


Primjena računara u biologiji



Vladimir Filipović

vladaf@matf.bg.ac.rs



Analysis of Variance and Non-parametric Methods

Elementary Statistics with R

- ▶ Qualitative Data

- ▶ Quantitative Data

- ▶ Numerical Measures

- ▶ Probability Distributions

- ▶ Interval Estimation

- ▶ Hypothesis Testing

- ▶ Type II Error

- ▶ Inference About Two Populations

- ▶ Goodness of Fit

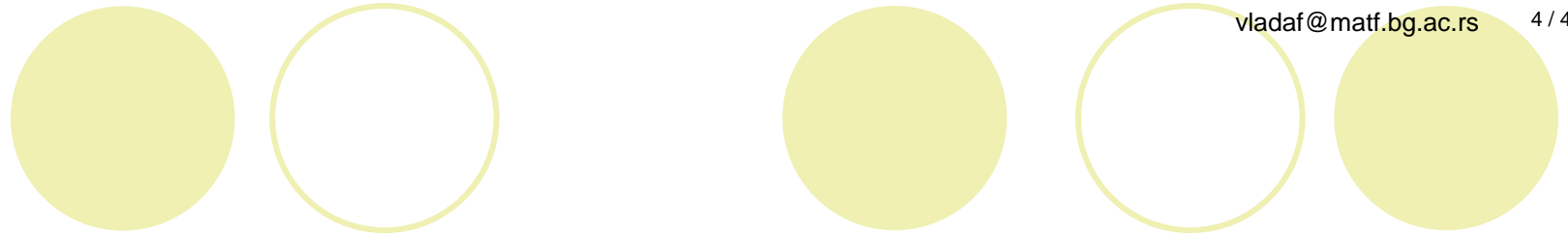
- ▶ Analysis of Variance

- ▶ Non-parametric Methods

- ▶ Simple Linear Regression

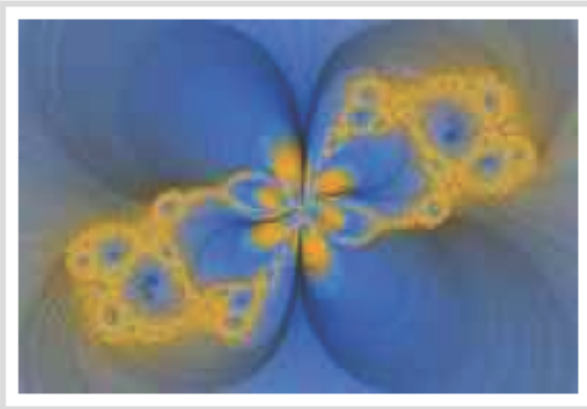
- ▶ Multiple Linear Regression

- ▶ Logistic Regression



Analysis of Variance

Analysis of Variance



In an experiment study, various treatments are applied to test subjects and the response data is gathered for analysis. A critical tool for carrying out the analysis is the **Analysis of Variance** (ANOVA). It enables a researcher to differentiate treatment results based on easily computed statistical quantities from the treatment outcome.

The statistical process is derived from estimates of the **population variances** via two separate approaches. The first approach is based on the variance of the **sample means**, and the second one is based on the mean of the sample variances. Under the ANOVA assumptions as stated below, the ratio of the two statistical estimates follows the **F distribution**. Hence we can test the null hypothesis on the equality of various response data from different treatments via estimates of critical regions.

- The treatment responses are independent of each other.
- The response data follow the **normal distribution**.
- The variances of the response data are identical.

In the following tutorials, we demonstrate how to perform ANOVA on a few basic experimental designs.

Analysis of Variance (2)

- Completely Randomized Design
- Randomized Block Design
- Factorial Design

Completely Randomized Design

In a **completely randomized design**, there is only one primary factor under consideration in the experiment. The test subjects are assigned to treatment levels of the primary factor at random.

Example

A fast food franchise is test marketing 3 new menu items. To find out if they the same popularity, 18 franchisee restaurants are randomly chosen for participation in the study. In accordance with the completely randomized design, 6 of the restaurants are randomly chosen to test market the first new menu item, another 6 for the second menu item, and the remaining 6 for the last menu item.

Completely Randomized Design (2)

Problem

Suppose the following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing. At .05 level of significance, test whether the **mean** sales volume for the 3 new menu items are all equal.

Item1	Item2	Item3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

Completely Randomized Design (3)

Solution

The solution consists of the following steps:

1. Copy and paste the sales figure above into a **table file** named "fastfood-1.txt" with a text editor.
2. Load the file into a **data frame** named df1 with the read.table function. As the first line in the file contains the column names, we set the header argument as TRUE.

```
> df1 = read.table("fastfood-1.txt", header=TRUE); df1
```

	Item1	Item2	Item3
1	22	52	16
2	42	33	24
3	44	8	19
4	52	47	18
5	45	43	34
6	37	32	39

Completely Randomized Design (4)

3. Concatenate the data rows of `df1` into a single vector `r`.

```
> r = c(t(as.matrix(df1))) # response data
> r
[1] 22 52 16 42 33 ...
```

4. Assign new variables for the treatment levels and number of observations.

```
> f = c("Item1", "Item2", "Item3") # treatment levels
> k = 3                             # number of treatment levels
> n = 6                             # observations per treatment
```

5. Create a vector of treatment factors that corresponds to each element of `r` in step 3 with the `gl` function.

```
> tm = gl(k, 1, n*k, factor(f)) # matching treatments
> tm
[1] Item1 Item2 Item3 Item1 Item2 ...
```

Completely Randomized Design (5)

6. Apply the function `aov` to a formula that describes the response `r` by the treatment factor `tm`.

```
> av = aov(r ~ tm)
```

7. Print out the ANOVA table with the `summary` function.

```
> summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tm	2	745	373	2.54	0.11
Residuals	15	2200	147		

Answer

Since the p-value of 0.11 is greater than the .05 significance level, we do not reject the null hypothesis that the mean sales volume of the new menu items are all equal.

Exercise

Create the response data in step 3 above along *vertical* columns instead of horizontal rows. Adjust the factor levels in step 5 accordingly.

Randomized Block Design

In a **randomized block design**, there is only one primary factor under consideration in the experiment. Similar test subjects are grouped into **blocks**. Each block is tested against all treatment levels of the primary factor at random order. This is intended to eliminate possible influence by other extraneous factors.

Example

A fast food franchise is test marketing 3 new menu items. To find out if they have the same popularity, 6 franchisee restaurants are randomly chosen for participation in the study. In accordance with the randomized block design, each restaurant will be test marketing all 3 new menu items. Furthermore, a restaurant will test market only one menu item per week, and it takes 3 weeks to test market all menu items. The testing order of the menu items for each restaurant is randomly assigned as well.

Randomized Block Design (2)

Problem

Suppose each row in the following table represents the sales figures of the 3 new menu in a restaurant after a week of test marketing. At .05 level of significance, test whether the **mean** sales volume for the 3 new menu items are all equal.

Item1	Item2	Item3
31	27	24
31	28	31
45	29	46
21	18	48
42	36	46
32	17	40

Randomized Block Design (3)

Solution

The solution consists of the following steps:

1. Copy and paste the sales figure above into a **table file** named "fastfood-2.txt" with a text editor.
2. Load the file into a **data frame** named df2 with the read.table function. As the first line in the file contains the column names, we set the header argument as TRUE.

```
> df2 = read.table("fastfood-2.txt", header=TRUE); df2
```

	Item1	Item2	Item3
1	31	27	24
2	31	28	31
3	45	29	46
4	21	18	48
5	42	36	46
6	32	17	40

Randomized Block Design (4)

3. Concatenate the data rows in df2 into a single vector r.

```
> r = c(t(as.matrix(df2))) # response data
> r
[1] 31 27 24 31 28 ...
```

4. Assign new variables for the treatment levels and number of control blocks.

```
> f = c("Item1", "Item2", "Item3") # treatment levels
> k = 3                             # number of treatment levels
> n = 6                             # number of control blocks
```

5. Create a vector of treatment factors that corresponds to the each element in r of step 3 with the gl function.

```
> tm = gl(k, 1, n*k, factor(f)) # matching treatment
> tm
[1] Item1 Item2 Item3 Item1 Item2 ...
```

Randomized Block Design (5)

6. Similarly, create a vector of blocking factors for each element in the response data `r`.

```
> blk = gl(n, k, k*n)           # blocking factor
> blk
[1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6
Levels: 1 2 3 4 5 6
```

7. Apply the function `aov` to a formula that describes the response `r` by both the treatment factor `tm` and the block control `blk`.

```
> av = aov(r ~ tm + blk)
```

8. Print out the ANOVA table with the `summary` function.

```
> summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tm	2	539	269	4.96	0.032 *
blk	5	560	112	2.06	0.155
Residuals	10	543	54		

Randomized Block Design (6)

Answer

Since the p-value of 0.032 is less than the .05 significance level, we reject the null hypothesis that the mean sales volume of the new menu items are all equal.

Exercise

Create the response data in step 3 above along *vertical* columns instead of horizontal rows. Adjust the factor levels in steps 5 and 6 accordingly.

Factorial Design

In a **factorial design**, there are more than one factors under consideration in the experiment. The test subjects are assigned to treatment levels of every factor combinations at random.

Example

A fast food franchise is test marketing 3 new menu items in both East and West Coasts of continental United States. To find out if they the same popularity, 12 franchisee restaurants from each Coast are randomly chosen for participation in the study. In accordance with the factorial design, within the 12 restaurants from East Coast, 4 are randomly chosen to test market the first new menu item, another 4 for the second menu item, and the remaining 4 for the last menu item. The 12 restaurants from the West Coast are arranged likewise.

Factorial Design (2)

Problem

Suppose the following tables represent the sales figures of the 3 new menu items after a week of test marketing. Each row in the upper table represents the sales figures of 3 different East Coast restaurants. The lower half represents West Coast restaurants. At .05 level of significance, test whether the **mean** sales volume for the new menu items are all equal. Decide also whether the mean sales volume of the two coastal regions differs.

East Coast:

=====

	Item1	Item2	Item3
E1	25	39	36
E2	36	42	24
E3	31	39	28
E4	26	35	29

Factorial Design (3)

West Coast:

=====

	Item1	Item2	Item3
w1	51	43	42
w2	47	39	36
w3	47	53	32
w4	52	46	33

Solution

The solution consists of the following steps:

Factorial Design (4)

1. Save the sales figure into a file named "fastfood-3.csv" in **CSV format** as follows.

```
Item1,Item2,Item3
E1,25,39,36
E2,36,42,24
E3,31,39,28
E4,26,35,29
W1,51,43,42
W2,47,39,36
W3,47,53,32
W4,52,46,33
```

2. Load the data into a **data frame** named df3 with the read.csv function.

```
> df3 = read.csv("fastfood-3.csv")
```

3. Concatenate the data rows in df3 into a single vector r .

```
> r = c(t(as.matrix(df3))) # response data
> r
[1] 25 39 36 36 42 ...
```

Factorial Design (5)

4. Assign new variables for the treatment levels and number of observations.

```
> f1 = c("Item1", "Item2", "Item3") # 1st factor levels
> f2 = c("East", "West")           # 2nd factor levels
> k1 = length(f1)                  # number of 1st factors
> k2 = length(f2)                  # number of 2nd factors
> n = 4                             # observations per treatment
```

5. Create a vector that corresponds to the 1th treatment level of the response data `r` in step 3 element-by-element with the `gl` function.

```
> tm1 = gl(k1, 1, n*k1*k2, factor(f1))
> tm1
[1] Item1 Item2 Item3 Item1 Item2 ...
```

Factorial Design (6)

6. Similarly, create a vector that corresponds to the 2nd treatment level of the response data `r` in step 3.

```
> tm2 = gl(k2, n*k1, n*k1*k2, factor(f2))  
> tm2  
[1] East East East East East ...
```

7. Apply the function `aov` to a formula that describes the response `r` by the two treatment factors `tm1` and `tm2` with interaction.

```
> av = aov(r ~ tm1 * tm2) # include interaction
```

8. Print out the ANOVA table with summary function.

```
> summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
tm1	2	385	193	9.55	0.0015	**
tm2	1	715	715	35.48	1.2e-05	***
tm1:tm2	2	234	117	5.81	0.0113	*
Residuals	18	363	20			

Factorial Design (7)

Answer

Since the p-value of 0.0015 for the menu items is less than the .05 significance level, we reject the null hypothesis that the mean sales volume of the new menu items are all equal. Moreover, the p-value of $1.2e-05$ for the east-west coasts comparison is also less than the .05 significance level. It shows there is a difference in overall sales volume between the coasts. Finally, the last p-value of 0.0113 (< 0.05) indicates that there is a possible interaction between the menu item and coast location factors, i.e., customers from different coastal regions have different tastes.

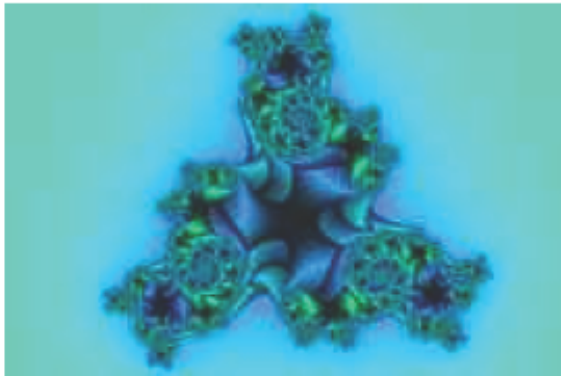
Exercise

Create the response data in step 3 above along *vertical* columns instead of horizontal rows. Adjust the factor levels in steps 5 and 6 accordingly.



Non-parametric Methods

Non-parametric Methods



A statistical method is called **non-parametric** if it makes no assumption on the population distribution or sample size.

This is in contrast with most parametric methods in elementary statistics that assume the data is quantitative, the population has a **normal distribution** and the sample size is sufficiently large.

In general, conclusions drawn from non-parametric methods are not as powerful as the parametric ones. However, as non-parametric methods make fewer assumptions, they are more flexible, more robust, and applicable to non-quantitative data.

-
- **Sign Test**
 - **Wilcoxon Signed-Rank Test**
 - **Mann-Whitney-Wilcoxon Test**
 - **Kruskal-Wallis Test**

Sign Test

A **sign test** is used to decide whether a **binomial distribution** has the equal chance of success and failure.

Example

A soft drink company has invented a new drink, and would like to find out if it will be as popular as the existing favorite drink. For this purpose, its research department arranges 18 participants for taste testing. Each participant tries both drinks in random order before giving his or her opinion.

Problem

It turns out that 5 of the participants like the new drink better, and the rest prefer the old one. At .05 significance level, can we reject the notion that the two drinks are equally popular?

Sign Test (2)

Solution

The null hypothesis is that the drinks are equally popular. Here we apply the `binom.test` function. As the p-value turns out to be 0.096525, and is greater than the .05 significance level, we do not reject the null hypothesis.

```
> binom.test(5, 18)
```

Exact binomial test

data: 5 and 18

number of successes = 5, number of trials = 18,

p-value = 0.09625

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.09695 0.53480

sample estimates:

probability of success

0.27778

Sign Test (3)

Answer

At .05 significance level, we do not reject the notion that the two drinks are equally popular.

Wilcoxon Signed-Rank Test

Two data samples are **matched** if they come from repeated observations of the same subject. Using the **Wilcoxon Signed-Rank Test**, we can decide whether the corresponding data population distributions are identical *without* assuming them to follow the **normal distribution**.

Example

In the built-in data set named **immer**, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the **data frame columns** Y1 and Y2.

```
> library(MASS)           # load the MASS package
> head(immer)
  Loc Var   Y1   Y2
1  UF  M  81.0 80.7
2  UF  S 105.4 82.3
  ....
```

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 in data set immer have identical data distributions.

Wilcoxon Signed-Rank Test (2)

Solution

The null hypothesis is that the barley yields of the two sample years are identical populations. To test the hypothesis, we apply the `wilcox.test` function to compare the matched samples. For the paired test, we set the "paired" argument as `TRUE`. As the p-value turns out to be 0.005318, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(immer$Y1, immer$Y2, paired=TRUE)
```

```
Wilcoxon signed rank test with continuity correction
```

```
data: immer$Y1 and immer$Y2
```

```
V = 368.5, p-value = 0.005318
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
Warning message:
```

```
In wilcox.test.default(immer$Y1, immer$Y2, paired = TRUE) :  
cannot compute exact p-value with ties
```

Wilcoxon Signed-Rank Test (3)

Answer

At .05 significance level, we conclude that the barley yields of 1931 and 1932 from the data set immer are *nonidentical* populations.

Mann-Whitney-Wilcoxon Test

Two data samples are **independent** if they come from distinct populations and the samples do not affect each other. Using the **Mann-Whitney-Wilcoxon Test**, we can decide whether the population distributions are identical *without* assuming them to follow the **normal distribution**.

Example

In the **data frame column** **mpg** of the data set **mtcars**, there are gas mileage data of various 1974 U.S. automobiles.

```
> mtcars$mpg  
[1] 21.0 21.0 22.8 21.4 18.7 ...
```

Meanwhile, another data column in **mtcars**, named **am**, indicates the transmission type of the automobile model (0 = automatic, 1 = manual). In other words, it is the differentiating factor of the transmission type.

```
> mtcars$am  
[1] 1 1 1 0 0 0 0 0 ...
```

In particular, the gas mileage data for manual and automatic transmissions are independent.

Mann-Whitney-Wilcoxon Test (2)

Example

In the **data frame column** **mpg** of the data set **mtcars**, there are gas mileage data of various 1974 U.S. automobiles.

```
> mtcars$mpg  
[1] 21.0 21.0 22.8 21.4 18.7 ...
```

Meanwhile, another data column in **mtcars**, named **am**, indicates the transmission type of the automobile model (0 = automatic, 1 = manual). In other words, it is the differentiating factor of the transmission type.

```
> mtcars$am  
[1] 1 1 1 0 0 0 0 0 ...
```

In particular, the gas mileage data for manual and automatic transmissions are independent.

Problem

Without assuming the data to have normal distribution, decide at .05 significance level if the gas mileage data of manual and automatic transmissions in **mtcars** have identical data distribution.

Mann-Whitney-Wilcoxon Test (3)

Solution

The null hypothesis is that the gas mileage data of manual and automatic transmissions are identical populations. To test the hypothesis, we apply the `wilcox.test` function to compare the independent samples. As the p-value turns out to be 0.001817, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(mpg ~ am, data=mtcars)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: mpg by am
```

```
W = 42, p-value = 0.001871
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
Warning message:
```

```
In wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8, ...):  
cannot compute exact p-value with ties
```

Mann-Whitney-Wilcoxon Test (4)

Answer

At .05 significance level, we conclude that the gas mileage data of manual and automatic transmissions in mtcars are *nonidentical* populations.

Kruskal-Wallis Test

A collection of data samples are **independent** if they come from unrelated populations and the samples do not affect each other. Using the **Kruskal-Wallis Test**, we can decide whether the population distributions are identical *without* assuming them to follow the **normal distribution**.

Example

In the built-in data set named **airquality**, the daily air quality measurements in New York, May to September 1973, are recorded. The ozone density are presented in the **data frame column** Ozone.

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1   41     190  7.4   67     5   1
2   36     118  8.0   72     5   2
  ....
```

Kruskal-Wallis Test (2)

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.

Solution

The null hypothesis is that the monthly ozone density are identical populations. To test the hypothesis, we apply the `kruskal.test` function to compare the independent monthly data. The p-value turns out to be nearly zero ($6.901e-06$). Hence we reject the null hypothesis.

```
> kruskal.test(Ozone ~ Month, data = airquality)
```

```
Kruskal-Wallis rank sum test
```

```
data: Ozone by Month
```

```
Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

Kruskal-Wallis Test (3)

Answer

At .05 significance level, we conclude that the monthly ozone density in New York from May to September 1973 are *nonidentical* populations.

Acknowledgments

Material in this presentation is taken from
<http://www.r-tutor.com/>