Amsterdam Public Health

Amsterdam UMC University Medical Centers VU Vrije Universiteit Amsterdam University of Amsterdam

Quantitative research - Data analysis - Initial Data Analysis	Set-up & Conduct- Process & Analyse data	
	VERSION	4.0

Aim

- To get a first impression of the data:
- To get an impression of the distribution properties of the continuous variables and the numbers in the subgroups of the categorical variables;
- To analyze the association between exposure and outcome

Requirements

- To investigate continuous variables for normal distribution;
- To check for percentages of missing values and outliers;
- To calculate Cronbach's alpha for sub scales.

Documentation

- Syntax which has been used to perform the initial data analysis;
- Percentages of missing values and outliers;
- Cronbach's alpha of new sub scales.
- Baseline characteristics of the study population (across treatment / exposure)
- Results from initial data analysis (odds ratios, hazard ratios, relative risks, sensitivity/specificity, etc.)

Responsibilities

- Executing researcher: To perform the initial data analysis, including:
    - Normal distribution of continuous variables;
    - Percentages of missing values and outliers;
    - Cronbach's alpha for new sub scales.
- Project leaders: To advice the executing researcher to perform initial data analysis.
- Research assistant: N.a.

How To

First impression

It is advisable to investigate the distribution of the variables that you are going to use. Frequencies are examined for all categorical variables (e.g. marital status, education). Descriptive statistics (percentage missing values, average, "trimmed" average, standard deviation, median, other possible percentiles, minimum, maximum) are calculated for continuous variables (e.g. body weight, blood pressure). It is advisable to create figures, e.g. boxplots or histograms in order to review their distribution.

Outliers and odd combinations

So-called outliers may occur in continuous variables. These are values that, theoretically, are not "out of range", but are extremely unlikely given the observed distribution. Generate a boxplot to check for outliers. Moreover, scatterplots can be created for continuous variables to reveal any unlikely combinations (simply reviewing correlations is not sufficient). For instance: A weight of 120 kg and height of 1.50 metres will be an outlier in most populations.

Cross-tabulations can be generated for categorical variables (e.g. gender x ADL limitations) in order to assess whether odd combinations are present. When it has been decided that a certain value or combination of values are outliers and the true value cannot be recovered from the raw data, then these need to be recoded as “missing”.

Missing values
It is important to carefully review missing values when evaluating the distributions. Often specific codes (e.g. -1 or 9) are used for missing values. This is not always necessary, but it could provide insight in why certain values are missing, especially when working with multiple datasets (e.g. participant forgot to fill out one question, or, missed the complete session due to dropping out) Note whether these codes have been defined as missing values. If there are missing values, consider whether these need to be imputed (filled in). There are a number of methods to impute missing data. More information on missing data and how to handle this can be found in Chapter handling missing data. Consulting a statistician can be valuable.

Normal distribution of outcome variables
Check for normal distribution of all relevant variables. Graphs can be used for this, such as histograms or Q-Q plots. If it is apparent that the variable is not normally distributed, then a transformation could be considered (for instance a logarithm transformation) to see whether this improves the distribution. However, note that results with transformed variables might be more difficult to interpret.

Distribution of categories
Categories can be combined if the numbers in one or more categories is/are too small. The need for this is not always evident from an ordinary frequency distribution. However, it can be apparent from a cross-tabulation. For instance in a study where there is stratification by rurality and density of AEDs (automated external defibrillators), the cross-tabulation of shows that for a rural area the highest category “high density” rarely occurs, whereas for an urban area the lowest category “low density” rarely occurs. The lowest and next lowest categories can then be added together, as well as the highest and second highest.

Evaluating the randomisation procedure
In order to evaluate whether the randomisation has been “successful”, the distribution of all the relevant (prognostic) variables needs to be reviewed separately for each treatment arm. Descriptive statistics (percentages, averages, median, standard deviation, range) can be used for this. Differences between groups can be tested (e.g. chi-square or t-test), although it needs to remembered that due to the randomisation procedure any differences found are, by definition, due to chance. So, if differences are found there is no need to change it, but it is important to remember it when evaluating and interpreting results.

Evaluating baseline characteristics (observational research)
In order to evaluate your data, the distribution of all relevant variables needs to be reviewed for each exposure category. Descriptive statistics (percentages, mean, median, standard deviation, range) can be used for this. Differences between groups can be tested by parametric tests (e.g. chi-square, t-test, ANOVA) or in the case of non-normal distributed variables a non-parametric test (e.g. Kruskal Wallis, Kolmogorov-Smirnov, Mann-Whitney U), to identify possible confounder variables.

Evaluating correlations between variables

By investigating interactions between variables (especially with the exposure) the models can be improved. For example, when researching health benefits from tea consumption the association might be different per sex (the regression has a different slope for men as for women), as women are more frequent consumers of tea and consume a larger volume of tea or there are biological differences in the absorption or metabolism of the elements that provide the possible health benefit(s). If this is not included in the analysis, the results might be generalized to men as well, although there is no evidence for this.

Investigating the association between exposure and outcome

After studying the characteristics of your research participants, it is time to test the hypothesis you made in the analysis plan. Depending on your research question there are several ways to analyze the association between your exposure and outcome, using univariable and/or multivariable models.

Depending on the outcome data different analyses need to be performed:

- dichotomous (binary) outcome: logistic regression
- continuous outcome: analysis of covariance, linear regression,

time-to-event outcome: survival analysis (Kaplan-Meier, cox proportional hazards model)

Audit questions

1. Has the distribution of all variables been reviewed?
2. Were there variables with a high percentage of missing values?
3. If so, how were these dealt with?
4. Have outliers been explored?
5. If so, how?
6. Where relevant: How were (large) deviations from normality solved?
7. Has been assessed whether the items belonging to a scale actually fit to the scale?
8. Has a correct analysis method(s) been chosen?
9. Have analyses been performed that have not been stated in the analysis plan?
10. If so, why, and is this reported?

LINKS

	Link
Wiki statistiek	https://wikistatistiek.amc.nl/index.php/Wiki_Statistiek

DOCUMENT HISTORY

Version	Status	Date	Name
4.0	Revision	30NOV2020	Laura van Dongen & Dr. Marieke Blom Elize Vlainic
3.1	Minor revision	23JAN2017	EMGO
3.0	Minor revision	13OCT2016	EMGO
2.0	Revision format	12MAY2015	EMGO
1.1	English translation	01JAN2010	EMGO
2.1	-	23APR2007	EMGO

DOCUMENT APPROVAL

Role	Name	Date

Project Leader	Dr. Seta Jahfari	12MAY2021