EMMA NEUROSCIENCE GROUP

GUIDELINE – PREPROCESSING VERSION 1.0

GENERAL

In data analysis we distinguish between pre-processing of data and the actual analysis where statistical tests are deployed. Pre-processing makes the data better suitable for statistical analysis. Pre-processing can involve integration, exploration, cleaning, imputation, transformation and reduction of data. Here we provide a detailed description of the preferred approach to data processing. We recommend the book by Field for a practical guide during pre-processing.

Field, A. (2022). *Discovering statistics using R and RStudio*. Sage.

Chicago

1. CHOOSE YOUR SOFTWARE

a. Before you start with your data, decide which software you are going to use.
    i. SPSS is the most widely used statistical software. SPSS is user-friendly and allows to perform almost all conventional analyses. SPSS is not suitable for more advanced analysis (e.g. artificial intelligence).
    ii. R is open source software that is often used for more advanced analyses. The software is controlled by the command line or in a script. When you are a beginner, this may be a little bit scary. You can use R with R studio for a user interface that makes working with R much more user-friendly. The main advantage of R is that is has a massive library for statistical packages, allowing you to perform almost all analyses that you can dream of. In the Emma Neuroscience Group, R is the preferred software to perform more advanced analyses.
    iii. Python is open source software that is popular in the data science field. Python is comparable to R, yet it may have higher efficiency for complex computational procedures. Often, data scientists work with R ánd Python, combining their favorite statistical packages from both sources.

b. The Emma Neuroscience Group advises to use SPSS for conventional statistics and to use R for more advanced analyses (e.g. cluster analysis). It is recommended to use R for visualization of your data (link). A template of a pre-processing pipeline in R is available from: . You can download the folder 'pre-processing', place it in your own environment and add your databases to start right away.

2. WORK WITH SYNTAXES

a. Always work with syntaxes or scripts. Preprocessing steps are labor intensive and often require tweaking during the process. Therefore, you may want to change something in the preprocessing steps that influences all downstream analyses. When you work with syntaxes, adapting preprocessing steps and repeating the subsequent analysis takes little time and effort.

3. REPLICABILITY OF THE WORK

a. Make sure that others can replicate your work by using descriptive comments that increase code readability. This is another advantage of working with syntaxes; others can look into the steps that you have undertaken before performing the main analysis.
b. Keep the original databases original. Make changes in copies of the dataset, and preferably save the dataset in a different file after each preprocessing step. For example
    i. After merging datasets you save a copy named 'data_merged'.
    ii. After Winsorizing you save a copy named 'data_merged_win'.
    iii. After normalization you save a copy named 'data_merged_win_norm'.
    iv. Etc.

    You may also number successively created data files.
c. Work in a shared folder (i.e. your folder in the Emma Neuroscience Group) in order to grant your supervisors easy access to your work.
d. Save the raw and processed datasets (the one's that you are using for actual statistical tests) and the pre-processing script in a folder together with the published manuscript.

PREPROCESSING STEPS

The order of the preprocessing steps can be variable. Here, we present a reasonable order for a range of pre-processing steps.

1. MERGING DATABASES

Often you have multiple data sources (e.g. Castor and Emma Toolbox) which need to be merged before you are able to perform analyses.

a. Always use an identifier: These databases can be merged by the use of a unique identifier that is present in all databases. Do not merge data without the use of an identifier; this puts you at risk of faulty merging which would mix up data within the dataset and make all subsequent analyses invalid.

b. Multiple identifiers: Sometimes you need multiple variables as identifiers. This can be the case when you have repeated measurements. For example, if you have two successive measurements at two different time points and these are registered in a long format (i.e. the measurements are displayed in the database as separate rows). If you would use only the subject number as identifier, the identifier is not unique for each row (you have two measurements per subject). This will lead to faulty merging.

c. Always perform a sanity check:
    i. Were errors reported after the merging of databases? If yes, you need to figure out where errors originate from. Do not proceed with other analyses because of the risk of faulty merging.
    ii. How many observations (rows) has the new dataset? Is this what you would expect?
    iii. Does the number of subjects in the merged database correspond to the number of included subjects?

d. Youtube tutorials about merging databases
    i. In SPSS: How to merge files in SPSS - YouTube
    ii. In R: Merge Data Frames by Column Names in R (Example) | Combine with merge Function - YouTube

2. EXPLORE YOUR DATA

a. When visiting your data for the first time, you need to thoroughly screen the data for validity. This may bring forward errors in the data that can have major influence on the results.

b. Is the data within valid range? Print the range of observations for each variables and determine whether the lower bound and upper bound are within realistic boundaries of the measurement. For example, IQ scores cannot have negative observations. Use the realistic range of variables to screen for errors in the data and correct them accordingly. Do not make changes to the original data (but rather in a copy, saved under a different name or a successive number) and keep a list of changes that you make to specific values (using the syntax). If you encounter more than a few errors, you should find out what the cause was. Did somebody in the research team make more errors than others? Then this is a reason to do a more proactive and systematic check on all the work performed by that member of the research team.
    i. In SPSS: SPSS Range, IQR, Standard Deviation - YouTube.
    ii. In R: Basic summary statistics in R - YouTube.

3. OUTLIERS

a. Outliers are observations with an extreme value relative to other observations. The definition of an outlier is variable. Often the cut-off is chosen at M +/- 3SD (at p <.001), alternatively the cut-off can be chosen at M +/- 2SD (at p < .05 level).

b. Outliers are problematic for most statistical analyses, especially parametric analyses (t-tests, ANOVA, regression). Outliers disproportionally influence the data distribution, and may invalidate statistical results.

c. Outliers can reflect errors in the data that were not identified by data exploration because they are within the realistic range, yet much smaller/larger than the majority of observations. The validity of an outlier should always be checked by returning to the source data and checking whether the value is correct.

d. Outliers can also be values that are extreme, yet not incorrect. When this is the case, there are different options.
    i. You could delete the outlier. This is not the preferred option because you are neglecting an observation for which there is no reason to assume that is an invalid measurement. When researchers delete outliers, they essentially assume that a population has no members with extreme deviations from the mean. This is a faulty assumption, because by definition of the outlier there will be extremes in the population. This is especially true for clinical populations, where variability is often higher and therefore the risk of outliers is also increased.
    ii. You could keep the outlier. This is also not the preferred option because this may affect the validity of your analysis.
    iii. Winsorize the outlier: This is an elegant comprise between the above options. Winsorizing keeps an outlier in the data, while moving the value of the outlier towards the less extreme direction in the data distribution. More specifically, Winsorizing revalues an outlier to the most extreme value in the dataset that does not qualify as an outlier.
    iv. In SPSS: Dealing with an outlier - Winsorize - YouTube.
    v. In R: Winsorize: Winsorize (Replace Extreme Values by Less Extreme Ones) in DescTools: Tools for Descriptive Statistics (rdrr.io).

4. NORMALITY

a. Normality of the data distribution for dependent variables is a central assumption for parametric tests (e.g. t-tests, ANOVA, regression). When you are planning to use parametric tests for your analyses, you should therefore check the shape of the data distribution for deviations of normality.

b. We recommend to do a visual inspection of normality using Q-Q plots:
   i. In SPSS: 6 ways to test for a Normal Distribution — which one to use? | by Joos Korstanje | Towards Data Science.
   ii. In R: Normal QQ Plots using R - Tutorial for beginners - YouTube.

c. Deviations from normality are not per definition problematic (also see Field, section 10.2.10 and for a more elaborate description see Pituch & Stevens, 2016). Since most parametric tests are relatively robust to violation of the assumption of normality, we recommend to keep the dependent variable as original in the case of modest violations.

Pituch, K. A., & Stevens, J. (2016). Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS. 6th Edition. New York: Routledge.

d. What to do if your dependent variable severely violates the assumption of normality:
   i. In the case of severe violation of normality, you can transform the variable. This means that you change the shape of the distribution thus better resembling a normal distribution. Changing the shape of the distribution can also influence the interpretation of the results; therefore you should be conservative in applying transformations.
   ii. The Emma Neuroscience Group recommends to use the van der Waerden transformation. This transformation places each observation in the original distribution to a new place in the normal distribution based on ranking (i.e. percentile scores). This also reduces the influence of outliers. Therefore, it may not be necessary to carry out Winsorizing if you use the van der Waerden transformation.
      1. In SPSS: Transformations > Rank Cases > Vink 'Rank' uit, Vink 'Normal scores' aan > Kies 'van der Waerden'.
      2. In R: blom function - RDocumentation.
   iii. In case the van der Waerden transformation is not suitable or is not effective, please consult Field, section 5.7 for an alternative transformation. For example, when your data contains two groups between which group differences are to be expected, Winsorizing can unintentionally eliminate or diminish the present group differences (i.e. by identifying poor performing patients as being outliers as compared to the control group distribution). In such cases, a Log transformation may be more suitable.

5. MISSING VALUES

a. Missing values are problematic because the children that have missing data will not be included in the analysis. This may affect statistical power and the generalizability of the findings. Most analyses are sensitive to missing data. Multilevel analysis is the exception and can handle missing data quite well. If you are using multilevel analyses, you may not need to handle missing data in preprocessing steps (Field, section 19.3).

b. We can handle missing values by replacing them with estimations of what the data would look like if they were not missing. This can only be done if the data are missing at random. This means that data are not systematically missing, i.e. the data are not missing for a specific reason that applies to multiple individuals with missing data. For example, you investigate children with traumatic brain injury. There is a proportion of children that cannot be tested at the six month follow-up because they have a disorder of consciousness. This data are not missing at random, because 'missingness' of the data are associated with the level of consciousness (which is in turn associated to injury severity). You cannot estimate these missing values, you have to exclude these children from your analyses and adapt your findings and generalization accordingly. See: https://stefvanbuuren.name/fimd/sec-MCAR.html.

c. The appropriateness of replacing randomly missing data is also dependent on the amount of missing data in the dataset. There are no strict guidelines for imputation, but you can use the following guidelines that we use in the Emma Neuroscience Group, assuming that data are missing at random:
   i. <10% generally not seen as problematic.
   ii. 10-15% generally seen as acceptable.
   iii. 15-20% generally seen as acceptable but questions may be asked.
   iv. > 20% probably seen as not acceptable.

d. The amount missing data is primarily calculated as a percentage of missing data per variable. However, this means that in theory you would impute all the observations of a participant that did not show up at the test assessment in the first place. Therefore, we recommend to impute data for variables with <20% missing data and for participants with <20% missing data.

e. Replacing missing data with estimations of missing data is most often performed using multiple imputation. Multiple imputation uses the correlation structure in your dataset to build a prediction model for the variables with missing data and replaces the missing values with the individual prediction for each missing data point. This process is repeated ('multiple' in 'multiple imputation') in different subsets of the database (often five imputations are performed). The end result is long format database with five different imputed datasets displayed below each other.
   i. SPSS has in-built software to perform a range of conventional tests on imputed datasets. This essentially means that the analysis is automatically performed on the different subsets, and afterwards the results are merged with meta-

analysic methods. How to Use SPSS-Replacing Missing Data Using Multiple Imputation (Regression Method) - YouTube.

ii. R packages that you may want to use often do not have the possibility to handle multiple datasets. Therefore, you may need to reduce the multiple datasets into one dataset with the missing data filled in. This can be done by aggregating the imputation datasets using the mean function. Please be aware that subsequent analyses do not account for the fact that data are imputed, which is not ideal but generally accepted in practice. Handle Missing Values: Imputation using R ("mice") Explained - YouTube.

f. Always check the result of imputation. Especially when imputing categorical variables, since not all imputation methods are suitable for imputation of non-numeric data.

6. DIRECTIONALITY TRANSFORMATIONS

a. Think about the directionality of your variables. This means, do higher scores reflect better performance or the other way around? This may vary within your data, which makes it very difficult for the reader to follow the coupling of a test result to an interpretation.

b. We advise to transform the direction of all variables to follow intuition:
   i. For all neurocognitive test scores: higher values reflect better performance.
   ii. For all measures of behavioral functioning: higher values reflect more problems.
   iii. Etc.

c. You can transform the direction of a variable by multiplying the variable with -1. This has the advantage that you retain the original values in the variable, but they are now shifted to the negative domain and this flips the higher/lower direction as in the example below.

Original variable		*Direction transformed*	
Lower	1	-1	Higher
	2	-2	
Higher	3	-3	Lower

7. DATA REDUCTION

a. In the Emma Neuroscience Group, we often carry out highly detailed measurements in one single domain of functioning, such as neurocognitive functioning. The advantage of this approach is that we can provide a highly detailed quantification of functioning.

b. The disadvantage of a detailed outcome assessment is that you end up with a large number of variables. If you use these variables as outcome measures (i.e. dependent variables), then you have to perform a lot of statistical tests. For every test, there is a risk of 5% of a false positive result (assuming test results with p-values < .05 significant). Therefore, running a lot of statistical tests is problematic; one expects one significant finding for every 20 tests performed. This is when assuming that there is no effect of interest! Therefore, it is generally not excepted to perform many different tests. There is no strict guideline on the maximum number of tests that you can perform without handling the problem of 'multiple comparisons', but you can use the following guidelines that we use in the Emma Neuroscience Group:
    i. <5 outcome measures: you are fine
    ii. 5-10 outcome measures: acceptable, but questions may be asked
    iii. >10 outcome measures: handle the multiple comparison problem

c. To handle the multiple comparison problem, you can adapt the p-value that is used as the benchmark to determine statistical significance. Bonferroni correction is the most well-known and straightforward option, but is very conservative for small samples and not very elegant in the sense that it does not use the structure of the data in order to handle the multiple comparison problem. Therefore, the Emma Neuroscience Group typically advises two other ways to handle the multiple comparison problem:
    i. Adapt p-values to the number of tests conducted using false-discovery rate (FDR) correction. This is a data-driven method to correct the p-value and the method of choice for p-value correction.
    ii. Use a strategy to reduce the number of outcome variables (component analysis) or statistical tests to be performed (step-wise data-driven analysis structure).

d. If you use detailed assessments as a source for predictors (e.g. MRI data in a regression), then the disadvantage is that the number of variables often exceeds the power of a statistical model. For prediction models, one needs 10-15 observations per variable in a regression model, https://aph-qualityhandbook.org/media/0n3pozyn/sample-size-and-power-calculations.pdf. In this case, p-value correction is no solution to the problem. Rather you should reduce the number of variables. There are several options:
    i. Theory-driven feature selection; use the literature to select certain predictors and discard others. This needs a solid framework in the existing literature and convincing reasoning. If reviewers are not convinced of your decisions, this

will be a delicate problem in the revision process that is likely to strongly impact your work in a late phase of the process.

ii. Data-driven feature selection; use the data to select certain predictors and discard other. This can be done on the basis of bivariate correlation strength to the outcome variable (selecting the top 10% most strongly correlating variables, or using a stepwise selection approach in the statistical model). When using this approach, there is a risk of multicollinearity (relations between predictors) causing flawed results in the regression models.
    1. In SPSS: Stepwise regression procedures in SPSS (new, 2018) - YouTube.
    2. In R: Statistics with R: Stepwise, backward elimination, forward selection in regression - YouTube.

iii. Principal component analysis (PCA); use the correlation structure in the data to create a set of summary variables (components). This procedure is elegant because you reduce the number of variables, while retaining as much of information from these variables as possible in the components (typically > 75% of the variance in the dataset). Another important advantage of PCA is that reliability of the resulting components is known to exceed reliability of the variables that are summarized in the components, thus reducing error variance in the measurements. PCA also provides a solution for multicollinearity, because interrelated variables will be captured in the same component, and the components typically have low correlation to each other. There is also a disadvantage to PCA: You will have to interpret what the components measure based on how variables load onto the components. Sometimes, this is difficult and can complicate the analysis. However, when using inter-related data (e.g. neurocognitive variables) the solutions are often intuitive and helpful to summarize the data in a smaller set of components.
    1. See Field, chapter 17.2 for a comprehensive background on PCA. Use this information to make a decision on what type of PCA you want to perform (e.g. varimax rotation).
        1 Also see Field, chapter 17.2 for how to apply PCA in SPSS.
        2 In R: Principal component analysis in R | PCA for genetic diversity assessment using varimax rotation | - YouTube