  - ci konfidenční interval
  - stats statistické výsledky
    - fstat výsledek testovací statistiky
    - df1 počet stupňů volnosti vektoru X
    - df2 počet stupňů volnosti vektoru Y

- Všechny parametry na levé straně se nemusí uvádět. Často se uvádí např. [h,p]

8.4.1 Test o shodě dvou rozptylů, výběrů z normálního rozdělení

- Př. Máte následující data a zjistěte, zda n.v. X má na hladině významnosti 0.05 prokazatelně větší rozptyl než Y.
- x=[10.5,10.8,10.9,11.0,11.1,11.3,11.5,11.6]
- y=[10.8,11.0,11.1,11.2,11.2,11.3,11.4]

- Var(x)=0.1355 var(y)=0.0395
- Matlab:
  - $H_0 : {s_X}^2 \leq {s_Y}^2$ $H_A: {s_X}^2 > {s_Y}^2$ 'right'
  - >> [h,p,ci,stats]=vartest2(x,y,0.05,'right')
  - h = 0
  - p = 0.0772
  - ci = 0.8152 Inf
  - stats = fstat: 3.4292 df1: 7 df2: 6
- Hypotézu H_0 o shodě rozptylů na hladině významnosti 5 % nezamítáme.

- Klasický výpočet
  - $T = \frac{0.1355}{0.0395} = 3.4292$
  - $F_{1-\alpha}(n_X - 1, n_Y - 1\ st. v.)$ $F_{0.95}(7,6) = finv(0.95,7,6) = 4.207$
  - Hypotézu H_0 na hladině významnosti 5 % nezamítáme.

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- Mějme dva nezávislé výběry $X_1, X_2, \dots, X_{n_1}$ a $Y_1, Y_2, \dots, Y_{n_2}$, které pocházejí z populací mající normální rozdělení $N(\mu_x, {\sigma_x}^2)$ a $N\left(\mu_y, {\sigma_y}^2\right)$.
- Označme výběrové průměry a rozptyly:

$$\bar{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}, \bar{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}, {s_X}^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{n_1 - 1}, {s_Y}^2 = \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_2 - 1}$$

- Chceme otestovat, zda $\mu_x = \mu_y$ oproti alternativě $\mu_x \neq \mu_y$. (Obdobně i alternativní hypotézy menší, větší)
- Mohou nastat následující případy:
  1. Předem jsou definovány rozptyly obou populací (velmi vzácný případ)
  2. Rozptyly populací nejsou známy, ale předpokládáme, že jsou shodné
  3. Rozptyly populací nejsou známy, ale předpokládáme, že nejsou shodné
  4. Párové testy jsou uvedeny v kapitole jednovýběrových testů – viz kapitola 8.3.3.
- Testu o shodě středních hodnot předchází test shody rozptylů.
- Je nutno ověřit, že data pochází z normálního rozdělení

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- **1) Předem jsou definovány rozptyly obou populací**
- Velmi vzácný případ
- Potom testovací kritérium je:

$$T(X,Y) = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{n_X} + \frac{{\sigma_Y}^2}{n_Y}}}$$

- Testovací kritérium má normované normální rozdělení.
- V matlabu není implementováno.

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- **2) Rozptyly populací nejsou známy, ale předpokládáme, že jsou shodné**
- Potom testovací kritérium je:

$$T(X,Y) = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n_1 - 1) \cdot {s_X}^2 + (n_2 - 1) \cdot {s_Y}^2}{n_X + n_Y - 2}}}$$

- Testovací kritérum má Studentovo rozdělení s $n_X + n_Y - 2$ stupni volnosti.
- $(\mu_X - \mu_Y)$ je často rovno 0. Nenulový případ nastane, když H0: $\mu_X + 5 \leq \mu_Y$ H1: $\mu_X + 5 > \mu_Y$
  - Například: Na hladině významnosti 5% ověřte, že změnou technologie z X na Y dosáhneme zvýšení střední životnosti výrobku, alespoň o 5 měsíců.

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- **3) Rozptyly populací nejsou známy, ale předpokládáme, že nejsou shodné**
- Potom testovací kritérium je:

$$T(X,Y) = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{s_X}^2}{n_X} + \frac{{s_Y}^2}{n_Y}}}$$

- Testovací kritérum má Studentovo rozdělení s

$$\upsilon = \frac{\left(\frac{{s_X}^2}{n_X} + \frac{{s_Y}^2}{n_Y}\right)^2}{\left(\frac{{s_X}^2}{n_X}\right)^2 \cdot \frac{1}{n_X - 1} + \left(\frac{{s_Y}^2}{n_Y}\right)^2 \cdot \frac{1}{n_Y - 1}}$$ stupni volnosti.

- Předpoklad shodnosti rozptylu ověříme testem uvedeným v kapitolu 8.4.1.

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- Funkce v matlabu: ttest2
  - Funkce ttest2 obsahuje více parametrů

- [h,p,ci,stats]=ttest2(x,y,alpha,tail,vartype)
  - x 1. vektor vstupních dat
  - y 2. vektor vstupních dat
  - alpha hladina významnosti
  - tail typ intervalového odhadu
    - 'both' $H_A\colon \mu_A \neq \mu_B$
    - 'left' $H_A\colon \mu_A < \mu_B$
    - 'right' $H_A\colon \mu_A > \mu_B$
  - Vartype shodnost/neshodnost rozptylu
    - 'equal' rozptyly v obou výběrech jsou shodné 'unequal' rozptyly nejsou shodné
  - h výsledek hypotézy
  - p p-value
  - ci konfidenční interval
  - Stats Tstat velikost testového kriteria df počet stupňů volnosti
    sd průměrná směrodatná odchylka dat

Seznam hypotéz:

$H_0\colon \mu_A = \mu_B$ $H_A\colon \mu_A \neq \mu_B$ both

$H_0\colon \mu_A \geq \mu_B$ $H_A\colon \mu_A < \mu_B$ left

$H_0\colon \mu_A \leq \mu_B$ $H_A\colon \mu_A > \mu_B$ right

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- Výrobce A říká, že jeho výrobky mají delší životnost než výrobky B. Otestujte na hladině významnosti 5 %.
- A=[27,28,29,31,32,34,35,36,38,42], B=[25,27,29,29,30,31,32,33,33,34,35].

- 1) ověříme shodu rozptylů
  - >> A=[27,28,29,31,32,34,35,36,38,42];
  - >> B=[25,27,29,29,30,31,32,33,33,34,35];
  - >> [h,p]=vartest2(A,B,0.05,'both')
  - h = 0
  - p = 0.1934
  - Nezamítáme na hladině významnosti 5 % hypotézu shody rozptylů

- 2) ověříme shodnost středních hodnot
  - H_0: $\mu_A \leq \mu_B$ H_A: $\mu_A > \mu_B$
  - >> [h,p,ci,stats]=ttest2(A,B,0.05,'right','equal')
  - h = 0
  - p = 0.0839
  - ci = -0.5082 Inf
  - stats = tstat: 1.4343 df: 19 sd: 3.9456
- Na základě vstupních dat zamítáme hypotézu na hladině významnosti 5 %, že výrobce A má výrobky s delší životností než výrobce B.

8.4.2 Test o shodě dvou středních hodnot, výběrů z normálního rozdělení

- 3) počítejme stejný příklad s předpokladem, že rozptyly nejsou shodné
  - $H_0\colon\ \mu_A \leq \mu_B \quad H_A\colon\ \mu_A > \mu_B$
  - >> [h,p,ci,stats]=ttest2(A,B,0.05,'right','unequal')
  - h = 0
  - p = 0.0900
  - ci = -0.6093 Inf
  - stats = tstat: 1.4052 df: 15.19 sd: [4.73,3.07]

- Všimněte si, že pvalue je podobná jako v případě shodnosti roztpylů.

8.4.3 Mann-Whitneyův test mediánů

- Mann Whitneyův test je neparametrickým testem o shodě mediánů. Mějme dva nezávislé výběry $X_1, X_2, \ldots, X_{n_1}$ a $Y_1, Y_2, \ldots, Y_{n_2}$ ze spojitých rozdělení se stejným rozptylem a tvarem rozdělení.
- Testujeme hypotézu:
  - H_0: $x_{0.5} = y_{0.5}$ $\quad$ H_A: $x_{0.5} \neq y_{0.5}$ $\quad$ both
  - H_0: $x_{0.5} \geq y_{0.5}$ $\quad$ H_A: $x_{0.5} < y_{0.5}$ $\quad$ left
  - H_0: $x_{0.5} \leq y_{0.5}$ $\quad$ H_A: $x_{0.5} > y_{0.5}$ $\quad$ right

8.4.3 Mann-Whitneyův test mediánů

- Postup testování:
  1. Data z obou výběrů seřadíme do jednoho výběru vzestupně a zaznamenáme k jakému výběru patří.
  2. Určíme pořadí R_i vektoru. V případě shodných naměřených hodnot se průměruje pořadí R_i.
  3. Označíme $T_X = \sum_{i=1, i \in X}^{n_1+n_2} R_i$ (sečteme pořadí dat z výběru X) a $T_Y = \sum_{i=1, i \in Y}^{n_1+n_2} R_i$
  4. Pro součet T_X+T_Y platí: $T_X + T_Y = \frac{(n_1+n_2)\cdot(n_1+n_2+1)}{2}$
  5. Vypočteme statistiky: $U_X = n_1 \cdot n_2 + \frac{n_1\cdot(n_1+1)}{2} - T_X$ $\quad U_Y = n_1 \cdot n_2 + \frac{n_2\cdot(n_2+1)}{2} - T_Y$
  6. Testové kritérium je : $T(X,Y) = min(U_X, U_Y)$

- Pro stanovení zamítnutí / nezamítnutí hypotézy H_0 lze použít tabelovaných hodnot, nebo pro větší množství naměřených dat převést výsledek na normované normální rozdělení použitím vztahu:

$$z = \frac{min(U_X, U_Y) - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}}$$

8.4.3 Mann-Whitneyův test mediánů

- Funkce v matlabu ranksum

- [p,h,stats]=ranksum(x,y,alpha,method,tail)
  - x první vektor vstupních dat
  - y druhý vektor vstupních dat
  - alpha hladina významnosti
  - method typ výpočtu pvalue.
    - Zadává se: 'method','exact‘ nebo 'method',‘approximate'
  - tail typ intervalového odhadu, zadává se: 'tail','both'
    - Nebo lze nahradit 'both' za 'right' nebo 'left'

  - p p-value
  - h výsledek hypotézy
  - stats $\sum_{Y_i \geq 0} R_i^{+}$
- Parametry „alpha“, „method“, „tail“ a u výsledků „h“ a „stats“ lze vynechat

8.4.3 Mann-Whitneyův test mediánů

- Příklad z předchozí kapitoly, kde jsme předpokládali normalitu dat. Rozptyly byly buď shodné nebo neshodné.

- Výrobce A říká, že jeho výrobky mají delší životnost než výrobky B. Otestujte na hladině významnosti 5 %.
- A=[27,28,29,31,32,34,35,36,38,42], B=[25,27,29,29,30,31,32,33,33,34,35].

- Řešení:
  - >> A=[27,28,29,31,32,34,35,36,38,42];
  - >> B=[25,27,29,29,30,31,32,33,33,34,35];
  - H_0: $x_{0.5} \leq y_{0.5}$ H_A: $x_{0.5} > y_{0.5}$ right

  - >> [p,h,stats]=ranksum(A,B,0.05,'method','exact','tail','right')
  - p = 0.1295
  - h = 0
  - stats = ranksum: 126.5000

- Výsledky metodou aproximace na normované normální rozdělení
  - >> [p,h,stats]=ranksum(A,B,0.05,'method','approximate','tail','right')
  - p = 0.1291
  - h = 0
  - stats = zval: 1.1304 ranksum: 126.50

8.4.3 Mann-Whitneyův test mediánů

- Na základě vstupních dat zamítáme hypotézu, že výrobce A má výrobky s delší životností než výrobce B.
- Porovnej s výsledky z předchozího příkladu
  - Normální rozdělení, shodné rozptyly $p = 0.0839$
  - Normální rozdělení, neshodné rozptyly $p = 0.0900$
  - Shodný typ rozdělení, ale nenormální $p = 0.1295$
  - Rozdílné hodnoty pval jsou způsobeny nestejnými požadavky na vstupní data

8.4.4 Testování relativních četností

- Předpokládejme, že v sérii n_1 nezávislých opakování pokusu se náhodný jev A vyskytl x −krát. Obdobně v sérii n_2 nezávislých opakování pokusu se vyskytl náhodný jev A y −krát.
- Pravděpodobnost výskytu jevu A je $p_1 = \frac{x}{n_1}$; $p_2 = \frac{y}{n_2}$.
- Předpoklad testu je, že $n_1 > \frac{9}{p_1 \cdot (1-p_1)}$ a $n_2 > \frac{9}{p_2 \cdot (1-p_2)}$
- Chceme testovat hypotézu shody dvou relativních četností:
  - $H_0: \pi_1 = \pi_2$ $H_A: \pi_1 \neq \pi_2$ 'both'
  - $H_0: \pi_1 \geq \pi_2$ $H_A: \pi_1 < \pi_2$ 'left'
  - $H_0: \pi_1 \leq \pi_2$ $H_A: \pi_1 > \pi_2$ 'right'

8.4.4 Testování relativních četností

- Potom testovým kritériem je statistika:

$$T(X,Y) = \frac{(p_1 - p_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}}}$$

- Testovací kritérium splňuje normované normální rozdělení.
  - Oboustranný test H0 nezamítám, jestliže $z_{\frac{\alpha}{2}} < T(X,Y) < z_{1-\frac{\alpha}{2}}$ pval=2*min(normcdf(T,0,1),1-normcdf(T,0,1))
  - Jednostranný test left H0 nezamítám, jestliže $T(X,Y) > z_{\alpha}$ pval=normcdf(T,0,1))
  - Jednostranný test right H0 nezamítám, jestliže $T(X,Y) < z_{1-\alpha}$ pval=1-normcdf(T,0,1))

- Není implementováno v matlabu.
  - T=(p1-p2)/sqrt((p1*(1-p1)/n1)+(p2*(1-p2)/n2))

- Proběhly dva výzkumy týkající se volby politické strany LEŽ. V červnu se dotázali 500 respondentů, v červenci 300. Sympatie měla strana v červnu u 8 % respondentů, v červenci u 6 %. Strana chce vědět, zda na hladině významnosti 5 % se změnila podpora voličů.
  - p_1=0.08 p_2=0.06 n_1=500 n_2=300
  - $H_0: \pi_1 = \pi_2$ $H_A: \pi_1 \neq \pi_2$
  - >> T=(0.08-0.06)/sqrt((0.08*0.92/500)+(0.06*0.94/300)) T = 1.0924
  - >> pvalue=2*min(normcdf(T,0,1),1-normcdf(T,0,1)) pvalue = 0.2747
  - Nelze zamítnout H_0, že sympatie strany jsou stále shodné.

8.5 Vícevýběrové testy hypotéz

- 8.5.1 Test shody rozptylů
- 8.5.2 Jednofaktorová ANOVA
- 8.5.3 Kruskal Wallisův test
- 8.5.4 Metody mnohonásobného porovnávání
- 8.5.5 Dvoufaktorová vyvážená anova
- 8.5.6 Dvoufaktorová nevyvážená anova
- 8.5.7 Vícefaktorová anova

8.5.1 Test shody rozptylů

- Předpokládejme, že máme k nezávislých výběrů z normálního rozdělení a chceme testovat hypotézu:
- H_0: $\sigma_1{}^2 = \sigma_2{}^2 = \cdots = \sigma_k{}^2$
- H_A: alespoň jedna dvojice rozptylů se liší
- Využíváme:
  - Bartlettův test (nutná normalita vstupních dat)
  - Leveneův test (méně citlivý na porušení normality)

8.5.1 Test shody rozptylů

- Bartlettův test
  - testovací kritérium popisujeme pomocí χ^2 rozdělení
  - Implementován v matlabu
- Leveneův test
  - Testovací kritérium popisujeme pomocí F rozdělení
  - Implementován v matlabu
- Výsledkem
  - P-value a statistické údaje
  - Tabulka – výsledek testové statistiky, stupně volnosti, p-value
  - Krabicový graf s naměřenými daty – z velikosti výšky krabice lze odhadnout změny rozptylů jednotlivých výběrů

8.5.1 Test shody rozptylů

- Funkce v matlabu: vartestn

- [p,stats]=vartestn(X,group,displayopt,testtype)
  - X sloupcový vektor naměřených dat
  - group sloupcový vektor s uvedením označení skupiny
  - displayopt tvorba krabicového grafu – zadává se 'on' nebo 'off'
  - testtype druh statistického testu
    - 'classical' Bartlettův test
    - 'robust' Levennův test

  - p p-value
  - stats
    - chisqstat velikost testovaného kriteria (u Levennova testu Fstat)
    - df počet stupňů volnosti

8.5.1 Test shody rozptylů

- Př. Stroj produkuje nepřetržitě výrobky. Seřízení stroje se provede jednou za 3 hodiny. Vlivem seřízení se zmenší rozptyl rozměrů a tím i zmetkovitost výroby. Pracovník - statistik chce ověřit, zda doba 3 hodin má/nemá vliv na celkový rozptyl. Naměřil následující data odchylek rozměrů:
  Data po 1 hodině - 10 hodnot [-0.34,-0.22,-0.17,-0.04,-0.02,-0.01,0.01,0.02,0.05,0.11]
  Data po 2 hodinách - 11 hodnot [-0.44,-0.32,-0.16,-0.11,-0.07,-0.05,-0.03,0.02,0.09,0.15,0.21]
  Data po 3 hodinách - 11 hodnot [-0.17,-0.09,-0.01,0.02,0.05,0.08,0.12,0.17,0.24,0.48,0.56]
- Ověřte na hladině významnosti 5 %, zda lze přijmout hypotézu shody rozptylů.

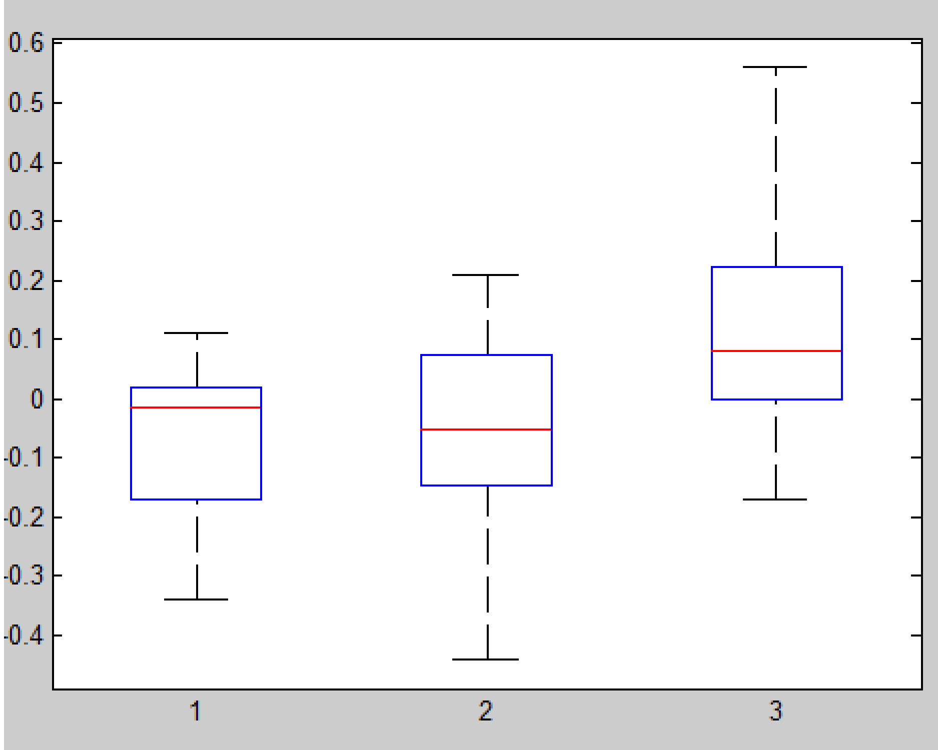
- Matlab:
  - >> data1=[-0.34,-0.22,-0.17,-0.04,-0.02,-0.01,0.01,0.02,0.05,0.11];
  - >> data2=[-0.44,-0.32,-0.16,-0.11,-0.07,-0.05,-0.03,0.02,0.09,0.15,0.21];
  - >> data3=[-0.17,-0.09,-0.01,0.02,0.05,0.08,0.12,0.17,0.24,0.48,0.56];
  - >> skupina1(1:10)=1;
  - >> skupina2(1:11)=2;
  - >> skupina3(1:11)=3;
  - >> data=[data1,data2,data3]'; % spojení dat do 1 vektoru,
    % ’ transponování dat
  - >> skupina=[skupina1,skupina2,skupina3]'; %transponování dat

8.5.1 Test shody rozptylů

1. př: Ověřena normalita dat – Bartlettův test
   - >> [p,stats]=vartestn(data,skupina,'on','classical')
   - p = 0.37266
   - stats = chisqstat: 1.9742 df: 2
   - Parametry 'on','classical' možno vynechat

2.př: Neověřena normalita dat – Leveneův test
   - >> [p,stats]=vartestn(data,skupina,'on','robust')
   - p = 0.3887
   - stats = fstat: 0.9764 df: [2 29]

Na hladině významnosti 5 % přijímáme hypotézu H_0 o shodě rozptylů.



Group Summary Table

Group	Count	Mean	Std Dev
1	10	-0.061	0.1386
2	11	-0.0645	0.19268
3	11	0.1318	0.22364
Pooled	32	0.0041	0.18977

Bartlett's statistic	1.97419
Degrees of freedom	2
p-value	0.37266

Group Summary Table

Group	Count	Mean	Std Dev
1	10	-0.061	0.1386
2	11	-0.0645	0.19268
3	11	0.1318	0.22364
Pooled	32	0.0041	0.18977

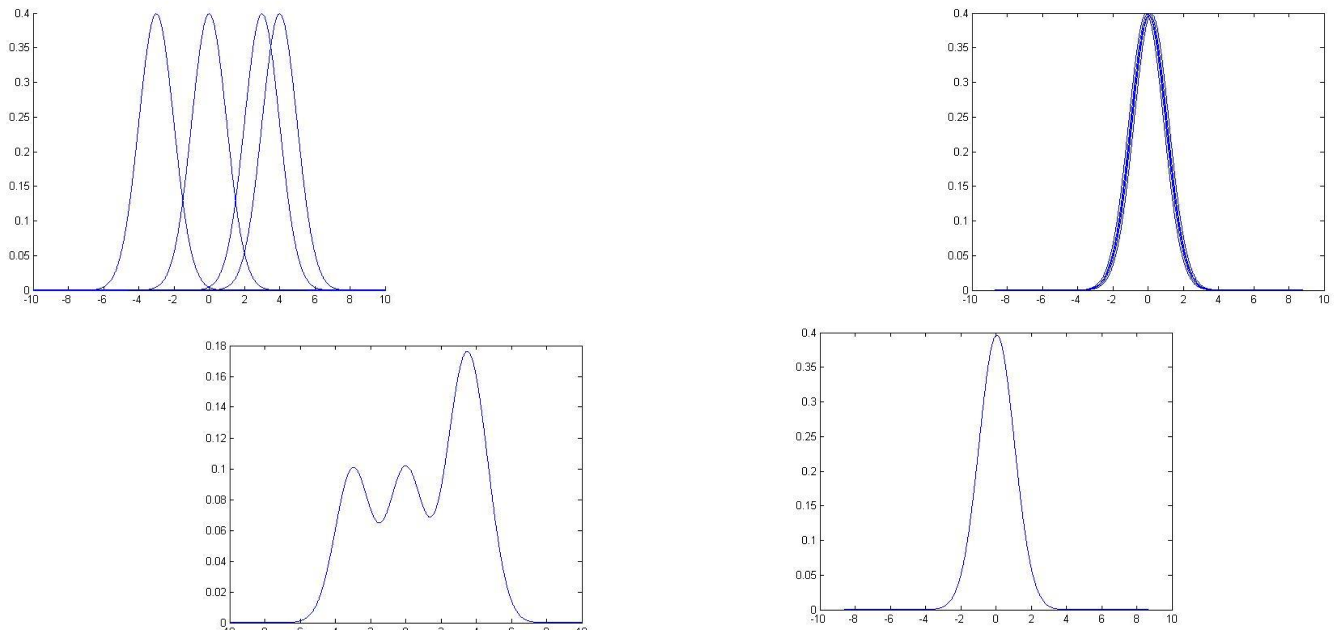
Levene's statistic	0.97642
Degrees of freedom	2, 29
p-value	0.3887

8.5.2 Jednofaktorová ANOVA

- ANOVA – analýza rozptylu
- Metoda ANOVA se používá pro porovnání shody průměrů více než dvou výběrů.
- Předpoklady:
  - Nezávislost výběrů.
  - Normalita rozdělení všech výběrů
  - Shodné rozptyly všech výběrů.
  - Pokud není splněno používá se Kruskal Wallisův test – kap. 8.5.3
- Mějme k ($k > 2$) nezávislých výběrů z normálního rozdělení, které mají normální rozdělení a shodný rozptyl. Potom lze pro určení shody středních hodnot použít metodu ANOVA. Testuje se hypotéza:
  - H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
  - H_A: alespoň dvě střední hodnoty nejsou rovny.

8.5.2 Jednofaktorová ANOVA

- Princip metody:
  - Zjišťujeme vliv velikosti rozptylu:
    - pro všechny sloučené naměřené hodnoty
    - pro naměřené hodnoty v každém výběru
    - Poměr velikosti rozptylů vzájemně porovnáváme.
  - Obrázek
    - Vlevo nahoře – čtyři rozdělení s rozdílnou střední hodnotou a shodným rozptylem o velikosti 1.
    - Vpravo nahoře – čtyři rozdělení s podobnou střední hodnotou a shodným rozptylem o velikosti 1
    - Vlevo dole – hustota pravděpodobnosti sloučených dat s rozdílnou střední hodnotou.
      Celkový rozptyl = 8.525 (rozptyl sloučených dat je výrazně vyšší než u základního výběru)
    - Vpravo dole – hustota pravděpodobnosti sloučených dat s podobnou střední hodnotou.
      Celkový rozptyl = 1.0475 (rozptyl sloučených dat je obdobný jako u základního výběru



8.5.2 Jednofaktorová ANOVA

- Postup metody ANOVA
  - Naměřené hodnoty: X_{ij}
    - i – označení výběru $i \in \langle 1, k \rangle$
    - j – naměřené hodnoty
  - 1) vypočteme střední hodnoty v každém výběru $\overline{X}_i$
  - 2) vypočteme střední hodnotu ze všech dat $\overline{\overline{X}}$.
  - 3) zjistíme rozsah naměřených hodnot v každém výběru n_i.
  - 4) zjistíme celkový součet čtverců

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{\overline{X}} \right)^2$$

    - výpočet obdobný jako u rozptylu, ale nedělí se počtem prvků.

8.5.2 Jednofaktorová ANOVA

- 5) reziduální součet čtverců

$$SS_E = \sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2\right)$$

- 6) součet čtverců mezi skupinami

$$SS_B = \sum_{i=1}^{k} n_{i\cdot}\left(\overline{X}_i - \overline{\overline{X}}\right)^2$$

- Platí: $SS_T = SS_E + SS_B$
- 7) vypočteme celkový rozptyl

$$MS_T = \frac{SS_T}{n-1}$$

- 8) vypočteme reziduální rozptyl

$$MS_E = \frac{SS_E}{n-k}$$

- 9) vypočteme rozptyl mezi skupinami

$$MS_B = \frac{SS_B}{k-1}$$

8.5.2 Jednofaktorová ANOVA

- 10) vypočteme testovací kriterium - $F poměr$, který splňuje Fisher Snedecorovo rozdělení s $(k-1, n-k)$ stupni volnosti.
- 11) Zjistíme pvalue
- 12) Sestavíme tabulku ANOVY

Součet čtverců	Počet stupňů volnosti	Rozptyl	F poměr	pvalue
SS_B	$df_B = k-1$	$MS_B = \frac{SS_B}{k-1}$	$F = \frac{MS_B}{MS_E}$	$1 - F(x)$
SS_E	$df_E = n-k$	$MS_E = \frac{SS_E}{n-k}$		
SS_T	$df_T = n-1$			

- 13) V případě zamítnutí H0 budou následovat metody mnohonásobého porovnávání.

8.5.2 Jednofaktorová ANOVA

- Funkce v matlabu anova1
- [p,anovatab,stats]=anova1(X,group,displayopt)
  - X sloupcový vektor naměřených dat
  - group sloupcový vektor s uvedením označení skupiny
  - displayopt tvorba krabicového grafu – zadává se ‘on’ nebo ‘off’

  - p p-value
  - anovatab vrátí výsledky v tabulce ANOVA
  - stats používá se jako vstup pro porovnání shody středních hodnot mezi výběry.

    vstup do funkce multcompare – viz kapitola 8.5.4

8.5.2 Jednofaktorová ANOVA

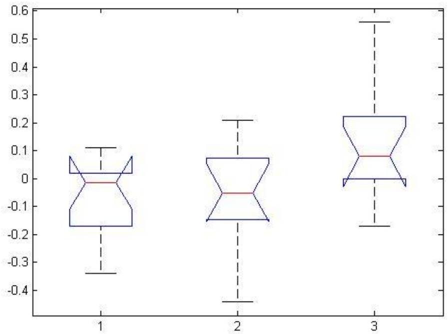
- Příklad obdobný jako v kapitole 8.5.1
- Př. Stroj produkuje nepřetržitě výrobky. Seřízení stroje se provede jednou za 3 hodiny. Vlivem seřízení se zmenší rozptyl rozměrů a tím i zmetkovitost výroby. Pracovník - statistik chce ověřit, zda doba 3 hodin má/nemá vliv na celkový rozptyl. Naměřil následující data odchylek rozměrů:
  - Data po 1 hodině - 10 hodnot [-0.34,-0.22,-0.17,-0.04,-0.02,-0.01,0.01,0.02,0.05,0.11]
  - Data po 2 hodinách - 11 hodnot [-0.44,-0.32,-0.16,-0.11,-0.07,-0.05,-0.03,0.02,0.09,0.15,0.21]
  - Data po 3 hodinách - 11 hodnot [-0.17,-0.09,-0.01,0.02,0.05,0.08,0.12,0.17,0.24,0.48,0.56]
- Ověřte na hladině významnosti 5 %, zda lze přijmout hypotézu shody středních hodnot.

- Matlab:
  - Vytvoříme dva vektory: „data“ a „skupina“ – viz kapitola 8.5.1
  - [p,anovatab,stats]=anova1(data,skupina, 'on')
- Výsledky
  - p = 0.0342
  - anovatab :

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F‘
'Groups'	[0.2736]	[2]	[0.1368]	[3.7994]	[0.0342]
'Error'	[1.0443]	[29]	[0.0360]	[]	[]
'Total'	[1.3180]	[31]	[]	[]	[]

  - stats =

gnames: {3x1 cell}	n: [10 11 11]	source: 'anova1‘
means: [-0.0610 -0.0645 0.1318]	df: 29	s: 0.1898



- Na hladině významnosti 0.05 hypotézu H0 o shodě středních hodnot zamítáme.
- (Pval je v anovatab označena červeně a je rovna 0.0342)
- Protože H0 zamítáme, bude následovat v kapitole 8.5.4 metoda mnohonásobného porovnávání, která zjistí, jaké výběry mají odlišné střední hodnoty.

8.5.2 Jednofaktorová ANOVA

- Výpočet otrocký pomocí papíru a tužky
- Data:
  - 1 hodina [-0.34, -0.22, -0.17, -0.04, -0.02, -0.01, 0.01, 0.02, 0.05, 0.11]
  - 2 hodiny [-0.44, -0.32, -0.16, -0.11, -0.07, -0.05, -0.03, 0.02, 0.09, 0.15, 0.21]
  - 3 hodiny [-0.17, -0.09, -0.01, 0.02, 0.05, 0.08, 0.12, 0.17, 0.24, 0.48, 0.56]

- 1) $\overline{x_1} = \frac{\sum_{i=1}^{n_1} x_i}{n_1} - 0.061;$ $\overline{x_2} = -0.0645;$ $\overline{x_3} = 0.1318$
- 2) $\bar{\bar{x}} = \frac{\sum_{i=1}^{n} x_i}{n} = 0.00406$
- 3) $n_1 = 10;$ $n_2 = 11;$ $n_3 = 11;$
- 4) $SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - 0.0406\right)^2 = 1.3180$
- 5) $SS_E = \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X_i}\right)^2\right) = (-0.34 - (-0.061))^2 + (-0.22 - (-0.061))^2 + \cdots = 1.0443$
- 6) $SS_B = SS_B = \sum_{i=1}^{k} n_i.\left(\overline{X_i} - \bar{\bar{X}}\right)^2 = 10 \cdot (-0.061 - 0.00406)^2 + 11 \cdot (-0.0645 - 0.00406)^2 + \cdots = 0.2736$
- 7) $MS_T = \frac{SS_T}{n-1} = \frac{1.3180}{31} = 0.042516$
- 8) $MS_E = \frac{SS_E}{n-k} = \frac{1.0443}{29} = 0.0360$
- 9) $MS_B = \frac{SS_B}{k-1} = \frac{0.2736}{2} = 0.1368$
- 10) $F = \frac{MS_B}{MS_E} = \frac{0.1368}{0.0360} = 3.7994$
- 11) $pval = 1 - fcdf(F, n-1, n-k) = 0.0342$

- Výsledky z bodů 4, 5, 6, 8, 9, 10 a 11 jsou vidět v anova tabulce. Nejdůležitější je výsledek pvalue.

8.5.3 Kruskal Wallisův test

- Kruskal Wallisův test je neparametrickou obdobou jednofaktorové metody ANOVA
- Kruskal Wallisův test je obdobou Mannova-Whitneyova testu pro více než 2 výběry

- Nechť máme k nezávislých výběrů z rozdělení se spojitou distribuční funkcí stejného typu.
- $H_0: x_{0.5_1} = x_{0.5_2} = \cdots = x_{0.5_k}$
- H_A: alespoň jedna shoda mediánů neplatí.

8.5.3 Kruskal Wallisův test

- Postup testu:
  - Naměřené hodnoty: X_{ij}
    - i – označení výběru $i \in \langle 1, k \rangle$
    - j – naměřené hodnoty
  - 1) Všechna data se setřídí vzestupně a zaznamenáme k jakému výběru patří.
  - 2) Určíme pořadí R_{ij}. V případě shodně naměřených hodnot se průměruje pořadí R_{ij}.
  - 3) Označíme T_i součtem pořadí každého i-tého výběru. Platí: $\sum_i T_i = \frac{n \cdot (n+1)}{2}$
  - 4) Testovací kritérium je:

$$T = -3 \cdot (n+1) + \frac{12}{n \cdot (n+1)} \sum_{i=1}^{k} \frac{{T_i}^2}{n_i}$$

- Kritické hodnoty testovacího kritéria jsou buď tabelovány, nebo jestliže každý výběr je větší než 5 prvků, potom má testová statistika přibližně χ^2 rozdělení s (k-1) stupni volnosti.

8.5.3 Kruskal Wallisův test

- Funkce v matlabu kruskalwallis
- [p,anovatab,stats]=kruskalwallis(X,group,displayopt)
  - X sloupcový vektor naměřených dat
  - group sloupcový vektor s uvedením označení skupiny
  - displayopt tvorba krabicového grafu – zadává se ‘on’ nebo ‘off’
  - p p-value
  - anovatab vrátí výsledky v obdobné tabulce ANOVA, kde částečným vstupem je pořadí.
  - stats používá se jako vstup pro porovnání shody středních hodnot mezi výběry.

    používá se funkce multcompare

8.5.3 Kruskal Wallisův test

- Příklad obdobný jako v kapitole 8.5.2
- Stroj produkuje nepřetržitě výrobky. Seřízení stroje se provede jednou za 3 hodiny. Vlivem seřízení se zmenší rozptyl rozměrů a tím i zmetkovitost výroby. Pracovník - statistik chce ověřit, zda doba 3 hodin má/nemá vliv na celkový rozptyl. Naměřil následující data odchylek rozměrů:

  Data po 1 hodině - 10 hodnot [-0.34,-0.22,-0.17,-0.04,-0.02,-0.01,0.01,0.02,0.05,0.11]

  Data po 2 hodinách - 11 hodnot [-0.44,-0.32,-0.16,-0.11,-0.07,-0.05,-0.03,0.02,0.09,0.15,0.21]

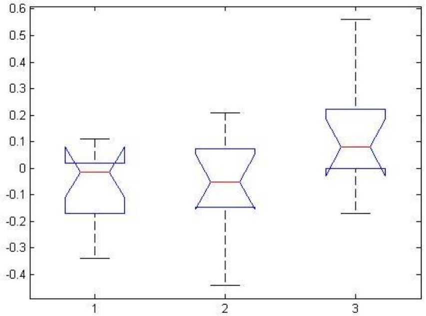
  Data po 3 hodinách - 11 hodnot [-0.17,-0.09,-0.01,0.02,0.05,0.08,0.12,0.17,0.24,0.48,0.56]
- Matlab:
  - Vytvoříme dva vektory: „data“ a „skupina“ – viz kapitola 8.5.1
  - [p,anovatab,stats]=kruskalwallis(data,skupina,'on')
- Výsledky
  - p = 0.0642
  - anovatab =

'Source'	'SS'	'df'	'MS'	'Chi-sq'	'Prob>Chi-sq'
'Groups'	[482.5932]	[2]	[241.2966]	[5.4911]	[0.0642]
'Error'	[2.2419e+03]	[29]	[77.3071]	[]	[]
'Total'	[2.7245e+03]	[31]	[]	[]	[]

  - stats =

    gnames: {3x1 cell} n: [10 11 11] source: 'kruskalwallis‘

    meanranks: [13.5500 13.8182 21.8636] sumt: 42
- Na hladině významnosti 0.05 hypotézu H0 o shodě mediánů nezamítáme.
- (Pval je v anovatab označena červeně a je rovna 0.0642)



8.5.3 Kruskal Wallisův test

- Stejný příklad, výpočet ručně
  - Data 1 hod [-0.34,-0.22,-0.17,-0.04,-0.02,-0.01,0.01,0.02,0.05,0.11]
  - Data 2 hod [-0.44,-0.32,-0.16,-0.11,-0.07,-0.05,-0.03,0.02,0.09,0.15,0.21]
  - Data 3 hod [-0.17,-0.09,-0.01,0.02,0.05,0.08,0.12,0.17,0.24,0.48,0.56]
- Určím pořadí:
  - Data 1=[2,4,5.5,12,14,15.5,17,19,21.5,25] $T_1 = 135.5$
  - Data2=[1,3,7,8,10,11,13,19,24,27,29] $T_2 = 152$
  - Data3=[5.5,9,15.5,19,21.5,23,26,28,30,31,32] $T_3 = 240.5$
  - $n = 32$
- Testovací kriterium:

$$T = -3 \cdot (n+1) + \frac{12}{n \cdot (n+1)} \sum_{i=1}^{k} \frac{{T_i}^2}{n_i} = -3 \cdot 33 + \frac{12}{32 \cdot 33} \cdot \left(\frac{135.5^2}{10} + \frac{152^2}{11} + \frac{240.5^2}{11} \right) = 5.484$$

- >> pvalue=1-chi2cdf(5.484,2)
- pvalue = 0.0644
- H0 nezamítáme

8.5.4 Metody mnohonásobného porovnávání

- V případě zamítnutí nulové hypotézy shody všech středních hodnot (mediánů) je třeba zjistit mezi kterými výběry dochází k rozdílům.
- Testuje se:
  - H_0: $\mu_I = \mu_J$, H_A: $\mu_I \neq \mu_J$ pro každou kombinaci výběrů.
  - Obdobně pro shodu mediánů.
- Existuje několik metod:
  - Tukeyova metoda
  - Bonferroniho metoda
  - Scheffeho metoda

- Tukeyova metoda
  - Defaultně předimplementována v Matlabu
  - Obtížné pro vysvětlování, nebude zde vysvětlena

8.5.4 Metody mnohonásobného porovnávání

- Bonferroniho metoda
  - Odhad střední hodnoty / mediánu i, j výběru je $\widetilde{x}_I, \widetilde{x}_J$
  - Hladinu významnosti α se upraví na novou hladinu významnosti $\alpha^* = \frac{\alpha}{\binom{k}{2}}$, kde k je počet výběrů
  - Pro každou kombinaci výběrů i,j vypočteme

$$\left|\widetilde{x}_I - \widetilde{x}_J\right| \geq t_{\left(1-\frac{\alpha^*}{2}\right)}(n-k\ st.vol.)\sqrt{MS_e}\sqrt{\frac{1}{n_I}+\frac{1}{n_J}}$$

  - Jestliže je splněna nerovnost, rozdíly mezi výběry i a j jsou významné.
- Scheffeho metoda
  - Pro každou kombinaci výběrů i,j vypočteme

$$\left|\widetilde{x}_I - \widetilde{x}_J\right| \geq \sqrt{MS_e}\sqrt{F_{1-\alpha}(k-1, n-k\ st.vol.) \cdot (k-1) \cdot \left(\frac{1}{n_I}+\frac{1}{n_J}\right)}$$

  - Jestliže je splněna nerovnost, rozdíly mezi výběry i a j jsou významné.

8.5.4 Metody mnohonásobného porovnávání

- Funkce v matlabu multcompare
  - Navazuje na funkce anova1, kruskalwallis
  - Vstupem je funkce stats

- [comparison,means]=multcompare(stats,'parametr1',...)
  - Stats výsledek stats z funkce anova1,kruskalwallis
  - Parametr
    - 'alpha' zadává se hladina významnosti, např. 0.1
    - 'display' 'on' nebo 'off' zobrazí graf s intervalovým rozpětím průměrů.
    - 'ctype' typ metody. Implicitně 'tukey-kramer', dále lze 'bonferroni' a 'scheffe'
  - Comparison vektor 5 sloupců
    - I označení i-tého vektoru
    - J označení j-tého vektoru
    - Konf.min minimum konfidenčního intervalu rozdílu
    - St.hod střední hodnota rozdílu i a j
    - Konf.max maximum konfidenčního intervalu rozdílu
  - Means střední hodnoty a směrodatná odchylka výběrů

8.5.4 Metody mnohonásobného porovnávání

- Pokračování příkladu 8.5.2 ANOVA
  - >> [p,anovatab,stats]=anova1(data,skupina,'on')
- Výsledky
  - p = 0.0342
  - anovatab :

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Groups'	[0.2736]	[2]	[0.1368]	[3.7994]	[0.0342]
'Error'	[1.0443]	[29]	[0.0360]	[]	[]
'Total'	[1.3180]	[31]	[]	[]	[]

  - stats =

gnames: {3x1 cell}	n: [10 11 11]	source: 'anova1'
means: [-0.0610 -0.0645 0.1318]	df: 29	s: 0.1898

  - s: 0.1898
- Mnohonásobné porovnání Tukeyova metoda
  - >> COMPARISON= multcompare(stats)
  - COMPARISON =

1.0000	2.0000	-0.2012	0.0035	0.2083	interval obsahuje 0, proto má 1. a 2. výběr stejnou střední hodnotu
1.0000	3.0000	-0.3976	-0.1928	0.0120	
2.0000	3.0000	-0.3962	-0.1964	0.0035	

- Hypotéza H_0 o shodnosti všech mediánů byla na hladině významnosti 5 % zamítnuta.
- Hypotéza H_0 o shodnosti mediánů mezi dvěma výběry však zamítnuta nebyla.
  - Způsobeno malým množstvím vstupních dat, kdy vzorce platí limitně pro nekonečné množství vstupů
  - Z výsledků porovnání 1. a 3. (2. a 3.) výběru se velmi blížíme hranici, kdy bychom H_0 zamítli.

8.5.4 Metody mnohonásobného porovnávání

- Mnohonásobné porovnání, Bonferroniho metoda
  - >> [COMPARISON,MEANS] = multcompare(stats,'alpha',0.1,'display','on','ctype','bonferroni')
  - COMPARISON =

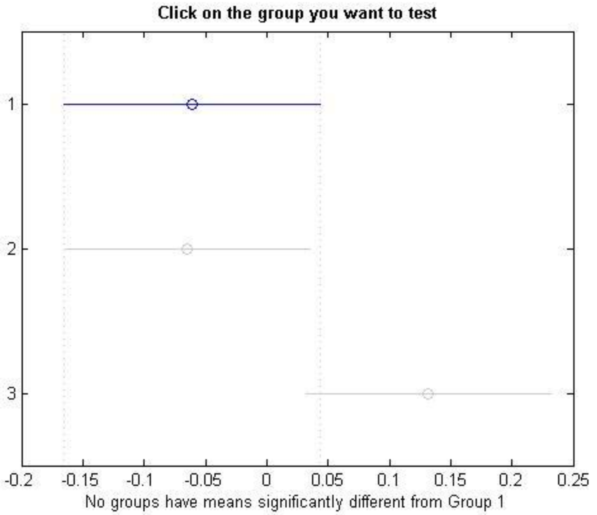
1. výběr	2. výběr	interval spolehlivosti rozdílů stř.hodnot/mediánů		
		min	rozdil medianů	max
1.0000	2.0000	-0.1817	0.0035	0.1888
1.0000	3.0000	-0.3781	-0.1928	-0.0076
2.0000	3.0000	-0.3772	-0.1964	-0.0156

jestliže interval neobsahuje 0, pak se přikláníme k H1

  - MEANS =

-0.0610	0.0600
-0.0645	0.0572
0.1318	0.0572

- Mezi 1. a 3., 2. a 3. výběrem byla prokázána na hladině významnosti 10 % neshodnost mediánů.



8.5.5 Dvoufaktorová vyvážená anova

- Jednofaktorová x Dvoufaktorová anova
  - Jednofaktorová (kap. 8.5.2) je určena pro porovnání shody průměrů více než dvou výběrů
  - Dvoufaktorová anova – je určena pro porovnání shody průměrů rozdělených podle dvou faktorů, kde každý faktor má více než dva výběry
- Příklad: Při měření byl zjišťován vliv teploty (1. faktor) při uskladnění a vlhkosti (2. faktor) na celkovou životnost výrobku (v měsících). Byly naměřeny následující hodnoty:

Teplota |Vlhkost	20 %	40 %	60 %	80 %	100 %
20 °C	32, 31	30, 28	30, 28	27, 26	20, 19
30 °C	30, 33	28, 25	26, 25	27, 22	18, 21
40 °C	27, 29	26, 24	25, 25	23, 21	21, 16

  - U jednofaktorové anovy bychom mohli zjistit pouze shodu středních hodnot v závislosti na teplotě (vlhkosti), ale nelze stanovit vliv obou faktorů. Proto se využívá dvoufaktorová anova.
- Vyvážená x nevyvážená anova – pokud je v každé buňce tabulky stejný počet měření, jedná se o vyváženou anovu.

8.5.5 Dvoufaktorová vyvážená anova

- Hypotéza:
  - 1) pro první faktor platí, že střední hodnota v řádku
    - H_0: $\mu_{1.} = \mu_{2.} = \mu_{3.} = \ldots = \mu_{p.}$
    - H_1: alespoň dvě střední hodnoty v řádcích nejsou shodné
  - 2) pro druhý faktor platí, že střední hodnota ve sloupci
    - H_0: $\mu_{.1} = \mu_{.2} = \mu_{.3} = \ldots = \mu_{.q}$
    - H_1: alespoň dvě střední hodnoty ve sloupcích nejsou shodné
  - 3) pro každý řádek a sloupec
    - H_0: interakce mezi faktory je nulová
    - H_1: interakce mezi faktory není nulová

8.5.5 Dvoufaktorová vyvážená anova

- Předpoklady:
  - 1) jednotlivá měření jsou nekorelovaná
  - 2) měření jsou normálně rozdělena
  - 3) shodné rozptyly všech rozptylů jednotlivých výběrů
- Způsob výpočtu je obdobný jako u jednofaktorové anovy. Vypočteme:
  - 1) S_i – součet čtverců mezi řádkovými průměry (p-1 stupňů volnosti)

$$S_i = n \cdot q \cdot \sum_{i=1}^{p} (\overline{x_{i.}} - \bar{x})^2$$

  - 2) S_j – součet čtverců mezi sloupcovými průměry (q-1 stupňů volnosti)

$$S_j = n \cdot p \cdot \sum_{j=1}^{q} \left(\overline{x_{.j}} - \bar{x}\right)^2$$

  - 3) $S_{i,j}$ – součet čtverců mezi interakcemi $S_{i,j}$ ((p-1)*(q-1) stupňů volnosti)

$$S_{i,j} = n \cdot \sum_{i,j} \left(\overline{x_{i,j}} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}\right)^2$$

  - Pokračování na dalším slidu

8.5.5 Dvoufaktorová vyvážená anova

- Způsob výpočtu je obdobný jako u jednofaktorové anovy. Vypočteme:
  - 4) S_r – reziduální součet čtverců (N-pq stupňů volnosti)

$$S_r = n \cdot p \cdot \sum_{i,j,k} \left(x_{i,j,k} - \overline{x_{i,j}}\right)^2$$

  - 5) S – celkový součet čtverců (N-1 stupňů volnosti)

$$S = \sum_{i,j,k} \left(x_{i,j,k} - \bar{x}\right)^2$$

  - $S = S_i + S_j + S_{i,j} + S_r$
  - Poznámka: Jestliže jsou naměřená data bez opakování (v každé buňce je právě jedna naměřená hodnota), nelze vypočítat součet čtverců mezi interakcemi.
- Vyhodnocení je obdobné jako u jednofaktorové anovy pomocí Fisher Snedecorova rozdělení.

8.5.5 Dvoufaktorová vyvážená anova

- Anova tabulka

	Součet čtverců	**Stupně volnosti**	**Podíl**	**Testovací kritérium**	**P-value**
Řádky	S_i	$p-1$	$MS_i = \frac{S_i}{p-1}$	$F = \frac{MS_i}{MS_r}$	Vypočte SW
Sloupce	S_j	$q-1$	$MS_j = \frac{S_j}{q-1}$	$F = \frac{MS_j}{MS_r}$	Vypočte SW
Interakce	$S_{i,j}$	$(p-1)(q-1)$	$MS_{ij} = \frac{S_{ij}}{(p-1)(q-1)}$	$F = \frac{MS_{ij}}{MS_r}$	Vypočte SW
Rezidua	S_r	$N - p \cdot q$	$MS_r = \frac{S_r}{N - p \cdot q}$		
Celkem	S	$N-1$			

- F podíl – testuje se pomocí jednostranného testu
  - F podíl >>1 – existuje vliv mezi řádky (sloupci, existuje interakce)
  - F podíl blízký 0 – neexistuje vliv mezi řádky (sloupci, neexistuje interakce)

8.5.5 Dvoufaktorová vyvážená anova

- Funkce v matlabu anova2

- [p,table,stats]=anova2(X,reps, displayopt)
  - X matice vstupních dat, viz další slide
  - Reps počet měření v každé ze skupin
  - displayopt vytvoří tabulku s výsledky, zadává se ‘on’ nebo ‘off’

  - p p-value
  - table vrátí výsledky v tabulce ANOVA
  - stats používá se jako vstup pro porovnání shody středních hodnot mezi výběry.

    vstup do funkce multcompare – viz kapitola 8.5.4

8.5.5 Dvoufaktorová vyvážená anova

- X definování matice vstupních dat

$$\begin{bmatrix} x_{111} & x_{121} & x_{131} \\ x_{112} & x_{122} & x_{132} \\ x_{211} & x_{221} & x_{231} \\ x_{212} & x_{222} & x_{232} \end{bmatrix}$$

- První index – i- tý řádkový faktor
  - 2 řádkové faktory
- Druhý index – j-tý sloupcový faktor
  - 3 sloupcové faktory
- Třetí index – k-té měření
  - 2 měření pro každou kombinaci faktorů
- Pozor při vyplňování parametru „reps“, pokud se nezadá uvažuje se že třetí index je 1.

8.5.5 Dvoufaktorová vyvážená anova

- Příklad:Při měření byl zjišťován vliv teploty (1. faktor) a vlhkosti (2. faktor) na celkovou životnost výrobku (v měsících). Byly naměřeny následující hodnoty:

Teplota |Vlhkost	20 %	40 %	60 %	80 %	100 %
20 °C	32, 31	30, 28	30, 28	27, 26	20, 19
30 °C	30, 33	28, 25	26, 25	27, 22	18, 21
40 °C	27, 29	26, 24	25, 25	23, 21	21, 16

- Zjistěte, zda existuje vliv teploty a vlhkosti na životnost výrobku.
- Letmým pohledem lze odhadnout, že výrobek má největší životnost pokud je využíván v málo vlhkém prostředí při teplotě 20°C.
- Vstupní matice:

32	30	30	27	20	20 °C
31	28	28	26	19	
30	28	26	27	18	30 °C
33	25	25	22	21	
27	26	25	23	21	40 °C
29	24	25	21	16	
20 %	40 %	60 %	80 %	100 %	

8.5.5 Dvoufaktorová vyvážená anova

- Matlab:
  - [p,tab,stats]=anova2(X,2,'on')
  - Výsledky:

Source	SS	df	MS	F	Prob>F
Columns	405.533	4	101.383	30.11	0
Rows	57.867	2	28.933	8.59	0.0033
Interaction	15.467	8	1.933	0.57	0.7835
Error	50.5	15	3.367		
Total	529.367	29			

  - Byl zjištěn na hladině významnosti 5 % vliv vlhkosti (pvalue = 0)
  - Byl zjištěn na hladině významnosti 5 % vliv teploty (pvalue = 0.0033)
  - Nebyl zjištěn vliv interakcí mezi sloupci a řádky.
- V úloze následuje porovnání pomocí funkce multcompare viz kapitoly 8.5.4

8.5.6 Dvoufaktorová nevyvážená anova

- Dvoufaktorovou anovu lze vytvořit i pro nevyvážená data.
  - Vznikne například odstraněním některých dat (odlehlé hodnoty)
- Odlišná funkce pro výpočet a odlišné zadávání vstupních dat oproti vyvážené anově.
- Výsledky jsou obdobné ve formě anova tabulky.

- Matlabovská funkce anovan

8.5.6 Dvoufaktorová nevyvážená anova

- Funkce v matlabu anovan
- [p,table,stats]=anovan(y,group,param)
  - y vektor naměřených hodnot
  - group skupina určující faktory (viz příklad)
  - Param označení zadefinovaného parametru
    - Alpha hladina významnosti testu
    - Display vytvoření externí anova tabulky (implicitně ‘on’, jinak ‘off’)
    - Model typ modelu: ‘linear’ – pouze vliv faktorů, ‘’interaction – zjistí i vliv základních interakcí, ‘full’ – zjistí vliv všech interakcí (možné výrazné rozšíření modelu)
  - p p-value
  - anovatab vrátí výsledky v tabulce ANOVA
  - stats používá se jako vstup pro porovnání shod středních hodnot mezi výběry. Vstup do funkce multcompare – viz kapitola 8.5.4

8.5.6 Dvoufaktorová nevyvážená anova

- Příklad: Shodné zadání jako v kapitole 8.5.5, jen bude vypočteno pomocí funkce anovan

y=[32,31,30,28,30,28,27,26,20,19, 30,33,28,25,26,25,27,22,18,21, 27,29,26,24,25,25,23,21,21,16];

%vlhkost

g1=[20 20 40 40 60 60 80 80 100 100 20 20 40 40 60 60 80 80 100 100 20 20 40 40 60 60 80 80 100 100]

%teplota

g2=[20 20 20 20 20 20 20 20 20 20 30 30 30 30 30 30 30 30 30 30 40 40 40 40 40 40 40 40 40 40]

- [p,table,stats]=anovan(y,{g1,g2},'alpha',0.01,'model','interaction')

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	405.533	4	101.383	30.11	0
X2	57.867	2	28.933	8.59	0.0033
X1*X2	15.467	8	1.933	0.57	0.7835
Error	50.5	15	3.367		
Total	529.367	29			

- Vypočten obdobný příklad bez interakcí
- [p,table,stats]=anovan(y,{g1,g2},'alpha',0.01,'model','linear')

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	405.533	4	101.383	35.35	0
X2	57.867	2	28.933	10.09	0.0007
Error	65.967	23	2.868		
Total	529.367	29			

- Výsledky hypotéz jsou shodné, tj. na hladině významnosti 5 % je výsledná životnost výrobků závislá nejen na teplotě, ale i na vlhkosti.

- Analýza může pokračovat příkazem „multcomare“, kde jsou porovnávány vlivem g1
  - Pokud chcete stanovit vliv podle faktoru g2, upravte příkaz: anovan(y,{g2,g1})

8.5.7 Vícefaktorová anova

- Způsob výpočtu je shodný s dvoufaktorovou nevyváženou anovou.
- Vstupují více než 2 faktory. Při výpočtu se výrazně zvyšuje počet vzájemných interakcí, proto při menším množství dat se využívá lineární model.
  - Například faktor teploty, vlhkosti, době uložení výrobků

- Matlabovská funkce: anovan
  - Viz kapitola 8.5.6

8.6 Přehled testů

		data z normálního rozdělení	data nejsou z normálního rozdělení
1 výběr	**rozptyl**	8.3.1 - **vartest**	
	střední hodnota/ medián	8.3.2 a 8.3.3 - **ttest**	8.3.4 - znaménkový test - **signtest**, 8.3.6 - Wilcoxonův test (nutná symetrie) – **signrank**
	relativní četnost	8.3.7 - výpočet vzorcem	
2 výběry	**rozptyl**	8.4.1 – **vartest2**	
	střední hodnota/ medián	8.4.2 - **ttest2**	8.4.3 - Mann-Whitneyův test - **ranksum**
	relativní četnost	8.4.4 - výpočet vzorcem	
více výběrů	**rozptyl**	8.5.1 - Bartlettův test - **vartestn**	8.5.1 - Leveneův test - **vartestn**
	střední hodnota/ medián	8.5.2 - ANOVA - **anova1** 8.5.4, 8.5.5 – více faktorů – **anova2, anovan**	8.5.3 - Kruskal Wallisův test - **kruskalwallis**

9 – Testy dobré shody

- V předchozích kapitolách jsme předpokládali, že data pochází z určitého rozdělení. Zatím jsme neřekli, jakým způsobem lze určité rozdělení testovat.
- Využíváme – testy dobré shody
  - H_0: Teoretické a empirické rozdělení se shoduje
  - H_A: Teoretické a empirické rozdělení se neshoduje
- Nejčastěji se využívá
  - χ^2-test dobré shody - ověření shody distribucí na základě rozdílů mezi skutečnou četností O_i a očekávanou E_i.
  - Kolmogorov-Smirnovův test - maximální rozdíl distribuce mezi očekávanou a zjištěnou distribuční funkcí.

9 – Testy dobré shody

- 9.1 χ^2- test dobré shody - ověření relativních četností
- 9.2 χ^2- test dobré shody - ověření shody s očekávaným rozdělením
- 9.3 Kolmogorov – Smirnovův jednovýběrový test rozdělení
- 9.4 Kolmogorov – Smirnovův dvouvýběrový test shody rozdělení

9.1 χ^2- test dobré shody ověření relativních četností

- Populaci roztřídíme podle nějakého znaku do k disjunktních skupin a chceme na základě náhodného výběru ověřit, zda jsou relativní četnosti jednotlivých variant rovny $\pi_1, \pi_2, \dots, \pi_k$.
- Test je založen na porovnávání očekávané četnosti $E_i = n \cdot \pi_i$ a naměřené četnosti O_i.
- Testovací kritérium je:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Testovací kritérium, jestliže se provádí dostatečně velký výběr, má přibližně χ^2 rozdělení s $k-1$ stupni volnosti.
  - V každé skupině musí být očekávaná četnost větší než 5.
  - V případě, že podmínka velikosti očekávané četnosti není splněna, sloučí se dva sousední intervaly v jeden.
  - Pro přibližné stanovení počtu intervalů se využívají dva vzorce:

$$interval = \sqrt{n}, nebo\ interval = 3.3ln(n)$$

  - $pvalue = 1 - F(x)$

9.1 χ^2- test dobré shody ověření relativních četností

- Př. Otestujte, zda šestistěnná kostka je „cinknutá“, když jsme naměřili následující výsledky hodů:

  1 – 15x   2 – 20x   3 – 9x   4 – 32x   5 – 24x   6 – 20x
- $H_0: \pi_1 = \pi_2 = \ldots = \pi_6 = \frac{1}{6}$ H_A: neplatí H_0
- $E_i = \frac{120}{6} = 20$
- $T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{25}{20} + 0 + \frac{121}{20} + \frac{144}{20} + \frac{16}{20} + 0 = \frac{306}{20} = 15.3$
- Rozdělení s 5 stupni volnosti
  - >> pvalue=1-chi2cdf(15.3,5)
  - pvalue = 0.0092
- H0 zamítáme

- Funkce a řešení pomocí matlabu – viz kapitola 9.2

9.2 χ^2- test dobré shody ověření shody s očekávaným rozdělením

- χ^2 - test dobré shody lze použít pro ověření, zda data pocházejí z určitého typu rozdělení.
- Testuje se hypotéza:
  - H_0: data pochází z daného rozdělení
  - H_A: data nepochází z daného rozdělení
- Používá se testu uvedeného v kapitole 9.1, přičemž optimální parametry daného rozdělení se určují pomocí střední hodnoty a rozptylu. Z těchto parametrů se následně vypočtou optimální parametry daného rozdělení.

9.2 χ^2- test dobré shody ověření shody s očekávaným rozdělením

- Funkce v matlabu chi2gof
- [h,p,stats]=chi2gof(x,'parametr1',hodnota,…)
  - x vstupní vektor
  - h výsledek hypotézy
  - p pvalue
  - stats výsledek statistiky
    - chi2stat velikost testové veličiny
    - df počet stupňů volnosti
    - edges hraniční body intervalů
    - O skutečná četnost na intervalech
    - E očekávaná četnost na intervalech
- POZOR: nutno dávat pozor, zda se intervaly výrazně neslučují. Může ovlivnit kvalitu výsledků. Lze rozpoznat podle stupňů volností „df“ v záložce stats. V případě, že je počet stupňů volnosti malý, vyskočí warningová hláška. Například:

  Warning: After pooling, some bins still have low expected counts.
  The chi-square approximation may not be accurate
  > In chi2gof>poolbins (line 304)
  In chi2gof (line 247)

9.2 χ^2- test dobré shody ověření shody s očekávaným rozdělením

- Parametry funkce chi2gof
  - Typ rozdělení
    - ‘cdf’,{@normcdf,mean(x),std(x)}
    - Lze zadávat i jina spojitá rozdělení wblcdf, expcdf apod.
    - Parametry mean(x),std(x) se nahrazují příslušnými parametry rozdělení
    - ‘cdf’,{@expcdf,50}
  - Hraniční body ‘edges’
    - Představuje dolní a horní mez
  - Očekávaný počet hodnot ‘expected’
    - Počet prvků v intervalech.
    - Parametr nelze použít, jestliže je definován typ rozdělení
    - Využití pro porovnání relativních četností

9.2 χ^2- test dobré shody ověření shody s očekávaným rozdělením

- Př. Vygenerujte 1000 dat z exponenciálního rozdělení se střední hodnotou rovnou 100. Otestujte tato data, zda jsou z exponenciálního rozdělení. Dále otestujte, zda mohou být data i z normálního rozdělení.
- Výpočet pro exponenciální rozdělení
  - >> x=exprnd(100,1,1000);
  - >> [h,p,stats]=chi2gof(x,'cdf',{@expcdf,100})
  - h = 0
  - p = 0.6104
  - stats = chi2stat: 3.5863 df: 5
    edges: [0.0522 79.8697 159.6873 239.5048 319.3223 399.1398 478.9573 798.2274]
    O: [561 256 105 46 17 7 8] E: [550 247 111 50 22.5 10.2 8.3]
- Výpočet pro normální rozdělení
  - >> [h,p,stats]=chi2gof(x,'cdf',{@normcdf,mean(x),std(x)})
  - h = 1
  - p = 3.0317e-27
  - stats = chi2stat: 122.1213 df: 2
    edges: [0.0522 79.8697 159.6873 239.5048 319.3223 798.2274]
    O: [561 256 105 46 32] E: [439.0340 308.0875 184.0187 58.2264 10.6334]

9.2 χ^2- test dobré shody ověření shody s očekávaným rozdělením

- Př. Otestujte, zda šestistěnná kostka je „cinknutá“, když jsme naměřili následující výsledky hodů:

  1 – 15x 2 – 20x 3 – 9x 4 – 32x 5 – 24x 6 – 20x
- $H_0: \pi_1 = \pi_2 = \ldots = \pi_6 = \frac{1}{6}$ H_A: neplatí H_0
- Matlab:
  - x=[1,1,…,1,1, 2,2,…,2,2, 3,3,…,3,3, 4,4,…,4,4, 5,5,…,5,5, 6,6,…,6,6];

    15x 20x 9x 32x 24x 20x
  - >> [h,p,stats]=chi2gof(x,'expected',[20,20,20,20,20,20])
  - h = 1
  - p = 0.0092
  - stats =

    chi2stat: 15.3000 df: 5

    edges: [1.0000 1.8333 2.6667 3.5000 4.3333 5.1667 6.0000]

    O: [15 20 9 32 24 20] E: [20 20 20 20 20 20]
- Na hladině významnosti 5 % H0 zamítáme
- Lze zadat i hranice, zde na výsledek nemá vliv
  - [h,p,stats]=chi2gof(x,'edges',[0.5,1.5,2.5,3.5,4.5,5.5,6.5],'expected',[20,20,20,20,20,20])

9.3 Kolmogorov – Smirnovův jednovýběrový test rozdělení

- Kolmogorov – Smirnovův test se používá k ověření hypotézy, zda výběr pochází z rozdělení se spojitou distribuční funkcí $F_0(x)$.
- Testuje se hypotéza:
  - H_0: data pochází z daného spojitého rozdělení
  - H_A: data nepochází z daného spojitého rozdělení
- Průběh testu:
  - 1) naměřený náhodný výběr setřídíme od nejmenšího k největšímu
  - 2) z dat vytvoříme distribuční funkci
  - 3) stanovíme rozdíl mezi distribuční funkcí z rozdělení a distribuční funkcí z naměřených dat
  - 4) testovací statistika je maximum z rozdílů distribučních funkcí
  - Výsledek testovací statistiky buď porovnáme s hodnotou v tabulkách, pro větší počet dat lze přijímací kritérium aproximovat vztahem:

$$maximální\ rozdíl_\alpha = \sqrt{\frac{1}{2n} ln \frac{2}{\alpha}}$$

, kde α je hladina významnosti

9.3 Kolmogorov – Smirnovův jednovýběrový test rozdělení

- Funkce v matlabu: kstest

- [h,p,ksstat,cv]=kstest(x,CDF,alpha,type)
  - X vstupní vektor
  - Cdf matice o dvou sloupcích
    - 1. sloupec – naměřené hodnoty,
    - 2. sloupec – hodnota porovnávané distribuční funkce
  - Alpha hladina významnosti
  - Type typ porovnávání

    'unequal' H0 distribuční funkce si jsou rovny

    'larger' H0 porovnávaná hypotetická distribuční funkce je větší než naměřená

    'smaller' H0 porovnávaná hypotetická distribuční funkce je menší než naměřená

  - h výsledek hypotézy
  - p pvalue
  - ksstat výsledek max. rozdílu mezi hypotetickou a skutečnou distribuční funkcí
  - cv maximální povolený rozdíl

9.3 Kolmogorov – Smirnovův jednovýběrový test rozdělení

- Mějme naměřeno 10 hodnot, ověřte zda data mohou být z normálního rozdělení s parametry $\mu = 10, \sigma = 5$.
- Data:
  - >> x=[-9,4,6,7,8,10,15,18,23,24];
- Výpočet:
  - >> a(:,1)=x‘;
  - >> a(:,2)=normcdf(a(:,1),10,5); %vypočte dist. fci z norm rozd.
  - >> [h,p,ksstat,cv]=kstest(x,a)

  - h = 0
  - p = 0.5085
  - ksstat = 0.2452
  - cv = 0.4093

9.3 Kolmogorov – Smirnovův jednovýběrový test rozdělení

- Funkce kstest se využívá především pro porovnání s distribuční funkcí, kterou předem známe, nebo není implementována ve funkci „lillietest“
- Funkce lillietest se využívá, pro ověření, zda data pocházejí z normálního nebo exponenciálního rozdělení.

- Funkce v matlabu lillietest

- [h,p,kstat,critval]=lillietest(x,alpha,distr)
  - x vstupní vektor
  - alpha hladina významnosti
  - distr typ distribuce: ‘norm’ data z normálního rozdělení
    ‘exp’ data z exponenciálního rozdělení

  - h výsledek hypotézy
  - P pvalue
  - kstat výsledek testového kritéria
  - critval kritická hodnota testu

9.4 Kolmogorov – Smirnovův dvouvýběrový test shody rozdělení

- Dvouvýběrový Kolmogorov-Smirnovův test se používá k ověření hypotézy, zda dva výběry pochází z rozdělení se shodnou distribuční funkcí.
- Testuje se hypotéza:
  - $H_0: F(x) = F(y)$
  - $H_A: F(x) \neq F(y)$

- Průběh testu:
  - 1) náhodné výběry setřídíme od nejmenšího k největšímu
  - 2) z dat vytvoříme distribuční funkce
  - 3) stanovíme rozdíl mezi oběma distribučními funkcemi
  - 4) testovací statistika je maximum z rozdílů distribučních funkcí
  - Výsledek testovací statistiky buď porovnáme s hodnotou v tabulkách. Pro větší počet dat, lze podle velikosti hladiny významnosti stanovit maximální rozdíl vzorcem:

$maximální\ rozdíl_\alpha = k\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}}$, kde parametr k je závislý na hladině významnosti α.

α	0.2	0.1	0.05	0.02	0.01
k	1.07	1.22	1.36	1.52	1.63

9.4 Kolmogorov – Smirnovův dvouvýběrový test shody rozdělení

- Funkce v matlabu: kstest2

- [h,p,kstest]=kstest2(x,y,alpha,type)
  - x 1. vstupní vektor
  - y 2. vstupní vektor
  - alpha hladina významnosti
  - type typ porovnávání

    'unequal' $H_1: F(x) \neq F(y)$

    'larger' $H_1: F(x) > F(y)$

    'smaller' $H_1: F(x) < F(y)$

  - h výsledek hypotézy
  - p pvalue
  - Kstest maximální rozdíl distribučních funkcí

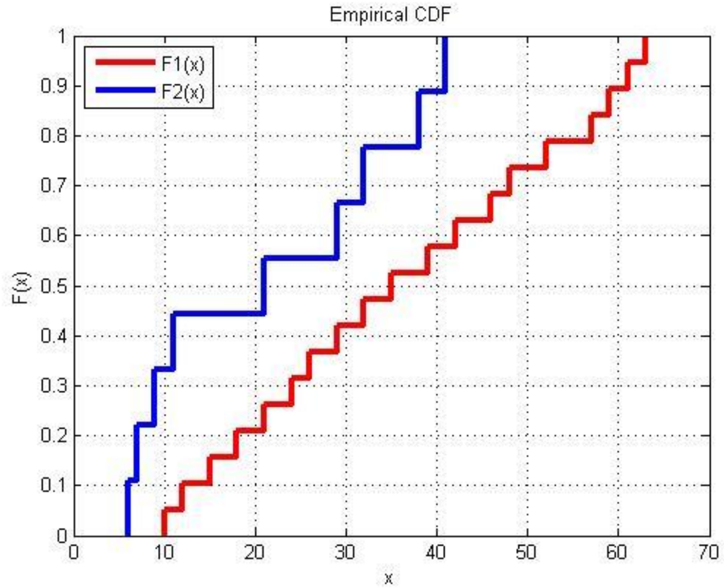
9.4 Kolmogorov – Smirnovův dvouvýběrový test shody rozdělení

- Př. Otestujte na hladině významnosti 1 %, zda data z vektoru x a y mohou mít shodnou distribuční funkci.

  x=[10,12,15,18,21,24,26,29,32,35,39,42,46,48,52,57,59,61,63]

  y=[6,7,9,11,21,29,32,38,41]
- Výpočet
  - >> [h,p,kstest]=kstest2(x,y,0.01)
  - h =   0
  - p =   0.1702
  - kstest =   0.4211
- Hypotézu H_0 o shodnosti distribučních funkcí na hladině významnosti 1 % nezamítáme.

```
F1 = cdfplot(x);
hold on
F2 = cdfplot(y)
set(F1,'Linewidth',3,'Color','r')
set(F2,'Linewidth',3)
legend([F1  F2],'F1(x)','F2(x)','Location','NW')
```



10 – Analýza závislostí

- 10.1 Kontingenční tabulky
- 10.2 Asociační tabulky
- 10.4 Pearsonův koeficient korelace
- 10.5 Spearmanův koeficient korelace

- Vhodné pro stanovení, zda naměřené hodnoty dvou výběrů jsou vzájemně nezávislé.
  - Existuje závislost velikosti mzdy na dosaženém vzdělání?
  - Existuje závislost mezi výškou a hmotností člověka?

10 – Analýza závislostí

- Mějme znaky X a Y, které nabývají určitých hodnot:
  - X a Y dvojstavová logika Asociační tabulka
  - X a Y diskrétní hodnoty Kontingenční tabulky
  - X a Y spojité hodnoty

    Pearsonův a Spearmanův korelační koeficient

10.1 Kontingenční tabulka

- Kontingenční tabulka se užívá k vizualizaci vzájemného vztahu dvou statistických znaků.
- Řádky kontingenční tabulky odpovídají možným hodnotám prvního znaku $x_1, \dots, x_i, \dots, x_r$. Sloupce pak možným hodnotám druhého znaku $y_1, \dots, y_j, \dots, y_s$. V příslušné buňce kontingenční tabulky je uveden počet případů n_{ij}, kdy nastala i-tá hodnota prvního a j-tá hodnota druhého znaku.
- V posledním řádku a sloupci jsou uvedeny součty výskytu jednotlivých znaků - $n_{i.}, n_{.j}$ a celkový rozsah výběru n.

Ukázka kontingenční tabulky

$X \backslash Y$	$y_{[1]}$	$y_{[2]}$	$\cdots$	$y_{[s]}$	Celkem
$x_{[1]}$	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
$x_{[2]}$	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Celkem	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	n

10.1 Kontingenční tabulka

- Chceme určit, zda znaky uvedené v kontingenční tabulce jsou vzájemně nezávislé.
- Nezávislé znaky

	Y1	Y2	Y3	
X1	0.4*0.5*n	0.3*0.5*n	0.3*0.5*n	0.5*n
X2	0.4*0.3*n	0.3*0.3*n	0.3*0.3*n	0.3*n
X3	0.4*0.2*n	0.3*0.2*n	0.3*0.2*n	0.2*n
	0.4*n	0.3*n	0.3*n	n

Závislé znaky

	Y1	Y2	Y3	
X1	0.2*n	0.3*n	0*n	0.5*n
X2	0.1*n	0*n	0.2*n	0.3*n
X3	0.1*n	0*n	0.1*n	0.2*n
	0.4*n	0.3*n	0.3*n	n

- Hypotézy:
  - H_0: znaky X a Y v kontingenční tabulce jsou statisticky nezávislé
  - H_A: znaky X a Y v kontingenční tabulce jsou statisticky závislé

10.1 Kontingenční tabulka

- Testování vychází z obdobného principu jako test dobré shody.
  - Porovnávání empirické četnosti s teoretickými.
  - Empirická četnost – naměřená data $O_{ij} = n_{ij}$
  - Teoretická četnost $E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$
- Testovací kritérium používáme náhodnou veličinu:

$$K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

- Podmínky:
  - Žádná z očekávaných četností E_{ij} nesmí být menší než 2.
  - Alespoň 80 % očekávaných četností E_{ij} musí být větší než 5.
- Testovací kritérium má v případě platnosti nulové hypotézy χ^2 rozdělení s $(r-1) \cdot (s-1)$ stupni volnosti.

10.1 Kontingenční tabulka

- Funkce v matlabu: crosstab

- [tbl,chi2,p]=crosstab(x1,x2)
  - X1 vektor hodnot prvního znaku
  - X2 vektor hodnot druhého znaku

  - Tbl kontingenční tabulka
  - Chi2 hodnota

Př. U 200 osob bylo zjišťováno jejich nejvyšší dosažené vzdělání, zároveň bylo zjišťováno nejvyšší dosažené vzdělání jejich otců. Určete závislost / nezávislost obou zjištěných údajů.

Naměřené hodnoty

syn/otec	ZŠ	učiliště	SŠ	VŠ	součet
ZŠ	12	8	5	1	26
učiliště	21	18	3	3	45
SŠ	15	33	30	15	93
VŠ	3	5	12	16	36
součet	51	64	50	35	200

Teoretická četnost

syn/otec	ZŠ	učiliště	SŠ	VŠ	součet
ZŠ	6.63	8.32	6.5	4.55	26
učiliště	11.475	14.4	11.25	7.875	45
SŠ	23.715	29.76	23.25	16.275	93
VŠ	9.18	11.52	9	6.3	36
součet	51	64	50	35	200

- Teoretické četnosti jsou: $T_{ZŠ,ZŠ} = \frac{26 \cdot 51}{200} = 6.63$
- Výsledek testového kritéria je: $K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{(12-6.63)^2}{6.63} + \cdots = 54.74$
- Porovnám výsledek testového kritéria s χ^2 rozdělením

10.1 Kontingenční tabulka

- Matlab:
  - >> x1(1:26)=1; >> x2(1:45)=2; >> x3(1:93)=3; >> x4(1:36)=4;
  - >> x=[x1,x2,x3,x4];
  - >> y11(1:12)=1; >> y12(1:8)=2; >> y13(1:5)=3; >> y14(1)=4;
  - >> y21(1:21)=1; >> y22(1:18)=2; >> y23(1:3)=3; >> y24(1:3)=4;
  - >> y31(1:15)=1; >> y32(1:33)=2; >> y33(1:30)=3; >> y34(1:15)=4;
  - >> y41(1:3)=1; >> y42(1:5)=2; >> y43(1:12)=3; >> y44(1:16)=4;
  - >> y=[y11,y12,y13,y14,y21,y22,y23,y24,y31,y32,y33,y34,y41,y42,y43,y44];
  - >> [table,chi2,p]=crosstab(x,y)
  - table =
  - 12 8 5 1
  - 21 18 3 3
  - 15 33 30 15
  - 3 5 12 16
  - chi2 = 54.7524
  - p = 1.3577e-08
- Zamítáme hypotézu, že data jsou vzájemně nezávislá.

10.2 Asociační tabulky

- Speciální typ kontingenčních tabulek, pro sledování závislosti dvou znaků.
  - Lze použít kontingenční tabulku, ale výsledek se porovnává pouze s 1 stupněm volnosti.

Znak	Y	$\bar{Y}$
X	a	b
$\bar{X}$	c	d

- Testovací statistika:

$$T = \frac{ad}{bc}$$

- Jestliže níže uvedený interval obsahuje 1, potom lze říci, že data jsou nezávislá; v opačném případě jsou závislá.
  - T je výsledek testovací statistiky, $z_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálního rozdělení.

$$\left\langle T \cdot e^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}}; T \cdot e^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle$$

10.2 Asociační tabulky

- Př. Byly sledovány výsledky při hodu dvou označených mincí, které jsou uvedeny v tabulce. Zjistěte na hladině významnosti 5 %, zda jednotlivé hody jsou vzájemně nezávislé.
- Př. 1

mince1/mince 2	panna	orel
panna	32	28
orel	30	27

Př. 2: ukázka závislých dat

mince1/mince 2	panna	orel
panna	40	20
orel	20	40

- $T = \frac{32 \cdot 27}{30 \cdot 28} = 1.0286$
- $\left\langle T \cdot e^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}}; T \cdot e^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle$
- $T_{min} = T \cdot e^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} = 1.0287 \cdot e^{-\sqrt{0.1373} \cdot 1.96} = 0.498$
- $T_{max} = T \cdot e^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} = 1.0287 \cdot e^{\sqrt{0.1373} \cdot 1.96} = 2.127$
- Data jsou nezávislá

- $T = \frac{40 \cdot 40}{20 \cdot 20} = 4$
- $T_{min} = 4 \cdot e^{-\sqrt{0.15} \cdot 1.96} = 1.87$
- $T_{max} = 4 \cdot e^{\sqrt{0.15} \cdot 1.96} = 8.55$
- Data jsou závislá

10.3 Kovariance

- Kovariancí lze stanovit míru lineární závislosti dvou náhodných veličin.
- Výběrová kovariance se vypočítá dle vzorce

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})\right) = \frac{n\sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot (n-1)}$$

- Jedná se o „smíšený“ rozptyl dvou vektorů

- Matlab: cov(x,y)

- Základní vlastnosti:
  - $cov(X,X) = D(X)$ lineární závislost
  - Jsou-li X,Y nezávislé náhodné veličiny, pak $cov(X,Y) = 0$
    - Obráceně neplatí – jestliže nám vyjde na datech $cov(X,Y) = 0$ neznamená to nutně, že data jsou nezávislá.
  - $cov(a_1X + b_1, a_2Y + b_2) = a_1 \cdot a_2 \cdot cov(X,Y)$

10.3 Kovariance

- Hodnota kovariance může být:
  - $cov(X,Y) > 0$ jestliže obě veličiny rostou, případně klesají, což může naznačovat lineární závislost
  - $cov(X,Y) < 0$ jestliže jedna veličina roste a druhá klesá , což může naznačovat lineární závislost
  - $cov(X,Y) \cong 0$ veličiny se neovlivňují, což může naznačovat lineární nezávislost
- Kovarianční matice
  - $var(X) = \begin{pmatrix} cov(X,X) & cov(X,Y) \\ cov(X,Y) & cov(Y,Y) \end{pmatrix} = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix}$
  - Kovarianční matice je symetrická
  - $D(X+Y) = D(X) + D(Y) + 2cov(X,Y)$
  - $D(X-Y) = D(X) + D(Y) - 2cov(X,Y)$

10.3 Kovariance

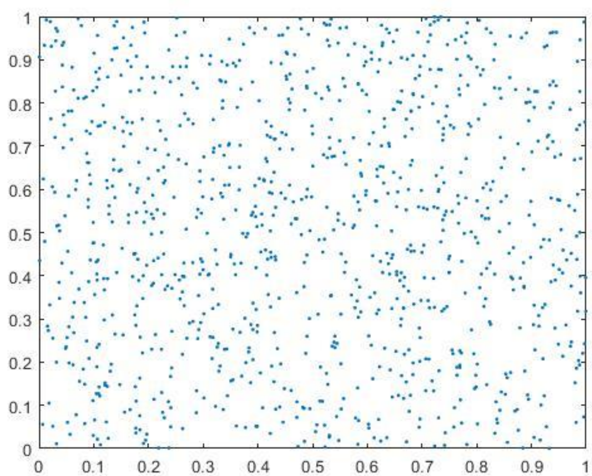
- Př. Sledujeme vliv mezi výškou otce a výškou syna. Zjistěte výslednou kovarianci
  - otec: 176 179 182 171 154 167 172 176
  - syn: 179 183 186 181 157 171 174 182
  - >> cov(otec,syn)

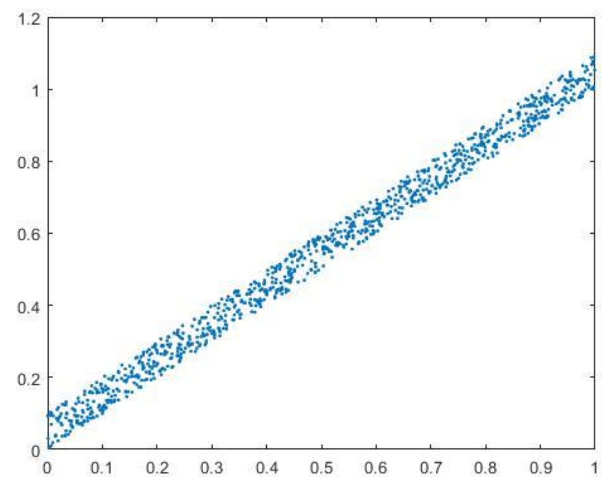
    ans = 75.8393 78.0536

    78.0536 86.5536
- Př. Vygenerujme 2 vektory o délce 100 s náhodnými čísly z rovnoměrného rozdělení a zjistěte jejich kovarianci.

10.3 Kovariance

- Př. Vygenerujme 2 vektory o délce 1000 s náhodnými čísly z rovnoměrného rozdělení a zjistěte jejich kovarianci. Jedná se o nezávislé jevy.
  - >> x=unifrnd(0,1,1,1000);
  - >> y=unifrnd(0,1,1,1000);
  - >> cov(x,y)
  - ans = 0.0797 0.0003
    0.0003 0.0814
- Př. Vygenerujem vektor o délce 1000. Druhý vektor bude popsán rovnicí Y=X+0.1*náhodné číslo. Zjistěte jejich kovarianci. Jedná se o závislé jevy.
  - >> a=unifrnd(0,1,1,1000);
  - >> b=a+0.1*unifrnd(0,1,1,1000);
  - >> cov(a,b)
  - ans = 0.0838 0.0838
  - 0.0838 0.0838
  - >> plot(a,b,'.')





10.4 Pearsonův korelační koeficient

- Pearsonův korelační koeficient ρ se používá, jestliže vstupní data mohou nabývat spojitých hodnot, a jsou zároveň normálně rozdělená.
- Výpočet z výběrové kovariance a výběrových rozptylů. Výhodou je, že korelační koeficient je omezen mezi -1 a 1.

$$\rho = \frac{cov(X,Y)}{\sqrt{s^2(X) \cdot s^2(Y)}} \qquad s^2(X), s^2(Y) \neq 0$$

- Funkce v matlabu: corrcoef
- Vlastnosti korelačního koeficientu
  - $-1 \leq \rho(X,Y) \leq 1$
  - $\rho(X,Y) = \rho(Y,X)$
  - $\rho(X,X) = 1$
  - Jsou-li X, Y nezávislé, pak $\rho(X,Y) = 0$
  - Je-li $\rho(X,Y) = \pm 1$, pak existuje lineární závislost mezi X a Y, taková že $Y = a \cdot x + b$.
- Je-li $\rho(X,Y) = 0$, říkáme, že X a Y jsou nekorelované.
  - Pozor, pokud náhodné veličiny jsou nekorelované, neznamená to, že jsou nezávislé.

10.4 Pearsonův korelační koeficient

- Výběrový korelační koeficient

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left((X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \right)$$

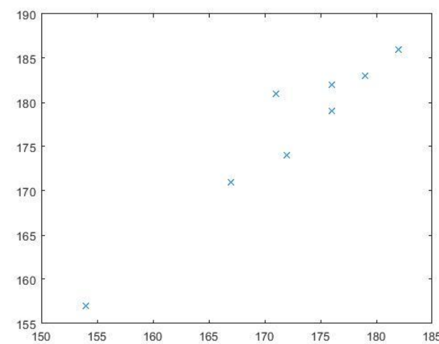
$$= \frac{\sum_{i=1}^{n} (X_i \cdot Y_i) - n \cdot \bar{X} \cdot \bar{Y}}{\left(\sum_{i=1}^{n} X_i{}^2 - n \cdot \bar{X}^2 \right) \cdot \left(\sum_{i=1}^{n} Y_i{}^2 - n \cdot \bar{Y}^2 \right)}$$

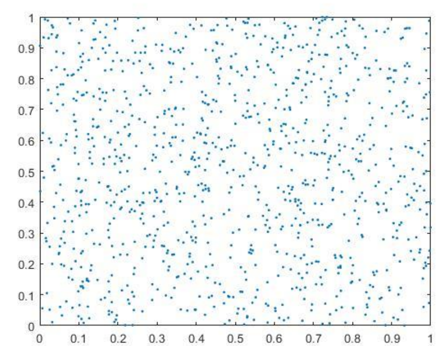
$$S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$$

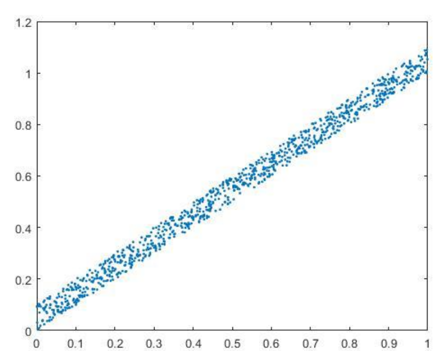
- Vstupní data jsou závislé $|r| \approx 1$, nezávislé $|r| \approx 0$.

10.4 Pearsonův korelační koeficient

- Př.: Výpočet korelace z příkladů uvedených v kapitole o kovarianci
  - Př.: výška otce a syna
    - >> corrcoef(otec,syn)
    - ans =
    - 1.0000 0.9634
    - 0.9634 1.0000
  - Př.: nezávislá data
    - >> corrcoef(x,y)
    - ans =
    - 1.0000 0.0036
    - 0.0036 1.0000
  - Př.: závislá data
    - >> corrcoef(a,b)
    - ans =
    - 1.0000 0.9946
    - 0.9946 1.0000







10.4 Pearsonův korelační koeficient

- Testování lineární závislosti / nezávislosti dat
  - H_0: $\rho = 0$
  - H_A: $\rho \neq 0$ (lze vytvořit i jednostrannou alternativní hypotézu)
- Testovací kritérium:

$$T = \frac{\varrho\sqrt{n-2}}{\sqrt{1-\varrho^2}}$$

- Testovací kritérium má Studentovo rozdělení s $n - 2$ stupni volnosti. Rozhodnutí o výsledku testu se provede na základě vypočtené p-value.
- Předpoklady:
  - Vstupní data obou vektorů musí být normálně rozdělená.

10.4 Pearsonův korelační koeficient

- Funkce v matlabu: corrcoef
- [R,P,RLO,RUP]=corrcoef(X,'alpha')
  - X vstupní data, X je matice, kde řádky jsou naměřená data a sloupce proměnné.

    Pokud jsou pouze 2 proměnné, může být nahrazeno corrcoef(x,y)
  - Alpha hladina významnosti
- Výsledek
  - R korelační matice
  - P hodnota pvalue testující hypotézu, že proměnné jsou nezávislé
  - RLO dolní intervalový odhad korelace mezi dvěma proměnnýma
  - RUP horní intervalový odhad korelace mezi dvěma proměnnýma

10.4 Pearsonův korelační koeficient

- Mějme naměřeny následující data, určete na hladině významnosti 1 %, zda jsou nezávislá
- (data byla vygenerována z rovnoměrného rozdělení)

- x = 0.8147 0.9058 0.1270 0.9134 0.6324 0.0975 0.2785 0.5469 0.9575 0.9649
- y = 0.1576 0.9706 0.9572 0.4854 0.8003 0.1419 0.4218 0.9157 0.7922 0.9595

- [R,P,RLO,RUP]=corrcoef(x,y,'alpha',0.01)

- $R = \begin{bmatrix} 1.000 & 0.2682 \\ 0.2682 & 1 \end{bmatrix}$ korelace mezi vektorem x a y je rovna 0.2682

- $P = \begin{bmatrix} 1.000 & 0.4537 \\ 0.4537 & 1 \end{bmatrix}$ pvalue na hypotézu, že vektory x a y jsou nezávislé je pval=0.4537

- $RLO = \begin{bmatrix} 1.000 & -0.6035 \\ -0.6035 & 1 \end{bmatrix}$ dolní mez korelačního koeficientu je rovna -0.6035

- $RUP = \begin{bmatrix} 1.000 & 0.8479 \\ 0.8479 & 1 \end{bmatrix}$ horní mez korelačního koeficientu je roven 0.8479
- Nezamítáme hypotézu H0, že data jsou nezávislá.

10.4 Pearsonův korelační koeficient

- Př. Z kapitoly o kovarianci na závislá data. Otestujte na hladině významnosti 5 %, že data jsou nezávislá.
  - >> a=unifrnd(0,1,1,1000);
  - >> b=a+0.1*unifrnd(0,1,1,1000);
  - [R,P,RLO,RUP]=corrcoef(a,b)

R $=\begin{pmatrix} 1 & 0.9946 \\ 0.9946 & 1 \end{pmatrix}$ P $=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

RLO$=\begin{pmatrix} 1 & 0.9939 \\ 0.9939 & 1 \end{pmatrix}$ RUP $=\begin{pmatrix} 1 & 0.9953 \\ 0.9953 & 1 \end{pmatrix}$

  - Zamítáme hypotézu H0, že data jsou nezávislá. (Hypotézu bychom zamítali, i kdybychom měli pouze 5 naměřených dat.

10.5 Spearmanův korelační koeficient

- Spearmanův korelační koeficient se používá, jestliže vstupní data mohou nabývat spojitých hodnot, a není splněn předpoklad o jejich normálním rozdělení.
- Mějme náhodný výběr $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ z dvourozměrného rozdělení. Označme $R_{X_1}, R_{X_2}, \dots, R_{X_n}$ pořadím veličin $X_1, X_2, \dots, X_n$. Obdobně označme $R_{Y_1}, R_{Y_2}, \dots, R_{Y_n}$ pořadím veličin $Y_1, Y_2, \dots, Y_n$.
  - V rámci testu se porovnává pořadí R_{X_i} a R_{Y_i}.
  - V případě nezávislosti vektorů X a Y bude jejich pořadí zcela náhodné.
  - Opačně v případě závislosti vektorů X a Y se například při vzrůstajícím x bude zvyšovat i hodnota y.

10.5 Spearmanův korelační koeficient

- Spearmanův korelační koeficient je definován

$$r_s = 1 - \frac{6}{n \cdot (n^2 - 1)} \sum_{i=1}^{n} \left(R_{X_i} - R_{Y_i}\right)^2$$

  – Při shodném pořadí nabývá koeficient r_s maximální hodnoty 1; při zcela opačném minimální hodnoty -1.

  – Je-li hodnota Spearmanova koeficientu $r_s = 0$, pořadí veličin jsou náhodně zpřeházená.

10.5 Spearmanův korelační koeficient

- Korekce Spearmanova koeficientu
  - Je-li velké množství naměřených hodnot shodných, je třeba provést korekci Spearmanova koeficientu.
  - Spearmanův koeficient se vypočte:
  - $r_s = 1 - \frac{6}{n \cdot (n^2 - 1) - T_X - T_Y} \sum_{i=1}^{n} \left(R_{X_i} - R_{Y_i} \right)^2$, kde $T_X = \frac{1}{2} \sum ({t_x}^3 - t_x)$ a $T_Y = \frac{1}{2} \sum \left({t_y}^3 - t_y \right)$, kde t_x a t_y je rozsah těchto shod.

10.5 Spearmanův korelační koeficient

- Testuje se hypotéza:
  - H_0: X a Y jsou nezávislé náhodné veličiny
  - H_A: X a Y jsou závislé náhodné veličiny
- Testování závislosti / nezávislosti náhodných výběrů se určuje pomocí testového kritéria:
  - $r_s{}^* = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}}$,
  - kde $z_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálního rozdělení
- Hypotézu H_0 zamítáme, jestliže $|r_s| \geq r_s{}^*$

10.5 Spearmanův korelační koeficient

- Funkce v matlabu: corr

- [rho,pval]=corr(x,y,'type','Spearman')
  - X vstupní sloupcový vektor
  - Y výstupní sloupcový vektor
  - Type druh korelačního koeficientu
    - Pearson defaultně, Pearsonův korelační koeficient
    - Spearman Spearmanův korelační koeficient

  - Rho korelační koeficient
  - Pval pvalue na základě pvalue lze zjistit nezávislost vektorů

10.5 Spearmanův korelační koeficient

- x = 0.4387 0.3816 0.7655 0.7952 0.1869 0.4898 0.4456 0.6463 0.7094 0.7547
- y = 0.2760 0.6797 0.6551 0.1626 0.1190 0.4984 0.9597 0.3404 0.5853 0.2238

- Numerický výpočet
  - Pořadí:
  - x: 3 2 9 10 1 5 4 6 7 8
  - y: 4 9 8 2 1 6 10 5 7 3
  - $\left(R_{X_i} - R_{Y_i}\right)^2$ 1 49 1 64 0 1 36 1 0 25
  - $\sum_{i=1}^{n}\left(R_{X_i} - R_{Y_i}\right)^2$ 178
  - Rho=1-$\frac{6\cdot 178}{10\cdot 99} = -0.0788$

- matlab
  - >> [rho,pval]=corr(x',y','type','Spearman')
  - rho = -0.0788
  - pval = 0.8380
- Nezamítáme hypotézu H0, že data jsou nezávislá.

11 – Regresní analýza

- 11.1 Lineární regrese
- 11.2 Verifikace modelu
- 11.3 Polynomiální regrese
- 11.4 Nelineární regrese

- Regrese umožňuje odhadovat hodnotu jisté spojité náhodné veličiny na základě znalosti (naměřených dat) jiných nezávislých veličin.
- Regresní analýza hledá funkční závislost mezi vysvětlující a vysvětlovanou proměnnou.

- Rozdíl mezi testováním hypotéz v kapitole 8 a regresní analýzou
  - Testování hypotéz – vstupem kategoriální výsledky
    - Například teplota 20, 30, 40, 50 °C, přidáno 1, 3, 10, 100 mg/l látky, porovnává metodu A, B a C
    - Nelze odhadnout výsledek pro 25, 35, 45 °C; pro 2, 5, 20 mg/l přidané látky; pro něco mezi metodou A a B
    - Pro dané kategorie máme obvykle více než 1 vstupní hodnotu.
  - Regresní analýza – vstupem spojité výsledky
    - Například přidal jsem 1, 2, π, 6, 12, 16 mg látky do roztoku. Naměřil jsem odpor 20, 24, 28, 35, 41, 45 Ω.
    - Nelze porovnávat vliv něco mezi metodou A a metodou B
    - Lze odhadnout výsledek pro libovolnou proměnnou ve spojitém rozmezí naměřených hodnot
    - Pro každou danou naměřenou hodnotu ze spojité n.v (nezávislá proměnná) máme obvykle 1 závislou hodnotu

11.1 Lineární regrese

- 11.1.1 Úvod do lineární regrese
- 11.1.2 Lineární regrese
- 11.1.3 Matematické odvození
- 11.1.4 Maticový způsob výpočtu

11.1.1 Úvod do lineární regrese

- $y = f(x)$
  - Matematická analýza určí množinu x z definičního oboru funkční hodnotu y.
    - Př. $y = x^3$, jestliže $y = 27, potom\ x = 3$
  - Ve statistice máme pro některá x naměřené hodnoty y a chceme odhadnout, jaká bude funkční hodnota pro změněné x.
  - Statistika – stochastická záležitost, pro jednu hodnotu x můžeme při opakovaných měřeních zjistit různé funkční hodnoty y.
- Příklad stochastických výsledků
  - Velikost hektarových výnosů plodiny v závislosti na nadmořské výšce,
  - Porovnání výšky a váhy člověka
  - Porovnání výšky platu otce a syna

- Naměřené hodnoty y jsou zatíženy chybou, snaha o proložení určitou funkční závislostí, která by minimalizovala součet kvadrátů chyby (obdoba součtu čtverců, rozptylu).
- Někdy nazývaná metoda nejmenších čtverců.

11.1.1 Úvod do lineární regrese

- Vstupem do úlohy
  - n počet naměřených dat
  - $[x_i, y_i]$ naměřené hodnoty
- Lineární regrese $\hat{Y} = ax + b$
  - a, b regresní koeficienty
  - x nezávislá proměnná, vysvětlující proměnná, regresor
  - y závislá proměnná, vysvětlovaná proměnná, regresand
- Polynomiální regrese

$$\hat{Y} = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$$

- Obecná funkční závislost - například

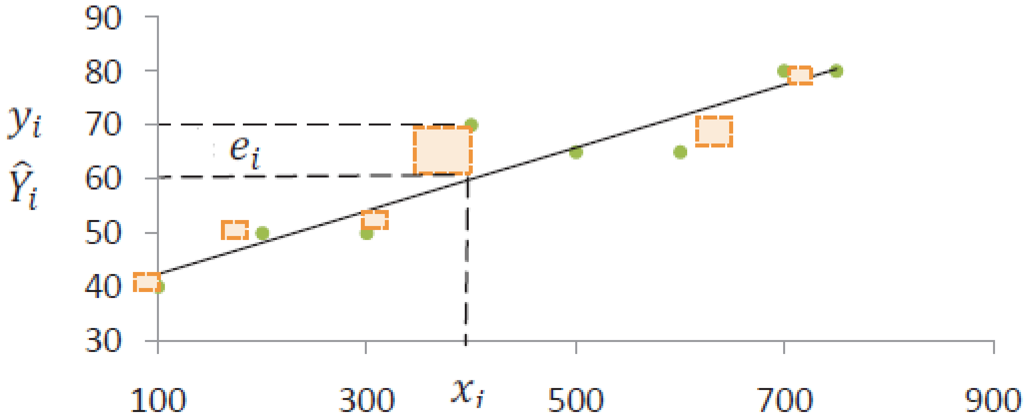
$$\hat{Y} = \sin(ax + b)$$

11.1.2 Lineární regrese

- Lineární regresní model
  - Matematická analýza: $y = ax + b$
  - Statistika $y_i = ax_i + b + \varepsilon_i$
  - ε_i je náhodná složka i-tého měření

- Hledáme parametry a a b tak, abychom minimalizovali sumu kvadrátů náhodných složek ε_i.
  - Metodě se někdy říká metoda nejmenších čtverců
  - Souvislost s rozptylem $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - Nejmenší čtverce $\varphi = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}\left(y_i - \widehat{Y}_i\right)^2$

11.1.2 Lineární regrese



- Model: $\hat{Y} = ax + b$
  - x_i naměřená hodnota nezávislé proměnné
  - $\hat{Y}_i$ odhadovaná hodnota
  - y_i pozorovaná hodnota
  - a, b parametry lineárního regresního modelu
  - e_i reziduum, rozdíl pozorované a odhadované hodnoty
- Minimalizujeme součet čtverců reziduí.

$$\varphi = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} \left(y_i - \hat{Y}_i\right)^2$$

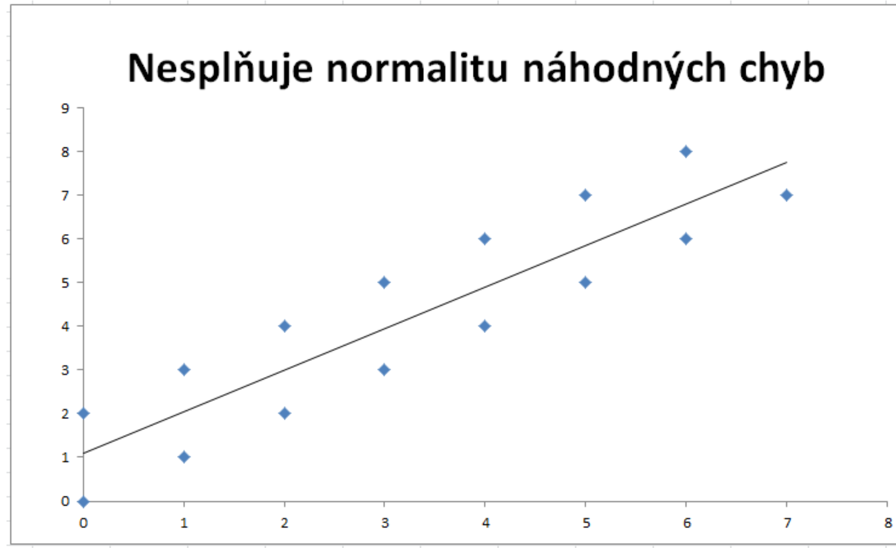
11.1.2 Lineární regrese

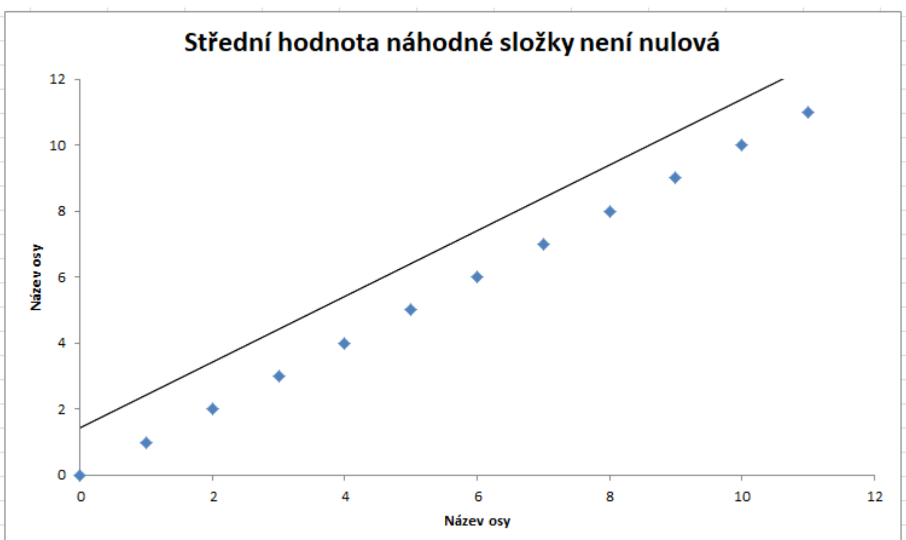
- Předpoklady:
  - 1) náhodné chyby ε_i mají normální rozdělení
  - 2) $E(\varepsilon_i) = 0$ střední hodnota náhodné složky je nulová.
  - 3) $D(\varepsilon_i) = \sigma^2$ rozptyl náhodné složky je konstantní
  - 4) navržený model nesmí být lineárně závislý.
    - Blíže viz maticový zápis modelu.
    - Např: $y = ax + bx + c$, nelze jednoznačně vyčíslit parametry a a b, protože jsou lineární závislé

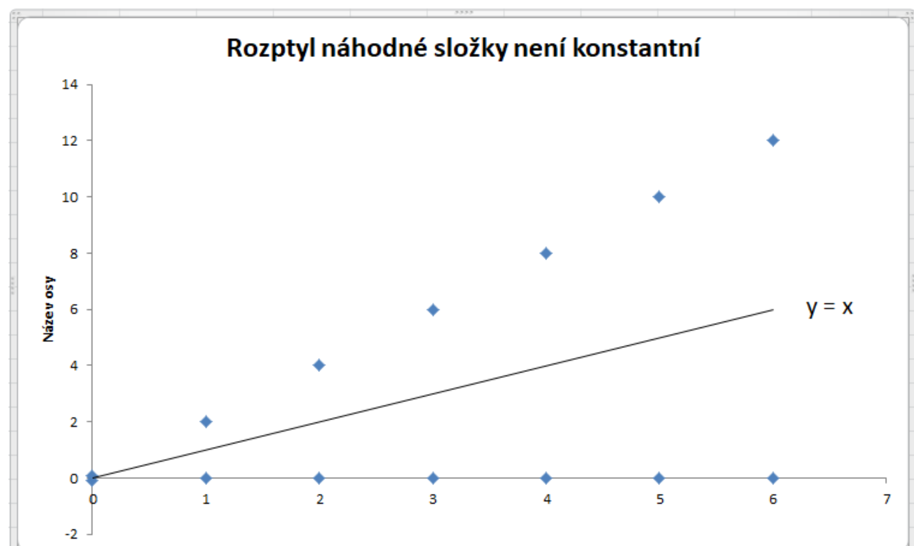
11.1.2 Lineární regrese

Předpoklady:

1. Nesplňuje normalitu náhodných chyb
2. Nesplňuje střední hodnotu náhodné složky rovnu 0
3. Nesplňuje konstantnost rozptylu







11.1.2 Lineární regrese

- Minimalizujeme součet čtverců reziduí.

$$\varphi = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} \left(y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} (y_i - ax_i - b\,)^2$$

- Suma se derivuje jako součet: $(\sum_i f_i(x))' = \sum_i f_i{}'(x)$
- Hledáme minimum funkce o dvou neznámých a a b.
  - Parciální derivace podle proměnných musí být rovny 0
  - Suma se derivuje jako součet: $(\sum_i f_i(x))' = \sum_i f_i{}'(x)$

$$\frac{\partial \varphi}{\partial a} = -2\sum_{i=1}^{n} (x_i \cdot (y_i - ax_i - b)) = 0$$

$$\frac{\partial \varphi}{\partial b} = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

  - Z těchto rovnice chceme vyjádřit a a b.

11.1.2 Lineární regrese

- Vzorce lineární regrese:

$$a = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x}) \cdot y_i\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

11.1.2 Lineární regrese

- Př. Nalezněte parametry lineární regrese pro následující data.
  - >> x=[10,20,25,30,35,40,45,50,60,70,80,85,90];
  - >> y=[7,14,14,19,22,21,27,29,32,39,43,48,48];

  - prum_x=mean(x) prum_x = 49.2308
  - prum_y=mean(y) prum_y = 27.9231

  - citatel=0; jmenovatel=0;
  - for i=length(x)

```
    citatel=citatel+(x(i)-prum_x)*y(i);
    jmenovatel=jmenovatel+(x(i)-prum_x).^2;
end
```

  - citatel= 3.8896e+03 jmenovatel = 8.0923e+03

  - a=citatel/jmenovatel a = 0.4807
  - b=prum_y-a*prum_x b = 4.2600
- $y = 0.4807\mathrm{x} + 4.26$

11.1.3 Matematické odvození

- 1) $$\frac{\partial \varphi}{\partial a} = -2\sum_{i=1}^{n}(x_i \cdot (y_i - ax_i - b)) = 0$$
$$\frac{\partial \varphi}{\partial b} = -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0$$
- 2) Druhou rovnici podělím (-2) a dále upravím:

$$\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - n \cdot b = 0$$

$$n \cdot b = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i$$

$$n \cdot b = n\bar{y} - an\bar{x}$$

$$b = \bar{y} - a\bar{x}$$

11.1.3 Matematické odvození

- 3) dosadím výsledek z bodu 2 do druhé rovnice

$$\frac{\partial \varphi}{\partial a} = -2\sum_{i=1}^{n}(x_i \cdot (y_i - ax_i - b)) = 0$$

$$\sum_{i=1}^{n}(x_i \cdot (y_i - ax_i - b)) = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a\sum_{i=1}^{n} x_i{}^2 - b\sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a\sum_{i=1}^{n} x_i{}^2 - (\bar{y} - a\bar{x}) \cdot \left(\sum_{i=1}^{n} x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a\sum_{i=1}^{n} x_i{}^2 - \bar{y}\sum_{i=1}^{n} x_i + a\bar{x}\sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - \bar{y}\sum_{i=1}^{n} x_i = a\sum_{i=1}^{n} x_i{}^2 - a\bar{x}\sum_{i=1}^{n} x_i$$

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i{}^2 - \bar{x}\sum_{i=1}^{n} x_i}$$

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{\sum_{i=1}^{n} y_i}{n} \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i{}^2 - \frac{\sum_{i=1}^{n} x_i}{n} \cdot \sum_{i=1}^{n} x_i}$$

$$a = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i{}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i{}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

11.1.3 Matematické odvození

- Pokračování odvození

$$a = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \frac{n \cdot \sum_{i=1}^{n}\left((x_i - \bar{x}) \cdot y_i\right)}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x}) \cdot y_i\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Výsledek

$$a = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x}) \cdot y_i\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

11.1.4 Maticový způsob výpočtu

- Platí pro tuto kapitolu:
  - a malé písmeno netučné skalár
  - $\boldsymbol{a}$ **malé písmeno tučné** **vektor**
  - $\boldsymbol{F}$ **velké písmeno tučné** **matice**
- Lze odvodit maticový vzorec, pomocí kterého lze zjistit parametry.

$$\boldsymbol{b} = (\boldsymbol{F}'\boldsymbol{F})^{-1} \cdot \boldsymbol{F}' \cdot \boldsymbol{y}$$

11.1.4 Maticový způsob výpočtu

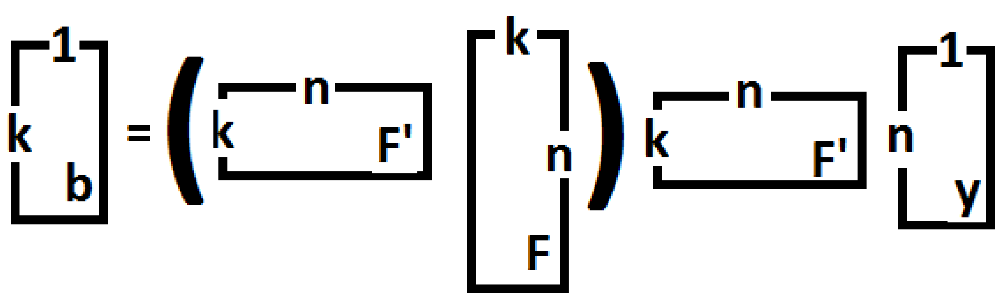
$$\boldsymbol{b} = (\boldsymbol{F}'\boldsymbol{F})^{-1} \cdot \boldsymbol{F}' \cdot \boldsymbol{y}$$

Model je například ve tvaru $y = ax^2 + bx + c$ k=3 parametrů

n= počet naměřených dat

Obtížný výpočet inverzní matice, proto se u složitějších modelů využívá numerika

- $\boldsymbol{b}$ vektor výsledných parametrů $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$
- $\boldsymbol{F}$ matice modelu nezávislých proměnných $\begin{bmatrix} {x_1}^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ {x_n}^2 & x_n & 1 \end{bmatrix}$
- $\boldsymbol{y}$ vektor naměřených hodnot (závislých proměnných) $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$



11.1.5 Výpočet v Matlabu

- Matlab funkce: fitlm
  - NLM=fitlm(x,y,modelfun,další parametry)
  - x,y naměřená data
  - x může být i matice, kde ve sloupcích jsou jednotlivé nezávislé proměnné
    například: y=f(výška, věk)
  - modelfun navržený typ modelu
    - Pokud není uvedeno, uvažuje se lineární model
    - V modelu nemusí být nezávislé proměnné x a y
    - 'constant' $y = a$
    - 'linear' přednastaveno $y = a + bx$ nebo $y = a + bx_1 + cx_2$
    - 'interactions' $y = a + bx$ $y = a + bx_1 + cx_2 + dx_1x_2$
    - 'purequadratic' $y = a + bx + cx^2$ $y = a + bx_1 + cx_2 + dx_1^2 + ex_2^2$
    - 'quadratic' $y = a + bx + cx^2$ $y = a + bx_1 + cx_2 + dx_1x_2 + ex_1^2 + fx_2^2$
  - Například fitlm(x,y,'linear') fitlm(tab,y,’linear’)

11.1.5 Výpočet v Matlabu

- Výsledky v matlabu ve formě tabulky

  – Linear regression model — uvedení matematického tvaru modelu
  – Estimated Coefficients — tabulka, kde řádky jsou parametry modelu a sloupce:
    - Estimate — odhad parametru
    - SE — směrodatné odchylky parametrů
    - tStat — výsledek testu, že daný parametr je roven 0. (směrodatná odchylka splňuje Studentovo rozdělení s n-počet param. st. vol.)
    - pvalue — pvalue hypotézy, že daný parametr je roven 0. Jestliže je pvalue menší než hladina významnosti alfa, potom je parametr důležitý, jinak lze upravit model bez uvedeného parametru.

  – Number of observations — počet pozorování
  – Error degrees of freedom — počet stupňů volnosti
  – R-Squared — koeficient determinace
  – Adjusted R-Squared — modifikovaný koeficient determinace
  – F statistic — výsledek stability modelu a jeho pvalue testuje se H0: model=konst (potom $pval > \propto$)

11.1.5 Výpočet v Matlabu

- Mějme naměřená data:>> x=[1,2,3,4,5,6,7,8]‘ a y=[1,2.01,3.04,4.1,5.15,6.2,7.3,8.4]‘. Proložte data parabolou.
- Načtení dat:
  - >> x=[1,2,3,4,5,6,7,8]';
  - >> y=[1,2.01,3.04,4.1,5.15,6.2,7.3,8.4]';
- Výsledky:
  - >> fitlm(x,y,'quadratic')
  - Linear regression model: y ~ 1 + x1 + x1^2
  - Estimated Coefficients:

	Estimate	SE	tStat	pValue	
(Intercept)	-0.0042857	0.013409	-0.3196	0.76219	parametr není třeba
x1	0.99583	0.0068367	145.66	2.8932e-10	nutný parametr
x1^2	0.0067857	0.00074154	9.1508	0.00026121	nutný parametr

  - Number of observations: 8, Error degrees of freedom: 5 — 8 měření, 5 stupňů volnosti
  - Root Mean Squared Error: 0.00961 — součet čtverců chyb
  - R-squared: 1, Adjusted R-Squared 1 — koeficient determinace – míra kvality modelu
  - F-statistic vs. constant model: 2.54e+05, p-value = 3.04e-13 — ověření kvality modelu, porovnává hypotézu shody modelu s konstantním modelem

- Výsledek je ve tvaru: $y = -0.0042857 + 0.99583x + 0.0067857x^2$
- Mohla by následovat ověření pomocí nelineární regrese ve tvaru $y = ax_1 + bx_2{}^2$. Viz kapitola 11.4

11.2 Verifikace modelu

- Po vypočtení parametrů lineární regrese se můžeme ptát:
  - Byl zvolen vhodný typ regresní funkce?
    - Například funkci $y = a \cdot x^2 + b \cdot x + c$ není vhodnější proložit pouze přímkou?
  - Jsou všechny parametry v modelu nutné?
  - Jak dokonale model charakterizuje naměřené výsledky?
- Využijeme výsledků příkladu z kapitoly 11.1.5 a dalších zjištěných výsledků ☺

11.2 Verifikace modelu

- 11.2.1 – F-test – ověřuje se správnost použitého typu modelu
- 11.2.2 – Intervalový odhad regresních koeficientů
- 11.2.3 – Testy hypotéz o koeficientech regresní funkce
- 11.2.4 – Koeficient determinace

```
Linear regression model:
    y ~ 1 + x1 + x1^2
Estimated Coefficients:
                   Estimate        SE          tStat       pValue

    (Intercept)    -0.0042857     0.013409     -0.3196       0.76219
    x1                0.99583    0.0068367      145.66    2.8932e-10
    x1^2            0.0067857   0.00074154      9.1508    0.00026121

Number of observations: 8, Error degrees of freedom: 5
Root Mean Squared Error: 0.00961
R-squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 2.54e+05, p-value = 3.04e-13
```

11.2.2 → SE (x1)
11.2.3 → tStat (x1^2)
11.2.4 → R-squared
11.2.1 → F-statistic

11.2.1 Verifikace modelu - Ftest

- **Řešení:** **F-statistic vs. constant model:**
- Pomocí Ftestu se ověřuje správnost použitého modelu
  - $H_0: všechny\ parametry = 0$ (kromě konstantního parametru)
  - $H_1: některý\ z\ parametrů\ \neq 0$
  - Obvykle se hypotéza H_0 zamítá, pval je velmi malá
- Výsledek se zapisuje do Anova tabulky

Variabilita	Součet čtverců	Stupňů volnosti	Rozptyl	F poměr
Model	$SS_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2$	k	$\frac{SS_{\hat{Y}}}{k}$	$\frac{\frac{SS_{\hat{Y}}}{k}}{\frac{SS_e}{n-(k+1)}}$
Reziduální	$SS_e = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2$	$n-(k+1)$	$\frac{SS_e}{n-(k+1)}$	
Celkový	$SS_Y = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$		

11.2.1 Verifikace modelu - Ftest

- F poměr splňuje Fisher-Snedecorovo rozdělení s $(k, n-(k+1))$ stupni volnosti
- $pvalue = 1 - F_0(x_{obs})$

- Obvykle vychází, že hypotézu H0 zamítáme, tzn. některý z parametrů (krom konstanty) je odlišný od 0.
  - Například u příkladu v kapitole 11.1.5 je $pval = 3 \cdot 10^{-13}$

```
Linear regression model:
    y ~ 1 + x1 + x1^2
Estimated Coefficients:
                   Estimate        SE           tStat       pValue

    (Intercept)    -0.0042857      0.013409     -0.3196       0.76219
    x1                0.99583      0.0068367     145.66     2.8932e-10
    x1^2            0.0067857      0.00074154    9.1508     0.00026121

Number of observations: 8, Error degrees of freedom: 5
Root Mean Squared Error: 0.00961
R-squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 2.54e+05, p-value = 3.04e-13
```

11.2.2 Intervalový odhad regresních koeficientů

- Řešení: **tStat** **pValue**
- Ptáme se, zda některý z parametrů může být roven 0.

  Funkce fitlm řeší hypotézu H0: parametr = 0; oproti H1: parametr ≠ 0.
- Pvalue < α zamítáme hypotézu, že parametr je roven 0

- Pro stanovení intervalu odhadů regresních koeficientů je nutné stanovit jejich rozptyl. Ten lze stanovit pomocí následujících vzorců:
  - $SS_e = {e_i}^2 = \sum_{i=1}^{n}\left(y_i - \widehat{Y}_i\right)^2$
  - ${S_e}^2 = \frac{\sum_{i=1}^{n} {e_i}^2}{n-(k+1)} = \frac{SS_e}{n-(k+1)} = \frac{\sum_{i=1}^{n}(y_i-\widehat{Y_i})^2}{n-(k+1)}$
  - ${S_a}^2 = \frac{{S_e}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
  - ${S_b}^2 = {S_e}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

11.2.2 Intervalový odhad regresních koeficientů

- Intervalový odhad regresních koeficientů lze zjistit pomocí následujících vzorců:

$$\left\langle a - t_{1-\frac{\propto}{2}} \cdot S_a ; a + t_{1-\frac{\propto}{2}} \cdot S_a \right\rangle$$

$$\left\langle b - t_{1-\frac{\propto}{2}} \cdot S_b ; b + t_{1-\frac{\propto}{2}} \cdot S_b \right\rangle$$

- kde $t_{1-\frac{\propto}{2}}$ je kvantil Studentova rozdělení s $n - počet\ param$. st. vol..

11.2.2 Intervalový odhad regresních koeficientů

- Příklad 11.1.5 pokračování
- Určete 95% IS pro konstantní parametr.
  - Odhad parametru -0.0042857
  - Směrodatná odchylka parametru 0.013409
  - Výpočet: $\left\langle a - t_{1-\frac{\propto}{2}} \cdot S_a; a + t_{1-\frac{\propto}{2}} \cdot S_a \right\rangle$

$$\left\langle -0.0042857 - t_{1-\frac{\propto}{2}}(5) \cdot 0.013409; -0.0042857 + t_{1-\frac{\propto}{2}}(5) \cdot 0.0013409 \right\rangle$$

$$t_{1-\frac{\propto}{2}}(5) = 2.5705$$

95% IS konstantního parametru je: $\langle -0.03875; 0.03182 \rangle$

```
Linear regression model:
    y ~ 1 + x1 + x1^2
Estimated Coefficients:
                   Estimate        SE           tStat       pValue
    (Intercept)   -0.0042857     0.013409     -0.3196      0.76219
    x1             0.99583       0.0068367    145.66       2.8932e-10
    x1^2           0.0067857     0.00074154   9.1508       0.00026121

Number of observations: 8, Error degrees of freedom: 5
Root Mean Squared Error: 0.00961
R-squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 2.54e+05, p-value = 3.04e-13
```

11.2.3 – Testy hypotéz o koeficientech regresní funkce

- Výběrovou statistiku $\frac{b-\beta}{s_b} \sim N(0,1)$ lze využít pro testování hypotéz o koeficientech regresní funkce.
  - b_i naměřená hodnota parametru
  - β_i hodnota parametru se kterým naměřenou hodnotu porovnáváme (obvykle je nula)
  - s_{b_i} směrodatná odchylka parametru.
- Testujeme hypotézu:
  - $H_0\colon \mathrm{a} = \alpha \quad H_0\colon \mathrm{a} \neq \alpha$
  - $H_0\colon b = \beta \quad H_0\colon b \neq \beta$
- Testovací kritérium:
  - $T = \frac{a-\alpha}{s_a} \qquad\qquad T = \frac{b-\beta}{s_b}$
  - Testovací kritérium odpovídá Studentovu rozdělení s $n - poč\ param.$ stupňů volnosti. Hypotézu H0 nezamítáme, jestliže $-t_{1-\frac{\propto}{2}}(st.v.) < T < t_{1-\frac{\propto}{2}}(st.v.)$
  - Obdobně lze řešit i jednostranné hypotézy

11.2.3 – Testy hypotéz o koeficientech regresní funkce

- Př. Otestujte z příkladu 11.1.5 na hladině významnosti 5 %, zda parametr x^2 může být roven 0.008.

- Testujeme hypotézu $y = ax^2 + bx + c$:
  - $H_0\!: \mathrm{a} = 0.008 \qquad H_0\!: \mathrm{a} \neq 0.008$
- Testovací kriterium $T = \frac{a-\alpha}{s_a} = \frac{0.0067857-0.008}{0.00074154} = -1.6375$
  - $t_{1-\frac{\propto}{2}}(5) = 2.5706$
- Hypotézu H0 nezamítáme.

```
Linear regression model:
    y ~ 1 + x1 + x1^2
Estimated Coefficients:
                   Estimate        SE          tStat       pValue

    (Intercept)    -0.0042857      0.013409    -0.3196       0.76219
    x1                0.99583     0.0068367     145.66    2.8932e-10
    x1^2            0.0067857    0.00074154     9.1508    0.00026121

Number of observations: 8, Error degrees of freedom: 5
Root Mean Squared Error: 0.00961
R-squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 2.54e+05, p-value = 3.04e-13
```

11.2.4 Koeficient determinace

- Řešení: R-squared, Adjusted R-Squared

- Pro určení síly závislosti se vychází z poměru proložených ku naměřeným součtu čtverců.
- Koeficient determinace je definován:

$$r^2 = 1 - \frac{SS_e}{SS_Y} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Koeficient determinace r^2 udává kvalitu regresního modelu, neboli jaká část součtu čtverců (rozptylu) je popsána modelem, a jak velká zbývající část je nevysvětlena.
- Koeficient nabývá hodnot od 0 k 1, přičemž:
  - $r^2 = 1$ data jsou přesně proloženy funkcí
  - $r^2 = 0$ proložení je zcela nevhodné
  - Obvykle model je uznán za vhodný, jestliže $r^2 > 0.8$

11.3 Polynomiální regrese

- Analyticky obtížně zjistitelné parametry – nutno využít numeriky
- V přednáškách bude uvedeno pouze softwarové řešení
- Matlab polyfit(x,y,n)
  - x sloupcový vektor vstupních dat
  - y sloupcový vektor vstupních dat
  - n stupeň polynomiální regrese

11.3 Polynomiální regrese

$$p = polyfit(x, y, n)$$

- Funkce vypočte vektor parametrů polynomiální regrese stupně n:

$$y = a_1 x^n + a_2 x^{n-1} + \cdots + a_{n+1}$$

- Pro bližší stanovení parametrů doporučuji použít model nelineární regrese. Pomocí funkce lze stanovit i další statistické parametry.

11.4 Nepolynomiální regrese

- Pomocí nelineární regrese lze numericky stanovit parametry uživatelem navrženého modelu.
- Rozšíření polynomiálního modelu o další funkce.
- Výpočet probíhá numericky. Nutno stanovit vstupní vektor přibližného řešení
  - Například funkce $y = \sin(ax + b)$ má pro
    - parametr a řešení $y = \sin(ax)$, ale také $y = -\sin(-ax)$
    - parametr b řešení $b = \beta + 2k\pi; k\epsilon Z$
    - z toho vyplývá nejednoznačnost řešení.

Jedná se o mocný numerický nástroj. Vlivem numerického výpočtu je však nutné mít u výsledků inženýrský náhled!

11.4 Nepolynomiální regrese

- Matlab funkce: fitnlm
  - NLM=fitnlm(x,y,modelfun,beta0,další parametry)
  - x,y naměřená data
  - modelfun navržený typ modelu;
    - Funkce se zapisuje:
    - Funkce pomocí znaku @ - např. @(b,x)b(1) + b(2)*x.^b(3) ,
    - Funkce je pro neznámý vektor parametrů **b** proměnné x
    - Pomocí řetězce označující rovnici 'y ~ b0+b1*sin(b2*X)'
  - beta0 vstupní vektor, ze kterého jsou numericky počítány optimální parametry
    - Řešení úlohy záleží na volbě počáteční iteraci beta0

11.4 Nepolynomiální regrese

fitnlm – výsledky

- Nonlinear regression model - uvedení tvaru modelu
- Estimated Coefficients – tabulka, kde řádky jsou parametry modelu a sloupce:
  - Estimate odhad parametru
  - SE směrodatné odchylky parametrů
  - tStat výsledek testu, že daný parametr je roven 0
  - pvalue pvalue hypotézy, že daný parametr je roven 0, jestliže je pvalue malá parametr je důležitý, jinak upravíme model bez uvedeného parametru.
- Number of observations počet pozorování
- Error degrees of freedom počet stupňů volnosti
- R-Squared koeficient determinace
- Adjusted R-Squared modifikovaný koeficient determinace
- F statistic výsledek ověření stability modelu a jeho pvalue

- Obdobné výsledky jako u funkce fitlm

11.4 Nepolynomiální regrese

- Příklad: Mějme naměřená následující data
- x=[1,2,3,4,5,6,7,8,9,10]';
- y=[0,12,50,92,168,291,435,639,889,1203]';

- Určete polynomiální model ve tvaru $y = ax^3 + bx^2 + cx + d$. Pokud některý z parametrů nebude nutný, vylučte ho a řešte model znovu bez vyloučeného parametru.

- Použijeme funkci fitnlm

- 1) načteme data
  - \>> x=[1,2,3,4,5,6,7,8,9,10]';
  - \>> y=[0,12,50,92,168,291,435,639,889,1203]';
- 2) budeme uvažovat navržený model:
  - \>> modelfun=@(b,x)b(4)*x.^3+b(3)*x.^2 + b(2)*x+b(1); definice prokládající funkce
  - \>> beta0=[1,1,1,1]; počáteční vektor $y = 1 + x + x^2 + x^3$
  - \>> NLM=fitnlm(x,y,modelfun,beta0) příkaz výpočtu

11.4 Nepolynomiální regrese

- 2) výsledky
  - Nonlinear regression model:
  - y ~ b4*x^3 + b3*x^2 + b2*x + b1
  - Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	-6.7667	8.6708	-0.7804	0.46482
b2	4.5627	6.4991	0.70205	0.50895
b3	0.90851	1.3406	0.67771	0.52321
b4	1.073	0.080387	13.348	1.09e-05

- Parametry b1, b2 a b3 vycházejí statisticky nevýznamné, budou vynechány
- Úloha bude znovu vypočtena znovu bez parametrů b1, b2 a b3

11.4 Nepolynomiální regrese

- 3) budeme uvažovat upravený model:
  - >> modelfun=@(b,x)b(1)*x.^3;
  - >> beta0=[1];
  - >> NLM=fitnlm(x,y,modelfun,beta0)
- 4) výsledky:
  - Nonlinear regression model: y ~ b1*x^3
  - Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	1.2225	0.011224	108.92	2.3523e-15

  - Number of observations: 10, Error degrees of freedom: 9
  - Root Mean Squared Error: 15.8
  - R-Squared: 0.999, Adjusted R-Squared 0.999
  - F-statistic vs. zero model: 1.19e+04, p-value = 2.35e-15

- Model $y = x^3$ je možno použít.

11.4 Nepolynomiální regrese

- Pozor při přeparametrování modelu, vyskočí warningová hláška, že vypočtené parametry nemusí být správné.
- Například:
  - Warning: The Jacobian at the solution is ill-conditioned, and some model parameters may not be estimated well (they are not identifiable). Use caution in making predictions.
- Snažte se zjednodušit model odstraněním funkce, která nejvíce ovlivňuje výsledné proložení (např. nejvyšší polynom) a úlohu zopakujte

Stejný příklad jako na minulých slidech, špatně určený model:

```
x=[1,2,3,4,5,6,7,8,9,10]';
y=[0,12,50,92,168,291,435,639,889,1203]';
modelfun=@(b,x)b(7)*x.^6+b(6)*x.^5+b(5)*x.^4+b(4)*x.^3+b(3)*x.^2 + b(2)*x+b(1);
beta0=[1,1,1,1,1,1,1];
NLM=fitnlm(x,y,modelfun,beta0)
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	17.8	76.142	0.23377	0.8302
b2	-47.602	142.93	-0.33304	0.76102
b3	39.178	94.09	0.41639	0.70512
b4	-11.768	29.051	-0.40506	0.71258
b5	2.1457	4.5682	0.4697	0.6706
b6	-0.17362	0.35387	-0.49064	0.65733
b7	0.0054167	0.010704	0.50606	0.64766

Number of observations: 10, Error degrees of freedom: 3
Root Mean Squared Error: 6
R-Squared: 1, Adjusted R-Squared 1
F-statistic vs. constant model: 7.09e+03, p-value = 2.59e-06

Stejný příklad jako na minulých slidech, špatně určený model:

```
x=[1,2,3,4,5,6,7,8,9,10]';
y=[0,12,50,92,168,291,435,639,889,1203]';
modelfun=@(b,x) b(6)*x.^5+b(5)*x.^4+b(4)*x.^3+b(3)*x.^2 +
b(2)*x+b(1);
beta0=[1,1,1,1,1,1];
NLM=fitnlm(x,y,modelfun,beta0)
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	-16	32.985	-0.48506	0.653
b2	19.088	49.938	0.38222	0.72174
b3	-6.329	24.98	-0.25336	0.81248
b4	2.6136	5.4428	0.4802	0.65617
b5	-0.14656	0.53466	-0.27412	0.79757
b6	0.0051282	0.019382	0.26459	0.8044

Number of observations: 10, Error degrees of freedom: 4
Root Mean Squared Error: 5.41
R-Squared: 1, Adjusted R-Squared 1
F-statistic vs. constant model: 1.04e+04, p-value = 2.56e-08

- Model je přeparametrován,
- pvalue jsou vysoké,
- Odstraněním parametru, který nejvíce ovlivňuje proložení se F statistika výrazně sníží.
- Skutečný model je ve tvaru $y = ax^3$

11.4 Nepolynomiální regrese

Mějme vygenerovaná data pomocí funkce y=sin(x)+0.01*normrnd(0,1) a proložte je sinusovkou.

Případ 1 – parametr beta0 je blízky skutečnému řešení

```
>> modelfun=@(b,x)b(1)+sin(b(2)*x+b(3));
>> beta0=[1,1,1]

>> NLM=fitnlm(x,y,modelfun,beta0)
```

- NLM = Nonlinear regression model: y ~ b1 + sin(b2*x + b3)
- Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	0.011174	8.1139e-12	1.3771e+09	2.3811e-146
b2	1	1.4266e-12	7.0096e+11	2.3033e-192
b3	8.8773e-11	9.4334e-12	9.4105	3.7441e-08

- Number of observations: 20, Error degrees of freedom: 17
- Root Mean Squared Error: 3.37e-11
- R-Squared: 1, Adjusted R-Squared 1
- F-statistic vs. constant model: 4.51e+21, p-value = 6.92e-177

Případ 2 – parametr beta0 je vzdálený od skutečného řešení

```
>> modelfun=@(b,x)b(1)+sin(b(2)*x+b(3));
>> beta0=[0,0,0]
```

numerický výpočet neví, zda u fce sin x je parametr + nebo -

```
>> NLM=fitnlm(x,y,modelfun,beta0)
```

NLM = Nonlinear regression model: y ~ b1 + sin(b2*x + b3)

Estimated Coefficients:

	Estimate	SE	tStat	pValue
b1	0.89096	0.2579	3.4546	0.0030275
b2	-0.10187	0.065162	-1.5634	0.13639
b3	-0.37851	0.70943	-0.53354	0.60057

Number of observations: 20, Error degrees of freedom: 17

Root Mean Squared Error: 0.756

R-Squared: 0.0522, Adjusted R-Squared -0.0593

F-statistic vs. constant model: 0.468, p-value = 0.634

- Jak lze zjistit správnost modelu:
  - Root Mean Squared Error pokud velké v porovnání s jiným model, tak špatně;
  - R-Squared pokud malé v porovnání s jiným modelem, tak špatné
  - F-statistic – velká hodnota špatně popsaný model
- Problémy z důvodu obtížného numerického výpočtu a hledání nikoliv lokálního, ale globálního minima

A to je vše

Děkuji za pozornost a přeji hodně úspěchu při zkoušce