Okruh 2 - Aplikovaná štatistika

Jakub Chalmovianský,
Tatiana Keseliová,
Vlastimil Reichel

4. januára 2021

Kapitola 2

Aplikovaná štatistika v Matlabe

V tejto kapitole sa zameriame na vybrané problémy spracovania aplikovanej štatistiky, ktoré možno vykonať v programe Matlab. Konkrétne zacielime na testovanie normality, testovanie hypotéz o zhode rozptylu a stredných hodnôt (t-test, ANOVA) a základy korelačnej analýzy.

2.1 Testovanie normality

Mnoho štatistických metód predpokladá, že základný súbor dát má normálne rozdelenie. Na určenie, či sa dá rozdelenie považovať za normálne, slúžia testy normality. Overenie normality môžeme uskutočniť viacerými spôsobmi. My si ukážeme overenie pomocou grafických metód a overenie normality pomocou formálnych štatistických testov.

K najčastejším grafickým metódam overenia patria histogramy (funkcia **histogram**) a probability plots (pravdepodobnostné grafy), ktoré overujú zhodnosť teoretického a empirického rozdelenia. Patria sem Q-Q ploty, ktoré porovnávajú empirické a teoretické kvantily daného rozdelenia. V prípade normálneho rozdelenia sa Q-Q plot označuje tiež ako N-P plot (normal-probability plot). V Matlabe na vykreslenie Q-Q plotu využívame funkciu **qqplot**, zápis: qqplot(X). Ďalšou, často využívanou grafickou metódou overenia je P-P plot, ktorý porovnáva empirické a teoretické distribučné funkcie. V Matlabe rozlišujeme 2 funkcie:

- **probplot** - klasický P-P plot s možnosťou voľby špecifického rozdelenia; zápis: probplot('rozdelenie',X). Za voľbu 'rozdelenie' je možné zvoliť napríklad: 'exponential','lognormal','normal'. Defaultne je nastavená voľba 'normal'.

- **normplot** - P-P pre normálne rozdelenie; zápis: normplot(X)

Na overenie normality formálne, pomocou štatistických testov, využívame Jarque-Berov test normality, ktorého nulová hypotéza znie: "Dáta sú normálne rozdelené". Zodpovedajúca funkcia v Matlabe sa nazýva **jbtest**. Syntax je: [h,p,stat,cv]=jbtest(x,'alpha',α), kde vstupnými parametrami je vektor dát x, voliteľný parameter *'alpha'*, ktorý slúži na voľbu hladiny významnosti α. Výstupnými parametrami sú: h (rozhodnutie o (ne)zamietnutí nulovej hypotézy), p (p-hodnota testu), *stat* (hodnota testovej štatistiky) a *cv* (kritická hodnota pre test). Na overenie, či dáta pochádzajú zo štandardizovaného normálneho rozdelenia je možné využiť Kolmogorovov-Smirnovov test, v Matlabe funkcia **kstest**. Zápis: [h,p,ksstat,cv] = kstest(x). Vysvetlenie jednotlivých parametrov je rovnaké ako v prípade **jbtest**.

Určitú predstavu o normalite nám môžu poskytnúť aj funkcie **skewness** a **kurtosis**, v preklade šikmosť a špicatosť. Nulová šikmosť značí, že hodnoty náhodnej veličiny sú rovnomerne rozdelené vľavo a vpravo od strednej hodnoty, t.j. symetrické rozdelenia majú šikmosť nula. Kladná šikmosť poukazuje na častejší výskyt odľahlejších hodnôt vpravo od strednej hodnoty a na väčšiu kumuláciu hodnôt v ľavom okolí strednej hodnoty. Pre zápornú šikmosť je to naopak. Kladná špicatosť značí, že väčšina hodnôt náhodnej veličiny leží blízko jej strednej hodnoty a hlavný vplyv na rozptyl majú málo pravdepodobné odľahlé hodnoty. Krivka rozdelenia je špicatejšia. Záporná špicatosť značí, že rozdelenie je rovnomernejšie a jeho krivka je plochšia. Normálne rozdelenie má šikmosť rovnú nule a špicatosť rovnú trom. Zápis: skewness(x) a kurtosis(x).

Cvičenie 2.1 - Generovanie dát

Zadanie: Vygenerujte štyri dátové vzorky X_1, X_2, X_3 a X_4 z normálneho rozdelenia.

a) $X_1 \sim N(20, 25^2)$ a $X_2 \sim N(10, 25^2)$.

b) $X_3 \sim N(10, 25^2)$ a $X_4 \sim N(10, 15^2)$.

c) Všetky 4 vzorky vykreslite pomocou histogramov.

Riešenie: Na generovanie vzoriek z normálneho rozdelenia využijeme funkciu **randn**.

a) Našou úlohou je vygenerovanie vzoriek X_1 a X_2 so strednou hodnotou $\mu_1 = 20$ a $\mu_2 = 10$ a rozptylom $\sigma^2 = 25^2$.

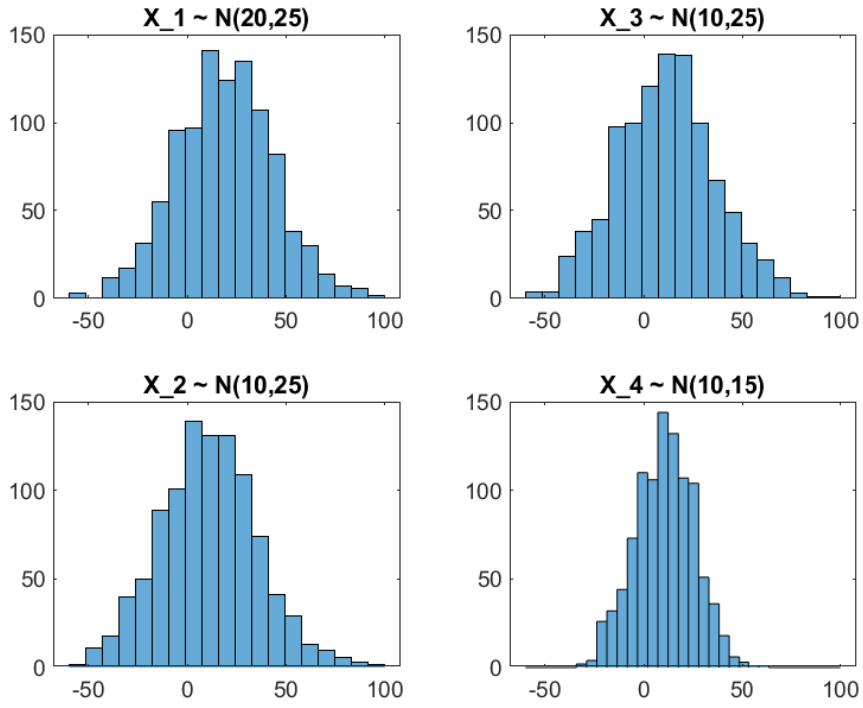
```
>> X_1 = 20 + 25*randn(1000,1);
```

```
>> X_2 = 10 + 25*randn(1000,1);
```

b) V druhej úlohe budeme generovať vzorky s rovnakou strednou hodnotou $\mu = 10$ a rôznym rozptylom $\sigma_1^2 = 25^2$ a $\sigma_2^2 = 15^2$.

```
>> X_3 = 10 + 25*randn(1000,1);
>> X_4 = 10 + 15*randn(1000,1);
```

c) Pre lepšiu ilustráciu si všetky 4 vzorky vykreslíme pomocou funkcie **histogram**. Využijeme taktiež funkciu **subplot**, kde v prvom stĺpci obrázku nájdeme riešenie úlohy a) (vzorky s rôznou strednou hodnotou) a v druhom stĺpci riešenie úlohy b) (vzorky s rôznym rozptylom). Výsledné grafy viď Obrázok 2.1.



Obrázok 2.1: Histogramy generovaných vzoriek dát

```
figure
subplot(2,2,1)
histogram(X_1,'BinLimits',[-60,100])
```

```
title('X\_1 ~ N(20,25)')
subplot(2,2,3)
histogram(X_2,'BinLimits',[-60,100])
title('X\_2 ~ N(10,25)')
subplot(2,2,2)
histogram(X_3,'BinLimits',[-60,100])
title('X\_3 ~ N(10,25)')
subplot(2,2,4)
histogram(X_4,'BinLimits',[-60,100])
title('X\_4 ~ N(10,15)')
```

Cvičenie 2.2

Zadanie: Rozhodnutie o (ne)zamietnutí hypotézy o normalite dát častokrát závisí od veľkosti vzorky. Ilustrujte túto skutočnosť na troch náhodných výberoch rôznych dĺžok.

a) Vygenerujte tri vzorky z normálneho rozdelenia s veľkosťou výberu $n = \{20, 200, 2000\}$.

b) Využite grafické metódy na overenie normality.

c) Využite štatistické testy na overenie normality.

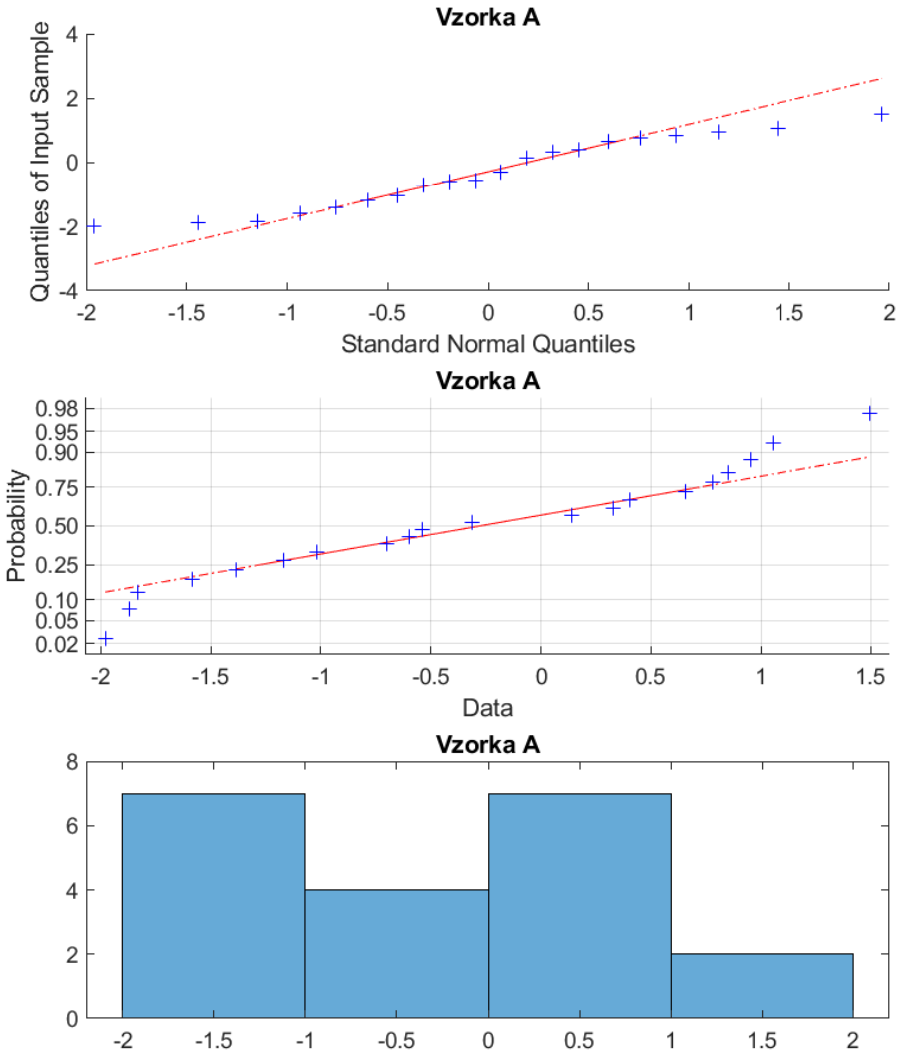
Riešenie:

a) Pomocou funkcie **randn** vygenerujeme tri vzorky dát rôznych dĺžok.

```
>> A = randn(20,1);
>> B = randn(200,1);
>> C = randn(2000,1);
```

b) Na overenie normality pomocou grafických metód využijeme Q-Q plot (v prípade normálneho rozdelenia nazývaný aj N-P plot) - funkcia **qqplot** a P-P plot - funkcia **normplot**. Na Obrázku 2.2 môžeme vidieť uvedené grafy pre vzorku A.

```
figure
subplot(3,1,1)
qqplot(A)
```



Obrázok 2.2: Q-Q plot, P-P plot a histogram pre vzorku A

```
title('Vzorka A')
subplot(3,1,2)
qqplot(B)
title('Vzorka B')
subplot(3,1,3)
qqplot(C)
title('Vzorka C')
```

```
figure
subplot(3,1,1)
normplot(A)
title('Vzorka A')
subplot(3,1,2)
normplot(B)
title('Vzorka B')
subplot(3,1,3)
normplot(C)
title('Vzorka C')

figure
subplot(3,1,1)
histogram(A)
title('Vzorka A')
subplot(3,1,2)
histogram(B)
title('Vzorka B')
subplot(3,1,3)
histogram(C)
title('Vzorka C')
```

c) V záverečnej úlohe využijeme Jarque-Berov test normality. Taktiež sa pozrieme na hodnoty šikmosti a špicatosti dátových vzoriek. Výsledky sú zhrnuté v Tabuľke 2.1, v závislosti na generovaných vzorkách sa môžu meniť. Z posledného stĺpca tabuľky je zrejmé, že hypotézu o normálnom rozdelení dát by sme nezamietli pre vzorku 20 a 200 pozorovaní, naopak by došlo k zamietnutí pre 2000 pozorovaní. Je dôležité poznamenať, že tento výsledok nie je platný vždy a častejšie sa ukáže, že hypotézu pre vyšší počet vzoriek nezamietame. Avšak občas preváži prílišná citlivosť testu a hypotéza o normalite dát je (pre dáta inak pochádzajúcich z normálneho rozdelenia) zamietnutá, preto je napríklad dobré využiť viac typov testov na túto otázku.

```
[Ha,Pa] = jbtest(A);
[Hb,Pb] = jbtest(B);
[Hc,Pc] = jbtest(C);

S_a = skewness(A);
K_a = kurtosis(A);
```

```
S_b = skewness(B);
K_b = kurtosis(B);
S_c = skewness(C);
K_c = kurtosis(C);

fprintf('Výsledky testov normality\n')
fprintf('Vzorka   Šikmosť   Špicatosť   p-hodnota JB test)
fprintf('  A       %1.3f     %1.3f       %1.3f
        \n',S_a,K_a,Pa)
fprintf('  B       %1.3f     %1.3f       %1.3f
        \n',S_b,K_b,Pb)
fprintf('  C       %1.3f     %1.3f       %1.3f
        \n',S_c,K_c,Pc)
```

Vzorka	Šikmosť	Špicatosť	p-hodnota JB testu
A	0.119	1.926	0.402
B	0.056	2.621	0.474
C	-0.066	2.796	0.083

Tabuľka 2.1: Zhrnutie výsledkov

2.2 T-testy a test zhody rozptylov

Táto podkapitola bude venovaná postupne jednoduchému, párovému a dvojvýberovému t-testu na ilustračných príkladoch. T-test (Studentov t-test) je metódou matematickej štatistiky, ktorá umožňuje overiť niektorú z nasledujúcich hypotéz:

- či normálne rozdelenie, z ktorého pochádza určitý náhodný výber, má určitú konkrétnu strednú hodnotu, pričom rozptyl je neznámy
- či dve normálne rozdelenia majúce rovnaký (aj keď neznámy) rozptyl, z ktorého pochádzajú dva nezávislé náhodné výbery, majú rovnaké stredné hodnoty (resp. rozdiel týchto stredných hodnôt je rovný určitému danému číslu)

V prvom prípade môže byť náhodný výber tvorený buď jednotlivými hodnotami (potom sa jedná o jednovýberový t-test), alebo dvojicami hodnôt, u ktorých sa skúmajú ich rozdiely (potom sa jedná o párový t-test). V druhom

prípade ide o dvojvýberový t-test. V praxi sa t-test často používa k porovnaniu, či sa výsledky meraní na jednej skupine významne líšia od výsledkov meraní na druhej skupine.

Jednoduchý t-test spočítame v Matlabe pomocou príkazu **ttest**. Funkcia má pevne daný zápis: [h,p,ci,stats] = ttest(x,'alpha',α,'tail','both'). Parameter *'alpha'* slúži na voľbu hladiny významnosti α. Parameter *'tail'* rozlišuje tvar alternatívnej hypotézy: obojstranná/ľavostranná/pravostranná. Výstupnými parametrami sú: h (rozhodnutie o (ne)zamietnutí nulovej hypotézy), p (p-hodnota testu), *ci* (konfidenčný interval), *stats* (hodnota testovej štatistiky). Párový t-test vykonávame pomocou už známej funkcie **ttest**, s tým rozdielom, že našimi vstupnými parametrami budú tentokrát dve náhodné vzorky: [h,p,ci,stats] = ttest(x,y).

V prípade, že pracujeme s dvoma nezávislými výbermi, využívame v Matlabe funkciu **ttest2**. Vstupné parametre a výstupné hodnoty sú analogické ako v prípade funkcie **ttest**.
Funkcia **vartest2** spočíta dvojvýberový F-test zhody rozptylov. Jej zápis je obdobný ako pre predošlé funkcie: [h,p,ci,stats] = vartest2(x,y).

Cvičenie 2.3 - Jednovýberový t-test

Zadanie: Autobusový dopravca uskutočnil pri jednej zahraničnej ceste pri 100 cestujúcich kontrolu hmotnosti batožiny. Maximálna povolená hmotnosť batožiny stanovená prepravcom je 20 kg. Predpokladajme, že hmotnosť batožiny sa riadi normálnym rozdelením. Na základe váženia bola vypočítaná priemerná váhová hodnota batožiny 20,3 kg a výberová smerodajná odchýlka 2,7 kg. Náhodne vygenerujte hmotnosť 100 batožín pomocou generátoru so zmieneným rozdelením.

a) Zistite počet batožín z vygenerovanej vzorky s váhou nad 20 kg.

b) Je možné na hladine významnosti 10 % usúdiť, že stredná hodnota váhy batožiny na vygenerovanej vzorke neprevýši 20 kg?

c) Vyskúšajte rovnakú analýzu pre veľkosť vygenerovanej vzorky 1000 batožín, líšia sa nejak výsledky?

Riešenie: Predtým, než začneme riešiť jednotlivé úlohy, potrebujeme si vygenerovať 100 údajov o hmotnosti batožín pomocou generátora náhodných čísiel. V Matlabe na generovanie vzoriek z pseudonormálnych čísiel slúži funkcia **randn**.

```
>> n = 100;
>> alfa = 0.1;
>> vaha_zavazadla = 20.3 + 2.7*randn(n,1);
```

a) Riešenie prvej úlohy je veľmi jednoduché.

```
>> nadmerna_zavazadla = sum(vaha_zavazadla>20)
```

b) Testujeme nulovú hypotézu H_0: $\mu = 20$, oproti alternatíve H_1: $\mu \neq 20$. Využijeme funkciu **ttest**.

```
>> mu = 20;
>> [H,P,CI] = ttest(vaha_zavazadla,mu, 'alpha', alfa);
```

Následne rozhodneme o (ne)zamietnutí nulovej hypotézy.

```
if H==1
    fprintf('Zamítáme nulovou hypotézu. \n')
else
    fprintf('Nezamítáme nulovou hypotézu. \n')
end

fprintf('Na základě vygenerovaného vzorku, se bude střední
hodnota váhy \nzavazadla pohybovat
v rozmezí %2.4f - %2.4f. \n',CI(1,1),CI(2,1));
```

c) Posledná úloha sa rieši analogicky. Jediným rozdielom je počet vygenerovaných vzoriek na začiatku príkladu, ktorý zmeníme na 1000. Následne už len zopakujeme postup v bode b).

Cvičenie 2.4

Zadanie: Na základe náhodného výberu 100 výrobkov chceme na hladine významnosti 0,05 overiť predpoklad, že produkcia výrobkov trvá viac ako 75 hodín. Dáta sú uložené v súbore *dlzka.mat*.

a) Vypočítajte p-hodnotu testu a rozhodnite o platnosti hypotézy.

b) Zistite hodnotu realizovanej testovej štatistiky, stupne voľnosti a odhad smerodajnej odchýlky.

c) Na základe konfidenčného intervalu učiňte príslušný záver o platnosti hypotézy.

Riešenie: Riešenie začneme opäť prípravou prostredia v Matlabe a načítaním dát.

```
>> load('dlzka.mat')
```

a) Testujeme nulovú hypotézu H_0: dĺžka ≤ 75 oproti jednostrannej alternatíve, H_1: dĺžka > 75. Uskutočníme t-test a z výstupných parametrov si vypíšeme ten, ktorý stojí pre p-hodnotu testu.

```
>> mu = 75;
>> [~,p,ci,stats] = ttest(dlzka,mu,'tail','right');

>> fprintf('\nP-hodnota testu: %1.4f\n',p)
```

b) Hodnotu testovej štatistiky, stupne voľnosti a hodnotu smerodajnej odchýlky získame vytiahnutím príslušných hodnôt zo štruktúry.

```
>> fprintf('\nHodnota testovej statistiky: %2.3f\n',stats.tstat)
>> fprintf('Stupne volnosti: %2.0f\n',stats.df)
>> fprintf('Smerodatna odchylka: %2.3f\n',stats.sd)
```

c) O (ne)platnosti našej hypotézy môžeme rozhodnúť aj na základe konfidenčného intervalu.

```
>> fprintf('\nKonfidencny interval: %2.3f - %2.3f\n',ci(1),ci(2))
```

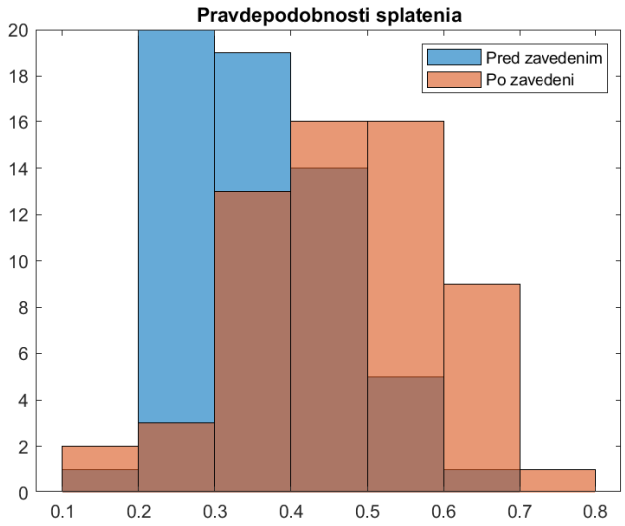
Cvičenie 2.5 - Párový t-test

Zadanie: Komerčná banka testuje nové pravidlá pre čerpanie úverov. Má k dispozícii odhady pravdepodobnosti splatenia šesťdesiatich klientov pred a po zavedení nových pravidiel. Na hladine významnosti 5% rozhodnite o tom, či existuje rozdiel v odhade splatenia úveru pri nových a starých pravidlách.

Dáta sú v dostupné v súbore *pravdepopodobnost_splaceni.mat*.

Riešenie: Dáta si načítame a vykreslíme si k nim príslušný histogram pomocou funkcie **histogram**. Viď Obrázok 2.3.

```
load('pravdepopodobnost_splaceni.mat')

figure
histogram(pred)
hold on
histogram(po)
title("Pravdepodobnosti splatenia")
legend("Pred zavedenim","Po zavedeni")
```



Obrázok 2.3: Histogramy

Vzhľadom k povahe dát, využijeme funkciu **ttest** na uskutočnenie párového t-testu. Následne opäť rozhodneme o (ne)zamietnutí nulovej hypotézy a vypíšeme si konfidenčný interval.

```
[H,P,CI,STATS] = ttest(pred,po);

if H==1
    fprintf('Zamítáme nulovou hypotézu. \n')
else
    fprintf('Nezamítáme nulovou hypotézu. \n')
```

```
end

fprintf('Na základě vzorku, se rozdíl v pravděpodobnosti
    čerpání úvěrů \npři nových pravidlech v rozmezí
    od %2.4f do %2.4f. \n',CI(1,1),CI(2,1));
```

Cvičenie 2.6 - Dvojvýberový t-test

Zadanie: Chceme otestovať, ktorá z dvoch metód učenia slovíčok je efektívnejšia. Máme k dispozícií dáta od 440 žiakov. Polovica z nich sa učí slovíčka metódou A a druhá polovica metódou B. Počet správne zodpovedaných slovíčok sumarizuje dátový súbor *slovicka.mat*.
Na hladine významnosti 0,05 skúmajte, či:

a) sa rozptyly daných metód od seba významne líšia,

b) sú obe metódy zameniteľné,

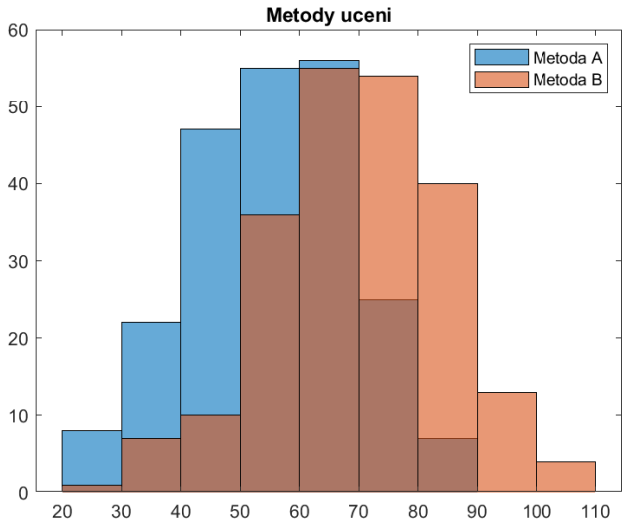
c) je metóda A efektívnejšia ako metóda B.

Riešenie:

a) Dáta si opäť načítame. O zhodnosti rozptylov budeme rozhodovať pomocou dvoch spôsobov. Vizuálne pomocou histogramov a formálne pomocou testu **vartest2**. Príslušné histogramy sú zobrazené na Obrázku 2.4. V prípade testu formulujeme nulovú hypotézu H_0: $\sigma_{mA} = \sigma_{mB}$, oproti alternatíve H_1: $\sigma_{mA} \neq \sigma_{mB}$.

```
load('slovicka.mat');

figure
histogram(Metoda_A)
hold on
histogram(Metoda_B)
title('Metody uceni')
legend("Metoda A","Metoda B")

[h,p,ci,stats] = vartest2(Metoda_A,Metoda_B);

if h==1
    fprintf('Zamítáme nulovou hypotézu o shodnosti
```

```
        rozptylů výběrů. \n')
else
    fprintf('Nezamítáme nulovou hypotézu o shodnosti
        rozptylů výběrů. \n')
end
```



Obrázok 2.4: Histogramy

b) V druhej úlohe testujeme pomocou dvojvýberového t-testu nulovú hypotézu H_0: $\mu_{mA} - \mu_{mB} = 0$, oproti alternatíve H_1: $\mu_{mA} - \mu_{mB} \neq 0$. Využijeme funkciu **ttest2**.

```
[h,p,ci,stats] = ttest2(Metoda_A,Metoda_B);

if h==1
    fprintf('Zamítáme nulovou hypotézu o rovnosti
        středních hodnot výběrů. \n')
else
    fprintf('Nezamítáme nulovou hypotézu  o rovnosti
        středních hodnot výběrů. \n')
end
```

c) V záverečnej úlohe si ukážeme zápis testovania jednostrannej alternatívy dvojvýberového t-testu. Testujeme H_0: $\mu_{mA} - \mu_{mB} \leq 0$, oproti pravostrannej alternatíve H_1: $\mu_{mA} - \mu_{mB} > 0$.

```
[h,p,ci,stats] = ttest2(Metoda_A,Metoda_B,'Tail','left')

if h==1
    fprintf('Zamítáme nulovou hypotézu, že by metoda
        učení B byla efektivnější než metoda A.\n')
else
    fprintf('Na 5% hladině významnosti nemůžeme
        zamítnut nulovou hypotézu, že by metoda učení B
        byla efektivnější než metoda A.\n')
end
```

2.3 Analýza rozptylu - ANOVA

Táto podkapitola bude venovaná metóde ANOVA. ANOVA je metódou matematickej štatistiky, ktorá umožňuje overiť, či na hodnotu náhodnej veličiny pre určitého jedinca má štatisticky významný vplyv hodnota niektorého znaku, ktorý sa u jedinca dá pozorovať. Ukážeme si taktiež použitie testu zhody rozptylov.

Funkcia **vartestn** spraví Bartlettov test zhody rozptylov pre stĺpce vstupnej matice X. Zápis: [P,STATS] = vartestn(X). Funkcia **anova1** vykonáva jednofaktorovú analýzu rozptylu, pričom porovnávame viacero skupín a hľadáme či sa štatisticky významne odlišujú len v 1 faktore. Zápis: p = anova1(X). Výstupná hodnota p je p-hodnota pre nulovú hypotézu, že skupiny sa od seba významne nelíšia.

Cvičenie 2.7

Zadanie: Firma *Nápoj v plechu* testovala nový design plechoviek. Bolo navrhnutých 5 designov: Kone, Psíkovia, Wombati, Pandy a Mačičky. Plechovky boli rozdistribuované do 100 obchodov a boli sledované čísla o ich predajoch. Výsledky predajov sú zaznamenané v súbore *design.mat*, tento súbor načítajte.

a) Otestujte či je splnený predpoklad zhody rozptylov jednotlivých výberov.

b) Rozhodnite, či má design plechovky štatisticky významný vplyv na počet predajov plechoviek. Aký design by mala firma zvoliť?

Riešenie:

a) Opäť naším prvým krokom bude načítanie dát. Potom vyskúšame použitie testu zhody rozptylov.

```
>> load('design.mat')

>> [P,STATS] = vartestn(Plechovky)
```

Výstup, ktorý nám funkcia **vartestn** vracia je nasledovný:

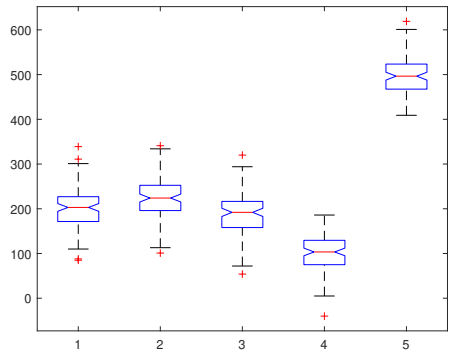
```
Group                    Count      Mean     Std Dev
-------------------------------------------------
1                        100        200.36    45.7582
2                        100        221.66    46.4945
3                        100        189.74    46.9103
4                        100        102.42    42.6994
5                        100        499.25    42.7753
Pooled                   500        242.686   44.9646

Bartlett's statistic     1.6356
Degrees of freedom       4
p-value                  0.8024
```

Keďže je p-hodnota testu väčšia ako 0,05, môžeme považovať rozptyly všetkých výberov za zhodné.

b) Riešenie úlohy b) je samotná analýza rozptylu a jej vyhodnotenie. Využijeme funkciu **anova1**.

```
>> [p,tbl,stats] = anova1(Plechovky);
```

Z obrázku 2.5 je zrejmé, že medzi prvými štyrmi druhmi plechoviek nebude pravdepodobne významný rozdiel v počte predaných kusov (intervaly sa prekrývajú). Narozdiel od zmienených troch má posledný druh výrazne vyššiu strednú hodnotu počtu predaných plechoviek oproti



Obrázok 2.5: Boxploty

predchádzajúcim druhom plechoviek. Formálne overenie môžeme vykonať s využitím tabuľkového výstupu funkcie **anova1**, ktorý zachytáva tiež štatistické vyhodnotenie testov.

Source	SS	df	MS	F	Prob>F
Columns	9.05e+06	4	2263412.47	1119.49	2.25e-246
Error	1.00e+06	495	2021.82		
Total	1.00e+07	499			

V našom prípade je výsledná p-hodnota (2,25031e-246) menšia ako 0,05 a teda hypotézu o zhode stredných hodnôt zamietame. Odpoveď na otázku b) je teda, že design má vplyv na počet predajov plechoviek, a v priemere najvyšší a štatisticky rozdielny počet predajov oproti iným designom ma design mačičiek.

Cvičenie 2.8

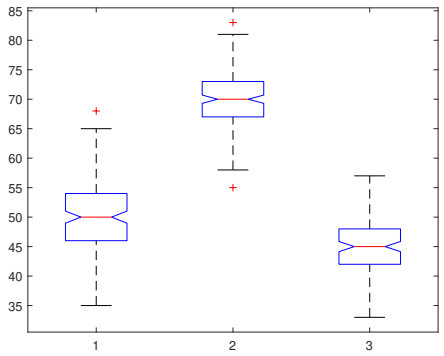
Zadanie: Úrad pre kontrolu kvality chcel porovnať kvalitu mliečnych výrobkov rôznych firiem v okrese. Za týmto účelom bola vytvorená hodnotiaca škála a pomocou nej boli hodnotené náhodne vybrané mliečne výrobky. Bodové výsledky mliečnych výrobkov a informácie o tom, ku ktorej firme údaj patrí, sú uvedené v matici s radením $[hodnoceni; firma]$. Túto maticu je možné nájsť v súbore *mlecne.mat*. Na hladine významnosti $\alpha = 0,05$ zistite, či je „priemerné“ hodnotenie mliečnych výrobkov rovnaké. Ako ovplyvní výsledok prieskumu zmena hladiny významnosti na $0,01$?

Riešenie:

Riešenie úlohy dosiahneme pomocou rozboru analýzy rozptylu, ktorú získame príkazom **anova1**.

```
>> anova1(hodnoceni,firma)
```

Ukážku grafického (obrázok 2.6) a tabuľkového výstupu je možné opäť vidieť nižšie. Z výstupu je opäť zrejmé, že môžeme hypotézu o zhode stredných hodnôt zamietnuť a to aj na 1 % hladine významnosti.



Obrázok 2.6: Boxploty

Source	SS	df	MS	F	Prob>F
Groups	52945.1	2	26472.6	1008.28	2.13359e-166
Error	11736	447	26.3		
Total	64681.1	449			

2.4 Korelačná analýza

Korelácia znázorňuje mieru lineárnej závislosti dvoch kvantitatívnych veličín (meria vzájomný vzťah dvoch premenných). Dve premenné sú korelované, ak určité hodnoty jednej premennej majú tendenciu sa vyskytovať spoločne s určitými hodnotami druhej premennej.

Funkcia corr(...) vracia maticu rozmeru PxP obsahujúcu hodnotu korelačného koeficientu medzi každou dvojicou stĺpcov v matici X. Zápis: R =

corr(X). Taktiež vracia maticu korelačných koeficientov medzi dvoma maticami X a Y. Zápis: R = corr(X,Y). Dodatočný výstupný parameter *'pval'* vracia maticu p-hodnôt testu hypotézy o žiadnej korelácií oproti alternatívne, že medzi prvkami existuje nenulová korelácia. Zápis: [R,pval] = corr(...). V prípade, že chceme testovanie obmedziť na niektorú z jednostranných hypotéz špecifikujeme dodatočný parameter a jeho hodnotu, pomocou príkazu [...] = corr(...,'PARAM',VAL), kedy za parameter dosadíme príkaz *'tail'* a za hodnotu *VAL* dodáme informáciu o tvare alternatívnej hypotézy. Tj., *'both'* v prípade, keď testujeme či je korelácie nenulová; *'right'* v prípade, keď testujeme či je korelácia väčšia ako nula (kladná) a *'left'* v prípade, keď testujeme či je korelácia menšia ako nula (záporná).

Cvičenie 2.9

Zadanie: Vygenerujme si dáta o predaji zmrzliny a teplote vody a vzduchu. Vieme, že sa teplota vzduchu riadi normálnym rozdelením $N(26, 3^2)$, ďalej vieme, že rozdelenie teploty vody je možné zapísať ako

$$teplota\_vody = round(teplota\_vzduchu - 10 + 3 * randn(n, 1), 1)$$

a počet predaných zmrzlín ako

$$pocet\_predajov = round(teplota\_vzduchu * 10 + 20 * randn(n, 1)),$$

kde n je počet vygenerovaných pozorovaní.

a) Vygenerujte si príslušné dáta pre $n = \{20, 200, 2000\}$ tak, aby vyšší počet pozorovaní vždy využíval všetky pozorovania z predošlého menšieho vzorku.

b) Vypočítajte korelačný koeficient pre všetky tri veľkosti vzoriek. Je možné sledovať nejakú zmenu v korelačnom koeficiente? Prečo?

c) Otestujte štatistickú významnosť korelačných koeficientov na hladine významnosti $0,01$. Sú výsledky pre všetky veľkosti výberov obdobné? Dopracujeme sa k rovnakým výsledkom pokiaľ skript spustíme niekoľkokrát za sebou?

Riešenie:

a) V prvom kroku si vygenerujeme príslušné dátové súbory. Najjednoduchším spôsobom je vygenerovať pre všetky tri premenné vektor s

veľkosťou 2000 pozorovaní a z toho potom urobiť tri matice tak, že matica X1 bude obsahovať prvých dvadsať pozorovaní všetkých premenných, matica X2 bude obsahovať prvých dvesto pozorovaní všetkých premenných a matica X3 bude obsahovať všetky pozorovania všetkých premenných.

```
>> n = 2000;

>> teplota_vzduchu = round(26 + 3*randn(n,1),1)
>> teplota_vody = round(teplota_vzduchu - 10 + 3*randn(n,1),1)
>> pocet_prodeju = round(teplota_vzduchu*10 + 20*randn(n,1))

>> X1 = [pocet_prodeju(1:20,:),teplota_vzduchu(1:20,:), ...
    teplota_vody(1:20,:)]
>> X2 = [pocet_prodeju(1:200,:),teplota_vzduchu(1:200,:), ...
    teplota_vody(1:200,:)]
>> X3 = [pocet_prodeju,teplota_vzduchu,teplota_vody]
```

b),c) Následne vypočítame korelačné koeficienty a zistíme ich štatistickú významnosť u jednotlivých podvýberov.

```
>> [rho1,pval1] = corr(X1)
>> [rho2,pval2] = corr(X2)
>> [rho3,pval3] = corr(X3)
```

Výstupom príkladu môžu byť nasledujúce korelačné matice (pri každom spustení sa budú čísla líšiť).

$$\rho_1 = \begin{pmatrix} 1 & 0,937 & 0,732 \\ & 1 & 0,8138 \\ & & 1 \end{pmatrix} \quad \rho_2 = \begin{pmatrix} 1 & 0,868 & 0,627 \\ & 1 & 0,736 \\ & & 1 \end{pmatrix}$$

$$\rho_3 = \begin{pmatrix} 1 & 0,832 & 0,599 \\ & 1 & 0,713 \\ & & 1 \end{pmatrix}$$

Vidíme, že každá korelačná matica vykazuje mierne odlišné výsledky príslušných korelačných koeficientov. To je spôsobené tým, že pri vyššom počte pozorovaní dochádza k spresneniu výpočtu korelačného koeficientu medzi danými dvoma data-generujúcimi procesmi. Štatistickú

významnosť korelačných koeficientov overíme výpisom matíc pval1, pval2 a pval3.

$$pval_1 = \begin{pmatrix} 1 & 0,0000 & 0,0002 \\ & 1 & 0,0000 \\ & & 1 \end{pmatrix} \quad pval_2 = \begin{pmatrix} 1 & 0,0000 & 0,0000 \\ & 1 & 0,0000 \\ & & 1 \end{pmatrix}$$

$$pval_3 = \begin{pmatrix} 1 & 0,0000 & 0,0000 \\ & 1 & 0,0000 \\ & & 1 \end{pmatrix}$$

Všetky p-hodnoty pre vybrané dvojice korelačných koeficientov sú menšie ako hladina významnosti 1 % a teda všetky korelačné koeficienty sú štatisticky významné. Odpoveď na otázku: „Dopracujeme sa k rovnakým výsledkom, pokiaľ skript spustíme niekoľkokrát za sebou?“ ponecháme pre čitateľa ako cvičenie.

Ďalšie funkcie

Nasledujúci zoznam zhŕňa ďalšie užitočné funkcie, ktoré pri štatistických aplikáciách v Matlabe možno využiť.

- **iqr** - príkaz vracia medzikvartilové rozpätie vstupných hodnôt vektoru alebo matice X. Zápis: Y = iqr(X).
- **range** - príkaz vracia rozpätie hodnôt vstupného vektoru alebo matice X. Zápis: Y = range(X).
- **moment** - počíta centrálne momenty všetkých stupňov pre vstup X. Zápis: Y = moment(X, S), kde S značí stupeň centrálneho momentu.
- **quantile** - počíta kvantily z hodnôt X. Zápis: Y = quantile(X,P), kde P je skalár alebo vektor, ktorý obsahuje kumulatívne hodnoty pravdepodobnosti.
- **mvnrnd** - funkcia generuje náhodné vektory z mnohorozmerného normálneho rozdelenia s vektorom priemerov mu a kovariančnou maticou sigma. Zápis: R = mvnrnd(mu, sigma), kde mu je matica tvaru N x D a sigma je symetrická, pozitívne semidefinitná matica D x D.
- **tcdf** - príkaz počíta kumulatívnu distribučnú funkciu Studentovho t-rozdelenia s v stupňami voľnosti na hodnotách vektora X. Zápis: P = tcdf(X,v). V prípade voľby *'upper'*, vracia komplement ku kumulatívnej distribučnej funkcii.

- **tstat** - príkaz vracia priemer m a rozptyl v Studentovho t-rozdelenia s n stupňami voľnosti. Zápis: [m,v] = tstat(n).
- **fcdf** - príkaz počíta kumulatívnu distribučnú funkciu F-rozdelenia s v1 a v2 stupňami voľnosti na hodnotách vektora X. Zápis: P = fcdf(X,v1,v2). V prípade voľby *'upper'* vracia komplement ku kumulatívnej distribučnej funkcii.
- **fstat** - príkaz vracia priemer m a rozptyl v F-rozdelenia s v1 a v2 stupňami voľnosti. Zápis: [m,v] = fstat(v1,v2).
- **chi2cdf** - príkaz počíta kumulatívnu distribučnú funkciu χ^2-rozdelenia s v stupňami voľnosti a v hodnotách vektoru X. Zápis: P = chi2cdf(X,v). Opäť voľba *'upper'* vracia komplement ku kumulatívnej distribučnej funkcii.
- **chi2stat** - príkaz vracia priemer m a rozptyl v χ^2-rozdelenia s v stupňami voľnosti. Zápis: [m,v] = chi2stat(v).
- **cdfplot** - funkcia zobrazuje empirickú kumulatívnu distribučnú funkciu pozorovaní z dátového vzorku vektora X. Zápis: [H, stats] = cdfplot(X). V prípade pripísania výstupného parametra *stats* vracia aj štatistický prehľad uložený v štruktúre, ktorého obsahom je minimum (*stats.min*), maximum (*stats.max*), výberový priemer (*stats.mean*), výberový medián (*stats.median*), výberová štandardná odchýlka (*stats.std*).
- **corrcoef** - funkcia počíta maticu R korelačných koeficientov pre pole X, kde každý stĺpec predstavuje premennú a každý riadok je pozorovanie. Zápis: R = corrcoef(X), prípadne R = corrcoef(X,Y).
- **corrplot** - vykresľuje korelácie premenných. Zápis: corrplot(X).

Neriešené príklady

1. Z hodín štatistiky vieme, že ľubovoľná lineárna transformácia dvoch nezávislých náhodných veličín pochádzajúcich z normálneho rozdelenia je taktiež normálna veličina.

   a) Vygenerujte si dve nezávislé náhodné veličiny rovnakých dĺžok X_1 a X_2 z normálneho rozdelenia.

   b) Pomocou výpočtových a grafických metód overte, že lineárna transformácia $Y = 2X_1 + 3X_2$ taktiež pochádza z normálneho rozdelenia.

2. V dátovom súbore *trojice.mat* nájdete tri náhodné vzorky. Pracujte so všetkými troma vzorkami a zistite, ktorá z nich má normálne rozdelenie. Pokúste sa odhadnúť rozdelenie aj ostatných vzoriek.

3. Učiteľ chce zistiť, či sú ním vytvorené dva testy rovnako náročné. Každému zo svojich 100 žiakov preto dal vyplniť obe verzie testu. Dosiahnute bodové výsledky sú obsiahnuté v súbore *testy.mat*. Na hladine významnosti 0,05 overte, či sa výsledné hodnotenie variant líši.

4. Chceme zistiť či rastliny, ktoré sú vystavené priamemu slnečnému žiareniu rastú rýchlejšie ako rastliny vysadené v tieni. Predpokladajme, že rast je normálne rozdelená náhodná veličina. V súbore *rostliny.mat* máme zaznamenaný náhodný výber rastu dvesto rastlín z každého prostredia.

   a) Zistite, či sa rozptyly u oboch výberov štatisticky významne líšia.

   b) Preukážte hypotézu, že rastliny na slnku rastú rýchlejšie ako rastliny v tieni. ($H_1 : \mu_{sl} > \mu_{st}$)

5. Výrobca práčok vyrába práčky s predným plnením, vrchom plnené práčky a práčky so sušičkou. Výrobca podrobil práčky testovaniu a pri každej si zapísal počet pracích cyklov pred prvou poruchou. Zistite, či sa poruchovosť práčok líši v závislosti na jej type, prípadne, ktoré typy práčok sa líšia (dáta sú uložené v súbore *pracky.mat*).

6. Marketingové oddelenie skúmalo, aký vplyv má druh reklamy na jej úspešnosť. Oddelenie vytvorilo tri typy obdobne sugestívnych reklám a náhodne ich sprostredkovalo medzi testovanú vzorku spotrebiteľov. Výsledky uložené v súbore *reklama.mat* uvádzajú bodové hodnotenie subjektov na škále 0-100. Na hladine významnosti $\alpha = 0,05$ zistite, či sú na základe hodnotiteľov všetky reklamy „priemerne“ rovnako sugestívne.

7. Z veľkého súboru domácností bolo náhodne vybraných 50 jednočlenných, 80 dvojčlenných, 100 trojčlenných, 100 štvorčlenných a 70 päťčlenných domácností, spolu teda 400 domácností a boli sledované ich mesačné výdavky za potraviny a nápoje pripadajúce na jedného člena domácnosti (v Kč). Overte pomocou analýzy rozptylu, či sa mesačné výdavky za potraviny (na osobu) líšia podľa počtu členov domácnosti. Dáta sú dostupné v súbore *domacnosti.mat*.

8. Existujú dve metódy, ktorými môže firma Rosnička vyrábať svoj nový produkt. Pre obe metódy uskutočnila 50 meraní celkovej doby potrebnej k výrobe nového produktu. Namerané hodnoty sú uložené v dátovom súbore *metody.mat*. Na hladine významnosti 1% testujte hypotézu o zhodnosti oboch metód výroby.

9. V dátovom súbore *alkohol.mat* je uvedená priemerná spotreba tvrdého alkoholu v decilitroch za deň (X) a úmrtnosť na cirhózu pečene a alkoholizmus v niektorých vymyslených krajinách (Y). Určte na hladine významnosti 0,05, či úmrtnosť na cirhózu pečene a alkoholizmus závisia na spotrebe alkoholu.

10. Sledovaním nákladov a predajných cien produktu u 200 výrobcov bol získaný dvojrozmerný náhodný výber (predpokladajme, že pochádza z dvojrozmerného normálneho rozdelenia). Dátový súbor je uložený ako *vyrobci.mat*. Vypočítajte realizáciu výberového korelačného koeficientu. Na hladine významnosti 5% testujte hypotézu, že náklady a cena produktu spolu nesúvisia, oproti alternatíve, že s klesajúcimi nákladmi klesá i cena.

11. V súbore *obrat.txt* je uvedený obrat zahraničného obchodu (Y) a počet obyvateľov niekoľkých vybraných štátov (X). Hint: pri riešení bodu b) a c) využite príkaz help corr.

    (a) Vhodne upravte a importujte dátový súbor do Matlabu.

    (b) Zmerajte tesnosť lineárnej závislosti pomocou Pearsonovho korelačného koeficientu a na hladine významnosti 5% testujte hypotézu $H_0 : \rho = 0$ oproti alternatívnej hypotéze $H_0 : \rho \neq 0$.

    (c) Zmerajte tesnosť lineárnej závislosti pomocou Spearmanovho korelačného koeficientu a na hladine významnosti 5% testujte hypotézu $H_0 : \rho_s = 0$ oproti alternatívnej hypotéze $H_0 : \rho_s \neq 0$.