Statistical and methodological entrance skills and knowledge

Upon entrance of the Research Master programme, the student is expected to have a sound basis on methodological and statistical principles and procedures as used in the social and behavioural sciences.

This implies active knowledge of a number of statistical procedures, both the theoretical aspects as experience in applying them. It is imperative that prospective students have experience with using statistical software for data analysis and can provide correct interpretations of the output of these software. We do not require proficiency for a specific software package.

An important, yet difficult to measure, skill is experience with and proficiency in reasoning under uncertainty. This includes a thorough understanding of different definitions of probability, critical reflection on e.g. different interpretations of p-values and confidence intervals, the replication crisis and questionable research practices, etc.

The prospective student should master the following topics:

- Descriptive statistics, both textual as graphical
- Basic concepts from probability theory ((in)dependence, conditional probability, marginal probability, sum and product rule)
- Confidence intervals
- Correlation
- Simple and multiple linear regression
- Between-subjects analysis of variance (ANOVA)
- Critical assessment of the assumptions of statistical procedures.

Furthermore, acquaintance is expected with the following more advanced topics (although it is not essential to master all these topics):

- Within-subject analysis of variance (repeated measure ANOVA)
- Logistic regression
- Analysis of covariance (ANCOVA)
- Non-linear regression
- Mediation and moderator analysis
- Model selection procedures.

In case some of this knowledge is lacking, prospective students can prepare themselves by studying one of the following text books:

- D.S. Moore, G. McCabe, B.A. Craig (2019). Introduction to the Practice of Statistics. 9th Edition. New York: Freeman
- A. Agresti (2018). Statistical Methods for the Social Sciences. 5th Edition. London: Pearson.
- R.G. Lomax & D.L. Hahs-Vaughn (2012). Statistical Concepts: A Second Course. 1st edition. London: Taylor & Francis.

Statistical Entrance Knowledge Diagnostic Test

1) We have scores of 100 males and 90 females on the variable length in centimeters. Which graph is most informative in summarizing the scores including possible gender differences?
   a) A stemplot of the variable length
   b) Two boxplots of length: one for the males and one for the females
   c) A scatterplot with on the x-axis the length of males and on the y-axis the length of females
   d) A histogram of length and a bar graph of gender

2) A statistic (e.g. the median) for a group of scores is called a resistant or robust statistic when
   a) a small change in all scores results in a substantial change in the value of the statistic
   b) a small change in one score results in an unsubstantial change in the value of the statistic
   c) a large change in all scores results in a substantial change in the value of the statistic
   d) a large change in one score results in an unsubstantial change in the value of the statistic

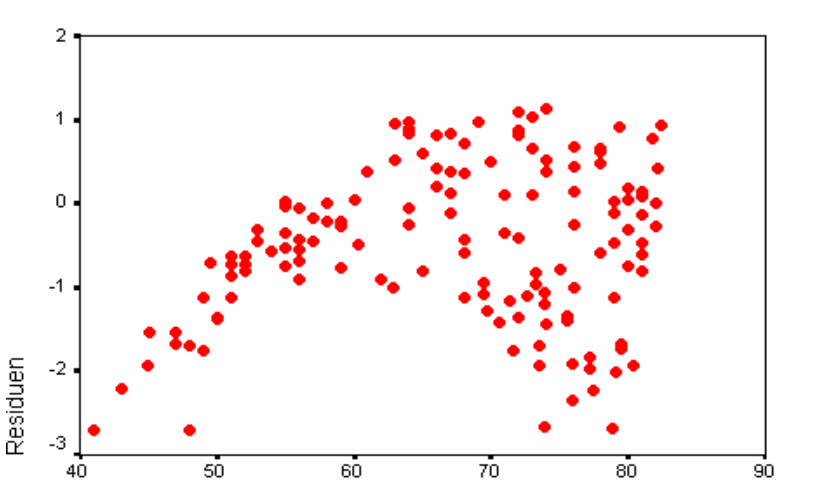
3) A researcher computes the variance of a series of scores and finds it to be 0. What can you say about the series of scores?
   a) All scores are 0
   b) All scores are the same, but not necessarily equal to 0
   c) The researcher has made a mistake, because a variance of 0 is impossible
   d) It is impossible to say anything about the series of scores based on this information only

4) Suppose the events A "Chelsea wins the Champions League" and B "The Netherlands will be hit by a serious earthquake in the next ten years" are statistically independent. P(A) denotes the probability that A will happen, P(B) denotes the probability that B will happen, P(A and B) denotes the probability that both A and B will happen, P(A or B) denotes the probability that A or B or both will happen. Then which of the following statements is true?
   a) P(A and B) = P(A) + P(B)
   b) P(A and B) = P(A)P(B)
   c) P(A and B) = P(A) + P(B) + P(A or B)
   d) P(A and B) = P(A or B) - P(A)P(B)

5) In the population, scores on a test are normally distributed with mean 125 and standard deviation 15. What is the probability that 25 randomly chosen respondents from the population have a mean score of more than 131 (rounded to 2 decimals)?
   a) 0.02
   b) 0.35
   c) 0.65
   d) 0.98

6) For the analysis of a continuous dependent variable and a categorical independent variable with 4 categories, you use
   a) The X^2-test for two-way tables (cross-tables)
   b) One-way analysis of variance
   c) Simple linear regression with one dummy variable
   d) Logistic regression analysis

7) Below, the residual plot for a multiple regression analysis of 122 countries is given. The dependent variable is life expectancy. The independent variables are GNP, number of doctors (per 10000 residents), grade of urbanization and the number of radios (per 100 residents). The residual plot shows that two assumptions in the multiple regression analysis are not satisfied. Which assumptions are these?
   a) Normally distributed residuals and independent observations
   b) Independent observations and constant variance
   c) Constant variance and a linear model
   d) A linear model and normally distributed residuals



DEPENDENT: life expectency

8) The relation between brain size and intelligence is examined by determining the regression line where IQ is predicted by the number of active pixels on an MRI-scan (indicator for brain size). Below, part of the SPSS output for 41 respondents is given. Determine the 95% confidence interval for the regression slope based on this information.

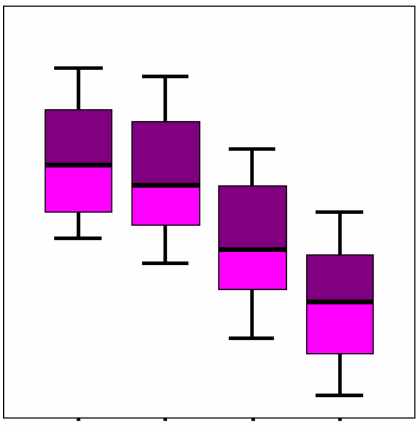
Coefficients(a)

Model	Unstandardized Coefficients		Standardized Coefficients	95% Confidence Interval for B	
	B	Std. Error	Beta	Lower Bound	Upper Bound
1 (Constant)	5,168	46,008			
MRI pixel count (hundred thousand)	11,915	5,047	,358		

A Dependent Variable: Full scale IQ

a) $5.168 \pm 1.96 \times 46.008$
b) $5.168 \pm 2.021 \times 46.008$
c) $11.915 \pm 1.96 \times 5.047$
d) $11.915 \pm 2.021 \times 5.047$

9) Beside, boxplots of scores for four groups of subjects on a particular variable are given. In an analysis of variance, the variation within groups is compared with the variation between groups in order to evaluate the differences between the four shown groups. What aspects of the boxplots show these types of variation?



a) Variation between groups: the length of each box Variation within groups: the length of corresponding tails
b) Variation between groups: the location of the median in each box
Variation within groups: the length of the corresponding box
c) Variation between groups: the differences between the four medians
Variation within groups: the length of the boxes
d) Variation between groups: the differences between the four box-lengths
Variation within groups: the length of the tails

10) Below is a table with SPSS results of a logistic regression analysis. The variable GESLACHT was coded as male = 0 and female = 1. The dependent variable is smoking (0=no, 1=yes). The estimated parameter β of GESLACHT represents
   a) how much the variable smoking changes when GESLACHT increases with 1 unit
   b) how much the chance of smoking increases when GESLACHT increases with 1 unit
   c) how much the odds of smoking increase when GESLACHT increases with 1 unit
   d) how much the log of the odds of smoking increase when GESLACHT increases with 1 unit

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	GESLACHT(1)	2.197	.516	18.104	1	.000	9.000	3.271	24.763
	Constant	-1.099	.365	9.052	1	.003	.333		

a. Variable(s) entered on step 1: GESLACHT.

11) A researcher wants to perform a regression of Y (continuous variable) on X (continuous variable) and A (categorical variable with three levels). She has doubts on whether to use dummy coding or unweighted effects coding to code variable A. She is afraid that the results concerning the overall fit of the regression model might differ a lot between the two analyses. Which of the following statements is correct?

   a) The researcher should not worry, since the choice of the coding never affects the overall fit of the regression model.
   b) The researcher should worry, because different codings can lead to different results for the overall fit of the regression model.
   c) The researcher should worry if Type I sum of squares is used in both analyses.
   d) The researcher should worry in case of an unbalanced design.

12) The assumption of sphericity requires:
   a) Equality of correlations between repeated measures.
   b) Equality of correlations between the code variables.
   c) Equality of variances of the original repeated measures.
   d) Equality of variances of the pairwise differences of repeated measures.

13) What is a feature of repeated measures ANOVA?
   a) Individual differences between subjects are removed from the between-treatments variance.
   b) Sphericity is one of the required assumptions for the RM-MANOVA multivariate test.
   c) The treatment effect explains part of the within-subjects variance.
   d) The within-subjects variance is reduced by removing individual differences between subjects.

14) In which situations are nonlinear regression models always more appropriate than linear regression models?
   a) When the assumption of homoscedasticity is violated.
   b) When the dependent variable is categorical.
   c) When the independent variable is categorical.
   d) When the sample size is large.

15) Consider the following experiment: All subjects in a sample are tested in four different occasions (factor `treatment'), immediately after an intervention. Gender and age (two levels: age 20-40, age 40-60) are taken into account. Repeated measures ANOVA is used to analyse these data. Which predictors are the between and the within factors?
   a) Between-factor: age, gender. Within-factor: treatment.
   b) Between-factor: gender. Within-factor: age, treatment.
   c) Between-factor: gender, treatment. Within-factor: age.
   d) Between-factor: treatment. Within-factor: age, gender.

Answers Statistical Entrance Knowledge Diagnostic Test

Did you do the complete Diagnostic Self Test Statistics, with 15 questions? If not, first go (back) to the test, and only after that verify your answers.
If you did, you can verify your answers below. Brief explanations are given each time. In several cases the explanation is simply "by definition". This means that you can look up why this is the correct answer by searching for the definition of this matter in any good introductory statistics book (or, although possibly less reliable, on the internet).

Evaluation of results

0-9 correct	Clearly insufficient
10-11 correct	Marginally sufficient
12-15 correct	Sufficient

If your result is clearly insufficient, you need to improve or at least refresh your knowledge of basic statistics.
If your result is marginally insufficient, you need to improve or at least refresh parts of your knowledge of basic statistics. Check in particular the sections related to your wrong answered questions.
If your result is sufficient, there is no indication that you need to improve or refresh your knowledge of basic statistics, although, obviously the test is only a minor indication on this.

Answers to test questions

1. b, suitable for 1 continuous (length) and 1 categorical variable (gender)
2. d, by definition
3. b, When all scores are the same, the mean score is equal to the scores. Therefore, the quadratic difference between scores and the mean is 0.
4. b, by definition
5. a, $$P\left(\overline{X} > 131 \middle| N\left(125, \frac{15}{\sqrt{25}} = 3\right)\right) = P\left(Z > \frac{131-125}{3}\right) = P(Z > 2) = 0.0228$$
6. b, by definition
7. c, the variance of the residuals is larger for larger values of the dependent variable. Therefore, no constant variance for all values of the dependent variable.
8. d, the estimated coefficient of the slope is 11.915 with standard error 5.047, using the t-distribution with df=40, the critical value is 2.021.
9. c, by definition
10. d, by definition
11. a, by definition
12. d, by definition
13. c, by definition
14. b, by definition
15. a, by definition