10. ANOVA – One Factor Analysis of Variance

Study time: 60 minutes

Aim - you will be able to

- explain structure of *F-ratio*
- conclude by the test named Analysis of Variance
- construct the ANOVA table
- realize the post hoc analysis

Explication

10.1. Introduction

We talked about one-sample and two-sample tests for mean in the previous lectures. The analysis of variance (ANOVA) is an extension of these tests. It enables compare any mean of independent random samples. The analysis of variance (in its parametric form) assumes normality of the distributions and homoscedasticity (identical variances). If these conditions are not executed then we must use nonparametric *Kruskall-Wallis test*. It is an analog of the one-factor sorting in the analysis of the variance. It doesn't assume distributions normality but its disadvantage is a smaller sensitivity.

10.2. Construction of the F-statistic

Let's have *k*-random samples that are independent on each other. These samples have the standard distribution with the same variation:

$(X_{11}, X_{12}, ..., X_{1n_1}) \rightarrow N(\mu_1, \sigma^2)$

$(X_{21}, X_{22}, ..., X_{2n_2}) \rightarrow N(\mu_2, \sigma^2)$

...

$(X_{k1}, X_{k2}, ..., X_{kn_k}) \rightarrow N(\mu_k, \sigma^2)$,

---

$$\sum_{i=1}^{k} n_i = N$$

Let n_i ... number of observations in i-th sample.

***Formulation of the problem*:**
The hypothesis of interest is H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$
The alternate hypothesis is: H_A: At least two μ_i's are different.

We want determine on H_0 in terms of one test. Cause we try to find such test statistic that enable not only H_0 implementation but also it was sensitive on the H_0 validity.

Define the **total sum of squares** (or **total variability**) as

$$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2 \text{ , where } \overline{X} \text{ is the mean of all observations.}$$

- the total sum of squares is our raw measure of variability in the data

This total sum of squares we can separate into 2 components:

$$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2 \Rightarrow SS_{TOTAL} = SS_W + SS_B \text{ ,}$$

where
SS_W ... the within group variation (the sum of squares within groups) - is the raw variability within samples

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

- the degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: N - k

S_i is a sample standard deviation of i-th random sample:

$$S_i = \sqrt{\frac{\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}{n_i - 1}}$$

and

$$\overline{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \quad \text{is a sample mean in } i\text{-th sample.}$$

SS_B ... the between group variation (the sum of squares between groups) - is the raw variability between samples:

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

The within group variance (mean square within groups): $S_W^2 = \frac{SS_W}{N-k}$

The between group variance (mean squares between groups): $S_B^2 = \frac{SS_B}{k-1}$

Properties of these variances:

1. $ES_W^2 = \frac{1}{N-k} E(\sum_{i=1}^{k} (n_i - 1) S_i^2) = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) E(S_i^2) = \sigma^2$

   because $E(S_i^2) = \sigma^2$.

   The within mean square is an unbiased estimate of the variance, independently of H_0.

2. $ES_B^2 = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^{k} n_i (E\overline{X} - E\overline{X}_i)^2$

   $ES_B^2 = \sigma^2 \Leftrightarrow$ when H_0 is true

Therefore the ratio of the two sums of squared divided by their degrees of freedom will have an F distribution under the hypothesis of equal population means.

$$F = \frac{SS_B / k-1}{SS_W / N-k} = \frac{S_B^2}{S_W^2}$$

Definition:
We call this F statistic as ***F-ratio***.

Why is useful use *F*-ratio as the test statistic?

We see that if H_0 is true then *F-ratio* is any random number close to 1 ... $F \approx 1$. If H_0 is faslse then this number is markedly bigger than 1 (see property 2). The statistic *F-ratio* is sensitive to validity of the hypothesis H_0. So we can use it during following testing as test statistic we have to determine its statistical behavior what means to determine its propability distribution.

We know that $\frac{S_w^2}{\sigma^2} \cdot (N-k) = \sum_{i=1}^{k} \frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(N-k)$,

because $\frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(n_i - 1)$, and further is known that sum of random variables $\chi^2(n_i - 1)$ is again a random variable of a same type with degrees of freedom number same as summarized variables.

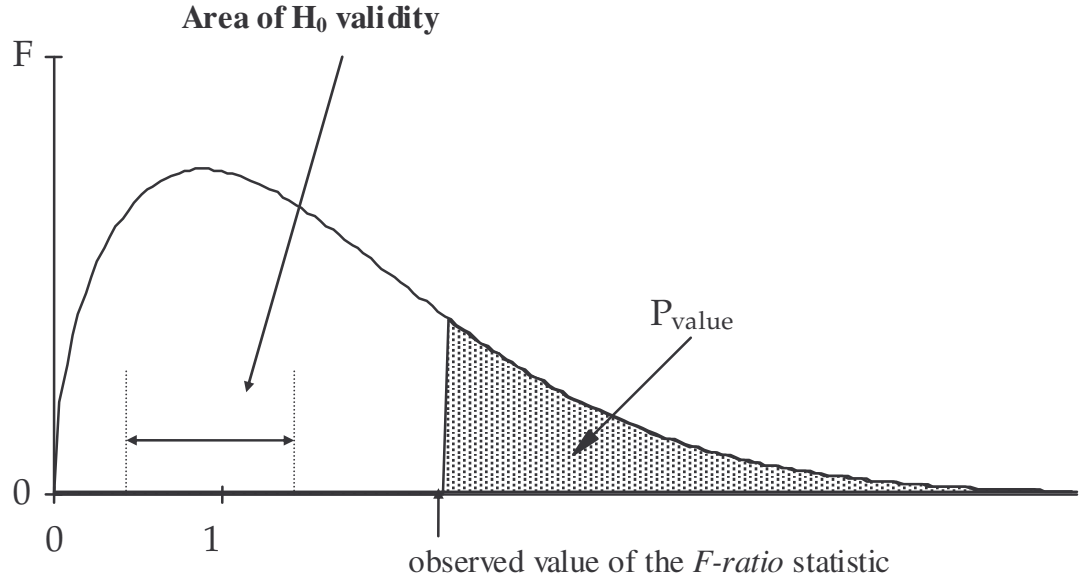
If H_0 is true then: $\frac{S_B^2}{\sigma^2} \cdot (k-1) = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2 = \sum_{i=1}^{k} (\frac{\overline{X}_i - \overline{X}}{\frac{\sigma}{\sqrt{n_i}}})^2 \rightarrow \chi^2(k-1)$

Then we know (Fisher-Snedecor distribution) that following ratio:

$$\frac{\frac{S_B^2}{\sigma^2} \cdot (k-1)}{k-1} = \frac{S_B^2}{S_W^2} = F_{k-1, N-k}$$
$$\frac{\frac{S_w^2}{\sigma^2} \cdot (N-k)}{N-k}$$

it must have *F distribution* with *(k-1)* and *(N-k)* degrees of freedom.
If we know a F-ratio statistical behavior we can use it for analysis and determination of previously stated problem in H_0. Following figure illustrated a usage of *F-ratio* to determine

a hypothesis H_0 validity.



10.3. ANOVA Table

We summarize the data in an ANOVA table:

Source	Sum of squares	Degrees of freedom	Mean squares	*F-ratios*	P-value
total	$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$	$N-1$			
between	$SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$	$k-1$	$S_B^2 = \frac{SS_B}{k-1}$	$F = \frac{S_B^2}{S_W^2}$	see definition
within	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$	$N-k$	$S_W^2 = \frac{SS_W}{N-k}$		

Analysis of variance table - ANOVA

The big values of F-ratio indicate small values of p_{value} what means rejection of H_0. The F-ratio value will be a big number if the within group variation is a negligible part of the total variability and equivalently if the between variation is a significant part of the total variability.

10.4. Solved example

We assume three data sets for illustration of *F-ratio* statistical behavior. In each data set, the sample means are the same but the variations within groups differs. When the within group variation is small than the *F-ratio* is large. When the within group variation is large than the *F-ratio* is small. The examples illustrate three cases: small within group variation; normal within group variation; and large within group variation.

Example 1:

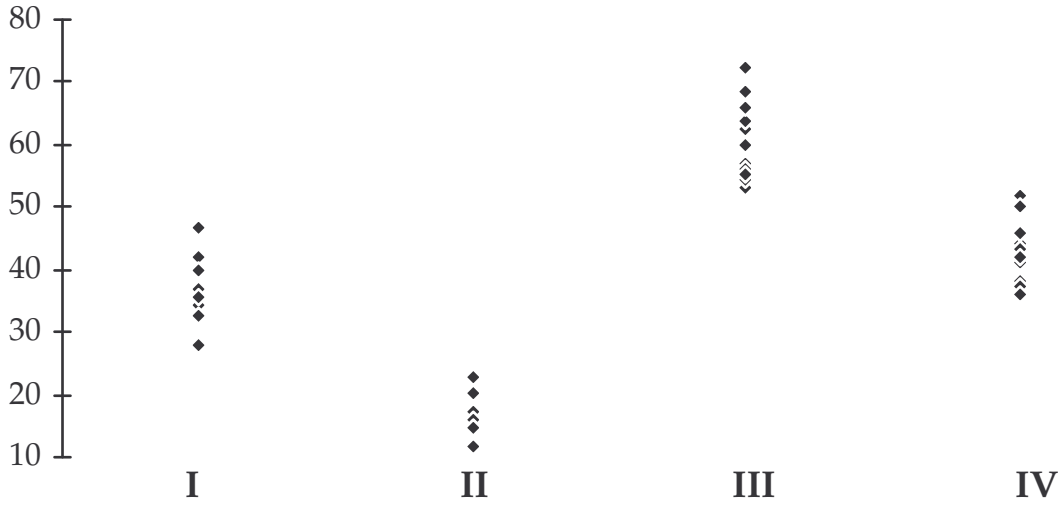
Small within group variation

Groups	I	II	III	IV
Data	42 34.5 32.5 40 46.5 28 37 35.5	17.5 12 16 15 20.5 23 15	68.5 72 53 64 57 56 54.5 62.5 63.5 60 66 55	38 44 52 50 43.5 41 42 46 37.5 36
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	5.78	3.71	6.06	5.27

ANOVA Table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	9872.7027		
between	3	8902.7027	2967.57	100.96
within	33	970	29.39	

P-value = 0.0000



Example 2:

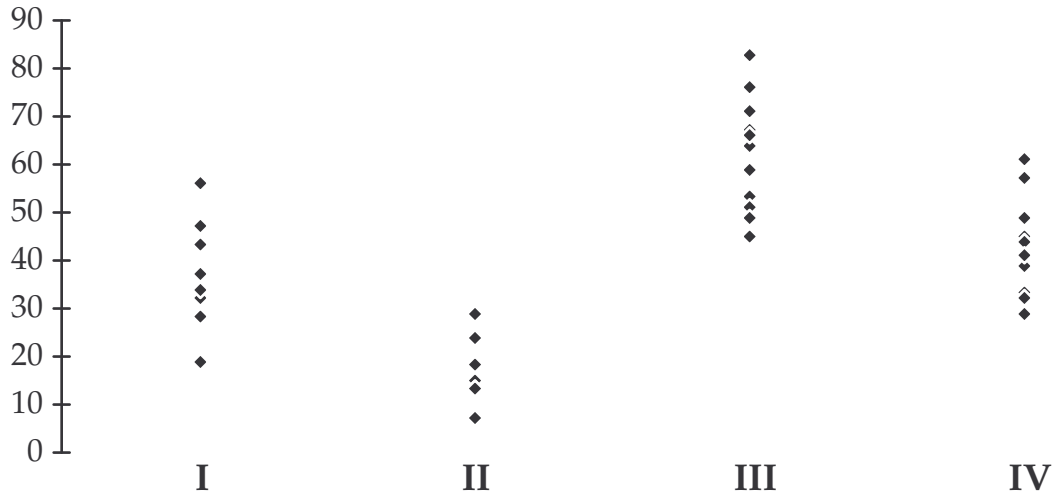
Normal within group variation

Groups	I	II	III	IV
Data	47 32 28 43 56 19 37 34	18 7 15 13 24 29 13	76 83 45 67 53 51 48 64 66 59 71 49	33 45 61 57 44 39 41 49 32 29
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	11.56	7.42	12.12	10.53

ANOVA table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	12782.7027		
between	3	8902.7027	2967.57	25.24
within	33	3880	117.58	

P-value = 0.0000



Example 3:

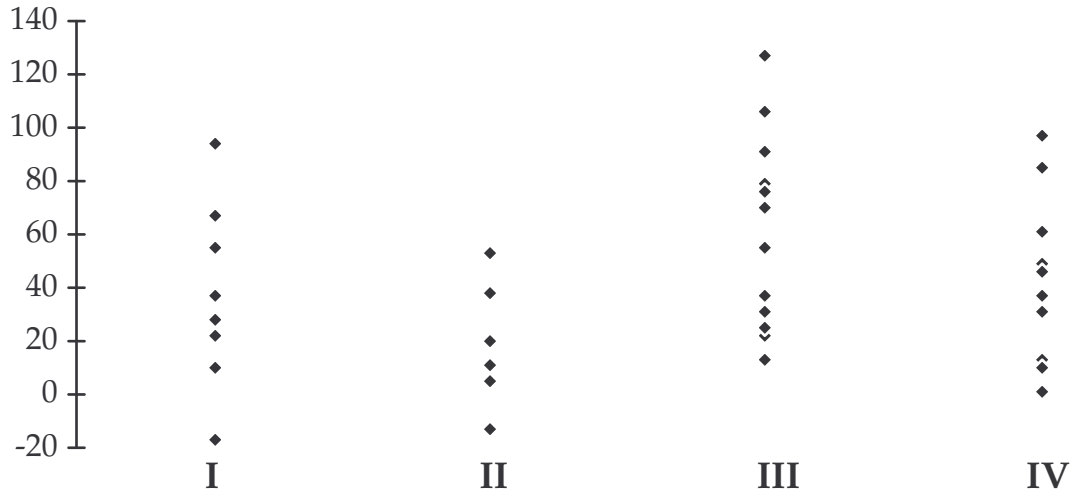
Large within group variation

Groups	I	II	III	IV
Data	67 22 10 55 94 -17 37 28	20 -13 11 5 38 53 5	106 127 13 79 37 31 22 70 76 55 91 25	13 49 97 85 46 31 37 61 10 1
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	34.69	22.25	36.36	31.59

ANOVA table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	43822.7027		
between	3	8902.7027	2967.57	2.804
within	33	34920	1058.18	

P-value = 0.0549



10.5. Post Hoc analysis

A large *F-ratio* indicates only that some differences exist among the group means, but not where those differences occur. If the *F-ratio* is large, our analysis would be incomplete without identifying which group means differ. This process is called **post hoc** analysis, and consists of comparing the means of all pairs of samples to determine if there is a difference of means.

Several methods are available for post hoc multiple comparisons. We will discuss the simplest method here, least significant differences. The Least Significant Difference or **LSD-method** consists of applying the two-sample *t* test to every pair of sample means. However, we make one adjustment and use the square root of mean square within rather than the pooled standard deviation from the two samples as our estimate of population standard deviation. Thus for any pair of sample means, we compute LSD as,

$$(LSD)_{i,j} = \frac{\overline{X}_i - \overline{X}_j}{S_W \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \rightarrow t_{N-k}$$

where $S_W = \sqrt{S_W^2} = \sqrt{\frac{SS_W}{N-k}}$.

This statistic has Student distribution with *N-k* degrees of freedom.

The LSD method is illustrated for the three examples given previously.

Example 1: Small within group variation

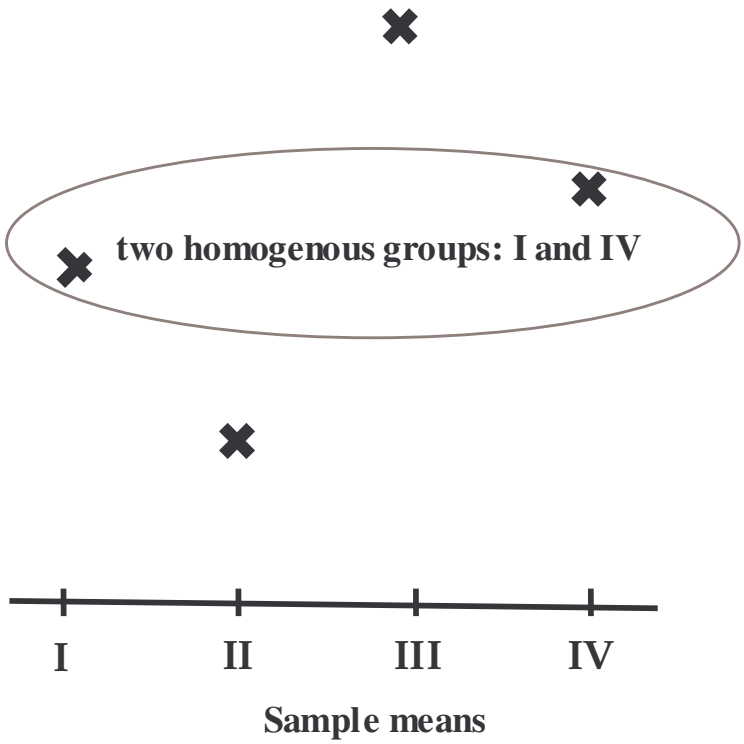
We determine $(LSD)_{i,j}$ for all pairs of given four groups and the obtained values we inscribe in the following table:

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-7.128	9.698	2.333
7	II	7.128	0	17.064	9.731
12	III	-9.698	-17.064	0	-7.754
10	IV	-2.333	-9.731	7.7541	0

In this case, there is very strong evidence of differences between all groups except I and IV where the evidence is not as strong.

Example 2: **Normal within group variation**

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-3.564	4.849	1.167
7	II	3.564	0	8.532	4.8656
12	III	-4.849	-8.532	0	-3.877
10	IV	-1.167	-4.866	3.877	0



In this case, even though the sample means are the same, there is no evidence of differences between the means of Groups I and IV. Therefore there are essentially three Groups: II; III; and I and IV together.

Example 3: **Large within group variation**

Since the *F-ratio* for this example is very small, we would normally conclude that there is no evidence against the null hypothesis of equal group means and not proceed further. Any two-sample *t* test which produces small p-values should be regarded as spurious. However, for illustration, we have produced the table of least significant differences.

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-1.188	1.616	0.389
7	II	1.188	0	2.844	1.622
12	III	-1.616	-2.844	0	-1.292
10	IV	-0.389	-1.622	1.292	0

In this example, the only least significant difference which has a small **p-value** is between groups II and III. However, because the overall *F-ratio* was too small, this difference would be disregarded and we would conclude that no differences exist between the means of any of the groups.

Note:
There exists other tests then LSD method which allows similar multiple comparisons what means a post hoc analysis. Also there were developed more flexible methods which are accessible thru the more advanced software (e.g. Duncan test, Tukey test for significant differences, Scheffe test and Bonferoni test). These tests are based on similar decision strategy and that's on setting of a critical difference requested for determination if two sample means from several groups are different. In many cases these tests are much more effective than LSD method.

10.6. Kruskal-Wallis test

The *F-ratio* test statistic used in the standard analysis of variance is known to be very sensitive to the assumption that the original observations are normally distributed. Because the test statistic is based on squared deviations from the mean, it can be badly distorted by outliers. For two-sample analysis, the Wilcoxon/Mann Whitney rank test was introduced as a nonparametric alternative which is less sensitive to outliers than the *t* test. For multiple samples, the Kruskal-Wallis rank test can be used for the same purpose. Like the Wilcoxon/Mann Whitney test, the Kruskal-Wallis test substitutes the ranks of the original data values and performs an analysis of variance on the ranks. For the large deviation data of the previous example the ranks for each group are listed in the following table.

Groups	I	II	III	IV
	28	11	36	9.5
Ranks	12.5	2	37	23
of original	6.5	8	9.5	35
data	25.5	4.5	31	32
	34	21	19	22
	1	24	16.5	16.5
	19	4.5	12.5	19
	15		29	27
			30	6.5
			25.5	3
			33	
			14	
Sample size	8	7	12	10
Mean rank	17.6875	10.7143	24.4167	19.35
Standard deviation	11.1674	8.5919	9.6668	10.6538

The test statistic is a modification of calculating the *F-ratio* for the ranks. In this example, the test statistic and p-value are:

K-W test statistics = 7.24325 p-value = 0.0645

The p-value for the Kruskal-Wallis test is slightly higher than for the *F-test*, but the conclusions are the same in both cases. The null hypothesis of equal group means is not rejected.

Summary of notions

Analysis of variance (ANOVA) is an extension of the two-sample tests for means and it enables compare any mean of independent random samples. ***F-ratio*** is the test statistics in analysis of variance. *F-ratio* statistics is sensitive to validity of the hypothesis H_0, which is formulate as an equality of the samples means. Particular interresults (that we execute during analysis of variance) are recorded into **ANOVA table**. The second step (in ANOVA) is **post hoc** analysis, and consists of comparing the means of all pairs of groups of purpose to choose homogenous groups. **LSD-statistics** is a criterion for assignment to homogenous groups.
Described procedure ANOVA is sensitive to the assumption that the original observations are normally distributed. If this condition is not executed then we must use nonparametric ***Kruskall-Wallis*** *rank test*.

Questions

1. Describe construction and statistical behavior of the *F-ratio* statistics.
2. What is the usual output from analysis of variance?
3. What is post hoc analysis?

Problems

Example 1: We made a research of dependency of earning and achieved education. In the table there are earnings in thousand CZK at 7 randomly selected men at each level of education. (B - basic, H - high, U - university).

	B	H	U
1	10.9	8.9	11.2
2	9.8	10.3	9.7
3	6.4	7.5	15.8
4	4.3	6.9	8.9
5	7.5	14.1	12.2
6	12.3	9.3	17.5
7	5.1	12.5	10.1

Do a simple sorting and determine if education does influence earning.

{*Answer:* p-value = 0.057}

Example 2: From a large set of homes we randomly selected 5 single homes, 8 couple, 10 three-member, 10 four-member and 7 five-member homes. We watched their month spending for food and drinks for one family member (in CZK). Confirm by analysis of variance if a month spending for food and drinks depends on a number of family members.

	Spending for one family member (in CZK)				
Number of family members	**1**	**2**	**3**	**4**	**5**
	3.440	2.350	2.529	2.137	2.062
	4.044	3.031	2.325	2.201	2.239
	4.014	2.143	2.731	2.786	2.448
	3.776	2.236	2.313	2.132	2.137
	3.672	2.800	2.303	2.223	2.032
		2.901	2.565	2.433	2.101
		2.656	2.777	2.224	2.121
		2.878	2.899	2.763	
			2.755	2.232	
			3.254	2.661	

{*Answer:* Use suitable software package.}