Statistics 1 and 2

Caspar J. Van Lissa

Table of contents

1 Overview

This GitBook covers the basics of statistics and data analysis. The ability to extract insights from data is an essential skill for both academic and non-academic work, and "data literacy" is increasingly important in a world where data are collected about every aspect of our lives. In this book, you will be able to independently analyze data, interpret and report your findings, and assess the results of analyses performed by others, such as you might find in scientific articles.

1.1 Learning goals

This book covers the following learning goals:

1. compute and interpret commonly used descriptive statistics such as the sample mean, the median, the mode, variance and standard deviation, the standard error, and the correlation coefficient.
2. recognize different probability distributions such as the normal distribution, and make computations for these probability distributions.
3. explain the essential aspects of null-hypothesis significance testing, including sampling distributions, Type I and Type II errors, one-tailed versus two-tailed testing, and statistical power.
4. apply different statistical tests such as the Z-test, the one sample t-test, the one way Between Subjects Analysis of Variance test, and statistical tests related to (multiple) linear regression analysis with continuous and categorical predictors; and clarify the statistical and/or methodological assumptions that apply to the techniques that are discussed in this course.
5. explain basic concepts in regression analysis, including: linear association, least-squares estimation, explained variance, Multiple R, multiple correlation, adjusted R-square, raw and standardized regression coefficients, model-comparison tests, predicted scores, residuals and the assumptions;
6. choose the appropriate analysis technique for answering a specific research problem from the range of techniques that are covered in the course.
7. use the software package SPSS to perform several statistical data analyses and be able to correctly interpret and report the output to an informed audience (e.g., Liberal arts students, researchers from the social sciences/business and economics/cognitive neuroscience).

8. draw valid conclusions from the results of empirical data analyses given specific research questions envisaged.
9. apply statistical tests in the context of multiple linear regression models with interaction terms and logistic regression models; interpret the corresponding output.
10. describe the concepts of probabilities, odds and logits; describe the relationship between the three scales; transform one into another (formulae are provided).
11. apply statistical tests in the context of factorial ANOVA, ANCOVA and Analysis of Repeated measures; interpret the corresponding output; and calculate and interpret effect size estimates relevant for these statistical techniques (e.g., (partial) eta squared)
12. apply statistical tests in the context of multiple linear regression models with interaction terms and interpret the corresponding output.
13. gauge the reliability of measurements from questionnaires and identify problematic items.
14. explore the dimensionality of questionnaire data.

1.2 Using this GitBook

This GitBook is "Open Educational Material". It is intended to replace conventional statistics textbooks, for free.

All essential information is contained within this GitBook. However, it is recommended that you download several materials to your local hard drive, and use them from your your local computer throughout the course:

1. Download all data files for the tutorials
    - Save the ZIP archive to your drive
    - Right-click the downloaded file, and select "Extract here" (or similar option, depending on your operating system)
    - You should now have a folder with the data files.
2. Download all lecture PDFs
3. *Optional:* Download the PDF of the book

1.3 Software

By default, it is assumed that you will be making the exercises in this book using the commercial program 'SPSS'.

It is important to note that there are free alternatives to SPSS; you might consider using these on your own computer, instead of buying a license for SPSS:

- PSPP, which is designed to be nearly identical to SPSS with all the same basic functionality: https://www.gnu.org/software/pspp/pspp.html

  - If the PSPP website is down, the Windows installer is here: https://sourceforge.net/projects/pspp4windows/

- JASP, which is more modern, looks nicer and is very easy to use – but looks less similar to SPSS: https://jasp-stats.org/

1.4 Contributing / Fixing Errors

This book is a work in progress, so you might find errors. Please help me fix them! The best way is to open an issue on github that describes the error. You are also welcome to suggest fixes directly by opening a pull request, if you know how to.

1.5 Credit

This book was authored by Caspar J. Van Lissa. Its code and layout are derived from Lisa DeBruine's "booktem" (DeBruine & Lakens, n.d.).

Also see: https://psyteachr.github.io/

2 Statistics 1

The course *Statistics 1* covers the basics of statistics and data analysis. This GitBook contains all relevant information about this course. It is assumed that every student reads it carefully. If you have any questions, first consult this GitBook, then ask a fellow student, and only if your question is still not answered, then contact the course coordinator.

Communication about the course occurs through Canvas (Login with your student ID and password).

2.1 Course Description

In the course, the following techniques will be discussed:

- Information on the use of SPSS and interpretation of the output.
- Descriptive statistics;
- Normal distribution; standard scores;
- Sampling distributions; Z and t distributions;
- Hypothesis tests and confidence intervals for the mean.
- The power of a statistical test.
- One way Between Subjects Analysis of Variance.
- Linear regression analysis

2.2 Learning goals

After taking this course, students will be able to...

1. compute and interpret commonly used descriptive statistics such as the sample mean, the median, the mode, variance and standard deviation, the standard error, and the correlation coefficient.
2. recognize different probability distributions such as the normal distribution, and make computations for these probability distributions.
3. explain the essential aspects of null-hypothesis significance testing, including sampling distributions, Type I and Type II errors, one-tailed versus two-tailed testing, and statistical power.

4. apply different statistical tests such as the Z-test, the one sample t-test, the one way Between Subjects Analysis of Variance test, and statistical tests related to (multiple) linear regression analysis with continuous and categorical predictors; and clarify the statistical and/or methodological assumptions that apply to the techniques that are discussed in this course.
5. explain basic concepts in regression analysis, including: linear association, least-squares estimation, explained variance, Multiple R, multiple correlation, adjusted R-square, raw and standardized regression coefficients, model-comparison tests, predicted scores, residuals and the assumptions;
6. choose the appropriate analysis technique for answering a specific research problem from the range of techniques that are covered in the course.
7. use the software package SPSS to perform several statistical data analyses and be able to correctly interpret and report the output to an informed audience (e.g., Liberal arts students, researchers from the social sciences/business and economics/cognitive neuroscience).
8. draw valid conclusions from the results of empirical data analyses given specific research questions envisaged.

2.3 Course Schedule

The official **course schedule** is available on TimeEdit. The information below might go out of date. For a general overview of the content, see below:

	Week	Activity	Date	Topic
8	35	Lecture	26-08-2025	Introduction to Statistics
23	35	Tutorial	28-08-2025	Introduction to Statistics
24	35	Tutorial	28-08-2025	Introduction to Statistics
25	35	Tutorial	28-08-2025	Introduction to Statistics
1	35	Exams	29-08-2025	Register group for Assignment 1

	Week	Activity	Date	Topic
9	36	Lecture	02-09-2025	NO LECTURE! But watch Descriptive Statistics
26	36	Tutorial	04-09-2025	Descriptive Statistics
27	36	Tutorial	04-09-2025	Descriptive Statistics
28	36	Tutorial	04-09-2025	Descriptive Statistics
10	37	Lecture	08-09-2025	Bivariate Descriptives
29	37	Tutorial	09-09-2025	Bivariate Descriptives
30	37	Tutorial	11-09-2025	Bivariate Descriptives
31	37	Tutorial	11-09-2025	Bivariate Descriptives
11	38	Lecture	17-09-2025	Probability Distributions
32	38	Tutorial	18-09-2025	Probability Distributions
33	38	Tutorial	18-09-2025	Probability Distributions
34	38	Tutorial	18-09-2025	Probability Distributions
12	39	Lecture	22-09-2025	Sampling Distribution

	Week	Activity	Date	Topic
35	39	Tutorial	25-09-2025	Sampling Distribution
36	39	Tutorial	25-09-2025	Sampling Distribution
37	39	Tutorial	25-09-2025	Sampling Distribution
13	40	Lecture	29-09-2025	Philosophy of Science
38	40	Tutorial	02-10-2025	Philosophy of Science
39	40	Tutorial	02-10-2025	Philosophy of Science
40	40	Tutorial	02-10-2025	Philosophy of Science
14	41	Lecture	06-10-2025	Hypothesis Testing
41	41	Tutorial	09-10-2025	Hypothesis Testing
42	41	Tutorial	09-10-2025	Hypothesis Testing
43	41	Tutorial	09-10-2025	Hypothesis Testing
2	41	Exams	10-10-2025	Deadline Assignment 1
5	41	Exams	11-10-2025	Exam 1

	Week	Activity	Date	Topic
3	42	Exams	17-10-2025	Register group for Assignment 2
15	43	Lecture	20-10-2025	General Linear Model (GLM) I: Bivariate regression
44	43	Tutorial	23-10-2025	General Linear Model (GLM) I: Bivariate regression
45	43	Tutorial	23-10-2025	General Linear Model (GLM) I: Bivariate regression
46	43	Tutorial	24-10-2025	General Linear Model (GLM) I: Bivariate regression
16	44	Lecture	27-10-2025	GLM II: Sums of squares, explained variance, and correlation
47	44	Tutorial	30-10-2025	GLM II: Sums of squares, explained variance, and correlation
48	44	Tutorial	30-10-2025	GLM II: Sums of squares, explained variance, and correlation
49	44	Tutorial	31-10-2025	GLM II: Sums of squares, explained variance, and correlation
17	45	Lecture	03-11-2025	Assumptions
50	45	Tutorial	06-11-2025	Assumptions
51	45	Tutorial	06-11-2025	Assumptions
52	45	Tutorial	07-11-2025	Assumptions

	Week	Activity	Date	Topic
18	46	Lecture	10-11-2025	GLM III: Differences between two groups: dummy variables and two-sample t-test
53	46	Tutorial	14-11-2025	GLM III: Differences between two groups: dummy variables and two-sample t-test
54	46	Tutorial	14-11-2025	GLM III: Differences between two groups: dummy variables and two-sample t-test
55	46	Tutorial	14-11-2025	GLM III: Differences between two groups: dummy variables and two-sample t-test
19	47	Lecture	17-11-2025	GLM IV: Differences between more groups: dummy variables and ANOVA
56	47	Tutorial	21-11-2025	GLM IV: Differences between more groups: dummy variables and ANOVA
57	47	Tutorial	21-11-2025	GLM IV: Differences between more groups: dummy variables and ANOVA
58	47	Tutorial	21-11-2025	GLM IV: Differences between more groups: dummy variables and ANOVA
20	48	Lecture	24-11-2025	Open science and questionable research practices
59	48	Tutorial	25-11-2025	Open science and questionable research practices
60	48	Tutorial	27-11-2025	Open science and questionable research practices
61	48	Tutorial	27-11-2025	Open science and questionable research practices
4	50	Exams	12-12-2025	Deadline Assignment 2

	Week	Activity	Date	Topic
6	51	Exams	19-12-2025	Exam 2

2.4 Attendance

Attendance is mandatory based on our experience that students who actively participate tend to pass the course, whereas those who do not tend to drop out or fail. All lectures and practicals 'build' on each other, so if you have to miss either one, absolutely make sure you have caught up with the materials before the next session.

2.5 Study Load

Below is a breakdown of the expected study load:

Activity	Duration	Times	Total
Synchronous			
Lectures	2.0	14	28
Tutorials	2.0	14	28
Asynchronous			
Knowledge clips	1.0	14	14
Formative tests	1.0	14	14
Reading time	1.0	14	14
Studying	1.0	1	37
Assessment type(s)			
Portfolio	15.0	2	30
Exam	1.5	2	3
Total			**168**
ECTS			**6**

2.6 Staff

Coordinator:

dr. Caspar J. van Lissa

Lab sessions

Amirali Rezazadeh

2.7 Teaching Philosophy

1. Student-paced learning: instead of having traditional lectures where you sit and listen for two hours, you will watch relatively short (~45 minutes) lecture videos to prepare for class. In class, we use the material from these videos to guide discussions, make exam questions, and work on your portfolios.
2. Challenge-based learning: a substantial part of your grade is based on your ability to apply the techniques you've learned to a real research question in several portfolio assignments. You can choose your own research question, can find your own dataset (or use a default dataset), and work on a topic that actually interests you.
3. Throughout the course, you will be working in small learning teams to promote interaction among students, peer support, and accountability. Learning to work effectively in groups is an important skill; we will focus on group skills in the first lecture.

2.7.1 Why group assignments?

Contact with fellow students is a key aspect of the university experience. We want to stimulate you to engage with the material and with one another. Therefore, the portfolio assignments are made in groups. There are also aspects of learning in groups that can really improve your knowledge, like peer feedback. To ensure that every group member pulls their weight, the final exam tests each student's individual comprehension of all material covered in the portfolios.

Groups comprise 3-5 members and are assigned randomly when the course starts. However, it is allowed to switch with a consenting member of another group, or to join/merge with another small group if your group has become smaller than 3 members. There are three portfolio registration deadlines. Before these deadlines, one group member must submit the definitive group composition via a Google form.

2.7.2 Why use portfolio assessment?

Portfolio assignments are well-suited for a skills-based course like Statistics 1. They also take a lot of the pressure off because you can work at your own pace, and keep improving the work until it is good enough. We entrust you with the responsibility of making these portfolio assignments in good faith, without instrumental assistance from outside your group or plagiarism, so I kindly ask you to make good on this trust, and hand in original work to show what you've learned.

2.8 Grading

Your grade is based on two components:

1. *A portfolio* composed of three assignments made in groups, and
2. *An individual exam*, split into three sessions, to test comprehension of the material covered in the portfolios.

A grade of 5.5 or higher is required for both components to pass the course.

The first occasion for the exam is split into three sessions, administered throughout the semester, for the following reasons:

- To reduce study load by administering small tests shortly after the material is taught
- To ensure continued engagement with the course
- To give students feedback on their current level of understanding

While you do receive an informal grade for each session, the final grade is simply calculated based on your correct answers in all sessions. If that grade falls below 5.5, you can take a resit which covers the material of the entire exam (all 3 sessions).

2.8.1 Portfolios 40% (2 x 20%)

You work on the portfolio assignments with your group, both during the lab sessions and outside of class. For each assignment, register your group membership before the set deadline at http://tiny.cc/stats12_portfolio. You hand in your group's portfolio assignment before the set deadline, at which point it is graded. If your grade is below the passing level of 5.5, your group will have the opportunity to revise the portfolio based on teacher feedback to receive a maximum grade of 6.

Groups should equally distribute the work load for the portfolio assignments. In case doubts are raised about the equal distribution of labor in a particular group, the portfolio assignment in question will be supplemented with individual oral examination and an individual grade, which can not exceed the original grade for the group assignment. In other words, failing to

distribute the work properly can not have positive effects, but it can have negative effects on your grade. To prevent this, make clear agreements about the distribution of work with your group mates.

2.8.2 Exam 60%

To make sure that all students are equally involved in the making of the portfolio assignments, an individual exam assesses comprehension of the material covered therein. It is a digital multiple choice exam, split into three sessions, to test comprehension of the material covered in the portfolios.

2.8.2.1 Exam 1

Covers Week 35 (Introduction to Statistics) up to Week 41 (Hypothesis Testing)

2.8.2.2 Exam 2

Covers Week 43 (General Linear Model (GLM) I: Bivariate regression) up to Week 47 (Open science and questionable research practices)

2.9 Assignments

Below is a description of the assignments. For each assignment, every element labeled with a lower case letter is graded fail (0 points), pass (1 point), or excellent (1.5 points). Grades are summed for each assignment, and rescaled from 1-10. The final grade is the average across assignments of the rescaled grades. Note the stated word limit for each section. If you can write a good report with fewer words, that's fine. If you exceed the word limit however, your grade for that section cannot exceed a pass (1 point).

The focus of the assignments should me on motivating, reporting, interpreting, and discussing your analyses. You will get a good grade for well-reasoned and discussed analyses.

See the Appendices section to access data sources for the assignment.

2.9.1 Assignment 1

Descriptive statistics and statistical inference

a. Select at least three variables for further analysis, and motivate your selection based on theory, using at least one reference to explain why are you interested in the properties of the selected variables (150 words)
    i. Include one continuous variable
    ii. Include one nominal variable
    iii. Include one ordinal variable
b. Describe the dataset (200 words + tables/figures)
    i. Use appropriate univariate descriptive statistics for all variables
    ii. Plot data using appropriate plots
    iii. Include at least one frequency- or crosstable
c. For a continuous variable:
    i. Select one or more values with clinical/societal/statistical relevance (i.e., provide some justification for the choice of value)
    ii. Using probability calculus, calculate and report the probability of observing values that fall below/between/exceed the chosen value(s)
d. For a continuous variable:
    i. Formulate a specific null- and alternative hypothesis
    ii. Report a one-sample t-test or Z-test for the specific null-hypothesis
    iii. Calculate the probability of comitting a Type II error
e. Discuss your analyses (300 words)
    i. Explain your rationale for important modeling decisions
    ii. Motivate your choice for the type of statistics and analyses
    iii. Discuss assumptions
    iv. Discuss what you have learned from it and how you might improve it
f. Use APA style throughout your report
g. Reflect on the group process (300 words). Note: I will grade your *reflection*, not your *process*. So: if your group's process is not working well, but you reflect on it properly, you can still get full marks for this component. Use Gibbs' Reflective Cycle:
    i. Describe what happened during the group work
    ii. Explain how you felt during the group work
    iii. Look at the good and bad aspects of the group work
    iv. What were the obstacles you experienced? What factors contributed to success?
    v. What could you have done differently to improve the situation?
    vi. What are your intentions to make the next group assignment work (even) better?

2.9.2 Assignment 2

General linear model

a. Select at least three variables for further analysis, and using at least one reference, explain what research questions you will investigate and what hypotheses you will test (150 words)
    i. Include one continuous outcome variable
    ii. Include one continuous predictor
    iii. Include one nominal or ordinal predictor
b. Construct a model with only the continuous predictor (200 words)
    i. Report and interpret the different sums of squares
    ii. Report and interpret the explained variance
    iii. Conduct a separate correlation analysis. Compare the results with the regression analysis.
c. Construct a model with only the categorical predictor (200 words)
    i. Report and interpret the model results
    ii. Conduct a separate ANOVA or t-test with the same variables, whichever one is suitable. Compare the results with the regression analysis.
d. Practice Open Science:
    i. Pick a continuous outcome, ideally one you have not used as an outcome before. Pick a predictor (continuous or categorical doesn't matter, but use the appropriate analysis).
    ii. Write down a hypothesis about this predictor and outcome, BEFORE looking at the analysis results. Treat this hypothesis as your "preregistration".
    iii. Conduct the appropriate analysis to test your hypothesis. In the spirit of reproducibility, use the "Paste" buttons in SPSS to save all analysis syntax, to make it reproducible. Report the SPSS version number and the syntax in your portfolio. After preparing the syntax, run it and report the results.
e. Practice P-hacking:
    i. Use the syntax from d. to conduct 5 more unplanned analyses. Swap out the predictor- and outcome variables in the syntax as you please to create these additional analyses.
    ii. Run all analyses, look at the results, and pick one with the most interesting (unexpected or significant) results.
    iii. Come up with a hypothesis to explain this finding.
    iv. Discuss why the analysis from d. should be trusted, and the one from e. should not (200 words)
f. Discuss your analyses (300 words)

i. Explain your rationale for important modeling decisions
ii. Motivate your choice for the type of statistics and analyses
iii. Discuss assumptions
iv. Discuss what you have learned from it and how you might improve it

g. Use APA style throughout your report
h. Reflect on the group process (300 words). Note: I will grade your *reflection*, not your *process*. So: if your group's process is not working well, but you reflect on it properly, you can still get full marks for this component. Use Gibbs' Reflective Cycle:

i. Describe what happened during the group work
ii. Explain how you felt during the group work
iii. Look at the good and bad aspects of the group work
iv. What were the obstacles you experienced? What factors contributed to success?
v. What could you have done differently to improve the situation?
vi. What are your intentions to make the next group assignment work (even) better?

2.10 Use of Large Language Models (LLMs)

Honestly: I advise against using ChatGPT and similar LLMs for this course. Here's why: LLMs learn from all text on the internet, which includes a *lot* of text posted by people who do not understand statistics. As a result, my experience is that LLMs produce a lot of plausible sounding nonsense for statistics assignments.

If you're worried about the quality of your writing: that is not graded here. I'd rather have a simple and clear report in imperfect English than a beautifully written AI-fluff piece full of hallucinated nonsense.

If you decide to use LLMs, it is your responsibility to thoroughly check its output for logical consistency and correctness. You may not yet have the level of expertise required to know when ChatGPT generates irrelevant nonsense - but the teacher who grades your work does. Consider this carefully when deciding what makes more sense: doing your work manually, making sure each step is correct - or outsourcing it to AI, and then checking its work before submitting.

3 Introduction to Statistics

Statistics are more relevant than ever in this digital age, where data about our entire lives is readily available, and software to analyze such data has become extremely user-friendly and freely available. We live in a world where organizations large and small collect data to tailor products and services, and being data literate is becoming increasingly important across industries.

Statistics allows us to make sense of data and gain valuable insights. It helps us better understand social phenomena, predict sales and optimize marketing strategies, and even explore the relationship between brain activity and behavior. Data analysis is one of the most marketable skills taught at universities.

Before we delve deeper into statistics, it's crucial to distinguish between methods and statistics. Methods refer to the procedures used in research, such as data collection, participant selection, and study design. Statistics, on the other hand, focuses on analyzing the data obtained from these methods.

Two fundamental branches of statistics covered in this course are descriptive statistics and inferential statistics. Descriptive statistics involves summarizing and describing the characteristics of a dataset, while inferential statistics allows us to make educated guesses about a larger population based on a smaller sample.

Statistical modeling is another aspect of statistics where theories are represented mathematically. This enables us to predict important outcomes, such as sales figures, well-being, or the likelihood of neurological disorders. Statistical modeling also allows us to explore data for interesting patterns or to perform tests to answer theoretically driven research questions.

In scientific research, statistics can help us test theories. The process of scientific knowledge acquisition is described by the empirical cycle: We start with a theory, from which we derive testable hypotheses. A theory is an abstract system of assumptions about the relationships between constructs. A hypothesis is a concrete statement, derived from the theory, about expected quantitative relationships between measured variables. We then collect data and test the hypothesis. If the hypothesis is refuted, we re-examine the theory and possibly amend it.

To lay a foundation for understanding statistics, it's essential to be familiar with some basic concepts. First, data in the social sciences often come in tabular format (e.g., spreadsheets), where each row represents an individual observation, and each column represents the individuals' scores on various variables.

3.0.1 Population and Sample

A crucial distinction is the one between population and sample. The population refers to the complete set of objects of interest, such as all people in a country or all students in a class. However, due to practical limitations, we usually do not have access to the population. Instead, we draw a sample from it, which is a subset of the population. Sampling theory establishes the rationale for drawing inferences about a population based on samples. Sample statistics serve as our best estimate of population parameters. If the sample is representative, those estimates will be unbiased. Moreover, we can estimate our uncertainty about the sample statistics as estimates of population parameters. The best way to ensure a representative sample is to use random sampling, where each individual in the population has an equal chance of being included—though in practice, constraints often lead researchers to rely on convenience or stratified sampling, which can limit the strength of our inferences.

The distinction between constructs and variables is also important. Constructs are abstract features of interest within a population, like short-term memory, intelligence, or education. Variables, on the other hand, are placeholders that represent specific values associated with these constructs—much like column headers in a spreadsheet. Data then refer to the specific values of a variable.

3.0.2 Measurement Levels

Measurement level refers to the kind of information contained in a variable. The four common measurement levels are nominal, ordinal, interval, and ratio (NOIR). Each subsequent level of the NOIR taxonomy carries more information than the previous level, thus permitting different comparisons and statistics.

Nominal variables sort cases into categories with no inherent order (e.g., academic major, country); you can count them, calculate proportions, and the mode). Ordinal variables have a rank order but not equal spacing (e.g., Likert agreement, socioeconomic status quintiles). You can use them to calculate medians and percentiles. Interval variables have equal spacing between units, but no true zero value (e.g., temperature in Celsius, calendar year). Differences between values on an interval scale are therefore meaningful, allowing for the calculation of means, standard deviations, and Pearson's correlation coefficient. You can apply a linear transformation to an interval scale (e.g., $a + b\ X$), but ratios between values on an interval scale are not interpretable (20 degrees Celsius is **not** twice as hot as 10 degrees Celsius). Ratio variables have equal units and a meaningful zero (e.g., income, reaction time, number of friends); both differences and ratios are meaningful.

In practice, researchers sometimes treat coarse ordinal scales (e.g., 5point Likert scales) as interval (or even ratio) for convenience's sake. This is a pretty strong assumption about measurement level.

3.0.3 Descriptive Statistics

Descriptive statistics are used to summarize and analyze data. They help us get a sense of the dataset and answer questions like the most common major among students or the average age of a group. Descriptive statistics can also be relevant in answering research questions, such as evaluating exam questions or determining if the proportion of correct answers on a multiple-choice question is greater than chance.

3.0.4 Study Design and Validity

Researchers employ several study designs to collect data, for example, experimental, quasi-experimental, and observational studies. Each has its own strengths and limitations. The quality of study designs can be evaluated along two dimensions:

Internal validity: the credibility of a finding *within* the context of the study. To what degree does the study design rule out alternative explanations for key findings? In the case of an experiment, for example, can observed differences in outcomes between the experimental and control condition be attributed to the manipulation rather than to alternative explanations such as confounding, selection bias, measurement error, history/maturation effects, demand characteristics, noncompliance, or interference between units? Methods to increase internal validity include: Random assignment, using pretests/pilot studies, reliable instruments, blinding experimentors, participants, and coders (where feasible), manipulation checks, preregistration, and management of participant drop-out.

External validity: the extent to which findings generalize beyond the specific sample, setting, and operationalizations used in the study. It is threatened by, for example, using convenience samples and artificial laboratory tasks that do not resemble the real world (low "ecological validity"). It can be strengthened through random sampling, conducting experiments with high ecological validity, such as field experiments, and conducting conceptual replications (using different methods to find the same effect). A tightly controlled laboratory experiment may score high on internal validity but lower on external validity, whereas a large-scale field survey often shows the opposite pattern. Balancing both is a central concern of scientific inquiry.

3.0.5 Ethics, Privacy, and Reproducible Workflows

Modern data analysis carries legal and ethical responsibilities. Ethical social science research involves protecting participants' autonomy, welfare, and privacy. In the EU, the GDPR legislation requires researchers to establish a lawful basis for processing participants' data (which can include informed consent). Informed consent - participants agreeing to participate in the study and share their data after receiving information about the study and purposes of data collection - must be freely given, specific, and unambiguous Finally, there have been concerns about a lack of reproducibility in the social sciences and many other fields. By some

estimates, as many as 70% of published research findings cannot be independently reproduced. Contemporary practices in social scientific data analysis seek to address this issue. For example, preregistration and registered reports are used to demarkate confirmatory hypothesis tests from exploratory data analysiss. Many researchers also produce reproducible code for their analyses (which can be done in SPSS by generating syntax, instead of running analyses interactively). They sometimes create "replication packages" or "reproducible research archives" that contain the data and code to produce the results, so that other researchers (and sometimes, anyone) can re-use the code, and reproduce and verify the results.

3.1 Lecture

https://www.youtube.com/embed/_TZIIANBt94

3.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the primary purpose of statistics? [1]

- (A) To collect and store data
- (B) To create data visualizations
- (C) To design research studies
- (D) To summarize and analyze data

Question 2

What is the difference between descriptive statistics and inferential statistics? [2]

[1] To summarize and analyze data

[2] Descriptive statistics summarize data, while inferential statistics involves making informed guesses about parameters in a larger population.

- (A) Descriptive statistics involves analyzing data, while inferential statistics involves collecting data.
- (B) Descriptive statistics is used in social sciences, while inferential statistics is used in natural sciences.
- (C) Descriptive statistics summarize data, while inferential statistics involves making informed guesses about parameters in a larger population.
- (D) Descriptive statistics deals with nominal variables, while inferential statistics deals with ratio variables.

Question 3

What is the purpose of statistical modeling? [3]

- (A) To predict outcomes and explore patterns in data
- (B) To make educated guesses about a larger population based on a smaller sample
- (C) To summarize data using graphs and charts
- (D) To represent a theory as a testable statistical model.
- (E) To describe the characteristics of a dataset

Question 4

What is the empirical cycle in scientific research? [4]

- (A) The process of summarizing and describing data using statistics
- (B) The process of repeatedly collecting and analyzing data
- (C) The process of formulating a theory, deriving hypotheses, testing these with data, and reflecting on theory
- (D) The process of designing research studies and selecting participants

Question 5

What is the distinction between population and sample in statistics? [5]

[3] To represent a theory as a testable statistical model.
[4] The process of formulating a theory, deriving hypotheses, testing these with data, and reflecting on theory
[5] Population refers to the complete set of potential participants, of which the sample is a subset.

- (A) Population refers to inferential statistics, while sample refers to descriptive statistics.
- (B) Population refers to the complete set of potential participants, of which the sample is a subset.
- (C) Population refers to all participants in a study, while sample refers all participants who provided complete answers.

Question 6

What is the variance? [6]

- (A) The average distance of observations to the mean.
- (B) The average squared distance of observations to the mean.
- (C) The measure of dispersion for scores on a continuous variable.
- (D) The distance between the lowest and highest value on a continuous variable.

Question 7

Six students work on a Statistics exam. They obtain the following grades: 8, 9, 5, 6, 7 and 8. The teacher calculates a measure of central tendency, which is equal to 7.5. Which statistic did the teacher calculate? [7]

- (A) Median
- (B) Standard deviation
- (C) Mean
- (D) Mode

Question 8

For which of the three scatterplots below is the correlation coefficient strongest? [8]

- (A) B
- (B) A
- (C) C

[6] The average squared distance of observations to the mean.
[7] Median
[8] A

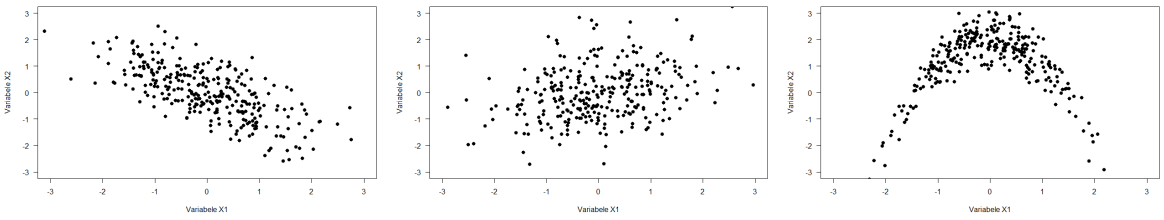
Show explanations

Question 1
Statistics is the science concerned with developing and studying methods for analyzing data.
Question 2
Descriptive statistics are calculated based on sample data; inferential statistics involves using those sample statistics to make best guesses about population parameters and quantify uncertainty about those guesses.
Question 3
Statistical modeling in particular refers to the process of translating a theoretical model into a statistical model whose coefficients can be estimated using data.
Question 4
The empirical cycle is a theoretical cyclical model of knowledge production through scientific research, whereby theory gives rise to hypotheses, which are tested in data, after which the theory is revisited based on the results.
Question 5
Population refers to the complete set of potential participants, of which the sample is a subset.
Question 6
The variance is the sum of squared distances of observations to the mean, divided by the number of observations minus one. So calculate:$S_X^2 = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(7+6+8+6+8)}{5} = 7$
Question 7
First rule out improbable answers; the variance is not a measure of central tendency, and all grades are pretty close to each other, so it would be impossible for the variance to be that high. We can see what the mode (most common value) is: it's 8. So we only choose between mean or median. Mean: calculate $X = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{8+9+5+6+7+8}{6} = 7.17$Median: order the numbers, note that there is an odd number, take the average of the two middle numbers. 5, 6, 7, 8, 8, 9 -> 7.5
Question 8
Correlation measures linear association, so eliminate option C. Option B shows a very small correlation - probably 0 or maybe .1. So the correct answer is A, which shows a moderate negative correlation.

3.3 In SPSS

3.3.1 Instruction Video

https://www.youtube.com/embed/bapuGcjwiLQ?si=51NN65ETr3evOBWb



{width = 30%} {width = 30%} {width = 30%}

Figure 3.1: Scatterplots

3.4 Tutorial

3.4.1 Introducing SPSS

Welcome everyone to your first lab session for Statistics 1 and 2. Today, we start working with an introduction to SPSS and we calculate a few basic descriptive statistics.

Each lab session consists of several assignments and includes explanations on how to carry out the analyses in SPSS.

You can work at your own pace. If you experience any problems, or if you have any questions, feel free to ask your teacher.

You will receive feedback to your answers after you have submitted the practical.

Good luck!

3.4.1.1 Step 1

Hi there!

During the lab sessions of this course you will learn how to work with the statistics program IBM SPSS (SPSS for short).

Background information is given throughout the exercises. We will occasionally refer to additional reading materials for this course, or other sources (e.g., youtube videos).

If you're working from a student workplace, SPSS is already installed. If you're working from your own computer, you either have to purchase SPSS, or you can use a free alternative (see **?@sec-software**) - but note that, at this point, the instruction text is still focused on SPSS so if there are any differences it will be your responsibility to figure out how to use your software.

Your first task is to start the SPSS program. You can easily find SPSS via the Windows Start Menu. SPSS may ask about the coding: use Unicode (button to the left). Then there may be another window open that you can close. In the end you should see an empty spread sheet.

3.4.1.2 Step 2

Now you've got SPSS running, we're ready to go!

We will start with a number of introductory exercises using the data file `stressLAS.sav`. To obtain this and other tutorial data files, download the GitBook, and open the `data` folder to find all files.

Open the file in SPSS. Proceed as follows: via the op menu follow the route: File -> open -> data. SPSS now opens a new window. Search for the file `stressLAS.sav` and open the file in SPSS.

3.4.1.3 Step 3

The file contains data about a study on - you guessed it - stress.

More precisely, it contains data on the following variables:

- `stress`: Measures whether the participant experiences stress, and where the stress comes from.
- `smoke`: Measures the smoking behavior of the participant.
- `relation`: Whether or not the participant is involved in a long-term romantic relationship.
- `optim`: Measures how optimistic the participant is on a scale of 0 to 50.
- `satis`: Measures of life satisfaction of the participant on a scale of 0 to 50.
- `negemo`: The amount of negative emotions on a scale of 0 to 50.

3.4.1.4 Step 4

After opening the data file, you will see the tabs Data View and Variable View at the bottom of your screen.

Make sure the tab Data View is selected.

Look at the Data View and describe the data file. What do the rows represent, and what do the columns represent?

3.4.1.5 Step 5

Now switch to the Variable View tab.

The Variable View lists the variables and their properties. We will not discuss all the columns in detail, but focus on the most important ones, which includes: name, label, values, and measure.

Explain for each of the columns name, label, values, and measure what aspect of the variable it describes. Also explain the difference between variable name and variable label.

3.4.1.6 Step 6

Value Labels

For nominal and ordinal variables we have to indicate what the scores represent; that is, we have to assign so called value labels. Value labels are specified under Values.

If you click on values for the variable of interest, and then on the blue button with the three dots on the right, SPSS opens a new window that allows you to view, define, or modify the value labels.

What are the value labels for the Stress and what are they for Smoke?

3.4.1.7 Step 6a

You may have noticed that the value labels are missing for the nominal variable Relation.

Add the value labels yourself in SPSS such that a score 1 represents "Single" and 2 represents "In a relationship".

3.4.1.8 Step 7

Every variable has a so-called Measurement Level.

First, summarize the measurement levels in your own words (as if you have to explain it to a fellow student). Then, indicate the measurement level for each of the variables of interest (Stress, Smoking, etc.).

3.4.1.9 Step 8

Congratulations, you have completed your first assignment!

Before we proceed make sure that you save the data file (via file > save). Because you changed the data, it is important to save the file under a different name. This way, you don't risk losing the original data.

In the next assignment we will generate descriptive statistics for this data.

3.4.2 Plotting data

3.4.2.1 Step 1

The first step in any statistical analysis involves inspection of the data at hand by means of descriptive statistics and/or graphical summaries. Descriptive statistics include the mean, standard deviation, minimum and maximum value. Examples of graphical summaries are bar charts, histograms, and scatter plots.

In this assignment we will look at graphical summaries. In particular, we will look at three: bar charts, histogram, and scatter plots.

You may use the same data file as for the previous assignments.

3.4.2.2 Step 2

First, we will create a bar chart for Stress.

Proceed as follows:

Graphs > legacy dialogs > bar Select Simple and click on define Select Stress under Category Axis (i.e., the variable at the x-axis) Then Click on OK and consult the graph in the output

3.4.2.3 Step 3

You may have noticed that SPSS by default creates a bar chart with the observed frequency depicted on the y-axis. We will now create a new bar chart and instead ask SPSS to show the percentages on the y-axis.

Proceed as follows:

Graphs > Legacy Dialogs > Bar Again choose Simple and click on Define Under "Bars represent" choose "% of cases" Click on OK. SPSS will now create a bar chart, where the heights of the bars represent percentages.

3.4.2.4 Step 4

Next, we will create a histogram for Negative Emotions.

Proceed as follows:

Graph > Legacy Dialogs > Histogram Select Negative Emotions under variable, and ask SPSS to Display normal curve (check the box). Click on OK.

Investigate the histogram; What is shown on the x-axis and what is shown on the y-axis?

How to read the histogram:

- x-axis: the scores on the negative emotions (here numbers between 0 and 50). bars represent score ranges; the more respondents with a score in that range, the higher the bar.
- y-axis: the observed number of respondents per score range.

3.4.2.5 Step 5

Finally, we will create a scatter plot for Negative Emotions and Life Satisfaction. Scatter plots are very useful to get a first impression of whether variables are associated.

Create a scatter plot as follows:

Graphs > legacy dialogs > scatter/dot Choose Simple Scatter Select Negative Emotions on the x-axis, and Life Satisfaction on the y-axis Click OK

Consult the output. Look at the scatter plot and see if you understand the graph.

How to read a scatter plot:

- x-axis represents the scores on Negative Emotions.
- y-axis represents the scores on Life Satisfaction.
- Each dot in the graph is a case, representing how the case scores on both Negative Emotions and Life Satisfaction.

3.4.3 Quiz

Describe the first bar chart; What is shown on the x-axis? [9]

- (A) Numeric scores of Stress
- (B) Categories of Stress

[9] Categories of Stress

- (C) Percentages of Responses
- (D) Frequency of Responses

In the first bar chart, what is shown on the y-axis? [10]

- (A) Frequency of Responses
- (B) Numeric scores of Stress
- (C) Categories of Stress
- (D) Percentages of Responses

What's the approximate proportion of people experiencing work-related stress? [11]

- (A) 66%
- (B) 70%
- (C) 33%

Based on the bar charts, what can you say about differences in stress levels in the sample? Are most people stressed or not? In other words: How is stress distributed across the three categories? [12]

- (A) Most people report work stress
- (B) Most people report no stress
- (C) Evenly distributed
- (D) Most people report life stress

Describe the distribution of Negative Emotions. Are the scores normally distributed (i.e., like a bell-shape)? Really consider why this is / is not the case before checking your answer. [13]

- (A) Not normal
- (B) Normal

[10] Frequency of Responses
[11] 33%
[12] Evenly distributed
[13] Not normal

Based on the scatter plot from Step 5, would you expect an association between Negative Emotions and Life Satisfaction? [14]

- (A) Strong positive
- (B) Small negative
- (C) No association
- (D) Small positive

3.4.4 Descriptive Statistics

3.4.4.1 Step 1

As explained before, the first step in any statistical analysis involves inspection of the data. In the previous assignment we looked at graphical summaries.

This assignment shows you how to explore data using descriptive statistics. Descriptive statistics include values such as the mean, standard deviation, the maximum value and the minimum value.

Use the same data file as for the previous assignments.

3.4.4.2 Step 2

We will first take a look at the descriptive statistics for Optimism, Life Satisfaction, and Negative Emotions.

Compute descriptive statistics as follows:

Analyze > Descriptive Statistics > Descriptives

Select the variables Optimism, Life Satisfaction and Negative Emotions Now click on OK SPSS will open a new window - the output window - including a table with the descriptives for the selected variables.

[14] No association

3.4.4.3 Step 3

In the previous step we computed the average value and standard deviations. However, for nominal and ordinal variables, the average value is meaningless. To explore nominal and ordinal variables we may produce Frequency tables. A frequency table shows the observed percentage for each level of the variable.

Let's generate a frequency table for variables Smoke and Relation.

Analyze > Descriptive Statistics > Frequencies

Select the variables for which you want to have the frequency distribution (i.e., Smoke and Relation) Click OK. SPSS now adds a table with the frequency distributions of the selected variables to the output file.

Note: SPSS reports percentages and valid percentages. Percentages differ when there are missing values. Because we don't have missing values here, the numbers are the same. Missing values will be discussed in the next assignment.

3.4.5 Quiz

How many participants are in the sample? _____[15]

What is the mean value of Optimism? _____[16]

For which of the variables is the spread in the scores highest? [17]

- (A) NEGEMO
- (B) SATIS
- (C) OPTIM

The minimum and maximum observed scores for Negative Emotions were: [_[18], ___[19]].

What percentage of participants is a non-smoker? _____[20]

What percentage of participants is in a relationship? _____[21]

---

[15] 780
[16] 19.13
[17] OPTIM
[18] 3
[19] 37
[20] 48.1
[21] 47.9

3.4.5.1 Step 4

One of the reasons to first inspect descriptive statistics is to have a first check if there are erroneous values in the data file. Erroneous values are values that are out of range, or impossible given the variable envisaged. For example, a person may have mistyped his/her age (e.g., 511 instead of 51).

Now it's your task to check for each variable whether there are erroneous values (out of range values) in the file using descriptive statistics and/or graphs.

Use the descriptive statistics to find any erroneous values.

One way to deal with missing values is by removing the entire case. This is not a recommended practice; however, at this point, it is the only method you have learned.

To find the cases that have missing values you may sort the data file on a variable with suspect values from high to low (or low to high).

This can be done as follows:

Data > Sort Cases Select the variable on which cases should be sorted Select the cases in descending order Click on OK Go the data view and verify that the cases are now ordered.

Remove the case(s) (i.e., delete the row from the data file) with invalid values.

3.4.5.2 Step 5

Now that we've "cleaned" the data file it's time to answer our first research question!

The question is: "Are non-smokers in our sample on average more satisfied with their life than smokers?"

To answer this question, we need the mean of life satisfaction per smoking group. In order to generate those, we will use the Split File option in SPSS. This is an option in SPSS that allows us to get results for separate groups.

Data > Split File > Compare Groups Select the groups based on the variable Smoke Click OK

Notice that you don't see any changes in the data file or anything in the output file yet (!). However, after running the Split file command, SPSS from now will do the analyses per group, as we will see next.

Compute the mean of Life Satisfaction (via descriptive statistics) and consult the output.

You may notice that SPSS provides the means of the non-smoking and smoking group separately. Compare the means for both groups to answer the following questions.

3.4.6 Quiz

Was there an erroneous value in the data file? If so, type the value of that erroneous value here: _____[22]

To answer this question, only use reasoning. If you delete that value, how do you think the mean of that variable will be affected? [23]

- (A) Becomes larger
- (B) Stays the same
- (C) Becomes smaller

To answer this question, only use reasoning. If you delete that value, how do you think the standard deviation of that variable will be affected? [24]

- (A) Stays the same
- (B) Becomes larger
- (C) Becomes smaller

To answer this question, only use reasoning. If you delete that value, how do you think the standard deviation of that variable will be affected? [25]

- (A) Becomes larger
- (B) Becomes smaller
- (C) Stays the same

In this sample, who are more satisfied with life? [26]

- (A) Non-smokers
- (B) Smokers

Do you think this also holds for the population of all persons? [27]

---

[22] 220
[23] Becomes smaller
[24] Becomes smaller
[25] Becomes smaller
[26] Smokers
[27] Can't tell

- (A) Yes
- (B) Can't tell
- (C) No

3.4.7 Missing Values

This is a short assignment about missing values.

Missing values are 'holes in the data matrix'. Missing data is a common issue in empirical research. Respondents may forget to fill in questions or refuse to answer questions (if the latter is the case, we are in trouble). It is important that missing data are adequately handled in data analysis.

Use the same data file as for the previous assignments.

In the previous assignment we activated the split file option. However, we don't need this split file in the remaining questions, therefore we have to undo the split file option.

Data > split file Choose "Analyze all cases, do not create groups"

Compute the frequency distribution of stress. Consult the output, and answer the following questions

3.4.8 Quiz

What is the percentage of respondents who experience No Stress? _____[28]

Which type of stress is most common in the sample? [29]

- (A) Life stress
- (B) No stress
- (C) Work stress

For educational purposes only, we will now create missing values in the data file.

Navigate to the Data view and delete the value for Stress for the first 10 cases. Notice that you only have to delete the scores for the variable Stress, and not the complete case.

[28] 33.6
[29] Life stress

Compute the frequency distribution for Stress again and compare the new table with the previous one.

Explain what has changed and why.

Answer

We can see that the values of Percent and Valid Percent have changed and that a 'missing' row has been added to the table. It makes sense that the percentages have changed, as there are now missing values. You may have noticed that the values for Percent and Valid Percent differ. Percentage is obtained by dividing the observed frequency by the total (including respondents with a missing value). Valid Percentage is obtained by dividing the observed frequency by the number of respondents with a valid score (thus, not counting the respondents who had a missing value).

Imagine I have a sample of 65 participants, with 3 missing value. Of these participants, 15 reported no stress. What is the percentage of no stress, calculated by hand? _____[30]

What is the percentage out of valid responses (i.e., valid percent), calculated by hand? _____[31]

By deleting the values we created empty cells in the data file. SPSS sees these empty cells as *system missing*. Some researchers instead use specific values to indicate missing values. For example, we may code missing values by 999 if the respondent *refused* to answer, and 998 if the respondent *accidentally* skipped the question. These are examples of *user missing* values, and we have to specify the values to be coded as missing in the Variable view.

Let's try this!

Go to the Data View, and fill in 999 in the cells that have no value on the variable Stress. Then go to the variable view, look for the column 'Missing' and click on Missing for Stress. A new window opens. Specify 999 as a discrete missing value. SPSS now knows that the value 999 stands for "missing observation". Click OK.

Re-compute the frequency distribution for Stress.

Examine how the table changed compared to the previous ones.

3.4.9 More Descriptive Statistics

In this final assignment, we will continue with descriptive statistics.

As mentioned in the lecture, describing the data is an important first step in any research situation.

[30] 23.1
[31] 24.2

For didactic reasons, we will do some computations by hand, but this is not something you have to do on the exam. However, it is good to experience at least once how the computations work and that the numbers in SPSS are not the result of magic.

Let's first look at measures of central tendency:

Consider the following grades for 10 students: 6, 3, 4, 6, 7, 6, 8, 9, 10, 9.

Compute (by hand) the mean, median, and mode.

Remind me how

- The mode is the most common value.
- The median is the middle value (or mean of two middle values for an equal number)
- The mean is calculated as the sum of all values, divided by the number of values: $\frac{\sum_{i=1}^{n} X_i}{n}$

3.4.10 Quiz

What is the mean? _____[32]

What is the median? _____[33]

What is the mode? _[34]

Measures of variation

Next we will look at a measure of variation (i.e., indicating the amount of spread in the observations).

Consider the grades of 6 students: 2, 7, 6, 7, 8, 9.

Compute the variance and standard deviation by hand.

[32] 6.8
[33] 6.5
[34] 6

Remind me how

The variance is the "average squared distance between observations and the mean": $\frac{\sum_{i=1}^{n}(X_i X)^2}{n1}$
The SD is the square root of the variance
Follow these steps:

1. Compute the mean, e.g., $X = 5$
2. For each observation, calculate the distance from the mean; e.g., $3\ 5 = 2$
3. Square these distances, e.g.: $(2)^2 = 4$
4. Add these distances for all observations
5. Divide by number of observations minus 1

3.4.11 Quiz

What is the variance? _____[35]

What is the standard deviation? _____[36]

We now will verify the answer to the question in the previous step using SPSS!

First, we have to enter the data in SPSS. Proceed as follows:

Open SPSS (use Unicode, and close the opening windows)

Make sure that you have the data view on the screen

Type in the grades in SPSS (i.e.: 2, 7, 6, 7, 8, 9):

Go to variable view and change the name of the variable and provide a meaningful label

Second, we can compute the variance and standard deviation in SPSS.

Proceed as follows:

Analyze > Descriptive statistics > Descriptives

Select the variable you just defined Now click op Options. A new window opens which shows many more descriptive options Enable the variance Click Continue and OK

Consult the table descriptive statistics in the output window.

Were your computations correct?

[35] 5.9
[36] 2.43

3.4.12 Correlation

For the next few questions we need the data file LAS_SocSc_DataLab2.sav. Open the file in SPSS. You will see that the file contains data for six variables, named X1 through X6. We will inspect the associations between pairs of variables (so called bivariate relationships).

First, generate a scatter plot for X1 and X2. Proceed as follows: Graphs > Legacy dialogs > Scatter/dot. Then ask for a Simple scatter. Put X1 on the X-Axis and X2 on the Y-Axis. Describe the association. Take into account whether the relationship follows a straight line (i.e., linearity), is positive or negative (i.e., direction), and whether the relationship seems to be weak, moderate or strong (i.e., strength).

Second, generate a scatter plot for X3 and X4. Make sure that X3 is shown on the X-axis and X4 on the Y-axis. Describe the association in terms of linearity, direction and strength.

Third, generate a scatter plot for X5 and X6. Describe the association in terms of linearity, direction and strength.

Is the relationship between X1-X2 positive? TRUE / FALSE[37]

Is the relationship between X5-X6 positive? TRUE / FALSE[38]

Is the relationship between X1-X2 linear? TRUE / FALSE[39]

Is the relationship between X3-X4 linear? TRUE / FALSE[40]

Give an indication of the strength of the relationship between X1-X2: [41]

- (A) strong
- (B) moderate
- (C) zero
- (D) weak

Give an indication of the strength of the relationship between X3-X4: [42]

- (A) moderate
- (B) zero

[37] TRUE
[38] FALSE
[39] TRUE
[40] FALSE
[41] moderate
[42] strong

- (C) strong
- (D) weak

Give an indication of the strength of the relationship between X5-X6: [43]

- (A) zero
- (B) moderate
- (C) weak
- (D) strong

Consider the relationship between X3 and X4, can you think of an example of two variables that would be associated in this way?

Show answer

Any cyclical process;

- Time in the day and how far the water reaches up the beach (ebb and flow)
- Location of the sun in the sky

3.4.12.1 Correlation Coefficient

In this step we will look at the correlation coefficient as numerical description of linear association.

Notice that in the previous step we found a non-linear association. The correlation coefficient would not be a valid measure to describe such an association, but nevertheless it is instructive to see why caution should be exercised in drawing conclusions about association from the correlation coefficient alone.

We will use SPSS to compute the correlation coefficient.

Analyze > Correlate > Bivariate Select X1, X2, ... X6 as the variables Click OK

Consult the table Correlations in the output.

There are several values in the table, but we are looking for the Pearson Correlation. The other numbers are the so called significance level, a concept we discuss soon, and the sample size.

[43] moderate

3.4.13 Quiz

What is the correlation coefficient for the variables X1 and X2? _____[44]

What is the correlation coefficient for the variables X2 and X6? _____[45]

What is the correlation coefficient for the variables X3 and X4? _____[46]

Can we interpret this correlation coefficient? [47]

- (A) Yes, otherwise SPSS would give an error
- (B) No, assumption of association violated
- (C) No, assumption of normality violated
- (D) No, assumption of linearity violated

Interpret the correlation between X5 and X6? [48]

- (A) Moderate negative
- (B) Weak negative
- (C) Weak positive
- (D) Moderate positive

[44] 0.5
[45] 0.06
[46] -0.8
[47] No, assumption of linearity violated
[48] Moderate negative

4 Descriptive Statistics

Descriptive statistics describe or summarize properties of data collected in a sample. If you collect data on three variables for five participants, you can still print the entire dataset as a table and maintain the overview:

Sex	Height	Age
M	167	21
F	176	17
F	188	18
M	176	19
X	171	17

Any time you collect data from more than just a handful of participants, however, this becomes unfeasible. Instead, we report descriptive statistics.

Descriptive statistics are almost always computed when data are collected, for a variety of reasons:

1. To describe properties of the sample (e.g., the demographic composition)
2. To check for mistakes in data entry; e.g., if the maximum value of the variable `age` is 124, the person who entered the data might have made a mistake
3. To check assumptions of a particular statistical model, which we will cover in later chapters
4. To answer research questions that do not require hypothesis tests, for example:
    - In which country are most of our sales conducted?
    - What is the most common major in my classroom?
    - Based on data collected from all inhabitants of the Netherlands (i.e.: a census, not a sample), what is the average income?

4.0.1 Measures of Central Tendency

Measures of central tendency are statistics that try to capture the "most common" value in a sample. The most common measure of central tendency is the "average", which statisticians would call the "mean". All measures of central tendency summarize the distribution of values of one particular variable as **one representative number**.

4.0.1.1 Mean: the "average" value

The most common measure of central tendency is the mean (or average). It is computed by adding all observed scores, and dividing that total by the number of observations.

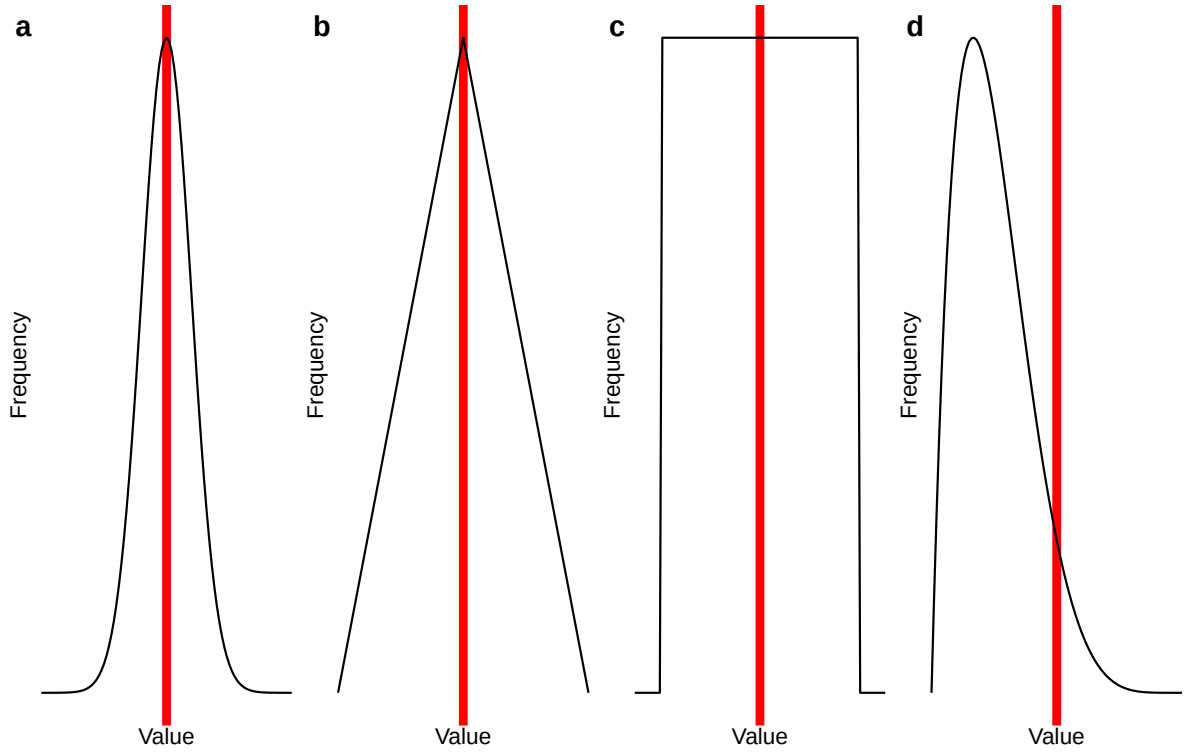
As a formula, this looks like:

$x = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$

An advantage of the mean is that every participant's score contributes to its value equally. This also implies that it is sensitive to extreme values (also called: outliers). If you calculate the mean income in a country where 99% of inhabitants live below the poverty level and 1% are ultra-rich oligarchs, then the mean income will make it look like, on average, people make good money. This sensitivity to extreme values implies that the mean is a good description of the distribution of scores if the distribution is approximately symmetrical (i.e., about 50% of scores are above the mean, and 50% are below it).

Take a look at the figures below; they show different possible distributions of scores in a sample. On the X-axis is the number line. Exact values are not given here because the important point is the *shape* of the distribution, but you can imagine that the X-axis is a scale of height from 150-210 centimeter, or a self-report questionnaire scale from 1-10. On the Y-axis is the frequency with which each number is reported by the participants; a higher value on this axis means that this response is more common. The red line indicates the location of the mean.

Notice that the distributions labeled a-c are all symmetrical: In distribution a, scores cluster around one common value and quickly drop off when you get further away from that common value. In distribution b, scores also cluster around one common value, but they drop off more gradually. In distribution c, every value is exactly equally common. For distributions a-c, the mean would be a good measure of central tendency - it gives you the middle of the distribution. However, also notice that the mean is a better representation of the "most common" value in distributions a-b, but not in distribution c.



4.0.1.2 Median: the middle milestone

If you were to order all scores of your variable from lowest to highest, then the median value is the value that splits your sample in half: half of the participants score lower than this value, and half score higher.

Another name for the median is the *50th percentile*. The *nth percentile* is the score that divides the sample so that $n\%$ of participants score lower. Thus, the 50th percentile means that 50% of participants score lower than the median (and, of course, 50% score higher).

Based on the explanation of the mean, you might already realize that this value should be equal to that of the mean in a perfectly symmetrical distribution. If there are outliers, though, the median is less strongly affected by them than the mean. We can thus say that the median is a measure of central tendency that is *more robust to outliers* than the mean.

The median is not really "calculated", but it is found by literally sorting all values in order, and then picking the middle value (if you have an odd number of observations), or calculating the mean of the two middle values (if you have an even number of observations).

If our variable has these values (which are already ordered):

```
2, 3, 6, 7, 100
```

Then the median value is the middle value, $Med = 6$. Note that the outlier with value `100` does not really affect it (the mean for this sample is much higher, $M = 23.6$).

If our variable has these values (which are already ordered):

```
1, 2, 3, 6, 7, 100
```

Then the median would be the mean of the middle two values: $(3 + 6)/2 = 4.5$.

Below is the picture of the means again, but now, the location of the median is indicated with a blue line. Note that for the symmetrical distributions a-c, the median is identical to the mean - but for the asymmetrical distribution d, the median is a much better representation of the "most common value" than the mean is.

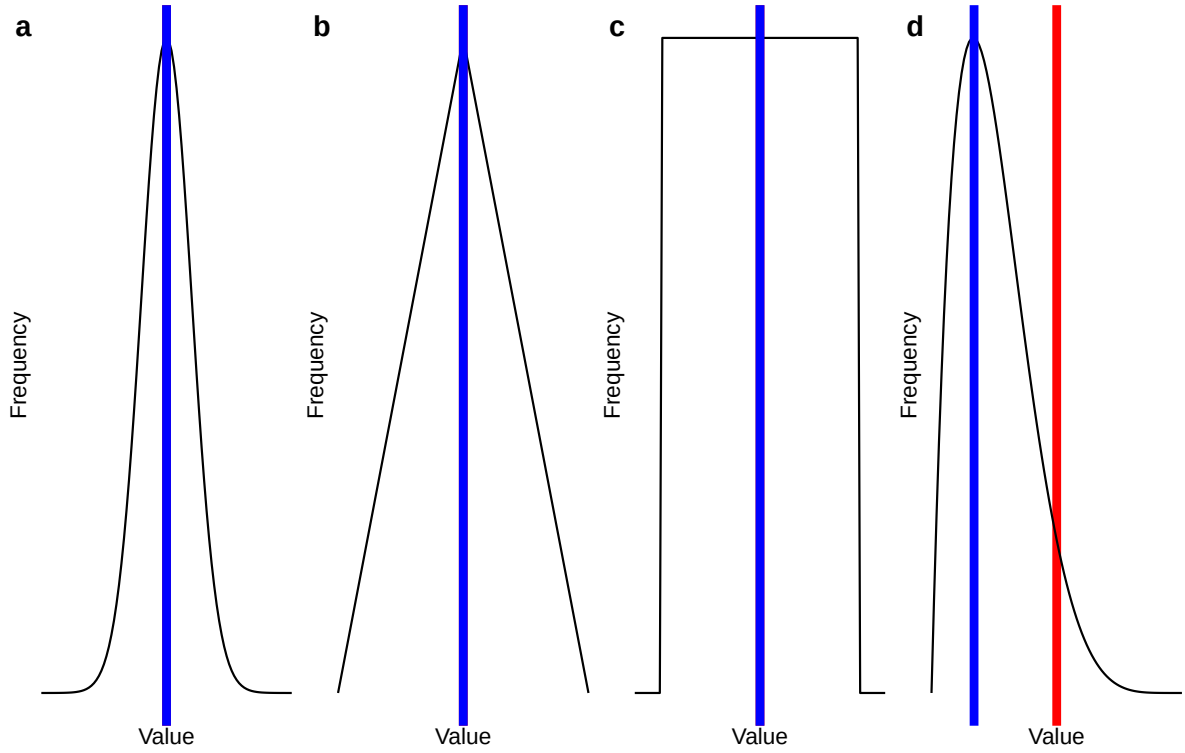


Figure 4.1: Location of the median in symmetrical and asymmetrical distributions.

4.0.1.3 Mode: the most common value

The mode is the most common value in a sample. We can calculate or find it by creating a frequency table, tabulating how often each score is observed in the sample, and then picking the score that occurs most frequently.

While the mode can be obtained for variables with any measurement level, it is the only valid measure of "central tendency" for *nominal* data (e.g., sex, major, favourite color). The other measures of central tendency are *not* valid for nominal data, because these lack a numerical value.

Again, note that in a perfectly symmetrical data distribution, the mode will be identical to the mean and the median.

Imagine, for example, that I have students from three majors:

Major	Frequency
Social Science	43
Cognitive Neuroscience	22
Business & Economics	11

The mode of the variable major, in this case, is "Social science". Do you see why we cannot calculate a mean or median for the variable major? Because the majors don't have a numerical value.

However, it would be perfectly reasonable for me to say that the mode grade obtained last year was a 6.5. This implies that 6.5 was the most common grade, but it doesn't tell you how many students got that grade, or whether the average grade was above or below the level required to pass.

Visually, the location of the mode is the same as the location of the median in plots a, b, and d in Figure Figure 4.1. Plot c does not have a mode; no score is more common than any other score.

4.0.2 Choosing a Measure of Central Tendency

Which measure to choose depends, in part, on the measurement level of the variable.

- **Nominal:** Mode
- **Ordinal:** Mode; if you calculate the mean or median, that means you assume that the distances between all categories are equal (i.e., you're treating your ordinal variable as interval).
- **Interval/Ratio:** Mode (but: it is rare for multiple interval/ratio scores to have identical values, unless they are integer), mean, and median

4.0.3 Measures of Dispersion

Measures of central tendency tell us what is a typical score; measures of dispersion tell *how typical* that score is. Dispersion simply means variability, so from now on, we will use this term.

Here are several measures of variability:

4.0.3.1 Range: full span

The range is the distance from the smallest to the largest value. You calculate it by subtracting the smallest value from the largest; for example, if your smallest value is 1 and the largest is 5, then the range is 5 1 = 4.

As a formula, this looks like:

$R = x_{\text{largest}}\ \ x_{\text{smallest}}$

Sometimes, the range is also reported as an interval, $[1, 5]$, or as minimum and maximum values. The range is an intuitive metric, but it is unstable because its value is fully determined by just two observations (the lowest and highest). The variability of all of the other observations does not affect it.

4.0.3.2 Sum of Squared Distances to the Mean

A metric of variability that does take all observations into account is the sum of squared distances to the mean - or "sum of squares".

To calculate it, follow these steps:

1. Calculate the mean of all observations, e.g. if our observations are `1,2,3`, then $M = 2$
2. For each observation, calculate the distance from that mean (subtract the mean), so for our observations `1,2,3`, we get 1 2 = 1, 2 2 = 0, and 3 2 = 1.
3. Square all these distances to get rid of negative values, so $1^2 = 1$, $0^2 = 0$, $1^2 = 1$.
4. Sum the squared distances, in this case $1 + 0 + 1 = 2$

As a formula, this looks like:

$SS = \sum_{i=1}^{n}(x_i\ \ x)^2$

Note that if we would not square the distances, the sum would always be zero because the mean is mathematically in the middle of all scores, so the negative distances of values below the mean exactly cancel out the positive distances of values above the mean.

The sum of squares has several important properties. First, note that its value depends on the sample size: sums of squares of larger samples tend to be larger than those of smaller samples.

Second, note that they are not on a very meaningful scale. Without further information, you cannot interpret what it means to say that the sum of squares for the variable age is 6524. Third, note that squaring distances does mean that high deviations become (quadratically) more influential:

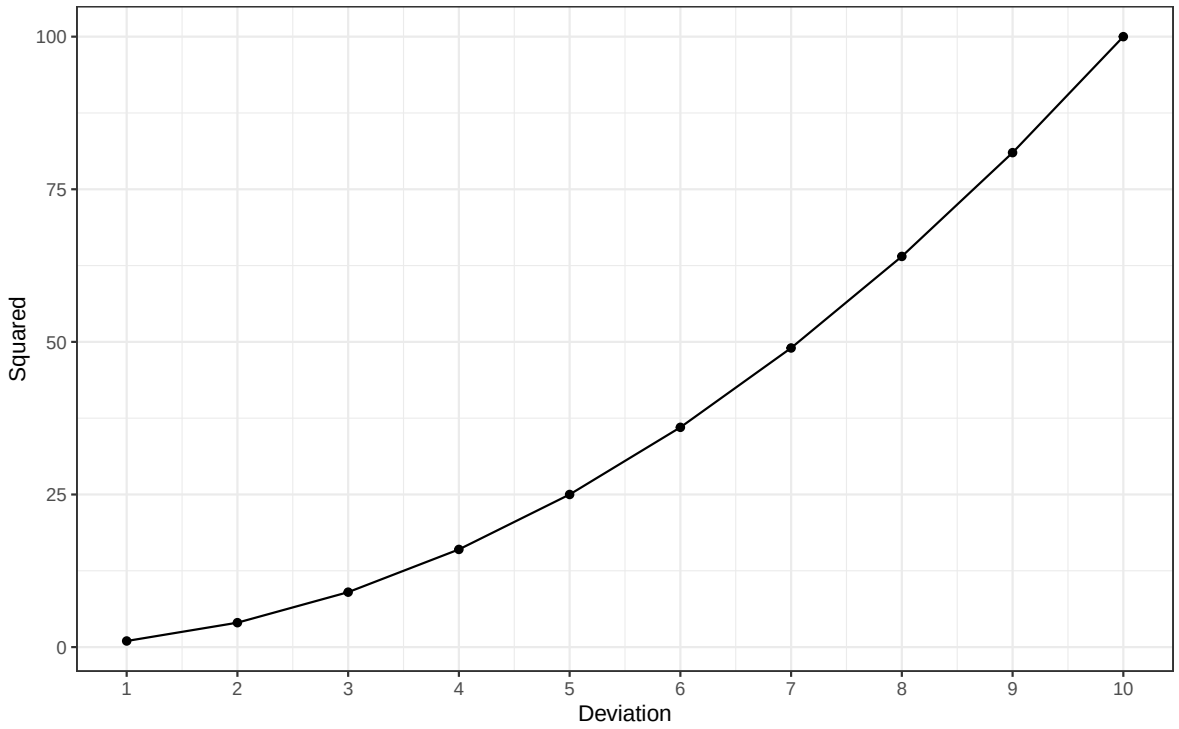


Figure 4.2: When squared, large values are more influential than small values.

4.0.3.3 Variance: mean squared distance

One way to make the sum of squares more interpretable is to divide it by the number of observations. This tells us how far away each observation is from the sample mean, on average.

Here are three formulas, that all describe the calculation of the variance. The first describes how you calculate the variance from the sum of squares (SS); the second includes the formula for the sum of squares, and the third describes how you calculate it by squaring the raw scores and subtracting a sum of n times the squared mean of X, $_x$:

$$s^2 = \frac{SS}{n} = \frac{\sum_{i=1}^{n}(x_{ix})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2 n_x^2}{n}$$

Note that here, we divide by the sample size n. When using the variance as a descriptive statistic, this is fine.

However, in later lectures, we will use sample statistics to make claims about the population (inferential statistics). Then, it becomes very important to divide by n 1 if the population mean is unknown. The formulas then look like this:

$s^2 = \frac{SS}{n1} = \frac{\sum_{i=1}^{n1}(x_i x)^2}{n1} = \frac{\sum_{i=1}^{n} x_i^2 n x^2}{n1}$

The consequence of dividing by n 1 is that we get a slightly higher value for the variance. We do this to account for the fact that we don't know the exact value of the population mean; we estimated it from the sample. If we just assume that the sample mean is a perfect representation of the population mean, we will systematically under-estimate the variance. By dividing by n 1, we get a slightly larger variance estimate, adjusted for our uncertainty about the value of the population mean.

4.0.3.4 Standard Deviation

One disadvantage of the variance is that it is still on the squared scale we obtained by squaring the distances. So, if your variable measures age in years, than the variance of age is expressed in years squared.

To restore the variance to the original units of the variable, we can simply take the square root. So if our variables are measured in euros, centimeters, and milliseconds - then the variances will be expressed in euros^2, centimeters^2, and milliseconds^2. Taking the square root restores the original units; we call the resulting statistic the *standard deviation*.

You can think of the standard deviation as the *average deviation* between individual scores and the sample mean. Why don't we just call it the "average deviation" then? Because that would be mathematically inaccurate - when we squared the deviations before taking the average, we allowed larger deviations to have a disproportionately larger impact on the value of the variance. Taking the square root of the end result, the variance, does not cancel out that disproportionate influence of large deviations.

So, intuitively it is fine to think of the standard deviation as the "average" deviation, as long as you're aware that mathematically, this is not exactly correct, because an average value should assign equal weight to each observation, whereas the standard deviation assigns greater weight to extreme observations.

Imagine I tell you that, in one class, the average grade is a 5, with a standard deviation of .5. You would know that most students scored close to a 5, and many of them failed the course. If I told you that the average grade is 5 with a standard deviation of 2, you would know that scores are much more spread out, and a large portion of the students must have passed the course as well.

4.0.4 Effects of Transformations & Outliers

Transformation	The **mean**...	The **SD**...
Add / subtract constant	Shifts by that constant	Doesn't change
Multiply / divide by constant	Scales by that factor	Scales by that factor
Inject one extreme score	Pulls center toward outlier	Increases

Example: Changing units (e.g., converting centimeters to inches) would rescale both the mean and SD.

4.0.5 Why Descriptives Matter

- **Data cleaning**: Outliers leap out when you know the usual range.
- **Analysis choices**: Skewed or heavy-tailed distributions may call for robust or non-parametric methods.
- **Transparency**: Readers can judge your results only if they see the data's headline features.
- **Communication**: "Participants averaged *8.9 hours of screen time per day* (SD = 1 hr)" paints an instant picture.

4.0.6 Descriptive VS Inferential Statistics

Descriptive statistics describe sample properties, while inferential statistics (which are covered in later chapters) give us a best guess for the corresponding population parameter. It is important to realize that some descriptive statistics are *calculated in the same way*, or very similar, as inferential statistics. The distinctive feature of descriptive statistics is how they are used: use them to describe the data observed in a sample, and not to make claims about the larger population.

4.0.7 Context of Discovery VS Justification

In the first chapter, we described De Groot's empirical cycle as a model of cumulative knowledge acquisition through scientific research. A crucial assumption of this cycle, highlighted by Wagenmakers and colleagues (see Figure 4.3), is the distinction between the context of discovery and the context of justification. The context of discovery is exploratory: we peruse data, looking for interesting patterns that might spark a new hypothesis. The context of justification is confirmatory: we test a theory-driven hypothesis. In order to obtain an unbiased test of a

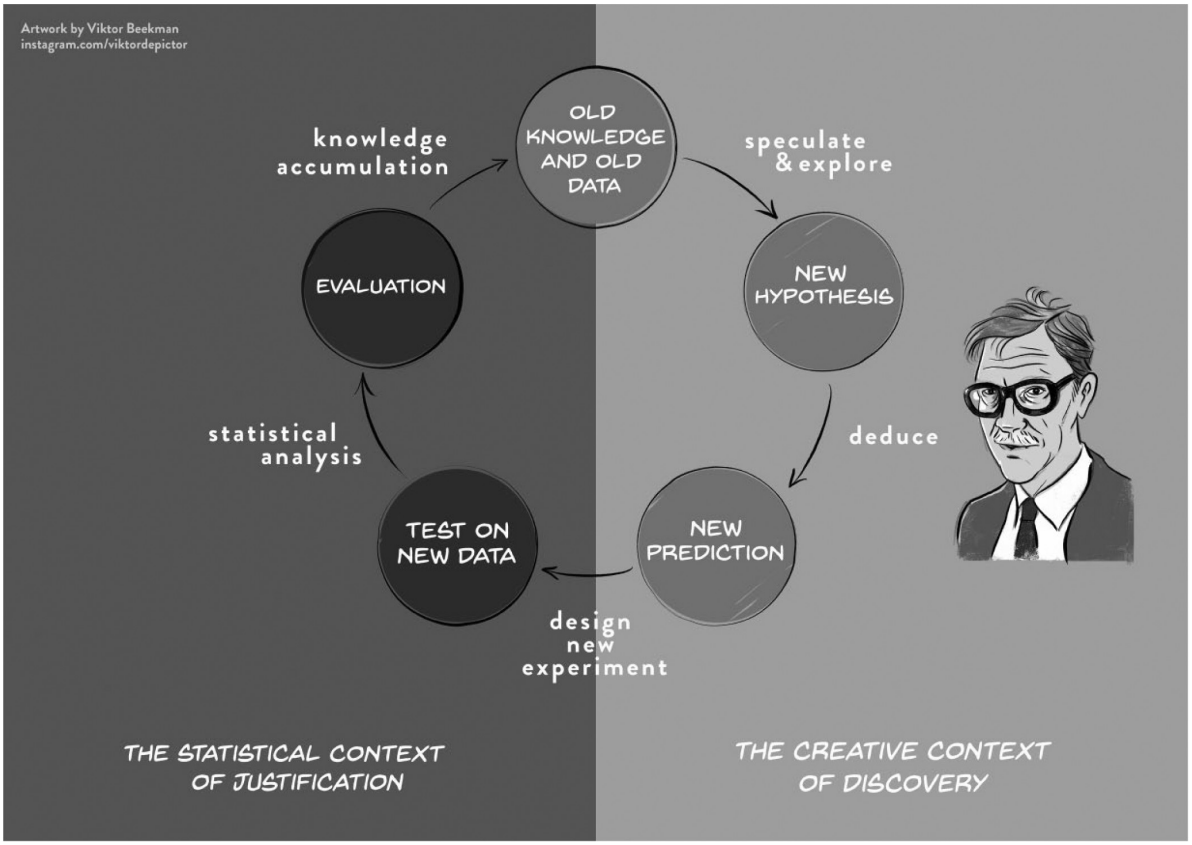


Figure 4.3: An interpretation of De Groot's empirical cycle, by Wagenmakers, Dutilh, & Sarafoglou, 2018. CC-BY: Artwork by Viktor Beekman, concept by Eric-Jan Wagenmakers

hypothesis, the hypothesis cannot be shaped by prior exploration of the data. If we first observe an interesting pattern in data (exploratory), and then conduct a test of that pattern (confirmatory), the test is more likely to confirm the pattern. There are legitimate ways to explore data looking for interesting patterns, and machine learning can be a helpful tool in this search (Van Lissa, 2022a). However, be careful of any cross-contamination between exploration and confirmation. Any pattern observed during exploration can introduce bias in subsequent confirmatory tests (Hoijtink et al., 2023).

Very often, the first thing researchers do when collecting or accessing a dataset is to engage in exploratory data analysis (EDA): visualizing and summarizing the data to detect errors, spot patterns, and generate new ideas or research questions. EDA was popularized by John Tukey, and it often involves histograms, boxplots, and scatterplots that reveal outliers or interesting trends worth modeling.

There is one important caveat to this common practice, namely that it introduces a potential risk of cross-contamination between exploration and confirmation. Observing the descriptive statistics may influence other downstream analysis decisions. This is not a problem when conducting purely exploratory analyses, and it is also not a problem if any confirmatory analyses have already been preregistered. Preregistration means that the analysis plans have been published in a time-stamped archive before collecting or accessing the data, so it's possible for others to check whether the reported analyses were executed as planned (Peikert et al., 2023), with changes made after seeing the descriptive analyses. In all other cases: be mindful of the risk of introducing bias. For an example of how preregistered hypothesis tests can be combined with rigorous exploration using machine learning, see Van Lissa (2022b).

4.1 Lecture

https://www.youtube.com/embed/zkEQmYjBHfE?si=DsDyXBztPmpgWTVQ

4.2 Formative Test

A formative test helps you gauge how well you've grasped the ideas and calculations from **Chapter 2 – Descriptive Statistics**. Try the quiz after working through the lecture slides but **before** our live meeting, so we can focus on any topics that still feel wobbly. If you miss a question you'll see a hint that points you back to the relevant slide or worked example.

Question 1

Which measure of central tendency is robust against extreme outliers? [1]

[1] Median

- (A) Mode
- (B) Mean
- (C) Weighted mean
- (D) Median

Question 2

For purely nominal data (e.g., eye color), the only valid measure of central tendency is the: [2]

- (A) Median
- (B) Mean
- (C) Mode
- (D) Geometric mean

Question 3

The range summarises spread by using: [3]

- (A) Only the minimum and maximum
- (B) Squared deviations
- (C) Every score in the set
- (D) Only scores below the mean

Question 4

Squaring deviations when computing variance ensures that: [4]

- (A) All deviations contribute positively to variability
- (B) Deviations stay in original units
- (C) Variance must be unbiased
- (D) Positive and negative deviations cancel

[2] Mode
[3] Only the minimum and maximum
[4] All deviations contribute positively to variability

Question 5

Adding a constant (e.g., +5) to every score will: [5]

- (A) Shift SD up by 5
- (B) Multiply SD by 5
- (C) Not affect the mean
- (D) Shift the mean up by 5, leave SD unchanged

Question 6

Multiplying every score by 3 will: [6]

- (A) Leave SD unchanged
- (B) Shift the mean by 3
- (C) Multiply both mean and SD by 3
- (D) Divide SD by 3

Question 7

The "degrees of freedom" for the sample variance (n-1) reflect that: [7]

- (A) The sample size is unknown
- (B) One piece of information is already used to estimate the mean
- (C) One piece of information is used to estimate the variance
- (D) Only n-1 scores are valid

Question 8

When a distribution is strongly right-skewed (e.g., income), which measure of central tendency best represents a typical observation? [8]

- (A) Mean

[5] Shift the mean up by 5, leave SD unchanged
[6] Multiply both mean and SD by 3
[7] One piece of information is already used to estimate the mean
[8] Median

- (B) Median
- (C) Mode
- (D) Range

Question 9

A distribution with two distinct peaks is called: [9]

- (A) Symmetric
- (B) Unimodal
- (C) Bimodal
- (D) Skewed

Question 10

Reporting the standard deviation alongside the mean helps readers understand: [10]

- (A) The measurement units
- (B) How tightly scores cluster around the mean
- (C) If there are outliers
- (D) The sample size

[9] Bimodal
[10] How tightly scores cluster around the mean

Show explanations

Question 1
The median depends only on rank order; extreme values cannot pull it up or down.
Question 2
Nominal categories lack numerical distance, so the most frequent category (mode) is the only appropriate centre.
Question 3
Range is computed as X_max and X_min, relying solely on the two extreme scores.
Question 4
Squares turn all deviations positive, preventing positive and negative differences from cancelling out.
Question 5
A constant shift moves the centre but does not change the spread of scores.
Question 6
Scaling stretches both centre and dispersion by the same factor.
Question 7
After fixing the sample mean, only n-1 unique pieces of information remain.
Question 8
The median is unaffected by the long tail of extreme high values.
Question 9
Two modes (peaks) indicate a bimodal distribution.
Question 10
SD converts variance back to original units, expressing average distance from the mean.

4.3 Tutorial

4.3.1 Descriptive Statistics

4.3.1.1 Step 1

As explained before, the first step in any statistical analysis involves **inspection of the data**. In the previous assignment we looked at graphical summaries.

This assignment shows you how to explore data using **descriptive statistics**—values such as the mean, standard deviation, maximum, and minimum.

Use the data file `stressLAS.sav`, as in the previous chapter.

4.3.1.2 Step 2 – Descriptives for Key Variables

We will first examine the descriptive statistics for **Optimism, Life Satisfaction,** and **Negative Emotions**.

Compute descriptive statistics as follows:

1. *Analyze > Descriptive Statistics > Descriptives*
2. Select **Optimism**, **Life Satisfaction**, **Negative Emotions**
3. Click **OK**

SPSS opens a new **Output** window with a table of descriptives for the selected variables.

4.3.1.3 Step 3 – Frequency Tables

In the previous step we computed the average value and standard deviations. However, for nominal and ordinal variables, the average value is meaningless. To explore nominal and ordinal variables we may produce **frequency tables**. A frequency table shows the observed percentage for each level of the variable.

Generate frequencies for **Smoke** and **Relation**:

1. *Analyze > Descriptive Statistics > Frequencies*
2. Select **Smoke** and **Relation**
3. Click **OK**

SPSS now adds a table with the frequency distributions of the selected variables to the output file.

Note: SPSS reports **Percent** and **Valid Percent**. These differ only when missing values are present (none in this dataset).

4.3.1.3.1 Extra – Spotting Multimodality

Sometimes a single mean or median masks sub-groups.

1. *Graphs > Legacy Dialogs > Histogram*
2. Choose **Life Satisfaction** for *Variable* and click **OK**

If you notice **two peaks**, color the bars by **Relation** (single vs. relationship):

1. *Graphs > Chart Builder*
2. Go to **Histogram** in the gallery section and drag **Stacked Histogram** onto the canvas
3. Place **Life Satisfaction** on the x-axis
4. Drag **Relation** into **Stack** box
5. Click **OK**

Take-away: multiple modes often reveal hidden clusters that may need separate analysis.

4.3.2 Quiz 1 – Basic Descriptives

How many participants are in the sample? _____[11]

What is the mean value of Optimism? _____[12]

For which of the variables is the spread in the scores highest? [13]

- (A) SATIS
- (B) OPTIM
- (C) NEGEMO

The minimum and maximum observed scores for Negative Emotions were: [__[14], ___[15]].

What percentage of participants is a non-smoker? _____[16]

What percentage of participants is in a relationship? _____[17]

4.3.2.1 Weighted Mean

Suppose Class A ($n = 12$, mean = 6) and Class B ($n = 8$, mean = 7) are merged.
SPSS effectively multiplies each mean by its n, sums those products, and divides by the **total** 20 students, yielding **6.4**.

Quick SPSS route

[11] 780
[12] 19.13
[13] OPTIM
[14] 3
[15] 37
[16] 48.1
[17] 47.9

- Merge the two files if separate (*Data > Merge Files*).
- Run *Analyze > Descriptive Statistics > Descriptives* on the combined score column.

4.3.2.2 Step 4 – Finding Erroneous Values

One reason to inspect descriptives first is to spot **erroneous values** (e.g., age 511 instead of 51).

Use the descriptives to find any out-of-range values, then:

1. *Data > Sort Cases*
2. Sort the suspect variable ascending or descending
3. Delete rows with invalid values

At this stage we remove entire cases; later you'll learn gentler missing-data techniques.

4.3.2.3 Step 5 – Group Comparison with Split File

Research question: *"Are non-smokers more satisfied with life than smokers?"*

1. *Data > Split File > Compare Groups* choose **Smoke**
2. Run *Analyze > Descriptives* on **Life Satisfaction**

SPSS now outputs separate means for smokers and non-smokers.

4.3.3 Quiz 2 – Group Means

Was there an erroneous value in the data file? Enter it here: _____[18]

If you delete that value, how will the **mean** change? [19]

- (A) Stays the same
- (B) Becomes larger
- (C) Becomes smaller

[18] 220
[19] Becomes smaller

If you delete that value, how will the **standard deviation** change? [20]

- (A) Becomes smaller
- (B) Becomes larger
- (C) Stays the same

Who is more satisfied in this sample? [21]

- (A) Smokers
- (B) Non-smokers

Does this difference necessarily hold in the population? [22]

- (A) No
- (B) Yes
- (C) Can't tell

4.3.3.1 Step 6 – Quick Check: How Recoding Affects Spread

Add 10 points to every Life-Satisfaction score:

1. *Transform > Compute Variable*
2. Target variable: `SATIS_plus10`
3. Numeric expression: `SATIS + 10` **OK**

Run Descriptives on both variables:

- Mean shifts up by 10
- **SD is unchanged**

Multiply by 3 (expression `SATIS * 3`):

- Mean Œ 3

[20] Becomes smaller
[21] Smokers
[22] Can't tell

- **SD Œ 3**

4.3.4 More Descriptive Statistics

Describing the data is an essential first step in any research context.

4.3.4.1 Central Tendency by Hand

Grades: 6 3 4 6 7 6 8 9 10 9

Compute **mean, median, mode** by hand.

Remind me how

- Mode = most common value
- Median = middle value (or midpoint)
- Mean = sum / n

4.3.4.1.1 Quiz 3 – Hand Computation

Mean _____[23]

Median _____[24]

Mode _[25]

4.3.4.2 Variation by Hand

Grades: 2 7 6 7 8 9

Compute **variance** and **standard deviation**.

Remind me how

Variance = average squared distance from mean
SD = variance

[23] 6.8
[24] 6.5
[25] 6

Why divide by $n - 1$?
After the mean is fixed, only $n - 1$ deviations are free to vary, so dividing by $n - 1$ keeps the sample variance unbiased.

4.3.4.2.1 Quiz 4 – Hand Computation

Variance _____[26]

Standard Deviation _____[27]

4.3.4.3 Verifying in SPSS

Enter the six grades, name the variable, then:

1. *Analyze* > *Descriptive Statistics* > *Descriptives*
2. **Options... > Variance Continue > OK**

Confirm SPSS matches your hand calculations.

[26] 5.9
[27] 2.43

5 Bivariate Descriptive Statistics

In the previous chapter, we covered univariate (= single variable) descriptive statistics. In this chapter, we introduce the first *bivariate* (= two variables) descriptive statistics. Let's take stock of the road so far, and set a goal for where we want to go. Last week, we ended with the variance. The variance is a statistic that tells us, on average, how much people's scores deviate from the mean. Today we will move into bivariate descriptive statistics. We will learn about the *covariance*, which tells us: If someone's score on one variable deviates positively from the mean, is their score on another variable also likely to deviate positively from the mean? We will also learn about the *correlation*, which tells us: How strong is the association between two variables, and is it positive or negative?

In this chapter, we will talk about two hypothetical variables, X and Y. In your mind, you can substitute any two variables you like; for example, X = *hours studied*, Y = *grade obtained*, or X = *extraversion*, Y = *number of friends*.

Before arriving at the correlation coefficient, statisticians often begin with **covariance**—a preliminary measure of how two variables vary together. Covariance reflects direction: it is positive when high values of X accompany high values of Y, and negative when they move in opposite directions. However, its numerical value is not directly interpretable because it is tied to the units of the measurement. A covariance expressed in centimeters and kilograms will differ from one computed in meters and pounds, even if the underlying association remains unchanged. As a result, covariance cannot meaningfully convey the *strength* of a relationship—only whether the variables tend to move in the same or opposite directions.

It provides a concise summary of the association between pairs of scores across individuals. For example, a researcher might retrieve each student's high school GPA (a measure of academic performance) and pair it with their family's annual income. The goal is to determine whether higher grades tend to correspond with higher income. In correlational studies, each individual contributes two measurements, commonly referred to as X and Y forming the foundation for analysis.

The correlation coefficient is a statistic that quantifies the strength and direction of association between two variables. It tells us the degree to which two variables move together. One way to think of the correlation coefficient is as a *bivariate (= two variables) descriptive statistic*.

To explore this relationship visually, researchers often rely on scatter plots. In a scatter plot, X values appear along the horizontal axis and Y values along the vertical. Each point on the plot corresponds to one participant's pair of scores. These plots allow immediate detection of

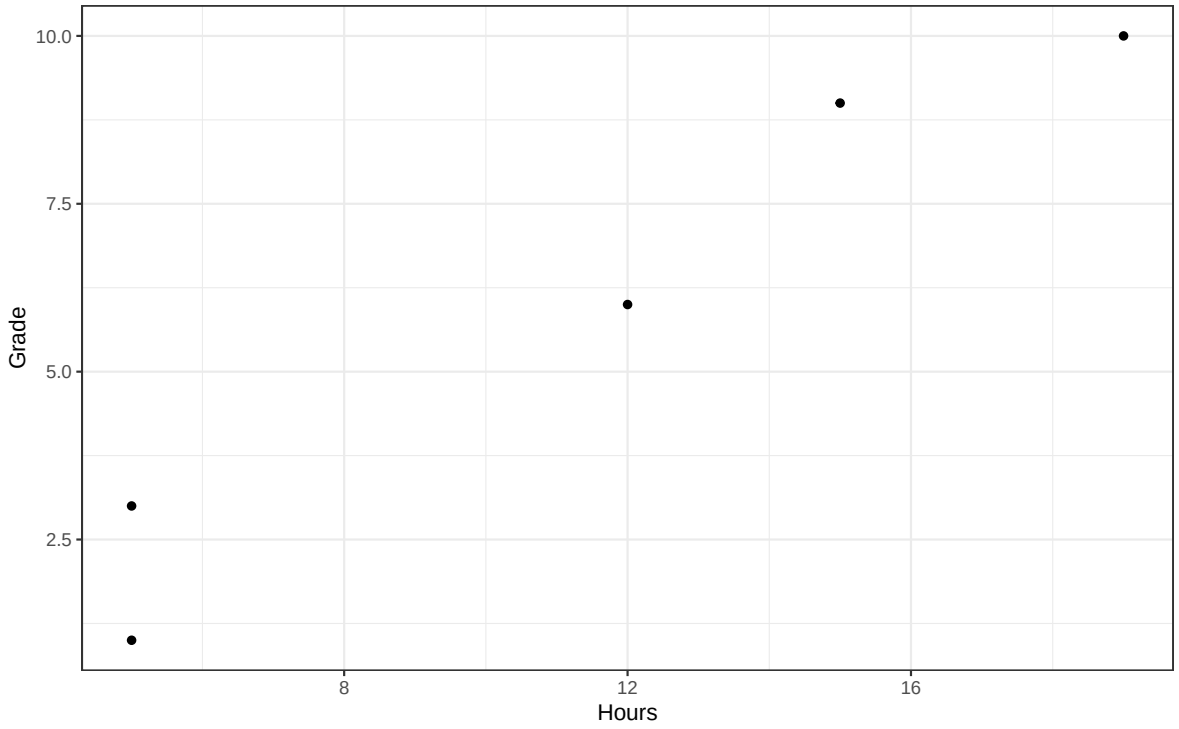
linear trends, and outliers—patterns that may remain obscured when examining data in purely numerical or tabular form.

5.0.1 Covariance

The word "covariance" means: varying, or moving, together. Let's have a look at mock data from five students on hours studied and final grade obtained:

Hours	Grade
5	3
15	9
12	6
5	1
19	10

We can visualize these data using a "scatterplot"; a simple graph where each observation is shown as a dot with X-coordinate determined by their value on the X variable (Hours), and Y-coordinate determined by the Y variable (Grade):



Notice that, if you squint, it appears like there might be some pattern in the data: more hours studied tends to go hand in hand with a higher grade. There might be a positive association between these variables! In the next sections, we go about quantifying this association numerically, step by step.

5.0.2 Sum of Products (SP)

The first stage in quantifying the association between two variables is to compute the **sum of products of deviations** (SP). The SP is similar to the sum of squares (SS), but whereas the SS captures the variability of one variable, the SP measures how two variables vary together.

To calculate the SP, take the following steps:

5.0.2.0.1 Step 1: Calculate the variables' means

Take the mean of each column (bold in the table below):

X	Y
5	3
15	9
12	6
5	1
19	10
11.2	**5.8**

5.0.2.0.2 Step 2: Calculate Deviations

For each variable, calculate the deviations by subtracting the mean from the observed scores:

X	Y	X-mean(X)	Y-mean(Y)
5	3	-6.2	-2.8
15	9	3.8	3.2
12	6	0.8	0.2
5	1	-6.2	-4.8
19	10	7.8	4.2

5.0.2.0.3 Step 3: Multiply Deviations

If we were to calculate the SS, we would now square the deviations and add them up in each column. To get the SP, instead of squaring the deviations - we multiply them across variables. Note that if the deviations for both variables have the same sign, then this will give a positive result (positive times positive is positive, and negative times negative is positive too). Moreover, if the deviations from both variables are high, the product will be a high number too. So the SP tends to be a large positive number if high positive (or negative) deviations on one variable go hand in hand with high positive (or negative) deviations on the other variable.

X	Y	X-mean(X)	Y-mean(Y)	Product
5	3	-6.2	-2.8	17.36
15	9	3.8	3.2	12.16
12	6	0.8	0.2	0.16
5	1	-6.2	-4.8	29.76
19	10	7.8	4.2	32.76

Now, we calculate the SP just by taking the sum of the column of products: 92.2.

Note that if the SP is positive, then there is a positive association between the variables; if it is negative, there is a negative association. In this case, the association is positive.

Here is a formula describing what we just did: we took the sum of the product ()() of the deviations of X from the mean of X, $X\ \ X$ times the deviations of Y from the mean of Y, $Y\ \ Y$:

$$SP = (X\ \ X)(Y\ \ Y)$$

5.0.2.1 Covariance

To get the covariance from the sum of products, we divide by the sample size, so in this case, $\frac{92.2}{5}$.

Another way to think about this is: we standardize the SP by the sample size n. This gives us the "average co-deviation" per participant. That number is called the covariance.

If the covariance is positive, there is a positive association between the two variables. If it is negative, there is a negative association.

But how strong is the association? It is hard to say, because the *size* of the covariance depends on the units and scale of the two variables involved.

5.0.3 Correlation

To answer the question of how strong the association is, we must standardize the covariance to drop the units of both variables. This gives us the so-called **Pearson correlation coefficient** (r). Specifically, the covariance is divided by the product of the standard deviations of X and Y. This standardization results in a number between -1 and +1, where 0 means no association, -1 means perfect negative association, and +1 means perfect positive association. This number, the correlation coefficient, tells us both the **direction** (-/+) and **strength** (value) of association between two variables. Because the correlation coefficients is unit-free, or standardized, it can also be compared across variables measured on different scales and across studies.

5.0.4 Limitations

While correlation coefficients are useful, they must be interpreted with care.

To illustrate the limitations of correlations, the statistician Anscombe (1973) created four data sets with identical correlation coefficients, $r = 0.82$. When plotting the data, however, it becomes clear that the correlation coefficient can only be meaningfully interpreted for the first dataset (figure a below).

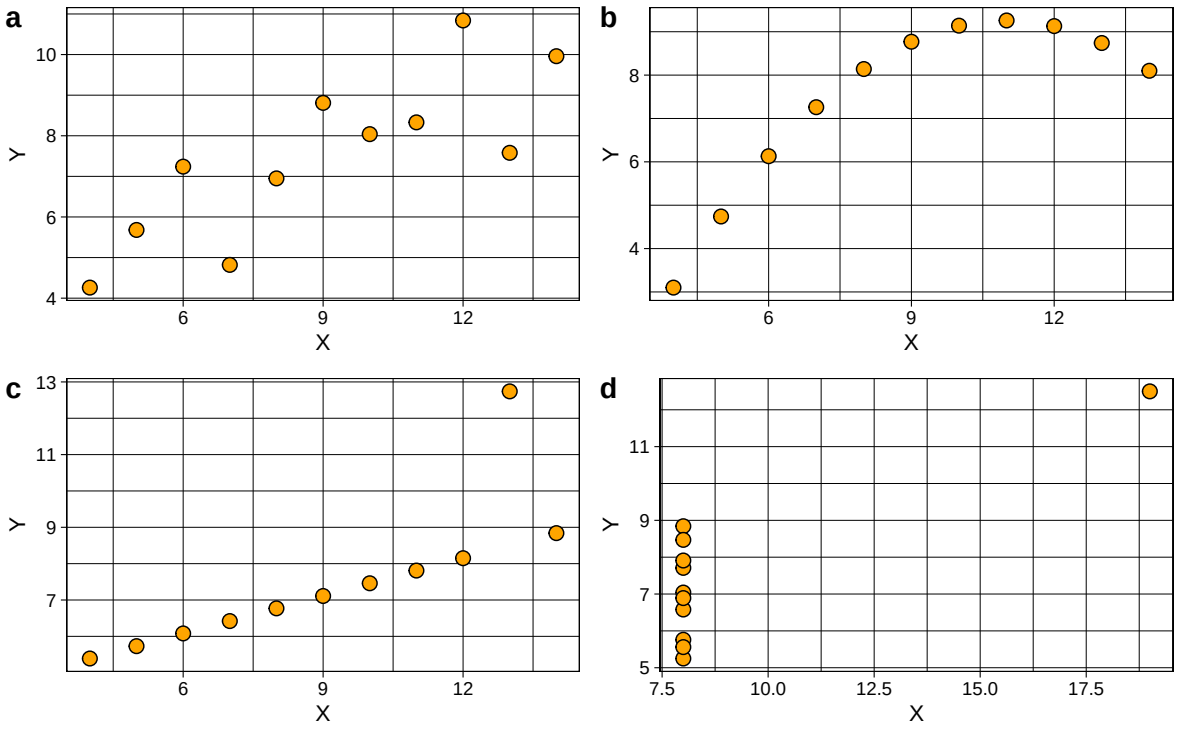


Figure 5.1: Anscombe's quartet, 1973

The first and most important limitation is that, Pearson's correlation coefficient only meaningfully captures *linear associations*, or: patterns that look like a straight line. Note that figure a in Figure 5.1 shows such a linear pattern of association; the correlation coefficient of $r = .82$ tells us that there is a strong - but not perfect - positive association.

Figure b in Figure 5.1 , on the other hand, shows a *perfect non-linear* association. All dots are perfectly in line; the line is just not straight. This illustrates that Pearson's correlation coefficient is not suited for capturing non-linear patterns, even if a strong relationship exists in another form.

Figure c shows a correlation of $r = 1$ for most of the points - but one outlier brings it down to $r = .82$.

Figure d shows no association at all for most of the points (they all have the same value for X, and if X does not vary, it cannot covary/correlate with Y) - but a single outlier makes it look like there is a strong correlation..

Secondly, these plots illustrate that **outliers** can have a disproportionate impact. In figures c and d, a single extreme observation artificially deflates (c) or inflates (d) the correlation coefficient, potentially leading to misleading conclusions.

Thirdly, a **restricted range** of scores can obscure or distort relationships. For example, if you were to examine the pattern in figure b of Figure 5.1 for values of X between [4, 9], you would conclude that $r = 0.99$, or near perfect positive correlation. If you examined the same pattern for values of X between (0, 13), you would conclude that $r = 0.07$, or near-zero. If you examined the same pattern for values of X between [13, 20), you would conclude that $r = 1.00$, or perfect negative correlation. Figure Figure 5.2 below zooms into the pattern from figure b, by restricting the range of variable X into three segments:

Restriction of range can easily happen in real life. For example, if your sample only consists of university students, you will probably have restriction of range on IQ.

Finally, you may have heard the phrase **correlation does not imply causation**. Observing a strong association between two variables does not mean that one causes the other. In general, it is not possible to conclude causality from statistics: causality is an assumption, which can be either supported by a theory, or by a particular methodology. In a randomized controlled experiment, participants are randomly assigned to receive either a treatment or control condition. Thus, any differences between the two groups should be due to the experimental treatment, or random chance. We will revisit the topic of causality later.

Anscombe's quartet is a good illustration of the limitations of causality, and also demonstrates the value of **visually inspecting your data (including with scatter plots)** before interpreting any statistics.

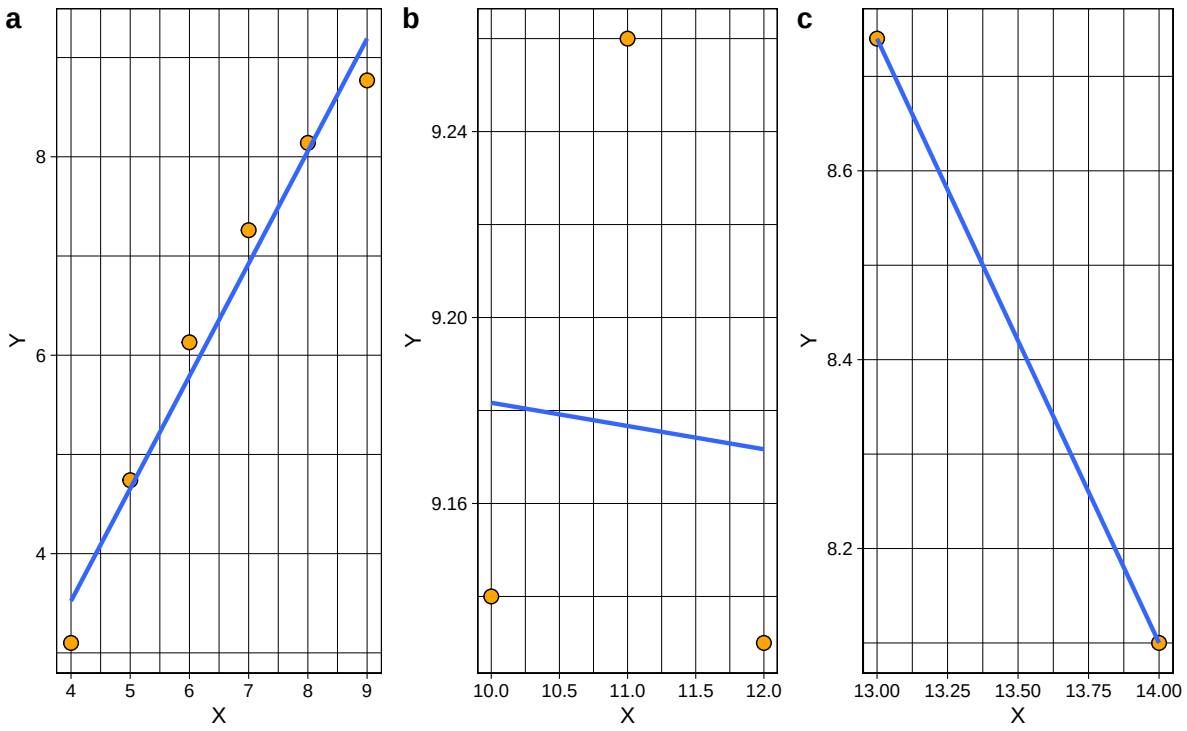


Figure 5.2: Zooming in on panel b of Anscombe's quartet, restricting the range of X

5.0.5 Summary

In summary, **covariance** offers an initial metric for gauging whether two variables tend to vary in the same or opposite direction. However, because its magnitude depends on the measurement units of the variables involved, it cannot be directly interpreted in terms of strength. The **Pearson correlation coefficient (r)** addresses this limitation by standardizing the covariance, yielding a unit-free statistic bounded between –1 and +1. This standardized measure expresses both the **direction** and **strength** of a linear relationship, enabling meaningful comparisons across contexts. Nevertheless, interpreting correlations requires caution, particularly with respect to restricted sampling ranges, the influence of outliers, and the fundamental distinction between correlation and causation.

5.1 Lecture

https://www.youtube.com/embed/jNI0YNjpvHs?si=hSdVTaoX1geeVMzn

5.2 Formative Test

This short quiz checks your grasp of **Chapter 3 – Covariance & Correlation**.
Work through it after you've studied the lecture slides (and before the next live session) so we can focus on anything that still feels uncertain. Each incorrect answer reveals a hint that sends you back to the exact slide or numerical example you need.

Question 1

A covariance of +120 cmkg tells you… [1]

- (A) 120 % of Y variance explained
- (B) A perfect linear trend
- (C) The units have been standardised
- (D) X and Y tend to rise together

Question 2

If the covariance between study hours and stress level is negative, what does that imply? [2]

[1] X and Y tend to rise together
[2] Longer study -> lower stress

- (A) No relationship
- (B) Longer study -> higher stress
- (C) Longer study -> lower stress
- (D) Units are incomparable

Question 3

Which statement about covariance magnitude is true? [3]

- (A) Its size depends on measurement units
- (B) It equals the regression slope
- (C) It ranges only from -1 to +1
- (D) A larger value always means a stronger relationship

Question 4

Converting temperatures from Celsius to Fahrenheit will make the covariance between temperature and ice-cream sales... [4]

- (A) Increase by a constant factor
- (B) Switch sign
- (C) Become unitfree
- (D) Stay exactly the same

Question 5

Pearson's r is best described as... [5]

- (A) Raw measure of joint variability
- (B) Mean of X and Y combined
- (C) Ratio of two variances

[3] Its size depends on measurement units
[4] Increase by a constant factor
[5] Standardised (unit-free) covariance

- (D) Standardised (unit-free) covariance

Question 6

If r = 0, we can conclude that... [6]

- (A) X causes Y
- (B) X and Y are unrelated in every way
- (C) No linear relationship is present
- (D) The data contain no outliers

Question 7

A positive covariance but r ~= 0.05 usually indicates that... [7]

- (A) Data range is restricted to zero
- (B) The variables move together only slightly
- (C) The relationship is strong
- (D) Units have been standardised

Question 8

Which scatterplot feature primarily determines the sign of covariance (and r)? [8]

- (A) Sample size
- (B) Overall slope direction
- (C) Point density
- (D) Presence of a mode

Question 9

You multiply every X score by 10 but leave Y unchanged. What happens? [9]

[6] No linear relationship is present
[7] The variables move together only slightly
[8] Overall slope direction
[9] Covariance x 10; r unchanged

- (A) Both covariance and r unchanged
- (B) Covariance unchanged; r x 10
- (C) Covariance x 10; r unchanged
- (D) Covariance x 10; r x 10

Question 10

A covariance of 0 implies that... [10]

- (A) They have opposite scales
- (B) Their linear relationship (r) is 0
- (C) X causes Z
- (D) X and Y are unrelated in every way

Question 11

Two variables show r = 0.85. Which conclusion is justified? [11]

- (A) X causes Y
- (B) X and Y are associated; causal direction unknown
- (C) A third variable is impossible
- (D) Y causes X

Question 12

Analysing data with a restricted range typically makes r... [12]

- (A) Smaller in magnitude
- (B) Exactly zero
- (C) Larger in magnitude
- (D) Change sign

[10] Their linear relationship (r) is 0
[11] X and Y are associated; causal direction unknown
[12] Smaller in magnitude

Question 13

An extreme outlier that follows the overall trend will most likely... [13]

- (A) Remove measurement error
- (B) Inflate the magnitude of r
- (C) Drive r toward zero
- (D) Make covariance negative

Question 14

A coefficient of determination (r2) of 0.49 means that... [14]

- (A) 49 % of X variance is explained by Y
- (B) The correlation is -0.70
- (C) Covariance is unitfree
- (D) 49 % of Y variance is explained by X

Question 15

After converting both X and Y to zscores, the covariance of those zvariables equals... [15]

- (A) Their geometric mean
- (B) Always zero
- (C) Sample size
- (D) Pearson's r

[13] Inflate the magnitude of r
[14] 49 % of Y variance is explained by X
[15] Pearson's r

Show explanations

Question 1
Positive sign = same-direction movement; magnitude is unit-dependent so strength not directly interpretable.
Question 2
Negative covariance means high X pairs with low Y and viceversa.
Question 3
Rescaling either variable rescales covariance; therefore magnitude alone is not comparable across units.
Question 4
Multiplying Celsius by 1.8 and adding 32 rescales covariance by 1.8; adding a constant does not affect it.
Question 5
Dividing covariance by the product of SDs removes units and bounds the result between -1 and +1.
Question 6
r only detects linear association; other patterns may still exist.
Question 7
Small r means weak linear association despite positive sign.
Question 8
Positive slope -> positive sign; negative slope -> negative sign.
Question 9
Scaling one variable scales covariance by that factor but leaves r (unitfree) unchanged.
Question 10
Zero covariance means no linear comovement; nonlinear links could still exist.
Question 11
Correlation quantifies association but cannot establish causality without experimental control.
Question 12
Less variability reduces covariance relative to the SDs, shrinking r.
Question 13
Trendconsistent outliers add leverage, increasing |r|.
Question 14
r2 translates correlation into varianceexplained terms.
Question 15
Standardising divides by SDs, so covariance in zspace equals r.

5.3 Tutorial

5.3.1 Load Data

Open LAS_SocSc_DataLab2.sav.
The file contains six variables (`X1` ... `X6`). You'll inspect three bivariate relationships.

5.3.1.1 Plot the pairs

Generate three simple scatterplots:

1. `Graphs Legacy Dialogs Scatter/Dot` **Simple Scatter**
2. Pairings & axis order
   - `X1` (X-axis) vs `X2` (Y-axis)
   - `X3` (X-axis) vs `X4` (Y-axis)
   - `X5` (X-axis) vs `X6` (Y-axis)
3. *Paste* and *Run* each syntax block.

Describe **linearity**, **direction**, and **strength** for each plot.

"The relationship between `X1` and `X2` is positive." TRUE / FALSE[16]

"The relationship between `X5` and `X6` is positive." TRUE / FALSE[17]

"The relationship between `X1` and `X2` is linear." TRUE / FALSE[18]

"The relationship between `X3` and `X4` is linear." TRUE / FALSE[19]

Strength of `X1`–`X2`:
[20]

- (A) zero
- (B) strong

[16] TRUE
[17] FALSE
[18] TRUE
[19] FALSE
[20] moderate

- (C) weak
- (D) moderate

Strength of `X3`–`X4`:
[21]

- (A) strong
- (B) weak
- (C) moderate
- (D) zero

Strength of `X5`–`X6`:
[22]

- (A) weak
- (B) zero
- (C) strong
- (D) moderate

5.3.1.2 Correlation coefficients

Even when the pattern is non-linear it's useful to see why Pearson r can mislead.

Analyze Correlate Bivariate
Select all six variables **OK**.

`X1`–`X2` correlation: _____[23]

`X2`–`X6` correlation: _____[24]

`X3`–`X4` correlation: _____[25]

---

[21] strong
[22] moderate
[23] 0.5
[24] 0.06
[25] -0.8

Can we interpret `X3`–`X4`'s r at face value?
[26]

- (A) No, assumption of normality violated
- (B) Yes, otherwise SPSS would give an error
- (C) No, assumption of linearity violated
- (D) No, assumption of association violated

Interpret `X5`–`X6`:
[27]

- (A) Moderate positive
- (B) Weak positive
- (C) Weak negative
- (D) Moderate negative

Take-away: Pearson's r is good at detecting linear patterns (like X1–X2), but it may be close to zero even when the variables have a strong curved pattern (like X3–X4).

5.4 Correlation – Work Dataset (`Work.sav`)

Having practiced on simulated data, let's now apply the same workflow to a real dataset related to the workplace.

Data File: `Work.sav`

5.4.0.1 Why inspect the plot first?

Before trusting Pearson r we check for

- an **approximately linear** pattern, and
- **extreme values** that could distort the statistic.

[26] No, assumption of linearity violated
[27] Moderate negative

Select the correct reason:

[28]

- (A) To check if the relationship is strong enough
- (B) To check if the relationship is positive
- (C) To check if the relationship is linear

5.4.0.2 Create the scatter-plot

`Graphs Legacy Dialogs Scatter/Dot` **Simple Scatter**

- X-axis =`scmental` (Mental Pressure)
- Y-axis =`scemoti` (Emotional Pressure)

Paste and *Run*.

The cloud of data points is roughly **linear**: TRUE / FALSE[29]

There are obvious **outliers**: TRUE / FALSE[30]

Approximate strength:

[31]

- (A) Moderately weak and negative
- (B) Moderately weak and positive
- (C) Moderately strong and positive
- (D) Moderately strong and negative

[28] To check if the relationship is linear
[29] TRUE
[30] FALSE
[31] Moderately strong and positive

5.4.0.3 Compute Pearson *r*

`Analyze Correlate Bivariate` (`scmental`, `scemoti`) **OK**

The correlation coefficient is (2 decimals): _____[32]

Interpretation:

[33]

- (A) There is no relationship between mental and emotional pressure.
- (B) There is a relationship between mental and emotional pressure.
- (C) We cannot draw a conclusion on whether or not there is a relationship.

Take-away: Mental and emotional pressure show a **moderately strong, significant positive relationship**—employees who feel more mentally pressured also tend to feel more emotionally pressured.

[32] 0.54
[33] There is a relationship between mental and emotional pressure.

6 Probability Distributions

Probability refers to the likelihood or chance of an outcome occurring in a random experiment. It is defined as the proportion of times that a particular outcome is expected to occur if the experiment is repeated an infinite number of times.

A random experiment is a process with multiple potential outcomes that could theoretically be repeated under similar conditions. For example, flipping a coin is a random experiment, and before flipping the coin, the outcome is a random experiment with a probability of getting heads or tails of 50% each. Once the coin is flipped, the outcome becomes fixed (the opposite of random), resulting in either heads or tails.

In a way, when you draw samples from a population and observe the values of particular variables (e.g., country of origin, height, age), you are performing random experiments. That means that, like with any random experiment, the values you are likely to observe also follow certain probability distributions. Discrete random variables have a finite or countable number of possible outcomes, such as the outcome of a coin toss. On the other hand, continuous random variables, such as the height of individuals, have an infinite number of possible outcomes.

For discrete (categorical) variables, we use discrete frequency and probability distributions, which summarize the observations and probabilities of each possible outcome, respectively. These distributions can be represented using frequency distributions, contingency tables, or bar charts.

Frequency distributions summarize observed outcomes in a sample. For example, a frequency distribution can tell us the proportion of Dutch students in a class or the number of times a particular number was rolled on a die.

Contingency tables (also called crosstables) are used to describe the join frequency distribution, and possibly relationship, between two categorical variables. They show the frequencies of different combinations of values for the two variables.

We can use frequency distributions to estimate the probabilities of observing those outcomes in the future. To calculate probabilities from frequencies, we can use different approaches depending on the type of probability distribution we want. In general, dividing frequencies by the total number of observations (grand total) gives us probabilities. In contingency tables, marginal probability distributions are obtained by dividing the marginal totals (row sums or column sums) by the grand total, which provides us with a probability distribution for each separate variable. Conditional probability distributions are derived by dividing a specific row

or column by the row- or column total (marginal total), and tells us the probabilities of one variable given a specific value of the other variable.

In continuous probability distributions, the possible outcomes are infinite and described by a continuous function. One common example is the normal distribution, also known as the bell curve. It is a symmetric distribution that extends from negative infinity to positive infinity, and it is characterized by two parameters: its mean (average) and standard deviation (measure of dispersion). The square of the standard deviation is called the variance.

The *standard* normal distribution, also known as the Z-distribution, is a standardized version of the normal distribution, rescaled to have a mean of 0 and a standard deviation of 1. Standardizing normal distributions allows us to calculate probabilities more easily using standard normal distribution tables or calculators. We can then convert these probabilities back to the original units if needed.

Probability distributions can be used as models to describe/approximate the distribution of real data. Behind the scenes, we do this any time we describe the distribution of scores on a variable using its mean and standard deviation. While we often assumpe that variables are normally distributed, that assumption is not always accurate. For example, depression symptoms do not follow a normal distribution: Most people score near-zero on depression symptoms, and few people have higher scores (but these are also not normally distributed). In such cases of violations of the assumption of normality, the mean and standard deviation are not very informative. You may use other descriptive statistics, consider different probability distributions (outside the scope of this course), or discuss the limitations of the assumption of normality.

In conclusion, probability distributions provide a way to represent the probabilities associated with different outcomes of a random variable, whether discrete or continuous. By using probability distributions, we can report descriptive statistics, calculate probabilities, and make predictions about future observations.

6.1 Lecture

VIDEO ERRATA: from 10:10 - 10:50 I talk about the probability of Being Dutch and Having a Tattoo, but I'm calculating the probability of Being Dutch and Not Having a Tattoo (I misread the column labels).

https://www.youtube.com/embed/p9QWOXR4OT8

6.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

An HR advisor is looking for new employees for LEGO. He thinks it is important for them to be creative. Creativity is normally distributed in the population, with a mean of 180 and a standard deviation of 25. A higher score indicates higher creativity. The advisor only wants to select applicants that belong to the 0.015 proportion of most creative people. What cut-off/boundary score for creativity should the HR advisor use? [1]

- (A) 206.13
- (B) 234.25
- (C) 125.75
- (D) 180.38

Question 2

What is probability? [2]

- (A) A measure of association between two categorical variables.
- (B) The proportion of times an outcome would be observed in a random experiment if it were repeated many times.
- (C) The subjective chance of observing a specific outcome in a single random experiment.
- (D) The proportion of times an outcome was observed in a sample.

Question 3

What is a random experiment? [3]

[1] 234.25

[2] The proportion of times an outcome would be observed in a random experiment if it were repeated many times.

[3] A process with multiple potential outcomes that could be repeated under similar conditions.

- (A) A naturally occurring experiment; for example comparing participants who grow up in an area where the drinking water is rich in a particular mineral with control participants from another area.
- (B) A process with multiple potential outcomes that could be repeated under similar conditions.
- (C) Any process with multiple potential outcomes where the probability of each outcome is unknown.
- (D) An example of the experimental method where people are assigned to a treatment- or control group.

Question 4

What are discrete random variables? [4]

- (A) Variables with an infinite number of possible outcomes.
- (B) Variables with a normal distribution.
- (C) Variables with a probability distribution.
- (D) Variables with a finite or countable number of possible outcomes.

Question 5

What information is contained in a frequency distribution? [5]

- (A) A summary of the observed counts of discrete outcomes in a sample.
- (B) A summary of the observed counts of discrete outcomes in the population.
- (C) A summary of observed probabilities of discrete outcomes in a sample.
- (D) A measure of association between two categorical variables.

Question 6

What is the standard normal distribution? [6]

- (A) A contingency table summarizing two categorical variables.

[4] Variables with a finite or countable number of possible outcomes.
[5] A summary of the observed counts of discrete outcomes in a sample.
[6] A standardized version of the normal distribution with a mean of 0 and a standard deviation of 1.

- (B) Any continuous distribution with infinite possible outcomes.
- (C) A standardized version of the normal distribution with a mean of 0 and a standard deviation of 1.
- (D) A probability distribution where the total probability sums to 1.

Question 7

A researcher is interested in the relationship between movie watched and popcorn consumption. She counts the number of people who consume popcorn during a movie, and whether they have watched Mean Girls or not. The results are presented in the table below this quiz. What is P(Popcorn|Mean Girls), rounded to 3 decimal places? [7]

- (A) 0.04
- (B) 0.63
- (C) 0.08
- (D) 0.06

[7] 0.63

Show explanations

Question 1
Find the Z-score that matches a right tail probability of .015 and calculate $(Z\ 25) + 180$
Question 2
(The frequentist definition of) probability builds upon the idea that you could theoretically repeat the random experiment many times and calculate the proportion of times a given outcome is observed
Question 3
A random experiment is a process with multiple potential outcomes that could theoretically be repeated many times under similar conditions.
Question 4
Discrete random variables have a finite (=discrete) or countable number of possible outcomes.
Question 5
Frequency distributions are used to summarize observed outcomes in a sample.
Question 6
The standard normal distribution is a standardized version of the normal distribution with a mean of 0 and a standard deviation of 1.
Question 7
The question is about the *conditional probability* of having popcorn given that (|) someone watched mean girls; divide the 151 pp who match this description by the total number of people who watched mean girls.

6.2.0.1 Popcorn consumption Table

		Popcorn consumption		
		Yes	No	Total
Movie type	Mean Girls	151	90	241
	Other	1678	2130	3808
	Total	1829	2220	4049

6.3 Tutorial

6.3.1 Normal Distribution

In this assignment you will practice with the normal distribution.

The normal distribution deserves special attention as it is commonly used in statistics for the social sciences. The normal distribution was already derived in the 18 century by DeMoivre,

Adrian, and also Gauss*, and since then it played a central role in statistics. More importantly, many attributes in the social sciences are by close approximation normally distributed, as was first discovered by Quetelet. Hence, the normal distribution has great empirical relevance, which comes in handy for our research!

- For that reason, the normal distribution is sometimes referred to as a Gaussian curve.

Before we start, the concept of a random variable needs to be introduced first. A random variable is a numerical outcome of a chance experiment.

For example, a random variable is the number of pips when throwing two dice (here the chance experiment is throwing two dice). Also, the proportion of girls in a random sample of 10 children is a random variable (here the chance experiment is the random selection of 10 children).

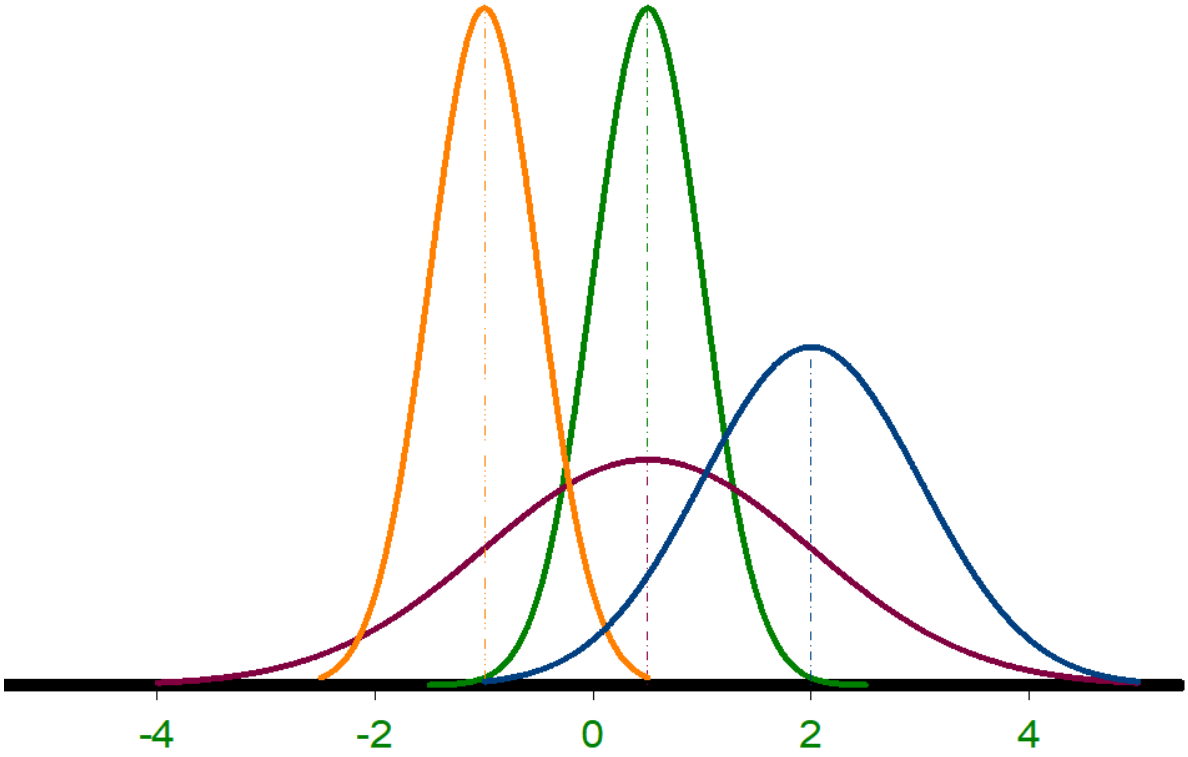
We also distinguish between continuous and discrete (categorical) random variables.

A continuous variable can take on infinitely many values. For example, the height of a person is a continuous variable. Take any two persons of different height, and we can always find a third person in between. Discrete variables can take on only particular values. For example, the number of correct of correct answers when a person blindly guesses all the answers is a discrete random variable.

- if you want to know, see for example Wikipedia

Normal distributions are distributions for continuous random variables.

Let's first have another look at different examples of normal distributions:



Dist.		
1	–1.0	0.5
2	0.5	0.5
3	0.5	1.5
4	2.0	1.0

Inspection of the graphs of the bell-shape distributions shows that it is a symmetric distribution. We also see that the distributions differ in location (the point on the x-axis where it reaches its maximum) and the spread. In other words, the distribution is characterized by two parameters: the mean and standard deviation. The mean is denoted by the Greek letter (pronounced as "moo") and standard deviation is denoted by the Greek letter (pronounced as "sigma").

Compare the distributions and see how the parameters determine the location (mean) and spread (standard deviation).

6.3.1.1 Quiz

Correctly complete the sentence below by filling in the gaps:

Given the example "A student guesses the correct answer to 10 multiple choice questions with four answer categories each" the random experiment is [8]

- (A) guessing the correct answer
- (B) the number of correct answers
- (C) the correct answer
- (D) the 4 answer categories

and the random variable is [9]

- (A) the number of correct answers
- (B) the correct answer
- (C) the 4 answer categories
- (D) guessing the correct answer

.

Are the following random variables discrete or continuous?

Number of heads in 10 throws with a fair coin is continuous. TRUE / FALSE[10]

Time by train from Tilburg to Eindhoven is continuous. TRUE / FALSE[11]

The number of correct answers on a test is continuous. TRUE / FALSE[12]

The mean height in a random sample is continuous. TRUE / FALSE[13]

The average number of correct answers in a random sample of 100 students is continuous. TRUE / FALSE[14]

If the standard deviation (SD) increases, the distribution becomes [15]

- (A) wider
- (B) narrower

[8] guessing the correct answer
[9] the number of correct answers
[10] FALSE
[11] TRUE
[12] FALSE
[13] TRUE
[14] TRUE
[15] wider

6.3.1.2 "The Empirical Rules"

As a first step we may consider some practical rules for working with the normal distribution, the so called "empirical rules". In particular, if a random variable is normally distributed the following empirical rules apply:

68% of the values lie within one standard deviation from the mean. 95% of the values lie within two standard deviations from the mean.

If we know that a variable is normally distributed, we can also compute probabilities of certain outcomes, so called events. For example, if we know that the scores on a test are normally distributed, we may want to know the probability that a randomly selected person has a score above a certain cut-off (i.e., satisfies a certain selection criterion).

In the next few steps, you will practice on how to compute probabilities under the normal distribution. To do so, we have to be able to work with the standard normal distribution (Z-distribution) and accompanying tables.

6.3.1.3 Quiz

Complete the following sentences:

IQ scores are normally distributed with mean 100 and an SD of 15. This means that 95% of the persons in the population has an IQ in between [16]

- (A) 70
- (B) 55
- (C) 15
- (D) 85

and [17]

- (A) 130
- (B) 115
- (C) 100
- (D) 145

[16] 70
[17] 130

.

Students loan after completing the bachelor is normally distributed with mean 1500 Euros and an SD of 150. This means that 68% of the students ends up with a loan between [18]

- (A) 1000
- (B) 1485
- (C) 1200
- (D) 1350

and [19]

- (A) 1800
- (B) 1515
- (C) 1450
- (D) 1650

.

What is the mean of the standard normal distribution? _[20]

What is the SD of the standard normal distribution? _[21]

6.3.1.4 Calculating probabilities

For the next series of exercises you need to use a Z-table or calculator (e.g., Excel, Google Sheets, R online, or the Z-table in this GitBook, Appendix B).

We have seen that probabilities are related to the area under the curve. This means that for continuous variables we can only find the probability that the outcome falls within a certain interval. For example, the time to complete a task is more than 10 minutes; the IQ is in between 70 and 90.

Note that there may be different ways to get to the correct answer.

Numerical Examples

[18] 1350
[19] 1650
[20] 0
[21] 1

6.3.1.5 Quiz

Consider a continuous variable X, which is normally distributed with $X\ (=30, =4)$.

Compute the following probabilities:

P(X>36.8): _____[22]

P(X<24): _____[23]

P(X<35): _____[24]

P(28<X<34): _____[25]

[22] 0.04
[23] 0.07
[24] 0.89
[25] 0.53

Explanation

P(X>36.8):

- Transform X into Z: (36.8 30)/4 = 1.7
- Read Upper Tail Area for Z = 1.7, which equals 0.0446
- Conclusion: P(X>36.8)= 0.0446.

P(X<24):

- Transform X into Z: (24 30)/4 = -1.5
- Because the Z-distribution is symmetric, we know that P(Z<1.5) is equal to P(Z>1.5). The latter probability can be found in the Z-table, which is 0.0668. This is also the probability we are looking for.
- Conclusion: P(X<24) = 0.0668.

P(X<35):

- The probability we are looking for equals 1 P(X>35). Thus, we first need P(X>35).
- Compute corresponding Z-value: Z = = 1.25.
- Look for the upper tail area in the Z-table: P(Z>1.25) = .1056.
- Thus, the area we are looking for is 1 .1056 = .8944
- Conclusion: P(Z<35) = 0.89444

P(28<X<34):

- Remark: there are different ways to come the answer. So the answer below is just one of few possibilities.
- We need the area under curve between 28 and 34. This area equals 1 minus the tail areas; that is, P(28<X<34) = 1 P(X<28) P(X>34)
- Lets start with P(X>34). First compute Z= = 1. Via the Z-table we find P(Z>1)= .1587.
- Now determine P(X<28). First compute Z= = 0.5. Via the Z-table we find P(Z<0.5) = P(Z>0.5) = .3085.
- Hence, we have 1 0.1587 0.3085 = 0.5328
- Conclusion: P(28<X<34) = 0.5328.

Students are looking for a new roommate. They read in an article that the time people spend under the shower is in the population normally distributed with mean of 10 and SD of 8 (measured in minutes). Suppose they randomly select a person as their new roommate.

What is the probability that this randomly selected person will spend more than 20 minutes under the shower? _____[26]

[26] 0.11

Suppose "confidence in society" is measured on a continuous scale from 0 (no confidence at all) to 100 (highly confident). Also suppose that confidence is normally distributed in the population with mean 52.6 and SD of 12. One speaks of low confidence if the score falls below 43.

What percentage of the population has low confidence? _____[27]

The scores on a test for aggressive behavior are normally distributed with mean 50 and SD of 10. The test is used to select police officers. In particular, only police officers with scores between 42 and 62 qualify for the job as they are not too aggressive and not too friendly either.

What percentage of the population qualifies as police officer? _____[28]

[27] 21.2
[28] 67.3

Explanations

Let X denote time people spend under the shower. We want to know the probability that the person showers for more than 20 minutes; that is, P(X>20) given = 10 and = 8.
Compute Z value: (20 10)/8 =1.25. Hence, we need P(Z>1.25); that is, the area beyond Z = 1.25.
Via the Z-table (look in the column labelled C) we find: P(Z>1.25) = .1056.
Conclusion: the probability that a random selected person will shower more than 20 minutes is 0.106 (about 11%).
Let X stands for the confidence level. X is normally distributed with mean = 52.6 and SD = 12. We need P(X<43). That is, we need the area under the curve to the left of 43.
First, transform to Z-scores: X=43 => Z = (43 52.6)/12 = 0.8. Thus, we need P(Z<.8).
The left-tail areas are not shown in the Z-table. Therefore, to find the area, we will first look for the area in the other tail; that is, we will look for P(Z>0.8) in the Z-table. The probability equals 0.2119. Because the distribution is symmetric, the left tail area is also 0.2119. This gives us the answer.
Conclusion: about 21.2% in the population has low confidence in society.
Let X be the test scores. We need to compute P(42<X<62). This area can not be directly found in the Z-table, so we have to take some additional steps. First, because the whole area under the curve is 1, we can say that the area we are looking for is equal to 1 minus the areas in the tail; thus, 1 P(X<42) P(X>62). These latter probabilities can be read from the Z-table!
Compute Z-values: 1 P(Z<0.8) P(Z>1.2). Remember that $Z = \frac{Observedmean}{SD}$.
Determine the tail areas: because the distribution is symmetric, we have P(Z<0.8)=P(Z>0.8). The latter can found in Z-table, which equals .2119. Probability P(Z>1.2) can be directly read from the Z-table, which is .1151.
Hence, 1 .2119 .1151 = 1.327 = .673.
Conclusion: 67.3% of the population qualifies as police officer.

Suppose the time to complete a certain task is normally distributed with mean 8.6 and standard deviation (SD) of 3.5.

What is the probability that a randomly selected person needs more than 11.4 minutes to complete the task? _____[29]

Again, suppose the time to complete a task is normally distributed with mean 8.6 and standard deviation of 3.5.

Complete the sentence:

"95% of the participants completes the task within 1.6 and _____[30] minutes."

[29] 0.21
[30] 15.6

Scores on a selection test are normally distributed with a mean of 500 and a standard deviation of 50. A person qualifies for the job if they score between 480 and 580.

What is the probability that a randomly selected person will qualify for the job? _____[31]

An IT company is looking for new programmers. To qualify for the job the programmers need to score high on conscientiousness. Therefore, the job applicants need to complete the Conscientiousness scale from the NEO-PI-R (a popular personality inventory for the Big Five personality traits*) as part of the selection procedure. Research has shown that in the population the scores on the scale are normally distributed with a mean of 133.4 and SD of 18.3. To qualify for the job, the conscientiousness of the programmer needs to be among the highest 20% in the population.

What cut-off should the company use to select new personnel? Round to a whole number. _____[32]

- The Big Five is a popular model for personality; see Wikipedia for more info on the Big Five.

Explanations

Question 1:

- Transform X into Z: (11.4 -8.6)/3.5 = 0.8
- Read Upper Tail Area for Z = 0.8, which equals 0.2119
- Conclusion: P(X>36.8)= 0.212

Question 2:
95% of the participants completes the task within 1.6 and 15.6 minutes. You can use the empirical rule that 95% of the observations falls within 2SDs from the mean. Thus, 95% of the observations lies within 8.62Œ3.5 and 8.6+2Œ3.5
And 8.6+2Œ3.5 = 15.6
Question 3:
We need P(480<X<580). This probability equals 1 P(X<480) P(X>580).
P(X<480) = P(Z<0.4) = P(Z>0.4) = 0.3446 (via Z-table)
P(X>580) = P(Z>1.6) = P(Z>1.6) = 0.0548 (via Z-table)
So the final answer equals: 10.34460.0548=0.6006 (0.601 when rounded at three decimal places).
Question 4:
To get to the correct answer, these are the steps: - First find the Z-value that marks the highest 20%. This value equals 0.84. - Then compute the corresponding cut-off on the X-scale: X = 0.84Œ18.3 + 133.4 = 148.772 - Rounded to the nearest integer equals 149.

[31] 0.601
[32] 149

6.3.2 Missing Values

For this assignment, and also later assignments, we will use a (real) data set on Type D personality and several background characteristics (age, gender, and education level (7 ordered levels)).

Type D personality is defined as the tendency towards negative affectivity (NA) (e.g., worry, irritability, gloom) and social inhibition (SI) (e.g., reticence and a lack of self-assurance). Theory suggests that Type D individuals have poorer health outcomes.

Type D is measured with the DS14 scale. The DS14 consists of 14 items, 7 measuring NA and 7 items measuring SI. Answers are given on a five point scale (scored 0 through 4).

- see also Type D personality on Wikipedia.

Open the data file `TypeDDataSSC.sav` in SPSS. The data file contains the scores on the DS14 items measuring Type D as well as the background variables for 80 respondents.

Go to the variable view. The content of the items are given under labels and it is indicated whether the item measures NA or SI.

6.3.2.1 Quiz

Is the first item in the DS14 an indicator of NA or SI? [33]

- (A) SI
- (B) NA

Go to the data view in SPSS and inspect the data.

Do you see any missings? [34]

- (A) Yes
- (B) No

What is the valid N of the variable age? ___[35]

How many system missings do we have on gender? _[36]

[33] SI
[34] Yes
[35] 77
[36] 8

6.3.2.2 Recoding missing values

Remember that Empty cells are called system missings. There are reasons to use user-specified missing codes instead; for example, this allows you to keep track of reasons for missingness (which enables you to report more comprehensively on your missing data).

So, for this exercise we will define a user-missing code for the missing values. Missing code is number that a researcher uses to designate that the value is missing. The code must be chosen such that it cannot be confused with actual scores. For example, for age the missing code can be 999, because 999 is an impossible age.

Now, we will first define missings for age.

Go to the data view; look at the values of age and whenever the value is missing fill in 999. (In total three persons had a missing on age; so you have to fill in 999 three times).

Go to the variable view. We have to define the missing code under missing. Click on the cell and [...], and SPSS opens a new window. Define 999 as missing code.

In the previous step we filled in the missing codes manually. For a small data set this is okay, but for large data sets (say thousands of persons on many variables) this would be problematic.

In the next few steps we will see how we can define the missings more easily.

To do so, we will use the function recode in SPSS. We will first apply the recoding to gender.

Before going into the recoding, let's first look at the frequency table for gender.

You may already have this output from answering the Quiz.

6.3.2.3 Recode into Same Variables

Replacing user missing values with a missing code using the recode option works as follows:

Navigate to Transform > Recode into the same variables. SPSS opens a new window.

Select gender as the variable to be recoded.

Click on Old and New Values. SPSS opens a new window. In this window we can specify the recoding. In our case we want to recode System Missing into 999. So, choose "system missing" as old Value, and specify 999 as new value, and click on Add below. (See the more info section below for the SPSS specifications.)

Click on Continue, and click on OK. SPSS now replaced system missing by 999. Go to the data view and verify that SPSS filled in 999 at the empty places.

Go to the variable view and specify 999 as the user missing code for gender. (In the same way as you did for age).

Compute the frequency table for gender again.

Verify that all "system missings" are now reported as "user missings" instead.

In the previous step we only recoded the missings for gender, but we could do that for all variables. It is most convenient to use a code that can be used for all variables. In this case we can use 999 as the missing code for all variables, it's easy because we can apply this recoding to all variables at once. Just follow the same steps as before, but now select all variables to recode.

Run the recode command for all variables.

Verify in the data file that SPSS replaced all system missings by 999.

Now we also have to specify in the variable view that 999 is the missing codes for all variables. We already changed it for age and gender. To do the same thing for other variables is easy; you can just use copy-paste! Click on Missing for gender, click on the right mouse button, choose copy. Go to the next variable, click on the right mouse button, and with paste you can define the missings for other variables.

Tip: If you like shortcuts: you can also click on missing, type Ctrl C, and then use Ctrl V to copy the information about the missings.

So, we specified the missing codes, but we also want to know for each respondent how many missings values he or she had. In other words, for each participant we want to count the number of missings. Participants with too many missings may be excluded from further analysis. Counting the number of missings per person can also be easily accomplished in SPSS.

Transform > Count Values within Cases. SPSS opens a new window. Specify the name of the target variable (e.g., CountMiss); this is the name of a new variable that gives the number of missings. You may also give the variable a label, say: "Number of Missings".

Select all variables.

Click on Define Values. SPSS opens a new window. Select System or User Missing at the left and click on add. Click on continue than OK.

SPSS will now create a new variable that shows how many missings there are for each person on the variables selected in the list.

Go to the data view and verify that SPSS added a new variable (i.e., a new column with values) named CountMiss.

6.3.2.4 Quiz

Compute the frequency table for the third DS14 item. How many missing values do we have on this item? _[37]

[37] 6

How many missings does person 8 have, using the variable CountMiss? _[38]

Compute the frequency table for CountMiss.

What is the maximum number of missings for the participants in this dataset? ___[39]

How many participants have this many missings? _[40]

How many participants have at least one missing value? ___[41]

Make sure you save the data file including the variable with the number of missings. We will use it in the next assignment. ‘

6.3.3 Select Cases and Split File

In this assignment we will take a closer look at selecting cases and how to do analyses for subgroups.

6.3.3.1 Selecting Cases

In the previous assignment we have seen that some respondents had one or more missing values. Suppose we want to discard these persons in the analysis, which means that for all the remaining analyses we only want to include participants with no missings. This method of handling missing data is called *listwise deletion*, and it is generally considered bad practice - but it’s also easy, so we will teach it in this course. More advanced courses cover expert methods of handling missing data.

Proceed as follows:

Data > Select cases. SPSS opens a new window.

Chose “If condition is satisfied” and chose ‘if’. SPSS opens a new window again.

Specify the condition `CountMiss = 0` to select cases with no missings.

Make sure that the output is specified as “Filter out unselected cases”. Click on OK.

Go to the data view.

Verify that SPSS crossed out cases with one or more missings.

Verify that SPSS added a new variable labelled ‘filter_$’. This is filter variable indicating who is included in the analyses (value = 1) or not (value is 0). If you remove the filter variable, SPSS will use all cases again.

[38] 3
[39] 14
[40] 1
[41] 17

6.3.3.2 Quiz

What is the mean for the variable age of the selected group? _____[42]

What is the valid sample size for that mean? ___[43]

Now, run the selection procedue again, but remove the incomplete cases from the data file.

Proceed as follows:

Data > Select cases Choose "Delete Unselected Cases" under output. Click OK. Verify that the incomplete cases are removed.

Because you have modified the data, it is prudent to save the new file under a different name (e.g., TypeD_selection.sav). Use this file with only the complete cases (i.e., TypeD_selection.sav) for the remaining steps.

6.3.3.3 Split File

Sometimes we want to do analyses for subgroups, especially when exploring the data for the first time. For example, we may want to have the descriptive statistics for males and females separately. One way to do this is to use the Split File option. With this option you can tell SPSS that you want to have tables for each subgroup separately.

Let's see how it works!

Proceed as follows:

Via menu follow Data > split file. SPSS opens a new window.

Choose 'Compare Groups' and choose gender as the variable for Groups Based on.

Click OK.

Important: Notice no output appears and no changes are made to the data. This makes sense because we haven't asked SPSS to generate any output nor to make changes in the data. Yet, SPSS now knows that he has to produce tables for males and females separately once we ask to generate output.

To undo the split file, proceed as follows.

Data > Split file Choose Analyze all cases, do not create groups. Click OK. SPSS now no longer produces the output per group.

[42] 42.68
[43] 63

6.3.3.4 Quiz

Have SPSS compute the mean and SD for age.

How many women are there in the sample? [44] **What is the mean age of men in the sample? _**[45] **What is the mean age of women in the sample? _____**[46]

One of the variables is Education Level. It is an ordinal variable with 7 levels, score 1 represents the lowest level of education, and score 7 the highest.

Compute the mean age per level of education.

For which education level was the mean age highest? _[47]

What was the value of the mean age for this education level? _____[48]

6.3.4 Recode and Compute

For this assignment we will continue with the data on Type D and the selection of complete cases.

Before you start, make sure that the Split File option is disabled.

In practice, you often have to do some data handling before you can actually start doing analyses. For example by coding missing values, or you may have questionnaire data for which some of the questions are formulated contra-indicative and therefore should be reverse coded. Another reason would be that you may have to compute the total score (e.g., sum or average) for a set of questions.

In this assignment we will practice some basic data handling skills.

6.3.4.1 Reverse Scoring Contra Indicative Items

If you read the item labels, you will see that the first two SI items (items DS14_1 and DS14_3 in the scale) are contra indicative. This means that for these items higher scores reflect low SI, while for other items higher scores reflect high SI. Therefore, the responses to these items should be recoded first to make sure that all items are scored in the same direction. To do so, we will create new variables that reflect the recoded items. Proceed as follows:

Transform > Recode into different Variables

---

[44] 38
[45] 41.64
[46] 43.37
[47] 2
[48] 54.5

Choose DS14_1 as the Numerical Variable to be recoded

Specify a name for the output variable (say DS14_1R)

Give a label, say: "SI item 1 (recoded)"

Click on change

Do not close the window yet, but continue to the next step...

To recode, we have to specify the Old Values and New values. Reverse scoring of the DS14 items means that

old value 0 -> new value 4

old value 1 -> new value 3

old value 2 -> new value 2 (*)

old value 3 -> new value 1

old value 4 -> new value 0

(*) You may think this line is superfluous but for the recoding in SPSS you need to specify for every possible value a recoded new value, even if the values remains the same.

Specify the old and new values. Each time you specified old and new values click on ADD such that the recoding scheme appears in the little dialog.

6.3.5 Using Syntax

To ensure that others can see exactly how you got from raw data to the final dataset used for analysis, it is **essential** to keep a complete record of any changes made to the data. This is also why we previously argued that you should **not overwrite** a datafile after altering it.

Up until now, we ran the analyses by "click-and-point" via the menu. This is a good starting point to explore the software SPSS, but it is not good practice for professional use because there's no record of what you did to the data in order to get your result.

NOTE: For all your portfolio assignments, providing syntax is mandatory so I can grade what you DID, not just what you reported!

To keep a record of changes made to your data, you can prepare a script that contains all instructions for the analysis instead. By evaluating this script on the data, you should consistently get the same results.

6.3.5.1 Why syntax?

Using syntax is important for several reasons:

- First, efficiency: it makes life easier. Once you have the syntax, you can easily redo the whole analysis without going through all the points-and-clicks again.
- Second, communication: When you work together on research projects, it is important that you exactly understand the analyses that were done, even if you didn't do the analyses yourself. By using syntax, all team members can see what has been done and how.
- Third, documentation & data management: As a researcher you are responsible for data storage and management (!). This not only includes storage of data, but also documentation of the all the steps and analyses you did to come to your results (e.g., handling missing values, detection of outlying values). Ideally, you should provide the materials such that other researchers can easily replicate your analyses starting from the raw data file. Working with SPSS syntax is a great way to do so.
- Fourth, necessity: some statistical procedures are only available via SPSS syntax (e.g., simple effects analysis in MANOVA).

6.3.5.2 Help!

You don't need to memorize the commands by heart. SPSS offers help functions. If you highlight the command (e.g., statistics) and click on the button with the paper and the question mark in the top menu. SPSS opens a help file.

Use the help function to modify the syntax such that SPSS produces a table that also reports the range (e.g., the difference between the largest and smallest value).

6.3.5.3 How to use syntax

There are two ways to use syntax. The first is to create an empty syntax script via File -> New -> Syntax, and then start adding the code from scratch. You can either write the code as text, or create it via SPSS' various dialog windows. For this course, we recommend using the dialogs:

- In any dialog window, click "Paste" instead of "OK".
- A new window opens (or existing window comes into focus) with a script file (or "syntax" file). The instructions for the requested analysis are added at the bottom of this file.
- Select all instructions you wish to execute, and press the green "Play" button.
- You can re-organize this script file, adding, or removing operations or changing their order. Keep it nice and organized!

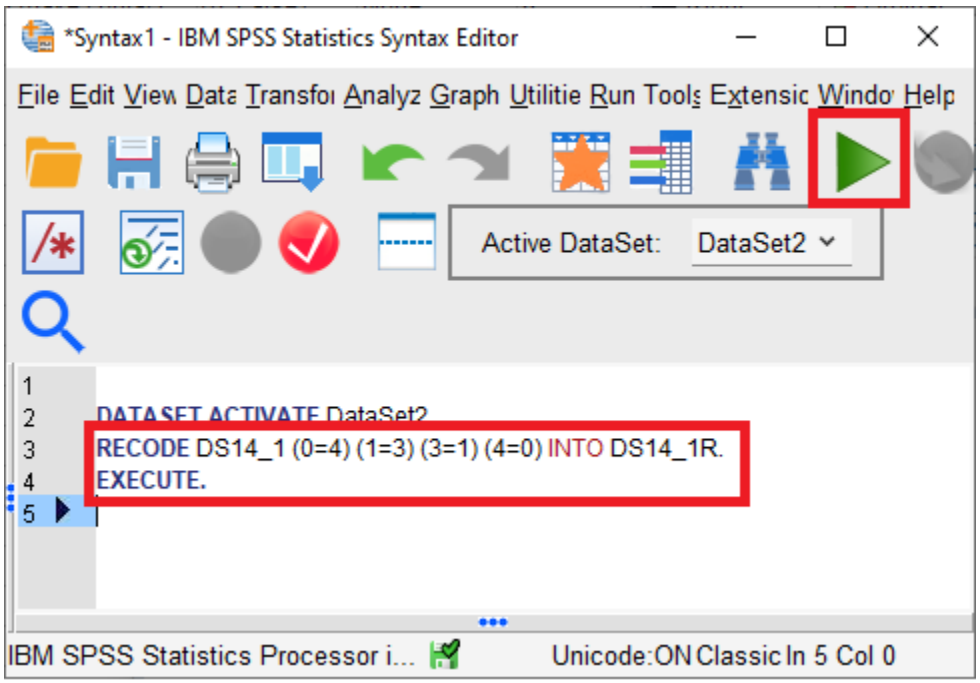


Figure 6.1: Syntax for recoding

Select the highlighted lines and press the green "Play" button. Verify in the data view that SPSS added a new variable (i.e., new column) with the recoded values for Item 1.

Syntax is also useful because it may be more straightforward than the graphical interface. For example, the recode syntax above took a lot of pointing and clicking to get:

```
RECODE DS14_1 (0=4) (1=3) (3=1) (4=0) INTO DS14_1R.
EXECUTE.
```

For recoding more variables, we can copy-paste these instructions and change the variable names.

For example, to recode DS14_3 in the same way as you did the recoding for DS14_1 we would write:

```
RECODE DS14_1 (0=4) (1=3) (3=1) (4=0) INTO DS14_1R.
RECODE DS14_3 (0=4) (1=3) (3=1) (4=0) INTO DS14_3R.
EXECUTE.
```

Alternatively, we could simply say:

```
COMPUTE DS14_1R = 4-DS14_1.
COMPUTE DS14_3R = 4-DS14_3.
EXECUTE.
```

Verify that this calculation also works.

6.3.5.4 Commenting

When working with syntax, it is highly recommended to add comments as reminders to your future self of the purpose of each step in the script. These comments should clarify the syntax and give general information (e.g., when was the syntax last modified, who did the modifications, etc.).

Comments are text lines that start with an * and ends with a dot. Comments are printed in grey.

REMARK: the dot at the end of your text line is really important. If you do not add it, your syntax will run work properly!

Add comments including the following information:

- When was the syntax made?
- Who made syntax?
- What does the syntax do?

For example (CJ is short for Caspar J. Van Lissa):

```
* CJ: This script also recodes Likert scales with integer values.
COMPUTE DS14_1R = 4-DS14_1.
COMPUTE DS14_3R = 4-DS14_3.
EXECUTE.
```

After including the comments, run the complete syntax including the comment line (by selecting and running it, or via top menu Run > All).

If the syntax runs correctly, the comments were correctly included.

6.3.5.5 Compute Variable

For the analyses we are not interested in the single item scores but in the summed scores. Because the DS14 contains two subscales (consult the item labels to see to which subscale the item belongs), we want to compute the summed scores for the NA and SI items separately.

Let's first do this for NA. Proceed as follows:

Transform > Compute variable. SPSS opens a new window.

Choose a name for the target variable, say: NAtotal.

Under numerical expression you have to say what you want to compute. In this case the sum of the NA items, which is DS14_2 + DS14_4 + ... etc. So, select the first item to be summed from the list at the left, type +, select the next item you want to add, and so on. Make sure that you only add up the NA items (in the variable view you can see which items measure NA and which measure SI).

Click Paste, and run the code.

Alternatively, copy-paste this syntax, complete it and run it:

```
COMPUTE NAtotal = DS14_2 + DS14_4 + .
COMPUTE SItotal = .
EXECUTE.
```

6.3.5.6 Quiz

Compute the frequency table for the recoded DS14_3 item.

What percentage of the respondents had the highest score on this item? _____[49]

What is the mean of NAtotal? _____[50]

Compute the mean of NAtotal for men and women separately. For which group is the mean highest? (Hint: Use the skills you've learned in the previous assignments). [51]

- (A) Women
- (B) Men

[49] 28.6
[50] 18.3175
[51] Women

Compute the total scores for SI. Make sure you use the reverse scored items for items 1 and 3 to compute the total score (this implies you have to leave the original item 1 and 3 out).

What is the mean of SI in the total sample? _____[52]

What is the mean for the subsample of men? _____[53]

[52] 15.5873
[53] 17.2

7 The Sampling Distribution

As explained in lecture 1, a sample is an observed subset of a larger population. We typically calculate statistics based on sample data, and use these as best guesses of the values of population parameters. This process is called statistical inference. A crucial insight is that sample statistics are not perfect estimates of population parameters. The discrepancy between the sample statistic and population parameter is known as sampling error.

We have some theoretical insight into theoretical behavior of sample statistics. For example, we can imagine constructing a probability distribution of the values we might see for a sample statistic, such as the mean, if we were to draw very many random samples from an identical population. This theoretical distribution of means is called the sampling distribution. The central limit theorem tells us that, regardless of the shape of the distribution of the data in the population, as the sample size increases, the sampling distribution of the mean approaches a normal distribution. This is an important realization, because it means that we can use probability calculus using the normal distribution to draw inferences about population parameters based on sample statistics.

The standard deviation of the sampling distribution plays a central role in inferential statistics. It is so important that we give it a unique name: we call this particular standard deviation the *standard error* (SE). The standard error quantifies the average, or expected, amount of sampling error when we use a sample statistic to estimate the population parameter. If the standard error is small, our estimates based on the sample are likely to be accurate, whereas a large standard error indicates greater uncertainty.

With the help of the normal distribution, and given a particular (hypothesized or known) population mean and standard error, we can calculate how likely it is to observe specific sample means. For example, if we want to determine the probability that the mean of a random sample exceeds a certain value, we can standardize the sample mean using the formula $Z = \frac{M}{SE_M}$, where M is the sample mean, is the known or hypothesized population mean, and SE is the standard error. By looking up the corresponding probability on the standard normal distribution table or using statistical software, we can assess the likelihood of observing a specific sample mean (or greater, or smaller).

Confidence intervals are a way to express our uncertainty about the sample statistic as estimator of the population parameter. A confidence interval is a range of values - a window - within which we expect the true population parameter to fall with a certain level of confidence. Typically, we select a 95% confidence interval, which means that if we could repeat the sampling process many times and calculated confidence intervals each time, 95% of those intervals would

contain the true population parameter. The width of the confidence interval is determined by the standard error and is proportional to the level of confidence desired. The formula for a confidence interval is often written as: M ś $Z_{95\%}$ SE_M. In practice, this comes down to approximately: M ś 2 SE_M.

7.1 Lecture

VIDEO ERRATA: from 19:40 - 19:50 I incorrectly report the probability of P(Z > 1) as .025, but it is .16.

https://www.youtube.com/embed/iNxF0Tr6R9M

7.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

Introversion is normally distributed with a mean of 50 and a standard deviation of 10. What is the probability that the mean introversion level of a randomly selected group of 16 people is smaller than 52? Round the answer to 3 decimal places. [1]

- (A) 0.788
- (B) 0.345
- (C) 0.655
- (D) 0.045

Question 2

Variable X is not normally distributed in the population. Variable X has a population mean of 30 and a population standard deviation of 6. A random sample of N = 36 scores is drawn from the population for variable X. The sample mean is equal to 32. Which of the following

[1] 0.788

statements about the sampling distribution of the sample means for this sample (n = 36) is incorrect? [2]

- (A) The standard error is smaller than the population standard deviation.
- (B) The standard deviation of the sampling distribution of sample means is equal to 1.
- (C) The mean of the sampling distribution of the sample means is equal to 32.
- (D) The sampling distribution of sample means is approximately normally distributed.

Question 3

What does the sampling distribution represent? [3]

- (A) A theoretical distribution of sample statistics
- (B) The distribution of population parameters
- (C) The distribution of sample means
- (D) The distribution of raw data

Question 4

Which of the following statements about the sampling distribution is true? [4]

- (A) The sampling distribution is identical to the population distribution
- (B) The sampling distribution is centered around the population parameter
- (C) The sampling distribution is based on a single sample
- (D) The sampling distribution of a skewed variable is also skewed

Question 5

What is the standard deviation of the sampling distribution called? [5]

- (A) Standard error

[2]The mean of the sampling distribution of the sample means is equal to 32.
[3]A theoretical distribution of sample statistics
[4]The sampling distribution is centered around the population parameter
[5]Standard error

- (B) Bias
- (C) Variance
- (D) Population standard deviation

Question 6

How does sample size affect the shape of the sampling distribution? [6]

- (A) Larger sample sizes increase the variability of sample statistics
- (B) Larger sample sizes make the sampling distribution more spread out
- (C) Larger sample sizes result in smaller standard errors
- (D) Larger sample sizes have no impact on the sampling distribution

Question 7

What is the probability that a sample mean falls within +/- 1 standard deviation of the population mean, assuming a normal distribution of sample means? [7]

- (A) 0.34
- (B) 0.68
- (C) 0.48
- (D) 0.95

Question 8

If the standard deviation of the population is 10 and the sample size is 25, what is the standard error of the sample mean? [8]

- (A) 0.2
- (B) 5
- (C) 2
- (D) 0.4

[6] Larger sample sizes result in smaller standard errors
[7] 0.68
[8] 2

Question 9

What is the probability that the sample proportion falls within +/- 2 standard deviations of the population proportion, assuming a large sample size? [9]

- (A) 0.48
- (B) 0.95
- (C) 0.68
- (D) 0.34

Question 10

If the standard deviation of the population is 5 and the sample size is 50, what is the standard error of the sample mean? [10]

- (A) 0.0707
- (B) 0.283
- (C) 0.141
- (D) 0.707

[9] 0.95
[10] 0.707

Show explanations

Question 1
Calculate the standard error as 10/sqrt(16). Then, calculate the Z-score as (52-50)/SE. Find the right tailed probability of that Z-score, then calculate 1 minus that probability.
Question 2
The sampling distribution will be approximately normal because n >= 30. The SE is indeed 1, because sigma/sqrt(n) = 6 /sqrt(36) = 1. The SE is always smaller than sigma, because it is calculated as sigma divided by square root of n.
Question 3
The sampling distribution represents the distribution of sample statistics, such as sample means or proportions, derived from multiple samples drawn from the same population. It provides insights into the variability and characteristics of these sample statistics.
Question 4
The sampling distribution is centered around the population parameter.
Question 5
The standard deviation of the sampling distribution is known as the standard error. It measures the average variability or spread of sample statistics around the population parameter, reflecting the precision of the estimation.
Question 6
Larger sample sizes result in smaller standard errors. As the sample size increases, the sampling distribution becomes more concentrated around the population parameter, leading to a decrease in the standard error. This implies that larger samples provide more precise estimates of the population parameter.
Question 7
The probability that a sample mean falls within +/- 1 standard deviations of the population mean, assuming a normal distribution of sample means, is 68%. This is based on 'the empirical rule'.
Question 8
The standard error of the sample mean is 2. The standard error can be calculated by dividing the standard deviation of the population by the square root of the sample size. In this case, it would be 10 / 25 = 2.
Question 9
The key lesson here is that everything you learned about the sampling distribution also applies to other statistics than the mean, so according to the empirical rule, 95% of sample proportions will fall within +/- 2 standard deviation of the population proportion.
Question 10
The standard error of the sample mean is SD/sqrt(n), so 5/sqrt(50) = .707

7.3 Tutorial

7.3.1 Sampling Distribution

Complete the following sentences:

IQ scores in the population of potential students are normally distributed with mean 120 and an SD of 10. If a cohort contains 75 students, 95% of cohorts will have an average IQ in between [11]

- (A) 100
- (B) 118.85
- (C) 117.69
- (D) 118.27

and [12]

- (A) 121.15
- (B) 122.31
- (C) 121.73
- (D) 130

.

After graduating, a cohort of 75 LAS students can expect to earn a starter salary of 2650 Euros, with an SD of 300 euros. What percentage of cohorts will have a mean starter salary greater than 2750 euros? [13]

- (A) 99%
- (B) 0.2%
- (C) 2%
- (D) 4%

[11] 117.69
[12] 122.31
[13] 0.2%

.

In a sample of 5000 babies, the average birthweight is 3.213 kg, with an SD of 254 grams. What is the mean birthweight of the sampling distribution? [14]

- (A) 3.213 kg
- (B) Can't say
- (C) Between 3202.22 and 3223.78 grams

Consider a continuous variable X, which is normally distributed with X ($= 30$, $= 4$). We draw a sample of 15 participants. What is the probability that the sample mean will be smaller than 32? _____[15]

The proportion of male babies is .51. Assume babies born in each hospital in a given month constitute a random sample of size 100. The standard error of a proportion is given by $(p\ (1\ p)/n) = (.51\ .49/100) = 0.05$. What proportion of hospitals will have more than 60% male babies? _____[16]

7.3.2 In SPSS

7.3.2.1 SE for Means

Open the file called `student_questionnaire.sav`.

These are data from a previous cohort of students. Note that we have data about biometric differences (e.g., age, height, shoesize), as well as school-related questions (which program they are enrolled in), variables about their love for statistics, and about moral preferences (based on the "Morality As Cooperation" questionnaire that I helped develop).

Go to Analyze -> Descriptives and ask for descriptive statistics on height and shoesize. Click Options, and notice that there's an option to request the standard error of the mean. Select this option, then paste and run your syntax. Check if it corresponds to the syntax below.

Answer

```
DESCRIPTIVES VARIABLES=height
  /STATISTICS=MEAN STDDEV MIN MAX SEMEAN.
```

Note the `SEMEAN` option was added by clicking that option!

[14] Can't say
[15] 0.974
[16] 0.04

The mean length in the population of Dutch people is 177.434. With this in mind, calculate the probability that a random sample of the same size as this sample would have the mean length you calculated for this sample or smaller. _____[17]

Explanation

The question asks for the lower-tail probability below a value of 174.62 in a distribution with mean 177.434 and SD .772 (the SE you obtained from SPSS).
$\frac{174.62177.34}{.772} = 3.65$
A Z-score of nearly -4, so this probability is going to be extremely small, $< .001$.

7.3.2.2 SE for Proportions

Go to Analyze -> Compare Means -> One Sample Proportions.

This procedure allows you to estimate proportions and their standard errors. It's not very common, in fact I learned about it by Googling "standard error for proportion spss"! Any time you need to know how to do something in SPSS, you can find advice on the internet.

Calculate the proportion for the variable sex, and paste and run your syntax.

```
PROPORTIONS
  /ONESAMPLE sex TESTVAL=0.5 TESTTYPES=MIDP SCORE  CITYPES=AGRESTI_COULL JEFFREYS WILSON_SCO
  /SUCCESS VALUE=LAST
  /CRITERIA CILEVEL=95
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE.
```

Note the table labelled "One-Sample Proportions Confidence Intervals". This table contains confidence intervals for the proportion, calculated according to three different procedures. In the lecture, you also learned a procedure to calculate confidence intervals.

Using the procedure from the lecture, calculate a 95% confidence interval for the proportion. You can round the Z-score for this confidence interval to 2.

The 95% CI for the proportion of male students is [_____[18], _____[19]].

Note that the differences between this procedure and the three procedures in the table only differ in the third decimal.

How do you interpret a confidence interval?

[17] 0.000133
[18] 0.359
[19] 0.507

[20]

- (A) A range of values that contains the population parameter with 95% certainty.
- (B) A range of values that, 95% of the time, contains the population parameter.
- (C) The range of likely population values.
- (D) The population value, with a 95% margin of error.

7.3.2.3 SE for Correlation

Recall from the first lecture that the correlation coefficient is a measure of linear association between two variables, or: a descriptive statistic that describes how strongly two continuous variables are associated.

Go to Analyze -> Correlate -> Bivariate. Add the variables work_hours and study_hours, paste and run the syntax.

The value of the correlation coefficient is labelled "Pearson Correlation". What value do you observe? _____[21]

The correlation coefficient ranges from 0-1 (or minus 1). With this in mind, answer the following question:

True or false: This correlation coefficient is near zero. TRUE / FALSE[22]

The calculation of a standard error is a bit more complicated, but there's an "approximation" (a quick approach that gives reasonable results in some cases, but could be wrong in other cases). It is calculated as:

$$SE_r = \frac{1\ r^2}{n\ 2}$$

Calculate the SE this way. What is its value? _____[23]

Assume for a moment that the true population correlation is zero (r = 0). Using the SE you calculated, what would then be the probability of observing a correlation between 0 and the correlation you actually observed? _____[24]

---

[20] A range of values that, 95% of the time, contains the population parameter.
[21] 0.057
[22] TRUE
[23] 0.07525529
[24] 0.2756014

Explanation

The question asks for the probability between the mean (0) and a value of .057 in a distribution with mean 0 and SD .075 (the SE you calculated).
So we first calculate the right-tailed probability for the value of .057.
$Z = \frac{.0570}{.075} = 0.76$
A Z-score of 0.76, so the right-tailed probability is 0.22.
Then, take .5 (the probability to the right of 0), and subtraCT .22: .28

8 Philosophy of Science

Students of statistics often have the impression that this course is all about cold, hard facts. Nothing could be further from the truth: statistics is a direct extension of philosophy of science. The numbers we calculate here only have relevance to real-world research questions thanks to some complex philosophical arguments. While most of these are outside the scope of the present course, we want you to be familiar with the main concepts before moving on.

8.0.1 Deduction and Induction in Hypothesis Testing

Hypothesis testing has roots in the philosophy of logic, or correct reasoning. In logic, arguments consist of a set of premises, which can be true or false, that together lead to a conclusion, which can also be true or false. The most famous example is:

1. **Premise:** All men are mortal
2. **Premise:** Socrates is a man
3. **Conclusion:** Therefore, Socrates is mortal.

This is a *deductive* argument, which has the property that if the premises are true, then the conclusion must also be true. Deductive arguments are also often thought of as arguments from the general to the specific. In this case, the general rule "all men are mortal" gives rise to the specific claim that one specific man, Socrates, is also mortal.

The counterpart to deductive reasoning is *inductive* reasoning, which proceeds from specific observations or claims to general rules. The most famous example is:

1. **Premise:** All swans I have ever seen are white
2. **Conclusion:** Therefore, all swans are white.

Unlike the deductive argument above, however, inductive reasoning does not guarantee that true premises always produce true conclusions. Even if it is true that I have only seen white swans, the conclusion is false - black swans exist. David Hume introduced another famous example:

1. **Premise:** The sun has risen in the east every morning up until now.

2. **Conclusion:** The sun will also rise in the east tomorrow.

Here, too, the conclusion is not supported by the premise. This might make us feel uncomfortable - which sane person would reject the conclusion that the sun will rise in the east tomorrow? This discomfort illustrates that people are naturally inclined to reason inductively. As scientists, however, we should be very cautious to remember that this way of reasoning is not guaranteed to produce true conclusions.

Inductive and deductive reasoning are both used in statistical hypothesis testing. Deduction is used when we derive a specific prediction (hypothesis) from a general theory. For example, we could say that:

1. **Premise:** More time spent studying causes better grades.

2. **Conclusion:** In my dataset, time spent studying should correlate positively with grades.

If the premise is true, then we would expect to observe the corresponding pattern in our data.

Induction comes into play when we draw general conclusions from observed data.

1. **Premise:** I observed a positive correlation between time spent studying and grades in my dataset

2. **Conclusion:** Therefore, time spent studying causes better grades.

This conclusion does not logically follow from the premise. The problem is not resolved by removing the word "causes":

1. **Premise:** I observed a positive correlation between time spent studying and grades in my dataset

2. **Conclusion:** Therefore, time spent studying correlates positively with grades in the general population of students.

It follows that we can never conclusively "prove" general conclusions from specific observations. The philosopher David Hume wrote extensively about this "problem of induction": How can we justify the assumption that unobserved cases will follow the same patterns as observed ones? Hume argued that this assumption cannot be logically justified. Our sense that generalization is justified might be based on intuition, but not logic. The problem of induction challenges the very foundation of science. We cannot escape the use of induction when seeking to learn general insights from specific observations, but Hume showed that induction lacks a purely rational justification.

8.0.2 Falsificationism

It follows from the problem of induction that it is impossible to definitively prove a theory to be true. No matter how much evidence I have observed that supports a theory (white swans), all it takes is one refuting observation (black swan) to reject is. Karl Popper sought to avoid the problem of induction by devising a scientific method that relies exclusively on deduction: falsificationism. Popper demarcated the distinction between "pseudo-science" and science by arguing that *"[scientific] statements [...] must be capable of conflicting with [...] observations"* (Popper 1962, 39). The core business of science, according to Popper, should be to try to reject theories. Note that - while Popper's work has been heavily criticized (for good reasons), his work is very influential in social science. Therefore, Popper is a good *starting point* for our course - even though he probably should not be the *endpoint* for students with a genuine interest in philosophy of science.

8.0.3 Falsificationism and Hypothesis Testing

The idea of falsificationism has been very influential in applied statistics in the social sciences, particularly in the practice of "Null-Hypothesis Significance Testing" (NHST). In NHST, researchers proceed as follows:

1. Develop a testable proposition about a population parameter; for example:
    - "On average, my students understand the course material. Their mean grade is $\quad 6$".
    - "On average, there is a positive association between hours studied and grade, $\quad 0$".
2. Develop a second hypothesis whose sole purpose is to be rejected, to pay lip service to falsificationism. Call this the "null hypothesis". The "null hypothesis" is often taken to be the exact opposite of the researcher's true belief, or "alternative hypothesis":
    - "On average, my students **DO NOT** understand the course material. Their mean grade is < 6".
    - "On average, there is **NO** positive association between hours studied and grade, < 0".
3. Execute a procedure to make a decision to reject (falsify) or not reject the null hypothesis (next chapters).
4. If the null hypothesis is rejected, act as if this finding supports the alternative hypothesis.

As argued by Andrew Gelman in this blog post, this approach only pays lip service to falsificationism. A true falsificationist would put their true belief (alternative hypothesis) to the test. Fake falsificationism is creating a meaningless, "straw man" null-hypothesis, whose sole purpose is to be rejected.

8.0.4 Moving Forward

Where does this leave us? The most important point is that there are important limitations to commonly used methods, including null-hypothesis significance testing. There is no real satisfying solution. Just keep in mind that no statistical test can give evidence in support of a theory or hypothesis; neither a null nor an alternative hypothesis.

8.0.5 Causality

Another crucial philosophical issue relevant for statistics is the question of causality. Scientists are often interested in causal questions. We assume causality, for example, any time we want to *act on* knowledge derived from scientific research. For example, say that I do find a strong correlation between hours studied and grade obtained. If, based on this finding, you increase your study hours in order to improve your grade - then you are assuming causality. The same applies for governments making evidence-based policy, companies adjusting sales strategies based on customer analytics, or drugs that replenish a particular neurotransmitter that has been found lacking in patients with a specific diagnosis.

Despite the fact that causality is so important in scientific research, it is rarely defined. Contemporary definitions are typically based on *counterfactuals*: A is a cause of B if B would not have happened if A had not happened. This definition has important limitations, but it is sufficient for our course (Halpern, 2015).

Most people have heard the phrase "correlation does not imply causation". What does this mean? One important misunderstanding is that the *correlation coefficient* is an inappropriate statistic for investigating causal research questions. This is not the case. This phrase warns us that, just because observe a statistical association between variables X and Y (for example, a correlation of $r = .43$), that does not mean that X caused Y. Importantly, observing this correlation is consistent with a causal effect of X on Y - but also with other explanations. For example, maybe Y caused X, or maybe a third variable caused both X and Y, and that is why they are correlated.

The problem is related to the previous section: observing a pattern in data that is consistent with a causal association between X and Y cannot conclusively prove that X caused Y, no matter what statistic you use to describe the pattern.

So where does causality come from? The short answer is: from theory or methodology. Causality can be assumed on theoretical grounds, or established using a randomized controlled experiment. In such an experiment, researchers randomly assign participants to either an experimental condition (e.g., receiving a drug, instruction, treatment, et cetera), or a control condition (e.g., receiving a placebo, no instruction, non-effective treatment, et cetera). The random assignment should, theoretically, result in two groups with no systematic differences that could explain between-group differences in the outcome of interest. Of course, due to pure chance, it could happen that there are systematic differences (more men in one group, taller people in one

group, et cetera). But there is no procedure that has a better chance of resulting in comparable groups than random assignment.

In the social sciences, experiments are not always feasible or ethical, so researchers often use observational data. Does this preclude all causal claims? It does not. It is perfectly legitimate to present a theory that predicts a causal effect in the Introduction section of your scientific writing, and then present empirical data that show a pattern *consistent with* that effect. For example, if I assume that hours studied causes improved grades, a correlation of $r = .43$ is consistent with that assumption. An alternative explanation might be that having receiving high grades in the past has motivated some students to study harder. Just make sure not to take the observed data as *evidence for* a causal effect.

What is required to argue that X has a causal effect on Y? Several philosophers have addressed this issue; most notably Hume and Mill (see Morabia, 2013). The necessary conditions for causality are sometimes summarized as association, temporal precedence, and non-spuriousness. Below, each is supported with quotes from Hume (1902), which were selected by Aaron Peikert:

1. **Association**: Cause and effect must be associated (this could be statistical association)
    - "When one particular species of event has always [...] been conjoined with another, we [...] call the one object, Cause; the other, Effect." (VII, II, 59)
    - "familiar objects or events to be constantly conjoined together" (V, I, 35) There must be a "constant union" between cause and effect, and they must be "contiguous in space and time" (Hume 1739,16, pp. 173–175).
2. **Temporal precedence**: The cause must occur before the effect.
    - observe a continual succession of objects, and one event following another (V, I, 35)
    - "[when] the same object is always followed by the same event; we then begin to entertain the notion of cause and connexion." (VII, II, 61)
3. **Non-spuriousness**: All alternative explanations of the effect are excluded.
    - "Their conjunction may be arbitrary and casual. There may be no reason to infer the existence of one from the appearance of the other." (describing spuriousness; V, I, 35)
    - "we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed." (VII, II, 61)

This latter definition resembles the aforementioned counterfactual definition of causality. Note that this definition does not *require* randomized experimentation, but randomized experiments do help us meet all three criteria. The field of *causal inference* focuses on developing methods that can estimate causal effects (like you would get from a randomized controlled experiment) from observational data (Pearl, 2009).

8.1 Lecture

TO DO

8.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the main difference between deductive and inductive reasoning? [1]

- (A) Deductive reasoning guarantees the conclusion if the premises are true, while inductive reasoning does not.
- (B) Deductive reasoning is used in science; inductive is not.
- (C) Inductive reasoning guarantees conclusions; deductive does not.
- (D) There is no real difference; both produce valid conclusions.

Question 2

Which of the following is an example of inductive reasoning? [2]

- (A) The sun has always risen in the east, so it will rise in the east tomorrow.
- (B) Water boils at 100 degrees at sea level.
- (C) Students who study more get better grades.
- (D) All humans are mortal, so Socrates is mortal.

Question 3

What is the role of deduction in statistical hypothesis testing? [3]

[1] Deductive reasoning guarantees the conclusion if the premises are true, while inductive reasoning does not.
[2] The sun has always risen in the east, so it will rise in the east tomorrow.
[3] It is used to derive specific predictions from general theories.

- (A) It is used to generalize observations to populations.
- (B) It tests if data supports any hypothesis.
- (C) It is used to derive specific predictions from general theories.
- (D) It determines the sample size needed.

Question 4

What was Karl Popper's proposed solution to the problem of induction? [4]

- (A) Renouncing inductively constructed theories as pseudoscience.
- (B) Falsification: trying to disprove theories rather than prove them.
- (C) Using more data to confirm theories.
- (D) Only using observational data.

Question 5

In Null-Hypothesis Significance Testing (NHST), what role does the null hypothesis play? [5]

- (A) It shows causality when rejected.
- (B) It provides definitive proof of a theory.
- (C) It merely exists to be refuted.
- (D) It reflects the researcher's true belief.

Question 6

What does the phrase 'correlation does not imply causation' mean? [6]

- (A) Don't use a correlation coefficient if the effect is causal.
- (B) Causal relationships cannot be studied using statistics.
- (C) A statistical association may not reflect a true causal relationship.
- (D) Causation can only be studied with regression.

[4]Falsification: trying to disprove theories rather than prove them.
[5]It merely exists to be refuted.
[6]A statistical association may not reflect a true causal relationship.

Question 7

According to a common interpretation of Hume, which three conditions are necessary for causality? [7]

- (A) Falsifiability, deduction, and experimentation.
- (B) Association, temporal precedence, and non-spuriousness.
- (C) Randomization, generalization, and verification.
- (D) Observation, correlation, and prediction.

Question 8

How do randomized controlled experiments support causal inference? [8]

- (A) They confirm inductive conclusions.
- (B) They eliminate alternative explanations through random assignment.
- (C) They always prove causality through replication.
- (D) They eliminate the need for statistical testing.

[7] Association, temporal precedence, and non-spuriousness.
[8] They eliminate alternative explanations through random assignment.

Show explanations

Question 1
Deductive reasoning leads to necessarily true conclusions if the premises are true, while inductive reasoning involves probable conclusions based on patterns.
Question 2
Assuming that the sun will rise tomorrow because it always has is an example of induction, and it is not logically justified.
Question 3
Deduction is used to derive derive specific, testable predictions from general theories, such as expecting a correlation between study time and grades if studying improves performance.
Question 4
Popper argued that science should focus on rejecting falsifiable hypotheses rather than confirming them, avoiding the pitfalls of induction.
Question 5
The null hypothesis is a 'straw man' hypothesis, only intended to be rejected. This interpretation of Popper's falsificationism actually never submits theories to the test, and aligns more closely with confirmationism.
Question 6
No statistical finding can prove causation, so a correlation (or other type of association) between two variables doesn't confirm a causal link.
Question 7
A common interpretation of Hume is that for a cause-effect relationship, the variables must be associated, the cause must precede the effect, and other explanations must be ruled out.
Question 8
Random assignment maximizes the probability that the treatment and control groups do not differ on any confounding variables, thus ensuring nonspuriousness. Furthermore, a well-designed experiment ensures association (if there is an effect) and temporal precedence.

8.3 Tutorial

8.3.1 Assignment 1: Induction and Deduction

Below are two texts like you might find in a social science publication. Carefully read both, and highlight examples of inductive- and deductive reasoning. Then discuss with your group mates:

a. Did you all highlight the same examples? Did you disagree about any of them?

b. Did you find examples of induction and deduction in both texts? Is one mode of reasoning more prominent in one text than the other?
c. Are all inductive and deductive inferences warranted? Do the fragments show sufficient awareness of the potential limitations of these modes of reasoning? Or conversely, are they *too* careful?

Fragment A:

> Using a crosssectional survey of adolescents from six urban schools (N = 2,184), we examined associations between evening screen time (selfreported minutes after 8 p.m.) and sleep quality (Pittsburgh Sleep Quality Index). Across schools and controlling for grade and gender, greater evening screen time was consistently associated with poorer sleep quality, r = .30, 95% CI [.25, .33]. Subgroup analyses by device type (phone vs. tablet) and by extracurricular workload yielded similar patterns. While these observational data cannot establish causation, finding a moderate effect across all schools suggests that higher evening screen exposure is related to adolescents' diminished sleep quality. Reducing evening device use could help improve adolescents' sleep outcomes.

Fragment B:

> Social norms theory posits that behavior is shaped by perceptions of typical peer behavior. We thus hypothesized that providing households with descriptive norm feedback about neighborhood electricity use would reduce individual electricity consumption, relative to a neutral informational control. We preregistered H1: households receiving monthly comparative reports will exhibit lower kWh usage over three billing cycles than controls. The null hypothesis H0 was that there was no difference between the experimental and control conditions. In a randomized field experiment (N = 3,042 households), treatment households received reports comparing their usage to that of "similar homes", alongside efficiency tips; control households received tips only. Mixed-effects models with random intercepts for household indicated a 2.4% reduction in usage for households in the experimental condition, $= 0.024$, SE = 0.007, $p < .001$. We rejected the null hypothesis of no difference. These results are consistent with H1, indicating that social norms theory is a relevant framework for understanding household electricity consumption.

8.3.2 Assignment X: Causality

Below are five examples of statements from social scientific papers, courtesy of Calvin Isch.

a. First, read all the examples, and sort them into causal claims/non-causal claims. Discuss with your groupmates: did you classify each claim the same way? What makes the difference for you?

b. With your group, choose one of these claims and examine the associated paper. Discuss:

   i. Do you still think the claim is causal/noncausal?
   ii. Would a causal claim be justified in this case?
   iii. Why/why not?

Example 1:

> Violence exposure hampers compromise among Israelis, emphasizing the importance of abstaining from violence for conflict resolution.

https://doi.org/10.1093/pnasnexus/pgae581

Example 2:

> We investigate both the role of gender and feminism in friends-with-benefits (FWB) relationships at a United States college, and ask whether identification with feminist ideology impacts students' motivations and assessments of their relationships.

https://link.springer.com/article/10.1007/s12119-014-9252-3

Exampe 3:

> This distrust of atheists is driven by religious predictors, social location, and broader value orientations.

https://doi.org/10.1177/000312240607100203

Example 4:

> we show that being in a dual-career household increases one's willingness and lowers the perceived risk of leaving their job and joining a startup venture—especially if the household prioritizes their spouse's career.

https://doi.org/10.1002/smj.3481

Example 5:

> We find that those who live in regions with a greater share of migrants from Eastern Europe have more positive attitudes towards the EU but that this positive influence diminishes in highly segregated areas.

https://doi.org/10.1080/13501763.2023.2271504

9 Hypothesis Testing

9.0.1 Falsificationism

In science, it is rarely possible to **prove** a general theory true. Instead, theories earn credibility by **surviving** serious attempts to refute them. This is the core of **falsificationism**, most closely associated with Karl Popper (Popper, 1959/2002). A scientific claim must be **testable** and **falsifiable**: it should make predictions that could, in principle, be shown false by observation. The proverbially simple example is "All swans are white." No number of white swans can verify the claim, but a **single** black swan would falsify it.

This logic connects directly to statistical practice. In hypothesis testing we do not "prove the theory"; rather, we pose a **precise** claim—the null hypothesis H_0—and ask whether the observed data are sufficiently incompatible with H_0 to warrant rejecting it. A small p-value is evidence **against** H_0, not proof **for** any particular alternative. Likewise, **failing to reject** H_0 does not verify H_0; it merely indicates that the data are not unusually discordant with it given the test's design and assumptions. Good scientific practice increases the *riskiness* of tests—deriving clear, prior predictions; minimizing researcher degrees of freedom; and using designs that would make discordant data likely **if** the theory were false (e.g., preregistration and prospective power analysis). In short, falsificationism reminds us that scientific conclusions are **provisional** and should be sharpened by attempts to refute, not by post hoc confirmation.

9.0.2 Hypothesis testing {#sec-Hypothesis testing}

Hypothesis testing is a method of inferential statistics which allows researchers to draw conclusions about the population based on sample data. It involves formulating hypotheses, calculating test statistics, determining p-values, and drawing conclusions about the null hypothesis.

Hypothesis testing builds upon previously covered topics like sampling theory and estimation, where sample statistics are used as the best estimate of population parameters; standard errors to express the uncertainty surrounding those estimates; and probability calculus, using probability distributions - like the standard normal distribution - to compute the probability of observing certain values based on the sampling distribution.

To introduce the concept of hypothesis testing, let's consider an intuitive example. Imagine your car won't start, and you hypothesize that the battery is dead. You then perform an

experiment by replacing the battery. If the car starts, you conclude that your initial hypothesis was correct - the battery was indeed dead.

In this thought experiment, you only need one piece of evidence. Statistical hypothesis instead rely on evidence from many observations, and use probability calculus to test hypotheses in the presence of uncertainty. Statistical tests use probability calculations to compute how probable it is to observe the sample data if the null hypothesis were true. If the resulting probability is very low, we may doubt whether the null hypothesis is indeed true.

The steps involved in hypothesis testing are as follows:

1. Formulate hypotheses: This involves stating a testable proposition about population parameters.
2. Calculate a test statistic: The test statistic describes how many standard errors away from the population statistic, under the null hypothesis, the sample statistic is.
3. Calculate the p-value: The p-value represents the probability of observing a value at least as extreme as the sample statistic, assuming the null hypothesis is true.
4. Draw a conclusion about the null hypothesis: Based on the p-value, we either reject or fail to reject the null hypothesis.

Hypotheses can be formulated as equality or inequality statements. Equality hypotheses state that a value, difference, or effect is equal to zero, while inequality hypotheses state that a value, difference, or effect is larger or smaller than a specific value. It's important to keep in mind that hypothesis testing does not provide evidence for hypotheses but rather helps in casting doubt on a null hypothesis.

In addition to the null hypothesis, we can also specify an alternative hypothesis. The specification of the alternative hypothesis depends on a bit of philosophy of science. Fisher's philosophy suggests using only a null hypothesis; if this null hypothesis is rejected, the "truth" must be anything other than the null hypothesis. We could thus say that, according to Fisher's philosophy, the alternative hypothesis is the negation of the null hypothesis. If $H_0 \quad = 0$, then $H_a \quad 0$; or, if $H_0 \quad > 0$, then $H_a \quad 0$. The alternative hypothesis is in both cases the "opposite" of the null hypothesis.

Neyman-Pearson's philosophy instead involves stating specific null and alternative hypotheses, with an explicit expected effect size for the alternative hypothesis. Assuming a specific expected effect size allows us to calculate the probabilities of drawing correct or incorrect conclusions.

In hypothesis testing, we calculate a test statistic, which measures the distance between the hypothesized population value and the sample statistic in terms of standard errors. The probability of observing a test statistic at least as extreme as the one we did observe is computed using an appropriate probability distribution. For many tests, we use either the Z-distribution or t-distribution, depending on whether we know the population standard deviation or not. This gives us a probability value (p-value), representing the probability of observing data as extreme as or more extreme than the sample data, assuming that the null hypothesis is true.

When interpreting p-values, it's crucial to understand that they give the probability of observing certain data assuming the null hypothesis is true, rather than providing the probability of the null hypothesis being true or false. The p-value is then compared to a pre-determined significance level (usually denoted as alpha) to make a decision about accepting or rejecting the null hypothesis.

Rejecting the null hypothesis indicates that the observed data is unlikely to occur if the null hypothesis were true. On the other hand, failing to reject the null hypothesis means that the observed data is not surprising or does not provide sufficient evidence to reject it.

When testing hypotheses, we can make two types of errors: A Type I error refers to rejecting the null hypothesis when it is true (a false-positive conclusion), while Type II error refers to accepting the null hypothesis when it is false (failing to detect a true effect).

9.0.3 Causality

Causal knowledge explains **what would change if we intervened**: "If we change X, Y will tend to change in a predictable direction." This is why causality matters for science and policy—causal claims connect description to **action** (Shadish, Cook, & Campbell, 2002). Two levels are useful to distinguish. **Type (general) causality** concerns population regularities (e.g., "Smoking causes lung cancer."). **Actual (token) causality** concerns whether a **specific** event in a concrete situation caused a particular outcome (e.g., "This crash occurred because the brakes failed."). For the latter, the modified Halpern–Pearl account captures the core intuition: X counts as a cause of Y when (i) both occurred; (ii) had X been different, and relevant background factors held as they actually were, Y would have been different; and (iii) nothing extraneous is needed for the claim (Halpern, 2015).

How do we justify causal claims in practice? A helpful everyday test traces to Mill's requirements: **covariation**, **temporal precedence**, and **non-spuriousness** (Oppewal, 2010). Randomized experiments are powerful because they **enforce** these conditions by design: random assignment severs links from unmeasured causes to treatment, outcomes are measured **after** assignment, and treatment–control contrasts establish covariation (Shadish et al., 2002). Outside experiments, causal inference relies on design and assumptions—e.g., careful measurement of confounders, quasi-experimental strategies, and transparent modeling. Hypothesis tests then assess whether the observed data are compatible with specific causal predictions, but **testing alone does not create causality**: the strength of a causal claim ultimately rests on research design, the plausibility of assumptions, and the theory's survival of severe attempts at refutation.

9.1 Lecture

https://www.youtube.com/embed/-1J2Ge0B3Ro

9.2 Statistical Power

9.2.1 Hypothesis Testing: Type I and Type II Errors

When we conduct a null-hypothesis significance test, we select the significance level . Alpha is the probability of committing a Type I error (drawing a false-positive conclusion). Since we select the alpha level, it is known. If we use = .05, that means that - by definition - we accept a 5% risk of committing a Type I error.

There is also the probability of committing a Type II error. This is called . We don't know the value of beforehand, but we can calculate it if we make some assumptions. The probability of committing a Type II error (drawing a false-negative conclusion) depends on a few factors:

9.2.1.1 How big the effect is

Big effects are harder to miss; imagine trying to detect a difference between two groups. If the mean of both groups is really close together, it will be harder to detect a difference (see below):

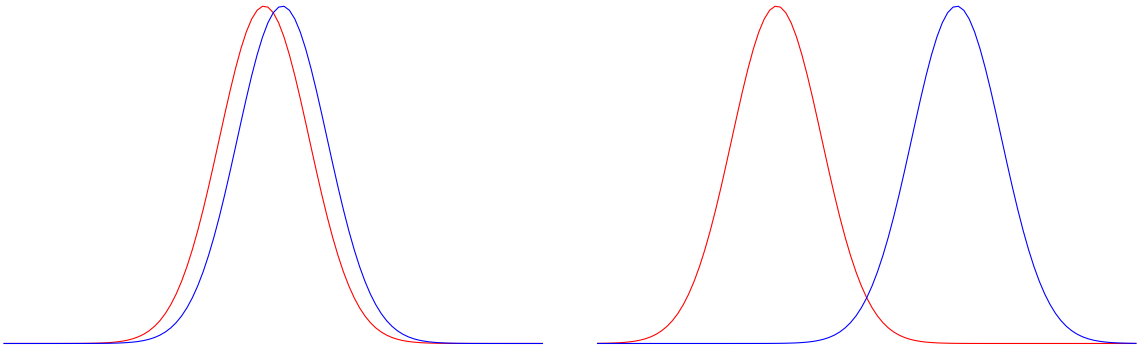


Figure 9.1: Effect of Effect Size

9.2.1.2 How big the sample is

Large samples make it easier to detect smaller effects; imagine that the two distributions below are sampling distributions for two groups with very small sample sizes (left) and very large sample sizes (right):

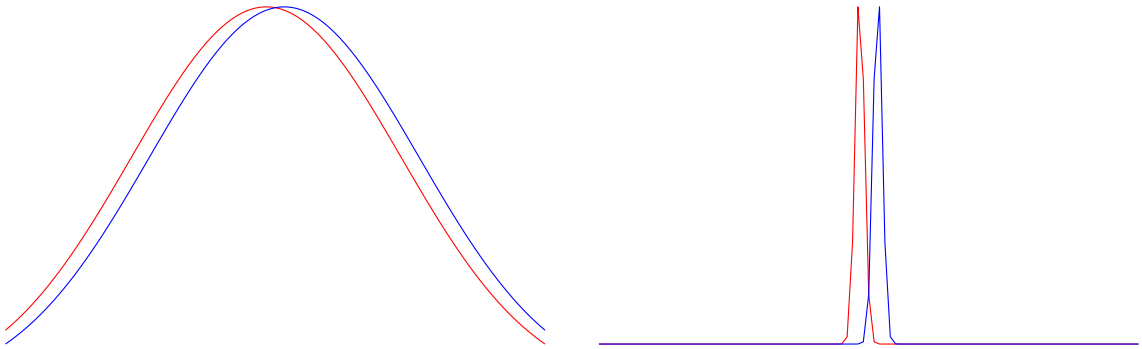


Figure 9.2: Effect of Sample Size

9.2.1.3 How 'noisy' the data are

The standard deviation is a measure of how "noisy" the data are. If observations are very spread out (high standard deviation), it will be harder to detect small differences. Consider that a small difference between two groups would be hard to detect if the two groups overlapped very much (= high standard deviation). Look at the same picture from the previous point (sample size); it illustrates this principle. The reason that both sample size and "noise in the data" have an impact on the probability of committing a Type II error is because they are used to calculate the standard error:

$$SE_M = \frac{SD}{n}$$

9.2.2 Power of a Test

The "power" of a test is the probability that it will correctly detect a true effect of a specific size. Since is the probability of *missing* a true effect, it follows that 1 must be the probability of *detecting* a true effect, or the power.

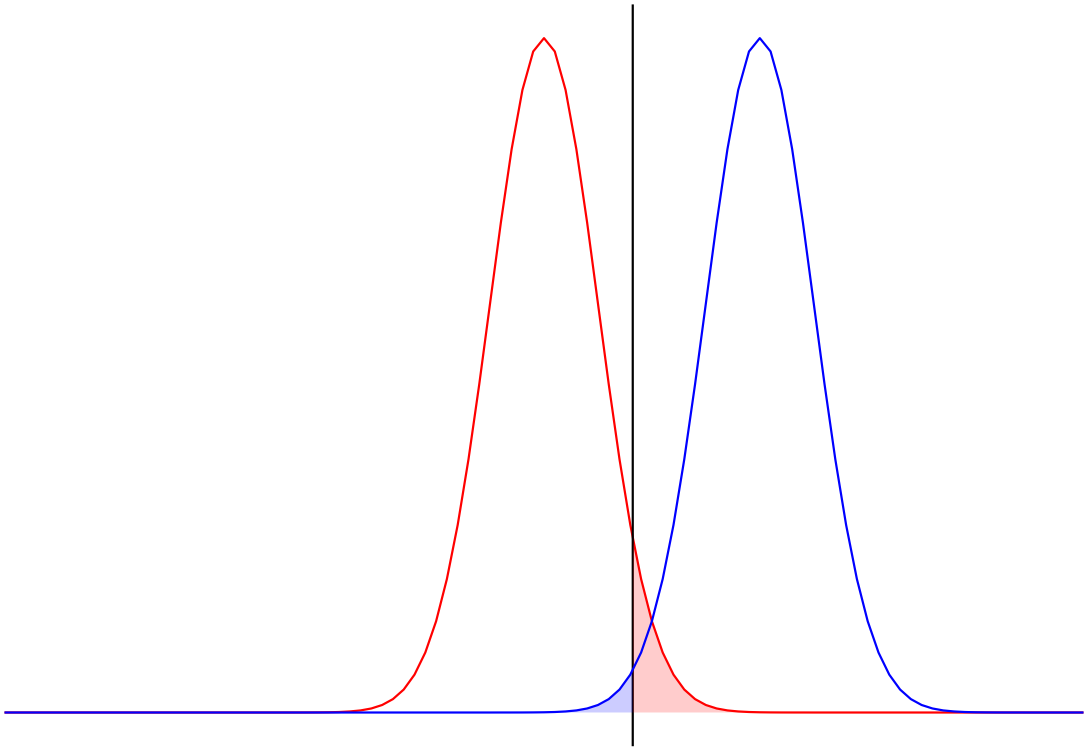
As explained in the previous paragraph, we must know a few pieces of information to be able to calculate :

1. Effect size
2. Sample size
3. Standard deviation

When we conduct a study, we often know the sample size and standard deviation. The effect size is unknown, but we can assume a specific effect size. Think of this as an "informative" alternative hypothesis. The standard alternative hypothesis in null-hypothesis significance testing is just "anything that's not the null hypothesis". So if $H_0 \quad = 0$, then $H_a \quad 0$. Now, we

must specify an exact value. For example, we could choose the smallest effect size of interest as the alternative hypothesis: Let's say we'd be interested in a mean value of $= 0.2$. Then we could set our informative alternative hypothesis as $H_i \quad = 0.2$.

Now we have all the information needed to calculate the power of the test. To do so, we draw two sampling distributions (see illustration below): One (in red) centered around the null hypothesis, $H_0 \quad = 0$, and one centered around the informative alternative hypothesis, $H_i \quad = 0.2$. We find the critical value in the red distribution around the null hypothesis; remember that is the 5% of probability in the right tail of the red distribution. But we can now also calculate , the unknown probability in the tail of the blue distribution **to the left** of the critical value. If the informative alternative hypothesis is true, then this is the probability of failing to detect that true effect. Although this example has no numeric values, we see that the blue shaded area representing is slightly smaller than the red shaded area representing , so the probability of committing a Type II error must be less than .05, and therefore the power 1 must be greater than 95%! If our assumptions are correct, we'd be really well able to detect a true effect of the size specified under H_i.



9.2.3 Try it Yourself

Now, let's calculate this by hand. Imagine that last year's average grade was $M = 5$, with a standard deviation of $SD = 1.5$. This year, we have 73 students. We've made some changes to

the teaching material, and we hope to reach an average grade of $M = 6$.

Assume that the standard deviation this year will be the same as last year, and calculate the power of being able to detect a mean grade of $H_i \ = 6$ when the null hypothesis is that the mean grade is the same as last year, $H_0 \ = 5$.

Step 1: Calculate the SE

We calculate the SE as $SE = \frac{SD}{n} = \frac{1.5}{73} = 0.18$

Step 2: Calculate Critical Value

The critical value is the boundary that corresponds to $\ = .05$ in the distribution centered around H_0. Looking at the t- or Z-table (because sample size is »30), we see that this corresponds to a Z-value of about 1.64.

Z	0	0.01	0.02	0.03	0.04
1.6	0.055	0.054	0.053	0.052	0.051

Converting this back to a score on the grades scale, we get:

$$\text{Grade}_{\text{critical}} = (Z_{\text{critical}} \ \ SE) + _{H_0} = (1.64 \ \ 0.18) + 5 = 5.3$$

Step 3: Get Left-Tail Probability for That Value

Now, we just need to get the left-tail probability for that critical value, in the blue distribution. Convert that critical value back to a Z-value, but now in the blue distribution which is centered around $_{H_i} = 6$:

$$Z = \frac{\text{Grade}_{\text{critical}} \ _{H_i}}{SE} = \frac{5.3 \ \ 6}{0.18} = 3.89$$

This is an extremely large (negative) Z-value; it's not even in our table. Thus, the left-tail probability will be tiny - $< .01$.

That means that our power to detect a true effect of 6 would be very high - $1 \ = 1 \ \ .01 = .99$**, 99%!**

9.3 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

For a two-tailed Z-test of the sample mean, the p-value is the probability of finding a more extreme sample mean than the observed sample mean, if the alternative hypothesis were true. [1]

- (A) TRUE
- (B) FALSE

Question 2

A researcher performs a Z-test to test the hypotheses H0: mu = 0 versus H1: mu > 0. She finds a test statistic of Z = 2.03 and a one-tailed p-value of 0.02. Statement: If the researcher had performed a two-tailed Z-test, the value of the two-tailed p-value would have been halved: 0.01. [2]

- (A) TRUE
- (B) FALSE

Question 3

Chris expects that people who have bungee jumped will score high on average on the Big 5 personality trait 'Openness to Experience'. Openness to Experience has been measured on a 10 point scale (1= not at all open, 10 = extremely open). He takes a random sample of 45 persons who have bungee jumped and observes a mean of 6 and SD of 1.954. He tests the following hypotheses: H0: mu =< 5.5 versus H1: mu > 5.5 with a one-sample t-test. He assumes that Openness to Experience is normally distributed in the population. What is the smallest significance level for which Chris can reject the null hypothesis? [3]

- (A) Cannot reject H0

[1] FALSE
[2] FALSE
[3] 0.05

- (B) 0.01
- (C) 0.1
- (D) 0.05

Question 4

What is the purpose of inferential statistics? [4]

- (A) Calculating sample statistics
- (B) Testing the null hypothesis
- (C) Estimation of population parameters
- (D) Using sample data to infer properties of the population

Question 5

What is the standard error? [5]

- (A) An estimate of the uncertainty about the average of the sample values
- (B) The spread of the sample data
- (C) An estimate of the average sampling error when estimating a population parameter using a sample statistic
- (D) An estimate of the uncertainty about the sample statistic

Question 6

What is a hypothesis in the context of statistical testing? [6]

- (A) An explanation for observed phenomena
- (B) A statement that some parameter is equal to zero
- (C) Something you want to know about the population
- (D) A proposition about the population that can be tested in a sample

[4] Using sample data to infer properties of the population
[5] An estimate of the average sampling error when estimating a population parameter using a sample statistic
[6] A proposition about the population that can be tested in a sample

Question 7

What is meant by 'power' in statistical testing? [7]

- (A) The probability of rejecting the null hypothesis
- (B) The probability of committing a Type I error
- (C) The probability of correctly finding a true effect
- (D) The probability of committing a Type II error

Question 8

You want to test if the mean height of a sample of 50 students is significantly different from the population mean of 65 inches. The sample mean is 68 inches, and the standard deviation is 2 inches. What is the calculated t-value for this hypothesis test? [8]

- (A) 10.61
- (B) 1.50
- (C) 240.42
- (D) 75.00

Question 9

You want to test if the average time spent on a particular task is different from 30 minutes. You collect a sample of 25 participants, and the sample mean time spent on the task is 28 minutes with a standard deviation of 3 minutes. Conducting a two-tailed t-test, what is the calculated t-value? [9]

- (A) -16.67
- (B) 3.33
- (C) -3.33
- (D) -0.67

[7] The probability of correctly finding a true effect
[8] 10.61
[9] -3.33

Show explanations

Question 1
When calculating a test statistic, we assume the null hypothesis to be true - not the alternative hypothesis.
Question 2
The p-value for a two-tailed test is twice as large as for a one-tailed test (because you have the same one-tailed probability in both tails). For two-tailed tests, if the observed effect is in the direction of the alternative hypothesis, you can half the two-tailed p-value.
Question 3
Divide the standard deviation by the square root of 45 to get the standard error. Then, divide the difference between the observed mean of 6 and the hypothesized mean of 5.5 by that standard error to get the test statistic. Then, find the critical t-values for a one-sided test with the three alpha levels mentioned in the answers in the t-distribution for 44 degrees of freedom (n - 1). Note that the answer is .05!
Question 4
Inferential statistics involves using sample data to make inferences or draw conclusions about the larger population from which the sample was drawn. It allows researchers to make educated guesses about population parameters based on the information collected from the sample. Calculating sample statistics is a step in the inferential process, but it is not the primary purpose of inferential statistics. Testing the null hypothesis is another inferential procedure, but it is a specific type of hypothesis testing, and not the overall purpose of inferential statistics.
Question 5
The standard error is a measure of the uncertainty associated with the sample statistic as estimator of the population parameter. It represents how much the sample statistic is expected to vary from one sample to another if multiple samples were drawn from the same population.
Question 6
In statistical testing, a hypothesis is a testable proposition about the population that can be examined using sample data. It is a statement or assumption that researchers put to the test to determine if there is evidence to support it or not. The hypothesis is formulated based on the theory or observations made about the population.
Question 7
Power in statistical testing refers to the probability of correctly detecting a true effect or relationship between variables. It is the likelihood of finding a significant result in a study when the effect being investigated truly exists in the population. It is important to have sufficient power in a study to avoid false-negative findings, where we fail to reject the null hypothesis when there is a real effect. Power is one minus the probability of committing a Type II error, or 1-beta.
Question 8
To calculate the t-value for this hypothesis test, you can use the formula: t = (sample mean - population mean) / (standard deviation / square root of sample size). Plugging in the values, t = (68 - 65) / (2 / sqrt(50)) = 10.61.
Question 9
The t-value for a two-tailed t-test can be calculated using the formula: t = (sample mean - population mean) / (standard deviation / square root of sample size). In this case, the population mean is 30 minutes. Plugging in the values, t = (28 - 30) / (3 / sqrt(25)) = -3.33.

9.4 Tutorial

9.4.1 Assignment 1: Hypothesis Testing - Formulating Hypotheses

Discuss with your portfolio group the logic behind hypothesis testing, and how it relates to your personal (and group's) research interests.

Consider the following three research descriptions. Formulate H0 and H1 in words. Discuss your answers with your group members.

Researchers want to know whether it matters for test performance if an exam is completed on a computer or using paper and pencil. Hence, the research question reads: Is there an effect of the type of administration (computer or paper and pencil) on the test performance?

What would be the H0 and HA for this study?

Show answer

This appears to be an undirected hypothesis about a mean difference for two independent samples, without a clearly specified alternative hypothesis. Thus, we could state:
H_0 $_{computer} = _{paper}$ H_A $_{computer}$ $_{paper}$

Researchers want to know whether the alcohol consumption among Dutch students differs from the alcohol consumption in the general Dutch population. Using CBS statistics, they know that in the general population the average alcohol consumption is 5.6 glasses a week. The question is whether the average alcohol consumption among students is different from this national average.

What would be the H0 and H1 for this study?

Show answer

This appears to be an undirected hypothesis about the difference between a mean and a hypothesized value, without a clearly specified alternative hypothesis. Thus, we could state:
H_0 $= 5.6$ H_A 5.6

Researchers want to study whether social isolation is associated with income.

What would be the H0 and H1 for this study?

Show answer

This appears to be an undirected hypothesis about an association between two variables, without a clearly specified alternative hypothesis. We could thus state:
H_0 $= 0$ H_A 0

Formulating the hypothesis is an important very first step in hypotheses testing. Continue with the next assignment, in which we will go through the steps of a hypothesis test.

9.4.2 Assignment 2: Test Statistics, Alpha and Significance

In this assignment we will go through the steps of a hypothesis test.

While going through the steps we will come across the most important concepts related to hypothesis testing.

For the next steps, we consider the following situation:

Suppose we are interested in the personality profile of musicians; that is, we want to know whether, on average, personality characteristics of musicians differ from those of the general population. For now, we'll only focus on Openness. We pretend that we have collected data among 25 musicians using a validated scale for which previous research has shown that in the general population the scores are normally distributed with mean 50 and SD 15. It is our task to test whether the mean of Openness for musicians differs from the mean in the general population. To keep things simple, we assume that in the population of musicians the SD is the same as in the general population; that is, we assume that LaTeX: $= 15$.

Let openness be the variable of interest. Let $_{musicians}$ represent the mean openness in the population of musicians. The hypothesis test amounts to testing:

$H0$ $_{musicians} = 50$

$H1$ $_{musicians}$ 50

Now, when we do the hypothesis test, we seek for evidence against the null hypothesis. More specifically, our testing procedure starts with the assumption that H0 represents the truth and as long as we don't have convincing evidence that our assumption is false we stick to that assumption.

The question is, however, when do we have convincing evidence against H0?

Finding evidence against H0 works as follows:

If H0 is true, we expect mean values close to H0. And, if we observe a mean value that is much different from the value under H0, we have convincing evidence against H0. If this happens, we reject H0 as representing the truth and accept the alternative hypothesis, H1.

Hypothesis testing fits Popper's philosophy of falsification. He introduced this well-known analogy to explain the logic of falsificationism:

1. Suppose we assume that all Swans are white, H_0 $Swans = white$
2. We would then not expect to observe black ones.
3. If we do observe black swans, our initial hypothesis is called into question.

4. The number of white swans we see (= observations consistent with the hypothesis) does not provide evidence for H_0, because there could always be a black swan out there we haven't observed yet.

So, the next questions are:

What are the sample values we can expect under H0? When is evidence "convincing" enough? To answer the first question we have to go back to sampling distributions!

For the second question, we need a criterion. We have to realize that even if H0 is true, sample values can be far off just by sampling fluctuations (i.e., by chance). The common criterion is: if the observed value is among the 5% most unlikely samples under H0 (i.e. if H0 is true), we reject the null hypothesis.

Let's go back to our example about musicians.

Let X be openness. Under H0 we assume that X is normally distributed with mean 50 and SD equal to 15.

What are the mean and standard deviation of the sampling distribution of the mean under H0 given that the sample size is 25? And what do we call the standard deviation of the sampling distribution?

(Use what you have learned in the previous lectures. Hint: first make a drawing of the situation, then do the computations).

Explanation

Sampling distribution:

- Mean: $= 50$
- Standard error (=SD of sampling distribution!): $_X = \frac{15}{25} = 3$

Suppose we want to indicate sample means that are unlikely if H0 would be true. In particular, we want to know how far the sample mean must be from the hypothesized mean to be among the 5% of all possible samples under H0 that are furthest away from the hypothesized means.

What should the value of the sample mean be to fall within the 5% most deviant samples if the sample size is 25?

Explanation

We are talking about the distribution of the mean; so we need to work with the sampling distribution. We want to know the cut offs that mark the 2.5% highest and 2.5% lowest means. We first have to find the Z-values: they are 1.96 for the highest 2.5%, and (by symmetry) -1.96 marks the 2.5% lowest.
Hence, to be among the 5% of all possible sample means that are most unlikely under H0, the sample mean should be:
larger than 50+1.96 x 3 = 55.88 or smaller than 50-1.96 x 3 = 44.12

Let's do some more exercises on the Z-test.

Suppose the mean for Openness we found in our sample was 59.

If we use a significance level of 5%, would we reject the null hypothesis? [10]

- (A) Yes
- (B) No

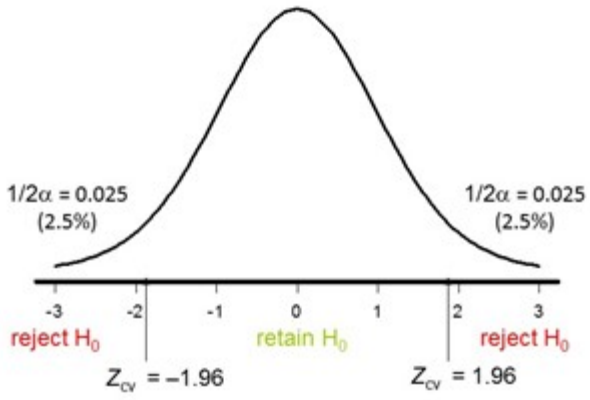
In the previous step we used cut offs for the sample means to decide about significance. The cut off scores were obtained via the Z-distribution. However, doing all these computations is not necessary (there's a shortcut!!). In fact, if we know the Z-value for the sample, we can easily find out if the sample is among the 5% of the most unlikely sample means. We only have to compare the value with 1.96 and -1.96 to see whether that is the case.

In this course, we will use Z-values for different purposes. In these specific calculations, Z is used as a *Test Statistic*. A test statistic quantifies evidence against the null hypothesis. In this case, the Z test statistic expresses how far away from the mean under the null hypothesis the observed mean is, in terms of the number of standard errors.

The Z test-statistic follows the standard normal distribution. The values 1.96 and -1.96 are called the critical values and they mark the 5% most unlikely sample means under H0. In other words, the critical values mark the reject region for H0.

So, if we compute the Z-value for the sample mean, and if that sample value of Z falls in the rejection region, we reject H0 (we found something that is unlikely enough to no longer believe H0 is true). If H0 is rejected we speak of a significant result. See the graph below:

[10] Yes



Following these steps to test a mean is one example of performing a *"Z-test"*!

We can use the Z-test to test hypotheses about the population mean if we know the population .

The test statistic for the Z-test is:

$z = \frac{X_{H_0}}{X}$

This statistic is computed using the mean from the sample, the hypothesized mean under H0 and .

H0 is rejected at the 5% significance level if z is either larger than 1.96 or smaller than -1.96.

So far, we rejected the null hypothesis if the sample is among the 5% most unlikely sample means under H0. This 5% was called the significance level, and is denoted as $= .05$. However, we could just as well choose 1% or .5%.

What would be the critical values for the Z-test if one tests at $= .01$? _____[11]

What would be the critical values for the Z-test if one tests at $= .005\%$? _____[12]

For historical reasons, social scientists tend to use $= 0.05$ as a default. So in this course, if alpha is not explicitly stated, assume $= 0.05$.

When we test hypotheses we reject H0 if the sample we find is unlikely if H0 is true. However, the flip side is that, even though H0 is true, we may find a sample that is much different by chance, and erroneously reject H0. Or, in other words, we could make an error. Rejecting H0 while it is true in reality is called a Type I error!

Consider the following:

1. If H0 is true, and you test at $= 0.05$, what is the probability of committing a Type I error?

[11] 2.58
[12] 2.81

2. What is the link between the -level and type I error rate?

Explanation

1. If H0 is really true (i.e., H0 should not be rejected), then the probability that the sample mean is among the 5% most unlikely is equal to 5%.
2. The alpha level specifies the risk of a Type I error. So if one tests at an alpha level of .05, it means that one accepts a risk of 5% to commit a Type I error.

Properties of the Z-test:

Used to test hypotheses about the mean in a population, assuming known.

The test-statistic equals $z = \frac{X}{x}$

The test statistic is normally distributed.

9.4.3 Assignment 3: Z-test

In this assignment we will apply the Z-test.

This assignment first presents an example, followed by two practice questions.

A researcher wants to test H_0 = 50 against H_1 50

Data are available from a random sample of 26 respondents. The mean was 53.7. The researcher assumes the SD in the population is 8.5. Perform all steps of the Z-test.

Explanation

Step 1: Formulate hypotheses
H_0 = 50 H_1 50
Step 2: Compute test statistic
Standard error: $\frac{8.5}{26} = 1.667$
Test statistic: $z = \frac{53.750}{1.667} = 2.212$
Step 3: Decide about significance
 = .05, so critical values +/- 1.96.
Our test statistic exceeds this critical value.
The sample mean thus falls in the rejection region, and we should conclude that the test is significant so H_0 s rejected.
Step 4: Draw conclusion
We have convincing evidence that the population mean differs from 50.

A researcher wants to test whether the population mean is equal to 80. Data are available from a random sample of 60 respondents. The mean was 74. The researchers assume the SD in the

population is 40. Perform and report all steps of the Z-test. What is the resulting p-value? [13]

A researcher wants to test whether the population mean is equal to 500. Data are available from a random sample of 75 respondents. The mean was 546. The researchers assume that the SD in the population is 200. Perform all steps of the Z-test. Use = .01. Perform and report all steps.

Show answer

Step 1: Hypotheses: H_0 = 500, H_1 500
step 2: Compute Statistic:

- standard error: $\frac{200}{75} = 23.094$
- test statistic: $z = \frac{546500}{23.094} = 1.992$

Step 3: Decide about significance.
Z does not exceed +/- 2.576. This means that Z does not fall in the reject region when tested at the 1% significance level. The test is not significant.
Step 4: Draw conclusion
H_0 is not rejected.

9.4.4 Quiz

"The null and alternative hypothesis are deduced from the data." TRUE / FALSE[14]

"When performing a hypothesis test, we start by assuming H_0 is true." TRUE / FALSE[15]

"If we reject H_0 with = 0.05, then we will also reject it at = 0.10, assuming all other quantities are held constant." TRUE / FALSE[16]

Explanation

The critical values of = 0.05 are +/- 1.96. Hence, if H_0 is rejected it means that z in the sample is larger than 1.96 or smaller than -1.96."
The critical values of = 0.1 are +/- 1.645. This means that for rejecting H_0 at this alpha level, that z should be larger than 1.645 or smaller than -1.645. That is implied by the fact that it exceeds +/- 1.96.

"If we reject H_0, then H_0 is surely wrong." TRUE / FALSE[17]

[13] 0.12
[14] FALSE
[15] TRUE
[16] TRUE
[17] FALSE

Explanation

We should always be aware of the possibility of making a Type I error. The probability of making a Type I error is equal to .

"Increasing the sample size n (and holding all the rest constant) decreases the probability of a Type I error." TRUE / FALSE[18]

Explanation

Increasing the sample size n (and holding all the rest constant) does not decrease the probability of a Type I error.
The Type I error is determined by the alpha level.
If our sample is among the 5% most unlikely sample means of all possible sample means with the same size under H_0, whatever that sample size N may be.
Increasing the sample size n (and holding all the rest constant) does not decrease the probability of a Type I error.

9.4.5 Assignment 4: Z-test and Alpha-levels

In this assignment we will practice some more with the Z-test, meanwhile we will review important concepts of hypothesis testing. In particular, we will look at significance levels.

To test hypotheses, we need to specify the "significance level", usually denoted by . The significance level is our *decision criterion* to reject H0.

The most common choice is .05. But what does this criterion exactly entail?

Discuss with your group what an level entails.

Explanation

If we test at an of .05 it means that we are willing to reject H0 in favor of H1 if our sample mean belongs to the 5% most extreme scores (2.5% in each tail) under the null hypothesis.
If indeed the sample mean is among this 5%, it means that we have observed a sample in a range that is quite unlikely if the null hypothesis would be true and, therefore, justifies rejection of the null hypothesis.

In the previous assignments you already used the critical values for the Z-test for specific alpha levels.

[18] FALSE

For two-tailed tests, it holds that if the absolute value of Z exceeds the critical value, we may reject H_0.

Let Z_{crit} be the critical value. For the Z-test it holds that:

- $Z_{\text{crit}} = 1.65$, if $= 0.10$ (two-tailed)
- $Z_{\text{crit}} = 1.96$, if $= 0.05$ (two-tailed)
- $Z_{\text{crit}} = 2.58$, if $= 0.01$ (two-tailed)

9.4.6 Quiz

Researchers want to test whether $= 70$. They assume that $= 10$. Researchers found a mean of 72 in a random sample of 40 persons.

True or false:

H_0 can be rejected at one of the three levels discussed above ($= .10, .05, .01$. TRUE / FALSE[19]

"If the two-tailed test is significant at the 5% level, it will also be significant at the 1% level (keeping everything else fixed)." TRUE / FALSE[20]

"If the two-tailed test is not significant at the 10% level, it won't be significant at the 5% level either (keeping everything else fixed)." TRUE / FALSE[21]

"If the two-tailed test is not significant at the 5% level, it could still be significant at the 10% level (keeping everything else fixed)." TRUE / FALSE[22]

"If the two-tailed test is significant at the 1% level, it might not be significant at the 5% level (keeping everything else fixed)." TRUE / FALSE[23]

9.4.7 Assignment 5: P-values

We will now focus on the interpretation of the p-values and how to use the p-values to decide about significance.

Consider the following situation:

Scores on a test measuring confidence in police are normally distributed in the general population, with $= 500$ and an $= 50$. Researchers want to know if the average confidence level is different for those who have been a victim of crime. They collect data for 60 victims. They find a sample

[19] FALSE
[20] FALSE
[21] TRUE
[22] TRUE
[23] FALSE

mean of 511. They test $H_0 \quad = 500$ against $H_1 \quad 500$, while assuming that the population variance is $\quad = 50$.

Compute the p-value. Draw a graph for the two-tailed p-value. Write down in your own words and as precise as possible the interpretation of the p-value in the answer box below. Then, discuss your response with your group.

Explanation

- The p-value represents the proportion of all possible sample means that are further away from our hypothesized mean than the observed sample mean is.
- We have the sampling distribution with $\quad = 500$ and $_X = \frac{50}{60} = 6.455$.
- First, we compute the right-tail area: $P(X > 511) = P(Z > 1.70) = 0.0446$.
- Hence, 4.66% of all possible samples is further away from H_0 on the right side.
- Second, we compute the left-tail area. These are the sample means that are more than 11 points from the hypothesized mean to the left $P(X < 489) = P(Z < 1.70) = 0.0446$.
- Hence, the two-tailed p-value is 0.0892.

Is the test significant at the 5% level? TRUE / FALSE[24]

Is it significant at the 1% level? TRUE / FALSE[25]

Researchers test whether $\quad = 90$. They assume that $\quad = 21$. The sample mean was 85. Sample size was 50.

What is the two-tailed p-value? _____[26]

What is the highest level at which the test is significant? [27]

- (A) 0.01
- (B) 0.005
- (C) 0.1
- (D) 0.05

Researchers test whether $\quad = 35$. They assume $\quad = 16$. The sample mean was 38. Sample size was 64.

Compute the two-tailed p-value and indicate which of the following statements is true.

[24] FALSE
[25] FALSE
[26] 0.09
[27] 0.05

[28]

- (A) The test is significant at the 10% and 5% level, but not at the 1% level.
- (B) The test is significant at the 10%, 5% level, and 1% level.
- (C) The test is significant at the 10% level, but not at 5% or 1% level.
- (D) The test is not significant at 10%, not significant at 5% and not significant at 1%.

Consider these true- or false statements:

If a two-tailed p-value is .0567 then the test is significant at the 10% level but not at the 5% level. TRUE / FALSE[29]

If a two-tailed test is significant at the 5% level but not at the 1% level, then the two-tailed p-value will be less than 0.01. TRUE / FALSE[30]

A two-tailed p-value of 0.060 indicates that we have 6% chance that the null hypothesis is true. TRUE / FALSE[31]

[28]The test is not significant at 10%, not significant at 5% and not significant at 1%.
[29]TRUE
[30]FALSE
[31]FALSE

10 GLM-I: Linear Regression

The General Linear Model (GLM) is a family of models used to analyze the relationship between an outcome variable and one or more predictors. In this lecture, we will focus on bivariate linear regression, which describes a linear relationship between a continuous outcome variable and a continuous predictor. However, it's important to note that the GLM encompasses other members that can handle predictors of any measurement level (continuous or categorical), multiple predictors, transformations of the outcome and predictors, and different error distributions.

Linear regression is based on the concept of using information about other variables associated with the outcome to improve predictions. It begins with the understanding that the mean is the best predictor (expected value) when no further relevant information is available. However, if we have information about other variables, such as the number of hours studied being strongly associated with exam grades, we can use that information to enhance our predictions. This process is known as regression.

To visually explore associations between two variables, we often use scatterplots. Scatterplots require both variables to be at least of ordinal measurement level. By plotting the data points, we can observe whether there is a linear pattern or trend. In linear regression, we aim to find a line that represents the best possible predictions. This line, called the regression line, goes through the middle of the cloud of data points.

The regression line is described by the formula Y = a + bX, where "a" is the intercept (the predicted value when X equals 0) and "b" is the slope (how steeply the line increases or decreases). The predictions made using the regression line are not identical to the observed values, as there is always some prediction error. The Ordinary Least Squares method is used to obtain the line that minimizes the sum of squared prediction errors.

In a bivariate regression, the regression formula expands to include the individual prediction error, assuming that the errors are normally distributed around the regression line with a mean of zero. The regression model is represented as Yi = a + b * Xi + ei, where Yi is the individual's score on the dependent variable, a is the intercept, b is the slope, Xi is the individual's score on the independent variable, and ei is the individual prediction error.

Hypothesis tests can be conducted on the regression coefficients to determine their significance. The default null hypothesis for the intercept is that it is equal to zero, while the null hypothesis for the slope is also zero. The t-test is commonly used, with the degrees of freedom being n - p, where n is the sample size and p is the number of parameters. By testing the coefficients,

we can determine the statistical significance of the relationship between the predictor and the outcome.

While linear regression offers valuable insights, it is essential to consider the assumptions underlying the model. These assumptions include linearity of the relationship between the predictor and the outcome, normality of residuals (prediction errors), homoscedasticity (equal variance of residuals), and independence of observations. Violations of these assumptions can affect the validity of the model and lead to misleading results. Checking and addressing these assumptions is crucial for accurate and reliable regression analysis.

Linear regression is a powerful tool for analyzing the relationship between variables, and a building block for many more advanced analysis techniques. It allows us to make predictions based on available information and understand the strength and significance of the relationship between a continuous predictor and continuous outcome. By considering the assumptions and conducting hypothesis tests, we can ensure the validity of our regression models and draw meaningful conclusions from the analysis.

10.1 Lecture

https://www.youtube.com/embed/Mkc17DG4KdI

10.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the General Linear Model (GLM)? [1]

- (A) A family of models to analyze the relationship between one outcome and one or more predictors
- (B) A family of models to analyze the relationship between continuous outcomes and categorical predictors

[1] A family of models to analyze the relationship between one outcome and one or more predictors

- (C) A family of models to analyze the relationship between categorical outcomes and continuous predictors

- (D) A family of models to analyze the relationship between multiple outcomes and multiple predictors

Question 2

What type of relationship does bivariate linear regression describe? [2]

- (A) A relationship of any shape between a continuous outcome variable and a predictor of any measurement level, with normally distributed prediction errors

- (B) A linear relationship between a categorical outcome variable and a continuous predictor

- (C) A linear relationship between a continuous outcome variable and a predictor of any measurement level with normally distributed prediction errors

- (D) A nonlinear relationship between a continuous outcome variable and a continuous predictor

Question 3

What does it mean when we say 'The mean is the best predictor when there's no further relevant information' in the context of regression? [3]

- (A) The mean is the best predictor regardless of whether we have additional information about predictors

- (B) The mean is the expected value when we have no additional information about predictors

- (C) The mean is the least accurate predictor when we have no additional information about predictors

- (D) The mean is only a good predictor when there's no variability in the outcome

Question 4

What is the purpose of a scatterplot in the context of regression analysis? [4]

[2] A linear relationship between a continuous outcome variable and a predictor of any measurement level with normally distributed prediction errors
[3] The mean is the expected value when we have no additional information about predictors
[4] To visualize associations between two variables

- (A) To visualize associations between two variables
- (B) To calculate the mean and standard deviation of two variables
- (C) To explore causal relationships between two variables
- (D) To determine the distribution of two variables

Question 5

What is the primary goal of ordinary least squares regression in linear modeling? [5]

- (A) To find the line that fits the data exactly by minimizing the sum of absolute prediction errors
- (B) To find the line that passes through the mean of the data points
- (C) To find the line that predicts the maximum number of data points correctly
- (D) To find the line that gives the best possible predictions by minimizing the sum of squared prediction errors

Question 6

In the formula 'Yi = a + bXi + ei', what are the parameters?, [6]

- (A) a and b
- (B) X and Y
- (C) Yi, Xi, and ei
- (D) Xi and ei

Question 7

How are the coefficients 'a' and 'b' interpreted in the context of linear regression? [7]

- (A) 'a' is the predicted value when b equals 0, and 'b' is the slope indicating how steeply the line increases or decreases

[5] To find the line that gives the best possible predictions by minimizing the sum of squared prediction errors
[6] a and b
[7] 'a' is the intercept where the line crosses the Y-axis, and 'b' is the slope indicating how steeply the line increases or decreases

- (B) 'a' is the slope indicating how steeply the line increases or decreases, and 'b' is the intercept where the line crosses the Y-axis
- (C) 'a' is the intercept where the line crosses the Y-axis, and 'b' is the slope indicating how steeply the line increases or decreases
- (D) 'a' is the intercept where the line crosses the Y-axis, and 'b' is the predicted value when X equals 0

Question 8

What is the purpose of checking assumptions in linear regression? [8]

- (A) To determine the significance of the predictors
- (B) To ensure that the model accurately represents the data and that inferences are valid
- (C) To find ways to manipulate the data to fit the model better
- (D) To improve the visualization of the scatterplot

Question 9

What is the assumption of homoscedasticity? [9]

- (A) That the effect of the predictor on the outcome is linear and monotonous
- (B) That the dependent variable is normally distributed
- (C) That prediction errors are equally distributed for all values of the predictor
- (D) That the prediction errors are normally distributed

Question 10

Given regression formula Yi = 65.13 + 95.27*Xi+ei, what is the predicted score for a person who scores 15 on X? [10]

- (A) 80.13
- (B) 1494.18

[8] To ensure that the model accurately represents the data and that inferences are valid
[9] That prediction errors are equally distributed for all values of the predictor
[10] 1494.18

- (C) 1072.22
- (D) Can't say

Question 11

Frank scores 22 on Yi and has a prediction error of 7.33. What was his predicted value, given regression formula Yi = 65.13 + 95.27*Xi+ei? [11]

- (A) -0.53
- (B) 22
- (C) 14.67
- (D) 27.33

[11] 14.67

Show explanations

Question 1
The GLM is used to analyze the relationship between a single outcome and one or more predictors.
Question 2
Bivariate linear regression specifically describes a linear relationship between two continuous variables.
Question 3
When there's no further information available, the mean is the most reasonable estimate for the outcome.
Question 4
Scatterplots visually depict the relationships and associations between two variables.
Question 5
Ordinary least squares regression aims to minimize the sum of squared prediction errors to find the best-fitting line.
Question 6
The parameters of a model are the quantities estimated from data. Yi and Xi are the data; ei is calculated based on the model-implied predictions.
Question 7
The coefficient 'a' represents the intercept, and the coefficient 'b' represents the slope of the regression line.
Question 8
Assumption checks ensure that the model accurately represents the data and that any inferences drawn from the model are valid.
Question 9
Homoscedasticity literally means: equal variances; this assumption means that the variance of prediction errors is equal at all values of the predictor.
Question 10
65.13 + 95.27*15 = 1494.18
Question 11
The observed score Yi is equal to the predicted score plus prediction error. If prediction error was 7.33, the predicted score must have been 22-7.33 = 14.67

10.3 In SPSS

10.3.1 Linear Regression

https://youtu.be/0AGLdgUtIJg?si=SZQxa2Qt9oOTEg-O

https://www.youtube.com/watch?v=VEQPX6d-EQw

10.4 Tutorial

10.4.1 Regression Analysis

In this assignment we will make a start with regression analysis.

We will go through the different steps of running and interpreting a regression analysis.

Open the file `Work.sav` to get started.

Consider the following research question: "Does variety at work predict pleasure at work?"

What is the dependent variable in this case? [12]

- (A) Pleasure
- (B) Variety

To answer the research question, we will run a linear regression analysis.

Select the following menu item: Analyze > Regression > Linear

Choose the dependent variable (scpleasure) and independent variable (scvariety). Paste and run the syntax.

If you look in the output, you will see that SPSS shows four tables in the output file.

In the table labeled "Model Summary" we can find the R2 value. R2 indicates the total proportion of explained variance in the dependent variable in the model; this is the focus of next week's class.

What proportion of the variance Emotional Pleasure (scpleasure) is explained by our single predictor Variety at work (scvariety)? _____[13]

Consider the unstandardized Coefficients in the table labeled "Coefficients".

What is the value of the intercept (b0) for the regression line? _____[14]

How should we interpret the intercept (or "constant") within the context of this analysis? [15]

- (A) Someone who reports zero Variety at work (meaning a score of 0 on scvariety) has an expected value of this many points on Pleasure.

[12] Pleasure
[13] 0.195
[14] -9.024
[15] Someone who reports zero Variety at work (meaning a score of 0 on scvariety) has an expected value of this many points on Pleasure.

- (B) For every point in Variety at work, we expect an increase of this many points in Pleasure.
- (C) Everyone who reports zero Variety at work (meaning a score of 0 on scvariety) has a value of this many points on Pleasure.
- (D) The sample average of Pleasure is this many points

Consider the unstandardized regression coefficients again.

What is the value of the regression coefficient of scpleasure on scvariety (b1)? _____[16]

How should we interpret the regression coefficient of scvariety within the context of this analysis?

[17]

- (A) This is the sample average of Pleasure
- (B) If someone's score on Variety at work increases with 1 point, their score on Pleasure increases by this many points.
- (C) This is the sample average score of Variety at work.
- (D) If someone's score on Variety at work increases with 1 SD, their score on Pleasure increases by this many SDs.

The "Coefficients" table also shows whether or not the effect of scvariety on scpleasure is significant.

What is the p-value for the regression coefficient for scvariety? _[18]

Can we conclude that the effect of scvariety on scpleasure is significant? (use $= .05$). [19]

- (A) Yes
- (B) No

[16] 0.618

[17] If someone's score on Variety at work increases with 1 point, their score on Pleasure increases by this many points.

[18] 0

[19] Yes

10.4.2 Assumptions

Recall that regression assumes linearity, normality of residuals, homoscedasticity (equal variance of residuals), and independence of observations. We will check each of these assumptions in turn, except for independence of observations because this is a property of our sampling method and cannot be checked statistically.

10.4.2.1 Scatterplot

A scatter plot can provide some insight into linearity.

To make a scatter plot: Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter

Place variety at work on the X axis and emotional pressure on the Y axis.

Is the assumption of linearity met in this case? [20]

- (A) Yes
- (B) No

10.4.2.2 Regression Diagnostics

Aside from the scatterplot, we can check the assumptions of regression by requesting additional options in the analysis.

Go back to the analysis dialog via Analyze –> Regression –> Linear. Verify that you still have the correct predictor and outcome.

Then, click the Plots button. You want a plot of the predicted values against the residual values, so put ZPRED in the X box and ZRESID in the Y box.

Also check the boxes for a Histogram and normal probability plot, then hit continue.

Now paste and run the syntax. You should see the following added to your previous regression syntax:

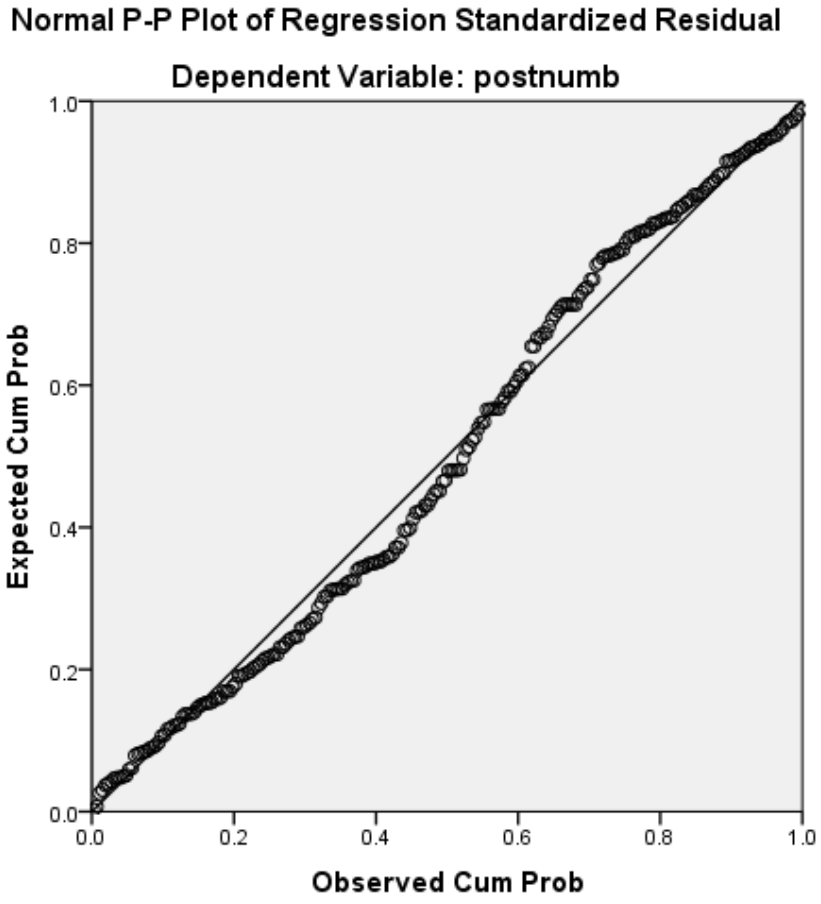
```
/SCATTERPLOT=(*ZRESID ,*ZPRED)
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
```

[20] Yes

10.4.2.3 Linearity

How can we test linearity using this additional output?

First, we can use the "Normal P-P plot". If the relationship is perfectly linear, all dots should be on the diagonal line. If the points are deviating from the line, the relationship is not perfectly linear. Small deviations are OK; for example, the plot below shows a linear association:



Does the P-P plot for your regression give cause for concern for violation of the assumption of linearity? [21]

- (A) Yes
- (B) No

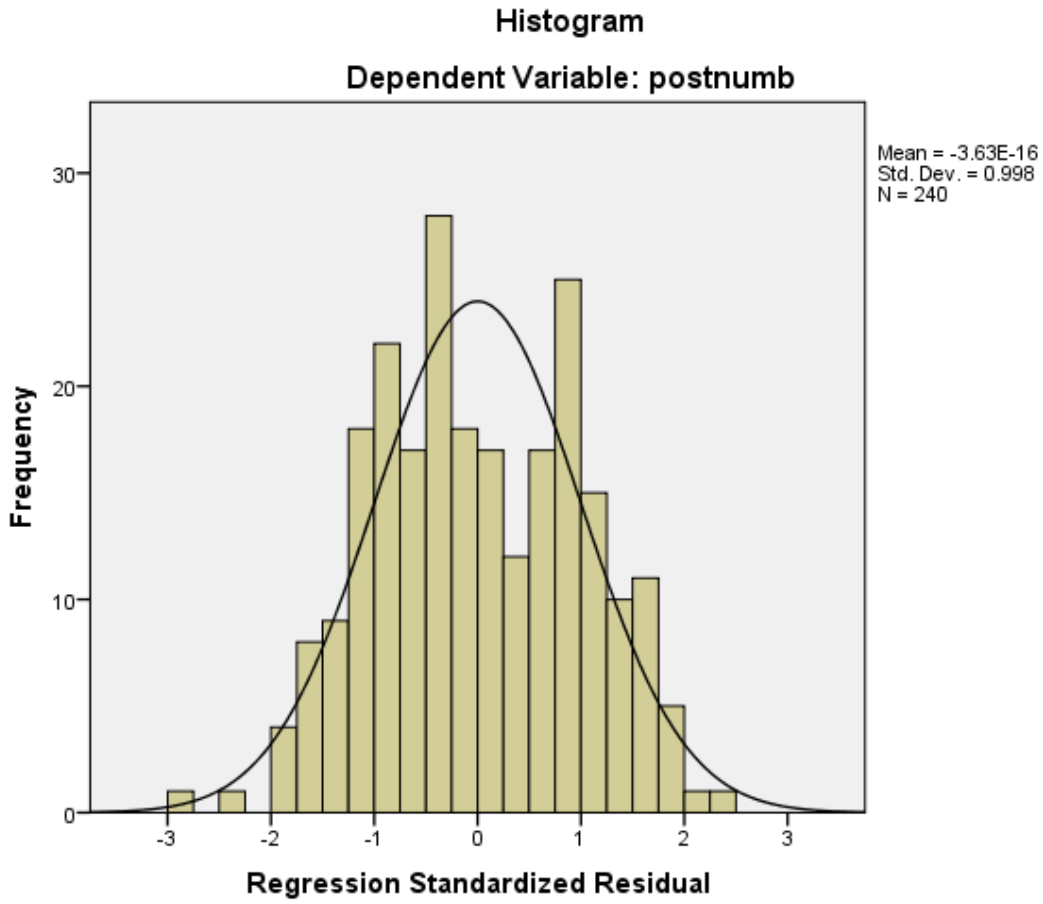
[21] No

- (C) Unclear

10.4.2.4 Normality

One way to check normality is by examining the histogram of residuals. This histogram displays a normal curve by default. If the observed residuals deviate strongly from this histogram, there may be a problem.

The plot below shows a residual histogram with some minor deviations from normality (too few scores near the mean). This is probably still fine:



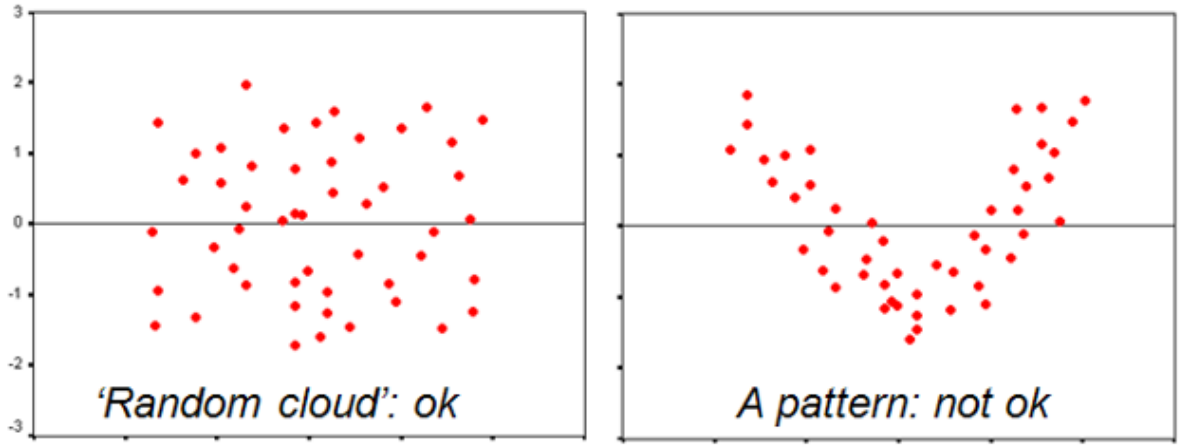
Does the residual histogram for your regression give cause for concern for violation of the assumption of normal residuals? [22]

[22] Yes

- (A) Yes
- (B) No
- (C) Unclear

10.4.2.5 Homoscedasticity

We examine homoscedasticity using a plot of standardized predicted values against standardized residuals. We want residuals to be identically distributed on the Y-axis for all values on the X-axs. In other words, this scatterplot should look like a dot cloud (no pattern) around the zero line (left picture below), and not like a pattern (right picture below).



Does the scatterplot for standardized predicted values against residual values for your regression give cause for concern for violation of the assumption of homoscedasticity? [23]

- (A) Yes
- (B) No
- (C) Unclear

Write up a discussion of potential violations of the assumptions for your regression, then check your answer.

[23] Yes

Explanation

We observed that the observed scores deviated from the P-P plot in an S-shaped pattern. We further observed that, in a histogram of standardized residuals, the observed residuals were right-skewed. Finally, we observed less variance around the regression line for low scores and more variance around the regression line for high scores.
These findings give cause for concern of violations of the assumptions of regression. *One potential explanation is that the effect might be quadratic instead of linear. (Optional)*

11 GLM-II: Sums of Squares

Last week we discussed how linear regression represents the relationship between a predictor variable (X) and an outcome variable (Y) as a diagonal line. This line will have some prediction error for each individual data point. The regression line, by definition, is the line with the smallest possible overall prediction error (across all participants). Today, we explore this concept of "smallest possible overall prediction error" in more detail.

The sum of prediction errors across all participants is always zero because the regression line passes through the "middle" of the data. So there's always an equal amount of negative prediction errors and positive ones, which cancel each other out. To calculate the "total prediction error", we must square the prediction errors, which eliminates the negative values and ensures that we can add them up to a positive number. We call the sum of squared prediction errors the the "sum of squared errors" (SSE). When we estimate a regression model in statistical software, we ask it to find the regression line that minimizes the SSE and give us the line with the smallest prediction errors. For bivariate linear regression, we can calculate this line using matrix algebra (outside the scope of this course); we call this the "ordinary least squares" method.

Now that we know the total amount of prediction error (SSE), we also have a basic measure of goodness of fit for the regression line. However, SSE is not readily interpretable because it lacks a meaningful scale. To assess the goodness of fit relative to a baseline, we compare the SSE of the regression line to the sum of squares we would obtain if we did not use the predictor variable - that is, if we just predicted the mean value for each individual. A model with only the mean and no predictor variables is called the null model. The sum of squared distances between the mean and individual observations is referred to as the Total Sum of Squares (TSS), which represents the average squared distance of individual observations from the mean of Y.

To understand how much of the TSS is explained by the regression line, we calculate the Regression Sum of Squares (RSS). This is the difference between the TSS and the SSE: the reduction in TSS achieved by using the regression line to predict observations instead of just the mean. It indicates how well the regression line explains the variance in the dependent variable.

We can standardize this RSS by dividing it by the SSE, which gives us the "explained variance" R^2, which ranges from 0 to 1. A higher R^2 value indicates that a larger portion of the total variance in the dependent variable is accounted for by the predictor variable. Explained variance is the proportion of the total sum of squares (TSS) that is explained by the regression line (RSS).

Understanding these sums of squares gives us a good foundation for understanding another statistic: the correlation coefficient r. The correlation coefficient describes the strength and direction of the linear relationship between two variables. It differs from regression because regression describes one of these variables as an outcome of the other: an asymmetrical relationship. Correlation instead just describes how strongly these two variables are associated without labeling one as the predictor and the other as the outcome: a symmetrical relationship. The correlation coefficient is a standardized measure of the strength of this association that ranges from -1 (perfectly negatively associated) to 1 (perfectly positively associated). A correlation coefficient of 0 means that there is no association between X and Y. Correlation and regression are very closely related, as the squared correlation coefficient (r, squared) is the same as the measure of explained variance from simple linear regression, R^2, and is also the same as the standardized regression coefficient.

11.1 Lecture

https://www.youtube.com/embed/okrRJXb0YT4

11.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What does SSE stand for in linear regression? [1]

- (A) Sum of Squared Errors
- (B) Sum of Squared Explained Variance
- (C) Sum of Squares for the Expectation
- (D) Sum of Standard Errors

[1] Sum of Squared Errors

Question 2

Which of the following describes the concept of Ordinary Least Squares? [2]

- (A) Minimizing the total sum of squares
- (B) Minimizing the sum of squared errors
- (C) Minimizing the regression sum of squares
- (D) Minimizing the correlation

Question 3

What does the Regression Sum of Squares (RSS) represent? [3]

- (A) The total sum of squares minus the predicted values
- (B) The reduction in sum of squares by using the regression line as a predictor, rather than the mean
- (C) The sum of squared prediction errors
- (D) The explained variance by the regression line

Question 4

What is the purpose of the Total Sum of Squares (TSS) in linear regression? [4]

- (A) It measures the standard deviation of the errors.
- (B) It measures the residual error in the model.
- (C) It measures the explained variance of the predictor.
- (D) It measures the total variability of the dependent variable.

Question 5

What does the term 'explained variance' refer to in linear regression? [5]

- (A) The portion of the regression sum of squares explained by the predictor.

[2] Minimizing the sum of squared errors
[3] The reduction in sum of squares by using the regression line as a predictor, rather than the mean
[4] It measures the total variability of the dependent variable.
[5] The portion of the total variance explained by the regression line.

- (B) The portion of the total variance explained by the regression line.
- (C) The portion of the standard deviation explained by the predictor.
- (D) The portion of the total variance explained by the null model.

Question 6

How is correlation different from regression? [6]

- (A) Correlation focuses on categorical variables, while regression focuses on continuous variables.
- (B) Correlation measures the total variance, while regression measures the explained variance.
- (C) Correlation predicts one variable from another, while regression measures the strength of linear association.
- (D) Correlation measures the strength of linear association, while regression predicts one variable from another.

Question 7

What is the range of the correlation coefficient (r) between two variables? [7]

- (A) [-1, 0]
- (B) [-1, 1]
- (C) [0, 1]
- (D) [-,]

Question 8

Which statistic is equivalent to the standardized regression coefficient in bivariate regression? [8]

- (A) Coefficient of determination (Rš)
- (B) Standard deviation of the predictor

[6] Correlation measures the strength of linear association, while regression predicts one variable from another.
[7] [-1, 1]
[8] Correlation coefficient (r)

- (C) Sum of Squared Errors (SSE)
- (D) Correlation coefficient (r)

Question 9

What is the relationship between the Total Sum of Squares (TSS) and the Regression Sum of Squares (RSS)? [9]

- (A) TSS - SSE = RSS
- (B) SSR- SSE = SST
- (C) TSS + SSE = SSR
- (D) TSS + RSS = SSE

Question 10

How is the proportion of explained variance (Rš) related to the correlation coefficient (r)? [10]

- (A) Rš = 1 / r
- (B) Rš = 2 * r
- (C) Rš = rš
- (D) Rš = 1 - rš

[9] TSS - SSE = RSS
[10] Rš = rš

Show explanations

Question 1
SSE stands for Sum of Squared Errors, which represents the sum of the squared differences between actual and predicted values.
Question 2
Ordinary Least Squares aims to minimize the total prediction error, which is achieved by finding the regression line.
Question 3
RSS is the reduction in sum of squares that occurs when using the regression line to predict observations instead of just the mean.
Question 4
TSS measures the total variability of the dependent variable around its mean.
Question 5
Explained variance refers to the proportion of the total variance in the dependent variable that is explained by the predictor.
Question 6
Correlation quantifies the strength and direction of the linear relationship between two variables, while regression aims to predict one variable from another using the regression line.
Question 7
The correlation coefficient (r) ranges from -1 to 1, where -1 represents a perfect negative association, 0 represents no association, and 1 represents a perfect positive association.
Question 8
In bivariate linear regression, the standardized regression coefficient is equivalent to the correlation coefficient (r) between the predictor and the outcome.
Question 9
The relationship between TSS and RSS is given by TSS - RSS = SSE, indicating that the difference between TSS and RSS accounts for the error sum of squares.
Question 10
The proportion of explained variance (Rš) is equal to the square of the correlation coefficient (r), meaning Rš = rš.

11.3 In SPSS

11.3.1 Correlation Analysis

https://www.youtube.com/watch?v=VOI5IlHfZVE

11.4 Tutorial

11.4.1 Bivariate Regression

Social science students were asked about their opinion towards Tilburg's nightlife, number of Facebook friends, and some other characteristics. The data are in the SocScSurvey.sav file.

Download the file to your computer and open it in SPSS.

Suppose researchers are interested in the relationship between personality and social media use. In particular, they want to know if extraversion explains the number of Facebook friends.

What is the independent variable here? [11]

- (A) Extraversion
- (B) Facebook friends

Remember that the independent variable is the variable that predicts the other variable (which we call the dependent variable). The dependent variable is influenced by the independent variable (its value depends on the independent variable).

Run a regression analysis in which you regress Facebook friends on extraversion (via analyze > regression > linear).

Keep in mind that we "regress the dependent variable Y on the independent variables (X)".

Consult the output.

Write down the estimated unstandardized regression equation.

Answer

$\text{Friends}_i = 62.377 + 26.788\ \text{Extraversion}_i + e_i$

Which of the following statements is true?

[12]

- (A) If Extraversion increases with 26.788 units, the number of facebook friends increases with 1 unit
- (B) If Extraversion increases with one unit, the number of facebook friends increases with 26.788 units

[11] Extraversion
[12] If Extraversion increases with one unit, the number of facebook friends increases with 26.788 units

- (C) If Facebook friends increases with one unit, extraversion increases with 26.788 units

Remember that the general form of interpretation of the unstandardized effect is: "If X increases with 1 unit, Y increases/decreases with 'unstandardized regression coefficient' units".

What is the value of the standardized regression coefficient? _____[13]

You can find the standardized regression coefficients in the column called 'Standardized Coefficients Beta'.

Which of the following statements about the standardized regression coefficients is correct? [14]

- (A) If extraversion increases with one SD, the number of facebook friends increases with 0.438 SDs
- (B) If extraversion increases with one SD, the number of facebook friends increases with 0.438 units
- (C) If extraversion increases with one unit, the number of facebook friends increases with 0.438 SDs

Remember that standardized regression coefficients are interpreted in a similar way as unstandardized regression coefficients are, with the one difference being they are interpreted in terms of standard deviations.

Consider the first person in the data file. The person had an extraversion score of 9.

What is the predicted number of Facebook friends for this person? _____[15]

Consider the first person again.

Given the predicted number of Facebook friends for this person, what is the prediction error (rounded to the nearest integer)? _____[16]

Prediction error = yobserved - ypredicted

Consider two people, one with an extraversion score of 10 and the other with an extraversion score of 15.

What is the difference in the predicted number of Facebook friends between the two persons? (report the absolute value) _____[17]

Consult the output of the regression analysis.

---

[13] 0.438
[14] If extraversion increases with one SD, the number of facebook friends increases with 0.438 SDs
[15] 179
[16] -159
[17] 134

What percentage of the total variance in number of Facebook friends can be explained by extraversion? _____[18]

Consult the ANOVA table.

The table shows the results of an F-test.

What is the default null hypothesis and alternative hypothesis for the reported test?

Answer

$H_0 \ ^2 = 0$, $H_A \ ^2 > 0$

Suppose three researchers test the significance of the R-square.

Researcher I tests at the 10% level, researcher II tests at the 5% level, and researcher III at the 1% level.

Which researcher will reject the null hypothesis? [19]

- (A) All three researchers
- (B) Only researcher III
- (C) Only researcher I
- (D) Only researcher II

When reporting the F-test for the model, you would report R^2, the F-test statistic, its degrees of freedom, and the p-value.

The F-test has two distinct degrees of freedom. The first refers to the degrees of freedom for the regression equation, and the second to the degrees of freedom for the residuals. The degrees of freedom are given in brackets. For example, if regression has 2 degrees of freedom and the residuals 100, we write the F-value as F(2,100) = ...

Which of the following F value and corresponding degrees of freedom should be reported? [20]

- (A) F(1,133) = 31.283
- (B) F(1,132) = 0.000
- (C) F(1,132) = 31.283

[18] 19.2
[19] All three researchers
[20] F(1,132) = 31.283

11.4.2 Correlation

Correlations and regression analyses can both be used to study the relationship between variables, but there is an important difference.

Discuss with your group mates what the similarities and differences between the two methods are.

Answer

A correlation is a symmetric measure of association, meaning we are agnostic about which is the predictor and which is the outcome (or neither are predictor/outcome). The correlation between X and Y is the same as the one between Y and X.
In regression analysis, we do define an independent and dependent variable. The goal is to predict the outcome using the predictor. Most of the time, this implies an assumption of causality - but not necessarily.
For example, we can use regression to predict sales based on customer characteristics without assuming that those characteristics CAUSE sales. But if we want to cause an increase in sales, and we look at the regression coefficients to decide where to intervene - then it suddenly matters a lot whether the predictors are causes of sales or not.
You see this a lot with online marketing when you are receiving a lot of adds for a product that you recently bought. Their regression model knows that looking at the product page is a great predictor of intention to buy it - but they don't know that the reason you were looking at that page is because you were already buying it.

Now, let's have a look at the correlation between these two variables.

Analyze > correlate > bivariate.

Choose as variables: Facebook Friends and Extraversion, and click OK.

What is the correlation between Extraversion and number of Facebook friends? _____[21]

Suppose three researchers test the significance of the correlation between Extraversion and Facebook friends. Researcher I tests at the 10% level, researcher II tests at the 5% level, and researcher III at the 1% level.

Which researcher will reject the null hypothesis? [22]

- (A) Only researcher II
- (B) All three researchers

[21] 0.438
[22] All three researchers

- (C) Only researcher I
- (D) Only researcher III

Which of the following interpretations is true?

[23]

- (A) We don't have convincing evidence that Facebook friends and extraversion are associated in the population.
- (B) We have convincing evidence that Facebook friends and extraversion are associated in the population.
- (C) It would be very unlikely to observe a sample correlation of .44 by chance if the population correlation would be zero.

Compare the correlation coefficient to the standardized regression coefficient from the bivariate regression you conducted previously.

Then, compare it to the value labeled "R" in the "Model Summary" table from the regression.

Square the correlation, and compare it to the value labeled "R Square".

What do you observe?

Answer

If you did everything correctly, you should observe that the bivariate correlation is identical to the standardized regression coefficient. This is only the case with *bivariate* regression.
Furthermore, the bivariate correlation should be identical to the R reported in the Model Summary table, because they are both just the correlation coefficient. R squared is the squared correlation coefficient, and we interpret it as the "proportion of variance in the outcome explained by the predictor". Only in bivariate regression is this identical to the squared correlation coefficient.

11.4.3 R squared

The R squared expresses how well the predictors explain variance in the outcome of a regression. In the next few steps we will look in more detail at this concept.

Consider the results of the regression model again.

[23] It would be very unlikely to observe a sample correlation of .44 by chance if the population correlation would be zero.

Write down the (unstandardized) regression equation based on your previous results, and use the raw data in the Data View to answer the following question.

Answer

$\text{pressure}_i = 37.863\ \ .320\ \ \text{variety}_i + e_i$

What is the predicted value (Y') for emotional pressure at work for the first person in the data file (i.e., the person with respondent number 1)? _____[24]

What is the prediction error (a.k.a. the residual) for the first person? _____[25]

Remember Residual = Yobserved - Ypredicted

We've just computed the predicted value and error by hand. It would be very tedious if we would have to do that for all respondents. Fortunately, SPSS offers the option to compute predicted values and errors for all cases for us!

Navigate to Analyze > Regression > Linear

Click on the 'Save' button. SPSS opens a new window.

Ask for the Unstandardized predicted values and the unstandardized residuals. Paste and run the syntax.

Let's inspect the Data View in SPSS again and verify that SPSS added two columns in the data file. One column is labeled PRE_1 and the other RES_1. These columns show the predicted values and residuals for each person, respectively.

You may verify this for the first person (i.e., the values should be the same as you computed in the previous steps).

Now we will look at the variance of the observed values of Pleasure at work, the variance of the predicted values of pleasure at work, and variance of the residuals.

Compute the variances of Emotional pressure, as well as for the predicted values of Emotional pressure, and for the residuals.

Navigate to Analyze > Descriptive statistics > Descriptives Select scemoti, PRE_1, and RES_1. Click on 'options' and ask for the Variance. Paste and run the syntax.

How large is the variance of the observed scores for Emotional pressure? _____[26]

How large is the variance of predicted values of Work pleasure? _____[27]

How large is the variance of the residuals? _____[28]

---

[24] 29.329
[25] 8.771
[26] 145.168
[27] 33.361
[28] 111.807

In the previous questions, we looked at three variance components.

Discuss with your group what the three variances represent.

Answer

The variance in observed values of Emotional pressure is the total variance in Y (i.e., the dependent variable). The variance in the predicted values of Emotional pressure reflects "differences in emotional pressure that can be explained because some persons have a job with a lot of variety and some have a monotonous job". This variance component is also known as the explained variance. The variance of the residuals, also known as the residual variance, represents differences in emotional pressure that cannot be attributed to differences in variety at work. Hence, the residual describes differences that are unrelated to variety at work.

In the previous step, we looked at the variances itself, but the numbers are not very informative. A more convenient way to look at the explained variance is proportion wise.

So, let's use the variances we just generated to calculate the proportion of variance in Emotional pressure that can be explained by Variety at work.

What percentage of the total variance in emotional pressure can be explained by variety at work? _____[29]

Consult the output of the regression analysis again, particularly the table Model Summary.

Verify that the R-square that is reported in the table is the same as the proportion of explained variance that you have calculated yourself.

Finally, independently go through all the steps of a simple regression analysis using the data file `Work.sav`.

Your theory suggests that independence at work predicts emotional pressure.

- Construct an appropriate research question and hypotheses.
- Conduct the analysis
- Describe the relationship (i.e., regression coefficient)?
- Discuss the effect size in terms of R2.
- Perform a significance test and report your results

Finally, compare the standardized regression coefficient to the R coefficient in the Model Summary table, and optionally to a correlation computed via the Correlation interface. Verify that these are all identical.

[29] 22.981

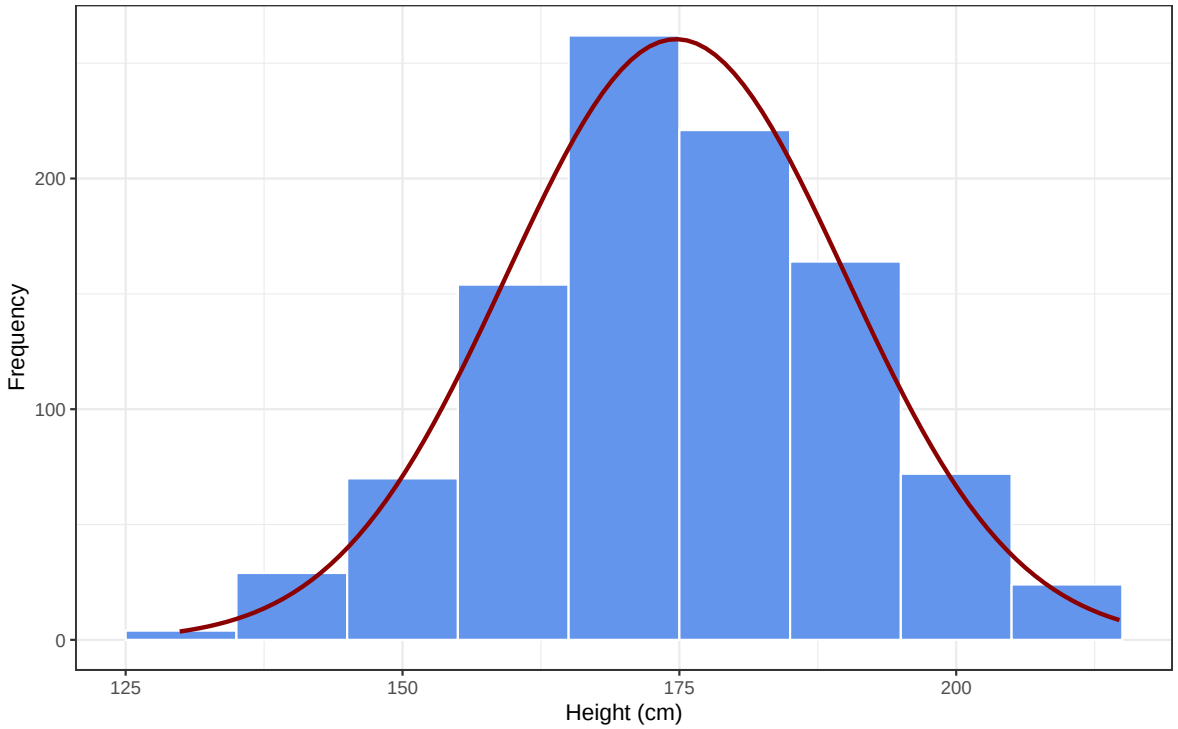
12 Assumptions

Every time we use a statistical model to describe data, we make certain simplifying assumptions. If these assumptions are met, the model is a good representation of the data (descriptive statistics), and we can make valid inferences about the population based on the model's parameters (inferential statistics). However, when these assumptions are violated, the model is a bad descriptor of the data, and inferences based on the model can be misleading or difficult to interpret.

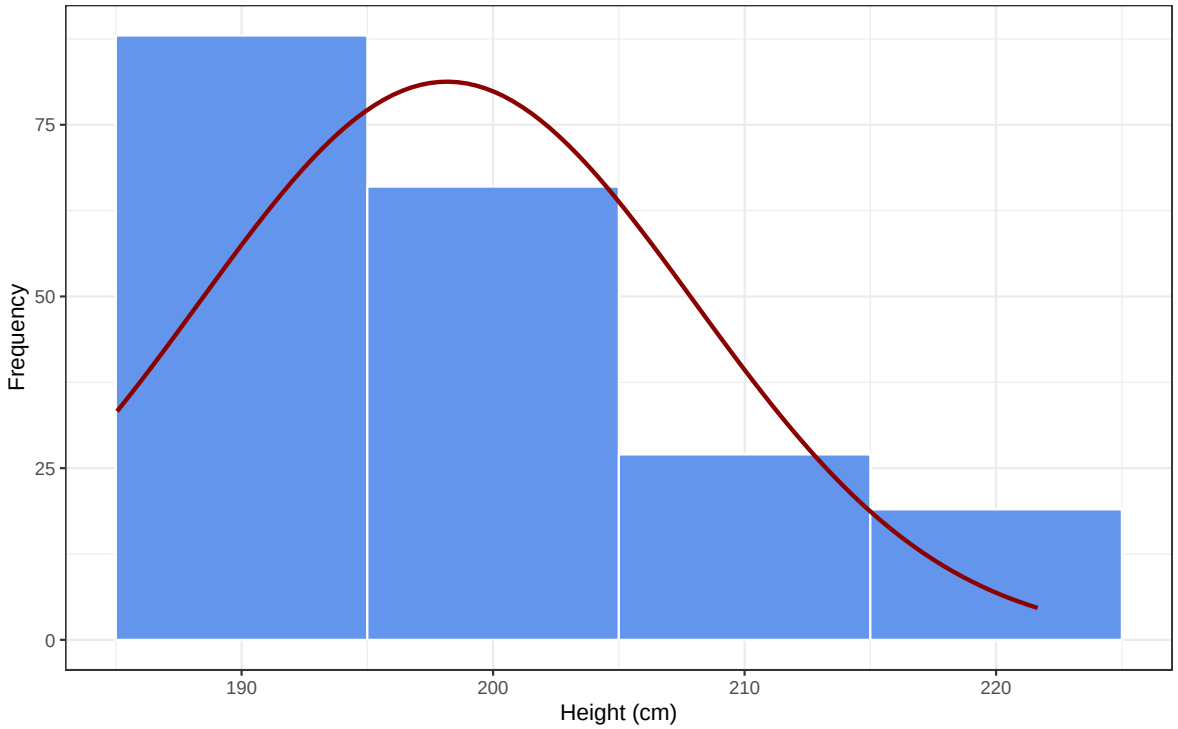
To de-mystify assumptions, let's examine one of the simplest statistical models possible: the normal distribution. The normal distribution is a statistical model to describe the distribution of scores on a variable (or: in the population), and its two parameters are the mean and standard deviation. If I draw a random sample of 1000 participants from the population of the Netherlands, their observed heights might be distributed as in the histogram below. I could use the normal distribution as a model for these data, and it would do a pretty good job (see the red normal distribution). In this case, my assumption is that *height is normally distributed around a mean and with standard deviation* , or:

$$\text{Height} \quad N(,)$$

If this assumption holds, the mean and standard deviations will be pretty good descriptive statistics of the distribution of data in the sample. If I assume that height is also normally distributed in the population, and that my sample is representative - then my sample statistics are also pretty good estimators for the population parameters.



Now imagine that I draw a convenience sample of 200 members of my local basketball association (figure below). Do you think I can assume that their heights will be normally distributed? Why (not)? Do you think these individuals will be representative of the Dutch population? Will they be representative of the population of Dutch basketball players? If the assumption that these scores are normally distributed is violated, then the mean and standard deviation of the normal distribution will be poor descriptive statistics. Moreover, these sample statistics will be poor estimators of the population parameters.



12.0.1 Assumptions for Linear Regression

The same principles apply to more complex models than the normal distribution - for example, linear regression. In fact, linear regression can be written *as* a normal distribution whose mean depends on the value of a predictor variable. If this equation says that height is normally distributed:

$$\text{Height} \quad N(,)$$

Then this equation says that height is normally distributed with a mean value that depends on age:

$$\text{Height}_i \quad N(+ \quad \text{Age}_i,)$$

Notice that the overall (population) mean of height is now replaced with a linear formula with population intercept and effect of Age . Another way to rewrite this formula without changing the meaning is:

$$\text{Height}_i = + \ \text{Age}_i + _i \tag{12.1}$$
$$_i \ N(0,) \tag{12.2}$$

This is the familiar notation of a regression equation. There are two points here: one, regression can be written as an extension of the normal distribution by plugging a linear formula in the place of the distribution mean. This means that regression inherits the assumptions of the normal distribution (e.g., no outliers), and gains a few more because of the added linear formula. Two, all of the assumptions are right there, in the formula itself: the fact that we specify height as a *linear function* of age means that we assume linearity. The fact that we use a little subscript $_i$ for height and age means that we assume independent observations from different individuals for these variables. The fact that we have one normal distribution for the error term $_i$ means that we do not expect the error distribution to vary at different values of the predictor, in other words, we expect homoscedasticity.

Below, we get deeper into the assumptions of linear regression, explains why each one matters, and shows how to check whether they are likely to hold in your data.

12.0.1.1 Independence of Observations

Linear regression assumes that every observation, or every row in the dataset, represents an independent observation, contributing unique information to the dataset. This means that observations should not be systematically related to each other. For example, participants should not be partners, friends, classmates, et cetera - any reason why participants might be more similar than randomly selected members of the population could introduce a violation of the assumption of independence.

Why it matters

When observations are *clustered* - for example, when data come from students in the same classroom, patients treated by the same clinician, or repeated measurements from the same individual - the assumption of independence is violated. In such cases, the residuals of these observations are correlated, which causes the model to underestimate the true variability, leading to overconfident conclusions.

How to check

- Consider the study design: Were the data collected from naturally grouped or repeated units, such as individuals within teams, families, schools, or measured over time?

When it is violated

This assumption is likely to be violated when:

- Observations are nested within a shared context (e.g., students within schools).
- The same individual or unit appears multiple times in the dataset.
- There is a known time-based or spatial structure to the data.

12.0.1.2 Correct Measurement Levels

Linear regression requires the outcome variable (Y) to have a continuous measurement level. This also follows from the linearity assumption: if we assume that an increase on X from 1 to 2 will have the same effect (regression slope) as an increase on X from 4 to 5, then that also means that Y must have a measurement level where the same distance has the same numerical value (interval or ratio). Examples of appropriate variables are test scores, height, or income. Predictors (X variables) can have any measurement level, but they must be *encoded* as continuous or binary (0 and 1-coded dummy variables). Some statistical software encodes nominal and ordinal variables as binary indicators behind the scenes, effectively doing this work for you.

Why it matters

If Y is nominal, it does not make sense to predict it numerically. If Y is ordinal, we cannot be sure that steps of equal numerical size have the same meaning. Sometimes a linear model works quite well for ordinal scales, but it is always important to check for indications of model violations when you use it for such variables.

How to check

First, review the codebook, metadata, or variable definitions in SPSS to confirm that:

- The outcome (Y) is coded as a continuous numeric variable.
- Categorical predictors are either dummy-coded or otherwise appropriately handled.

When it is violated

The assumption is violated when:

- Y is nominal (categories like “red”, “green”, “blue”) or ordinal (e.g., Likert scales).
- An ordinal X is treated as numeric without justification, leading the model to assume equal spacing between categories.

Note that the operationalization of a variable is not the only factor that matters; its true measurement level also matters. For example, if you want to measure height, you could put a mark on the doorpost and rate everyone taller than the mark as “tall”, and anyone shorter than the mark as “short” - but the fact that you operationalized height this way doesn’t negate the fact that it is inherently a continuous variable.

More pertinently: gender is often operationalized as binary. This does not mean that gender *is* binary; its true measurement level is more complex. If you are interested in gender as a social

construct, then there are more then two discrete categories. If you are interested in gender for its biological aspects, then there is both nominal variability. Nominal variability includes aneuploidy of sex chromosomes, and continuous variability occurs in various biological aspects of male-ness and female-ness, like hormone balances and -sensitivities.

12.0.1.3 Linearity – the "straight-line" assumption

Linear regression assumes that the relationship between each predictor X and the outcome Y is linear, that is, that the same change in X corresponds to the same change in Y for all values of X. The slope , or its sample estimate b, tells us how much Y is expected to increase (or decrease) for a one-unit increase in X, but this only holds if the relationship is approximately linear.

Why it matters

If the relationship is actually non-linear, fitting a straight line will misrepresent the nature or strength of the association. Remember Anscombe's quartet in **?@sec-anscombe**. A straight line fitted to a pattern in data that, in reality/in the population, is non-linear (quadratic, S-shaped, etc) will result in inaccurate or meaningless slope estimates and misleading conclusions about the predictor's effect.

How to check

1. Create a scatterplot of Y against each X variable.

2. Add a straight trend line (e.g., "fit line at total" in your software).
3. Visually assess whether the line aligns with the overall pattern of the data points.

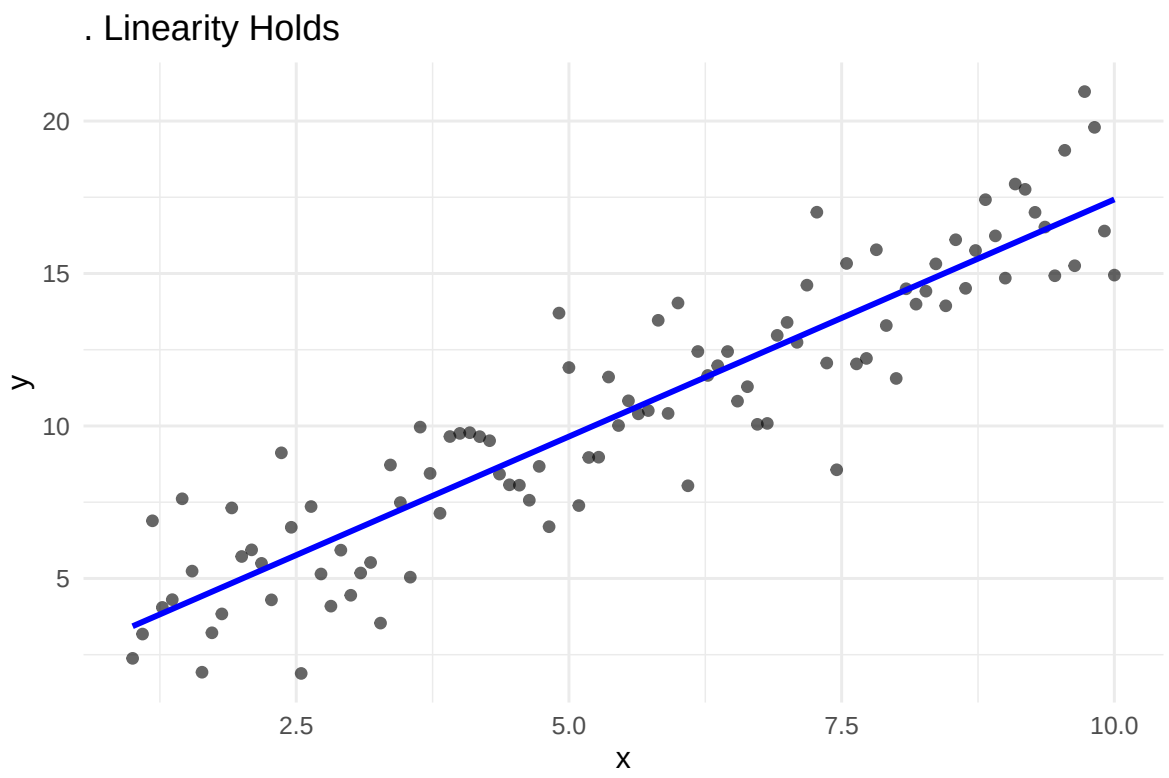
When it is violated

Linearity is likely violated if the plotted points form a clear curve, wave, or other systematic pattern that deviates from a straight path.

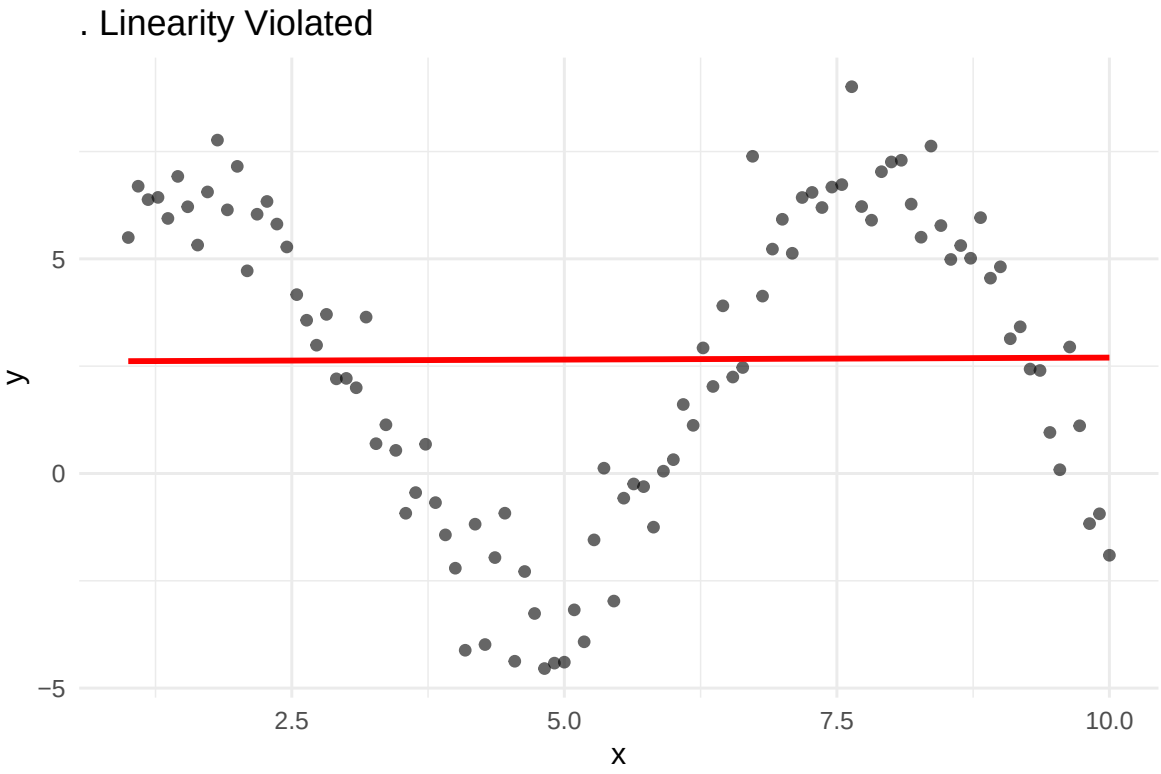
Alternatively, outside the scope of this course:

1. Estimate a linear model
2. Estimate a model with a different functional form (e.g., quadratic)
3. Compare the fit of the models using the BIC model fit index (lower is better)

Visual example of when linearity holds



Visual example of when linearity is violated



12.0.1.4 Normality of Residuals

As evident from the $_i \ N(0,)$ part of the regression equation, linear regression assumes that the residuals, the differences between the observed and predicted values, are normally distributed.

Why it matters

When the residuals deviate strongly from normality (e.g., they are skewed or have heavy tails), inferences based on the regression model may be misleading. Standard errors and metrics derived from them, like p-values, and confidence intervals, depend on this assumption.

How to check

- Save the residuals and inspect their distribution using a **histogram** or a **Q–Q plot**.
- For very small datasets (e.g., $N < 50$), formal tests such as the **Shapiro–Wilk test** can be used to assess normality.

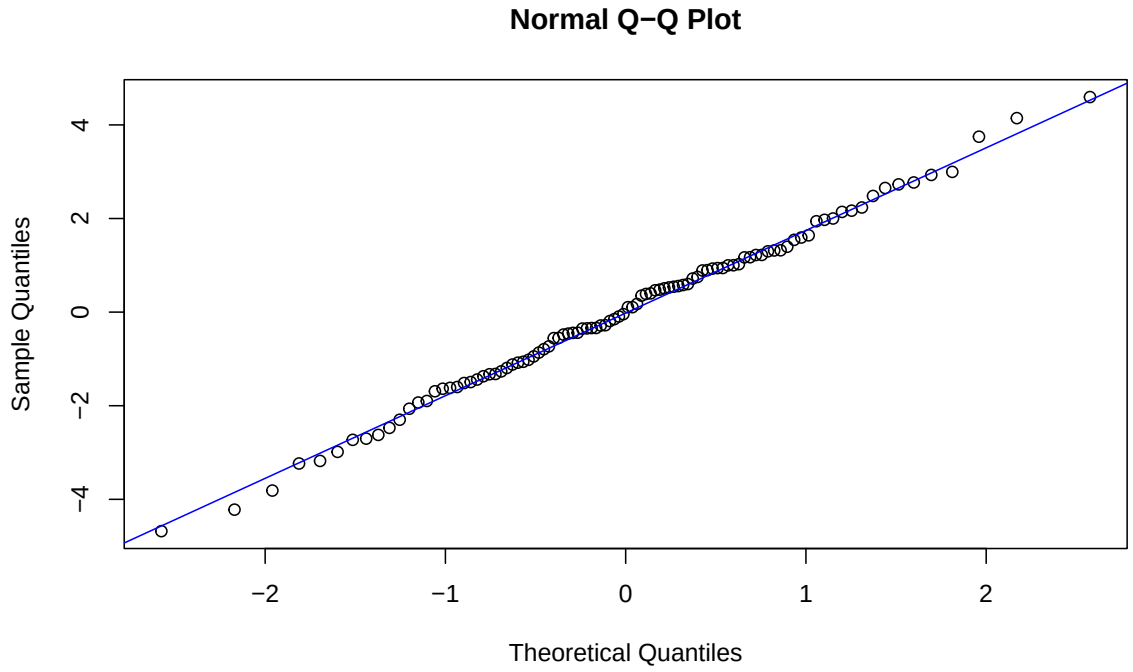
When it is violated

This assumption is often violated when:

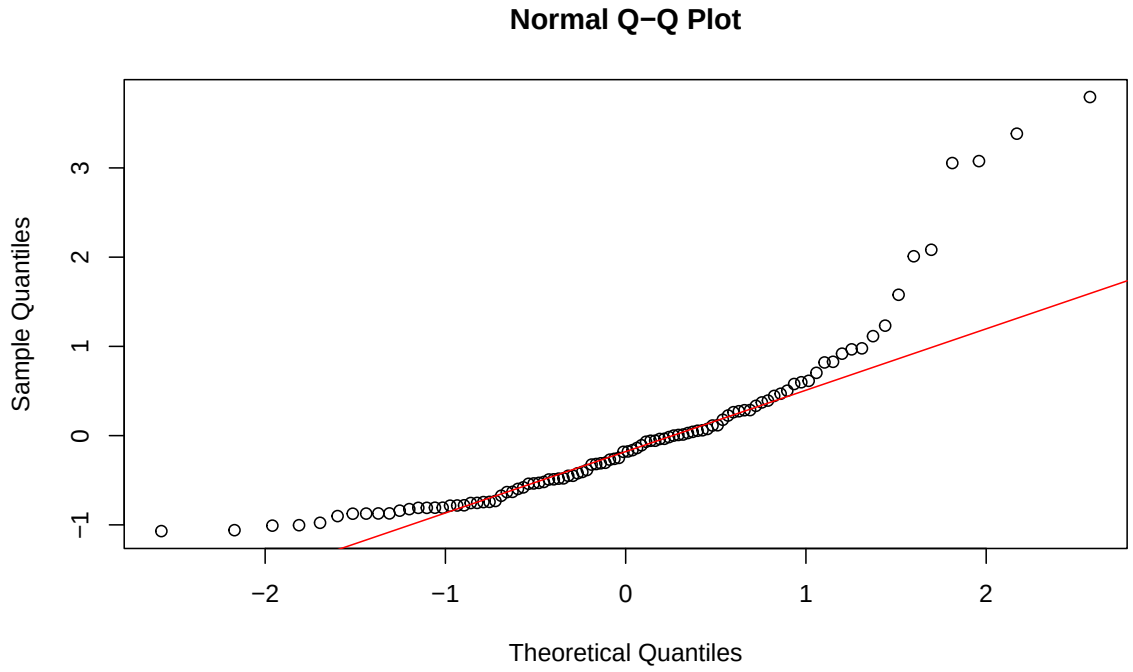
- The outcome variable is highly skewed or bounded
- There are extreme values in the outcome that disproportionately influence the residuals.

- The sample size is small and the residual pattern does not resemble a bell-shaped curve.

Visual example of residuals following a normal distribution



Visual example of residuals deviating from a normal distribution



12.0.1.5 Homoscedasticity

The fact that linear regression has a single error term implies an assumption that the variance of the residuals remains constant across all levels of the predicted values. This condition, known as *homoscedasticity* (homo = equal, scedasticity = variance), implies that the model has equal predictive accuracy across the full range of the outcome.

Why it matters

When residuals fan out, contract, or otherwise vary as the predicted values increase, this is called *heteroscedasticity* (hetero = different, scedasticity = variability). In such cases, the standard errors may be inaccurate, which undermines the reliability of p-values and confidence intervals.

How to check

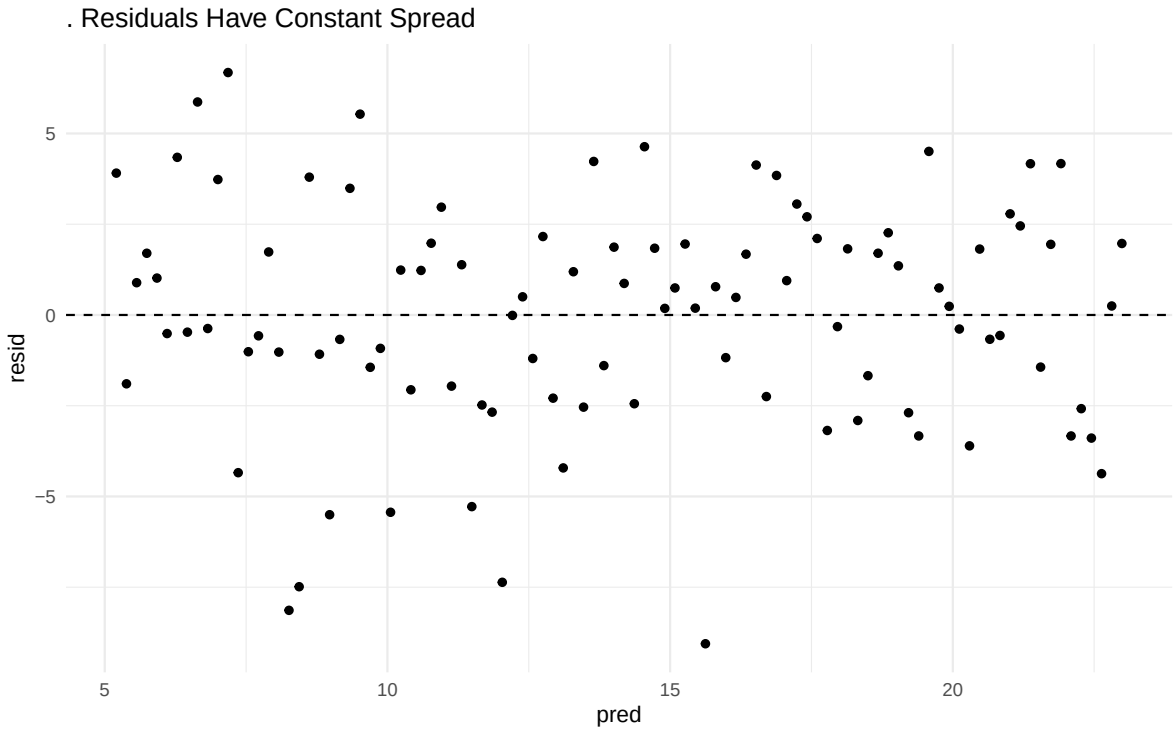
1. Save the residuals from the fitted model.

2. Create a scatter plot of residuals (Y-axis) and predicted values (X-axis).
3. Visually determine whether the spread of residuals appears approximately constant across the range of predicted values.

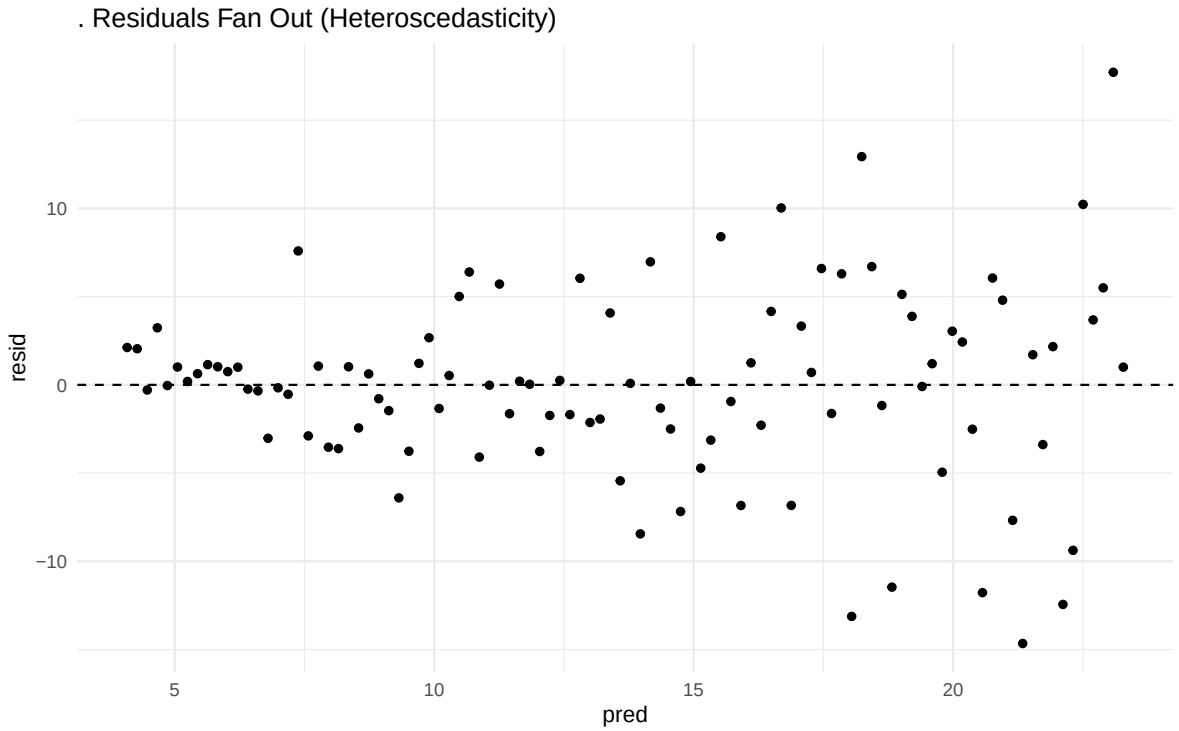
When it is violated

This assumption is violated when: - The residuals become more dispersed or more concentrated as the predicted value increases. - The residual-versus-predicted plot reveals a funnel-like or cone-shaped pattern rather than a uniform band. - The residuals shows any pattern, other than a random dot cloud.

Visual example of homoscedasticity



Visual example of violation of homoscedasticity



12.0.1.6 No Outliers

The assumption of no outliers is related to the assumption of linearity and the assumptions of normal, homoscedastic residuals. An extreme case can distort slope estimates (as in Anscombe's quartet, figure c) and standard errors, and consequently, confidence intervals and p-values which are based on those standard errors.

Why it matters

Outliers can pull the regression line toward themselves, leading to misleading interpretations. Even a single influential point can change the direction, strength, or significance of a predictor's effect.

How to check

- Calculate diagnostic statistics such as **leverage**, **Cook's distance**, and **standardized residuals** to detect potential outliers.

When it is violated

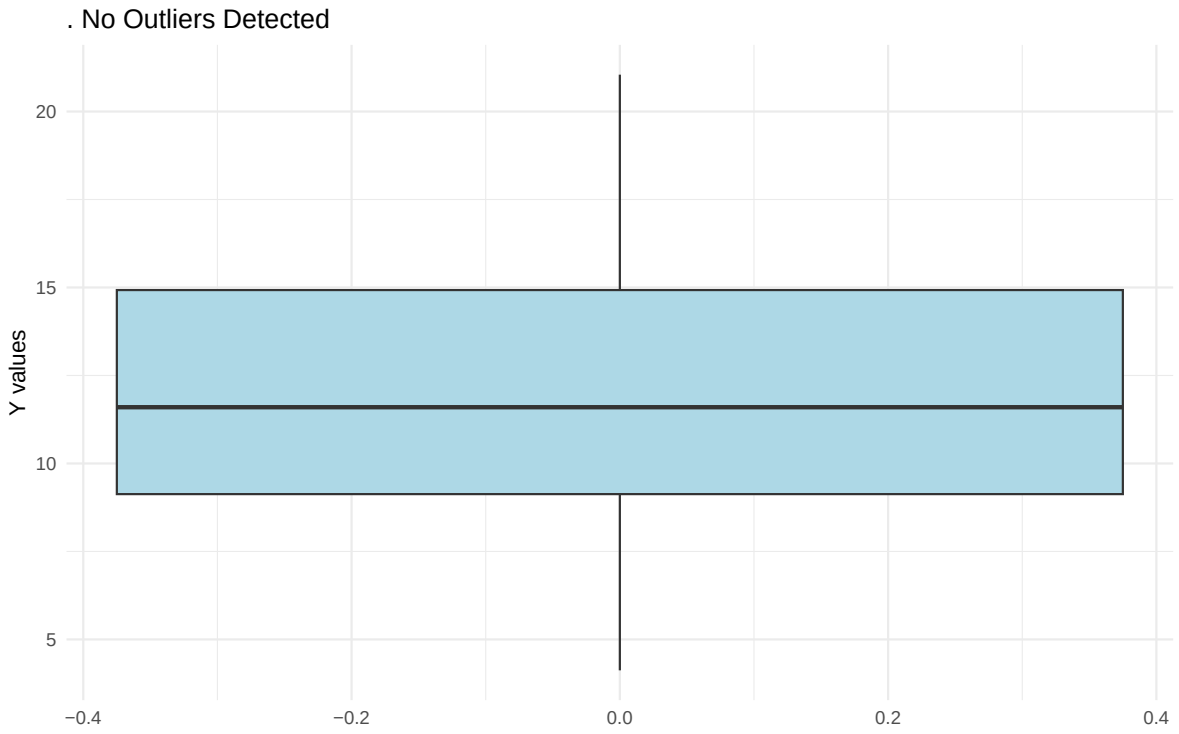
This assumption may be violated when:

a. A case lies far from the bulk of the data on one or more predictors.

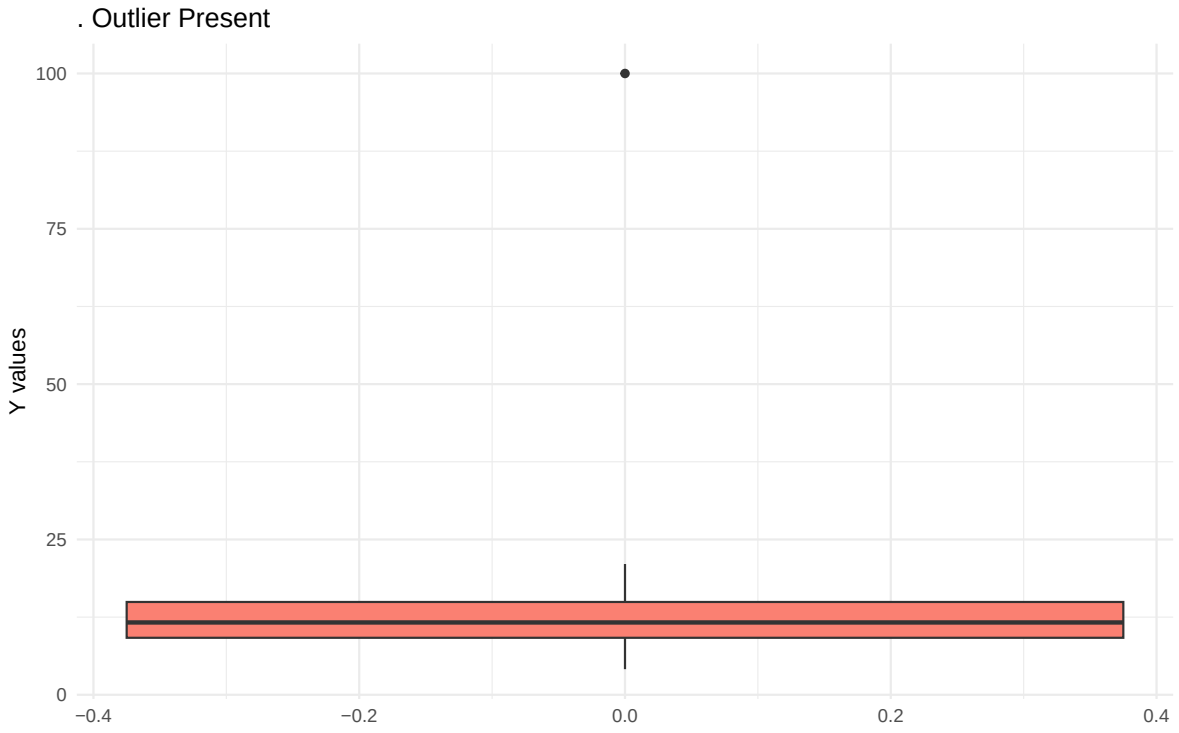
b. The residual for a single observation is large relative to others.

c. Diagnostic measures flag a case as both high-leverage and high-influence.

Visual example of no influential outliers present



Visual example of presence of an outlier



12.0.1.7 Reliable Predictors

In the regression equation, the outcome has an error term, $_i$. If the model makes imperfect predictions for any reason, these prediction errors contribute to the error term. One reason for prediction error is measurement error in the outcome. Note, however, that while the outcome has an error term - the predictor does not. This implies an assumption that predictor variables (X) are measured without error. Inaccurate or inconsistent measurement introduces noise, which can attenuate the estimated relationship between X and Y. As a result, regression coefficients may be biased toward zero, and the model may attribute true effects to random error.

Why it matters

When predictors are measured with error (unreliable), the estimated slopes become less trustworthy. Even in large samples, measurement error in X can severely compromise the interpretability of regression results, leading to underestimation of effect sizes and increased standard errors.

How to check

- For predictors based on multiple items (e.g., survey scales), compute internal consistency (e.g., Cronbach's alpha). Values below 0.70 often indicate problematic measurement.

- If repeated measurements of the same predictor are available, examine the test–retest correlation. High correlation supports reliability.

When it is violated

This assumption is violated when: - A predictor contains high random measurement error. - Multi-item scales exhibit low internal consistency. - Temporal stability of repeated measures is weak (e.g., inconsistent responses across time).

12.0.1.8 *No Multicollinearity*

As a preview of what is to come: it is possible to include more than one predictor in a regression model. This is called *multiple regression*, and it will be covered in Statistics 2. Multiple regression additionally assumes that each predictor contributes uniquely to the explanation of the outcome variable. When two or more predictors explain the *same* variance in the outcome, the model struggles to estimate their unique effects. This overlap in explained variance makes coefficient estimates unstable and difficult to interpret, and inflates the individual predictors' standard errors.

Why it matters

Multicollinearity undermines the precision of regression coefficients. When predictors convey redundant information, the model's ability to estimate each slope independently deteriorates. This can lead to wide confidence intervals, non-significant p-values, or coefficients with counterintuitive signs.

How to check

- Examine the **Variance Inflation Factor (VIF)** for each predictor. A common guideline is that values above 5 may indicate multicolinearity, values above 10 indicate severe multicolinearity.

When it is violated

This assumption may be violated when:

- Two predictors are strongly correlated (e.g., income and years of education).
- More than two predictors explain the same variance in the outcome (e.g., mother's income, father's income, and family Socio-Economic Status might explain the same variance in children's educational attainment)
- VIF values are unusually high.

- The inclusion of additional predictors drastically alters the estimated slopes or increases their standard errors.

12.0.1.9 Putting It All Together

Before interpreting regression results, it is essential to check for evidence of violations of assumptions. If the assumptions are violated, interpretation of the results might not be straightforward. Note, however, that assumption checks are subject to the same caveats as other statistical inferences:

- Assumptions are statements about the *population*; even if they appear to be violated in the *sample*, they might be met in the population
- Consequently, it is possible to draw false positive (incorrectly concluding that an assumption is violated, while in reality it is not) or false negative (incorrectly concluding that an assumption is not violated, when in reality it is) conclusions about assumptions
- Making data-driven analysis decisions incurs researcher degrees of freedom (see the chapter on questionable research practices). You run the risk of overfitting (customizing) your analysis so much to the sample at hand that it no longer generalizes well to the general population, or other samples from that population.

None of this diminishes the importance of checking assumptions, it is merely a call to exercise critical thinking when doing so. For example, regression assumes that the dependent variable is normally distributed. If your dependent variable is a Likert-type scale, sometimes, you may get away with making this assumption (figure **?@fig-figviolate** panel a below). Nevertheless, you should discuss this *potential* violation of the assumption of normality in the Discussion of your paper or report. However, if you notice after data collection that your dependent variable is distributed as in figure b below, the assumption of normality is so eggregiously violated that analysis results are probably meaningless. In this case, you might still analyze the results as planned - but that analysis will likely be meaningless. You might want to present a second analysis that treats the outcome as ordinal (which is possible, but outside the scope of this course), and emphasize that this was a data-driven analysis decision.

In sum:

- Always check assumptions
- Always report the result of assumption checks and discuss (in the Discussion) how potential violation might affect your conclusions
- In case of strong evidence of violations, you might report a secondary analysis that is robust to the violation of the assumption, but make it clear that this analysis was performed *after* seeing the data.
- Optionally, compare the results of both the planned analysis and the robust analysis. If the conclusions are the same (e.g., both analyses provide results consistent with your hypothesis), this is reassuring.

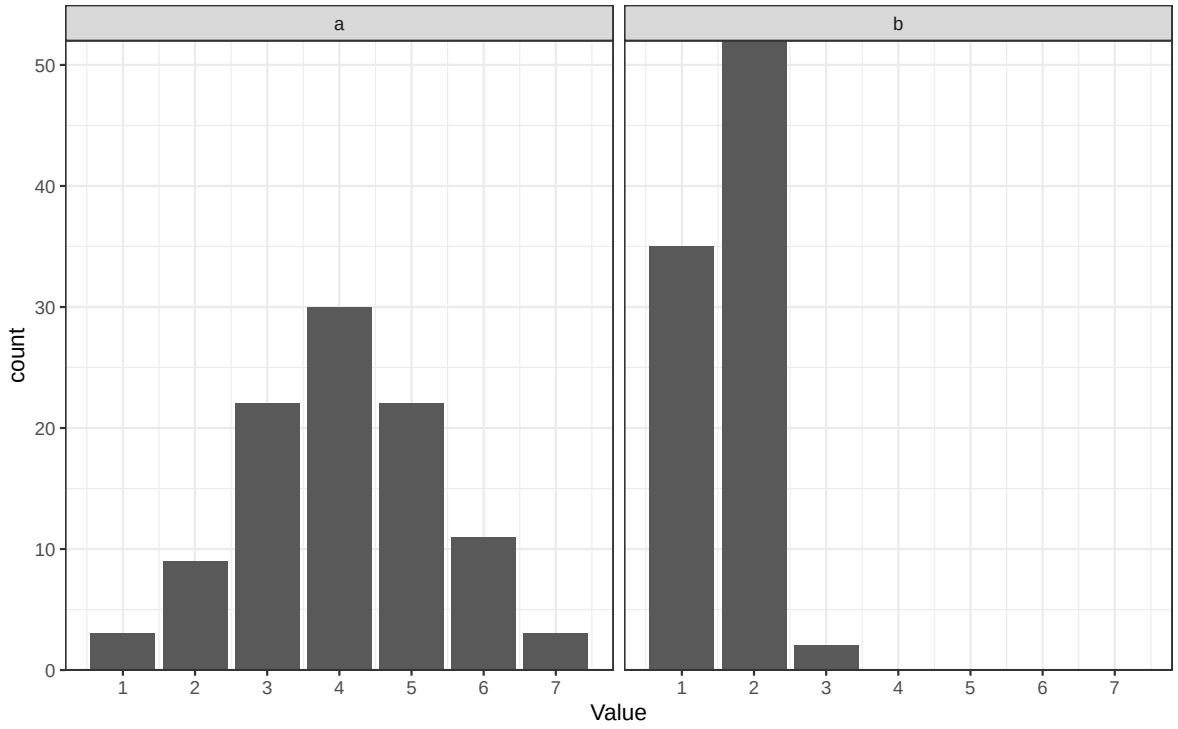


Figure 12.1: Figure a shows a Likert-type scale that is approximately normally distributed. Figure b shows a Likert scale with extreme censoring

1. **Visual diagnostics** – Begin with graphical checks for linearity, constant variance, outliers, and normality of residuals. Plots often reveal violations at a glance.

2. **Statistical diagnostics** – Follow up with numerical checks, such as the Variance Inflation Factor (VIF) for multicollinearity and formal tests for heteroskedasticity or autocorrelation when appropriate.

3. **Assessment of model structure** – Consider whether the data meet requirements for independence and correct scale of measurement.

12.1 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

Which statement best captures the linearity assumption in OLS regression? [1]

- (A) The residuals must have constant variance
- (B) The predictors must be normally distributed
- (C) The expected change in the outcome is proportional to changes in each predictor across its range
- (D) The outcome must be measured without error

Question 2

Homoscedasticity means: [2]

- (A) Residuals have constant variance across all levels of the predicted values
- (B) Predictors are uncorrelated with each other
- (C) Residuals are centered at zero
- (D) Predictors are measured on a continuous scale

Question 3

A residuals versus predicted plot shows a clear funnel shape. Which assumption is most likely violated? [3]

- (A) Independence of observations
- (B) Normality of predictors
- (C) Homoscedasticity
- (D) Linearity

Question 4

Why does normality of residuals matter most for small samples in OLS? [4]

[1] The expected change in the outcome is proportional to changes in each predictor across its range
[2] Residuals have constant variance across all levels of the predicted values
[3] Homoscedasticity
[4] It underpins the accuracy of t tests and confidence intervals for coefficients

- (A) It underpins the accuracy of t tests and confidence intervals for coefficients
- (B) It guarantees unbiased slope estimates
- (C) It eliminates the need to check other assumptions
- (D) It ensures predictors are measured reliably

Question 5

Your data consist of students nested within classrooms but you fit a single level OLS model that treats all rows as independent. Which assumption is threatened? [5]

- (A) Independence of observations
- (B) No outliers
- (C) Normality of predictors
- (D) Correct scale of measurement for the outcome

Question 6

Two predictors are highly correlated. What is the most common consequence for the regression coefficients? [6]

- (A) They become unbiased and more precise
- (B) They remain unchanged but model fit worsens
- (C) They become systematically too large
- (D) They become unstable with inflated standard errors and may change sign with small data changes

Question 7

Which outcome variable violates the correct scale of measurement assumption for standard OLS regression? [7]

- (A) Height in centimeters

[5] Independence of observations
[6] They become unstable with inflated standard errors and may change sign with small data changes
[7] A binary pass or fail indicator coded 0 and 1

- (B) A binary pass or fail indicator coded 0 and 1
- (C) Test score from 0 to 100
- (D) Annual income in dollars

Question 8

What is the typical effect of measurement error in a predictor on its estimated slope in OLS? [8]

- (A) Bias toward zero
- (B) No bias but larger p values
- (C) No effect if the outcome is normal
- (D) Bias away from zero

Question 9

You see a systematic curve in the residuals versus predicted plot. Which assumption is most suspect? [9]

- (A) Linearity
- (B) Independence
- (C) Homoscedasticity
- (D) No outliers

Question 10

Which statement about outliers and influence in OLS is correct? [10]

- (A) Only points with extreme outcome values can be influential
- (B) Any point far from the regression line is necessarily highly influential
- (C) A point can be influential if it has unusual predictor values and substantially changes the fitted line

[8] Bias toward zero
[9] Linearity
[10] A point can be influential if it has unusual predictor values and substantially changes the fitted line

- (D) Influential points affect p values but never change coefficient signs

Show explanations

Question 1
Linearity is about the average of Y at each X forming a straight line. The other options describe different assumptions.
Question 2
Homoscedasticity is constant spread of residuals across the range of fitted values.
Question 3
A funnel pattern indicates changing residual variance across fitted values.
Question 4
With small samples inference for slopes uses normality. In large samples asymptotics help.
Question 5
Clustering creates correlated errors. Rows are not independent.
Question 6
High correlation among predictors creates multicollinearity and unstable estimates.
Question 7
OLS assumes a continuous outcome. A binary outcome calls for a different model.
Question 8
Classical measurement error in X attenuates the slope.
Question 9
A patterned residual plot suggests a wrong functional form. The relation is not linear.
Question 10
Influence depends on leverage and residual. Such points can change slopes and even signs.

12.2 Tutorial

12.2.1 Before you start

- In each exercise, fit an OLS model with y as the outcome and the listed predictors.

12.2.2 Assignment 1

File: `reg_assump_check.sav`
Variables: y, x1, x2

Steps (SPSS GUI):

1) **Check linearity (for each predictor)**

   - Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter
     - Y axis: **y**; X axis: **x1**. Create plot.
     - Repeat for **x2**.
   - In the Chart Editor: Element > Fit Line at Total.

2) Analyze > Regression > Linear

   - Dependent: y
   - Independents: x1 x2
   - Statistics: Estimates, Confidence intervals, Collinearity diagnostics
   - Plots: ZPRED on X, ZRESID on Y
   - Save: Standardized residuals (ZRESID), Predicted values (ZPRED)

3) Graphs > Legacy Dialogs > Histogram

   - Variable: ZRESID; check "Display normal curve"

4) Graphs > Legacy Dialogs > Q-Q

   - Variable: ZRESID

Questions:

- Is there a linear association among these variables?
- Is the measurement scale of these variables appropriate for regression analysis?
- Are residuals normally distributed?
- Is homoscedasticity supported?
- Are there any multicollinearity concerns?

12.2.3 Assignment 2 — Linearity check

File: `reg_linearity_check.sav`
Variables: y, x1

Steps:

1) Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter

   - Y axis: y; X axis: x1
   - In Chart Editor, add a straight fit line (Fit Line at Total).

2) Analyze > Regression > Linear

   - y on x1; request the same plots and saves as in Assignment 1.

Questions:

- Does the y vs x1 scatterplot suggest a straight-line relation?
- Do residual plots show a pattern (e.g., systematic bends) inconsistent with linearity?
- Based on the diagnostics, is a linear specification for x1 adequate in this dataset?

12.2.4 Assignment 3 — Homoscedasticity

File: reg_homoscedasticity_check.sav
Variables: y, x1

Steps: 1) Fit OLS as in Assignment 1 and save ZRESID and ZPRED.
2) Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter
- X axis: ZPRED; Y axis: ZRESID.

Questions:

- Do residuals show roughly constant spread across predicted values, or a cone/funnel?

12.2.5 Assignment 4 — Normality of residuals

File: reg_normality_check.sav
Variables: y, x1

Steps:

1) Fit the regression and save residuals

   - Analyze > Regression > Linear
     - Dependent: y
     - Independent(s): your predictor(s)
     - Save > Standardized residuals
     - Plots > Tick Histogram

2) To produce Q-Q plot

- Analyze > Descriptive Statistics > Explore
- Add the saved standardized residuals to the dependent list
- Plots > Check Normality plots with tests

Questions:

- Is the residual distribution approximately symmetric and bell-shaped?
- Do Q-Q points track the diagonal?

12.2.6 Assignment 5 — Outliers

File: reg_outliers.sav
Variables: y, x1

Steps:

1) Analyze > Regression > Linear
   - Dependent: y
   - Independent: x1
   - Statistics: Estimates, Casewise diagnostics (e.g., standardized residuals > 3), Collinearity diagnostics (optional here)
   - Save: Cook's distance, and Leverage (Hat)
2) Graphs > Legacy Dialogs > Boxplot
   - Summaries for separate variables: **y**, **x1**
3) Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter
   - Y axis: **y**; X axis: **x1**

Questions:

- Are any cases flagged in Casewise diagnostics (e.g., |Std. Residual| > 3)?
- Do any observations show high leverage* or large Cook's distance relative to others?
- Based on these diagnostics, could a single case plausibly dominate the fitted line? Identify the case ID if so.

12.2.7 Exercise 6 — Multicollinearity

File: reg_multicollinearity.sav

Variables: y, x1, x2, x3

Steps: 1) Analyze > Correlate > Bivariate
- Inspect correlations among x1, x2, x3.
2) Analyze > Regression > Linear
- y on x1 x2 x3
- Statistics: Collinearity diagnostics, Estimates.

Questions:

- Are any predictor pairs highly correlated in the correlation matrix?
- What are the VIF and Tolerance values for each predictor?
- Do the signs and standardized Betas align with the simple correlations, or do you see suppression patterns?

What to look for:

- VIF substantially above 5 (or Tolerance, which is $\frac{1}{VIF}$, below .20) suggests collinearity.
- Large divergence between simple r and Beta can signal overlap among predictors.

12.3 Assignment 7 — Putting it together (choose two datasets)

Files: pick any two from the set

Task:

- For each dataset, run the standard diagnostic workflow from Assignment 1. - Summarize, in a short paragraph per dataset, which assumptions are reasonably met and which are doubtful, citing the specific plot or statistic you used.

Reminder:

- Focus on diagnosis only. Do not apply remedies or re-specify models here.

13 GLM-III: Binary Predictors

We can examine group differences in a continuous outcome variable using bivariate regression. To do this, group membership must be represented as a binary variable (e.g., gender or ethnicity). To ensure meaningful results, we use dummy coding to represent the binary variable. Dummy coding assigns the value 0 to one category, which serves as the reference category, and the value 1 to the other category. When we include this dummy variable as the predictor in a bivariate linear regression analysis, it will estimate the mean value of the reference category and test the difference between the means of the two categories.

Regression with a binary predictor is completely equivalent to the independent samples t-test. The independent samples t-test is also used to compare the means of two independent groups. In regression, we estimate the slope (b) for the binary predictor, which represents the difference between the means of the two groups. This t-test of the slope in regression is the same as an independent samples t-test.

Both regression with a binary predictor and the independent samples t-test rely on certain assumptions. These include the linearity of the relationship between the binary predictor and the outcome variable, the normality of residuals (the outcome variable should be normally distributed within each group), homoscedasticity (equal variances in both groups), and independence of observations. To check for equal variances, we can use Levene's test - but keep in mind that "assumption checks" are questionable. If you assume equal variances, report the normal t-test; if you do not assume equal variances, you can report a corrected t-test that allows for different variances. Both are included in SPSS output by default.

To determine the practical significance of a mean difference, we can calculate an effect size measure. Cohen's d is a commonly used effect size for mean differences. It standardizes the difference between the two group means by the pooled standard deviation. A larger Cohen's d indicates a greater magnitude of difference between the groups. As a rule of thumb, a small effect size is typically considered around d 0.2, a medium effect size around d 0.5, and a large effect size around d 0.8.

13.1 Lecture

https://www.youtube.com/embed/QeKr2R8Eyhk

13.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the purpose of dummy coding in regression with binary predictors? [1]

- (A) To increase the number of predictors in the model.
- (B) To convert binary variables into continuous variables.
- (C) To estimate the mean of the reference category and test the difference between categories.
- (D) To simplify the model by removing categorical predictors.

Question 2

What does the slope coefficient (b) represent in regression with a binary predictor? [2]

- (A) The difference in means between the two categories.
- (B) The intercept of the reference category.
- (C) The mean of the second category.
- (D) The ratio of the two categories' means.

Question 3

Which assumption is not relevant for the independent samples t-test and bivariate linear regression with only a binary predictor? [3]

- (A) Normality of residuals
- (B) Linearity of relationship between X and Y

[1] To estimate the mean of the reference category and test the difference between categories.
[2] The difference in means between the two categories.
[3] Linearity of relationship between X and Y

- (C) Homoscedasticity
- (D) Independence of observations

Question 4

What does the Levene's test check in the context of the independent samples t-test? [4]

- (A) Normality of residuals.
- (B) Equality of variances in both groups.
- (C) Linearity of relationship between X and Y.
- (D) Normality of variances in both groups.

Question 5

How is the independent samples t-test related to the t-test of the slope in regression with a binary predictor? [5]

- (A) The t-test of the slope is a subset of the independent samples t-test.
- (B) The independent samples t-test is a subset of the t-test of the slope.
- (C) They are equivalent.
- (D) They are not related.

Question 6

What does the p-value in the context of the independent samples t-test indicate? [6]

- (A) The probability of observing a statistically significant result.
- (B) The probability that the null hypothesis is true.
- (C) The probability of observing a group difference at least as extreme as the one observed, if the null hypothesis is true.
- (D) The probability that the null hypothesis is rejected.

[4] Equality of variances in both groups.
[5] They are equivalent.
[6] The probability of observing a group difference at least as extreme as the one observed, if the null hypothesis is true.

Question 7

What does Cohen's d represent? [7]

- (A) An effect size for the mean difference, expressed in number of standard deviations.
- (B) The difference between the two categories' means.
- (C) A measure of explained variance for mean differences.
- (D) An effect size for the difference between standardized group means.

Question 8

What is the recommended approach when assumption checks for homoscedasticity are significant? [8]

- (A) It depends - in confirmatory analyses, you may switch to a robust test; in exploratory analyses, you would report the violation and proceed as planned.
- (B) Exclude outliers to ensure homoscedasticity.
- (C) Use a robust t-test.
- (D) It depends - in exploratory analyses, you may switch to a robust test; in confirmatory analyses, you would report the violation and proceed as planned.

Question 9

The observed mean difference between two groups is 2.50, and Cohen's D is 1.25. What is the pooled standard deviation? [9]

- (A) 1.25
- (B) 2
- (C) 2.50
- (D) Can't say based on this information.

[7] An effect size for the mean difference, expressed in number of standard deviations.
[8] It depends - in exploratory analyses, you may switch to a robust test; in confirmatory analyses, you would report the violation and proceed as planned.
[9] 2

Question 10

A researcher accidentally coded a dummy variable as 0 and 2, instead of 0 and 1. The regression equation is Y = 5.66 + 3*D. What is the mean value of the group coded 2? [10]

- (A) 7.16
- (B) 8.66
- (C) 11.66
- (D) 3

[10] 11.66

Show explanations

Question 1
Dummy coding allows regression to include binary predictors by assigning numerical values to each category, estimating the mean of the reference category, and testing the difference between categories.
Question 2
The slope coefficient (b) in regression with a binary predictor represents the difference in means between the two categories, indicating how much the dependent variable changes when the binary predictor changes from 0 to 1.
Question 3
The assumption of linearity is not relevant, because the difference between two binary values of the predictor is linear by definition.
Question 4
Levene's test checks the assumption of equality of variances in both groups for the independent samples t-test.
Question 5
The independent samples t-test and the t-test of the slope in regression with a binary predictor are equivalent tests that compare means between two independent groups.
Question 6
The p-value indicates the probability of observing a group difference at least as extreme as the one observed, assuming that the null hypothesis is true.
Question 7
Cohen's d is an effect size that standardizes the difference between group means by the (pooled) standard deviation, making it interpretable on a meaningful scale.
Question 8
Assumption checks can alert you that important assumptions of the test are violated, but you should not blindly adapt analyses based on their results either - particularly in confirmatory research. You can always perform a sensitivity analysis in which you report both the planned analysis and the robust version.
Question 9
Cohen's D = mean difference/pooled sd.
Question 10
The slope tells you how much the predicted value goes up for a 1-unit increase in the predictor D. Since D is coded 0 and 2, a 1-unit increase only gets you halfway!

13.3 In SPSS

13.3.1 Independent Samples t-test

As a t-test and as regression with a dummy predictor:

https://www.youtube.com/watch?v=PbXMMN4YTQc

13.4 Tutorial

13.4.1 Independent Samples T-Test

In this assignment we will use the data file `5groups.sav`. Download the file and open it in SPSS.

This time, we will compare the means of the variable y of two specific groups: group 1 and group 4. To test the difference between two sample means, we will use the t-test for independent samples.

What is the null hypothesis of this test? And what is the alternative hypothesis?

Answer

H_0 $_1 = _4$, against H_1 $_1$ $_4$

Create the necessary syntax for the t-test that compares the means of group 1 and group 4.

You can find the dialog for the two-sample t-test under Analyze > Compare Means > Independent Samples T Test

In the SPSS dialog you have to specify which two groups you want to compare. In our case, it's group 1 and group 4. After placing the variable in the box named "Grouping Variable", click the button named "Define Groups" to define the groups.

Compare your syntax to the correct syntax:

Answer

T-TEST GROUPS=group(1 4) /MISSING=ANALYSIS /VARIABLES=y /CRITERIA=CI(.95).

One of the assumptions of the independent samples t-test is homoscedasticity (equal variances for all levels of the predictor). We can compare the sizes of the variances of the two groups with a simple F-test, which we call Levene's test.

Have a look at Levene's test and try to interpret it. Discuss with your group what null-hypothesis is being tested here.

What is the p-value of the Levene's test? _____[11]

What do you conclude from this? What's the practical use of the outcome of this test?

[11] 0.05

Explanation

Levene's test is not significant. Remember that the null hypothesis of Levene's test is that the population variances of the group are equal. As the p-value is not significant, we cannot reject the null hypothesis. Consequently, there is no evidence that the population variances of two groups are unequal. Thus, there is no reason to question the assumption.

Now you will have to decide on the outcome of the actual t-test. SPSS reports two versions: one that assumes equal variances (top row) and one that relaxes this assumption (bottom row).

You should pick one of these. In principle, you should decide which one you will use before seeing the results - although if there is clear evidence of violation of assumptions, you might want to discuss in your report whether the results change if you use the robust version (bottom row).

For now remember: we assume equal variances.

What is the two-sided p-value? _____[12]

Do you reject the null hypothesis of this t-test at alpha 0.05? [13]

- (A) Yes
- (B) No

13.4.2 Regression with dummies

We will now perform the exact same analysis, but with regression and dummies.

To test the difference between group 1 and group 4, we first create a dummy variable to distinguish these two groups. Use group 1 as reference category. You can use either Transform -> Recode into different variables, or syntax:

```
RECODE group (1=0) (4=1) INTO dgroup4.
EXECUTE.
```

Note that all other groups are coded as missing on this variable, which is exactly what we want!

We will use regression to perform our t-test. The hypothesis is the same as in the previous assignment, but you could also rewrite it in terms of regression coefficient(s). What is the

[12] 0.021
[13] Yes

null hypothesis of this test in terms of regression coefficient(s)? And what is the alternative hypothesis?

Answer

H_0 $_{group1vsgroup2} = 0$ which is the same as H_0 $_{group1} = {}_{group2}$, versus H_1 $_{group1vsgroup2}$ 0 which is the same as H_0 $_{group1}$ $_{group2}$

Create the necessary syntax for a regression with the dummy variable that compares the means of group 1 and group 4.

You can find the dialog under Analyze > Regression > Linear

In the SPSS dialog you have to specify the Dependent and Independent variable. In our case, the independent variable is the dummy we created.

Compare your syntax to the correct syntax:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER dgroup4.
```

Note that, unlike the t-test interface, the regression interface does not provide a Levene's test. This is one reason you might want to use the t-test interface. The regression interface provides a more generic way to test the assumption of homoscedasticity: a residual plot.

Go back through the regression interface, but this time click the Plots button and plot the predicted value (X = ZPRED) against the residual value (Y = ZRESID).

Your syntax will now say:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER dgroup4
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
```

If the assumption of homoscedasticity is met, we should see that the dots in this plot are equally distributed around the zero line for all values on the X-axis. In this case, we see much narrower spread on the right side than on the left side.

What can you conclude from this, and does it match your conclusion from Levene's test?

Now you will have to decide on the outcome of the actual t-test.

Remember that the t-test of the dummy variable should be the same as the t-test we conducted before. Verify that this is true.

What is the two-sided p-value? _____[14]

We see one more t-test: for the "(Constant)" or intercept. How do we interpret this?
[15]

- (A) The mean difference between both Groups is 7, and this differs significantly from zero.
- (B) The mean value in Group 4 is 7, and this differs significantly from zero.
- (C) The mean value across both Groups is 7, and this differs significantly from zero.
- (D) The mean value in Group 1 is 7, and this differs significantly from zero.

[14] 0.021
[15] The mean value in Group 1 is 7, and this differs significantly from zero.

14 GLM-IV: ANOVA

We can also use regression analysis to examine mean differences between the categories of a nominal or ordinal predictor with more than two categories. Suppose we have a categorical predictor, such as socioeconomic status (SES), with three categories: Low, Medium, and High. We want to predict fathers' involvement in child rearing based on these SES categories. We can use regression analysis to model this relationship by using dummy variables.

We previously discussed how *bivariate* linear regression allows us to model the effect of a binary categorical predictor by using dummy coding. We code one variable as the reference group (giving it the value 0), and estimate the mean difference between the reference group and the other category.

When we have two or more categories, we can use the same principle - but we need to expand the model. For our example with SES, we can select one reference category (say, High SES), and we would create **two** dummy variables to estimate the mean differences between the reference category and the Medium and Low SES categories. Our regression model then includes both dummy variables as predictors, along with the intercept term.

This regression model is completely equivalent to one-way ANOVA (Analysis of Variance). Think of ANOVA as a different interface to the same analysis, which presents the results in a slightly different way that is more common in some subfields of social science.

When we perform an ANOVA, we conduct an omnibus test of differences between group means. The default null hypothesis is that all group means are equal, and the alternative hypothesis suggests that at least two group means differ. We test this hypothesis with an F-test for the overall significance of the model. You are already familiar with this test from the lecture on sums of square.

One way to think about the F-test in the context of ANOVA is that it compares the size of the variance (differences) in group means, relative to the error variance in the data. If the differences between group means are large relative to the spread of the data, we observe a significant test. In ANOVA, the regression sum of squares is also called the "between-group sum of squares", and the error sum of squares is also called "within-group sum of squares".

When interpreting the results of ANOVA, it's common to use eta squared [2] as an effect size. It is simply another name for the familiar R^2. It reflects the proportion of variance in the outcome variable that can be explained by the categorical predictor.

14.1 Lecture

https://www.youtube.com/embed/2Z8fhhs69N0

14.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

How can you model a categorical predictor with more than 2 categories in regression? [1]

- (A) Create dummy variables for each category and include them in the regression equation.
- (B) Exclude the categorical predictor from the regression model.
- (C) Convert the categorical predictor into a binary predictor.
- (D) Use a single continuous variable to represent all categories.

Question 2

What does the intercept term represent in a regression model with dummy variables? [2]

- (A) The difference between the means of all categories.
- (B) The mean difference between all categories.
- (C) The standard deviation of the predictor variable.
- (D) The mean value of the reference category.

Question 3

How many dummy variables are created for a categorical predictor with three categories? [3]

[1] Create dummy variables for each category and include them in the regression equation.
[2] The mean value of the reference category.
[3] Two dummy variables are created, each representing membership in one of the non-reference categories.

- (A) For three categories, you don't need dummy variables.
- (B) Three dummy variables are created, each representing membership in a category.
- (C) Two dummy variables are created, each representing membership in one of the non-reference categories.
- (D) One dummy variable is created, representing membership in the reference category.

Question 4

What does the F-value represent in ANOVA? [4]

- (A) How large the variance between group means is relative to variance within groups.
- (B) How large the difference between the group means is, relative to the error variance.
- (C) An overall test of the difference between group means.
- (D) The proportion of variance explained by the predictor variable.

Question 5

What is the purpose of follow-up analyses after a significant ANOVA? [5]

- (A) To understand which specific group means differ significantly from each other.
- (B) To understand which groups are significant.
- (C) To calculate the effect size.
- (D) To adjust the p-value for multiple comparisons.

Question 6

How are the degrees of freedom calculated for the F-distribution in ANOVA? [6]

- (A) Numerator df: Total number of observations - 1; Denominator df: Number of groups
- (B) Numerator df: Total number of observations - Number of groups; Denominator df: Number of groups - 1

[4] How large the variance between group means is relative to variance within groups.
[5] To understand which specific group means differ significantly from each other.
[6] Numerator df: Number of groups - 1; Denominator df: Total number of observations - Number of groups

- (C) Numerator df: Number of groups - 1; Denominator df: Total number of observations
- (D) Numerator df: Number of groups - 1; Denominator df: Total number of observations - Number of groups

Question 7

What is the correct interpretation of a small eta squared (š) value in ANOVA? [7]

- (A) A small proportion of the total variance is due to individual differences.
- (B) A small proportion of the total variance is due to error.
- (C) A small proportion of the total variance is explained by the group differences.
- (D) A small proportion of the error variance is explained by the group differences.

Question 8

Given the regression equation Y = 20 + 5*D1* - *3*D2 for an ANOVA model with dummy coded predictors, what is the predicted value of Y when D1 = 2 and D2 = 1? [8]

- (A) Y = 20 + 5*1* - *3*1 = 22
- (B) Y = 20 - 5*0* - *3*1 = 17
- (C) Y = 20 + 5*1* + *3*0 = 25
- (D) This cannot happen as the dummy variables are orthogonal.

Question 9

In an ANOVA model with 4 groups and 300 observations, calculate the degrees of freedom for the numerator and denominator for the F-test. [9]

- (A) Numerator df: 4 - 1 = 3; Denominator df: 300 - 4 = 296
- (B) Numerator df: 300 - 4 = 296; Denominator df: , 4 - 1 = 3
- (C) Numerator df: 300 - 1 = 299; Denominator df: , 4 - 1 = 3
- (D) Numerator df: 4 - 1 = 3; Denominator df: 300 - 1 = 299

[7] A small proportion of the total variance is explained by the group differences.
[8] This cannot happen as the dummy variables are orthogonal.
[9] Numerator df: 4 - 1 = 3; Denominator df: 300 - 4 = 296

Question 10

In an ANOVA model, the variation of individual observations with respect to the grand mean (SST) is 1200, and the variation of individuals with respect to group means (SSW) is 800. Calculate the proportion of variance explained by the group means (š). [10]

- (A) š = SSB/SST = (1200 + 800)/1200 = 2.0
- (B) š = SSB/SSW = (1200 - 800)/800 = 0.5
- (C) š = SSW/SST = 800/1200 = 0.667
- (D) š = SSB/SST = (SST - SSW)/SST = (1200 - 800)/1200 = 0.333

[10] š = SSB/SST = (SST - SSW)/SST = (1200 - 800)/1200 = 0.333

Show explanations

Question 1
To model a categorical predictor with more than 2 categories, you create dummy variables for each category and include them as predictors in the regression equation.
Question 2
The intercept term in a regression model with dummy variables represents the mean value of the reference category.
Question 3
For a categorical predictor with three categories, two dummy variables are created, each representing membership in one of the non-reference categories.
Question 4
The F-test in ANOVA measures how large the variance in group means is relative to the error variance, helping us determine if there are significant differences between group means.
Question 5
Follow-up analyses are conducted after a significant ANOVA to understand which specific groups differ significantly from each other, as the omnibus ANOVA only tells us there are differences among groups but not which ones.
Question 6
The degrees of freedom for the F-distribution in ANOVA are calculated as follows: Numerator df = Number of groups - 1; Denominator df = Total number of observations - Number of groups.
Question 7
A small eta squared (š) value in ANOVA indicates that a small proportion of the total variance is explained by the group differences, suggesting weaker group effects.
Question 8
It is not possible for an observation to score 1 on two dummies.
Question 9
The degrees of freedom for the F-test are calculated as follows: Numerator df = Number of groups - 1; Denominator df = Total number of observations - Number of groups.
Question 10
The proportion of variance explained by the group means (š) is calculated as š = SSB/SST = (SST - SSW)/SST = (1200 - 800)/1200 = 0.333.

14.3 In SPSS

14.3.1 ANOVA

Using the ANOVA interface and the regression interface:

https://www.youtube.com/watch?v=LXkytSgHl6c

14.4 Tutorial

14.4.1 ANOVA

For this assignment, we will use the data file `5groups.sav`. Please open it in SPSS.

As you have seen in the previous assignments, this file contains the measurements of the y variable for 5 different groups. So far, we have - at most - compared two groups at once. This time, we will compare the means of all 5 groups simultaneously using an Analysis of Variance (ANOVA).

ANOVA is often used to examine the results of experimental research where different groups receive different manipulations of an independent variable, and a continuous dependent variable is measured. In that case, rejecting the null hypothesis indicates a causal effect of the manipulated independent variable on the dependent variable.

Let's say the 5 groups in our data file have received 5 different types of training, and the dependent variable y measures the effect of the training.

We will now perform an ANOVA to see if these trainings have an equal effectiveness.

The null hypothesis of the ANOVA with 5 groups is as follows:

$H_0 \ _1 = _2 = _3 = _4 = _5$

That is, the null hypothesis states that the population means of all five groups are the same. Rejecting the null hypothesis implies that at least one one of these means is different from the rest.

Let's run the analysis!

Navigate to Analyze > General Linear Model > Univariate. Choose y as the dependent variable. Choose group as the fixed factor. Now click on the "Options" button and check the two boxes named "Descriptive statistics" and "Homogeneity tests". Finally, click "Paste" to paste the syntax into the syntax editor, and run it from there.

You should end up with the following syntax:

```
UNIANOVA y BY group
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=group.
```

The first table in the output we will inspect is the "Descriptive Statistics" table. This table displays the means and standard deviations of the variable y for each of the five groups.

Do you think the population standard deviations are different for each group? If they are, why could that pose a problem for our analysis?

One of the assumptions of ANOVA is homoscedasticity. In this case, that means that the population of each group has the same variance (and hence, the same standard deviation). This assumption is also called "homogeneity of variance" (= translation of homoscedasticity). Of course the variance in each sample will differ somewhat. If these differences are significant, there is evidence to doubt our assumption. That's why we asked SPSS to perform "Homogeneity tests".

Note that the SD's of groups 1 and 2 are quite different from the SD's of groups 3, 4, 5.

The next table shows the output of Levene's test. You might remember using Levene's test for comparing the variances of two groups in the context of the independent samples t-test. This time, it tests whether the variance of all 5 groups should be considered equal.

Is there a reason to doubt the assumption of homoscedasticity based on Levene's test? Note: Use the Levene's test "Based on mean". `mcq(c(answer = "Yes", "No"))`

If Levene's test is significant, there is evidence that the population variances of at least 2 of the groups differ. This is evidence against our assumption. This could pose a problem for our analysis. We may choose to use a version of the analysis that is robust to violations of this assumption instead, but that makes our analysis dependent on the data (= no longer confirmatory). Instead, we could discuss the violation, and compare results with and without a robust test.

For now, we continue with interpreting the output of the final table. There's a lot of information, but for now we are only interested in the Sig. value of the "Corrected Model" in the first row. This is the two-sided p-value we can use to test our null hypothesis.

What is the two-sided p-value? Do you reject the null hypothesis? What does that mean?

Answer

The p-value is <.001. This is smaller than 0.05. Therefore, we reject the null hypothesis, which was: H_0 $1 = 2 = 3 = 4 = 5$
This means that the means of at least two groups are different. Note that we do not yet know for which groups the means differ!

Finally, there's an interesting nugget of information below the final table. It's called R Squared. This shows the total amount of variance in y that is explained by group membership.

What is the value of R Squared? _____[11]

[11] 0.57

By rule of thumb, what is the magnitude of this value (small, medium, or large)?

Cohen (1988) proposed the following guidelines for interpreting the magnitude of R2:

Size	R2
Small	0.01
Medium	0.06
Large	0.138

14.4.2 ANOVA using regression

This time, we will conduct the ANOVA using the regression interface.

When we conduct ANOVA using regression, we still test the null hypothesis mentioned before:

$H_0 \ _1 = _2 = _3 = _4 = _5$

A different way to phrase this when using regression is to state that all regression coefficients will be zero:

$H_0 \ _0 = _1 = _2 = _3 = _4$

Or to say that the explained variance will be zero:

$H_0 \ ^2 = 0$

All of these hypotheses are interchangeable and imply that the means of all five groups are the same. If this is not the case, each of these null-hypotheses would be rejected. Rejecting these null hypothesis implies that at least one one of these means is different from the rest.

To test the differences between groups, we first create dummy variables. Let's make them for all categories, but we will mostly use group 1 as reference category. When making dummies, it's most convenient to use syntax:

```
RECODE group (1=1) (2=0) (3=0) (4=0) (5=0) INTO dgroup1.
RECODE group (1=0) (2=1) (3=0) (4=0) (5=0) INTO dgroup2.
RECODE group (1=0) (2=0) (3=1) (4=0) (5=0) INTO dgroup3.
RECODE group (1=0) (2=0) (3=0) (4=1) (5=0) INTO dgroup4.
RECODE group (1=0) (2=0) (3=0) (4=0) (5=1) INTO dgroup5.
EXECUTE.
```

Create the necessary syntax for a regression with the dummy variable that compares the means of group 1 against all other groups.

You can find the dialog for the regression under Analyze > Regression > Linear

In the SPSS dialog you have to specify the Dependent and Independent variable. In our case, the independent variables are all dummies we created, except for the reference category! Use category 1 as reference category.

Compare your syntax to the correct syntax:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER dgroup2 dgroup3 dgroup4 dgroup5.
```

Also compare this syntax to the one we used for t-test using regression. What is the only difference?

Which test statistic do you use to determine the significance of your ANOVA? [12]

- (A) R2
- (B) F
- (C) All of the ts
- (D) t for the Constant

What is the value of the test statistic? _____ [13]

Compare this output to the output of your previous ANOVA!

Which parts are the same? Which parts are different?

Answer

The F-test is identical to the one from the ANOVA. What's different is that the regression also gives us t-tests for the difference between each group and the reference group. By changing the reference group, we can make all possible comparisons.

[12] F

[13] 14.783

Note that, unlike the t-test interface, the regression interface does not provide a Levene's test. This is one reason you might want to use the t-test interface. The regression interface provides a more generic way to test the assumption of homoscedasticity: a residual plot.

Go back through the regression interface, but this time click the Plots button and plot the predicted value (X = ZPRED) against the residual value (Y = ZRESID).

Alternatively, just add this line to your syntax (make sure to remove the period . from what was previously the last line):

```
/SCATTERPLOT=(*ZRESID ,*ZPRED).
```

If the assumption of homoscedasticity is met, we should see that the dots in this plot are equally distributed around the zero line for all values on the X-axis. In this case, we see some differences that could lead us to question the assumption. However, we don't get an actual test, which is a pity. Thus, you could use the ANOVA interface if you want this test.

14.4.3 One-Way ANOVA

We have prepared the following data file for this assignment: `hiking.sav`. Please download it and open it in SPSS.

The data file describes the result of a fictitious experiment in which a hiking guide has displayed five different types of behavior towards different groups of hikers. The treatment that each person received from the guide is recorded in the variable behavior.

The dependent variable of this experiment is feeling. Higher scores on this variable indicate a more positive attitude of a participant towards the guide. In this assignment, we will use ANOVA to determine whether the mean score on the dependent variable differs between the five experimental conditions.

What type of design do we use for this experiment? [14]

- (A) Within-subjects design
- (B) Between-subjects design
- (C) Combination of the two

[14] Between-subjects design

As you will have noticed, the data file contains a third variable named weather, which can be either good or bad. For now, we will only look at the results obtained during good weather. Hence, we will use "Select cases" to select only those participants with a value of 1 on the weather variable.

Click Data > Select Cases and select "If condition is satisfied" and click the "If"-button. Now enter the following condition into the equation box:

weather = 1

Now click "Continue" and "Paste" to paste the resulting syntax into the syntax editor. Select Run > All to run it.

Verify that half of the participants have been crossed out in the Data View.

We are now ready to perform an ANOVA with the 50 remaining participants.

To run an ANOVA in SPSS there are multiple options. We will use the module "General Linear Model", which encompasses ANOVA and all of its extensions which we will discuss later on (factorial ANOVA, ANCOVA, and repeated measurements).

Anyway, let's first create the basic syntax.

Analyze > General Linear Model > Univariate Choose feeling as the dependent variable and behavior as the fixed factor Click on the "Options" button and check the two boxes named "Descriptive" and "Homogeneity tests". After clicking "Paste" you should get the following syntax:

```
UNIANOVA feeling BY behavior
    /METHOD=SSTYPE(3)
    /INTERCEPT=INCLUDE
    /PRINT=DESCRIPTIVE HOMOGENEITY
    /CRITERIA=ALPHA(.05)
    /DESIGN=behavior.
```

What is the p-value of the Levene's test? Use the Levene's test "Based on mean" again. _____ [15]

Do we have reason to question the assumption of equal population variances? [16]

- (A) Yes
- (B) No

[15] 0.611
[16] No

In ANOVA, we distinguish between three sources of variation: the Sums of Squares total (SSt), the Sums of Squares between (SSb, or SSR) and the Sums of Squares within (SSw, or SSE).

What does the Sum of squares total mean (phrase your answer in your own words)? Look up the value of the SSt in the ANOVA output and write it down as well.

What does the Sum of squares between entail (phrase your answer in your own words)? Look up the value of the SSb in the ANOVA output, and write it down as well.

What does the Sums of squares within entail (phrase your answer in your own words)? Look up the value of the SSw in the ANOVA output, and write it down as well.

Answer

The Between Groups Sum of squares or SSb is equal to 18.330 and simply give the squared distance of individual scores to the mean, summed together. In other words, it shows how much variability there is in the group means. If all group means would be equal to each other, the SSb equals 0.
The SSw here is 38.405. It tells us how much the individual scores within a group deviate from the group mean. In other words, it shows how much variability there is within the groups. The is the variation that is independent from the experimental effect (because variation within groups cannot be caused by differences in experimental conditions).
The SSt tells us how much the invidual scores deviate from the grand mean. In other words, it shows how much variability there is in the dependent variable in total.
Recall that SSB is the same as SSR; it can be found in the row "Corrected Model", column Type III Sums of Squares.
Recall that the SSW is the same as SSE; it islabeled "Error" in the column Type III Sums of Squares.

How do we use the different types of Sum of Squares to calculate the F statistic?

[17]

- (A) (SSB/dfb)/(SSW/dfw)
- (B) (SSW/dfw)/(SSB/dfb)
- (C) (SSW)/(SSB)
- (D) (SSB)/(SSW)

[17] (SSB/dfb)/(SSW/dfw)

Answer

We can calculate F using the following formula:
$F = \frac{MSb}{MSw}$
The MSb and MSw give the between group variance and within group variance, respectively. They can be calculated using the following formula's:
$MSb = \frac{SSb}{k1}$, $MSw = \frac{SSw}{Nk}$

Again, consider the table Tests of Between Subjects, which represents the results of ANOVA.

What is the F-value of the ANOVA? _____[18]

The degrees of freedom between (dfb) are _[19] and the degrees of freedom within (dfw) are ___[20].

dfb = k-1

dfw = N-k

Again consider the table Tests of Between Subjects

What is the p-value of the ANOVA? _____[21]

You can find the p-value of the ANOVA in the table named "Tests of Between-Subjects Effects". The p-value is equal to the Sig.-value in the first row of this table.

What can you conclude from this?

Write down a statistical conclusion and a conclusion within the context of this research example.

Answer

The p-value is smaller than our alpha level (0.05). Therefore, we can conclude that there was a statistically significant difference in positive attitude between the groups, based on the behaviour the guide displayed towards them, (F(4,45) = 5.369, p = .001).

What is the proportion of variance explained by behavior? _____[22]

How would you describe this number in words? So what does it mean?

Remark: Cohen formulated some rules of thumb for interpreting the R^2 How would you qualify the strength of the effect based on Cohen's rules of thumb? And why should we not take the rules of thumb too seriously?

[18] 5.369
[19] 4
[20] 45
[21] 0.001
[22] 0.323

Cohen (1988) proposed the following guidelines for interpreting the magnitude of R^2

Size	R^2
Small	0.01
Medium	0.06
Large	0.14

Note that, in ANOVA, R^2 is also called ^{2} (eta squared) is a measure of effect size, it indicates the amount of variance in thedependen t variable that is explained by the independent variable(s). In our case, 32.3% of the variance in feeling is explained by behaviour. According to Cohen's guidelines this is a large effect size (see slide 28). However, these guidelines are rather arbitrary, which Cohen himself also stresses.

The correct conclusion so far is that the five groups differ significantly on the dependent variable feeling. However, we do not yet know which groups differ.

14.4.4 One-Way ANOVA using regression

In this assignment, we will conduct the same ANOVA using the regression interface.

To test the differences between groups, we first create dummy variables. Let's make them for all categories, but we will mostly use group 1 as reference category. When making dummies, it's most convenient to use syntax. Let's give the dummies informative names this time:

```
RECODE behavior (1=1) (2=0) (3=0) (4=0) (5=0) INTO rushing.
RECODE behavior (1=0) (2=1) (3=0) (4=0) (5=0) INTO stories.
RECODE behavior (1=0) (2=0) (3=1) (4=0) (5=0) INTO insulting.
RECODE behavior (1=0) (2=0) (3=0) (4=1) (5=0) INTO jokes.
RECODE behavior (1=0) (2=0) (3=0) (4=0) (5=1) INTO singing.
EXECUTE.
```

Create the necessary syntax for a regression with the dummy variable that compares the means of the rushing group against all other groups.

You can find the dialog for the regression under Analyze > Regression > Linear

In the SPSS dialog you have to specify the Dependent and Independent variable. In our case, the independent variables are all dummies we created, except for the reference category! Use category 1 as reference category.

Compare your syntax to the correct syntax:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT feeling
  /METHOD=ENTER stories insulting jokes singing.
```

Can you find all three of the aforementioned sources of variation in the output? The Sums of Squares total (SSt), the Sums of Squares between (SSb, or SSR) and the Sums of Squares within (SSw, or SSE).

Compare the results to those of the One-Way ANOVA interface.

Answer

In the ANOVA table, the Regression Sum of Squares is identical to the "Between Groups Sum of Squares" from the One-Way ANOVA interface. The Residual Sum of Squares is identical to the "Within Groups Sum of Squares" from the One-Way ANOVA interface. And the Total Sums of Squares are also the same.

15 GLM-V: Multiple regression

15.0.1 Multiple regression

Multiple regression is a statistical technique that allows us to examine the relationship between one outcome and multiple predictors. It extends the concept of bivariate linear regression, where we model the relationship between two variables, to include more predictors. In the context of social science research, multiple regression helps us answer the question: What is the unique effect of one predictor, while controlling for the effects of all other predictors?

As a matter of fact, last week's analyses for categorical variables with more than two categories were already an example of multiple regression. We included two dummy variables to represent a categorical variable with three categories. All that's new today is that we also consider the case where the multiple predictors are continuous variables. An important realization is that a regression model can be expanded to include as many predictors as needed. The general formula for multiple regression is $Y = a + b_1X_1 + b_2X_2 + \ + b_KX_K$, where Y represents the predicted value of the dependent variable Y, a is the intercept, and b_{1K} are the slopes for each predictor.

When interpreting the regression coefficients, the intercept (a) represents the expected value of the dependent variable when all predictors are equal to 0. For dummy variables, this is the mean value of the reference category, while for continuous predictors, it represents the expected value for someone who scores 0 on all predictors. The regression coefficients (b1, b2, ..., bK) indicate how many units the dependent variable Y is expected to change when the corresponding predictor X increases by 1 unit, while holding all other predictors constant.

Centering predictors can be useful in multiple regression. By centering, we shift the zero-point of the predictor to a meaningful value, such as the mean value on that predictor. This helps in interpretation, because the intercept now gives us the mean value on the outcome for someone who has an average score on all predictors.

As previously explained, standardized regression coefficients drop the units of the predictor and outcome variable. They are calculated by transforming the predictors and outcome variable into z-scores with a mean of 0 and a standard deviation of 1, and performing the (multiple) regression analysis on those z-scores. Because the units of the variables are dropped, standardized coefficients make the effects of predictors comparable across different studies or variables with different measurement units. They represent the change in the dependent variable in terms of standard deviations when the corresponding predictor increases by 1 standard deviation.

15.0.2 Causality

A statistical association between variables does not necessarily imply a causal relationship. Instead, causality is either assumed on theoretical grounds, or established using the experimental method. In an experiment, researchers manipulate an independent variable and observe its effects on the dependent variable. However, in many social science studies, experiments are not feasible or ethical, so researchers rely on observational data. In these cases, establishing causal relationships relies on theory and careful statistical analysis.

One important concept in causal inference is the direction of effects. While statistical methods can identify associations between variables, determining the direction of causality is a causal assumption that cannot be estimated using statistics alone. The assumed direction of effects is often based on theory and prior knowledge of the subject matter. Researchers make informed assumptions about which variable is likely to have a causal effect on the other based on theoretical reasoning and empirical evidence.

In the process of analyzing causal relationships, it is essential to consider the presence of confounders, mediators, and colliders. Confounders are variables that are associated with both the independent and dependent variables and can create a spurious association or distort the true causal relationship. Identifying and controlling for confounders is crucial to ensure accurate causal inference.

Mediators, on the other hand, are variables that explain the relationship between the independent and dependent variables. They act as intermediate steps or process variables in the causal pathway. Understanding and analyzing mediation effects help us understand the underlying mechanisms through which the independent variable affects the dependent variable.

Colliders are variables that are caused by both the independent and dependent variables. Controlling for colliders can lead to spurious statistical relationships between unrelated variables. It is essential to be cautious when including variables in the analysis and consider the causal structure of the variables involved.

One important take-home message is that, in multiple regression, the distinction between confounders and colliders is crucial. Including confounders as control variables in multiple regression improves our inferences - but accidentally including a collider as control variable (severely) biases our inferences. You therefore have to carefully reason about each variable's role in relation to the other variables in the model.

15.1 Lectures

https://www.youtube.com/embed/lT-6YoSfwC4

https://www.youtube.com/embed/wBB4sed9ku0

15.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the primary advantage of using multiple regression analysis? [1]

- (A) To understand the unique effect of each predictor while accounting for others
- (B) To group predictors based on their significance
- (C) To identify the most significant predictor
- (D) To find a single predictor that explains all the variance

Question 2

What is the purpose of centering in multiple regression analysis? [2]

- (A) To ensure the assumption of normality is met
- (B) To standardize all predictors
- (C) To remove outliers from the data
- (D) To choose a meaningful zero-point for predictors

Question 3

What does the intercept (a) in a multiple regression model represent? [3]

- (A) The effect of the first predictor
- (B) The effect of the last predictor
- (C) Expected value when all predictors are equal to 0

[1] To understand the unique effect of each predictor while accounting for others
[2] To choose a meaningful zero-point for predictors
[3] Expected value when all predictors are equal to 0

- (D) The difference between the two groups

Question 4

What do the b-coefficients in a multiple regression model represent? [4]

- (A) The change in the dependent variable for a one-unit change in the predictor while other predictors are held constant
- (B) The average of all predictor values
- (C) The change in the predictor for a one-unit change in the dependent variable while other predictors are held constant
- (D) The change in the dependent variable for a one-unit change in the predictor without considering other predictors

Question 5

When is multicollinearity a concern in multiple regression analysis? [5]

- (A) When predictor variables are independent
- (B) When there are multiple outcome variables
- (C) When predictor variables are highly correlated with each other
- (D) When there are outliers in the data

Question 6

What is the role of standardized regression coefficients in multiple regression analysis? [6]

- (A) To compare the effect sizes of predictors on a common scale
- (B) To calculate the intercept
- (C) To determine the unique effect of each predictor
- (D) To convert categorical predictors to continuous ones

[4]The change in the dependent variable for a one-unit change in the predictor while other predictors are held constant
[5]When predictor variables are highly correlated with each other
[6]To compare the effect sizes of predictors on a common scale

Question 7

What is the potential bias introduced when controlling for a collider in multiple regression analysis? [7]

- (A) It biases the estimate of the association between the two variables that are caused by the collider
- (B) It biases the estimate of the association between the two variables that cause the collider
- (C) It creates a causal relationship between the variables that cause the collider
- (D) It has no effect on the causal inference between the variables that form the collider

Question 8

Given the multiple regression equation: Y = 12.5 + 2.3*X1 + 1.8*X2 - 0.5*X3, calculate the predicted value of Y when X1 = 5, X2 = 3, and X3 = 2. [8]

- (A) 23.1
- (B) 20.5
- (C) 28.4
- (D) 27.6

Question 9

In a multiple regression model, if the coefficient of determination (Rš) is 0.75 and the SSE is 150, what is the value of SST? [9]

- (A) 0.25
- (B) 112.5
- (C) 450
- (D) 600

[7] It biases the estimate of the association between the two variables that cause the collider
[8] 28.4
[9] 600

Question 10

Given the following standardized regression equation: Y = 0.6 + 0.35*X1 + 0.25*X2 - 0.15*X3, what is the correct conclusion about the effect of X2? [10]

- (A) A one unit increase in X2 is associated with a 0.25 increase in Y, keeping all other predictors constant.
- (B) X2 is associated with a 0.25 increase in Y, keeping all other predictors constant.
- (C) A one SD increase in X2 is associated with a 0.25 SD increase in Y, keeping all other predictors constant.
- (D) A one unit increase in X2 is associated with a 0.25 unit increase in Y.

[10] A one SD increase in X2 is associated with a 0.25 SD increase in Y, keeping all other predictors constant.

Show explanations

Question 1
Multiple regression analysis helps to understand the unique effect of each predictor while controlling for the effects of other predictors.
Question 2
Centering is used in multiple regression analysis to choose a meaningful zero-point for predictors.
Question 3
The intercept (a) in a multiple regression model represents the expected value when all predictors are equal to 0.
Question 4
The b-coefficients in a multiple regression model represent the change in the dependent variable for a one-unit change in the predictor while other predictors are held constant.
Question 5
Multicollinearity is a concern in multiple regression analysis when predictor variables are highly correlated with each other.
Question 6
Standardized regression coefficients are used to compare the effect sizes of predictors on a common scale, especially when the units of predictors are not meaningful.
Question 7
Controlling for a collider in multiple regression analysis can introduce bias in the estimated association between the variables that form the collider.
Question 8
Plug in the values of X1, X2, and X3 into the regression equation: Y = 12.5 + 2.3*5* + *1.8*3 - 0.5*2 = 28.4.
Question 9
Use the formula R2 = 1-(SSE/SST) or R2 = SSR/SST
Question 10
The correct interpretation is in the original units of the variables, and emphasizing the fact that other predictors were controlled for.

15.3 In SPSS

15.3.1 Multiple Regression

https://www.youtube.com/watch?v=ueNrP5TyZaE

15.4 Tutorial

15.4.1 Multiple Regression

Social science students were asked about their opinion towards Tilburg's nightlife, number of Facebook friends, and some other characteristics. The data are in the SocScSurvey.sav file.

In a previous assignment we predicted Facebook friends by extraversion.

In this question we will add another predictor, peer pressure.

The variable peer pressure refers to the tendency to be influenced by close friends. Higher scores reflect higher sensitivity to peer pressure.

Before we proceed with the regression analysis, we will first look at the correlations between the variables.

Analyze > correlate > bivariate.

Now choose as variables: Facebook Friends, Extraversion and Peer Pressure, and click OK.

What is the correlation between peer pressure and number of Facebook friends? _____[11]

Suppose three researchers test the significance of the correlation between peer pressure and Facebook friends. Researcher I tests at the 10% level, researcher II tests at the 5% level, and researcher III at the 1% level.

Which researcher will reject the null hypothesis? [12]

- (A) All three researchers
- (B) Only researcher I
- (C) Only researcher III
- (D) Only researcher II

Now, run the regression analysis in which the number of Facebook friends is regressed on extraversion and peer pressure.

Proceed as follows: via analyze > regression > linear. Choose Facebook friends as dependent and extraversion and peer pressure as independents.

Consult the output and write down the regression equation.

[11] 0.145
[12] Only researcher I

Answer

$\text{Friends}_i = 158.012 + 26.560\ \text{Extraversion}_i + 12.056\ \text{Peer}_i + e_i$

Consider the second person in the sample. The person had an extraversion score of 11 and a score of 9 on peer pressure.

What is the predicted number of Facebook friends for this person? _____[13]

Consult the output.

Researchers conclude that – in the sample – as peer pressure increases with one unit, the predicted number of Facebook friends increases with 12.056 units.

Is this a valid conclusion? [14]

- (A) Yes
- (B) No

In multiple regression, the regression coefficients show us the expected changes in the dependent variable, while keeping the other independent variables constant.

With this in mind, what is the correct conclusion?

[15]

- (A) As peer pressure increases with one unit the predicted number of Facebook friends increases with 12.056 units. This is added to the constant of -158.01.
- (B) As peer pressure increases with one unit the predicted number of Facebook friends increases with 12.056 units.
- (C) As peer pressure increases with one unit the predicted number of Facebook friends increases with 12.056 units, while keeping extraversion constant.
- (D) As peer pressure increases with one unit the predicted number of Facebook friends increases with 12.056 units, while extraversion changes with 26.56 units.

Consult the table Coefficients. The table shows the results of t-tests.

What are the null hypotheses and alternative hypotheses that are tested here?

[13] 242.652
[14] No
[15] As peer pressure increases with one unit the predicted number of Facebook friends increases with 12.056 units, while keeping extraversion constant.

Answer

The t-tests test significance of the individual regression coefficients. In particular, for each coefficient we can use the t-tests to test the following hypotheses:
$H_0 \quad = 0$, $H_1 \quad 0$

What is the value of the test-statistic for the significance test for extraversion? _____[16]

Consider the t-tests for the regression coefficients again.

How many degrees of freedom do the t-tests have? _____[17]

Explanation

Note that:
Degrees of freedom = N - p N = number of participants; p = number of parameters in the model (intercept + two regression slopes)

Suppose three researchers test the significance of peer pressure as a predictor of Facebook friends, while controlling for extraversion. Researcher I tests at the 10% level, researcher II tests at the 5% level, and researcher III at the 1% level.

Which researcher(s) will reject the null hypothesis? [18]

- (A) Only researcher I
- (B) Only researcher II
- (C) All three researchers
- (D) Only researcher III

What *percentage* of the total variance in Facebook friends can be explained by both extraversion and peer pressure? _____[19]

Compare the R^2 of the regression model with both predictors with the R^2 of a model with only extraversion as the predictor.

What is the difference? _____[20]

In the previous step we compared the R^2 of two so called nested models.

---

[16] 5.583
[17] 131
[18] Only researcher I
[19] 20.9
[20] 0.017

Two models are nested if the larger model (i.e., the model with the most predictors) contains all predictors of the smaller model.

In the next lecture we will learn more about nested models, model comparisons, and how useful they are for researchers!

15.4.2 Multiple Regression II

For this assignment we need the file HealthyFood.sav.

This file contains hypothetical data on three variables:

Eating healthy food (the higher the score, the healthier a person's diet) Knowledge about food (the higher the score, the more a person knows about healthy food and risks of unhealthy food) Income (higher scores = more income).

Let's first look at the associations (correlations) between the three variables.

Compute the correlations and summarize the relationships between all pairs of variables. Include in your answer the strength of the relationship (i.e., weak, moderate, or strong), the direction of the relationship (i.e., positive or negative), and generalizability to the population (i.e., is the correlation significant at the 5% level).

Cohen's rules of thumb:

- r = 0.00-0.30 (none to weak)
- r = 0.30-0.50 (weak to moderate)
- r = 0.50-0.70 (moderate to strong)
- r = 0.70-0.90 (strong to very strong)
- r = 0.90-1.00 (very strong)

Answer

- Income and eating have a weak positive correlation, which is significant at the 5% level.
- Income and knowledge have a weak to moderate positive correlation, which is significant at the 5% level.
- Knowledge and eating have a moderate to strong positive correlation, which is significant at the 5% level.

Researchers may be interested in explaining differences in eating healthy food: in other words, they want to know why some people eat very healthy, while others tend to eat unhealthy.

One of the hypotheses is that healthy food is on average more expensive than unhealthy food, so one of the explanatory variables may be income.

Run a regression analysis using eating as the dependent variable and income as the independent variable.

Consult the output.

Which of the following conclusions is correct?

[21]

- (A) The effect of Income on eating healthy food is positive and significantly different from zero.
- (B) The effect of Income on eating healthy food is positive but not significantly different from zero.

Simple regression analysis suggests a positive relationship between income and healthy food.

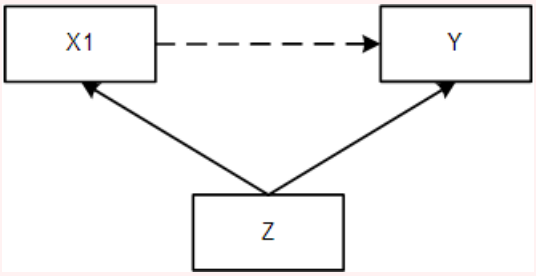
However, other researchers (say Team B) came up with an alternative explanation. They hypothesized that the relationship between income and healthy food can be explained by a confounder; knowledge. People with more knowledge will have better jobs (on average), and, as a result more, income. As the result of their knowledge they also prefer to eat healthy food. I.e., Team B thinks knowledge is a common cause of income and eating healthy food.

In other words, the researchers of Team B hypothesize that the relationship between income and eating healthy food is [22]

- (A) Indirect
- (B) Spurious

Draw (on a piece of paper) the conceptual model that reflects the hypotheses of the researchers.

Answer



Now let's see if the data support the hypotheses of the researchers.

[21] The effect of Income on eating healthy food is positive and significantly different from zero.
[22] Spurious

Run a multiple regression analysis using eating healthy food as the dependent variable and income and knowledge as independent variables.

Consult the output. Look at the effect of income, controlled for knowledge (both the coefficient and the significance test).

- What happened with the effect of income if you control for knowledge?
- Does knowledge predict eating healthy food (controlled for income)?
- Do the data support the hypothesis that the relationship between income and healthy food is confounded by knowledge?

What is the p-value of Income when you control for Knowledge? _____[23]

What is the p-value of Knowledge, controlling for Income? _____[24]

Do the data support the hypothesis that the relationship between income and healthy food is confounded by knowledge? [25]

- (A) Yes
- (B) No

Finally, interpret the output. Write down the answers to the following questions:

1. How well can we predict the variance in healthy eating with the predictors income and knowledge? Interpret R2, report the appropriate test and its significance
2. Interpret the regression coefficients (size, direction, significance)
3. Which predictor is the most important predictor of healthy eating behavior? Inspect the standardized regression coefficients

Answer

Income and Knowledge together predict 44.9% of the variance in Eating healthy food, which is significantly different from zero, $R^2 = .45, F(2, 347) = 141.179, p < .001$.
Controlling for Knowledge, Income has a positive, but non-significant effect on Eating healthy food, t(347) = .283, p = .778.
Controlling for Income, Knowledge has a positive, significant effect on Eating healthy food, t(347) = 15.287, p <.001.
Knowledge is the most important predictor (= .665) (compare it with = .012 of Income).

[23] 0.778
[24] 0.001
[25] Yes

16 GLM-VI: Nested models

Nested models refer to models that are identical, except for the fact that some parameters are constrained to zero in one of them, while all parameters are free in the other. The smaller or constrained model model is said to be "nested in" the larger or unconstrained model, meaning that all predictors in the smaller model are also present in the larger model. By definition, the unconstrained model always provides a better fit than the constrained model.

Incremental F-tests are used to determine whether the increase in R^2 between two nested models is statistically significant. Recall that R^2 represents the proportion of variance in the dependent variable explained by the predictors in the model. The incremental F-test compares the variation in the dependent variable explained by an unconstrained model with the variation explained by a constrained model. The F-test assesses whether the increase in R-squared is greater than what would be expected by chance alone, indicating the significance of adding additional predictors to the model.

Hierarchical regression refers to an approach where predictors are added to a regression model in blocks, allowing us to assess the additional variance explained by each block of predictors with an incremental F-test. Each block represents a set of predictors, and they are added to the model in a stepwise manner. This approach is useful when we want to determine the significance of adding a single categorical predictor which is represented by multiple dummy variables, or when we want to test whether adding some predictors (e.g., theoretically relevant variables) explain significant variance above and beyond other variables (e.g., demographic covariates). By conducting incremental F-tests at each step, we can assess the significance of adding each block of predictors and their contribution to the overall model.

In summary, nested models allow us to compare models with different levels of complexity, incremental F-tests help us determine the significance of adding predictors to a model, and hierarchical regression enables the examination of additional variance explained by different blocks of predictors. These concepts are valuable tools in statistical analysis and can provide insights into the relationships between variables and the predictive power of a model.

16.1 Lecture

https://www.youtube.com/embed/s15b3KvfnO4

16.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

Which of the following statements is true about nested models? [1]

- (A) Model A is nested in Model B if they are identical, except some parameters in model A are constrained.
- (B) Model A is nested in Model B if they are identical, except Model A uses a subsample of the sample used to estimate Model B.
- (C) Nested models always have the same number of predictors.
- (D) Model A is nested in Model B if they are identical, except some parameters in model B are constrained.

Question 2

What can we say about the relative fit of Model A and Model B if Model A is nested in Model B? [2]

- (A) Model A will always have better fit than Model B.
- (B) Nested models always have equal fit.
- (C) Model B will always have better fit than Model A.
- (D) In nested models, one model always has better fit than the other.

Question 3

In the context of nested models, what does 'constrained' mean? [3]

- (A) The set of predictors is constrained.

[1] Model A is nested in Model B if they are identical, except some parameters in model A are constrained.
[2] Model B will always have better fit than Model A.
[3] Some parameters are fixed to 0.

- (B) The R-squared value of the smaller model is constrained to be lower than that of the larger model.
- (C) Some variables are fixed to 0.
- (D) Some parameters are fixed to 0.

Question 4

What is the purpose of performing a nested model test? [4]

- (A) To compare two different models.
- (B) To determine whether adding specific predictors significantly improves the model's fit.
- (C) To determine if the models are independent of each other.
- (D) To check if the predictors are correlated with each other.

Question 5

Given the nested models: Y = a + b1*X1 and* $Y = a + b1$X1 + b2*X2, which model is the 'constrained' one? [5]

- (A) Y = a + b1*X1 + b2*X2
- (B) Y = a + b1*X1
- (C) It cannot be determined from the information given.
- (D) Neither is constrained.

Question 6

How does the F-test for the R-squared of a single model relate to the F-test for the delta R-squared of two nested models? [6]

- (A) Both are F-tests; these are ratios comparing the explained variance to the unexplained variance.

[4] To determine whether adding specific predictors significantly improves the model's fit.
[5] Y = a + b1*X1
[6] Both can be seen as delta R-squared tests; the former compares the model of interest to a model with only an intercept, which is a nested model test.

- (B) Both can be seen as R-squared tests; both compare the model of interest to a model with only an intercept, which is a nested model test.
- (C) The similarity is only superficial; although both are F-tests, the calculation is different.
- (D) Both can be seen as delta R-squared tests; the former compares the model of interest to a model with only an intercept, which is a nested model test.

Question 7

What can we conclude from an incremental F-test in nested models? [7]

- (A) Whether the smaller model is significantly better than the larger model.
- (B) Whether adding specific predictors significantly improves the model's fit.
- (C) Whether the predictors of the nested models are significantly different.
- (D) Whether the larger set of predictors has significant multicollinearity.

Question 8

When performing an incremental F-test, what does the numerator in the F ratio represent? [8]

- (A) The increase in residual mean square when adding predictors to the model.
- (B) The reduction in residual mean square when adding predictors to the model.
- (C) The reduction in the number of predictors from the larger to the smaller model.
- (D) The increase in the number of degrees of freedom when adding predictors to the model.

Question 9

Which scenario is suitable for using hierarchical regression? [9]

- (A) To determine if theoretically relevant factors explain variance beyond demographic characteristics.

[7] Whether adding specific predictors significantly improves the model's fit.
[8] The reduction in residual mean square when adding predictors to the model.
[9] To determine if theoretically relevant factors explain variance beyond demographic characteristics.

- (B) To test whether each dummy of a categorical predictor with >2 categories explains more variance than the other dummies.
- (C) To compare two unrelated models.
- (D) To check the multicollinearity of the predictors.

Question 10

Given two nested models, one has regression sum of squares = 283.96 and residual sum of squares = 958.58; the other has regression sum of squares = 202.76 and residual sum of squares = 1039.78. What is the delta R-squared between these models? [10]

- (A) .065
- (B) .085
- (C) .078
- (D) Can't say with this information

Question 11

In a nested model test, numerator degrees of freedom for the F ratio is 4, how many parameters were added to the model? [11]

- (A) 3
- (B) 4
- (C) Depends on the model
- (D) 5

Question 12

In a nested model test, one model has regression sum of squares = 202.76 and 3 parameters and residual sum of squares = 1039.78, the other model has regression sum of squares = 283.96 and 4 parameters and residual sum of squares = 958.58. There are 387 participants. What is the F-value for the nested model test? [12]

- (A) 24.90

[10] .065
[11] 4
[12] 32.44

- (B) 32.44
- (C) Can't say based on this information.
- (D) 28.29

Show explanations

Question 1
Nested models are identical except for some constrained parameters.
Question 2
The bigger model always has a better fit than the smaller model in a nested pair.
Question 3
In the context of nested models, a 'constrained' model has some parameters fixed to 0.
Question 4
A nested model test is used to determine whether adding specific predictors significantly improves the model's fit.
Question 5
The model Y = a + b1*X1 is the 'smaller' or 'constrained' model.
Question 6
Both F-tests can be seen as nested model tests; the F-test for the R-squared of a single model tests against the variance explained by a model including only the intercept.
Question 7
The incremental F-test determines if adding specific predictors significantly improves the model's fit.
Question 8
The numerator in the F ratio of an incremental F-test represents the reduction in residual mean square when adding predictors to the model.
Question 9
Hierarchical regression is suitable for determining if theoretically relevant factors explain variance beyond demographic characteristics.
Question 10
The delta R-squared is the difference in regression sums of squares, divided by the total sum of squares. The latter can be calculated by adding the error sum of squares to the regression sum of squares for either model.
Question 11
The numerator degrees of freedom is equal to the difference in number of parameters.
Question 12
The F-value is obtained by first calculating the regression mean square: dividing the difference in regression sum of squares by the difference in model DF: (283.957-202.758)/1, and calculating the residual mean square for the larger model: 958.58/(387-4). The F-value is then ((283.957-202.758)/1)/(958.58/(387-4)).

16.3 In SPSS

16.3.1 Multiple Regression

https://www.youtube.com/watch?v=pkM4nXvP5Bo

16.4 Tutorial

16.4.1 Multiple Regression

This assignment revolves around multiple regression with more than two predictors.

Use the data file called `Work.sav`. These data were about work characteristics.

Open the date file in SPSS to get started!

Consider the following research question(s):

"To what extent do variety at work, learning possibilities, and independence at work explain pleasure at work, and which of these explanatory characteristics is most important?"

What are the independent variables in this research question?

[13]

- (A) variety, learning possibilities, independence, pleasure
- (B) variety
- (C) pleasure
- (D) variety, learning possibilities, independence

Let's now run a multiple regression analysis to get the output we need to answer our research question.

Navigate to Analyze > Regression > Linear. Enter the dependent variable (pleasure at work). Enter the independent variables (variety at work, learning possibilities, independence at work). Paste and run the syntax.

In the next steps we will go over the output from this analysis.

What percentage of the total variance in work pleasure is explained by variety at work, learning possibilities, and independence at work altogether? _____[14]

[13] variety, learning possibilities, independence
[14] 26.9

Which of the following statements correctly summarizes the F-test for the R2 in this analysis? [15]

- (A) R-square is significant, F(3, 121) = 14.455, p < .001
- (B) R-square is significant, F(3, 118) = 14.455, p < .001
- (C) R-square is not significant, F(3, 118) = 14.455, p < .001
- (D) R-square is not significant, F(3, 121) = 14.455, p < .001

Consider the "Coefficients" table.

What is the unstandardized regression coefficient for the effect of variety at work on work pleasure? _____ [16]

Describe the effect of variety at work on work pleasure as precise as possible.

In other words, how should we interpret the unstandardized regression coefficient for variety at work?

Answer

When variety at work increases with one unit and the other two variables stay constant, pleasure at work increases with .272 units.

Now, consider the effect of independence at work on work pleasure.

What is the standardized regression coefficient of the effect of independence at work on work pleasure? _____ [17]

Describe the effect the effect of independence at work on work pleasure as precise as possible using the standardized regression coefficient.

In other words, how should we interpret the standardized coefficient for this predictor?

Answer

When independence at work increases with one SD and the other two variables stay constant, pleasure at work increases with .107 SDs.

Which of the predictors has a significant partial effect on pleasure at work when using =.10? [18]

[15] R-square is significant, F(3, 118) = 14.455, p < .001
[16] 0.272
[17] 0.107
[18] Variety at work and Learning possibilities

- (A) Variety at work
- (B) Variety at work and Learning possibilities
- (C) Independence at work
- (D) None
- (E) Learning possibilities

In the final couple of steps we went through the output displayed in the "Coefficients" table. Before you proceed with the next assignment, please review the following aspects in the answers your gave:

- When you wrote down the effect of variety at work on pleasure at work in unstandardized form, did you include that it is the effect of the IV on the DV controlled for the other two variables? If not, please keep in mind that this is very important! In multiple regression, all effects are partial (unique) effects, controlling for the other predictors.
- When you wrote down the effect of independence at work on pleasure at work in standardized form, did you indicate that it is the effect in standard deviations, and, again, did you include that it is the effect of IV on the DV controlled for the other two variables (i.e., 1 SD change in independent variable independence at work leads to a .107 SDs change in the predicted score for pleasure at work, controlled for the other variables)?

16.4.2 Unique Contributions

Use the data file called `Work.sav`. These data were about work characteristics.

In this assignment, we will look more closely at how to evaluate the "added value" of one predictor in addition to the other predictors in an analysis explaining the dependent variable. Or, put the other way around, how much we lose if we would remove a certain predictor from the model.

We will again consider the analysis in which we predicted pleasure at work by the three predictors learning possibilities, independence at work, and variety at work.

We want to know how much the variable variety at work adds to predicting work pleasure.

Run the two analyses specified below:

Run the regression analysis using all three predictors.

Run the regression analysis analysis with only learning possibilities and independence at work as predictors.

Inspect the R2 of both regression analyses.

What happened with the R2 when variety was removed?

Answer

With all three predictors included, R-square equals .269 (see slide 3) With Variety at Work removed, R-square equals .248. So, R-square decreased with .021, indicating that Variety at Work uniquely explains 2.1% of the total variance in Pleasure at Work.

Let's do the same for the other two predictors; that is, check how much the R2 changes if you would remove the predictor from the model.

Write down the changes below.

Answer

With Learning Possibilities removed, R square equals .211. R-square decreased with .058, which means that Learning Possibilities uniquely explains 5.8% of the total variance in Pleasure at Work. With Independence at Work removed, R square equals .260. R-square decreased with .009, which means that Independence at Work uniquely explains 0.9% of the total variance in Pleasure at Work.

Which predictor explained the least unique variance, controlling for the other two? [19]

- (A) Independence at work
- (B) Learning possibilities
- (C) Variety at work

16.4.3 Hierarchical Regression Analysis

In this assignment we will carry out a hierarchical regression analysis, which is a structured way to compare nested models in linear regression analysis.

With hierarchical regression analysis we are able to compare two (or more) nested models. Consider the model below:

$Y = b0 + b1X1 + b2X2 + b4X4 + b5X5 + b7X7$

Which of the following models is/are nested within the larger model presented above?

[20]

[19] Independence at work
[20] $Y = b0 + b1X1 + b2X2 + b4X4$

- (A) $Y = b0 + b1X1 + b3X3 + b4X4$
- (B) $Y = b0 + b1X1 + b2X2 + b4X4$
- (C) $Y = b0 + b1X1 + b2X2 + b6X6$

We will now get back to our data on work characteristics using the data file `Work.sav`.

Consider the following situation in which researchers are interested in clusters of variables:

Researchers are interested in the effect of "Health" (i.e., mental pressure, physical demands, and emotional pressure) and "Social Environment" (i.e., relationship with coworkers, and relationship with supervisors) on the Need for recovery.

We will address this research question in a number of steps.

First, we will look at the full model (i.e., the model that has all health and social environment predictors).

Run the regression analysis in which you regress Need for recovery on the three health variables (i.e., mental pressure, physical demands, and emotional pressure) and the two social environment variables (i.e., relationship with coworkers, and relationship with supervisors).

What is the R2 of the full model? _____[21]

Now we want to test the significance of the clusters of variables. We will first look at the Health variables (emotional pressure, physical demands, & mental pressure). We want to test whether Health has a direct effect on Need for recovery controlled for the social environment predictors.

To accomplish this, we have to do a hierarchical regression analysis, in which we compare two nested models: a small one, and a larger one.

Write down what these two models look like.

How many predictors does the smaller model contain? _[22]

How many predictors does the larger model contain? _[23]

[21] 0.22
[22] 2
[23] 5

Answer

- Small model: $Recover_i = b0 + b1\ \ sccowork_i + b2\ \ scsuperv_i + e_i$
- Large model: $Recover_i = b0 + b1\ \ sccowork_i + b2\ \ scsuperv_i + b3\ \ scmental_i + b4\ \ scphys_i + b5\ \ scemoti_i + e_i$

The small model has 2 predictors and the large model has 5 predictors. The large model has three predictors more than the small model.

Now run the hierarchical regression analysis in SPSS!

Navigate to Analyze > Regression > Linear

Click on "Reset" to start from scratch.

Select Need for Recovery as the dependent variable.

Select only the Social Environment variables (i.e., Relationship with coworkers and Relationship with supervisors) as the independent variables (this is the small model). This is block 1 of 1.

Click on "Next". Select the three Health variables (i.e., emotional pressure, physical demands, and mental pressure) and add those to this second block.

Important: Now click on the button "Statistics" to request the R2 change.

Paste and run the syntax. Note the following new elements:

```
/STATISTICS COEFF OUTS R ANOVA CHANGE
/METHOD=ENTER sccowork scsuperv
/METHOD=ENTER scmental scphys scemoti.
```

The `CHANGE` command asks for R^2 change statistics, and the two rows of `/METHOD=ENTER` sequentially add blocks of predictors to the model (enter them into the model).

Let's inspect the table labeled "Model Summary" first.

What percentage of the total variance in Need for recovery is explained by the small model? _____[24]

How much of the total variance in Need for recovery is explained by the large model? ___[25]

What is the R2-change from Model 1 to Model 2? _____[26]

The Model Summary table also reports the results of the F-tests, which tests whether the change in R2 is significant.

[24] 10.7
[25] 22
[26] 11.3

Write down the null and alternative hypothesis that are evaluated by these F-tests, then check your answer?

Answer

$H_0 \;\; R^2 = 0$
$H_0 \;\; R^2 \;\; 0$

Suppose three researchers use the output to see whether the R2-change is significant. Researcher I tests at the 10% level, Researcher II tests at the 5% level, and Researcher III tests at the 1% level.

Which of the researchers will conclude that there is a significant result? [27]

- (A) All three researchers
- (B) Only researcher I
- (C) Only researcher II
- (D) Only researcher III

Suppose the researchers summarized the result as follows:

"Health variables explain 11.3% of the variance in need for recovery." [28]

- (A) Yes
- (B) No

Explanation

The correct interpretation is:
"The health variables explain an additional 11.3% of the total variance in Need to recover, on top of what is already explained by the social environment variables."
Alternatively, you may summarize the result as:
"The health variables uniquely explain 11.3% of the total variance in Need to recover, while controlling for the social environment variables."
Since the R2-change was significant, we have evidence that "Health" had a direct effect on the Need to recover.

Now test whether the social environment variables have an effect on need to recover (while controlling for the health variables). Summarize the results (include all relevant statistics: test statistics, degrees of freedom, p-values).

[27] All three researchers
[28] No

Include in your answer:

The R2 of the Small and Full model.

The R2-change and whether or not it is significant (use a significance level of .05).

A substantive interpretation of the R2-change (within the context of this assignment).

The first step is to run the hierarchical regression analysis, with the health variables in Block 1 and the social environment variables in Block 2.

Answer

The model with social environment variables explains 14.2% of the total variance in Need for Recovery. Health variables explain an additional 7.8% of variance in Need for Recovery, and this difference significantly differs from zero, $R^2 = .08, F(2, 116) = 5.800, p = .004$. This means that the social environment variables uniquely explain 7.8% of the total variance in Need for Recovery, while controlling for the health variables. The total explained variance in Need for Recovery is 22.0%, which is significantly different from zero, $F(5, 116) = 6.536, p < .001$.

16.4.4 Dummies and Continuous Predictors

In this assignment we will predict a continuous outcome variable from a dichotomous predictor (i.e. gender).

The data file that we will use is PublicParticipation.sav.

This data file contains data on the following variables:

- income (higher score = higher income)
- public participation (i.e. being a member of school boards, municipal councilor, etc.)
- education
- age
- gender (0 = females; 1 = males)

We want to test the following research question:

"Are there gender differences in Public Participation?"

To answer this research questions, we can use an independent samples t-test.

Run an independent samples t-test to answer the research question (consult the information button in case you forgot how to run an independent samples t-test.)

Consult the output to answer the questions in the next steps.

Analyze > Compare means > Independent samples t-test.

Choose Public Participation as the dependent variable.

The grouping variable is Gender. Define the groups: group 1 = 0 (i.e., women) and group 2 = 1 (i.e., men).

Paste and run the syntax.

What is the mean difference between men and women? _____[29]

Consult the table Group Statistics.

"In the sample, men score on average _____[30] points higher on public participation than women."

Which of the following statements correctly summarizes the results of Levene's test (use =.05)?

[31]

- (A) Levene's test is not significant, no evidence against the assumption of homoscedasticity
- (B) Levene's test is significant, evidence against assumption of homoscedasticity.

The researchers conclude:

"We have convincing evidence that the population means of public participation differs between males and females."

Is this a valid conclusion? [32]

- (A) Yes
- (B) No

Report the test results, then check your answer.

Explanation

The mean level of public participation differs between males ($M = 16.54$) and females ($M = 9.95$), $t(41) = 4.942, p < .001$.

What is the (absolute) value of the t-statistic? _____[33]

---

[29] 6.502
[30] 6.5
[31] Levene's test is not significant, no evidence against the assumption of homoscedasticity
[32] Yes
[33] 4.942

From the results of the t-test the researchers would conclude that males on average have higher levels of public participation than females.

But, of course, we don't know for sure, because we did not test the entire population. Which error type could the researchers have made? [34]

- (A) Type I error
- (B) Type II error
- (C) Type I or Type II error

16.4.4.1 Examining Income

In the previous steps, we used the independent sample's t-test to test for mean differences in public participation and found a significant result. Results suggested that males show higher levels of public participation. However, males and females may differ not only in gender, but also in other characteristics that explain differences in public participation. In particular, the researchers hypothesize that income differences may play a role as well. Research has shown that income is positively associated with public participation. So if males and females differ in income (on average), the gender differences in public participation may be partly due to income differences.

In other words, according to the researchers income may be a [35]

- (A) moderator
- (B) collider
- (C) confounder
- (D) predictor

of the effect of gender.

Before we proceed, let's first check whether the groups differ on income. Request the means for the variable income for both males and females separately.

Do the males and females in the sample differ in average income?

Take your existing syntax for gender differences in Participation, and change it to test for income differences instead:

[34] Type I error
[35] confounder

```
T-TEST GROUPS=Gender(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Income
  /ES DISPLAY(TRUE)
  /CRITERIA=CI(.95).
```

Is there a difference in income? [36]

- (A) Yes
- (B) No

The next step is to study gender differences controlled for income differences. This is not possible via the independent samples t-test interface, but it is possible when specifying our model using regression analyses. Regression analysis allows us to study gender differences controlling for other variables.

16.4.4.2 Add continuous predictor

Before we proceed with the regression analysis including income (we will do this in the next assignment), we will first see what the regression model looks like if we would only use gender as the only predictor in the model.

Run a regression analysis using Public Participation as the dependent variable, and gender as the independent variable (Analyze > regression > linear). Note that the variable gender is already dummy coded.

Inspect the table with Coefficients.

What is the regression slope of gender? _____ [37]

Compare the value to the mean difference from ; what do you see?

Recall that the independent samples t-test is the exact same test as the t-test for the regression slope of a dummy variable.

Compare the value of the t-statistic in the regression with the value of t-statistic in the independent samples t-test (you may ignore the minus signs). What do you see?

The regression coefficient of gender is 6.502. This is the same as the difference in means of Public Participation between males and females. The t-statistic of the regression coefficient is 4.942. The t-statistic in the independent samples t-test was -4.942. As we can ignore the

[36] Yes
[37] 6.502

minus sign (just make sure to check which group is higher than the other), we can conclude they are equal.

What percentage of the total variance in public participation is explained by gender? _____[38]

Recall that the independent samples t-test and regression analysis with a dummy variable predictor are completely equivalent; SPSS just offers different interfaces to the same linear model.

We need to use the regression interface when we want to add other predictors (e.g., income) to see if the gender differences are confounded by other characteristics, and we can calculate the R2 value to evaluate the size of the effect of gender.

Next, we want to know whether there are differences in public participation between men and women with the same income level.

In other words: let's look at gender differences in public participation controlled for income.

Run a regression analysis of public participation on income and gender (analyze -> regression -> linear). Consult the output and answer the following questions:

Gender and income jointly explain _____[39]% of the variance in public participation.

Which of the following alternatives correctly report the F-statistic testing the significance of the R2? [40]

- (A) $F(2,40)=26.845$, $p < 0.001$
- (B) $F(2,42)=26.845$, $p < 0.001$

The researchers conclude:

"On average, women score 4.818 points higher on public participation than men."

Is this a valid conclusion? [41]

- (A) No
- (B) Yes

[38] 37.3
[39] 57.3
[40] $F(2,40)=26.845$, $p < 0.001$
[41] No

Explanation

I hope you concluded that the researchers' conclusion is incorrect.
They should have said:
"On average, women score 4.818 points less on public participation than men of an equal income level"
So, if we take two groups of men and women with the same income (= "controlling for income"), then we expect a mean difference of 4.818 in public participation.

Is there evidence of gender differences, while controlling for income differences when testing at = .01? [42]

- (A) Yes
- (B) No

Write down the complete regression equation, then check your answer below.

Answer

The regression equation is as follows:
$\text{Participation}_i = 5.762 + 3.475\ \text{Income}_i + 4.818\ \text{Gender}_i + e_i$

Suppose we have a man with an income of 3.0 and a woman with an income of 2.0, what is the predicted difference in public participation? _____ [43]

Explanation

Man with an income of 3 has a predicted score of:
Public Participation=5.762+3.4753+4.818 1
Public Participation=21.005
Woman with an income of 2 has a predicted score of:
Public Participation=5.762+3.4752+4.818 0
Public Participation=12.712
Difference = 21,005 – 12,712
Difference = 8.293

We previously computed the mean difference in public participation for men and women, which was 6.50.

Why is this previous mean difference not the same as the effect of gender that we find in this most recent regression analysis?

[42] Yes
[43] 8.293

Answer

The mean difference as observed in the sample is the mean difference without controlling for income differences.
However, as we have seen, men and women also differ in the average income, and income differences between males and females may also explain differences in public participation. In the regression analyses, we controlled for income differences, which means that we "filtered out" the effect of income (the correct way to say it: we partialed out income effects). After controlling for income the effect of gender is somewhat smaller. Hence, the difference in public participation in the sample means is partly due to income differences.

One researcher argues that there are gender differences in public participation, and gender differences in income, and that income additionally has an effect on public participation.

In this theoretical model, income is a [44]

- (A) Mediator
- (B) Confounder
- (C) Moderator
- (D) Common cause

16.4.5 Nested Models

Finally, we might want to know how much unique variance is explained by gender, *after* controlling for income. To this end, we would perform a nested model test.

Conduct this test as you have done before, using the graphical interface or syntax. Remember: Enter Income first, then Gender - and ask for the R^2 change statistics.

Let's inspect the Model Summary table.

What percentage of the total variance in Participation is explained by Income (only)? _____[45]

How much of the total variance in Participation is explained by both Income and Gender? _____[46]

What is the R2-change (in percentage) from Model 1 to Model 2? _____[47]

[44] Mediator
[45] 39.1
[46] 57.3
[47] 18.2

The Model Summary table also reports the results of the F-tests, which tests whether the change in R2 is significant.

Suppose three researchers use the output to see whether the R2-change is significant. Researcher I tests at the 10% level, Researcher II tests at the 5% level, and Researcher III tests at the 1% level.

Which of the researchers will conclude that there is a significant result? [48]

- (A) All three researchers
- (B) Only researcher II
- (C) Only researcher I
- (D) Only researcher III

Report the results of your analysis (include all relevant statistics: test statistics, degrees of freedom, p-values).

Include in your answer:

- The R2 of the Small and Full model.
- The R2-change and whether or not it is significant (use a significance level of .05).
- A substantive interpretation of the R2-change (within the context of this assignment).

Answer

Income explains 39.1% of variance in Participation, which is significantly greater than zero, $F(1, 41) = 26.32, p < .001$. Gender explains an additional 18.2% of variance in Participation, and this difference significantly differs from zero, $R^2 = .18, F(1, 40) = 17.06, p < .001$. The total explained variance in Participation by Income and Gender is 55.2%, which is significantly different from zero, $F(2, 40) = 26.85, p < .001$.

16.4.6 One more Categorical Variable

Let's finish this assignment by adding one more categorical variable, and seeing whether it has a significant unique effect after controlling for Income and Gender. We will use Education level, which has three categories. We thus have to create two dummy variables to account for this variable.

First, make the dummy variables, using low Education status as reference category.

[48] All three researchers

Remember that these dummies together code for one variable, so you must add them to the model in the same step together.

What is the F test statistic for the R^2 test for adding Education to a model that already contains Income and Gender? _____[49]

What is the difference in expected Participation between someone with Low education status and High education status? _____[50]

[49] 0.352
[50] -1.242

17 GLM-VII: Interaction

The foundation of the regression model is the equation that describes the relationship between two variables. In a simple bivariate linear regression, we use the equation:

$Y_i = a + b\ X_i + e_i$

Where:

- Y_i represents an individual's score on the dependent variable Y.
- a is the intercept coefficient of the regression line.
- b is the slope coefficient of the regression line.
- X_i is an individual's score on the independent variable X.
- e_i is the prediction error for individual i.

The regression line provides us with predicted values based on the model. It can be represented as:

$Y_i = a + b\ X_i$

Where Y^i is the predicted score for individual i on the dependent variable Y.

17.0.1 Introducing Interactions

Now, let's take our regression analysis to the next level by introducing the concept of interactions. An interaction implies that the effect of one predictor variable depends on the level of another predictor variable.

To incorporate interactions, we add a special building block to our regression equation:

$Y = a + b_1X_1 + b_2X_2 + b_3(X_1 \text{Œ} X_2)$

To include an interaction term in your regression model:

1. Calculate a new variable as the product of the two interacting variables.
2. Add this interaction term to the regression equation along with the original variables. **Note:** You should never include an interaction term in the model without including its constituent terms!

In summary, interactions add a new layer of complexity to regression analysis by acknowledging that the relationships between variables can be contingent on other factors. By incorporating interactions, we can gain deeper insights into the nuanced dynamics between predictors and outcomes in our statistical models.

17.0.2 Interaction between one Continuous and one Binary Predictor

Interaction is easiest to explain using one binary predictor, coded as 0 and 1, and one continuous predictor.

Recall that using this type of "dummy coding" allows us to represent different intercepts for the two groups in our regression model. The intercept of the regression equation applies to the group coded 0 (the reference group), and the intercept of the group coded 1 is equal to the overall intercept plus the effect of the dummy variable. In formulas this is represented as:

$$Y = a + bD, \text{ where D } (0, 1) \tag{17.1}$$
$$Y_{D=0} = a + b\ 0 = a \tag{17.2}$$
$$Y_{D=1} = a + b\ 1 = a + b \tag{17.3}$$

An interaction implies that not only the intercept but also the regression slope differs between two groups. Imagine we add continuous predictor X to the model, as well as the interaction term $D\ X$:

$$Y = a + b_1 D + b_2 X + b_3 (D\ X) \tag{17.4}$$

In some cases, we might want to estimate not only distinct intercepts but also distinct slopes for different groups. For instance, we might be interested in understanding how the effect of gender role attitudes on involvement differs for men and women.

The complete formula for the regression line varies based the value of dummy variable D:

$$Y_{D=0} = a + b_1\ 0 + b_2 X + b_3 (0\ X) = a + b_2 X \tag{17.5}$$
$$Y_{D=1} = a + b_1\ 1 + b_2 X + b_3 (1\ X) = (a + b_1) + (b_2 + b_3) X \tag{17.6}$$

Note that both groups' regression equations can be simplified into a basic linear formula of the form $a + bX$, except that they each have a unique value for the intercept and regression slope! This is how interaction terms allow you to make the effect of one variable contingent on the value of another.

If the interaction term is significant (i.e., the slope for the product of the interacting variables), we conclude that there is significant interaction and the slope of the effect of one interacting variable depends on the value of the other interacting variable.

17.0.3 Simple Effects

When the interaction between a binary moderator and a continuous predictor is significant, we often want to know how big the regression effect of the continuous predictor is in each group of the binary predictor. This is called simple effects analysis.

One straightforward way to perform simple effects analysis involves creating dummy variables for both categories of the binary moderator and computing interaction terms with these dummies. Then, specifying two regression models with different reference categories. This gives the effect of the continuous moderator for each group, along with a significance test.

17.0.4 Interaction with Two Continuous Predictors

When we previously discussed interaction effects involving one binary and one continuous predictor, we discovered that such interactions result in distinct regression lines for each unique value of the binary predictor. Now, consider an interaction between two continuous predictors. In this scenario, each variable can theoretically take infinite possible values. Therefore, we can no longer think of this as a distinct regression line for each value of the moderator. Instead, we can imagine how an increase in the value of one interacting variable leads to an adjustment of the effect of the other interacting variable.

To grasp the concept of interaction effects with two continuous predictors, let's dive into a concrete example. Imagine we're investigating the relationship between outcome Y, and continuous predictors X1 and X2. We've determined the coefficients for our regression model:

$Y = 12.50 + 1.50\ X_1\ 0.20\ X_2 + 0.07\ (X_1 \text{Œ}\, X_2)$

Let's do the same we did for understanding the regression equation with a binary moderator, and simplify it by plugging in a specific value for one of the interacting variables. Suppose we want to know the effect of X1 for someone who scores 0 on the continuous variable X2:

$Y = 12.50 + 1.50\ X_1\ 0.20\ 0 + 0.07\ (X_1 \text{Œ}\, 0) = 12.50 + \mathbf{1.50}\ X_1$

The effect for such a person is 1.50. Now, let's compare this to a person who scores 1 on the continuous variable X2:

$Y = 12.50 + 1.50\, X_1\, 0.20\, 1 + 0.07\, (X_1 \text{Œ} 1) = (12.50\, 0.20) + (1.50 + 0.07)\, X_1 = 12.30 + \mathbf{1.57}\, X_1$

Now, the effect of X_1 has increased by 0.07 - which was exactly the size of the regression slope for the interaction term.

17.0.5 Centering for Interpretability

When working with interactions between two continuous predictors, it's essential to center the variables. Centering aids interpretability - the effect of one predictor is now given for the average value of the other predictors. Moreover, centering avoids artificial multicollinearity between the two interacting variables and their interaction term.

17.0.5.1 Simple Slopes

If the interaction effect between two continuous predictors is significant, we might want to understand how the effect of one of the interacting predictors varies across levels of the other interacting predictor. This is similar to the simple effects approach from before, except now it's called simple slopes.

Instead of computing the effect of one variable for all unique values of a binary moderator, we pick specific values of the continuous moderator - typically +/- 1SD - and calculate the effect of the other predictor at those specific values.

By centering the interacting predictors at their mean value +/- 1SD and re-computing the interaction term using those transformed predictors, we obtain simple slopes at different levels of the moderator. Note that centering at $M + 1SD$ gives us the effect for people who score 1SD **below** the mean (you're sliding the distribution to the right on the number line, until people who used to score -1SD are centered at 0).

17.1 Lecture

https://www.youtube.com/embed/fGj_t72VJPk

17.2 Formative Test

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

An interaction effect is when... [1]

- (A) The effect of one predictor is added to the effect of another predictor.

[1] The effect of one predictor depends on the value of another predictor.

- (B) The effect of two predictors is cumulative.
- (C) You interact with your partipants.
- (D) The effect of one predictor depends on the value of another predictor.

Question 2

What is the purpose of centering continuous predictors in interaction analysis? [2]

- (A) Centering continuous predictors increases interpretability.
- (B) Centering continuous predictors is necessary for regression analysis to provide correct results.
- (C) Centering continuous predictors helps in enhancing interpretability and reducing multicollinearity.
- (D) Centering continuous predictors prevents multicollinearity.

Question 3

When two continuous predictors interact, how many unique regression lines are generated? [3]

- (A) Only one unique regression line is generated regardless of the interaction.
- (B) The number of unique regression lines depends on the number of unique values of the predictors.
- (C) A theoretically infinite number of unique regression lines, as both predictors can take on an infinite number of values.
- (D) Two unique regression lines are generated, one for each predictor.

Question 4

What is the concept of simple slopes in interaction analysis? [4]

- (A) Simple slopes are the slopes of predictors before they are centered.

[2] Centering continuous predictors helps in enhancing interpretability and reducing multicollinearity.
[3] A theoretically infinite number of unique regression lines, as both predictors can take on an infinite number of values.
[4] Simple slopes refer to the effect of one predictor variable, evaluated at specific levels of another variable.

- (B) Simple slopes refer to the effect of one predictor variable, evaluated at specific levels of another variable.
- (C) Simple slopes refer to the linear effects of a predictor on the outcome variable.
- (D) Simple slopes are the partial effects of two predictors, controlling for their interaction effect.

Question 5

In a multiple regression model with two continuous predictors, how would you assess the effect of one predictor at different levels of the other predictor? [5]

- (A) By standardizing the predictors.
- (B) By excluding the interaction term from the model.
- (C) By centering one predictor at a specific level and re-computing the interaction term.
- (D) By computing the mean of both predictors.

Question 6

What does it mean when an interaction term in a regression model is not significant? [6]

- (A) An insignificant interaction term indicates that the model is overfitting.
- (B) An insignificant interaction term indicates a problem with the data collection process.
- (C) An insignificant interaction term means that the outcome variable is not related to the predictor variables.
- (D) When an interaction term is not significant, it suggests that the effect of one predictor variable on the outcome variable is consistent across all levels of the other predictor variable.

Question 7

What is the predicted value for an individual with a score of 2.5 on X1 and score of 35 on X2, given the following regression equation: Y = 10 + 2 * X1 - 0.5 * X2 + 0.1 * (X1 * X2)? [7]

[5] By centering one predictor at a specific level and re-computing the interaction term.
[6] When an interaction term is not significant, it suggests that the effect of one predictor variable on the outcome variable is consistent across all levels of the other predictor variable.
[7] 6.25

- (A) -3.75
- (B) 8.75
- (C) -2.4
- (D) 6.25

Question 8

What is the effect of a one-unit increase in X1 on the predicted value for an individual who scores 40 on X2, given the following regression equation: Y = 8 + 1.2 * X1 - 0.3 * X2 + 0.05 * (X1 * X2)? [8]

- (A) 0.1
- (B) 1.2
- (C) 3.2
- (D) 1.5

Question 9

Assume that X1 and X2 are centered around the mean. Given the following regression equation: Y = 6 + 1.5 * X1 - 0.4 * X2 + 0.08 * (X1 * X2), what is the simple slope of X1 on Y for people who score 1SD above the mean on X2, if the SD of X2 is 0.5? [9]

- (A) 1.5
- (B) 1.54
- (C) 1.58
- (D) 1.46

[8] 3.2
[9] 1.54

Show explanations

Question 1
An interaction effect involves the combined influence of two or more predictor variables on the outcome variable, which is different from their individual effects.
Question 2
The two purposes of centering are to avoid artificial multicollinearity between the interacting variables and their product, and aids the model's interpretability.
Question 3
In interactions involving continuous predictors, the relationship between the predictors and the outcome variable can vary infinitely, leading to an infinite number of possible regression lines.
Question 4
Simple slopes allow us to assess how the relationship between two predictors changes at different levels of the moderator variable, helping to understand conditional effects.
Question 5
Centering and re-computing the interaction term allows us to obtain the slopes of the predictor of interest at different levels of the moderator, helping us understand its conditional effects.
Question 6
When an interaction term is not significant, it implies that the relationship between the predictors and the outcome remains relatively constant regardless of the values of the interacting predictors.
Question 7
Plug in the values: Y = 10 + 2 * 2.5 - 0.5 * 35 + 0.1 * (2.5 * 35)
Question 8
You have to add the interaction term to the effect of X1; when X2 has the value 0, the effect of X1 is 1.2. When X2 has the value 40, add 0.05*40 to that effect.
Question 9
Calculating the simple slope works just the same as calculating the effect of X1 for specific values of X2, so you calculate the effect for +1 SD of .05: $1.5X1 + 0.08.5X1 = 1.54$X1.

17.3 In SPSS

17.3.1 Multiple Regression

https://www.youtube.com/watch?v=l3Aoikhaxtg

https://www.youtube.com/watch?v=aeT8MkG3bx8

https://www.youtube.com/watch?v=aVV7KnAr-qY

https://www.youtube.com/watch?v=vYsjJpyrHFc

https://www.youtube.com/watch?v=SdOrkPn7d8Y

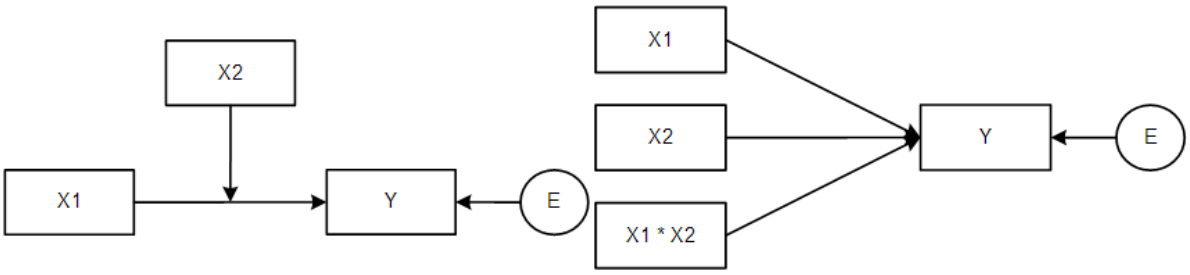
17.4 Tutorial

17.4.1 Interaction

In this assignment we work with the `PublicParticipation.sav` data. It contains (fictional) data on the following variables: income (higher scores, more income), public participation, education, age, and gender (0 = females; 1 = males). Public participation involves being member of school boards, municipal councillor, etc.

In this assignment we will see how we can model interaction between a continuous predictor and a dichotomous predictor.

Suppose we are interested in relationship between age and public participation, and we want to know if the relationship is moderated by gender. An interaction model is conceptually represented as follows (these two diagrams are interchangeable):



Modeling Interactions

The regression model for testing the interaction is:

$Y = b_0 + b_1X + b_2D_g + b_3XD_g$

where X = age, and D_g = gender (0 = women; 1 = men). Notice that women are our reference group.

To model interaction we need to create a new variable, which is the product of the dummy variable (gender in our case) and (age in our case).

This is best done via syntax, but to use the graphical interface proceed as follows:

via Transform > compute variable

Give the new variable a name (i.e., the target variable), say GenderTAge.

Then specify the product at the right (see more information button). Click on Paste, select and run the code. Check in Data View whether the product term was added correctly.

Alternatively, the syntax is:

```
COMPUTE GenderTAge = Gender * Age.
EXECUTE.
```

Now run the regression analysis that includes the interaction effect.

Important: Just like with dummies you must include all dummies that belong to the same variable in the model together, with an interaction term, you must always include its constituent variables as well. This is because the interaction term only *modifies* the effect of its constituent variables; the effect of those constituent variables must thus also be in the model.

So, if you add variable intXTZ into the model, you must also include X and Z.

Via analyze > regression > linear; choose age, gender and GenderTAge as the independent variables, and public participation as the dependent variable.

Consult the table Regression coefficients. Write down the general estimated model.

Finish the following equation, then check your answer.

Public Participation' = ...

Answer

Public Participation $= 3.252 + 0.137$ Age $+ 12.439$ Gender 0.116 Gender Age

Now write down the estimated models down for women and men separately. Hint: fill in 0 and 1 in the general estimated model mentioned in the previous step, then simplify the formula.

Complete the equations for women (W) and men (M):

$PP_W =$ _____[10] + _____[11] Age

$PP_M =$ _____[12] + _____[13] Age

Now draw (on a piece of paper) a graph of the results. That is, put age on the x-axis, the predicted public participation on the y-axis, and draw separate regression lines for males and females.

True or false

[10] 3.252
[11] 0.137
[12] 15.691
[13] 0.021

In the sample, age has a positive effect on public participation for women but a negative effect for men? TRUE / FALSE[14]

The researchers tested at the 5% level and concluded:

"We have convincing evidence that the population effect of age on public participation is different for men and women." TRUE / FALSE[15]

The estimated regression model was:

$Y = 3.252 + 0.137Age + 12.439D_g\ \ 0.116(Age\ \text{Œ}\ D_g)$

What would the regression equation look like if we would have used the men as the reference group? Use logic to answer this question, instead of re-running the analysis.

$Y =$ _____[16] + _____[17] Age+ _____[18] D_g+ _____[19] (Age Œ D_g)

To verify our answer to the previous question, we will recode the variable Gender such that males are scored 0 (= reference group) and females are scored 1.

Proceed as follows:

- via Transform > Recode into different variables
- Select Gender.
- Give a name to the new output variable (say GenderFem), give a label (say: "Gender (ref=males)" click on change.
- Specify old and new values: old value 0 becomes 1 and old value 1 becomes 0 (don't forget to click on add in between).
- Click OK. Verify that SPSS added a new column with a dummy variable where males are the reference group.
- Compute the product variable for the interaction between age and gender but now use the dummy having males as reference group.
- Rerun the regression analysis, but now using the new gender variable and interaction term. If you're answer in the previous step is correct you should find the values back in the table Regression Coefficients.

[14]FALSE
[15]FALSE
[16]15.691
[17]0.021
[18]-12.439
[19]0.116

17.4.2 Categorical Predictors with Three or more Categories

The categorical predictor Education has three levels (low, middle, high). If we want to include such a variable we need to use dummies.

Code the dummy variables as follows:

Value	D1	D2
Low	0	0
Middle	1	0
High	0	1

Which group is the reference group according to this coding? [20]

- (A) Low
- (B) Middle
- (C) High

Use syntax to create the dummies.

We are now ready for the regression analysis.

Run a hierarchical regression analysis with public participation as dependent variable. Model 1 only includes age. Model 2 includes age and the dummies. So we have the following nested models:

This model does not include the interaction effects yet! This means that we assume that the regression lines are parallel to one another. In the next assignment we check whether this assumption is reasonable.

Proceed as follows:

- via analyze > regression > linear.
- Select public participation as the dependent variable and only age as the independent variable. Click on next.
- Now select the two dummies we have created in the previous step. The two dummies together represent education. Always enter dummies into the model together!
- Via Statistics ask for the R-change statistics.

[20] Low

Consult the output and answer the questions in the next few steps.

Education and age together explain _____[21] % of the total variance.

What is the value of the test statistic that tests the unique effect of education, controlled for age? _____[22]

Report the results for the unique effect of education, then check your answer.

Answer

Education does not have a significant unique effect on public participation after controlling for age, $R^2 = .04, F(2, 38) = 0.895, p = .417$.

Consult the table with the regression coefficients.

Write down the estimated regression equation of Model 2.

Answer

$PublicParticipation = 10.478 + .097 \quad Age \quad 2.042 \quad D1 \quad 3.071 \quad D2$

Write down the estimated model for each of the three groups.

Then make a graph of the regression equations. Put age on the x-axis, the predicted public participation on the y-axis, and draw the lines for each education group.

Answer

The models were:
$PP_l = 10.478 + .097Age$
$PP_m = (10.478 \ \ 2.042) + .097Age = 8.436 + .097Age$
$PP_h = (10.478 \ \ 3.071) + .097Age = 7.407 + .097Age$
Did you get it right?

Suppose we have two persons, both are 40 years old, but one had middle level education and the other had high-level education.

What is the expected (absolute) difference in public participation between these two persons? _____[23]

The researchers conclude:

"Controlled for age, low educated people in the sample show highest level of public participation".

[21] 9.7
[22] 0.895
[23] 1.029

Is this a valid conclusion? TRUE / FALSE[24]

17.4.3 Interaction with more than Two Categories

In the previous assignment, we assumed that the effect of Age on Public participation was equal for each of the education level groups. However, we do not know whether this assumption is reasonable. In this assignment, we will check whether the interaction effect between Age and Education level is statistically significant or not.

Create the two interaction terms using syntax, with the Compute variable command. Note that we need two interaction terms: D1Tage and D2Tage.

We are now ready for the regression analysis.

Run a hierarchical regression analysis. Model 1 only includes age and the two dummy variables. Model 2 additionally includes the interaction terms.

Write down the formulas for the two nested models, then check your answer.

Answer

- Model 1: $Y = b_0 + b_1 Age + b_2 D_1 + b_3 D_2$
- Model 2: $Y = b_0 + b_1 Age + b_2 D_1 + b_3 D_2 + b_4 D_1 Age + b_5 D_2 Age$

Proceed as follows (or, preferably, use syntax):

- via analyze > regression > linear.
- Select public participation as the dependent and age, D1 and D2 as the independent variables. Click on next.
- Now select the two interaction terms we have created in the previous step. The two interaction terms together represent the interaction effect between education and age.
- Via Statistics ask for the R-change statistics.

Consult the output and answer the questions in the next few steps.

Before we carry out any of the significance tests, let's take a look at the coefficients table. Look at the unstandardized coefficients in Model2. First, write down the entire estimated model.

Complete the following equation:

[24] FALSE

Y = _____[25] + _____[26] Age+ _____[27] D_{middle}+ _____[28] D_{high}+ _____[29] (D_{middle} Age)+ _____[30] (D_{high} Age)

Next, write down the estimated model for each of the three education groups.

Remember, fill in 0 and 1 for the dummy variables, then simplify:

Y_{low} = _____[31] + _____[32] Age

Y_{middle} = _____[33] + _____[34] Age

Y_{high} = _____[35] + _____[36] Age

Now, answer the following questions.

True or False?

The effect of Age on Publication Participation in the sample is positive for all education groups. TRUE / FALSE[37]

For which group is the effect of Age on publication participation the strongest? [38]

- (A) Low
- (B) Middle
- (C) High

We inspected the estimated model. But is there a significant interaction effect to begin with? To answer that question we inspect the Model Summary Table.

First of all, write down the R^2 for the model without- and with interactions. What do these numbers mean?

Without interactions: _____[39] With interactions: _____[40]

---

[25] 11.426
[26] 0.073
[27] -5.19
[28] 1.577
[29] 0.067
[30] -0.088
[31] 11.426
[32] 0.073
[33] 6.236
[34] 0.14
[35] 13.003
[36] -0.015
[37] FALSE
[38] Middle
[39] 0.097
[40] 0.127

Finish the following sentence:

Model 2 with the interaction effects explains an additional _[41] % of the variance in Public Participation compared to Model 1 (on top of what was already explained by the main effects of Age and Education).

We will now carry out the F-change test. Write down the null hypothesis and alternative hypothesis that we test with this F-change test.

Answer

H_0 $R^2 = 0$
H_1 R^2 0

Write down the F-value, the df and the p-value.

- F-value: _____[42]
- df: (_[43] , ___[44])
- p-value: _____[45]

True or false: there is a significant interaction effect: TRUE / FALSE[46]

True or False: As a follow-up analysis, we should perform a simple effects analysis. TRUE / FALSE[47]

Interpret the results of Model 1 (without interaction) and report your results.

Answer

There is no evidence for a significant effect of Age and Education on Participation, $R^2 = .10, F(3, 38) = 1.36, p = .27$.

17.4.4 Interaction Effects

In this assignment, you will examine whether the effect of relationship with coworkers (sccowork; higher score = better relationship) on the emotional pressure at work (scemoti) has an interation effect with gender (0 = male, 1 = female).

If there is an interaction effect, the effect of sccowork on scemoti depends on the value of the variable gender.

[41] 3
[42] 0.618
[43] 2
[44] 36
[45] 0.545
[46] FALSE
[47] FALSE

Open Work.sav.

To be able to examine the interaction effect, you should first create a product variable.

- Go to Transform > Compute Variable
- Give a name to the new product variable in Target Variable (GenderTRelco for example).
- In Nummeric Expression you need to specify how the new variable should be computed. You have to enter gender * sccowork to compute the product of gender and sccowork.
- Paste and run the syntax, and check whether the product variable was added

Conduct a multiple regression analysis (using Analyze > Regression > Linear) with scemoti as dependent variable. The independent variables are the main effects (gender and sccowork) and the interaction effect (genderTsccowork).

What is the p-value of the interaction effect? _____[48]

True or false: The interaction effect is significant at = .10 TRUE / FALSE[49]

The regression equation for the entire sample is:

scemoti = _____[50] + _____[51] Gender+ _____[52] Relationship+ _____[53] (Gender Relationship)

For males, the value of Gender is 0. That means that GenderTRelco is also 0. The regression equation for males then becomes:

scemoti = _____[54] _____[55] Relationship

For females, the value of Gender is 1. What is the regression equation for females?

scemoti = _____[56] + _____[57] Relationship

Draw (on paper, not in SPSS) a schematic graph of the interaction effect. Put relationship with coworkers on the X-axis, and emotional pressure on the Y-axis. Draw a schematic regression line for each group.

In what group is the effect of relationship with coworkers on emotional pressure the strongest: males or females? [58]

---

[48] 0.083
[49] TRUE
[50] 27.166
[51] -7.103
[52] -0.237
[53] 0.439
[54] 27.166
[55] -0.237
[56] 20.063
[57] 0.202
[58] Males

- (A) Males
- (B) Females

In practice, you'd often want to know whether the effects within the groups are significant.

Can you use the output of this regression analysis to draw conclusions about the significance of the effect within each group? [59]

- (A) No
- (B) Yes, but only for the group of males
- (C) Yes, but only for the group of females
- (D) Yes, for both groups

At this moment, we don't have enough information in the output yet to test the effect within the female group. But we can test the effect within the male group!

What is the p-value of the effect of sccowork on scemoti within the male group? _____[60]

To test the significance within the the group of females, we can simply switch the reference groups.

- Make a new dummy variable called `male`, on which males score 1, and females 0
- Compute a new product variable: `COMPUTE maleTsccowork = male * sccowork.`

Perform a new regression analysis with these predictors. This is exactly the same analysis, but now with women as reference group instead of men.

Look at the table with the estimated coefficients. What is the p-value of the effect of sccowork on scemoti within the female group? _____[61]

[59] Yes, but only for the group of males
[60] 0.237
[61] 0.187

18 GLM VIII - Logistic Regression

When dealing with binary dependent variables (nominal or ordinal), we could model the probability of observing the outcome (coded as 0 or 1) with a conventional linear regression model. The problem with this approach is that linear regression predicts probabilities outside the range of [0, 1], and will have heteroscedastic and non-normal residuals because there are only two discrete values for the dependent variable.

Logistic Regression overcomes these limitations of linear regression. The core idea behind logistic regression is to predict a transformation of the dependent variable, Y, rather than the raw scores. Specifically, we model the **log odds** of the probability of Y being one category (e.g., 1) versus the other category (e.g., 0). The logit function, denoted as `log(p/(1-p))`, models the probability p using an s-shaped curve bounded by 0 and 1. Furthermore, logistic regression assumes a Bernoulli error distribution, instead of a normal error distribution, which accounts for the fact that observed outcomes can only take the values 0 or 1.

18.0.1 Introducing the logit

Logistic regression predicts the **logit function** of the individual probabilities of observing the outcome, pi. The logit of the probability (pi) is given by `log(pi/(1-pi))`, ensuring that the predicted values remain within the valid probability range. The outcome, Yi, is assumed to follow a Bernoulli distribution with an individual probability of success, pi. The logit of this success probability is modeled as a linear function of the predictors, allowing us to use the familiar linear regression model to predict the logit of pi.

Understanding the distinction between **probability**, **odds**, and the **logit** is crucial in logistic regression. Probability is defined as the long-run proportion of outcomes of a random experiment in which a particular outcome is observed. Odds describe the ratio of the probability of an event occurring relative to the probability of it not occurring. Finally, the logit transforms odds into a linear function, enabling us to use the regression model. The transformations from probability to odds and logit, and back, are given by:

Operation	Formula
Probability to odds	$\text{odds} = \frac{P}{1P}$
Odds to probability	$P = \frac{\text{odds}}{1+\text{odds}}$
Odds to logit	$\text{logit} = \ln(odds)$

Operation	Formula
Logit to odds	$\text{odds} = e^{\text{logit}}$
Probability to logit	$\text{logit} = \ln(\frac{p}{1p})$
Logit to probability	$p = \frac{e^{logit}}{1+e^{logit}}$

18.1 Maximum Likelihood Estimation (MLE)

In traditional linear regression, we use "ordinary least squares" (OLS) estimation to obtain the model parameters. This method involves simple matrix algebra and always yields a unique solution. However, for logistic regression, there is no OLS solution due to the binary nature of the dependent variable. Instead, we turn to **Maximum Likelihood Estimation (MLE)**. A complete explanation is beyond the scope of this course, but here is a basic intuitive explanation of the procedure:

1. Start with random values for the coefficients (a and b)
2. Using those parameter values, calculate the individual probabilities predicted by the logistic regression formula
3. For each individual, calculate the likelihood of observing their true outcome in a Bernoulli distribution with model-implied probability pi
4. Multiply these probabilities across all individuals to get the overall likelihood of observing these data, given the chosen coefficient values
5. Adjust the values of a and b a little bit
6. Check if the likelihood has become larger
7. Repeat steps 2-6 until until we find the coefficient values that maximize the likelihood and no further improvement can be found.

In other words, we look for the values of a and b that maximize the likelihood of observing the observed outcome values.

18.1.1 Interpreting Coefficients

The **intercept (a)** represents the log odds of the outcome (Y) for someone who scores 0 on all predictors. We can convert this log odds to the probability of the outcome for an individual who scores 0 on all predictors using the formula `P = e^(a) / (1 + e^(a))`. We can also solve for the inflection point at which the model stops predicting 0 and starts predicting 1, or vice versa, using $X_{p=.5} = \frac{a}{b}$.

The **slope (b)** of the logistic regression equation determines how steeply the logistic function switches from predicting 0 to predicting 1 as the predictor variable (Xi) increases. Larger absolute values of b indicate a steeper transition between the two outcomes. If the slope

is positive, the function ascends (starts at 0 and goes to 1), resulting in an S-shaped curve. Conversely, if the slope is negative, the function descends (starts at 1 and goes to 0), resulting in a Z-shaped curve.

18.1.2 Odds Ratio

The **odds ratio** is another important concept when interpreting logistic regression coefficients. It represents the odds of the outcome occurring given a one-unit increase in the predictor variable (Xi), relative to the odds of the outcome occurring when Xi remains unchanged. For binary predictors (e.g., conditions), the odds ratio provides a sensible effect size. For continuous predictors, the odds ratio is a multiplyer by which the odds increase when the predictor increases by one unit.

To calculate the odds ratio for a logistic regression coefficient (b), we use the formula `OR = e^(b)`. A value greater than 1 indicates that the predictor is associated with higher odds of the outcome, while a value less than 1 indicates lower odds of the outcome. For example, if the odds ratio for the test score coefficient (b = 2.12) is 8.35, it means that for each unit increase in the test score, the odds of the outcome are multiplied by 8.35.

18.1.3 Model Fit

To assess how well the logistic regression model fits the data, we can use the likelihood obtained from maximum likelihood estimation (MLE). By multiplying the log likelihood by -2, we obtain the $2LL$, which is a chi-square distributed test statistic. Performing a chi-square test allows us to determine if the overall model is significant. The null hypothesis is that the model does not significantly differ from a model with no predictors. If the chi-square test is significant, it indicates that the model provides a better fit than a null model.

18.1.4 Likelihood Ratio Test

In logistic regression, we can also conduct a **Likelihood Ratio (LR) test**, which is a chi-square test for the difference in log likelihood between two nested models, $2LL_0$ $2LL_1$. The first model is the restricted model with fewer parameters, and the second is the full model with more parameters. The LR test helps us compare the two models and determine if the additional predictors in the full model significantly improve its fit. The degrees of freedom for the LR test are equal to the difference in the number of parameters between the two models.

18.1.5 Pseudo R2

Unlike linear regression, logistic regression doesn't have a traditional R-squared to measure explained variance. However, researchers have proposed several **Pseudo R2** statistics to approximate the concept of explained variance in logistic regression. These Pseudo R2 statistics rescale the -2 log likelihood of the model. Two common Pseudo R2 statistics are Cox & Snell and Nagelkerke.

Cox & Snell is a generalization of the "normal" R2, which provides the same value for ordinary least squares regression. For logistic regression, however, its value can never reach 1; it will be somewhere between 0 and < 1. Nagelkerke aims to "fix" this property by rescaling Cox & Snell to a range of [0, 1], by dividing it by its maximum possible value. While these statistics can provide a measure of relative model fit and help compare models on the same dataset, they do not represent absolute model fit or effect size.

18.1.6 Classification Accuracy

One way to evaluate the model's predictive performance is by using a **classification table**. The classification table compares the predicted outcomes with the actual outcomes to determine how well the model predicts true positives and true negatives. The table is constructed by calculating the predicted probability for each individual and then dichotomizing these probabilities using a specific cutoff, typically 0.5. Individuals with predicted probabilities above the cutoff are classified as "1," and those below as "0." The observed outcomes are then cross-tabulated against the dichotomized predictions.

By examining the classification table, we can assess the accuracy of the model's predictions and identify areas of improvement. Researchers could choose a different cutoff to optimize the trade-off between false positives and false negatives, depending on the specific goals of the analysis.

In summary, evaluating logistic regression models involves assessing model fit through chi-square tests, using Pseudo R2 statistics for relative fit comparison, and examining classification accuracy to understand the model's predictive performance.

18.2 Lecture

18.3 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What type of dependent variable is suitable for logistic regression? [1]

- (A) Ordinal
- (B) Binary
- (C) Interval
- (D) Categorical

Question 2

What is the primary purpose of maximum likelihood estimation in logistic regression? [2]

- (A) To minimize the sum of squared errors
- (B) To maximize the parameter values
- (C) To maximize the likelihood of observing the data given the model
- (D) To maximize the variance explained by the predictors

Question 3

How is the likelihood ratio test used in logistic regression? [3]

- (A) To test for normality of residuals
- (B) To compare the fit of nested models
- (C) To check for multicollinearity
- (D) To assess the effect size of predictors

Question 4

Which of the following is a Pseudo R2 statistic commonly used in logistic regression? [4]

[1] Binary
[2] To maximize the likelihood of observing the data given the model
[3] To compare the fit of nested models
[4] Cox & Snell R2

- (A) Cox & Snell R2
- (B) Pearson's R2
- (C) Adjusted R2
- (D) -2LL

Question 5

What does the logit function do in logistic regression? [5]

- (A) Transforms the predicted probabilities to log odds
- (B) Standardizes the dependent variable
- (C) Calculates the Wald test statistic
- (D) Converts continuous predictors to categorical

Question 6

What is the range of the Cox & Snell Pseudo R2 statistic in logistic regression? [6]

- (A) -Infinite to 1
- (B) 0 to < 1
- (C) 0 to Infinite
- (D) 0 to 1

Question 7

How is the classification table used in logistic regression evaluation? [7]

- (A) To assess the model's predictive accuracy
- (B) To compare model fit using likelihood ratio test
- (C) To determine effect size of predictors
- (D) To test for multicollinearity

[5] Transforms the predicted probabilities to log odds
[6] 0 to < 1
[7] To assess the model's predictive accuracy

Question 8

What does a significant chi-square test in logistic regression indicate? [8]

- (A) The model provides a better fit than a null model
- (B) The model's predicted probabilities are accurate
- (C) The model's predictors are collinear
- (D) The model's residuals are normally distributed

Question 9

Which parameter represents the odds ratio associated with a one-unit increase in the predictor? [9]

- (A) The logit function
- (B) The coefficient b
- (C) The p-value
- (D) The exponent of the coefficient b, e^b

Question 10

When calculating the likelihood in logistic regression, what does a high value of likelihood imply? [10]

- (A) The model has a high R-squared value
- (B) The observed outcome values are very likely given the parameters
- (C) The model is overfitting the data
- (D) The model's residuals are normally distributed

Question 11

What is the main purpose of the likelihood ratio test in logistic regression? [11]

[8] The model provides a better fit than a null model
[9] The exponent of the coefficient b, e^b
[10] The observed outcome values are very likely given the parameters
[11] To compare model fit between two nested models

- (A) To assess the normality of residuals
- (B) To test the significance of individual predictors
- (C) To evaluate multicollinearity among predictors
- (D) To compare model fit between two nested models

Question 12

What can help researchers optimize the trade-off between false positives and false negatives in logistic regression? [12]

- (A) The odds ratio
- (B) The -2LL
- (C) The pseudo R2
- (D) The classification table

Question 13

What is the range of the Nagelkerke Pseudo R2 statistic in logistic regression? [13]

- (A) -1 to 1
- (B) 0 to < 1
- (C) 0 to 1
- (D) -Infinity to 1

[12] The classification table
[13] 0 to 1

Show explanations

Question 1
Logistic regression is used for binary categorical outcomes.
Question 2
Maximum likelihood estimation aims to maximize the likelihood of observing the data given the model.
Question 3
The likelihood ratio test compares the fit of nested models to determine if additional predictors significantly improve the model.
Question 4
Cox & Snell R2 is a Pseudo R2 statistic used in logistic regression.
Question 5
The logit function transforms the predicted probabilities to odds.
Question 6
The Cox & Snell Pseudo R2 statistic never reaches 1 for logistic regression, so it ranges from 0 to < 1.
Question 7
The classification table helps assess the model's predictive accuracy by comparing predicted outcomes with actual outcomes.
Question 8
A significant chi-square test indicates that the model provides a better fit than a null model.
Question 9
The exponent of the coefficient b represents the odds ratio associated with a one-unit increase in the predictor.
Question 10
A high value of likelihood indicates that the observed outcome values are very likely given the parameters.
Question 11
The likelihood ratio test is used to compare model fit between two nested models.
Question 12
The classification table helps researchers optimize the trade-off between false positives and false negatives in logistic regression.
Question 13
The Nagelkerke Pseudo R2 statistic rescales the Cox & Snell statistic to range from 0 to 1.

18.4 Tutorial

18.4.1 Probability, Odds, and Logits

We will start this session on logistic regression with some theoretical exercises. This way, you will learn how to work with probability, odds, and logits.

The given probability is P = 0.36.

What are the corresponding odds? _____[14]

Please find the formula's below:

Goal	Function
Probability -> Odds	$odds = P/(1\ \ P)$
Odds -> Probability	$P = odds/(1 + Odds)$
Odds -> Logit	$logit = \ln(odds)$
Logit -> Odds	$odds = e^{logit}$

Again, the given probability is P = 0.36.

What is the corresponding logit? _____[15]

The given logit is – 2.7.

What are the corresponding odds? _____[16]

Again, the given logit is – 2.7. _____[17]

Discuss with your group mates when and why we should carry out logistic regression analysis.

Explanation

We carry out logistic regression analysis when we want to carry out regression analysis and we have a dichotomous outcome variable (i.e., a dependent variable with two answer categories). In that case we cannot use linear regression because several of the assumptions of linear regression are violated.

In a study concerning the smoking behavior amongst adolescents, a logistic regression analysis is conducted to check the effect of the image of smoking (Image) on smoking behavior (Smoking).

[14] 0.563
[15] -0.575
[16] 0.067
[17] 0.063

Image: to what degree the young adult thinks smoking is perceived as "cool". The variable image is measured on a scale from 10 to 30, in which higher scores indicate that smoking is perceived as cooler.

Smoking: whether or not the adolescent smokes.

- 0 = the adolescent is a non-smoker
- 1 = the adolescent is a smoker

Below you can find part of the output the researchers retrieved.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Image	.284	.043	42.864	1	.000	1.328
	Constant	-5.647	.835	45.690	1	.000	.004

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

First, write down the estimated regression model.

Answer

$Logit(Smoking = 1) = 5.65 + 0.28\ \ Image$

What is the probability that an adolescent with a score of 15 on image smokes? _____[18]

Take a look at the regression coefficient.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Image	.284	.043	42.864	1	.000	1.328
	Constant	-5.647	.835	45.690	1	.000	.004

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

Imagine that one's score on Image will increase from 15 to 16.

To what degree will the logit and odds change?

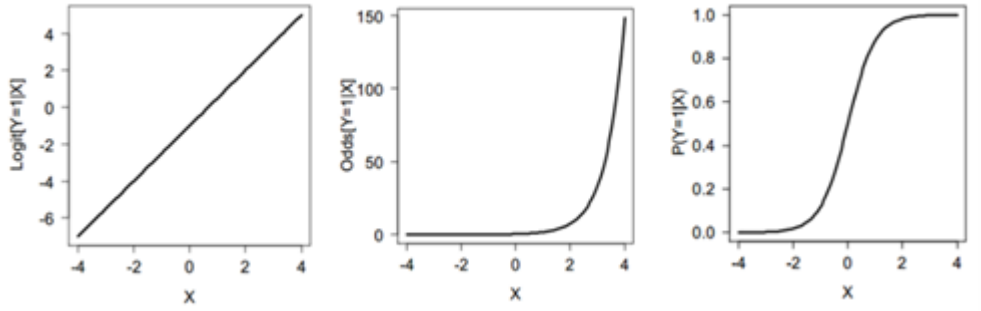
And what can you conclude about the increase in probability?

[18] 0.33

Answer

The logit increases with 0.284. So 1.387+0.284=1.103
The odds increase with a factor 1.328. So, the odds become 0.25Œ1.328=0.332
By just looking at the output, it is unclear what *exactly* will happen with the probability. We DO know that the probability will increase because the logits and the odds increase. If we would calculate the probability by hand, we would see that the probability increases with 0.05.

As you can see, the logit follows a linear function, the odds follow an exponential function, and probability follows a logistic function (perhaps less clear from the example but see the graphs below for an illustration). It is nice to work with a linear model (i.e., work with the logits), but from an interpretation point of view odds and probabilities are nicer to work with.



18.4.2 Logistic Regression

Today we will study whether the probability to pass a Statistics course depends on exam fear and the math grade obtained in high school (prior math ability).

Open LAS BE LR.sav.

In the table below you find the variables included in this fictional data set.

We want to study the following research question:

Does the probability of passing the statistics exam depend on exam stress and math ability?

To answer this question, we first need to dichotomize the dependent variable. This is an unusual step, because dichotomizing variables loses information! Still, for the purpose of the present exercise, we will do so. We will create the new dependent variable Passed (0 = fail, 1 = pass), where a 5.5 or higher counts as a pass. Use the following steps:

Navigate to Transform Recode into different variables. Select GradeStats as input variable and use Passed as the name for the new recoded variable. Do not forget to click on CHANGE. Click on Old and new values and specify the recoding in such a way that all grades of 5.5 or

higher will get the new value 1 and all grades lower than 5.5 (i.e., all other grades) the value 0. Paste and run the syntax. Go to variable view and specify the value labels.

Obtain the frequency distribution for Passed (via Analyze –> Descriptive Statistics –> Frequencies).

What percentage of students passed? _____[19]%

We will use logistic regression to study the effect of exam stress and math ability on the probability to pass. Take the following steps:

Navigate to Analyze –> Regression –> Binary logistic regression Select Passed as dependent variable and ExamStress and Math as independent variables (in SPSS called "covariates"). Paste and run the syntax.

Inspect the output corresponding to Block 1. Take a look at the table "Variables in the Equation".

What would be good description of the effects on the sample level? (i.e., what the effect of ExamStress and Math on the probability to pass looks like).

Answer

Turns out Exam Stress has a negative effect on the probability to pass (controlled for Math grade), and Math a positive effect (controlled for Exam Stress).

What is the value of the regression coefficient for the variable Exam Stress? _____[20]

Give a detailed interpretation of this number.

Answer

If we would increase one unit in Exam Stress the logits to pass the exam will decrease with 0.025, while keeping the variable Math constant.

True or false: The independent variables Math and Exam Stress together have a significant effect on the probability to pass. TRUE / FALSE[21]

Explanation

When we inspect the model test, we see that, together, Exam Stress and Math grade have a significant effect on the probability to pass, $2(2) = 81.043$, $p < .001$.

Second, carry out the significance tests for the individual predictors as well.

[19] 82.2
[20] -0.025
[21] TRUE

True or false: There is a significant effect of Math on the probability to pass, controlled for Exam stress. TRUE / FALSE[22]

True or false: There is a significant effect of Exam stress on the probability to pass, controlled for Math. TRUE / FALSE[23]

Explanation

When we inspect the separate regression coefficients together, we see that:
Controlled for Exam stress, Math grade has a significant effect on the probability to pass, 2(1)=54.500, p<.001. Controlled for Math grade, Exam stress has no significant effect on the probability to pass, 2(1)=0.040, p=.841.

18.4.3 Logistic Regression with Categorical Predictor

Now, we will carry out a logistic regression analysis to study whether the probability to pass can be predicted from the amount of preparation (Preparation), controlled for prior math ability (Math).

But... the variable Preparation is a nominal (categorical) variable. Use dummy coding to include it in the model. The variable Preparation has three categories. Create all three dummy variables!

Take the following approach to carry out the analysis.

Navigate to Analyze –> Regression –> Binary Logistic

Select Math as covariate.

Select the dummies for Preparation as covariates. At this stage, use "only reading the book" as the reference group.

Paste and run the syntax.

Take a look at the table Variables in the Equation.

Inspect the estimated regression coefficients. Consider the results on the sample level, ignoring inferential statistics for now.

Which group has the highest estimated probability to pass? [24]

- (A) Students who read the book and did the exercises.

[22] TRUE
[23] FALSE
[24] Students who read the book and did the exercises.

- (B) Student who only read the book.
- (C) Students who only did the exercises.

Which group has the lowest estimated probability to pass (controlled for math ability)? [25]

- (A) Student who only read the book.
- (B) Students who only did the exercises.
- (C) Students who read the book and did the exercises.

Write down the full model equation, filling in the values of coefficients. Then check your answer.

Answer

$logit_{passed} = 5.833 + .300\ D_{exercises} + .946\ D_{both} + .999\ Math$

Controlled for Math Grade, what is the difference in predicted odds between people who only read the book and people who read the book and made the exercises? _____[26]

Explanation

Take the exponent of the regression coefficient for the dummy for people who read the book and made the exercises (see table).

By hand, calculate the probability that the third person in the file (respID=3) passes the exam. _____[27]

Imagine that you do not know whether this person passed the exam or not.

True or false: Based on your answer in the previous step and using a decision threshold of .5, you would predict this student to pass. TRUE / FALSE[28]

SPSS can easily calculate the probability to pass.

Either use the visual interface:

Click on the SAVE button in the menu for logistic regression.

Select Probabilities under the header Predicted Values.

Click on continue.

[25] Student who only read the book.
[26] 2.576
[27] 0.752
[28] TRUE

Or add this line of code to your model:

```
/SAVE=PRED
```

SPSS adds a new variable to the data file, which gives the predicted probability for each person. Check whether you calculated the predicted probability for person 3 correctly.

Imagine that we would have used this model to predict the probability to pass for the students in this sample before they even made the exam. Use a decision threshold of .5 to classify students as those likely to pass vs fail.

Out of all students that were classified as "pass", which proportion actually failed? _____[29]

Explanation

Of the 370 students who were predicted to pass, 52 failed the exam (.141). As you can see, the model is not completely flawless in making predictions.

What is worse in your opinion? Incorrectly predicting that students will pass (when they end up failing), or incorrectly predicting that students will fail (when they end up passing)? Why?

If you chose a different decision criteria to predict passing - say .7, what would happen to the proportion of students that were classified as "pass", but actually failed? [30]

- (A) can't say
- (B) stays the same
- (C) gets smaller
- (D) gets bigger

18.4.4 Hierarchical Logistic Regression

In this assignment we will compare the following two models against each other.

- Model1: $Logit(Pass) = b0 + b1\ Math$
- Model2: $Logit(Pass) = b0 + b1\ Math + b2\ ExamStress + b3\ Evaluation$

[29] 0.141
[30] gets smaller

Later on, we want to carry out a model comparison test to check whether the larger model predicts the probability of passing the exam better than the smaller model.

What would be the regression df for the model comparison test, in which we compare the larger model 2 to the smaller model 1? _[31]

Now we will carry out the model comparison test. Take the following steps.

Navigate to Analyze -> Regression -> Binary Logistic

Select Math as predictor (covariates).

Click on Next (upper right); we can now indicate which block of predictors we like to add in addition to the variables added in the first model.

Enter ExamStress and Evaluation as predictors (covariates).

Paste and run the syntax.

Inspect the output. SPSS organizes the output in three blocks. The results in Block 0 refer to the null model without any predictors. Note that this null model also exists when you use the "Linear Regression" interface, but it's not featured in the output.

Block 1 refers to the results of Model 1 and Block to the results of Model 2.

Take a look at the results of the model comparison test.

What test statistic is used to compare nested logistic regression models? [32]

- (A) Z
- (B) t
- (C) F
- (D) Chi squared

What is the value of the appropriate test statistic for comparing models 1 and 2? _____[33]

True or false: Adding predictors ExamStress and Evaluation lead to significantly better predictions, compared to a model with only high school math grade as predictor. TRUE / FALSE[34]

[31] 2
[32] Chi squared
[33] 1.297
[34] FALSE

19 GLM: Contrasts

Recall that in one-way ANOVA, we have a categorical predictor and a continuous outcome variable. The overall F-test of an ANOVA provides an omnibus (overall) test of differences between group means. The null hypothesis tested in this case is that all k groups have the same mean, $H_0 \ _1 = _2 = \ _k$. Today, we delve deeper into tests you can perform when you have more complex hypotheses about group means, or want to perform follow-up tests to determine which groups differ significantly.

First, we already covered that we can use dummy variables to incorporate this categorical predictor in a regression model. We pick one reference category and create dummy variables to indicate membership in the other groups. These dummy variables take binary values (0 or 1) to represent group membership.

The formula for regression with dummies is:

$Y_i = a + b_1 D_{1i} + b_2 D_{2i} + b_3 D_{3i} + b_4 D_{4i}$

The intercept represents the mean value of the reference category, and the coefficients b represent the differences between each group and that reference category.

A different way to specify the same model is to omit the intercept, and include a dummy for each of the k groups, so we represent k groups with k dummies:

$Y_i = b_1 D_{1i} + b_2 D_{2i} + b_3 D_{3i} + b_4 D_{4i} + b_5 D_{5i}$

Both of these models are mathematically identical. The advantage of estimating all group means is that this model provides a standard error for each group mean, allowing us to test each group mean against hypothesized values.

19.0.1 Effects Coding

Another way to include a categorical predictor is via effects coding. Effects coding compares each group to the grand mean. When we have unequal group sizes, the coding scheme should account for relative group size.

19.0.2 Contrast Coding

Contrast coding is yet another coding scheme; it compares groups of means. This allows us to test specific hypotheses about differences between groups. Contrast coding is a very advanced technique that requires you to perform some basic matrix algebra in Excel.

19.0.3 'Post-Hoc' Tests

The notion of post-hoc tests is a bit outdated; it essentially refers to making all possible comparisons between group means. The name is based on the fact that such tests are rarely hypothesized beforehand. They can be considered an exploratory procedure to look for differences between groups. Note that performing many tests inflates the risk of drawing false-positive conclusions (Type I error).

19.0.4 Adjusting for Multiple Comparisons

When conducting multiple tests, we face an increased risk of committing Type I errors (false positives). If we perform m tests within one study, the experiment-wise Type I error rate is $1\ (1\ )^m$. To control the experiment-wise Type I error rate, we can apply a Bonferroni correction, which divides the significance level by the number of tests m. This trades off fewer false positive results for more false negative results.

Planning to test specific hypotheses before conducting the study also helps control Type I error.

19.1 Lecture

https://www.youtube.com/embed/8UDpTwXoINU

19.2 Tutorial

19.2.1 Group Means

In this tutorial, you will learn to use the general linear model to estimate means and to test the difference between two group means, the difference between individual group means and the overall mean, and between groups of means.

Open `hiking_long.sav` in SPSS.

The data file describes the result of a fictitious experiment. A hiking guide has displayed five different types of behavior towards different groups of hikers. The treatment that each person received from the guide is recorded in the variable `behavior`.

The dependent variable of this experiment is `feeling`. Higher scores on this variable indicate a more positive attitude of a participant towards the guide. In this assignment, we will use ANOVA to determine whether the mean score on the dependent variable differs between the five experimental conditions.

The data file contains a third variable named `weather` which can be either good or bad. For now, we will only look at the results obtained during good weather. Hence, we will use "Select cases" to select only those participants with a value of 1 on the weather variable.

Additionally, the data contains a variable named `balanced` which distinguishes between data resulting from a balanced experiment (with equal sample sizes in all groups), and from an unbalanced experiment (with unequal group sizes). For now, just ignore this variable.

Click Data > Select Cases and select "If condition is satisfied" and click the "If"-button. Now enter the following condition into the equation box:

```
weather = 1
```

Now click "Continue" and "Paste" to paste the resulting syntax into the syntax editor. Select Run > All to run it. You should now see in the Data View tab that half of the participants have been crossed out.

First, let's compute the overall mean of feeling and tabulate the group means.

What is the overall mean of feeling? _____[1]

What are the group means:

- What is the mean of the rushing group? _____[2]
- What is the mean of the stories group? _____[3]
- What is the mean of the insulting group? _____[4]

[1] 5.74
[2] 5.47
[3] 6.12
[4] 4.62

Answer

To get the mean of feeling, use Analyze -> Descriptive Statistics -> Descriptives.
To get the group means, use Analyze -> Compare Means -> Means

```
DESCRIPTIVES VARIABLES=feeling
  /STATISTICS=MEAN STDDEV MIN MAX.

MEANS TABLES=feeling BY behavior
  /CELLS=MEAN COUNT STDDEV.
```

You have previously learned to include categorical variables in a linear model by using dummy coding. Today, we will build upon this principle of encoding the information from a categorical variable into several numerical variables.

First, recall that a linear model with a five-group nominal predictor can be written as follows:

$Y = b_0 + b_1\ D_1 + b_2\ D_2 + b_3\ D_3 + b_4\ D_4$

What is b_0 in this equation?

[5]

- (A) The intercept; it is the mean of the reference category.
- (B) The average of the group means
- (C) The overall sample mean.
- (D) The slope of the reference category.

To estimate the model above using regression, you could code dummy variables as follows:

behavior	D1	D2	D3	D4
rushing	1	0	0	0
telling stories	0	1	0	0
insulting	0	0	1	0
making jokes	0	0	0	1
singing	0	0	0	0

What is the reference category in the coding scheme above? [6]

[5] The intercept; it is the mean of the reference category.
[6] singing

- (A) singing
- (B) rushing
- (C) none
- (D) jokes

Specify the dummies as described in the table, and estimate the model.

Answer

To get the mean of feeling, use Analyze -> Descriptive Statistics -> Descriptives.
To get the group means, use Analyze -> Compare Means -> Means

```
RECODE behavior (1=1) (2=0) (3=0) (4=0) (5=0) INTO rushing.
RECODE behavior (1=0) (2=1) (3=0) (4=0) (5=0) INTO stories.
RECODE behavior (1=0) (2=0) (3=1) (4=0) (5=0) INTO insulting.
RECODE behavior (1=0) (2=0) (3=0) (4=1) (5=0) INTO jokes.
EXECUTE.


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT feeling
  /METHOD=ENTER rushing stories insulting jokes.
```

What's the value of the intercept? _____[7]

In this analysis, the intercept is the mean value on feeling for the reference category (singing). Verify that this is true by comparing the intercept of this regression to the Means table you made previously.

What is the value of the coefficient for `stories`? _____[8]

How can we interpret this coefficient?

[9]

[7] 6.35
[8] -0.23
[9] The difference between the mean of the singing group and the mean of the stories group.

- (A) The difference between the mean of the singing group and the mean of the stories group.

- (B) The mean of the stories group.

19.2.2 More Dummies

As we've previously established, dummy variables allow us to test the significance of mean differences between one reference group and all other groups.

Now, imagine we expect rushing to have a negative effect on behavior, and we want to know which other behaviors are "better" (i.e., result in a higher score on behavior) than rushing.

Specify your hypotheses, then check your answer.

Answer

H_0 $rushing$ $(stories, insulting, joking, singing)$
H_1 $rushing < (stories, insulting, joking, singing)$

Use dummy variables to test this hypothesis. Note: you will need to specify one additional dummy.

Answer

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT feeling
  /METHOD=ENTER stories insulting jokes singing.
```

What is the R^2 of this model? _____[10]

Compare this to the R^2 of your previous model. They should be identical, as should be the overall F-test and p-values. Changing the reference category doesn't change what information the dummies convey.

Do we perform one-sided or two sided tests? [11]

- (A) one-sided

[10] 0.37
[11] one-sided

- (B) two-sided

Use this information to test your hypotheses.

Which behaviors are "better" than rushing?

[12]

- (A) jokes and singing
- (B) no behaviors
- (C) all behaviors
- (D) stories, jokes, and insulting

You can use this technique any time you want to test the significance of the difference between one reference group and other groups.

Note that, in the first assignment, we used a different set of dummy variables than in the second assignment. This means that you can always use different sets of dummy variables when want to compare against multiple reference groups.

19.2.3 Estimating group means

In the first assignment, we computed the group means. But remember that the ANOVA model allows us to estimate them using the general linear model. In this assignment, we will do so by hand. After the previous assignment, you should have five dummy variables to represent the five groups of `behavior`.

Until now, you've always included four dummy variables to represent five categories, as the last category is represented by the intercept.

However, it is also possible to represent five groups as follows:

$Y = b_1\ D_1 + b_2\ D_2 + b_3\ D_3 + b_4\ D_4 + b_5\ D_5$

What is b_5 in this equation?

[13]

- (A) The mean value of group 5
- (B) The intercept; it is the mean of the reference category.

[12] stories, jokes, and insulting
[13] The intercept; it is the mean of the reference category.

- (C) The slope of the reference category.
- (D) The overall sample mean.

To estimate the model above using regression, you could code dummy variables as follows (note that you should already have all these dummies from the previous assignment):

behavior	D1	D2	D3	D4	D5
rushing	1	0	0	0	0
telling stories	0	1	0	0	0
insulting	0	0	1	0	0
making jokes	0	0	0	1	0
singing	0	0	0	0	1

Now, go to Analyze -> Regression -> Linear, and add **all five** dummies as predictors.

Then, click the Options button, and notice the option titled "Include Constant in Equation".

Turn this option off to remove the intercept from the regression equation, then paste your syntax. Notice a new line that says `/ORIGIN` instead of `/NOORIGIN`. This command removes the intercept.

Run your syntax, and examine the results.

You might notice that the R^2 and F-test changed. This is because these are computed relative to a null-model with only the intercept - but you told SPSS not to include an intercept, so it can't compute that null model here. It's not a big deal. As soon as you estimate models with an intercept again, the R^2s will be identical again, regardless of the dummy coding.

What's the value of the dummy for singing? _____[14]

Compare all coefficients to the table of means from the first assignment. They should all be identical.

This is how you estimate means using the linear model!

Now, what do the t-tests and p-values in the Coefficients table tell us?

[15]

- (A) Whether the means are significantly different from the reference category.
- (B) Whether the means are significantly different from each other.

[14] 6.35

[15] Whether the means are significantly different from zero.

- (C) Whether the means are significantly different from zero.
- (D) They are not meaningful.

Keep in mind that you can use the standard errors from the coefficients table to perform t-tests against other values than 0; for example, what would the test statistic be when testing whether the mean of the insulting group is significantly different from 5, so $H_0 \quad _{insulting} = 5$? t = _____[16]

Is the difference significant? [17]

- (A) Yes
- (B) No

You can use regression without an intercept any time you wish to estimate *all* group means in a single analysis and/or test the group means against specific hypothesized values.

19.2.4 Comparing to Overall Mean

Until now, we've represented levels of a categorical variable using *dummy variables* with values 0 or 1.

In this assignment, we introduce an alternative coding system: *effects coding*.

The main difference with dummy coding is that the reference category does not receive 0 values on all indicator variables, but instead, receives a negative value. In a balanced design (with equal sizes for each group), this value is -1.

So in a balanced design, with equal group sizes, the coding scheme for effects coding is (I now use the letters E1-E4 to clarify that these are not dummies but effect coded indicators):

behavior	E1	E2	E3	E4
rushing	1	0	0	0
telling stories	0	1	0	0
insulting	0	0	1	0
making jokes	0	0	0	1
singing	-1	-1	-1	-1

The reference category is still "singing".

The resulting model will give us the following information:

[16] -1.52
[17] No

- The overall sample mean for feeling
- The difference between each group mean, except for singing, compared to the overall mean

Most of the time, however, we will **not** have balanced designs. With this in mind, it is more useful to learn the general way to construct effect coding.

Specifically, the weights assigned for the singing category (reference category) differ for each dummy, and are computed as:

1 $n_{\text{this category}}/n_{\text{reference category}}$

Check the group sizes in the output from assignment 1.

What is the sample size for the rushing group? ___[18]

What is the sample size for the reference category? ___[19]

With this in mind, what should the weight be for the singing group, on the dummy that codes for membership of the rushing group? _____[20]

Complete the following syntax, then run it:

RECODE behavior (1=1) (2=0) (3=0) (4=0) (5= . . .) INTO Erushing. RECODE behavior (1=0) (2=1) (3=0) (4=0) (5=[21]**) INTO Estories. RECODE behavior (1=0) (2=0) (3=1) (4=0) (5=**____[22]) INTO Einsulting. RECODE behavior (1=0) (2=0) (3=0) (4=1) (5=. . .) INTO Ejokes. EXECUTE.

Note the correct answer for the effect code for the stories group. Compare the number of people in the stories group and in the reference group. Then recall that I explained that *In a balanced design (with equal sizes for each group), this value is -1.* You see that this is true now, and why.

Calculate the effect indicators, then specify a regression model with these four effect indicators. Make sure to include the intercept again!

[18] 13
[19] 11
[20] -1.18
[21] -1
[22] -1.18

Check correct syntax

```
RECODE behavior (1=1) (2=0) (3=0) (4=0) (5= -1.18) INTO Erushing.
RECODE behavior (1=0) (2=1) (3=0) (4=0) (5=-1) INTO Estories.
RECODE behavior (1=0) (2=0) (3=1) (4=0) (5=-1.18) INTO Einsulting.
RECODE behavior (1=0) (2=0) (3=0) (4=1) (5=-1.18) INTO Ejokes.
EXECUTE.


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT feeling
  /METHOD=ENTER Erushing Estories Einsulting Ejokes.
```

Run the syntax. What is the F-value of the model? _____[23]

Verify that this is the same value you got before in models with an intercept.

What is the value of the intercept? _____[24]

Verify that this is identical to the overall mean of the dependent variable.

Which group means differ significantly from the overall mean?

[25]

- (A) no behaviors
- (B) insulting
- (C) all behaviors
- (D) jokes and insulting

Using the coefficients table, calculate the mean of the jokes group. What value do you get? _____[26]

This should be identical to the mean you observed in the previous assignment (using regression to estimate means), and in the first assignment (just computing the means).

[23] 8.34
[24] 5.74
[25] jokes and insulting
[26] 6.3

19.2.5 Comparing Groups of Means

Extending the methods above, it is also possible to compare *groups* of means. For example, we might wonder whether negative behaviors (rushing and insulting) differ significantly from positive behaviors (stories, jokes, and singing).

This approach builds upon the logic of effects coding, where the weights for the reference category were based on the relative sample size of the reference category. This time, however, the weights for the category to be compared to the reference category are *also* based on a sample size.

We are going to perform several steps, as explained in the lecture.

19.2.5.1 Step 1: Plan Contrasts

Keep in mind these rules:

1. The possible values of each indicator variable must sum to 0.
2. Each group must be uniquely identified by a particular combination of the contrast variables.

Assume for a moment that we have equal group sizes and want to compare groups 1 and 2 to groups 3, 4, and 5.

Appropriate contrasts would then be (I'm using the letter C to indicate that these are not dummies or effect indicators):

behavior	C1	C2	C3	C4
rushing	1	1	0	0
insulting	1	-1	0	0
telling stories	$-\frac{2}{3}$	0	2	0
making jokes	$-\frac{2}{3}$	0	-1	1
singing	$-\frac{2}{3}$	0	-1	-1

Note that:

- Each column sums to 0
- Every level of behavior is uniquely identified by some combination of contrasts

In this case, we only care about C1; we created C2, C3 and C4 to ensure that every level of behavior is uniquely identified. But what do C2-C4 test?

C2 compares the two negative behaviors; C3 compares stories against jokes and singing. C4 compares jokes and singing.

19.2.5.2 Step 2: Account for Group Size

Now, we have to account for the relative sample sizes of these groups to ensure that we can interpret the coefficients as the difference between the means of those combinations of groups.

Use the descriptive statistics you previously obtained to weight the contrasts from step 1.

E.g., contrast C3 below is already completed. Which other contrasts do you still need to change? [27]

- (A) C1
- (B) C1, C2, C4
- (C) none
- (D) C2 and C4

behavior	C1	C2	C3	C4
rushing	1	1	0	0
insulting	1	-1	0	0
telling stories	$-\frac{2}{3}$	0	1	0
making jokes	$-\frac{2}{3}$	0	-13/(11+13)	1
singing	$-\frac{2}{3}$	0	-11/(11+13)	-1

19.2.5.3 Step 3: Do Matrix Algebra

Enter the complete matrix into a spreadsheet program. Add one column before the contrasts with an intercept for each group, equal to $1/k$.

What's the value of this intercept for this study? _____[28]

- Click an Empty cell
- Paste `=MINVERSE(TRANSPOSE(`
- Select your contrast matrix
- Finish the formula by typing closing brackets `))`

[27] C1
[28] 0.2

These are the values you will use for your indicators!

Now, write syntax to create the contrasts using the values you calculated in a spreadsheet. Give these contrasts informative names to help remind yourself of their interpretation. Here is one example; complete the rest yourself:

```
RECODE behavior (1=.6) (2=.6) (3=-.4) (4=-.4) (5= -.4) INTO posVneg.
```

Answer

```
RECODE behavior (1=.6) (2=.6) (3=-.4) (4=-.4) (5= -.4) INTO posVneg.
RECODE behavior (1=.5) (2=-.5) (3=0) (4=0) (5= 0) INTO rushVinsult.
RECODE behavior (1=-.01) (2=-.01) (3=.67) (4=-.33) (5= -
.33) INTO storyVjokesing.
RECODE behavior (1=.02) (2=.02) (3=.02) (4=.48) (5= -.53) INTO ? .
EXECUTE.
```

What does the final contrast encode? [29]

- (A) positive versus negative
- (B) all levels
- (C) joke versus singing
- (D) story vs singing

19.2.6 Run the Analysis

Create the indicator variables and run the regression analysis.

What is the mean difference in feeling between rushing and insulting behaviors? _____ [30]

Which effects are significant? [31]

- (A) all contrasts
- (B) positive V negative behaviors
- (C) story V jokes, singing
- (D) rushing V insulting

[29] joke versus singing
[30] -0.65
[31] story V jokes, singing

19.2.7 Adjusting for Multiple Comparisons

In these assignments, we have been conducting many tests. You have learned that the significance level indicates the probability of drawing a false-positive conclusion (Type I error). However, these probabilities add up for multiple tests! So when you perform many tests, you can be in a situation where you have a very high probability of comitting at least one Type I error.

We call the total probability of committing at least one Type I error across multiple tests in the same study the "family-wise" or experiment-wise Type I error. You compute it as:

$P(1 + TypeIerror) = 1\ (1\ )^{\text{number of tests}}$

So if we perform 3 comparisons, the probability of committing at least one Type I error is: _____[32]

And if we perform 10 tests? _____[33]

If this makes you uncomfortable - you're not alone! People often seek to maintain a low risk of drawing any false-positive conclusions, and we can do so simply by lowering .

19.2.7.1 Bonferroni correction

Bonferroni proposed a simple correction of $= {}_{EW}/m$, where ${}_{EW}$ is the desired experiment-wise Type I error rate (e.g., .05), and m is the number of tests.

What alpha level would you use per test if you want to achieve an experiment-wise alpha of .05 and conduct 7 tests? _____[34]

The Bonferroni correction is quite conservative; in other words - although Bonferroni helps you avoid false-positive conclusions, it becomes much harder to detect true effects.

19.2.8 Compare All Groups

Through the ANOVA interface, you can compare all groups to one another. This is equivalent to repeating a regression analysis multiple times, making each category the reference category in turn.

Go to Analyze -> Compare Means -> One Way ANOVA. Enter Feeling as dependent variable and behavior as Factor.

Now, click post-hoc. Note that you can select many different tests. Select LSD; this corresponds to "normal" p-values.

[32] 0.14
[33] 0.4
[34] 0.007142857

The other tests in this menu will either apply a penalty to the p-value, or compute the test statistic in a different way, with the purpose of adjusting for multiple comparisons.

We will manually apply the correction for multiple comparisons instead, because the fact that SPSS performs the correction behind the scenes has a high risk of user error.

Assuming we perform two-sided tests at $= .05$, how many significant differences between group means are there? _[35]

Now, apply a Bonferroni correction to the alpha level. How many tests are you performing? ___[36]

What is the new alpha level? _____[37]

How many comparisons are still significant when using this new alpha level? _[38]

[35] 6
[36] 10
[37] 0.005
[38] 3

20 GLM: Factorial ANOVA

In previous chapters, we have explored various regression techniques, including bivariate linear regression, independent samples t-tests, and analysis of variance (ANOVA). We've also delved into the concept of interaction between binary and continuous predictors. In this chapter, we build upon that prior knowledge when explaining the Factorial ANOVA.

Factorial ANOVA is used to examine the effects of multiple categorical predictors and their interactions on a continuous outcome variable. Despite its historical development as a separate method, factorial ANOVA can be conceptualized as a special case of multiple regression. It combines the concepts of dummy coding and interaction that we've previously encountered. Each factor is represented via dummy coding, and creating interaction terms are computed by multiplying those dummies.

The reason ANOVA is often considered a separate technique is that it has different historical roots from regression, and because of those roots, researchers typically focus on different output when reporting factorial ANOVA versus regression. For example, ANOVA focuses more on variance explained by each factor, and overall tests of the effect of each Factor across all of its levels.

A factorial ANOVA involves two or more factors, each with multiple levels. Each combination of factor levels creates a unique condition or group in the study. For instance, if we have Factor A with 3 levels and Factor B with 2 levels, the factorial design will have a total of 3 Œ 2 = 6 groups.

Factorial designs can be visually represented in a matrix-like structure, where each cell represents a unique combination of factor levels. This representation helps us understand the experimental conditions and the interactions between factors.

In factorial ANOVA, we examine main effects and interaction effects. As in multiple regression, main effects represent the influence of each factor on the dependent variable, controlling for all other predictors. Interaction effects capture how the effects of one factor depend on the levels of another factor.

Because each factor may be represented by multiple dummies, we can't use individual t-tests for the dummies to determine the siginficance of the Factor they belong to. Instead, we use F-tests to test the significance of the effect of a Factor. These F-tests compare the variance explained by the factor to the unexplained variance.

In addition to significance tests, effect size measures help us understand the practical importance of the effects observed in factorial ANOVA. Eta squared and partial eta squared are common effect size measures that quantify the proportion of variance explained by each factor or interaction, relative to the total variance or unexplained variance, respectively. Eta squared is just a different name for R squared; partial eta squared is something different, and typically only reported for ANOVA.

20.1 Lecture

https://www.youtube.com/embed/fBQoxrleFoM

20.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the primary objective of factorial ANOVA? [1]

- (A) To determine the correlation between two continuous variables.
- (B) To analyze the variance between different levels of a single categorical predictor.
- (C) To examine the effects of multiple categorical predictors and their interactions on a continuous outcome variable.
- (D) To compare the means of two independent groups.

Question 2

In a 2x2 factorial design, how many unique conditions or groups are there? [2]

- (A) 3

[1] To examine the effects of multiple categorical predictors and their interactions on a continuous outcome variable.
[2] 4

- (B) 4
- (C) 6
- (D) 2

Question 3

What does a main effect represent in factorial ANOVA? [3]

- (A) The combined effect of all factors on the dependent variable.
- (B) The unexplained variance in the model.
- (C) The influence of a single factor on the dependent variable, controlling for other factors.
- (D) The interaction between two factors on the dependent variable.

Question 4

What does an interaction effect capture in factorial ANOVA? [4]

- (A) The unexplained variance in the model.
- (B) How the effects of one factor depend on the levels of another factor.
- (C) The influence of a single factor on the dependent variable, controlling for other factors.
- (D) The combined effect of all factors on the dependent variable.

Question 5

Which effect size measure quantifies the proportion of variance explained by each factor or interaction relative to the total variance? [5]

- (A) Eta
- (B) Sum of Squares
- (C) Partial eta squared
- (D) Eta squared

[3]The influence of a single factor on the dependent variable, controlling for other factors.
[4]How the effects of one factor depend on the levels of another factor.
[5]Eta squared

Question 6

What does it mean when lines representing different conditions on a means plot cross each other? [6]

- (A) There is a significant interaction between the factors.
- (B) The factors have equal effects on the dependent variable.
- (C) The main effects of the factors are significant.
- (D) The model is poorly specified.

Question 7

In a 3x2 factorial design, how many dummies would be required in total to represent both factors? [7]

- (A) 3
- (B) 6
- (C) 4
- (D) 5

Question 8

Which term describes the variance remaining unexplained by the factors and interactions in factorial ANOVA? [8]

- (A) Explained variance
- (B) Total variance
- (C) Interaction variance
- (D) Residual variance

Question 9

What does a partial eta squared measure in factorial ANOVA? [9]

[6] There is a significant interaction between the factors.
[7] 4
[8] Residual variance
[9] The proportion of variance explained by each factor, controlling for other factors.

- (A) The interaction between two factors.
- (B) The total variance explained by each factor.
- (C) The proportion of variance explained by each factor, controlling for other factors.
- (D) The proportion of variance explained by each factor relative to the unexplained variance.

Show explanations

Question 1
Factorial ANOVA allows us to explore how multiple categorical predictors and their interactions impact a continuous outcome variable.
Question 2
In a 2x2 factorial design, there are 2 levels for Factor A and 2 levels for Factor B, resulting in a total of 2 x 2 = 4 unique conditions.
Question 3
A main effect reflects the impact of a specific factor on the dependent variable while considering the influence of other factors.
Question 4
An interaction effect in factorial ANOVA describes how the effects of one factor change based on the levels of another factor.
Question 5
Eta squared is an effect size measure in factorial ANOVA that indicates the proportion of variance explained by a factor or interaction relative to the total variance.
Question 6
Crossing lines on a means plot indicate a significant interaction between the factors, suggesting that the effect of one factor depends on the levels of another factor.
Question 7
For a 3x2 factorial design, Factor A would require 2 dummies, and Factor B would require 1 dummy. Therefore, the total number of dummies needed is 2 + 1 = 3.
Question 8
Residual variance refers to the unexplained variability in the dependent variable after accounting for the effects of factors and interactions in factorial ANOVA.
Question 9
Partial eta squared quantifies the proportion of variance explained by a specific factor while controlling for the influence of other factors in the model.

20.3 Tutorial

20.3.1 Factorial ANOVA

Consider the following research situation: A psychologist wants to study if and to what extent different behavior of waiters affect the amount of tip money they get, and whether it matters if the behavior is shown by a waiter or waitress.

The researcher distinguishes the following types of behavior: neutral behavior, drawing a smiley on the bill, or making small talk.

They ran a fully crossed experiment with 3 (behaviors) x 2 (gender: waiter or waitress) = 6 conditions. For each condition they collected data for 10 customers who were helped by a waiter showing neutral behavior, 10 helped by a waiter drawing a smiley on the bill, and so forth.

Is the design balanced? [10]

- (A) Yes
- (B) No

What is/are the independent variable(s) in this experiment? [11]

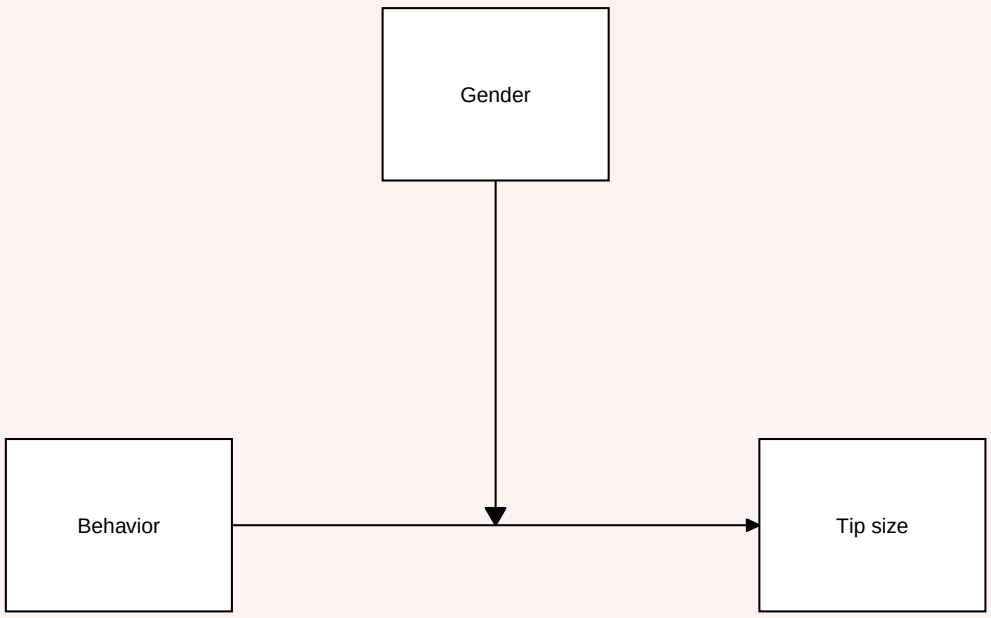
- (A) Behavior
- (B) Gender
- (C) Tip size

Draw the conceptual model of the experiment. Then, check your answer.

[10] Yes
[11] Behavior

Answer

```
library(tidySEM)
library(ggplot2)
lo <- get_layout("", "Gender", "",
                 "Behavior", "", "Tip size", rows = 2)
edges <- data.frame(from = "Behavior", to = "Tip size")
p <- prepare_graph(layout = lo, edges= edges)
plot(p) + geom_segment(aes(x = p$nodes$x[p$nodes$name == "Gender"], xend =  p$nodes$x[p$no
```



As you can see, the dependent variable is Tip money, and the independent variables are Type of behavior and Gender. More specifically. Gender is the moderator, as is expected to influence the relationship between Behavior and Tip money.

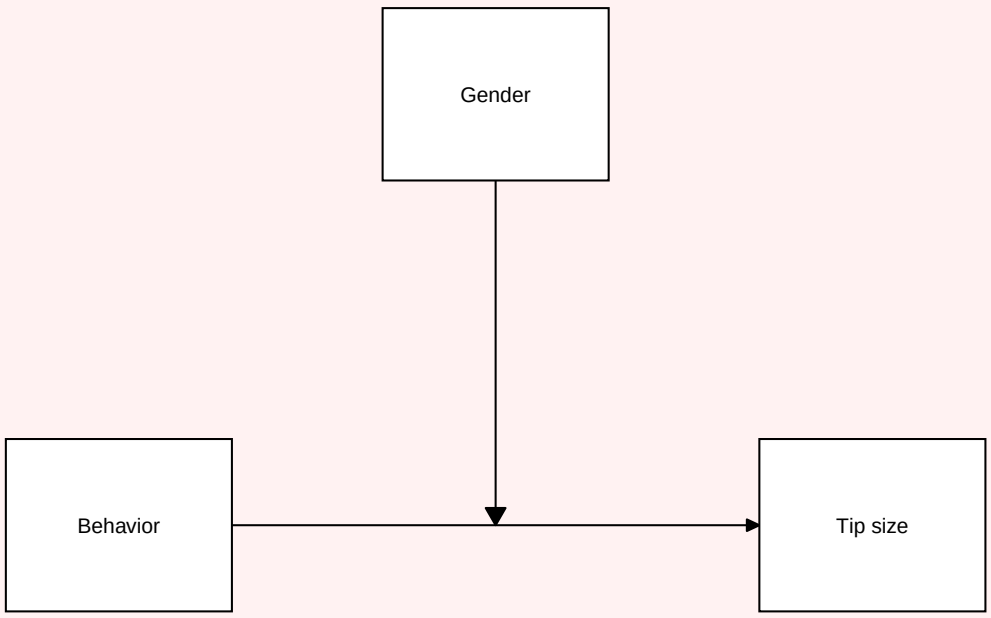
Now, open the dataset WaiterBehavior.sav.

We will run the analysis to find out what the results of the experiment are. We can do so with a factorial ANOVA.

First, create all the dummy variables you need, using syntax. Check your answer below.

Answer

```
library(tidySEM)
library(ggplot2)
lo <- get_layout("", "Gender", "",
                 "Behavior", "", "Tip size", rows = 2)
edges <- data.frame(from = "Behavior", to = "Tip size")
p <- prepare_graph(layout = lo, edges= edges)
plot(p) + geom_segment(aes(x = p$nodes$x[p$nodes$name == "Gender"], xend =  p$nodes$x[p$no
```



As you can see, the dependent variable is Tip money, and the independent variables are Type of behavior and Gender. More specifically. Gender is the moderator, as is expected to influence the relationship between Behavior and Tip money.

20.3.2 Regression with dummies

First, we analyze these data using regression with dummies.

You will need to dummy code both categorical predictors.

Do this in the familiar way, then check your syntax.

Answer

```
RECODE behavior (1=1) (2=0) (3=0) INTO talking.
RECODE behavior (1=0) (2=1) (3=0) INTO smiling.
RECODE behavior (1=0) (2=0) (3=1) INTO neutral.

RECODE Gender (1=1) (2=0) INTO waitress.
RECODE Gender (1=0) (2=1) INTO waiter.
EXECUTE.
```

Estimate a regression model with a main effect for gender and behavior. Create the syntax, then check your answer below.

Answer

Using neutral behavior and waiter as reference categories, the code is:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Tip
  /METHOD=ENTER waitress smiling talking.
```

Now, we want to examine whether there is an interaction between gender and behavior. To do so, we want to specify a "full factorial" model. This simply means that we want to know if there is an interaction effect between the two categorical variables. The way to include an interaction effect is to multiply the predictor. As each predictor is represented by multiple dummies, we have to multiply each dummy for each variable with all dummies from the other variable.

In this case, that means multiplying the `waiter` dummy with the `smiling` and `talking` dummies.

Do this via syntax, then check your work.

Answer

```
COMPUTE smilingXwaitress = smiling*waitress.
COMPUTE talkingXwaitress = talking*waitress.
EXECUTE.
```

Conduct a hierarchical regression analysis that includes these new interaction terms.

What proportion of the total variance in Tip money is explained by the full factorial model? _____[12]

True or false: The full factorial model explains a significant amount of variance. TRUE / FALSE[13]

Is there a significant interaction effect? [14]

- (A) Yes
- (B) No
- (C) Can't tell

Explanation

Even though we see significant interaction terms in the coefficients table, determining whether there is a significant interaction between categorical variables requires more than just "eyeballing" whether those terms are significant or not. You need to perform a nested model test.

Based on the output, complete the following table:

Behavior	Gender	Mean
Neutral	Waiter	4.600
Neutral	Waitress	_____[15]
Smiley	Waiter	5.100
Smiley	Waitress	7.600
Small talk	Waiter	_____[16]
Small talk	Waitress	10.000

Draw a rough plot of these means on a piece of paper.

Put the type of behavior on the x-axis and draw separate lines for waiters and waitresses.

True or false: The graph suggests a potential interaction. TRUE / FALSE[17]

If so, describe the interaction effect (i.e., what can we say about the effect of behavior on amount of tip money for waiters and waitresses?).

[12] 0.659
[13] TRUE
[14] Can't tell
[15] 2.2
[16] 5.5
[17] TRUE

Answer

The lines are not parallel. Hence, also from the graph we see that there is interaction. The effect of type of behavior on the amount of tip money depends on the gender of the waiter/waitress.
It seems that for waiters the tip money does not depend much on the behavior.
For waitresses the effect is stronger; neutral behavior produces the least amount of tip money, whereas small talk is most beneficial.

See if you can answer the following question by yourself:

True or false: The interaction effect is significant. TRUE / FALSE[18]

Answer

To answer this question, you need to perform a nested model test.
The syntax is:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Tip
  /METHOD=ENTER smiling talking waitress
  /METHOD=ENTER smilingXwaitress talkingXwaitress.
```

Note two things: we request `CHANGE` statistics, and we enter all "interaction effects" in a separate step, so we can use the R-squared change test to determine overall significance.

What is the value of the test statistic for the significance of the **interaction effect**? _____[19]

Report the effect, then check your answer.

Answer

There was a significant interaction effect between waiters' sex and behavior, $F(2,54) = 18.315$, $p < .001$.
This means that the effect of Type of behavior on Tip money depends on the Gender of the waiter/waitress.

Out of curiosity - how much variance is explained by the main effects only? _____[20]

[18] TRUE
[19] 18.32
[20] 0.427

20.3.3 ANOVA interface

SPSS also has a dedicated "ANOVA interface". It specifies exactly the same model as we have been investigating, but its output is more focused on information that is typically requested when estimating a model with only categorical predictors and (optionally) interactions between them.

We will now use the ANOVA interface and pay special attention to output it gives us that is not directly available via the Regression interface.

You can either use the graphical user interface, or copy-paste this syntax. Either way, pay particular attention to the following:

- Under Options, ask for Homogeneity tests: `/PRINT=HOMOGENEITY`
- Under Options, ask for Estimates of effect size: `/PRINT=ETASQ`
- Under Options, ask for Parameter Estimates: `/PRINT=PARAMETER`
- Under EM Means (expected marginal means), ask for the means for the interaction, and compare Main effects: `/EMMEANS=TABLES(Behavior*Gender) COMPARE(Behavior) ADJ(LSD)`

```
UNIANOVA Tip BY Behavior Gender
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(Behavior*Gender)
  /PRINT=ETASQ HOMOGENEITY DESCRIPTIVE PARAMETER
  /CRITERIA=ALPHA(.05)
  /EMMEANS=TABLES(Behavior*Gender) COMPARE(Behavior) ADJ(LSD)
  /EMMEANS=TABLES(Behavior*Gender) COMPARE(Gender) ADJ(LSD)
  /DESIGN=Behavior Gender Behavior*Gender.
```

Copy and run this syntax.

Note that the table labelled **Parameter Estimates** is identical to the Coefficients table from the previous regression analysis.

Also note that, for example, the F-value for the explained variance for the interaction between Gender an Behavior is identical to the F-value for the R^2 test you performed in the previous hierarchical regression analysis.

20.3.3.1 Homoscedasticity

Regression models have an assumption of homoscedasticity. If all predictors are categorical, that means that we assume equal variances in all groups.

The ANOVA output offers us a Levene's test for homogeneity of variance. Find it, and answer the following question:

True or false: There is reason to doubt the assumption of homogeneity. TRUE / FALSE[21]

What is the value of the appropriate test statistic? _____[22]

True or false: As a rule, the assumption of homogeneity is more likely to be violated in a factorial ANOVA with an unbalanced design (as compared to a balanced design). TRUE / FALSE[23]

20.3.3.2 Effect size

We can also calculate effect sizes ^{2} or partial ^{2}, for the two factors and the interaction effect separately.

Recall that ^{2} is just the explained variance, R^2. In other words: What proportion of the total sum of squares is explained by the factor of interest?

We obtain ^{2} for Factor A by dividing the sum of squares for factor A, SS_A, by the SST, which is labeled "Corrected total" in the ANOVA output: $^2 = \frac{SS_A}{SST}$

What is ^{2} for the interaction effect? _____[24]

Go back to your previous nested model test, where you determined whether adding the interaction terms led to a significant improvement in explained variance, R^2. Verify that this number is identical to the ^{2} for the interaction! They are the same thing.

Another measure of effect size is the partial ^{2}. It tells us what proportion of the variance not explained by other factors is explained by the factor of interest.

We obtain ^2_p for Factor A by dividing the sum of squares for factor A, SS_A, by SS_A plus the residual sum of squares SSE: $^2_p = \frac{SS_A}{SS_A + SSE}$.

Note that SPSS allows us to request partial ^{2}. We did so by including the line `/PRINT=ETASQ` in our syntax.

What is the value of partial ^{2} for the factor Behavior? _____[25]

---

[21] FALSE
[22] 1.644
[23] TRUE
[24] 0.23
[25] 0.515

20.3.3.3 Pairwise comparisons

Another unique feature of the ANOVA interface is that it gives us all pairwise comparisons when we ask for them using the code `/EMMEANS=TABLES(Behavior*Gender) COMPARE(Behavior) ADJ(LSD)`

Note that this line asks for comparisons of the three levels of behavior for each gender separately. You can also ask for comparisons of the two levels of gender for each behavior separately by specifying `COMPARE(Gender)` instead.

Inspect the table Pairwise Comparisons. The three experimental conditions are compared in a pairwise manner, split over the factor Gender.

For which pair of groups do the means differ significantly from one another (at the 5% level)? [26]

- (A) Waiter: Neutral-Smiley
- (B) Waiter: Small talk-Smiley
- (C) Waitress: Neutral-Smiley
- (D) Waiter: Neutral-Small talk

Look at the note under the table. True or false: the p-values in this table are adjusted for multiple comparisons. TRUE / FALSE[27]

Why do we need to apply a correction like the Bonferroni correction?

Answer

When doing multiple tests on one sample, like doing these pairwise comparisons, the level of risk of a Type I error increases. To correct for this, we should use an adjusted alpha-level, such as the Bonferroni correction.
Although it is possible to ask for SPSS to adjust the p-values in the table, I have expressly not instructed you to do so. The reason for this is that, strictly speaking, Bonferroni is an adjustment of the significance level - not of the p-values. Moreover, sometimes you conduct many more tests in a study aside from the pairwise comparisons performed here. In that case, you might want to include those in your Bonferroni correction too, and SPSS does not know about their existence when it applies a Bonferroni correction to the p-values.
In other words: It is better to set the alpha level yourself, and apply it consistently across all tests in a study.

[26] Waitress: Neutral-Smiley
[27] FALSE

20.3.3.4 Simple Effects test

The significant interaction effect tells us that the effect of behavior differs by gender, or conversely, that gender differences vary across the three behavoirs.

Simple effects analysis allows us to test the significance of these marginal effects.

Simply put: They give us an overall test of the mean differences across levels of one factor, within each level of a different factor.

Consult the table Univariate Tests (still for the contrast `/EMMEANS=TABLES(Behavior*Gender) COMPARE(Behavior) ADJ(LSD)`). This table displays the results of the simple effects tests.

Does the Behavior of waiters have an influence on the amount of tip money people give?

What is the appropriate p-value? _____[28]

Report the simple effect test for waiters and interpret the finding, then check your answer.

Answer

There is no significant difference among the three behaviors within waiters. Hence, for waiters we don't have evidence that the behavior has an effect on average tips received, $F(2,52) = 0.591$, $p = .557$.

Again, consult the table Univariate Tests. This table gives the results of the simple effects tests.

What p-value do we see here? _____[29]

True or false: the type of behavior of waitresses has an influence on the tip people give. TRUE / FALSE[30]

Does it make sense to adjust your behavior as a waiter/waitress if you want to increase your tip?

[31]

- (A) For both waiters and waitresses behavior does not affect amount of tip money received.
- (B) For both waiters and waitresses behavior does affect amount of tip money received.

[28] 0.557
[29] 0.001
[30] TRUE
[31] For waitresses behavior does affect amount of tip money received, but for waiters it does not.

- (C) For waitresses behavior does affect amount of tip money received, but for waiters it does not.

- (D) For waiters behavior does affect amount of tip money received, but for waitresses it does not.

Report your results, then check the answer.

Answer

We do see a significant effect of behavior on average tip money for waitresses, $F(2,52) = 46.385$, $p < .001$. Hence, we have convincing evidence that the type of behavior by waitresses affects the average amount of tips. Results suggest that in order to have high tips, waitresses best can make small talk.

20.3.4 Optional: do it yourself!

Open the datafile `hiking.sav`. The data file also contains data on weather.

Examine the effect of weather and behavior and their potential interaction, with feelings as the dependent variable.

What is the explained variance of the main effects and interaction effects together? _____[32]

True or false: the explained variance of the whole model is significant. TRUE / FALSE[33]

True or false: the interaction effect is significant. TRUE / FALSE[34]

Request a plot from SPSS that allows you to describe what the effects look like.

Describe the trends in your own words, then check your answer.

Answer

The lines in the plot are not parallel, also pointing to an interaction effect.

Perform a simple effects analysis of the effect of behavior by weather.

Report your findings and conclusion, then check your answer.

[32] 0.239
[33] TRUE
[34] TRUE

Answer

We see a significant effect of behavior when the weather was good, $F(4,90) = 3.864$, $p = .006$, while the effect of behavior is not significant when the weather was bad, $F(4,90) = 1.320$, $p = .269$.
Results suggest that when the weather is good, joking and singing significantly improves the participants' feelings about the guide.
When the weather is bad, the behavior of the guide does not have much influence.

21 GLM: ANCOVA

ANCOVA, which stands for Analysis of Covariance, is an extension of the concepts we've covered in bivariate linear regression and multiple regression. It is essentially a multiple regression with a categorical predictor and one or more continuous predictors. What's "special" about this technique is that it is commonly used when the predictor of interest is that categorical variable, and the continuous predictor(s) are so-called "covariates": predictors that are only included to improve our estimate of the effect of the categorical predictor of interest.

You will often see this technique used to analyze data from experiments or "natural experiments", where participants self-select into a treatment group.

While ANCOVA is a useful technique, it comes with some serious pitfalls: any time control variables are used, we are making assumptions about causality. If these assumptions are incorrect, our estimates of the effect of interest will be (severely) biased.

21.0.1 Covariates and Their Role

Covariates are variables that have a relationship with the dependent variable but are not the primary focus of the study. They are often referred to as control variables, as they help control for unwanted variability and improve the precision of the analysis. Examples of common covariates include age, gender, education level, or any other variables that might influence the dependent variable.

In terms of causality, it's crucial to consider the relationships between covariates, predictors, and the outcome variable. Control variables should ideally be confounders – variables that influence both the predictor of interest and the outcome. It's essential to avoid controlling for colliders, which are variables *caused by* both the predictor and the outcome. A thorough understanding of causal relationships is crucial for proper interpretation.

One reason why researchers use control variables in ANCOVA is because they reduce the residual variance in the outcome variable, which in turn increases the power to detect the effect of the predictor of interest. Another reason to use covariates is when the goal is making causal inferences, especially in quasi-experimental designs. The proper selection of covariates that enable causal inference requires careful consideration and is beyond the scope of this course.

21.0.2 Good, Neutral, and Bad Controls

Covariates fall into different categories based on their relationship with the predictor of interest and the outcome. An example of a good control is a confounder: a variable that causes both the predictor and the outcome. These need to be controlled to avoid spurious relationships. An example of a neutral control is a covariate that is unrelated to the predictor but can reduce error variance in the outcome, thereby increasing statistical power. Bad controls, on the other hand, can introduce biases, such as collider bias (controlling for an outcome of predictor and outcome), case control bias (controlling for an outcome of the outcome), or overcontrol bias (controlling for a mediator of the effect of the focal predictor on the outcome).

One crucial insight is that in randomized controlled experiments, the random assignment of participants to different groups breaks the relationship between confounders and the treatment variable. This makes control variables related to the confounders unnecessary. Controlling for them could even introduce bias into the analysis.

21.0.3 Calculating Adjusted Means

Adjusting for covariates involves calculating adjusted means – the means that groups would have had if they scored equally on the covariate. There are two ways to mathematically calculate the adjusted means. One way is to fill in the regression equation for the desired value of the covariate.

The other way is to calculate the adjusted means from the group means:

$$Y_g^{adj} = Y_g \; b(X_g \; X)$$

Where:

- Y_g^{adj}: Adjusted mean of the outcome for group g
- Y_g: Unadjusted mean of the outcome for group g
- b: Regression coefficient of the covariate
- X_g: Group mean of covariate X
- X: Overall mean of covariate X

In sum, ANCOVA is a different name for regression with a categorical predictor of interest, and continuous predictor(s) that are included to improve our estimate of the effect of the predictor of interest. ANCOVA can enhance statistical power and help make more accurate (potentiall causal) inferences. However, the careful selection of covariates and an understanding of causal relationships are paramount to its proper implementation and interpretation.

21.1 Lecture

https://www.youtube.com/embed/MewqUBfQYok

21.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is a valid reason for using covariates in ANCOVA? [1]

- (A) To increase statistical power
- (B) To replace categorical predictors
- (C) To create more complex models
- (D) To increase the total variance

Question 2

What is an example of a 'good control' variable in ANCOVA? [2]

- (A) A variable unrelated to the predictor
- (B) A confounder
- (C) A variable related to the predictor
- (D) A variable affected by the outcome

Question 3

Which of the following is an example of a 'bad control' variable in ANCOVA? [3]

[1] To increase statistical power
[2] A confounder
[3] A variable affected by the outcome

- (A) A variable unrelated to the predictor or outcome
- (B) A variable that causes the predictor
- (C) A variable affected by the outcome
- (D) A confounder

Question 4

In a randomized controlled experiment, are there benefits to controling for any observed differences between the two experimental groups? [4]

- (A) No, random assignment breaks the predictor's relationships with confounders
- (B) Yes, controlling for relevant covariates is essential
- (C) Only if the covariates are significantly related to the outcome
- (D) Only if the covariates are significantly related to the predictor

Question 5

What is the purpose of calculating adjusted means in ANCOVA? [5]

- (A) To replace the original means
- (B) To improve the model's fit
- (C) To account for covariate effects
- (D) To create new variables

Question 6

Which of the following is a key factor in choosing appropriate covariates for ANCOVA? [6]

- (A) Number of covariates
- (B) Causal relationships
- (C) Covariates with the highest correlations

[4]No, random assignment breaks the predictor's relationships with confounders
[5]To account for covariate effects
[6]Causal relationships

- (D) Covariates with the smallest p-values

Question 7

What is the distinguishing feature of ANCOVA relative to multiple regression? [7]

- (A) Multiple regression includes only continuous predictors
- (B) Multiple regression includes only covariates
- (C) ANCOVA always includes a categorical predictor and control variable(s)
- (D) ANCOVA controls for confounding variables

Question 8

In which of these situations should you avoid controlling a particular covariate in ANCOVA? [8]

- (A) When the covariate is a mediator
- (B) When the covariate is irrelevant
- (C) When the covariate is a confounder
- (D) When the covariate is perfectly correlated with another covariate

Question 9

An ANCOVA model compares statistics grades for two classes of students, controlling for number of hours studied per week. The regression model is Grade $= 1.27 + 0.50\ D_{class2} + .30\ Hours$. The unadjusted mean grades are 5.2 for class 1 and 7.3 for class 2. The mean hours studied were 21.56 for class 1 and 30.23 for class 2. The overall mean hours studied was 25.90. What are the adjusted means? [9]

- (A) 6.50 and 6.00
- (B) 9.35 and 2.97
- (C) -1.27 and -1.77

[7] ANCOVA always includes a categorical predictor and control variable(s)
[8] When the covariate is a mediator
[9] 6.50 and 6.00

Show explanations

Question 1
Covariates can be used in ANCOVA to reduce error variance caused by factors other than the predictor of interest.
Question 2
A confounder is a 'good control' variable in ANCOVA – a variable that influences both the predictor of interest and the outcome.
Question 3
A variable that is affected by the outcome is a 'bad control' variable in ANCOVA and can introduce biases.
Question 4
In a randomized controlled experiment, random assignment breaks the relationships between confounders and the treatment. Therefore, controlling for covariates related to the predictor is unnecessary.
Question 5
Adjusted means in ANCOVA are controlled for the effects of covariates, allowing us to understand how groups would differ on the outcome variable if they had equal scores on the covariate.
Question 6
Causal relationships are crucial when selecting covariates for ANCOVA. It's important to choose covariates that are related to the outcome and predictor in a meaningful way.
Question 7
ANCOVA is one example of the broader category of models known as multiple regression; it is characterized by having a categorical predictor of interest and (continuous) covariates.
Question 8
Avoid controlling for covariates that are mediators, as controlling for mediators could lead to overcontrol bias and distort the relationships between the predictor and outcome.
Question 9
Use the formula $Y_g^{adj} = Y_g \ b(X_g \ X)$

21.3 Tutorial

21.4 Bivariate Regression (RECAP)

Researchers are interested in the relationship between age and depression.

They hypothesized that older people are more vulnerable to depressive thoughts than younger people.

To test their research hypothesis, they collected data in a random sample of 164 persons from the general population. Open the dataset `HADShealthyGroup.sav`.

Run a linear regression analysis using age as the independent variable and depression as the dependent variable.

Proceed as follows:

Navigate to Analyze > Regression > Linear

Select the correct dependent and independent variable.

Paste and run the syntax.

How much of the total variance in Depression is explained by Age? _____[10]

What can you say about the effect size? Would you say it's a lot?

Answer

To me, 2% of explained variance does not seem like a lot. There are probably better predictors of depression (i.e., predictors that explain more of the variance in depression).

True or false: The explained variance is significant at the 10% level. TRUE / FALSE[11]

Answer

To conclude anything about the significance of the proportion explain variance of this model, we can look at the ANOVA table or ask SPSS to show the R2-change. This shows us that the explained variance in this model is not significant compared to an empty model (i.e. including no predictors), $F(1,140) = 2.836$, $p = .094$.

Write down the estimated regression line using the unstandardized coefficients.

Y'= _____[12] + _____[13] *Age

How can we interpret the constant?

[14]

- (A) The average level of depression.
- (B) The average age in the sample.

[10] 0.02
[11] FALSE
[12] 1.89
[13] 0.02
[14] The predicted level of depression when age would be 0.

- (C) The predicted level of depression for the average age.
- (D) The predicted level of depression when age would be 0.

Consult the table with the coefficients again

True or false: We can conclude from this table that the effect of age is significant at the 10% level. TRUE / FALSE[15]

Answer

To conclude whether the effect of Age on Depression is significant, we look at the t-test for the estimated coefficient.
We should conclude that the effect of Age on Depression is significant when using =.10, t(140) = 1.684, p = .094.

One of the assumptions of bivariate regression analysis is that the relationship between the independent and dependent variable is linear.

State in your own words what this assumption entails.

Answer

The assumption of a linear relationship entails that the relationship between the variables can be described with a straight line.

How would you evaluate the assumption of linearity graphically? Do it for the data at hand.

True or false: The relationship is linear. TRUE / FALSE[16]

If the assumption is not met, speculate about other possible relationships between age and depression.

Answer

The scatter plot does not have the shape of a cigar, so it does not unambiguously suggest a linear relationship. Thus, you may doubt whether the relationship between age and depression is best described by a linear model. Perhaps the relationship is quadratic. Especially persons in middle ages may be vulnerable to depressive thoughts. Next to the plot, you find both the estimated linear trend and a non-linear trend. It seems that the quadratic curve fits better with the data. Moreover, the quadratic model explains 6% of the variance, whereas the linear model only 2%.

Run a regression analysis using Age as the independent variable and Anxiety as the dependent variable.

[15] TRUE
[16] FALSE

Summarize the results.

Include in your answer the proportion of explained variance (R-square), a description of the effect based on the estimated regression coefficients, and evaluate the significance of the effect of age on anxiety.

Answer

The proportion explained variance in Anxiety by Age is .071. The explained variance in this model is not significant compared to an empty model (i.e. including no predictors), F(1,140) = 0.710, p = .401.
The effect is Age on Anxiety is negative (= -0.013), meaning that anxiety decreases with age. However, the effect of Age on Anxiety is not significant when using =.05, t(140) = -0.842, p = .401.

21.5 ANOVA (RECAP)

Consider the following hypothetical research situation...

Researchers are interested in effects of stereotyping on cognitive performance. For their research they performed a quasi-experiment. They selected three schools and asked girls from eight grade to do a math test. However, the teacher in School A says that boys do particularly well on the test (i.e., negative stereotyping for girls). In School B the teachers says that girls do particularly well on the test (i.e., positive stereotyping for girls). In the third school, the teacher gives gender-neutral information (control group).

Afterwards the researchers compare the average math grades across the three groups. Because the schools may also differ in the student population, researchers also measured scholastic aptitude and use that as a covariate in the analysis.

Open the dataset `stereotyping.sav`.

In your own words, explain what a covariate is and give two examples of covariates that should/can be included in neuro-psychological research.

Answer

Covariates are "nuissance" variables that we are not directly interested in, but that allow us to better estimate the effect of another variable of interest.
This is related to causal inference.
We must thus justify our covariates by reference to their putative causal role with relation to our predictor and outcome.
A covariate that *causes* both our predictor and our outcome is a *confounder*, and should be controlled for. This sometimes happens in "natural experiments", where the factor is not randomly assigned to participants.
A covariate that *causes* our outcome, but is unrelated to our predictor, can be controlled for to reduce error variance in the outcome and increase statistical power of the effect of interest. This typically happens in randomized controlled experiments, but it may also happen in natural experiments.
As a counter-example: A variable that *is caused by* our predictor and our outcome is a collider, and should never be controlled for! This will bias the effect of our predictor.

Mention two often-used covariates in research.

Answer

Gender and Age are two covariates that are often used in research.

Let's start analyzing the math scores using an ANOVA, thus ignoring any covariates for now.

- Compute the means across the three groups (Analyze à Compare means à Means.
- Select MATH for the dependent list and STEREO as the independent.
- Paste and run the syntax.
- Inspect the table that displays the mean differences between the groups.

What is the first impression of stereotyping that we have from the mean differences?

Answer

The table shows that no stereotyping results in the lowest mean score on the math test. Positive stereotyping results in the highest mean score, and negative stereotyping is in between.

Run an ANOVA: (Analyze -> general linear model -> univariate). Select MATH as the dependent variable and STEREO as fixed factor.

Write down the null and alternative hypothesis of the ANOVA, then check your answer.

Answer

H0: $1 = 2 = 3$ H1: not $1 = 2 = 3$

True or false: The effect of stereotyping is significant (use =0.05). TRUE / FALSE[17]

Answer

Yes, the F-test for the effect of Stereotyping on Math performance is significant, F(2,27) = 5.614, p = .009. Hence, we reject H0.
We have convincing evidence that, also at the population level, the means differ.
In other words, we have convincing evidence that the mean differences in math performance are not the result of sampling fluctuations but reflect true differences due to the manipulation (i.e. stereotyping).

How large is the R-square? _____[18]

How do you interpretat this value of the R-square?

Answer

The R2 is 0.294.
This means that 29.4% of the variance in Math performance is explained by the Stereotyping.

The R-square should be equal to:

[19]

- (A) The ratio of the mean square for STEREO to the (corrected) total mean square.
- (B) The ratio of the sum of squares for STEREO to the error sum of squares.
- (C) The ratio of the mean square for STEREO to the error mean square.
- (D) The ratio of the sum of squares for STEREO to the (corrected) total sum of squares.

Test whether there is an effect of stereotyping (regardless of whether it is positive or negative stereotyping).

True or false: The effect of stereotyping (regardless of whether it is positive or negative stereotyping) is significant. TRUE / FALSE[20]

Report Levene's test, significance of the contrast that you tested, and an interpretation of the difference between the two means.

[17] TRUE
[18] 0.294
[19] The ratio of the sum of squares for STEREO to the (corrected) total sum of squares.
[20] TRUE

Answer

The Levene's test is not significant (p = .805), so there is no evidence for violation of the assumption of homoscedasticity.

Test the mean math score for each experimental group against the control group; that is, you have to test two planned contrasts.

Are the means different when tested at an experiment-wise alpha of .05 and using a Bonferroni corrected alpha per test? Substantiate your answer.

Answer

The contrast is significant at = .05.
These results suggest that stereotyping (positive/negative) results in better math performance than non-stereotyping.

21.5.1 ANCOVA

The next step is to add scholastic aptitude as a covariate in our analysis.

In the lecture, we considered two situations in which ANCOVA is used; one in which the covariate was not related to the grouping variable, and one in which the covariate is associated.

Which situation do we have in this study? To answer the question, you need to ask for some additional statistics in SPSS.

Answer

We can check if the covariate is associated with the experimental factor by inspecting the means of the experimental groups on the covariate, or by running a one-way ANOVA. The covariate is associated with the grouping variable, and thus mean differences in math between the three groups may be confounded by mean differences in scholastic aptitude between the three groups.

21.5.1.1 Assumption or not?

Some texts state that ANCOVA has an additional assumption, beyond those of multiple regression. This "extra assumption" is the assumption of "homogeneity in regression slopes".

The assumption of homogeneity in regression slopes states that the within-group effect of the covariate is the same across groups. That is, the covariate does not interact with the grouping variable.

For example, in this assignment, the assumption would imply that the effect of scholastic aptitude is independent of the experimental condition.

But: ANCOVA is just a special name for multiple regression with a categorical predictor of interest and control variables that we're not interested in ("nuissance variables"). If ANCOVA is multiple regression, then how can it have differen assumptions than multiple regression?

The answer is that ANCOVA does **not** have different assumptions, but if we were to allow for interaction between the factor and covariate, we would simply no longer call the resulting model an ANCOVA.

So instead of saying "ANCOVA has an assumption of homogeneity in regression slopes", we could say: "it is conventional to call a model ANCOVA if it contains a factor and some control variables, but no interaction". Of course we can add interaction terms in such a model - but then we just call it multiple regression with an interaction, not ANCOVA.

Omitting an interaction between the factor and the covariate means that we force the within-group effect of the covariate to be the same across levels of the factor. In this study, that means that we do not allow the effect of scholastic aptitude to depend on the experimental condition.

Before we carry out the actual ANCOVA, we can check whether this model is correctly specified, or whether we are missing a significant interaction. We can check whether there is a significant interaction effect using the following syntax:

```
UNIANOVA SA BY STEREO WITH MATH
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE PARAMETER HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=STEREO MATH MATH*STEREO.
```

Copy and run the syntax. Go to the table "Tests of Between-Subjects Effects". Take a look at the row STEREO*MATH.

True or false: There is a significant interaction effect. TRUE / FALSE[21]

If there is a significant interaction effect, what do we do?

[21] FALSE

Answer

Remember we can check for a significant interaction effect for two reasons:

1. Our hypothesis is about the interaction effect
2. As an assumption check for correct model specification (i.e., we're interested in main effects, but we want to make sure that we're not ignoring an interaction when we do). This is similar to checking for linearity.

In the first case, we should not be conducting ANCOVA at all; we should start with multiple regression with interaction, because the interaction effect is our effect of interest. In the second case, we can do two things: We can change our analysis based on the results of the "assumption check" - but such data-dependent decisions increase the risk of overfitting and Type I errors. Alternatively, we can just report the results of our planned ANCOVA analysis, but mention in the discussion that we observed a significant interaction, which means that the model may have been misspecified. We can also report results from both models, and see if/how the conclusions change if we allow for interaction (this is called a *sensitivity analysis*).

21.5.1.2 Back to ANCOVA

Let's run an ANCOVA, proceed as follows:

- Navigate to Analyze > General linear model ? Univariate
- Select MATH as the dependent variable, STEREO as fixed factor, and SA as the covariate.
- Also, via OPTIONS ask for the Parameter estimates. Paste and run the syntax.

Consider the Tests of Between Subjects Effects table (henceforth referred to as the "ANCOVA table") and the FF-test for the grouping factor (STEREO).

What's the p-value for the overall test of model fit? _____[22]

What conclusions can be drawn from the F-test?

[22] 0.117

Answer

The F-test is a test of the effect of Stereotyping on Math performance controlled for Scholastic aptitude. Conceptually, it tests whether differences in the adjusted means in Math performance are significant. Adjusted means are the means we would expect if the group had an average level of Scholastic aptitude.
In other words, it tests the differences in the hypothetical situation we would have had three groups that had exactly the same level of Scholastic aptitude.
The F-test is not significant, F(2,26) = 2.333, p = .117, which means that, controlled for Scholastic aptitude, we don't have convincing evidence that Stereotyping had an effect on Match performance.

Compare the results ANOVA and ANCOVA. What important difference do we see and how would you explain those?

Answer

The ANOVA suggested a significant effect, whereas once controlled for Scholastic aptitude (ANCOVA) the effect was no longer significant.
Thus, the mean differences between the experimental groups we saw before were indeed confounded with differences in Scholastic aptitude!

Consult the table parameter estimates.

What is the regression slope for scholastic aptitude? _____[23]

Explain the meaning of estimated parameter for scholastic aptitude.

Answer

The parameter estimate for Scholastic aptitude is 0.470. The effect is significant when using = .05.
It is the pooled within-group regression effect of Scholastic aptitude on Math performance, controlled for Stereotyping.
Thus, if Scholastic aptitude increases by one unit, the predicted Math score increases by .470 units, while controlling for Stereotyping.

Based on the Parameters Estimates and the group-specific means, compute the adjusted group means (for each of the groups!) on MATH for an average scholastic aptitude.

What is the adjusted group mean for the group that received the Negative Stereotype manipulation? _____[24]

$Y_k^{adj} = Y_k \; b_w(X_k \; X)$

[23] 0.47
[24] 6.359

To use the formula, you need to know the group means on MATH (you computed them before), you have to know the group means and overall mean SA (you can compute them via means), and the regression effect which is given in the table with parameter estimates.

Check your adjusted means against this answer model:

Answer

Group	MATH ($\bar{Y}_k$)	SA ($\bar{X}_k$)	($\bar{X}_k - \bar{X}$)	$\bar{Y}_k^{adj}$
1	6.5	6.5	0.3	6.359
2	7.5	6.9	0.7	7.171
3	5.1	5.2	-1	5.570
	6.367	*6.2* (= $\bar{X}$)		

For example, for group 2 (k=2) we have: $7.5 - 0.470 \times 0.7 = 7.171$

Rerun the ANCOVA. Now via OPTIONS also ask for the estimated means. You do so by selecting stereo in the list of Display Means for (at the top of the menu). Look in the table Estimated Marginal Means and verify your answer to the previous question.

Write down in your own words – and as precise as possible – the meaning of adjusted means.

Answer

The estimated marginal means (i.e., the adjusted means) are the group means if all groups would've had an average of 6.20 on the covariate.

Finally, we want to look at several effect size estimates.

How much of the variance in Math do SA and STEREO explain? _____[25]

Controlled for SA, how much of the remaining variance in Math does STEREO explain? Use the formula mentioned in the lecture slides to calculate the partial ^{2}: _____[26]

Verify your answer by running the ANCOVA again. Now, in options, select the box Estimates if effect size.

True or false: SPSS reports the same partial 2. TRUE / FALSE[27]

Could you summarize your findings of the ANCOVA in a few brief sentences?

Mention the significance tests, (un)adjusted means and the effect size estimates.

[25] 0.421
[26] 0.152
[27] TRUE

22 GLM: Repeated Measures ANOVA

Repeated Measures ANOVA is used to analyze data collected in within-participants designs, where the same outcome measure is collected from the same individuals multiple times.

A study design in which the same participants are assessed repeatedly is called a *Within-Participants Design*. Within-participants designs have distinct advantages in comparison to between-participants designs. In these designs, participants serve as their own control, eliminating variability due to individual differences from the error term. This intrinsic control enhances statistical power and efficiency. These designs are used, for example, in longitudinal studies, test-retest designs, diary studies, and repeated physiological assessments.

While within-participants designs offer significant advantages, they also present challenges that require careful consideration. Order effects, where the sequence of experimental conditions influences results, are a common concern. Differential order effects, where the influence of order varies across different sequences, can further complicate data interpretation. An example of a differential order effect is wen the effect of a drug administered before a placebo condition persists into the placebo phase of the experiment. To mitigate order effects, researchers often employ the Latin square design. This experimental design ensures each condition appears once in every position within the order, thus minimizing the influence of sequence on outcomes. By controlling for order effects, researchers enhance the internal validity of their experiments.

Beyond order effects, within-participants designs are also affected by learning- and historical effects. A learning effect occurs when participants' increasing familiarity with questionnaires affects their subsequent responding. Historical effects occurr when external events happen during the study, and influence participants' responses. Finally, the effect of time is often confounded with the effect of experimental conditions.

22.0.1 Two Repeated Measurements

The paired samples t-test is suitable for scenarios where participants are measured before and after an intervention. This technique simply analyzes the difference score between pretest and posttest scores.

22.0.2 More Than Two Measurements

For scenarios with more than two repeated measurements, there are two potential solutions: the linear mixed model, and the multivariate approach. The linear mixed model, treats all repeated measurements as a single variable with multiple observations per participant. Thus, if one participant gave four repeated measurements, we would have four rows in the data for that participant. The multivariate approach treats the repeated measurements as correlated outcomes. Each measurement occasion is analyzed while controlling for the other measurement occasions.

22.0.3 Sphericity Assumption

The linear mixed model assumes *sphericity*, which is analogous to the assumption of homogeneity of error variance. Sphericity implies that the variances of the differences between all combinations of repeated measures are equal.

If you do not, or can not, assume sphericity, you can use a corrected test for the linear mixed model, or switch to the multivariate approach.

22.0.4 Mixed Designs

A mixed design involves both within-participants and between-participants factors. This factorial design allows researchers to examine interactions between these factors, such as the interplay between time and exposure conditions. Post hoc analyses can be used to understand the direction and significance of these interactions.

22.1 Lecture

https://www.youtube.com/embed/K1elAF1SFrM

22.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

In Repeated Measures ANOVA, what type of experimental design involves the same participants being exposed to multiple conditions? [1]

- (A) Longitudinal design
- (B) Within-participants design
- (C) Cross-sectional design
- (D) Between-participants design

Question 2

Which of these is NOT a methodological concern in within-participants designs? [2]

- (A) Historical effects
- (B) Order effects
- (C) Learning effects
- (D) Interaction effects

Question 3

How does the Latin square design address order effects in experiments? [3]

- (A) Manipulates the order of conditions for experimental purposes
- (B) Controls for order effects by ensuring each condition appears once in every position
- (C) Selects a single condition for all participants
- (D) Randomly assigns participants to different conditions

Question 4

What is the primary advantage of within-participants designs in comparison to between-participants designs? [4]

- (A) Lower cost

[1] Within-participants design
[2] Interaction effects
[3] Controls for order effects by ensuring each condition appears once in every position
[4] Elimination of variability due to individual differences from the error term

- (B) Greater external validity
- (C) Larger sample sizes
- (D) Elimination of variability due to individual differences from the error term

Question 5

What statistical technique is commonly used to analyze data collected in within-participants designs with two repeated measurements? [5]

- (A) Paired samples t-test
- (B) Independent samples t-test
- (C) Chi-square test
- (D) Analysis of variance

Question 6

What assumption of the general linear model is violated when analyzing data with repeated measurements? [6]

- (A) Independence of errors
- (B) Linearity of relationships
- (C) Normal distribution of errors
- (D) Homogeneity of variances

Question 7

What is the purpose of using a multivariate approach in analyzing data with more than two repeated measurements? [7]

- (A) To perform a test that is robust to violations of sphericity
- (B) To calculate mean differences between measurements
- (C) To assess the order effects in the data

[5] Paired samples t-test
[6] Independence of errors
[7] To perform a test that is robust to violations of sphericity

- (D) To ignore the repeated measurements

Question 8

What is the key assumption in the multivariate approach for analyzing data with repeated measurements? [8]

- (A) Homoscedasticity assumption
- (B) Sphericity assumption
- (C) Independence assumption
- (D) Normality assumption

Question 9

In mixed design ANOVA, what type of factors are considered? [9]

- (A) Categorical and continuous factors
- (B) Both within-participants and between-participants factors
- (C) Only within-participants factors
- (D) Only between-participants factors

Question 10

What does the term 'sphericity' refer to in the context of repeated measures ANOVA? [10]

- (A) The shape of the distribution of the outcomes
- (B) The distribution of the residuals
- (C) Equal variances and correlations of differences scores of all pairs of repeated measurements
- (D) The distribution of the predictor variables

[8] Sphericity assumption
[9] Both within-participants and between-participants factors
[10] Equal variances and correlations of differences scores of all pairs of repeated measurements

Show explanations

Question 1
In a within-participants design, the same participants are exposed to different conditions, allowing for the comparison of outcomes within the same individuals.
Question 2
Order effects refer to the potential impact of the sequence in which conditions are presented on the observed outcomes. Learning effects imply that participants respond to a questionnaire differently when they already know the questions. Historical effects mean that something external happens while you are running the experiment. Interaction effects are a statistical term.
Question 3
The Latin square design helps control for order effects by ensuring that each condition appears in each position within the order an equal number of times.
Question 4
Within-participants designs allow each participant to serve as their own control, effectively removing variability due to individual differences from the error term. The cost is often indeed lower, but that's not the primary advantage.
Question 5
The paired samples t-test is used to analyze the differences between two related measurements, such as pretest and posttest scores.
Question 6
Repeated measurements within the same individuals violate the assumption of independence of errors, as observations from the same participant are likely to be correlated.
Question 7
A multivariate approach is robust to the assumption of sphericity because it considers the interrelationships between different repeated measurements, treating them as correlated outcomes.
Question 8
The sphericity assumption assumes that the variances and correlations among all pairs of repeated measurements are equal, which is essential for accurate results.
Question 9
Mixed design ANOVA involves the consideration of both within-participants and between-participants factors to understand the interactions between these factors on the outcomes.
Question 10
Sphericity refers to the assumption that the variances and correlations among all difference scores between pairs of repeated measurements are equal, which is crucial for accurate analysis.

22.3 Tutorial

22.3.1 Repeated Measures ANOVA

In this tutorial, we will explore how to perform a repeated-measures ANOVA using SPSS to assess the effect of repeated measurements of depression symptoms in a sample of military veterans. The primary objective is to determine whether there are significant changes in depression symptom scores across multiple time points.

Load the dataset called `depression.sav` containing depression symptom scores at different time points for each participant.

3. Click on "Analyze" in the top menu and select "General Linear Model" and then "Repeated Measures."

22.3.1.1 Defining the Within-Subjects Factor

1. In the "Repeated Measures" dialog box, name your within-subjects factor as "time."
2. Specify the number of levels as 4 (since there are four repeated measurements).
3. Click the "Add" button.

22.3.1.2 Defining Within-Subjects Variables

1. Click on the "Define" button to configure within-subjects variables.
2. In the "Repeated Measures" dialog box, move the variables corresponding to each time point (e.g., scl1, scl2, scl3, scl4) to the "Within-Subjects Variables" box while maintaining their correct order.

Configuring Options

1. Click the "Options" button.
2. Check the boxes for "Descriptive statistics" and "Estimate of effect size."
3. Click "Continue."

Running the Test

1. Click "OK" to run the repeated-measures ANOVA.
2. The result will appear in the Output Viewer.

Interpreting the Result

Descriptive Statistics

The descriptive statistics provide insight into the direction of any potential effect. The means comparison shows the average depression symptom scores at different time points.

True or false: There is an increase in symptoms over time. TRUE / FALSE[11]

Assumption of Sphericity

SPSS tests assumption of sphericity using Mauchly's test of sphericity.

True or false: In this analysis, the assumption of sphericity is met. TRUE / FALSE[12]

True or false: According to the Huyn-Feldt estimate of epsilon, the deviation from sphericity is small. TRUE / FALSE[13]

Let's assume sphericity for now. Choose the appropriate test and correction based on this assumption.

What is the appropriate F-value for the chosen test? _____[14]

What is the appropriate df for the chosen test? _[15]

22.3.2 Pairwise Comparisons

Examine the table of pairwise comparisons.

Which difference is smallest? [16]

- (A) T1 v T2
- (B) T2 v T3
- (C) T3 v T4
- (D) T2 v T4

[11] TRUE
[12] FALSE
[13] TRUE
[14] 7.29
[15] 3
[16] T2 v T3

If you were to use Bonferroni correction to control for multiple comparisons, you would divide the experiment-wise alpha level by the number of comparisons. How many comparisons are you making here? _[17]

Report your results. Make sure to reference both the RM-ANOVA test, and post hoc comparisons with Bonferroni correction. Then, check your answer.

Answer

"A repeated-measures ANOVA revealed a significant effect of time on depression symptom scores, $F(3, 2931) = 7.29$, $p < .001$. For post hoc pairwise comparisons, we applied a Bonferroni correction. Since there are 6 comparisons between 4 time points, we established the alpha level as $.05/6 = .008$. Using this alpha level, we found that the mean depression symptom score increased significantly from T1 to T3 (Mean difference $= .29$, $p = .003$), and from T1 to T4 (Mean difference $= .41$, $p < .001$). These results suggest that depression symptoms increased significantly over time for the military veteran sample."

[17] 6

23 Reliability and Validity

Questionnaires are widely used in the social sciences to measure a variety of constructs, including self-reported behavior, beliefs, knowledge, opinions, values, attitudes, and attributes. Crucially, oftentimes the same construct is measured with multiple questions. Researchers may want to know whether these questions do a good job at measuring the construct of interest. This is the topic we will explore this week.

Recall that a construct is an abstract feature of interest within a population, such as intelligence, perseverance, or education. Some constructs can be either observed (directly measured); other constructs are latent (measured indirectly through observed indicators, which can be questions).

Examples of observed constructs are height and weight. Latent constructs cannot be directly measured and require observed indicators to capture their underlying meaning. Latent constructs are commonly used in social sciences to measure attitudes, beliefs, opinions, and other complex attributes. The items, more often than not, are questions in a questionnaire - although they could also be tests, performance assessments, behavioral observations, multiple choice questions, puzzles, et cetera.

23.0.1 Classical Test Theory

Classical Test Theory posits that observed test scores are a function of the true score on the latent construct, plus measurement error. Different sources of measurement error can influence observed scores, including instrument properties and individual factors.

23.0.2 Reliability

Reliability refers to the consistency or stability of test scores over time or across different measurement occasions. It is related to the proportion of true score variance to total score variance. Reliability can be estimated using methods such as test-retest reliability (consistency across repeated measurements), internal consistency (consistency across items within a test), and inter-rater reliability (consistency across different raters).

23.0.2.1 Test-Retest Reliability

Test-retest reliability involves administering the same test to the same participants on two separate occasions and calculating the correlation between their scores. Test-retest reliability is suitable for stable traits and can provide insight into the stability of a construct over time. However, learning effects, memory effects, and change over time need to be considered when determining the appropriate interval between test administrations.

23.0.2.2 Internal Consistency

Internal consistency measures the association among items within a test. It can be estimated using methods such as split halves or Cronbach's alpha. Split halves involve dividing the test into two halves and correlating the scores between them. Cronbach's alpha is a measure of internal consistency that indicates how closely related items are to each other within a scale. It is affected by the number of items, their average covariance, and their average variance. Note that this means you can artificially inflate Cronbach's alpha by using very similar items, or using very many items.

When diagnosing the internal consisteny of a scale, we can compute item-total correlations to examine the correlation between individual items and the total scale score (minus that item). Low item-total correlations may indicate problematic items, and Cronbach's alpha can be recalculated with items removed to assess their impact on reliability.

23.0.3 Validity

Validity refers to the extent to which an instrument measures what it is intended to measure. There are several types of validity, including face validity (whether the items appear relevant to the construct), content validity (whether the instrument adequately covers all aspects of the construct), and criterion validity (whether the instrument is associated with relevant outcomes or indicators of the construct).

23.0.3.1 Face Validity

Face validity assesses whether items in an instrument are clearly related to the construct of interest. It considers the clarity, readability, and unambiguity of wording and answer options. Face validity is subjective and relies on a first-glance assessment of whether the items seem relevant to the construct being measured.

23.0.3.2 Content Validity

Content validity is usually determined by involving experts in the field, and having them define the construct's scope, generating items for subdomains of the construct, and rating the relevance of items. An instrument has content validity if it adequately covers all aspects of the construct. Content validity is essential for ensuring that the instrument comprehensively captures the intended construct.

23.0.3.3 Criterion Validity

Criterion validity assesses whether an instrument is associated with outcomes or indicators of the construct it is designed to measure. This can involve correlations between the instrument and external measures (e.g., other validated scales) or predictions of behavior related to the construct. Criterion validity provides evidence that the instrument measures what it is intended to measure.

In sum, effective measurement instruments in the social sciences must have both high reliability and high validity. Reliability ensures that the instrument consistently measures the same construct with low measurement error, and validity ensures that the instrument accurately measures the intended construct (not something else). By considering these principles, researchers can enhance the quality and meaningfulness of their questionnaire-based measurements.

23.1 Lecture

https://www.youtube.com/embed/LuxXUQCI_XI

23.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What does reliability refer to in psychometrics? [1]

[1] Consistency or stability of test scores over time or across measurement occasions.

- (A) The average covariance between items and the average variance of items.
- (B) The extent to which an instrument measures what it is intended to measure.
- (C) Consistency or stability of test scores over time or across measurement occasions.
- (D) The clarity and readability of questionnaire items.

Question 2

Which type of reliability assesses consistency across repeated administrations of the same test? [2]

- (A) Inter-rater reliability
- (B) Internal consistency
- (C) Content reliability
- (D) Test-retest reliability

Question 3

What is the primary purpose of content validity? [3]

- (A) To assess whether items in the instrument are clearly related to the construct of interest.
- (B) To determine the association between the instrument and external measures.
- (C) To ensure that the instrument adequately covers all aspects of the construct being measured.
- (D) To examine the correlation between individual items and the total scale score.

Question 4

Which type of validity assesses whether an instrument is associated with relevant outcomes or indicators of the construct? [4]

- (A) Face validity

[2] Test-retest reliability
[3] To ensure that the instrument adequately covers all aspects of the construct being measured.
[4] Criterion validity

- (B) External consistency
- (C) Content validity
- (D) Criterion validity

Question 5

What is the key focus of face validity? [5]

- (A) The association between repeated assessments of the test.
- (B) The apparent relevance and clarity of the instrument's items.
- (C) The relationship between items within a test.
- (D) The stability of test scores over time.

Question 6

Which measure of internal consistency is based on the average covariance between items and the average variance of items? [6]

- (A) Test-retest correlation
- (B) McDonald's Omega
- (C) Cronbach's alpha
- (D) Variance

Question 7

What is the relationship between reliability and validity? [7]

- (A) Reliability and validity are independent.
- (B) Validity is a necessary condition for reliability.
- (C) You can't have reliability without validity, and vice versa.
- (D) Reliability is a necessary condition for validity.

[5]The apparent relevance and clarity of the instrument's items.
[6]Cronbach's alpha
[7]Reliability is a necessary condition for validity.

Question 8

If an item is contra-indicative, which statement is likely true about the item-total correlation for that item? [8]

- (A) It is positive
- (B) It is near-zero
- (C) It is negative
- (D) It is small (e.g., below .30)

Question 9

What is true about alpha-if-item-deleted? [9]

- (A) It is a diagnostic tool to help you identify items that are not valid.
- (B) It is a decision criterion to eliminate items from a questionnaire.
- (C) It is a diagnostic tool that can help you identify items that don't work well with the rest of a questionnaire.
- (D) It is a diagnostic tool that can help you identify unreliable items.

[8] It is negative
[9] It is a diagnostic tool that can help you identify items that don't work well with the rest of a questionnaire.

Show explanations

Question 1
Reliability refers to the consistency or stability of test scores over time or across different measurement occasions.
Question 2
Test-retest reliability assesses consistency across repeated administrations of the same test over time.
Question 3
Content validity involves ensuring that the instrument comprehensively covers all aspects of the construct being measured.
Question 4
Criterion validity assesses whether an instrument is associated with outcomes or indicators of the construct it is designed to measure.
Question 5
Face validity focuses on the apparent relevance and clarity of the instrument's items.
Question 6
Cronbach's alpha estimates the internal consistency of a scale by considering the average covariance between items and the average variance of items.
Question 7
Reliability is a necessary condition for validity, meaning that an instrument must be consistent and stable to accurately measure the intended construct.
Question 8
Contra-indicative items measure the opposite of what the scale measures, so they should correlate negatively with the scale.
Question 9
Alpha-if-item-deleted tells you what Cronbach's alpha of a questionnaire would be if you left out that item. You should not follow this blindly, but you can use it to identify items that don't work well with the rest of the scale. This says nothing about the reliability of individual items, nor their validity.

23.3 Tutorial

23.3.1 Norm Violating Behaviors

In this assignment we are going to take a look at scale that measures whether people engage in norm violating behaviors.

The scale consists of the following items:

- Joyriding
- Taking soft drugs

- Accepting a bribe
- Throwing away litter
- Driving under influence of alcohol
- Smoking in public places
- Speeding over limit
- Euthanasia

Open the datafile `evs.sav`.

In SPSS, navigate to Analyze –> Scale –> Reliability Analysis.

Select the seven items that are in this scale.

Then go to statistics and select the options "Item", "Scale" and "Scale if item deleted". Click on continue.

Now paste and run the syntax.

What is the value of Cronbach's Alpha? _____[10]

Finish the following sentence.

This Cronbach's Alpha is [11]

- (A) Poor
- (B) Questionable
- (C) Adequate
- (D) Good

Cronbach's Alpha is an estimate of the scale reliability.

Describe in your own words what scale reliability entails.

Answer

The reliability of a scale shows the internal consistency of the answers on all items on a scale.

Look at the "Item-Total Statistics" table.

What does the last column "Cronbach's alpha if Item Deleted" tell you?

[10] 0.693
[11] Questionable

Answer

The values in the column "Cronbach's Alpha if item deleted" shows the Cronbach's Alpha of the scale if one of the items would be deleted. In other words, it shows what the impact on the reliability of the scale would be if a certain item would be excluded from the scale.

If you would were examining the psychometric properties of this scale, which items would you consider to give cause for concern?

Answer

We look at the Cronbach's Alpha if item deleted to see the internal consistency of a scale without that item.
Given the Cronbach's Alpha of our original scale (.693), the table shows the only deleting the item "Euthenasia" would result in a higher Cronbach's Alpha.
We might thus question whether this item belongs in the scale or not.

23.3.2 Machiavellianism

In this assignment we will look into the latent concept Machiavellianism.

The personality trait of Machiavellianism is part of what's called the "dark triad", a personality type that is characterized by deceitfulness, cynicism, and an absence of morality and empathy.

Open the datafile `shortmach2.sav`. The data file contains data on Machiavellianism and some other variables.

The included Machiavellianism scale consists of the following 20 items. These items are scored on a 1-5 Likert scale ranging from "Disagree" to "Agree".

Q1. Never tell anyone the real reason you did something unless it is useful to do so.

Q2. The best way to handle people is to tell them what they want to hear.

Q3. One should take action only when sure it is morally right.

Q4. Most people are basically good and kind.

Q5. It is safest to assume that all people have a vicious streak, and it will come out when they are given a chance.

Q6. Honesty is the best policy in all cases.

Q7. There is no excuse for lying to someone else.

Q8. Generally speaking, people won't work hard unless they're forced to do so.

Q9. All in all, it is better to be humble and honest than to be important and dishonest.

Q10. When you ask someone to do something for you, it is best to give the real reasons for wanting it rather than giving reasons which carry more weight.

Q11. Most people who get ahead in the world lead clean, moral lives.

Q12. Anyone who completely trusts anyone else is asking for trouble.

Q13. The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught.

Q14. Most people are brave.

Q15. It is wise to flatter important people.

Q16. It is possible to be good in all respects.

Q17. P.T. Barnum was wrong when he said that there's a sucker born every minute.

Q18. It is hard to get ahead without cutting corners here and there.

Q19. People suffering from incurable diseases should have the choice of being put painlessly to death.

Q20. Most people forget more easily the death of their parents than the loss of their property.

Items can be indicative or contra-indicative of a certain trait.

Which items are contra-indicative to the trait Machiavellianism?

Answer

In this assignment, the following questions are contra-indicative: Q3, Q4, Q6, Q7, Q9, Q10, Q11, Q14, Q16, and Q17.
If an item is contra indicative, low scores on these items indicate a lot of Machiavellianist traits, whereas high scores indicate not so much of an endorsement of Machiavellianist traits in one's personality.

We need to recode contraindicative items before we carry out reliability analysis.

However, just to see what happens, let's first carry out a reliability analysis with all original (i.e. not recoded) variables.

Take the following steps:

Analyze –> Scale –> Reliability Analysis

Select the 20 items (Q1A - Q20A)

Click on Statistics

Under Descriptives, select Items, Scale, and Scale if item deleted

Under Inter-Item, select Correlations

Paste and run the syntax

What is the estimated reliability? _____[12]

True or false: This scale can be considered reliable. TRUE / FALSE[13]

This low reliability might be a result of the fact that we have not recoded our contra-indicative items yet. We can also see this in the inter-item correlation table; many of the items are negatively correlated! Let's rectify this.

We have to recode all contraindicative items. Use syntax to do so, and check your answer below.

Answer

```
COMPUTE Q3r = 6-Q3A.
EXECUTE.
```

For the second person in the dataset, the original score on Q4A was [14] *and the score on the recoded variable Q4r was* [15].

Based on this comparison, do you think you successfully recoded the variable?

We will now run the reliability analysis including the recoded items. Chnge your syntax, or re-run the analysis via the visual interface.

What is the value of Cronbach's Alpha? _____[16]

This Cronbach's Alpha is [17]

- (A) Adequate
- (B) Poor
- (C) Questionable
- (D) Good

Check the corrected item total correlations.

Which of these items has the smallest association with other items in the scale? (Type its name) ____[18]

---

[12] 0.233
[13] FALSE
[14] 4
[15] 2
[16] 0.887
[17] Good
[18] Q19

Explain why you think it makes sense (or not) that this item is correlated the least with the rest of the scale.

Answer

This item-total correlation of Q19 is .255.
Item Q19 reads: "People suffering from incurable diseases should have the choice of being put painlessly to death".
One could argue that a high score on this item should relate to a low score on the scale for agreeing with this item nowadays might actually show empathy with those suffering.

Inspect Cronbach's alpha if item deleted. If you had to remove one item based on this Cronbach's alpha if item deleted, which item would it be? (Type its original name) _____[19]

Explanation

Based on the Cronbach's Alpha if item deleted, we would delete the item that would result in the greatest increase in Cronbach's Alpha if it would be deleted.
We find the highest Cronbach's Alpha if item deleted for the item Q17r (.889).
Therefore, based on the Cronbach's Alpha if item deleted, we would delete item Q17r.

Taking everything into consideration, would you remove any items from the Machiavellianism scale?

Explanation

Taking the statistical output into consideration, we might want to consider removing item Q17r or item Q19.
We do so for two reasons:

1. They both correlate less than .3 with the other items on the scale, and
2. have a Cronbach's Alpha if item deleted higher than .887 (the current Cronbach's Alpha of the scale).

Please note that we should always take into consideration theoretical reasons as well when deciding to delete an item from a scale. In this assignment, however, we base our conclusions on statistical reasons only.
You might want to consider removing item Q17 or Q19. They both correlate less than .3 with the rest of the scale, and have an Alpha if item removed higher than .887.

We always remove items one by one. We start with the worst item. After removing a bad item from a scale, the item-statistics will change a little. It might be that the item-statistics improve, and that there is no need to remove the second item.

[19] Q17

However, since the differences are not that big and the scale reliability is pretty high, we will keep both items in our scale for the remainder of this assignment.

Once we finished reliability analysis, we can use the scale in other analyses.

In order to do that, we need to arrive at a total scale score. So, we need one score for each person in the dataset that tells what their score is on the personality trait Machiavellianism.

There are several methods of obtaining such a total score, but one straightforward and easy way is to calculate the sum score.

Navigate to Transform -> Compute Variable.

Give a new name to the sum score in the box Target Variable, such as Mach.

In the box Numeric Expression, enter all variables and add together (i.e., Q1A + Q2A + Q3r .... + Q20A). Ensure that you use the recoded variables for the contra-indicative items!

Paste and run the syntax.

We will now use our newly developed scale in a regression analysis! In this analysis we will try to explain Machiavellianism based on the variables Gender (0=men; 1=women), Age and Voted (0=Voted in past election, 1= Not voted in past election).

Navigate to Analyze –> Regression –> Linear

Enter the sumscore Mach as dependent variable

Enter Gender, Age, and Voted as independent variables

Paste and run the syntax

Inspect the Model Summary table. How much of the variance in Machiavellianism do the variables Gender, Age and Voted explain? _____[20]

True or false: Gender, Age and Voting together explain a significant amount of variance in the variable Machiavellianism. TRUE / FALSE[21]

Inspect the table Coefficients and take a look at the partial effects. Which of the variables does not have a significant partial effect on Machiavellianism? [22]

- (A) Gender
- (B) Age
- (C) Voting

[20] 0.126
[21] TRUE
[22] Voting

Controlled for Age and Voting, which group (men or women) scores higher on the Machiavellianism scale? [23]

- (A) Men
- (B) Women

Explanation

Given that in the variable Gender men are coded as 0 and women are coded as 1, and that the regressions coefficient for Gender is negative (-7.440), we should conclude that men score higher on Machiavellianism than women, controlled for Age and Voting.

Finish the following sentence.

Controlled for Gender and Voting, if we would increase one year in age, the predicted score on the Machiavellianism scale would [24]

- (A) Decrease with 0.294 units.
- (B) Increase with 0.294 units.
- (C) Decrease with 0.243 units.
- (D) Increase with 0.243 units.

23.3.3 Solidarity

In this assignment, you will evaluate the reliability of a scale measuring solidarity.

Download and open the following data file: `solidarity.sav`.

You've practiced with reliability analysis in the previous assignments. Now you can use that knowledge to evaluate the reliability of the solidarity scale more independently.

Include all eleven items that are part of the scale (v266 to v276).

Inspect the output of the reliability analysis.

True or false: You have to recode questions. TRUE / FALSE[25]

[23] Men
[24] Decrease with 0.294 units.
[25] FALSE

Explanation

We check if we should recode any items by looking at how they are phrased. To do so, we don't necessarily need to look at the output.
We could, however, also use the inter-item correlations displayed in the output to check if we should recode items. If any of the inter-item correlations are negative, this should be in indication for contra-indicative items.

What is the reliability of the scale? _____[26]

If you had to remove one item from the scale based on Cronbach's Alpha if item deleted, which one would you pick? Type the name: _____[27]

Can you think of other reasons for removing this item?

Explanation

Comparing the content of item Q80E to the content of the other questions in the scale, we can conclude that the content of this item is off-topic. This is a theoretical reason for excluding item Q80E from the scale.

Which item is most typical for the scale? Type its name: _____[28]

Explanation

We can tell from the higher item-total correlation, that tells us that this item has the strongest correlation with all other variables on the scale.
Also, the reliability of the scale would decrease most if this item would be deleted from the scale.

Last, construct sum scores on the scale for all individuals (via Transform -> Compute). Make sure you do not include the one item we discussed previously!

Once you created the sum score, have SPSS show the mean for this total score.

What is the mean value of the sum score? _____[29]

[26] 0.847
[27] Q80E
[28] Q79D
[29] 27.897

24 Dimension Reduction

In the previous section, we explored how multiple items can be used to measure a single underlying construct. Today, we will delve into three powerful techniques for reducing multiple items to a smaller number of variables: Principal Components Analysis (PCA), Exploratory Factor Analysis (EFA), and a little bit of Confirmatory Factor Analysis (CFA). Our focus will be on PCA and EFA, as they are particularly useful for understanding underlying structures in social science research. This section will primarily discuss Principal Components Analysis (PCA) and Exploratory Factor Analysis (EFA). These techniques help us explore relationships among items and identify latent constructs that explain the observed patterns in data. They serve as effective tools for dimensionality reduction, enabling us to summarize complex datasets with a smaller set of variables. While we do introduce CFA and reflect on its relationship to EFA, a more in-depth discussion of CFA is beyond the scope of this course.

Throughout this lecture, assume that we have k items and n participants. Let's now dive into the details of different data reduction methods.

24.1 Principal Components Analysis (PCA)

PCA is a data rotation technique designed to transform original items into uncorrelated components. These components represent linear combinations of the original items. The primary goal of PCA is dimension reduction, where a small number of components are used to explain most of the variance in the items. This allows us to represent the variance in the items more efficiently. For instance, if ten items measure extraversion, and one component explains most of the variance, we can retain that one component and discard the remaining nine.

24.2 Exploratory Factor Analysis (EFA)

Unlike PCA, EFA is a latent variable method that assumes that latent variables (factors) cause people's responses to the items. For example, extraversion may cause individuals to respond positively to questions about partying and socializing. EFA models the item covariance matrix as a function of a fixed number of factors. It is called "exploratory" because all items are allowed to load on (contribute to) all factors, without a predefined structure. In practice, well-constructed questionnaires will exhibit high loadings of items on one factor and low loadings on others.

24.3 Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis (CFA) tests a theory about the specific associations between latent variables and observed indicators. Unlike Principal Components Analysis (PCA) and Exploratory Factor Analysis (EFA), which are exploratory, CFA is a confirmatory approach that tests how well a hypothesized measurement model fits the data. In CFA, researchers specify a theoretical model that defines the relationships between observed variables and latent constructs (factors). These latent constructs are not directly measured but are assumed to explain the correlations among the observed variables. The primary goal of CFA is to evaluate whether the data support the hypothesized model. By doing so, researchers can determine if their theoretical model fits the observed data well, providing evidence for the validity of the underlying construct and the measurement instrument. CFA is part of a family of statistical modeling techniques known as "Structural Equation Modeling" (SEM).

24.3.1 Comparing Method

Purpose:

- PCA: Dimensionality reduction.
- EFA: Exploration of relationships among items and identification of latent constructs.
- CFA: Testing a predefined theory about which items relate to specific latent constructs.

Assumption:

- PCA: Does not assume latent variables; dropping components assumes they are irrelevant or represent error variance.
- EFA: Assumes all items are caused by a smaller number of latent variables (factors).
- CFA: Assumes specific items are caused by specific latent variables.

Interpretation:

- PCA: Components are mathematical constructs with no further meaning.
- EFA: Factors represent theoretical latent constructs.
- CFA: Factors represent known theoretical latent constructs.

24.3.2 Principal Components Analysis

PCA is a data rotation technique that aligns the largest amount of variance with the first component, the second-largest variance with the second component, and so on. These components are uncorrelated by definition, and they serve as linear combinations of the original items. The primary use of PCA is dimension reduction by retaining only components that explain a significant amount of variance, thus providing a lower-dimensional representation of the data.

We can understand PCA in different ways. Firstly, as rotation of the data. PCA rotates the data so that the first component best reproduces the correlation matrix, and each subsequent component improves the reproduction. Secondly, we can understand PCA as a way to summarize k items using fewer than k components, without significant information loss (lossy compression of data).

Selecting the Number of Components:

Various strategies exist to determine the number of components to retain, including Kaiser's criterion (Eigenvalue > 1), Cattell's scree plot (inflection point), and Horn's Parallel Analysis (comparison with random data's Eigenvalues). Additionally, theoretical knowledge about the underlying data can guide the choice of components.

Interpreting PCA Loadings:

Interpreting PCA loadings can be challenging, especially in cases where multiple components are correlated. Orthogonal rotation, such as Varimax, can be employed to simplify the pattern of loadings and improve interpretability. However, it is essential to remember that rotated loadings should not be directly interpreted as correlations between items and factors as in PCA.

24.3.3 Exploratory Factor Analysis

EFA is a model-based approach that assumes the existence of latent variables that cause item responses. It is suitable when you expect clusters of items to be correlated (multicollinear) and seeks to explain correlations between items. EFA assumes that unexplained variance in the items can be attributed to measurement error. This aligns with test theory, where it is assumed that observed items measure latent constructs with error. EFA is particularly suitable when there is a theoretical basis for assuming the existence of latent variables, such as when developing a new questionnaire that has not been validated yet. However, if a theoretical model already exists, Confirmatory Factor Analysis (CFA) may be more appropriate.

To conduct EFA, we estimate the unknown factor loadings. Two common estimation methods are Principal Axis Factoring (PAF) and Maximum Likelihood (ML). PAF is a default method in SPSS and is based on an iterative procedure involving matrix algebra. It provides a solution even when the model is complex or the data are non-normal. On the other hand, ML is the same estimator used for CFA and works well when the data are multivariate normal. However, ML may not perform well when the model is overly complex (which is not necessarily a bad thing). ML estimation also allows for a test of model fit, which is useful for evaluating the appropriateness of the chosen model.

Factor loadings represent the correlations between each item and the extracted factors. They indicate the strength and direction of the relationship between the observed item and the underlying factor. Factor loadings range from -1 to +1, with values closer to 1 indicating a stronger relationship. In our example, we can see the factor loadings in a factor matrix, where

each row corresponds to an item and each column corresponds to a factor. The factor loadings help us identify which items load more strongly on specific factors.

We can compute Eigenvalues in EFA just as in PCA by taking the column sums of the squared loadings and indicate the amount of variance explained by each factor. Eigenvalues are always smaller than the initial eigenvalues obtained in Principal Component Analysis (PCA) because some variance is now attributed to error variance. Consequently, the sum of the Eigenvalues is also less than the number of indicators, and some Eigenvalues may even be negative.

Similarly, communalities in EFA are always < 1 because EFA assumes the existence of error variance.

24.3.3.1 Selecting the Number of Factors

Determining the appropriate number of factors to extract is a critical step in EFA. Researchers often use eigenvalues as a cue to determine the number of factors to extract, similar to the Kaiser's criterion and Scree plot used in PCA - but note that in EFA, this can be misleading as Eigenvalues now depend on the number of extracted factors. Also, by default, SPSS applies Kaiser's criterion and the Scree plot to PCA Eigenvalues, even if you request EFA!

An alternative criterion for determining the number of factors is using theoretical knowledge to guide the decision. For example, if emotions are believed to break down into positive and negative emotions, we may choose to extract two factors. Additionally, the chi-square test can be used to evaluate the appropriateness of different factor solutions and assist in selecting the best-fitting model. To directly compare models, one can compute the Bayesian Information Criterion (BIC) - a relative model fit index designed for comparing models, which balances model fit and complexity. It is computed from the chi square as follows:

$$BIC = ^2 \ df \ log(n)$$

24.3.4 EFA Assumption Checks

Before conducting exploratory factor analysis (EFA), it is good practice to perform several assumption checks to ensure the validity and appropriateness of the analysis. One critical aspect to consider is multicollinearity. While factor analysis aims to identify clusters of items that are correlated, excessive multicollinearity can lead to issues. This occurs when multiple items are perfectly linearly dependent, meaning that one item's score can be exactly reproduced using other variables. In such cases, it becomes difficult to discern the unique contribution of collinear items to the underlying factor model. To detect multicollinearity, researchers can examine the determinant, a value between 0 and 1. It has been argued that the determinant should be greater than 0.00001, which indicates multicollinearity is not too high.

Another assumption check for EFA is the proportion of common variance among items. The Kaiser-Meyer-Olkin (KMO) statistic provides an estimate of this proportion. A higher KMO value indicates that more of the variance among items can be explained by common factors, making the data more suitable for factor analysis. Researchers can interpret the KMO value as follows:

Value	Interpretation
0.00 to 0.49	unacceptable
0.50 to 0.59	miserable
0.60 to 0.69	mediocre
0.70 to 0.79	middling
0.80 to 0.89	meritorious
0.90 to 1.00	marvelous

24.3.5 Rotating Factor Loadings

In factor analysis, we aim to interpret the underlying structure of observed variables. The pattern of factor loadings is crucial in this process, helping us identify items that load highly on specific factors and potentially naming those factors based on high-loading indicators. In a perfect world, factor loadings would be clear and straightforward, with each item loading highly on only one factor. However, real-life factor loadings are not always so clear-cut, making interpretation more challenging.

To improve interpretability, we use rotation, which applies a linear transformation to the original factor loadings. Two main types of rotation are orthogonal and oblique rotation. Orthogonal rotation produces uncorrelated factors. The most common technique is VARIMAX rotation, which maximizes the variance of the squared loadings within each factor. Oblique rotation allows factors to correlate; the most common technique is oblimin rotation. In the social sciences, it is often sensible to allow factors to correlate (e.g., different personality dimensions are probably associated).

One-Factor EFA and One-Factor CFA:

Although this course is not about confirmatory factor analysis, it is nevertheless useful to know that a one-factor EFA model is identical to a one-factor CFA model. In other words, if our theory implies a one-factor model, we can use exploratory factor analysis (EFA) with maximum likelihood (ML) estimation to test that model. While EFA aims to identify underlying factors without any preconceived hypotheses about their association with items, CFA tests a hypothesized model - in this case, that one factor explains all item scores. CFA with ML estimation produces a chi-square test that can be used to assess model fit. Note, however, that this test can be sensitive to sample size and may reject good models. Researchers can also use the Root Mean Square Error of Approximation (RMSEA) as an alternative model fit index, where values below 0.08 indicate good fit. RMSEA is calculated from the chi square as:

$$RMSEA = \frac{^2\ df}{(n\ 1)\ df}$$

Treating a one-factor EFA as CFA also allows us to estimate latent variable reliability. Recall that Cronbach's alpha assumes that all items are equally important. This means that it assumes that all factor loadings are the same. Factor analysis tests this assumption. Especially when factor loadings differ, it may be useful to compute latent variable reliability instead, using McDonald's Omega (or composite reliability). It allows for different factor loadings, making it more appropriate for cases where items have varying contributions to the latent variable. The formula for McDonald's Omega is:

$$= \frac{SSL}{SSL + SSR} = \frac{\text{Sum of Squared Loadings}}{SSL + \text{Sum of Squared Residuals}}$$

Calculate SSL as: $SSL = (_{j=0}^{k} L_{1,k})^2$ (first sum loadings, then square sum)

Calculate SSR as: $SSR = 1\ _{j=0}^{k} L_{1,k}^2$ (first square loadings, then sum)

24.3.6 Estimating Factor Scores

In many cases, researchers want to conduct further analyses using idividuals' scores on components or latent variables. In previous sections, we learned about two common methods for obtaining scale scores from multiple items: sum scores and mean scores. In sum scores, we add up the responses from each item to create a total score for each individual. Similarly, in mean scores, we take the average of the responses from all items to obtain a score. In both cases, all items contribute equally to the final scale score. However, this approach assumes that all items are equally important, which might not always be the case. PCA and EFA both allow us to determine whether items are indeed equally important. We can also try to compute scale scores that take differences in item loadings into account.

For PCA, computing such scores is straightforward; these are simply given by multiplying the loadings for one component with the observed item scores. Since this is not a latent variable technique, there is only one possible solution to this calculation. To compute a PCA score for a specific individual, we multiply their standardized item scores by the corresponding factor loadings and then sum the results. For instance, if an individual has standardized item scores of 1, 3, and 2 on items with factor loadings of 0.85, 0.80, and 0.14, respectively, their PCA score would be calculated as $(0.85\ \ 1 + 0.80\ \ 3 + 0.14\ \ 2)/(0.85^2 + 0.80^2 + 0.14^2) = 2.44$. This score represents the individual's relative level on the component.

Estimating latent variable scores in exploratory factor analysis (EFA) is more complex compared to PCA. Unlike PCA, which provides unique factor scores for each individual, EFA does not uniquely determined factor scores. An infinite number of latent variable datasets is consistent

with the same EFA model. To estimate factor scores, researchers use methods like the regression method and the Bartlett method. The regression method involves ordinary least squares estimates and aims to maximize the multiple correlation between factor scores and common factors. However, these estimates are biased and the estimated factor scores correlate with one another and with the different latent variables. The Bartlett method produces factor scores that only correlate with their own latent variable but still correlate with estimated scores for other factors. Both methods thus have shortcomings. Some (see references below) have argued that it might be preferable to simply use mean scores instead of factor scores. In cases where factor loadings are approximately equal, this is probably fine.

Further reading:

Everitt, B. S., & Howell, D. C. (2005). Encyclopedia of Statistics in Behavioral Science. DOI:10.1002/0470013192.bsa726 DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. DOI:10.7275/da8t-4g52

24.4 Lecture

https://www.youtube.com/embed/EF2Jcsh4OqA

24.5 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What is the primary goal of Principal Components Analysis (PCA)? [1]

- (A) Model confirmation
- (B) Data classification
- (C) Dimension reduction
- (D) Factor extraction

[1] Dimension reduction

Question 2

In Exploratory Factor Analysis (EFA), what is the role of latent variables? [2]

- (A) They are directly measured
- (B) They confirm pre-defined structures
- (C) They represent linear combinations of original items
- (D) They cause people's responses to the items

Question 3

What distinguishes Confirmatory Factor Analysis (CFA) from PCA and EFA? [3]

- (A) It is used for dimension reduction
- (B) It is an exploratory method
- (C) It tests a hypothesized measurement model
- (D) It assumes that factors cause item responses

Question 4

How are components ordered in PCA? [4]

- (A) Randomly assigns variance to components
- (B) Equally distributes variance across all components
- (C) Largest variance with the first component, second-largest with the second, and so on
- (D) Aligns smallest variance with the first component

Question 5

What is one way to understand PCA? [5]

- (A) As a technique for classification of individuals

[2]They cause people's responses to the items
[3]It tests a hypothesized measurement model
[4]Largest variance with the first component, second-largest with the second, and so on
[5]As a method of lossy compression of data

- (B) As a process for data duplication
- (C) As a method of lossy compression of data
- (D) As a method for data expansion

Question 6

What is the purpose of using orthogonal rotation, like Varimax, in PCA? [6]

- (A) To increase the correlation between items and factors
- (B) To simplify the pattern of loadings and improve interpretability
- (C) To reduce the number of components retained
- (D) To directly interpret loadings as correlations

Question 7

What distinguishes Principal Axis Factoring (PAF) from Maximum Likelihood (ML) estimation? [7]

- (A) ML is only suitable for CFA
- (B) PAF provides a solution even in complex models or non-normal data
- (C) ML always results in higher factor loadings
- (D) PAF allows for a test of model fit

Question 8

How are Eigenvalues computed differently in EFA compared to PCA? [8]

- (A) EFA does not compute Eigenvalues
- (B) Eigenvalues are smaller in EFA as some variance is attributed to error variance
- (C) Eigenvalues are the same in both EFA and PCA
- (D) Eigenvalues are larger in EFA

[6] To simplify the pattern of loadings and improve interpretability
[7] PAF provides a solution even in complex models or non-normal data
[8] Eigenvalues are smaller in EFA as some variance is attributed to error variance

Question 9

What indicates a problem with multicollinearity in Exploratory Factor Analysis (EFA)? [9]

- (A) A high Kaiser-Meyer-Olkin (KMO) statistic
- (B) No presence of multicollinearity
- (C) A determinant value equal to 1
- (D) A determinant value lower than 0.00001

Question 10

Which of these is a method for estimating latent variable scores in EFA? [10]

- (A) Regression method
- (B) Sum scores
- (C) Assigning equal weights to all items
- (D) Using mean scores of all items

[9] A determinant value lower than 0.00001
[10] Regression method

Show explanations

Question 1
The primary goal of PCA is dimension reduction, where a small number of components are used to explain most of the variance in the items.
Question 2
In EFA, latent variables (factors) are assumed to cause people's responses to the items, unlike PCA where components represent linear combinations of original items.
Question 3
CFA is a confirmatory approach that tests how well a hypothesized measurement model fits the data, unlike PCA and EFA, which are exploratory.
Question 4
PCA aligns the data such that the largest amount of variance is with the first component, and each subsequent component accounts for the next largest variance.
Question 5
PCA can be understood as a way to summarize k items using fewer than k components, which can be seen as a form of lossy compression of data.
Question 6
Orthogonal rotation, such as Varimax, is employed in PCA to simplify the pattern of loadings, making them easier to interpret.
Question 7
PAF is an iterative method suitable for complex models or non-normal data, whereas ML, used for both EFA and CFA, is better for data that are multivariate normal.
Question 8
In EFA, Eigenvalues are always smaller than in PCA because some variance is attributed to error variance, leading to Eigenvalues that are less than the number of indicators and sometimes negative.
Question 9
In EFA, a determinant value lower than 0.00001 indicates excessive multicollinearity, making it difficult to discern the unique contribution of collinear items to the factor model.
Question 10
In EFA, factor scores can be estimated using methods like the regression method or the Bartlett method, each with its own advantages and shortcomings.

24.6 Tutorial

24.6.1 PCA

Open the data file: `emotions.sav`.

The data file consists of data from the International College Survey 2001 (Diener and colleagues, 2001). In this survey, data on emotions was collected for 41 countries. The data you'll analyze in this assignment is about norms for experiencing/expressing 12 emotions in Belgium.

Let's look at the data. The first two columns contain the number of the participant and the nation, so you don't need to include them in the analysis.

True or false: There are missing data. TRUE / FALSE[11]

Suppose we are only interested in reducing the number of dimensions of the data, which method would you use? [12]

- (A) Confirmatory Factor Analysis
- (B) Explanatory Factor Analysis
- (C) Path Analysis
- (D) Principal Component Analysis

Navigate to Analyze Dimension Reduction Factor in SPSS

In the tab Extraction: choose the correct method.

Also enable the option Scree plot and specify which variables need to be included in the analysis.

Check the Options tab. Can you determine what method is used to deal with missing data?

[13]

- (A) All correlations are computed based on available data for that pair of variables
- (B) All cases with missing values are removed prior to analysis
- (C) All missing values are removed prior to analysis
- (D) No action is taken

Paste the syntax and run the analysis.

Take a look at the output.

What number of component have an Eigenvalue greater than 1 (Kaiser's criterion)? _[14]

[11] TRUE
[12] Principal Component Analysis
[13] All cases with missing values are removed prior to analysis
[14] 3

How many components does the scree plot suggest?

_[15]

Redo the analysis with the number of components you need to retain according to the scree plot.

You can specify the number of components in the Extraction menu of the Factor Analysis window.

Click Fixed number of factors and enter the number of components (2).

Run the analysis and look at the loadings in the Component matrix.

True or false: This solution is easy to interpret. TRUE / FALSE[16]

The two principal components seem to correspond with positive emotions (appropriate and valued), and negative emotions (inappropriate and not valued), but there is not enough simple structure (too many variables have a high loading on both components).

To aid interpretation, you could rotate the solution. Which type of rotation is most appropriate here? [17]

- (A) orthogonal
- (B) oblique

Answer

It is unlikely that positive and negative emotions are uncorrelated! An oblique rotation seems by far the most sensible choice.

Regardless of your previous answer, redo the analysis and choose Direct Oblimin in the Rotation menu.

Take a look at the component loadings in the Pattern matrix.

Which component would you label Positive Emotions? Number.. _[18]

Compare the component loadings in the Pattern Matrix with the loadings in the Component Matrix.

We now observe that the loadings resemble a simple structure more closely than before the rotation: the low loadings are lower and the high loadings are higher.

[15] 2
[16] FALSE
[17] oblique
[18] 2

Note: Due to the oblique rotation, the loadings are no longer equal to item-component correlations.

What is the correlation between the two rotated components? _____[19]

Redo the Principal Component Analysis again one last time to save the component scores in the data set. Open Scores in the Factor Analysis window, check the Save as variables checkbox. Have a look at these component scores (now added to your data set): these are the scores for each person on the two components.

Alternatively, add this syntax:

```
/SAVE REG(ALL)
```

What is the component score for the first person on the first component? _____[20]

Take a look at the table Total Variance Explained.

How much of the variance do the two components together account for? _____[21]%

What proportion of the variance in the item stress is accounted for by the two components? _____[22]

Which item has the highest unicity? [23]

- (A) Happy
- (B) Anger
- (C) Pride
- (D) Cheerful

24.6.2 Exploratory Factor Analysis

We will move on to work with Exploratory Factor Analysis.

For this second assignment you will perform an Exploratory Factor Analysis (EFA) in SPSS on a set of 18 items. These items measure Tolerance and are part of the European Value Survey (EVS).

[19] 0.143
[20] 1.937
[21] 51.575
[22] 0.505
[23] Pride

Discuss with your group when we decide to use Exploratory Factor Analysis and when we decide to use Principal Component Analysis.

Explanation

PCA is a data reduction technique. We use it when we want to summarize information in the items.
EFA is used to identify latent variables underlying the measured items. EFA is typically used when a questionnaire has not been validated yet. When we use EFA, we usually do not know exactly which item belongs to which dimension (although we might have an idea based on our theory).

Discuss with your group: When do we use Confirmatory Factor Analysis?

Explanation

CFA is used when we DO know which items belong to which dimension. With CFA we can then check whether the model that we have in mind corresponds with what we see in the data.

Open the file `evs.sav` in SPSS.

Select Factor via Analyze -> Dimension Reduction.

Which extraction method should we use if we want a test of model fit? [24]

- (A) Maximum Likelihood
- (B) Principal Components Analysis
- (C) Principal Axis Factoring
- (D) Unweighted Least Squares

Drag all items of the tolerance scale (i.e., V225 - V2242) into the 'items' window. Go to Descriptives and select the options "Coefficients", "Determinant", and "KMO and Bartlett's test of sphericity". Then, go to extraction and select "unrotated factor solution" and "scree plot". Paste and run the syntax.

What is the Determinant? _____ [25]

True or false: The determinant indicates that multicollinearity might be a problem for these data. TRUE / FALSE[26]

[24] Maximum Likelihood
[25] 0.004
[26] FALSE

The factorability, as determined by the KMO index, is [27]

- (A) Marvelous
- (B) Mediocre
- (C) Middling

How many factors would you want to select based on the scree plot? _[28]

How many factors would you want to select based on Kaiser's criterion? _[29]

What are the limitations of using these criteria?

Answer

Both are based on eigenvalues computed for PCA, but you are performing EFA now. Although you can also compute eigenvalues for EFA, SPSS doesn't use those for the scree plot and Kaiser's criterion - and moreover, eigenvalues for EFA depend on the number of extracted factors, which defeats the purpose of using them to determine how many factors to extract.
Furthermore, EFA is a theory-driven technique; it makes sense to use theory to determine how many factors to retain.

Assume that we're extracting two factors for now. Re-do your analysis with the appropriate number of factors.

In the tab Extraction: choose the number of factors you want to extract.

In the tab Rotation: Tick the box Direct Oblimin.

In the tab Options: The interpretation of the pattern matrix is easier if you suppress all coefficients in that table that are small (e.g., values < 0.30). To do so, click on options and ask SPSS to suppress the small coefficients.

In the tab Descriptives: Ask for the reproduced matrix.

Paste and run the syntax.

When we interpret the output of the factor analysis, we inspect 4 tables: the pattern matrix, the communalities, the factor correlation matrix, and the reproduced correlation matrix.

We will start with the pattern matrix.

Inspect the factor loadings in the pattern matrix.

[27] Marvelous
[28] 2
[29] 3

Which item has the highest absolute factor loading on Factor 2? Type the variable label from the table: _____[30]

Decide for yourself: are the two factors clearly interpretable? Then check your answer.

Answer

The solution almost follows a simple structure where each item loads on one factor. Only for the item Having casual sex do we see high factor loadings on both factors.

Inspect the communalities table.

How much of the variance in the item "suicide" do the factors explain? _____[31]

Check the correlations between the three factors.

How substantial is the correlation between the factors? [32]

- (A) weak
- (B) moderate
- (C) large

Inspect the residual correlations.

Which residual correlation is most concerning?

[33]

- (A) Between speeding over the limit and smoking in public places.
- (B) Between taking soft drugs and joyriding.
- (C) Between Cheating on tax and Paying cash
- (D) Between driving under the influence and claiming state benefits.

Take a look at the pattern matrix again.

Can you think of a meaningful label for each of the factors? (Take into consideration whether the loadings are positive or negative). Then check your answer.

[30] divorce
[31] 0.317
[32] moderate
[33] Between Cheating on tax and Paying cash

Answer

There appears to be a distinction between legal and religious issues.

24.6.3 Exploratory Factor Analysis II

Open the dataset called student_questionnaire.sav.

It contains data on moral judgment in a variety of domains of social life (variables whose names start with MACJ). Note that you need the variable `MACJ13_imputed`, not `MACJ13`.

24.6.3.1 Model Selection using the BIC

When conducting EFA with ML estimation, we obtain a chi-square test of model fit that allows us to compute the BIC, a comparative fit index that can help us choose the number of factors that best balances model fit and complexity.

Run an EFA analysis for 1-3 and 7-9 factors. Using syntax can help you do this easily - just copy-paste the basic syntax below four times and change the number of classes:

```
FACTOR
  /VARIABLES MACJ1 MACJ2 MACJ3 MACJ4 MACJ5 MACJ6 MACJ7 MACJ8 MACJ9 MACJ10 MACJ11 MACJ12 MACJ
    MACJ14 MACJ15 MACJ16 MACJ17 MACJ18 MACJ19 MACJ20 MACJ21
  /MISSING LISTWISE
  /CRITERIA FACTORS(1) ITERATE(100)
  /EXTRACTION ML
  /ROTATION NOROTATE.
```

Open a spreadsheet in Excel or Google Sheets, and copy-paste the chi-square values and degrees of freedom into the first two columns. Obtain the number of (valid) observations using whatever procedure you want (for example, Descriptives).

Assuming that you have used the first two columns, paste the following formula into the fourth column. Replace "n" with the number of valid observations. Drag the formula down to copy it the all cells in its column:

```
= A1 - B1 * LOG(n)
```

What is the BIC for 3 factors? _____[34]

[34] 83.39268271

Based on the BIC, out of the set of models compared, which number of factors would you choose? _[35]

True or false: This finding corresponds to the conclusion you would draw from the Scree plot. TRUE / FALSE[36]

True or false: This finding corresponds to the conclusion you would draw from Kaiser's criterion. TRUE / FALSE[37]

True or false: KMO suggests that there is insufficient common variance for factor analysis. TRUE / FALSE[38]

True or false: The determinant suggests a potential problem with multicollinearity. This might be because there are so many similar items. TRUE / FALSE[39]

If I told you that the theory specified 7 factors, how many factors would you prefer? Explain why, then check your answer.

Answer

The BIC for 7 factors is almost identical to the one for 8 factors. If theory dictates 7 factors, you might prefer to stick with 7, as the evidence for 8 factors is not overwhelmingly stronger.

24.6.3.2 Latent Variable Reliability

Regardless of your previous answer, perform EFA with one factor. Recall that this is equivalent to performing CFA with one factor.

Cronbach's alpha assumes that all items have equal factor loadings. Examine the factor loadings matrix.

True or false: it looks like the factor loadings are indeed all equivalent. TRUE / FALSE[40]

Compute Cronbach's alpha for these items, and report the value: _____[41]

Copy-paste the factor loadings into a spreadsheet.

Use the spreadsheet function `=SUM()` to sum the loadings, then square the sum to get the SSL, $SSL = (\sum_{j=0}^{k} L_{1,k})^2$

[35] 8
[36] FALSE
[37] FALSE
[38] FALSE
[39] TRUE
[40] FALSE
[41] 0.931

Create a new column with the squared factor loadings. Use the function `=A1^2` (assuming that cell A1 contains your first factor loading). Then sum these squared loadings to get the SSR, $SSR = 1 \sum_{j=0}^{k} L_{1,k}^2$.

Finally, calculate McDonald's Omega:

$$= \frac{SSL}{SSL + SSR}$$

Report McDonald's Omega: _____[42]

Note that McDonald's Omega is larger than Cronbach's alpha. This is a rule; Cronbach's alpha underestimates reliability compared to McDonald's omega, and the underestimation becomes worse as the assumption of equal factor loadings is more violated.

24.6.3.3 Model Fit

Finally, calculate the RMSEA model fit index for this one-factor model. The cutoff for acceptable fit is RMSEA $< .08$.

$$RMSEA = \frac{^2\ df}{(n\ 1)\ df}$$

True or false: The one-factor model has acceptable fit. TRUE / FALSE[43]

[42] 0.952571482
[43] TRUE

25 BE2 - Confidence Intervals

25.1 Lecture

There is no lecture for this topic, but you can re-watch part of the lecture on the sampling distribution to refresh your knowledge about confidence intervals!

25.2 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

What does a confidence interval represent in statistics? [1]

- (A) The average value of the sample data.
- (B) The highest value in the sample data.
- (C) A range of values within which the true population parameter is likely to fall.
- (D) The exact value of the population parameter.

Question 2

How is the confidence level of a confidence interval chosen? [2]

- (A) By the range of values in the sample data.

[1] A range of values within which the true population parameter is likely to fall.
[2] The researcher determines the desired level of confidence in capturing the true parameter.

- (B) By the size of the sample data.
- (C) The researcher determines the desired level of confidence in capturing the true parameter.
- (D) By using standard values such as 95% or 99%.

Question 3

What does a wider confidence interval indicate? [3]

- (A) Greater uncertainty or less precision in estimating the population parameter.
- (B) Higher confidence in the accuracy of the estimate.
- (C) A narrower range of values in the sample data.
- (D) Smaller sample size.

Question 4

In general, as the sample size increases, what happens to the width of the confidence interval? [4]

- (A) It remains unchanged.
- (B) It becomes equal to the population parameter.
- (C) It increases.
- (D) It decreases.

Question 5

What is the relationship between the width of a confidence interval and the level of confidence? [5]

- (A) Inverse relationship: Higher confidence leads to a wider interval.
- (B) Width and confidence level are equal.
- (C) Direct relationship: Higher confidence leads to a narrower interval.

[3] Greater uncertainty or less precision in estimating the population parameter.
[4] It decreases.
[5] Inverse relationship: Higher confidence leads to a wider interval.

- (D) No relationship: Width and confidence level are independent.

Question 6

What is the purpose of a confidence interval in hypothesis testing? [6]

- (A) One can reject a null hypothesis whose hypothesized value lies outside of a X% confidence interval, with a p-value of 1-(X/100).
- (B) One can reject a null hypothesis whose hypothesized value lies outside of a X% confidence interval, with a p-value smaller than = X/100.
- (C) One can reject a null hypothesis whose hypothesized value lies outside of a X% confidence interval, with a p-value smaller than = 1-(X/100).
- (D) One can reject an alternative hypothesis whose hypothesized value lies outside of a X% confidence interval, with a p-value smaller than = 1-(X/100).

Question 7

A researcher collected data from a sample of 80 participants. The sample mean is 72, and the standard deviation is 2.5. What is the 95% confidence interval? [7]

- (A) [67.10, 76.90]
- (B) [72.45, 72.55]
- (C) [71.72, 72.28]
- (D) [69.50, 74.50]

Question 8

What is the correct statement about a 95% confidence interval? [8]

- (A) The 95% confidence interval contains the population value with 95% probability.
- (B) 95% of all confidence intervals contain the population value.

[6] One can reject a null hypothesis whose hypothesized value lies outside of a X% confidence interval, with a p-value smaller than = 1-(X/100).
[7] [72.45, 72.55]
[8] The 95% confidence interval is a procedure with 95% probability of providing an interval that contains the population value.

- (C) There is a 95% probability that the population value is larger than the lower bound and smaller than the upper bound of the confidence interval.

- (D) The 95% confidence interval is a procedure with 95% probability of providing an interval that contains the population value.

Show explanations

Question 1
A confidence interval represents a range of values within which the true population parameter is likely to fall.
Question 2
The confidence level of a confidence interval is chosen by the researcher to determine the desired level of confidence in capturing the true parameter.
Question 3
A wider confidence interval indicates greater uncertainty or less precision in estimating the population parameter.
Question 4
As the sample size increases, the width of the confidence interval decreases, indicating increased precision in estimating the parameter.
Question 5
The relationship between the width of a confidence interval and the level of confidence is inverse: higher confidence leads to a wider interval.
Question 6
The purpose of a confidence interval in hypothesis testing is to assess the range of values where the population parameter might lie and make decisions about hypotheses.
Question 7
The standard error of the sample mean is calculated by dividing the standard deviation (2.5) by the square root of the sample size; add +/- 1.96* the standard error to the mean.
Question 8
The probability is in the procedure of calculating a confidence interval. We cannot make any probability statements about the population value, or about the results of specific confidence intervals.

25.3 Tutorial

25.3.1 Confidence Intervals

In this assignment we will work on some questions regarding the confidence interval. We focus on the confidence interval around the mean, but everything you learn can also be applied to

confidence intervals around the mean difference or regression coefficients.

Finish the following sentence. The confidence interval around the mean is constructed around the [9]

- (A) Population mean (under H0)
- (B) Sample mean

In the population the variable IQ is normally distributed with $IQ \sim N(\mu = 100, \sigma = 15)$.

Imagine that we drew 2000 samples from the population. For each of the samples we would calculate a 90% confidence interval around the sample mean. If you had to make a guess, how many intervals would you expect to contain the value 100? _____[10]

Imagine we drew a sample from the population and we calculated the 95% confidence interval around the sample mean for a particular variable. The lower bound of the confidence interval is equal to 85 and the upper bound to 95.

Which of the following statements is correct?

[11]

- (A) There is a 95% probability that the population mean lies between 85 and 95.
- (B) The probability that the population mean lies between 85 and 95 is 95%.
- (C) There is a 95% probability that if we would draw a new sample for the same population, the true population value lies between 85 and 95.
- (D) There is a 95% probability that a confidence interval calculated based on a sample from this population contains the true population value.

Confidence intervals are interpreted in terms of long-run probability. IF we could draw a huge number of samples from the population, 95% of those samples would provide a confidence interval that contains the population mean.

We can never know whether one specific confidence interval contains the population value, however.

So we can NEVER draw a conclusion like "there is a 95% probability that the population mean lies between 85 and 95".

[9] Sample mean
[10] 1800
[11] There is a 95% probability that a confidence interval calculated based on a sample from this population contains the true population value.

Recall the first lecture, in which I explained the idea of a "random experiment". Think of a 95% confidence interval as a random experiment with a 95% probability of containing the population value. One specific confidence interval is **not** a random experiment. Whether the population mean lies within the interval is not a matter of probability. It either does or it does not. We just don't know which of these is true.

Imagine a population with variable X, where $X \sim N(50, = 10)$

Assume a confidence level of 95% for all intervals.

You plan to draw a sample with $n = 20$ and compute a 95% confidence interval. What's the probability that this interval will contain 50? ___[12]%

Your colleague has already drawn a sample of $n = 20$. What's the probability that their confidence interval includes 50? [13]

- (A) 95%
- (B) Can't say
- (C) 100%
- (D) 0%

If you would draw 20 samples, how many samples would you expect the confidence interval to contain the value 50? ___[14]

True or false: if you draw 100 samples, 95 of them will provide a confidence interval that contains the population value. TRUE / FALSE[15]

Explanation

This is false because the phrase "will provide" is not a probability statements, but a deterministic one.
"The number of 95% confidence intervals out of 100 samples that contain the population value" is a random experiment. We expect an outcome of 95, but if we conduct this random experiment, the observed outcome may differ a little, e.g. 93, 94, 97 times are all fine.

All else being equal, what would you expect to happen to the confidence intervals of smaller samples? [16]

[12] 95
[13] Can't say
[14] 19
[15] FALSE
[16] They become wider

- (A) They stay the same
- (B) They become narrower
- (C) They become wider

Explanation

By increasing the sample size, our estimate becomes more precise. This will lead to more narrow confidence intervals.
Mathematically it also makes sense, because the confidence interval is based on the standard error. Remember that the formula for the standard error is $SE = \frac{}{n}$. A smaller sample size leads to a smaller standard error, which leads to a narrower interval.
Note that this does affect the probability of confidence intervals containing .

If the standard deviation increases (and everything else stays the same) the confidence interval will [17]

- (A) stay the same
- (B) become narrower
- (C) become wider

.

If we change the confidence level to 90%, the interval will [18]

- (A) become narrower
- (B) stay the same
- (C) become wider

and [19]

- (A) the same number of
- (B) fewer
- (C) more

intervals will contain .

[17] become wider
[18] become narrower
[19] fewer

26 Open science and questionable research practices

26.1 Introduction — Open Science and Questionable Research Practices

Over the past decade, large-scale replication efforts have shown that many published effects cannot be replicated, or finds much weaker effects when reproducing published work with high-powered designs (e.g., replication effects were, on average, about half the size of originals; 36% of replications were statistically significant versus 97% of originals) (Open Science Collaboration, 2015). One reason might be that scientific journals select for "newsworthy" (novel and exciting) findings, rather than unsurprising but diligently produced facts. In other words, there is a misalignment between what advances careers and what advances knowledge (Nosek et al., 2012). Scientific claims should earn credibility because they are based on **transparent and reproducible** research, not because results are surprising or the narrative is compelling.

This week examines how research practices and incentive structures can **inflate false positives** and **distort effect estimates**. Psychology has reported an unusually high share of significant findings (96%) despite typical studies being underpowered—conditions that favor publication bias and exaggerated effects (Bakker, van Dijk, & Wicherts, 2012). When samples are small, hypotheses are numerous, and analytic choices are flexible, the **positive predictive value** of a single significant result is low (Ioannidis, 2005). These problems motivate **open-science** reforms that reward accuracy over novelty: preregistration to reduce researcher degrees of freedom and to separate confirmatory from exploratory work; routine sharing of data, code, and materials; and explicit valuing of replication (Nosek et al., 2012; Bakker et al., 2020).

26.2 Scientific Fraud

While it is unlikely that scientific fraud is a main cause of the replication crisis in psychology and other fields, a highly publicized case of fraud did act as a catalyst, drawing attention to potential shortcomings in the way science was being conducted. Tilburg University professor Diederik Stapel, a prominent social psychologist, was found by the Levelt Committee to have fabricated and manipulated data across many studies, often supplying such fake datasets to PhD students and coauthors, unwittingly involving them in the fraud. The resulting papers

reported striking effects that were considered exemplary, passed peer review, were widely cited, and shaped research agendas - until independent teams failed to reproduce them. The case exposed structural failures of the scientific enterprise, including incentives for newsworthy findings, a lack of auditing of data and analysis code, and tolerance of researcher degrees of freedom that are also common in non-fraudulent questionable research practices (phacking, HARKing, low power). This case motivated the adoption of open science practices that seek to improve academic rigor, such as preregistration of study- and analysis plans prior to data collection, and transparent sharing of data, code, and materials.

More recently, the Francesca Gino scandal revived this debate. In 2023, Data Colada research auditors published a report of irregularities in several of Gino's papers on dishonesty (ironically). One striking finding was that the user history of an Excel spreadsheet seemed to indicate that several cases were moved from the control- to the experimental condition - and that this edit led to statistically significant findings.

While these cases are thought-provoking, it is unlikely that they are the sole cause for the replication crisis. The main contributing factors might be much more mundane.

26.3 Questionable Research Practices (QRPs)

Definition: QRPs are choices made in research and reporting that may be defensible but, when guided by the pursuit of statistical significance (motivated reasoning), end up inflating false-positive findings and effect sizes, thus distorting the published record. In psychology, unusually high rates of "positive" findings alongside typically low power indicate conditions under which QRPs and publication bias can thrive (Bakker et al., 2012).

Questionable research practices are especially problematic in relation to researcher degrees of freedom, because nearly any dataset can be tortured and sliced until it serves up a significant effect. This raises the proportion of false-positive findings in the literature.

26.3.1 Examples of QRPs:

- **Optional stopping / sequential testing:** Conducting an analysis, looking at the results, and adding more participants if the result is not (yet) significant ends up inflating the Type I error rate. This is especially problematic if you consider that, when the null hypothesis is true, p-values are uniformly distributed (all values equally likely). Thus, you can keep adding participants until by pure chance you find a significant result.
- **Outcome Switching / Selective reporting:** Measuring multiple dependent variables but reporting only those that are significant.
- **Flexible data cleaning:** Post hoc exclusions, outlier rules, and transformations chosen after seeing results and applied such that a hypothesized difference becomes significant, or an inconvenient significant difference disappears.

- **Researcher degrees of freedom in modeling:** After inspecting the data, trying multiple reasonable model specifications (e.g., control variables, mathematical transformations, exploring interaction effects) and picking the model with the "most interesting" results, or the one that supports the researcher's hypothesis.
- **HARKing (hypothesizing after results are known):** Constructing a hypothesis after seeing a surprising result, and then presenting this post hoc hypothesis as if it was specified before seeing the data. This blurs the distinction between confirmatory and exploratory research. Keep in mind that unexpected things do happen by chance, and creating a hypothesis after observing something unexpected does mean that the hypothesis will be significant, but does not mean that it is likely to be true. For example, Lucia de B. and Lucy Ledby were both convicted after an unusual event was observed (high proportion of infant deaths at their hospitals), and the post-hoc hypothesis was constructed that this must be the result of murder. However, rare events do happen, and they are not always the result of murder.
- **Publication bias ("file drawer" effect):** Studies with significant findings are more likely to be submitted and published than null results. When typical power is low (e.g., 20–40%), only about 20–40% of studies should be significant **even if effects are real**; yet many literatures report mostly significant findings. This indicates publication bias.

26.4 Well-Intended Flawed Practices

Even well-intended practices can misfire when applied uncritically. As briefly mentioned in the chapter on philosophy of science, NullHypothesis Significance Testing (NHST) is a prime example of how deeply entrenched scientific practices can be misleading. Gerd Gigerenzer called this the "null ritual": set up a straw man null hypothesis which states that the effect is zero, adopt $= .05$ as default significance level, reject the null hypothesis if $p < .05$ and interpret the result as positive evidence for some finding. This practice encourages dichotomous thinking, neglects effect sizes and uncertainty, and mixes the incompatible philosophies of testing from Fisher and Neyman–Pearson. This ritual ignores important principles - such as, that even trivial effects become "significant" in very large samples, and that extreme results (including effects that are significant by pure chance) are more common in small samples. What is the alternative? To move beyond the ritual and emphasize effect size estimation, uncertainty quantification, power analysis and sample size justification, and open science practices that allow others to replicate and audit findings.

26.5 Open Science Practices — Preregistration and Registered Reports

Open science practices aim to improve scientific rigor, and the efficiency with which knowledge can accumulate and errors can be corrected, by making many aspects of the scientific enterprise

open, accessible, transparent, and reproducible. We examine several open science practices.

26.5.1 Data Sharing

In a way, you have already enjoyed the benefits of data sharing: the exercises and portfolio assignments for this course make use of either real "open data", or fake data that was generated based on open data accessed by the author (Caspar van Lissa). Data sharing helps people reproduce published findings, and makes it possible to reuse existing datasets to test novel research questions.

26.5.2 Reproducibility

Reproducibility means being able to re-perform the same analysis with the same code using a different analyst (Patil et al., 2019). This can be achieved by creating research archives that include (Van Lissa et al., 2021):

1. Well-documented analysis code; for example, in the form of a dynamic document that combines the written text of a paper or research report (Introduction, Discussion) with the code required to generate the analysis results (e.g., SPSS syntax, R- or Python code). Such dynamic documents can be exported to any format, including PDF, or a website. This GitBook is an example of a dynamic document; both text and analysis results/figures are dynamically generated.
2. A complete time line of the research archive's historical development. Akin to a lab notebook that documents decisions made during the research process, modern "version control" systems make it possible to track, document, and preserve all changes to data and code from the moment a project is conceived, until it is preregistered, and data are collected, and it is ultimately published.
3. A time capsule of the computer environment (exact versions of software used), because the same analysis code can sometimes give different results on different systems. This is called "dependency management"; the most extreme form of dependency management is "containerization": creating a virtual computer to run the code.

26.5.3 Preregistration

Purpose: Reduce undisclosed flexibility by **specifying key decisions in advance**—hypotheses, primary outcomes, sampling plan and stopping rule, inclusion/exclusion criteria, randomization, and the primary analysis plan. Preregistration **separates confirmatory from exploratory work**; it does not forbid exploration. Deviations are permitted, but they should be **documented and justified**, with the original, time-stamped plan remaining visible. This transparency lets readers follow the study's evolution, evaluate analytic degrees of

freedom, and interpret results accordingly; it also encourages **follow-up confirmation** of exploratory findings with new data (Bakker et al., 2020; Nosek et al., 2012).

Quality matters. Effective preregistrations are **specific, precise, and comprehensive**. Structured templates with itemized prompts constrain opportunistic degrees of freedom better than unstructured formats, though **neither eliminates** flexibility entirely (Bakker et al., 2020).

What preregistration is not. It is not a ban on creativity. Exploratory analyses remain valuable, preregistration simply makes their status transparent and encourages follow-up confirmation with **new data** (Bakker et al., 2020).

26.5.4 Registered Reports

Model. A two-stage publication track in which the **study rationale, design, and analysis plan** are peer-reviewed **before data collection** (Stage 1). Upon **in-principle acceptance**, the journal commits to publish the results regardless of outcome **if** authors follow the approved protocol (Stage 2). This shifts incentives away from "significance" toward **design quality and theoretical contribution**, reducing publication bias and HARKing pressure (Nosek et al., 2012; Bakker et al., 2020).

Practical payoff. Registered Reports move the credibility test **upstream**. Peer review and in-principle acceptance occur **before** data collection, which (a) locks key design and analysis decisions, (b) commits publication regardless of outcome, and (c) requires transparent documentation of any deviations. By decoupling publication from statistical significance and limiting undisclosed flexibility, Registered Reports reduce publication bias, HARKing, and selective reporting, complementing preregistration with editorial enforcement at the design stage (Nosek et al., 2012; Bakker et al., 2020).

26.6 Lecture

Please watch this conference presentation by Dr. Amy Orben on questionable research practices:

https://www.youtube.com/embed/EIDY6TAy52M?si=SvbSK86hFga3JLq6

27 Formative Test

A formative test helps you assess your progress in the course, and helps you address any blind spots in your understanding of the material. If you get a question wrong, you will receive a hint on how to improve your understanding of the material.

Complete the formative test ideally after you've seen the lecture, but before the lecture meeting in which we can discuss any topics that need more attention

Question 1

A field reports about 90 percent significant findings while typical study power is about 35 percent. What is the most plausible interpretation? [1]

- (A) Replications will usually find larger effects
- (B) The field studies exceptionally large true effects
- (C) The nominal alpha is lower than 0.05 across studies
- (D) Publication bias and selective reporting are likely inflating the visible discovery rate

Question 2

Which practice best describes collecting an initial sample then checking results and adding participants until p is below 0.05 without a prespecified stopping rule or correction [2]

- (A) Proper sequential design
- (B) HARKing
- (C) Optional stopping without error control
- (D) Publication bias

[1] Publication bias and selective reporting are likely inflating the visible discovery rate
[2] Optional stopping without error control

Question 3

A study measures five outcomes but reports only the one that is significant as the primary outcome while omitting the rest. Which QRP is this [3]

- (A) Appropriate parsimony
- (B) Flexible data cleaning
- (C) Model fishing
- (D) Selective outcome reporting

Question 4

After inspecting results the authors exclude outliers using a post hoc rule that increases the effect size. Which problem is most salient [4]

- (A) Correct robustness analysis
- (B) Registered Report compliance
- (C) Flexible data cleaning that capitalizes on chance
- (D) Predefined exclusion criteria

Question 5

Authors try many plausible models with different covariates and transforms and present only the one that yields p below 0.05. This primarily illustrates [5]

- (A) Correct multiple testing adjustment
- (B) Proper sensitivity analysis
- (C) Model fishing
- (D) Randomization failure

Question 6

A moderation effect is described as predicted but the interaction was considered only after seeing the data. What is this called [6]

[3] Selective outcome reporting
[4] Flexible data cleaning that capitalizes on chance
[5] Model fishing
[6] HARKing

- (A) Preregistration
- (B) HARKing
- (C) Optional stopping
- (D) Outcome switching

Question 7

With ten independent tests at alpha equals 0.05 and no adjustment what is the approximate probability of at least one false positive [7]

- (A) About 40 percent
- (B) About 95 percent
- (C) About 10 percent
- (D) About 5 percent

Question 8

Which statement best captures the purpose of preregistration in this course [8]

- (A) Specify key decisions in advance to separate confirmatory from exploratory work and reduce undisclosed flexibility
- (B) Guarantee replication success
- (C) Ban exploratory analyses
- (D) Prevent all deviations from the plan

Question 9

Which preregistration description is strongest [9]

- (A) We will choose outcomes after data collection to reduce noise

[7] About 40 percent

[8] Specify key decisions in advance to separate confirmatory from exploratory work and reduce undisclosed flexibility

[9] Primary outcome PSQI total; hypothesis treatment lowers PSQI versus control; N equals 200 with a fixed stopping rule; exclusions preregistered attention check; analysis linear model with prespecified covariates; alpha equals 0.05; secondary outcomes labeled exploratory

- (B) Primary outcome PSQI total; hypothesis treatment lowers PSQI versus control; N equals 200 with a fixed stopping rule; exclusions preregistered attention check; analysis linear model with prespecified covariates; alpha equals 0.05; secondary outcomes labeled exploratory

- (C) We predict positive effects on several variables and will include covariates as needed

- (D) We will analyze outcomes using appropriate models and stop when effects stabilize

Question 10

Which statement about Registered Reports is accurate [10]

- (A) Methods are reviewed only after results are known

- (B) In principle acceptance before data collection commits publication if the approved protocol is followed regardless of outcome

- (C) Acceptance depends on obtaining p below 0.05

- (D) They eliminate the need to share data or code

[10] In principle acceptance before data collection commits publication if the approved protocol is followed regardless of outcome

Show explanations

Question 1
If average power is about 35 percent then only about 35 percent of studies should be significant even when effects are real. Much higher rates suggest selection on significance and QRPs
Question 2
Peeking and topping up until a threshold inflates the Type I error unless a sequential design with error spending is prespecified
Question 3
Reporting only significant outcomes increases false positives and overstates effects in the visible record
Question 4
Data dependent exclusion rules raise false positives and bias estimates
Question 5
Undisclosed multiple testing across specifications inflates the effective Type I error and effect estimates
Question 6
Hypothesizing after results are known presents post hoc explanations as if they were a priori predictions
Question 7
The family wise error is 1 minus 0.95 to the power of 10 which is about 0.40
Question 8
Preregistration clarifies plans and reduces hidden degrees of freedom. It does not forbid exploration or ensure positive results
Question 9
Specific precise and comprehensive plans reduce undisclosed flexibility and clarify confirmatory versus exploratory analyses
Question 10
Registered Reports decouple publication from results and move review upstream to focus on theory and design quality

28 Tutorial

28.1 Assignment 1: Spot the Practice — QRPs or Good Methods?

Below are short methods/results fragments from fictional studies. Read each one closely. For each fragment:

1) Identify whether it (potentially) shows a **questionable research practice (QRP)** or **good practice**.

2) Name the specific issue (e.g., optional stopping, outcome switching, flexible data cleaning, model fishing, HARKing, or good practice).

3) Explain **why** it matters.

4) Propose a **minimal fix** (what the authors should have done or reported).

5) Decide whether the claim should be labeled **confirmatory** or **exploratory** in the write-up.

Fragment A

> Our sample consisted of two cohorts of first-year bachelor's students: 74 students enrolled in 2023, and 82 students enrolled in 2024. We found a statistically significant effect, t(154), p = .047, which we interpret as evidence that the intervention improves well-being.

Fragment B

> **Experimental Design** We randomly assigned participants to be in the social support or control condition. The social support condition received daily encouraging messages, supposedly sent by another participant in the study. The control condition received daily informative messages, taken from WikiPedia. **Measurements** We measured participants' stress, mood, sleep, productivity, and affect variability using validated self-report questionnaires. **Results** In line with predictions, participants in the social support condition showed a significant improvement in mood, p = .03.

Fragment C

Participants with unusually fast reaction times (log reaction time < M 2 SD; n = 4) were considered outliers and were excluded from the study.

Fragment D

We tested several plausible specifications (adding/removing covariates such as age, SES, hours worked; linear vs. log transforms). The model controlling for age and using a log transform yielded a significant effect (p = .041), so we focus on this specification below. Other models are not shown.

Fragment E

Introduction: The present study hypothesized that herbal tea would increase sleep quality. Our sample consisted of patients who presented with disturbed sleep but were not considered eligible for other medical treatment. **Results**: Consistent with expectations, we found a significant effect of herbal tea on sleep quality for participants who scored high on baseline stress, p = .02.

Fragment F

Before data collection, we preregistered our hypotheses, primary outcome (PSQI total score), stopping rule (N = 200), and analysis plan (linear model with preregistered covariates: age, gender). We also specified exclusion criteria (failed attention check; PSQI missingness > 20%). Deviations: We added one robustness check using a median-split sensitivity analysis (exploratory; reported in Supplement). Data, code, and materials are available on OSF (anonymized).

Fragment G

The study protocol (theory, design, and analysis) received in-principle acceptance prior to data collection. Results were published regardless of outcome, provided fidelity to the approved plan. The primary effect was not significant (p = .21); exploratory analyses are labeled and reported in the Supplement.

Discuss with your group:

a. Did you all agree on the label (QRP vs. good practice) and the specific issue?

b. Which fixes would you prioritize if the authors could change only one thing?

c. How would preregistration or a Registered Report have changed the design, analysis, or write-up?

28.2 Assignment 2: Preregistration Audit — Make It Specific

Below is an excerpt from a **vague** preregistration. Your task is to (i) spot ambiguities and (ii) rewrite the excerpt so it is **specific, precise, and comprehensive**.

Vague preregistration excerpt:

> We will test whether our workshop improves student success. We'll recruit around 150 students and stop once effects stabilize. We'll measure several outcomes related to performance and well-being and analyze them using appropriate models. Outliers will be excluded. Demographic control variables will be included.

Your tasks:

1) Underline every ambiguity, pinpoint what the risk of QRPs is
2) Rewrite the preregistration to make it more robust against QRPs. Pinpoint how each change improves the preregistration.

Hints

Consider specifying:

- **Primary outcome** (exact variable/scale and scoring); **secondary outcomes** (if any).
- **Hypotheses** (directional; one line per test).
- **Sampling plan and stopping rule** (target N; any interim looks; conditions for stopping).
- **Inclusion/exclusion** and **outlier rules** (exact thresholds, applied blindly if feasible).
- **Analysis plan** (model form, covariates, sidedness, alpha, multiple-testing plan).
- **Deviations policy** (how you will document any changes).
- Distinguishing which analyses are **confirmatory** and which are **exploratory**.

28.3 Psilocybin Liberates the Entrenched Brain?

In recent years, enthusiasm has grown for psychedelics (like magic mushrooms) as a treatment for depression and other mental health problems. However, the methodological rigor of studies

that find support for psychedelics' efficacy has been severely criticized. Professor Eiko Fried, one of the main contributors to the methodological critiques, engaged in a real-life exercise, very similar to the one you just completed.

With your group, pick one of the following sources:

- Eiko Fried's blog post, summarizing the critiques
- The original authors' rebuttal of the critiques (which is separated into seven "Points"; pick one point)

Discuss: Do you find the arguments (for or against the presence of QRPs) persuasive? Which QRPs do you recognize?

28.4 Mini Registered Report Pitch

With your group, prepare a single presentation slide for an **imaginary study** that you could use for your portfolio assignment (i.e., you can use the same hypothesis as for your portfolio).

Describe the:

- **Theory** (1–2 sentences) and **confirmatory hypothesis** (directional).
- **Study Design** (sampling strategy, assignment or not, sample size, stopping rule, exclusion criteria).
- **Primary outcome** (exact measure) and **analysis plan** (model, covariates, sidedness, alpha, multiple-testing plan if applicable).
- **Transparency** (data/code/materials; prereg link to be created; planned deviations policy).

If there is time, groups present and receive peer feedback focusing on **clarity**, **testability**, and **reducing flexibility** at design time.

A Data for Portfolio

You will need to use an appropriate data source for your portfolio assignments. Although you are welcome to use any data that you consider to be suitable for making the assignments - including data sets that you have previously collected, or open access data sources - we want to make sure that everybody has a backup option that meets the course requirements. Below, we introduce three data sources that have been customized for the three major tracks. You can follow your own interests in selecting one of these data sources.

NOTE: Make sure to enter your correct group number when you generate the data; you must use a file that is unique to your group.

To generate a synthetic data set based on the sources described below, visit this link.

A.1 SS: Values and Beliefs about Individuals and Collectives

This synthetic dataset was inspired by Wave 7 of the World Values Survey (Haerpfer et al., 2022).

The World Values Survey (WVS) is a global research project that explores people's values and beliefs and what social and political impact these have. Among topics covered are support for democracy, tolerance of ethnic minorities, support for gender equality, the role of religion and changing levels of religiosity, the impact of globalization, attitudes toward the environment, work, family, politics, national identity, culture, diversity, insecurity, and subjective well-being. This data source is used by governments, scholars, and international organizations like the United Nations.

Examples of research questions:

- What proportion of participants considers work to be very important in life? (Q5)
- What proportion of participants score more extreme than 9/10 on a left-right political ideology scale? (Q240)
- Is trust in the government significantly higher than the neutral middle of the scale (3)? (Q292O)
- Does participants' age predict the attitude that children should take care of their parents? (Q38 and Q262)

Data documentation: https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

Reference: Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi:10.14281/18241.20.

A.2 CN: Behavioral and Neural Correlates of Empathy in Adolescents

This synthetic dataset was inspired by a study by Overgaauw and colleagues (2014).

Adolescence is characterized by significant changes in how individuals perceive and interact with others, both cognitively and emotionally. Empathy is a crucial element in appropriately responding to the emotions and actions of others. It is often described as the capacity to understand and share the emotional experiences of others, enabling us to comprehend and anticipate their intentions. Children who possess higher levels of empathy demonstrate greater emotional regulation and engage in more prosocial behavior towards others. This experimental study presented adolescents with either positive or negative social situations, and asked them to focus either on person A or person B in those situations (in negative situations, person A was the perpetrator and person B was the victim). They then measured how many coins participants were willing to give to the focal person. Empathy was measured using a scale with three sub-dimensions of empathy (Contagion, Understanding, and Support), and brain activation in several regions of interest was measured.

Reference: Overgaauw, S., Gürolu, B., Rieffe, C., & Crone, E. A. (2014). Developmental Neuroscience, 36 (3-4). Behavior and Neural Correlates of Empathy in Adolescents. https://doi.org/10.1159/000363318

A.3 BE: Sustainable Food Choices

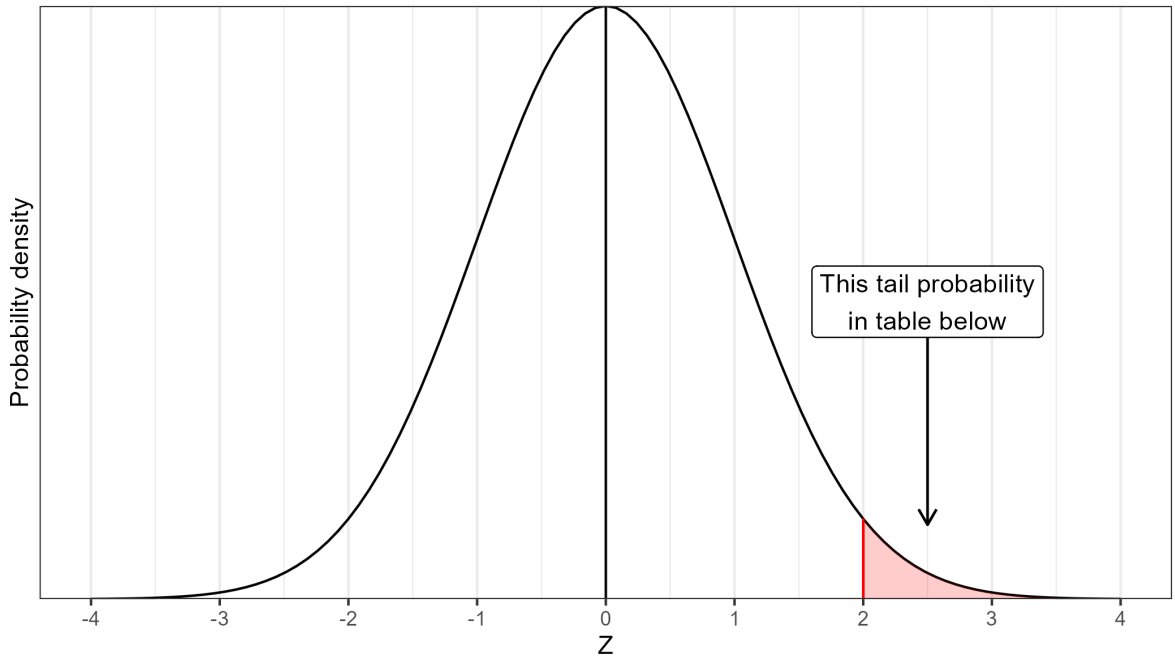
This synthetic dataset was inspired by a study by De Boer and colleagues (2007).

Sustainability goals may require people in Western countries to reduce their meat consumption. This study investigated which values motivate sustainable food choices related to meat consumption. The researchers surveyed 1530 Dutch consumers and found that various human values were related to different food choice motives. Universalism, in particular, had a unique impact on food choices that favored reduced meat, or free-range meat consumption. This study provided insight into the way values, motives and attitudes influencing sustainable food choices and shape individuals' dietary decisions.

Reference: Joop de Boer; Carolien T. Hoogland; Jan J. Boersema (2007). Towards more sustainable food choices: Value priorities and motivational orientations. Food Quality and Preference, 18(7), 0–996. doi:10.1016/j.foodqual.2007.04.002.

B Z-table

Table gives the right-tail probability corresponding to a Z-value of the value in the Z-column plus the value in the column name, which indicates the second digit of the Z-score.



Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

B.1 t-table

Table gives the t-value corresponding to the right-tail probability indicated by the columns.

df	p = 0.4	p = 0.25	p = 0.1	p = 0.05	p = 0.025	p = 0.01	p = 0.005	p = 0.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144

df	p = 0.4	p = 0.25	p = 0.1	p = 0.05	p = 0.025	p = 0.01	p = 0.005	p = 0.001
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385

C Formula sheet

C.1 General Part

Mean: $X = \frac{\sum_{i=1}^{n} x_i}{N}$

Variance: $S_x^2 = \frac{\sum_{i=1}^{n} (x_i x)^2}{n1}$

Standardized values (Z-values): $Z = \frac{X}{}$

Z-statistic in one sample Z-test: $Z = \frac{x_x}{x}$

Standard error of the mean: $_x = \frac{}{n}$

Cohen's d: $\frac{X_1 X_2}{s_{pooled}}$

$s_{pooled}^2 = \frac{(n_1 1)s_1^2 + (n_2 1)s_2^2}{n_1 + n_2 2}$

$s_{pooled} = s_{pooled}^2$

F-statistic in one-way ANOVA: $F(df_b, df_w) = \frac{(SS_b/df_b)}{(SS_w/df_w)} = \frac{MS_b}{(MS_w}$

Simple regression model: $Y = b_0 + b_1 X$

Multiple regression model: $Y = b_0 + b_1 X_1 + b_2 X_2$

Explained variance: $R^2 = \frac{s_y^2}{s_y^2}$

t-statistic in a one sample t-test: $t = \frac{X_{H0}}{se_x}$, where $se_x = \frac{s_x}{n}$, $df = n\ 1$

t-statistic in an independent samples t-test: $t = \frac{(X_1 X_2)(_1 _2)_{H0}}{se_{x_1 x_2}}$

$se_{x_1 x_2} = s_{pooled}^2 (\frac{1}{n_1} + \frac{1}{n_2})$

C.2 Business and economics

Logistic function: $P(Y = 1|X) = \frac{e^{(b_0+b_1X)}}{1+e^{(b_0+b_1X)}}$

From probability to odds: odds $= \frac{P}{1P}$

From odds to probability: $P = \frac{\text{odds}}{1+\text{odds}}$

From odds to logit: logit $= \ln(\text{odds})$

From probability to logit: logit $= \ln\left(\frac{P}{1P}\right)$

From logit to odds: odds $= e^{\text{logit}}$

From logit to probability: $P = \frac{e^{\text{logit}}}{1+e^{\text{logit}}}$

Wald test statistic: $W = (\frac{b}{se_b})^2$

C.3 Cognitive neuroscience

Number of Possible Pairwise Comparisons: k Œ $\frac{(k1)}{2}$

Factorial ANOVA Linear Model: $Y_{jkl} = {}_Y + {}_k + {}_l + {}_{kl} + {}_{jkl}$

Eta-squared for Factor A: ${}^2_A = \frac{SS_A}{SS_{total}}$

Partial eta-squared for Factor A: ${}^2_{partial.A} = \frac{SS_A}{SS_A+SS_w}$

Adjusted Mean: $Y_{i(adj)} = Y_i \;\; b_w(X_i \;\; X)$

t-Statistic in Paired Samples t-Test: $t = \frac{d}{\frac{s_d}{n}}, \quad \text{df} = n \;\; 1$

C.4 Social Sciences

Reliability: $r_{xx} = \frac{\text{var}(T)}{\text{var}(X)} = \frac{\text{var}(T)}{\text{var}(T)+\text{var}(E)}$

Eigenvalue of Component 1 for 6 Items: ${}_1 = a_{11}^2 + a_{21}^2 + a_{31}^2 + a_{41}^2 + a_{51}^2 + a_{61}^2$

The proportion of Variance Accounted For by component 1 (when there are J items) is: Proportion VAF $= \frac{1}{\text{TotalVar}} = \frac{1}{J}$

Component loadings for component 1 and item j are represented as: $a_{j1} = r_{X_jC_1}$

Communality for 2 Components: $h_{j2} = r^2_{XjC1} + r^2_{XjC2} = a^2_{j1} + a^2_{j2}$

Unicity for 2 Components: $b_{j2} = 1 \;\; h_{j2}$

D References

DeBruine, L. M., & Lakens, D. (n.d.). *Methods Book Template*. Retrieved June 4, 2025, from https://debruine.github.io/booktem/

Halpern, J. Y. (2015, May 1). *A Modification of the Halpern-Pearl Definition of Causality*. https://doi.org/10.48550/arXiv.1505.00162

Hoijtink, H., Bruin, J. de, Duken, S. B., Flores, J., Frankenhuis, W., & Lissa, C. J. van. (2023). *The Open Empirical Cycle for Hypothesis Evaluation in Psychology*. https://doi.org/10.31234/osf.io/wsxbh

Morabia, A. (2013). Hume, Mill, Hill, and the Sui Generis Epidemiologic Approach to Causal Inference. *American Journal of Epidemiology*, *178*(10), 1526–1532. https://doi.org/10.1093/aje/kwt223

Patil, P., Peng, R. D., & Leek, J. T. (2019). A visual tool for defining reproducibility and replicability. *Nature Human Behaviour*, *3*(7), 650–652. https://doi.org/10.1038/s41562-019-0629-z

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-SS057

Peikert, A., Ernst, M. S., & Brandmaier, A. M. (2023). *Why does preregistration increase the persuasiveness of evidence? A Bayesian rationalization* [Preprint]. https://osf.io/cs8wb. https://doi.org/10.31234/osf.io/cs8wb

Van Lissa, C. J. (2022a). Developmental data science: How machine learning can advance theory formation in Developmental Psychology. *Infant and Child Development*, *32*(6), 1–12. https://doi.org/10.1002/icd.2370

Van Lissa, C. J. (2022b). Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Religion, Brain & Behavior*, *0*(0), 1–5. https://doi.org/10.1080/2153599X.2022.2070254

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A.-L., Peikert, A., Struiksma, M. E., & Vreede, B. M. I. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, *4*(1), 29–49. https://doi.org/10.3233/DS-210031