Vybrané funkce v programu Scilab z oblasti pravděpodobnost a statistika

Pavla Pecherková, Ivan Nagy

15. dubna 2017

Tento materiál byl podpořen grantem

Obsah

1 Úvod **4**
- 1.1 Úvod do pravděpodobnosti a statistiky . . . 4
- 1.2 Úvod do Scilabu . . . 4

2 Kombinatorika **5**
- 2.1 Kombinace bez opakování . . . 6
- 2.2 Kombinace s opakováním . . . 7
- 2.3 Variace bez opakování . . . 8
- 2.4 Variace s opakováním . . . 9
- 2.5 Permutace . . . 10

3 Pravděpodobnost **11**
- 3.1 Statistická pravděpodobnost . . . 12
- 3.2 Klasická pravděpodobnost . . . 13
- 3.3 Podmíněná pravděpodobnost . . . 14
- 3.4 Nezávislost jevů . . . 15
- 3.5 Úplná pravděpodobnost . . . 16
- 3.6 Bayesův vzorec . . . 17

4 Popisná statistika **18**
- 4.1 Střední hodnota . . . 19
- 4.2 Rozptyl . . . 20
- 4.3 Směrodatná odchylka . . . 21
- 4.4 Modus . . . 22
- 4.5 Medián . . . 23
- 4.6 Kvantil . . . 24
- 4.7 Mezikvartilové rozpětí . . . 25
- 4.8 Variační rozpětí . . . 26
- 4.9 Centrální moment . . . 27
- 4.10 Obecný moment . . . 28
- 4.11 Kovariance . . . 29
- 4.12 Kovarianční matice . . . 31
- 4.13 Korelace . . . 33

5 Intervalové odhady **35**
- 5.1 Interval spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 . . . 35
- 5.2 Interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu σ^2 . . . 37
- 5.3 Interval spolehlivosti pro rozptyl σ^2 . . . 39
- 5.4 Interval spolehlivosti pro podíl π . . . 41

6 Testy hypotéz **43**
- 6.1 Testy jedné veličiny . . . 44
  - 6.1.1 Test hypotéz o střední hodnotě μ při známém rozptylu σ^2 . . . 44
  - 6.1.2 Test hypotéz o střední hodnotě μ při neznámém rozptylu σ^2 . . . 46
  - 6.1.3 Test hypotéz o podílu π . . . 48

6.1.4 Test hypotéz o rozptylu σ^2 . . . 50
6.1.5 Test mediánu (znaménkový test) . . . 51
6.1.6 Wilcoxonův test . . . 53
6.1.7 Test normality (triviální) . . . 54
6.1.8 χ^2 test normality . . . 55
6.1.9 KS test . . . 56
6.1.10 χ^2 test - test dobré shody . . . 57
6.2 Testy dvou veličin . . . 59
6.2.1 Test hypotéz o shodě dvou středních hodnot $\mu_1 = \mu_2$ při známém rozptylu σ_1^2, σ_2^2 . . . 59
6.2.2 Nezávislý T-test . . . 61
6.2.3 Párový T-test . . . 63
6.2.4 Test hypotéz o shodě dvou podílů $p_1 = p_2$ . . . 65
6.2.5 Test hypotéz o shodě dvou rozptylů $\sigma_1^2 = \sigma_2^2$ . . . 67
6.2.6 Mann-Whitneyův test . . . 68
6.2.7 Znaménkový test . . . 69
6.2.8 McNemarův test . . . 71
6.3 Testy více veličin . . . 72
6.3.1 Analýza rozptylu při jednoduchém třídění . . . 72
6.3.2 Barlettův test . . . 74
6.3.3 Scheffého test . . . 75
6.3.4 Analýza rozptylu při dvojném třídění . . . 76
6.3.5 Kruskal-Wallisův test . . . 79
6.3.6 Friedmanův test . . . 80
6.4 Testy nezávislosti . . . 81
6.4.1 χ^2 test - test nezávislosti . . . 81
6.4.2 Test nezávislosti výběrů - Pearsonův test . . . 83
6.4.3 Test nezávislosti výběrů - Spearmanův test . . . 85
6.4.4 Test pořadové nezávislosti prvků výběru . . . 87

7 Regresní analýza **89**
7.1 Lineární regrese . . . 90
7.2 Lineární predikce . . . 92
7.3 Vícenásobná lineární regrese . . . 93
7.4 Vícenásobná lineární predikce . . . 95
7.5 Polynomiální regrese . . . 96
7.6 Polynomiální predikce . . . 99
7.7 Exponenciální regrese . . . 100
7.8 Exponenciální predikce . . . 102
7.9 T-test korelačního koeficientu . . . 103
7.10 F-test poměru vysvětleného a nevysvětleného rozptylu pro predikci . . . 104
7.11 Test bělosti reziduí . . . 106
7.12 Test autoregrese residuí . . . 107

Kapitola 1

Úvod

1.1 Úvod do pravděpodobnosti a statistiky

Pravděpodobnost (někdy nazývána teorie pravděpodobnosti) je matematická disciplína, která se zabývá studiem zákonitostí v náhodných pokusech. První zmínky o pravděpodobnosti se objevují v pracích dvou významných matematiků, a to Blaise Pascala a Pierra de Fermata (oba Francouzi). První úloha, kterou tito matematici řešili, bylo vlastně řešení hráčské otázky, zda se vyplatí či nevyplatí vsadit na výhru za následujících podmínek: provedeme 24 hodů párem kostek, výhercem se staneme v případě, že dvě šestky padly alespoň 1x.....

1.2 Úvod do Scilabu

Kapitola 2

Kombinatorika

Kombinatorika je jednou z nejpoužívanějších částí matematiky. Zabývá se hledáním všech kombinací (jejich počtu), které lze vytvořit za určitých podmínek (konečné nebo spočitatelné hodnoty). Kombinatorikou se zabýval již v 17. století Blaise Pascal a Pierre de Fermat při studiu pravděpodobnosti. Termín kombinatorika byl poprvé použit v moderní matematice německým filosofem a matematikem Gottfriedem Wilhelme Leibnizem v jeho práci „Dissertation de Arte Combinatoria" - Disertační práce zabývající se uměním kombinací.....

Kombinatorika se využívá jak v mnoha teoretických oblastí, jako je pravděpodobnost, algebra, geometrii, tak jako aplikace v optimalizaci, fyziky atd.

2.1 Kombinace bez opakování

TEORIE

Kombinace bez opakování (dále jen kombinace) použijeme v případě, že vybíráme určitý počet objektů za předpokladu, že *nezáleží na pořadí* v jakém jsme je vybrali a *žádný objekt nemůžeme vybrat více než jedenkrát.* Vybíráme k-tice z n prvkové množiny $n \geq k$ a platí, že

$$C_k(n) = \underbrace{\binom{n}{k}}_{\substack{kombinační \\ číslo}} = \frac{n!}{k!\,(n-k)!}$$

SCILAB - `kombinace()`

```
[vysledek]=kombinace(k,n)
   k ... počet k-členných skupin
   n ... počet prvků
```

Příklad. Určete kolik dvojjazyčných slovníků je třeba vydat, aby byla zajištěna možnost vzájemného přímého překladu z ruského, anglického, německého a francouzského jazyka?

1. Teorie

   $C_2(4) = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6.$

2. Scilab

   ```
   [vysledek]=kombinace(2,4)

   vysledek=6
   ```

2.2 Kombinace s opakováním

TEORIE

Kombinace s opakováním použijeme v případě, že vybíráme určitý počet objektů za předpokladu, že *nezáleží na pořadí* v jakém jsme je vybrali a *objekt můžeme vybrat více než jedenkrát*. Vybíráme k-tice z n prvkové množiny a platí, že

$$C'_k(n) = \begin{pmatrix} (n+k-1) \\ k \end{pmatrix}$$

V tomto případě nemusí být $n \geq k$, ale naopak často je situace, kdy $n < k$.

SCILAB - `kombinace_s_opakovanim()`

```
[vysledek]=kombinace_s_opakovanim(k,n)
   k ... počet k-členných skupin
   n ... počet prvků
```

Příklad. Jste vedoucí grantové agentury a máte rozdělit 10 milionů korun mezi 12 žadatelů. Kolika způsoby lze rozdělit tato částka mezi žadatele za předpokladu, že částka lze dělit pouze na celé miliony?

1. Teorie

   Je třeba si uvědomit, že možnost, že všechny peníze přidělíte jednomu žadateli je stejně správná jako když je rozdělíte mezi 10 žadatelů. Ale zároveň nezáleží, jestli první žadatel dostane 1. milión nebo 10.

   $C'_{12}(10) = \begin{pmatrix} (10+12-1) \\ 12 \end{pmatrix} = \begin{pmatrix} 21 \\ 12 \end{pmatrix} = 293\,930$

2. Scilab

```
[vysledek]=kombinace_s_opakovanim(12,10)


vysledek=293930
```

2.3 Variace bez opakování

TEORIE

Variace bez opakování použijeme v případě, že vybíráme určitý počet objektů za předpokladu, že *záleží na pořadí* v jakém jsme je vybrali a *žádný objekt nemůžeme vybrat více než jedenkrát.* Vybíráme k-tice z n prvkové množiny a platí, že

$$V_k(n) = \frac{n!}{(n-k)!} \text{ pro } k \leq n$$

SCILAB - `variace()`

```
[vysledek]=variace(k,n)
   k ... počet k-členných skupin
   n ... počet prvků
```

Příklad. Při vytváření rozvrhu je dáno, že každý den lze mít maximálně 7 hodin. Rozvrháři při vytváření středečního rozvrhu ještě zbývá 7 hodin. Kolik možných způsobů rozvrhů na středu lze vytvořit?

1. Teorie

   Důležité je si uvědomit, že při vytváření rozvrhu záleží na pořadí a zároveň nelze duplikovat žádný předmět.

   $V_7(12) = \frac{12!}{(12-7)!} = 3\,991\,680$

2. Scilab

   ```
   [vysledek]=variace(7,12)
   ```

   ```
   vysledek=3991680
   ```

2.4 Variace s opakováním

TEORIE

Variace s opakováním použijeme v případě, že vybíráme určitý počet objektů za předpokladu, že *záleží na pořadí* v jakém jsme je vybrali a *objekt můžeme vybrat více než jedenkrát.* Vybíráme k-tice z n prvkové množiny a platí, že

$$V'_k(n) \quad = \quad n^k$$

SCILAB - `variace_s_opakovanim()`

```
[vysledek]=variace_s_opakovanim(k,n)
   k ... počet k-členných skupin
   n ... počet prvků
```

Příklad. Na lístek Vám kamarád napsal své telefonní číslo. Bohužel se Vám ho podařilo vyprat a tak je čitelné pouze prvních pět čísel z devíti. Kolik telefonních čísel je potřeba zkusit, aby bylo jisté, že jedno z volaných čísel bude kamarádovo?

1. Teorie

   $V'_4(10) = 10^4 = 10\,000$

2. Scilab

   ```
   [vysledek]=variace_s_opakovanim(4,10)
   ```

   ```
   vysledek=10000
   ```

2.5 Permutace

TEORIE

Permutace použijeme v případě, že vybíráme všechny objekty za předpokladu, že *záleží na pořadí* v jakém jsme je vybrali a *žádný objekt nemůžeme vybrat více než jedenkrát.* Vybíráme n-tice z n prvkové množiny a platí, že

$$P(n) = n!$$

SCILAB - `permutace()`

```
[vysledek]=permutace(n)
   n ... počet prvků
```

Příklad. Určete kolika způsoby může jít 10 studentů na zkoušku.

1. Teorie

   $P(10) = 10! = 3\,628\,800$

2. Scilab

   ```
   [vysledek]=permutace(10)
   ```

   ```
   vysledek=3628800
   ```

Kapitola 3

Pravděpodobnost

Během 19. století význam pravděpodobnosti a statistiky rostla zejména v oblasti sociálního poznání o státu. Byla shromažďována čísla o obyvatelstvu a hospodářství. I když se jednalo o nesystematický sběr dat, již se z těchto hodnot dalo, za pomocí právě teorie pravděpodobnosti, získat mnoho užitečných informací. První takový sběr dat udělal již v roce 1662 angličan John Graunt, který sledoval úmrtnost v Londýně a na základě těchto dat se pokusil vysledovat (předpovědět) chování dalších vln moru, který toto město často sužovaly....

3.1 Statistická pravděpodobnost

TEORIE

Jestliže provedeme N pokusů a jev J při nich nastane M krát, definujeme pravděpodobnost P jevu J jako

$$P(J) = \frac{M}{N}.$$

Poměr $\frac{M}{N}$ se označuje jako poměrná či relativní četnost jevu J.
Za předpokladu, že

1. při rostoucím velkém počtu pokusů se bude relativní četnost jevu J blížit k nějakému číslu,
2. experiment bude probíhat za ustálených podmínek,

lze považovat relativní četnost jevu J za skutečnou pravděpodobnost daného jevu J.

Statistická pravděpodobnost se při výpočtu pravděpodobnosti opírá o experimenty, což má za následek různé pravděpodobnosti při opakování experimentu. Na druhou stranu, jejich výhodou je jednoduchost.

SCILAB - `statisticka_pr()`

```
[vystup]=statisticka_pr(M,N)
   vystup ... relativní četnost jevu J
   M...počet pokusů, kdy jev nastoupil
   N...celkový počet pokusů
```

Příklad. Údaje o 100 narozených dětech jsou v tabulce:

	Váha do 3 kg	Váha nad 3 kg
Výška do 50 cm	60	20
Výška nad 50 cm	15	5

S jakou pravděpodobností bude vybrané dítě lehčí než 3 kg a zároveň vyšší než 50 cm.

1. Teorie

   $p = \frac{M}{N} = \frac{15}{100} = 0,15$

2. Scilab

```
[vysledek]=statisticka_pr(15,100)
```

```
vysledek=0.15
```

Příklad.

3.2 Klasická pravděpodobnost

TEORIE

Za předpokladu, že

1. náhodný pokus má konečný počet bezprostředních výsledků,
2. každý bezprostřední výsledek je stejně pravděpodobný,

je pravděpodobnost P jevu J dána vzorcem

$$P(J) = \frac{m}{n},$$

kde m je počet možných bezprostředních výsledků, při kterých nastane jev J a n je počet všech bezprostředních výsledků.

Klasická pravděpodobnost se při výpočtu pravděpodobnosti opírá o hodnoty pravděpodobnosti jednotlivých jevů, a to na základě teoretického rozboru pokusu, což má za následek stejné pravděpodobnosti při opakování experimentu. Na druhou stranu, nevýhodou se ukazuje výpočet při komplikovaných pokusech.

SCILAB - `klasická_pr()`

```
[vystup]=klasicka_pr(m,n)
   vystup...pravděpodobnost jevu J
   m...počet jevů, kdy se splní zadané pravidlo
   n...počet jevů, které mohou nastoupit
```

Příklad. Jaká je pravděpodobnost, že ve skupině 25 lidí jsou alespoň dva lidé, kteří budou mít narozeniny ve stejný den (narozeninový paradox).

1. Teorie
   (a) 1. člověk má narozeniny v libovolný den, tzn. $p(1) = 1$,
   (b) 2. člověk má narozeniny v jiný den je $p_{ne}(2) = p(1) \cdot \frac{364}{365}$,
   (c) 3. člověk má narozeniny v jiný den než 1. člověk a 2. člověk je $p_{ne}(3) = p(1) \cdot p_{ne}(2) \cdot \frac{363}{365} \Rightarrow$ lze odvodit obecný vzorec $p(n) = \frac{364 \cdot 363 \cdot \ldots \cdot (365-(n-1))}{365^{n-1}} = \frac{365!}{365^n (365-n)!}$
   (d) pravděpodobnost, že žádní dva lidé nemají narozeniny ve stejný den $p_{ne}(25) = \frac{365!}{365^{25} 340!} = 0,4313$
   (e) pravděpodobnost, že alespoň dva lidé budou mít narozeniny ve stejný den je $p_{ano}(25) = 1 - p_{ne}(25) = 0,5687$
2. Scilab

```
n=25;        //skupina 25 lidi
vysledek=1; //pomůcka pro for cyklus
for i=1:n
   vysledek=vysledek*klasicka_pr(365-i+1,365); //postupný součin: 365/365 · 364/365 · ... · 340/365
end
pravdepodobnost=1-vysledek //pravdepodobnost, ze nebudou mit narozeniny ve stejny den
```

3.3 Podmíněná pravděpodobnost

TEORIE

Pří počítání pravděpodobnosti lze k náhodnému pokusu přidat i podmínku. Zjednodušeně lze říci, že nastoupení jevu J_1 je podmíněno výsledkem jevu J_2.

Nechť J_1 je sledovaný jev a J_2 je pozorovaný jev pro který je $P(J_2) > 0$. Potom pravděpodobnost podmíněného jevu $P(J_1|J_2)$ je dána vzorcem

$$P(J_1|J_2) = \frac{P(J_1 \cap J_2)}{P(J_2)}$$

kde $P(J_1 \cap J_2)$ je pravděpodobnost průniku jevů, tedy pravděpodobnost výsledků, které jsou společné oběma jevům.

SCILAB - `podminena_pr()`

```
[J1podminenoJ2,J2podminenoJ1]=podminena_pr(J1,J2,n)
   J1podminenoJ2...J1|J2
   J2podminenoJ1...J2|J1
   J1...vektor hodnot jevu J1
   J2...vektor hodnot jevu J2
   n...počet jevů, které mohou nastoupit, pokud není n zadáno, n=J1∪J2
```

Příklad. Házíme šestistrannou kostkou. S jakou pravděpodobností hodíme číslo liché, když víme, že jsme hodili číslo menší jak 5?

1. Teorie
   $p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{1}{2}$
2. Scilab

```
[vysledek]=podminena_pr({1,3,5},{1,2,3,4},6)
```

```
vysledek=0.5
```

3.4 Nezávislost jevů

TEORIE

Pokud opakujeme n pokusů několikrát po sobě za stejných podmínek a přitom platí, že pravděpodobnost každého takového pokusu nezávisí na předchozím pokusu, tzn., můžeme tvrdit, že dané jevy jsou nezávislé.

Pro nezávislé jevy tedy platí, že $P(J_1 \cap J_2) = P(J_1)\,P(J_2)$ a z předpisu podmíněné pravděpodobnosti pak platí, že $P(J_1|J_2) = P(J_1)$.

Opačně lze o jevech J_1, J_2,..., J_n tvrdit, že jsou nezávislé, platí-li

$$P(J_1, J_2, ..., J_n) \quad = \quad P(J_1)\,P(J_2)\,...P(J_n)$$

SCILAB - `nezavislost_jevu()`

```
[vysledek]=nezavislost_jevu(pJ1,pJ2,pJ3,pJ4,pJ5)
   vysledek...pravděpodobnost jevů
   pJ1...pravděpodobnost jevu J1
   pJ2...pravděpodobnost jevu J2
   pJ3...pravděpodobnost jevu J3
   pJ4...pravděpodobnost jevu J4
   pJ5...pravděpodobnost jevu J5
   maximální počet hodnot na vstupu je 5, minimální počet hodnot je 2
```

Příklad. Při zkoušce statistiky má student 3 pokusy. Protože se neučí, zkoušku složí v každém pokusu s pravděpodobností 0,3. S jakou pravděpodobností zkoušku (a) složí na poslední pokus, (b) nesloží?

1. Teorie
   (a) $(1-0,3)^2\,0,3 = 0,147$
   (b) $(1-0,3)^3 = 0,343$
2. Scilab
   (a)
   ```
   [vysledek]=nezavislost_jevu(0.7,0.7,0.3)

   vysledek=0.147
   ```
   (b)
   ```
   [vysledek]=nezavislost_jevu(0.7,0.7,0.7)

   vysledek=0.343
   ```

3.5 Úplná pravděpodobnost

TEORIE

Podmíněnou pravděpodobnost používáme k výpočtu pravděpodobnosti složitějších jevů, které se realizují v několika různých možnostech. Pokud známe pravděpodobnosti těchto možností a pravděpodobnosti jevu v jednotlivých možnostech, pak lze snadno celkovou pravděpodobnost spočítat. Vzorec, který je z tohoto schématu odvozen, se nazývá *vzorec pro úplnou pravděpodobnost.*

Věta: Nechť P je pravděpodobnost na jevovém poli J a systém náhodných jevů $\{B_i \in J, \quad 1 \leq i \leq n\}$ splňuje tyto podmínky:

a) náhodné jevy jsou po dvou disjunktní

b) systém jevů je rozkladem jevového pole

pak pro náhodné jevy $A \in J$ je

$$P(A) \quad = \quad \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

!!!!Výpočet se provádí PŘED provedením pokusu!!!!

SCILAB - `uplna_pr()`

```
[vysledek]=uplna_pr(pAzaB1,pB1,pAzaB2,pB2,pAzaB3,pB3)
   pAzaB1 = p(A|B1),
   pB1=p(B1)
   pAzaB2 = p(A|B2),
   pB2=p(B2)
   pAzaB3 = p(A|B3)
   pB3=p(B3)[a]
```

[a] `pB3 pro 3 náhodné jevy, resp. pB2 pro 2 náhodné jevy nemusí být zadána`

Příklad. V dílně pracuje 20 dělníků, kteří vyrábějí stejné součástky. Každý z nich vyrobí za směnu stejné množství. Deset z nich vyrobí 94% výrobků I. třídy, šest 90% a čtyři 85%. Jaká je pravděpodobnost, že náhodně vybraný výrobek bude I. třídy?

1. Teorie

   $p(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i) = 0,94 \cdot \frac{1}{2} + 0,90 \cdot \frac{3}{10} + 0,85 \cdot \frac{2}{10} = 0,91$

2. Scilab

```
[vysledek]=uplna_pr(0.94,0.5,0.9,0.3,0.85,0.2)


vysledek=0.91
```

3.6 Bayesův vzorec

TEORIE

Nechť P je pravděpodobnost na jevovém poli J a systém náhodných jevů $\{B_i \in J, \quad 1 \leq i \leq n\}$ splňuje tyto podmínky:

a) náhodné jevy jsou po dvou disjunktní

b) systém jevů je rozkladem jevového pole

pak

$$P(B_i|A) \quad = \quad \frac{P(A|B_i) \cdot P(B_i)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

!!!!Výpočet se provádí PO provedení pokusu!!!!

SCILAB - `bayesuv_vzorec()`

```
[vysledek]=bayesuv_vzorec(PA,pAzaB1,pB1,pAzaB2,pB2,pAzaB3,pB3)
   pA... =1 pro p(B1|A), =2 pro p(B2|A), =3 pro p(B3|A)
   pAzaB1 = p(A|B1),
   pB1=p(B1)
   pAzaB2 = p(A|B2),
   pB2=p(B2)
   pAzaB3 = p(A|B3)
   pB3=p(B3)
   pB3 pro 3 náhodné jevy, resp. pB2 pro 2 náhodné jevy nemusí být zadána.
```

Příklad. V dílně pracuje 20 dělníků, kteří vyrábějí stejné součástky. Každý z nich vyrobí za směnu stejné množství. Deset z nich vyrobí 94% výrobků I. třídy, šest 90% a čtyři 85%. Vybrali jsme výrobek a ten byl I. třídy, s jakou pravděpodobností ho vyrobil jeden ze šestice dělníků, kteří vyrábí 90% výrobků I. třídy

1. Teorie

   $p(B2|A) = \frac{p(A|B_2)p(B_2)}{p(A)} = \frac{0,9 \cdot 0,3}{0,91} = 0,2967$

2. Scilab

```
[vysledek]=bayesuv_vzorec(2,0.94,0.5,0.9,0.3,0.85,0.2)


vysledek=0.2967
```

Kapitola 4

Popisná statistika

Příklad. Následující posloupnost čísel zobrazuje výsledky série 100 hodů poctivou šestistrannou kostkou (kostka.dat):

2	5	1	2	4	4	6	5	6	1	4	4	5	2	4	2	2	2	6	4
2	6	2	2	3	2	4	3	2	4	4	3	2	4	3	6	1	3	2	3
2	1	5	2	1	5	1	5	6	3	3	6	1	2	4	4	5	6	4	3
3	6	6	3	3	5	2	3	2	4	4	1	2	4	5	1	5	2	3	5
4	3	2	6	1	2	6	6	4	6	4	6	1	5	3	4	6	1	5	6

Tabulka 4.1: Tabulka hodů kostkou

Pro další práci s daty je vhodné data převést do tabulky četností těchto hodů. Četnosti jsou zaznamenány v tabulce četností, kde x_i označuje hodnotu hodu a n_i je počet (četnost) hodů, kdy daný jev nastoupil.

x_i	1	2	3	4	5	6
n_i	12	22	16	20	13	17

Tabulka 4.2: Tabulka četností

4.1 Střední hodnota

TEORIE

Pro diskrétní náhodnou veličinu X s hodnotami $\{x_1, x_2, ..., x_n\}$ s pravděpodobnostní funkcí , resp. spojitou náhodnou veličinu X s hodnotami z množiny X^* a hustotou pravděpodobnosti $f(x)$ definujeme střední hodnotu $E[X]$ vztahem:

1. $E[X] = \sum_{i=1}^{n} x_i f(x_i)$ pro diskrétní náhodnou veličinu.
2. $E[X] = \int_{X^*} x f(x)\, dx$ pro spojitou náhodnou veličinu.

SCILAB - `mean()`

```
[y]=mean(x,orient)
   x ... vektor nebo matice prvků
   orient... orientace po řádcích nebo sloupcích pro výpočet střední hodnoty 'r', 'c'
```

Příklad. Spočítejte střední hodnotu hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Převedeme tabulku četností na tabulku pravděpodobností výskytu jednotlivých hodů, tedy

x_i	1	2	3	4	5	6
n_i	0,12	0,22	0,16	0,20	0,13	0,17

Tabulka 4.3: Tabulka pravděpodobnosti výskytu

   Poté dosadíme do vzorce pro výpočet střední hodnoty pro diskrétní náhodnou veličinu

$$E[X] = \sum_{i=1}^{6} x_i f(x_i) = 1 \cdot 0,12 + 2 \cdot 0,22 + 3 \cdot 0,16 + 4 \cdot 0,20 + 5 \cdot 0,13 + 6 \cdot 0,17 = 3,51$$

2. Scilab

```
load("kostka_data.sod")
vysledek=mean(kostka,'c')

vysledek=3.51
```

4.2 Rozptyl

TEORIE

1. Souborový rozptyl je druhý centrální moment výběru. Souborový rozptyl je možno použít v případě, že je splněn předpoklad, že se nejedná o výpočet z výběru. Rozptyl je dán vzorcem

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

kde X_i jsou hodnoty náhodných veličin a $\bar{X}$ je průměr všech jejich hodnot.

2. Výběrový rozptyl udává proměnlivost hodnot datového souboru.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

SCILAB - `variance()`

```
[s2] = variance(x, orient, w)
   x ... vektor nebo matice prvků
   orient... orientace po řádcích nebo sloupcích pro výpočet rozptylu 'r', 'c'
   w ... určení typu rozptylu: 0 pro výběrový rozptyl, 1 pro souborový rozptyl
```

Příklad. Spočítejte rozptyl hodů kostkou, viz4 tabulka 4.1.

1. Teorie

Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Podle vzorce pro střední hodnotu, si spočteme střední hodnotu, která nám vyjde $\bar{x} = 3,51$. Poté dosadíme do vzorce pro rozptyl (v tomto případě výběru) pro diskrétní náhodnou veličinu

$$\begin{aligned} s^2 &= \frac{1}{100-1}\left((1-3,51)^2\cdot 12 + (2-3,51)^2\cdot 22 + \ldots + (6-3,51)^2\cdot 17\right) = \\ &= \frac{1}{100-1}\left(76,6012 + 50,1622 + 4,1616 + 4,8020 + 28,8613 + 105,4017\right) = 2,7271 \end{aligned}$$

2. Scilab

```
load("kostka_data.sod")
vysledek=variance(kostka,'c',0)

vysledek=2.7170
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

4.3 Směrodatná odchylka

TEORIE

Směrodatná odchylka je druhá odmocnina z rozptylu

$$\sigma = \sqrt{D[X]}$$

SCILAB - `st_deviation(x,orient)`

```
[s] = st_deviation(x, orient)
   x ... vektor nebo matice prvků
   orient... orientace po řádcích nebo sloupcích pro výpočet rozptylu 'r', 'c'
```

Příklad. Spočítejte směrodatnou odchylku hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Nejprve spočteme rozptyl, který nám vyjde $s^2 = 2,7271$ (nadále používáme teoretický výsledek). Poté dosadíme do vzorce pro směrodatnou odchylku pro diskrétní náhodnou veličinu

   $$s^2 = \sqrt{2,7271} = 1,6514$$

2. Scilab

   ```
   load("kostka_data.sod")
   vysledek=st_deviation(kostka,'c')


   vysledek=1.6484
   ```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

4.4 Modus

TEORIE

Modus náhodné veličiny $\hat{x}$je nejčetnější hodnota statistického souboru (hodnota s největší relativní četností). Modus nemusí být vždy určen jednoznačně, tzn. je více hodnot s nejvyšší četností.

Vlastnosti modu:

1. není ovlivněn případnými odlehlými (extrémními) hodnotami souboru (výběru). tzn. modus není ovlivněn všemi hodnotami dané proměnné,
2. v případě, že je modus určen jednoznačně, lze ho při neznalosti typu rozdělení považovat za střed souboru.

SCILAB - `modus()`

```
[vystup]=modus(x)
   vystup ... hodnota modu
   x ... vektor vstupních hodnot
```

Příklad. Spočítejte modus hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Pokud pracujeme s tabulkou četností, tak modus tohoto výběru je hodnota s největší četností.

$$\hat{x} = 2$$

2. Scilab

```
load("kostka_data.sod")
vysledek=modus(kostka)


vysledek=2
```

4.5 Medián

TEORIE

Medián $\tilde{x}_{0,5}$ je prostřední hodnota uspořádaného souboru, tedy taková hodnota, která rozděluje soubor na dvě stejně velké části co do počtu prvků. V případě lichého počtu prvků je mediánem prostřední hodnota, v případě sudého počtu prvků je to aritmetický průměr dvou prostředních hodnot.

Vlastnosti mediánu:

1. není ovlivněn případnými odlehlými (extrémními) hodnotami souboru (výběru),
2. lze ho použít i při neznalosti typu rozdělení jako prostřední hodnotu, a to i v případě, že se jedná o rozdělení s více vrcholy.

SCILAB - `median()`

```
y=median(x,orient)
   y ... hodnota medianu
   x ... vektor vstupních hodnot
   orient ... upřesnění, zda hledáme medián po řádcích nebo po sloupcích
```

Příklad. Spočítejte medián hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Protože máme 100 hodů, medián bude aritmetických průměr 50. a 51. prvku u uspořádaného souboru.

$$\tilde{x}_{0,5} = \frac{3+4}{2} = 3,5$$

2. Scilab

```
load("kostka_data.sod")
vysledek=median(kostka)


vysledek=3.5
```

4.6 Kvantil

TEORIE

Pro náhodnou veličinu x s hustotou pravděpodobnosti $f(x)$ se definuje kvantil ζ_α jako $\alpha \cdot 100\%$ nejmenších realizací náhodné veličiny. Platí tedy

$$\int_{-\infty}^{\zeta_\alpha} f(x)\,dx \quad = \quad \alpha$$

Speciální typy kvantilu

1. horní kvartil $\tilde{x}_{0,75} = 75\%$ kvantil,
2. medián $\tilde{x}_{0,5} = 50\%$ kvantil,
3. dolní kvartil $\tilde{x}_{0,25} = 25\%$ kvantil.

SCILAB - `percentil()`

```
[per]=percentil(x,r)
   per ... vysledek percentilu
   x ... vstupní vektor dat
   r ... hodnota percentilu
```

Příklad. Spočítejte horní a dolní kvartil hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Protože máme 100 hodů, horní kvartil $\tilde{x}_{0,75}$ bude aritmetickým průměrem mezi 75. a 76. prvkem, u dolního kvartilu $\tilde{x}_{0,25}$to bude mezi 25. a 26. prvkem.

$$\begin{aligned} \tilde{x}_{0,25} &= \frac{2+2}{2} = 2 \\ \tilde{x}_{0,75} &= \frac{5+5}{2} = 5 \end{aligned}$$

2. Scilab

```
load("kostka_data.sod")
dolni_kvartil=percentil(kostka,0.25)
horni_kvartil=percentil(kostka,0.75)

horni_kvartil=5
dolni_kvartil=2
```

4.7 Mezikvartilové rozpětí

TEORIE

Mezikvartilové rozpětí je rozdíl horního a dolního kvartilu, tedy

$$Iqr \quad = \quad \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

SCILAB - `mezikvartilove_rozpeti()`

```
[vystup]=mezikvartilove_rozpeti(x)
   vystup ... hodnota mezikvartilového rozpětí
   x ... vstupní vektor dat
```

Příklad. Spočítejte mezikvartilové rozpětí hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Vypočítáme horní kvartil $\tilde{x}_{0,75}$ a dolní kvartil $\tilde{x}_{0,25}$. Mezikvartilové rozpětí je rozdíl těchto dvou kvartilů, tedy

   $$Iqr \quad = \quad \tilde{x}_{0,75} - \tilde{x}_{0,25} = 5 - 2 = 3$$

2. Scilab

   ```
   load("kostka_data.sod")
   vysledek=mezikvartilove_rozpeti(kostka)

   vysledek=3
   ```

4.8 Variační rozpětí

TEORIE

Variační rozpětí je rozdíl mezi největší a nejmenší hodnotou statistického souboru, tedy

$$R = x_{max} - x_{min}$$

SCILAB - `variacni_rozpeti()`

```
[vystup]=variacni_rozpeti(X)
   vystup ... hodnota variačního rozpětí
   X ... hodnoty náhodné veličiny X
```

Příklad. Spočítejte variační rozpětí hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Nejdeme největší a nejmenší hodnotu statistického souboru. Je výhodnější použít tabulku četností 4.2 než procházet celý neuspořádaný výběr. Platí tedy

   $$R = x_{max} - x_{min} = 6 - 1 = 5$$

2. Scilab

   ```
   load("kostka_data.sod")
   vysledek=variacni_rozpeti(kostka)

   vysledek=5
   ```

4.9 Centrální moment

TEORIE

K-tý centrální moment náhodné veličiny X definujeme vztahem

$$\mu_k = E\left[\left(X - E[X]^k\right)\right].$$

Druhým centrálním momentem je rozptyl souboru.

SCILAB - `cmoment()`

```
[mom] = cmoment(x,k,orient)
   mom ... hodnota centralního momentu
   x ... vstupní vektor dat
   k ... k-tý moment. Například 2, znamená výpočet souborového rozptylu
   orient ... upřesnění, zda hledáme centrální moment po řádcích nebo po sloupcích
```

Příklad. Spočítejte druhý centrální moment hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Podle vzorce pro střední hodnotu, si spočteme střední hodnotu, která nám vyjde $\bar{x} = 3,51$. Poté dosadíme do vzorce pro druhý centrální moment náhodné veličiny, tedy do vzorce $\mu_2 = E\left[\left(X - E[X]^2\right)\right]$

$$\begin{aligned}\mu_2 &= \frac{1}{100}\left((1-3,51)^2 \cdot 12 + (2-3,51)^2 \cdot 22 + ... + (6-3,51)^2 \cdot 17\right) = \\ &= \frac{1}{100}\left(76,6012 + 50,1622 + 4,1616 + 4,8020 + 28,8613 + 105,4017\right) = 2,6999\end{aligned}$$

2. Scilab

```
load("kostka_data.sod")
vysledek=cmoment(kostka,2,'c')

vysledek=2.6899
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

4.10 Obecný moment

TEORIE

K-tý obecný moment náhodné veličiny X definuje vtahem

$$\mu'_k = E\left[X^k\right]$$

Prvním obecným momentem je střední hodnota.

SCILAB - `moment()`

```
[mom] = moment(x,k,orient)
   mom ... hodnota obecného momentu
   x ... vstupní vektor dat
   k ... k-tý moment. Například 1, znamená výpočet střední hodnoty
   orient ... upřesnění, zda hledáme centrální moment po řádcích nebo po sloupcích
```

Příklad. Spočítejte první obecný moment hodů kostkou, viz 4 tabulka 4.1.

1. Teorie

   Pro výpočet je výhodnější použít tabulku četností 4.2 než pracovat s jednotlivými hody. Převedeme tabulku četností na tabulku pravděpodobností výskytu jednotlivých hodů, tedy

x_i	1	2	3	4	5	6
n_i	0,12	0,22	0,16	0,20	0,13	0,17

Tabulka 4.4: Tabulka pravděpodobnosti výskytu

Poté dosadíme do vzorce pro výpočet první obecný moment, tedy do vzorce $\mu'_1 = E\left[X^1\right] = E\left[X\right]$. Hledáme střední hodnotu, tedy

$$E\left[X\right] = \sum_{i=1}^{6} x_i f\left(x_i\right) = 1 \cdot 0,12 + 2 \cdot 0,22 + 3 \cdot 0,16 + 4 \cdot 0,20 + 5 \cdot 0,13 + 6 \cdot 0,17 = 3,51$$

1. Scilab

```
load("kostka_data.sod")
vysledek=moment(kostka,1,'c')


vysledek=3.51
```

4.11 Kovariance

TEORIE

$$cov\left[X,\,Y\right]=E\left[\left(X-E\left[X\right]\right)\left(Y-E\left[Y\right]\right)\right]=E\left[XY\right]-E\left[X\right]E\left[Y\right]$$

kde X a Y jsou náhodné veličiny a $E\left[X\right]$ resp. $E\left[Y\right]$ jsou střední hodnoty náhodné veličiny X resp. Y.

Základní vlastnosti kovariance:

1. Náhodné veličiny X a Y jsou nezávislé $\Rightarrow cov\left(X,Y\right)=0$.
2. $cov\left(X,Y\right)=cov\left(Y,X\right)$.
3. $cov\left(X,X\right)=D\left[X\right]$.

SCILAB - `kovariance()`

```
[vystup]=kovariance(X,Y,typ)
   vystup ... hodnota kovariance
   X ... hodnoty náhodné veličiny X
   Y ... hodnoty náhodné veličiny Y
   typ ... kovariance souboru (typ=1 nebo typ='s'), kovariance výběru (typ=0 nebo typ='v')
```

Příklad. Máme poctivou minci a poctivou šestistrannou kostku. Nejdříve hodíme mincí. V případě, že padne líc, hodíme 2x kostkou v opačném případě hodíme kostkou 1x. Náhodná veličina X nabývá hodnoty 0 (padle líc) nebo hodnoty 1 (padne rub). Náhodná veličina Y nabývá hodnoty 0 (ani jednou nepadne šestka), nabývá hodnoty 1 (právě jednou padne šestka) nebo hodnoty 2 (šestka padla v obou hodech). Provedli jsme 100 pokusů a výsledné údaje zapsali do kontingenční tabulky, viz tabulka 4.5.

	$Y=0$	$Y=1$	$Y=2$
$X=0$	37	17	2
$X=1$	38	6	0

Tabulka 4.5: Kontingenční tabulka pro hod mincí a kostkou

Spočtěte kovarianci veličin X a Y.

1. Teorie

   Pro výpočet je výhodnější převést kontingenční tabulku 4.5 do sdruženého rozdělení pravděpodobnosti f(x,y), kde dopočítáme i marginály f(x) a f(y) tedy

$f\left(x,y\right)$	$y=0$	$y=1$	$y=2$	$f\left(x\right)$
$x=0$	0,37	0,17	0,02	0,66
$x=1$	0,38	0,06	0	0,44
$f\left(y\right)$	0,75	0,23	0,02	

Tabulka 4.6: Sdružené rozdělení pravděpodobnosti

$$\begin{aligned}
E\left[X\right] &= \sum_{i=1}^{2}x_i f\left(x_i\right)=0\cdot 0,66+1\cdot 0,44=0,44\\
E\left[Y\right] &= \sum_{i=1}^{3}y_i f\left(y_i\right)=0\cdot 0,75+1\cdot 0,23+2\cdot 0,02=0,27\\
E\left[X,Y\right] &= \sum x_i y_j f\left(x_i,y_j\right)=0\cdot 0\cdot 0,37+1\cdot 0\cdot 0,38+0\cdot 1\cdot 0,17+1\cdot 1\cdot 0,06+0\cdot 2\cdot 0,02+1\cdot 2\cdot 0=0,06
\end{aligned}$$

$$cov\left[X,Y\right] = E\left[X,Y\right]-E[X]E[Y]=0,06-0,44\cdot 0,27=-0,0588$$

2. Scilab

```
load("mince_kostka.sod");
x=[zeros(1,56) ones(1,44)];
y=[zeros(1,37) ones(1,17) 2 2 zeros(1,38) ones(1,6)];
vysledek=kovariance(x,y,'s');

vysledek=-0.0588
```

4.12 Kovarianční matice

TEORIE

Nechť X_1, X_2, ..., X_k popisují k vektorů náhodných veličin. Kovarianční matice je zobecněním rozptylů mezi těmito náhodnými veličinami. Lze psát

$$C = \begin{pmatrix} D[X_1] & Cov[X_1, X_2] & \cdots & Cov[X_1, X_k] \\ Cov[X_1, X_k] & D[X_k] & \cdots & Cov[X_2, X_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[X_1, X_k] & Cov[X_1, X_2] & \cdots & D[X_k] \end{pmatrix}$$

SCILAB - `kovariancni_matice()`

```
[vystup]=kovariancni_matice(X,Y)
   vystup ... hodnota kovariance
   X ... hodnoty náhodné veličiny X
   Y ... hodnoty náhodné veličiny Y
```

Příklad. Máme poctivou minci a poctivou šestistrannou kostku. Nejdříve hodíme mincí. V případě, že padne líc, hodíme 2x kostkou v opačném případě hodíme kostkou 1x. Náhodná veličina X nabývá hodnoty 0 (padle líc) nebo hodnoty 1 (padne rub). Náhodná veličina Y nabývá hodnoty 0 (ani jednou nepadne šestka), nabývá hodnoty 1 (právě jednou padne šestka) nebo hodnoty 2 (šestka padla v obou hodech). Provedli jsme 100 pokusů a výsledné údaje zapsali do kontingenční tabulky, viz tabulka 4.7.

	$Y=0$	$Y=1$	$Y=2$
$X=0$	37	17	2
$X=1$	38	6	0

Tabulka 4.7: Kontingenční tabulka pro hod mincí a kostkou

Spočtěte kovarianci veličin X a Y.

1. Teorie

   Pro výpočet je výhodnější převést kontingenční tabulku 4.7 do sdruženého rozdělení pravděpodobnosti f(x, y), kde dopočítáme i marginály f(x) a f(y) tedy

$f(x,y)$	$y=0$	$y=1$	$y=2$	$f(x)$
$x=0$	0,37	0,17	0,02	0,66
$x=1$	0,38	0,06	0	0,44
$f(y)$	0,75	0,23	0,02	

Tabulka 4.8: Sdružené rozdělení pravděpodobnosti

Spočteme kovarianci $cov[X, Y]$ a rozptyly $D[X]$ a $D[Y]$

$$\begin{aligned} cov[X,Y] &= E[X,Y] - E[X]E[Y] = 0,06 - 0,44 \cdot 0,27 = -0,0588 \\ D[X] &= E[X^2] - (E[X])^2 = 0,44 - (0,44)^2 = 0,2464 \\ D[Y] &= E[Y^2] - (E[Y])^2 = 0,31 - (0,27)^2 = 0,2371 \end{aligned}$$

$$C = \begin{pmatrix} 0,2464 & -0,0588 \\ -0,0588 & 0,2371 \end{pmatrix}$$

2. Scilab

```
load("mince_kostka.sod")
x=[zeros(1,56) ones(1,44)];
y=[zeros(1,37) ones(1,17) 2 2 zeros(1,38) ones(1,6)];
vysledek=kovariancni_matice(x,y);

vysledek=[0.2488889 -0.0593939; -0.0593939 0.2394949]
```

4.13 Korelace

TEORIE

Korelační koeficient je míra vztahu mezi dvěma proměnnými. Nabývá hodnot mezi -1 a $+1$. Tyto hraniční hodnoty znamenají, že proměnné jsou silně korelované. V případě, že hodnota $r = 0$ hovoříme o nekorelovaných veličinách, tedy o veličinách bez vzájemného lineárního vztahu. Korelační koeficient lze vypočítat podle následujícího vzorce

$$r = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \sum_{i=1}^{n} (y_i - \mu_y)^2}}$$

SCILAB - `korelace()`

```
[vystup]=korelace(X,Y)
   vystup ... korelační koeficient
   X ... hodnoty náhodné veličiny X
   Y ... hodnoty náhodné veličiny Y
```

Příklad. V továrně byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). Byly zaznamenány následující údaje

produkce	532	297	378	121	519	613	592	497
náklady	48	32	42	27	45	51	53	48

Určete korelační koeficient r.

1. Teorie

   Pro výpočet korelace je nutné vypočítat následující charakteristiky náhodné veličina: (i) střední hodnotu, (ii) rozptyl a (iii) kovarianci.

   (a) Spočteme střední hodnotu.

$$\begin{aligned} E[X] &= \frac{1}{8}(532 + 297 + 378 + 121 + 519 + 613 + 592 + 497) = 443,625 \\ E[Y] &= 43,375 \end{aligned}$$

   (b) Spočteme rozptyl:

$$\begin{aligned} D[X] &= \frac{1}{8}\left((532 - 443,625)^2 + (297 - 443,625)^2 + ... + (497 - 433,625)^2\right) = 24616,984 \\ D[Y] &= 74,234 \end{aligned}$$

   (c) Spočteme kovarianci

$$cov[X,Y] = \frac{1}{8}((532 - 443,625)(48 - 43,375) + ... + (497 - 443,625)(48 - 43,375)) = 1496,1607$$

   (d) Spočteme korelační koeficient

$$r = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \sum_{i=1}^{n} (y_i - \mu_y)^2}} = \frac{cov[X,Y]}{\sqrt{D[X]\,D[Y]}} = \frac{1309,1406}{\sqrt{24616,984 \cdot 74,234}} = 0.9684$$

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
vysledek=korelace(x,y);

vysledek=0.9687
```

Kapitola 5

Intervalové odhady

5.1 Interval spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2

TEORIE

Nechť $\bar{x}$ je střední hodnota z náhodného výběru o rozsahu n, který byl vybrán ze souboru se známým rozptylem σ^2.

1. Oboustranný$(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \,;\, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

   kde $z_{\alpha/2}$ je rovna $\frac{\alpha}{2}$-té kritické hodnotě normovaného normálního rozdělení $N(0,1)$.

2. Levostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha} \,;\, \infty \right)$$

   kde z_{α} je rovna α-tékritické hodnotě normovaného normálního rozdělení $N(0,1)$.

3. Pravostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(-\infty \,;\, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha} \right)$$

   kde z_{α} je rovna α-té kritické hodnotě normovaného normálního rozdělení $N(0,1)$.

SCILAB

```
[dolni_mez, horni_mez]=z_int (stredni_hodnota,rozptyl,n,strana,alpha)
   horni_mez ... horní mez intervalu
   dolni_mez ... dolní mez intervalu
   stredni_hodnota ... střední hodnota výběru
   rozptyl ... rozptyl souboru (musí být zadán)
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Předpokládejme, že výška chlapců ve věku 9 až 10 roků má normální rozdělení $N\left(\mu, \sigma^2\right)$ s neznámou střední hodnotou a rozptylem rovným 39,112. Změřili jsme výšku 15 chlapců a vypočítali průměr 139,13. Určete 99% oboustranný IS pro skutečnou výšku chlapců.

1. Teorie

   $\sigma^2 = 39,112$; $\bar{x} = 139,13$; $n = 15$; $\alpha = 0,01$; oboustranný IS, $z_{\alpha/2} = 2,576$, $\mu = ?$,

$$\begin{aligned} \mu &\in \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\,;\, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) \\ \mu &\in \left(139,13 - \frac{\sqrt{39,112}}{\sqrt{15}} 2,58\,;\, 139,13 - \frac{\sqrt{39,112}}{\sqrt{15}} 2,58\right) \\ \mu &\in (134,9639\,;\, 143,2961) \end{aligned}$$

2. Scilab

```
[DM,HM]=z_int(139.13,39.112,15,'o',0.01)

HM=143.2894
DM=134.9706
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

5.2 Interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu σ^2

TEORIE

Nechť $\bar{x}$je střední hodnota a s^2 je rozptyl vypočtený z náhodného výběru o rozsahu n, který byl vybrán ze souboru s neznámým rozptylem σ^2.

1. Oboustranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(\overline{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}\,;\, \overline{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}\right)$$

   kde $t_{\alpha/2}$ je rovna $\frac{\alpha}{2}$-té kritické hodnotě studentova rozdělení $St\,(n-1)$.

2. Levostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(\overline{x} - \frac{s}{\sqrt{n}} t_{\alpha}\,;\, \infty\right)$$

   kde z_{α} je rovna α-té kritické hodnotě studentova rozdělení $St\,(n-1)$.

3. Pravostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečnou střední hodnotu lze zapsat jako

$$\mu \in \left(-\infty\,;\, \overline{x} + \frac{s}{\sqrt{n}} t_{\alpha}\right)$$

   kde z_{α} je rovna α-té kritické hodnotě studentova rozdělení $St\,(n-1)$.

Pozn.: Pokud pracujeme s výběrem, kde $n > 100$, lze nahradit studentovo rozdělení přesnějším normovaným normálním rozdělením $N\,(0,1)$.

SCILAB

```
[dolni_mez, horni_mez]=t_int (stredni_hodnota,rozptyl,n,strana,alpha)
   horni_mez ... horní mez intervalu
   dolni_mez ... dolní mez intervalu
   stredni_hodnota ... střední hodnota výběru
   rozptyl ... rozptyl výběru
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Měřili jsme průměr klikové hřídele na 250 součástkách. Předpokládáme, že naměřené veličiny mají rozdělení $N\left(\mu, \sigma^2\right)$. Z výsledků měření jsme vypočetli průměrnou hodnotu 995,6 a rozptyl $s^2 = 134,7$. Určete 95% oboustranný IS pro skutečný průměr klikové hřídele.

1. Teorie

   $s^2 = 134,7$; $\bar{x} = 995,6$; $n = 250$; $\alpha = 0,05$; oboustranný IS, kritická hodnota $t_{0,025}\,(249) = 1,96$, μ =?

$$\begin{aligned}
\mu &\in \left(\overline{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}\,;\, \overline{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}\right) \\
\mu &\in \left(995,6 - \frac{\sqrt{134,7}}{\sqrt{250}} 1,96\,;\, 995,6 + \frac{\sqrt{134,7}}{\sqrt{250}} 1,96\right) \\
\mu &\in (994,1613\,;\, 997,0387)
\end{aligned}$$

2. Scilab

```
[DM,HM]=t_int(995.6,134.7,250,'o',0.05)

HM=997.0457
DM=994.1543
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

5.3 Interval spolehlivosti pro rozptyl σ^2

TEORIE

Nechť je náhodný výběr o rozsahu n vybrán ze souboru s normálním rozdělením $N(\mu, \sigma^2)$. Hledáme interval spolehlivosti neznámého parametru σ^2 (rozptyl souboru) za předpokladu znalosti parametru μ(střední hodnota).

1. Oboustranný $(1-\alpha)\,100\,\%$interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný rozptyl lze zapsat jako

$$\sigma^2 \in \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} ; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

   kde $\chi_{\frac{\alpha}{2}}$ resp. $\chi_{1-\frac{\alpha}{2}}$je rovna $\frac{\alpha}{2}$-tému resp. $\left(1-\frac{\alpha}{2}\right)$-té kritické hodnotě Chi-kvadrát rozdělení $\chi^2\,(n-1)$.

2. Levostranný $(1-\alpha)\,100\,\%$interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný rozptyl lze zapsat jako

$$\sigma^2 \in \left(\frac{(n-1)s^2}{\chi^2_{\alpha}} ; \infty \right)$$

   kde χ_α je rovna α-té kritické hodnotě Chi-kvadrát rozdělení $\chi^2\,(n-1)$.

3. Pravostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný rozptyl lze zapsat jako

$$\sigma^2 \in \left(0 ; \frac{(n-1)s^2}{\chi^2_{1-\alpha}} \right)$$

   kde $\chi_{1-\alpha}$ je rovna $(1-\alpha)$-té kritické hodnotě Chi-kvadrát rozdělení $\chi^2\,(n-1)$.

SCILAB

```
[dolni_mez, horni_mez]=var_int(rozptyl,n,strana,alpha)
   horni_mez ... horní mez intervalu
   dolni_mez ... dolní mez intervalu
   rozptyl ... rozptyl výběru
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Pro zjišťování přesnosti metody pro stanovení obsahu manganu v oceli byla provedena 4 nezávislá měření vzorků. Chceme stanovit hranici, pro níž platí, že rozptyl větší než tato hranice se bude objevovat jen v 5% pokusů. Výsledky měření jsou: 0,31; 0,30; 0,29; 0,32.

1. Teorie

   pravostranný IS$\chi^2_{1-\alpha}(3) = 0,352$, n=4

   $\bar{x} = \frac{1}{4}\,(0,31 + 0,30 + 0,29 + 0,32) = 0,305$

   $s^2 = \frac{1}{3}\left((0,31-0,305)^2 + (0,30-0,305)^2 + (0,29-0,305)^2 + (0,32-0,305)^2\right) = 0,00017$

$$\begin{aligned} \sigma^2 &\in \left(0\,;\, \frac{(n-1)s^2}{\chi^2_{1-\alpha}}\right) \\ \sigma^2 &\in \left(0\,;\, \frac{(4-1)\cdot 0,00017}{0,352}\right) \\ \sigma^2 &\in (0\,;\, 0,0014) \end{aligned}$$

2. Scilab

```
x=[0.31 0.30 0.29 0.32];
v=variance(x);
n=size(x,2);
[DM,HM]=var_int(v,n,'p',0.05)

HM=0.0014
DM=0
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

5.4 Interval spolehlivosti pro podíl π

TEORIE

Nechť náhodná veličina X pochází z binomického rozdělení $Bi(\pi, n)$, kde π je pravděpodobnost úspěchu a n je počet opakování. Přestože binomické rozdělení není svou podstatou normální, lze ho za určitých podmínek ($n > 30$ a $np > 5$ resp. $n(1-p) > 5$), kde n je počet dat ve výběru a p je pravděpodobnost úspěchu získaných z výběru, aproximovat normální rozdělením.

Interval spolehlivosti se určuje pro parametr π, tedy pravděpodobnost úspěchu jako oboustranný test.

1. Oboustranný$(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný podíl lze zapsat jako

$$\pi \in \left(p - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}\,;\, p + \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2},\right)$$

   kde $z_{\alpha/2}$ je rovna $\frac{\alpha}{2}$-té kritické hodnotě normovaného normálního rozdělení $N(0,1)$.

2. Levostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný podíl lze zapsat jako

$$\pi \in \left(p - \sqrt{\frac{p(1-p)}{n}} z_{\alpha}\,;\, 1\right)$$

   kde z_{α} je rovna α-té kritické hodnotě normovaného normálního rozdělení $N(0,1)$.

3. Pravostranný $(1-\alpha)\,100\,\%$ interval spolehlivosti (tzn. 99%, 95% IS atd.) pro skutečný podíl lze zapsat jako

$$\pi \in \left(0\,;\, p - \sqrt{\frac{p(1-p)}{n}} z_{\alpha}\right)$$

   kde z_{α} je rovna α-té kritické hodnotě normovaného normálního rozdělení $N(0,1)$.

SCILAB

```
[dolni_mez, horni_mez]=prop_int(p,n,strana,alpha)
   horni_mez ... horní mez intervalu
   dolni_mez ... dolní mez intervalu
   p ... pravděpodobnost úspěchu nebo počet úspěchů ve výběru
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. V jakém intervalu lze s pravděpodobností 0,99 očekávat podíl nekvalitních výrobků, jestliže v náhodném výběru o rozsahu 1000 ks bylo zjištěno 15 nekvalitních výrobků?

1. Teorie

   $n = 1000$, $p = \frac{15}{1000} = 0,015$, oboustranný IS $z_{\alpha/2} = 1,576$, $p = ?$

$$\pi \in \left(p - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}\,;\, p + \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2},\right)$$

$$\pi \in \left(0,015 - \sqrt{\frac{0,015(1-0,015)}{1000}} 2,576\,;\, 0,015 + \sqrt{\frac{0,015(1-0,015)}{1000}} 2,576\right)$$

$$\pi \in (0,0051\,;\, 0,0249)$$

1. Scilab

```
[DM,HM]=prop_int(15,1000,'o',0.01)

HM=0.0249
DM=0.0051
```

2. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

Kapitola 6

Testy hypotéz

Na základě výběru srovnáváme dvě tvrzení o hodnotě určitého parametru θ rozdělení$f(x, \theta)$. První tvrzení (které většinou obhajuje stávající stav věcí) se nazývá **nulová hypotéza** a značí se H_0, druhé tvrzení (které většinou prosazuje, že věci se změnily) je **alternativní hypotéza** označená H_A. Nulová hypotéza něco tvrdí: např., že střední hodnota μ je rovna μ_0 a alternativní hypotéza ji odporuje. To může mít tři různé podoby:

hypotéza	příklad
I. parametr má podle H_A větší hodnotu než podle H_0	$\mu > \mu_0$
II. parametr má podle H_A menší hodnotu než podle H_0	$\mu < \mu_0$
III. parametr se podle H_A nerovná hodnotě parametru podle H_0	$\mu \neq \mu_0$

6.1 Testy jedné veličiny

6.1.1 Test hypotéz o střední hodnotě μ při známém rozptylu σ^2

TEORIE

Na základě náhodného výběru ze souboru s libovolným rozdělením[1] testujeme předpoklad, že střední hodnota náhodné veličiny μ se rovná nějaké námi předpokládané střední hodnotě μ_0. Předpokládáme, že rozptyl souboru je znám. Testujeme na hladině významnosti α.

Testová statistika pro hypotézu je definována jako:

$$T = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad T \sim N(0,1)$$

1. Oboustranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu \neq \mu_0$,
   (c) kritický obor $W = (-\infty\,; -z_{\alpha/2}) \cup (z_{\alpha/2}\,; \infty)$,
   (d) obor přijetí $OP = (-z_{\alpha/2}\,;\, z_{\alpha/2})$
2. Levostranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu < \mu_0$,
   (c) kritický obor $W = (-\infty\,; -z_{\alpha})$,
   (d) obor přijetí $OP = (z_{\alpha}\,; \infty)$
3. Pravostranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu > \mu_0$,
   (c) kritický obor $W = (z_{\alpha}\,; \infty)$,
   (d) obor přijetí $OP = (-\infty\,; z_{\alpha})$

SCILAB

```
[p_hodnota,T,z_alpha]=z_test(Mu0,stredni_hodnota,rozptyl,n,strana,alpha)
  p_hodnota ... p-hodnota
  T ... statistika náhodné veličiny
  z_alpha ... kritická hodnota
  Mu0 ... předpokládaná střední hodnota
  stredni_hodnota ... střední hodnota výběru
  rozptyl ... rozptyl souboru (musí být zadán)
  n ... počet prvků výběru
  strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
  alpha ... hladina významnosti
```

Příklad. Ze souboru ocelových nosníků stejné nominální délky 6,5 m jsme náhodně vybrali 6 ks. Výrobce zaručuje, že rozptyl délek nosníků je menší než 0,1 m. Naměřili jsme následující data

$$6,2 \quad 7,5 \quad 6,9 \quad 8,9 \quad 6,4 \quad 7,1$$

Na hladině významnosti $\alpha = 0,1$ testujte tvrzení výrobce, že ocelové nosníky mají průměrnou délku 6,5 m.

[1] pokud je hodnot v náhodném výběru méně než 30, musí pocházet z normálního rozdělení $N(\mu, \sigma^2)$

1. Teorie

   oboustranný test, $\alpha = 0,1$, $\sigma^2 = 0,1$, $\mu_0 = 6,5$

   (a) $\bar{x} = \frac{1}{6}(6,2 + 7,5 + 6,9 + 8,9 + 6,4 + 7,1) = 7,1667$

   $$T = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{7,1667 - 6,5}{\frac{\sqrt{0,1}}{\sqrt{6}}} = 5,1642$$

   (b) $z_{\alpha/2} = 1,645$
   $W = (-\infty, -1,645) \bigcup (1,645, \infty)$

   (c) $T \in W \Rightarrow$ Zamítáme tvrzení výrobce, že průměrná délka nosníků je 6,5 m.

2. Scilab

```
x=[6.2 7.5 6.9 8.9 6.4 7.1];
str_hodnota=mean(x);
n=size(x,2);
[p_hodnota,T,z_alpha]=z_test(6.5,str_hodnota,0.1,n,'o',0.1)

Kritický obor W=(-Inf,-1.644854)U(1.644854,Inf)
Hodnota testové statistiky T=5.163978
Hypotézu zamítáme

z_alpha = 1.6448536
T = 5.1639778
P_hodnota = 2.403D-8
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.1.2 Test hypotéz o střední hodnotě μ při neznámém rozptylu σ^2

TEORIE

Na základě náhodného výběru ze souboru s libovolným rozdělením[2] testujeme předpoklad, že střední hodnota náhodné veličiny μ se rovná nějaké námi předpokládané střední hodnotě μ_0. Rozptyl souboru není znám. Testujeme na hladině významnosti α.

Testová statistika pro hypotézu je definována jako:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad T \sim St\,(n-1)$$

1. Oboustranný test hypotéz o střední hodnotě μ při neznámém rozptylu souboru σ^2 (známém rozptylu výběru s^2) má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu \neq \mu_0$,
   (c) kritický obor $W = \left(-\infty\,;\, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-t_{\alpha/2}\,;\, t_{\alpha/2}\right)$

2. Levostranný test hypotéz o střední hodnotě μ při neznámém rozptylu souboru σ^2 (známém rozptylu výběru s^2) má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu < \mu_0$,
   (c) kritický obor $W = (-\infty\,;\, -t_{\alpha})$,
   (d) obor přijetí $OP = (t_{\alpha}\,;\, \infty)$

3. Pravostranný test hypotéz o střední hodnotě μ při neznámém rozptylu souboru σ^2 (známém rozptylu výběru s^2) má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu = \mu_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu > \mu_0$,
   (c) kritický obor $W = (t_{\alpha}\,;\, \infty)$,
   (d) obor přijetí $OP = (-\infty\,;\, t_{\alpha})$

SCILAB

```
[p_hodnota,T,t_alpha]=t_test(Mu0,stredni_hodnota,rozptyl,n,strana,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   t_alpha ... kritická hodnota
   Mu0 ... předpokládaná střední hodnota
   stredni_hodnota ... střední hodnota výběru
   rozptyl ... rozptyl výběru
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Ze souboru ocelových nosníků stejné nominální délky jsme provedli náhodný výběr 50 nosníků a vypočetli průměr $\bar{x} = 5,77$ m a směrodatnou odchylku s=0,8. Na 95% hladině významnosti testujte tvrzení výrobce, že nominální délka nosníků není větší než 6 m.

1. Teorie

[2] pokud je hodnot v náhodném výběru méně než 30, musí pocházet z normálního rozdělení $N\left(\mu, \sigma^2\right)$

(a) $n = 50$,$\bar{x} = 5,77$, $s = 0,8$, $\alpha = 0,05$, $\mu_0 = 6$

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{5,77 - 6}{\frac{0,8}{\sqrt{50}}} = -2,0329$$

(b) $t_\alpha = 1,676$
$W = (1,676\,,\,\infty)$

(c) $T \notin W \Rightarrow$ Nezamítáme tvrzení výrobce, že průměrná délka nosníků není větší než 6 m.

2. Scilab

```
[p_hodnota,T,t_alpha]=t_test(6,5.77,0.8^2,50,'p',0.05)

Kritický obor W=(1.676551,Inf)
Hodnota testové statistiky T=-2.032932
Hypotézu nezamítáme

t_alpha = 1.6765509
T = - 2.032932
P_hodnota = 0.9762521
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.1.3 Test hypotéz o podílu π

Nechť p je odhad podílu pravděpodobnosti úspěchu v souboru na základě výběru o rozsahu n. Testujeme předpoklad, že pravděpodobnost úspěchu p je rovna nějaké námi předpokládané pravděpodobnosti úspěchu p_0. Testujeme na hladině významnosti α.

Testová statistika pro test hypotéz o podílu je definována takto

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad T \sim N(0,1)$$

TEORIE

1. Oboustranný test hypotéz o podílu π má:
   (a) nulovou hypotézu definovanou jako $H_0: \pi = p_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \pi \neq p_0$,
   (c) kritický obor $W = (-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; \infty)$,
   (d) obor přijetí $OP = (-z_{\alpha/2}; z_{\alpha/2})$
2. Levostranný test hypotéz o podílu π má:
   (a) nulovou hypotézu definovanou jako $H_0: \pi = p_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \pi < p_0$,
   (c) kritický obor $W = (-\infty; -z_{\alpha})$,
   (d) obor přijetí $OP = (z_{\alpha}; \infty)$
3. Pravostranný test hypotéz o podílu π má:
   (a) nulovou hypotézu definovanou jako $H_0: \pi = p_0$,
   (b) alternativní hypotézu definovanou jako $H_A: \pi > p_0$,
   (c) kritický obor $W = (z_{\alpha}; \infty)$,
   (d) obor přijetí $OP = (-\infty; z_{\alpha})$

SCILAB

```
[p_hodnota,T,z_alpha]=prop_test(p0,p,n,strana,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   p0 ... předpokládaná hodnota pravděpodobnosti
   p ... pravděpodobnost úspěchu nebo počet úspěchů ve výběru
   n ... počet prvků výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Ve výběru z výrobků o rozsahu 100 bylo nalezeno 12 vadných. Je tato skutečnost v souladu s tvrzením, že v produkci je nejvýše 5% vadných výrobků. Testujte na hladině významnosti 0,05.

1. Teorie
   (a) $n = 100$, $p = 12$, $p_0 = 0,05$, $\alpha = 0,05$

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,12 - 0,05}{\sqrt{\frac{0,05(1-0,05)}{100}}} = 3,2218$$

(b) $t_\alpha = 1,645$
$W = (1,645\,,\,\infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení výrobce, že v produkci je nejvýše 5% vadných výrobků.

2. Scilab

```
[p_hodnota,T,z_alpha]=prop_test(0.05,12,100,'p',0.05)

Kritický obor W=(1.644854,Inf)
Hodnota testové statistiky T=3.211820
Hypotézu zamítáme

z_alpha = 1.6448536
T = 3.2118203
p_hodnota = 0.0006595
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.1.4 Test hypotéz o rozptylu σ^2

TEORIE

Na základě náhodného výběru ze souboru s libovolným rozdělením[3] testujeme předpoklad, že rozptyl náhodné veličiny σ^2 se rovná nějaké námi předpokládané hodnotě rozptylu σ_0^2. Testujeme na hladině významnosti α.

Testová statistika pro test hypotéz o rozptylu je definována takto

$$T = \frac{(n-1)\,s^2}{\sigma_0^2} \quad T \sim \chi^2\,(n-1)$$

1. Oboustranný test hypotéz o rozptylu souboru σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma^2 = \sigma_0^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma^2 \neq \sigma_0^2$,
   (c) kritický obor $W = \left(0\,;\,\chi^2_{1-\alpha/2}\right) \cup \left(\chi^2_{\alpha/2}\,;\,\infty\right)$,
   (d) obor přijetí $OP = \left(\chi^2_{1-\alpha/2}\,;\,\chi^2_{\alpha/2}\right)$

2. Levostranný test hypotéz o rozptylu souboru σ^2má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma^2 = \sigma_0^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma^2 < \sigma_0^2$,
   (c) kritický obor $W = \left(0\,;\,\chi^2_{1-\alpha}\right)$,
   (d) obor přijetí $OP = \left(\chi^2_{1-\alpha}\,;\,\infty\right)$

3. Pravostranný test hypotéz o rozptylu souboru σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma^2 = \sigma_0^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma^2 > \sigma_0^2$,
   (c) kritický obor $W = \left(\chi^2_{\alpha}\,;\,\infty\right)$,
   (d) obor přijetí $OP = \left(0\,;\,\chi^2_{\alpha}\right)$

SCILAB

```
[p_hodnota,T]=var_test(sigma0,rozptyl,strana,n,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   sigma0 ... předpokládaná hodnota rozptylu
   rozptyl ... rozptyl výběru
   n ... počet prvků výběru/prvky výběru
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Ze souboru ocelových nosníků stejné nominální délky 6,5 m jsme náhodně vybrali 6 ks. Výrobce zaručuje, že rozptyl délek nosníků je menší než 0,1 m. Naměřili jsme následující data

$$6,2 \quad 7,5 \quad 6,9 \quad 8,9 \quad 6,4 \quad 7,1$$

Na hladině významnosti $\alpha = 0,1$ testujte tvrzení výrobce, že rozptyl délek nosníků je menší než 0,1.

1. Teorie

   oboustranný test, $\alpha = 0,1$, $\sigma^2 = 0,1$, $\mu_0 = 6,5$

[3] pokud je hodnot v náhodném výběru méně než 30, musí pocházet z normálního rozdělení $N\left(\mu, \sigma^2\right)$

(a) $\bar{x} = \frac{1}{6}(6,2 + 7,5 + 6,9 + 8,9 + 6,4 + 7,1) = 7,1667$
$s^2 = \frac{1}{5}\left((6,2 - 7,1667)^2 + (7,5 - 7,1667)^2 + (6,9 - 7,1667)^2 + (8,9 - 7,1667)^2 + (6,4 - 7,1667)^2 + (7,1 - 7,1667)^2\right) = 0,9427$

$$T = \frac{(n-1)\,s^2}{\sigma_0^2} = \frac{(6-1)\,0,9427}{0,1} = 47,135$$

(b) $\chi^2_\alpha = 9,236$
$W = (9,236\,,\,\infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení výrobce, že rozptyl délek nosníků je menší než 0,1.

2. Scilab

```
x=[6.2 7.5 6.9 8.9 6.4 7.1];
rozptyl=variance(x);
n=size(x,2);
[p_hodnota,T]=var_test(0.1,rozptyl,n,'p',0.1)

Kritický obor W=(9.236357,Inf)
Hodnota testové statistiky T=47.133333
Hypotézu zamítáme

T = 47.133333
p_hodnota = 5.337D-09
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.1.5 Test mediánu (znaménkový test)

TEORIE

Speciální případ znaménkového testu je test **mediánu**, kdy výběr X porovnáváme s neznámým mediánem $\tilde{x}_{0,5}$.
Předpokládejme, že máme výběr X a testujeme shodu se zadaným mediánem $\tilde{x}_{0,5}$. Vypočteme rozdíl

$$D_i = X_i - \tilde{x}_{0,5}, \quad i = 1, 2, \ldots, n$$

a písmenem b označíme počet kladných D_i. b je statistika s binomickým rozdělením, tzn. $b \sim Bi\,(n, \pi)$.
Testová statika má tvar

$$T = \frac{2b - n}{\sqrt{n}} \quad T \sim N(0;1)$$

a lze ji testovat pomocí z-testu.

1. Oboustranný znaménkový test má:
   (a) nulovou hypotézu definovanou jako $H_0 :\ X = \tilde{x}_{0,5}$,
   (b) alternativní hypotézu definovanou jako $H_A :\ X \neq \tilde{x}_{0,5}$,
   (c) kritický obor $W = \left(-\infty\,;\,-z_{\alpha/2}\right) \cup \left(z_{\alpha/2}\,;\,\infty\right)$,
   (d) obor přijetí $OP = \left(-z_{\alpha/2}\,;\,z_{\alpha/2}\right)$

2. Levostranný znaménkový test má:
   (a) nulovou hypotézu definovanou jako $H_0 :\ X = \tilde{x}_{0,5}$,

(b) alternativní hypotézu definovanou jako $H_A: \ X < \tilde{x}_{0,5}$,

(c) kritický obor $W = (-\infty\,;\, -z_\alpha)$,

(d) obor přijetí $OP = (z_\alpha\,;\, \infty)$

3. Pravostranný znaménkový test má:

(a) nulovou hypotézu definovanou jako $H_0: \ X = \tilde{x}_{0,5}$,

(b) alternativní hypotézu definovanou jako $H_A: \ X > \tilde{x}_{0,5}$,

(c) kritický obor $W = (z_\alpha\,;\, \infty)$,

(d) obor přijetí $OP = (-\infty\,;\, z_\alpha)$

SCILAB

```
[p_hodnota,T,z_alpha]=sign_test(X,Y,strana,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   X ... náhodný výběr č. 1
   Y ... medián
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

6.1.6 Wilcoxonův test

Je to neparametrický test mediánu pro jednu veličinu nebo test shody dvou veličin s párovými výběry. Jedná se o alternativu k jedno-výběrovému nebo dvouvýběrovému párovému t-testu. Není vyžadována normalita. Jedná se o neparametrický test (bez předpokladu normality).

SCILAB

```
[pv,W,kr1,kr2]=wilcoxon_test(x1,x2,sm,al)
   pv ... p-hodnota
   W ...
   kr1 ...
   kr2 ...
   x1,x2 ... náhodné výběry
   sm ... směr testu
   al ... hladina významnosti α
```

6.1.7 Test normality (triviální)

Testuje, zda výběr pochází z náhodné veličiny s normálním rozdělením. Je velmi jednoduchý a ne moc spolehlivý.

SCILAB

```
[pv,q,krD,krH]=norm_test(x,al)
   pv ... p-hodnota
   q ...
   krD ...
   krH ...
   x ... náhodný výběr
   al ... hladina významnosti α
```

6.1.8 χ^2 test normality

Testuje, zda výběr pochází z náhodné veličiny s normálním rozdělením. Je to velmi účinný test, použitelný jak pro diskrétní, tak i pro spojitou veličinu.
Předpoklad: všechny četnosti větší než 0, jen 20% menších než 5. Jedná se o neparametrický test (bez předpokladu normality).

SCILAB

```
pv=normCh2_test(x,n)
   pv ... p-hodnota
   x ... náhodný výběr
   n ... počet hodnot (intervalů)
```

6.1.9 KS test

Test typu rozdělení (Kolmogorovův-Smirnovův test)

TEORIE

Tento test slouží k ověření, zda se rozdělení náhodné veličiny liší od určitého teoretického rozdělení. Je založen na porovnání distribuční funkce $F(x)$ testovaného rozdělení a výběrové distribuční funkce $F_n(x)$, určené z výběru o rozsahu n.

Statistika testu je definována vztahem

a má různá rozdělení podle typu testované distribuční funkce. Nulová hypotéza H_0je, že náhodná veličina má testované rozdělení.

Tento test by měl předcházet všechny testy, které vyžadují normalitu dat případně jiné rozdělení.

Statistika je

$$T = \sup_{x_i \in X} |F_n(x_i) - F(x_i)|$$

kde $F_n(x)$ je teoretická distribuční funkce testovaného rozdělení a $F(x)$ je distribuční funkce náhodné veličiny a *sup* je supremum, tedy soubor vzdáleností.

Test typu rozdělení - Kolmogorovův-Smirnovův test je oboustranný a má:

1. nulovou hypotézu definovanou jako H_0 : náhodná veličina má testované rozdělení,
2. alternativní hypotézu definovanou jako H_A : náhodná veličina nemá testované rozdělení,
3. kritický obor W, je podle testovaného rozdělení,
4. obor přijetí OP, je podle testovaného rozdělení.

SCILAB

1. KS test pro spojitá rozdělení

```
[p_hodnota,T,z_alpha]=ks_test_spojity(F,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   F ... vektor hodnot distribuční funkce spočtených v hodnotách dat
   alpha ... hladina významnosti
```

6.1.10 χ^2 test - test dobré shody

TEORIE

Testy dobré shody testují, zda, na hladině významnosti α, data pocházejí ze zvoleného rozdělení. Zjistíme četnosti dat O_i, $i = 1, 2, ..., n$ v jednotlivých námi zvolených intervalech a teoretické četnosti dat E_i, $i = 1, 2, ..., n$, které by měly v jednotlivých intervalech být, pokud by data měla testované rozdělení. Testová statistika má tvar:

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad T \sim Chi^2(n-1)$$

pokud jsou splněny následující předpoklady: (i) všechny očekávané četnosti jsou alespoň rovny 1, (ii) nejvýše 20% očekávaných četností je menší než 5.

Statistika měří vzdálenost mezi pozorovanými a teoretickými četnostmi (je nezáporná). Jsou-li četnosti stejné, rovná se nule. Čím více jsou četnosti jiné, tím je hodnota statistiky větší.

Test dobré shody má:

1. nulovou hypotézu definovanou jako H_0 : *data mají testované rozdělení*,
2. alternativní hypotézu definovanou jako H_A : *data nemají testované rozdělení*,
3. kritický obor $W = \left(\chi^2_\alpha\,;\,\infty\right)$,
4. obor přijetí $OP = \left(0\,;\,\chi^2_\alpha\right)$

SCILAB

```
[p_hodnota,T,alpha]=chisquare_test(O,E,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   O ... pozorované četnosti
   E ... očekávané četnosti/očekávané pravděpodobnosti četností/délky intervalů
   alpha ... hladina významnosti
```

Příklad. Rodiče s krevní skupinou AB mají děti s krevními skupinami AA, AB, BB. Jestliže hypotéza o dědičnosti podle Mendela je pravdivá, pak by se u potomků měly tyto krevní skupiny vyskytovat v poměru 25%, 50% a 25%. Následující tabulka ukazuje krevní skupiny u 284 dětí jejichž rodiče měli krevní skupiny AB.

skupina	AA	AB	BB
počet	65	152	67

Potvrzují tato data na hladině významnosti 0,05 Mendelovu hypotézu?

1. Teorie

   (a) pravostranný test, $\alpha = 0,05$, n=3

skupina	O_i	p_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
AA	65	0,25	$0,25 \cdot 284 = 71$	$(65-71)^2 = 36$	$\frac{36}{71}$
AB	152	0,50	$0,50 \cdot 284 = 142$	$(152-142)^2 = 100$	$\frac{100}{142}$
BB	67	0,25	$0,25 \cdot 284 = 71$	$(67-71)^2 = 16$	$\frac{16}{71}$
$\sum$	284	1	284		$\frac{204}{142}$

$$\mathrm{T} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{204}{141} = 1,44$$

   (b) $\chi^2_\alpha = 5,991$
   $W = (5,991\,,\,\infty)$

   (c) $T \notin W \Rightarrow$ Nezamítáme Mendelovu hypotézu, že krevní skupiny potomků by se měly vyskytovat v zadaném poměru.

2. Scilab

```
[T,p_hodnota,alpha]=chisquare_test([65 152 67],[0.25 0.5 0.25])

alpha = 0.05
T = 1.4366197
p_hodnota = 0.4875756
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2 Testy dvou veličin

6.2.1 Test hypotéz o shodě dvou středních hodnot $\mu_1 = \mu_2$ při známém rozptylu σ_1^2, σ_2^2

TEORIE

Předpokládejme, že máme dva nezávislé náhodné výběry o rozsahu n_1 resp. n_2, kdy oba pochází z normálního rozdělení $N\left(\mu_1, \sigma_1^2\right)$ resp. $N\left(\mu_2, \sigma_2^2\right)$. Testujeme shody dvou středních hodnot těchto výběrů, tedy $\mu_1 = \mu_2$, za předpokladu, že oba souborové rozptyly σ_1^2 a σ_2^2 jsou známy. Testujeme na hladině významnosti α.

Tento test se používá i pro test hypotéz o shodě dvou středních hodnot $\mu_1 = \mu_2$ při neznámém rozptylu σ_1^2, σ_2^2, pokud jsou rozsahy n_1 a $n_2$dostatečně velké.

Testová statistika pro hypotézu je definována jako:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{R_1 + R_2}} \quad T \sim N(0,1)$$

kde $R_1 = \frac{\sigma_1^2}{n_1}$, $R_2 = \frac{\sigma_2^2}{n_2}$ a

1. Oboustranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 \neq \mu_2$,
   (c) kritický obor $W = \left(-\infty\,;\, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-z_{\alpha/2}\,;\, z_{\alpha/2}\right)$

2. Levostranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 < \mu_2$,
   (c) kritický obor $W = (-\infty\,;\, -z_\alpha)$,
   (d) obor přijetí $OP = (z_\alpha\,;\, \infty)$

3. Pravostranný test hypotéz o střední hodnotě μ při známém rozptylu σ^2 má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 > \mu_2$,
   (c) kritický obor $W = (z_\alpha\,;\, \infty)$,
   (d) obor přijetí $OP = (-\infty\,;\, z_\alpha)$

SCILAB

```
[p_hodnota,T,z_alpha]=z_test_2(stredni_hodnota_1,rozptyl_1,n1,stredni_hodnota_2,rozptyl_2,
                               n2,strana,alpha)
  p_hodnota ... p-hodnota
  T ... statistika náhodné veličiny
  z_alpha ... kritická hodnota
  str_hod_1 ... střední hodnota výběru č. 1
  rozptyl_1 ... rozptyl souboru č. 1
  n1 ... počet prvků výběru č. 1
  str_hod_2 ... střední hodnota výběru č. 2
  rozptyl_2 ... rozptyl souboru č. 2
  n2 ... počet prvků výběru č. 2
  strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
  alpha ... hladina významnosti
```

Příklad. Lékaři chtějí zjistit, zda nově vyvinutý lék má vliv i na pozornost při řízení motorového vozidla. Testovací osoby rozdělili do dvou skupin. První skupině, ve které je 900 osob podají nový lék, druhé skupině, ve které je 1000 osob podají placebo. První skupina zvládla test v průměru na $\bar{x}_1 = 9,78$ bodů, se směrodatnou odchylkou $\sigma_1 = 4,05$. Druhá skupina zvládla test v průměru na $\bar{x}_2 = 15,10$ bodů se směrodatnou odchylkou $\sigma_2 = 4,28$. Na hladině významnosti $\alpha = 0,05$ testujte tvrzení, že nový lék neovlivňuje schopnost řídit motorové vozidlo.

Pozn. V tomto případě je směrodatná odchylka vypočtena z výběru, ale protože je vypočtena z takto obrovského množství dat, je možno jí brát jako směrodatnou odchylku souboru a pro testování se použije z_test_2.

1. Teorie

   (a) oboustranný test, $\alpha = 0,05$, $n_1 = 900$,$\bar{x}_1 = 9,78$, $\sigma_1 = 4,05$, $n_2 = 1000$,$\bar{x}_2 = 15,10$, $\sigma_2 = 4,28$.
   $R_1 = \frac{\sigma_1^2}{n_1} = \frac{4,05^2}{900} = 0,018$,
   $R_2 = \frac{\sigma_2^2}{n_2} = \frac{4,28^2}{1000} = 0,018$

   $$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{R_1 + R_2}} = \frac{-5,32}{\sqrt{0,018 + 0,018}} = -28$$

   (b) $z_{\alpha/2} = 1,960$
   $W = (-\infty\,,\, -1,960)\,(1,960\,,\, \infty)$

   (c) $T \in W \Rightarrow$ Zamítáme tvrzení, že nový lék neovlivňuje pozornost při řízení motorového vozidla.

2. Scilab

```
[p_hodnota,T,z_alfa]=z_test_2(9.78,4.05^2,900,15.1,4.28^2,1000,'o')


z_alpha = 1.959964
T = - 27.829612
p_hodnota = 1.90D-170
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.2 Nezávislý T-test

TEORIE

Předpokládejme, že máme dva náhodné výběry o rozsahu n_1 resp. n_2, kdy oba pochází z normálního rozdělení $N\left(\mu_1, \sigma_1^2\right)$ resp. $N\left(\mu_2, \sigma_2^2\right)$. Testujeme shody dvou středních hodnot těchto výběrů, tedy $\mu_1 = \mu_2$, za předpokladu, že oba souborové rozptyly σ_1^2 a σ_2^2 neznáme. Testujeme na hladině významnosti α.

1. Oboustranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 \neq \mu_2$,
   (c) kritický obor $W = \left(-\infty\,;\, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-t_{\alpha/2}\,;\, t_{\alpha/2}\right)$
2. Levostranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 < \mu_2$,
   (c) kritický obor $W = (-\infty\,;\, -t_\alpha)$,
   (d) obor přijetí $OP = (t_\alpha\,;\, \infty)$
3. Pravostranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 > \mu_2$,
   (c) kritický obor $W = (t_\alpha\,;\, \infty)$,
   (d) obor přijetí $OP = (-\infty\,;\, t_\alpha)$

Za předpokladu, že náhodné výběry jsou nezávislé a nepárové, je testová statistika pro hypotézu definována jako:

$$T \quad = \quad \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{R_1 + R_2}} \quad T \sim St\left(\nu\right)$$

kde $R_1 = \frac{s_1^2}{n_1}$, $R_2 = \frac{s_2^2}{n_2}$ a

$$\nu \quad = \quad \frac{\left(R_1 + R_2\right)^2}{\frac{R_1^2}{n_1 - 1} + \frac{R_2^2}{n_2 - 1}}$$

SCILAB

```
[p_hodnota,T,t_alpha]=t_test_2n(stredni_hodnota_1,rozptyl_1,n1,stredni_hodnota_2,rozptyl_2,
n2,strana,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   str_hod_1 ... střední hodnota výběru č. 1
   rozptyl_1 ... rozptyl výběru č. 1
   n1 ... počet prvků výběru č. 1
   str_hod_2 ... střední hodnota výběru č. 2
   rozptyl_2 ... rozptyl výběru č. 2
   n2 ... počet prvků výběru č. 2
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Je třeba porovnat dva technologické postupy A a B. Proto je 7 výrobků zhotoveno technologií A a 6 technologií B. Pro proměření stejné charakteristiky na všech výrobcích máme porovnat kvalitu obou technologií (která je dána průměrnou hodnotou měřené charakteristiky). Testujte shodu obou technologií na hladině významnosti $\alpha = 0,05$. Naměřené údaje jsou

technologie A	62	54	55	60	53	58	57
technologie B	52	52	49	50	51	52	

1. Teorie

   (a) oboustranný test, $\alpha = 0,05$, $n_1 = 7$, $n_2 = 6$
   $\bar{x}_1 = \frac{1}{7}(62 + 54 + 55 + 60 + 53 + 58 + 57) = 57$
   $\bar{x}_2 = \frac{1}{6}(52 + 52 + 49 + 50 + 51 + 52) = 51$
   $s_1^2 = \frac{1}{6}\left((62-57)^2 + (54-57)^2 + \ldots + (57-57)^2\right) = 10,67$
   $s_2^2 = \frac{1}{5}\left((52-51)^2 + (52-51)^2 + \ldots + (52-51)^2\right) = 1{,}6$
   $R_1 = \frac{\sigma_1^2}{n_1} = \frac{10,67}{7} = 1,524,$
   $R_2 = \frac{\sigma_2^2}{n_2} = \frac{1,6}{6} = 0,267$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{R_1 + R_2}} = \frac{57-51}{\sqrt{1,524 + 0,267}} = 4,484$$
$$\nu = \frac{(R_1 + R_2)^2}{\frac{R_1^2}{n_1 - 1} + \frac{R_2^2}{n_2 - 1}} = \frac{(1,524 + 0,267)^2}{\frac{1,524^2}{6} + \frac{0,267^2}{5}} = 7,998 \doteq 8$$

   (b) $t_{\alpha/2} = 2,306,$
   $W = (-\infty\,,\ -2,306)\,(2,306\,,\ \infty)$

   (c) $T \in W \Rightarrow$ Zamítáme tvrzení, že pevnost materiálu je stejná..

2. Scilab

```
x=[62 54 55 60 53 58 57];
y=[52 52 49 50 51 52];
SH1=mean(x);
SH2=mean(y);
R1=variance(x);
R2=variance(y);
n1=size(x,2);
n2=size(y,2);
[p_hodnota,T,t_alpha]=t_test_2n(SH1,R1,n1,SH2,R2,n2,'o',0.05)

t_alpha = 2.3060041
T = 4.4840142
p_hodnota = 0.0020449
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.3 Párový T-test

TEORIE

Předpokládejme, že máme dva náhodné výběry o rozsahu n_1 resp. n_2, kdy oba pochází z normálního rozdělení $N\left(\mu_1, \sigma_1^2\right)$ resp. $N\left(\mu_2, \sigma_2^2\right)$. Testujeme shody dvou středních hodnot těchto výběrů, tedy $\mu_1 = \mu_2$, za předpokladu, že oba souborové rozptyly σ_1^2 a σ_2^2 neznáme. Testujeme na hladině významnosti α.

1. Oboustranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 \neq \mu_2$,
   (c) kritický obor $W = \left(-\infty\,;\, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-t_{\alpha/2}\,;\, t_{\alpha/2}\right)$
2. Levostranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 < \mu_2$,
   (c) kritický obor $W = \left(-\infty\,;\, -t_{\alpha}\right)$,
   (d) obor přijetí $OP = \left(t_{\alpha}\,;\, \infty\right)$
3. Pravostranný T-test má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \mu_1 = \mu_2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \mu_1 > \mu_2$,
   (c) kritický obor $W = \left(t_{\alpha}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-\infty\,;\, t_{\alpha}\right)$

Za předpokladu, že náhodné výběry jsou závislé a testujeme střední hodnotu párových výběrů, je testová statistika pro hypotézu definována jako:

$$T = \frac{\bar{x}_d}{s_d}\sqrt{n} \quad T \sim St\left(n-1\right)$$

kde n je počet párů, $\bar{x}_d$ je průměr a s_d je směrodatná odchylka diferencí.

$$\bar{x}_d = \frac{\sum_{i=1}^{n}\left(d_{1,i} - d_{2,i}\right)}{n}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_{1,i} - d_{2,i}\right)^2}{n} - \bar{x}_d^2}$$

SCILAB

```
[p_hodnota,T,t_alpha]=t_test_2p(vyber1,vyber2,strana,alpha)
  p_hodnota ... p-hodnota
  T ... statistika náhodné veličiny
  t_alpha ... kritická hodnota
  vyber11 ... prvky výběru č. 1
  vyber2 ... prvky výběru č. 2
  strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
  alpha ... hladina významnosti
```

Příklad. Osm vzorků chemické látky jsme postupně analyzovali titrační metodou a polarograficky. Výsledky jsou v tabulce

Vzorek	1	2	3	4	5	6	7	8
Polarografická metoda	18,6	27,6	27,5	25,0	24,5	26,8	29,7	26,5
Titrační metoda	18,58	27,37	27,27	24,64	24,10	26,33	29,33	26,63

Zjistěte, zda na hladině významnosti $\alpha = 0,05$ dávají obě metody v průměru podobné výsledky.

1. Teorie

   (a) oboustranný test, $\alpha = 0,05$, $n = 8$

$$\bar{x}_d = \frac{\sum_{i=1}^{n}(d_{1,i} - d_{2,i})}{n} = \frac{(18,6 - 18,58) + (27,6 - 27,37) + \ldots + (26,5 - 26,63)}{8} = \frac{1,95}{8} = 0,24$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_{1,i} - d_{2,i})^2}{n} - \bar{x}_d^2} = \sqrt{\frac{(18,6 - 18,58)^2 + (27,6 - 27,37)^2 + \ldots + (26,5 - 26,63)^2}{8} - 0,24^2} =$$

$$= \sqrt{\frac{0,7705}{8} - 0,24^2} = \sqrt{0,039} = 0,20$$

$$T = \frac{\bar{x}_d}{s_d}\sqrt{n} = \frac{0,24}{0,20}\sqrt{8} = 3,39$$

   (b) $t_{\alpha/2} = 2,306$,
   $W = (-\infty\,,\, -2,365)\,(2,365\,,\, \infty)$

   (c) $T \in W \Rightarrow$ Zamítáme tvrzení, že obě metody dávají v průměru podobné výsledky.

2. Scilab

```
vyber1=[18.6 27.6 27.5 25.0 24.5 26.8 29.7 26.5]
vyber2=[18.58 27.37 27.27 24.64 24.10 26.33 29.33 26.63]

[p_hodnota,T,t_alfa]=t_test_2p(vyber1,vyber2,'o')

t_alfa = 2.3646243
T = 3.3572962
p_hodnota = 0.0121291
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.4 Test hypotéz o shodě dvou podílů $p_1 = p_2$

TEORIE

Nechť p_1 resp. p_2 je odhad podílu pravděpodobnosti úspěchu v souboru na základě výběru o rozsahu n_1 resp. n_2. Testujeme předpoklad, že pravděpodobnost úspěchu p_1 je rovna pravděpodobnosti úspěchu p_2, tedy $p_1 = p_2$. Testujeme na hladině významnosti α.

Testová statistika pro hypotézu je definována jako:

$$T = \frac{p_1 - p_2}{\sqrt{p_P}} \quad T \sim N(0,1)$$

kde $p_1 = \frac{n_1^+}{n_1}$, $p_2 = \frac{n_2^+}{n_2}$ a $p_p = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$.

1. Oboustranný test hypotéz o shodě dvou podílů $p_1 = p_2$ má:
   (a) nulovou hypotézu definovanou jako $H_0 : \ p_1 = p_2$,
   (b) alternativní hypotézu definovanou jako $H_A : \ p_1 \neq p_2$,
   (c) kritický obor $W = \left(-\infty\,;\, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-z_{\alpha/2}\,;\, z_{\alpha/2}\right)$

2. Levostranný test hypotéz o shodě dvou podílů $p_1 = p_2$ má:
   (a) nulovou hypotézu definovanou jako $H_0 : \ p_1 = p_2$,
   (b) alternativní hypotézu definovanou jako $H_A : \ p_1 < p_2$,
   (c) kritický obor $W = (-\infty\,;\, -z_\alpha)$,
   (d) obor přijetí $OP = (z_\alpha\,;\, \infty)$

3. Pravostranný test hypotéz o shodě dvou podílů $p_1 = p_2$ má:
   (a) nulovou hypotézu definovanou jako $H_0 : \ p_1 = p_2$,
   (b) alternativní hypotézu definovanou jako $H_A : \ p_1 > p_2$,
   (c) kritický obor $W = (z_\alpha\,;\, \infty)$,
   (d) obor přijetí $OP = (-\infty\,;\, z_\alpha)$

SCILAB

```
[p_hodnota,T,z_alpha]=prop_test_2(p1,n1,p2,n2,strana,alpha)
  p_hodnota... p-hodnota
  T ... statistika náhodné veličiny
  z_alpha ... kritická hodnota
  p1 ... pravděpodobnost úspěchu nebo počet úspěchů ve výběru č. 1
  n1 ... počet prvků výběru č. 1
  p2 ... pravděpodobnost úspěchu nebo počet úspěchů ve výběru č. 2
  n2 ... počet prvků výběru č. 2
  strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
  alpha ... hladina významnosti
```

Příklad. Na křižovatce jsme opakovaně zaznamenávali počty vozidel jedoucích přímo a odbočujících vlevo nebo vpravo. Zjistili jsme, že přímo jelo 46 vozidel, vpravo 62 a vlevo 39. Na hladině významnosti $\alpha = 0,1$ testujte tvrzení, že podíly automobilů odbočujících doprava a doleva, vztažené ke všem vozidlům, která křižovatkou projela, jsou stejné.

1. Teorie

(a) oboustranný test, $n_1 = n_2 = 147$, $\alpha = 0,1$
$p_1 = \frac{62}{46+62+39} = \frac{62}{147} = 0,422$
$p_2 = \frac{39}{46+62+39} = \frac{39}{147} = 0,265$

$$p_p = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = \frac{0,422(1-0,422)}{147} + \frac{0,265(1-0,265)}{147} = \frac{0,244+0,195}{147} = 0,003$$

.

$$T = \frac{p_1 - p_2}{\sqrt{p_P}} = \frac{0,422-0,265}{\sqrt{0,003}} = \frac{0,157}{0,055} = 2,855$$

(b) $t_{(1+\alpha)/2} = 1,645$
$W = (-\infty\,,\,-1,645)\,(1,645\,,\,\infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení, že obě metody dávají v průměru podobné výsledky.

2. Scilab

```
vlevo=39;
vpravo=62;
celkem=46+62+39;
[p_hodnota,T,z_alfa]=prop_test_2(vpravo,celkem,vlevo,celkem,'o',0.1)

z_alfa = 1.6448536
p_hodnota = 2.8246339
T = 0.0047335
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.5 Test hypotéz o shodě dvou rozptylů $\sigma_1^2 = \sigma_2^2$

TEORIE

Na základě dvou náhodných výběrů s libovolným rozdělením[4] testujeme předpoklad, že rozptyl náhodné veličiny z prvního souboru σ_1^2 se rovná rozptylu z druhého souboru σ_2^2. Testujeme na hladině významnosti α.

$$T = \frac{s_1^2}{s_2^2} \quad T \sim F(n_1 - 1, n_2 - 1)$$

kde s_1^2 resp. s_2^2 jsou rozptyly dvou náhodných výběrů.

1. Oboustranný test hypotéz o shodě dvou rozptylů $\sigma_1^2 = \sigma_2^2$ má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma_1^2 = \sigma_2^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma_1^2 \neq \sigma_2^2$,
   (c) kritický obor $W = (F_\alpha\,; \infty)$,
   (d) obor přijetí $OP = (0\,; F_\alpha)$

2. Levostranný test hypotéz o shodě dvou rozptylů $\sigma_1^2 = \sigma_2^2$ má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma_1^2 = \sigma_2^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma_1^2 < \sigma_2^2$,
   (c) kritický obor $W = \left(0\,; \frac{1}{F_\alpha}\right)$,
   (d) obor přijetí $OP = \left(\frac{1}{F_\alpha}\,; \infty\right)$

3. Pravostranný test hypotéz o shodě dvou rozptylů $\sigma_1^2 = \sigma_2^2$ má:
   (a) nulovou hypotézu definovanou jako $H_0: \ \sigma_1^2 = \sigma_2^2$,
   (b) alternativní hypotézu definovanou jako $H_A: \ \sigma_1^2 > \sigma_2^2$,
   (c) kritický obor $W = (F_\alpha\,; \infty)$,
   (d) obor přijetí $OP = (0\,; F_\alpha)$

SCILAB

```
[p_hodnota,T,f_alpha]=var_test_2(Rozptyl_S1,n1,Rozptyl_S2,n2,strana,alpha)
   p_hodnota... p-hodnota
   T ... statistika náhodné veličiny
   Rozptyl_S1 ... výběrový rozptyl z výběru č. 1
   n1 ... počet prvků výběru č. 1
   Rozptyl_S2 ... výběrový rozptyl z výběru č. 2
   n2 ... počet prvků výběru č. 2
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Ze dvou souborů ocelových nosníků stejné nominální délky 6,5 m jsme náhodně vybrali 6 resp. 10 ks.

soubor č. 1	6,2	7,5	6,9	8,9	6,4	7,1				
soubor č. 2	6,5	6,8	7,2	5,9	7,5	8	6,2	5,8	6,5	6,3

Výrobce zaručuje, že rozptyl délek nosníků z obou souborů je stejný. Testujte toto tvrzení na hladině významnosti 0,05.

1. Teorie

[4] pokud je hodnot v náhodném výběru méně než 30, musí pocházet z normálního rozdělení $N(\mu, \sigma^2)$

(a) oboustranný test, $\alpha = 0,95$, $n_1 = 6$, $n_2 = 10$
$\bar{x}_1 = \frac{1}{6}(6,2 + 7,5 + 6,9 + 8,9 + 6,4 + 7,1) = 7,1667$
$s_1^2 = \frac{1}{5}\left((6,2 - 7,1667)^2 + (7,5 - 7,1667)^2 + (6,9 - 7,1667)^2 + (8,9 - 7,1667)^2 + (6,4 - 7,1667)^2 + (7,1 - 7,1667)^2\right) = 0,9427$
$\bar{x}_2 = \frac{1}{10}(6,5 + 6,8 + ... + 6,3) = 6,67$
$s_2^2 = \frac{1}{10}\left((6,5 - 6,67)^2 + (6,8 - 6,67)^2 + ... + (7,1 - 7,1667)^2\right) = 0,5023$

$$T = \frac{s_1^2}{s_2^2} = \frac{0,9427}{0,5023} = 1,8768$$

(a) $F_{\alpha/2} = 0,1497$, $F_{(1-\alpha)/2} = 4,4844$
$W = (-\infty\,,\, 0,1497) \bigcup (4,4844\,,\, \infty)$
(b) $T \notin W \Rightarrow$ Nezamítáme tvrzení výrobce, že rozptyl délek nosníků v obou souborech je stejný

2. Scilab

```
vektor1=[6.2 7.5 6.9 8.9 6.4 7.1]
vektor2=[6.5 6.8 7.2 5.9 7.5 8 6.2 5.8 6.5 6.3]
R1=variance(vektor1);
n1=size(vektor1,2);
R2=variance(vektor2);
n2=size(vektor2,2);
[p_hodnota,T,alpha]=var_test_2(R1,n1,R2,n2,'o')

alpha = 0.1496770   4.4844113
T = 1.876576
p_hodnota = 0.3882733
```

3. Rozdíl výsledků
Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.6 Mann-Whitneyův test

Nepárový neparametrický test shody dvou středních hodnot. Obdoba t-test pro dva nezávislé výběry.

SCILAB

```
[pv,U,Kr]=mannwhit_test(x1,x2)
   pv ... p-hodnota
   U ...
   Kr ...
   x1,x2 ... náhodný výběr č. 1 a č. 2
```

6.2.7 Znaménkový test

TEORIE

Předpokládejme, že máme dva párové výběry X a Y a testujeme shodu párových prvků výběru X a Y. Vypočteme rozdíl

$$D_i = X_i - Y_i, \quad i = 1, 2, \ldots, n$$

a písmenem b označíme počet kladných D_i. b je statistika s binomickým rozdělením, tzn. $b \sim Bi\,(n, \pi)$. Testová statika má tvar

$$T = \frac{2b - n}{\sqrt{n}} \quad T \sim N(0; 1)$$

a lze ji testovat pomocí z-testu.

Speciální případ znaménkového testu je test **mediánu**, kdy výběr X porovnáváme s neznámým mediánem $x_{0,5}$.

1. Oboustranný znaménkový test má:
   (a) nulovou hypotézu definovanou jako $H_0: \; X = Y$,
   (b) alternativní hypotézu definovanou jako $H_A: \; X \neq Y$,
   (c) kritický obor $W = \left(-\infty\,;\, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}\,;\, \infty\right)$,
   (d) obor přijetí $OP = \left(-z_{\alpha/2}\,;\, z_{\alpha/2}\right)$

2. Levostranný znaménkový test má:
   (a) nulovou hypotézu definovanou jako $H_0: \; X = Y$,
   (b) alternativní hypotézu definovanou jako $H_A: \; X < Y$,
   (c) kritický obor $W = (-\infty\,;\, -z_{\alpha})$,
   (d) obor přijetí $OP = (z_{\alpha}\,;\, \infty)$

3. Pravostranný znaménkový test má:
   (a) nulovou hypotézu definovanou jako $H_0: \; X = Y$,
   (b) alternativní hypotézu definovanou jako $H_A: \; X > Y$,
   (c) kritický obor $W = (z_{\alpha}\,;\, \infty)$,
   (d) obor přijetí $OP = (-\infty\,;\, z_{\alpha})$

SCILAB

```
[p_hodnota,T,z_alpha]=sign_test(X,Y,strana,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   X ... náhodný výběr č. 1
   Y ... náhodný výběr č. 2 / medián
   strana ... typ IS ('l'=levostranný, 'p'=pravostranný, 'o' oboustranný)
   alpha ... hladina významnosti
```

Příklad. Děti na základní škole skákaly do dálky z místa, na začátku a na konci školního roku. Na hladině významnosti $\alpha = 0,95$ testujeme tvrzení, že děti se ve školním roce ve skoku do dálky zlepšily. Délky skoků jsou v tabulce

	délka v cm										
začátek roku	135	157	180	164	200	220	198	111	164	175	180
konec roku	136	150	190	170	220	218	180	113	210	160	180

1. Teorie
   (a) levostranný test, $\alpha = 0,05$, n=11

$$D_i = X_i - Y_i, \quad i = 1, 2, \ldots, n$$

	délka v cm										
Y	135	157	180	164	200	220	198	111	164	175	180
X	136	150	190	170	220	218	180	113	210	160	180
D	+1	-7	+10	+9	+20	-2	-18	+2	+46	-15	0
	+	-	+	+	+	-	-	+	+	-	$\emptyset$

Počet kladných znamének b=6

$$T = \frac{2b - n}{\sqrt{n}} = \frac{2 \cdot 6 - 11}{\sqrt{11}} = \frac{1}{\sqrt{11}} = 0,30$$

(b) $z_\alpha = 1,64$
$W = (-\infty\,;\, -1,64)$

(c) $T \notin W \Rightarrow$ Nezamítáme tvrzení, že se děti ve skoku do dálky zlepšily během školního roku.

2. Scilab

```
vektor1=[135 157 180 164 200 220 198 111 164 175 180]
vektor2=[136 150 190 170 220 218 180 113 210 160 180]

[p_hodnota,T,z_alpha]=sign_test(vektor2,vektor1,'l')

z_alpha = -1.6448539
T = 0.301511
p_hodnota = 0.6184877
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.2.8 McNemarův test

Testuje, zda po provedení nějaké akce nastala změna. Pracuje s binárními daty: 1- má vlastnost, 0 - nemá vlastnost.
Analyzuje 2x2 tabulku četností veličiny před a veličiny po akci
Jedná se o neparametrický test (bez předpokladu normality).
Jde o párový test - před a po musí být stejné objekty.

SCILAB

```
pv=mcnemar_test(Tab)
   pv ... p-hodnota
   Tab ... tabulka četností 2x2 pro výběr před a výběr po.
```

6.3 Testy více veličin

6.3.1 Analýza rozptylu při jednoduchém třídění

Analýza rozptylu - ANOVA (Analysis of variance) je soubor statistických modelů použitých k analýze rozdílů mezi skupinami. Tyto rozdíly mohou být způsobeny působením různých podmínek (faktorů). Analýza rozptylu může být použita v případě, že máme více než dvě skupiny. V případě, že skupiny jsou právě dvě, použijeme pro porovnání jednodušší test dvou středních hodnot s neznámým rozptylem. Použití tohoto testu i analýzy rozptylu dává stejné výsledky.

Předpoklady k použití analýzy rozptylu:

1. výběry jsou vzájemně nezávislé,
2. výběry pochází ze souboru s normálním rozdělením $N\left(\mu, \sigma^2\right)$,
3. rozptyly uvnitř skupin jsou homogenní.

TEORIE

Analýza rozptylu při jednoduchém třídění (jedno-faktorová analýza rozptylu) je nejjednodušší případ analýzy rozptylu, kdy zkoumáme vliv jednoho faktoru na závisle proměnnou. Příkladem může být situace, kdy zkoumáme několik podobných zdrojů dat a chceme se přesvědčit, že všechny tyto zdroje fungují stejně, tj. průměry dat změřených na jednotlivých zdrojích jsou stejné.

Při analýze postupujeme následujícím postupem:

1. Z dat sestavíme kontingenční tabulku, kde sloupce budou jednotlivé třídy (počet tříd označíme a) a řádky budou jednotlivá měření.
2. Spočteme:
   (a) průměry dat od j-té třídy (průměry ze sloupců)

$$\bar{x}_{\cdot,j} = \frac{1}{b}\sum_{i=1}^{a} x_{i,j}$$

   kde $x_{i,j}$ je měření j-té třídy při i-té třídě (prvek tabulky na pozici (i,j))

   (b) průměry dat od j-té třídy (průměry ze sloupců)

$$\bar{x}_{i,\cdot} = \frac{1}{a}\sum_{j=1}^{a} x_{i,j}$$

   (c) $\bar{\bar{x}}$, kde $\bar{\bar{x}}$ je průměrná hodnota celé tabulky
3. Definujeme:
   (a) celkový součet čtverců změřených dat

$$S_C = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2$$

   (b) součet čtverců uvnitř tříd, který po normování dává rozptyl uvnitř tříd - tj. nevysvětlený rozptyl

$$S_U^2 = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot,j})^2$$

   normalizační faktor, stupně volnosti, je

$$df_U = N - a$$

(c) součet čtverců mezi třídami, která po normování dává rozptyl mezi třídami - tj. vysvětlený rozptyl

$$S_M^2 = \sum_{j=1}^{a} n_j \cdot (\bar{x}_j - \bar{\bar{x}})^2$$

normalizační faktor, stupně volnosti, je

$$df_M \quad = \quad a - 1$$

Testová statistika pro test ANOVA je definována jako

$$T = \frac{S_M^2}{S_U^2} \cdot \frac{N-1}{a-1} \quad T \sim F\,(a-1,\, N-1)$$

1. Pravostranný test ANOVA má:

   (a) nulovou hypotézu definovanou jako H_0 : střední hodnoty tříd jsou stejné,

   (b) alternativní hypotézu definovanou jako H_A : střední hodnoty tříd nejsou stejné,

   (c) kritický obor $W = (f_\alpha\,;\,\infty)$,

   (d) obor přijetí $OP = (0\,;\,f_\alpha)$.

SCILAB

```
[p_hodnota,T]=anova_1(y,n)
   p_hodnota... p-hodnota
   T ... statistika náhodné veličiny
   y ... matice dat - skupiny po sloupcích (zadat jen y)') nebo vektor vyberu uspořádaných za
sebou')
   n ... vektor delek jednotlivych vyberu')
```

Příklad. Sledujeme tři stroje. Náhodně zjišťujeme jejich hodinové produkce ($P1$, $P2$ a $P3$). Je pravdivé tvrzení, že na hladině významnosti $\alpha = 0,05$ jsou průměrné produkce všech tří strojů shodné?

$P1$	53	55	49	58	52	61	56	55
$P2$	49	56	52	45	51	56	44	51
$P3$	52	53	52	54	55	53	53	52

1. Teorie

   Teoretický výpočet je příliš složitý a zdlouhavý, proto není u této kapitoly uveden.

2. Scilab

```
y=[53 55 49 58 52 61 56 55; 49 56 52 45 51 56 44 51; 52 53 52 54 55 53 53 52];

[p_hodnota,T]=anova_1(y') //třídy musí být ve sloupcích, proto se transponuje

T = 3.3601453
p_hodnota = 0.0541908
```

Vyhodnocení: p_hodnota $> \alpha$ $\Rightarrow$Nezamítáme hypotézu, že průměrná produkce všech tří strojů je shodná.

6.3.2 Barlettův test

pre-analýza k testu ANOVA - testuje shodu rozptylů několika souborů.
Vyžaduje normalitu.

SCILAB

```
pv=batlett_test(L)
   pv ... p-hodnota
   L ... list výběrů
```

6.3.3 Scheffého test

post-analýza k testu ANOVA. V případě, že střední hodnoty nejsou stejné, určuje které z nich se liší. Výsledkem je tabulka (počet výběrů) x (počet výběrů). Jedničky indikují veličiny, které se odlišují.

SCILAB

```
[MgtF,M,F]=scheffe_test(L,al)
   MgtF ... tabulka, indikující odlišné veličiny
   M ...
   F ...
   L ... list výběrů
   al ... hladina významnosti
```

6.3.4 Analýza rozptylu při dvojném třídění

Analýza rozptylu - ANOVA (Analysis of variance) je soubor statistických modelů použitých k analýze rozdílů mezi skupinami. Tyto rozdíly mohou být způsobeny působením různých podmínek (faktorů). Analýza rozptylu může být použita v případě, že máme více než dvě skupiny. V případě, že skupiny jsou právě dvě, použijeme pro porovnání jednodušší test dvou středních hodnot s neznámým rozptylem. Použití tohoto testu i analýzy rozptylu dává stejné výsledky.

Předpoklady k použití analýzy rozptylu:

1. výběry jsou vzájemně nezávislé,
2. výběry pochází ze souboru s normálním rozdělením $N\left(\mu, \sigma^2\right)$,
3. rozptyly uvnitř skupin jsou homogenní.

TEORIE

Analýza rozptylu při dvojném třídění (dvou-faktorová analýza rozptylu) je analýza, kdy zkoumáme vliv dvou faktorů na závisle proměnnou. Příkladem může být situace, kdy zkoumáme několik podobných zdrojů dat a chceme se přesvědčit, zda rozdíly mezi daty jsou způsobeny působením dvou faktorů nebo zdroje jsou skutečně různé.

1. Z dat sestavíme kontingenční tabulku, kde ve sloupcích budou třídy jednoho faktoru (počet tříd ve sloupcích (počet sloupců) označíme a) a v řádcích budou třídy druhého faktoru (počet tříd v řádcích (počet řádků) označíme b)
2. Spočteme:
   (a) celkový počet měření (počet všech prvků tabulky)

$$N = \sum_{j=1}^{a}$$

   (b) průměry dat od j-té třídy (průměry ze sloupců)

$$\bar{x}_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$$

   kde $x_{i,j}$ je měření j-té třídy při i-tém měření (prvek tabulky na pozici (i,j))

   (c) průměr z průměrů pro jednotlivé třídy (celkový průměr tabulky)

$$\bar{\bar{x}} = \frac{1}{a}\sum_{i=1}^{a} \bar{x}_j$$

   kde $x_j = \left[\begin{array}{cccc} x_{1,j} & x_{2,j} & \ldots & x_{n,j} \end{array}\right]$ je měření od j-té třídy (j-tý sloupec tabulky).
3. Definujeme:
   (a) celkový rozptyl dat

$$S_C = \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \bar{\bar{x}})^2$$

(b) součet čtverců mezi průměry tříd ve sloupcích

$$S_A = b \cdot \sum_{i=1}^{a} (\bar{x}_{i,\cdot} - \bar{\bar{x}})^2$$

normalizační faktor, stupně volnosti, je

$$df_A = a - 1$$

(c) součet čtverců mezi průměry tříd v řádcích

$$S_B = a \cdot \sum_{j=1}^{b} (\bar{x}_{\cdot,j} - \bar{\bar{x}})^2$$

normalizační faktor, stupně volnosti, je

$$df_B = b - 1$$

(d) reziduální součet čtverců (uvnitř tříd)

$$S_U^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (x_{ij} - \hat{x}_{i,,j})^2$$

kde

$$\hat{x}_{i,j} = (\bar{x}_{\cdot,\mathrm{j}} - \bar{\bar{x}}) + (\bar{x}_{i,\cdot} - \bar{\bar{x}}) + \bar{\bar{x}}$$

a normalizační faktor, stupně volnosti, je

$$df_U = (a-1)(b-1)$$

Testová statistika pro test ANOVA pro sloupce je

$$T = \frac{(b-1)\, S_A^2}{S_U^2} \quad T \sim F\left(a-1,\, (a-1)(b-1)\right)$$

1. Pravostranný test ANOVA pro sloupce má:
   (a) nulovou hypotézu definovanou jako H_0 : střední hodnoty tříd v sloupcích jsou stejné,
   (b) alternativní hypotézu definovanou jako H_A : střední hodnoty tříd v sloupcích nejsou stejné,
   (c) kritický obor $W = (f_\alpha\,;\,\infty)$,
   (d) obor přijetí $OP = (0\,;\,f_\alpha)$.

Testová statistika pro test ANOVA pro řádky je

$$T = \frac{(a-1)\, S_B^2}{S_U^2} \quad T \sim F\left(b-1,\, (a-1)(b-1)\right)$$

.

$$F = \frac{(a-1)\, S_B}{S_U}.$$

Rozdělení této statistiky je Fisherovo $F\left(b-1,\, (a-1)(b-1)\right)$.

1. Pravostranný test ANOVA pro řádky má:
   (a) nulovou hypotézu definovanou jako H_0 : střední hodnoty tříd v řádcích jsou stejné,
   (b) alternativní hypotézu definovanou jako H_A : střední hodnoty tříd v řádcích nejsou stejné,
   (c) kritický obor $W = (f_\alpha\,;\,\infty)$,
   (d) obor přijetí $OP = (0\,;\,f_\alpha)$.

SCILAB

```
[pv_col, pv_row]=anova_2(s)
   pv_col ... p-hodnota pro sloupce (1. faktor)')
   pv_row ... p-hodnota pro řádky (2. faktor)')
   s ... matice dat - 1. faktor sloupce', 2. faktor radky
```

Příklad. V různých městech jsme se dotazovali mužů a žen, kolik hodin denně v průměru stráví za volantem. Odpovědi jsme zaznamenali do následující tabulky

	Praha	Plzeň	Brno	Ostrava	Cheb
ženy	50	35	48	32	16
muži	58	25	47	30	18

Na hladině významnosti $\alpha = 0,05$ testujte, zda muži i ženy ve zkoumaných městech stráví v průměru stejně času za volantem. V opačném případě určete, zda rozdíly jsou způsobeny rozdílností měst nebo typem řidiče.

1. Teorie

   Teoretický výpočet je příliš složitý a zdlouhavý, proto není u této kapitoly uveden.

2. Scilab

```
s=[50 35 48 32 16; 58 25 47 30 18];

[pv_sloupce, pv_radky]=anova_2(s)

pv_radky = 0.8475261
pv_sloupce = 0.0062780
```

   Vyhodnocení:

   - pv_radky $> \alpha$ $\Rightarrow$Nezamítáme hypotézu, že muži i ženy ve městech stráví v průměru stejně času za volantem.
   - pv_sloupce $< \alpha$ $\Rightarrow$Zamítáme hypotézu, že průměrná doba strávená za volantem je pro všechny města stejná.

6.3.5 Kruskal-Wallisův test

Neparametrická obdoba jedno-faktorové Anovy - bez předpokladu normality. Testuje shodu středních hodnot několika souborů.

```
pv=kruskal_test(L)
   pv ... p-hodnota
   L ... list výběrů
```

6.3.6 Friedmanův test

Neparametrická bloková obdoba jedno-faktorové Anovy. Testuje shodu středních hodnot několika souborů, ze kterých jsme provedli blokové výběry - to znamená např., že výběry jsou hodnocení několika posuzovatelů. Jsou provedeny tak, že v každém výběru je postupně hodnocení všech posuzovatelů. Hodnocení od posuzovatelů se nazývá blok.

```
pv=friedman_test(L)
   pv ... p-hodnota
   L ... list výběrů (v každém je postupně hodnocení všech posuzovatelů).
```

6.4 Testy nezávislosti

6.4.1 χ^2 test - test nezávislosti

TEORIE

Test nezávislosti pro dvě diskrétní veličiny používá kontingenční tabulku absolutních četností dvou náhodných veličin, jejichž nezávislost testujeme. Úkolem testu je rozhodnout, zda jsou obě náhodné veličiny na sobě závislé či nezávislé. Testová statistika má tvar

$$T = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-n_{ij}^{o}\right)^2}{n_{ij}^{o}} \quad T \sim Chi^2((r-1)(s-1))$$

kde r resp. s je počet řádků resp. sloupců, n_{ij} jsou pozorované četnosti a n_{ij}^o jsou očekávané četnosti. χ^2test nezávislosti lze použít pokud jsou splněny následující předpoklady: (i) všechny očekávané četnosti jsou alespoň rovny 1, (ii) nejvýše 20% očekávaných četností je menší než 5.

χ^2 test nezávislosti má:

1. nulovou hypotézu definovanou jako H_0 : náhodné veličiny jsou nezávislé,
2. alternativní hypotézu definovanou jako H_A : náhodné veličiny jsou závislé,
3. kritický obor $W = \left(\chi^2_\alpha\,;\,\infty\right)$,
4. obor přijetí $OP = \left(0\,;\,\chi^2_\alpha\right)$

SCILAB

```
[p_hodnota,T,chi_alpha]=chisquare_test_i(KT,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   KT ... kontingenční tabulka
   alpha ... hladina významnosti
```

Příklad. Zjišťovala se závislost mezi barvou vlasů a barvou očí u mužů. V náhodném výběru jsme se dotazovali 6800 mužů a získali jsme následující údaje

oči\vlasy	světlé	hnědé	hnědočerné	černé
modré	1768	807	189	47
šedé	946	1387	746	53
hnědé	115	438	288	16

Testujte nezávislost barvy vlasů a očí na hladině významnosti 0,05.

1. Teorie

   (a) pravostranný test, $\alpha = 0,05$, n=6800

oči\vlasy (E)	světlé	hnědé	hnědočerné	černé
modré	1768	807	189	47
šedé	946	1387	746	53
hnědé	115	438	288	16

$\Downarrow$

oči\vlasy	světlé	hnědé	hnědočerné	černé	$\sum$
modré	$\frac{1768}{6800}$ =0,260	0,119	0,028	0,007	0,414
šedé	0,139	0,204	0,110	0,008	0,461
hnědé	0,017	0,064	0,042	0,002	0,125
$\sum$	0,416	0,387	0,180	0,017	1

$$\Downarrow$$

oči\vlasy (O)	světlé	hnědé	hnědočerné	černé
modré	$0,416 \cdot 0,414 \cdot 6800 = 1171$	1090	507	48
šedé	$0,416 \cdot 0,461 \cdot 6800 = 1304$	1213	564	53
hnědé	$0,416 \cdot 0,125 \cdot 6800 = 354$	329	153	14

$$\mathrm{T} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_{ij}^{o}\right)^2}{n_{ij}^{o}} = \frac{(1768-1171)^2}{1171} + \frac{(807-1090)^2}{1090} + ... + \frac{(16-14)^2}{14} = 1073,5$$

1. $\chi^2_\alpha = 12,592$
   $W = (12,592\,,\,\infty)$

   (a) $T \in W \Rightarrow$ Zamítáme hypotézu o nezávislosti barvy vlasů na barvě očí.

2. Scilab

```
KT=[1768 807 189 47; 946 1387 746 53; 115 438 288 16]

[p_hodnota,T,alpha]=chisquare_test_i(KT)

alpha = 0. 12.591587
T = 1073.5076
p_hodnota = 1.12D-228
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.4.2 Test nezávislosti výběrů - Pearsonův test

TEORIE

Pearsonův korelační test nám říká, jak silná je lineární závislost mezi dvěma párovými výběry X a Y o rozsahu n. Pearsonův test lze použít v případě, že výběry pochází z dat s normálním rozdělením $N\left(\mu, \sigma^2\right)$[5].

Statistika je

$$T = r_P\sqrt{\frac{n-2}{1-r_P^2}} \quad T \sim St\left(n-2\right)$$

kde n je rozsah výběrů a r_P je výběrový korelační koeficient. Výběrový korelační koeficient r_p se spočte jako

$$r_P = \frac{s_{xy}^2}{\sqrt{s_x^2 \cdot s_y^2}}$$

kde s_{xy}^2 je kovariance mezi náhodnými veličinami X a Y a s_x^2 resp. s_y^2 je výběrový rozptyl náhodné veličiny X resp. Y.

Test nezávislosti prvků výběru - Pearsonův korelační test je oboustranný a má:

1. nulovou hypotézu definovanou jako H_0 : data jsou lineárně nezávislá,
2. alternativní hypotézu definovanou jako H_A : data jsou lineárně závislá,
3. kritický obor $W = \left(-\infty\,;\, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}\,;\, \infty\right)$,
4. obor přijetí $OP = \left(-t_{\alpha/2}\,;\, t_{\alpha/2}\right)$

SCILAB

```
[p_hodnota,T,z_alpha]=pearson_test(X,Y,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   t_alpha ... kritická hodnota
   X ... náhodný výběr č. 1
   Y ... náhodný výběr č. 2
   alpha ... hladina významnosti
```

Příklad. V továrně byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). byly zaznamenány následující údaje

produkce	532	297	378	121	519	613	592	497
náklady	48	32	42	27	45	51	53	48

Na hladině významnosti 0,05 testujte tvrzení, že data jsou lineárně nezávislá.

1. Teorie

   (a) oboustranný test, $\alpha = 0,05$, n=8,
   vypočteno: $E\left[X\right] = 443,625$, $E\left[Y\right] = 43,25$, $s_{xy}^2 = 1505,536$, $s_x^2 = 28133,697$, $s_y^2 = 85,072$

$$r_P = \frac{s_{xy}^2}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{1505,536}{\sqrt{28133,697 \cdot 85,072}} = 0,973$$

$$T = r_P\sqrt{\frac{n-2}{1-r_P^2}} = 0,973\frac{\sqrt{8-2}}{\sqrt{1-0,973^2}} = \frac{2,383}{0,231} = 10,316$$

[5] Pokud neplatí podmínka normality dat, pro testování se doporučuje použít Spearmanův korelační test.

(b) $t_{\alpha/2} = 2,262$
$W = (-\infty\,;\,-2,447)\,(2,447\,;\,\infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení, že data jsou lineárně nezávislá. Data je možné použít k lineární regresi.

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
[p_hodnota,T,t_alpha]=pearson_test(x,y)

t_alpha = 2.4469119
T = 10.358974
p_hodnota = 0.0000474
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.4.3 Test nezávislosti výběrů - Spearmanův test

TEORIE

Spearmanův korelační test, někdy též nazývaný test koeficientem pořadové korelace, nám říká, jaký je stupeň závislosti mezi dvěma náhodnými veličinami X a Y o rozsahu n. Na rozdíl od Pearsonova testu se zde nepočítá s hodnotami obou veličin, ale s jejich pořadím. To umožňuje testovat závislost i v takovém případě, že výběry nepochází z dat s normálním rozdělením $N(\mu, \sigma^2)$.

Statistika je

$$T = r_S\sqrt{\frac{n-2}{1-r^2}} \quad T \sim St(n-2)$$

kde n je rozsah výběrů a r_S je výběrový korelační koeficient, který se spočte jako

$$r_S = 1 - \frac{6}{n(n^2-1)}S,$$

kde $S = \sum_{i=1}^{n}(p_i - q_i)^2 = \sum_{i=1}^{n} d_i^2$, p_i a q_i jsou pořadová čísla a d_i je rozdíl pořadových čísel.

Test nezávislosti prvků výběru - Pearsonův korelační test je oboustranný a má:

1. nulovou hypotézu definovanou jako H_0 : data jsou nezávislá,
2. alternativní hypotézu definovanou jako H_A : data jsou závislá,
3. kritický obor $W = (-\infty\,;\,-t_{\alpha/2}) \cup (t_{\alpha/2}\,;\,\infty)$,
4. obor přijetí $OP = (-t_{\alpha/2}\,;\,t_{\alpha/2})$.

SCILAB

```
[p_hodnota,T,z_alpha]=spearman_test(X,Y,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   t_alpha ... kritická hodnota
   X ... náhodný výběr č. 1
   Y ... náhodný výběr č. 2
   alpha ... hladina významnosti
```

Příklad. V továrně byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). Byly zaznamenány následující údaje

produkce	532	297	378	121	519	613	592	497
náklady	48	32	42	27	45	51	53	48

Na hladině významnosti 0,05 testujte tvrzení, že data jsou nezávislá.

1. Teorie

   (a) oboustranný test, $\alpha = 0,05$, n=8,

produkce (x_i)	532	297	378	121	519	613	592	497	
náklady (y_i)	48	32	42	27	45	51	53	48	
pořadí x_i	6	2	3	1	5	8	7	4	
pořadí y_i	5	2	3	1	4	7	8	5	
d_i	1	0	0	0	1	1	1	1	$\sum$
d_i^2	1	0	0	0	1	1	1	1	5

$$r_S = 1 - \frac{6}{n(n^2-1)}S = 1 - \frac{6}{8\,(8^2-1)} \cdot 5 = 1 - \frac{30}{504} = 0,940$$
$$T = r\sqrt{\frac{n-2}{1-r^2}} = 0,940\frac{\sqrt{8-2}}{\sqrt{1-0,940^2}} = \frac{2,303}{0,341} = 6,754$$

(b) $t_{\alpha/2} = 2,262$
$W = (-\infty\,;\,-2,447)\,(2,447\,;\,\infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení, že data jsou nezávislá. Data je možné použít k regresi.

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
[p_hodnota,T,t_alpha]=spearman_test(x,y)

Kritický obor W=(-Inf,-2.446912)U(2.446912,Inf)
Hodnota testové statistiky T=6.778349
Hypotézu zamítáme

t_alfa = 2.4469119
T = 6.7783488
p_hodnota = 0.0005040
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

6.4.4 Test pořadové nezávislosti prvků výběru

TEORIE

Předpokládejme, že máme výběr X o rozsahu n. Testujeme rozdílnost mezi prvky výběru X a výběrového mediánu $\hat{x}_{0,5}$

$$D_i = X_i - \hat{x}_{0,5}, \quad i = 1, 2, \ldots, n$$

Na těchto rozdílech definuje série, tj. souvislé posloupnosti prvků se stejným znaménkem mezi dvěmi změnami znaménka. b definujeme jako počet sérií v posloupnosti rozdílů D.

Testová statistika má tvar

$$T = \frac{2b - (n-2)}{\sqrt{n-1}} \quad T \sim N(0;1)$$

kde n je počet dvojic $[x, y]$.

Test pořadové nezávislosti prvků výběru je levostranný a má:

1. nulovou hypotézu definovanou jako H_0 : prvky výběru jsou nezávislé ,
2. alternativní hypotézu definovanou jako H_A : prvky výběru jsou závislé,
3. kritický obor $W = (-\infty\,; -z_\alpha)$,
4. obor přijetí $OP = (z_\alpha\,; \infty)$

SCILAB

```
[p_hodnota,T,z_alpha]=ordinal_test(X,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   X ... náhodný výběr
   alpha ... hladina významnosti
```

Příklad. Na hladině významnosti $\alpha = 0,05$ testujeme tvrzení, že prvky výběru x jsou nezávislé.

$$x \;=\; \{\; 2,4 \quad 2,2 \quad 1,6 \quad 1,8 \quad 1,5 \quad 1,8 \quad 2,2 \quad 2,3 \quad 2,3 \quad 2,5 \;\}$$

1. Teorie

   (a) levostranný test, $\alpha = 0,05$, n=10, $\hat{x}_{0,5} = 2,2$

$$D_i = X_i - \hat{x}_{0,5}, \quad i = 1, 2, \ldots, n$$

x	2,4	2,2	1,6	1,8	1,5	1,8	2,2	2,3	2,3	2,5
$\hat{x}_{0,5}$	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2
D	0,2	0,0	-0,6	-0,4	-0,7	-0,4	0,0	0,1	0,1	0,3
	+	0	-	-	-	-	0	+	+	+

Počet sérií se stejnými znaménky $b = 3$.

$$T = \frac{2b - (n-2)}{\sqrt{n-1}} = \frac{2 \cdot 3 - (10-2)}{\sqrt{10-1}} = -\frac{2}{3}$$

(b) $z_{\alpha/2} = 1,645$
$W = (-\infty, -1,645)$

(c) $T \notin W \Rightarrow$ Nezamítáme tvrzení, že prvky výběru jsou nezávislé.

2. Scilab

```
x=[2.4 2.2 1.6 1.8 1.5 1.8 2.2 2.3 2.3 2.5]
[p_hodnota,T,z_alpha]=ordinal_test(x)

z_alpha = - 1.6448536
T = - 0.6666667
P_hodnota = 0.2524925
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

Kapitola 7

Regresní analýza

7.1 Lineární regrese

TEORIE

Předpokládejme, že máme data ve tvaru uspořádaných dvojic $[x, y]$. Nějakým testem (např. testem nezávislosti výběrů - Pearsonovým testem) jsme ověřili závislost dat. Myslíme si, že závislost mezi x a y by měla být lineární ve tvaru

$$y_i = b_1 \cdot x_i + b_o$$

kde x_i je nezávisle proměnná, y_i je závisle proměnná, b_0 je absolutní člen a b_1 je směrnice přímky.

V našem případě lineární regresí nazýváme aproximaci bodů přímkou pomocí metody nejmenších čtverců, tzn. známe nezávisle proměnnou x_i a závisle proměnnou y_i a hledáme optimální nastavení koeficientů b_1 a b_0. Optimálním nastavením rozumíme minimalizaci součtu rozdílů pozorovaných závisle proměnných y_i a odhadovaných (teoretických) závisle proměnných $\hat{y}_i$.

Bodové odhady parametrů b_1 a b_0 lze vypočítat dle vzorců:

$$b_1 = \frac{\sum x_i \sum y_i - n \sum x_i \cdot y_i}{\left(\sum x_i\right)^2 - n \sum x_i^2},$$
$$b_0 = \frac{\sum x_i \sum x_i \cdot y_i - \sum x_i^2 \sum y_i}{\left(\sum x_i\right)^2 - n \sum x_i^2}.$$

Přes index i se sčítá. Symbol n znamená počet dvojic $[x, y]$.

Po provedení regrese bychom měli provést ještě test bělosti reziduí, abychom ověřili, že přímka je pro data x a y opravdu vhodnou regresní křivkou.

SCILAB

```
p=lin_reg(x,y)
   p ... hodnoty parametrů [b1 b0]
   x ... nezávisle proměnná
   y ... závisle proměnná
```

Příklad. Byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). Byly zaznamenány následující údaje

produkce(x_i)	532	297	378	121	519	613	592	497
náklady(y_i)	48	32	42	27	45	51	53	48

Určete koeficienty regresní přímky aproximující tato data.

1. Teorie

   (a) mezivýpočty: $n = 8$

$$\sum x_i = 532 + 297 + ... + 497 = 3549$$
$$\sum y_i = 48 + 32 + 42 + ... + 48 = 346$$
$$\sum x_i \cdot y_i = 532 \cdot 48 + 297 \cdot 32 + ... + 497 \cdot 48 = 164033$$
$$\sum x_i^2 = 532^2 + 297^2 + ... + 497^2 = 1771361$$

   (b) směrnice přímky

$$b_1 = \frac{\sum x_i \sum y_i - n \sum x_i \cdot y_i}{\left(\sum x_i\right)^2 - n \sum x_i^2} = \frac{3549 \cdot 346 - 8 \cdot 164033}{3549^2 - 8 \cdot 1771361} = \frac{-84310}{-1575487} = 0,054$$

(c) absolutní člen

$$b_0 = \frac{\sum x_i \sum x_i \cdot y_i - \sum x_i^2 \sum y_i}{(\sum x_i)^2 - n \sum x_i^2} = \frac{3549 \cdot 164033 - 1771361 \cdot 346}{3549^2 - 8 \cdot 1771361} = \frac{-30737789}{-1575487} = 19,510$$

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
p=lin_reg(x,y)

p = [0.0535136 19.510024] //[b1 b0]
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.2 Lineární predikce

TEORIE

Předpokládejme, že máme data ve tvaru uspořádaných dvojic $[x, y]$. Nějakým testem (např. testem nezávislosti výběrů - Pearsonovým testem) jsme ověřili závislost dat. Myslíme si, že závislost mezi x a y by měla být lineární ve tvaru

$$y_i = b_1 \cdot x_i + b_o$$

kde x_i je nezávisle proměnná, y_i je závisle proměnná, b_0 je absolutní člen a b_1 je směrnice přímky.

V našem případě lineární regresí nazýváme aproximaci bodů přímkou pomocí metody nejmenších čtverců, tzn. známe nezávisle proměnnou x_i a závisle proměnnou y_i. Touto aproximací jsme našli optimální nastavení koeficientů b_1 a b_0 a my hledáme bodový odhad závisle proměnné y_p v čase $i+n$. Pokud $n = 1$ hovoříme o jednokrokové predikci, pro $n = 2$ hovoříme o dvoukrokové predikci atd. Je důležité na základě dat posoudit maximální krok predikce.

SCILAB

```
yp=lin_pred(x,p)
   yp ... predikce závisle proměnné
   x ... nezávisle proměnná
   p ... hodnoty parametrů [b1 b0]
```

Příklad. Byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). Byly zaznamenány následující údaje

produkce(x_i)	532	297	378	121	519	613	592	497
náklady(y_i)	48	32	42	27	45	51	53	48

Predikujte hodnotu nákladů při produkci 800 000 kusů.

1. Teorie
   (a) nejdříve spočítáme parametry (viz lineární regrese): $b_0 = 19,510$ a $b_1 = 0,054$.
   (b) predikce pro $x = 800$

$$\begin{aligned} y_p &= b_1 \cdot 800 + b_0 \\ y_p &= 0,054 \cdot 800 + 19,510 \\ y_p &= 62.71 \end{aligned}$$

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
xp=800;
p=lin_reg(x,y);
yp=lin_pred(xp,p)

yp=62.320913
```

3. Rozdíl výsledků
   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.3 Vícenásobná lineární regrese

TEORIE

Předpokládejme, že závisle proměnná y závisí na více nezávisle proměnných x_n s tím, že závislost je lineární. Tuto závislost můžeme zapsat jako

$$y_t = b_n x_{t,n} + b_{n-1} x_{t,n-1} + ... + b_1 x_{t,1} + b_0$$

.

kde $x_{t,n}$ je n-tá nezávisle proměnná v čase t, y_t je závisle proměnná v čase t, b_0 je absolutní člen a $b_1 - b_n$ jsou parametry směrnice přímky.

V našem případě lineární regresí nazýváme aproximaci bodů přímkou, tzn. známe nezávisle proměnné $x_{t,1} - x_{t,n}$ a závisle proměnnou y_i a hledáme optimální nastavení koeficientů $b_n, b_{n-1}, ..., b_0$. Optimálním nastavením rozumíme minimalizaci součtu rozdílů pozorovaných závisle proměnných y_i a odhadovaných (teoretických) závisle proměnných $\hat{y}_i$. Koeficienty získáme následujícím postupem:

1. Z dat sestrojíme rozšířenou datovou matici:

$$D = \begin{pmatrix} y_1 & x_{1,n} & x_{1,n-1} & \dots & x_{1,1} & 1 \\ y_2 & x_{2,n} & x_{2,n-1} & \dots & x_{2,1} & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_t & x_{t,n} & x_{t,n-1} & \dots & x_{t,1} & 1 \end{pmatrix}.$$

2. Spočteme informační matici V

$$V = D' \cdot D,$$

kde D' značí matici transponovanou.

3. Matici V rozdělíme na 4 submatice

$$V = \begin{pmatrix} V_y & V'_{xy} \\ V_{xy} & V_x \end{pmatrix}.$$

kde V_y je jednoprvková matice (prvek $V_{1,1}$), V'_{xy} je první řádek matice V bez prvního prvku, V_{xy} je první sloupec matice V bez prvního prvku, V_x je zbytek.

4. Bodové odhady regresních koeficientů získáme dle vzorce:

$$\begin{pmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \\ b_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy}.$$

SCILAB

```
p=lin_reg_n(x,y)
   p ... hodnoty parametrů [bn bn-1 ... b1 b0]
   x ... nezávisle proměnná (matice několika nezávisle proměnných)
   y ... závisle proměnná
```

Příklad. Na sledovaném procesu byla naměřena data, kde x_1 resp. x_2 jsou nezávisle proměnné a y je závisle proměnná

x_1	15	12	11	9	9	8	5	3
x_2	3	9	5	11	28	14	32	58
y	9	7	22	12	27	31	44	36

Zjistěte koeficienty regrese

1. Teorie

   (a) Protože jsou zde nezávisle proměnné x_1, x_2 jedná se o vícenásobnou regresi s rovnicí

$$y = b_2x_2 + b_1x_1 + b_0$$

   (b) Dále pokračujeme podle výše uvedeného postupu a sestrojíme datovou matici

$$D = \begin{bmatrix} y_1 & x_{1,2} & x_{1,1} & 1 \\ y_2 & x_{2,2} & x_{2,1} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ y_8 & x_{8,2} & x_{8,1} & 1 \end{bmatrix} = \begin{bmatrix} 9 & 3 & 15 & 1 \\ 7 & 9 & 12 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 36 & 58 & 3 & 1 \end{bmatrix}$$

   a spočteme informační matici V

$$V = D' \cdot D = \begin{bmatrix} 9 & 7 & \dots & 36 \\ 3 & 9 & \dots & 58 \\ 15 & 12 & \dots & 3 \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 9 & 3 & 15 & 1 \\ 7 & 9 & 12 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 36 & 58 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 5680 & 5018 & 1388 & 188 \\ 5018 & 5604 & 1005 & 160 \\ 1388 & 1005 & 750 & 72 \\ 188 & 160 & 72 & 8 \end{bmatrix},$$

   (c) Matici rozdělíme na 4 submatice. Pro výpočet koeficientů potřebujeme inverzní matici V_x^{-1} a vektor V_{xy}

$$V_x^{-1} = \begin{bmatrix} 5604 & 1005 & 160 \\ 1005 & 750 & 72 \\ 160 & 72 & 8 \end{bmatrix}^{-1} = \begin{bmatrix} 0,0018 & 0,0078 & -0,1064 \\ 0,0078 & 0,0429 & -0,5419 \\ -0,1064 & -0,5419 & 7,1293 \end{bmatrix}$$
$$V_{xy} = \begin{bmatrix} 5018 & 1388 & 188 \end{bmatrix}'$$

   Vypočteme parametry

$$\begin{pmatrix} b_2 \\ b_1 \\ b_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy} = \begin{bmatrix} -3,2793 \\ -0,0701 \\ 54,4157 \end{bmatrix}.$$

2. Scilab

```
x1=[15 12 11 9 9 8 5 3];
x2=[3 9 5 11 28 14 32 58];
x=[x1; x2];
y=[9 7 22 12 27 31 44 36];
p=lin_reg_n(x,y)

p = [-3.2793169  -0.0700927  54.415707]' //[b2 b1 b0]
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.4 Vícenásobná lineární predikce

TEORIE

Předpokládejme, že závisle proměnná y závisí na více nezávisle proměnných x_n s tím, že závislost je lineární. Tuto závislost můžeme zapsat jako

$$y_t = b_n x_{t,n} + b_{n-1} x_{t,n-1} + ... + b_1 x_{t,1} + b_0$$

.

kde $x_{t,n}$ je n-tá nezávisle proměnná v čase t, y_t je závisle proměnná v čase t, b_0 je absolutní člen a $b_1 - b_n$ jsou parametry směrnice přímky.

V našem případě vícenásobnou lineární regresí nazýváme aproximaci bodů přímkou, tzn. známe nezávisle proměnné $x_{t,1} - x_{t,n}$ a závisle proměnnou y_i a hledáme optimální nastavení koeficientů $b_n, b_{n-1}, ..., b_0$. Touto aproximací jsme našli optimální nastavení koeficientů $b_n, b_{n-1}, ..., b_0$ a my hledáme bodový odhad závisle proměnné y_p v čase $i + n$. Pokud $n = 1$ hovoříme o jednokrokové predikci, pro $n = 2$ hovoříme o dvoukrokové predikci atd. Je důležité na základě dat posoudit maximální krok predikce.

SCILAB

```
yp=lin_pred_n(x,p)
   yp ... predikce závisle proměnné
   x ... nezávisle proměnná
   p ... hodnoty parametrů [b1 b0]
```

Příklad. Na sledovaném procesu byla naměřena data, kde x_1 resp. x_2 jsou nezávisle proměnné a y je závisle proměnná

x_1	15	12	11	9	9	8	5	3
x_2	3	9	5	11	28	14	32	58
y	9	7	22	12	27	31	44	36

Predikujte hodnotu y_p pro $x_1 = 2$ a $x_2 = 60$.

1. Teorie
   (a) nejdříve spočítáme parametry (viz vícenásobná lineární regrese): $b_0 = 54,4157$,$b_1 = -3,2793$ a $b_2 = -0,0701$.
   (b) predikce pro $x_1 = 2$ a $x_2 = 60$

$$\begin{aligned} y_p &= b_2 \cdot 60 + b_1 \cdot 2 + b_0 \\ y_p &= -0,0702 \cdot 60 - 3,2793 \cdot 2 + 54,4157 \\ y_p &= 43,6451 \end{aligned}$$

2. Scilab

```
xp=[60 2]; //xp=[x2 x1]
p = [-0.0700927  -3.2793169  54.415707]' //[b2 b1 b0]
yp=lin_pred_n(xp,p)

yp=43.65151
```

3. Rozdíl výsledků
   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.5 Polynomiální regrese

TEORIE

Předpokládáme polynomickou závislost veličiny y na proměnné x:

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k.$$

V našem případě polynomiální regresí nazýváme aproximaci bodů křivkou pomocí metody nejmenších čtverců, tzn. známe nezávisle proměnnou x_i a závisle proměnnou y_i a hledáme optimální nastavení koeficientů $b_n, b_{n-1}, ..., b_0$. Optimálním nastavením rozumíme minimalizaci součtu rozdílů pozorovaných závisle proměnných y_i a odhadovaných (teoretických) závisle proměnných $\hat{y}_i$. Koeficienty určíme následujícím postupem:

1. Z dat sestrojíme datovou matici:

$$D = \begin{pmatrix} y_1 & x_1^k & x_1^{k-1} & \ldots & x_1 & 1 \\ y_2 & x_2^k & x_2^{k-1} & \ldots & x_2 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ y_n & x_n^k & x_n^{k-1} & \ldots & x_n & 1 \end{pmatrix}.$$

2. Spočteme informační matici V

$$V = D' \cdot D,$$

kde D' značí matici transponovanou.

3. Matici V rozdělíme na 4 submatice

$$V = \begin{pmatrix} V_y & V'_{xy} \\ V_{xy} & V_x \end{pmatrix}.$$

kde V_y je jednoprvková matice (prvek $V_{1,1}$), V'_{xy} je první řádek matice V bez prvního prvku, V_{xy} je první sloupec matice V bez prvního prvku, V_x je zbytek.

4. Bodové odhady regresních koeficientů získáme dle vzorce:

$$\begin{pmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \\ b_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy}.$$

SCILAB

```
p=pol_reg(x,y,k)
   p ... hodnoty parametrů [bn bn-1 ... b1 b0]
   x ... nezávisle proměnná
   y ... závisle proměnná
   k ... stupeň regresního polynomu
```

Příklad. Na sledovaném procesu byla naměřena data, kde x je nezávisle proměnná a y je závisle proměnná

x_i	5	12	20	26	29	38	40	45
y_i	19	17	12	1	27	35	44	76

Pro tato data proveďte polynomiální regresi 4. řádu

1. Teorie

(a) Pro polynomiální regresi 4. řádu platí rovnice

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4$$

(b) Dále pokračujeme podle výše uvedeného postupu a sestrojíme datovou matici

$$D = \begin{bmatrix} y_1 & x_1^4 & x_1^3 & x_1^2 & x_1 & 1 \\ y_2 & x_2^4 & x_2^3 & x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_n^4 & x_n^3 & x_n^2 & x_n & 1 \end{bmatrix} = \begin{bmatrix} 19 & 625 & 125 & 25 & 5 & 1 \\ 17 & 20736 & 1728 & 144 & 12 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 76 & 4100625 & 91125 & 2025 & 45 & 1 \end{bmatrix}$$

a spočteme informační matici V

$$V = D' \cdot D = \begin{bmatrix} 19 & 17 & \dots & 76 \\ 625 & 20736 & \dots & 4100625 \\ 125 & 1728 & \dots & 91125 \\ 25 & 144 & \dots & 2025 \\ 5 & 12 & \dots & 45 \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 19 & 625 & 125 & 25 & 5 & 1 \\ 17 & 20736 & 1728 & 144 & 12 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 76 & 4100625 & 91125 & 2025 & 45 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 10461 & 5,191 \cdot 10^8 & 12465850 & 305946 & 7858 & 231 \\ 5,191 \cdot 10^8 & 2,845 \cdot 10^3 & 6,785 \cdot 10^{11} & 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 \\ 12465850 & 6,785 \cdot 10^{11} & 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 & 261815 \\ 305946 & 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 & 261815 & 7155 \\ 7858 & 4,020 \cdot 10^8 & 10091379 & 261815 & 7155 & 215 \\ 231 & 10091379 & 261815 & 7155 & 215 & 8 \end{bmatrix}$$

(c) Matici rozdělíme na 4 submatice. Pro výpočet koeficientů potřebujeme inverzní matici V_x^{-1} a vektor V_{xy}

$$V_x^{-1} = \begin{bmatrix} 2,845 \cdot 10^3 & 6,785 \cdot 10^{11} & 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 \\ 6,785 \cdot 10^{11} & 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 & 261815 \\ 1,638 \cdot 10^{10} & 4,020 \cdot 10^8 & 10091379 & 261815 & 7155 \\ 4,020 \cdot 10^8 & 10091379 & 261815 & 7155 & 215 \\ 10091379 & 261815 & 7155 & 215 & 8 \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} 4,176 \cdot 10^{-10} & -4,225 \cdot 10^{-8} & 0,0000014 & -0,0000183 & 0,0000656 \\ -4,225 \cdot 10^{-8} & 0,0000043 & -0,0001491 & 0,0019349 & -0,0070611 \\ 0,0000014 & -0,0001491 & 0,0052305 & -0,0694200 & 0,2594320 \\ -0,0000183 & 0,0019349 & -0,0694200 & 0,9487243 & -3,6821933 \\ 0,0000656 & -0,0070611 & 0,2594320 & -3,6821933 & 15,412431 \end{bmatrix}$$

$$V_{xy} = \begin{bmatrix} 0,0000584 & 0,0022618 & 0,0921514 & 3,8460258 \end{bmatrix}$$

(d) Vypočteme parametry

$$\begin{pmatrix} b_4 \\ b_3 \\ b_2 \\ b_1 \\ b_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy} = \begin{bmatrix} 0,0000139 \\ 0,0011045 \\ -0,0516896 \\ 0,0400101 \\ 20,387983 \end{bmatrix}.$$

2. Scilab

```
x=[5 12 20 26 29 38 40 45];
y=[19 17 12 1 27 35 44 76];
k=4;
p=pol_reg(x,y,k)

p = [0.0000139 0.0011045 - 0.0516896 0.0400101 20.387983]' //[b4 b3 b2 b1 b0]
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.6 Polynomiální predikce

TEORIE

Předpokládáme polynomickou závislost veličiny y na proměnné x:

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k.$$

V našem případě polynomiální regresí nazýváme aproximaci bodů křivkou pomocí metody nejmenších čtverců, tzn. známe nezávisle proměnnou x_i a závisle proměnnou y_i. Touto aproximací jsme našli optimální nastavení koeficientů $b_n, b_{n-1}, ..., b_0$ a my hledáme bodový odhad závisle proměnné y_p v čase $i + n$. Pokud $n = 1$ hovoříme o jednokrokové predikci, pro $n = 2$ hovoříme o dvoukrokové predikci atd. Je důležité na základě dat posoudit maximální krok predikce.

SCILAB

```
yp=pol_pred(x,p)
   yp ... predikce závisle proměnné
   x ... nezávisle proměnná
   p ... hodnoty parametrů [bn bn-1 ... b1 b0]
```

Příklad. Na sledovaném procesu byla naměřena data, kde x je nezávisle proměnná a y je závisle proměnná

x_i	5	12	20	26	29	38	40	45
y_i	19	17	12	1	27	35	44	76

Pro tato data proveďte polynomiální predikci pro nezávisle proměnnou $x_p = 60$ za předpokladu polynomiální regrese 4. řádu.

1. Teorie

   (a) nejdříve spočítáme parametry (viz polynomiální regrese): $b_0 = 20,3880$,$b_1 = 0,0400$, $b_2 = -0,0517$, $b_3 = 0,0011$ a $b_4 = 0,00001$.

   (b) predikce pro $x_P = 60$

$$\begin{aligned} y_p &= b_4 \cdot 60^4 + b_3 \cdot 60^3 + b_2 \cdot 60^2 + b_1 \cdot 60 + b_0 \\ y_p &= 0,00001 \cdot 60^4 + 0,0011 \cdot 60^3 - 0,0517 \cdot 60^2 + 0,04 \cdot 60 + 20,3880 \\ y_p &= 203,868 \end{aligned}$$

2. Scilab

```
xp=60;
p = [0.0000139 0.0011045 - 0.0516896 0.0400101 20.387983]; //[b4 b3 b2 b1 b0]
yp=pol_pred(xp,p)

yp=255.02454
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.7 Exponenciální regrese

TEORIE

Předpokládejme, že máme data ve tvaru uspořádaných dvojic $[x, y]$. Myslíme si, že závislost mezi x a y je nelineární, přesněji exponenciální. Platí tedy, že hledaná regresní přímka má tvar

$$y_i = b_o \exp\{b_1 \cdot x_i\} = b_0 \mathrm{e}^{b_1 \cdot x}$$

kde x_i je nezávisle proměnná, y_i je závisle proměnná, b_0 je absolutní člen a b_1 je směrnice přímky.

V tomto případě lze použít metodu nejmenších čtverců, ale až po linearizaci rovnice, tedy pokud ji zlogaritmujeme jako

$$\ln y = \ln b_0 + b_1 \cdot x \Rightarrow \tilde{y} = \beta_0 + b_1 \cdot x$$

Koeficienty regrese získáme následujícím postupem:

1. Z dat sestrojíme rozšířenou datovou matici:

$$D = \begin{pmatrix} y_1 & x_{1,n} & x_{1,n-1} & \dots & x_{1,1} & 1 \\ y_2 & x_{2,n} & x_{2,n-1} & \dots & x_{2,1} & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_t & x_{t,n} & x_{t,n-1} & \dots & x_{t,1} & 1 \end{pmatrix}.$$

2. Spočteme informační matici V

$$V = D' \cdot D,$$

kde D' značí matici transponovanou.

3. Matici V rozdělíme na 4 submatice

$$V = \begin{pmatrix} V_y & V'_{xy} \\ V_{xy} & V_x \end{pmatrix}.$$

kde V_y je jednoprvková matice (prvek $V_{1,1}$), V'_{xy} je první řádek matice V bez prvního prvku, V_{xy} je první sloupec matice V bez prvního prvku, V_x je zbytek.

4. Bodové odhady regresních koeficientů získáme dle vzorce:

$$\begin{pmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \\ b_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy}.$$

SCILAB

```
p=exp_reg(x,y)
   p ... hodnoty parametrů [b1 b0]
   x ... nezávisle proměnná
   y ... závisle proměnná
```

Příklad. Na sledovaném procesu byla naměřena data, kde x je nezávisle proměnná a y je závisle proměnná

x_i	5	12	20	26	29	38	40	45
y_i	19	17	12	1	27	35	44	76

Pro tato data proveďte exponenciální regresi.

1. Teorie

   (a) Pro exponenciální regresi má exponenciální přímka tvar

$$y_i = b_o \exp\{b_1 \cdot x_i\}$$

   (b) Dále pokračujeme podle výše uvedeného postupu a sestrojíme datovou matici

$$D = \begin{bmatrix} \ln(y_1) & x_1 & 1 \\ \ln(y_2) & x_2 & 1 \\ \vdots & \vdots & \vdots \\ \ln(y_8) & x_8 & 1 \end{bmatrix} = \begin{bmatrix} \ln(19) & 5 & 1 \\ 2.83 & 12 & 1 \\ \vdots & \vdots & \vdots \\ 4.33 & 45 & 1 \end{bmatrix}$$

   a spočteme informační matici V

$$V = D' \cdot D = \begin{bmatrix} 2.94 & 2.83 & \dots & 4.33 \\ 5 & 12 & \dots & 45 \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 3,94 & 5 & 1 \\ 2,83 & 12 & 1 \\ \vdots & \vdots & \vdots \\ 4,33 & 45 & 1 \end{bmatrix} = \begin{bmatrix} 79,45 & 675,35 & 23,23 \\ 675,35 & 7155 & 215 \\ 23,23 & 215 & 8 \end{bmatrix},$$

   (c) Matici rozdělíme na 4 submatice. Pro výpočet koeficientů potřebujeme inverzní matici V_x^{-1} a vektor V_{xy}

$$V_x^{-1} = \begin{bmatrix} 7155 & 215 \\ 215 & 8 \end{bmatrix}^{-1} = \begin{bmatrix} 0,0007 & -0,0195 \\ -0,0195 & 0,6496 \end{bmatrix}$$
$$V_{xy} = \begin{bmatrix} 675,35 & 23,23 \end{bmatrix}'$$

   Vypočteme parametry b_1 a β_0

$$\begin{pmatrix} b_1 \\ \beta_0 \end{pmatrix} = V_x^{-1} \cdot V_{xy} = \begin{bmatrix} 0,037 \\ 1,907 \end{bmatrix}.$$

   (d) Vypočteme skutečný parametr b_0

$$b_0 = \exp\{\beta_0\} = \exp\{1,907\} = 6,733$$

2. Scilab

```
x=[5 12 20 26 29 38 40 45];
y=[19 17 12 1 27 35 44 76];
p=exp_reg(x,y)

p = [0.0370996 6.7297024]' //[b1 b0]
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.8 Exponenciální predikce

TEORIE

Předpokládejme, že máme data ve tvaru uspořádaných dvojic $[x, y]$. Myslíme si, že závislost mezi x a y je nelineární, přesněji exponenciální. Platí tedy, že hledaná regresní přímka má tvar

$$y_i = b_o \exp\{b_1 \cdot x_i\} = b_0 \mathrm{e}^{b_1 \cdot x}$$

kde x_i je nezávisle proměnná, y_i je závisle proměnná, b_0 je absolutní člen a b_1 je směrnice přímky.

V tomto případě lze použít metodu nejmenších čtverců, ale až po linearizaci rovnice, tedy pokud ji zlogaritmujeme jako

$$\ln y = \ln b_0 + b_1 \cdot x \Rightarrow \tilde{y} = \beta_0 + b_1 \cdot x$$

Touto aproximací jsme našli optimální nastavení koeficientů b_1 a b_0 a my hledáme bodový odhad závisle proměnné y_p v čase $i + n$. Pokud $n = 1$ hovoříme o jednokrokové predikci, pro $n = 2$ hovoříme o dvoukrokové predikci atd. Je důležité na základě dat posoudit maximální krok predikce.

SCILAB

```
yp=exp_pred(x,p)
   yp ... predikce závisle proměnné
   x ... nezávisle proměnná
   p ... hodnoty parametrů [b1 b0]
```

Příklad. Na sledovaném procesu byla naměřena data, kde x je nezávisle proměnná a y je závisle proměnná

x_i	5	12	20	26	29	38	40	45
y_i	19	17	12	1	27	35	44	76

Pro tato data proveďte polynomiální predikci pro nezávisle proměnnou $x_p = 60$.

1. Teorie

   (a) nejdříve spočítáme parametry (viz exponenciální regrese): $b_0 = 6,733$ a $b_1 = 0,037$.

   (b) predikce pro $x_P = 60$

$$\begin{aligned} \ln y_p &= b_1 \cdot 60 + \ln b_0 \\ \ln y_p &= 0,037 \cdot 60 + 1,907 \\ \ln y_p &= 4,127 \\ y_p &= 61,99 \end{aligned}$$

2. Scilab

```
xp=60;
p = [0.0370996 6.7297024]; //[b1 b0]
yp=exp_pred(xp,p)

yp=62.334089
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.9 T-test korelačního koeficientu

TEORIE

Jedná se vlastně o Pearsonův test, který je vysvětlen v kapitole 6.4.2. T-test korelačního koeficientu oproti klasickému Pearsonovu testu testuje zda jsou data vhodná či nevhodná k lineární regresi. Tento test je potřeba v lineární analýze, aby nám řekl, zda lze ze získaných interpretovat nějaké závěry ze získaných výsledků - data musí být vhodná k regresi.

T-test korelačního koeficientu test lze použít v případě, že výběry pochází z dat s normálním rozdělením $N(\mu, \sigma^2)$[1].

Statistika je

$$T = r_P \sqrt{\frac{n-2}{1-r_P^2}} \quad T \sim St(n-2)$$

kde n je rozsah výběrů a r_P je výběrový korelační koeficient, který se spočte jako

$$r_P = \frac{s_{xy}^2}{\sqrt{s_x^2 \cdot s_y^2}}$$

kde s_{xy}^2 je kovariance mezi náhodnými veličinami X a Y a s_x^2 resp. s_y^2 je výběrový rozptyl náhodné veličiny X resp. Y.

T-test korelačního koeficientu - Pearsonův korelační test je oboustranný a má:

1. nulovou hypotézu definovanou jako H_0 : data nejsou vhodná k regresi,
2. alternativní hypotézu definovanou jako H_A : data jsou vhodná k regresi,
3. kritický obor $W = (-\infty\,;\, -t_{\alpha/2}) \cup (t_{\alpha/2}\,;\, \infty)$,
4. obor přijetí $OP = (-t_{\alpha/2}\,;\, t_{\alpha/2})$

SCILAB

```
[p_hodnota,T,z_alpha]=pearson_test(X,Y,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   t_alpha ... kritická hodnota
   alpha ... hladina významnosti
   X ... náhodný výběr č. 1
   Y ... náhodný výběr č. 2
```

Příklad. Viz příklad v kapitole 6.4.2.

[1] Pokud neplatí podmínka normality dat, pro testování se doporučuje použít Spearmanův korelační test.

7.10 F-test poměru vysvětleného a nevysvětleného rozptylu pro predikci

TEORIE

Tento test je velmi kvalitním testem vhodnosti regrese, který lze použít i v případě více nezávislých proměnných. Jeho postup je následující

Definujme:

1. celkový součet čtverců odchylek dat od průměru, který vypovídá o rozptýlenosti změřených dat:
$$S_y = \sum (y_i - \overline{y})^2$$
2. regresní součet čtverců odchylek predikcí od průměru, který ukazuje, kolik původního rozptylu lze vysvětlit na základě regrese (vysvětlený rozptyl):
$$S_{\hat{y}} = \sum (\hat{y}_i - \overline{y})^2$$
3. reziduální součet čtverců odchylek dat od predikcí, který vypovídá o zbylém rozptylu (nevysvětlený rozptyl):
$$S_{\hat{e}} = \sum (y_i - \hat{y}_i)^2 = S_y - S_{\hat{y}}.$$

Testová statistika pro hypotézu je definována jako:

$$T = \frac{(n-2)\, S_{\hat{y}}}{S_{\hat{e}}} \quad T \sim F\,(1,\, n-2)$$

1. Pravostranný test hypotéz vysvětleného a nevysvětleného rozptylu má:
   (a) nulovou hypotézu definovanou jako H_0 : ‚data nejsou vhodná k regresi,
   (b) alternativní hypotézu definovanou jako H_A : data jsou vhodná k regresi,
   (c) kritický obor $W = (f_\alpha\,;\,\infty)$,
   (d) obor přijetí $OP = (0\,;\,f_\alpha)$.

SCILAB

1. F-test pro predikci (lineární i nelineární regrese)

```
[p_hodnota,T, df_num, df_den]=f_test_pred(y,yp,np)
   p_hodnota... p-hodnota
   T ... statistika náhodné veličiny
   df_num ... stupně volnosti pro čitatele
   df_num ... stupně volnosti pro jmenovatele
   y ... závisle proměnná
   yp ... predikce závisle proměnné
   np ... počet parametrů regrese (pro b1*x+b0 je np=2)
```

Příklad. V továrně byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). byly zaznamenány následující údaje

produkce	532	297	378	121	519	613	592	497
náklady	48	32	42	27	45	51	53	48

Pomocí srovnání vysvětleného a nevysvětleného rozptylu testujte na hladině významnosti $\alpha = 0,05$ vhodnost dat pro lineární regresi.

1. Teorie
   (a) pravostranný test, $\alpha = 0,05$, n=8,
   - pomocí lineární regrese vypočteme regresní koeficienty: $b_0 = 19,51$ a $b_1 = 0,05$

- a následně pomocí lineární predikce vypočteme predikci závisle proměnné yp: $\begin{bmatrix} 48 & 35,4 & 39,7 & 26 & 47,3 & 52,3 & 51,2 \end{bmatrix}$
- vypočteme celkový součet čtverců odchylek dat od průměru $\overline{y} = 43,25$:

$$S_y = \sum (y_i - \overline{y})^2 = (48 - 43,25)^2 + (32 - 43,25)^2 + ... + (48 - 43,25)^2 = 595,5$$

- vypočteme regresní součet čtverců odchylek predikcí od průměru $\overline{y} = 43,25$:

$$S_{\hat{y}} = \sum (\hat{y}_i - \overline{y})^2 = (48 - 43,25)^2 + (35,4 - 43,25)^2 + ... + (46,1 - 43,25)^2 = 564$$

- vypočteme reziduální součet čtverců odchylek dat od predikcí:

$$S_{\hat{e}} = \sum (y_i - \hat{y}_i)^2 = S_y - S_{\hat{y}} = 595,5 - 564 = 31,5$$

$$T = \frac{(n-2)\, S_{\hat{y}}}{S_{\hat{e}}} = \frac{(8-2)\, 564}{31,5} = 107,4$$

(b) $f_{\alpha/2} = 6,61$
$W = (6,61\,,\, \infty)$

(c) $T \in W \Rightarrow$ Zamítáme tvrzení, že data jsou lineárně nezávislá. Data je možné použít k lineární regresi.

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
p=lin_reg(x,y);
yp=lin_pred(x,p);
[p_hodnota,T,df_num,df_den]=f_test_pred(y,yp,2)

df_den = 5
df_num = 2
T = 107.30835
p_hodnota = 0.0000782
```

3. Rozdíl výsledků

Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.11 Test bělosti reziduí

TEORIE

Reziduum je rozdíl mezi skutečnou hodnotou y_n a ideální hodnotou $\hat{y}_n$ ležící na regresní křivce. Obojí pro stejné x_n. Pokud regresní křivku zvolíme dobře, měla by být rezidua kladná i záporná, pořadově nezávislá.

Vhodnost křivky testujeme testem pořadové nezávislosti reziduí. Tento test vychází z pořadového testu nezávislosti prvků výběru, viz kapitola 6.4.2.

Předpokládejme, že máme vektor reziduí R o rozsahu n. Testujeme rozdílnost mezi rezidui R a výběrového mediánu $\hat{x}_{0,5}$

$$D_i = R_i - \hat{R}_{0,5}, \quad i = 1, 2, \ldots, n$$

kde $\hat{R}_{0,5}$ je medián reziduí.

Na těchto rozdílech definuje série, tj. souvislé posloupnosti prvků se stejným znaménkem mezi dvěmi změnami znaménka. b definujeme jako počet sérií v posloupnosti rozdílů D.

$$T = \frac{2b - (n-2)}{\sqrt{n-1}} \quad T \sim N(0;1)$$

kde n je počet reziduí.

Test pořadové nezávislosti prvků výběru je levostranný a má:

1. nulovou hypotézu definovanou jako H_0 : byla použita nesprávná regresní metoda,
2. alternativní hypotézu definovanou jako H_A : byla použita správná regresní metoda,
3. kritický obor $W = (-\infty\,; -z_\alpha)$,
4. obor přijetí $OP = (z_\alpha\,; \infty)$

SCILAB

```
[p_hodnota,T,z_alpha]=wz_test(y,yp,alpha)
   p_hodnota ... p-hodnota
   T ... statistika náhodné veličiny
   z_alpha ... kritická hodnota
   alpha ... hladina významnosti
   y ... skutečný výstup
   yp ... predikovaný výstup
```

Příklad. V továrně byla sledována závislost celkových nákladů (desítky tis. Kč) na produkci (tis. ks). byly zaznamenány následující údaje

produkce	532	297	378	121	519	613	592	497
náklady	48	32	42	27	45	51	53	48

Testujte vhodnost lineární regrese pomocí testu na bělost reziduí (prvky výběru jsou nezávislé).

1. Teorie

2. Scilab

```
x=[532 297 378 121 519 613 592 497];
y=[48 32 42 27 45 51 53 48];
p=lin_reg(x,y);
yp=lin_pred(x,p);

[p_hodnota,T,z_alpha]=wz_test(y,yp)
```

3. Rozdíl výsledků

   Rozdíl ve výsledcích není způsoben použitím jiné metody, ale zaokrouhlováním při teoretickém postupu.

7.12 Test autoregrese residuí

Test správnosti lineární regrese. Testuje se, zda mezi rezidui neexistuje dynamická vazba - tj. zda některé reziduum není funkcí předchozích.

SCILAB

```
[res,a]=autoreg_test(x,y)
   res ... slovní výsledek (ano, ne)
   a ... regresní koeficient
   x ... nezávisle proměnná
   y ... závisle proměnná
```