Statistical Inference

ANOVA - Nonparametric Statistics

Sonia Malefaki

Department of Mechanical Engineering & Aeronautics
University of Patras, Greece.

☎: 2610 997673, ✉: smalefaki@upatras.gr

December 21, 2020

Outline

Analysis of Variance

Introduction

Goal: To test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

where k is the number of independent populations

We can do all the possible tests in pairs.
In this case, the probability of Type I Error increases, i.e. the possibility to obtain statistically significant differences when in fact there are not.

For example if $k = 10$, there are 45 possible pairs for testing. By using $\alpha = 5\%$ in each of these tests, $0.05 \times 45 \approx 2$ tests may give us statistically significant differences in means completely at random.

- Thus, a new methodology is required.

One-way ANOVA

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs } H_1: \ \text{at least one } \mu_i \text{ differs}$$

- Y_{ij} i-th observation that belongs to j-th group, $i = 1, \ldots, n_j$ and $j = 1, \ldots, k$
- $Y_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}$ the sum of observations of the j-th group.
- $\overline{Y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ sample mean of j-th group
- $Y_{\cdot\cdot} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}$ the sum of all the observations
- $\overline{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}$ the sample mean, where $n = \sum_{j=1}^{k} n_j$

One-way ANOVA

Two sources of variability in the data:

- variation among observations within a population
- variation among populations that is due to the differences in the characteristics of the population,

Part of the goal of the analysis of variance is to determine if the differences among the k sample means are what we would expect due to random variation alone or, rather, due to variation beyond merely random effects.

- Assumptions

The k populations are independent and normally distributed with means $\mu_1, \ \ldots, \ \mu_k$ and common variance σ^2

One-way ANOVA

$$y_{ij} - \overline{y}_{..} = y_{ij} - \overline{y}_{.j} + \overline{y}_{.j} - \overline{y}_{..}$$

- $y_{ij} - \overline{y}_{.j}$ variability within group
- $\overline{y}_{.j} - \overline{y}_{..}$ variability between groups .

- If the variability between groups is large and the variability within groups is small, the null hypothesis must be rejected.
- If the variability between groups is small and the variability within groups is large, the null hypothesis must not be rejected.

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{.j})^2 + \sum_{j=1}^{k} n_j(\overline{y}_{.j} - \overline{y}_{..})^2$$

where

- SST Total Sum of Squares
- SSW Sum of Squares Within group
- SSB Sum of Squares Between groups

$$MSW = \frac{SSW}{n-k}, \qquad MSB = \frac{SSB}{k-1},$$

$$F = \frac{MSB}{MSW} \sim F_{k-1,n-k}$$

Thus, at α% significance level, the hull hypothesis is rejected if $F > F_{k-1,n-k \cdot a}$.

One-way ANOVA

Variability	Degrees of Freedom	Sum of Squares	Mean Square
Between groups	$k-1$	$\sum_{j=1}^{k} n_j(\overline{y}_{.j} - \overline{y}_{..})^2 = SSB$	$\frac{SSB}{k-1} = MSB$
Within groups	$n-k$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{.j})^2 = SSW$	$\frac{SSW}{n-k} = MSW$
Total	$n-1$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{..})^2 = SST$	$\frac{SST}{n-1} = MST$

$$F = \frac{MSB}{MSW} \sim F_{k-1,n-k}$$

Reject H_0 when $F > F_{k-1,n-k,a}$.

One-way ANOVA

- Homoscedasticity. Levene's test is applied for testing the equality of variances in the k populations which is a generalization of the standard F-test for testing the equality of variances in two populations.

- When the above assumptions are not valid, either the response variable is transformed, or the corresponding nonparametric test, named Kruskal Wallis test is applied.

- If H_0 is rejected, i.e. if it is assumed that there is a statistically significant difference between three or more groups, then it is important to find between which groups there is difference.

- Bonferroni's method. N tests are applied and the $p-$value is readjusted to $p' = N\,p$ with the assumption that $p' \leq 1$.

One-way ANOVA

22 patients who underwent heart surgery were divided into three groups.

- Group 1. Patients who were treated with a mixture of 50% nitrous oxide and 50% oxygen for 24 hours.
- Group 2. Patients who were treated with a mixture of 50% nitrous oxide and 50% oxygen Only during surgery.
- Group 3. Patients who were treated with a mixture of $35 - 50\%$ nitrous oxide and 50% oxygen for 24 hours.

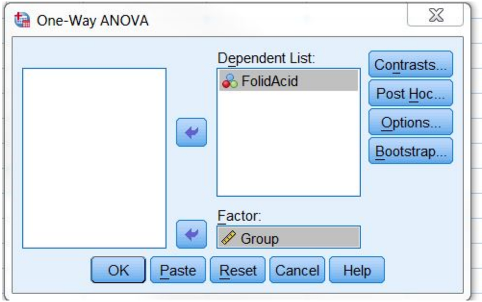
In the following table the value of folic acid mg/l in erythrocytes of patients is presented.

Patients	Group 1	Group 2	Group 3
1	243	206	241
2	251	210	258
3	275	226	270
4	291	249	293
5	347	255	328
6	354	273	
7	380	285	
8	392	295	
9		309	

Do the patients of the three groups have the same mean level of folic acid mg/l in their erythrocytes?

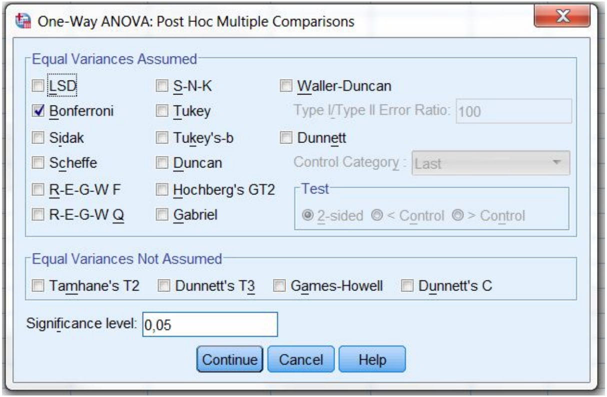
One-way ANOVA in SpSS

`Analyze > Compare Means > One-Way ANOVA...`



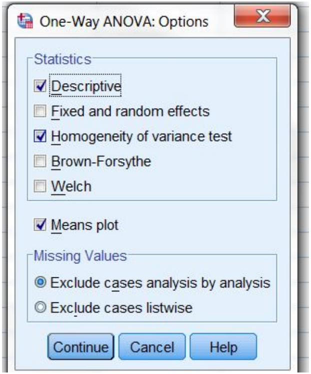
One-way ANOVA in SpSS

Post Hoc...



One-way ANOVA in SpSS

Options...



One-way ANOVA in SpSS

Output

Descriptives

olicAcid

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1,00	8	316,6250	58,71709	20,75963	267,5363	365,7137	243,00	392,00
2,00	9	256,4444	37,12180	12,37393	227,9101	284,9788	206,00	309,00
3,00	5	278,0000	33,75648	15,09636	236,0858	319,9142	241,00	328,00
Total	22	283,2273	51,28439	10,93387	260,4890	305,9655	206,00	392,00

Test of Homogeneity of Variances

FolicAcid

Levene Statistic	df1	df2	Sig.
3,823	2	19	,040

ANOVA

FolicAcid

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	15515,766	2	7757,883	3,711	,044
Within Groups	39716,097	19	2090,321		
Total	55231,864	21			

Multiple Comparisons

Dependent Variable: FolicAcid

Bonferroni

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1,00	2,00	60,18056*	22,21594	,042	1,8614	118,4998
	3,00	38,62500	26,06443	,464	-29,7969	107,0469
2,00	1,00	-60,18056*	22,21594	,042	-118,4998	-1,8614
	3,00	-21,55556	25,50141	1,000	-88,4995	45,3884
3,00	1,00	-38,62500	26,06443	,464	-107,0469	29,7969
	2,00	21,55556	25,50141	1,000	-45,3884	88,4995

Chi-Square test

Goodness-of-fit test

- a test to determine if a population has a specified theoretical distribution.

  H_0 : The population has a specified theoretical distribution vs
  H_1 : The population has NOT the specified theoretical distribution

- The test is based on how good is the obtained fit between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

- For example, we consider the tossing of a die. We hypothesize that the die is honest.

$$H_0 : f(x) = \frac{1}{6},\ x = 1, \ldots, 6 \text{ vs } H_1 : \text{not } H_0 :$$

The die is tossed 120 times and each outcome is recorded

x	1	2	3	4	5	6
observed	18	22	30	21	17	12
expected	20	20	20	20	20	20

Test at a 5% significant level if the dice is unbiased.

Goodness-of-fit test

- The observed frequencies O_i
- Compute the expected frequencies under the null hypothesis.
  $E_i = np_i$
  - n : the sample size
  - p_i : the probability of the value x_i.
- Compute the quantity

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim X^2_{k-1} \text{ under } H_0$$

- If $X^2 > X^2_{k-1 \cdot \alpha}$, then H_0 is rejected at α% significant level.
- The decision criterion described here should not be used unless each of the expected frequencies is greater or equal to 5.

- $\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7$
  $\chi^2_{5,0.05} = 11.07$

Goodness-of-fit test

- Example

The level of noise in a village near to a racetrack is rated during a race at a scale of 1-5 from all adult residents of the population. The results of the responses of the people presented in the following table.

Noise Level	1	2	3	4	5
Frequencies	90	88	85	104	128

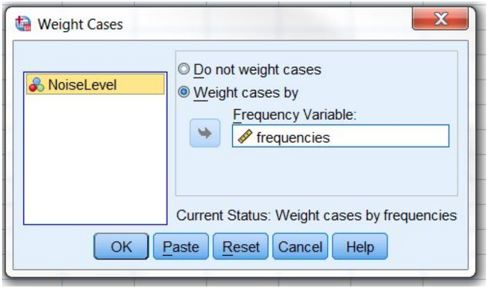
At $\alpha = 5\%$ significance level, test if there is a difference between the percentage of responses of residents about how they experience the noise level.

Goodness-of-fit test

	H0 p1=p2=p3=p4=p5 = 0.2		
Noise Level	**Oi**	**Ei**	**(Oi-Ei)^2/Ei**
1	90	99	0,81818182
2	88	99	1,22222222
3	85	99	1,97979798
4	104	99	0,25252525
5	128	99	8,49494949
N=	495	X2=	**12,7676768**
		X2(4,0.05)=	**9,48772904**
		απορρίπτω H0	

Goodness-of-fit test

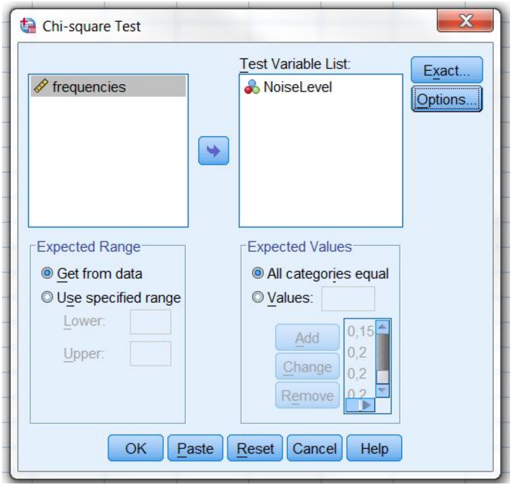
`Data > Weight Cases...`



	NoiseLevel	frequencies
1	1,00	90,00
2	2,00	88,00
3	3,00	85,00
4	4,00	104,00
5	5,00	128,00

Goodness-of-fit test

`Analyze > Nonparametric Tests > Legacy Dialogs > Chis-square...`



Goodness-of-fit test

Chi-Square Test

Frequencies

Noise Level

	Observed N	Expected N	Residual
1,00	90	99,0	-9,0
2,00	88	99,0	-11,0
3,00	85	99,0	-14,0
4,00	104	99,0	5,0
5,00	128	99,0	29,0
Total	495		

Test Statistics

	Noise Level
Chi-Square	12,768[a]
df	4
Asymp. Sig.	,012

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 99,0.

Goodness-of-fit test

▶ Example

The Weibull distribution with cumulative distribution function (cdf)

$$F(t) = 1 - e^{-(t/\beta)^{\alpha}}, \qquad t > 0$$

contains as a spacial case for $\alpha = 2$ the Rayleigh distribution which is used to describe the error in the determination of the location of an item by its programmer. When the errors are analyzed in rectangular coordinate systems, they are described by two independent normal distributions with mean 0 and common standard deviation. A company that produces positioning devices, argues that the determination error (in meters) of the location that a new device presents when used in open area follows the Rayleigh distribution and at 50% of cases, the error is less than 0.832555 meters. The following table presents the measurements of determination error (in meters) of the location in a random sample of 66 selected users of the new device.

Determining error	<0.3	0.3-0.5	0.5-0.7	0.7-0.9	0.9-1.1	>1.1
Frequencies	6	8	11	12	10	19

Based on the above data, check the validity of the company's claim.

Goodness-of-fit test

Determining error	Oi	pi	Ei	(Oi-Ei)^2/Ei
<0.3	6	0,086069	5,680542	0,01796546
0.3-0.5	8	0,13513	8,918607	0,09461545
0.5-0.7	11	0,166174	10,96751	9,625E-05
0.7-0.9	12	0,167768	11,07271	0,07765646
0.9-1.1	10	0,146661	9,679612	0,01060461
>1.1	19	0,298197	19,68102	0,02356528
N=	66		X2=	**0,22450352**
			X2(5,0.05):	**11,0704977**
		δεν μπορώ να απορρίψω την H0		

Goodness-of-fit test

- When the X^2 test is applied to continuous variables, it is influenced by the grouping of the data. So, X^2 goodness of fit test is preferred when we have categorical variables with finite state space.

In SpSS, we cannot apply X^2 goodness of fit test to continuous variables. In cases where continuous data is available, Kolmogorov - Smirnov test is preferable where it is based on the empirical cumulative distribution function.

- The X^2 test can be applied even when the parameters of the population distribution are unknown. In this case the degrees of freedom of the statistical test are reduced according to the number of parameters under estimation. The unknown parameters of the distribution are estimated by the observations and in this case the test statistic has the following form

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - \widehat{E}_i)^2}{\widehat{E}_i} \sim X^2_{k-1-r}$$

and H_0 is rejected when $X^2 > X_{k-1-r,\alpha}$

Kolmogorov - Smirnov test

Empirical cumulative distribution function

Let $x_1, x_2, \ldots, x_n$ be a random sample.

$$F_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

If the sample is derived from the assumed distribution then the empirical cumulative distribution function should not differ significantly from the theoretical cdf.

It holds

$$P(\lim_{n\to\infty} |F_n(x) - F(x)| = 0) = 1, \quad \forall x$$

Kolmogorov - Smirnov test is based on the observed differences of $F_n(x)$ and $F(x)$.

Kolmogorov - Smirnov test

$$D_n^+ = \sup\{F_n(x) - F(x)\}$$
$$D_n^- = \sup\{F(x) - F_n(x)\}$$

The test statistic is

$$\begin{aligned} D &= \sup\{|F_n(x) - F(x)|\} \\ &= \max\{D_n^+, D_n^-\} \end{aligned}$$

It is based on the maximum observed difference of the theoretical and the empirical cdf.

Kolmogorov - Smirnov test

Under H_0, it holds

$$P(\sqrt{n}\; D < d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{k-1}e^{-2k^2/d^2}, \quad 0 \leq d \leq 1$$

This is true for any theoretical distribution assumed.

H_0 is rejected at significance level a if $D > D_{n,a}$ where $D_{n,a}$ the value of the corresponding table.

- Kolmogorov - Smirnov test requires the theoretical distribution under the null hypothesis to be fully determined.

- If the theoretical distribution is not known, its parameters are estimated by the data. But in this case there are no tables to give the critical values so simulations are needed to identify them.

Example:

Test if the observations given in the file `KS1.sav` come from the normal distribution.

Test of independence

The chi-square test can also be used to test the hypothesis of independence of two variables of classification.

H_0 The variables A and B are independent
H_1 The variables A and B are dependent

	B_1	B_1	$\cdots$	B_c
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}

H_0 The probability in the cell (i, j) is equal to the product of the probabilities to be in the group i of the variable A and in group j of the variable B

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \ \forall i, j$$

H_1 Not H_0 i.e. $p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}$, for at least one (i, j).

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \mathcal{X}^2_{(r-1)(c-1)}$$

Test of independence

Example:

The aeronautical aluminum alloy 2024-73 and 2024-74 are often used in the fuselage and wings of the aircraft. During the repair of 250 damages in the fuselage and the wings of aircraft (a) the type of aluminum alloy used to manufacture, and (b) the type of failure were recorded.

	failure		
Alloy	fatigue	corrosion	crash
2024 - 73	47	76	27
2024 - 74	41	42	17

Is the type of failure independent of the type of alloy?

Test of Homogeneity

- It can be used to test whether different populations have the same percentage of people with the same characteristic

- $r \geq 2$ populations are divided into $c \geq 2$ groups based on some characteristic and examine if the rate of each group is the same across all populations.

$\mathbf{H_0}$ The percentage of each group is the same across all populations

$\mathbf{H_1}$ The percentage of at least one group is not the same across all populations

Test of Homogeneity

Example:

A telecommunication company has 5 factories that operate with the same specifications. In order to check the company if there is not any difference in the offered quality between the factories, it got a sample of phones manufactured and submitted to check in order to see how many of them are defective. The collected data are shown in the table below.

Factory	defective,	not defective
A	20	80
B	7	153
C	8	152
D	6	74
E	19	201

Test if the percentage of defective telephones is the same across the five factories.

Test for several proportions

The chi-square statistic for testing homogeneity is also applicable for testing the hypothesis that k binomial parameters have the same value $H_0: \; p_1 = p_2 = \cdots = p_k$

Samples	1	2	...	k
successes	x_1	x_2	...	x_k
failures	$n_1 - x_1$	$n_2 - x_2$	...	$n_k - x_k$

- Compute the quantity

$$X^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} \sim X^2_{k-1} \text{ under } H_0$$

- If $X^2 > X^2_{k-1 \cdot \alpha}$, then H_0 is rejected at $\alpha\%$ significance level.

Nonparametric Tests

Sign Test

- $H_0: \quad \delta_X - \delta_Y = 0$ vs $H_1: \delta_X - \delta_Y \neq 0$

Test the equality of the medians of two continuous dependent random variables X and Y (paired samples)

Or equivalent

$$H_0 \quad p = 0.5 \text{ vs } H_1 \quad p \neq 0.5$$

where $p = P(X > Y)$.
Let

$$n_+ = \sum_{i=1}^{n} I(x_i > y_i), \;\; n_- = \sum_{i=1}^{n} I(x_i < y_i)$$

$R = \max\{n_+, n_-\}$

Under H_0, $R \sim \mathcal{B}(n, 0.5)$ thus $E(R) = n/2 \;\; Var(R) = n/4$.
The probability $P(R \geq r)$ can be computed.

If n is large then under H_0, $R \sim \mathcal{N}(n/2, n/4)$
In this case, the test statistic is

$$\frac{R - n/2 - 0.5}{\sqrt{n/4}} \sim \mathcal{N}(0.1)$$

Wilcoxon signed-rank test

- $H_0: \quad \delta_X - \delta_Y = 0$ vs $H_1: \quad \delta_X - \delta_Y \neq 0$

Wilcoxon signed-rank test step by step:
1. Compute the ranks of the absolute differences $|x_i - y_i|$ ignoring the cases where $x_i - y_i = 0$.
2. Compute the sums of the ranks S_p and S_n that correspond to positive and negative differences respectively.
3. If the sample size n is large, then

$$Z = \frac{\min\{S_n, S_p\} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0.1)$$

under H_0.

Wilcoxon signed-rank test

- $H_0: \quad \delta_X - \delta_Y = 0$ vs $H_1: \quad \delta_X - \delta_Y \neq 0$

Wilcoxon signed-rank test step by step:

1. Compute the ranks of the absolute differences $|x_i - y_i|$ ignoring the cases where $x_i - y_i = 0$.
2. Compute the sums of the ranks S_p and S_n that correspond to positive and negative differences respectively.
3. If the sample size n is large, then

$$Z = \frac{\min\{S_n, S_p\} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{L} \frac{t_i^3 - t_i}{48}}} \sim N(0.1)$$

under H_0.

L the number of cases that we have equal observations and t_i the number of observations with the same rank.

Wilcoxon signed-rank test

Example

To evaluate the results of the retraining seminar in engineers, 35 engineers were randomly selected and asked to answer a series of questions before and a month after the seminar. The scores are presented in the file entitled `seminar.sav`.

Test if there is a difference in the median performance of engineers before and after participation in the seminar.

Kruskal–Wallis Test

- $H_0: \quad \delta_{X_1} = \delta_{X_2} = \cdots = \delta_{X_k}$ and H_1 not H_0

Kruskal–Wallis Test step by step

1. Compute the ranks of the observations assuming the k samples as one.
2. Compute the sums of the ranks R_i, the number of observations n_i in each group and the quantity $T_i = t_i^3 - t_i$ in cases where they are observations with the same rank where t_i is the number of observations with the same rank.
3. If n_i's are large, i.e. $n_i > 5$.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \sim X_{k-1}^2$$

under H_0.
In the case that they are observations with the same rank, the above quantity is corrected as follows

$$H' = \frac{H}{1 - \sum_{i=1}^{L} \frac{t_i^2(t_i - 1)}{N^2(N-1)}}$$

where L is the number of cases that we have equal observations.

Kruskal–Wallis Test

Example:

The file `tires.sav` displays the results of laboratory measurements of the tread wear indicator (TWI) presenting the tires of three different companies when used under the same conditions.

Test, at a significance level of 5%, if there are differences between the tires of the three companies.

Run Test

- Randomness test

$\mathbf{H}_0$ The sample is random $\mathbf{H}_1$ The sample is not random

The random variable is of the form $\{0, 1\}$

- defective or not defective product.
- component type A or Type B
- a characteristic above or bellow a value

If the population is expressed by a binary variable then the sample is a run sequence.

Definition: In a symbol sequence of at least two types of symbols, run is called a sequence of the same symbol which is blocked by symbols of other type. The number of symbols of the run is called the length of the run.

If the random variable is continuous then the sample is a device of real numbers, but it can be converted to a run sequence.

A symbol is associated with each sample depending on whether or not it is greater than the median or the mean. Price equal to the median or the average is omitted.

Run Test

$$0, 0, 0, 1, 0, 1, 1, 0, \ldots$$

H_0 The sample is random **H_1** The sample is not random

If the sample is not random then there are trends expressed with symbol accumulations and either the number of runs is small but their length is large, or there are symbol recycling expressed by systematic variations of symbols and the number of runs is very large.

So according to the above we reject the null hypothesis, if the number of runs is very small or very large.

- Let R_0 and R_1 the number of 0 and 1 runs and n_0 and n_1 the number of the corresponding symbols in the sample.

H_0 is rejected at significance level $\alpha\%$ if

$$P(R \leq k_1) = \alpha/2 \text{ or } P(R \geq k_2) = \alpha/2$$

where $R = R_0 + R_1$,

k_1 and k_2 the corresponding values in the tables of this test.

$$\mu_R = E(R) = \frac{2n_0 n_1}{n_0 + n_1} + 1$$

$$\sigma_R^2 = Var(R) = \frac{2n_0 n_1 (2n_0 n_2 - n_0 - n_2)}{(n_0 + n_1)^2 (n_0 + n_2 - 1)}$$

$$Z = \frac{R - \mu_R}{\sigma_R} \sim N(0, 1)$$

▶ Example.

During a laboratory experiment, the friction coefficient between two metals was measured. The results are presented in the following table.

0.59	0.61	0.56	0.58	0.60	0.52	0.54
0.53	0.51	0.62	0.58	0.61	0.57	0.56
0.63	0.62	0.55	0.51	0.50	0.61	0.57
0.55	0.53	0.49	0.62	0.63	0.60	0.53

The head of the laboratory suspects that measurements shots were not correct and therefore wants to examine whether the measures can be considered random or not. With the help of the median, test if there are significant indications at significance levels $a = 0.05\%$ that the measurements can not be considered random.