

**SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE  
STAVEBNÁ FAKULTA**

**Analýza rozptylu a kovariancie**

**Bakalárska práca**

EVIDENČNÉ ČÍSLO: SVF-5342-50366

Študijný program: Matematicko-počítačové modelovanie

Číslo a názov študijného odboru: 9.1.9 aplikovaná matematika

Školiace pracovisko: Katedra matematiky a deskriptívnej geometrie

Vedúci záverečnej práce/školiťel': RNDr. Jana Kalická, PhD.

**Bratislava 2010**

**Ivona Fojtíková**

## **Pod'akovanie**

Ďakujem vedúcej práce RNDr. Jane Kalickej, PhD., za jej pomoc, odbornú radu a pripomienky pri vypracovaní predkladanej práce.

## Súhrn

Predkladaná bakalárska práca sa zaoberá testami analýzy rozptylu a kovariancie. Analýza rozptylu (ANOVA) skúma závislosť intervalovej premennej  $Y$  na jednej alebo viacerých nominálnych premenných (faktoroch). Umožňuje testovať hypotézy o rovnostiach stredných hodnôt pri viacerých výberoch. Testy jednofaktorovej a dvojfaktorovej analýzy sú implementované v systéme Mathematica. Práca dopĺňa tieto testy o testy za špecifických podmienok – pri nerovnosti rozptylov, pri malom počte meraní v jednotlivých triedach a pod. Vhodnosť použitia testov je overená v simulačných štúdiách.

Druhá časť je venovaná analýze kovariancie (ANCOVA). Analýza kovariancie predstavuje rozšírenie ANOVA o prvky lineárnej regresie. Bakalárska práca obsahuje aj programy testov v softvéri Mathematica. Všetky spomínané testy sú podrobnejšie popísané v teoretickej časti. V softvéri Mathematica sú naprogramované vo forme užívateľsky prijateľných procedúr a prakticky vyskúšané na príkladoch.

**Kľúčové slová:** testy hypotéz, hladina významnosti, analýza rozptylu, analýza kovariancie

## **Abstract**

The baccalaureate work deals with ANOVA and ANCOVA tests. Analysis of variance (ANOVA) describes the dependence  $Y$  of one or more nominal variables (factors). In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal. One-factor and two-factor ANOVA tests are implemented in system Mathematica. The work deals with tests under specific conditions (unequal variances in the groups, a small number of measurements in individual classes). Applicability of the tests is verified in simulation studies.

The second part deals with analysis of covariance (ANCOVA). Analysis of covariance represents an extension of ANOVA. The theory of ANOVA is completed by linear regression. The baccalaureate work also contains programs for ANCOVA tests. These tests are programmed in Mathematica system. They are programmed in user-friendly procedures and practically tested on examples.

**Key words:** hypothesis testing, significance level, analysis of variance, analysis of covariance

# Obsah

1. Úvod	1
2. Dvojvýberové testy	3
2.1. Dvojvýberový nepárový t – test	3
2.2. Dvojvýberový t – test v softvéri Mathematica	4
3. Analýza rozptylu	4
3.1. Jednofaktorová analýza rozptylu	4
3.2. Jednofaktorová ANOVA v systéme Mathematica	6
4. Testy pre strednú hodnotu	7
4.1. Cochranov test	7
4.2. Welchov test	8
4.3. Modifikovaný Welchov test	9
4.4. Brown-Forsytheov test	10
5. Simulačné štúdie	10
6. Dvojfaktorová analýza rozptylu bez interakcií	12
6.1. Dvojfaktorová ANOVA bez interakcií v softvéri Mathematica	14
7. Dvojfaktorová ANOVA s interakciami	15
7.1. Dvojfaktorová ANOVA s interakciami v softvéri Mathematica	17
7.2. Tukeyho test aditivity	18
8. Analýza kovariancie	19
9. Záver	23
10. Použitá literatúra	24
Prílohy	25
Príloha č. 1: CD	25

# 1. Úvod

Štatistika je veda zaoberajúca sa zberom, analýzou, interpretáciou a prezentáciou dát. Používa sa v širokej škále vedeckých disciplín od prírodných vied, techniky až po sociálne a humanitné vedy.

Ekonomom slúži na modelovanie ekonomiky, predpovedanie ekonomického vývoja. V bankovníctve slúži zase pri rozhodovaní o poskytnutí úveru. Meteorológovia využívajú štatistiku pri čoraz presnejšom predpovedaní počasia a v medicíne na včasné odhaľovanie rôznych ochorení, určovanie správnej liečby a pri vývoji a testovaní nových liekov.

Štatistické metódy môžu byť použité na zhrnutie zozbieraných dát, hovoríme o deskriptívnej štatistike. Okrem toho vzory zistené v dátach môžu byť modelované spôsobom, ktorý umožňuje zohľadniť neurčitost' a náhodnosť a následne vyvodzovať závery na celú populáciu. To sa nazýva indukčná alebo inferenčná štatistika. Deskriptívna a indukčná štatistika tvoria aplikovanú štatistiku. Okrem nej existuje disciplína nazvaná matematická štatistika, ktorá sa zaoberá aj teoretickými základmi štatistiky [8].

V praxi často nastáva situácia, že je treba porovnať stredné hodnoty viacerých výberov ako napríklad testovanie rôznych druhov liekov. Pri dvoch výberoch ešte možno použiť dvojvýberový, alebo párový t-test. Pri viacerých výberoch treba použiť analýzu rozptylu. Táto metóda vychádza z potrieb biologickej, technickej, medicínskej praxe.

Analýza rozptylu (ANOVA) skúma vzťah medzi závislou intervalovou premennou  $Y$  a jednou alebo viacerými nominálnymi premennými (faktormi). ANOVA možno použiť ak napríklad chceme zistiť, či je vzťah medzi priemerným dosiahnutým obratom obchodného zástupcu ( $Y$ ) a jeho vzdelaním (faktor 1), vekom (faktor 2), pohlavím (faktor 3) atď [8].

Analýza rozptylu bola neformálne použitá výskumníkmi už v prvom desaťročí 19. storočia metódou najmenších štvorcov. Vo svojej modernej podobe bola jedna z mnohých dôležitých štatistických inovácií Ronalda A. Fishera. Fisherov formálny návrh analýzy rozptylu je obsiahnutý v jeho článku Korelácia medzi príbuznými za predpokladu Mendelovho zákona dedičnosti z roku 1918. Jeho prvé použitie analýzy rozptylu bolo uverejnené až v roku 1921. Analýza rozptylu sa stala všeobecne známou potom, ako bola zahrnutá v roku 1925 do Fisherovej knihy Štatistické metódy pre výskumných pracovníkov [9].

Najjednoduchším prípadom ANOVA je jednofaktorová analýza rozptylu. ANOVA, ktorá obsahuje viac ako jeden faktor sa nazýva viacfaktorová analýza rozptylu. Práve popis a naprogramovanie týchto druhov ANOVA v systéme Mathematica je jedným z cieľov tejto práce.

V samotnom softvéri Mathematica je už zadefinovaná funkcia na výpočet analýzy rozptylu. Na jej správnu funkčnosť je potrebné načítanie príslušného balíka. Predpokladá sa však rovnosť rozptylov jednotlivých výberov, čo v praxi nie je vždy možné. Preto sa práca venuje testom, ktoré tento fakt berú do úvahy a taktiež testom, ktoré skúmajú rovnosť rozptylov. V systéme Mathematica je tiež implementovaná funkcia pre viacfaktorovú analýzu rozptylu. Nie je však zohľadnený prípad, keď je len jedno meranie v triede a napriek tomu je nutné počítať s interakciami. V týchto prípadoch treba použiť iné testy napr. Tukeyho test, ktoré sú obsiahnuté v práci.

Prvá časť práce sa venuje už spomínanej analýze rozptylu s využitím systému Mathematica. Druhá časť je venovaná analýze kovariancie (ANCOVA) a taktiež naprogramovaním jej testov v softvéri Mathematica. Vo všeobecnosti analýza kovariancie predstavuje rozšírenie ANOVA o prvky lineárnej regresie.

V ďalšej časti práce sú popisy programov ako aj ich výsledky prezentované na konkrétnych príkladoch. Poslednou časťou je záver, ktorý sumarizuje dosiahnuté výsledky práce.

## 2. Dvojvýberové testy

V praxi sa často stretáme so situáciou, keď treba porovnať stredné hodnoty dvoch výberov.

Najprv predpokladajme situáciu, že na  $n$  nezávislých objektoch zisťujeme dve navzájom súvisiace hodnoty (tzv. párové). Ako príklad môžu slúžiť merania prahu počuteľnosti ľavého a pravého ucha, ostrosť videnia ľavého a pravého oka atď. Máme teda k dispozícii dvojicu náhodných veličín  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Položme  $z_1 = x_1 - y_1, z_2 = x_2 - y_2, \dots, z_n = x_n - y_n$ . Takouto úpravou dostávame nové dáta  $z_1, z_2, \dots, z_n$ . Ak sa dá predpokladať, že  $z_i$  sú normálne rozdelené testujeme namiesto hypotézy o rovnosti stredných hodnôt hypotézu, že stredná hodnota je nula pomocou jednovýberového testu [5].

Do skupiny dvojvýberových testov patria okrem párových i nepárové. Pri nepárových testoch namerané údaje z dvoch rôznych objektov nemusia mať rovnakú dĺžku. Nasledujúca kapitola sa zaoberá práve jedným z týchto testov tzv. dvojvýberovým nepárovým  $t$  – testom.

### 2.1. Dvojvýberový nepárový $t$ – test

Majme namerané údaje z dvoch rôznych objektov. Merania nemusia byť rovnakej dĺžky, ale musia spĺňať predpoklad normality výberu a rovnosti disperzií oboch výberov. Nech  $X_1, \dots, X_m$  je výber z  $N(\mu_1, \sigma^2)$  a  $Y_1, \dots, Y_n$  výber z  $N(\mu_2, \sigma^2)$ . Nech tieto dva výbery sú na sebe nezávislé. Predpokladajme, že  $n \geq 2, m \geq 2, \sigma^2 > 0$ . Dvojvýberový  $t$  – test je testom hypotézy

$$H_0 : \mu_1 - \mu_2 = \delta \text{ proti } H_1 : \mu_1 - \mu_2 \neq \delta,$$

kde  $\delta$  je dané číslo (najčastejšie býva  $\delta = 0$ ). Označme

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (2.1.1)$$

$$S_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2 \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (2.1.2)$$

Potom náhodná veličina

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}} \quad (2.1.3)$$

má Studentovo  $t$  - rozdelenie s  $(m + n - 2)$  stupňami voľnosti, ktoré sú parametrom rozdelenia.

Ak  $|T| \geq t_{m+n-2}(1-\alpha)$  (príslušný kvantil  $t$  - rozdelenia), zamietame hypotézu  $H_0$  na hladine významnosti  $\alpha$ .

Dvojvýberový  $t$  – test používame v prípadoch, keď sa napr. na  $m$  pacientoch skúša pôsobenie lieku  $A$  a na iných  $n$  pacientoch pôsobenie lieku  $B$ . Účelom je zistiť, či pôsobenie oboch liekov je rovnaké.

Medzi predpoklady dvojvýberového  $t$  – testu patria normalita a rovnaký rozptyl každého výberu. Porušenie týchto predpokladov máva len malý vplyv na výsledok testu.



Napriek tomu však v prípade výraznej nenormality dávame prednosť niektorému neparametrickému testu (napr. dvojvýberový Wilcoxonov test).

## 2.2. Dvojvýberový t – test v softvéri Mathematica

V systéme Mathematica je na výpočet dvojvýberového t – testu už implementovaná procedúra. Na jej správnu funkčnosť je potrebné načítať balík

`<<HypothesisTesting`` .

Pre samotný test použijeme príkaz `MeanDifferenceTest` v tvare:

`MeanDifferenceTest` [dáta1, dáta2,  $\delta$ , ďalšie príkazy].

## 3. Analýza rozptylu

Situácia je teraz podobná ako pri dvojvýberovom t – teste, ale počet výberov môže byť vyšší. Majme  $I$  navzájom nezávislých výberov, všetky nech sú normálne rozdelené a s rovnakou disperziou  $D(X)$ . Nech dĺžka  $i$ -teho výberu je  $n_i$  a nech  $n = n_1 + n_2 + \dots + n_i$ . Úlohou je overiť hypotézu o rovnosti stredných hodnôt, čiže

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I.$$

Ak by sme testovali výbery po dvojiciach a každý test by mal tú istú hladinu významnosti  $\alpha$ , potom by sme vykonali  $I*(I-1)$  testov, pričom by sa hladina významnosti zvýšila a test by sa znehodnotil. Preto bol navrhnutý iný postup [5].

### 3.1 Jednofaktorová analýza rozptylu

Jednofaktorová (One-Way ANOVA) je najjednoduchšou formou ANOVA, ktorá skúma vzťah medzi intervalovou a nominálnou premennou. Testuje nulovú hypotézu o zhode stredných hodnôt, pričom predpokladá, že výbery majú rovnaký rozptyl. Nulová hypotéza vyjadruje, že medzi intervalovou a nominálnou premennou nie je vzťah.

Nech  $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$  je náhodný výber z rozdelenia  $N(\mu_i, \sigma^2)$ . Označme  $Y_{ij}$   $j$ -te meranie v  $i$ -tom výbere. Samotné merania, súčty v jednotlivých triedach a priemery pre jednotlivé triedy zapíšeme do nasledujúcej tabuľky:

<b>Trieda</b>	<b>Hodnoty</b>	<b>Početnosť</b>	<b>Súčet hodnôt</b>	<b>Priemer</b>
1	$Y_{11}, Y_{12}, \dots, Y_{1n_1}$	$n_1$	$Y_{1.}$	$y_{1.}$
.....	.....	.....	.....	.....
i	$Y_{i1}, Y_{i2}, \dots, Y_{in_i}$	$n_i$	$Y_{i.}$	$y_{i.}$
.....	.....	.....	.....	.....
I	$Y_{I1}, Y_{I2}, \dots, Y_{In_I}$	$n_I$	$Y_{I.}$	$y_{I.}$
Celkom		$n$	$Y_{..}$	$y_{..}$

Pred samotnou realizáciou testu treba spočítať nasledujúce štatistiky

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij} \quad Y_{..} = \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}, \quad (3.1.1)$$

ako aj priemery v jednotlivých triedach a celkový priemer

$$y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{i.}}{n_i}, \quad i = 1, 2, \dots, I \quad y_{..} = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij} = \frac{Y_{..}}{n}. \quad (3.1.2)$$

Predpoklady o veličinách  $Y_{ij}$  môžeme ekvivalentne vyjadriť takto:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (j = 1, 2, \dots, n_i; i = 1, 2, \dots, I),$$

kde  $\varepsilon_{ij}$  sú nezávislé náhodné veličiny s rozdelením  $N(0, \sigma^2)$ . Hypotézu  $H_0$  môžeme teraz vyjadriť v tvare:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_I = 0.$$

K testom tejto hypotézy treba spočítať:

$$\text{celkový súčet štvorcov: } S_T = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - y_{..})^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{..}^2}{n} \quad (3.1.3)$$

$$\text{riadkový súčet štvorcov: } S_A = \sum_{i=1}^I n_i (y_{i.} - y_{..})^2 = \sum_{i=1}^I \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n} \quad (3.1.4)$$

$$\text{reziduálny súčet štvorcov: } S_e = S_T - S_A \quad (3.1.5)$$

Výsledky môžeme opäť zapísať do tabuľky:

<b>Variabilita</b>	<b>Súčet štvorcov</b>	<b>Stupne voľnosti</b>	<b>Podiel</b>	<b>F</b>
riadková	$S_A$	I-1	$\frac{S_A}{I-1} = (1)$	$\frac{(1)}{(2)}$
reziduálna	$S_e$	n-I	$\frac{S_e}{n-I} = (2)$	
celková	$S_T$	n-1		

Ak vypočítaná štatistika  $F$  je väčšia ako príslušný kvantil Fisherovho-Snedecorovho rozdelenia  $F_{1-1, n-1}$  zamietame hypotézu o rovnosti stredných hodnôt.

### 3.2 Jednofaktorová ANOVA v systéme Mathematica

V softvéri Mathematica je už implementovaná metóda na počítanie jednofaktorovej ANOVA. Kvôli správne použitiu tejto metódy je však najskôr potrebné načítať balík

```
<<"ANOVA`".
```

Po načítaní balíka môžeme použiť príkaz na výpočet ANOVA a to v tvare:

**ANOVA** [vstupné dáta, ďalšie príkazy].

Vstupné dáta sú v tvare množiny dvojíc  $\{\{1, Y_{11}\}, \{1, Y_{12}\}, \dots, \{l, Y_{li}\}\}$ .

Ako ďalšie doplnujúce príkazy môžeme použiť nasledujúce:

príkaz	vysvetlenie
CellMeans → True	Vo výstupnej tabuľke budú zahrnuté priemery v triedach ako aj celkový priemer
CellMeans → False	Priemery tried a celkový priemer nebudú počítané
WorkingPrecision → 10	Výsledky budú s presnosťou na 10 desatinných miest
WorkingPrecision → MachinePrecision	Softvér sám určí vhodný počet desatinných miest vo výsledkoch
PostTests → {}	Zoznam testov, ktoré sa majú vykonať, keď ANOVA zistí výrazne rozdiely medzi triedami. Na výber máme: Bonferroni, Tukey, Duncan, Dunnett, StudentNewmanKeuls
SignificanceLevel → 0.05	Test bude počítaný na hladine významnosti 0.05

#### Príklad 1:

Boli sledované výnosy 4 odrôd zemiakov A, B, C, D. Každá bola pestovaná na 7 rovnako veľkých poliach. Testujme, či výnosy zemiakov sú závislé na odrode [2].

#### Vstup:

```
<<"ANOVA`"
```

```
data={{1,19.3},{1,18.0},{1,21.6},{1,22.4},{1,20.9},{1,20.1},{1,24.0},
{2,23.1},{2,26.5},{2,25.2},{2,25.0},{2,24.3},{2,21.4},{2,26.7},
{3,23.7},{3,20.8},{3,19.8},{3,24.1},{3,22.2},{3,22.6},
{3,22.9},{4,17.2},{4,16.6},{4,16.9},{4,17.7},{4,21.3},{4,15.2},
{4,19}};
```

```
ANOVA[data, CellMeans→True, WorkingPrecision→MachinePrecision,
SignificanceLevel→0.05, PostTests→{Bonferroni, Tukey, Duncan,
Dunnnett, StudentNewmanKeuls}]
```

### Výstup:

```

      DF      SumOfSq      MeanSq      FRatio      PValue
{ANOVA → Model    3      174.913      58.3042      17.0148      3.87199 × 10-6,
      Error    24      82.24      3.42667
      Total    27      257.153

      All      21.375
      Model[1] 20.9
CellMeans → Model[2] 24.6 , PostTests →
      Model[3] 22.3
      Model[4] 17.7

      Bonferroni      {{1, 2}, {1, 4}, {2, 4}, {3, 4}}
      Tukey          {{1, 2}, {1, 4}, {2, 4}, {3, 4}}
{Model → Duncan    {{1, 2}, {1, 4}, {2, 3}, {2, 4}, {3, 4}}}
      StudentNewmanKeuls {{1, 2}, {1, 4}, {2, 3}, {2, 4}, {3, 4}}
      Dunnnett      {2, 4}
```

Výstupom je kompletná tabuľka analýzy rozptylu. Nájde sa v nej stupne voľnosti (v stĺpci DF), súčty štvorcov (SumOfSq), príslušné podiely (MeanSq) a výsledok testovacej štatistiky (FRatio). Okrem toho v nej nájdeme aj celkový priemer a priemery v triedach (CellMeans). Keďže v príkaze boli vyžiadané aj dodatočné testy, tak tabuľka obsahuje aj ich výsledky. Tieto testy ukazujú, ktoré triedy sa od seba výrazne odlišujú.

Výsledok testovacej štatistiky pre daný príklad je 17,0148, ktorý musíme porovnať s kvantilom Fisherovho-Snedecorovho rozdelenia  $F_{3,24}(0.95) = 3,01$ . Na základe teórie analýzy rozptylu hypotézu  $H_0$  zamietame, t.j. odrody majú vplyv na hektárové výnosy.

Procedúra zabudovaná v softvéri Mathematica nie je optimálna. Jej základným predpokladom je rovnosť rozptylov jednotlivých tried, čo v praxi však nie je vždy možné. Preto sa častokrát v praxi využívajú aj iné vhodnejšie testy, ktorým sa venujú nasledujúce kapitoly práce.

## 4. Testy pre strednú hodnotu

Jedným z predpokladov, ktoré sme využili k odvodeniu vzorcov pre analýzu rozptylu jednoduchého triedenia, bola rovnosť rozptylov všetkých  $I$  výberov. Pri splnení tohto predpokladu, mohol byť použitý štandardný test analýzy rozptylu. Variančná homogenita však nie je vždy splnená a preto musia byť na výpočet použité testy zohľadňujúce nerovnosť rozptylov.

### 4.1. Cochranov test

Uvažujme opäť náhodný výber  $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ ,  $i = 1, \dots, I$  z rozdelenia  $N(\mu_i, \sigma_i^2)$ . Cochranov test patrí medzi testy, ktoré uvažujú rovnosť stredných hodnôt pri nerovnakých

rozptyloch. Testujeme hypotézu  $H_0 : \mu_1 = \mu_2 = \dots = \mu_I = \mu$  oproti alternatíve  $H_1 : \exists$  také  $i, j$ , že  $\mu_i \neq \mu_j$ .

Štatistika

$$Q = \sum_{i=1}^I \omega_i \left( \bar{Y}_i - \sum_{i=1}^I h_i \bar{Y}_i \right)^2, \quad (4.1.1)$$

$$\omega_i = n_i / s_i^2 \quad h_i = \omega_i / \sum_{j=1}^I \omega_j, \quad (4.1.2)$$

má  $\chi^2$  rozdelenie s  $(I - 1)$  stupňami voľnosti pre veľký rozsah meraní. Hypotézu  $H_0$  zamietame na hladine významnosti  $\alpha$ , keď  $Q > \chi_{I-1}^2(1 - \alpha)$  (kvantil  $\chi^2$  rozdelenia).

Program počítajúci Cochranov test, ktorý sa nachádza na priloženom CD vyžaduje na vstup:

CochranTest [data, alfa]

kde: - **data** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$

- **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare

Výstupom tohto programu je výsledok Cochranovej testovacej štatistiky, kvantil, s ktorým výsledok porovnávame a samotné rozhodnutie o prijatí, či neprijatí hypotézy  $H_0$ .

## 4.2. Welchov test

Pre ten istý prípad ako Cochran odvodil Welch testovaciu štatistiku:

$$W = \frac{\sum_{i=1}^I w_i \left( \bar{y}_i - \sum_{j=1}^I h_j \bar{y}_j \right)^2}{(I - 1) + 2(I - 2)(I + 1)^{-1} \sum_{i=1}^I (n_i - 1)^{-1} (1 - h_i)^2} \quad (4.2.1)$$

Welch aproximoval rozdelenie  $W$  pomocou  $F$ - rozdelenia s  $(I - 1)$  a  $v_w$  stupňami voľnosti, kde

$$v_w = \frac{I^2 - 1}{3 \sum_{i=1}^I (n_i - 1)^{-1} (1 - h_i)^2} \quad (4.2.2)$$

$$\omega_i = n_i / s_i^2 \quad h_i = \omega_i / \sum_{j=1}^I \omega_j \quad (4.2.3)$$

Hypotéza  $H_0$  sa zamietá na hladine významnosti  $\alpha$  ak  $W > F_{I-1, v_w}(1 - \alpha)$ .

Program, ktorý využíva Welchovu štatistiku, nachádzajúci sa v prílohe, potrebuje na vstup:

WelchTest [data, alfa]

kde: - **data** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{ dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$   
 - **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare

Výstupom je výsledok Welchovej testovacej štatistiky, kvantil, s ktorým výsledok porovnáваме a samotné rozhodnutie o prijatí, či neprijatí hypotézy  $H_0$ .

### 4.3. Modifikovaný Welchov test

Tento test je modifikovanou verziou štandardného Welchovho testu. Test je liberálnejší voči malým rozsahom výberov a vzrastajúcim počtom tried.

Testovacia štatistika je daná:

$$W^* = \frac{\sum_{i=1}^I w_i^* \left( \bar{y}_i - \sum_{j=1}^I h_j^* \bar{y}_j \right)^2}{(I-1) + 2(I-2)(I+1)^{-1} \sum_{i=1}^I (n_i-1)^{-1} (1-h_i^*)^2}, \quad (4.3.1)$$

$$h_i^* = w_i^* / \sum_{m=1}^I w_m^*, \quad w_i^* = n_i / (\varphi_i s_i^2), \quad \varphi_i = (n_i + \delta_1) / (n_i + \delta_2), \quad (4.3.2)$$

kde  $\delta_1$  a  $\delta_2$  sú reálne čísla vybrané tak, aby platilo  $1 \leq \varphi_i \leq c_i$ ,

$$c_i = (n_i - 1) / (n_i - 3) \quad (4.3.4)$$

pre  $i = 1, \dots, I$ . Testovacia štatistika je aproximovaná F- rozdelením s  $I - 1$  a  $\nu_w^*$  stupňami voľnosti

$$\nu_w^* = \frac{I^2 - 1}{3 \sum_{i=1}^I (n_i - 1)^{-1} (1 - h_i^*)^2}. \quad (4.3.5)$$

Hypotézu  $H_0$  zamietame na hladine významnosti  $\alpha$ , ak  $W^* > F_{I-1, \nu_w^*}(1-\alpha)$ . Keď sú výbery rozsahov väčšie, Modifikovaný Welchov test sa nahradí klasickým Welchovým testom. Správanie sa modifikovaného testu je možné kontrolovať pomocou vhodného výberu parametra  $\varphi_i$ .

Program, ktorý využíva Modifikovaný Welchov test, nachádzajúci sa v prílohe, potrebuje na vstup:

ModifiedWelchTest [data, alfa, volba]

kde: - **data** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{ dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$   
 - **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare  
 - **volba** určuje čomu sa budú rovnať čísla  $\delta_1$ ,  $\delta_2$ . Užívateľ má na výber z troch možností, ak za voľbu zadá číslo iné ako 1, 2 alebo 3, program ho upozorní a za voľbu vezme číslo 1 kedy sa  $\delta_1$ ,  $\delta_2$  rovnajú nule.

Výstupom je výsledok testovacej štatistiky, kvantil, s ktorým ho porovnáваме a samotné rozhodnutie o prijatí, či neprijatí hypotézy  $H_0$  a aj to čomu sa rovnajú  $\delta_1$ ,  $\delta_2$  podľa voľby užívateľa.

## 4.4. Brown – Forsytheov test

Brown – Forsytheov test je modifikáciou ANOVA F – testu. Test je daný testovacou štatistikou:

$$B = \frac{\sum_{i=1}^I n_i (\bar{y}_i - \bar{y}_{..})^2}{\sum_{i=1}^I (1 - n_i/N) s_i^2} . \quad (4.4.1)$$

Keď hypotéza  $H_0$  platí, testovacia štatistika B je aproximovaná F – rozdelením s  $(I - 1)$  a v stupňami voľnosti

$$v = \frac{\left[ \sum_{i=1}^I (1 - n_i/N) s_i^2 \right]^2}{\sum_{i=1}^I (1 - n_i/N)^2 s_i^4 / (n_i - 1)} . \quad (4.4.2)$$

$H_0$  zamietame, ak platí  $B > F_{I-1, v; (1 - \alpha)}$ .

Program, ktorý využíva Brown-Forsytheov test, nachádzajúci sa v prílohe, potrebuje na vstup:

BrownForsytheTest [data, alfa]

kde: - **data** sú užívateľom zadané v tvare {{dáta triedy 1},{ dáta triedy 2},..., {dáta triedy I}}  
 - **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare

Výstupom je výsledok testovacej štatistiky, kvantil, s ktorým výsledok porovnáваме a samotné rozhodnutie o prijatí, či neprijatí hypotézy  $H_0$ .

## 5. Simulačné štúdie

V tejto simulačnej štúdií chceme porovnať sily už známych testov pre testovanie hypotézy o homogenite stredných hodnôt. K simuláciám boli použité testy uvedené v kapitole 4.1 tejto práce a test pre jednofaktorovú ANOVA.

Simulácie boli realizované 1000 krát pre dáta s malým počtom opakovaní v triedach a parametre, ako stredná hodnota a rozptyl sa vyberali týmito spôsobmi:

- stredná hodnota rovnaká, disperzia rôzna
- stredná hodnota aj disperzia rôzne

- stredná hodnota rôzna, disperzia rovnaká
- stredná hodnota aj disperzia rovnaké

Všetky testy boli počítané na hladine významnosti  $\alpha = 0.05$ . Výsledky simulácií sú spracované v nasledujúcich tabuľkách.

**Poznámka:** Kvôli lepšej prehľadnosti tabuliek boli použité iba stručné názvy testov. Najmä pre Modifikovaný Welchov test, ktorého výsledky závisia od zvolenia  $\delta_1$  a  $\delta_2$ . Pri teste označenom ako Modified Welch 1, boli zvolené  $\delta_1 = 0$  a  $\delta_2 = 0$ , Modified Welch 2  $\delta_1 = -1$  a  $\delta_2 = -3$  a pri teste Modified Welch 3 boli  $\delta_1 = 2$  a  $\delta_2 = 1$ .

Pod percentuálnou úspešnosťou sa rozumie správne vyhodnotenie testu t.j. ak  $H_0$  platí nezamietam, ak  $H_0$  neplatí, zamietam.

	Cochran	Welch	Modified Welch 1	Modified Welch 2	Modified Welch 3	Brown-Forsythe	Anova
Počet zamietnutí	141	55	55	20	47	57	72
Počet nezamietnutí	859	945	945	980	953	943	928
Percentuálna úspešnosť	85.9	94.5	94.5	98.	95.3	94.3	92.8

Tabuľka (5.1) : Výsledky testov pri rovnakých stredných hodnotách použitých dát a rôznych disperziách.

Simulácie v prvom prípade ukázali, že Cochranov test a štandardný ANOVA test dávajú horšie výsledky. Čo môže byť spôsobené tým, že Cochranov test je určený na testovanie dát s veľkým počtom opakovaní. Brown-Forsytheov test, Welchov a Modifikovaný Welchov test 1 dávajú približne rovnaké výsledky. Modifikovaný Welchov test 3 ešte dáva porovnateľné výsledky so spomínanými tromi testami. Modifikovaný Welchov test 2 je pri malom počte opakovaní v triedach najliberálnejší.

	Cochran	Welch	Modified Welch 1	Modified Welch 2	Modified Welch 3	Brown-Forsythe	Anova
Počet zamietnutí	934	887	887	784	877	873	889
Počet nezamietnutí	66	113	113	216	123	127	101
Percentuálna úspešnosť	93.4	88.7	88.7	78.4	87.7	87.3	88.9

Tabuľka (5.2): Výsledky testov pri rôznych stredných hodnotách aj disperziách.

Podľa tabuľky (5.2) pri rôznych stredných hodnotách i disperziách dát je úspešnosť testov presne opačná ako v prvom prípade. Najlepšie výsledky tentokrát dáva Cochranov test, ktorý najviac zohľadňuje to, že dáta majú rôznu strednú hodnotu a tak hypotézu  $H_0$  o rovnosti stredných hodnôt zamietam na 93,4 %. Najhoršie výsledky dáva Modifikovaný Welchov test 2. Ostatné testy sú približne rovnako úspešné. Ich úspešnosť sa pohybuje v rozmedzí 88,9 % - 87,3 %.

	Cochran	Welch	Modified Welch 1	Modified Welch 2	Modified Welch 3	Brown-Forsythe	Anova
Počet zamietnutí	837	768	768	598	738	775	787
Počet nezamietnutí	163	232	262	402	232	225	213
Percentuálna úspešnosť	83.7	76.8	76.8	59.8	73.8	77.5	78.7

Tabuľka (5.3): Výsledky testov pri rôznych stredných hodnotách použitých dát a rovnakých disperziách.

Výsledky simulácií v tomto prípade (tabuľka (5.3)) sú obdobné ako v predchádzajúcom. Cochranov test dáva opäť najlepšie výsledky, i keď v porovnaní s predchádzajúcim prípadom je percentuálne menej úspešný. Najnižšiu percentuálnu úspešnosť má opäť Modifikovaný Welchov test 2, ktorý hypotézu  $H_0$  o rovnosti stredných hodnôt zamietam len na 59,8 % napriek



tomu, že samotné dáta sú nastavené na rôzne stredné hodnoty. Výsledky zvyšných testov sú porovnateľné.

	Cochran	Welch	Modified Welch 1	Modified Welch 2	Modified Welch 3	Brown-Forsythe	Anova
Počet zamietnutí	106	48	48	11	38	39	43
Počet nezamietnutí	894	952	952	989	962	961	957
Percentuálna úspešnosť	89.4	95.2	95.2	98.9	96.2	96.1	95.7

Tabuľka (5.4): Výsledky testov pri rovnakých stredných hodnotách aj disperziách.

V poslednom prípade, ktorý bol simulovaný (tabuľka (5.4)), sa takisto potvrdili vlastnosti testov. Cochranov test, ktorý je určený pre veľký rozsah dát, dáva najhoršie výsledky. Opäť najlepšie dáva Modifikovaný Welchov test 2, ktorý je určený práve pre malý počet opakovaní v triedach.

## 6. Dvojfaktorová analýza rozptylu bez interakcií

Niekedy sa celkový výber rozpadá na skupiny, pretože sa musí prihliadať na dva rôzne triediace znaky. Napríklad sa môžu zisťovať výnosy určitej plodiny v závislosti na tom, na akom spomedzi troch typov pôdy bola pestovaná a zároveň, aký zo štyroch druhov hnojiva bol použitý. Ak je možné, že sa vplyv týchto znakov prejavuje súčtom ich efektov na strednú hodnotu, dospejeme k modelu

$$Y_{ijp} = \mu + \alpha_i + \beta_j + e_{ijp}, \quad (6.1)$$

kde  $\mu$ ,  $\alpha_i$ ,  $\beta_j$  pre  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ ,  $p = 1, \dots, P$  sú neznáme parametre a  $e_{ijp}$  sú nezávislé veličiny s rozdelením  $N(0, \sigma^2)$ . Parameter  $\sigma^2 > 0$  je takisto neznámy. Označme  $n = IJP$  celkový počet veličín  $Y_{ijp}$  a položeme

$$Y_{ij.} = \sum_p Y_{ijp}, \quad Y_{i..} = \sum_j \sum_p Y_{ijp}, \quad Y_{...} = \sum_i \sum_j \sum_p Y_{ijp}, \quad (6.2)$$

$$y_{ij.} = \frac{Y_{ij.}}{P}, \quad y_{i..} = \frac{Y_{i..}}{JP}, \quad y_{...} = \frac{Y_{...}}{n}. \quad (6.3)$$

Podrobnejším rozpísaním modelu (6.1) sa zistí, že obsahuje nadbytočné parametre v tom zmysle, že mu zodpovedá matica  $\mathbf{X}$  s neúplnou stĺpcovou hodnotou. Preto sa zavádzajú reparametrizačné rovnice

$$\sum \alpha_i = 0, \quad \sum \beta_j = 0. \quad (6.4)$$

Ďalej zavedme

$$S_A = \frac{1}{JP} \sum_i Y_{i..}^2 - \frac{1}{n} Y_{...}^2, \quad f_A = I - 1, \quad (6.5)$$

$$S_B = \frac{1}{IP} \sum_j Y_{.j.}^2 - \frac{1}{n} Y_{...}^2, \quad f_B = J - 1, \quad (6.6)$$

$$S_T = \sum_i \sum_j \sum_p Y_{ijp}^2 - \frac{1}{n} Y_{\dots}^2, \quad f_T = n - 1, \quad (6.7)$$

$$S_e = S_T - S_A - S_B, \quad f_e = n - I - J + 1. \quad (6.8)$$

Keby nezáležalo na riadkovom triediacom znaku, mali by sme  $\alpha_1 = \dots = \alpha_I = 0$ . Dospeli by sme k podmodelu

$$Y_{ijp} = \mu + \beta_j + e_{ijp}, \quad (6.9)$$

ktorý zodpovedá jednoduchému triedeniu.

Reziduálny súčet štvorcov za platnosti modelu (6.1) je rovný  $S_e$  a rozdiel reziduálnych súčtov štvorcov je rovný  $S_A$ . Hypotézu o možnosti redukcie modelu (6.1) na podmodel (6.9) zamietneme, keď

$$F_A = \frac{S_A/f_A}{S_e/f_e} \geq F_{f_A, f_e}(1 - \alpha). \quad (6.10)$$

Podobne sa postupuje, keď chceme overiť, či záleží na stĺpcovom triediacom znaku, t.j. či je možné model (6.1) redukovať na podmodel

$$Y_{ijp} = \mu + \alpha_i + e_{ijp}. \quad (6.11)$$

Pritom

$$F_B = \frac{S_B/f_B}{S_e/f_e} \geq F_{f_B, f_e}(1 - \alpha). \quad (6.12)$$

Výsledky sa môžu prehľadne zapísať do tabuľky analýzy rozptylu dvojného triedenia:

Variabilita	Súčet štvorcov S	Počet stupňov voľnosti f	Podiel S/f	F
riadková	$S_A$	$f_A$	$\frac{S_A}{f_A}$	$F_A$
stĺpcová	$S_B$	$f_B$	$\frac{S_B}{f_B}$	$F_B$
reziduálna	$S_e$	$f_e$	$\frac{S_e}{f_e}$	-
celková	$S_T$	$f_T$	-	-

Položme  $s^2 = S_e/f_e$ . Ak hodnota  $F_A$  je štatisticky významná, bude spravidla nutné zistiť, ktoré dvojice riadkov sa od seba významne odlišujú. Podľa Tukeyho metódy rovnosť  $\alpha_i = \alpha_t$  zamietame, ak

$$|y_{i..} - y_{t..}| \geq \sqrt{\frac{s^2}{JP}} q_{I, n-I-J+1}(\alpha), \quad (6.13)$$

kde  $q_{I, n-I-J+1}(\alpha)$  sú tabelované kritické hodnoty studentizovaného rozdelenia [1].

Podobne pri významnej hodnote  $F_B$  sa zamietajú rovnosť  $\beta_j = \beta_t$ , ak

$$|y_{.j} - y_{.t}| \geq \sqrt{\frac{s^2}{IP}} q_{J, n-I-J+1}(\alpha). \quad (6.14)$$

## 6.1. Dvojfaktorová ANOVA bez interakcií v softvéri Mathematica

V systéme Mathematica sa dá počítať dvojfaktorová analýza rozptylu bez interakcií pomocou už implementovanej metódy. Tak ako v predchádzajúcich prípadoch aj teraz je potrebné načítanie balíka

```
<<"ANOVA`".
```

Po načítaní môžeme použiť podobný príkaz ako pre jednofaktorovú ANOVA v tvare:

**ANOVA**[vstupné dáta, {faktor 1, faktor 2}, {faktor 1, faktor 2}, ďalšie príkazy].

Ďalšie príkazy, ktoré možno použiť sú obdobné ako pre jednofaktorovú ANOVA:

príkaz	vysvetlenie
CellMeans → True	Vo výstupnej tabuľke budú zahrnuté priemery v triedach ako aj celkový priemer
CellMeans → False	Priemery tried a celkový priemer nebudú počítané
WorkingPrecision → 10	Výsledky budú s presnosťou na 10 desatinných miest
WorkingPrecision → MachinePrecision	Softvér sám určí vhodný počet desatinných miest vo výsledkoch
PostTests → {}	Zoznam testov, ktoré sa majú vykonať, keď ANOVA zistí výrazne rozdiely medzi triedami. Na výber máme: Bonferroni, Tukey, Duncan, Dunnett, StudentNewmanKeuls
SignificanceLevel → 0.05	Test bude počítaný na hladine významnosti 0.05

### Príklad 2:

Bol sledovaný predaj určitého tovaru počas troch rovnako dlhých časových období. Pritom bol sledovaný jednak vplyv balenia (balenie vo vrecku, balenie v krabíčke) a jednak druh reklamy (bez reklamy, reklama v novinách, reklama v tv aj v novinách). Je treba zistiť vplyv reklamy i vplyv balenia tovaru na jeho predaj. Hodnoty predaného tovaru sú v miliónoch Eur [1].

### Vstup:

```
<<"ANOVA`"
```

```
data={{1,1,1},{1,2,1},{1,3,6},{2,1,3},{2,2,4},{2,3,9}};
```

```
ANOVA[data, {balenie, reklama}, {balenie, reklama}, CellMeans→True,
WorkingPrecision→MachinePrecision, SignificanceLevel→0.05, PostT
ests→{Bonferroni, Tukey, Duncan, Dunnett, StudentNewmanKeuls}]
```

## Výstup:

```

      balenie      DF      SumOfSq      MeanSq      FRatio      PValue
{ANOVA → reklama  1      10.6667      10.6667      64.         0.0152681
      Error      2       0.333333      0.166667      111.        0.00892857,
      Total      5        48.
}

      All      4.
CellMeans → balenie[1] 2.66667
      balenie[2] 5.33333
      reklama[1] 2.
      reklama[2] 2.5
      reklama[3] 7.5

      Bonferroni      {1, 2}
      Tukey           {1, 2}
PostTests → {balenie → Duncan      {1, 2},
      StudentNewmanKeuls {1, 2}
      Dunnett           {2}

      Bonferroni      {{1, 3}, {2, 3}}
      Tukey           {{1, 3}, {2, 3}}
reklama → Duncan      {{1, 3}, {2, 3}}}}
      StudentNewmanKeuls {{1, 3}, {2, 3}}
      Dunnett         {3}

```

Výstupom je tabuľka analýzy rozptylu. Nájde v nej stupne voľnosti (v stĺpci DF), súčty štvorcov (SumOfSq), príslušné podiely (MeanSq) a výsledok testovacej štatistiky (FRatio). Okrem toho v nej nájdeme aj celkový priemer a priemery v triedach (CellMeans). Keďže v príkaze boli vyžiadané aj dodatočné testy, tak tabuľka obsahuje aj ich výsledky. Tieto testy ukazujú, ktoré triedy jednotlivých faktorov sa od seba výrazne odlišujú.

Pretože  $64 > F_{1,2}(0.95) = 18,5$ , zamietneme hypotézu, že sa pri predaji neuplatňuje spôsob balenia tovaru. Ďalej máme  $111 > F_{2,2}(0.95) = 19$ , takže zamietame aj hypotézu, že druh reklamy nemá vplyv na predaj.

## 7. Dvojfaktorová ANOVA s interakciami

Pri dvojfaktorovej analýze rozptylu sa často stáva, že efekty riadkov a stĺpcov nemožno sčítať tak ako tomu bolo v kapitole 6., pretože pri určitých kombináciách riadkov a stĺpcov môžu nastať špecifické účinky. Preto je v týchto situáciách realistickejší model

$$Y_{ijp} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijp}, \quad (7.1)$$

kde  $\mu$ ,  $\alpha_i$ ,  $\beta_j$  a  $\lambda_{ij}$  sú neznáme parametre a  $e_{ijp}$  sú nezávislé náhodné veličiny s rozdelením  $N(0, \sigma^2)$ . V práci sa obmedzíme na rovnaký počet pozorovaní v podtriedach, takže budeme predpokladať  $i = 1, \dots, I$ ,  $j = 1, \dots, J$  a  $p = 1, \dots, P$ . Položíme  $n = IJP$ .

Parametrom  $\lambda_{ij}$  sa hovorí interakcie. Na rozdiel od dvojfaktorovej ANOVA bez interakcií, tu budeme uvažovať  $P \geq 2$ .

Najprv sa zavedú reparametrizačné rovnice

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \lambda_{ij} = 0 \text{ pre každé } j, \quad \sum_j \lambda_{ij} = 0 \text{ pre každé } i.$$

Spravidla nás zaujímajú tri rôzne nulové hypotézy:

$$\begin{aligned} H_A &: \alpha_1 = \dots = \alpha_I = 0 \\ H_B &: \beta_1 = \dots = \beta_J = 0 \\ H_{AB} &: \lambda_{ij} = 0 \text{ pre všetky dvojice } (i, j). \end{aligned}$$

Na výpočet štatistiky je potrebné vypočítať

$$S_A = \frac{1}{JP} \sum_i Y_{i..}^2 - \frac{1}{n} Y_{...}^2, \quad f_A = I - 1 \quad (7.2)$$

$$S_B = \frac{1}{IP} \sum_j Y_{.j.}^2 - \frac{1}{n} Y_{...}^2, \quad f_B = J - 1 \quad (7.3)$$

$$S_e = \sum_i \sum_j \sum_p Y_{ijp}^2 - \frac{1}{P} \sum_i \sum_j Y_{ij.}^2, \quad f_e = n - IJ \quad (7.4)$$

$$S_T = \sum_i \sum_j \sum_p Y_{ijp}^2 - \frac{1}{n} Y_{...}^2, \quad f_T = n - 1 \quad (7.5)$$

$$S_{AB} = S_T - S_A - S_B - S_e, \quad f_{AB} = (I - 1)(J - 1). \quad (7.6)$$

Zostaví sa tabuľka analýzy rozptylu:

**Tabuľka 7.1**

Variabilita	Súčet štvorcov	Stupne voľnosti	Podiel	F
riadková	$S_A$	$f_A$	$S_A/f_A$	$F_A$
stĺpcová	$S_B$	$f_B$	$S_B/f_B$	$F_B$
interakcií	$S_{AB}$	$f_{AB}$	$S_{AB}/f_{AB}$	$F_{AB}$
reziduálna	$S_e$	$f_e$	$S_e/f_e$	-
celková	$S_T$	$f_T$	-	-

Položme

$$s^2 = \frac{S_e}{f_e}, \quad (7.7)$$

potom máme

$$F_A = \frac{S_A/f_A}{s^2}, \quad F_B = \frac{S_B/f_B}{s^2}, \quad F_{AB} = \frac{S_{AB}/f_{AB}}{s^2}. \quad (7.8)$$

Ak  $F_A \geq F_{f_A, f_e}(1-\alpha)$ , zamietame hypotézu  $H_A$ . V prípade  $F_B \geq F_{f_B, f_e}(1-\alpha)$ , zamietneme hypotézu  $H_B$  a keď  $F_{AB} \geq F_{f_{AB}, f_e}(1-\alpha)$  zamietneme  $H_{AB}$ .

Ak hodnota  $F_A$  je štatisticky významná, potom rovnosť  $\alpha_i = \alpha_t$  zamietame v prípade, že

$$|y_{i..} - y_{t..}| \geq \sqrt{\frac{s^2}{JP}} q_{I, f_e}(\alpha). \quad (7.9)$$

Keď hodnota  $F_B$  je štatisticky významná, potom rovnosť  $\beta_j = \beta_t$  zamietame v prípade, že

$$|y_{.j.} - y_{.t.}| \geq \sqrt{\frac{s^2}{IP}} q_{J, f_e}(\alpha). \quad (7.10)$$

## 7.1. Dvojfaktorová ANOVA s interakciami v Mathematice

V systéme Mathematica sa dá počítať aj dvojfaktorová analýza rozptylu s interakciami pomocou už implementovanej metódy. Tak ako v predchádzajúcich prípadoch aj teraz je potrebné načítanie balíka

```
<<"ANOVA`".
```

Po načítaní môžeme použiť podobný príkaz ako pre dvojfaktorovú ANOVA v kapitole 6.1. v tvare:

```
ANOVA[vstupné dáta, {faktor 1, faktor 2, All}, {faktor 1, faktor 2}, ďalšie príkazy].
```

Ďalšie príkazy, ktoré možno použiť sú obdobné ako pre iné druhy ANOVA, preto v tejto kapitole, už nebudú spomínané bližšie. Uvedieme len názorný príklad.

### Príklad 3:

Boli skúmané výnosy sena v t/ha v závislosti na type pôdy (normálna, kyslá) a v závislosti na type hnojenia (bez hnojenia, mrva, vápenaté hnojivo). Každá kombinácia bola realizovaná dvakrát [2].

### Vstup:

```
<<"ANOVA`"
data={{1,1,2.9},{1,1,3.1},{1,2,3.6},{1,2,3.9},{1,3,3.2},{1,3,3.5},
{2,1,2.7},{2,1,3.0},{2,2,3.5},{2,2,3.4},{2,3,3.8},{2,3,4.2}}

faktor1=TypPody;
faktor2=TypHnojenia;
```

```
ANOVA[data, {faktor1, faktor2, All}, {faktor1, faktor2}, SignificanceLevel -> 0.05,
PostTests -> {Bonferroni, Duncan, Dunnett, Tukey, StudentNewmanKeuls}
]
```

## Výstup:

```

ANOVA ->
      TypPody          DF      SumOfSq      MeanSq      FRatio      PValue
      TypHnojenia      2      1.365      0.6825      17.0625     0.00334356
      TypHnojenia TypPody  2      0.521667    0.260833    6.52083    0.0312852
      Error            6      0.24      0.04
      Total            11     2.14

      All              3.4
      TypPody[1]       3.36667
      TypPody[2]       3.43333
      TypHnojenia[1]   2.925
      TypHnojenia[2]   3.6
      TypHnojenia[3]   3.675
CellMeans -> TypHnojenia[1] TypPody[1] 3.      PostTests ->
TypHnojenia[1] TypPody[2] 2.85
TypHnojenia[2] TypPody[1] 3.75
TypHnojenia[2] TypPody[2] 3.45
TypHnojenia[3] TypPody[1] 3.35
TypHnojenia[3] TypPody[2] 4.

      Bonferroni      {}      Bonferroni      {{(1, 2), (1, 3)}}
      Tukey           {}      Tukey           {{(1, 2), (1, 3)}}
      {TypPody -> Duncan      {}, TypHnojenia -> Duncan      {{(1, 2), (1, 3)}}}
      StudentNewmanKeuls {}      StudentNewmanKeuls {{(1, 2), (1, 3)}}
      Dunnett        {}      Dunnett        {2, 3}

```

Výstupom je opäť prehľadná tabuľka analýzy rozptylu, v ktorej nájdeme súčty štvorcov, počet stupňov voľnosti i výsledky testovacích štatistík.

Keďže  $F_A = 0,3333 < F_{1,6}(0,95) = 5,99$ , závislosť výnosu sena na type pôdy nie je dokázaná. Je však dokázaný vplyv hnojenia, pretože  $F_B = 17,0625 \geq F_{2,6}(0,95) = 5,14$ . Takisto je dokázaná prítomnosť interakcií, keďže  $F_{AB} = 6,52083 < F_{2,6}(0,95) = 5,14$ .

Procedúra implementovaná v softvéri Mathematica pre dvojfaktorovú ANOVA s interakciami nepočíta s možnosťou, že pre každú kombináciu je vykonané len jediné meranie. Túto možnosť je však možné testovať pomocou iných testov napr. Tukeyho test, ktorý je popísaný v ďalšej kapitole.

## 7.2. Tukeyho test aditivity

Veľmi často sa stáva, že pri dvojfaktorovej analýze rozptylu je len jedno pozorovanie v triede ( $P = 1$ ). Napriek tomu, je však možné počítať s prítomnosťou interakcií. Túto situáciu popisuje model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijp}, \quad (7.2.1)$$

kde  $e_{ij}$  sú nezávislé veličiny s rozdelením  $N(0, \sigma^2)$  a  $i = 1, \dots, I, j = 1, \dots, J$ . Tak ako pri klasickej dvojfaktorovej analýze rozptylu s interakciami aj pri tomto teste sa zavedú reparametrizačné rovnice

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0. \quad (7.2.2)$$

Pretože nie je možné uvažovať všeobecný tvar interakcií  $\lambda_{ij}$ , Tukey navrhol model obsahujúci len jediný parameter  $\lambda$ .

$$\lambda_{ij} = \lambda \alpha_i \beta_j. \quad (7.2.3)$$

Možno testovať hypotézy:

$$\begin{aligned} H_A: \alpha_1 = \dots = \alpha_I = 0 \\ H_B: \beta_1 = \dots = \beta_J = 0 \\ H_{AB}: \lambda = 0. \end{aligned}$$

Položme  $n = IJ$  a vypočíta sa

$$S_A = J \sum_i y_{i.}^2 - n y_{..}^2, \quad f_A = I - 1 \quad (7.2.4)$$

$$S_B = I \sum_j y_{.j}^2 - n y_{..}^2, \quad f_B = J - 1 \quad (7.2.5)$$

$$S_{AB} = \frac{IJ}{S_A S_B} \left[ \sum_i \sum_j Y_{ij} (y_{i.} - y_{..})(y_{.j} - y_{..}) \right]^2, \quad f_{AB} = 1 \quad (7.2.6)$$

$$S_T = \sum_i \sum_j Y_{ij}^2 - n y_{..}^2, \quad f_T = n - 1 \quad (7.2.7)$$

$$S_e = S_T - S_A - S_B - S_{AB}, \quad f_e = (I - 1)(J - 1) - 1. \quad (7.2.8)$$

Tabuľka analýzy rozptylu má rovnaký tvar ako tabuľka (7.1). Mnohonásobné porovnávanie sa počítajú pomocou vzorcov (7.9), (7.10) [2].

Program počítajúci Tukeyho test potrebuje na vstup:

TukeyTest[data,alfa]

kde: - **data** sú užívateľom zadané v tvare {{dáta triedy 1},{ dáta triedy 2},..., {dáta triedy I}}

- **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare

Výstupom je prehľadná tabuľka obdobná tabuľke štandardnej analýzy rozptylu a rozhodnutia o prítomnosti interakcií, o vplyve jednotlivých faktorov.

## 8. Analýza kovariancie

Analýza kovariancie (ANCOVA) spája prvky analýzy rozptylu a viacnásobnej regresie. Do modelu ANOVA, ktorý skúma vzťah medzi intervalovou a nominálnymi



faktormi, sa pridá jeden alebo viac intervalových faktorov. Model jednofaktorovej analýzy rozptylu možno zapísať:

$$Y_{ij} = \mu_i + e_{ij} \quad (8.1)$$

kde  $e_{ij}$  je chyba merania aproximovaná normálnym rozdelením  $N(0, \sigma^2)$ . Ak intervalová premenná  $X$  s hodnotami  $X_{ij}$  má lineárny vzťah k premennej  $Y$  a regresný koeficient  $\beta$  medzi  $Y$  a  $X$  je rovnaký vo všetkých skupinách, t.j. regresné priamky majú rovnaký sklon, potom model ANCOVA má tvar:

$$Y_{ij} = \mu_i + \beta(X_{ij} - X) + u_{ij} \quad (8.2)$$

kde  $X$  je priemer premennej  $X$  a  $u_{ij} \sim N(0, \sigma^2)$ . Ak je vzťah medzi  $X$  a  $Y$  významný, potom  $u_{ij} < e_{ij}$ , teda model ANCOVA vysvetľuje viac variability premennej  $Y$  ako model ANOVA. Ak vzťah medzi  $X$  a  $Y$  nie je významný, ANCOVA nemožno použiť. Analýza kovariancie testuje či upravené priemery skupín sú rozdielne. Priemery sú upravené tak, ako keby vo všetkých skupinách bola rovnaká (priemerná) hodnota intervalového faktora.

ANCOVA si kladie rovnaké požiadavky ako ANOVA a viacnásobná regresia. Tradičná ANCOVA okrem toho predpokladá, že regresný koeficient  $\beta$  je rovnaký vo všetkých skupinách, teda neexistuje interakcia intervalového a nominálneho faktora [8].

Rovnice na výpočet analýzy kovariancie:

### Výpočet súčtu štvorcov pre nezávislú premennú $X$ a závislú $Y$

celkový súčet štvorcov pre premennú  $Y$ : 
$$S_{T_Y} = \sum_{i=1}^n \sum_{j=1}^k Y_{ij}^2 - \left( \frac{\sum_{i=1}^n \sum_{j=1}^k Y_{ij}}{n_T} \right)^2 \quad (8.3)$$

reziduálny súčet štvorcov pre  $Y$ : 
$$S_{e_Y} = \sum_{i=1}^n \sum_{j=1}^k Y_{ij}^2 - \sum_{i=1}^n \left( \frac{\sum_{j=1}^k Y_{ij}^2}{n_k} \right) \quad (8.4)$$

riadkový súčet štvorcov pre  $Y$ : 
$$S_{A_Y} = \sum_{i=1}^n \left( \frac{\sum_{j=1}^k Y_{ij}^2}{n_k} \right) - \frac{\left( \sum_{i=1}^n \sum_{j=1}^k Y_{ij} \right)^2}{n_T} \quad (8.5)$$

celkový súčet štvorcov je rovný aj reziduálnemu a riadkovému súčtu štvorcov:

$$S_{T_Y} = S_{e_Y} + S_{A_Y} \quad (8.6)$$

Súčty štvorcov pre nezávislú premennú  $X$  sú totožné s tými pre  $Y$ . Mení sa len ich označenie  $S_{T_X}$ ,  $S_{e_X}$ ,  $S_{A_X}$ .

### Výpočet kovariancie $X$ a $Y$

celkový súčet štvorcov premenných určuje kovariancia  $X$  a  $Y$  vo všetkých dátových vzorkách:

$$SCT = \sum_{i=1}^n \sum_{j=1}^k X_{ij} Y_{ij} - \frac{\left( \sum_{i=1}^n \sum_{j=1}^k X_{ij} \right) \left( \sum_{i=1}^n \sum_{j=1}^k Y_{ij} \right)}{n_T} \quad (8.7)$$

$$SCE = \sum_{j=1}^k \left( \sum_{i=1}^n X_{ij} Y_{ij} - \frac{\sum_{i=1}^n (X_{ij} Y_{ij})}{n_n} \right) \quad (8.8)$$

**Prispôsobený celkový súčet štvorcov**  $S_{T_Y}$

korelácia medzi X a Y: 
$$r_T^2 = \frac{SCT^2}{S_{T_x} S_{T_y}} \quad (8.9)$$

$$r_n^2 = \frac{SCE^2}{S_{e_x} S_{e_y}} \quad (8.10)$$

Pomer kovariancie je odčítaný od závislých hodnôt jednotlivých súčtov štvorcov:

$$S_{T_{Yadj}} = S_{T_Y} (1 - r_T^2) \quad (8.11)$$

$$S_{e_{Yadj}} = S_{e_y} (1 - r_n^2) \quad (8.12)$$

$$S_{A_{Yadj}} = S_{T_{Yadj}} - S_{e_{Yadj}} \quad (8.13)$$

**Prispôsobený priemer každej triedy**

Priemer každej triedy sa prispôsobuje nasledovne: 
$$\bar{Y}_{i.adj} = \bar{Y}_i - \frac{SCE_y}{SCE_x} (\bar{X}_i - \bar{X}_{..}) \quad (8.14)$$

**Analýza využíva prispôsobené hodnoty súčtov štvorcov**

stupne voľnosti: 
$$df_{Tr} = N_T - k - 1, \quad df_E = k - 1 \quad (8.15)$$

priemery štvorcov: 
$$MS_A = \frac{S_A}{df_{Tr}} \quad (8.16)$$

$$MS_e = \frac{S_e}{df_E} \quad (8.17)$$

**Testovacia štatistika**

$$F_{df_E, df_{Tr}} = \frac{MS_A}{MS_e} \quad (8.18)$$

Výsledok testovacej štatistiky porovnáваме s F – rozdelením so stupňami voľnosti  $df_{Tr}$ ,  $df_E$ . Ak je testovacia štatistika väčšia ako príslušný kvantil tohto rozdelenia zamietame hypotézu o rovnosti.

Samotná ANCOVA sa dá rozdeliť na tri časti:

1. Vizualne overenie splnenia podmienky rovnosti smerníc regresných priamok.
2. Test homogenity regresie – keďže analýza kovariancie predpokladá, že smernice (sklony) regresných priamok jednotlivých skupín nie sú výrazne odlišné od sklonu

celkovej regresie v skupinách. Ak sú výrazne odlišné, potom ANCOVA nie je schopná dávať pravdivé výsledky. Test homogenity preto testuje, či je tento základný predpoklad splnený. Výsledok testu je porovnávaný s kvantilom F- rozdelenia s  $(k - 1)$  a  $(N_T - 2k)$ , kde  $k$  je počet tried a  $N_T$  celkový počet meraní. Hypotézu o rovnosti smerníc regresných priamok zamietame ak výsledok testu je väčší ako príslušný kvantil.

3. Samotný test pre analýzu kovariancie, ktorého potrebné rovnice sú uvedené vyššie v tejto kapitole.

Program, ktorý spracúva analýzu kovariancie a test homogenity sa nachádza v prílohe a na vstup vyžaduje:

Ancova [dataX, dataY, alfa]

- kde:
- **dataX** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$ . Predstavujú hodnoty nezávislej premennej X.
  - **dataY** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$ . Predstavujú hodnoty závislej premennej Y.
  - **alfa** je hladina významnosti, na ktorej chce užívateľ test počítať, v desatinnom tvare

Výstupom je výsledok testu homogenity spolu s príslušným kvantilom a rozhodnutie o prijatí, alebo zamietnutí hypotézy  $H_0$ . Keďže program počíta i samotnú ANCOVA, tak jeho výstupom je i prehľadná tabuľka analýzy kovariancie, kvantil, s ktorým porovnáваме testovaciu štatistiku a výsledné rozhodnutie o prijatí resp. zamietnutí  $H_0$ .

Program, ktorý slúži na vykreslenie regresných priamok vyžaduje zadať na vstup:

Grafy [dataX, dataY]

- kde:
- **dataX** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$ . Predstavujú hodnoty nezávislej premennej X.
  - **dataY** sú užívateľom zadané v tvare  $\{\{\text{dáta triedy 1}\}, \{\text{dáta triedy 2}\}, \dots, \{\text{dáta triedy I}\}\}$ . Predstavujú hodnoty závislej premennej Y.

Výstupom je popísaný graf všetkých dát, graf dát jednotlivých tried spolu s regresnými priamkami, graf všetkých dát aj všetkých regresných priamok a napokon graf vykresľujúci všetky regresné priamky.

## 9. Záver

Predkladaná práca popisuje viaceré druhy analýzy rozptylu ako sú jednofaktorová analýza rozptylu, dvojfaktorová analýza bez interakcií a dvojfaktorová analýza rozptylu s interakciami. Medzi základné predpoklady jednofaktorovej analýzy rozptylu patrí predpoklad rovnosti rozptylov. Tento predpoklad však nie je možné v praxi vždy dodržať. Preto sa používajú iné testy, ktoré tento predpoklad neuvažujú ako napríklad Welchov, Cochranov, Brown-Forsytheov, Modifikovaný Welchov test. Práve týmto testom sa práca venuje a porovnáva ich silu so štandardným testom jednofaktorovej analýzy rozptylu, ktorý je zabudovaný i v používanom softvéri Mathematica. Úspešnosť týchto testov závisí aj od rozsahu dát. Cochranov test je vhodné používať pri veľkom rozsahu a naopak Modifikovaný Welchov test pri malých rozsahoch dát. Dvojfaktorová analýza rozptylu je v práci obohatená o tzv. Tukeyho test, ktorý uvažuje prítomnosť interakcií i napriek tomu, že máme len jedno meranie v každej triede.

V ďalšej časti práce je rozobratá problematika analýzy kovariancie, ktorá úzko súvisí so spomínanou jednofaktorovou analýzou rozptylu. ANOVA skúma vzťah medzi intervalovou premennou a nominálnou. Ak intervalová premenná  $X$  má lineárny vzťah k premennej  $Y$  a regresný koeficient  $\beta$  medzi  $Y$  a  $X$  je rovnaký vo všetkých skupinách, potom model ANCOVA má tvar (8.2). ANCOVA vysvetľuje viac variability premennej  $Y$  a dá sa použiť len ak vzťah medzi  $X$  a  $Y$  je významný. Keďže softvér Mathematica, ktorý bol používaný, neobsahuje procedúru na počítanie ANCOVA, bolo treba túto procedúru doplniť.

Všetky spomínané testy sú podrobnejšie popísané v teoretickej časti. Samotné sú naprogramované v softvéri Mathematica a prakticky vyskúšané na príkladoch. Testy umožňujú užívateľovi interaktívny prístup (zadanie dát, hladiny významnosti a pri Modifikovanom Welchovom teste možnosť výberu  $\delta_1, \delta_2$ ).

## 10. Použitá literatúra

- [1] Anděl, J.: Matematická statistika. Praha SNTL/ Alfa 1985
- [2] Anděl, J.: Statistické metody. Praha Matfyzpress 1993
- [3] Argaç, D.: Testing for homogeneity in a general one – way classification with fixed effects: power simulations and comparative study, Computational statistics & Data Analysis, 44 (2004), 603 – 612
- [4] Hartung, J., Argaç, D., Makambi, K. H., Small sample properties of tests on homogeneity in one – way ANOVA and meta – analysis, Statistical Papers, 43 (2002), 197 - 235
- [5] Kalická, J., Krivá, Z.: Praktická štatistika v Exceli. Vydavateľstvo STU 2005
- [6] Savin, A.: Testy a konfidenčné intervaly pre strednú hodnotu v modeloch jednoduchého triedenia. Robust 2004, zborník príspevkov, 355 – 362
- [7] <http://faculty.vasaar.edu/lowry>
- [8] [http:// rimarcik.com](http://rimarcik.com)
- [9] [http:// www.wikipedia.org](http://www.wikipedia.org)

# **Prílohy**

## **Príloha č. 1**