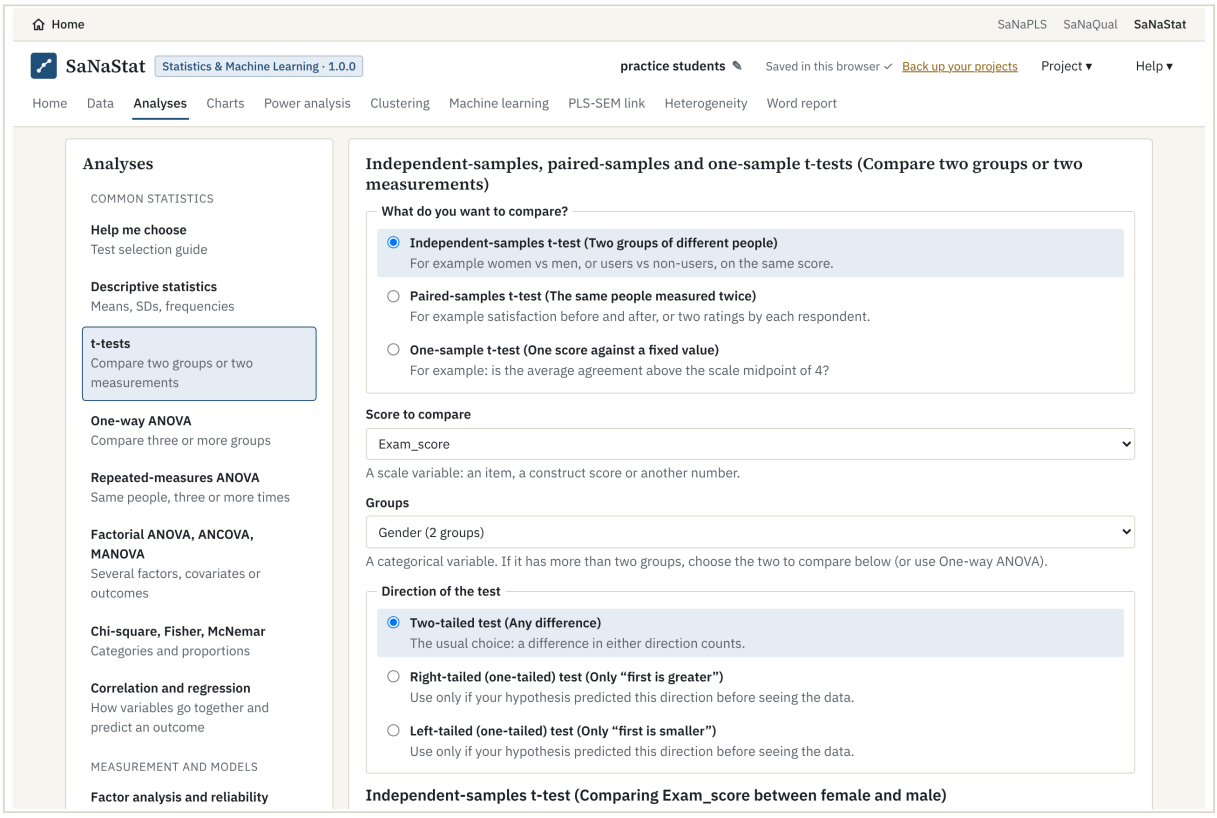
SANASOFT · USER MANUAL

SaNaStat

Common Statistical Tests: a step-by-step manual with practice data

Version 1.0.0 · October 2026



Worked example: practice_students.csv, 400 simulated students

Prof. S. Sabraz Nawaz and Prof. Ghazanfar Ali
SaNaSoft · 2026

Contents

About this manual....5
Who this manual is for....5
The practice data are simulated....6
What you need....7
How to read the instructions....7
The practice files....7
1. Getting started....9
1.1 Open SaNaStat....9
1.2 Open the practice file....10
1.3 Check the variable types....10
1.4 The variable editor....11
1.5 Missing data and missing-value codes....12
1.6 View the data....13
1.7 Scale scores and reliability....14
2. Preparing variables....17
2.1 Your own formula....17
2.2 If ... then ... else....19
2.3 Recode with a table....20
2.4 Reverse-code several items at once....20
2.5 A mean across items, with a minimum number answered....21
2.6 Transform a scale variable: a logarithm of Anxiety_score....22
2.7 Groups from a scale variable....23
2.8 An indicator and merged categories....24
2.9 A gain score from two variables....24
2.10 Select cases....24
2.11 Split file and weight cases....26
2.12 Restructure the data....28
3. Checking data quality....30
3.1 The screening settings....30
3.2 Reading the screening report....31
3.3 Outliers and missing values....32
3.4 Your decision....33
4. Describing your data....35
4.1 Descriptive statistics....35
4.2 The Charts screen....37
4.3 Histogram: the shape of one variable....37
4.4 Box plot: medians, spread and unusual values....38
4.5 Means with confidence intervals....38
4.6 Scatter plot with a fitted line....39
4.7 Bar chart of percentages....40
4.8 Further statistics and stem-and-leaf plots....41

5. Choosing the right test....43
5.1 The Help me choose guide....43
5.2 The decision table....43
5.3 Parametric or rank-based?....45
5.4 Multiple tests and error rates....45
5.5 The significance level α....45
6. Comparing two groups or two measurements....47
6.1 Independent-samples t-test: Exam_score by Gender....47
6.2 Effect sizes and their confidence intervals....49
6.3 Paired-samples t-test: Pre_score and Post_score....50
6.4 Is the outcome normal enough? Shapiro–Wilk and Q–Q plots....52
6.5 One-sample t-test: Exam_score against the pass mark....53
7. Comparing three or more groups....55
7.1 One-way ANOVA: Post_score by Teaching_method....55
7.2 Effect sizes with confidence intervals....56
7.3 Which groups differ? Choosing the post hoc test....57
7.4 The rank-based alternative: Kruskal–Wallis and Dunn....60
7.5 Unequal variances: Welch's ANOVA and Games–Howell....61
7.6 The gain score....62
8. The same people measured repeatedly....64
8.1 Run the analysis....64
8.2 Where is the change? Pairwise comparisons....65
8.3 The rank-based alternative: Friedman....66
9. Factorial ANOVA, ANCOVA and MANOVA....67
9.1 Factorial ANOVA: Exam_score by Teaching_method and Gender....67
9.2 ANCOVA: adding the pre-test as a covariate....69
9.3 One-way MANOVA: Exam_score and Project_score together....70
10. Categories and proportions....72
10.1 Frequencies of one categorical variable....72
10.2 Chi-square test of independence: Gender and Programme....74
10.3 Teaching_method and Passed: reading row percentages as rates....76
10.4 Two-by-two tables: Gender and Passed....77
10.5 Goodness of fit: are the programmes equally common?....77
10.6 McNemar test: did the answer change?....78
10.7 A third variable: layered tables and the Mantel–Haenszel test....79
11. Correlation and regression....82
11.1 Pearson correlation....82
11.2 Spearman and Kendall: rank correlations....84
11.3 Partial correlation....84
11.4 Multiple linear regression....85
11.5 Hierarchical regression: what does a second block add?....86
11.6 Binary logistic regression: predicting Passed....86
11.7 Ordinal logistic regression: predicting a 1 to 5 rating....88
11.8 Multinomial logistic regression: predicting the programme....91

11.9 Stepwise selection: letting the data choose the predictors....93
12. Measurement: reliability and validity of scales....96
12.1 Reliability: Cronbach's α and McDonald's ω....96
12.2 Exploratory factor analysis....97
12.3 Confirmatory factor analysis, AVE, HTMT and Fornell–Larcker....102
12.4 Agreement between raters: Cohen's kappa and the ICC....105
13. Mediation and moderation....108
13.1 Mediation: does the post-test carry the effect of the pre-test?....108
13.2 Moderation and the Johnson–Neyman technique....111
13.3 Moderated mediation: models 7 and 14....114
14. Missing data and multiple imputation....116
14.1 Set up the imputation....116
14.2 Diagnostics: did the imputation work?....117
14.3 Pooled results....119
15. Power analysis and sample size planning....122
15.1 The three kinds of power analysis....122
15.2 A priori: how many students do I need?....124
15.3 Sensitivity: what can I detect with the students I have?....125
15.4 Post hoc power: use with care....125
15.5 Planning a hierarchical regression or a factorial design....125
16. Machine learning as a check on regression....129
16.1 Set up the comparison....129
16.2 Reading the comparison....130
16.3 Which predictors matter, and how?....131
16.4 Calibration....132
16.5 A yes/no outcome: cut-offs, precision and recall....132
17. Reporting and exporting your results....136
17.1 Copying and exporting one table or chart....136
17.2 Collecting results in the Word report....137
17.3 A reporting checklist....141
Appendix A. Codebook of practice_students.csv....143
Appendix B. Assumptions and benchmarks of the tests....145
Appendix C. Glossary....148
References....151

About this manual

This manual teaches the common statistical tests with SaNaStat, the statistics and machine learning program of the SaNaSoft suite. It follows one practice data file, practice_students.csv, from opening the file to a finished Word report: data checking and preparation, descriptive statistics and charts, t-tests, analysis of variance, repeated measures, chi-square tests, correlation and regression, the reliability and validity of scales, mediation and moderation, missing data, power analysis, and a short machine learning check on a regression. Every chapter says what to click, what you will see, which rule of thumb applies, how to read the result and how to report it. All screenshots and numbers come from real runs of SaNaStat 1.0.0, so you can follow along and obtain the same results.

New in this edition (SaNaStat 1.0.0)

SaNaML is now called **SaNaStat**, at the address /stat/ (old links lead there). Projects saved in your browser stay, and project files are now saved as .sanastat; older .sanaml files still open. The methods and results are those of SaNaML 0.15.0. That version added the following, which this edition covers:

Sections 2.11 and 2.12 show the new Data cards: split file, weight cases and restructuring the data. Section 7.3 adds the Scheffé, Šidák, LSD, Dunnett and Student–Newman–Keuls post hoc tests, Section 4.8 the further statistics and stem-and-leaf plots, Section 10.7 layered cross-tables with the Mantel–Haenszel and Breslow–Day tests, Section 11.9 stepwise selection, Section 12.2 saved factor scores and Section 12.4 the new Agreement screen (Cohen's kappa and the intraclass correlation).

Who this manual is for

Postgraduate students and researchers who need to choose, run and report common statistical tests. You do not need to know any statistics software. Each statistical idea is explained in plain language in a "Concept" box when it first appears; if you already know the idea, skip the box. The chapters can be read in order, or you can go straight to the test you need using the contents list, the chapter map below and the test-choice table in Chapter 5.

Table 1

What each chapter covers

Chapter	Topic	What you learn
1	Getting started	Open the data, check the variable types (including Ordinal and ID), missing values and missing-value codes, the variable editor and the data view, scale scores with Cronbach's α and McDonald's ω

Chapter	Topic	What you learn
2	Preparing variables	Formulas, if–then rules, recode tables, reversing several items, means across items, logarithms, groups and indicators, Select cases, split file and weights, restructuring
3	Checking data quality	Careless responding, outliers, missing data, whom to exclude
4	Describing your data	Descriptive statistics, further statistics, stem-and-leaf plots and the six chart types
5	Choosing the right test	The test selection guide, the decision table, parametric or rank-based, the significance level α
6	Two groups or two measurements	t-tests with Shapiro–Wilk, Q–Q plots and effect sizes with confidence intervals
7	Three or more groups	One-way ANOVA, the choice among ten post hoc tests, homogeneous subsets, Kruskal–Wallis
8	Repeated measures	Sphericity, Greenhouse–Geisser and Huynh–Feldt, Friedman
9	Factorial ANOVA, ANCOVA and MANOVA	Main effects, interactions, covariates, several outcomes
10	Categories and proportions	Frequencies, chi-square, Fisher's exact and McNemar tests, layered tables (Mantel–Haenszel)
11	Correlation and regression	Pearson, Spearman, Kendall and partial correlation; linear, binary, ordinal and multinomial logistic regression; stepwise selection
12	Measurement	Reliability (α and ω), exploratory factor analysis with saved factor scores, confirmatory factor analysis, AVE, HTMT and Fornell–Larcker, Cohen's kappa and the ICC
13	Mediation and moderation	Indirect effects with bootstrap intervals, interactions with Johnson–Neyman, moderated mediation
14	Missing data	Multiple imputation by chained equations, diagnostics and pooled results
15	Power analysis	Sample size, sensitivity and power
16	Machine learning as a check	Out-of-sample prediction, explanations, cut-offs
17	Reporting and exporting	Copying and exporting tables and charts, the report builder, a reporting checklist

The practice data are simulated

The file practice_students.csv describes 400 students of a made-up university. **The data are simulated.** No real person appears in them, and the effects in the file were planted on purpose (for example, women score a little higher than men, and a flipped classroom

raises the post-test). You can therefore check whether your analysis finds what was built in. Do not report these numbers as findings about real students. The exercises at the end of most chapters use the other practice files that ship with SaNaStat (see "The practice files" below).

What you need

- A computer with an up-to-date browser (Google Chrome, Microsoft Edge, Mozilla Firefox or Safari). SaNaStat runs inside the browser; your data are read on your own computer and are not uploaded.
- The practice file practice_students.csv and, for the exercises, practice_marketing.csv, practice_brand_survey.csv, practice_hr.csv, practice_engagement.csv, segments_demo.csv and careless_demo.csv. Your own data should be a CSV, .xlsx spreadsheet or .sav / .zsav file with one row per person and one column per variable, with the column names in the first row. Missing values are blank cells (or codes such as −99 that you declare, Section 1.5).
- A day or two: a chapter takes between 20 and 45 minutes if you repeat each step.

How to read the instructions

Table 2
Conventions used in this manual

Convention	Meaning	Example
Bold coloured text	A button, tab, menu item, box or list on the screen. Click or choose exactly this.	**Create the construct score**
→	A sequence of clicks.	**Data** → **Choose a file…**
Numbered steps	Actions to carry out in this order.	1. Open … 2. Click …
Italics	A technical term explained nearby, or text exactly as it appears in the results.	*Welch's t-test*

The practice files

Table 3
The practice files that ship with SaNaStat

File	What it describes	Used in this manual
practice_students.csv (400 rows)	A university study: tests, ANOVA, regression, scales, mediation, missing data, machine learning	All chapters
practice_marketing.csv (1,500 rows)	A retailer's customers: A/B test, spend, churn	Exercises in chapters 2, 4, 6, 7, 10, 11, 15, 16 and 17
practice_brand_survey.csv (500 rows)	A brand perception survey, 1 to 7 ratings	Exercises in chapters 1 and 3

File	What it describes	Used in this manual
practice_hr.csv (1,200 rows)	A company's employees: pay, training, attrition	Chapter 7 (Games–Howell) and exercises
practice_engagement.csv (600 rows)	An engagement survey, three scales	Exercises in chapters 1 and 7
careless_demo.csv (300 rows)	A survey with planted careless responders	Exercise in chapter 3
segments_demo.csv (360 rows)	Customer segments	Charts exercise in chapter 4
practice_counts.csv (36 rows)	A table of counts: year, gender, programme, result and the number of students	Section 2.11 (weights)

Each file has an identifier column (StudentID, CustomerID and so on). SaNaStat recognises it and gives it the type ID: it is shown and kept, but never offered to an analysis.

Feedback on this manual is welcome at sabraz@seu.ac.lk.

1. Getting started

1.1 Open SaNaStat

Open SaNaStat in your browser at the address of your installation. It opens on its Home screen (Figure 1). The header shows the program name with its version ("Statistics & Machine Learning · 1.0.0"), the **Project ▾** and **Help ▾** menus and, below them, the screens: **Home**, **Data**, **Analyses**, **Charts**, **Power analysis**, **Clustering**, **Machine learning** and **Word report** (in the SaNaSoft suite also **PLS-SEM link** and **Heterogeneity**). Once you have added results to the report, the last item shows their number, for example **Word report (3)**. The **Home** link in the thin bar at the very top (and at the bottom left) leads to the start page; your work is kept. The Getting Started guide explains projects, the help menu and the guided tour; this manual does not need them.

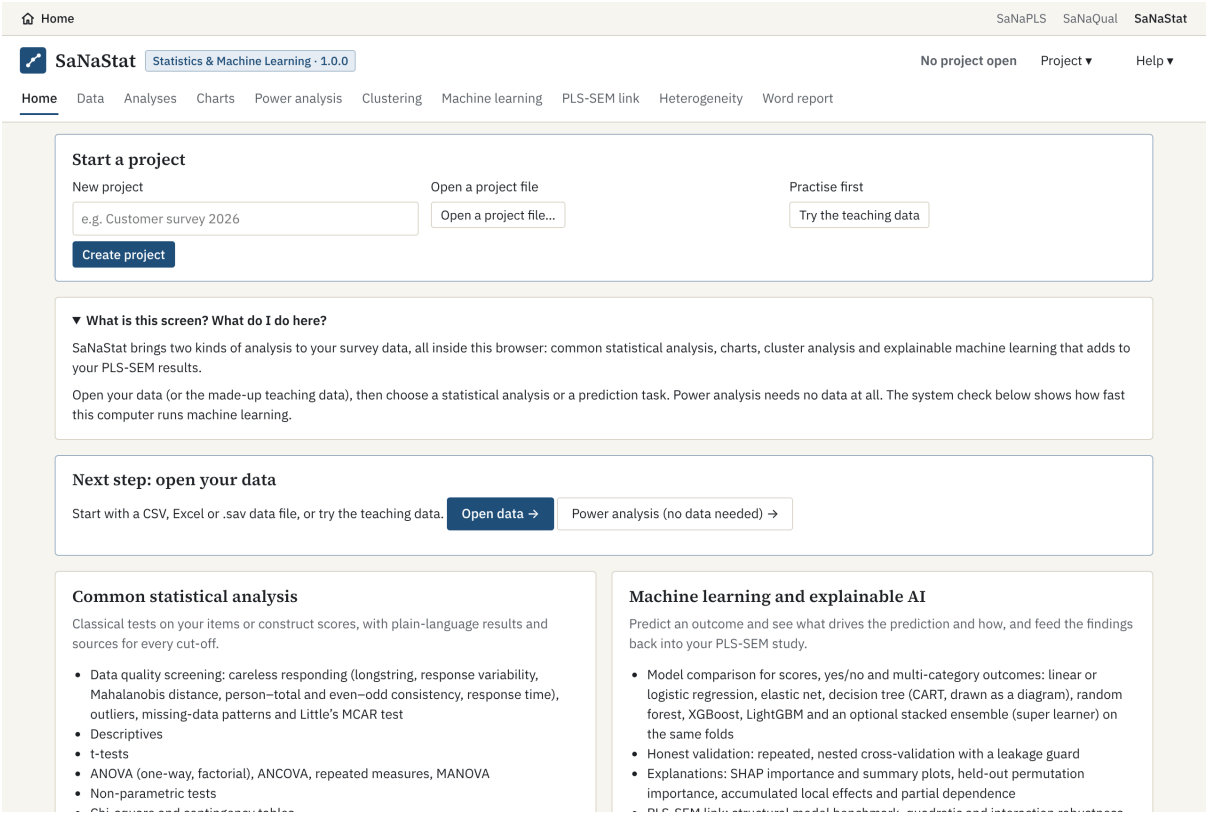


Figure 1. *The Home screen of SaNaStat.*

Every screen starts with a box *What is this screen? What do I do here?*, which explains the screen in two or three sentences. Screens that need data show "No data yet" with a **Go to Data →** button until you open a file. SaNaStat keeps your data, results and report in a project in this browser and restores them on your next visit; they are lost if you clear the browser's site data, so download the Word report (chapter 17) and export the project file to keep your work.

Note: what this manual covers

This manual covers the common statistical tests, the way to prepare data for them, the measurement of scales, mediation and moderation, and missing data. The Analyses screen also lists HR and marketing analyses (survival analysis, adverse impact and pay gaps, A/B tests, RFM, market basket and conjoint analysis), and SaNaStat has screens for clustering and a full machine learning comparison; the HR Analytics and Marketing Analytics manuals cover them. This manual uses machine learning only as a short check on a regression (chapter 16).

1.2 Open the practice file

1. Click **Data** in the menu.
2. Click **Choose a file…** and select practice_students.csv. SaNaStat reads the file in the browser.
3. SaNaStat reports "Open practice_students.csv: 400 respondents, 33 variables." **Close this data** closes the file again. If a project already holds data, SaNaStat asks whether to start a new project with the new file or to replace the data; start a new project.

1.3 Check the variable types

The second card, *2 · Check the variable types*, lists every variable with its type, the number of valid and missing values, and the categories of each categorical variable (Figure 2). SaNaStat guesses the type from the file and counts the result above the list: *25 scale · 7 categorical · 1 ID*. There are four types:

- **Scale:** numbers you can average (scores, hours, ratings).
- **Categorical:** groups or answers without a meaningful average (gender, programme, yes/no).
- **Ordinal:** groups in a meaningful order (low / medium / high, or a 1 to 5 answer treated as categories). An ordinal variable is offered wherever groups are, and as the outcome of an ordinal regression (Section 11.7).
- **ID:** a respondent number or code. It is shown and kept, but never offered to an analysis.

The type decides which analyses offer the variable: scale variables appear as outcomes and correlation variables, categorical and ordinal ones as groups, factors and chi-square variables. Change any type that is wrong in its list. **Find a variable** narrows a long list.

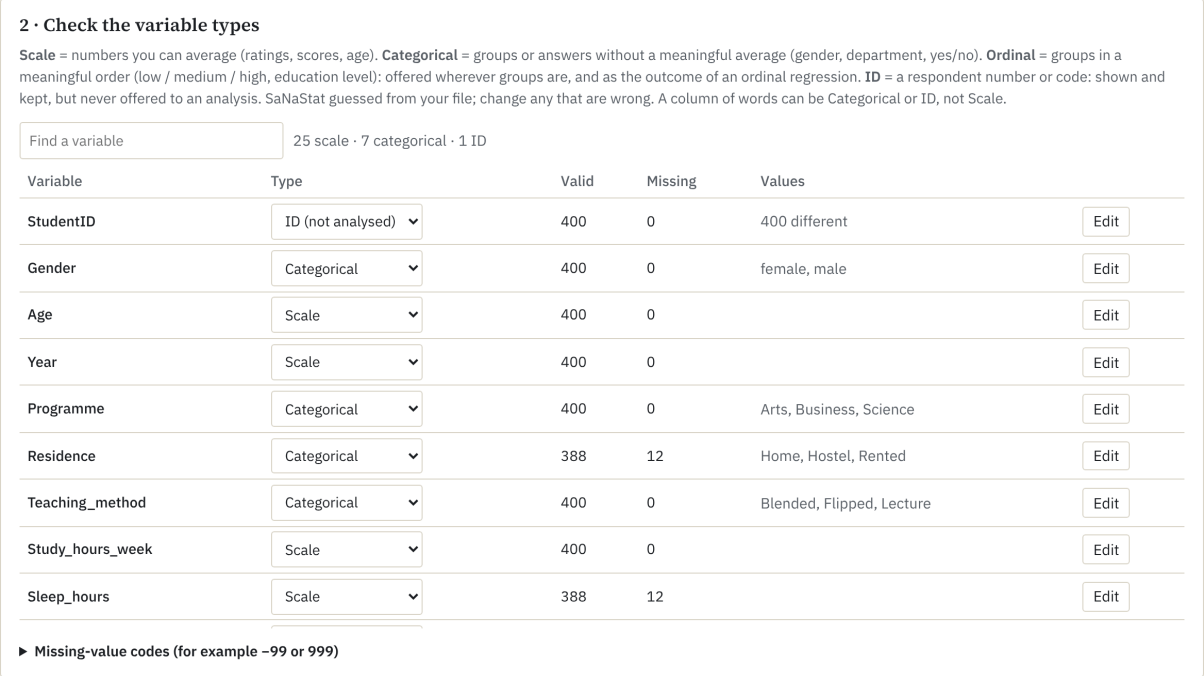


Figure 2. *Step 2: the variable types of practice_students.csv.*

What it means for our example

Gender (female, male), Programme (Arts, Business, Science), Residence (Home, Hostel, Rented), Teaching_method (Blended, Flipped, Lecture), Passed, Pass_before and Pass_after (No, Yes) are categorical. StudentID is an ID ("400 different" values): SaNaStat will never offer it as a variable. Everything else is scale, including Age, Year (1 to 4), the three repeated scores, the satisfaction rating and the items Mot1 to Anx6.

Some variables can be treated either way. Year has only four values, 1 to 4. Treated as scale, it can be correlated or used as a covariate; treated as categorical, it can define groups (Figure 3). Choose according to the question, not the number of values.

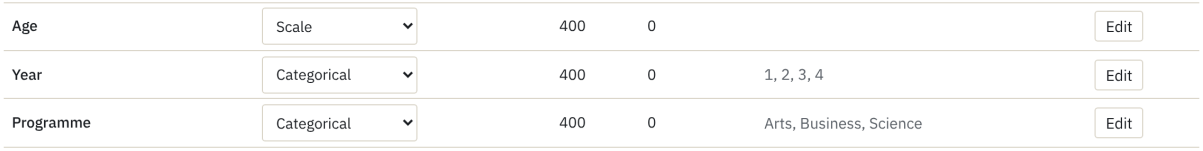


Figure 3. *Year changed from scale to categorical: SaNaStat now lists its categories 1, 2, 3, 4.*

In the practice file, change Year back to scale before you continue.

1.4 The variable editor

Edit at the end of each row opens the variable editor (Figure 4). In it you can:

- rename the variable (**Rename**); construct scores, formulas and filters that use it follow the new name;
- give it a label in **Label (what the question asked)** and click **Save label**; the label is shown next to the name in the lists;

- change its **Type**; for a categorical or ordinal variable, rename its categories (**Save category names**), put them in order with the arrows (for an ordinal variable, lowest first) and merge ticked categories (**Merge**);
- delete it with **Delete this variable…** (refused while an analysis, construct or formula uses it; undo is offered).

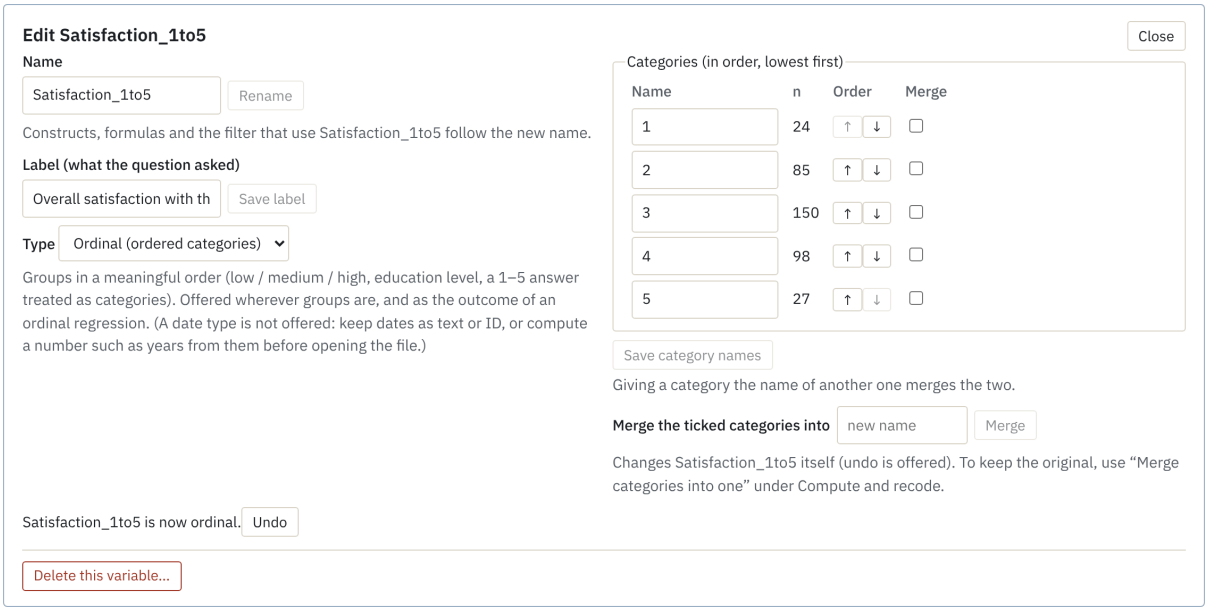


Figure 4. *The variable editor for Satisfaction_1to5, with a label and the type Ordinal.*

Try it: make the satisfaction rating ordinal

Click **Edit** next to Satisfaction_1to5. Type *Overall satisfaction with the course* as its label and click **Save label**; choose **Ordinal (ordered categories)** as its **Type**. SaNaStat reports *Satisfaction_1to5 is now ordinal.* and lists the five categories 1 to 5 with their counts (24, 85, 150, 98 and 27). Click **Close**. The row now reads "Satisfaction_1to5 · Overall satisfaction with the course", with 384 valid and 16 missing values. Section 11.7 uses this ordinal variable as an outcome; for the other chapters, set it back to **Scale** in the list.

1.5 Missing data and missing-value codes

Blank cells are read as missing values, and the *Missing* column counts them. In practice_students.csv Residence (12), Sleep_hours (12), Anxiety_score (22), Exam_score (13) and Satisfaction_1to5 (16) have missing values, between 3% and 6% of the 400 students.

Some files use a number such as −99, 99 or 999 for "no answer". Such a code must not be analysed as a real answer: a mean that includes −99 is meaningless. Open **Missing-value codes (for example −99 or 999)** under the variable list, type the codes in **Codes** and SaNaStat lists the variables in which they occur and how often; **Set as missing** turns them into missing values (undo is offered). A .sav file brings its own missing-value codes.

practice_students.csv uses blank cells only: typing −99, 99, 999 gives *These codes do not occur in the data.* (Figure 5).

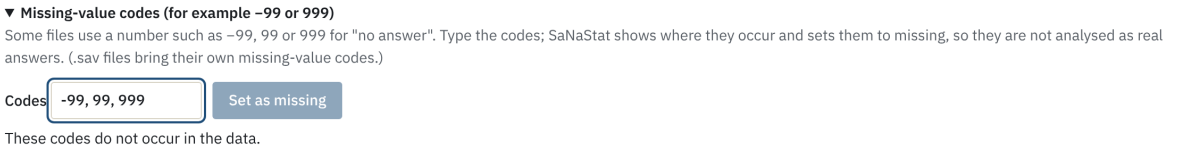


Figure 5. *Missing-value codes: the practice file contains none of −99, 99 and 999.*

Concept: how SaNaStat treats missing values

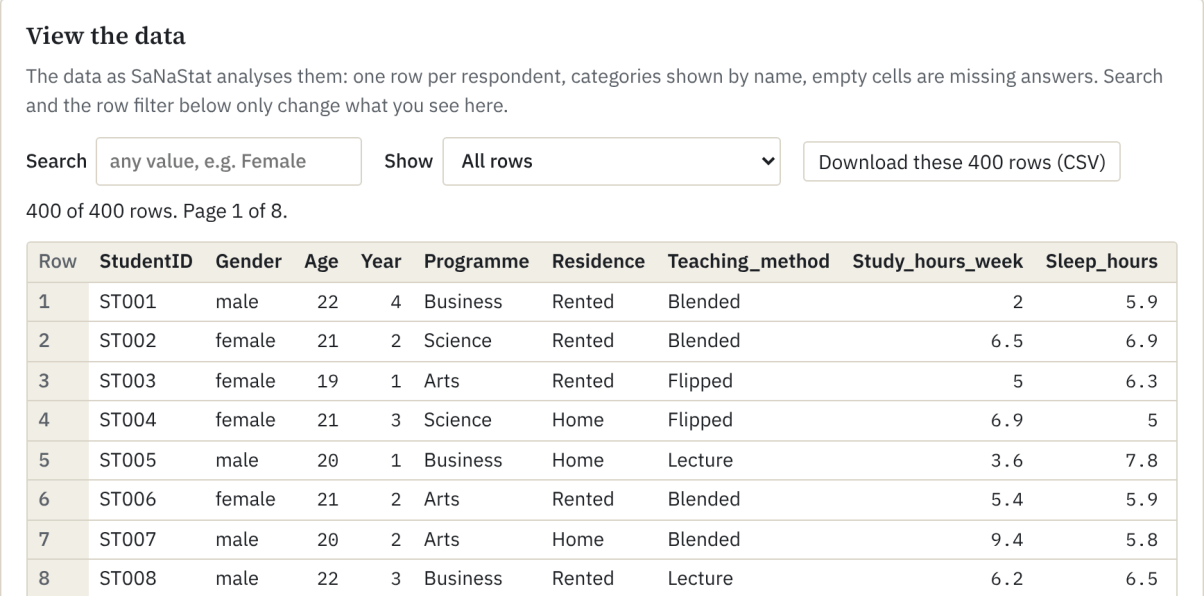
SaNaStat does not fill in missing values for the classical tests. Each analysis uses the cases that have the values it needs, and tells you how many that was:

- A t-test, ANOVA or descriptive table uses every case with a value on the variables involved (for example n = 387 for Exam_score).
- A correlation matrix uses, for each pair, every case with both values; a partial correlation uses complete cases only.
- A regression, repeated-measures ANOVA, factorial ANOVA or MANOVA uses only complete cases, and says so ("46 respondents with missing values left out").
- The machine learning screen fills in missing values inside each training set (chapter 16), and multiple imputation fills them in for the classical analyses (chapter 14).

Under about 10% missing per variable is usually harmless if the values are missing at random (Hair et al., 2019). The data quality screen of chapter 3 adds Little's test of whether the missing values are missing completely at random (Little, 1988b), and chapter 14 shows how to analyse incomplete data properly. A missing value is different from a zero: never type 0 for "no answer".

1.6 View the data

View the data (rows and values) → under *1 · Open your data* shows the data as SaNaStat analyses them: one row per respondent, categories shown by name, empty cells for missing answers, 50 rows per page (Figure 6). **Search** finds any value, and **Show** offers **Only rows with missing answers** (71 of the 400 students have at least one) or **Only rows where …** a condition holds. A button downloads the rows shown as a CSV file. Searching and filtering here only change what you see; they do not change the data.



View the data

The data as SaNaStat analyses them: one row per respondent, categories shown by name, empty cells are missing answers. Search and the row filter below only change what you see here.

Search any value, e.g. Female Show All rows Download these 400 rows (CSV)

400 of 400 rows. Page 1 of 8.

Row	StudentID	Gender	Age	Year	Programme	Residence	Teaching_method	Study_hours_week	Sleep_hours
1	ST001	male	22	4	Business	Rented	Blended	2	5.9
2	ST002	female	21	2	Science	Rented	Blended	6.5	6.9
3	ST003	female	19	1	Arts	Rented	Flipped	5	6.3
4	ST004	female	21	3	Science	Home	Flipped	6.9	5
5	ST005	male	20	1	Business	Home	Lecture	3.6	7.8
6	ST006	female	21	2	Arts	Rented	Blended	5.4	5.9
7	ST007	male	20	2	Arts	Home	Blended	9.4	5.8
8	ST008	male	22	3	Business	Rented	Lecture	6.2	6.5

Figure 6. *View the data: the first rows of practice_students.csv.*

1.7 Scale scores and reliability

Many questionnaires measure one idea with several items. practice_students.csv has two such sets: Mot1 to Mot6 (motivation) and Anx1 to Anx6 (anxiety), each answered from 1 to 5. A **construct score** (scale score) is the mean of the items, which is more reliable than any single item. SaNaStat builds it on the **Data** screen in the card *3 · Build construct scores (optional)*, and first checks that the items hang together.

Concept: Cronbach's alpha, McDonald's omega and reverse-coded items

Cronbach's α estimates how consistently the items measure the same thing (Cronbach, 1951). It rises when the items correlate and when there are more of them. **McDonald's ω** comes from a one-factor model of the items; unlike α it does not assume that every item measures the construct equally well, so it is the better estimate when the items differ (McDonald, 1999; McNeish, 2018). The **corrected item–total correlation** is the correlation of one item with the mean of the others. If one item is worded in the opposite direction ("I feel calm"), its correlation is negative until you **reverse-code** it: on a 1 to 5 scale a 2 becomes 4 (the lowest plus the highest answer, minus the value).

Build Motivation

1. In **Name of the construct**, type **Motivation**.
2. In **Quick pick: items starting with**, type **Mot** and press Tab. The items Mot1 to Mot6 are ticked and **Lowest answer** and **Highest answer** are filled in from the data (1 and 5).
3. Read Cronbach's α, McDonald's ω and the item statistics.

4. Click **Create the construct score**. Motivation appears in the variable list with a "construct" label.

> **What it means for our example**
>
> Motivation has α = .86 on 400 complete responses and ω = .86. The corrected item–total correlations are between .55 (Mot5) and .72 (Mot1), and deleting an item would never raise α above .86 (deleting Mot5 would leave it at .86): the six items form a consistent scale.

Build Anxiety, with two reversed items

In this file Anx3 and Anx5 are worded in the opposite direction and must be reversed.

1. In **Name of the construct**, type **Anxiety** (the name Anxiety_score is already used by another variable).
2. In **Quick pick: items starting with**, type **Anx** and press Tab. The quick pick matches every variable that starts with "Anx", so it also ticks **Anxiety_score**. Untick Anxiety_score.
3. Before reversing anything, look at the result. SaNaStat warns: *Anx3 loads negatively on Anxiety: reverse-code it, otherwise ω is too low*, and the same for Anx5.
4. Tick the **R** box next to Anx3 and Anx5 (Figure 7).
5. Click **Create the construct score**.

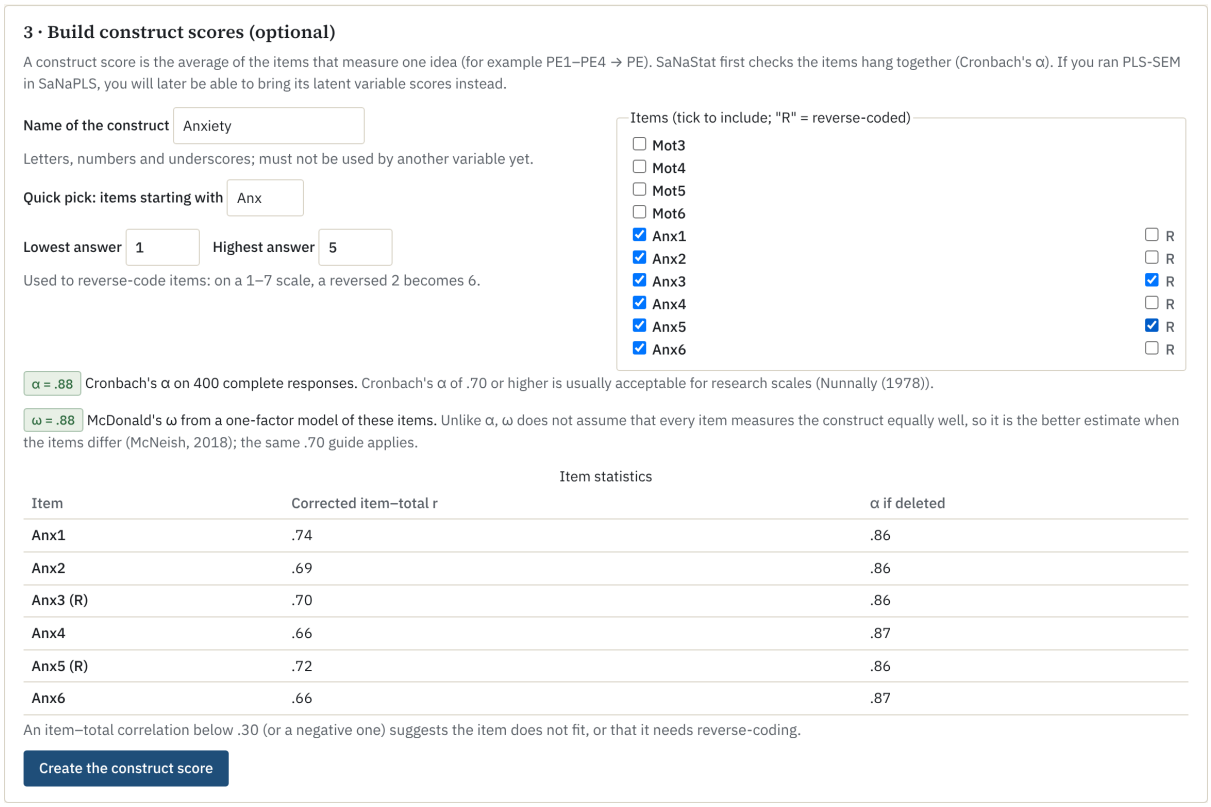


Figure 7. *Anxiety with Anx3 and Anx5 reverse-coded.*

What it means for our example

Without reversal, Anx3 and Anx5 have corrected item–total correlations of −.58 and −.61, α = −.25 and ω = .43: a negative α is a clear sign that items point in different directions. With Anx3 and Anx5 reversed, α = .88, ω = .88 and every corrected item–total correlation lies between .66 and .74. Always read the item statistics before you create a score.

Cut-off values: internal consistency

Cronbach's α or McDonald's ω of .70 or higher is usually acceptable for research scales (Nunnally, 1978). The pill is green at .70 or more, amber from .60 and red below.

Corrected item–total correlation below .30, or negative, suggests that the item does not fit or needs reverse-coding.

How to report it

"Motivation (six items, 1 to 5) was reliable, Cronbach's α = .86, McDonald's ω = .86, and so was Anxiety (six items, two reverse-coded), α = .88, ω = .88."

Chapter 12 checks these two scales more thoroughly, with a factor analysis and the validity measures.

Exercise: scales in practice_brand_survey.csv and practice_engagement.csv

Open practice_engagement.csv and build a scale for the three Vigor items (Vigor1 to Vigor3; they are rated 1 to 7, so check the lowest and highest answer). Do the same for Dedication and Absorption. Which scale has the highest α? Then open practice_brand_survey.csv and build Trust (Trust1 to Trust4), Value (Value1 to Value4) and Loyalty (Loyalty1 to Loyalty3). None of these items needs reversing. Answers: Vigor α = .83, Dedication α = .84 and Absorption α = .79 (so Dedication is the most consistent of the three); Trust α = .87, Value α = .81 and Loyalty α = .84.

2. Preparing variables

Raw data rarely have exactly the variables an analysis needs. Card *4 · Compute and recode variables (optional)* on the Data screen makes new variables from existing ones, and card *7 · Select cases (optional)* restricts the analyses to some respondents. The original variables are never changed (except a recode you choose to make into the same variable, which you can undo): each new variable is added to the list and every analysis offers it like any other. Choose what you want to make in **What do you want to make?** (Table 4).

Table 4
What card 4 can make

Choice	What it makes
Your own formula	Any expression of the variables: arithmetic, logarithms, means, comparisons, if … then (Section 2.1)
If … then … else	Groups or scores by rules, without typing a formula (Section 2.2)
Recode with a table of old → new values	New values for old values, lists or ranges (Section 2.3)
Ranks, percentiles, or z-scores within groups	Ranks or percentiles, also within groups; z-scores and centred scores within groups
Transform a scale variable	ln(x), ln(1 + x), square root, square, z-score, centred score, reverse code (Section 2.6)
Make groups from a scale variable	A median split, tertiles, quartiles or your own cut points (Section 2.7)
Make a 0/1 indicator of a category	1 for the ticked categories, 0 for the others (Section 2.8)
Combine two scale variables	A − B, A + B, A × B, A ÷ B or the percentage change (Section 2.9)
Merge categories into one	Two or more categories become one (Section 2.8)
Reverse-code several items at once	A reversed copy of each ticked item (Section 2.4)
Statistic across several variables	The mean, sum, SD, lowest or highest answer, a count, or the number answered or missing (Section 2.5)

2.1 Your own formula

A student's course mark is the mean of the exam and the project. Choose **Your own formula (+ − * / ^, ln, mean, if …)**.

1. Type **(Exam_score + Project_score) / 2** in **Formula**. **Insert a variable** adds a variable name where the cursor is, so you need not type it.

2. SaNaStat checks the formula at once and shows the first eight students with the values used and the result (Figure 8). A missing value (shown as ·) gives a missing result: student 6 has no exam score.
3. Type **Course_mark** in **Name of the new variable** and click **Add the new variable**. SaNaStat reports *Added Course_mark (scale): 387 of 400 respondents have a value.*

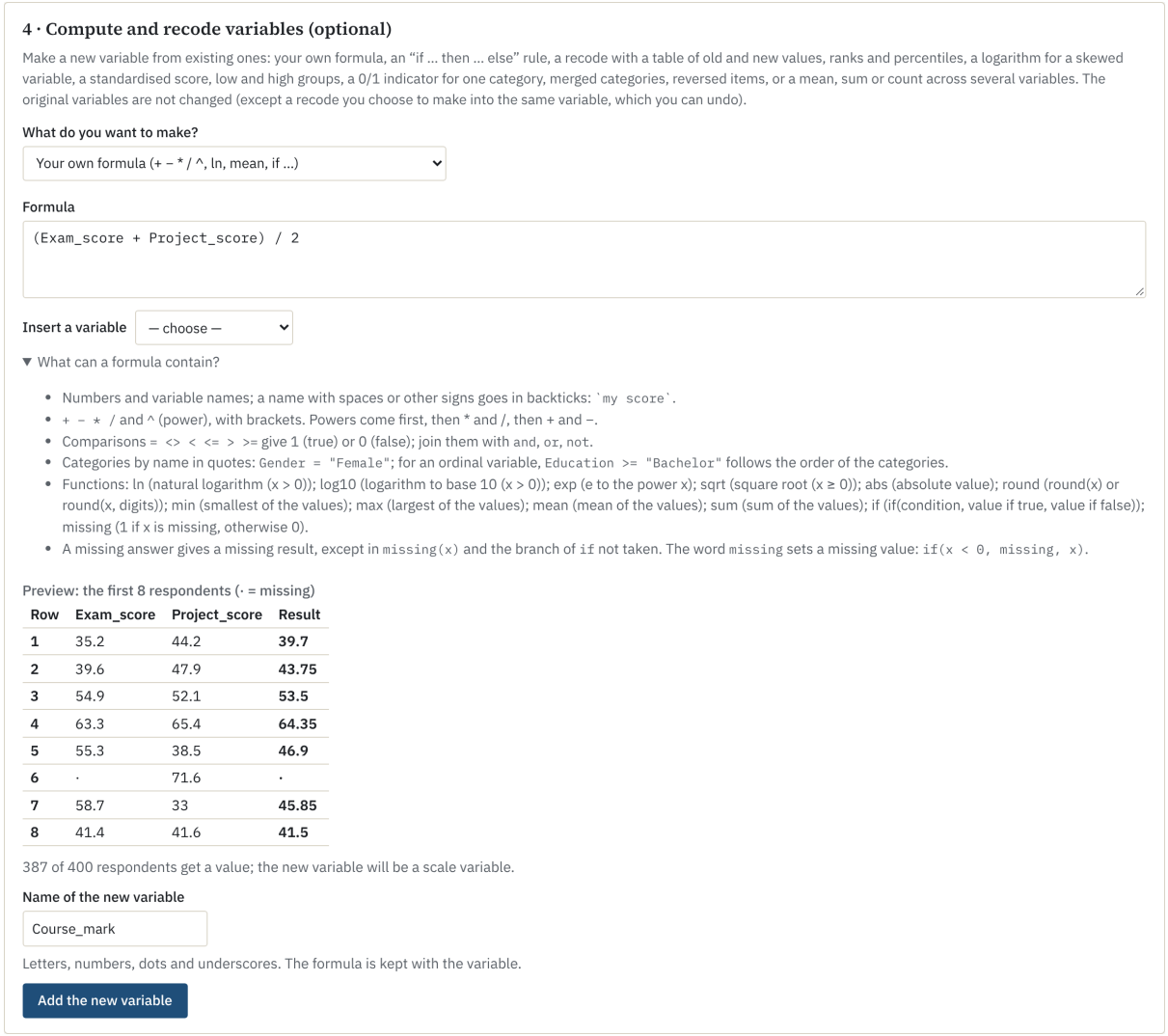


Figure 8. *A formula: the course mark as the mean of the exam and the project, with the preview of the first eight students.*

What can a formula contain? lists everything a formula understands: numbers and variable names (a name with spaces goes in backticks); + − * / and ^ with brackets; comparisons such as = <> < <= > >=, which give 1 or 0, joined with and, or, not; categories by name in quotes (Gender = "female"; for an ordinal variable a comparison such as >= follows the order of the categories); and the functions ln, log10, exp, sqrt, abs, round, min, max, mean, sum, if and missing. The word missing sets a missing value: if(x < 0, missing, x). The formula is kept with the variable and shown in the variable list.

2.2 If … then … else

Choose **If … then … else (groups or scores by rules)** to build a variable from rules without typing a formula. Each line is a condition and the value it gives; the lines are checked in turn and the first that holds gives the value. To grade the exam:

1. In the *If* line choose **Exam_score**, the comparison **is at least**, the value **70** and, after **then**, **Distinction**.
2. Click **Add an "else if" line** and make it **Exam_score is at least 50** then **Pass**.
3. Type **Fail** in **Otherwise (no line holds)** and **Exam_grade** as the name (Figure 9).

SaNaStat shows the formula it wrote, if(Exam_score >= 70, "Distinction", if(Exam_score >= 50, "Pass", "Fail")), and the preview. Because the results are words, the new variable is categorical: *Added Exam_grade (categorical: Distinction, Pass, Fail): 387 of 400 respondents have a value.* The 13 students without an exam score stay missing; they do not become "Fail". A frequency table (Section 10.1) shows 24 Distinction (6.2% of 387), 244 Pass (63.0%) and 119 Fail (30.7%).

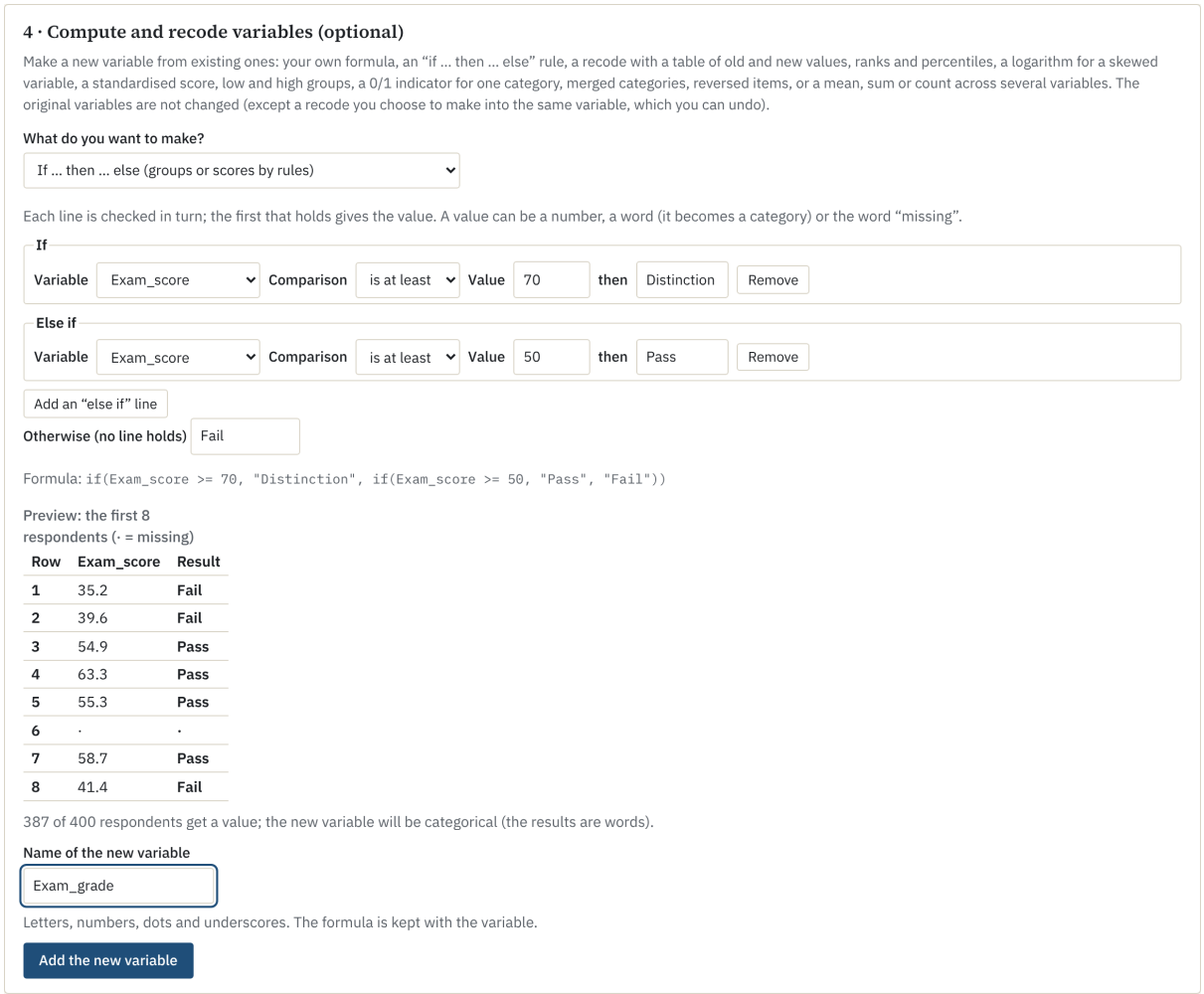


Figure 9. *If … then … else: a grade from the exam score.*

2.3 Recode with a table

Choose **Recode with a table of old → new values** and, in **Variable to recode**, **Year**. SaNaStat lists the four values 1 to 4, each with an empty new value. Type **Lower** for 1 and 2 and **Upper** for 3 and 4, and **Year_group** as the name (Figure 10). An old value can also be a list (1, 2) or a range (1 to 3, lowest to 2, 6 to highest); the word missing makes a value missing. **All other values** decides what happens to values not in the table (**Keep them as they are**, set them missing or give them one value), and **Where to put the result** is **Into a new variable (recommended)** or the variable itself (with undo). SaNaStat reports *Added Year_group: categorical with the categories Lower, Upper; 400 of 400 respondents have a value* (214 Lower and 186 Upper).

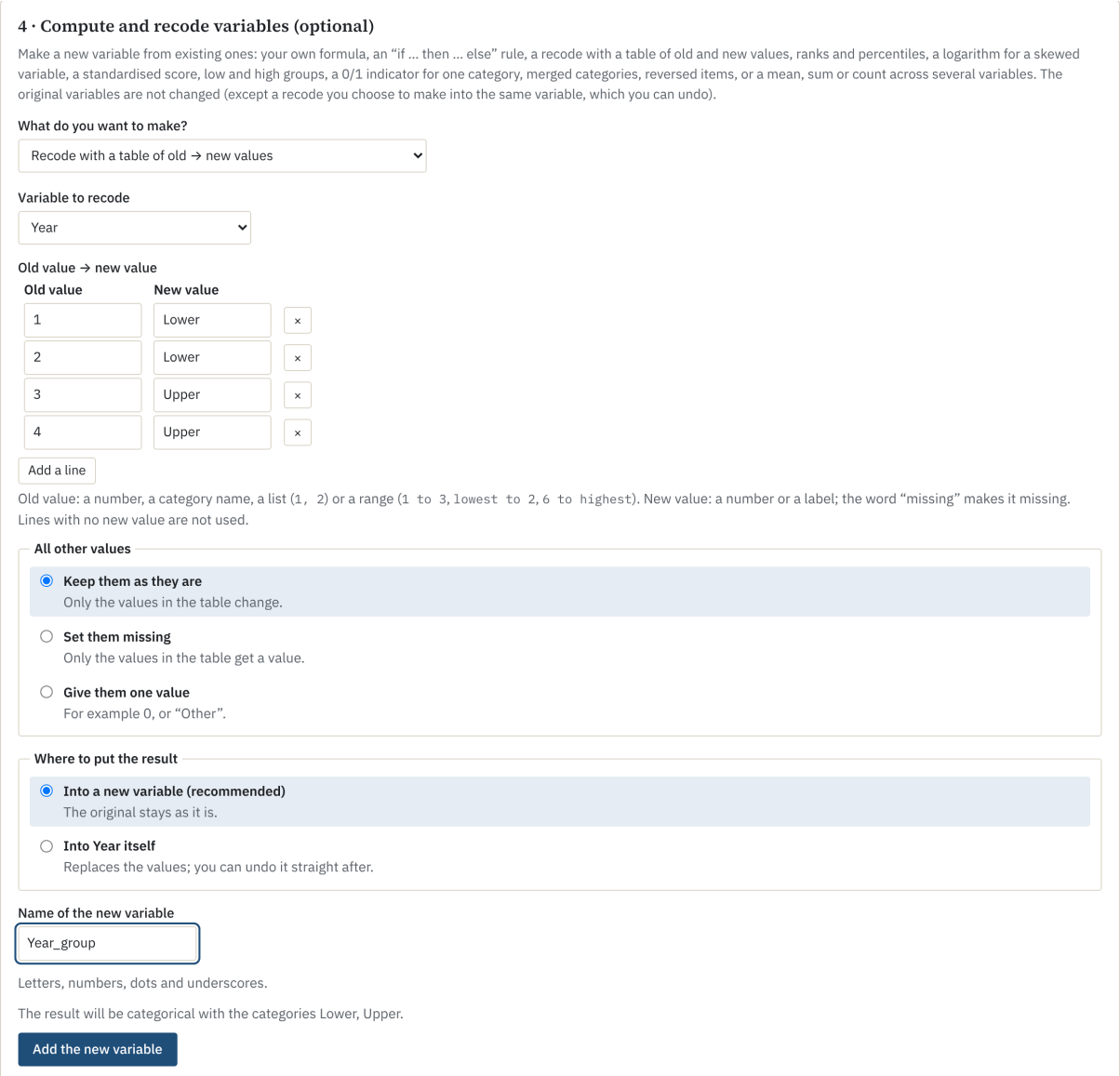


Figure 10. *A recode table: Year 1 and 2 become Lower, 3 and 4 Upper.*

2.4 Reverse-code several items at once

Section 1.7 reversed Anx3 and Anx5 inside the construct builder. When you need the reversed items themselves, for example for a factor analysis or another program, choose

Reverse-code several items at once, tick **Anx3** and **Anx5** under **Items to reverse-code (tick one or more)** and check the ending in **Add to the names** (*_r*); SaNaStat lists the new names (Figure 11). Leave **Lowest answer** and **Highest answer** empty to use the lowest and highest observed answers, or type them when an extreme answer was never given. **Add the reversed items** reports *Added 2 reverse-coded variables: Anx3_r, Anx5_r (1 + 5 − x). Use them in place of the originals when you build constructs.*

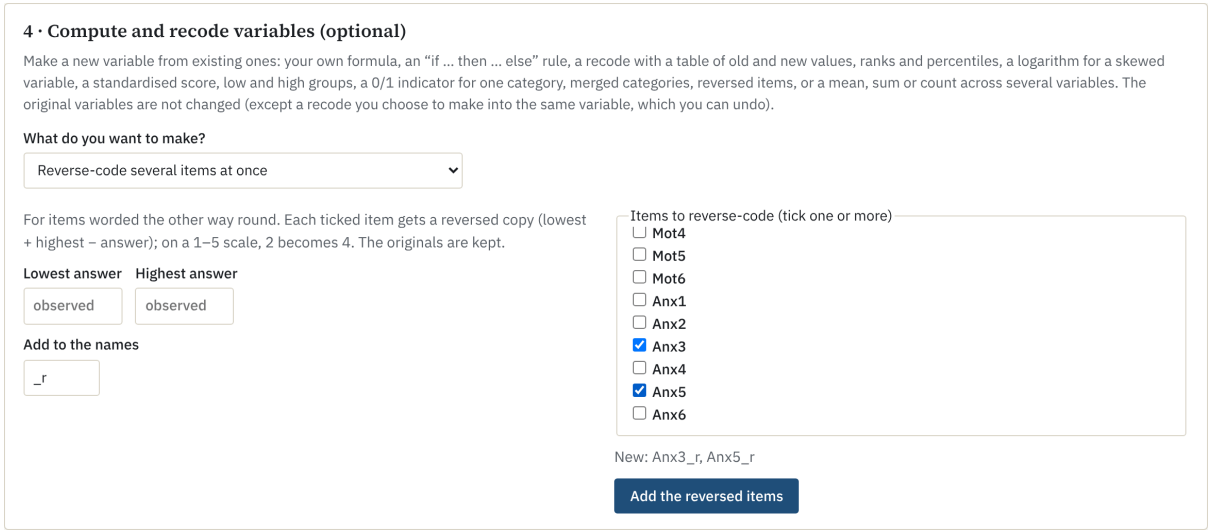


Figure 11. *Reversing Anx3 and Anx5 in one step.*

2.5 A mean across items, with a minimum number answered

Choose **Statistic across several variables (mean, sum, count …)** and **Mean of the answers** as the **Statistic**. Tick Anx1, Anx2, Anx3_r, Anx4, Anx5_r and Anx6, type **5** in **Respondents need at least this many answers** and **Anxiety_items** as the name (Figure 12). A respondent who answered at least five of the six items gets the mean of the answers given; with fewer answers the result is missing. This is the usual way to score a questionnaire with a few skipped items: one missing item does not throw the whole respondent away, but a respondent who skipped most of the scale gets no score. In the practice file every student answered every item: *Added Anxiety_items: 400 of 400 respondents have a value* (M = 2.97, SD = 0.79), the same score as the construct Anxiety of Section 1.7.

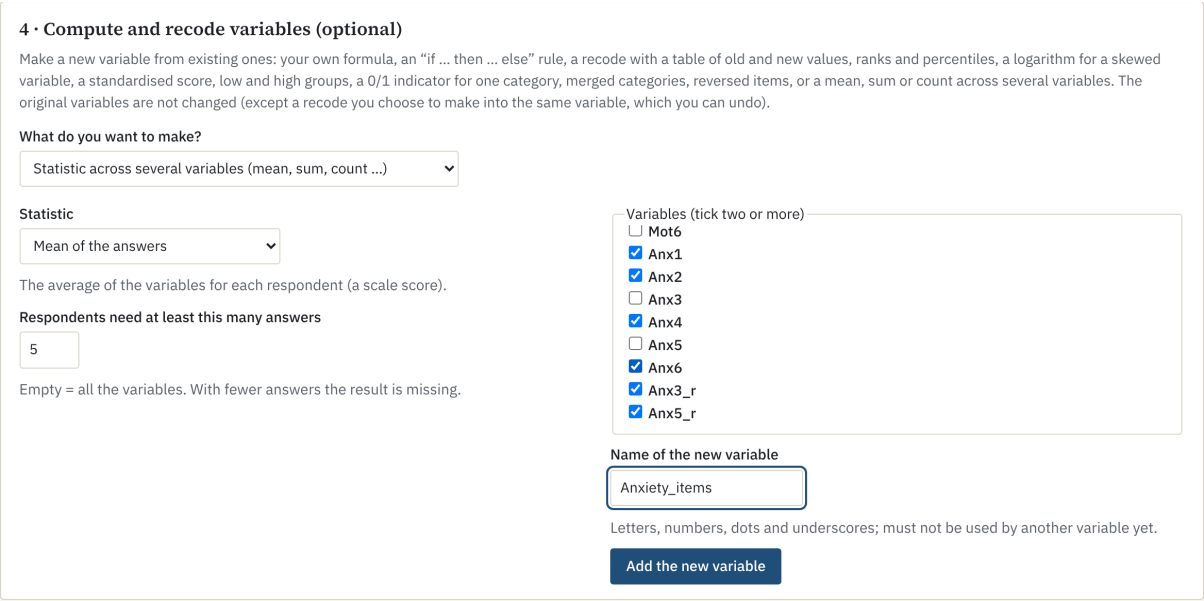


Figure 12. *The mean of six items for respondents with at least five answers.*

> **Tip: mean, not sum**
>
> With missing answers allowed, a sum is the sum of the answers given, so a respondent who skipped an item gets a lower total. Prefer the mean, which stays on the scale of the items (here 1 to 5).

2.6 Transform a scale variable: a logarithm of Anxiety_score

Choose **Transform a scale variable (logarithm, standardise, reverse …)**. The transformations are the natural logarithm ln(x), ln(1 + x), the square root, the square, a standardised score (z-score), the centred score (mean subtracted) and a reverse-coded score (lowest + highest − x). The logarithm needs x > 0 and the square root x ≥ 0; values that cannot be transformed become missing, and SaNaStat says how many respondents have a value.

1. In **Variable** choose **Anxiety_score** and in **Transformation** choose **Natural logarithm, ln(x)** (Figure 13).
2. Leave the suggested name and click **Add the new variable**. SaNaStat reports: "Added Anxiety_score_ln: 378 of 400 respondents have a value." The 22 students with a missing anxiety score keep a missing value.

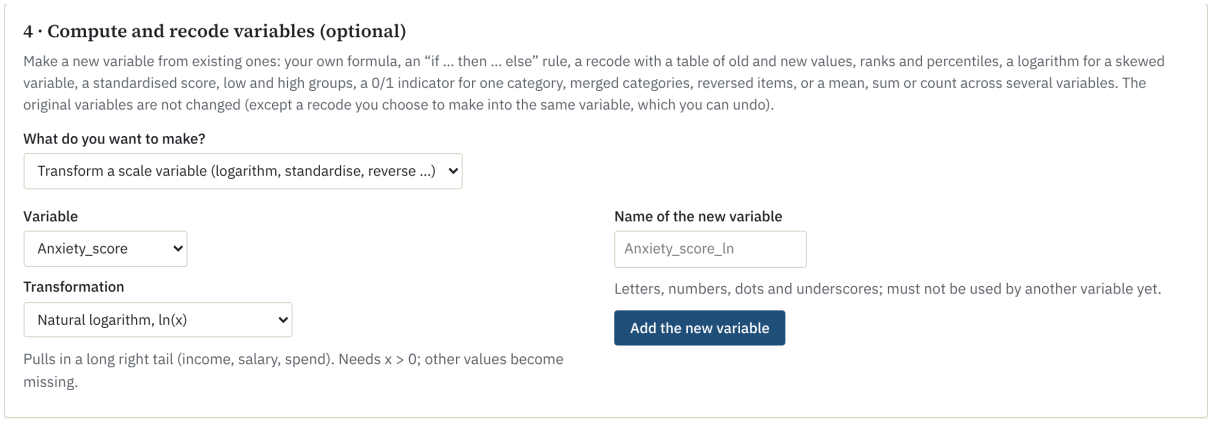


Figure 13. *Card 4 with Anxiety_score and the natural logarithm chosen.*

Now describe both variables (**Analyses**, **Descriptive statistics**). The skewness of Anxiety_score is 0.97 (kurtosis 1.74); that of Anxiety_score_ln is 0.26 (kurtosis 0.20), with M = 2.58, SD = 0.21 and n = 378 (Figure 14). The logarithm pulled in the long right tail. Results for the transformed variable are on the logarithmic scale: report them as such, and take the logarithm only when the distribution or the research question calls for it.

What it means: No variable is strongly skewed (all |skewness| ≤ 1).

Descriptive statistics

Variable	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Anxiety_score	378	22	13.50	2.90	13.10	7.10	26.20	0.97 (0.13)	1.74 (0.25)
Anxiety_score_ln	378	22	2.58	0.21	2.57	1.96	3.27	0.26 (0.13)	0.20 (0.25)
Passed_Yes	400	0	0.69	0.46	1.00	0.00	1.00	−0.85 (0.12)	−1.28 (0.24)

Skewness and kurtosis are bias-corrected (G1, G2; kurtosis is excess kurtosis); standard errors in brackets. Median = weighted average at (n + 1)p.

Figure 14. *Descriptive statistics of Anxiety_score, its logarithm and the indicator Passed_Yes.*

2.7 Groups from a scale variable

Choose **Make groups from a scale variable (median split, tertiles ...)**, **Study_hours_week** and **Two groups at the median**. The other choices are three (tertiles) or four (quartiles) equal-sized groups, or your own cut points; a value equal to a cut point goes to the lower group. SaNaStat shows the cut point (6.3 hours) before you add the variable; the new variable Study_hours_week_group has the groups "Low (≤ median)" and "High (> median)". Compare the exam scores of the two groups with an independent-samples t-test (chapter 6): Low M = 51.36 (SD = 8.69, n = 195), High M = 58.63 (SD = 9.55, n = 192), Welch's t(380.42) = −7.82, p < .001, d = −0.80, 95% CI [−1.00, −0.59]; the negative sign only means that the first group (Low) scored lower. The same groups are related to passing: 56% of the 202 low-hours students passed against 83% of the 198 high-hours students, $\chi^2(1, N = 400) = 35.40$, p < .001, Cramér's V = .30.

Note: think before you split a scale variable

Splitting a scale variable throws information away and the result depends on where the cut point falls. Use groups when the question is about groups (heavy against light users, tertiles of a score) and keep the variable as it is when you can: the correlation of Study_hours_week with Exam_score in chapter 11 uses every student's exact hours.

2.8 An indicator and merged categories

Choose **Make a 0/1 indicator of a category**, **Passed** and tick **Yes**: Passed_Yes is 1 for the students who passed and 0 for the others (400 of 400 have a value). Its mean is the proportion who passed: M = 0.69 (278 of 400 students). Choose **Merge categories into one**, **Residence**, tick **Hostel** and **Rented**, and type *Away from home* in **Name of the merged category**. Residence_merged has the categories Home and Away from home; the 12 students with a missing residence stay missing (388 of 400). Exam_score does not differ between them: Home M = 54.83 (SD = 9.58, n = 189) and Away from home M = 55.24 (SD = 10.04, n = 186), Welch's t(371.49) = −0.41, p = .682, d = −0.04, 95% CI [−0.24, 0.16].

2.9 A gain score from two variables

Choose **Combine two scale variables (difference, change, ratio ...)**, **Post_score** as **Variable A**, **Difference, A − B** and **Pre_score** as **Variable B** (Figure 15). Call the new variable Gain and add it: "Added Gain: 400 of 400 respondents have a value." The one-way ANOVA of this gain by teaching method is in Section 7.6.

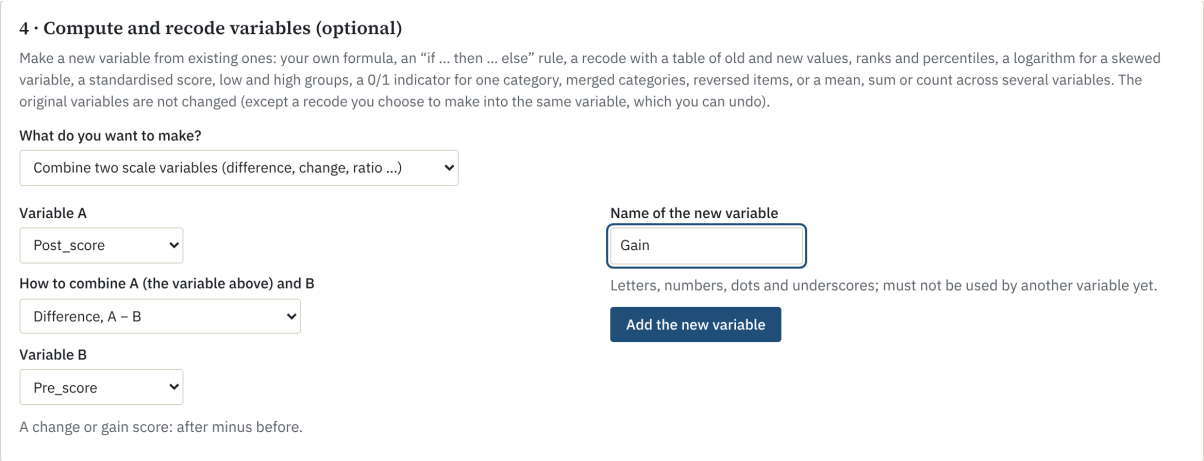


Figure 15. *Combining Post_score and Pre_score into the gain.*

2.10 Select cases

Sometimes an analysis should use only some respondents: only women, only final-year students, only those who finished the course. Card *7 · Select cases (optional)* keeps the other respondents with the data, so you can remove the filter at any time.

1. Under **Analyse only respondents where**, choose **Gender** in **Variable**, **is** in **Comparison** and **female** in **Value** (Figure 16). SaNaStat counts the matching respondents at once: *Gender = female: 246 of 400 meet this.*
2. **Add a second condition** joins a second condition with **Both (and)** or **Either (or)**.
3. Click **Use this filter**.

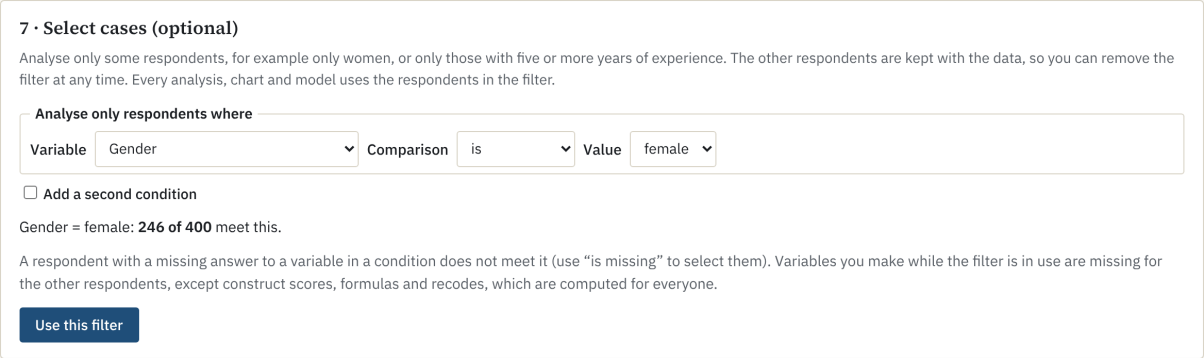


Figure 16. *Select cases: only the women.*

While the filter is in use, every analysis, chart and model uses only the selected respondents, and a banner at the top of the Analyses, Charts, Clustering and Machine learning screens says so (Figure 17): *Filter in use: Gender = female (246 of 400 respondents)*, with **Remove filter** and **Change it**. With the filter, the descriptive statistics of Exam_score are those of the women, M = 56.57, SD = 9.64, n = 239 (7 of the 246 women have no exam score); after **Remove filter** they are those of everyone again, M = 54.97, SD = 9.81, n = 387.

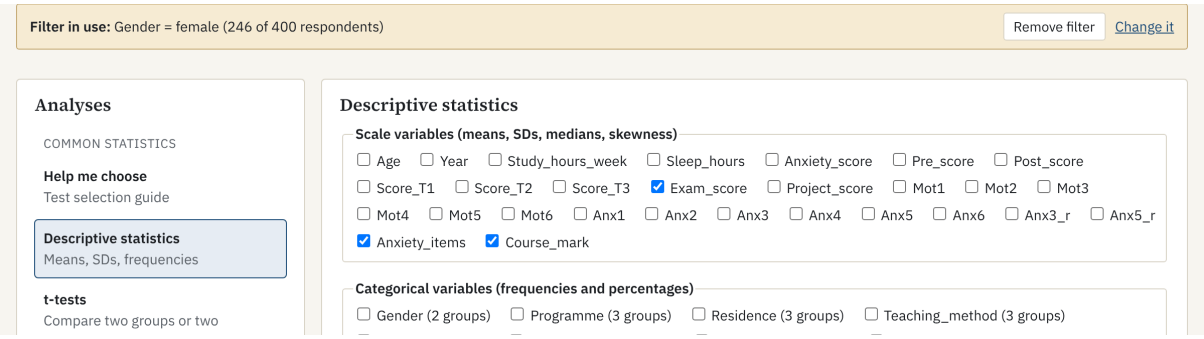


Figure 17. *The banner shows that only the women are analysed.*

Note: filters and new variables

A respondent with a missing answer to a variable in a condition does not meet it (choose **is missing** to select them). Variables you make while a filter is in use are missing for the other respondents, except construct scores, formulas and recodes, which are computed for everyone. Remove the filter before you continue with this manual.

How to report it

Say in the method section which variables you computed and how: "The course mark was the mean of the exam and project scores. Exam grades were Distinction (70 or more), Pass (50 to 69.9) and Fail (below 50). Anxiety was the mean of six items (two reverse-coded) for students who answered at least five of them. Analyses of women only used the 246 female students."

Exercise: transformations in practice_marketing.csv

Open practice_marketing.csv. Take the natural logarithm of Annual_spend and of Income_k and describe the four variables. Answer: the skewness of Income_k falls from 1.38 to 0.04; that of Annual_spend from 1.07 to −0.30, but its logarithm exists for only 1,418 customers, because the 82 customers with a spend of 0 become missing (ln(1 + x) would keep them). Then write a formula that is 1 for customers whose Annual_spend is above 5000 and 0 otherwise, and check its mean against a frequency table.

2.11 Split file and weight cases

Card *8 · Split file and weight cases (optional)* on the Data screen (Figure 18) has two tools that, like Select cases, change which rows every analysis uses.

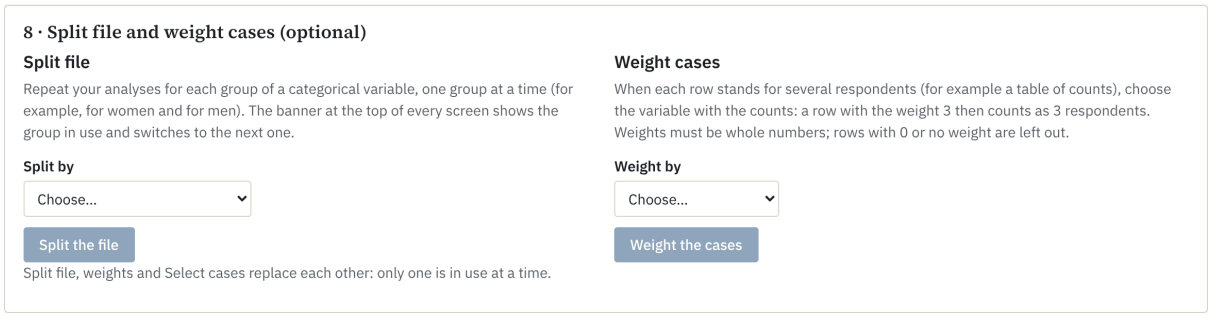


Figure 18. *Split file and weight cases.*

Split file repeats your analyses for each group of a categorical variable, one group at a time.

1. Under **Split by**, choose **Gender (2 groups)** and click **Split the file**.
2. The banner at the top of every screen now says *Split file: Gender = female (246 of 400 respondents)* (Figure 19). Run an analysis, for example the descriptive statistics of Exam_score: for the women M = 56.57, SD = 9.64, n = 239 (Figure 20).
3. Choose **male** in the banner: the same analysis now uses the men, M = 52.38, SD = 9.57, n = 148. Click **Show all groups** when you have finished.

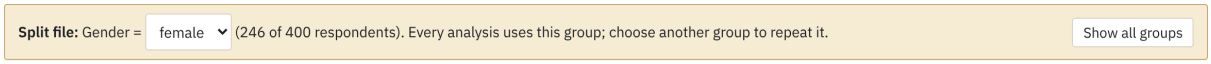


Figure 19. *The banner of a split file: choose the group to analyse.*

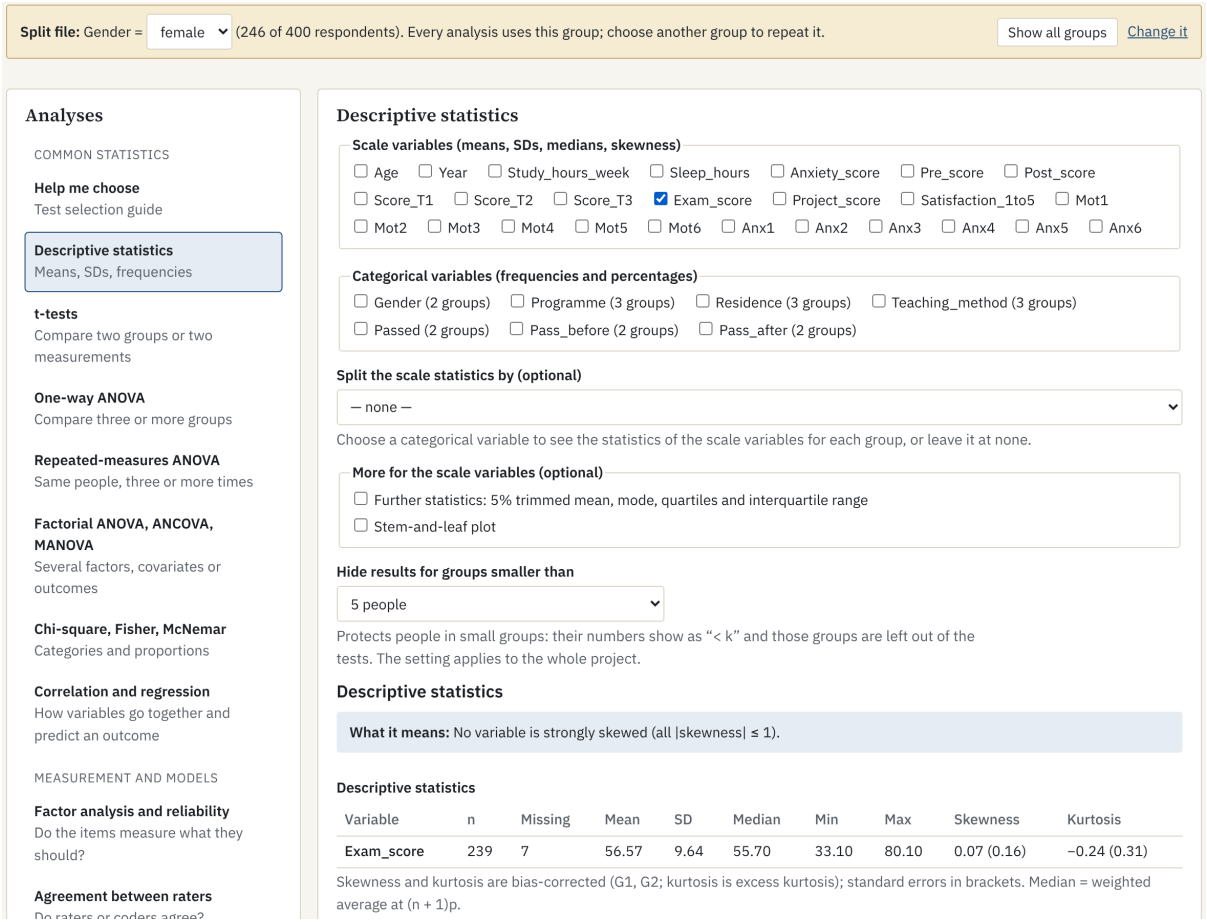


Figure 20. *The descriptive statistics of Exam_score for the women.*

Weight cases is for data in which a row stands for several respondents, for example a table of counts with one row per combination of answers. Choose the variable that holds the counts under **Weight by**: a row with the weight 3 then counts as 3 respondents. Weights must be whole numbers; rows with a weight of 0 or no weight are left out. Before you click **Weight the cases**, the card says how many respondents the weights give (Figure 21). practice_students.csv has no count variable; choosing Year only shows the message (*951 weighted respondents from 400 rows*), so choose **Choose...** again. To try weights on a real table of counts, open practice_counts.csv (36 rows, one per year, gender, programme and result) and weight it by **Count**.

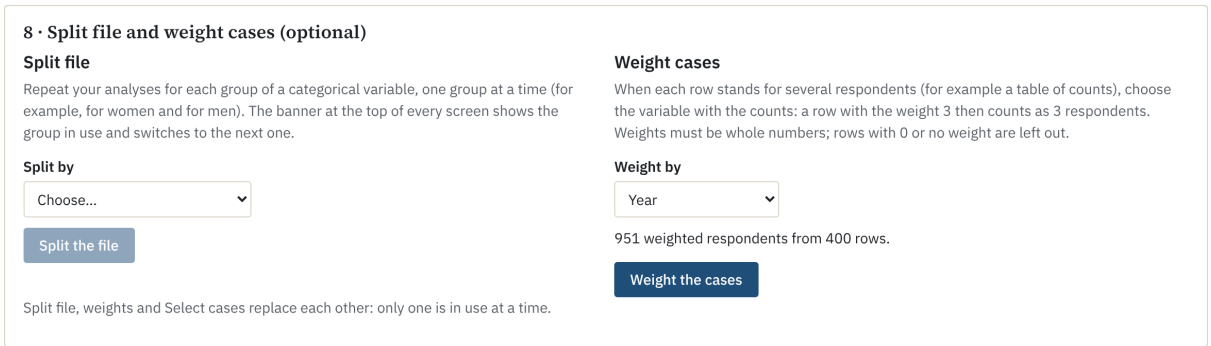


Figure 21. *The card counts the weighted respondents before the weights are used.*

Note: one at a time

Split file, weights and Select cases replace each other: only one is in use at a time, and the banner always says which. Remove them before you continue with this manual.

2.12 Restructure the data

Some analyses need one row per respondent, others one row per measurement. Card *9 · Restructure the data (optional)* changes the shape of the table (Figure 22); the data are replaced, and Undo brings them back.

- **Wide to long: repeated columns into rows**: Score_T1, Score_T2 and Score_T3 become one column, with a column saying which measurement each row holds;
- **Long to wide: rows into columns**: one row per respondent again, each measurement in its own column;
- **Transpose: rows become columns**: every scale variable becomes a row and every respondent a column, for small tables of up to 300 rows.

1. Choose **Wide to long: repeated columns into rows** and tick **Score_T1**, **Score_T2** and **Score_T3**, in that order.
2. Type **Time** as the name of the index column and **Score** as the name of the value column, then click **Restructure**.

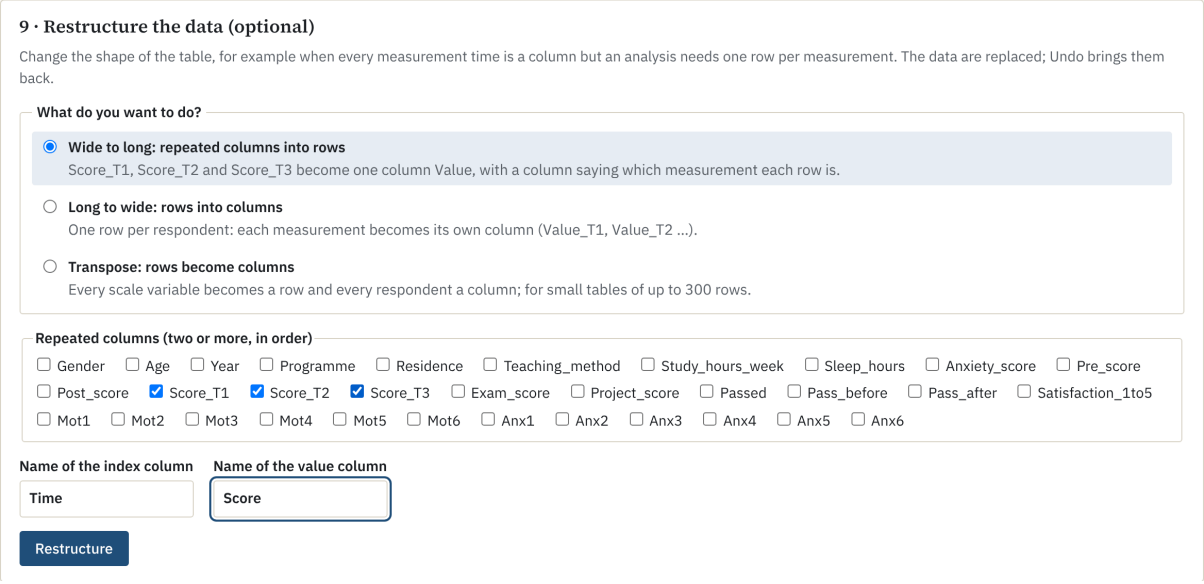
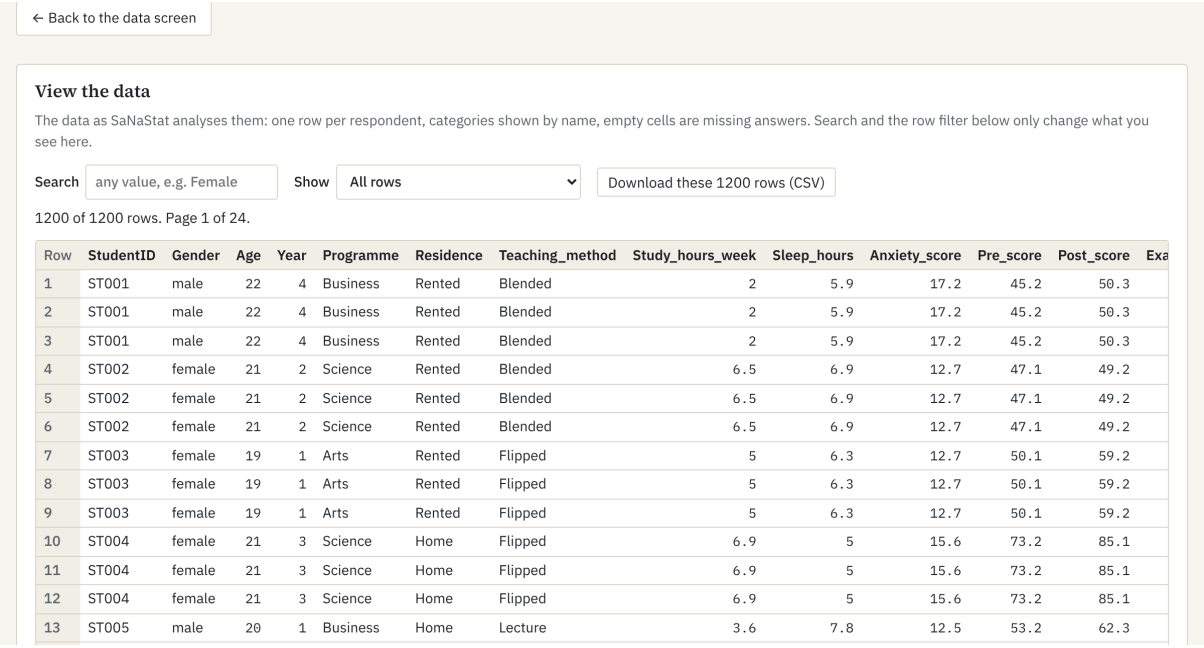


Figure 22. *Wide to long: the three measurements become one column.*

The file now has 1200 rows (400 students × 3 measurements) and 32 variables: StudentID and the other variables are repeated on each of a student's three rows, Time says which measurement the row holds and Score holds the value (Figure 23). Click **Undo** in the message, or open practice_students.csv again, before you continue.



Row	StudentID	Gender	Age	Year	Programme	Residence	Teaching_method	Study_hours_week	Sleep_hours	Anxiety_score	Pre_score	Post_score	Exa
1	ST001	male	22	4	Business	Rented	Blended	2	5.9	17.2	45.2	50.3	
2	ST001	male	22	4	Business	Rented	Blended	2	5.9	17.2	45.2	50.3	
3	ST001	male	22	4	Business	Rented	Blended	2	5.9	17.2	45.2	50.3	
4	ST002	female	21	2	Science	Rented	Blended	6.5	6.9	12.7	47.1	49.2	
5	ST002	female	21	2	Science	Rented	Blended	6.5	6.9	12.7	47.1	49.2	
6	ST002	female	21	2	Science	Rented	Blended	6.5	6.9	12.7	47.1	49.2	
7	ST003	female	19	1	Arts	Rented	Flipped	5	6.3	12.7	50.1	59.2	
8	ST003	female	19	1	Arts	Rented	Flipped	5	6.3	12.7	50.1	59.2	
9	ST003	female	19	1	Arts	Rented	Flipped	5	6.3	12.7	50.1	59.2	
10	ST004	female	21	3	Science	Home	Flipped	6.9	5	15.6	73.2	85.1	
11	ST004	female	21	3	Science	Home	Flipped	6.9	5	15.6	73.2	85.1	
12	ST004	female	21	3	Science	Home	Flipped	6.9	5	15.6	73.2	85.1	
13	ST005	male	20	1	Business	Home	Lecture	3.6	7.8	12.5	53.2	62.3	

Figure 23. *The long file: three rows per student.*

3. Checking data quality

Before you test anything, look for respondents who answered carelessly, for extreme values and for patterns of missing data. SaNaStat's data quality screen does this and suggests whom to exclude; you decide, and nothing is deleted from your file.

> **Concept: careless responding**
>
> Some respondents answer without reading: the same answer everywhere (straight-lining), random answers, or far too fast. Their answers weaken correlations and can create or hide effects. No single index catches every kind of careless answering, so the usual advice is to compute several indices and exclude only respondents flagged by more than one (Curran, 2016; Meade & Craig, 2012).

The practice file contains eight students who answered every item with 3. Let us find them.

3.1 The screening settings

1. Build the Motivation and Anxiety scores first (Section 1.7). SaNaStat takes the questionnaire items, the scales and the reverse-coded items from the construct scores.
2. On **Data**, click **Check data quality →** in the card *6 · Data quality screening*.
3. Check the card *Screening settings (What to check)* (Figure 24).
4. Click **Screen the data**.

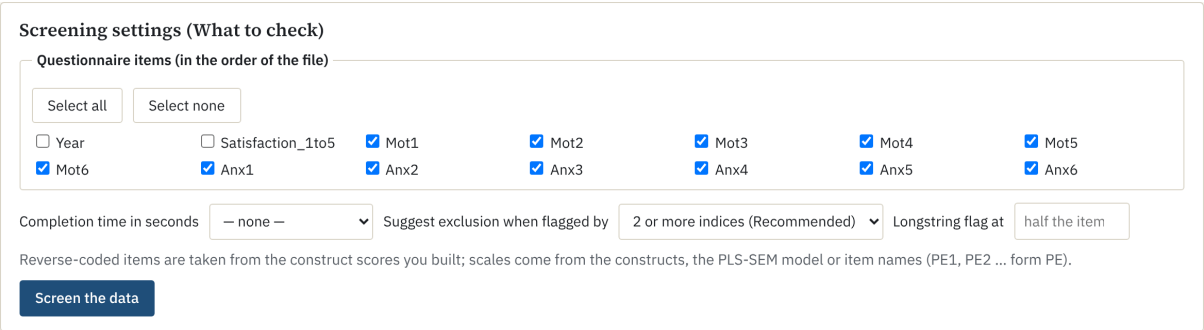


Figure 24. *The screening settings: the twelve questionnaire items are ticked.*

The list *Questionnaire items (in the order of the file)* offers every variable with whole-number answers and few values. With the constructs built, the twelve items Mot1 to Anx6 are ticked and Year and Satisfaction_1to5 are not. If you have not built constructs, SaNaStat ticks all the candidates: untick everything that is not an item of the questionnaire block you want to screen, because the indices compare a person's answers across the items you tick. **Completion time in seconds** is left at none (the practice file has no timing variable), **Suggest exclusion when flagged by** stays at **2 or more indices**

(Recommended), and **Longstring flag at** is left empty, which means half the items (here 6 of 12).

3.2 Reading the screening report

The report checks four things: careless responding, outliers ($|z| > 3.29$ on the questionnaire items, the construct scores and the other numeric variables of the file, and multivariate outliers on the construct scores by Mahalanobis D^2 with $p < .001$; Tabachnick & Fidell, 2019), the amount of missing data, and, when there are missing values among the items, Little's MCAR test (Little, 1988b).

Table 5

The careless-response indices

Index	Flag when
Longstring (same answer many times in a row)	as many identical answers in a row as the longstring flag (here 6 of 12)
Intra-individual response variability	the standard deviation of the person's answers is 0
Mahalanobis distance (unusual answer pattern)	D^2 with $p < .001$
Person–total correlation (answers unlike everyone else's)	$r < 0$; used only when the item means differ enough (SD of the item means at least 0.25)
Even–odd consistency (inconsistent within scales)	needs at least four scales with two or more items
Response time (too fast to have read the items)	needs a completion-time variable

What it means for our example

Three indices were computed on 12 items; the person–total correlation was not used because the item means are too similar (SD 0.04), the even–odd index needs at least four scales and we have two, and there is no completion time. Eight respondents (2.0% of 400) were flagged by two or more indices and are suggested for exclusion. Longstring flagged 28 respondents (7.0%), intra-individual response variability (an SD of zero) flagged 8 (2.0%), and Mahalanobis distance flagged none. The table *Overlap of the careless-response flags* shows that 20 respondents (5.0%) raised exactly one flag and 8 (2.0%) exactly two, all of them by longstring together with no variation. The 20 respondents flagged only by longstring have runs of 6 to 9 identical answers but vary elsewhere; this is why one index alone is not enough. Eight values are beyond $|z| = 3.29$, in two variables (Section 3.3), and nobody is a multivariate outlier on the construct scores.

Respondents suggested for exclusion

Respondent	Flags	Longstring	IRV	D² (p)	Person–total r	Even–odd r
ST072	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST109	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST119	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST172	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST278	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST293	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST299	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–
ST315	Longstring, Intra-individual response variability	12	0.00	0.0 (1.000)	–	–

Figure 25. *The respondents suggested for exclusion: all eight have a longstring of 12 and no variation.*

Outliers

Variable	Values with |z| > 3.29	Largest |z|	Respondents
Study_hours_week	5	9.09	ST037, ST114, ST122, ST236, ST333
Anxiety_score	3	4.38	ST103, ST172, ST286

Checked: the questionnaire items, the construct scores and the other numeric variables of the file (Study_hours_week, Sleep_hours, Anxiety_score, Pre_score, Post_score, Score_T1, Score_T2, Score_T3, Exam_score, Project_score). An outlier is not an error by itself: check it against the questionnaire; keep genuine extreme answers, correct typing errors, and report what you did. |z| > 3.29 corresponds to p < .001 (two-sided).

Figure 26. *The outliers table: values beyond |z| = 3.29 in numeric variables, with the respondents.*

The eight respondents (Figure 25) answered all 12 items with the same value: the students ST072, ST109, ST119, ST172, ST278, ST293, ST299 and ST315. SaNaStat labels respondents by the ID variable of the file, here StudentID; a file without one is labelled by row number. The dashed line in the longstring histogram marks the flag. The finding ends with a ready-made sentence for the flow of respondents: "Of the 400 respondents, 8 (2.0%) were excluded as careless responders because at least 2 indices flagged them ..., leaving 392 for analysis."

3.3 Outliers and missing values

The screening also looks for outliers in every numeric variable of the file, not only in the questionnaire items. The *Outliers* table (Figure 26) lists, for each variable, the number of values beyond |z| = 3.29, the largest |z| and the respondents concerned. Study_hours_week has five (largest |z| = 9.09: ST037, ST114, ST122, ST236 and ST333; up to 60 hours a week against a median of 6.30) and Anxiety_score has three (largest |z| = 4.38: ST103, ST172 and ST286). The check lists them; it does not remove them. An outlier is not an error by itself: check it against the questionnaire, keep genuine extreme answers, correct typing errors and report what you did. A box plot (Section 4.4) shows these extreme study times as well. Note that ST172 is both an outlier on Anxiety_score and one of the eight straight-liners.

To see the missing-data checks at work, tick Satisfaction_1to5 as a thirteenth item and screen again. SaNaStat then reports that 0.3% of the answers are missing (at most 4.0% for one item), that 384 respondents answered every item, and that Little's test does not reject missing completely at random, $\chi^2(12) = 6.18$, $p = .907$. Only two of the eight straight-liners are then suggested for exclusion (ST072 and ST299, whose satisfaction rating is also 3): for the other six the run of identical answers ends at the thirteenth item and their

response variability is no longer zero. The lesson is to screen the items of one questionnaire block, not unrelated variables. Untick Satisfaction_1to5 and screen again before you continue.

Cut-off values: missing data

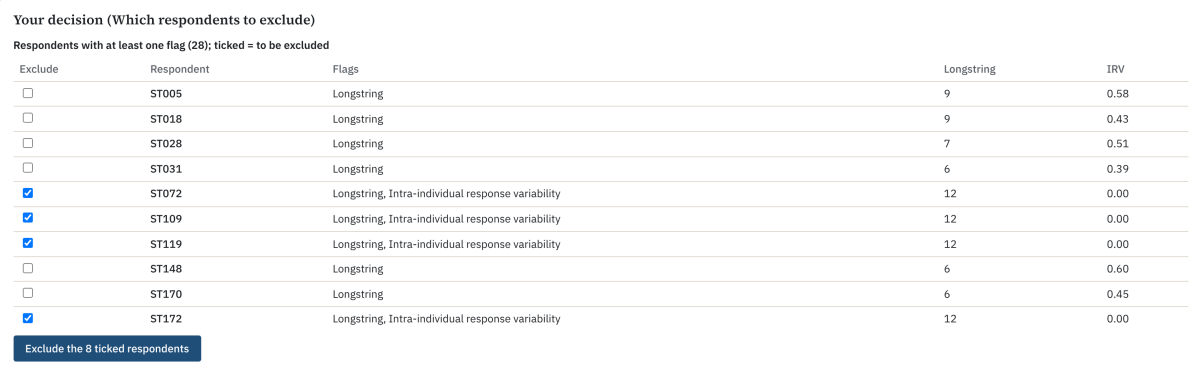
Under 10% missing per variable is usually ignorable if the missingness is random; more needs a principled method such as multiple imputation (Hair et al., 2019; chapter 14).

Little's MCAR test p ≥ .05: no evidence against missing completely at random (Little, 1988b). A non-significant result does not prove it, and a significant one does not rule out data that are missing at random. The EM estimates behind the test are for this diagnosis only; for analysis, use multiple imputation.

3.4 Your decision

The card *Your decision (Which respondents to exclude)* lists every respondent with at least one flag, 28 here; those suggested by the rule are ticked (Figure 27).

1. Look at the flagged respondents, and untick anyone you want to keep (or tick others).
2. Click **Add to Word report** under the screening report, so that the rule, the cut-offs and the number excluded are documented.
3. Click **Exclude the 8 ticked respondents**. The card *Screened data in use* confirms the exclusion: every analysis now uses the remaining 392 students.



Your decision (Which respondents to exclude)

Respondents with at least one flag (28); ticked = to be excluded

Exclude	Respondent	Flags	Longstring	IRV
☐	ST005	Longstring	9	0.58
☐	ST018	Longstring	9	0.43
☐	ST028	Longstring	7	0.51
☐	ST031	Longstring	6	0.39
☑	ST072	Longstring, Intra-individual response variability	12	0.00
☑	ST109	Longstring, Intra-individual response variability	12	0.00
☑	ST119	Longstring, Intra-individual response variability	12	0.00
☐	ST148	Longstring	6	0.60
☐	ST170	Longstring	6	0.45
☑	ST172	Longstring, Intra-individual response variability	12	0.00

Exclude the 8 ticked respondents

Figure 27. *Your decision: the flagged respondents, with the eight suggested for exclusion ticked.*

Undo the exclusion (use all 400 respondents) brings everyone back.

Note: the rest of this manual uses all 400 students

The eight straight-liners answered only the Mot and Anx items; their other variables (exam scores, hours, groups) are probably fine, and excluding them would change every number below. To let you reproduce the numbers in this manual, undo the exclusion before you continue. In your own study you would keep the exclusion, report it, and run every analysis on the screened data.

How to report it

"Data were screened for careless responding with three indices (longstring, intra-individual response variability, Mahalanobis distance; Curran, 2016; Meade & Craig, 2012). Eight of 400 respondents were flagged by at least two indices (each had answered all 12 items identically) and were excluded. Eight values on two variables lay beyond |z| = 3.29; they were checked and kept."

Exercise: find the straight-liners in practice_brand_survey.csv and careless_demo.csv

The brand survey has ten respondents who answered every item with 5. Open practice_brand_survey.csv, build Trust, Value and Loyalty scores, screen the data and check that ten respondents are suggested. Then open careless_demo.csv (300 respondents, 20 items and a Duration variable): choose Duration as **Completion time in seconds** and compare the number of respondents suggested when you ask for 1, 2 or 3 flags. Which setting would you report, and why? Answers: in the brand survey ten respondents are suggested (2.0% of 500; the even–odd index is not used because it needs at least four scales and there are three). In careless_demo.csv, five indices are computed on 20 items; 1 flag suggests 40 respondents (13.3%), 2 flags suggest 13 (4.3%) and 3 flags suggest nobody. The screen also lists three outliers on Duration (P011, P012 and P013; largest |z| = 4.01).

4. Describing your data

Describe your data before you test anything. Descriptive statistics tell you the typical values and the spread; charts show shape, unusual values and relationships that a single number hides. Both are on the **Analyses** and **Charts** screens.

4.1 Descriptive statistics

Click **Analyses** in the menu. The list on the left shows the analyses in four groups (*Common statistics*, *Measurement and models*, *HR and people analytics* and *Marketing analytics*), each with a one-line description; the result appears on the right. Choose **Descriptive statistics**, tick the scale variables to describe under *Scale variables (means, SDs, medians, skewness)* (here Study_hours_week, Sleep_hours, Anxiety_score and Exam_score) and, optionally, a categorical variable in **Split the scale statistics by (optional)**. The second list, *Categorical variables (frequencies and percentages)*, gives frequency tables (Section 10.1).

There is **no Run button** in most analyses: the result appears as soon as the variables are chosen, and changes when you change them (the slower analyses of chapters 12 to 14 have a Run button). Every result has the same parts:

- a title, for example *Descriptive statistics by Gender*;
- **What it means:** the finding in one or two plain sentences;
- the **checks**, each with a verdict (OK, Note, Check or Info), an explanation and the rule with its source;
- the **tables** and **figures**, each with buttons to copy or export it (Section 17.1);
- **How to report it (APA style):** a ready-made sentence, with **Copy this sentence**, **All tables (Excel)** and **Add to Word report**;
- **References**, collapsed at the bottom.

The table gives the number of valid values (n), the number missing, the mean, standard deviation (SD), median, minimum, maximum, and bias-corrected skewness and kurtosis with their standard errors in brackets.

Concept: mean, median, skewness and kurtosis

The **mean** is the arithmetic average; the **median** is the middle value. In a symmetric distribution they are close. A few very large values pull the mean up but hardly move the median, so a mean well above the median signals a right-skewed distribution. **Skewness** measures this asymmetry (0 = symmetric; positive = long tail to the right). **Kurtosis** here is *excess* kurtosis (0 = like a normal curve; large positive values = heavy tails or extreme values). SaNaStat flags a variable when |skewness| > 1.

What it means for our example

Study_hours_week has M = 7.28 and SD = 5.80 but a median of only 6.30, a maximum of 60.00 hours, skewness 5.63 and excess kurtosis 41.62: a few extreme values dominate its shape. SaNaStat therefore says that 1 of the 4 rows shows clear skewness and advises reporting medians and reading the rank-based tests. Anxiety_score is right-skewed too (skewness 0.97, M = 13.50, Mdn = 13.10, SD = 2.90, n = 378 with 22 missing), just under the flag. Sleep_hours (M = 6.68, SD = 1.07, n = 388) and Exam_score (M = 54.97, SD = 9.81, n = 387) are close to symmetric.

Choose **Gender** in **Split the scale statistics by (optional)** to see each variable for women and men (Figure 28). The mean Exam_score is 56.57 (SD = 9.64, n = 239) for women and 52.38 (SD = 9.57, n = 148) for men: a difference of about four points that chapter 6 will test. The APA sentence under the table now lists each group (for example "Exam_score: Gender = female, M = 56.57, SD = 9.64, n = 239; Gender = male, M = 52.38, SD = 9.57, n = 148"). Note that a descriptive table only describes; it does not say whether a difference could be due to chance.

What it means: 3 of the 8 rows show clear skewness (|skewness| > 1): report medians as well, and read the rank-based tests next to the tests of means.

Descriptive statistics

Variable	Gender	n	Missing	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Study_hours_week	female	246	0	7.47	6.35	6.40	1.40	60.00	5.40 (0.16)	36.38 (0.31)
Study_hours_week	male	154	0	6.96	4.79	6.20	0.70	52.50	5.87 (0.20)	53.24 (0.39)
Sleep_hours	female	238	0	6.71	1.03	6.70	3.90	9.90	0.08 (0.16)	0.29 (0.31)
Sleep_hours	male	150	0	6.61	1.14	6.70	3.50	9.80	−0.13 (0.20)	0.18 (0.39)
Anxiety_score	female	233	0	13.30	2.75	13.00	7.10	26.20	1.04 (0.16)	2.75 (0.32)
Anxiety_score	male	145	0	13.81	3.11	13.20	8.40	24.50	0.85 (0.20)	0.68 (0.40)
Exam_score	female	239	0	56.57	9.64	55.70	33.10	80.10	0.07 (0.16)	−0.24 (0.31)
Exam_score	male	148	0	52.38	9.57	51.35	29.00	75.10	0.13 (0.20)	−0.35 (0.40)

Skewness and kurtosis are bias-corrected (G1, G2; kurtosis is excess kurtosis); standard errors in brackets. Median = weighted average at (n + 1)p.

Figure 28. *The same variables split by Gender.*

Note: small groups

Hide results for groups smaller than protects people in small groups: their numbers are shown as "< k" and those groups are left out of the tests. The setting applies to the whole project. With 400 students it can stay at **Show every group (no hiding)**.

4.2 The Charts screen

Click **Charts** in the menu. The card *1 · Choose the chart* has a **Chart type** list and the variables the chart needs; the chart appears below with a short description, a summary table and a caption you can add to the Word report. Six chart types are available: histogram, box plot, means with confidence intervals, scatter plot with a fitted line, correlation heatmap and bar chart of percentages. Every chart can be copied or saved as a PNG or SVG picture (Section 17.1).

> **Note: charts show patterns, not proof**
>
> An apparent difference in a chart needs a test, and a pattern in a scatter plot can come from a few unusual respondents.

4.3 Histogram: the shape of one variable

Choose **Histogram (How are the values spread?)** and Anxiety_score as the **Variable** (Figure 29). Each bar counts the respondents whose value falls in a band; the lines mark the mean and the median.

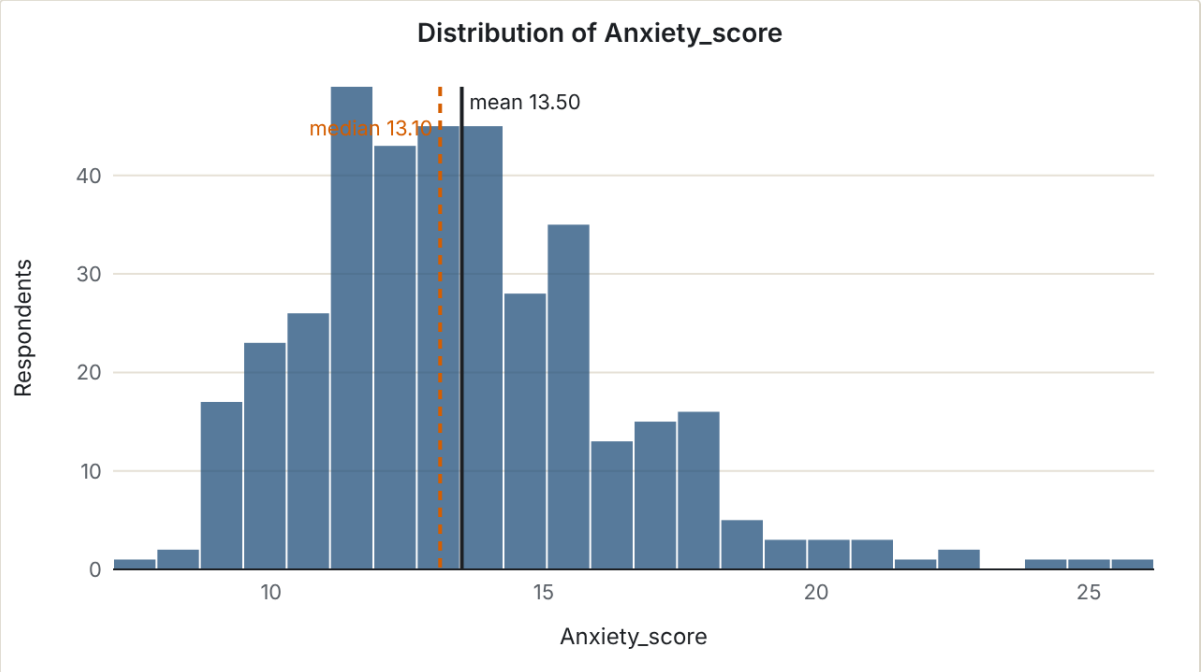


Figure 29. *Histogram of Anxiety_score with the mean (13.50) and median (13.10).*

SaNaStat summarises it: Anxiety_score has 378 values with mean 13.50, median 13.10 and SD 2.90, and the distribution is skewed to the right (a long tail of high values). Many tests of means tolerate this with 30 or more cases per group; for a skewed outcome in small groups, prefer the rank-based test.

4.4 Box plot: medians, spread and unusual values

Choose **Box plot (Medians and spread, by group)** and Study_hours_week (Figure 30). The box spans the middle half of the values (the quartiles), the line is the median, the whiskers reach the most extreme values within 1.5 times the interquartile range, and the circles are unusual values beyond that.

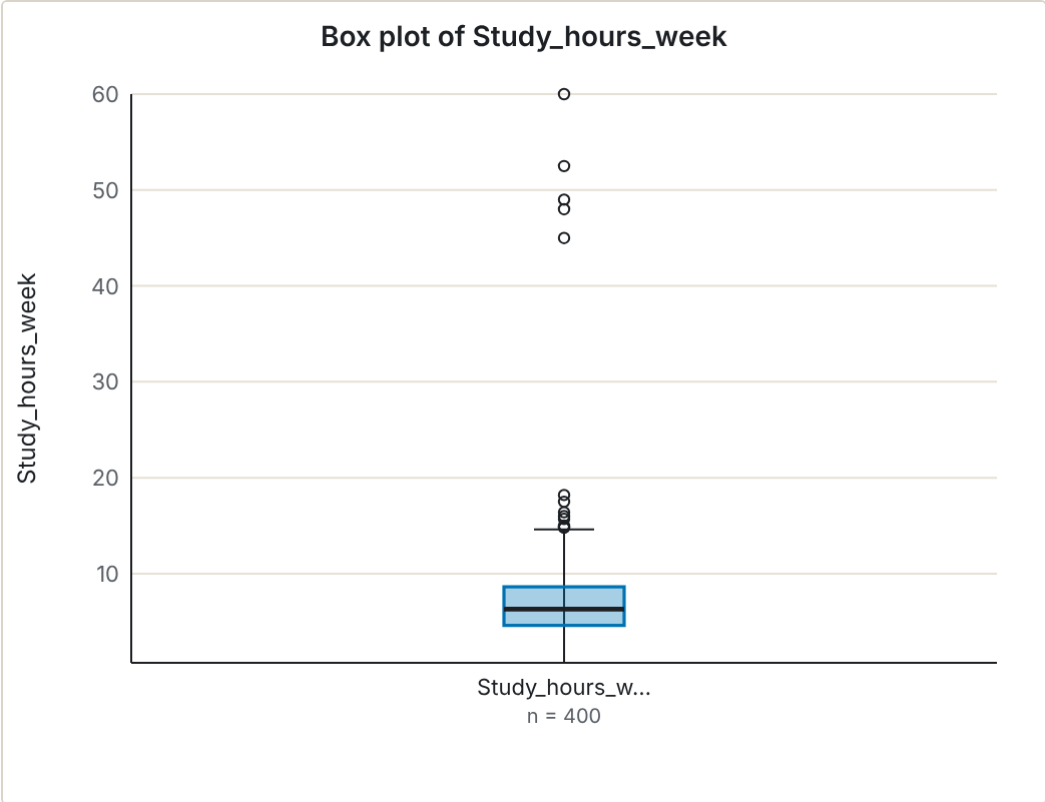


Figure 30. *Box plot of Study_hours_week: the circles are unusual values.*

The median is 6.30 hours and the quartiles are 4.60 and 8.70. Five students (with 45 to 60 hours a week) lie far above the rest, and several more lie above the whisker at about 15 hours. Before you analyse such a variable, ask whether the extreme values are errors (a typing mistake, hours per month instead of per week), a different kind of student, or a true feature of the data. Do not delete them just because they are inconvenient: compare the result with and without them, and use rank-based methods (Section 11.2), which are not pulled by single extreme values.

Choose Exam_score and **Groups (optional)** Gender to compare groups. The medians are 55.70 for women and 51.35 for men.

4.5 Means with confidence intervals

Choose **Means with confidence intervals (Group averages)**, Exam_score and **Groups** Teaching_method (Figure 31).

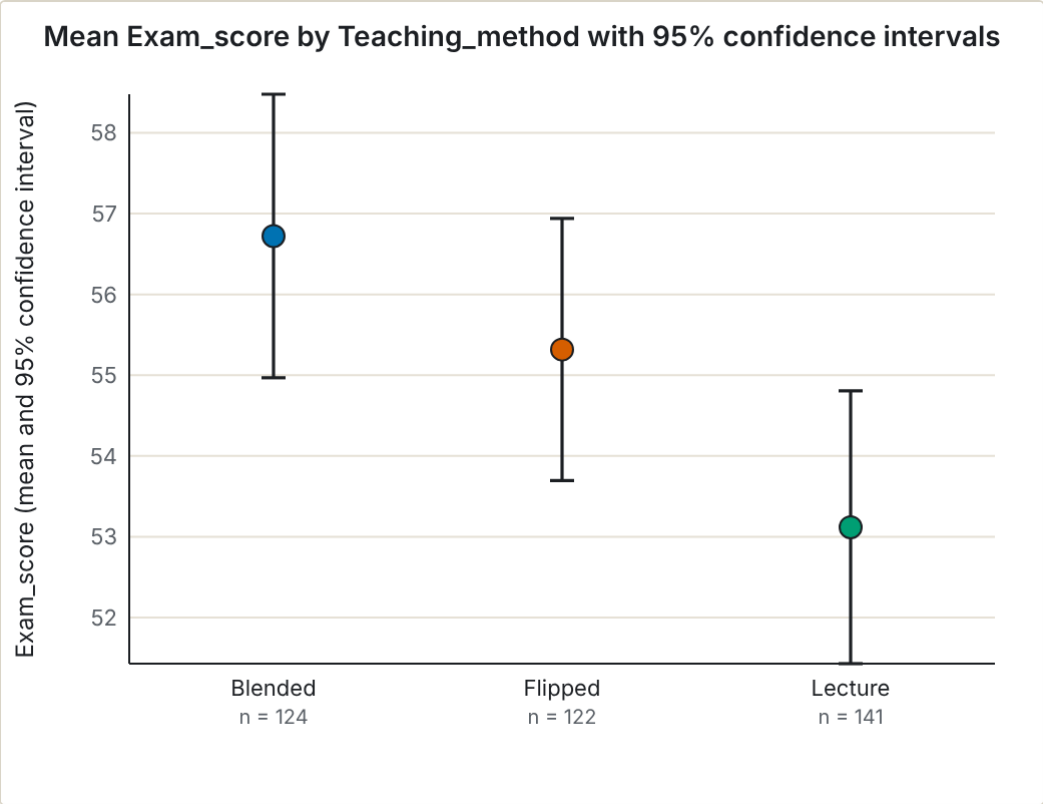


Figure 31. *Mean Exam_score by Teaching_method with 95% confidence intervals.*

Concept: confidence interval

A 95% confidence interval is a range computed so that, in repeated samples, 95% of such intervals would contain the true population mean. It shows how precisely each group mean is estimated. Where two intervals do not overlap, the groups very likely differ; overlapping intervals do not by themselves show that they are the same. Run the comparison in the Analyses to test it.

The means are Blended 56.72 [54.97, 58.48], Flipped 55.32 [53.70, 56.94] and Lecture 53.12 [51.43, 54.81]. The intervals of Blended and Lecture hardly separate; chapter 7 tests this properly.

4.6 Scatter plot with a fitted line

Choose **Scatter plot with line (Do two variables go together?)**, Study_hours_week on the **Horizontal axis** and Exam_score on the **Vertical axis** (Figure 32). Optionally choose **Groups (optional)** to colour the dots by group.

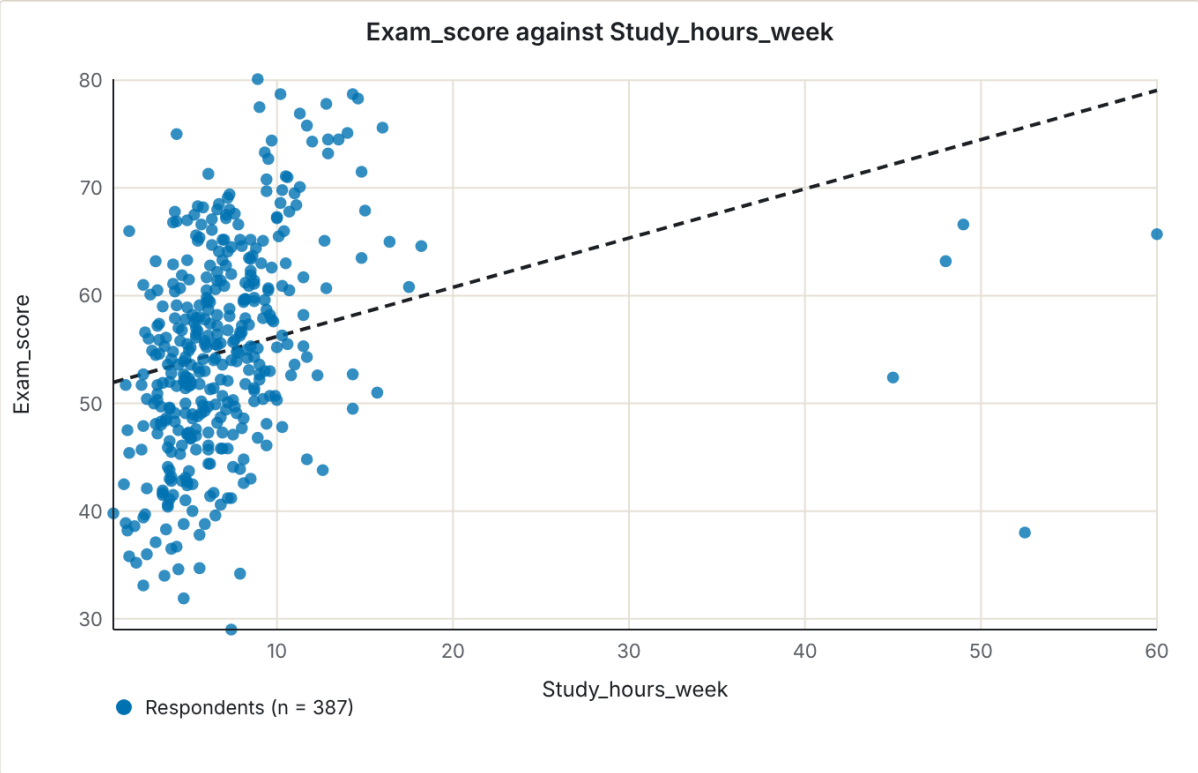


Figure 32. *Exam_score against Study_hours_week with the least-squares line.*

The relationship is small and positive (Pearson r = .27, n = 387). The dashed line is the straight-line fit: Exam_score changes by 0.457 for each one-unit increase in Study_hours_week (intercept 51.632). Look at the dots before you trust r: a few extreme study hours on the right pull the line, and the cloud may curve rather than follow a straight line. Chapter 11 compares Pearson's r with rank correlations on these data.

4.7 Bar chart of percentages

Choose **Bar chart of percentages (Categories)** and Programme. The table gives counts and percentages.

Programme (n = 400): Business 40.8% (163 students), Science 36.3% (145) and Arts 23.0% (92). Add any chart to the Word report with **Add to Word report**; its caption sentence ("Figure. Distribution of ...") goes with it.

How to report it

"Weekly study time was strongly right-skewed (M = 7.28, SD = 5.80, Mdn = 6.30, skewness = 5.63), with a few students studying more than 40 hours; medians are therefore reported alongside means."

Exercise: charts on practice_marketing.csv and segments_demo.csv

Open practice_marketing.csv. Draw a histogram of Annual_spend and a box plot of Income_k: both are skewed (the file note says they were built that way), so a logarithm is often taken before a linear regression (Section 2.6). Draw a correlation heatmap of Income_k, Visits_per_month, Avg_basket and Annual_spend. Answer: Visits_per_month and Annual_spend correlate at r = .87, Income_k and Avg_basket at r = .48 and Avg_basket and Annual_spend at r = .42, while Income_k and Visits_per_month are unrelated (r = −.00). Then open segments_demo.csv and draw a scatter plot of Brand_loyalty against Price_sensitivity, grouped by Channel: the relationship is negative.

4.8 Further statistics and stem-and-leaf plots

Under **More for the scale variables (optional)** on the Descriptive statistics screen, tick **Further statistics** and **Stem-and-leaf plot**, then choose **Sleep_hours**.

Further statistics (Figure 33) adds the 5% trimmed mean (the mean after removing 5% of the cases at each end), the mode, the quartiles, the interquartile range and the range. For Sleep_hours the trimmed mean (6.68) equals the mean, the median is 6.70 and the middle half of the students sleep between 5.90 and 7.40 hours (IQR = 1.50): the distribution is symmetric.

Further statistics

Variable	5% trimmed mean	Mode	Q1	Median	Q3	IQR	Range
Sleep_hours	6.68	6.70	5.90	6.70	7.40	1.50	6.40

5% trimmed mean: the mean after removing 5% of the cases at each end (the cases at the cut count in part). Mode: the smallest value if there are several. Quartiles: weighted average at (n + 1)p; IQR = Q3 − Q1.

Figure 33. *Further statistics of Sleep_hours.*

The stem-and-leaf plot (Figure 34) is a histogram made of the data themselves (Tukey, 1977). Each line is a stem (here half a unit wide, so 6 appears twice: 6.0 to 6.4 and 6.5 to 6.9), and each leaf digit is the first decimal of a value. With 388 values, each leaf stands for 2 cases, and & marks a leaf with fewer cases. Two very short and three very long sleepers are listed as extremes at the top and bottom.

Stem-and-leaf plot of Sleep_hours

Frequency	Stem	Leaf
2	Extremes	(=<3.60)
2	3	&&
5	4	&&&4
11	4	&677&9&
27	5	0001222&333&44&
51	5	55&66666677788888&99999999&
62	6	0000000111111&222222&33334444444
68	6	5555556666666&777777777888&99999999
67	7	000000&111111&22222222&333333444444
53	7	555555566666677777&888899999
26	8	0000&11&2&3344&
6	8	56&&
3	9	&&&
2	9	&&
3	Extremes	(>=9.70)

Stem width: 0.5; each leaf: 2 cases (leaf unit 0.1); & stands for a leaf with fewer than 2 cases. Extremes are values beyond 1.5 hinge spreads from the hinges (Tukey, 1977).

Figure 34. *Stem-and-leaf plot of Sleep_hours.*

5. Choosing the right test

Choosing a test is a decision about your data, not about the software. Four questions settle most cases.

1. **What kind of variable is the outcome?** A number or a rating you can average (scale), or a category (categorical)?
2. **How many groups or measurements are compared?** Two, or three or more?
3. **Are the measurements independent or paired?** Different people in each group (independent), or the same people measured more than once (paired or repeated)?
4. **What are you asking?** A difference between groups, a relationship between variables, or a prediction?

Concept: independent and paired data

Data are **independent** when each person contributes one value and what one person gives tells you nothing about another (men and women in a survey). Data are **paired** when the same people (or matched pairs) are measured again (a pre-test and a post-test). Ignoring the pairing throws away power; treating paired data as independent gives wrong p-values.

5.1 The Help me choose guide

SaNaStat asks the first and the last question for you. Click **Analyses**, then **Help me choose** at the top of the list (*Test selection guide*).

1. Answer *1 · What kind of answer are you analysing?*: a continuous outcome (a number or a rating) or a categorical outcome (a category).
2. Answer *2 · What do you want to compare or test?* (for a categorical outcome: *What do you want to know?*).
3. SaNaStat shows the suggested test, for example **McNemar test**, and **Open it →** opens it with the right design already chosen.

Each test also computes its rank-based or exact alternative and tells you which one to read, based on your data, so the guide only needs to settle the design.

5.2 The decision table

Table 6
Which test for which question (the tests of SaNaStat)

Outcome	Question and design	Test in SaNaStat	Rank-based or exact alternative
Scale	Two independent groups	Independent-samples t-test (Welch)	Mann–Whitney U

Outcome	Question and design	Test in SaNaStat	Rank-based or exact alternative
Scale	Same people, two measurements	Paired-samples t-test	Wilcoxon signed-rank
Scale	One group against a fixed value	One-sample t-test	Wilcoxon signed-rank
Scale	Three or more independent groups	One-way ANOVA (Welch; Tukey–Kramer, Games–Howell, Bonferroni or Holm)	Kruskal–Wallis with Dunn
Scale	Same people, three or more times	One-way repeated-measures ANOVA (Mauchly, Greenhouse–Geisser, Huynh–Feldt)	Friedman
Scale	Groups from two or more factors, or adjusted for a covariate	Factorial ANOVA, ANCOVA	none
Scale (several)	Several outcomes at once, one factor	One-way MANOVA	none
Categorical	Two categorical variables related?	Chi-square test of independence	Fisher's exact (2 × 2)
Categorical	Categories equally common?	Chi-square goodness-of-fit	none
Categorical	Same people, yes/no, two times	McNemar test	Exact binomial
Scale	How strongly do variables go together?	Pearson, Spearman, Kendall, partial correlation	Spearman or Kendall
Scale	Which variables predict a score?	Multiple linear regression (hierarchical)	none
Categorical (two)	Which variables predict a yes/no outcome?	Binary logistic regression	none
Ordinal (3 to 10 ordered)	Which variables predict an ordered outcome?	Ordinal logistic regression (Brant test)	none
Categorical (3 to 10)	Which variables predict a choice among categories?	Multinomial logistic regression	none
Items	Do the items measure what they should?	Reliability, exploratory and confirmatory factor analysis	none

Outcome	Question and design	Test in SaNaStat	Rank-based or exact alternative
Scale	Does X act through M, or does W change the effect of X?	Mediation and moderation (Hayes models 1, 4, 7, 14)	bootstrap intervals

5.3 Parametric or rank-based?

Concept: assumptions of the tests of means

The t-test and ANOVA assume independent observations, a roughly normal distribution of the sample *mean* (which the central limit theorem delivers for moderate to large groups even when the data are skewed) and, for the classic versions, equal variances. SaNaStat's default tests, Welch's t-test and Welch's ANOVA, do not need equal variances. The rank-based tests (Mann–Whitney, Wilcoxon, Kruskal–Wallis, Friedman, Spearman) need no normality, are not pulled by outliers, and are preferable for small samples of clearly skewed or ordinal data; with large samples both families usually agree. SaNaStat reports both and tells you which one to read, and shows the Shapiro–Wilk test and normal Q–Q plots so that you can see the shape of the data (Section 6.4).

5.4 Multiple tests and error rates

Every test you run has a 5% chance of a false positive when there is no effect at $\alpha = .05$. Run twenty tests and you should expect one. Decide your tests before you look at the data, report all of them, and use the adjusted p-values SaNaStat gives for families of comparisons: Tukey–Kramer, Games–Howell, Bonferroni or Holm after an ANOVA, Dunn with the Holm adjustment after Kruskal–Wallis, Holm-adjusted pairwise tests after repeated measures and factorial designs, and Holm-adjusted p-values in a correlation matrix (Holm, 1979).

5.5 The significance level α

The significance level is a decision you make once, for the whole project, before you look at the results. The box *Significance level α* under the list of analyses (Figure 35) offers .05 (the usual choice, with 95% intervals), .01 (stricter: fewer false alarms, 99% intervals) and .10 (lenient: for small pilot studies, 90% intervals). Every analysis uses it: a result counts as significant when p is below α, and confidence intervals are $100(1 - \alpha)\%$. Effect sizes of the F and χ^2 family (η^2, ω^2, partial η^2, Cramér's V, φ, w) get $100(1 - 2\alpha)\%$ intervals, 90% at $\alpha = .05$, whose lower limit is above zero exactly when the test is significant (Steiger, 2004).

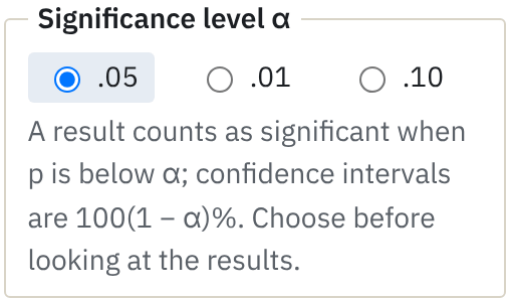


Figure 35. *The significance level α, under the list of analyses.*

What changes with α = .01? Choose .01 and look again at the t-test of Section 6.1: the test statistics and p-values are the same (t(313.32) = 4.17, p < .001), but every interval widens to 99%: the difference of the means has the 99% CI [1.59, 6.79], d = 0.44 has the 99% CI [0.16, 0.71], and the sentence says "d = 0.44, 99% CI [0.16, 0.71]". A result with p = .03 would no longer count as significant. This manual uses α = .05 throughout; set it back to .05.

Common mistake: choosing α after the results

Switching to .10 because p = .07, or to .01 to make a result look stronger, invalidates the test. Choose α before you collect or look at the data, and report it in the method section.

Exercise: choose the test

For each question below, use the guide (or Table 6) to name the test: (a) Do students in the three programmes differ in Exam_score? (b) Did the same students' Pre_score rise by the time of Post_score? (c) Is the proportion who passed related to Gender? (d) Does Sleep_hours predict Exam_score? (e) Does teaching method predict the 1 to 5 satisfaction rating? Answers: (a) one-way ANOVA; (b) paired-samples t-test; (c) chi-square test of independence; (d) correlation or multiple linear regression; (e) ordinal logistic regression (or, treating the rating as a score, a one-way ANOVA).

6. Comparing two groups or two measurements

This chapter covers the t-tests and their rank-based alternatives: two independent groups (Exam_score for women and men), the same students measured twice (Pre_score and Post_score) and one group against a fixed value.

Concept: the t-test and its effect size

A t-test divides the difference between two means by its standard error: the larger the ratio, the less likely a difference this large arises by chance if the true difference is zero. The p-value is that probability. It says nothing about how large or important the difference is. For that you need an **effect size**; for two means, Cohen's d: the difference in standard deviations (Cohen, 1988). Hedges' g corrects d for its small-sample bias (Hedges, 1981). Always report the effect size with a confidence interval.

Choose **t-tests** (*Compare two groups or two measurements*) in the Analyses menu. Under *What do you want to compare?* pick the design: **independent-samples**, **paired-samples** or **one-sample**. **Direction of the test** is two-tailed by default; choose a one-tailed test only if your hypothesis predicted the direction before you saw the data.

6.1 Independent-samples t-test: Exam_score by Gender

1. Choose **Independent-samples t-test (Two groups of different people)**.
2. In **Score to compare**, choose **Exam_score**; in **Groups**, choose **Gender** (Figure 36). If the grouping variable has more than two groups, two lists ask which two to compare; use One-way ANOVA (chapter 7) for all of them.

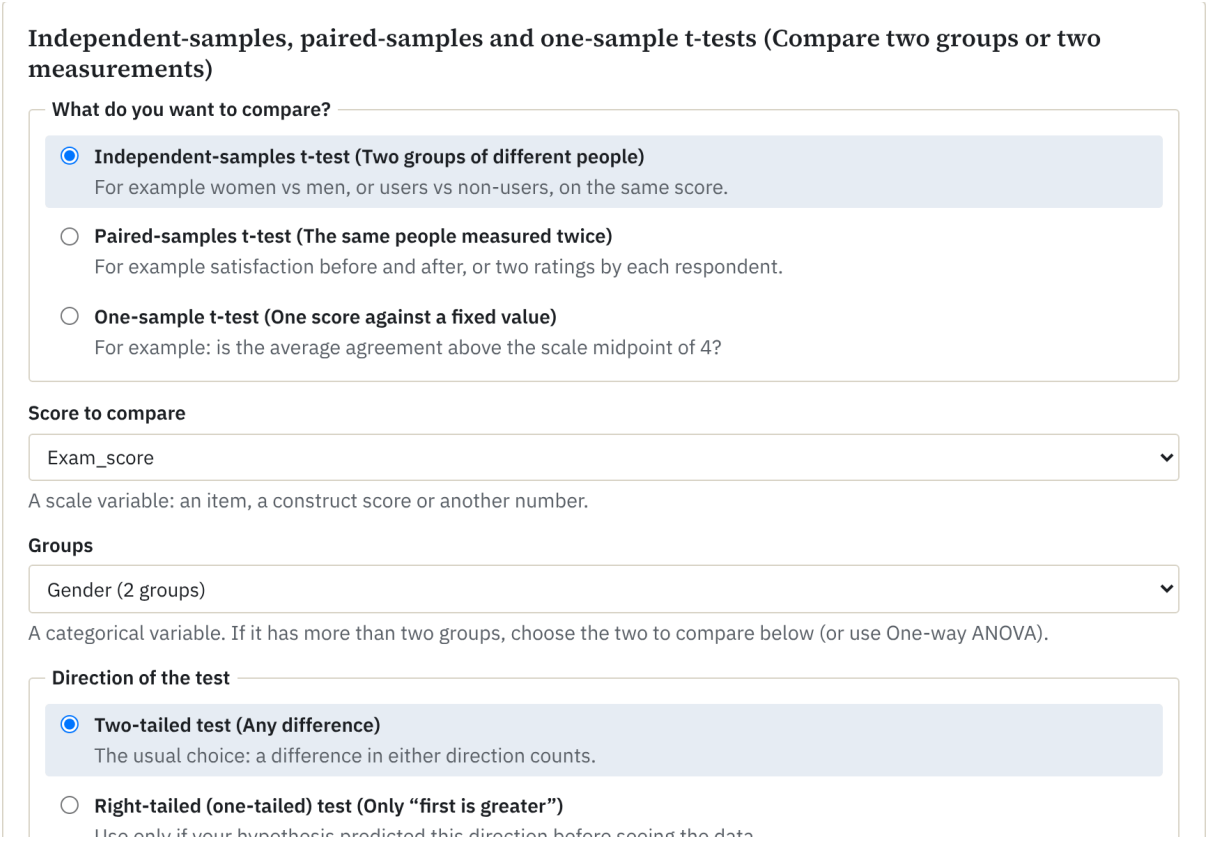


Figure 36. *The t-test screen with the independent-samples design: Exam_score by Gender.*

Group descriptives

Group	n	Mean	SD	Median	Skewness	Shapiro–Wilk W	p (normality)
female	239	56.57	9.64	55.70	0.07	.992	.265
male	148	52.38	9.57	51.35	0.13	.992	.608

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965): $p < \alpha$ suggests the scores are not normally distributed.

Copy table | Copy as picture | PNG | CSV

t-tests (two-sided)

Test	t	df	p	Mean difference	95% CI
Welch's t (recommended)	4.17	313.32	< .001	4.19	[2.21, 6.16]
Student's t (equal variances)	4.17	385	< .001	4.19	[2.21, 6.16]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d	0.44	[0.23, 0.64]	small
Hedges' g	0.43	[0.23, 0.64]	small

CIs of d and g are 95% intervals from the noncentral t distribution (Steiger & Fouladi, 1997); Hedges' g corrects d for small-sample bias (Hedges, 1981). The CI of the rank-biserial r uses Fisher's z (Ben-Shachar et al., 2020).

Figure 37. *Group descriptives with Shapiro–Wilk, the t-tests and the effect sizes with their confidence intervals.*

The result (Figure 37) reports Welch's t-test first, because it is accurate whether or not the variances are equal (Delacre et al., 2017), followed by Student's t-test, the effect sizes

(Cohen's d and Hedges' g, each with its confidence interval), the Mann–Whitney U test as the rank-based alternative with the rank-biserial correlation and its interval, Levene's and the Brown–Forsythe tests of equal variances, and normal Q–Q plots of each group.

What it means for our example

Women (M = 56.57, SD = 9.64, n = 239) scored higher than men (M = 52.38, SD = 9.57, n = 148) on Exam_score. The difference of 4.19 points is statistically significant, Welch's t(313.32) = 4.17, p < .001, 95% CI [2.21, 6.16], and small by Cohen's benchmarks, d = 0.44, 95% CI [0.23, 0.64], Hedges' g = 0.43, 95% CI [0.23, 0.64]. SaNaStat states the size as |d| = 0.44 and explains that the sign of d follows the order of the groups: here d = 0.44 means that female scored higher than male. Levene's test does not suggest unequal variances, F(1, 385) = 0.04, p = .834, so Student's t gives the same result, t(385) = 4.17. The Mann–Whitney test agrees, U = 21966.5, z = 4.00, p < .001, rank-biserial r = .24, 95% CI [.13, .35].

How to report it

"An independent-samples Welch's t-test showed that Exam_score differed between female (M = 56.57, SD = 9.64) and male (M = 52.38, SD = 9.57), t(313.32) = 4.17, p < .001, d = 0.44, 95% CI [0.23, 0.64], Hedges' g = 0.43, 95% CI [0.23, 0.64]."

One-tailed tests. Choose **Right-tailed (one-tailed) test (Only “first is greater”)** if you predicted before seeing the data that the first group (female) would be higher. The p-value then refers to that one direction only and the confidence interval of the difference becomes one-sided (here [2.53, ∞]), while t and d are unchanged. SaNaStat's sentence now says so: "An independent-samples Welch's t-test (one-tailed: female greater than male) showed that Exam_score was higher for female (M = 56.57, SD = 9.64) than for male (M = 52.38, SD = 9.57), t(313.32) = 4.17, p < .001, d = 0.44, 95% CI [0.23, 0.64] ...". When the result is not significant it says "was not significantly higher". The confidence interval of d stays two-sided. Never switch to a one-tailed test after seeing the direction of the result.

6.2 Effect sizes and their confidence intervals

An effect size without an interval hides how precisely it is known. SaNaStat gives an interval with every effect size (Figure 38): for d and g from the noncentral t distribution (Steiger & Fouladi, 1997), for the rank-biserial r through Fisher's z. Here d = 0.44 could be anything from 0.23 (small) to 0.64 (medium) in the population: the data show that women score higher, but not exactly by how much.

Effect size

Measure	Value	95% CI	Size
Cohen's d	0.44	[0.23, 0.64]	small
Hedges' g	0.43	[0.23, 0.64]	small

CIs of d and g are 95% intervals from the noncentral t distribution (Steiger & Fouladi, 1997); Hedges' g corrects d for small-sample bias (Hedges, 1981). The CI of the rank-biserial r uses Fisher's z (Ben-Shachar et al., 2020).

Figure 38. *The effect sizes of the independent-samples t-test, each with its 95% confidence interval.*

Table 7

Effect sizes for two means and their benchmarks (Cohen, 1988)

Effect size	What it is	Small / medium / large
Cohen's d	Difference of two means in pooled SDs	0.2 / 0.5 / 0.8
Hedges' g	d corrected for small-sample bias	as d
d_z, g_z	Mean change divided by the SD of the changes (paired data)	as d
Rank-biserial r	Effect size of the Mann–Whitney and Wilcoxon tests, −1 to 1	.1 / .3 / .5

6.3 Paired-samples t-test: Pre_score and Post_score

In a paired design every student has two scores, so the analysis is of the *differences*.

1. Choose **Paired-samples t-test (The same people measured twice)**.
2. In **First measurement**, choose **Pre_score**; in **Second measurement**, choose **Post_score**. The change is second minus first.

Descriptives (complete pairs)

Measure	n	Mean	SD	Median	Shapiro–Wilk W	p (normality)
Pre_score	400	59.37	11.65	59.05		
Post_score	400	66.09	13.47	65.70		
Difference (Post_score – Pre_score)	400	6.72	6.37	6.85	.996	.308

The paired t-test assumes that the differences are roughly normal, so Shapiro–Wilk (Shapiro & Wilk, 1965) is shown for the differences.

Copy table | Copy as picture | PNG | CSV

Paired t-test (two-sided)

t	df	p	Mean difference	95% CI
21.12	399	< .001	6.72	[6.10, 7.35]

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	95% CI	Size
Cohen's d_z	1.06	[0.93, 1.18]	large
Hedges' g_z	1.05	[0.93, 1.18]	large

d_z = mean change / SD of the changes. CIs of d and g are 95% intervals from the noncentral t distribution (Steiger & Fouladi, 1997); Hedges' g corrects d for small-sample bias (Hedges, 1981). The CI of the rank-biserial r uses Fisher's z (Ben-Shachar et al., 2020).

Copy table | Copy as picture | PNG | CSV

Rank-based alternative

Test	V	z	p	Matched rank-biserial r	95% CI	Method
Wilcoxon signed-rank	73510.5	14.87	< .001	.86	[.83, .89]	normal approximation

3 pairs with no change were left out of the signed-rank test.

Figure 39. *The descriptives of the complete pairs with Shapiro–Wilk for the differences, the paired t-test, the effect sizes and the Wilcoxon signed-rank test.*

What it means for our example

Post_score (M = 66.09, SD = 13.47) is higher than Pre_score (M = 59.37, SD = 11.65) by 6.72 points on average (SD of the differences 6.37, n = 400 complete pairs). The change is significant and large, t(399) = 21.12, p < .001, 95% CI of the difference [6.10, 7.35], d_z = 1.06, 95% CI [0.93, 1.18], Hedges' g_z = 1.05. The Wilcoxon signed-rank test agrees, V = 73510.5, z = 14.87, p < .001, matched rank-biserial r = .86, 95% CI [.83, .89]; three students with no change were left out of it.

Effect size d_z is the mean change divided by the SD of the changes (here 6.72 / 6.37 ≈ 1.06). It is not comparable with Cohen's d from independent groups, because its denominator is smaller when the two measurements are strongly correlated (here r = .88). The same benchmarks (.2, .5, .8) are applied.

How to report it

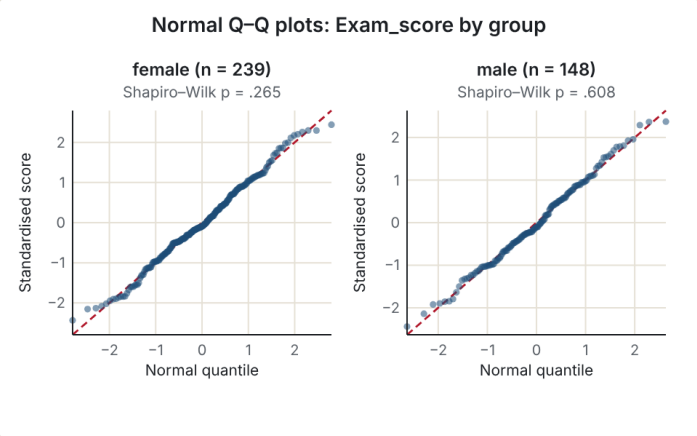
"A paired-samples t-test showed that Post_score (M = 66.09, SD = 13.47) differed from Pre_score (M = 59.37, SD = 11.65), t(399) = 21.12, p < .001, d_z = 1.06, 95% CI [0.93, 1.18], Hedges' g_z = 1.05, 95% CI [0.93, 1.18]."

Note: a rise from pre to post is not proof of a teaching effect

Without a comparison group, a pre-post rise can reflect practice, maturation or the test itself. The comparison of teaching methods in chapter 7 is the stronger design because students were assigned at random.

6.4 Is the outcome normal enough? Shapiro–Wilk and Q–Q plots

The t-tests assume a roughly normal distribution of the scores in each group (of the differences, for paired data). SaNaStat checks it in two ways. The **Shapiro–Wilk test** (Shapiro & Wilk, 1965) is given for each group in the descriptives table; p below α suggests that the scores are not normally distributed. A **normal Q–Q plot** (Figure 40) plots each respondent's standardised score against the value expected under a normal distribution (Wilk & Gnanadesikan, 1968): points close to the dashed line mean a normal shape, a bend at one end means skew, an S-shape heavy or light tails, and Likert-type scores form steps, which is expected.



Normal Q–Q plots: Exam_score by group. Each point is one respondent's standardised score against the value expected under a normal distribution; the dashed line marks a perfect fit. A bend at one end means skew, an S-shape heavy or light tails; Likert-type scores form steps, which is expected. (Wilk & Gnanadesikan, 1968)

Figure 40. *Normal Q–Q plots of Exam_score for women and men.*

For Exam_score both groups look normal: Shapiro–Wilk W = .992, p = .265 for women and W = .992, p = .608 for men, and the points follow the line. SaNaStat's check *Normality* combines this with the sample size: every group has 30 or more cases, so the test of means is robust to non-normal data anyway (Lumley et al., 2002).

Note: Shapiro–Wilk with large samples

With hundreds of cases the Shapiro–Wilk test flags even small, harmless departures from normality, and SaNaStat says so. In practice_marketing.csv the paired spend differences of 1,500 customers give W = .737, p < .001: the test is significant, but with this many cases the t-test is still accurate. Look at the Q–Q plot and the skewness, and read the rank-based test when the shape is clearly skewed and the groups are small.

6.5 One-sample t-test: Exam_score against the pass mark

Passing needs an Exam_score of 50 or more. Is the average above the pass mark?

1. Choose **One-sample t-test (One score against a fixed value)**.
2. In **Score to compare**, choose **Exam_score**, and set **Test value** to **50**.

What it means for our example

The mean of Exam_score (54.97, SD = 9.81, n = 387) is significantly above 50: t(386) = 9.96, p < .001, 95% CI of the mean [53.99, 55.95], d = 0.51, a medium effect (95% CI [0.40, 0.61]); Shapiro–Wilk W = .996, p = .486. The Wilcoxon signed-rank test agrees, V = 56396.0, z = 8.81, p < .001, matched rank-biserial r = .52, 95% CI [.43, .60].

How to report it

"A one-sample t-test showed that Exam_score (M = 54.97, SD = 9.81) differed from the test value of 50.00, t(386) = 9.96, p < .001, d = 0.51, 95% CI [0.40, 0.61], Hedges' g = 0.51, 95% CI [0.40, 0.61]."

Tip: choose a meaningful test value

The test value should come from your question, not from your data: a pass mark, the scale midpoint (4 on a 1 to 7 scale), a published norm. Testing a mean against a value just below it will always find a "significant" result with a large sample, so also look at the confidence interval and d.

Exercise: t-tests on practice_marketing.csv and practice_hr.csv

(a) In practice_marketing.csv (1,500 customers), run a paired-samples t-test of Monthly_spend_before and Monthly_spend_after. Answer: spend rose from M = 282.64 (SD = 198.64) to M = 304.35 (SD = 215.75), t(1499) = 28.78, p < .001, d_z = 0.74, 95% CI [0.69, 0.80], a medium effect. (b) Compare Annual_spend between loyalty members and non-members (Loyalty_member). Annual_spend is right-skewed and Levene's test is significant, so read Welch's test and the Mann–Whitney test: members spent more, M = 4389.38 (SD = 2591.22) against M = 2637.02 (SD = 1893.02), Welch's t(1131.47) = −14.51, p < .001, d = −0.79, 95% CI [−0.90, −0.68] (the sign shows the order of the groups), Mann–Whitney z = −14.09, p < .001. (c) In practice_hr.csv (1,200 employees), run a paired-samples t-test of Skills_test_before and Skills_test_after: M = 58.13 (SD = 11.07) rose to M = 63.74 (SD = 12.74), t(1199) = 31.37, p < .001, d_z = 0.91.

7. Comparing three or more groups

This chapter compares the mean of one outcome across three or more independent groups. Students in practice_students.csv were assigned at random to one of three teaching methods (Lecture, Blended, Flipped), so a difference between the methods can be read as an effect of teaching, apart from chance.

Concept: why not several t-tests?

With three groups there are three pairs. Testing each pair at α = .05 raises the chance of at least one false positive to about 14%, and with five groups (ten pairs) to about 40%. A one-way **analysis of variance (ANOVA)** first asks one question: do the group means differ at all? It compares the variation *between* the group means with the variation *within* the groups (the ratio F). Only if that omnibus test is significant do you look at which pairs differ, with a method that keeps the overall error rate at 5% (post hoc tests).

Note: the post-test and the gain

The planted effect in practice_students.csv is a larger *gain* (Post_score minus Pre_score) with the flipped classroom. This chapter compares Post_score across the methods; Section 7.6 analyses the gain itself, computed in Section 2.9, and Section 9.2 uses the pre-test as a covariate, which answers the same question and is usually more powerful.

7.1 One-way ANOVA: Post_score by Teaching_method

1. Choose **One-way ANOVA** (*Compare three or more groups*) in the Analyses menu.
2. In **Score to compare**, choose **Post_score**; in **Groups**, choose **Teaching_method**.

The result gives the group descriptives with Shapiro–Wilk for each group, the ANOVA table, Welch's F and Levene's test, the effect sizes η^2 and ω^2 with their confidence intervals, the pairwise comparisons, the Kruskal–Wallis test with Dunn's pairwise comparisons as the rank-based alternative, and normal Q–Q plots of each group (Figure 41).

Group descriptives

Group	n	Mean	SD	Skewness	Shapiro–Wilk W	p (normality)
Blended	129	68.34	11.97	0.09	.990	.491
Flipped	129	68.49	12.53	0.06	.994	.843
Lecture	142	61.87	14.59	−0.03	.984	.093

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965).

Copy table | Copy as picture | PNG | CSV

One-way ANOVA

Source	SS	df	MS	F	p
Between groups	3926.91	2	1963.46	11.38	< .001
Within groups	68473.40	397	172.48		
Total	72400.32	399			

Copy table | Copy as picture | PNG | CSV

Robust test and equal variances

Test	Result
Welch's F	F(2, 264.52) = 10.27, p < .001
Levene's test	F(2, 397) = 1.76, p = .173

Copy table | Copy as picture | PNG | CSV

Effect sizes

Measure	Value	90% CI	Size
η²	.054	[.022, .091]	small
ω² (less biased)	.049	[.022, .091]	small

CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). η² and ω² share one interval because both estimate the same population share of variance (Ben-Shachar et al., 2020).

Figure 41. *Group descriptives, the ANOVA table, the robust test and the effect sizes.*

What it means for our example

Post_score differs across the three groups: Blended M = 68.34 (SD = 11.97, n = 129), Flipped M = 68.49 (SD = 12.53, n = 129) and Lecture M = 61.87 (SD = 14.59, n = 142). F(2, 397) = 11.38, p < .001, η^2 = .054, ω^2 = .049, a small effect by the benchmarks. Levene's test does not suggest unequal variances, F(2, 397) = 1.76, p = .173, so the classic ANOVA and the Tukey–Kramer comparisons are used; Welch's F agrees, F(2, 264.52) = 10.27, p < .001. Shapiro–Wilk finds no departure from normality in any group (p = .491, .843 and .093), and every group has more than 30 students.

7.2 Effect sizes with confidence intervals

The table *Effect sizes* (Figure 42) gives η^2 (the share of the variance in Post_score explained by the teaching method in this sample) and ω^2 (a less biased estimate of the same share in the population), each with a confidence interval. For the F and χ^2 family

SaNaStat gives 100(1 − 2α)% intervals, 90% at α = .05: such an interval excludes zero exactly when the F test is significant (Steiger, 2004). Here η^2 = .054 and ω^2 = .049, both with the 90% CI [.022, .091] (the two share one interval because they estimate the same population share). The teaching method explains between about 2% and 9% of the differences in the post-test: a small effect, but clearly more than none.

Effect sizes

Measure	Value	90% CI	Size
η^2	.054	[.022, .091]	small
ω^2 (less biased)	.049	[.022, .091]	small

CIs of η^2, ω^2 and partial η^2 are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). η^2 and ω^2 share one interval because both estimate the same population share of variance (Ben-Shachar et al., 2020).

Figure 42. *The effect sizes of the one-way ANOVA with their 90% confidence intervals.*

Cut-off values: η^2 and ω^2

.01 small, .06 medium, .14 large (Cohen, 1988). Prefer ω^2 for the population effect; η^2 overestimates it, most with small groups.

7.3 Which groups differ? Choosing the post hoc test

Pairwise comparisons: Tukey–Kramer (chosen automatically: variances similar)

Pair	Mean difference	95% CI	p (adjusted)
Blended – Flipped	−0.15	[−4.00, 3.70]	.995
Blended – Lecture	6.47	[2.71, 10.23]	< .001
Flipped – Lecture	6.62	[2.86, 10.38]	< .001

Tukey–Kramer (Kramer, 1956) assumes equal variances and keeps the family-wise error at α; the intervals are simultaneous 95% intervals.

Figure 43. *The Tukey–Kramer pairwise comparisons.*

The Tukey–Kramer comparisons (Figure 43) give the difference between each pair of means with a simultaneous confidence interval and a p-value adjusted for the three comparisons. Blended and Flipped do not differ (difference −0.15, 95% CI [−4.00, 3.70], p = .995); both are higher than Lecture (Blended − Lecture 6.47, 95% CI [2.71, 10.23], p < .001; Flipped − Lecture 6.62, 95% CI [2.86, 10.38], p < .001). Read the confidence intervals: a difference of 6 to 7 points is the best estimate, but the data are compatible with anything from about 3 to 10.

Pairwise comparisons (post hoc tests) above the result opens the choice of method (Figure 44):

- **Automatic (recommended)**: Tukey–Kramer when Levene's test finds similar variances, otherwise Games–Howell;
- **Tukey–Kramer (Kramer, 1956)**: every pair, overall error rate kept at α; assumes similar variances;

- **Games–Howell (Games & Howell, 1976)**: like Tukey–Kramer, but safe when the variances or group sizes differ;
- **Bonferroni (Welch t-tests)** and **Holm (Welch t-tests)**: a Welch t-test for each pair, with each p multiplied by the number of pairs (Bonferroni) or with the step-down correction of Holm (1979), which gives the same protection with more power.
- **Scheffé**, **Šidák**, **Fisher's LSD (no adjustment)**, **Dunnett (each group against the first)** and **Student–Newman–Keuls**: described below.

A second choice sets the adjustment of Dunn's rank-based comparisons: **Holm (recommended)** or **Bonferroni**.

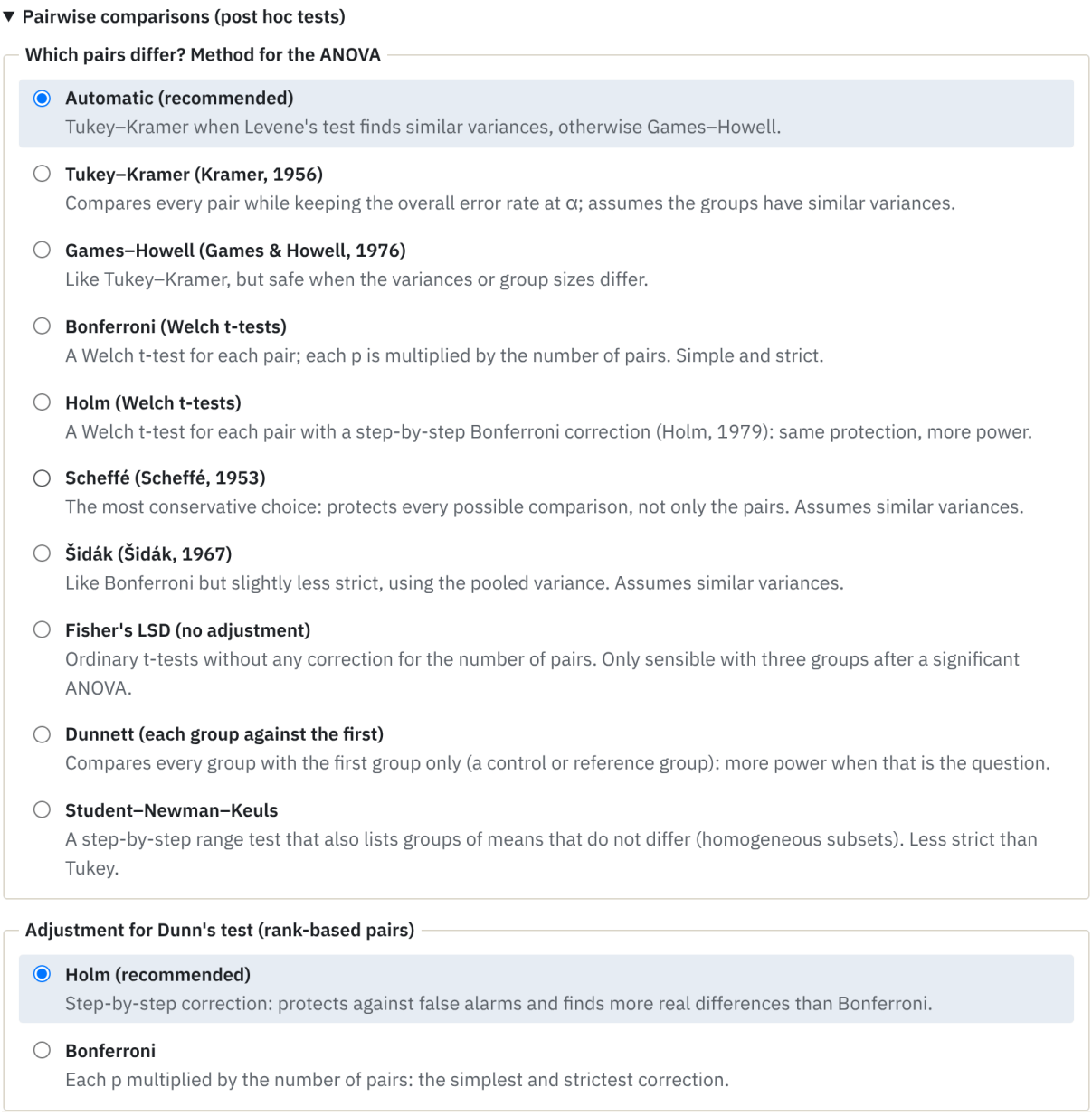


Figure 44. *The choice of post hoc test and of the adjustment for Dunn's test.*

With equal variances the methods agree. Choosing **Games–Howell (Games & Howell, 1976)** gives Blended − Lecture 6.47, 95% CI [2.66, 10.28], and Flipped − Lecture 6.62, 95% CI [2.74, 10.51], both $p < .001$, and Blended − Flipped $p = .995$. Choosing **Holm**

(Welch t-tests) gives the same conclusion (Blended − Flipped t(255.45) = −0.10, Holm p = .921; the other two pairs p < .001), with Bonferroni-adjusted 98.3% intervals, because Holm's method has no matching interval (Figure 45). Set the choice back to **Automatic (recommended)**.

Pairwise comparisons: Welch's t-tests, Holm-adjusted

Pair	Mean difference	98.3% CI (Bonferroni)	t	df	p (unadjusted)	p (Holm)
Blended – Flipped	−0.15	[−3.83, 3.53]	−0.10	255.45	.921	.921
Blended – Lecture	6.47	[2.58, 10.36]	4.01	266.28	< .001	< .001
Flipped – Lecture	6.62	[2.65, 10.59]	4.02	268.17	< .001	< .001

Holm (1979): a step-down version of Bonferroni that is never less powerful. The intervals are Bonferroni-adjusted 98.3% intervals (1 − α/3), because Holm has no matching interval. Welch's t does not assume equal variances.

Figure 45. *Pairwise comparisons with Welch's t-tests and the Holm adjustment.*

Five further methods use the pooled variance of the ANOVA (MSE = 172.48, df = 397), so they assume similar variances:

- **Scheffé (Scheffé, 1953)** protects every possible contrast, not only the pairs, and is the most conservative: Lecture − Blended −6.47, 95% CI [−10.40, −2.55], p < .001; Blended − Flipped p = .996 (Figure 46);
- **Šidák (Šidák, 1967)** is a little less strict than Bonferroni (Blended − Flipped p = 1.000);
- **Fisher's LSD (no adjustment)** uses ordinary t-tests without any correction (Blended − Flipped p = .926); use it only with three groups after a significant ANOVA;
- **Dunnett (each group against the first)** compares every group with one control group only, which gives more power when that is the question;
- **Student–Newman–Keuls** is a step-by-step range test that also lists the groups whose means do not differ.

Pairwise comparisons: Scheffé

Pair	Mean difference	95% CI	F	p (adjusted)
Lecture – Blended	−6.47	[−10.40, −2.55]	8.21	< .001
Lecture – Flipped	−6.62	[−10.55, −2.70]	8.59	< .001
Blended – Flipped	−0.15	[−4.17, 3.87]	0.00	.996

Scheffé (1953) protects every possible contrast, not only the pairs, and is therefore the most conservative test for pairs; the intervals are simultaneous. Based on the pooled within-group variance (MSE = 172.48, df = 397), so it assumes equal variances.

Figure 46. *Scheffé comparisons.*

Dunnett uses the first group as the control. To compare the two new methods with the lecture, make Lecture the first group: on the Data screen, click **Edit Teaching_method**, click **Move Lecture up** twice and close the editor. Then Dunnett gives Blended − Lecture 6.47, 95% CI [2.92, 10.02], and Flipped − Lecture 6.62, 95% CI [3.07, 10.17], both p < .001 (Figure 47).

Pairwise comparisons: Dunnett (each group against the first)

Pair	Mean difference	95% CI	t	p (adjusted)
Blended – Lecture	6.47	[2.92, 10.02]	4.05	< .001
Flipped – Lecture	6.62	[3.07, 10.17]	4.15	< .001

Dunnett (1955): every group is compared only with the first group (the control or reference), which gives more power than comparing all pairs; the intervals are simultaneous. Change the order of the groups on the Data screen to choose another control. Based on the pooled within-group variance (MSE = 172.48, df = 397), so it assumes equal variances.

Figure 47. *Dunnett: each method against the lecture.*

Student–Newman–Keuls (Figure 48) orders the means and tests each pair against the studentized range for the number of means it spans (r): Lecture differs from Blended (q = 5.73) and from Flipped (q = 5.86), both p < .001, while Blended and Flipped do not differ (q = 0.13, p = .926). The homogeneous subsets (Figure 49) say the same in one line each: subset 1 holds Lecture, subset 2 Blended and Flipped. Student–Newman–Keuls has no confidence intervals and, with four or more groups, does not keep the overall error rate at α.

Pairwise comparisons: Student–Newman–Keuls

Pair	Mean difference	r	q	p	Significant
Lecture – Blended	−6.47	2	5.73	< .001	Yes
Lecture – Flipped	−6.62	3	5.86	< .001	Yes
Blended – Flipped	−0.15	2	0.13	.926	No

Student–Newman–Keuls (Newman, 1939; Keuls, 1952): a step-down range test; r is the number of means spanned by the pair. A pair inside a range that is not significant is not significant either. It has no confidence intervals and does not keep the family-wise error at α with four or more groups.

Figure 48. *Student–Newman–Keuls comparisons.*

Homogeneous subsets (Student–Newman–Keuls)

Subset	Groups whose means do not differ
1	Lecture
2	Blended, Flipped

Groups listed in one subset do not differ significantly; the subsets are ordered from the lowest to the highest mean.

Figure 49. *Homogeneous subsets: groups whose means do not differ.*

> **Tip: decide the post hoc test in advance**
>
> Keep the automatic choice unless you have a reason: Tukey–Kramer is the most powerful when variances are equal, Games–Howell protects you when they are not. Choose the method before you look at the pairs, and never try several methods to find the one that makes a pair significant.

7.4 The rank-based alternative: Kruskal–Wallis and Dunn

The Kruskal–Wallis test compares the groups' rank distributions: H = 18.32, df = 2, p < .001, ε^2 = .046. Dunn's pairwise comparisons, adjusted by Holm's method, give the

same picture: Blended and Flipped do not differ (p = .953), and both differ from Lecture (p < .001). When the ANOVA and Kruskal–Wallis tests agree, as here, your conclusion does not depend on the assumptions.

How to report it

"A one-way ANOVA showed a significant effect of Teaching_method on Post_score, $F(2, 397) = 11.38$, $p < .001$, $\eta^2 = .054$, 90% CI [.022, .091], $\omega^2 = .049$. Tukey–Kramer comparisons showed that Blended (M = 68.34) and Flipped (M = 68.49) students scored higher than Lecture students (M = 61.87), both $p < .001$, whereas Blended and Flipped did not differ, $p = .995$."

7.5 Unequal variances: Welch's ANOVA and Games–Howell

Variances were equal in practice_students.csv. To see what SaNaStat does when they are not, open practice_hr.csv (1,200 simulated employees) and compare Annual_salary across Department. Levene's test is significant, $F(4, 1195) = 5.10$, $p < .001$, so SaNaStat switches to Welch's ANOVA and the Games–Howell comparisons (Figure 50) and says so in the check *Equal variances*.

Pairwise comparisons: Games–Howell (chosen automatically: variances differ)

Pair	Mean difference	95% CI	p (adjusted)
Finance – HR	3172.99	[−1453.77, 7799.74]	.329
Finance – IT	−4565.05	[−9160.06, 29.96]	.052
Finance – Operations	6616.43	[3077.50, 10155.37]	< .001
Finance – Sales	3839.12	[220.50, 7457.75]	.031
HR – IT	−7738.04	[−12633.35, −2842.72]	< .001
HR – Operations	3443.45	[−486.21, 7373.10]	.116
HR – Sales	666.14	[−3334.72, 4666.99]	.991
IT – Operations	11181.48	[7294.28, 15068.69]	< .001
IT – Sales	8404.17	[4444.37, 12363.98]	< .001
Operations – Sales	−2777.31	[−5424.41, −130.21]	.034

Games–Howell (Games & Howell, 1976) does not assume equal variances; the intervals are simultaneous 95% intervals.

Figure 50. *Games–Howell pairwise comparisons of Annual_salary across five departments (practice_hr.csv).*

What it means for our example

Welch's F(4, 439.57) = 18.43, p < .001. IT employees earn the most (M = 52486.00, SD = 17533.10, n = 195) and Operations the least (M = 41304.52, SD = 12491.95, n = 363). The largest difference is IT − Operations = 11181.48, 95% CI [7294.28, 15068.69], Games–Howell p < .001. The Kruskal–Wallis test agrees, H = 87.86, df = 4, p < .001. The classic ω^2 is .066 (medium), 90% CI [.046, .091]; the sentence reports the approximate ω^2 from Welch's F, .136, 90% CI [.091, .188]. Salaries are right-skewed in every department (Shapiro–Wilk p < .001, skewness 1.14 to 2.28), but with 113 or more employees per department the test of means is robust; SaNaStat notes that with this many cases Shapiro–Wilk flags even harmless departures.

Games–Howell is the post hoc test of choice when variances or group sizes differ, because it does not pool the variances (Games & Howell, 1976). When the variances are equal, Tukey–Kramer is slightly more powerful.

7.6 The gain score

The variable Gain (Post_score − Pre_score, Section 2.9) is the outcome that the planted effect was built on. Choose **One-way ANOVA**, **Gain** as **Score to compare** and **Teaching_method** as **Groups** (Figure 51). The mean gain is 7.51 points for Blended (SD = 6.03, n = 129), 8.93 for Flipped (SD = 6.11, n = 129) and 4.01 for Lecture (SD = 5.94, n = 142); F(2, 397) = 24.15, p < .001, η^2 = .108, 90% CI [.063, .155], ω^2 = .104, a medium effect. Levene's test does not suggest unequal variances, F(2, 397) = 0.08, p = .922. The Tukey–Kramer comparisons show that Flipped gains 4.92 points more than Lecture, 95% CI [3.20, 6.64], and Blended 3.50 points more, 95% CI [1.78, 5.23] (both p < .001), while Blended and Flipped do not differ, −1.42, 95% CI [−3.18, 0.35], p = .142. The effect is larger than for Post_score alone (ω^2 = .049): taking the pre-test into account removes the differences between students that were there before the course.

Group descriptives

Group	n	Mean	SD	Skewness	Shapiro–Wilk W	p (normality)
Blended	129	7.51	6.03	0.12	.990	.502
Flipped	129	8.93	6.11	−0.36	.983	.119
Lecture	142	4.01	5.94	−0.21	.989	.350

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965).

Copy table | Copy as picture | PNG | CSV

One-way ANOVA

Source	SS	df	MS	F	p
Between groups	1753.60	2	876.80	24.15	< .001
Within groups	14416.26	397	36.31		
Total	16169.86	399			

Copy table | Copy as picture | PNG | CSV

Robust test and equal variances

Test	Result
Welch's F	F(2, 263.04) = 24.24, p < .001
Levene's test	F(2, 397) = 0.08, p = .922

Copy table | Copy as picture | PNG | CSV

Effect sizes

Measure	Value	90% CI	Size
η²	.108	[.063, .155]	medium
ω² (less biased)	.104	[.063, .155]	medium

CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). η² and ω² share one interval because both estimate the same population share of variance (Ben-Shachar et al., 2020).

Figure 51. *The one-way ANOVA of Gain by Teaching_method.*

How to report it

"A one-way ANOVA showed a significant effect of Teaching_method on Gain, F(2, 397) = 24.15, p < .001, η² = .108, 90% CI [.063, .155], ω² = .104."

Exercise: ANOVA on three practice files

(a) practice_marketing.csv: does Satisfaction_1to10 differ by Main_channel? Answer: no, F(2, 1412) = 1.55, p = .213, η² = .002, 90% CI [.000, .007], ω² = .001. (b) practice_hr.csv: does Engagement differ by Department? Answer: no, F(4, 1195) = 0.28, p = .891. (c) practice_engagement.csv: build the scale Dedication (Section 1.7, exercise) and compare it across Department. Answer: it differs, F(4, 595) = 10.84, p < .001, η² = .068, 90% CI [.035, .097], ω² = .062; HR has the highest mean (4.67) and Operations the lowest (3.57). Which pairs differ in the Tukey–Kramer table? When a result is not significant, what does the confidence interval of the difference tell you?

8. The same people measured repeatedly

In practice_students.csv each student sat the same test three times: Score_T1, Score_T2 and Score_T3. The scores of one student are correlated, so they are not independent groups. The one-way repeated-measures ANOVA compares the three means while using each student as their own control.

Concept: sphericity

Repeated-measures ANOVA assumes **sphericity**: the variances of the differences between every pair of measurements are equal. If it fails, the F test is too liberal. **Mauchly's test** (Mauchly, 1940) checks it, and the **epsilon (ε)** of Greenhouse–Geisser (1959) or Huynh–Feldt (1976) measures how far it fails (1 = perfect; lower is worse). The corrections multiply the degrees of freedom by ε. With only two measurements sphericity is always met.

8.1 Run the analysis

1. Choose **Repeated-measures ANOVA** (*Same people, three or more times*).
2. Tick **Score_T1**, **Score_T2** and **Score_T3** under *Measurements*. SaNaStat orders them as they appear in the file, which should be the order of measurement. Respondents with any missing measurement are left out (here nobody: n = 400).

The result gives the descriptives of the complete cases, the repeated-measures ANOVA with the correction and the partial η^2 with its interval, Mauchly's test, pairwise paired t-tests (Holm-adjusted) with d_z and its interval, and the Friedman test with Wilcoxon pairwise tests.

Descriptives (complete cases)

Measure	n	Mean	SD
Score_T1	400	59.37	11.65
Score_T2	400	62.55	12.66
Score_T3	400	66.09	13.47

Copy table | Copy as picture | PNG | CSV

Repeated-measures ANOVA

Correction	df	F	p	ε	Partial η²	90% CI
None (sphericity assumed)	2, 798	346.88	< .001	1	.465	[.425, .500]
Greenhouse–Geisser	1.53, 610.27	346.88	< .001	.765	.465	[.419, .505]
Huynh–Feldt	1.53, 612.14	346.88	< .001	.767	.465	[.420, .505]

Partial η² = .465; generalised η² = .045 (Bakeman, 2005; no interval is given for it). CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004). Each row's interval uses that row's degrees of freedom; the APA sentence uses the Huynh–Feldt-corrected row.

Copy table | Copy as picture | PNG | CSV

Mauchly's test of sphericity

W	χ²	df	p
0.692	146.32	2	< .001

Figure 52. *The descriptives, the repeated-measures ANOVA and Mauchly's test.*

What it means for our example

The mean score rises from 59.37 (SD = 11.65) at T1 through 62.55 (SD = 12.66) at T2 to 66.09 (SD = 13.47) at T3. Mauchly's test is significant, W = 0.692, $\chi^2(2) = 146.32$, $p < .001$, and ε is .765 (Greenhouse–Geisser) and .767 (Huynh–Feldt). Because ε ≥ .75, SaNaStat uses the Huynh–Feldt correction: $F(1.53, 612.14) = 346.88$, $p < .001$, partial $\eta^2 = .465$, 90% CI [.420, .505], generalised $\eta^2 = .045$.

Note: partial and generalised η²

Partial $\eta^2 = .465$ is large, but SaNaStat labels the effect "small" because it applies the benchmarks to the generalised η^2 of .045. The generalised η^2 keeps the differences between students in its denominator and is therefore always much smaller than the partial η^2 in a within-subject design. It is meant for comparing designs (Bakeman, 2005); SaNaStat gives no interval for it. Report both, and judge the size of the change from the pairwise effect sizes below.

8.2 Where is the change? Pairwise comparisons

Each pair of time points differs: T1 − T2 = −3.18, $t(399) = -14.94$, $d_z = -0.75$, 95% CI [−0.86, −0.64]; T1 − T3 = −6.72, $t(399) = -21.12$, $d_z = -1.06$, 95% CI [−1.18, −0.93]; T2 − T3 = −3.54, $t(399) = -16.00$, $d_z = -0.80$, 95% CI [−0.91, −0.69]; all Holm-adjusted p

< .001. The negative sign only shows that the earlier measurement is lower. The scores rise steadily, by about three to four points between occasions.

8.3 The rank-based alternative: Friedman

The Friedman test ranks each student's three scores and compares the mean ranks: $\chi^2(2) = 327.05$, $p < .001$, Kendall's W = .409, with all three Wilcoxon pairwise tests significant after Holm's adjustment (r between −.72 and −.86). It leads to the same conclusion, so the result does not depend on the sphericity or normality assumptions.

How to report it

"A one-way repeated-measures ANOVA with the Huynh–Feldt correction (ε = .767) showed a significant difference across Score_T1, Score_T2, Score_T3, $F(1.53, 612.14) = 346.88$, $p < .001$, partial $\eta^2 = .465$, 90% CI [.420, .505], generalised $\eta^2 = .045$. Holm-adjusted paired t-tests showed that every measurement was higher than the one before (all $p < .001$)."

Note: what this screen cannot do

SaNaStat offers the one-way repeated-measures ANOVA only. A design that combines repeated measures with a between-group factor (a mixed ANOVA, for example the three scores by teaching method) is not available; for such designs compare the groups at each occasion with the tests of chapter 7, or analyse the gain (Section 7.6).

Exercise: repeated measures in practice_hr.csv

Open practice_hr.csv and tick Engagement_Q1 to Engagement_Q4. Answer: the quarters differ, $F(3, 3597) = 15.79$, $p < .001$, partial $\eta^2 = .013$, 90% CI [.007, .019], generalised $\eta^2 = .002$, which SaNaStat calls negligible. With 1,200 employees even a tiny change is significant, so read the effect size, not the p-value, to judge whether the change matters. Which pairs of quarters differ?

9. Factorial ANOVA, ANCOVA and MANOVA

Real outcomes usually depend on more than one thing. This chapter covers three extensions of the ANOVA: two or more grouping factors (factorial ANOVA), a covariate (ANCOVA), and two or more outcomes (MANOVA). All three are under **Factorial ANOVA, ANCOVA, MANOVA** (*Several factors, covariates or outcomes*).

Concept: main effects and interaction

A **main effect** is the effect of one factor averaged over the levels of the other. An **interaction** means that the effect of one factor depends on the level of the other (for example, the flipped classroom helps women but not men). If there is an interaction, interpret the cell means, not the main effects alone. SaNaStat reports Type III sums of squares with sum-to-zero coding, so each effect is tested after all the others, which suits groups of unequal size, and gives partial η^2 (the share of the variance not explained by other effects that the effect explains) with its confidence interval.

9.1 Factorial ANOVA: Exam_score by Teaching_method and Gender

1. Under *What is your design?* choose **Factorial ANOVA / ANCOVA**.
2. As **Score (outcome)** choose **Exam_score**, and tick **Teaching_method** and **Gender** under **Grouping factors**.
3. Keep **Full factorial model (Include the interaction; usual)**. **Main-effects model** leaves the interaction out and is for very small cells.

Cell descriptives

Cell	n	Mean	SD
female · Blended	78	59.01	9.41
female · Flipped	72	56.34	8.32
female · Lecture	89	54.61	10.43
male · Blended	46	52.85	9.51
male · Flipped	50	53.84	9.91
male · Lecture	52	50.56	9.17

Copy table | Copy as picture | PNG | CSV

Factorial ANOVA (Type III sums of squares)

Source	SS	df	MS	F	p	Partial η²	90% CI
Gender	1633.48	1	1633.48	18.06	< .001	.045	[.017, .084]
Teaching_method	757.63	2	378.81	4.19	.016	.022	[.002, .048]
Gender × Teaching_method	195.86	2	97.93	1.08	.340	.006	[.000, .021]
Error	34460.78	381	90.45				
Total (corrected)	37159.76	386					

R^2 = .073, adjusted R^2 = .060, n = 387. CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004).

Figure 53. *The cell descriptives and the ANOVA table with partial η².*

The result (Figure 53) checks that the variances are equal across the six cells (Levene's test on the medians) and that the cells are large enough, and then gives the cell descriptives, the ANOVA table, and for each factor the estimated marginal means and Holm-adjusted pairwise comparisons.

What it means for our example

Levene's test does not suggest unequal variances across the six cells, F(5, 381) = 1.03, p = .401, and the smallest cell has 46 students. There is a significant main effect of Gender, F(1, 381) = 18.06, p < .001, partial η^2 = .045, 90% CI [.017, .084], and of Teaching_method, F(2, 381) = 4.19, p = .016, partial η^2 = .022, 90% CI [.002, .048], but no interaction, F(2, 381) = 1.08, p = .340, partial η^2 = .006, 90% CI [.000, .021]: the advantage of women is about the same under all three methods (the cell means are 59.01, 56.34 and 54.61 for women and 52.85, 53.84 and 50.56 for men under Blended, Flipped and Lecture). The model explains 7.3% of the variance (R^2 = .073, adjusted R^2 = .060, n = 387 complete cases).

The **estimated marginal means** are the group means after allowing for the other factor: 56.65 for women (95% CI [55.44, 57.87]) and 52.42 for men (95% CI [50.88, 53.96]), a difference of 4.24 (SE = 0.997), t(381) = 4.25, p < .001 (Holm). For teaching methods the marginal means are Blended 55.93, Flipped 55.09 and Lecture 52.59; only Blended − Lecture differs (3.34, p = .018), while Blended − Flipped (p = .502) and Flipped − Lecture (p = .077) do not.

How to report it

"A 2 (Gender) × 3 (Teaching method) factorial ANOVA on Exam_score showed a significant effect of Gender, $F(1, 381) = 18.06$, $p < .001$, partial $\eta^2 = .045$, 90% CI [.017, .084]; a significant effect of Teaching_method, $F(2, 381) = 4.19$, $p = .016$, partial $\eta^2 = .022$, 90% CI [.002, .048]; no significant effect of Gender × Teaching_method, $F(2, 381) = 1.08$, $p = .340$, partial $\eta^2 = .006$, 90% CI [.000, .021]."

9.2 ANCOVA: adding the pre-test as a covariate

Tick **Pre_score** under **Covariates (optional: makes it an ANCOVA)** and the title changes to *ANCOVA*.

What it means for our example

Pre_score is a strong covariate, $F(1, 380) = 201.33$, $p < .001$, partial $\eta^2 = .346$, 90% CI [.285, .402], and the error sum of squares falls from 34460.78 to 22526.27 (R^2 rises from .073 to .394). Gender becomes more important, $F(1, 380) = 31.64$, $p < .001$, partial $\eta^2 = .077$, 90% CI [.039, .122]; the adjusted means are 56.74 for women and 52.20 for men (difference 4.54, SE = 0.807, $t(380) = 5.63$, $p < .001$). Teaching_method is no longer significant, $F(2, 380) = 2.62$, $p = .074$, partial $\eta^2 = .014$, 90% CI [.000, .036], and neither is the interaction, $F(2, 380) = 2.82$, $p = .061$, partial $\eta^2 = .015$.

For the post-test, the outcome the teaching methods were meant to change, the ANCOVA with the pre-test as covariate gives a clear and medium-sized effect of the method, $F(2, 396) = 23.82$, $p < .001$, partial $\eta^2 = .107$, 90% CI [.062, .154], with adjusted means of 66.87 (Blended), 68.29 (Flipped) and 63.38 (Lecture): the same picture as the gain score of Section 7.6.

The exam score is a single test, taken after the course, and the pre-test explains most of why some students score higher; after that, the teaching method adds only a little (the comparison of chapter 7, which used the post-test, found a clearer effect). Covariates change what a factor's p-value answers: it is now the effect of the method *for students with the same pre-test score*. Do not choose between a factorial ANOVA and an ANCOVA by which one gives the p-value you prefer: decide the model before you look.

Cut-off values: ANCOVA assumptions

The covariate must be measured *before* the groups were treated (the pre-test is), it should be related to the outcome, and the relationship must have the same slope in every group (no covariate × group interaction). SaNaStat does not test equal slopes: check it with the scatter plot of Post_score against Pre_score grouped by Teaching_method (Section 4.6) and look for roughly parallel lines (Maxwell et al., 2018).

9.3 One-way MANOVA: Exam_score and Project_score together

Students have two outcomes, Exam_score and Project_score. A **multivariate analysis of variance (MANOVA)** tests whether the groups differ on the outcomes *taken together*, which controls the error rate across the outcomes and uses their correlation.

1. Under *What is your design?* choose **One-way MANOVA (Several scores at once, one grouping factor)**.
2. Tick **Exam_score** and **Project_score** under **Outcomes**, and choose **Teaching_method** as **Grouping factor**.

Multivariate tests

Test	Value	F	df1	df2	p
Pillai's trace	0.126	12.91	4	768	< .001
Wilks' lambda	0.876	13.09	4	766	< .001
Hotelling–Lawley trace	0.139	13.27	4	764	< .001
Roy's largest root	0.118	22.74	2	384	< .001

n = 387 complete cases. Roy's largest root gives an upper bound for F.

Copy table | Copy as picture | PNG | CSV

Follow-up: each outcome separately

Outcome	F	df	p	Partial η²	90% CI
Exam_score	4.65	2, 384	.010	.024	[.003, .051]
Project_score	21.03	2, 384	< .001	.099	[.054, .145]

Consider a Bonferroni-adjusted α (α divided by the number of outcomes) for these follow-up tests. CIs of η², ω² and partial η² are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral F distribution. This level matches the F test: the lower limit is above 0 exactly when the test is significant (Steiger, 2004).

Figure 54. *The multivariate tests and the follow-up ANOVAs.*

What it means for our example

The three teaching methods differ on the two outcomes taken together: Pillai's trace V = .126, F(4, 768) = 12.91, p < .001 (n = 387 complete cases; Wilks' lambda = .876, F(4, 766) = 13.09, p < .001; the other two tests agree). The follow-up ANOVAs (Figure 54) show an effect on both outcomes, but a much larger one on Project_score, F(2, 384) = 21.03, p < .001, partial η² = .099, 90% CI [.054, .145], than on Exam_score, F(2, 384) = 4.65, p = .010, partial η² = .024, 90% CI [.003, .051].

Cut-off values: MANOVA

Report Pillai's trace: it is the most robust of the four multivariate tests when groups are unequal or the assumptions are doubtful (Olson, 1976). SaNaStat suggests a Bonferroni-adjusted α for the follow-up tests, the α divided by the number of outcomes: here .05 / 2 = .025, and both follow-up p-values (.010 and < .001) are below it. SaNaStat does not test the equality of the covariance matrices; with groups of similar size (124, 122 and 141 complete cases) the MANOVA is reasonably robust.

How to report it

"A one-way MANOVA showed a significant multivariate effect of Teaching method on Exam_score and Project_score, Pillai's V = .126, F(4, 768) = 12.91, p < .001. Follow-up ANOVAs (α = .025) showed significant effects on Exam_score, F(2, 384) = 4.65, p = .010, partial η^2 = .024, and on Project_score, F(2, 384) = 21.03, p < .001, partial η^2 = .099."

Exercise: factorial designs

In practice_students.csv, run a factorial ANOVA of Project_score by Teaching_method and Gender. Is there an interaction? Then add Pre_score as a covariate: does the effect of Teaching_method change? Compare the cell means with the main effects before you decide how to describe the result. In practice_hr.csv you can run a factorial ANOVA of Annual_salary by Gender and Department; the file note says that women earn about 4% less after controls, so look at Gender with Department in the model.

10. Categories and proportions

This chapter covers variables that are counts of categories rather than scores: how often each category occurs (frequencies), whether two categorical variables are related (the chi-square test of independence), whether the categories of one variable are equally common (the goodness-of-fit test) and whether a yes/no answer changed in the same people (the McNemar test).

Concept: expected counts and the chi-square statistic

A chi-square test compares the counts you observed in a table with the counts you would *expect* if the variables were unrelated. The expected count of a cell is its row total times its column total divided by N. The larger the gap between observed and expected, the larger χ^2, and the smaller the p-value. As with every test, p says nothing about the size of the association: report an effect size, Cramér's V (φ for a 2 × 2 table), with its confidence interval.

10.1 Frequencies of one categorical variable

Frequencies are on the **Descriptive statistics** screen. Tick **Programme** and **Residence** under *Categorical variables (frequencies and percentages)*. For each variable SaNaStat gives a table with the count (n), the percentage of all respondents, the *valid* percentage (of those who answered) and the cumulative percentage, a bar chart of the valid percentages and a sentence for each variable (Figure 55).

Frequencies of Programme

Category	n	Percent	Valid percent	Cumulative percent
Arts	92	23.0	23.0	23.0
Business	163	40.8	40.8	63.8
Science	145	36.3	36.3	100.0
Missing	0	0.0		
Total	400	100.0		

Percent: of all respondents. Valid percent and cumulative percent: of those with an answer.

Copy table | Copy as picture | PNG | CSV

Frequencies of Residence

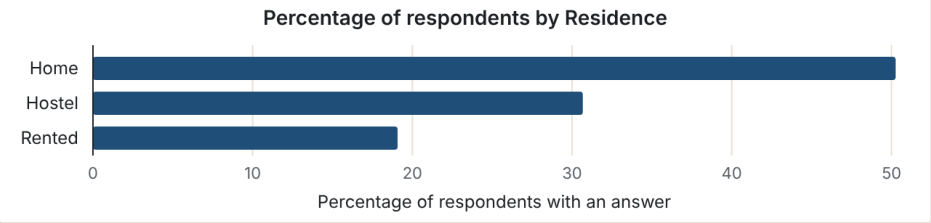
Category	n	Percent	Valid percent	Cumulative percent
Home	195	48.8	50.3	50.3
Hostel	119	29.8	30.7	80.9
Rented	74	18.5	19.1	100.0
Missing	12	3.0		
Total	400	100.0		

Percent: of all respondents. Valid percent and cumulative percent: of those with an answer.

Figure 55. *Frequency tables of Programme and Residence.*

What it means for our example

Programme has no missing values: Business is the largest programme, 163 of 400 students (40.8%), followed by Science (145, 36.3%) and Arts (92, 23.0%). Residence is missing for 12 students, so its percentages and valid percentages differ: 195 students live at home, 48.8% of all 400 but 50.3% of the 388 who answered; 119 (30.7% valid) live in a hostel and 74 (19.1% valid) in rented rooms. Report valid percentages, and say how many did not answer.



Percentage of respondents in each category of Residence. n = 388.

Figure 56. *The bar chart of the valid percentages of Residence.*

How to report it

"Of the 388 respondents who answered Residence, 50.3% were Home (n = 195), 30.7% Hostel (n = 119), and 19.1% Rented (n = 74); 12 gave no answer."

10.2 Chi-square test of independence: Gender and Programme

Choose **Chi-square, Fisher, McNemar** (*Categories and proportions*) in the Analyses menu. The question **What do you want to know?** offers the three tests of this chapter.

1. Choose **Chi-square test of independence (Are two categorical variables related?)**.
2. In **Categorical variable** (the rows), choose **Gender**; in **Second categorical variable** (the columns), choose **Programme** (Figure 57).

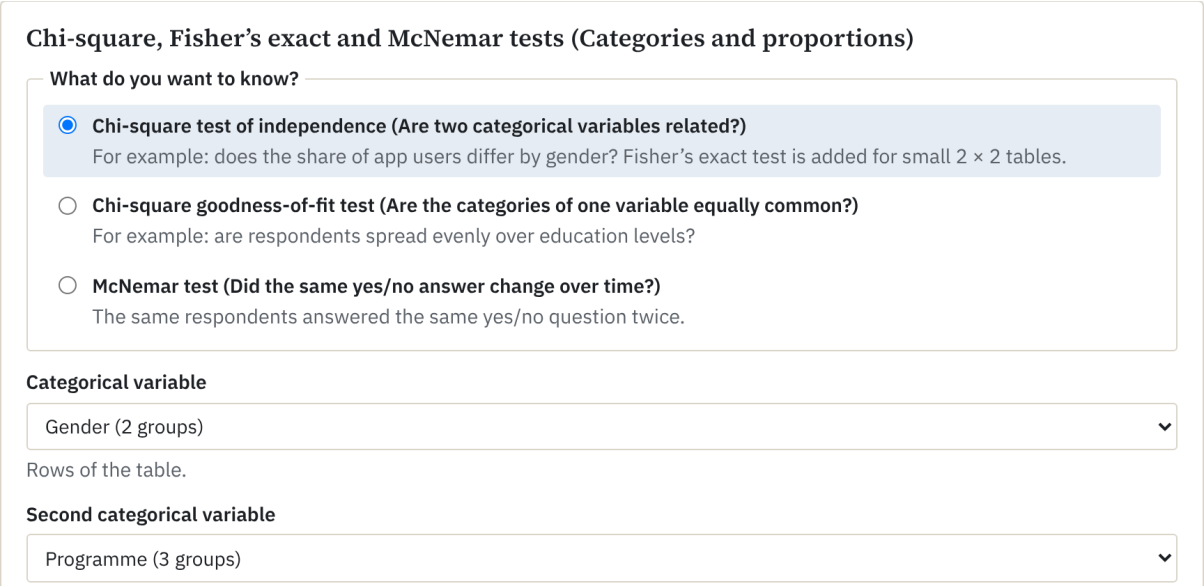


Figure 57. *The chi-square test of independence: Gender by Programme.*

The check on the expected counts comes first: no expected count should be below 1 and no more than 20% should be below 5 (Cochran, 1954). Here the smallest expected count is 35.42, so the test is trustworthy. The tables (Figure 58) show the counts with row percentages, Pearson's chi-square and the likelihood-ratio test, and the effect size.

Counts (row %)

Gender	Arts	Business	Science	Total
female	66 (27%)	91 (37%)	89 (36%)	246
male	26 (17%)	72 (47%)	56 (36%)	154

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	6.29	2	.043
Likelihood ratio	6.43	2	.040

Figure 58. *The counts with row percentages, and the tests.*

Row percentages answer the question "what share of each row falls in each column?". Of the 246 women, 27% study Arts, against 17% of the 154 men. The adjusted standardised residuals (Figure 59) show where the pattern comes from: a value beyond ±1.96 marks a cell with clearly more (positive) or fewer (negative) cases than expected. Women are over-represented in Arts (2.30) and men under-represented (−2.30).

Adjusted standardised residuals

Gender	Arts	Business	Science
female	2.30 *	−1.93	−0.04
male	−2.30 *	1.93	0.04

* beyond ±1.96: this cell holds clearly more (positive) or fewer (negative) cases than expected.

Copy table | Copy as picture | PNG | CSV

Effect size

Measure	Value	90% CI	Size
Cramér's V	.125	[.016, .200]	small

CIs of Cramér's V, φ and Cohen's w are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral χ^2 distribution; the lower limit is above 0 exactly when the χ^2 test is significant (Steiger, 2004).

Figure 59. *The adjusted standardised residuals and the effect size with its confidence interval.*

What it means for our example

Gender and Programme are related, $\chi^2(2, N = 400) = 6.29$, $p = .043$ (likelihood ratio 6.43, $p = .040$), but weakly: Cramér's V = .125, 90% CI [.016, .200], a small effect. The interval of V is a 90% interval (100(1 − 2α)%), so its lower limit is above zero exactly when the test is significant at α = .05 (Steiger, 2004). The relationship is carried by the Arts programme, which has proportionally more women than the other two. Business comes close to the cut-off (women −1.93, men 1.93) and Science does not differ from the expected counts at all (−0.04 and 0.04).

How to report it

"A chi-square test of independence showed a significant association between Gender and Programme, $\chi^2(2, N = 400) = 6.29$, $p = .043$, Cramér's V = .13, 90% CI [.02, .20]."

10.3 Teaching_method and Passed: reading row percentages as rates

Change **Categorical variable** to **Teaching_method** and **Second categorical variable** to **Passed** (Figure 60). Now the row percentages are pass rates: 78% of the 129 students taught in the blended format passed, 71% of the 129 in the flipped format and 60% of the 142 in the lecture format. The association is significant, $\chi^2(2, N = 400) = 11.13$, $p = .004$, with Cramér's V = .167, 90% CI [.073, .243] (small). The residuals show that the lecture group has more students who failed than expected (3.11) and the blended group fewer (−2.64); the flipped group is close to the expected counts (−0.54).

Counts (row %)

Teaching_method	No	Yes	Total
Blended	28 (22%)	101 (78%)	129
Flipped	37 (29%)	92 (71%)	129
Lecture	57 (40%)	85 (60%)	142

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	df	p
Pearson chi-square	11.13	2	.004
Likelihood ratio	11.15	2	.004

Copy table | Copy as picture | PNG | CSV

Adjusted standardised residuals

Teaching_method	No	Yes
Blended	−2.64 *	2.64 *
Flipped	−0.54	0.54
Lecture	3.11 *	−3.11 *

* beyond ±1.96: this cell holds clearly more (positive) or fewer (negative) cases than expected.

Figure 60. *Counts and residuals for Teaching_method by Passed.*

Put the grouping variable in the rows and the outcome in the columns, so that the row percentages are the rates you want to compare.

How to report it

"A chi-square test of independence showed a significant association between Teaching_method and Passed, $\chi^2(2, N = 400) = 11.13$, $p = .004$, Cramér's V = .17, 90% CI [.07, .24]. The pass rate was 78% for the blended, 71% for the flipped and 60% for the lecture format."

10.4 Two-by-two tables: Gender and Passed

With two rows and two columns SaNaStat adds the continuity-corrected (Yates) chi-square, Fisher's exact test and φ. Of the 246 women, 76% passed, against 58% of the 154 men. Pearson's chi-square is $\chi^2(1, N = 400) = 14.45$, $p < .001$; the Yates correction gives 13.61 and Fisher's exact test is also below .001; Cramér's V = .190, 90% CI [.108, .272] (φ = −.190: the sign shows the direction of the 2 × 2 association). All versions agree, as they do whenever the expected counts are large (the smallest is 46.97). Fisher's exact test matters when the expected counts are small, for example below 5 in one cell.

10.5 Goodness of fit: are the programmes equally common?

1. Choose **Chi-square goodness-of-fit test (Are the categories of one variable equally common?)**.
2. In **Categorical variable**, choose **Programme** (Figure 61).

Chi-square goodness-of-fit test (Are the Programme categories equally common?)

What it means: The Programme categories are not equally common (w = 0.23, small).

OK **Expected counts.** All expected counts are 5 or more.
Rule: Chi-square: no expected count below 1 and no more than 20% below 5; otherwise use the exact test (Cochran (1954)).

Observed and expected counts

Category	Observed	Expected
Arts	92	133.3
Business	163	133.3
Science	145	133.3

Copy table | Copy as picture | PNG | CSV

Test

χ^2	df	p	Cohen's w	90% CI	Size
20.44	2	< .001	0.23	[0.14, 0.30]	small

Cohen's w = √(χ^2/N); about .10 small, .30 medium, .50 large (Cohen, 1988). CIs of Cramér's V, φ and Cohen's w are two-sided 90% intervals (100(1 − 2α)% with α = .05), from inverting the noncentral χ^2 distribution; the lower limit is above 0 exactly when the χ^2 test is significant (Steiger, 2004).

Figure 61. *The goodness-of-fit test for Programme.*

The test compares the observed counts (Arts 92, Business 163, Science 145) with equal shares, 133.3 each. The categories are not equally common, $\chi^2(2, N = 400) = 20.44$, $p < .001$, with Cohen's w = 0.23, 90% CI [0.14, 0.30], small by Cohen's benchmarks of .10, .30 and .50 (w can exceed 1 when there are many categories, so it keeps its leading zero). SaNaStat tests against equal proportions only; to test against other known proportions, use the observed and expected counts to compute the statistic by hand.

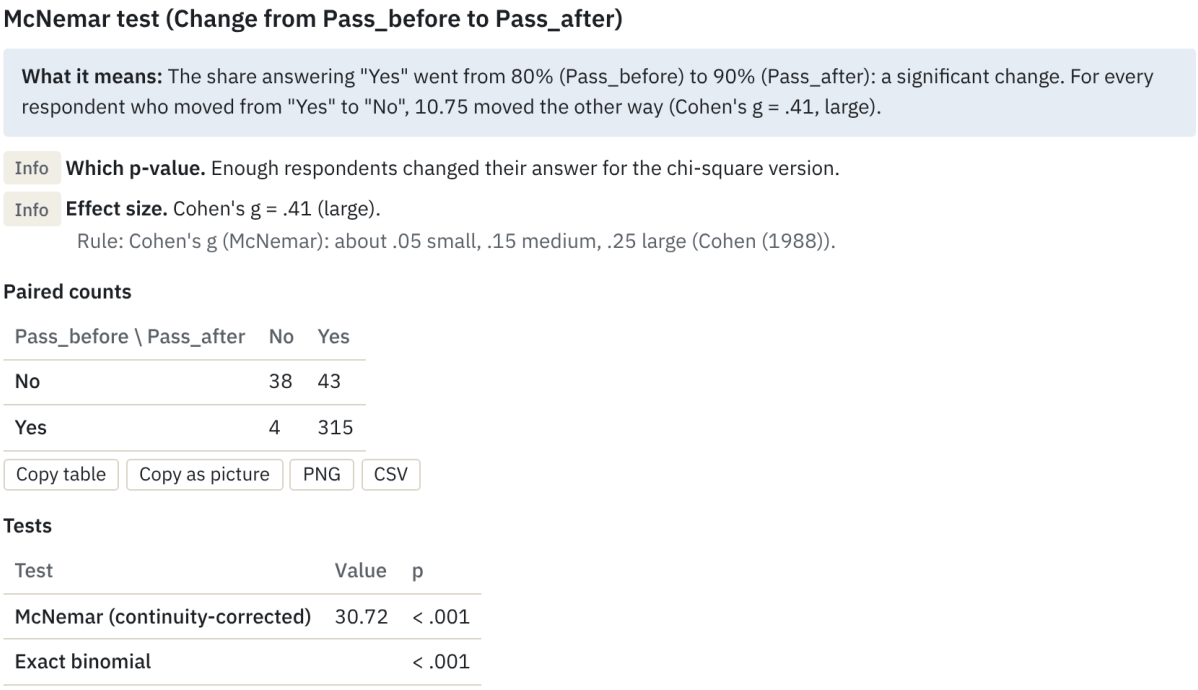
How to report it

"A chi-square goodness-of-fit test against equal proportions was significant, $\chi^2(2, N = 400) = 20.44$, $p < .001$, Cohen's w = 0.23, 90% CI [0.14, 0.30]."

10.6 McNemar test: did the answer change?

In practice_students.csv, Pass_before and Pass_after are the pass/fail versions (pass mark 50) of Pre_score and Post_score, so they describe the same 400 students before and after the course. A chi-square test of independence would be wrong here because the two answers come from the same people.

1. Choose **McNemar test (Did the same yes/no answer change over time?)**.
2. In **First time (yes/no)**, choose **Pass_before**; in **Second time (yes/no)**, choose **Pass_after**.



McNemar test (Change from Pass_before to Pass_after)

What it means: The share answering "Yes" went from 80% (Pass_before) to 90% (Pass_after): a significant change. For every respondent who moved from "Yes" to "No", 10.75 moved the other way (Cohen's g = .41, large).

Info **Which p-value.** Enough respondents changed their answer for the chi-square version.

Info **Effect size.** Cohen's g = .41 (large).
Rule: Cohen's g (McNemar): about .05 small, .15 medium, .25 large (Cohen (1988)).

Paired counts

Pass_before \ Pass_after	No	Yes
No	38	43
Yes	4	315

Copy table | Copy as picture | PNG | CSV

Tests

Test	Value	p
McNemar (continuity-corrected)	30.72	< .001
Exact binomial		< .001

Figure 62. *The finding, the paired counts and the tests.*

Only the students who changed their answer carry information. Of the 400, 43 changed from No to Yes and 4 from Yes to No; 38 said No and 315 said Yes both times. The share answering Yes went from 80% to 90%. Enough students changed for the chi-square version, $\chi^2(1) = 30.72$, $p < .001$ (continuity-corrected); the exact binomial test also gives p

< .001. SaNaStat says which p-value to use: the exact test is the one to report when only a few respondents changed. Two effect sizes describe the change: the odds ratio b / c = 43 / 4 = 10.75, 95% CI [3.86, 29.95] (for every student who moved from Yes to No, 10.75 moved the other way), and Cohen's g = .41, 95% CI [.30, .47], large by Cohen's benchmarks of .05, .15 and .25 (Cohen, 1988).

How to report it

"A McNemar test showed a significant change in the share answering Yes, from 80% to 90%, $\chi^2(1) = 30.72$, $p < .001$, odds ratio = 10.75, 95% CI [3.86, 29.95], Cohen's g = .41, 95% CI [.30, .47]."

Exercise: categories in the other practice files

(a) In practice_marketing.csv, test Campaign_group (rows) against Responded. Answer: 44% of the 738 customers who received the new offer responded, against 26% of the 762 who received the standard offer, $\chi^2(1, N = 1500) = 54.50$, $p < .001$, Cramér's V = .19, 90% CI [.15, .23]. (b) Test Churned against Loyalty_member. Answer: 46% of the 1,250 customers who did not churn were loyalty members, against 26% of the 250 who did, $\chi^2(1, N = 1500) = 35.64$, $p < .001$, V = .15. (c) In practice_hr.csv, test Department against Left_company. Answer: no significant association, $\chi^2(4, N = 1200) = 8.42$, $p = .077$, V = .08, although the residual for Operations (2.80) is beyond ±1.96: an example of a cell that should not be interpreted when the whole table is not significant.

10.7 A third variable: layered tables and the Mantel–Haenszel test

A relationship between two categorical variables can come from a third one. **Layer (optional: control for a third variable)** repeats the cross-table within each group of the third variable (Figure 63). For two yes/no variables it also gives the Mantel–Haenszel common odds ratio and tests whether the odds ratio is the same in every group.

1. Choose **Chi-square test of independence**, **Gender** in rows and **Passed** in columns.
2. Under **Layer**, choose **Teaching_method (3 groups)**.

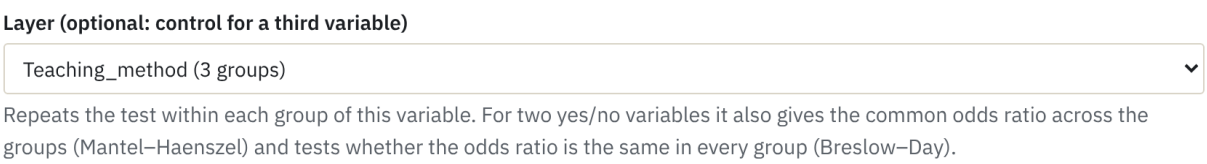


Figure 63. *A layer variable.*

Within each teaching method, gender and passing are related (Figure 64): Lecture $\chi^2(1) = 4.74$, $p = .029$; Blended $\chi^2(1) = 6.08$, $p = .014$; Flipped $\chi^2(1) = 4.73$, $p = .030$.

Gender × Passed in each group of Teaching_method

Teaching_method	n	χ^2	df	p
Lecture	142	4.74	1	.029
Blended	129	6.08	1	.014
Flipped	129	4.73	1	.030

Pearson chi-square test of independence within each group of Teaching_method; '–' where a group has fewer than two rows or columns with respondents.

Figure 64. *Gender × Passed in each teaching method.*

The common odds ratio (Figure 65) summarises the three tables: the odds of failing (Passed = No) for women are 0.418 times those for men (95% CI [0.268, 0.651]), allowing for the teaching method (Mantel & Haenszel, 1959). The Mantel–Haenszel test confirms the relationship, $\chi^2(1) = 15.12$, $p < .001$. The Breslow–Day test finds no difference between the odds ratios of the three groups ($\chi^2(2) = 0.26$, $p = .878$; Breslow & Day, 1980), so one common odds ratio describes them well.

Common odds ratio and risk ratio across Teaching_method (Mantel–Haenszel)

Measure	Value	95% CI
Common odds ratio	0.418	[0.268, 0.651]
Common risk ratio	0.563	[0.421, 0.751]
Odds ratio, Teaching_method = Lecture	0.463	[0.230, 0.931]
Odds ratio, Teaching_method = Blended	0.348	[0.148, 0.820]
Odds ratio, Teaching_method = Flipped	0.426	[0.196, 0.927]

Odds ratio: the odds of Passed = No for Gender = female relative to Gender = male; risk ratio likewise for the risk. Variances by Robins, Breslow and Greenland (1986) and Greenland and Robins (1985). The odds ratio of a layer with a zero cell uses 0.5 added to every cell.

Copy table | Copy as picture | PNG | CSV

Tests across the layers

Test	χ^2	df	p
Mantel–Haenszel (conditional independence)	15.12	1	< .001
Mantel–Haenszel, with continuity correction	14.26	1	< .001
Breslow–Day (equal odds ratios)	0.26	2	.878
Tarone's adjustment of Breslow–Day	0.26	2	.878

Mantel–Haenszel: is there a relationship after allowing for the layers? Breslow–Day and Tarone: is the odds ratio the same in every layer?

Figure 65. *The common odds ratio and the tests across the layers.*

Cut-off values: layered tables

When Breslow–Day (or Tarone's adjustment) gives $p < \alpha$, the odds ratio differs between the groups: report the odds ratio of each group rather than the common one alone (Breslow & Day, 1980).

How to report it

"A Mantel–Haenszel test, stratified by teaching method, showed a significant association between gender and passing, $\chi^2_MH(1) = 15.12$, $p < .001$; common OR = 0.42, 95% CI [0.27, 0.65]. The Breslow–Day test did not indicate heterogeneity of the odds ratios, $\chi^2(2) = 0.26$, $p = .878$."

11. Correlation and regression

This chapter asks how variables go together and which of them predict an outcome. Correlation describes the strength and direction of a relationship; regression predicts an outcome from one or more predictors and says how much each contributes. Both are under **Correlation and regression** (*How variables go together and predict an outcome*).

Concept: correlation is not causation

A correlation of r = .27 between study hours and exam scores says that students who study more tend to score higher. It does not say that studying more *causes* higher scores: a third variable (ability, motivation) can produce the same pattern, and the direction could be reversed. Correlation and regression describe association; causal claims need a design that supports them (such as the random assignment of teaching methods).

11.1 Pearson correlation

1. Choose **Bivariate and partial correlation (How strongly do variables go together?)**.
2. Tick **Study_hours_week**, **Sleep_hours**, **Anxiety_score** and **Exam_score** under *Variables*, and keep **Pearson product-moment correlation (Pearson r)**.

Correlation matrix

	1	2	3	4
1. Study_hours_week	—			
2. Sleep_hours	−.20***	—		
3. Anxiety_score	.12*	−.31***	—	
4. Exam_score	.27***	.20***	−.12*	—

* p < .05, ** p < .01, *** p < .001 (Holm-adjusted). Each pair uses every respondent with both values.

Copy table | Copy as picture | PNG | CSV

Every pair

Pair	r	95% CI	n	p	p (Holm)
Study_hours_week – Sleep_hours	−.201	[−.29, −.10]	388	< .001	< .001
Study_hours_week – Anxiety_score	.124	[.02, .22]	378	.016	.033
Study_hours_week – Exam_score	.273	[.18, .36]	387	< .001	< .001
Sleep_hours – Anxiety_score	−.306	[−.40, −.21]	366	< .001	< .001
Sleep_hours – Exam_score	.204	[.10, .30]	376	< .001	< .001
Anxiety_score – Exam_score	−.122	[−.22, −.02]	365	.020	.033

Figure 66. *The correlation matrix and the table of every pair (Pearson).*

The result shows the correlation matrix with stars for Holm-adjusted significance and a table of every pair with its 95% confidence interval, n, p and Holm-adjusted p. Each pair uses every student with both values (so n differs between pairs: 365 to 388).

What it means for our example

All six correlations are significant after Holm's adjustment. The strongest is Sleep_hours with Anxiety_score, r = −.31 (medium; students who sleep more are less anxious). Exam_score correlates with Study_hours_week, r = .27, 95% CI [.18, .36], n = 387, and with Sleep_hours, r = .20, 95% CI [.10, .30], and negatively with Anxiety_score, r = −.12, 95% CI [−.22, −.02], Holm-adjusted p = .033. Study_hours_week and Anxiety_score are barely related, r = .12 (Holm-adjusted p = .033).

How to report it

"Exam_score correlated positively with Study_hours_week, r(385) = .27, p < .001, 95% CI [.18, .36], and Sleep_hours, r(374) = .20, p < .001, 95% CI [.10, .30], and negatively with Anxiety_score, r(363) = −.12, Holm-adjusted p = .033 (Pearson correlations, Holm-adjusted p-values)."

11.2 Spearman and Kendall: rank correlations

Pearson's r measures a *straight-line* relationship and is pulled by extreme values. Spearman's ρ and Kendall's τ-b use ranks: they describe any steadily rising or falling relationship and are not affected by how extreme the extreme values are. Choose **Spearman's rank-order correlation (Spearman's ρ)** or **Kendall's rank correlation (Kendall's τ-b)**.

What it means for our example

The rank correlation of Study_hours_week with Exam_score is much larger than the Pearson correlation: ρ = .46, 95% CI [.37, .54], against r = .27 (Kendall's τ-b = .32, 95% CI [.26, .38]). The relationship is steadily rising but not a straight line, and the five extreme study times (Section 4.4) probably have a large influence on Pearson's r. In the other direction, the small Pearson correlation of Study_hours_week with Anxiety_score (r = .12, adjusted p = .033) is not supported by the ranks (ρ = .03, p = .611), which suggests that it came from the extreme values. When the two kinds of correlation disagree like this, report both and say why.

Use Spearman's ρ for ordinal data (single Likert items), for skewed variables and when outliers are present. Kendall's τ-b is preferable with many tied values or small samples; its values are smaller than ρ for the same data.

11.3 Partial correlation

A partial correlation is the correlation of two variables after removing what a third variable explains in both. Tick **Pre_score** under **Control for (optional: partial correlation)**; the title becomes *Partial correlation (Pearson r, controlling for Pre_score)*.

Partial correlations use complete cases only (n = 354 for every pair here). Controlling for the pre-test, Sleep_hours correlates more strongly with Exam_score (r = .29, 95% CI [.19, .39], against .20 before) and Anxiety_score correlates more negatively with it (r = −.19, against −.12): part of the noise in Exam_score was prior ability. Study_hours_week with Exam_score remains r = .27. Control only for variables that make sense as common causes or confounders, not for variables that lie on the path between the two (chapter 13).

11.4 Multiple linear regression

Concept: regression

Regression fits the equation *outcome = intercept + B_1 × predictor$_1$ + B_2 × predictor$_2$ + … + error*. Each unstandardised coefficient **B** is the change in the outcome for a one-unit change in that predictor, holding the others constant; the standardised **β** puts the predictors on a common scale (in SD units) so that they can be compared. **R^2** is the share of the variance in the outcome that the predictors explain together; adjusted R^2 corrects it for the number of predictors. Effect size for the model: $f^2 = R^2 / (1 - R^2)$.

1. Choose **Multiple linear regression (Which variables predict a score?)**.
2. As **Outcome (scale)** choose **Exam_score**.
3. Tick **Study_hours_week**, **Sleep_hours** and **Anxiety_score** under **Block 1 predictors (e.g. control variables)**. Categorical variables are marked "(groups)" and are dummy-coded against a reference group, which you choose under *Reference groups* (by default the first group).

The result checks the sample size, collinearity (VIF), normality of the residuals, constant error variance (Breusch–Pagan), influential cases (Cook's distance) and independent errors (Durbin–Watson), and then gives the model summary and the coefficients (Figure 67). The regression uses only complete cases: n = 354 of 400 (46 students with a missing value on any of the four variables are left out; chapter 14 shows how to keep them).

Model summary

Model	Predictors	R^2	Adjusted R^2	F	df	p	ΔR^2	p (ΔR^2)
1	Study_hours_week, Sleep_hours, Anxiety_score	.142	.134	19.28	3, 350	< .001		

Copy table | Copy as picture | PNG | CSV

Coefficients (model 1)

Predictor	B	SE	β	t	p	95% CI of B	VIF
(Intercept)	39.625	4.555		8.70	< .001	[30.668, 48.583]	
Study_hours_week	0.500	0.081	.315	6.20	< .001	[0.342, 0.659]	1.05
Sleep_hours	2.257	0.474	.252	4.76	< .001	[1.324, 3.191]	1.14
Anxiety_score	−0.266	0.171	−.081	−1.55	.121	[−0.602, 0.071]	1.11

n = 354 (46 respondents with missing values left out). β = standardised coefficient.

Figure 67. *The model summary and the coefficients of model 1.*

What it means for our example

The three predictors together explain 14.2% of the variance in Exam_score, R^2 = .142, adjusted R^2 = .134, F(3, 350) = 19.28, p < .001, f^2 = 0.17 (medium). Study_hours_week (B = 0.500, SE = 0.081, β = .315, 95% CI of B [0.342, 0.659]) and Sleep_hours (B = 2.257, SE = 0.474, β = .252, 95% CI [1.324, 3.191]) are significant predictors: each additional weekly hour of study goes with half a point more, and each additional hour of sleep with 2.3 points more, holding the other predictors constant. Anxiety_score is not significant, B = −0.266, β = −.081, p = .121. The largest VIF is 1.14, so collinearity is no problem; the residuals are normal (Shapiro–Wilk W = 0.997, p = .645); no Cook's distance exceeds 1 (largest 0.773). One check is marked *Note*: the Breusch–Pagan test, $\chi^2(3)$ = 30.37, p < .001, suggests that the error spread is not constant.

11.5 Hierarchical regression: what does a second block add?

In a hierarchical regression, predictors enter in blocks in an order decided by theory, and the change in R^2 (ΔR^2) tests what the later block adds. Tick **Gender (groups)** under **Block 2 predictors (optional: what do they add?)**. The title becomes *Multiple linear regression (two blocks)* and the model summary has two rows.

What it means for our example

Adding Gender raises R^2 from .142 to .172 (adjusted R^2 .163), ΔR^2 = .031, p < .001; the whole model is F(4, 349) = 18.18, p < .001. Men score 3.48 points lower than women on Exam_score, holding study hours, sleep and anxiety constant: B = −3.480, SE = 0.969, β = −.176, t(349) = −3.59, p < .001, 95% CI [−5.386, −1.575]. The coefficients of study hours (B = 0.481, β = .302) and sleep (B = 2.154, β = .241) hardly change, and anxiety remains non-significant (p = .190).

How to report it

"A hierarchical multiple linear regression predicted Exam_score from Study_hours_week, Sleep_hours and Anxiety_score (block 1, R^2 = .142) and Gender (block 2). Gender added a significant increment, ΔR^2 = .031, p < .001, final F(4, 349) = 18.18, p < .001, R^2 = .172, adjusted R^2 = .163. Study_hours_week (β = .30, p < .001), Sleep_hours (β = .24, p < .001) and Gender (male vs female, β = −.18, p < .001) were significant predictors."

11.6 Binary logistic regression: predicting Passed

When the outcome has two categories, linear regression is not appropriate; logistic regression models the probability of one of them.

Concept: odds and odds ratios

The **odds** of an event are the probability of the event divided by the probability of its absence (a 75% chance is odds of 3 to 1). A logistic regression coefficient is the change in the *log odds* for a one-unit change in the predictor; its exponential is the **odds ratio (OR)**. OR = 1 means no effect; above 1 the odds of the event rise; below 1 they fall. The *event* is the answer coded 1.

1. Choose **Binary logistic regression (Which variables predict a yes/no outcome?)**.
2. As **Outcome (two groups)** choose **Passed**, and choose "Yes" as the event.
3. Tick **Study_hours_week**, **Sleep_hours**, **Anxiety_score** and **Gender (groups)** under *Predictors*. **Cut-off for the classification table** stays at 0.50.

Model fit

χ^2	df	p	Nagelkerke R^2	McFadden R^2	Tjur R^2	AUC	95% CI of AUC
63.56	4	< .001	.225	.141	.186	.780	[.73, .83]

Copy table | Copy as picture | PNG | CSV

Coefficients

Predictor	B	SE	Wald	p	Odds ratio	95% CI of OR
(Intercept)	−2.740	1.267	4.68	.031	0.065	[0.005, 0.774]
Gender: male (vs female)	−0.906	0.251	13.06	< .001	0.404	[0.247, 0.661]
Study_hours_week	0.254	0.050	25.69	< .001	1.289	[1.168, 1.422]
Sleep_hours	0.485	0.132	13.58	< .001	1.624	[1.255, 2.101]
Anxiety_score	−0.066	0.045	2.19	.139	0.936	[0.858, 1.022]

"Yes" = 1, "No" = 0. Categorical predictors are dummy-coded against a reference group: Gender = female.

Figure 68. *The model fit and the coefficients with odds ratios.*

What it means for our example

The model uses 366 students and predicts passing better than chance: $\chi^2(4) = 63.56$, $p < .001$, Nagelkerke $R^2 = .225$ (McFadden .141, Tjur .186), AUC = .780, 95% CI [.73, .83] (acceptable). The events-per-variable check (28.0) and the Hosmer–Lemeshow test, $\chi^2(8) = 5.80$, $p = .669$, are fine. Each additional weekly hour of study multiplies the odds of passing by 1.29 (95% CI [1.17, 1.42], $p < .001$) and each additional hour of sleep by 1.62 (95% CI [1.25, 2.10], $p < .001$); the odds of passing for men are 0.40 times those of women (95% CI [0.25, 0.66], $p < .001$), equivalent to saying that women's odds are 2.47 times men's. Anxiety_score has no significant effect, OR = 0.94, 95% CI [0.86, 1.02], $p = .139$.

The classification table shows what the model does with each student: of the 112 who failed, it predicts 44 correctly (39.3%); of the 254 who passed, it predicts 232 correctly (91.3%); overall 75.4% are classified correctly, against 69.4% if the larger group ("Yes") were guessed for everyone. The overall accuracy looks good mostly because most students pass: the model finds failing students poorly. For an outcome like this, read the AUC and the sensitivity for the rarer group, not the overall accuracy alone; Section 16.5 shows how a different cut-off changes the picture.

How to report it

"A binary logistic regression predicting passing (Passed = Yes) from Gender, Study_hours_week, Sleep_hours and Anxiety_score was significant, $\chi^2(4, N = 366) = 63.56$, $p < .001$, Nagelkerke $R^2 = .23$, AUC = .78. Men had lower odds of passing than women, OR = 0.40, 95% CI [0.25, 0.66]; each weekly hour of study, OR = 1.29, 95% CI [1.17, 1.42], and each hour of sleep, OR = 1.62, 95% CI [1.25, 2.10], raised the odds."

11.7 Ordinal logistic regression: predicting a 1 to 5 rating

Satisfaction_1to5 is a single rating from 1 to 5. Its categories are ordered, but the distance between 2 and 3 need not equal that between 4 and 5, and a linear regression would treat them as equal. **Ordinal logistic regression** (the proportional-odds model; McCullagh, 1980) models the probability of being at or below each category.

Concept: proportional odds and the Brant test

The model has one coefficient per predictor and one threshold (cut-point) for each step between neighbouring categories (1 | 2, 2 | 3, 3 | 4, 4 | 5). The odds ratio of a predictor says how much the odds of being in a *higher* category change, and it is assumed to be the same at every cut: the **proportional-odds assumption**. The **Brant test** (Brant, 1990) checks it for each predictor and for all together; a significant test means that the effect differs between the cuts, and a partial proportional-odds model or a multinomial model is then safer (Peterson & Harrell, 1990).

1. On the Data screen, make Satisfaction_1to5 ordinal (Section 1.4). (A scale variable with 3 to 10 whole-number values is also accepted; its values are used in their numerical order.)
2. Choose **Ordinal logistic regression (Which variables predict an ordered outcome such as a 1–5 rating or low/medium/high?)** and **Satisfaction_1to5** as the outcome. SaNaStat shows the order it uses (1 < 2 < 3 < 4 < 5); **Reverse the order (if the first category is the highest)** turns it round.
3. Tick **Teaching_method (groups)** and **Gender (groups)** under *Predictors*, and choose **Lecture** as the **Reference group of Teaching_method**, so that both other methods are compared with the lecture (Figure 69).

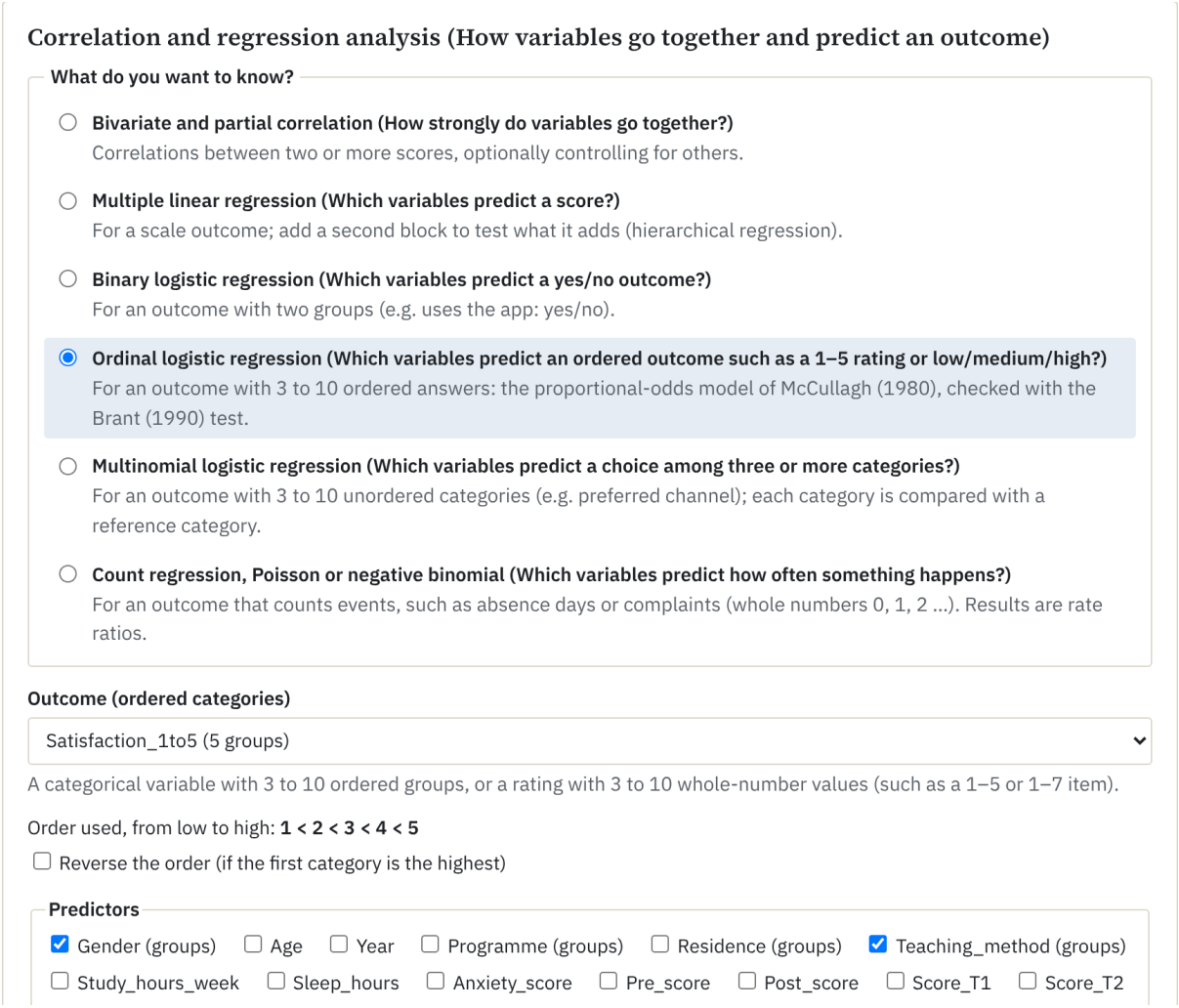


Figure 69. *Ordinal logistic regression: Satisfaction_1to5 as the outcome, with the order of its categories.*

Model fit

LR χ^2	df	p	−2LL	AIC	BIC	McFadden R^2	Cox–Snell R^2	Nagelkerke R^2
40.65	3	< .001	1041.83	1055.83	1083.49	.038	.100	.107

N = 384. LR χ^2 compares the model with a model without predictors. −2LL = −2 × log-likelihood; smaller AIC and BIC mean a better balance of fit and simplicity.

Copy table | Copy as picture | PNG | CSV

Coefficients

Predictor	B	SE	z	p	Odds ratio	95% CI of OR
Gender: male (vs female)	−0.055	0.193	−0.29	.774	0.946	[0.648, 1.381]
Teaching_method: Blended (vs Lecture)	0.875	0.228	3.84	< .001	2.399	[1.534, 3.750]
Teaching_method: Flipped (vs Lecture)	1.473	0.239	6.16	< .001	4.360	[2.728, 6.968]

Order of Satisfaction_1to5 (low to high): 1 < 2 < 3 < 4 < 5. A positive B (odds ratio above 1) means respondents tend to be in higher categories. Categorical predictors are dummy-coded against a reference group: Gender = female, Teaching_method = Lecture.

Figure 70. *The model fit and the coefficients with odds ratios.*

What it means for our example

The model uses the 384 students with a rating and explains satisfaction better than chance, LR $\chi^2(3) = 40.65$, $p < .001$, Nagelkerke $R^2 = .107$ (Figure 70). Compared with the lecture, blended teaching goes with 2.40 times the odds of a higher rating (B = 0.875, 95% CI of the OR [1.534, 3.750], $p < .001$) and the flipped classroom with 4.36 times the odds (B = 1.473, 95% CI [2.728, 6.968], $p < .001$). Gender has no effect, OR = 0.946, $p = .774$. The proportional-odds assumption holds: the Brant test is not significant for any term (for Flipped $\chi^2(3) = 6.50$, $p = .090$) or for all together, $\chi^2(9) = 8.73$, $p = .462$ (Figure 71). One check is marked *Check*: the smallest category ("1") has 24 respondents for 3 predictor terms, 8.0 per term, below the guideline of 10 events per variable (Peduzzi et al., 1996); with more predictors, merge sparse categories or collect more data.

Thresholds (cut-points)

Cut	Estimate	SE	z	p
1 | 2	−2.132	0.244	−8.73	< .001
2 | 3	−0.287	0.176	−1.63	.103
3 | 4	1.502	0.193	7.76	< .001
4 | 5	3.460	0.266	12.99	< .001

Logit of being at or below the category on the left when every predictor is 0.

Copy table | Copy as picture | PNG | CSV

Brant test of proportional odds

Term	χ^2	df	p
Gender: male (vs female)	0.82	3	.845
Teaching_method: Blended (vs Lecture)	2.20	3	.531
Teaching_method: Flipped (vs Lecture)	6.50	3	.090
All terms (omnibus)	8.73	9	.462

A significant test means the effect of that term differs between the cuts of the outcome.

Figure 71. *The thresholds and the Brant test of proportional odds.*

How to report it

"An ordinal logistic regression (proportional-odds model) of Satisfaction_1to5 on Gender, Teaching_method was significant, $\chi^2(3, N = 384) = 40.65$, $p < .001$, Nagelkerke $R^2 = .11$. Teaching_method = Blended (vs Lecture) went with 2.40 times the odds of a higher Satisfaction_1to5, OR = 2.40, 95% CI [1.53, 3.75], $p < .001$. Teaching_method = Flipped (vs Lecture) went with 4.36 times the odds of a higher Satisfaction_1to5, OR = 4.36, 95% CI [2.73, 6.97], $p < .001$. The proportional-odds assumption held, Brant $\chi^2(9) = 8.73$, $p = .462$."

11.8 Multinomial logistic regression: predicting the programme

When the outcome has three or more categories *without* an order (a programme, a preferred channel), **multinomial logistic regression** compares each category with a reference category. Its coefficients are relative risk ratios (RRR): how the odds of that category rather than the reference change for a one-unit increase in the predictor.

1. Choose **Multinomial logistic regression (Which variables predict a choice among three or more categories?)** and **Programme** as the outcome.
2. Choose **Business**, the largest programme, as **Reference category of Programme**.
3. Tick **Gender (groups)** and **Study_hours_week** under *Predictors* (Figure 72).

Model fit

LR χ^2	df	p	−2LL	AIC	BIC	McFadden R^2	Cox–Snell R^2	Nagelkerke R^2
7.05	4	.133	850.30	862.30	886.25	.008	.017	.020

N = 400. LR χ^2 compares the model with a model without predictors. −2LL = −2 × log-likelihood; smaller AIC and BIC mean a better balance of fit and simplicity.

Copy table | Copy as picture | PNG | CSV

Coefficients

Outcome (vs reference)	Predictor	B	SE	z	p	RRR	95% CI of RRR
Arts vs Business	(Intercept)	−0.218	0.246	−0.89	.376		
Arts vs Business	Gender: male (vs female)	−0.704	0.280	−2.51	.012	0.495	[0.286, 0.857]
Arts vs Business	Study_hours_week	−0.014	0.026	−0.55	.580	0.986	[0.937, 1.037]
Science vs Business	(Intercept)	−0.059	0.208	−0.28	.776		
Science vs Business	Gender: male (vs female)	−0.226	0.233	−0.97	.331	0.798	[0.506, 1.258]
Science vs Business	Study_hours_week	0.005	0.019	0.25	.799	1.005	[0.968, 1.043]

Reference category of Programme: "Business". RRR = relative risk ratio exp(B): how the odds of that category rather than the reference change for a one-unit increase. Categorical predictors are dummy-coded against a reference group: Gender = female.

Figure 72. *Multinomial logistic regression of Programme: the model fit and the coefficients against Business.*

What it means for our example

The model uses all 400 students but does not distinguish the programmes significantly, LR $\chi^2(4) = 7.05$, $p = .133$, Nagelkerke $R^2 = .020$. One coefficient stands out: men have about half the odds of studying Arts rather than Business, RRR = 0.495, 95% CI [0.286, 0.857], $p = .012$, the pattern found by the chi-square test of Section 10.2. Weekly study hours do not separate the programmes (Arts vs Business RRR = 0.986, $p = .580$; Science vs Business RRR = 1.005, $p = .799$), although Science students were simulated to study a little more: the five extreme study times hide that small difference. The classification table (Figure 73) shows how weak the model is: it assigns almost everyone to Business and classifies 40.5% correctly, no better than the 40.8% obtained by guessing the largest programme for everyone; SaNaStat marks this check as a *Note*.

How to report it

"A multinomial logistic regression of Programme (reference category "Business") on Gender, Study_hours_week was not significant, $\chi^2(4, N = 400) = 7.05$, $p = .133$, Nagelkerke $R^2 = .02$; 40.5% of respondents were classified correctly. Gender = male (vs female) lowered the odds of "Arts" rather than "Business", RRR = 0.49, 95% CI [0.29, 0.86], $p = .012$."

Classification

Observed	Predicted Arts	Predicted Business	Predicted Science	% correct
Arts	0	87	5	0.0
Business	0	156	7	95.7
Science	0	139	6	4.1
Overall				40.5

Each respondent is assigned to the category with the highest predicted probability. Guessing the largest category for everyone would be 40.8% correct.

Figure 73. *The classification table of the multinomial model.*

Note: independence of irrelevant alternatives

The multinomial model assumes that the odds between two categories do not depend on the other categories offered. Keep the categories distinct and meaningful; SaNaStat notes this in its checks.

Exercise: regression on practice_marketing.csv

Open practice_marketing.csv. (a) Run a multiple linear regression of Annual_spend on Income_k, Tenure_months, Loyalty_member, Visits_per_month and Satisfaction_1to10. Answer: $R^2 = .81$, $F(5, 1335) = 1108.12$, $p < .001$ ($n = 1341$); Visits_per_month dominates ($\beta = .86$), followed by Income_k ($\beta = .17$) and Satisfaction_1to10 ($\beta = .08$); Loyalty_member adds a little ($\beta = .04$); Tenure_months is not significant ($p = .747$). (b) Run a binary logistic regression of Churned (event "Yes") on Satisfaction_1to10, Complaints_12m, Tenure_months and Loyalty_member. Answer: AUC = .75 (acceptable); each satisfaction point has OR = 0.61 (95% CI [0.54, 0.68]), each complaint OR = 1.39 (95% CI [1.21, 1.60]) and loyalty membership OR = 0.69 (95% CI [0.49, 0.96]). Which of the checks does the regression flag? (c) Repeat (a) with the logarithms of Annual_spend and Income_k (Section 2.6). Answer: $R^2 = .77$ ($n = 1{,}268$ complete cases, because the 82 zero-spend customers and the missing incomes drop out); Visits_per_month is again the strongest predictor ($\beta = .84$), followed by Income_k_ln ($\beta = .17$), and the Breusch–Pagan test no longer flags the error variance ($p = .431$).

11.9 Stepwise selection: letting the data choose the predictors

Normally your theory says which predictors go into a regression (**Enter: all chosen predictors at once (recommended)**). When you only want to explore which of many variables matter, **How are the predictors entered?** offers **Forward selection**, **Backward elimination** and **Stepwise** (Figure 74).

1. Choose **Multiple linear regression**, **Exam_score** as the outcome and **Stepwise**.
2. Tick the candidate predictors **Pre_score**, **Study_hours_week**, **Sleep_hours**, **Anxiety_score**, **Age** and **Project_score**.

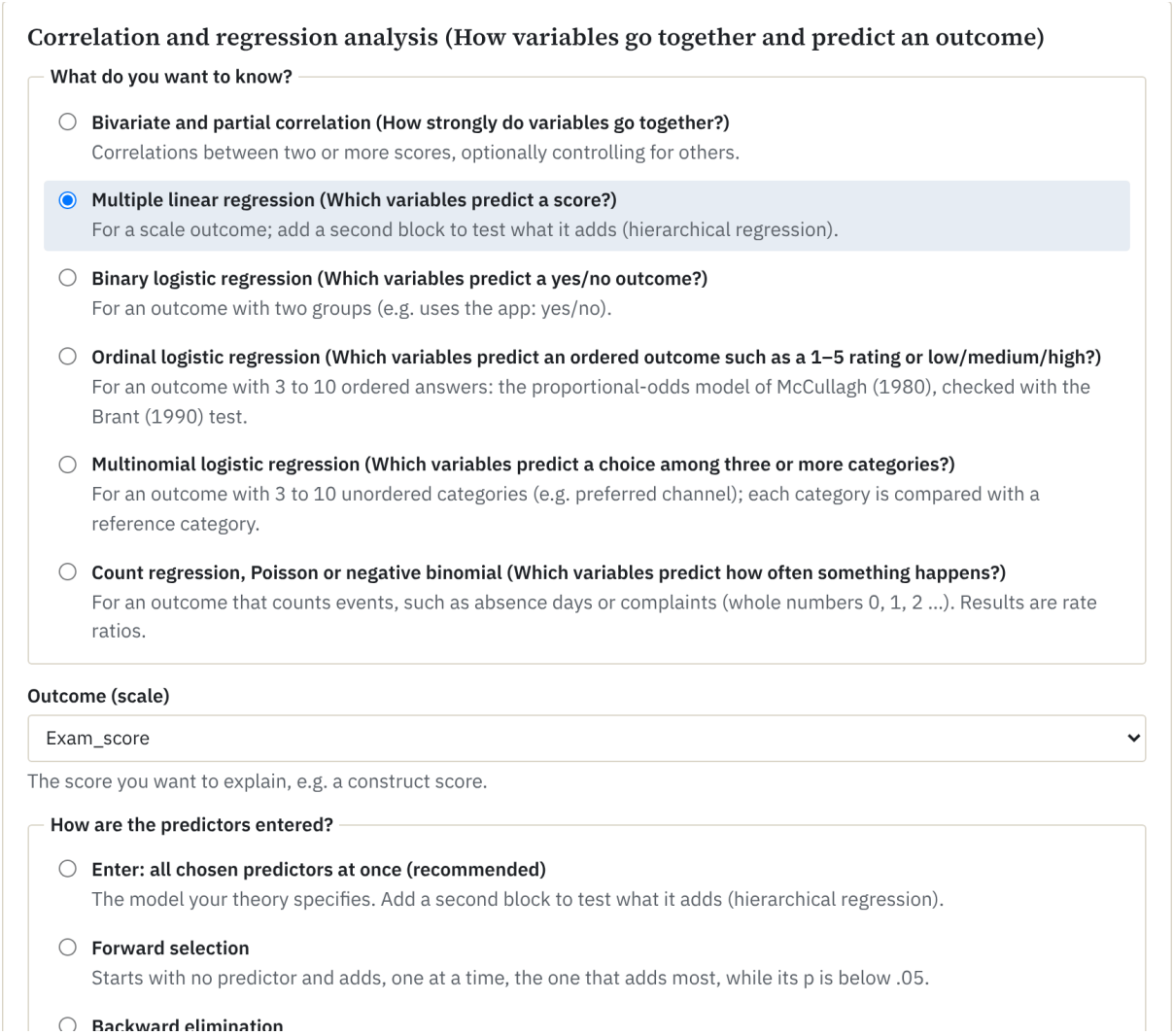


Figure 74. *Stepwise selection.*

SaNaStat adds the most useful predictor while its p is below .05 and, after each step, removes a predictor whose p has risen above .10 (Efroymson, 1960). The steps (Figure 75) enter Pre_score (R^2 = .341), Project_score (ΔR^2 = .088), Sleep_hours (ΔR^2 = .048), Study_hours_week (ΔR^2 = .069) and Anxiety_score (ΔR^2 = .009, p = .008). Age is left out (F to enter = 2.20, p = .138).

Steps of the stepwise selection

Step	Action	Variable	F	df	p	R^2	ΔR^2	Adjusted R^2
1	Entered	Pre_score	182.35	1, 352	< .001	.341	.341	.339
2	Entered	Project_score	54.43	1, 351	< .001	.430	.088	.426
3	Entered	Sleep_hours	32.46	1, 350	< .001	.478	.048	.474
4	Entered	Study_hours_week	52.76	1, 349	< .001	.547	.069	.541
5	Entered	Anxiety_score	7.01	1, 348	.008	.556	.009	.549

Entry when p < .05, removal when p > .10 (partial F test). After each entry, variables that are no longer needed are removed (Efroymson, 1960).

Figure 75. *The steps of the stepwise selection.*

The final model (Figure 76) explains 55.6% of the variance (R^2 = .556, adjusted R^2 = .549, $F(5, 348) = 87.01$, $p < .001$, $n = 354$); Pre_score has the largest standardised coefficient (β = .499) and Anxiety_score a small negative one (β = −.100).

Final model

Predictor	B	SE	β	t	p
(Constant)	3.320	3.867		0.86	.391
Pre_score	0.415	0.031	.499	13.24	< .001
Project_score	0.225	0.030	.282	7.49	< .001
Sleep_hours	2.226	0.343	.249	6.49	< .001
Study_hours_week	0.436	0.058	.274	7.47	< .001
Anxiety_score	−0.327	0.124	−.100	−2.65	.008

n = 354. R^2 = .556, adjusted R^2 = .549, $F(5, 348) = 87.01$, $p < .001$. β = standardised coefficient.

Figure 76. *The final model.*

Common mistake: testing a theory with stepwise selection

The variables were chosen by looking at the data, so the p values and confidence intervals of the final model are too small and R^2 is too large (Harrell, 2015); SaNaStat says so above the result (Figure 77). Use stepwise selection to explore, confirm the model on new data, and do not use it to test hypotheses.

Multiple linear regression, stepwise selection (Predicting Exam_score)

What it means: Stepwise selection kept 5 of 6 predictors: Pre_score, Project_score, Sleep_hours, Study_hours_week, Anxiety_score. Together they explain 55.6% of the variance in Exam_score (R^2 = .56). Because the predictors were chosen by looking at the data, treat this model as exploratory.

Note **Selected by the data.** Variables were chosen by looking at the data, so the p values, confidence intervals and R^2 of the final model are optimistic (p values too small, R^2 too large) and the selected set may not replicate. Treat the result as exploratory; confirm it on new data or with cross-validation.

Rule: Variables chosen by stepwise selection give optimistic p values and R^2; treat the model as exploratory (Harrell (2015)).

OK **Sample size.** 354 respondents for 6 candidate predictors (needed: 98, counting every candidate).

Rule: At least 50 + 8k respondents for k predictors to test the whole model (Green (1991)).

Figure 77. *The warning that the p values are optimistic.*

12. Measurement: reliability and validity of scales

A scale score is only as good as its items. Section 1.7 checked that the items of Motivation and Anxiety hang together; this chapter asks three further questions with **Factor analysis and reliability** (*Do the items measure what they should?*) in the *Measurement and models* group: how reliable is each scale (12.1), which items belong together when you do not tell the program (12.2), and does the intended measurement model fit, with scales that are both convergent and distinct (12.3)? Build the constructs Motivation and Anxiety first (Section 1.7).

12.1 Reliability: Cronbach's α and McDonald's ω

Choose **Reliability of scales (Cronbach's α and McDonald's ω)**. With **The construct scores built on the Data screen**, each construct is one scale, with its items and reverse-coding; **A set of items I pick here** checks a scale you define on the screen. Tick **Also show ω if an item is deleted (scales with four or more items)** (Figure 78).

Reliability of the scales

Scale	Items	n	Cronbach's α	McDonald's ω	Mean inter-item r
Motivation	6	400	.86	.86	.51
Anxiety	6	400	.88	.88	.56

ω (total) from a one-factor confirmatory factor analysis by maximum likelihood (McDonald, 1999). Complete cases per scale; (R) items are reverse-coded first.

Copy table | Copy as picture | PNG | CSV

Item statistics

Scale	Item	Corrected item–total r	α if deleted	ω if deleted	Standardised loading
Motivation	Mot1	.72	.83	.83	.78
Motivation	Mot2	.66	.84	.84	.72
Motivation	Mot3	.65	.84	.84	.70
Motivation	Mot4	.71	.83	.83	.77
Motivation	Mot5	.55	.86	.86	.60
Motivation	Mot6	.65	.84	.84	.71
Anxiety	Anx1	.74	.86	.86	.79
Anxiety	Anx2	.69	.86	.87	.74
Anxiety	Anx3 (R)	.70	.86	.86	.75
Anxiety	Anx4	.66	.87	.87	.71
Anxiety	Anx5 (R)	.72	.86	.86	.78
Anxiety	Anx6	.66	.87	.87	.71

Corrected item–total r: the item with the sum of the other items. "If deleted" values higher than the scale's own value mark items that weaken the scale. ω if deleted needs at least four items.

Figure 78. *Reliability of the scales and the item statistics.*

What it means for our example

Motivation has α = .86 and ω = .86, Anxiety α = .88 and ω = .88: both scales are reliable enough for research use (≥ .70). α and ω are close, so the items measure each construct about equally well. Every corrected item–total correlation is .30 or more (the lowest is Mot5, .55, with a standardised loading of .60), and no item would raise α or ω if deleted. One check is a *Note*: the mean inter-item correlations are .51 and .56, slightly above .50, where items may be near-duplicates and the scale narrow (Clark & Watson, 1995). That is a reason to read the item wording, not to delete items.

How to report it

"Internal consistency was ω = .86, α = .86 for Motivation (6 items); ω = .88, α = .88 for Anxiety (6 items), with McDonald's ω estimated from a one-factor model."

12.2 Exploratory factor analysis

Concept: exploratory factor analysis

An **exploratory factor analysis (EFA)** looks for a small number of factors that explain the correlations among the items, without being told which item belongs where. Each item has a **loading** on each factor (its correlation with the factor, for an orthogonal solution), and its **communality** h^2 is the share of its variance that the factors explain. **Parallel analysis** keeps the factors that explain more than factors found in random data of the same size (Horn, 1965), and is more trustworthy than the rule "eigenvalue greater than 1" (Kaiser, 1960), which keeps too many. A **rotation** makes the loadings easier to read; an oblique rotation (oblimin, promax) lets the factors correlate, as constructs in survey research usually do.

1. Choose **Exploratory factor analysis (Which items belong together?)**.
2. Type **Mot, Anx** in **Quick pick: items starting with** and press Tab. As in Section 1.7, the quick pick also ticks Anxiety_score; untick it, so that the twelve items Mot1 to Anx6 are chosen.
3. Keep the settings (Figure 79): **Parallel analysis (recommended)** under *How many factors?*, **Minimum residual (recommended)** extraction, **Oblimin (oblique, recommended)** rotation, **Pearson (usual choice)** correlations (choose **Polychoric (for Likert items)** for items with few answer options or skewed answers; Holgado-Tello et al., 2010) and **Hide loadings smaller than** .30.
4. Click **Run the factor analysis**.

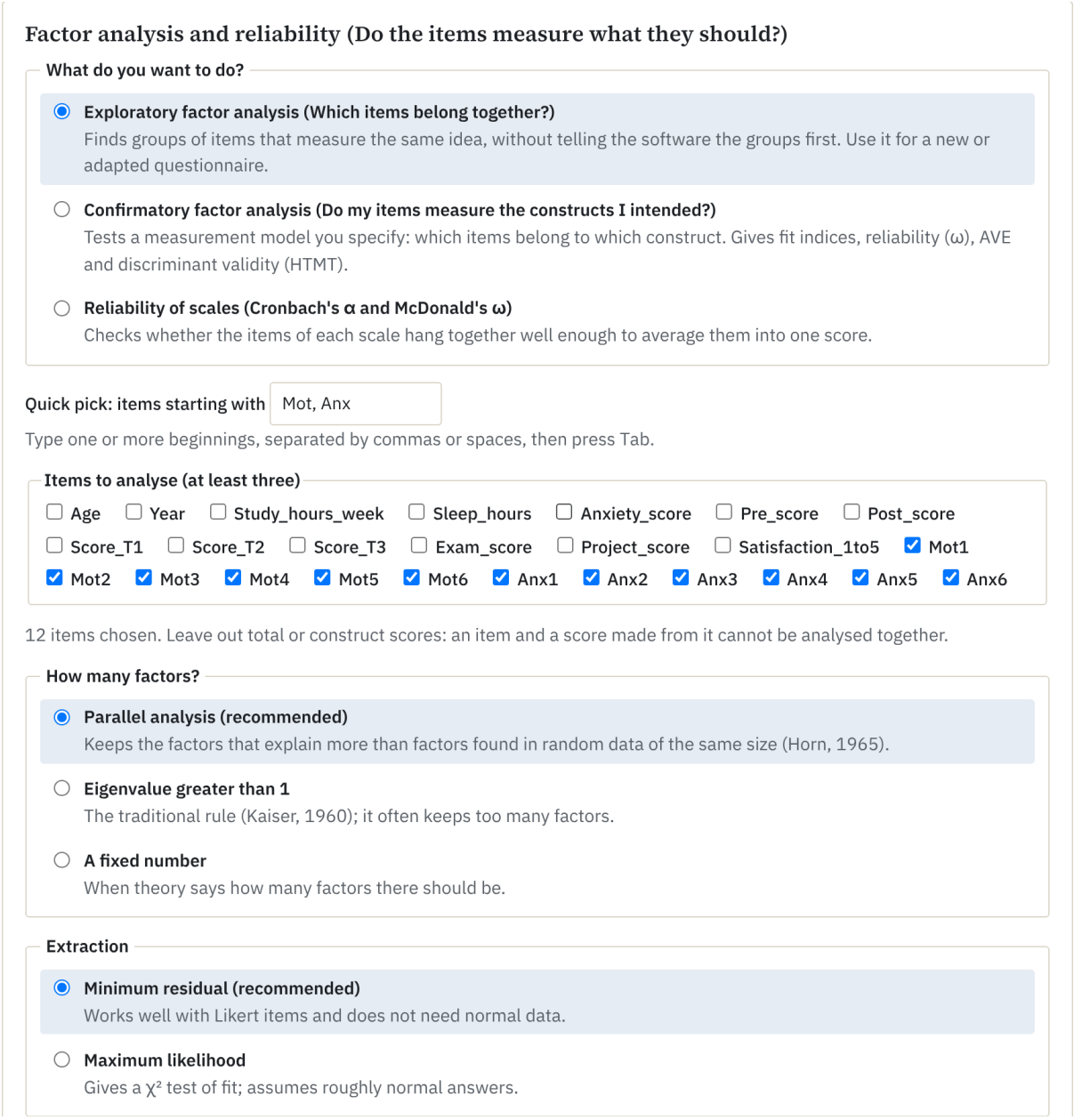


Figure 79. *The settings of the exploratory factor analysis.*

Suitability of the data

Statistic	Value
Kaiser–Meyer–Olkin measure (KMO)	.915 (marvelous)
Bartlett's test of sphericity	$\chi^2(66) = 2142.54$, $p < .001$
Determinant of the correlation matrix	.0044
Respondents analysed	400 of 400
Correlations	Pearson

Missing answers: respondents with any missing item are left out (listwise).

Copy table | Copy as picture | PNG | CSV

Measure of sampling adequacy per item (MSA)

Item	MSA
Mot1	.904
Mot2	.918
Mot3	.920
Mot4	.892
Mot5	.931
Mot6	.907
Anx1	.912
Anx2	.918
Anx3	.914
Anx4	.924
Anx5	.919
Anx6	.926

MSA below .50: the item shares too little with the others (Kaiser, 1974).

Copy table | Copy as picture | PNG | CSV

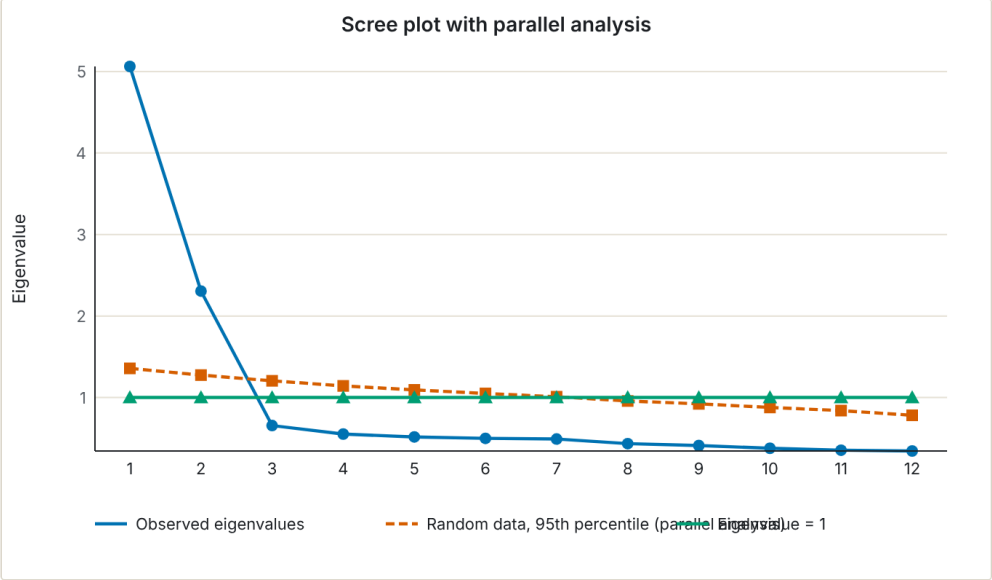
Eigenvalues and parallel analysis

Factor	Eigenvalue	Random 95th percentile	Keep?	Reduced-matrix eigenvalue	Random 95th percentile (reduced)
1	5.062	1.357	Yes	4.544	0.400
2	2.306	1.275	Yes	1.782	0.318
3	0.656	1.205		0.050	0.238
4	0.551	1.142		0.031	0.176
5	0.516	1.093		−0.025	0.121
6	0.499	1.050		−0.033	0.074
7	0.491	1.009		−0.057	0.033
8	0.434	0.958		−0.084	−0.020
9	0.411	0.922		−0.103	−0.054
10	0.377	0.878		−0.124	−0.098
11	0.353	0.839		−0.134	−0.139
12	0.344	0.782		−0.161	−0.192

Parallel analysis with 100 random data sets of 400 × 12 (95th percentile; Horn, 1965; Glorfeld, 1995). "Keep?" marks leading eigenvalues above the random ones. The reduced matrix has squared multiple correlations on the diagonal.

Figure 80. *The suitability of the data and the eigenvalues with parallel analysis.*

The report first checks whether the data are suitable (Figure 80): the Kaiser–Meyer–Olkin measure of sampling adequacy (KMO; Kaiser, 1974) for all items and per item, and Bartlett's test that the items are correlated at all (Bartlett, 1950). The scree plot (Figure 81) shows the eigenvalues of the data against those of random data.



Scree plot with parallel analysis. Factors above the dashed random-data line are worth keeping (Horn, 1965). Parallel analysis suggests 2 factors; 2 eigenvalues are above 1.

Figure 81. *Scree plot with parallel analysis: two factors lie above the random-data line.*

Pattern loadings (sorted by factor)

Item	F1	F2	h²	u²	Complexity	Note
Anx1	.79		.63	.37	1.00	
Anx5	−.76		.61	.39	1.01	
Anx2	.75		.55	.45	1.00	
Anx3	−.74		.57	.43	1.00	
Anx6	.72		.50	.50	1.00	
Anx4	.72		.50	.50	1.00	
Mot1		.79	.62	.38	1.00	
Mot4		.78	.60	.40	1.00	
Mot2		.73	.51	.49	1.00	
Mot6		.71	.50	.50	1.00	
Mot3		.69	.50	.50	1.00	
Mot5		.59	.36	.64	1.01	

Minimum residual (unweighted least squares) extraction, direct oblimin (oblique) rotation. Loadings below .30 in absolute value are left blank. † = cross-loading of .30 or more. h² = communality (variance shared with the factors), u² = uniqueness. Complexity = Hofmann's index (1 = the item loads on one factor only).

Figure 82. *The pattern loadings, sorted by factor: Anx3 and Anx5 load negatively.*

What it means for our example

The items are suitable: KMO = .915 ("marvelous"), every item has an MSA of at least .89, and Bartlett's test is significant, $\chi^2(66) = 2142.54$, $p < .001$. Parallel analysis keeps 2 factors (eigenvalues 5.062 and 2.306 against 1.357 and 1.275 in random data; the third, 0.656, lies below its random value of 1.205), which together explain 53.8% of the item variance. The pattern loadings (Figure 82) are clean: the six Anx items load .72 to .79 in absolute value on F1 and the six Mot items .59 to .79 on F2, with no cross-loading of .30 or more. Anx3 (−.74) and Anx5 (−.76) load negatively: they are the reverse-worded items of Section 1.7, and the table *Suggested constructs* proposes to reverse-code them. The two factors correlate at −.42: more motivated students are less anxious. The fit of the two-factor solution is good, $\chi^2(43) = 39.59$, $p = .620$, RMSEA = .000, TLI = 1.003.

Make construct scores from these factors turns the suggested constructs into scale scores, with the negative-loading items reversed; we already have them.

Cut-off values: exploratory factor analysis

KMO of .60 or more is the minimum and .70 or more is better; items with an MSA below .50 are candidates for removal (Kaiser, 1974).

Number of factors: parallel analysis with the 95th percentile of random eigenvalues (Horn, 1965; Glorfeld, 1995).

Clean loadings: at least .40 on one factor, below .30 on every other, and a gap of at least .20 (Howard, 2016).

Items per factor: at least three, better four or more; at least 100 respondents and 5 to 10 per item (Hair et al., 2019).

How to report it

"An exploratory factor analysis of 12 items (N = 400) used minimum residual (unweighted least squares) extraction with direct oblimin (oblique) rotation. The Kaiser–Meyer–Olkin measure was .91 and Bartlett's test of sphericity was significant, $\chi^2(66) = 2142.54$, $p < .001$. Parallel analysis (Horn, 1965) with 100 random data sets and the 95th percentile criterion (Glorfeld, 1995) suggested 2 factors, which were retained; together they explained 53.8% of the variance. All items loaded at least .40 on their factor with no cross-loadings of .30 or more (Howard, 2016)."

Saving factor scores

Below the results, **Save factor scores as variables** (Figure 83) adds one score per factor to the data. Factor scores weight every item by its loadings, instead of taking the plain mean of the items; they are standardised (mean 0). **Regression (Thurstone, 1935) (recommended)** gives the most accurate scores; **Bartlett (Bartlett, 1937)** gives scores

that each reflect only their own factor. Click **Save the factor scores**: SaNaStat saves *Anx_reg* and *Mot_reg*, named after the items of each factor.

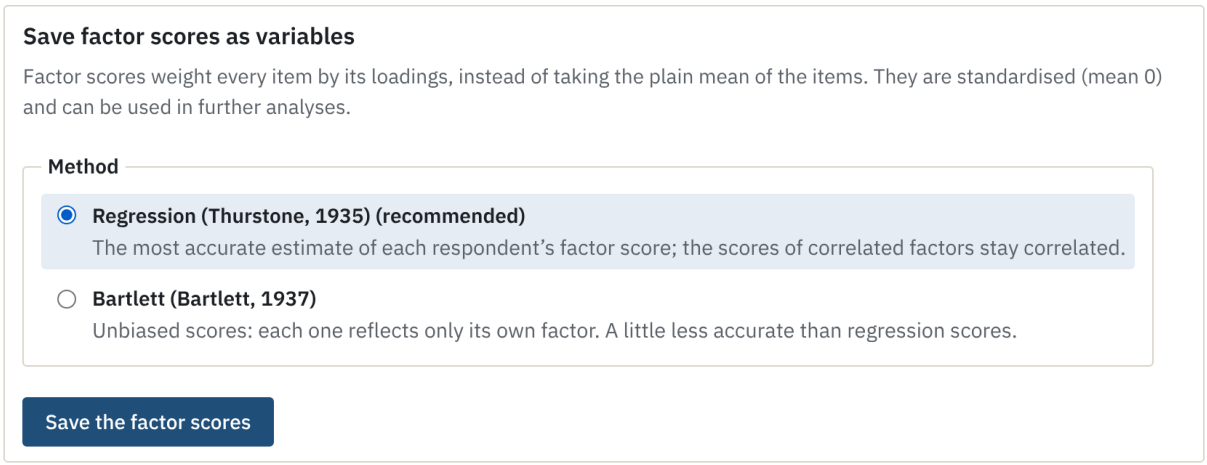


Figure 83. *Saving the factor scores.*

12.3 Confirmatory factor analysis, AVE, HTMT and Fornell–Larcker

Concept: confirmatory factor analysis and validity

A **confirmatory factor analysis (CFA)** tests a measurement model you specify: which items measure which construct. It reports how well the model reproduces the item covariances (**fit indices**: χ^2, CFI, TLI, RMSEA and SRMR), the loadings, the composite reliability ω, and two kinds of validity. **Convergent validity**: the items of a construct share most of their variance with it; the **average variance extracted (AVE)** should be at least .50 (Fornell & Larcker, 1981). **Discriminant validity**: the constructs are distinct; the **heterotrait–monotrait ratio (HTMT)** should be below .85 (Henseler et al., 2015), and each construct's √AVE should exceed its correlations with the other constructs (the **Fornell–Larcker criterion**).

1. Choose **Confirmatory factor analysis (Do my items measure the constructs I intended?)**.
2. Keep **Use the construct scores built on the Data screen**: Motivation and Anxiety become the two factors, each measured by its items, with Anx3 and Anx5 reversed (Figure 84). **Build the factors here** defines factors on the screen instead.
3. Click **Run the confirmatory factor analysis**.

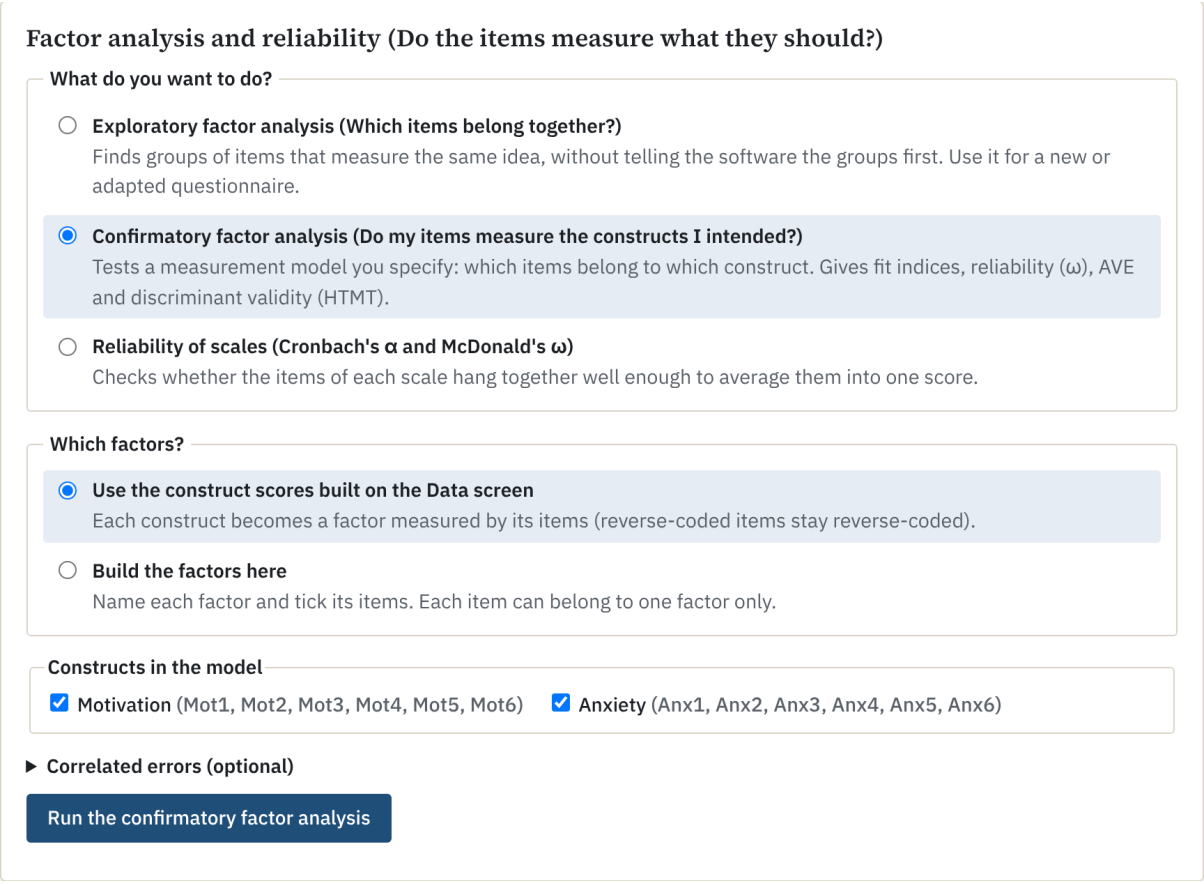


Figure 84. *The confirmatory factor analysis of Motivation and Anxiety.*

Model fit

χ^2	df	p	CFI	TLI	RMSEA	90% CI of RMSEA	p-close	SRMR	AIC	BIC
45.40	53	.761	1.000	1.004	.000	[.000, .023]	1.000	.022	11381.5	11481.3

Maximum likelihood, N = 400. RMSEA and its 90% CI use N (not N − 1) in the denominator, the convention of most current SEM programs. p-close tests RMSEA ≤ .05. SRMR is the standardised root mean square residual. Lower AIC/BIC = better when comparing models on the same items. Cut-offs: CFI/TLI ≥ .95 good, ≥ .90 acceptable; RMSEA ≤ .06 good, ≤ .08 acceptable; SRMR ≤ .08 (Hu & Bentler, 1999; Browne & Cudeck, 1993).

Figure 85. *The fit of the two-factor model.*

What it means for our example

The two-factor model fits very well: $\chi^2(53) = 45.40$, $p = .761$, CFI = 1.000, TLI = 1.004, RMSEA = .000, 90% CI [.000, .023], SRMR = .022; every index meets the strict cut-offs of Hu and Bentler (1999) (Figure 85). All standardised loadings lie between .60 (Mot5) and .79 and are significant. Convergent validity holds: ω = .86 and AVE = .52 for Motivation, ω = .88 and AVE = .56 for Anxiety (Figure 86). Discriminant validity holds too. The latent correlation of the two factors is −.429, smaller in size than √AVE (.72 and .75), so the Fornell–Larcker criterion is met, and the HTMT is −0.43. It is negative because the two constructs are negatively related; the thresholds apply to its absolute value, 0.43, far below .85.

Reliability and convergent validity

Factor	Items	Cronbach's α	McDonald's ω	AVE	√AVE
Motivation	6	.86	.86	.52	.72
Anxiety	6	.88	.88	.56	.75

ω (composite reliability) and AVE come from the standardised loadings of this model. Rules: α and ω ≥ .70; AVE ≥ .50.

Copy table | Copy as picture | PNG | CSV

Factor correlations

Pair	r	SE (covariance)	z	p
Motivation – Anxiety	−.429	0.037	−6.59	< .001

Latent correlations (corrected for measurement error); the test is for the covariance.

Copy table | Copy as picture | PNG | CSV

Fornell–Larcker criterion

	Motivation	Anxiety
Motivation	.72	
Anxiety	−.43	.75

Diagonal: √AVE; below it: factor correlations. Each diagonal value should be larger than the correlations in its row and column (Fornell & Larcker, 1981).

Copy table | Copy as picture | PNG | CSV

Heterotrait–monotrait ratio (HTMT)

	Motivation	Anxiety
Motivation	—	
Anxiety	−0.43	—

! = .85 or more (above the strict threshold); !! = .90 or more (above the lenient threshold) (Henseler et al., 2015). A negative HTMT means that the two constructs are negatively related; the thresholds apply to its absolute value.

Figure 86. *Reliability and convergent validity, the factor correlation, the Fornell–Larcker criterion and the HTMT.*

The table *Largest modification indices (hints only)* lists the changes that would most improve the fit; the largest here is 6.52 (errors of Mot4 and Mot6). With a model that already fits this well, there is nothing to change: free a parameter only with a theoretical or wording reason, never because a modification index is large (MacCallum et al., 1992).

Cut-off values: confirmatory factor analysis

CFI and TLI ≥ .95 good, .90 to .95 acceptable; **RMSEA** ≤ .06 good, up to .08 reasonable; **SRMR** ≤ .08 (Hu & Bentler, 1999; Browne & Cudeck, 1993). Judge the indices together, with the loadings and residuals (Marsh et al., 2004).

Standardised loadings ≥ .50 (better ≥ .70); **ω** ≥ .70; **AVE** ≥ .50 (Fornell & Larcker, 1981).

HTMT < .85 (strict) or < .90 (for conceptually close constructs) (Henseler et al., 2015); **√AVE** larger than the correlations with the other constructs.

How to report it

"A confirmatory factor analysis (maximum likelihood, N = 400) of a 2-factor model showed good fit, $\chi^2(53) = 45.40$, $p = .761$, CFI = 1.000, TLI = 1.004, RMSEA = .000, 90% CI [.000, .023], SRMR = .022. Standardised loadings ranged from .60 to .79 and all free loadings were significant ($p < .05$). McDonald's ω was .86 (Motivation), .88 (Anxiety). The largest HTMT was 0.43 in absolute value." Add the AVE (.52 and .56) and the Fornell–Larcker result if your field expects them.

Exercise: the engagement scales

Open practice_engagement.csv and build Vigor, Dedication and Absorption (Section 1.7, exercise). Run the reliability analysis, an EFA of the nine items with parallel analysis, and a CFA of the three constructs. How many factors does parallel analysis suggest? Do the three scales meet the AVE and HTMT criteria? If the HTMT of two constructs is close to .90, what would you write about their distinctness?

12.4 Agreement between raters: Cohen's kappa and the ICC

When two coders classify the same answers, or several raters score the same people, the question is whether they agree. The **Agreement between raters** screen (Figure 87) gives Cohen's kappa for categories and the intraclass correlation (ICC) for scores.

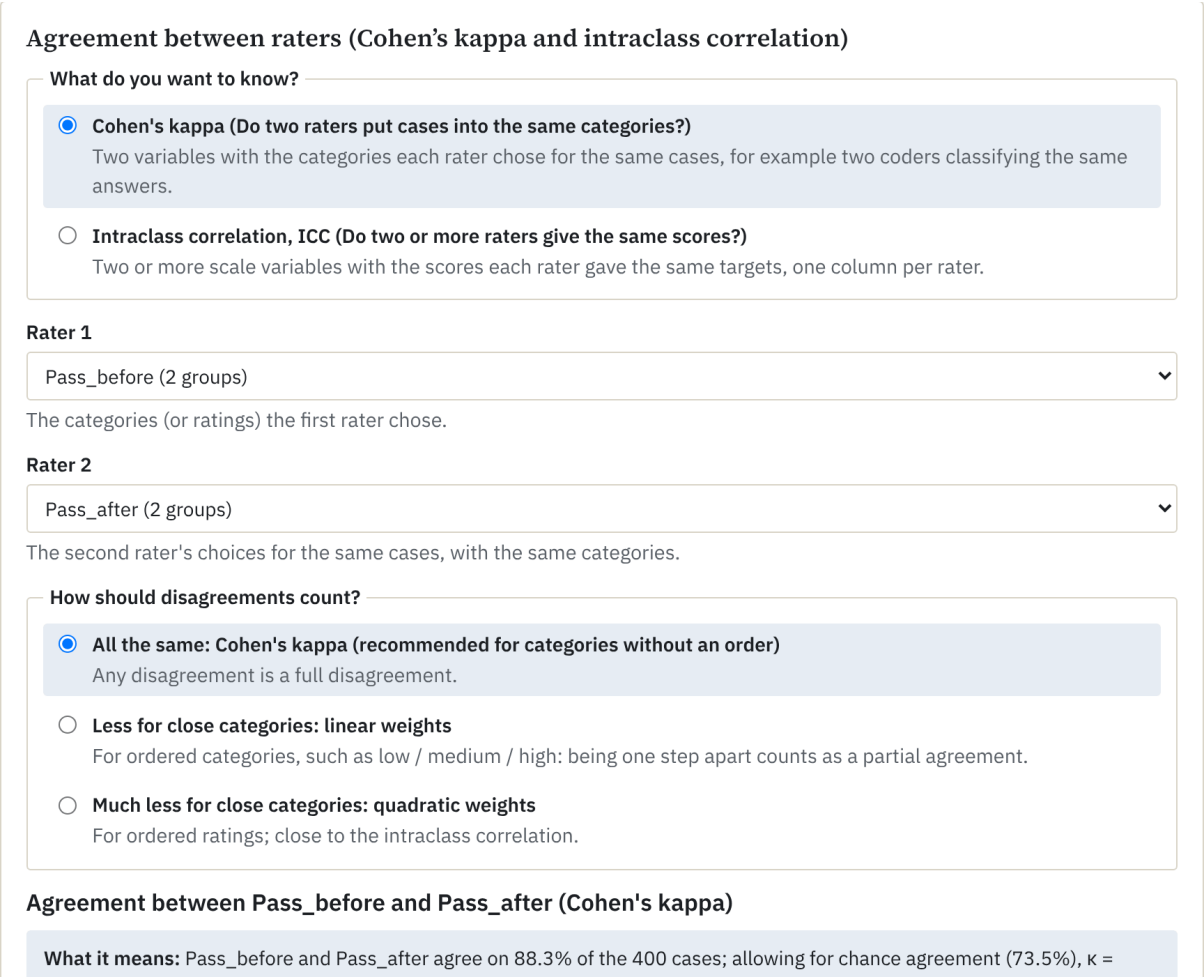


Figure 87. *The Agreement screen.*

Cohen's kappa. The practice file has no second coder, so the example treats the pass/fail classifications before and after the course (Section 10.6) as two "raters" of the same 400 students.

1. Choose **Cohen's kappa**, **Pass_before** as **Rater 1** and **Pass_after** as **Rater 2**. Keep **All the same: Cohen's kappa**; for ordered categories, linear or quadratic weights count close categories as partial agreement.

The two classifications agree for 88.3% of the students, but 73.5% agreement would be expected by chance alone, because most students pass both times. Kappa compares the two: κ = .557, 95% CI [.448, .665], z = 11.97, p < .001 (Figure 88), a moderate agreement (Cohen, 1960; Landis & Koch, 1977).

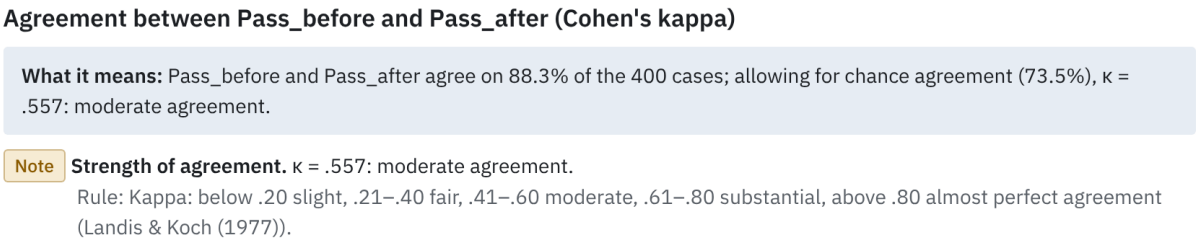


Figure 88. *Cohen's kappa of Pass_before and Pass_after.*

The intraclass correlation. Score_T1, Score_T2 and Score_T3 are three measurements of the same test, so their ICC is a test–retest reliability.

1. Choose **Intraclass correlation, ICC** and tick **Score_T1**, **Score_T2** and **Score_T3** as raters.

SaNaStat gives all six forms of Shrout and Fleiss (1979) with the names of McGraw and Wong (1996) (Figure 89). The consistency of a single measurement is excellent, ICC(3,1) = .918, 95% CI [.904, .930]: the students keep their ranking. Absolute agreement is lower, ICC(2,1) = .857, with a much wider interval, [.641, .927], because the scores rise from one measurement to the next (M = 59.37, 62.55 and 66.09), and absolute agreement counts that rise as disagreement.

Intraclass correlation coefficients

Form	Model	ICC	95% CI	F	df	p
ICC(1,1)	One-way random, single rater	.854	[.830, .875]	18.56	399, 800	< .001
ICC(2,1) · ICC(A,1)	Two-way random, absolute agreement, single rater	.857	[.641, .927]	34.61	399, 798	< .001
ICC(3,1) · ICC(C,1)	Two-way mixed, consistency, single rater	.918	[.904, .930]	34.61	399, 798	< .001
ICC(1,3)	One-way random, average of k raters	.946	[.936, .955]	18.56	399, 800	< .001
ICC(2,3) · ICC(A,3)	Two-way random, absolute agreement, average of k raters	.947	[.843, .975]	34.61	399, 798	< .001
ICC(3,3) · ICC(C,3)	Two-way mixed, consistency, average of k raters	.971	[.966, .976]	34.61	399, 798	< .001

Shrout and Fleiss (1979) forms with the McGraw and Wong (1996) names. Choose the form before looking at the values: two-way random when the raters are a sample of possible raters, two-way mixed when they are the only raters of interest; absolute agreement when the scores themselves must match, consistency when only the ranking matters; single or average depending on whether one rater or the mean of all raters will be used.

Figure 89. *The six intraclass correlations of the three measurements.*

Cut-off values: agreement

Kappa: below .20 slight, .21–.40 fair, .41–.60 moderate, .61–.80 substantial, above .80 almost perfect agreement (Landis & Koch, 1977). ICC: below .50 poor, .50–.75 moderate, .75–.90 good, above .90 excellent reliability; judge the lower limit of the 95% interval, not only the value (Koo & Li, 2016). Choose the ICC form before you look at the values.

How to report it

"Test–retest reliability was assessed with a two-way random-effects model for absolute agreement, single measurement (McGraw & Wong, 1996): ICC(2,1) = .86, 95% CI [.64, .93], F(399, 798) = 34.61, p < .001, indicating good reliability (Koo & Li, 2016)."

13. Mediation and moderation

Many research questions are not "does X go with Y?" but *how* (through what?) or *when* (for whom?). A **mediator** M carries part of the effect of X to Y (X → M → Y); a **moderator** W changes the size of the effect of X on Y. **Mediation and moderation** (*Indirect effects and interactions*) in the *Measurement and models* group fits the four regression models of Hayes (2022) that answer these questions most often (Table 8). These analyses run after a click on **Run**, because the bootstrap takes a few seconds.

Table 8
The four models of the Mediation and moderation screen (Hayes, 2022)

Model	Design	Question
4	Mediation: X → M → Y, one to four parallel mediators	Does X go with Y partly through M?
1	Moderation: X × W → Y	Is the effect of X on Y stronger, weaker or reversed for some values of W?
7	First-stage moderated mediation: W moderates X → M	Does the indirect effect depend on W, through the X → M path?
14	Second-stage moderated mediation: W moderates M → Y	Does the indirect effect depend on W, through the M → Y path?

Build the construct Motivation first (Section 1.7); Section 13.2 uses it as a moderator.

13.1 Mediation: does the post-test carry the effect of the pre-test?

Students who start the course with a higher pre-test score do better in the exam. Is that because they also finish the course with a higher post-test score?

Concept: indirect effect and bootstrap interval

Mediation is estimated with two regressions: M on X (the path **a**) and Y on X and M (the paths **b** and **c′**). The **indirect effect** is a × b: how much Y changes through M when X rises by one unit. The **direct effect** c′ is the effect of X that does not go through M, and the **total effect** c = c′ + ab. The indirect effect is not normally distributed, so its test is a **percentile bootstrap confidence interval**: SaNaStat draws 5,000 resamples of the respondents, computes ab in each and takes the middle 95%. The indirect effect is significant when this interval does not include zero (Preacher & Hayes, 2008; Hayes, 2022). The old normal-theory Sobel test is shown for comparison only.

1. Choose **Mediation (X → M → Y; Hayes 2022 model 4, one to four parallel mediators)** under *Which design?*.

2. As **X: predictor (scale or two groups)** choose **Pre_score**; as **Y: outcome (scale)** choose **Exam_score**; tick **Post_score** under **Mediators M (one to four; more than one = parallel mediation)** (Figure 90). **Covariates (optional: held constant in every equation)** adds control variables.
3. Under **Options**, keep **5,000 (recommended)** bootstrap samples and the **Random seed** 20221 (the same seed gives the same numbers every time; report it). **Heteroscedasticity-consistent standard errors (HC3)** is safer when the spread of the residuals differs across respondents (Long & Ervin, 2000).
4. Click **Run**.

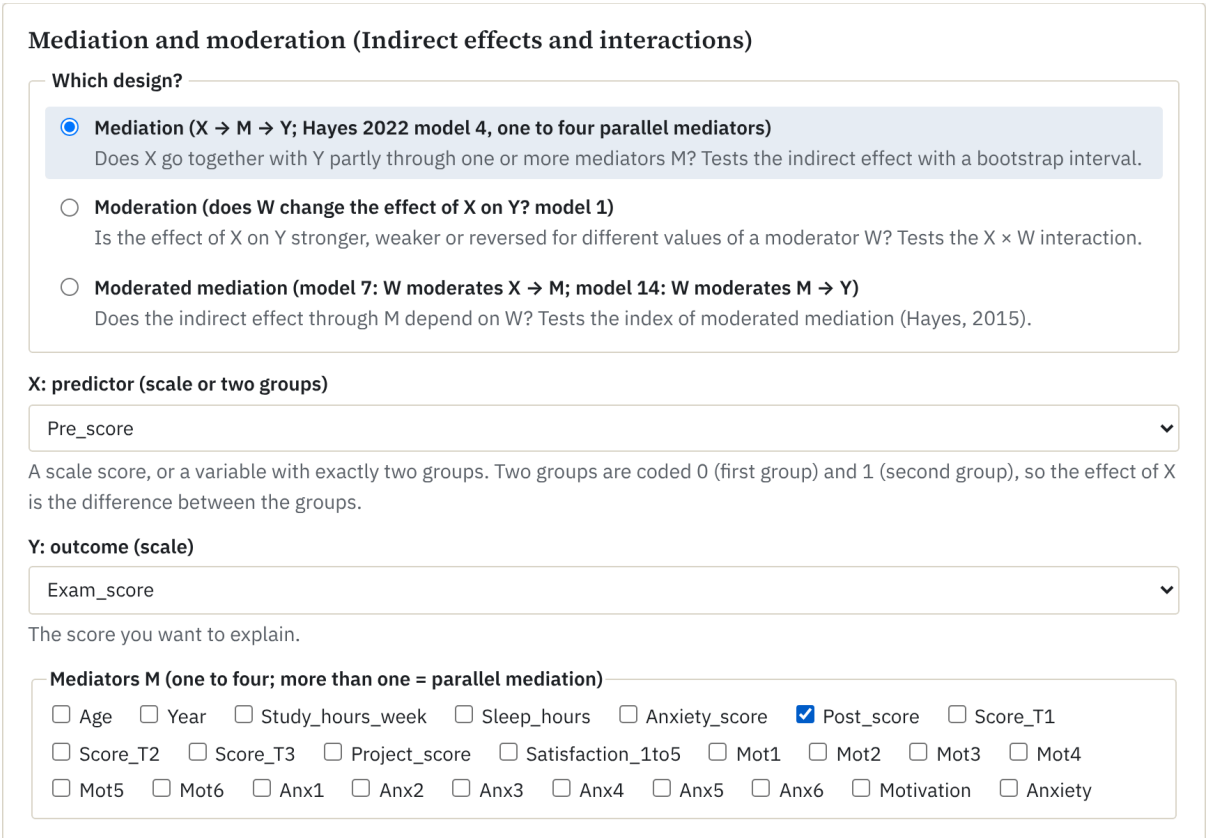


Figure 90. *Mediation, model 4: Pre_score → Post_score → Exam_score.*

Total and direct effect of Pre_score on Exam_score

Effect	b	SE	t	p	95% CI	Completely standardised
Total effect (c)	0.484	0.036	13.59	< .001	[0.414, 0.554]	0.569
Direct effect (c′)	−0.052	0.069	−0.75	.453	[−0.187, 0.084]	−0.061

c = c′ + the sum of the indirect effects. Completely standardised = effect when X and Y are in SD units.

Copy table | Copy as picture | PNG | CSV

Indirect effects of Pre_score on Exam_score (bootstrap)

Path	Effect	Boot SE	95% bootstrap CI	Excludes zero?	Completely standardised	95% CI (compl. std.)
Pre_score → Post_score → Exam_score	0.536	0.062	[0.416, 0.662]	Yes: significant	0.630	[0.495, 0.764]

Percentile bootstrap, 5,000 resamples, seed 20221.

Copy table | Copy as picture | PNG | CSV

Paths through each mediator

Mediator	a (Pre_score → M)	p (a)	b (M → Exam_score)	p (b)	ab
Post_score	1.029	< .001	0.521	< .001	0.536

a = effect of X on the mediator; b = effect of the mediator on Y holding X constant; ab = indirect effect.

Figure 91. *The total, direct and indirect effects and the paths through the mediator.*

What it means for our example

The model uses the 387 students with an exam score. The total effect of the pre-test on the exam is c = 0.484 (95% CI [0.414, 0.554], p < .001): one point more in the pre-test goes with about half a point more in the exam. The pre-test strongly predicts the post-test (a = 1.029, p < .001), and the post-test predicts the exam when the pre-test is held constant (b = 0.521, p < .001). The indirect effect ab = 0.536 has a 95% bootstrap CI of [0.416, 0.662], which excludes zero: it is significant (completely standardised 0.630). The direct effect is not significant, c′ = −0.052, 95% CI [−0.187, 0.084], p = .453 (Figure 91). The whole effect of the pre-test on the exam runs through the post-test: in Baron and Kenny's older language, full mediation. SaNaStat adds that the pattern is consistent with mediation but does not prove cause and effect on its own (Maxwell & Cole, 2007).

How to report it

"A mediation analysis (Hayes, 2022, model 4; 5,000 bootstrap samples) tested whether Post_score mediated the relationship between Pre_score and Exam_score (N = 387). The indirect effect through Post_score was significant, ab = 0.536, 95% bootstrap CI [0.416, 0.662], completely standardised ab = 0.630. The direct effect of Pre_score on Exam_score was not significant, c′ = −0.052, t(384) = −0.75, p = .453, and the total effect was significant, c = 0.484, t(385) = 13.59, p < .001." Add the seed (20221) in the method section.

Note: what the design can and cannot show

The pre-test was taken before the course and the post-test at its end, so the order X → M → Y is plausible here. With data collected at one time point, a significant indirect effect is consistent with mediation but cannot show that X causes M and M causes Y; say "consistent with", and justify the order from theory and design.

13.2 Moderation and the Johnson–Neyman technique

Does motivation change how much sleep matters for the exam?

Concept: moderation, simple slopes and Johnson–Neyman

Moderation is tested by the product term X × W in a regression of Y on X, W and X × W. If the product is significant, the effect of X depends on W. **Simple slopes** show the effect of X at a few values of W (one SD below the mean, the mean and one SD above). The **Johnson–Neyman technique** finds the values of W at which the effect of X changes between significant and not significant, so no arbitrary probe values are needed (Johnson & Neyman, 1936; Bauer & Curran, 2005). Interactions in field data are usually small, and detecting them needs large samples (McClelland & Judd, 1993).

1. Choose **Moderation (does W change the effect of X on Y? model 1)**.
2. As X choose **Sleep_hours**, as Y **Exam_score** and as **W: moderator (scale or two groups)** **Motivation** (Figure 92).
3. Keep **Mean-centre the variables in the product** (the coefficients of X and W become effects at the average of the other; the interaction test does not change) and **Mean and ± 1 SD** as the values of W at which to show the effect. Click **Run**.

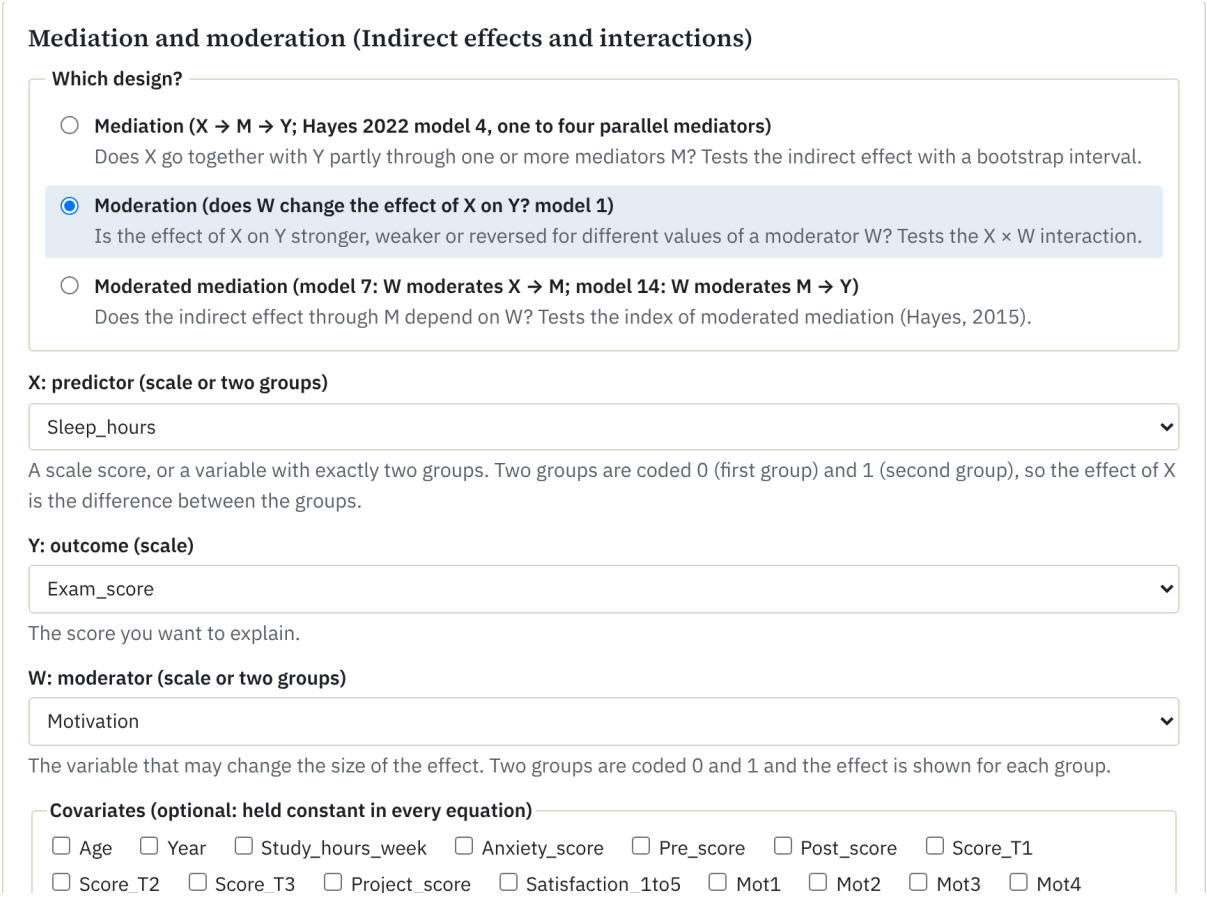


Figure 92. *Moderation, model 1: does Motivation change the effect of Sleep_hours on Exam_score?*

Test of the interaction (Sleep_hours × Motivation)

Outcome	ΔR²	F	df1	df2	p
Exam_score	.000	0.17	1	372	.677

ΔR² = the variance in Exam_score the product term adds.

Copy table | Copy as picture | PNG | CSV

Conditional effect of Sleep_hours on Exam_score at values of Motivation

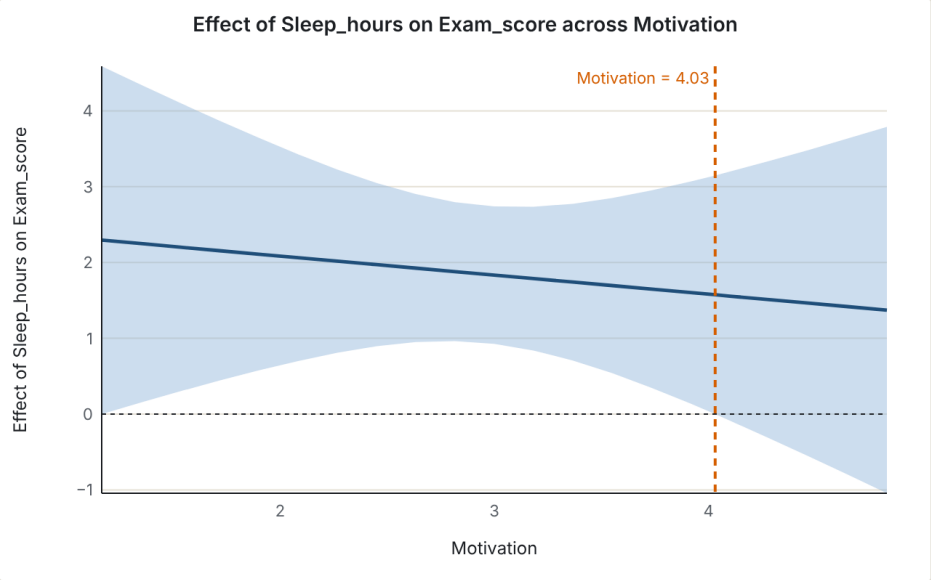
Value of the moderator	Motivation	Effect	SE	t	p	95% CI
Low (mean − 1 SD)	2.26	2.020	0.619	3.27	.001	[0.804, 3.236]
Average (mean)	3.02	1.829	0.462	3.96	< .001	[0.922, 2.737]
High (mean + 1 SD)	3.78	1.638	0.680	2.41	.017	[0.301, 2.976]

Values of Motivation on the original scale (mean and ± 1 SD).

Figure 93. *The test of the interaction and the conditional effects of Sleep_hours at three values of Motivation.*

What it means for our example

The model uses 376 students. Sleep matters: at the average motivation, each extra hour of sleep goes with 1.829 more exam points ($p < .001$). But motivation does not change this effect: the product term Sleep_hours × Motivation is $b = -0.251$, 95% CI [−1.437, 0.934], $p = .677$, and adds $\Delta R^2 = .000$ (Figure 93). The conditional effects at low (2.26), average (3.02) and high (3.78) motivation are 2.020, 1.829 and 1.638, all significant and close to one another. The Johnson–Neyman boundary lies at Motivation = 4.03: below it (91.2% of the students) the effect of sleep is significant and positive, above it not (Figure 94). That does not mean that sleep stops mattering for the most motivated students: the confidence band simply widens where there are few students, and the interaction itself is not significant. SaNaStat therefore advises to describe one overall effect of Sleep_hours.



Johnson–Neyman plot: effect of Sleep_hours on Exam_score across Motivation. Shaded band: 95% confidence interval. Where the band does not include zero, the effect is significant. Dashed vertical lines: Johnson–Neyman boundaries.

Figure 94. *The Johnson–Neyman plot: the effect of Sleep_hours on Exam_score across Motivation, with its 95% confidence band.*

How to report it

"A moderation analysis (Hayes, 2022, model 1) showed that the interaction between Sleep_hours and Motivation in predicting Exam_score was not significant, $b = -0.251$, $t(372) = -0.42$, $p = .677$, 95% CI [−1.437, 0.934], $\Delta R^2 = .000$, $F(1, 372) = 0.17$."

Note: the practice data have no planted interaction

practice_students.csv was built without interactions, so a non-significant product term is the expected result. Report it as it is: describe one overall effect of X, and do not interpret the conditional effects as if they differed.

13.3 Moderated mediation: models 7 and 14

Moderated mediation asks whether an *indirect* effect depends on a moderator. Choose **Moderated mediation (model 7: W moderates X → M; model 14: W moderates M → Y)** and, under **Where does W act?**, **First stage: W moderates X → M (model 7)** or **Second stage: W moderates M → Y (model 14)**. Choose X, Y, the mediator in **M: mediator (scale)** and the moderator W. The key result is the **index of moderated mediation** (Hayes, 2015): how much the indirect effect changes when W rises by one unit; for a two-group W, the difference between the indirect effects of the two groups. Moderated mediation is shown when its bootstrap interval excludes zero.

Is the mediation of Section 13.1 the same for women and men? Choose model 14, X **Pre_score**, Y **Exam_score**, M **Post_score** and W **Gender** (two groups, coded 0 = female and 1 = male), and click **Run** (Figure 95).

Conditional indirect effects of Pre_score on Exam_score through Post_score

Value of the moderator	Gender	Indirect effect	Boot SE	95% bootstrap CI	Excludes zero?
female	female	0.529	0.062	[0.412, 0.658]	Yes: significant
male	male	0.568	0.068	[0.436, 0.706]	Yes: significant

Percentile bootstrap, 5,000 resamples, seed 20221. Values of Gender on the original scale.

Copy table | Copy as picture | PNG | CSV

Index of moderated mediation

Moderator	Index	Boot SE	95% bootstrap CI	Moderated mediation?
Gender	0.039	0.052	[−0.065, 0.137]	No

How much the indirect effect changes when Gender rises by one unit (Hayes, 2015).

Figure 95. *Model 14: the conditional indirect effects for women and men and the index of moderated mediation.*

What it means for our example

The indirect effect of the pre-test through the post-test is significant for both groups: 0.529 for women, 95% CI [0.412, 0.658], and 0.568 for men, 95% CI [0.436, 0.706]. The index of moderated mediation is 0.039, 95% bootstrap CI [−0.065, 0.137]: the interval includes zero, so there is no clear evidence that the indirect effect differs between women and men, and it should be reported as one indirect effect. Model 7 (Gender moderating the path from the pre-test to the post-test) gives the same answer: index = −0.035, 95% bootstrap CI [−0.093, 0.024].

How to report it

"A moderated mediation analysis (Hayes, 2022, model 14; 5,000 bootstrap samples) tested whether Gender moderated the second stage (Post_score → Exam_score) of the indirect effect of Pre_score on Exam_score through Post_score (N = 387). The interaction Post_score × Gender was not significant, ΔR^2 = .001, F(1, 382) = 0.47, p = .495. The index of moderated mediation was not significant, index = 0.039, 95% bootstrap CI [−0.065, 0.137]."

Exercise: a mediation that is not there

In practice_students.csv, test whether anxiety carries the effect of sleep on the exam: X Sleep_hours, M Anxiety_score, Y Exam_score. Is the indirect effect significant? Is the direct effect? Then test Gender as a moderator of the effect of Sleep_hours on Exam_score (model 1). Answers: the indirect effect is not significant, ab = 0.171, 95% bootstrap CI [−0.148, 0.507] (N = 354): more sleep goes with less anxiety (a = −0.843, p < .001), but anxiety does not predict the exam once sleep is held constant (b = −0.203, p = .259). The direct effect of sleep is significant, c′ = 1.713, p < .001. Gender does not moderate the effect of sleep either, b = 0.640, p = .483, ΔR^2 = .001; the effect of sleep is 1.482 for women and 2.122 for men, both significant.

14. Missing data and multiple imputation

Chapters 6 to 11 analysed the cases that had the values each analysis needed. That is fine when few values are missing and they are missing completely at random, but it throws information away (the regression of Section 11.4 lost 46 of 400 students) and it can bias the results when whether a value is missing depends on other variables. **Multiple imputation** is the principled alternative (Rubin, 1987; van Buuren, 2018).

Concept: multiple imputation by chained equations

Multiple imputation fills in each missing value several times (m times) with plausible values predicted from the other variables, so that there are m completed data sets. Each is analysed in the usual way and the m results are **pooled with Rubin's rules**: the estimate is the mean of the m estimates, and its standard error adds the variation *between* the imputations to the usual standard error, so the uncertainty about the missing values shows in the result. **Chained equations** (MICE) fill in one variable at a time from all the others and repeat this for several iterations. For numbers and ordered answers SaNaStat uses **predictive mean matching**: each missing value is taken from one of five observed respondents with a similar predicted value, so imputed values are always real answers (Little, 1988a). Multiple imputation assumes that the data are **missing at random** (MAR): whether a value is missing may depend on the variables in the model, but not on the missing value itself.

14.1 Set up the imputation

Choose **Missing data: multiple imputation** (*Analyse incomplete data properly*) in the *Measurement and models* group.

1. Under **Variables to impute (the ones you will analyse)** tick **Sleep_hours** (12 missing), **Anxiety_score** (22 missing) and **Exam_score** (13 missing).
2. Under **Auxiliary variables (optional: help to predict the missing values)** tick **Gender**, **Study_hours_week**, **Pre_score**, **Post_score** and **Passed** (Figure 96). Auxiliary variables are related to the incomplete variables or to who skipped questions; they make the missing-at-random assumption more plausible and the imputations more precise (Collins et al., 2001). Passed is a particularly good one: it is known for all 13 students without an exam score, and it says whether their score was above 50.
3. Keep **Number of imputations (m)** at 20, **Iterations per imputation** at 10 and the **Random seed**. SaNaStat shows the expected time.
4. Click **Run multiple imputation**. The imputation runs in the background, with a progress bar and a **Stop** button.

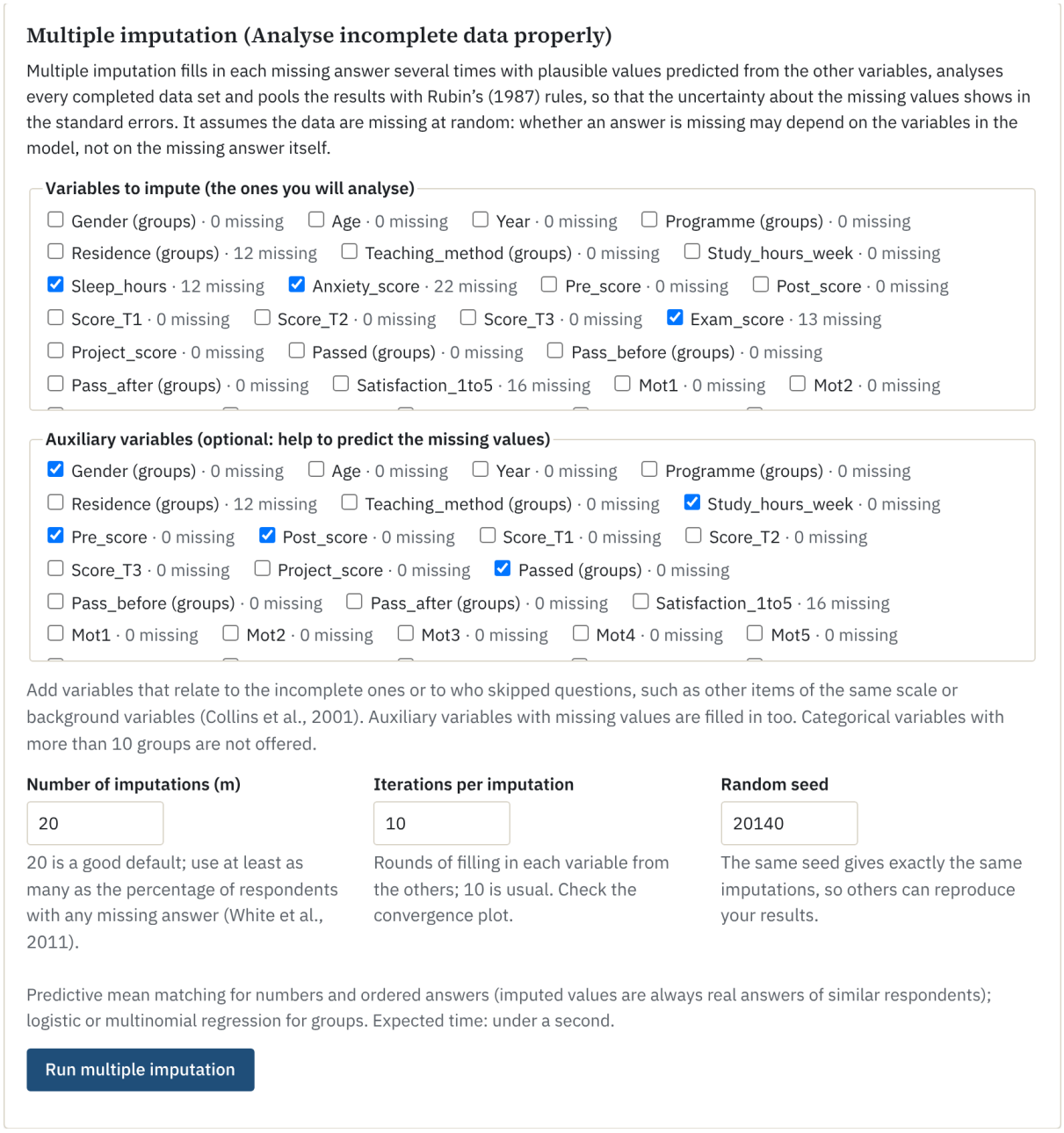


Figure 96. *The multiple imputation screen: three variables to impute and five auxiliary variables.*

Cut-off values: how many imputations?

Use at least as many imputations as the percentage of respondents with any missing value (White et al., 2011): 12% incomplete needs m ≥ 12, so the default 20 is enough. The more information is missing, the more imputations are needed: about 20 when the fraction of missing information (FMI) is up to .30, 40 when it is about .50 (Graham et al., 2007).

14.2 Diagnostics: did the imputation work?

The diagnostics report (Figure 97, Figure 98, Figure 99) answers three questions: what was imputed, did the chains settle, and do the imputed values look plausible?

What it means for our example

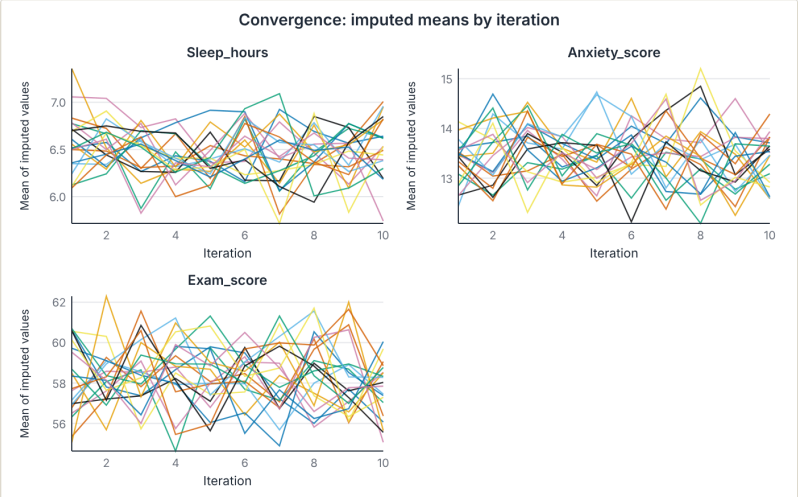
Three variables were imputed 20 times; 46 of 400 students (12%) had at least one missing value, most often in Anxiety_score (5.5% missing). The table *Variables in the imputation model* shows the method for each variable (predictive mean matching with 5 donors for the three imputed variables; the five auxiliary variables are complete). The *Missing-data patterns* show 354 complete students, 22 with only Anxiety_score missing, 12 with only Exam_score, 11 with only Sleep_hours and 1 with both Sleep_hours and Exam_score. The convergence plot shows the mean of the imputed values in each of the 20 chains across the 10 iterations: the lines overlap without a trend, and the largest $\hat{R}$ is 1.01, well below the warning level of 1.1 (Gelman & Rubin, 1992). The plot of observed and imputed values shows that the imputed values lie within the range of the observed ones.

Missing-data patterns

Respondents	Sleep_hours	Anxiety_score	Exam_score	Gender	Study_hours_week	Pre_score	Post_score	Passed	Missing in pattern
354	✓	✓	✓	✓	✓	✓	✓	✓	0
22	✓	missing	✓	✓	✓	✓	✓	✓	1
12	✓	✓	missing	✓	✓	✓	✓	✓	1
11	missing	✓	✓	✓	✓	✓	✓	✓	1
1	missing	✓	missing	✓	✓	✓	✓	✓	2

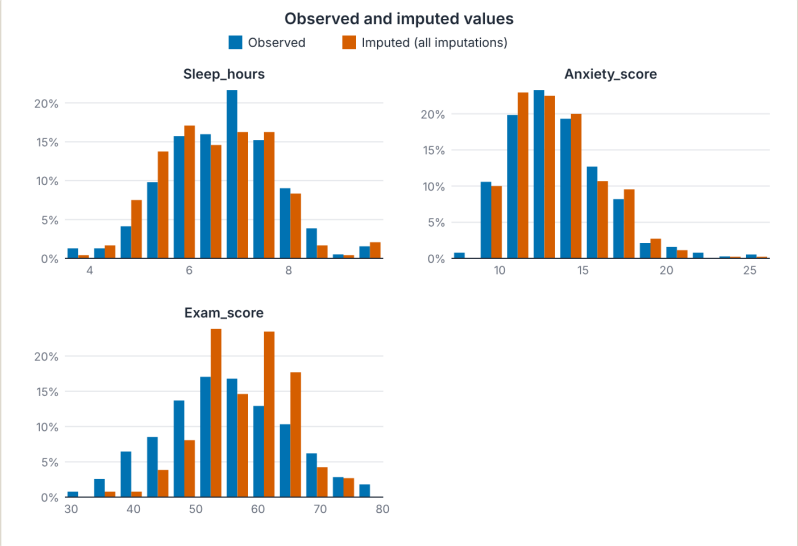
✓ = answered. 354 of 400 respondents (88.5%) are complete.

Figure 97. *The missing-data patterns: 354 complete students and four patterns of missing values.*



Convergence: imputed means by iteration. One line per chain (imputation). Lines should mix and show no trend after the first iterations.

Figure 98. *Convergence: the mean of the imputed values in each chain, iteration by iteration.*



Observed and imputed values. Share of respondents per range. Imputed values come from observed respondents with similar predicted values (predictive mean matching), so they look like real answers; their distribution may differ when missingness depends on other variables.

Figure 99. *Observed and imputed values of the three imputed variables.*

14.3 Pooled results

Under *Pooled analyses* choose the analysis to run on every completed data set: **Descriptive statistics (pooled means, SDs and percentages)**, **Multiple linear regression (Which variables predict a score?)**, **Binary logistic regression (Which variables predict a yes/no outcome?)** or **Pearson correlations (How strongly do scores go together?)**. Only variables in the imputation model can be used: the analysis model should never contain variables the imputation model left out.

The pooled descriptive statistics come first. The pooled mean of Exam_score is 55.06, 95% CI [54.09, 56.02], against 54.97 among the 387 students with a score; Sleep_hours 6.67 (6.68 answered only) and Anxiety_score 13.49 (13.50). The fraction of missing information (FMI) is at most .04, so 20 imputations are plenty. The small rise of the exam mean shows that the 13 students without an exam score were, as predicted from the other variables (especially Passed: 11 of them passed), slightly better than average.

Now repeat the regression of Section 11.5 on the imputed data: choose **Multiple linear regression (Which variables predict a score?)**, **Exam_score** as the outcome and **Study_hours_week**, **Sleep_hours**, **Anxiety_score** and **Gender (groups)** as predictors (Figure 100).

Pooled regression coefficients

Term	B	SE	t	df	p	95% CI	FMI	B (complete cases)
(Intercept)	41.91	4.46	9.40	324.30	< .001	[33.14, 50.68]	.08	41.20
Sleep_hours	2.08	0.46	4.54	338.30	< .001	[1.18, 2.98]	.07	2.15
Anxiety_score	−0.25	0.17	−1.47	345.90	.144	[−0.58, 0.08]	.06	−0.22
Gender: male (vs female)	−3.40	0.93	−3.67	386.20	< .001	[−5.22, −1.58]	.02	−3.48
Study_hours_week	0.54	0.08	6.80	389.50	< .001	[0.38, 0.69]	.01	0.48

Rubin's rules over 20 imputations; df by Barnard and Rubin (1999). FMI = fraction of missing information. Unstandardised coefficients.

Copy table | Copy as picture | PNG | CSV

Pooled model fit

Statistic	Value
R^2 (pooled, Fisher z)	.175
Adjusted R^2 (pooled)	.167
F test of the model (D2)	F(4, 15464.90) = 20.30, p < .001
Respondents	400
R^2 (complete cases, n = 354)	.172

R^2 pooled through Fisher's z of R (Harel, 2009); the model test combines the F statistics of the imputations (D2; Li et al., 1991), which is somewhat conservative.

Figure 100. *The pooled regression coefficients with their FMI, next to the complete-case coefficients, and the pooled model fit.*

What it means for our example

The pooled regression uses all 400 students. The four predictors together explain Exam_score significantly, F(4, 15464.90) = 20.30, p < .001 (the D2 test of Li et al., 1991), pooled R^2 = .175 (Harel, 2009), against .172 with the 354 complete cases. The pooled coefficients are close to the complete-case ones, shown in the last column: Study_hours_week B = 0.54 (complete cases 0.48), Sleep_hours 2.08 (2.15), Gender (male vs female) −3.40 (−3.48), all p < .001, and Anxiety_score −0.25 (−0.22), p = .144, not significant. The FMI of each coefficient is at most .08. With only 12% of the students incomplete and data that look missing at random, multiple imputation confirms the complete-case result and makes it slightly more precise; with more missing data, or missingness related to the outcome, the two can differ much more.

How to report it

"Missing values were handled by multiple imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011) with m = 20 imputations and 10 iterations, using predictive mean matching (Little, 1988a; 5 donors) for numeric and ordinal variables; Gender, Study_hours_week, Pre_score, Post_score, Passed were included as auxiliary variables. Results were pooled with Rubin's (1987) rules and Barnard–Rubin (1999) degrees of freedom. In the pooled multiple linear regression of Exam_score on Sleep_hours, Anxiety_score, Gender, Study_hours_week, the model was significant, F(4, 15464.90) = 20.30, p < .001, pooled R^2 = .18 (Harel, 2009)." Then report each coefficient with its B, SE, t, df, p, 95% CI and FMI, as SaNaStat does.

Note: two papers by Little from 1988

SaNaStat cites Little (1988) both for predictive mean matching (Little, 1988a, in the imputation sentence) and for the MCAR test (Little, 1988b, in the data quality screen). If you cite both, add the letters a and b in your text and reference list, as this manual does.

Exercise: impute the satisfaction rating

Add Satisfaction_1to5 (16 missing) to the variables to impute and Teaching_method as an auxiliary variable, run the imputation again and compare the pooled mean of Satisfaction_1to5 with the mean of those who answered. Then run a pooled binary logistic regression of Passed on Study_hours_week, Sleep_hours and Gender and compare it with Section 11.6.

15. Power analysis and sample size planning

Power analysis answers two questions: how many people do I need? and how likely is my study to find an effect that is there? It needs no data file: click **Power analysis** in the menu.

Concept: errors, power and effect size

A **Type I error** is a false positive (finding an effect that is not there); its probability is the significance level α, usually .05. A **Type II error** is a false negative (missing an effect that is there); its probability is β. **Power** = 1 − β is the probability of finding an effect that really exists; .80 is the usual target. Power rises with the sample size, the size of the true effect and α. The **effect size** is the smallest effect that would matter to you, or one taken from earlier studies; Cohen's (1988) conventions (small, medium, large) are a fallback, not a substitute for thinking.

15.1 The three kinds of power analysis

1. *1 · Which analysis will you run?* Choose the test: **Independent-samples t-test (Compare two groups)**, **Paired-samples t-test**, **One-sample t-test**, **One-way ANOVA, fixed effects**, **Chi-square test**, **Pearson correlation test**, **Multiple linear regression, F-test of R^2**, **Multiple linear regression, F-test of the R^2 increase** and **ANOVA or ANCOVA, one effect of a factorial design** (both in Section 15.5), or the PLS-SEM inverse square root method (not covered here).
2. *2 · What do you want to find?* **A priori power analysis (The sample size I need)** before data collection, **Post hoc power analysis (The power of a given sample size)** for a given sample, or **Sensitivity power analysis (The smallest effect I can detect)**.
3. Enter the effect size, α (.05), the power (.80) and the details of the design (the number of groups, predictors or degrees of freedom). The result and a ready-made sentence appear immediately (Figure 101), with **Copy this sentence** and **Add to Word report**; the report entry adds a table of the inputs and the result.

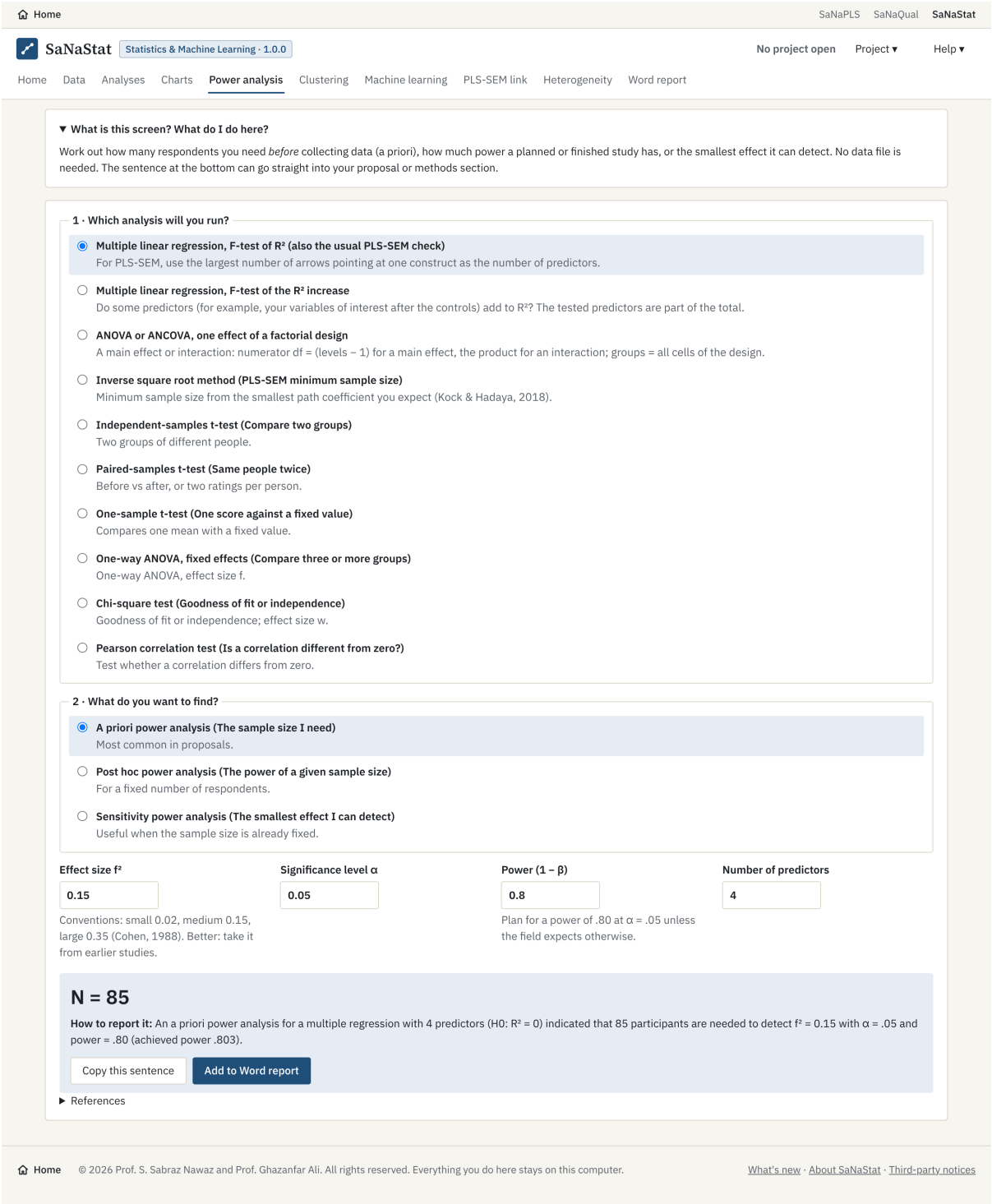


Figure 101. *The power analysis screen as it opens: a regression with four predictors needs N = 85.*

The effect-size measure depends on the test: Cohen's d (t-tests; d_z for paired data), f (ANOVA), f^2 (regression), w (chi-square) and r (correlation). Table 9 shows how to convert the effect sizes found in the earlier chapters.

Table 9

Effect sizes used for planning, and how they were obtained from the results of this manual

Planned analysis	Effect size input	From
Independent-samples t-test	Cohen's d = 0.44	Chapter 6, Exam_score by Gender
One-way ANOVA, 3 groups	Cohen's f = 0.24	Chapter 7: $f = \sqrt{(\eta^2 / (1 - \eta^2))} = \sqrt{(.054 / .946)} = 0.24$
Chi-square, 2 df	Cohen's w = 0.17	Chapter 10: w = Cramér's V × √(smaller table dimension − 1), here V = .17 for the 3 × 2 table, because the smaller dimension is 2
Pearson correlation	r = .27	Chapter 11, Study_hours_week and Exam_score
Multiple regression, 4 predictors	$f^2 = 0.21$	Chapter 11: $f^2 = R^2 / (1 - R^2) = .172 / .828 = 0.21$
Paired-samples t-test	$d_z = 0.50$	A medium effect, as a convention

15.2 A priori: how many students do I need?

Suppose you want to repeat the comparison of women and men (d = 0.44) in a new study, with a two-sided test, α = .05 and power .80. Choose **Independent-samples t-test**, **A priori** and type 0.44 for the effect size. SaNaStat reports N = 166, 83 per group (achieved power .805).

The other analyses in Table 9 give: one-way ANOVA with 3 groups and f = 0.24, N = 171 (57 per group, achieved power .802); chi-square with 2 degrees of freedom and w = 0.17, N = 334 (.801); a correlation of r = .27, N = 105 (.802); a regression with 4 predictors and $f^2 = 0.21$, N = 62 (.802); and a paired-samples t-test with $d_z = 0.50$, N = 34 (.808). A medium effect ($f^2 = 0.15$) with four predictors needs N = 85 (Figure 101). A one-sided test of d = 0.50 needs N = 102 (51 per group, achieved power .806) because the test is concentrated in one direction; use it only with a directional hypothesis made in advance.

How to report it

"An a priori power analysis for an independent-samples t-test indicated that 166 participants (83 per group) are needed to detect Cohen's d = 0.44 with α = .05 (two-sided) and power = .80 (achieved power .805)."

Tip: plan for drop-out and missing data

The numbers SaNaStat gives are the sample you need *after* exclusions. If you expect 5% of your respondents to drop out, give inaccurate answers or leave variables blank (as in this practice file), recruit the target divided by .95. Also remember that SaNaStat's power analyses assume the standard conditions of each test; the sample size for a multilevel or non-random sample may need more.

15.3 Sensitivity: what can I detect with the students I have?

With N = 400 and power .80, which effects can a study detect? Choose **Sensitivity power analysis** and type the sample size. With four predictors, a regression can detect f^2 = 0.030 or larger; an independent-samples t-test can detect Cohen's d = 0.281 or larger. These are small effects: with 400 students almost anything of a practical size will be found, and a non-significant result is informative.

15.4 Post hoc power: use with care

Post hoc power analysis computes the power of a given sample size for an assumed effect: with N = 387 and d = 0.44, an independent-samples t-test has a power of .991. Use it to *plan* (what could a study of this size detect?), not to explain a result after the fact. Power computed from the effect size observed in the same data is a function of the p-value and adds no information; a non-significant result is better described by its confidence interval (Hoenig & Heisey, 2001).

Exercise: plan a study from the marketing data

In practice_marketing.csv the new offer was answered by 44% of customers and the standard offer by 26%, Cramér's V = .19 (chi-square test, df = 1, N = 1500). Use w = 0.19 and 1 degree of freedom to find the sample size a new test of this effect needs, then the sensitivity of a study with 200 customers. Use the paired result of chapter 6 (d_z = 0.74) to find how many customers a before-and-after comparison needs.

15.5 Planning a hierarchical regression or a factorial design

The tests above plan a whole model. Two more F tests plan the analyses of chapters 9 and 11 more exactly: the R^2 that added predictors contribute (a hierarchical regression, Section 11.5), and one effect of a factorial ANOVA or ANCOVA (Section 9.1). In both, power comes from the noncentral F distribution with $\lambda = f^2 \times N$ (Faul et al., 2007, 2009).

The R^2 increase from added predictors

In Section 11.5, Gender added to three predictors raised R^2 from .142 to .172. How many students would a new study need to detect this increase?

1. Choose **Multiple linear regression, F-test of the R^2 increase**.

2. Type 1 under **Tested predictors** (Gender is one predictor) and 4 under **Total number of predictors** (the three of block 1 and Gender).
3. Open **I know the R^2 values**, type .172 under **R^2 of the full model** and .142 under **R^2 without the tested predictors**, and click **Use this f^2**. SaNaStat computes f^2 = (.172 − .142) / (1 − .172) = 0.0362 and enters it as the effect size. (The two rounded R^2 values give an increase of .030; the ΔR^2 = .031 of Section 11.5 comes from the unrounded values. Typed directly as f^2 = .031 / .828 = 0.0374, it gives N = 212.)

The study needs N = 219 (achieved power .801) (Figure 102). Compare this with the 62 students that the whole model needs (Section 15.2): a test of what one added predictor contributes needs many more people than a test of the whole R^2, because the increase is much smaller than R^2 itself. Use this test whenever the hypothesis is about the added predictors, for example your variables of interest entered after control variables.

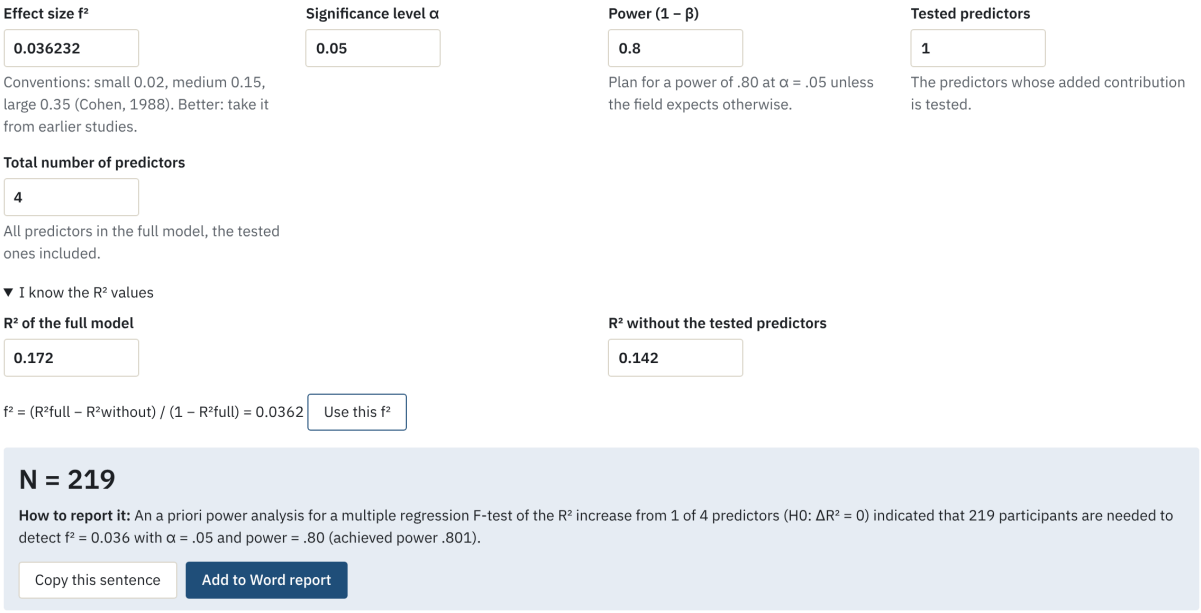


Figure 102. *An a priori power analysis for the R^2 increase of the hierarchical regression in Section 11.5.*

One effect of a factorial ANOVA or ANCOVA

Concept: partial η^2 and Cohen's f

For one effect of a design with several factors, Cohen's f comes from the partial η^2 of that effect: f = √(partial η^2 / (1 − partial η^2)). The **Numerator df** is the number of levels − 1 for a main effect and the product of these for an interaction; the **Number of groups** is the number of cells of the whole design (2 × 3 = 6 for Gender and Teaching_method). An ANCOVA adds its covariates under **Number of covariates**; each one costs a degree of freedom of the error term.

In Section 9.1, Teaching_method had partial η^2 = .022 in the 2 × 3 design, so f = √(.022 / .978) = 0.15.

1. Choose **ANOVA or ANCOVA, one effect of a factorial design**.

2. Type 0.15 for the effect size, 2 under **Numerator df** (three methods − 1), 6 under **Number of groups** and 0 under **Number of covariates**.

SaNaStat reports N = 432 (achieved power .801) (Figure 103). Table 10 plans the other effects of chapter 9 in the same way.

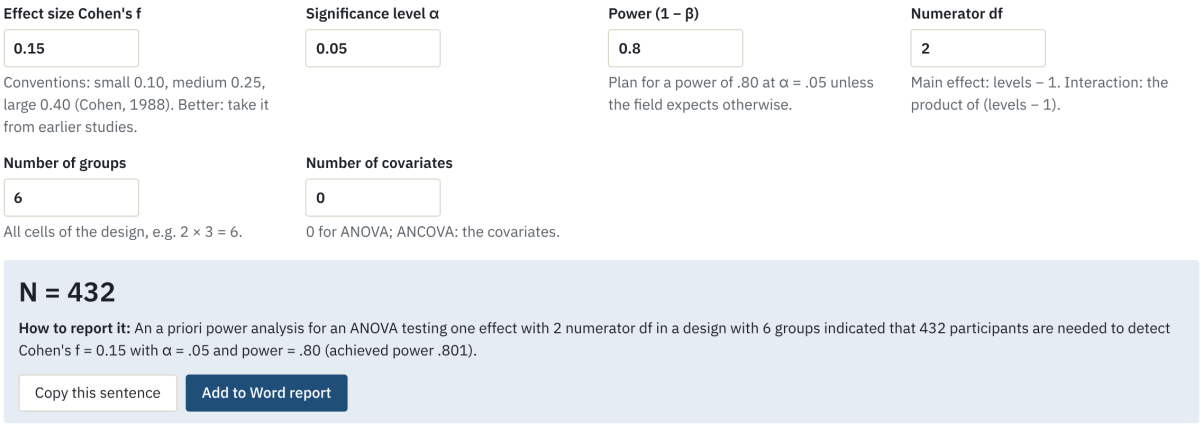


Figure 103. *An a priori power analysis for the effect of Teaching_method in the 2 × 3 factorial design of Section 9.1.*

Table 10

Sample sizes for the effects of chapter 9 (α = .05, power .80, 6 groups)

Effect	Partial η^2	Cohen's f	Numerator df	N
Teaching_method (Section 9.1)	.022	0.15	2	432
Gender (Section 9.1)	.045	0.217	1	169
Gender × Teaching_method (Section 9.1)	.006	0.078	2	1587
Teaching_method, ANCOVA with Pre_score (Section 9.2)	.014	0.119	2 (1 covariate)	684

An interaction is usually much smaller than the main effects, and this one would need about 1,600 students: plan an interaction hypothesis with its own effect size, not with that of a main effect. With the 387 students of chapter 9 and power .80, a **Sensitivity power analysis** shows that the design can detect a two-df effect of f = 0.158 or larger; a study of this size has a power of .753 for an effect of f = 0.15 (**Post hoc power analysis**, used here to plan, as in Section 15.4).

How to report it

"An a priori power analysis for an ANOVA testing one effect with 2 numerator df in a design with 6 groups indicated that 432 participants are needed to detect Cohen's f = 0.15 with α = .05 and power = .80 (achieved power .801)."

Note: designs not covered

SaNaStat does not plan repeated-measures designs (chapter 8), a MANOVA (Section 9.3), mediation or moderation (chapter 13). Their power depends on the correlation between the repeated measures, outcomes or paths, and SaNaStat offers a design only after its results have been checked against published examples. For mediation, Fritz and MacKinnon (2007) give sample sizes for the bootstrap test: about 80 respondents when both paths are medium, more than 500 when both are small.

16. Machine learning as a check on regression

Linear regression assumes that each predictor acts in a straight line and that predictors do not interact. The Machine learning screen of SaNaStat tests that assumption from the outside: it fits flexible models that need no such assumption and asks whether any of them predicts *new* students better than the linear model does. This chapter is a short introduction for that single purpose. It is not a course in machine learning, and a flexible model that predicts well does not tell you why.

Concept: out-of-sample prediction and cross-validation

A model always fits the data it was built on better than it will fit new data, because it partly learns the noise (overfitting). The honest measure is *out-of-sample* performance: fit the model on most of the students, predict the rest, and repeat. In 10-fold cross-validation the students are divided into 10 parts; each part is predicted in turn by a model fitted to the other nine. Repeating this 5 times with different divisions makes the result stable. **RMSE** is the typical size of the prediction error in the units of the outcome (lower is better); **R^2** here is the share of variance predicted in the unseen students. SaNaStat repeats every data-dependent step inside each training set, so nothing leaks from the test rows (nested cross-validation).

16.1 Set up the comparison

1. Click **Machine learning** in the menu.
2. In **Outcome to predict**, choose **Exam_score** (a score, so the task is regression).
3. Under **Predictors**, click **Select none** and tick **Gender**, **Study_hours_week**, **Sleep_hours** and **Anxiety_score**, the same four predictors as the regression of Section 11.5.

SaNaStat reports what it does with missing data: the 13 students without an Exam_score are left out, leaving 387, and the other missing values are filled in within each training fold. It also protects you from a respondent identifier: StudentID, an ID variable, is not offered as a predictor at all (*Not offered: StudentID (type ID: a respondent number would let a model memorise respondents)*). Card 2 sets the validation plan: the recommended plan for 387 respondents is 10-fold cross-validation repeated 5 times (the quick 5-fold plan is for trying things out, the thorough plan with 10 repeats for a result you will report), and SaNaStat estimates the time of each plan before you start. The **Random seed** makes the splits reproducible. Card 3 lists the models: a null model that always predicts the average, linear regression, elastic net, a decision tree (CART), a random forest and two boosting methods, XGBoost and LightGBM; a stacked ensemble can be added.

Click **Run the comparison**. The run took about two minutes on the practice file. The results open in tabs: **Comparison**, **Explanations**, **Calibration & intervals**, **Fairness** and **Model card** (and **Cut-off** for a yes/no outcome, Section 16.5).

16.2 Reading the comparison

Out-of-sample performance (nested cross-validation)

Model	RMSE M (SD)	95% CI	MAE	R^2	Q^2predict	Δ vs best (p)
LightGBM	8.279 (0.886)	[7.633, 8.926]	6.702	.247	.267	best
XGBoost	8.294 (0.917)	[7.625, 8.964]	6.721	.245	.265	0.015 (.821)
Random forest	8.449 (0.919)	[7.778, 9.120]	6.908	.219	.240	0.169 (.165)
Decision tree (CART)	8.684 (0.960)	[7.983, 9.385]	7.041	.172	.195	0.404 (.055)
Elastic net	9.012 (1.120)	[8.194, 9.831]	7.202	.116	.141	0.733 (.013)
Linear regression	9.021 (1.186)	[8.155, 9.887]	7.182	.112	.136	0.741 (.019)
Null model	9.766 (1.077)	[8.980, 10.552]	7.887	−.031	.000	1.487 (< .001)

RMSE M (SD) over 50 test folds; 95% CI with the Nadeau–Bengio corrected variance. Δ vs best: mean per-fold difference from the best model, corrected resampled t-test p.

Figure 104. *Out-of-sample performance of each model.*

Table 11 gives the main numbers of the **Comparison** tab. The null model, which ignores the predictors, has an R^2 of −.031: any useful model must beat it. Linear regression explains 11.2% of the variance in students it has not seen (RMSE = 9.021). LightGBM does best, RMSE = 8.279, 95% CI [7.633, 8.926], R^2 = .247, and its advantage over linear regression is statistically significant, ΔRMSE = −0.741, p = .019 (the corrected resampled t-test, which allows for the overlap between training sets; Nadeau & Bengio, 2003).

Table 11

Out-of-sample prediction of Exam_score (50 test folds; lower RMSE is better)

Model	RMSE (SD)	R^2	Difference from best	p
LightGBM	8.279 (0.886)	.247	best	—
XGBoost	8.294 (0.917)	.245	0.015	.821
Random forest	8.449 (0.919)	.219	0.169	.165
Decision tree (CART)	8.684 (0.960)	.172	0.404	.055
Elastic net	9.012 (1.120)	.116	0.733	.013
Linear regression	9.021 (1.186)	.112	0.741	.019
Null model	9.766 (1.077)	−.031	1.487	< .001

What it means for our example

Two things follow. First, the regression of Section 11.5 explained 17.2% of the variance of Exam_score in the sample it was fitted to, but only 11.2% in students it had not seen: the in-sample figure is optimistic, and the out-of-sample one is the honest estimate. Second, the boosting methods and the random forest predict better than the straight-line model, which suggests that part of the relationship may not be a straight line. That is consistent with chapter 11, where the rank correlation of Study_hours_week with Exam_score was larger than the Pearson correlation. The gain is modest in absolute terms: even the best model leaves about three-quarters of the variance unexplained. SaNaStat also names the simplest model that is not significantly worse than the best, here the decision tree (CART, p = .055): prefer it when interpretability matters.

16.3 Which predictors matter, and how?

Flexible models are opaque, so the **Explanations** tab explains one model with three tools. It opens on the recommended model (the decision tree); choose **LightGBM (best)** in **Explain this model** to explain the best one, as here. **SHAP values** give each predictor's average contribution to the predictions, in points of the outcome (Figure 105). **Permutation importance** shuffles one predictor in the test rows and measures how much the error rises. **Accumulated local effects (ALE)** show how the prediction changes as one predictor changes, allowing for correlated predictors (Apley & Zhu, 2020). The tab repeats that these are *not causal*: they show what the model uses to predict, not what would change Exam_score if a predictor were changed.

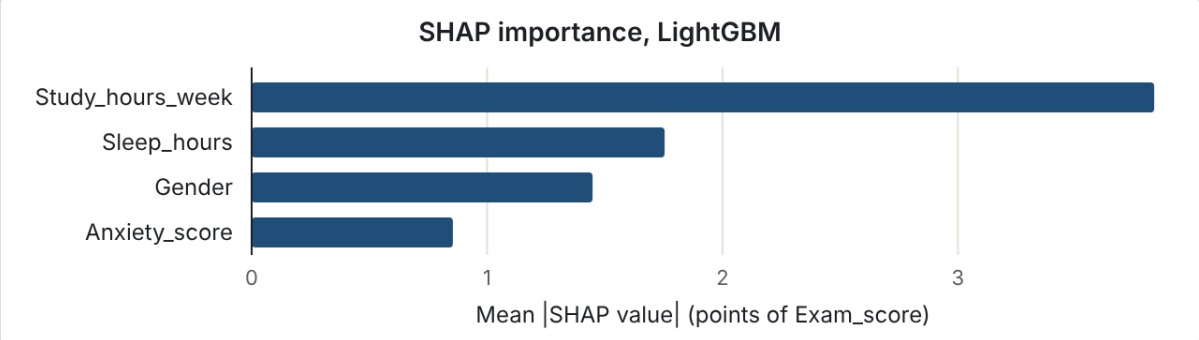


Figure 105. *The SHAP importance of each predictor in the LightGBM model.*

Study_hours_week is the strongest driver (mean |SHAP| = 3.834 points, permutation importance 2.402), followed by Sleep_hours (1.754; 0.513), Gender (1.448; 0.298) and Anxiety_score (0.855; 0.007). Anxiety_score adds almost nothing to prediction on unseen students, which matches its non-significant coefficient in the regression ($\beta = -.08$, $p = .121$). The two rankings agree (Spearman's $\rho = 1.00$), and the other models rank the predictors similarly; the linear model ranks Study_hours_week third (Spearman's ρ between the LightGBM and linear-regression rankings = .40), possibly because of the extreme study times noted in chapter 11. Importance describes what the model uses; it is not evidence of a causal effect (Molnar, 2022).

The ALE plots show how the prediction moves. For Study_hours_week the effect is about −8 points at the lowest values, close to zero between 6 and 8 hours, and about +10 points at the longest times; for Gender the average prediction is 56.10 for women and 53.03 for men.

Note: read the plot as well as the sentence

SaNaStat writes a one-sentence summary of each ALE curve. For Study_hours_week it says that the relationship "is not strictly monotonic: the prediction rises overall (from −8.443 to 10.358) with a small reversal along the way", and for Sleep_hours that "the prediction rises to a peak near 9.70 and then falls". The sentences are still brief: the last interval of the Study_hours_week curve spans 10.7 to 60 hours, so it summarises the few extreme study times in one step. Use the table and the plot to describe the shape.

16.4 Calibration

For a score outcome, the tab **Calibration & intervals** checks calibration: do the predictions have the right scale? Regressing the actual Exam_score on the cross-validated predictions of LightGBM gives a slope of 0.93, 95% CI [0.78, 1.08], and the mean prediction error is −0.011 points. A slope near 1 means the predictions are about as spread out as they should be; a slope well below 1 would be a sign of overfitting. The same tab can compute prediction intervals: a range for each new respondent's score.

How to report it

"Seven prediction models were compared with 10-fold cross-validation repeated 5 times (50 train/test splits), with every model tuned by an inner 5-fold cross-validation inside each training set (nested cross-validation; seed 2026); missing values were imputed and predictors standardised within each training fold. LightGBM performed best (RMSE = 8.279, 95% CI [7.633, 8.926]; R^2 = .247), significantly lower than linear regression, ΔRMSE = −0.741, corrected resampled t(49) = −2.42, p = .019."

16.5 A yes/no outcome: cut-offs, precision and recall

The same screen predicts a yes/no outcome. Choose **Passed** as **Outcome to predict**; SaNaStat asks **Which answer is the event (coded 1)?**, so choose **Yes**. Tick the same four predictors (Gender, Study_hours_week, Sleep_hours and Anxiety_score; never Exam_score, from which Passed was made) and choose the **Quick look** plan, 5-fold cross-validation without repeats, with only the null model, logistic regression and the random forest (Figure 106). The quick plan is fast but noisy; use it to try things out, not to report.

Out-of-sample performance (nested cross-validation)

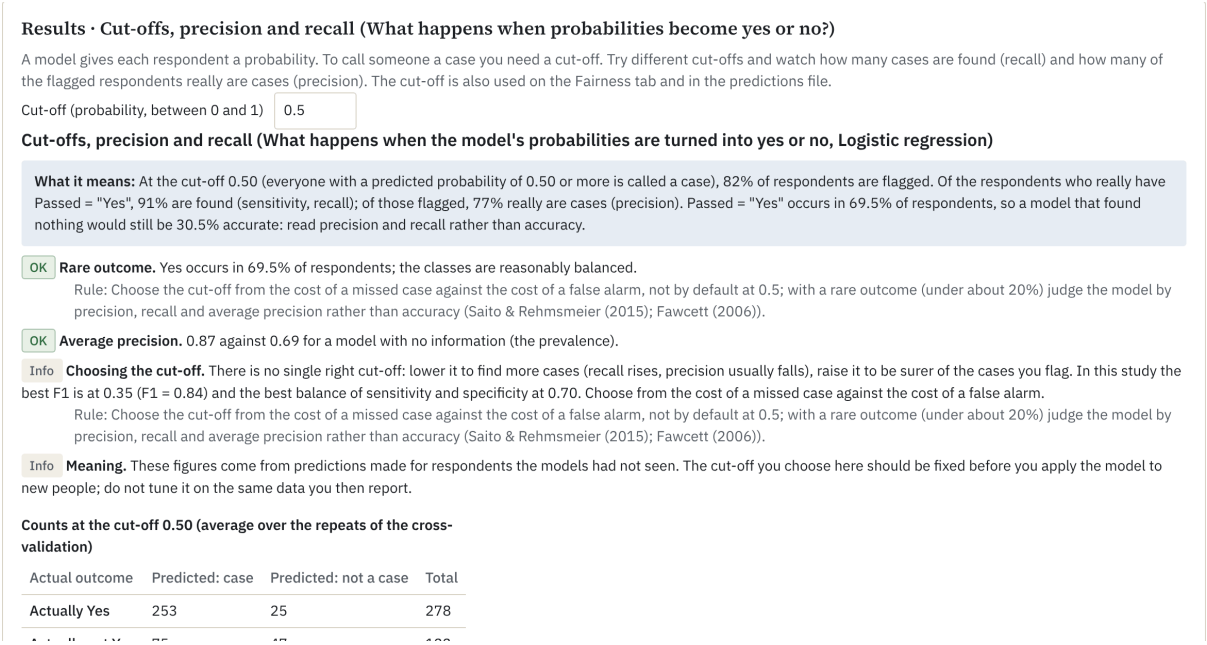
Model	AUC M (SD)	95% CI	Accuracy	Balanced accuracy	F1	MCC	Brier score	Δ vs best (p)
Logistic regression	.782 (.060)	[.671, .892]	.750	.647	.835	.357	0.170	best
Random forest	.748 (.031)	[.690, .806]	.737	.661	.819	.350	0.183	−.034 (.500)
Null model	.500 (.000)	[.500, .500]	.695	.500	.820	.000	0.212	−.282 (.002)

AUC M (SD) over 5 test folds; 95% CI with the Nadeau–Bengio corrected variance. Δ vs best: mean per-fold difference from the best model, corrected resampled t-test p.

Figure 106. *Out-of-sample performance of the three models predicting Passed = Yes (quick look).*

Logistic regression is the best model: AUC = .782 (SD = .060), 95% CI [.671, .892], accuracy = .750 and balanced accuracy = .647. The random forest is not significantly different (AUC = .748, difference −.034, p = .500) and the null model, with an AUC of .500, is clearly worse (p = .002). The interval is wide because there are only 5 test folds. SaNaStat adds that flexible models often need many more respondents than linear models before they pay off, and with 400 students a flexible model does not beat the logistic regression here.

A model gives each student a *probability* of passing. To call a student a case you need a **cut-off**. The tab **Cut-off** (Figure 107) has a number box, **Cut-off (probability, between 0 and 1)**; type a value and everything updates. At 0.50, 82% of the students are flagged; of the students who really passed, 91% are found (**sensitivity**, also called **recall**) and of those flagged, 77% really passed (**precision**). The count table gives the same story in numbers: 253 passers correctly called, 25 missed, 75 failers wrongly flagged and 47 correctly left alone. **Specificity** (failers correctly left alone, 47 of 122) is only 0.39. The table *Measures at every cut-off* and the chart (Figure 108) show how the measures trade off. Raising the cut-off to 0.70 flags 56% of the students, finds 70% of the passers and raises the precision to 87%, with a specificity of 0.75.



Results · Cut-offs, precision and recall (What happens when probabilities become yes or no?)

A model gives each respondent a probability. To call someone a case you need a cut-off. Try different cut-offs and watch how many cases are found (recall) and how many of the flagged respondents really are cases (precision). The cut-off is also used on the Fairness tab and in the predictions file.

Cut-off (probability, between 0 and 1) 0.5

Cut-offs, precision and recall (What happens when the model's probabilities are turned into yes or no, Logistic regression)

What it means: At the cut-off 0.50 (everyone with a predicted probability of 0.50 or more is called a case), 82% of respondents are flagged. Of the respondents who really have Passed = "Yes", 91% are found (sensitivity, recall); of those flagged, 77% really are cases (precision). Passed = "Yes" occurs in 69.5% of respondents, so a model that found nothing would still be 30.5% accurate: read precision and recall rather than accuracy.

OK **Rare outcome.** Yes occurs in 69.5% of respondents; the classes are reasonably balanced.
Rule: Choose the cut-off from the cost of a missed case against the cost of a false alarm, not by default at 0.5; with a rare outcome (under about 20%) judge the model by precision, recall and average precision rather than accuracy (Saito & Rehmsmeier (2015); Fawcett (2006)).

OK **Average precision.** 0.87 against 0.69 for a model with no information (the prevalence).

Info **Choosing the cut-off.** There is no single right cut-off: lower it to find more cases (recall rises, precision usually falls), raise it to be surer of the cases you flag. In this study the best F1 is at 0.35 (F1 = 0.84) and the best balance of sensitivity and specificity at 0.70. Choose from the cost of a missed case against the cost of a false alarm.
Rule: Choose the cut-off from the cost of a missed case against the cost of a false alarm, not by default at 0.5; with a rare outcome (under about 20%) judge the model by precision, recall and average precision rather than accuracy (Saito & Rehmsmeier (2015); Fawcett (2006)).

Info **Meaning.** These figures come from predictions made for respondents the models had not seen. The cut-off you choose here should be fixed before you apply the model to new people; do not tune it on the same data you then report.

Counts at the cut-off 0.50 (average over the repeats of the cross-validation)

Actual outcome	Predicted: case	Predicted: not a case	Total
Actually Yes	253	25	278

Figure 107. *The cut-off box and the summary of the results at a cut-off of 0.50.*

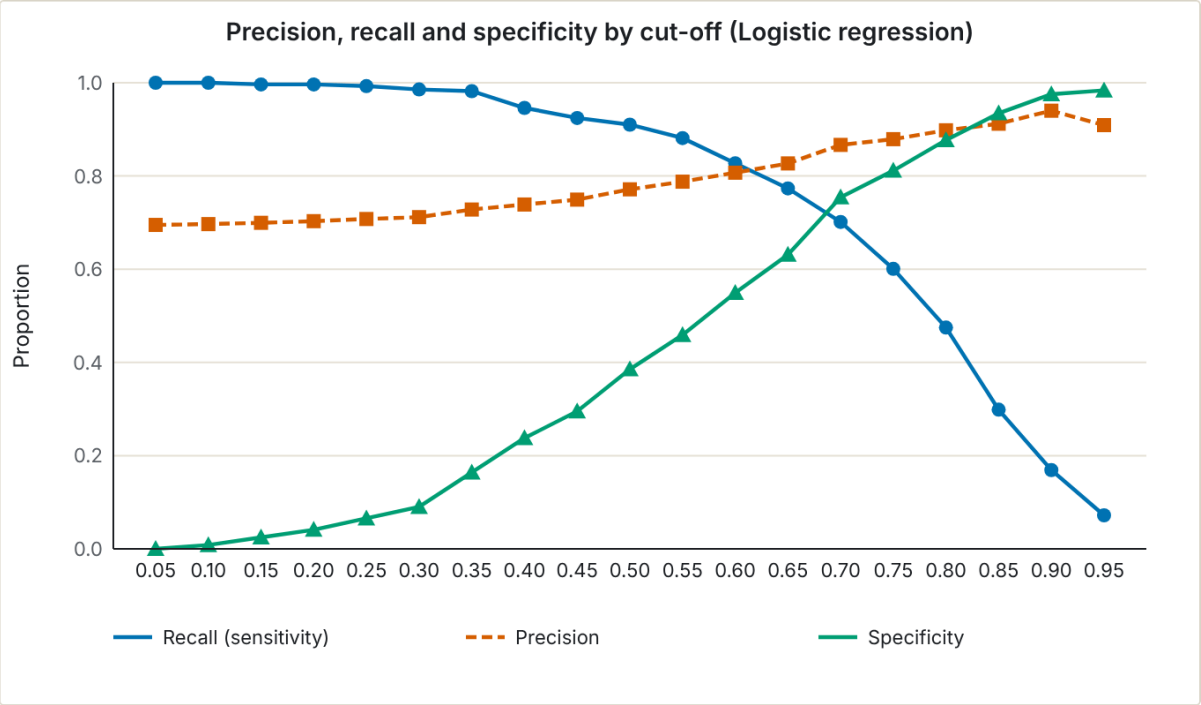


Figure 108. *Precision, recall and specificity at every cut-off.*

SaNaStat also reports the average precision, the area under the precision–recall curve: 0.87, against 0.69 for a model with no information (the prevalence). It checks whether the outcome is rare: passing occurs in 69.5% of the students, so the classes are reasonably balanced; with a rare outcome (under about 20%) accuracy misleads and you should judge a model by precision, recall and average precision (Saito & Rehmsmeier, 2015). In this study the best F1 is at 0.35 (F1 = 0.84) and the best balance of sensitivity and specificity at 0.70. There is no single right cut-off: choose it from the cost of a missed case against the cost of a false alarm (Fawcett, 2006), fix it before you apply the model to new students, and do not tune it on the data you then report. The **Fairness** tab compares the model's errors between groups, for example women and men, and the **Model card** tab summarises the study for a reader and saves the predictions.

How to report it

"On cross-validated predictions of Passed = "Yes", logistic regression at a cut-off of 0.70 had sensitivity = 0.70, specificity = 0.75, precision = 0.87 and F1 = 0.78; the average precision (area under the precision–recall curve) was 0.87, against 0.69 expected by chance (the prevalence; Saito & Rehmsmeier, 2015)."

Exercise: machine learning on another file

In practice_marketing.csv, predict Annual_spend from Income_k, Tenure_months, Loyalty_member, Visits_per_month and Satisfaction_1to10. In the regression of the exercise in chapter 11 these predictors explain 81% of the variance in the sample (R^2 = .81). Compare that figure with the out-of-sample R^2 of linear regression, then with that of the best model, and look at the ALE plot of Visits_per_month, the strongest predictor in the regression. Do the flexible models gain over the linear model?

17. Reporting and exporting your results

Statistics are only useful when other people can read them. This chapter shows how to take single tables and charts out of SaNaStat, how to collect results into a report with the report builder, and gives a short checklist for reporting tests in APA style.

17.1 Copying and exporting one table or chart

Every table of a result has a row of buttons under it (Figure 109):

- **Copy table** copies the table as formatted text: paste it into Word or Excel and it arrives as a table, not as a picture;
- **Copy as picture** copies an image of the table, and **PNG** saves it as a picture file;
- **CSV** saves the numbers as a CSV file (for example *Group_descriptives.csv*), for your own calculations or another program.

Group descriptives

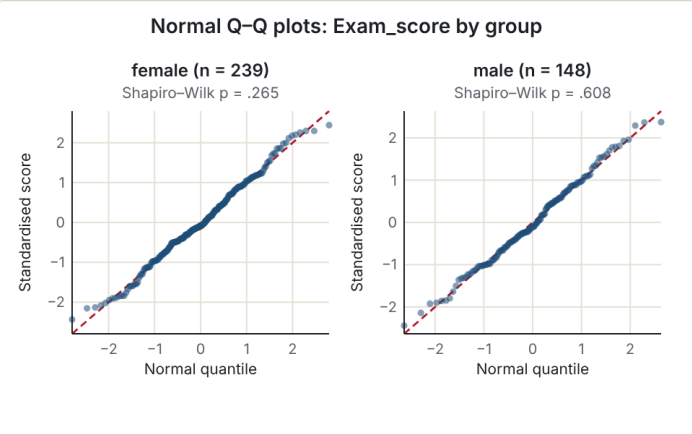
Group	n	Mean	SD	Median	Skewness	Shapiro–Wilk W	p (normality)
female	239	56.57	9.64	55.70	0.07	.992	.265
male	148	52.38	9.57	51.35	0.13	.992	.608

Shapiro–Wilk test of normality per group (Shapiro & Wilk, 1965): p < α suggests the scores are not normally distributed.

Copy table | Copy as picture | PNG | CSV

Figure 109. *A table with its export buttons: Copy table, Copy as picture, PNG and CSV.*

Every chart has **Copy picture**, **PNG** and **SVG** (Figure 110). SVG is a vector format that stays sharp at any size; use it for a thesis or a journal that asks for vector graphics. Under the APA sentence, **Copy this sentence** copies the sentence and **All tables (Excel)** saves every table of the result in one Excel workbook, one sheet per table (here *Independent-samples_t-test_Comparing_Examscore_between_female_and_male.xlsx*).



Normal Q–Q plots: Exam_score by group. Each point is one respondent's standardised score against the value expected under a normal distribution; the dashed line marks a perfect fit. A bend at one end means skew, an S-shape heavy or light tails; Likert-type scores form steps, which is expected. (Wilk & Gnanadesikan, 1968)

Copy picture | PNG | SVG

Figure 110. *A chart (the Q–Q plots of the t-test) with Copy picture, PNG and SVG.*

17.2 Collecting results in the Word report

Every result in SaNaStat (an analysis, a chart or a power analysis) ends with the sentence that you would put in your text, with **Add to Word report** (Figure 111). Adding a result marks it with the label *In the Word report* and a link to the report. Add the results as you go; the report keeps them in the order added.

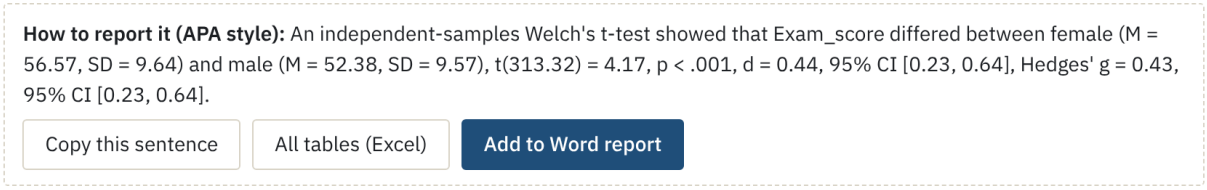


Figure 111. *The APA sentence with Copy this sentence, All tables (Excel) and Add to Word report.*

Click **Word report** in the menu to open the report builder (Figure 112). The first card, *Word report (5 results)*, lists the collected results with their finding:

- untick a result to leave it out (it stays in the list, crossed out);
- the arrows move a result up or down (tables and figures are numbered in this order), and **Remove** drops it;
- **Parts of this result** chooses what goes in for that result alone, for example the APA sentence, the interpretation and the tables (5) but not the figures (1) (Figure 113).

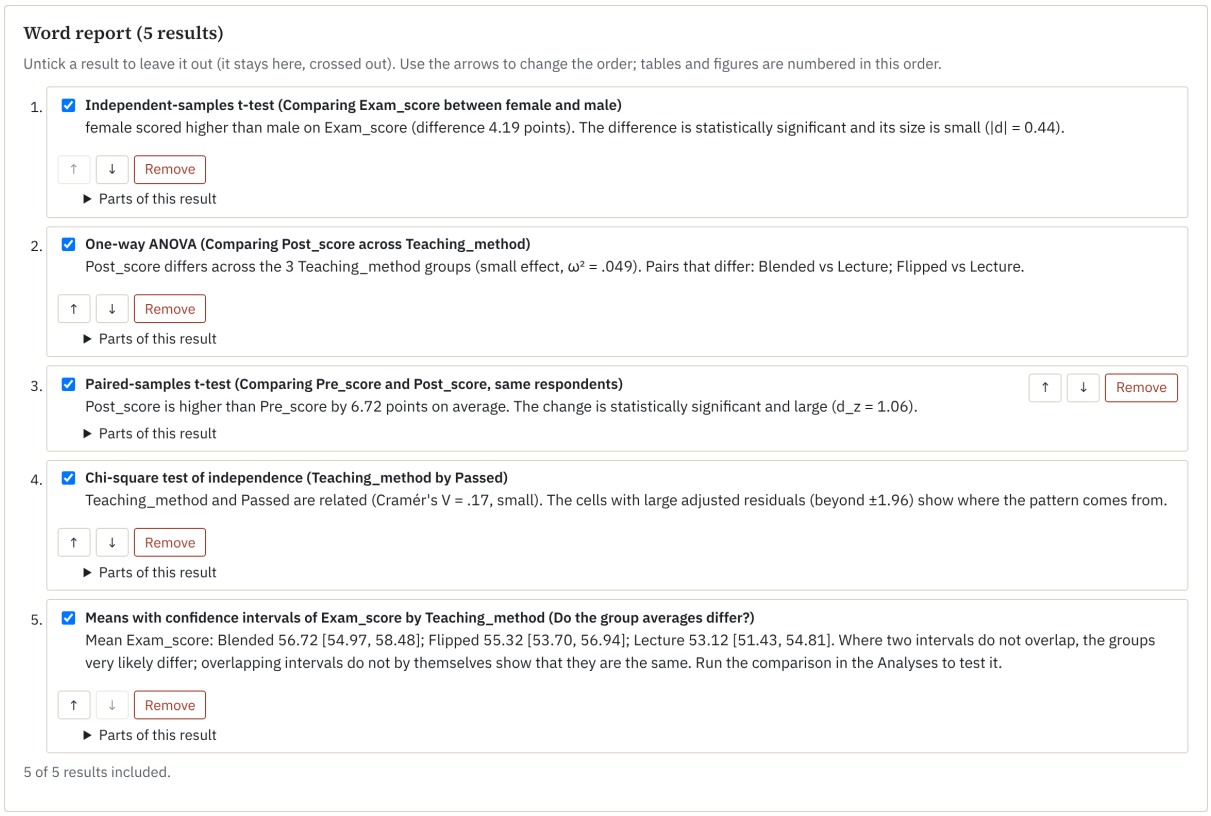


Figure 112. *The report builder: five results in the order chosen.*

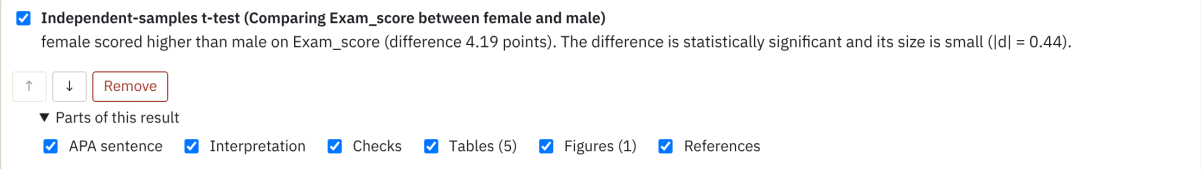


Figure 113. *The parts of one result: here the t-test has five tables and one figure.*

The cards below build the report as a whole:

- **What goes in** (Figure 114): the parts of every result (APA sentence, Interpretation, Checks, Tables, Figures, References) and the **Methods and reporting checklist**, left out or added before or after the results. The checklist covers the data, the flow of respondents, missing data, the constructs and their reliability, α, the program and the random seeds; items only you know are marked "To complete by the author".
- **Style** (Figure 115): the **Report title**, the font (Times New Roman 12 pt, Calibri 11 pt or Arial 11 pt, all accepted by APA 7), the **Line spacing** (single, 1.5 or double, as APA 7 asks for manuscripts and student papers), the **Paper** (A4 or US Letter) and an optional **Title page** with the title, authors, affiliation and date.
- *Preview and download* (Figure 116): **Download Word report (.docx)**, **Download all tables (Excel)**, **Download as a web page (HTML, prints as PDF)** and **Show the preview**, which shows the report as it will look. The Word file is named after the title (here "Common statistical tests practice students (SaNaStat).docx").

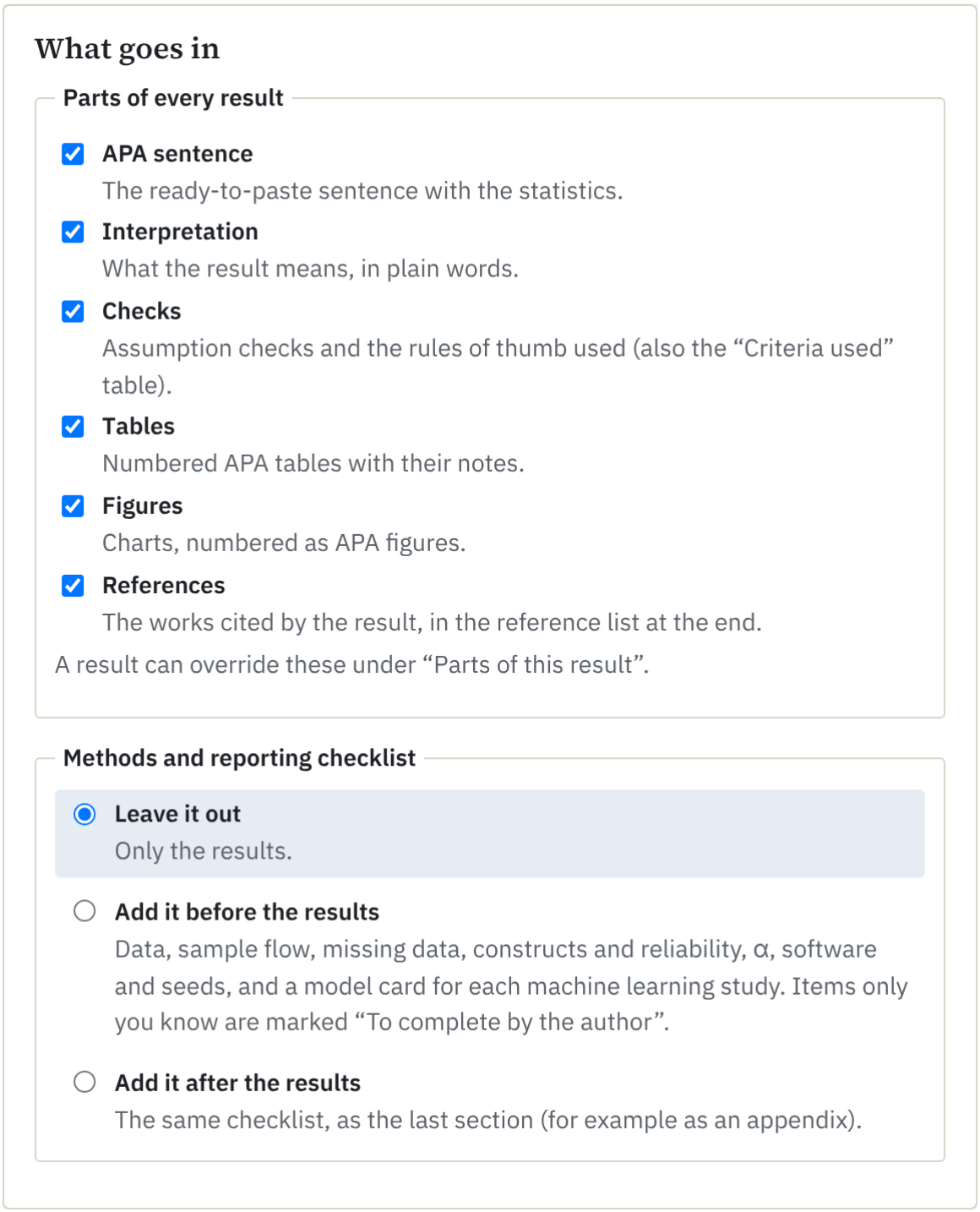


Figure 114. *What goes in: the parts of every result and the methods and reporting checklist.*

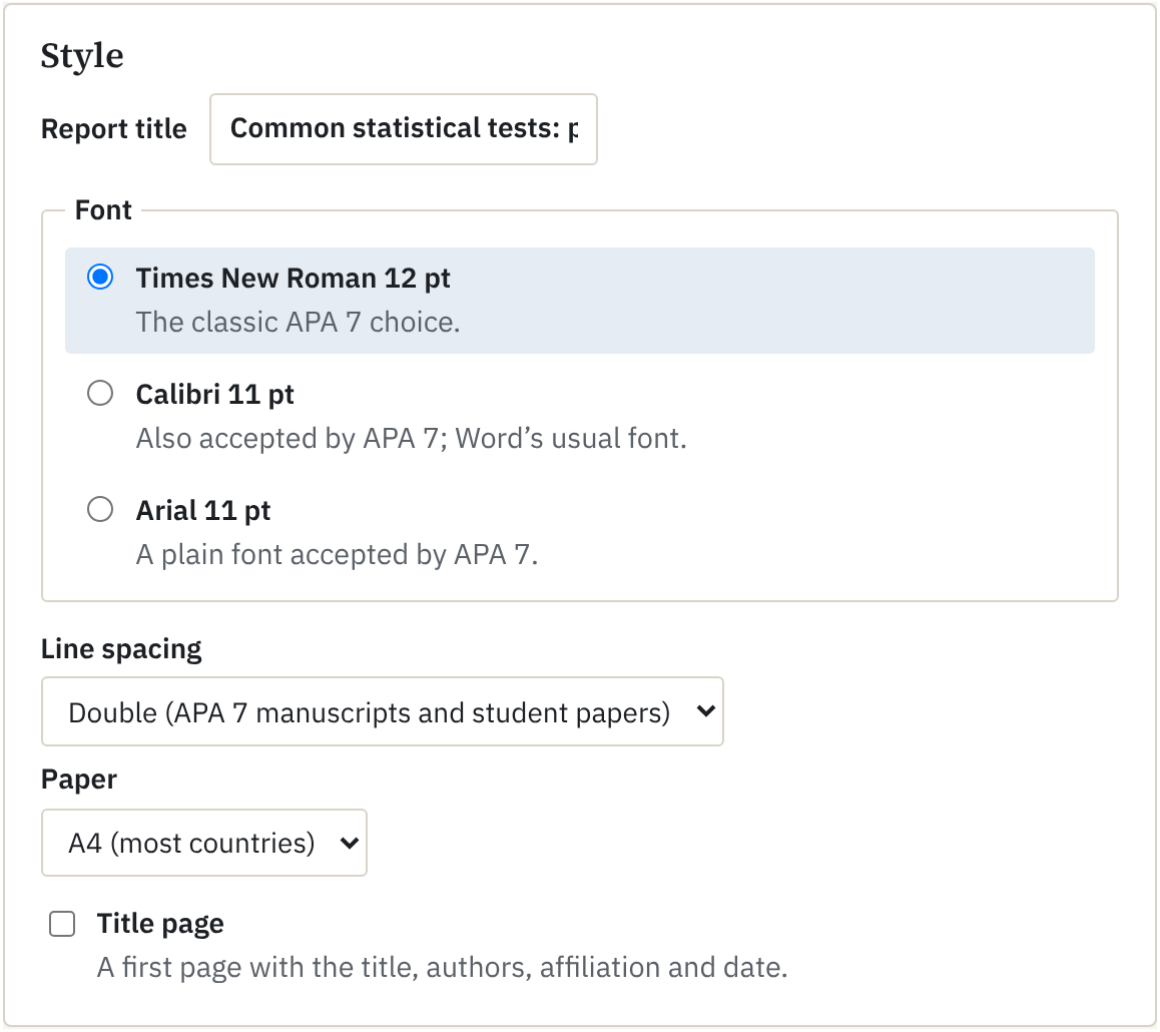


Figure 115. *Style: title, font, line spacing, paper and title page.*

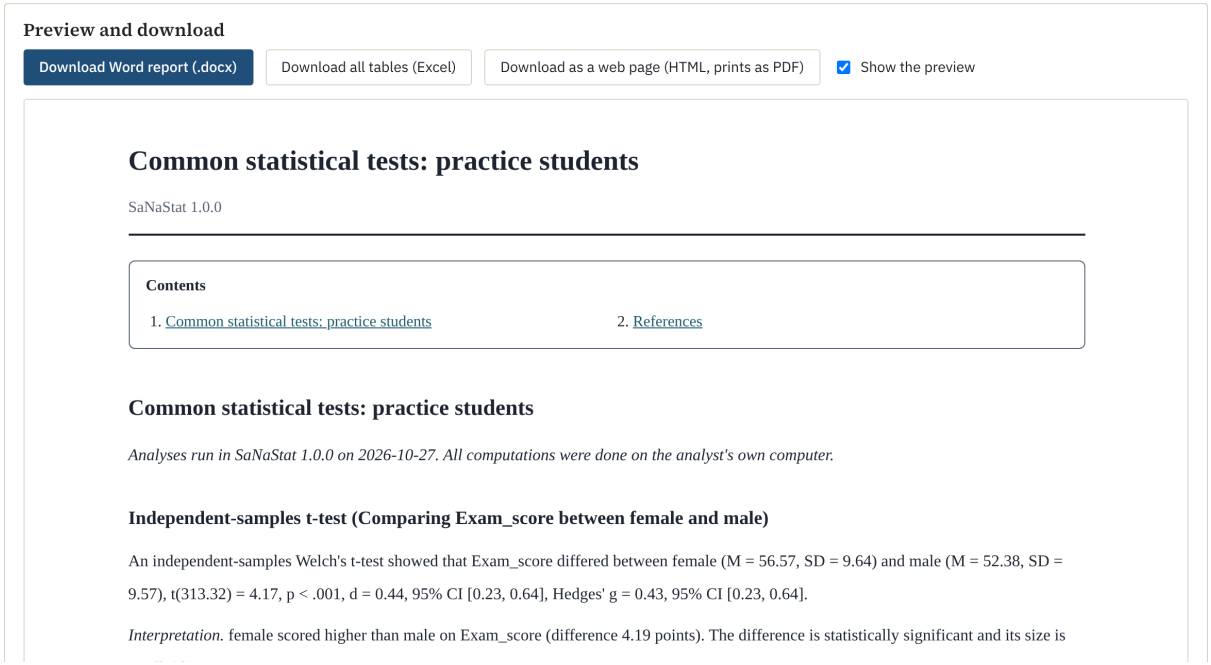


Figure 116. *Preview and download, with the preview of the report.*

The document for the five results of this manual's practice session (an independent-samples t-test, a one-way ANOVA, a paired-samples t-test, a chi-square test and a chart of means) contains, in this order:

- the report title, and a line stating that the analyses were run in SaNaStat 1.0.0 on the date of the download and that all computations were done on the analyst's own computer;
- for each result, a heading, the APA sentence, an *Interpretation*, the *Checks* that were run, the numbered tables (descriptives, the tests, the effect sizes with their confidence intervals, the rank-based alternative and the notes under each table) and the figures with their captions;
- a table of *Criteria used*, listing each rule of thumb that was applied in the report and its source;
- the *References* in APA style, in alphabetical order.

Your choices (parts, checklist, style) are kept with the project in this browser.

Tip: check before you submit

The sentences and tables are a starting point. Read each against your design: the sentence names the groups in the file's own wording, a one-tailed test is marked and worded as a direction ("was higher for A than for B"), and the sentence cannot know which hypotheses you set in advance. Edit the document in your word processor, keep the numbers as SaNaStat gives them and add the context that only you have: the sample, the design and what the result means for your question.

17.3 A reporting checklist

1. **Describe the data.** State the sample size, the population, how the sample was obtained and whether the data are real or simulated. Give the descriptive statistics (M and SD, or the median and range for skewed variables) for each variable analysed, and frequencies with valid percentages for categorical variables.
2. **Report how you handled problems.** Say how many respondents were excluded and why (chapter 3), how missing values were treated (each analysis states its n; chapter 14 for multiple imputation), whether scales were reversed or combined, which variables were computed, transformed or recoded, and whether cases were selected (chapters 1 and 2).
3. **Report the measurement.** Give α and ω for each scale and, for a new or adapted questionnaire, the factor analysis with its fit and validity (chapter 12).
4. **Name the test and the reason.** For example, "Welch's t-test, because the group sizes were unequal", and state α if it is not .05.
5. **Report the statistic, the degrees of freedom, the exact p-value and an effect size with its confidence interval.** For example, t(313.32) = 4.17, p < .001, d = 0.44, 95% CI [0.23, 0.64].

6. **Report adjustments.** Say which correction was used for the follow-up tests (Tukey–Kramer, Games–Howell, Bonferroni, Dunn or Holm) and for the family of tests, and report how many tests you ran.
7. **Interpret the size and check the assumptions.** Use the benchmarks (Appendix B) as a guide, judge importance in the context of your field, and say in one sentence which checks were made and which alternative was used when one failed.
8. **Make it reproducible.** Report the number of bootstrap samples, imputations and the random seeds (chapters 13, 14 and 16).

Cut-off values: formatting numbers in APA style

Report p to three decimals and write "p < .001" for smaller values. Leave out the leading zero for values that cannot exceed 1 (p, r, V, φ, η^2, ω^2, R^2, α, AVE) and keep it for d, g, f, w, B and the HTMT. Italicise the statistical symbols in the final document. SaNaStat's sentences and tables already follow this style. Use the same number of decimals across a table.

Note: simulated data

The results in this manual come from a simulated file made for practice. The p-values, effect sizes and conclusions describe the simulated students only and must not be quoted as findings about real students. When you report your own data, replace every example with your own output.

Exercise: build a report

Open practice_marketing.csv. Run (a) the independent-samples t-test of Annual_spend between the two Loyalty_member groups, (b) the chi-square test of Campaign_group by Responded and (c) the multiple linear regression of Annual_spend of chapter 11. Add the three to the Word report, give the report a title, move the chi-square test to the first place, add the methods and reporting checklist after the results, choose double line spacing and download the document. Check that the chi-square result in the file reads $\chi^2(1, N = 1500) = 54.50$, $p < .001$, Cramér's V = .19, 90% CI [.15, .23]. Then copy the coefficient table of the regression and paste it into a word processor.

Appendix A. Codebook of practice_students.csv

The file holds 400 simulated students. The values were generated by a computer program; there are no real persons in it. Table 12 lists every variable with its type in SaNaStat, its values or range and the number of missing values.

Table 12

The variables of practice_students.csv (N = 400)

Variable	Type	Values or range	Missing
StudentID	ID	ST001 to ST400; a label only	0
Gender	categorical	female (246), male (154)	0
Age	scale	17 to 24 years; M = 20.23, SD = 1.40	0
Year	scale	1 to 4 (treated as categorical in Section 1.3 and recoded in Section 2.3); M = 2.38	0
Programme	categorical	Arts (92), Business (163), Science (145)	0
Residence	categorical	Home (195), Hostel (119), Rented (74)	12
Teaching_method	categorical	Blended (129), Flipped (129), Lecture (142)	0
Study_hours_week	scale	0.7 to 60; M = 7.28, SD = 5.80	0
Sleep_hours	scale	3.5 to 9.9; M = 6.68, SD = 1.07	12
Anxiety_score	scale	7.1 to 26.2; M = 13.50, SD = 2.90	22
Pre_score	scale	25.3 to 95.0; M = 59.37, SD = 11.65	0
Post_score	scale	28.3 to 100.0; M = 66.09, SD = 13.47	0
Score_T1, Score_T2, Score_T3	scale	Three measurements of the same test; M = 59.37, 62.55 and 66.09	0
Exam_score	scale	29.0 to 80.1; M = 54.97, SD = 9.81	13
Project_score	scale	32.2 to 99.2; M = 59.85, SD = 12.02	0
Passed	categorical	Yes (278), No (122); the pass/fail version of Exam_score, pass mark 50	0
Pass_before	categorical	Yes (319), No (81); the pass/fail version of Pre_score	0
Pass_after	categorical	Yes (358), No (42); the pass/fail version of Post_score	0
Satisfaction_1to5	scale (ordinal in Sections 1.4 and 11.7)	One rating, 1 to 5; M = 3.05, SD = 1.00; 1 (24), 2 (85), 3 (150), 4 (98), 5 (27)	16
Mot1 to Mot6	scale	Motivation items, 1 to 5	0
Anx1 to Anx6	scale	Anxiety items, 1 to 5; Anx3 and Anx5 are worded in the opposite direction	0

Exam_score is missing for 13 students, but Passed is filled in for all of them (11 Yes and 2 No), so Passed has no missing values although Exam_score has. For analyses that involve Passed, N is 400; for analyses that involve Exam_score, at most 387. The file uses blank cells for missing values and contains no missing-value codes.

Note: what the practice file does not contain

The file has no timing variable (so the response-time index of chapter 3 cannot be used), no missing-value codes (Section 1.5), no planted interaction (Section 13.2) and no repeated measures combined with a group factor (chapter 8). The exercises point to the other practice files where these are needed.

Appendix B. Assumptions and benchmarks of the tests

Table 13

The tests of this manual: assumptions, effect sizes and benchmarks

Test	Main assumptions	Effect size and benchmarks
Independent-samples t (Welch)	Independent observations; roughly normal sample means (about 30 or more per group is enough; Shapiro–Wilk and Q–Q plots); Welch needs no equal variances	Cohen's d, Hedges' g: .2 small, .5 medium, .8 large; 95% CI
Paired-samples t	Independent pairs; roughly normal differences	d_z, g_z, same benchmarks
One-sample t	Independent observations; roughly normal sample mean	d, same benchmarks
Mann–Whitney U, Wilcoxon	Independent cases or pairs; similar shapes if medians are compared	Rank-biserial r: .1, .3, .5
One-way ANOVA	Independent observations; roughly normal means; equal variances (otherwise Welch and Games–Howell)	η^2, ω^2: .01, .06, .14; 90% CI
Kruskal–Wallis	Independent observations; similar shapes	ε^2
Repeated-measures ANOVA	Sphericity (Mauchly); if violated, Greenhouse–Geisser or Huynh–Feldt	Partial and generalised η^2
Factorial ANOVA, ANCOVA	As ANOVA; for ANCOVA, parallel slopes of the covariate in all groups	Partial η^2; 90% CI
MANOVA	Independent observations; multivariate normality; similar covariance matrices	Pillai's V; partial η^2 for each follow-up ANOVA
Chi-square	Independent observations; expected counts not below 1 and at most 20% below 5	Cramér's V (φ for 2 × 2): .1, .3, .5; w: .1, .3, .5; 90% CI
McNemar	Paired yes/no answers	Odds ratio b / c; Cohen's g: .05, .15, .25
Pearson r	Linear relationship; no extreme values	r: .1, .3, .5
Spearman ρ, Kendall τ-b	Monotonic relationship	ρ and τ read as r

Test	Main assumptions	Effect size and benchmarks
Linear regression	Linear effects; independent, normal errors with constant spread; no strong collinearity (VIF below 5); no influential cases	R^2; f^2: .02, .15, .35
Logistic regression	Independent observations; at least 10 cases in the smaller outcome group per coefficient	Odds ratios; Nagelkerke R^2; AUC: .7 acceptable, .8 excellent, .9 outstanding
Ordinal logistic regression	Proportional odds (Brant test not significant); at least 10 cases in the smallest category per coefficient	Odds ratios; Nagelkerke R^2
Multinomial logistic regression	Independence of irrelevant alternatives; at least 10 cases in the smallest category per coefficient	Relative risk ratios; Nagelkerke R^2; classification
Reliability	Items measure one construct; reverse-worded items reversed	α and ω ≥ .70; item–total r ≥ .30
Exploratory factor analysis	Correlated items (KMO ≥ .60, Bartlett p < .05); 100 or more respondents	Loadings ≥ .40, cross-loadings < .30; parallel analysis
Confirmatory factor analysis	Each item on one factor; complete cases; roughly normal items	CFI, TLI ≥ .95; RMSEA ≤ .06; SRMR ≤ .08; AVE ≥ .50; HTMT < .85
Mediation, moderation	As linear regression; for mediation, a defensible causal order	Bootstrap CI of ab excludes 0; ΔR^2 of the interaction
Multiple imputation	Missing at random; imputation model contains every analysed variable	FMI; m at least the percentage of incomplete cases

The rules behind the checks are: Welch's t-test by default (Delacre et al., 2017); about 30 or more cases per group before non-normality stops mattering (Lumley et al., 2002); Shapiro–Wilk for normality (Shapiro & Wilk, 1965) with Q–Q plots (Wilk & Gnanadesikan, 1968); Levene's test for equal variances (Levene, 1960), with Games–Howell comparisons when it fails (Games & Howell, 1976); 100(1 − 2α)% intervals for η^2, ω^2, partial η^2, V, φ and w (Steiger, 2004); Greenhouse–Geisser when ε < .75 and Huynh–Feldt otherwise (Girden, 1992); expected counts of at least 1, and at most 20% below 5, in a chi-square table (Cochran, 1954); at least 50 + 8k respondents for k regression predictors (Green, 1991); VIF below 5 (Hair et al., 2019; Kutner et al., 2005); Breusch–Pagan (Koenker) for constant error variance (Breusch & Pagan, 1979; Koenker, 1981); Cook's distance below 1 (Cook & Weisberg, 1982); at least 10 cases in the smaller outcome group per coefficient (Peduzzi et al., 1996); AUC and Hosmer–Lemeshow (Hosmer et al., 2013); the Brant (1990) test of proportional odds; α and ω of .70 (Nunnally, 1978; McNeish, 2018); KMO (Kaiser, 1974), parallel analysis (Horn, 1965; Glorfeld, 1995) and clean loadings (Howard, 2016); fit cut-offs (Hu & Bentler, 1999; Browne & Cudeck, 1993); AVE and the Fornell–Larcker criterion (Fornell & Larcker, 1981); HTMT (Henseler et al., 2015); the percentile bootstrap for indirect effects (Preacher & Hayes, 2008; Hayes, 2022); the index of moderated mediation (Hayes,

2015); Johnson–Neyman regions (Johnson & Neyman, 1936; Bauer & Curran, 2005); the number of imputations (White et al., 2011; Graham et al., 2007) and Rubin's (1987) rules. The benchmarks are Cohen's (1988) conventions as SaNaStat applies them; they are rules of thumb, not laws, and the importance of an effect depends on the field. For Kruskal–Wallis, partial η^2 and Pillai's V, SaNaStat reports the value and the other statistics, and its screens do not print a benchmark for all of them, so use the field's own conventions.

Appendix C. Glossary

Table 14

Terms used in this manual

Term	Meaning
Adjusted p-value	A p-value corrected for the number of tests in the family (Holm, Tukey–Kramer, Games–Howell and others).
Alpha (α)	The significance level, the false-positive rate you accept; also Cronbach's α, a reliability coefficient (the context shows which).
ANCOVA	ANOVA with a covariate, which adjusts the group means for a related scale variable.
Auxiliary variable	A variable added to an imputation model to predict the missing values better, without being analysed.
Average variance extracted (AVE)	The mean squared standardised loading of a construct's items; .50 or more shows convergent validity.
Bootstrap	Resampling the respondents with replacement many times to obtain the sampling distribution of a statistic, such as an indirect effect.
Brant test	A test of the proportional-odds assumption of an ordinal logistic regression.
Calibration	Whether predictions have the right scale; a calibration slope of 1 is ideal.
Communality (h^2)	The share of an item's variance explained by the factors.
Confidence interval	A range of values compatible with the data at the stated confidence (here 95%, or 90% for η^2, V and similar effect sizes).
Covariate	A scale variable that is related to the outcome and is held constant in an ANCOVA or regression.
Cramér's V	Effect size for a table of counts, from 0 to 1.
Cut-off (probability)	The predicted probability from which a respondent is called a case; it decides sensitivity, specificity and precision.
Direct effect (c')	The effect of X on Y that does not go through the mediator.
Effect size	The size of a difference or relationship, independent of the sample size.
Estimated marginal mean	A group mean adjusted for the other terms of the model.
Fraction of missing information (FMI)	The share of the uncertainty of an estimate that is due to the missing values.
Hedges' g	Cohen's d corrected for its small-sample bias.
Holm adjustment	A step-down correction for multiple tests that is never less powerful than Bonferroni's.
HTMT	The heterotrait–monotrait ratio of correlations; below .85 supports discriminant validity.
ID variable	A respondent number or code; SaNaStat shows it but never offers it to an analysis.

Term	Meaning
Index of moderated mediation	How much an indirect effect changes per unit of the moderator.
Indicator variable	A variable that is 1 for a chosen category and 0 for all others.
Indirect effect (ab)	The part of the effect of X on Y that passes through a mediator.
Interaction	An effect of one factor that depends on the level of another.
Johnson–Neyman technique	Finds the values of a moderator at which the effect of X changes between significant and not significant.
Levene's test	A test of equal variances in groups.
Loading	The relationship of an item with a factor.
McDonald's ω	A reliability coefficient from a one-factor model; it does not assume equal loadings.
Median split	Two groups made at the median of a scale variable; a value equal to the median goes to the lower group.
Missing at random (MAR)	Missingness that may depend on observed variables but not on the missing value itself.
Missing completely at random	Missingness that is unrelated to the data (Little's test).
Missing-value code	A number such as −99 that stands for "no answer" and must be set to missing.
Multiple imputation	Filling in each missing value several times and pooling the analyses with Rubin's rules.
Natural logarithm	ln(x); a transformation that pulls in a long right tail and needs positive values.
Odds ratio	The multiplicative change in the odds of an event for one unit of a predictor.
Ordinal variable	Categories in a meaningful order, such as a 1 to 5 rating treated as categories.
Parallel analysis	Keeps the factors whose eigenvalues exceed those of random data of the same size.
Power	The probability of finding an effect that really exists.
Precision	The share of the respondents flagged as cases who really are cases.
Predictive mean matching	An imputation method that takes each missing value from an observed respondent with a similar predicted value.
Q–Q plot	A plot of the observed scores against the values expected under a normal distribution.
Relative risk ratio (RRR)	In a multinomial model, the change in the odds of a category rather than the reference for one unit of a predictor.
Select cases	A filter that restricts every analysis to the respondents meeting a condition.
Sensitivity (recall)	The share of the real cases that the model finds.

Term	Meaning
Shapiro–Wilk test	A test of whether scores come from a normal distribution.
Specificity	The share of the non-cases that the model correctly leaves alone.
Sphericity	The assumption that the differences between all pairs of repeated measures have equal variance.
Straight-lining	Giving the same answer to every item of a questionnaire.
Type III sums of squares	Each effect is tested after all the others; the usual choice for unbalanced factorial designs.
VIF	Variance inflation factor, a measure of collinearity among predictors.

References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. Behavior Research Methods, 37(3), 379–384. https://doi.org/10.3758/BF03192707

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4), 948–955. https://doi.org/10.1093/biomet/86.4.948

Bartlett, M. S. (1937). The statistical conception of mental factors. British Journal of Psychology, 28(1), 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Statistical Psychology, 3(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. Multivariate Behavioral Research, 40(3), 373–400. https://doi.org/10.1207/s15327906mbr4003_5

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 46(4), 1171–1178. https://doi.org/10.2307/2532457

Breslow, N. E., & Day, N. E. (1980). Statistical methods in cancer research: Vol. 1. The analysis of case-control studies. International Agency for Research on Cancer.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287–1294. https://doi.org/10.2307/1911963

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Sage.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10(4), 417–451. https://doi.org/10.2307/3001616

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. Chapman & Hall.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology, 66, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology, 30(1), 92–101. https://doi.org/10.5334/irsp.82

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), 1096–1121. https://doi.org/10.1080/01621459.1955.10501294

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. Biometrika, 37(3/4), 409–428. https://doi.org/10.2307/2332391

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), Mathematical methods for digital computers (pp. 191–203). Wiley.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. https://doi.org/10.3758/BF03193146

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. Psychological Science, 18(3), 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.x

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n’s and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1(2), 113–125. https://doi.org/10.3102/10769986001002113

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7(4), 457–472. https://doi.org/10.1214/ss/1177011136

Girden, E. R. (1992). ANOVA: Repeated measures. Sage.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. Educational and Psychological Measurement, 55(3), 377–393. https://doi.org/10.1177/0013164495055003002

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24(2), 95–112. https://doi.org/10.1007/BF02289823

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. Journal of Applied Statistics, 36(10), 1109–1118. https://doi.org/10.1080/02664760802553000

Harrell, F. E., Jr. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

Hayes, A. F. (2015). An index and test of linear moderated mediation. Multivariate Behavioral Research, 50(1), 1–22. https://doi.org/10.1080/00273171.2014.962683

Hayes, A. F. (2022). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (3rd ed.). Guilford Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. https://doi.org/10.3102/10769986006002107

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19–24. https://doi.org/10.1198/000313001300339897

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. Quality & Quantity, 44(1), 153–166. https://doi.org/10.1007/s11135-008-9190-y

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185. https://doi.org/10.1007/BF02289447

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? International Journal of Human–Computer Interaction, 32(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1(1), 69–82. https://doi.org/10.3102/10769986001001069

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs, 1, 57–93.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31–36. https://doi.org/10.1007/BF02291575

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. Comprehensive Psychology, 3, 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

Keuls, M. (1952). The use of the "studentized range" in connection with an analysis of variance. Euphytica, 1(2), 112–122. https://doi.org/10.1007/BF01908269

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. Journal of Econometrics, 17(1), 107–112. https://doi.org/10.1016/0304-4076(81)90062-2

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. Biometrics, 12(3), 307–310. https://doi.org/10.2307/3001469

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (5th ed.). McGraw-Hill/Irwin.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics (pp. 278–292). Stanford University Press.

Li, K. H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. Statistica Sinica, 1(1), 65–92.

Little, R. J. A. (1988a). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3), 287–296. https://doi.org/10.1080/07350015.1988.10509663

Little, R. J. A. (1988b). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician, 54(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual Review of Public Health, 23, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, 111(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling, 11(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. The Annals of Mathematical Statistics, 11(2), 204–209. https://doi.org/10.1214/aoms/1177731915

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. Psychological Methods, 12(1), 23–44. https://doi.org/10.1037/1082-989X.12.1.23

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). Designing experiments and analyzing data: A model comparison perspective (3rd ed.). Routledge.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. Psychological Bulletin, 114(2), 376–390. https://doi.org/10.1037/0033-2909.114.2.376

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109–142. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17(3), 437–455. https://doi.org/10.1037/a0028085

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Independently published.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. Biometrika, 31(1–2), 20–30. https://doi.org/10.1093/biomet/31.1-2.20

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 83(4), 579–586. https://doi.org/10.1037/0033-2909.83.4.579

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. Applied Statistics, 39(2), 205–217. https://doi.org/10.2307/2347760

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley. https://doi.org/10.1002/9780470316696

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40(1–2), 87–110. https://doi.org/10.1093/biomet/40.1-2.87

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62(318), 626–633. https://doi.org/10.1080/01621459.1967.10482935

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology, 13, 290–312. https://doi.org/10.2307/270723

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological Methods, 9(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 221–257). Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson.

Thurstone, L. L. (1935). The vectors of mind. University of Chicago Press.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

van Buuren, S. (2018). Flexible imputation of missing data (2nd ed.). CRC Press. https://doi.org/10.1201/9780429492259

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38(3/4), 330–336. https://doi.org/10.2307/2332579

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377–399. https://doi.org/10.1002/sim.4067

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. Biometrika, 55(1), 1–17. https://doi.org/10.1093/biomet/55.1.1