SOCIETY. INTEGRATION. EDUCATION
Proceedings of the International Scientific Conference. Volume V, May 24th -25th, 2019. 374-382

SAMPLES DISTINCTION BY PARAMETRIC AND NONPARAMETRIC STATISTICS IN SPSS

Viktor Koshmak
Pskov State University, Russian Federation

Aleksandr Hvatcev
Pskov State University, Russian Federation

Ina Astahova
Pskov State University, Russian Federation

Aleksandr Zuev
Pskov State University, Russian Federation

Abstract. *Testing samples distinction is necessary in a wide range of practical tasks. Medicine, sociology, psychology, marketing - this is a short list of industries where it is required to conduct tests that establish effectiveness or inefficiency of a certain technology.*
Diversity of situations and techniques applied to sample distinction create a problem for compliance of testing procedures. The problem rises for tests including large and small samples (dependent or independent) with various distributions. The article proposes a list of problems created by testing differences between two samples. Limits of applicability of parametric and non-parametric tests are established based on selected distribution. Informative examples are included based on simulated data.
SPSS software was used for sample distinction tests. It is important to double-check the operation of the machine computing procedure "manually" to understand the nature of tests and in educational purposes. The article provides mathematical illustration for the algorithms used, which can be considered as supplementary information for SPSS help.
Keywords: *t-test, independent samples, paired samples, SPSS.*

Introduction

Testing of samples distinction is a well-studied and described in detail problem in the literature. The reference substantiating this statement is practically the entire bibliography of this article. Nevertheless, there are new publications on this issue (Luke, Corrine, & Ismail, 2012). And, due to the branching of testing cases, the task of systematization of existing methods, solved with examples of illustrations, is always relevant. This problem is solved in this article.

http://dx.doi.org/10.17770/sie2019vol5.3779

With minimal math justification, formulas are given for calculating the corresponding statistics. A more detailed description can be founded at IBM Knowledge Center. This is done in order to be able to repeat the calculations almost "manually" and is intended as an accompaniment of the course that is taught to students. It is assumed that the examples in this article will help in the conduct and understanding of the calculations carried out in SPSS (Official website SPSS).

Nature of the data

Two samples are tested - control and investigated. The main task is to determine whether the first and second samples belong to the same general population, i.e. whether the control and test samples are equally distributed. A condition equally distributed may relate to a certain type of distribution, for example, a normal distribution. If a distribution is established this can be related to its numerical characteristics such as expectation or variance for example.

Samples can be dependent or independent. The sample data is discrete or continuous. Discrete data can be nominal or ordinal. Total we get 12 terminal options for which certified procedures are established testing the hypothesis about the difference of samples. The very fact of the difference between the samples can be clarified by the nature of the difference, for example, the expectation is greater for the control sample. Or the variance of the investigated and control samples are equal.

There are some examples of testing the samples distinction in the paper. The data were modeled in SPSS. The main task of the generated examples is their reproducibility in any computing environment, including an ordinary calculator. This allows you to study the technique of testing the samples distinction, contributes to the understanding of the corresponding algorithm and useful in educational purposes.

Known distributions

Let the distributions of the control and investigated samples be known. There are F_{θ_x} and F_{θ_y} , where θ_x and θ_y parameter vectors. Then the task of testing the samples distinction is reduced to testing the hypothesis H_0: $\theta_x = \theta_y$.

Independent samples

T-tests

To conduct a t-test, it is necessary that the data obey the normal distribution. Or the sample should be large, and its parameter estimates are asymptotically normal.

Assume that the control sample: $X_1, X_2, ..., X_{n_x}; X_i$ obeys the normal distribution with parameters μ_x, σ_x. Shortly $X_i \sim N(\mu_x, \sigma_x^2)$. Investigated sample is normal too: $Y_1, Y_2, ..., Y_{n_y}; Y_j \sim N(\mu_y, \sigma_y^2)$. Null hypothesis: H_0: $\mu_x=\mu_y$. The decision: “There is no reason to reject the null hypothesis” is taken according to the p-value, at a given level of significance α. Hypothesis H_0 is not rejected if $p>\alpha$. When H_0 is true

$$Z = (\overline{X} - \overline{Y}) / \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \sim N(0,1) \qquad (1)$$

The variances of control and investigated samples are unknown as rule. If variances are equal $\sigma_x^2 = \sigma_y^2 = \sigma^2$ then statistic

$$\chi^2 = \frac{(n_x - 1)S_x^2}{\sigma^2} + \frac{(n_y - 1)S_y^2}{\sigma^2} \sim \chi^2(n_x + n_y - 2), \qquad (2)$$

where $S_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (X_i - \overline{X})^2$,

$S_y^2 = \frac{1}{n_y - 1} \sum_{j=1}^{n_y} (Y_j - \overline{Y})^2$,

$\chi^2(k)$ - is χ^2 distribution with k degrees of freedom.

Dividing (1) by the root of (2), normalized to the number of degrees of freedom gives us Student statistics with n_x+n_y-2 degrees of freedom (Gosset [Student, pseud.], 1908):

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}} \cdot \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}} \sim T(n_x + n_y - 2). \qquad (3)$$

If the variances are not equal, then the Student statistics is

$$T = (\overline{X} - \overline{Y}) / \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \sim T(k), \qquad (4)$$

where degrees of freedom k is calculated by the formula (Satterthwaite’s, 1946)

$$k = \left(S_x^2 / n_x + S_y^2 / n_y\right)^2 / \left(\frac{\left(S_x^2 / n_x\right)^2}{n_x - 1} + \frac{\left(S_y^2 / n_y\right)^2}{n_y - 1} \right).$$

In table 1 are shown the data and calculation of Student statistics for the case of equal and different variances.

Table 1 ***T-test for equal and unequal variances***
(simulated data, rounded to integer, p-value two-sided)

$\sigma_x^2 = \sigma_y^2$	X = rnd(RV.NORMAL(50,10))								Y = rnd(RV.NORMAL(60,10))						
i	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7
Var	50	64	48	47	57	40	50	61	46	57	52	57	49	71	74
Gr_Var	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
$\overline{X}$	52.125			S_x^2	63.268				$\overline{Y}$	58		S_y^2	114.667		
$\overline{X} - \overline{Y}$	-5.875			T	-1.217				p	0.245					
$\sigma_x^2 \neq \sigma_y^2$	X = rnd(RV.NORMAL(50,10))								Y = rnd(RV.NORMAL(60,15))						
i	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7
Var	51	47	58	51	51	53	38	26	72	76	66	45	53	95	50
Gr_Var	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
$\overline{X}$	46.875			S_x^2	10.190				$\overline{Y}$	65.286		S_y^2	17.509		
$\overline{X} - \overline{Y}$	-18.411			T	-2.443				k	9.377		p	0.0372		

Paired or matched samples

Paired or matched samples are a special case of dependent samples. Paired samples are appeared when the same object is measured twice, before and after exposure. For example, weight before and after the diet. Or two different objects have agreed characteristics. For example, visual acuity of the left and right eyes of the same person. For paired samples, the control and investigated variables have the same number of observations (n_x=n_y=n). With the validity of the hypothesis H_0 the average deviation

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \sim N(0, \sigma^2 / n), \qquad (5)$$

where σ^2 - variance of differences $D_i = X_i - Y_i$.

For deciding on the hypothesis calculated Student statistics

$$T = \frac{\overline{D}}{\sqrt{S_d^2 / n}} \sim T(n-1), \qquad (6)$$

where $S_d^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \overline{D})^2$ sample variance of differences.

Table 2 presents an example of paired samples, and t-test results. In accordance with the data of Table 2, there is no reason to reject the null hypothesis that samples are different.

Table 2 ***T-test for paired and matched samples***
(simulated data, rounded to integer, p-value two-sided)

X = rnd(RV.NORMAL(50,10))										
Y = X + rnd(RV.NORMAL(2,5))										
i	1	2	3	4	5	6	7	8	9	10
X_i	60	52	48	36	53	58	39	43	62	28
Y_i	70	62	48	36	52	54	41	49	60	20
D_i	-10	-10	0	0	1	4	-2	-6	2	8
$\overline{D}$	-1.3		S_d^2	34.233		T	-0.703		p	0.500

Nonparametric tests

When conducting non-parametric tests, the differences of the samples allowed arbitrary distributions of the control and test samples. The null hypothesis is that the distributions are equal. Most often, non-parametric tests are applied to relatively small samples.

Independent observations

Mann and Whitney U-test

The Mann and Whitney test or the U test refers to ranking criteria. To carry out a rank test, it is necessary to combine the X and Y samples. Then arrange the resulting sequence. Its elements are numbered 1, 2,...,n_x+n_y. If all values of a sequence are different (not the same), then the rank is equal to the element number of the sequence. Consecutive identical elements are assigned a rank equal to the arithmetic mean of their numbers. We obtain a sequence of ranks R_i. The belonging of X and Y in the combined sequence is fixed by additional grouping variable with two categories. The first category in this variable belongs to the control group, and the second to the investigated group.

Since, with the validity of the hypothesis H_0: $F_x = F_y$, all combinations of ranks are equally likely, the significance level of the rank criterion does not depend on the distribution of X and Y. The sum of the ranks X gives us statistics (Wilcoxon, 1945)

$$W = \sum R_i^x, \qquad (7)$$

where R_i^x – are ranks for testing variable X.

Too small W values indicate that $F_x<F_y$. Too big on $F_x>F_y$. U statistics is calculated using the formula (Mann & Whitney, 1947)

$$U = W - n_x(n_x + 1)/2. \qquad (8)$$

Z statistics is calculated by the formula

$$Z = \frac{W - E(W)}{\sqrt{Var(W)}} \xrightarrow{H_0} N(0,1), \qquad (9)$$

where $E(W) = n_x(n+1)/2$,

$n = n_x + n_y$,

$Var(W) = n_x n_y S_R^2 / n$,

$S_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R})^2$.

Moses extreme reactions test

Similar to the U-test, the Moses test is a ranking criterion. But if using the U test, systematic offsets are estimated. When using the Moses test, it is possible to estimate the multidirectional offsets of the tested variable. The Moses test is used for extreme responses compared to a control group. To carry out the test you combine samples X and Y and arrange the resulting sequence. Values of the combined sequence are assigned the ranks of R. Statistics is calculated (Moses, 1952)

$$SPAN = round(\max R_i^x - \min R_i^x + 1), \qquad (10)$$

where round(z) – is rounded z to nearest integer,

R_i^x – are ranks of control variable X.

The X variable outliers can distort the test results. To exclude it SPAN statistics are also calculated with the “Outliers Trimmered from each End” option. In this option the upper and lower 5% quantiles are cut off from the X variable. After which the ranks are assigned again and the SPAN is calculated from (10).

Wald-Wolfowitz runs test

During the Wald-Wolfowitz test, samples X and Y are also combined. The resulting sequence is ordered with the group variable indicating X or Y. For example, if X, then 0, and if Y, then 1, as in table 3. The concept of different runs is introduced. These are consecutive zeroes, or ones. Different runs have different lenght. For example run with lenght 2 means 00 or 11. Length of run can be equal to 1. This means that after X immediately follows Y, or vice versa. Then is calculated (Wald & Wolfowitz, 1940)

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}} \xrightarrow{H_0} N(0,1), \qquad (11)$$

where R - is equal a sum of runs,

$$E(R) = \frac{2n_x n_y}{n} + 1,$$

$$n = n_x + n_y,$$

$$Var(R) = \frac{(E(R)-1)(E(R)-2)}{n-1}.$$

If number of samples n<50 and $|R - E(R)| \geq 0.5|$ equation (11) is calculated with corrector (IBM Knowledge Center)

$$Z = \frac{R - E(R) + 0.5}{\sqrt{Var(R)}}, \qquad (12)$$

when $|R - E(R)| < 0.5|$ and $n \geq 50$ then Z=0.

Table 3 presents examples of Mann and Whitney U-test, Moses extreme reactions test and Wald-Wolfowitz runs test. All tests show that there is no reason to reject the null hypothesis that samples are different.

Table 3 ***Rank tests***
(simulated data from Table 1, rounded to integer, p-value two-sided)

	X = rnd(RV.NORMAL(50,10))								Y = rnd(RV.NORMAL(60,10))							
i	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	
Var	50	64	48	47	57	40	50	61	46	57	52	57	49	71	74	
Gr_Var	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
Sorted	*Mann and Whitney U-test*															*Sum*
Var	40	46	47	48	49	50	50	52	57	57	57	61	64	71	74	
Gr_Var	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1	
Num	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	120
R_i	1	2	3	4	5	6.5	6.5	8	10	10	10	12	13	14	15	120
R_i^x	1	0	3	4	0	6.5	6.5	0	10	0	0	12	13	0	0	56
R_i^y	0	2	0	0	5	0	0	8	0	10	10	0	0	14	15	64
W	64		E(W)			36			U	20						
S_R^2	19.821		Var(W)			74			Z	-0.930			p	0.352		
Moses extreme reactions test																
max R_x^i	13		min R_i^x			0		Span		13			p	0.446		
Sorted	*Wald-Wolfowitz runs test*															*Sum*
Var	40	46	47	48	49	50	50	52	57	57	57	61	64	71	74	
Gr_Var	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1	
Runs	R_1	R_1	R_2		R_1	R_2		R_1	R_1	R_2		R_2		R_2		
R	1	1	1		1	1		1	1	1		1		1		10
E(R)	8.467		Var(R)			3.449		Z	1.095				p	0.863		

Kolmogorov-Smirnov Z-test

The Kolmogorov-Smirnov test is based on the maximum absolute difference between the observed cumulative distribution functions for both samples. When this difference is significantly large, the two distributions are considered different. For the test, the sample distribution functions of random variables X and Y are calculated. This is $\hat{F}_x(t)$ and $\hat{F}_y(t)$. Then the maximum difference between them is determined (Kolmogorov, 1933)

$$D = \sup_t \left| \hat{F}_x(t) - \hat{F}_y(t) \right|. \quad (13)$$

During the test, statistics are calculated (Smirnov, 1933)

$$Z = \sqrt{\frac{n_x n_y}{n_x + n_y}} \cdot D \quad (14)$$

Paired or matched samples

Wilcoxon signed ranks test

The null hypothesis of the test is that the distribution $F_x = F_y$. With the same distributions of X and Y, the distribution of differences X_i-Y_i is symmetrical. During the test, the differences $D_i = X_i - Y_i$ are calculated. The sequence of absolute values of differences $|D_i|$ is being ordered. Members of $|D_i|$ are assigned numbers and ranks starting from D_i that are not equal to zero. The ranks are the same as the numbers if all members of the sequence are not equal. Or the ranks are equal to the arithmetic mean of the numbers for the matching members. The sequence of ranks R_i is converted to a sequence

$$r_i = \begin{cases} -R_i, & \text{if } D_i < 0; \\ 0 \quad , & \text{if } D_i = 0; \\ R_i, & \text{if } D_i > 0. \end{cases} \quad (15)$$

A statistic is computed that is asymptotically normal when the null hypothesis is valid (Sprent & Smeeton, 2007, 72)

$$Z = \frac{S^+ - E(S^+)}{\frac{1}{2}\sqrt{\sum_{i=1}^{n} r_i^2}} \xrightarrow[n \to \infty]{} N(0,1), \quad (16)$$

where $S^+ = \sum_{r_i > 0} r_i$ - sum of positive ranks; $E(S^+) = \sum |r_i|$.

Summary

SPSS offers a wide range of statistical calculations. The size of the article does not allow presenting even a small part of these possibilities with proper quality. The article describes only one direction of statistical research. This is testing the difference of two samples or two groups. A description is given of the corresponding test algorithms for independent and paired samples with a normal distribution. Algorithms for conducting non-parametric tests included in SPSS in case of unknown distributions are also described. The above algorithms are illustrated by examples with simulated data. Sample size allows you to repeat the calculation manually. It helps to remember the work of the test and contributes to the understanding of the material.

References

IBM Knowledge Center. https://www.ibm.com/support/knowledgecenter

Gosset [Student, pseud.], W.S. (1908). The probable error of a mean. *Biometrika, 6,* 1–2.

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari 4*, 83–91.

Luke, K., Corrine, M., & Ismail, W. (2012). Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity *American Journal of Political Science, 56,* 2, 484-499.

Mann, H.B., & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60.

Moses, L.E. (1952). Non-parametric statistics for psychological research. *Psychological Bulletin, 49*, 122-143.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2,* 110–114.

Smirnov, N.V. (1933). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University, 2,* 3–16.

Sprent, P., & Smeeton, N. (2007). *Applied Nonparametric Statistical Methods.* 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

Official website SPSS. https://www.ibm.com/products/spss-statistics

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math Statist., 11,* 147-162.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80–83.