Statistics for EMCL students

Sample Test (slightly more questions than tomorrow)

Tamas Biro

1. Basic concepts: which statement is true?

Multiple choice:

1.1. The *inter-quartile range* of a dataset is given by...
   a. the difference of the median and the minimum value.
   b. the difference of the third quartile and the first quartile.
   c. the difference of the maximum value and the minimum value.
   d. the difference of the maximum value and the mean.

1.2.A *resistant measure* is a measure that...
   a. will not change (or not change much) if we slightly change some of the values or remove outliers.
   b. will not change (or not change much) if we take a slightly different sample.
   c. will not change (or not change much) if we slightly change the scale.
   d. will not change (or not change much) if we slightly change the variable.

1.3. Two variables are associated *if and only if...*
   a. changing one variable causes a change in the other variable.
   b. knowing the value of one variable helps us predict the value of the second variable.
   c. an increase in one of the variables is typically linked to an increase in the second variable.
   d. correlation r^2 is close to 0

1.4. A Normal quantile plot is useful in checking...
   a. the correlation of two variables.
   b. whether the values of one variable are distributed approximately Normally.
   c. whether the cases of one variable have percentiles that are suitable to be used in an ANOVA.
   d. whether outliers would influence the mean, median and standard deviation of the sample.

1.5. The sampling distribution of any statistic is...
   a. the distribution of the statistic if we draw all possible samples.
   b. a standard Normal curve, as proven by the Central Limit Theorem.
   c. the distribution of the parameter if we draw all possible samples.
   d. approximately a Normal curve (if n is large), as proven by the Central Limit Theorem.

1.6. A *z*-procedure can be used if
   a. the number of observations is larger than 30.
   b. the number of cases is much larger than 30.
   c. the standard deviation of the population is known.
   d. the standard deviation of the sample is known.

1.7. A 90% confidence interval ... a 95% confidence interval:
   a. is always larger than
   b. is always narrower than
   c. can be larger or narrower than
   d. is usually the same as

1.8. If $p = 0.048$ in an ANOVA, then
   a. the null hypothesis can be rejected at the 1% significance level.
   b. one of the population means is certainly different from the rest.
   c. the probability of rejecting a true null hypothesis is 0.048.
   d. the probability of accepting a false null hypothesis is 0.048.

1.8. You have just rejected the null hypothesis of ANOVA. How do you go on?
   a. I run pairwise *z*-tests.
   b. I use Bonferroni, for example with $\alpha=0.05$.
   c. I use a contrast, for example with coefficients 0.5, -0.5 and 1.
   d. I call my supervisor to help me out.

1.9. The density curve of a Normal distribution is...
   a. Assymetric and Gaussian.
   b. Unimodal and bell-shaped.
   c. Symmetric and bimodal.
   d. Skewed to the left.

1.10. The chi-square statistic...
   a. can get both positive and negative values.
   b. follow an F-distribution.
   c. is used to test the independence of two categorical variables.
   d. can be used both with one-sided and two-sided null hypotheses.

Please define the following concepts in one sentence each:

1.11. Histogram:
1.12. When can the chi-square test be applied?
1.13. Confidence interval with a confidence level *C*:
1.14. Informed consent:
1.15. Non-parametric test:

2. Paper-and-pen calculations

2.1. You have collected the following observations: 2, 7, -2, 1, 6, -1, 0, 3, 2.

Of this dataset give
2.1.1. the median:

2.1.2. the mode:

2.1.3. the mean:

2.1.4. the interquartile range:

2.1.5. the variance with *n*:

2.1.6. the standard deviation with *n-1*:

2.1.7. You would like to test if these observation originate from a population whose mean is μ $\mu = 4.5$. Calculate the t-statistic.

2.1.8. What degree of freedom are you going to use in a t-test?

2.2. Each subject in your experiment toss a coin 100 times and you count the number of heads. It is known that the number of heads follow a *binomial distribution*, which can be approximated with a Normal distribution N((μ = 50, σ = 5). You have 1000 subjects.

2.2.1. Approximately how many of your subjects will have more heads than tails?

2.2.2. What is the Normal z-score of 65 heads?

2.2.3. How many of your subjects to you expect to have at most 65 heads?

2.2.4. How many of your subjects to you expect to have at least 45 heads?

2.2.5. What is the score above which you only expect 25 people?

IMPORTANT:
Could you solve this assignment using the 68-95-99.7 rule, without a table?

3. Graphics and their interpretation
See assignments at the end of this file.

4. Tests (choice, interpretation)

4.1. Take the following two-way table:

	N	Prep	V	Adj
Spanish	4	6	7	6
Portuguese	6	9	14	5
Catalan	8	12	9	18

a. Describe what type of data are usually plotted in such a table.
b. You are interested in testing the independence of the two variables. What does it mean?
c. What test are you using? How do we call the statistic on which this test is based?
d. What is the sampling distribution of this statistic? In particular, what is the degree of freedom that defines the distribution?
e. Imagine that the observed counts and the expected counts differ very much. What would you preliminarily conclude?

4.2. You are conducting a survey to compare whether political preferences are different in the countryside from those in towns.

4.2.1. Describe a few possible mistakes to be avoided that you should consider while designing the experiment.
4.2.2. What is the working hypothesis you would like to test?
4.2.3. What statistical procedure are you going to perform? (NB: there are more possibilities!)
4.2.4. What is the null hypothesis of the statistical procedure(s) you have chosen?
4.2.5. What is the alternative hypothesis?
4.2.6. What do you have to check before running the statistical procedure?
4.2.7. Suppose you get p = 0.039. How do you proceed? What can be your conclusion?

(The following questions were originally composed by Eleonora Rossi)

Interpretation of tests

1. Given the following SPSS output:

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
	Statistic students	10	100.0%	0	.0%	10	100.0%
	Non statistic students	10	100.0%	0	.0%	10	100.0%

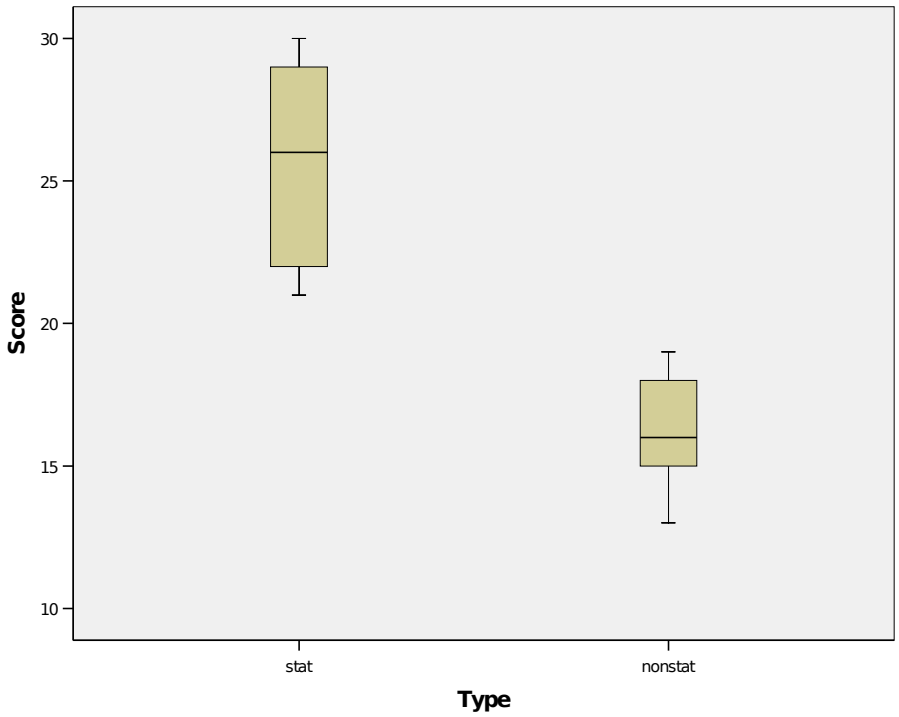
		Levene's Test for Equality of Variances							95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Score	Equal variances assumed	8.129	.011	8.019	18	.000	9.600	1.197	7.085	12.115
	Equal variances not assumed			8.019	13.797	.000	9.600	1.197	7.029	12.171

a. What do you suppose the researcher has confronted to give rise to this output?

b. Formulate the research questions that led to this test

c. Formulate an appropriate H_0 and H_a.

d. Looking at the following boxplots, what can you say about the two distributions?



e. Using the results given by the SPSS output (and using the given p-values), explain which of the two Hypotheses you have to accept, and which one you have to reject. Explain why.

f. What are your conclusions regarding the research question?

2. Given the following 2X2 table:

a. Complete the table by adding the row and column totals and the grand total.

b. Calculate the % of correctly produced verbs for both aphasic and SLI participants.

			verb_correct		Total
			correct	incorrect	
typeimpairment	aphasic	Count	67	35	
		Expected Count	64.2	37.8	
		% within verb_correct			
	SLI	Count	50	34	
		Expected Count	52.8	31.2	
		% within verb_correct			
Total		Count			

c. What kind of test do you run for this sort of data?

d. What will be the research question for this data set?

e. State the suitable H_0 and H_a

f. Given the following output, and given the given p-value, what are your conclusions? Which hypothesis can you confirm and which one will you disconfirm?

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.750(b)	1	.387		
Continuity Correction(a)	.509	1	.476		
Likelihood Ratio	.749	1	.387		
Fisher's Exact Test				.446	.238
Linear-by-Linear Association	.746	1	.388		
N of Valid Cases	186				

a Computed only for a 2x2 table
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.16.