Humboldt-Universität zu Berlin

Sprach- und literaturwissenschaftliche Fakultät

Institut für deutsche Sprache und Linguistik

Übung: Einführung in die Statistik mit SPSS

Modul 2: Methoden der Datenerhebung

Sommersemester 2017

Seminarleiterin: Prof. Dr. Pia Knoeferle

HAUSARBEIT

Statistische Datenanalyse mit SPSS

Verfasserin: Ana Lorena Ciutacu

Matrikelnummer: 584413

E-Mail: ciutacua@hu-berlin.de

Berlin, am 20. Juli 2017

Table of contents

Introduction

The aim of this term paper is to illustrate the process of statistical data analysis in experimental research, by analyzing three experimental studies using IBM SPSS Statistics 23. The first two experiments, presented in chapters 2 and 3 are related, in that they examine the same topic (math skills of women and men) and are analyzed with the *t*-test. Chapter 2 presents a between-subjects experiment in which men's and women's results on a math test are analyzed with the independent *t*-test. Chapter 3 presents a within-subjects experiment with repeated measures, in which women's results on two math tests are analyzed with the dependent *t*-test. Chapter 4 presents a within-groups experiment, in which the level of sportiness of people from three age groups is analyzed with one-way ANOVA. The results of the three experiments are summarized in chapter 5.

(21.270 characters)

1. Experiment 1

This chapter includes the analysis of the experiment described in the handout *Practice session III – Part 1 (1).*

1.1. Abstract

This experiment aimed to determine whether males and females differ in their performance on a math test. To test this, two groups of men and women completed a math test for which they received credits from 0 to 100. Women obtained higher results compared to men, however this difference was not significant.

1.2. Introduction

Gender stereotypes are still present in many domains, one of which is knowledge of mathematics. The aim of this study is to test empirically whether men and women differ in math skills.

1.3. Method

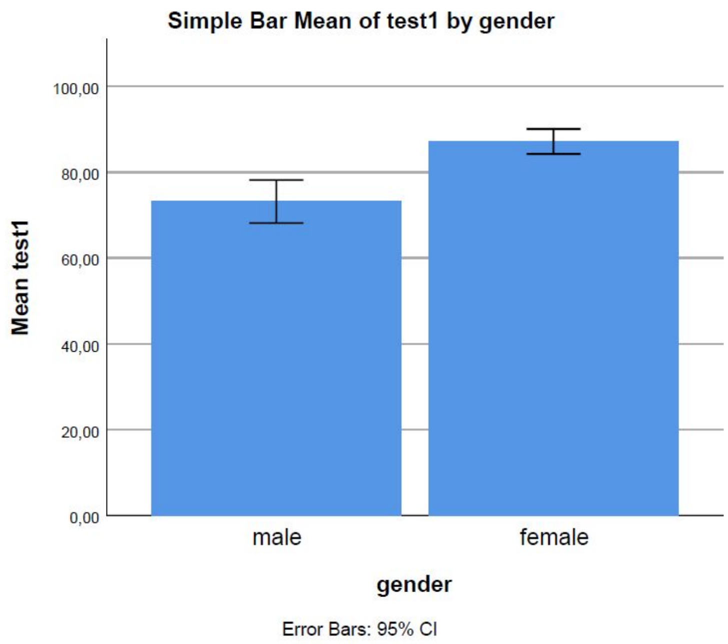
Participants: 40 students (20 females, 20 males).

Materials: One test of mathematics.

Design: The experiment was designed as between-group, with one independent variable on nominal scale (gender) and one dependent variable on metric scale (test credits).

Procedure: Participants had to complete a math test for which they received credits depending on their performance, from 0 (lowest) to 100 (highest).

Analysis: First, the mean credits of men and women were compared. Figure 1 shows that women obtained on average higher credits ($M = 87.15$, $SE = 6.20$) than men ($M = 73.15$, $SE = 10.75$). The error bars indicate the confidence interval



(CI) from the mean (i.e., the limits within which the population mean is likely to fall). The CIs of different means indicate whether the means are from the same population or different populations (Field 2009:47). In this case, the mean of men's credits has a CI of 68.11 to 78.18, and the mean of women's credits has a CI of 84.24 to 90.05. The error bars of the two means do not overlap, so the means come from the different populations.

In order to establish whether the mean credits differ significantly for men compared with women, the means were compared using the independent *t*-test, since all assumptions for this test were met:

- The independent variable is nominal, the dependent variable is metric, measured at interval level.
- The scores come from independent groups (error bars do not overlap).
- Variances are not significantly different, according to Levene's test for equality of variances: $F = 3.74$, *Sig*. = .06 (Table 1).
- The sampling distribution is normally distributed for both groups, as indicated by the non-significant Kolmogorov-Smirnov test $D(40) < .13$, $p > .19$, and Shapiro-Wilk test $W(40) < .98$, $p > .05$, and as see in Figure 2 and Figure 3.

Table 1

Independent Samples Test

		Levene's Test for Equality of Variances		*t*-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Lower	95% Confidence Upper
Test 1	Equal variances assumed	3.748	.060	-5.043	38	.000	-14.00000	2.77636	-19.62044	-8.37956
	Equal variances not assumed			-5.043	30.405	.000	-14.00000	2.77636	-19.66692	-8.33308

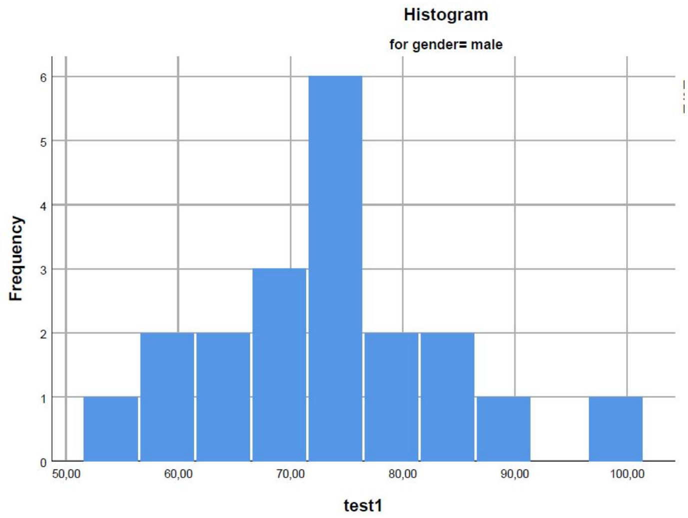


Figure 2

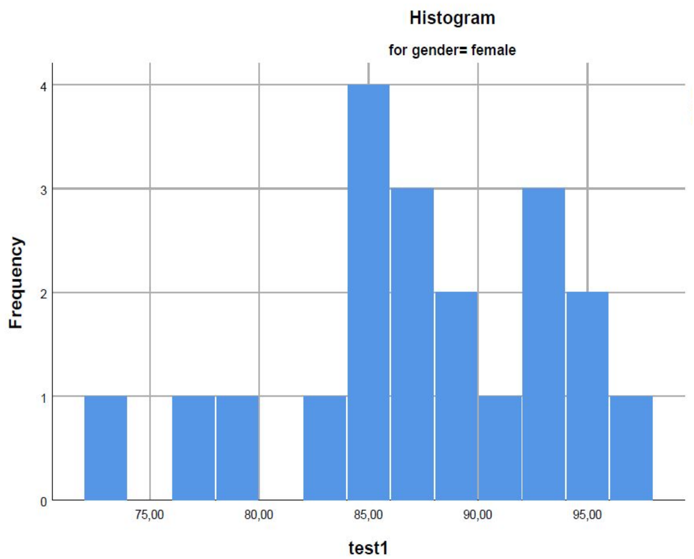


Figure 3

1.4. Results

The *p*-value is .06, thus not significant, and $t = -5.04$. Overall, on average women obtained higher scores (M = 87.15, SE = 6.20) than men (M = 73.15, SE = 10.75) on the math test, but this difference was not significant: $t(38) = -5.04$, $p > .05$.

1.5. Discussion

The results of this study do not support the gender stereotypes in mathematic skills, since the difference between the scores of women (M = 87.15, SE = 6.20) and men (M = 73.15, SE = 10.75) was not significant. However, this study has several limitations. First, the sample of 20 participants is small and the results cannot be applied to the general population. Second, the degree (BA/MA/Bsc/Msc) and field of study (e.g., Physics, Visual Arts) of the students have not been taken into consideration, though they are relevant for math skills (e.g., students of scientific degrees would have better knowledge of mathematics due to their specific training, compared to arts students). Third, general intelligence would also influence the performance on the test and this factor has not been taken into consideration. Controlling for the above-mentioned factors in a larger sample is important for future studies investigating gender differences in math skills.

2. Experiment 2

This chapter includes the analysis of the experiment described in the handout *Practice session III – Part 1 (2).*

2.1. Abstract

This study examined the influence of math lessons on student's performance on a test. For this, a group of female students took two math tests within a week: one on Monday, with little preparation, and another one on Friday, after additional lessons. Students scored significantly better on the first test than on the second one.

2.2. Introduction

Generally, the amount of lessons on a specific task improves the skills in the respective domain. The aim of this study is to examine whether additional math lessons improve students' in tests.

2.3. Method

Participants: 20 female students.

Materials: Two different math tests.

Design: The experiment was designed as within-participant, with two independent variables on nominal scale (gender, exam day), one dependent variable on metric scale (test credits).

Procedure: Participants were asked to complete two math tests within one week: one on Monday, after the students had had little training, and the second one on Friday, after they had had eight additional math lessons as preparation for the test. On each test, participants received credits from 0 (lowest) to 100 (highest).

<u>Analysis:</u> First, the results of the participants were adjusted, to eliminate variation due to individual differences (e.g., IQ, age). The mean credits of the participants on both tests were compared using the dependent *t*-test, based on the following assumptions:

- The independent variable is nominal, the dependent variable is metric measured at interval level.
- Variances are not significantly different.
- The sampling distribution is normally distributed in both tests, as indicated by the non-significant Kolmogorov-Smirnov test $D(20) < .13$, $p > .19$, and Shapiro-Wilk test $W(20) < .97$, $p > .05$.

2.4. Results

Participants obtained significantly higher scores on the first test (M = 93.14, SE = 2.15) compared with the second test (M = 81, SE = 2.15) (Figure 4, Table 2).

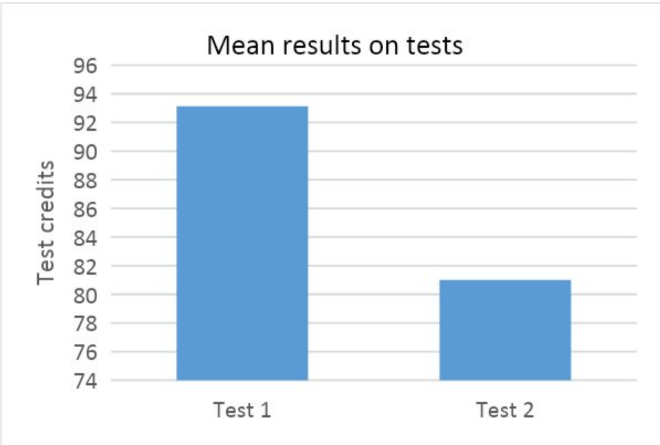


Figure 4

Table 2
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	test 1	93.14	20	9.62	2.15
	test 2	80.99	20	9.62	2.15

The Pearson correlation indicates that the scores in the two tests are significantly correlated ($r = -1$, $p < .05$) (Table 3).

Table 3
Paired Samples Correlations

	N	Correlation	Sig.

Pair 1 test1 & test2	20	-1.00	.000

The mean difference between the two experimental conditions is 12.15 and this difference was significant $t(19) = 2.83$, $p < .05$. The positive value of t indicates that the first condition had a higher mean value than the second. The CI shows that in 95 of the possible samples to be drawn, the true mean lies between 3.15 and 21.15 (Table 4).

Table 4

Paired Samples Test

				95% CI of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1 test1-test2	12.15	19.23	4.300	3.15	21.15	2.826	19	.011

2.5. Discussion

Surprisingly, participants in this experiment obtained significantly higher results on the first test with little preparation (M = 93.14, SE = 2.15) than on the second test after additional math lessons (M = 81, SE = 2.15), $t(19) = 2.83$, $p < .05$. However, this study has similar limitations to the first experiment. A possible factor is the time of day when the tests took place (e.g., early morning or evening), which could have influenced students' motivation and performance on the test. This should be considered in future studies.

3. Experiment 3

This chapter includes the analysis of the experiment described in the handout *Practice session III – Part 2.*

3.1. Abstract

This experiment examined the level of sportiness of individuals from three age groups: young, middle-aged, and old. Thirty participants (10 from each age group) rated their level of sportiness in the past year on a scale from 1 to 10. Age was found to have a significant effect on sportiness levels, with the youngest individuals having significantly higher levels than the older groups, but this difference was not significant between the middle-aged and the old individuals.

3.2. Introduction

Generally, the level of physical activity decreases with age. To evaluate this assumption in the general population, individuals from different ages (young, middle aged, old) were asked to rate their general sportiness during the past year on a scale from 1 (lowest) to 10.

Null hypothesis (H0): Mean level of sportiness is the same for all age groups.

Alternative hypothesis (H1): Mean level of sportiness is not the same for all age groups.

Alternative hypothesis (H2): There is a difference in sportiness levels between the young group and the old group.

3.3. Method

Participants: 30 subjects (10 young, 10 middle-aged, 10 old).

Materials: Online survey including questions about age and self-evaluation of physical activity on a scale from 1 to 10, where 10 represents the highest level of sportiness.

Design: The outcome (dependent) variable is the level of sportiness, measured at level from 1 to 10. The grouping factor (independent variable) is age, at three levels (young, middle-aged, old). The middle-aged group is the control group. The experiment has one condition.

Procedure: Participants from each age group rated their general sportiness during the past year on a scale from 1 to 10.

Analysis: Since the experiment has one factor and one outcome variable with three conditions, a one-way ANOVA was conducted. The assumptions of ANOVA (i.e., under which the *F* statistic is reliable) are the same as for all parametric tests (Field 2009:359):

- The dependent variable (in this case, the level of sportiness) is metric, measured on an interval scale;
- Homogeneity of variances: the variances in each experimental condition should be equal;
- Independence of observations: the samples should be independent;
- Normality: The distributions within groups should be normally distributed.

Therefore, the selected options for ANOVA in SPSS are (Field, 2009:379):

- Descriptive, to obtain the means, standard deviations, standard errors, ranges and confidence intervals of each mean, which assist in interpreting the results.
- Homogeneity of variance, to test this assumption.
- Brown-Forsythe and Welch, which are alternative versions of the F-ratio, useful in case the assumption of homoscedasticity is broken.
- Means plot, to obtain a line graph of the group means before the analysis.

- Exclude cases analysis by analysis, to exclude all cases that have a missing value for either the independent or the dependent variable used in this analysis.

The *F*-ratio of ANOVA indicates whether the model fitted to the data accounts for more variation than extraneous factors, but it does not specify where the differences between groups lie (Field 2009:360). Therefore, it is necessary to run planned comparisons or post-hoc tests, which consist of pairwise comparisons that are designed to compare all different combinations of the treatment groups 2009:372). In this case, a post-hoc test was run, because there were no hypotheses prior to collecting the data. The following post-hoc tests were selected:

- R-E-G-W-Q, which controls the Type I error rate, has good statistical power, and is used when the sample sizes are equal and population variances are similar;
- Tukey, which is similar to R-E-G-W-Q, but with less statistical power;
- Dunnett, which controls both experimental groups against the control group and controls Type I error rate. The parameters in SPSS were:
  - Control category: First (i.e., the control group is coded first);
  - Test: 2-tailed (because Hypotheses 1 and 2 are two-tailed);
- Games-Howell, which has good statistical power and is run if the population variances might not be similar (Field 2009:375).

A planned comparison was run as well for practice reasons, depicted in Table 5 In this test, the groups to be compared are coded with positive or negative values (weights), so that groups coded with positive weights are compared against groups coded with negative weights, and the sum of weights for a comparison should be 0 (Field 2009:365).

In Contrast 1, the control group (middle-aged) was compared to the two test groups (young and old). The groups were coded with values so that their sum equals 0. The control group was coded first (because it was the group coded as 1 in the variables list) as -2. Next, the young group (coded as 2 in the variables list) was coded as 1. Finally, the old group (coded as 3 in the variables list) was coded as 1.

In Contrast 2, only the two test groups (young and old) were compared. Like in Contrast 1, the groups were coded in the same order and with values so that their sum equals 0. First, the control group was coded as 0 (to mark that it is excluded from the calculation), then the young group was coded as 1, and finally the old group was coded as -1.

Table 5

Group	Contrast 1	Contrast 2
Middle-aged	-2	0
Young	1	1
Old	1	-1

3.4. Results

First, the mean of sportiness by age group was observed. The line graph depicted in Figure 5 indicates that the mean of sportiness is different between the three groups, namely the young group shows the highest level of sportiness (M = 7.80, SD = 1.317), followed by the middle-aged group (M = 5.20, SD = 1.54), and lastly the old group (M = 2, SD = .816). The error bars do not overlap, which indicates that there are between-group differences.

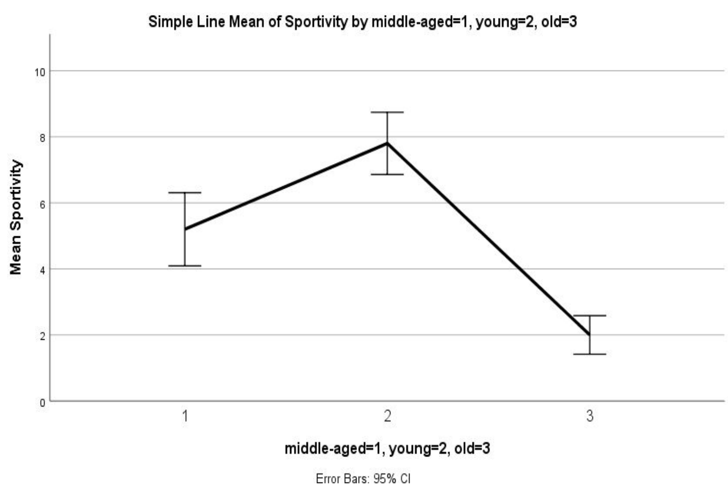


Figure 5

The assumption of homoscedasticity is met, as indicated by Levene's test *Sig*. > .05 (Table 6)

Table 6

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Sportivity	Based on Mean	1,812	2	27	,183
	Based on Median	1,527	2	27	,235
	Based on Median and with adjusted df	1,527	2	22,821	,239
	Based on trimmed mean	1,823	2	27	,181

Table 7 shows the ANOVA summary table. *Within Groups* shows the unsystematic variation within the data (Field, 2009:383), due to natural individual differences indicated by the residual sum of squares (SS_R = 43.20) and the average amount of unsystematic variation is indicated by the mean squares (MS_R = 1.60). The between-group effect labeled *Combined* is the overall experimental effect, represented by the sum of squares for the model (SS_M = 168.80) and mean squares (MS_R = 84.40) The *F*-ratio for the combine between-group effect indicates whether the group means are the same ($F(2)$= 52.75). The likelihood of an *F*-ratio of 52.75 to occur if there was no effect in the population is 0 (*Sig*. < .001), which means that the level of sportiness is significantly different between the age groups. For the linear trend (*Linear Term*, tests whether the means increase across groups in a linear way) $F(1) = 32$, *Sig*. < .001, which indicates that the level of sportiness decreases with age. For the quadratic trend (*Quadratic Term*, tests whether the pattern of trends is curvilinear) $F = 73.50$, *Sig*. < .001, so the contrast is significant (Field, 2009:383). Thus, the null hypothesis is rejected.

Table 7

ANOVA

Sportivity

		Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)	168,80	2	84,40	52,75	,000

			Sum of Squares	df	Mean Square	F	Sig.
	Linear Term	Contrast	51,20	1	51,20	32,00	,000
		Deviation	117,60	1	117,60	73,50	,000
	Quadratic Term	Contrast	117,60	1	117,60	73,50	,000
Within Groups			43,20	27	1,60		
Total			212,00	29			

Table 8 reports the contrasts, where *Assume equal variances* represents the value of the contrasts. This table is considered because the assumption of homogeneity of variance was met. For Contrast 1 (control group vs. test groups) *Sig*. (2-tailed) = .545, which means that the levels of sportiness do not differ significantly between middle-aged individuals and old and young individuals. For Contrast 2 (young vs. old) *Sig*. (2-tailed) < .001, which means that sportiness differs significantly between old and young individuals.

Table 8

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Sportivity	Assume equal variances	1	-,60	,980	-,612	27	,545
		2	5,80	,566	10,253	27	,000
	Does not assume equal variances	1	-,60	1,095	-,548	13,555	,593
		2	5,80	,490	11,839	15,031	,000

The post-hoc tests (Table 9) Tukey HSD and Games-Howell indicate that all groups differ significantly from each other (*Sig*. < .05). However, Dunnett’s test indicates that the old group and middle-aged group do not differ significantly (*Sig*. = 1), which is in contradiction with the results in the planned contrasts.

Table 9

Multiple Comparisons

Dependent Variable: Sportivity

	(I) middle-aged=1, young=2, old=3	(J) middle-aged=1, young=2, old=3	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	1	2	-2,600*	,566	,000	-4,00	-1,20
		3	3,200*	,566	,000	1,80	4,60
	2	1	2,600*	,566	,000	1,20	4,00
		3	5,800*	,566	,000	4,40	7,20

	3	1	-3,200*	,566	,000	-4,60	-1,80
		2	-5,800*	,566	,000	-7,20	-4,40
Games-Howell	1	2	-2,600*	,643	,002	-4,24	-,96
		3	3,200*	,554	,000	1,75	4,65
	2	1	2,600*	,643	,002	,96	4,24
		3	5,800*	,490	,000	4,53	7,07
	3	1	-3,200*	,554	,000	-4,65	-1,75
		2	-5,800*	,490	,000	-7,07	-4,53
Dunnett t (>control)[b]	2	1	2,600*	,566	,000	1,47	
	3	1	-3,200	,566	1,000	-4,33	

*. The mean difference is significant at the 0.05 level.

b. Dunnett t-tests treat one group as a control, and compare all other groups against it.

In Table 10, Tukey's test and the R-E-G-W-Q test are two tests that display subsets of groups that have the same means. In each subset only one group is displayed, which means that its mean is different than the means of the other two groups. The groups in the subsets have unequal means (*Sig*. = 1), so there are no two groups that are similar. The *Harmonic Mean Sample Size* represents the relation between variance and sample size, in order to reduce bias that might arise due to unequal sample sizes (Field, 2009:387).

Table 10

Homogeneous Subsets

Sportivity

	middle-aged=1, young=2, old=3	N	subset for alpha = 0.05 1	2	3
Tukey HSD[a]	3	10	2,00		
	1	10		5,20	
	2	10			7,80
	Sig.		1,000	1,000	1,000
Ryan-Einot-Gabriel-Welsch Range	3	10	2,00		
	1	10		5,20	
	2	10			7,80
	Sig.		1,000	1,000	1,000

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 10,000.

3.5. Discussion

The data provided statistically significant evidence that mean levels of sportiness are significantly different across age groups and there was a significant effect of age on levels of sportiness, $F(2) = 52.75$, $p < .05$. There was a significant linear trend ($F(1) = 32$, $p < .05$), indicating that the level of sportiness decreases with age: the older, the lower physical activity. The level of sportiness of the young group was significantly higher than that of than the old group, $t(27) = 10.25$, $p < .05$, whereas the level of sportiness of the middle-aged group was not significantly higher than that of the old group, $t(27) = -.612$, $p > .05$. However, this study has several limitations. First, the sample size in each group (N = 10) is small, so the results cannot be generalized. Second, the results are based on self-reported data, which allows for over-/underestimates of individual sportiness levels by the participants. These factors should be controlled for in future studies investigating the effect of age on sportiness.

4. Conclusion

In the previous chapters, three studies with different experimental designs have been analyzed in IBM SPSS Statistics 23. Experiment 1 yielded no significant difference between the results of men and women on a math test; experiment 2 yielded a significant difference between the results of female participants on two math tests, with better results obtained on the first test with little preparation; experiment 3 yielded a significant difference in the levels of sportiness of individuals from three age groups, with youngest individuals having the highest levels. These examples illustrated the process of statistical data analysis in experimental research and the interpretation of the results, necessary for drawing correct conclusions.

References

Field, A. (2009). Discovering statistics using IBM SPSS Statistics (3rd ed.). London: Sage Publications Ltd.