Preliminary Data Screening and Data Cleaning

Prepared & Presented by
Dr. Anas Khalifeh, PhD
Assistant Professor

Learning Outcomes

By the end of this class, you will be able to:

- Define missing data and outliers and understand how these can affect an analysis.
- Understand and apply remedies for missing data and outliers.
- Understand the importance of checking assumptions of statistical tests prior to analysis.

Introduction

- Have my data been entered correctly?
- Do I have missing data? If yes, now what?
- Are the assumptions of the statistical procedure met?
- Do I need to transform my variables? And how?
- Do I have pesky outliers?

Introduction

- Real datasets often contain errors, inconsistencies in responses or measurements, outliers, and missing values.
- Before conducting hypothesis testing/analyze your data, the investigator must ensure that data are free of any errors.
- Data cleaning is an important first step in every data analysis, to prevent the introduction of error in the results.

Introduction

- Some of the potential problems with data are as follows: errors in data coding and data entry, inconsistent responses, missing values, extreme outliers, and nonnormal distribution shapes.
- Problems with data should be **identified and remedied** (as adequately as possible) prior to analysis.

Introduction

- A thorough understanding of how the investigator has managed missing data and outliers will help us determine the quality of the findings.
- Once the data has been entered into a database, cleaning the data involves making certain that all of the variables have valid and usable values.

Possible Data Entry Problems at the Design Stage

1. **Questionnaires**:
   - Use simple, consistent formats for response options.
   - Ensure that each side of the questionnaire has clear labels or instructions, such as "Page 1" and "Page 2“ (Double sided)
   - Assign a unique identifier (e.g., a code) to each questionnaire.
2. **Entering data into a statistical package:**
   - Define codes in advance and stick to them (a clear codebook).
   - Do double check after the first data entry.
3. **Participant number (identity):**
   - Assign each participant a unique, anonymous ID (e.g., P001, P002)
   - Ensure the ID appears on every page of their questionnaire.

Screening and Cleaning the Data

- The data screening process involves **Two steps**:

Step 1: Check for errors: Check each of your variables for scores that are out of range (i.e., not within the range of possible scores, such as wild codes & outliers). Then, check for missing cases.

Step 2: Find and correct the error in the data file: Find where in the data file this error occurred (i.e., which case is involved) and correct or delete the value.

Screening and Cleaning the Data

- **Step 1** is completed by running frequencies and descriptive statistics or histograms/bar charts for categorical and continuous variables and examining carefully for invalid values, unusual values, large amounts of missing data, and adequate variability. It should become obvious to you if virtually every participant in the research gives much the same answer to a question.

Screening and Cleaning the Data

- So, what do we do if we find some out-of-range responses (e.g., a value of 3 for sex)? **First**, we need to find the error in the data file.
- In this situation, the researcher would need to identify cases with the improper codes, determine the correct codes, and then make the needed corrections. After that, a new set of frequency distributions should be run to make sure that the problems have been corrected as intended.

Missing Data

- Missing data occur when a study participant or subject deliberately or accidentally omits responses to a variable, when data are inadvertently left out, or when an investigator has not found a way of measuring a variable.
- **How missing data occurs:-**

✓ Overlooked questions, missing scores not related to other measured variables. e.g., a participant doesn't notice a question asking about their age and leaves it blank.

✓ Missing values a function of other measured variables, e.g., missing income scores for low education participants.

- The amount and the pattern of missing data are important (Random or Systematic).
- Understanding why and how the data are missing can help us find the correct solution.

Amount and Pattern of Missing Data

- Pattern is more important than amount of missing data. In general, the problem gets worse as the amount of missing data increases.

Patterns of missing data:

1) Missing Completely at Random (MCAR):

- The probability that the observation is missing is completely unrelated to either the value of the missing case, or the value of any other variables (No patterns of missingness).
- e.g., Participant accidentally skips a question about their age because the page is folded, but this has no relation to their actual age or their responses to other variables like income or education.
- How to handle MCAR: Listwise Deletion, Pairwise Deletion, and Mean/Median imputation.

https://youtu.be/-wz93kyQx8M?si=Huc-_vogITrUXpKV
https://youtu.be/DM14S_zB49o?si=ArQ2fpFbsP-efxVg

Amount and Pattern of Missing Data

2) Missing at random (MAR):

- Missingness does not depend on the value of the variable with the missing data after controlling for another variable.
- Missingness is unrelated to the value of the variable itself, but it is related to other variables that can be identified.
- e.g., Participants fail to report their income. The missingness is more likely among participants with lower education levels, which is recorded in the dataset. However, within each education level, the missingness of income is unrelated to the actual income values.
- How to handle MCAR: Multiple Imputation, Regression Models and Maximum Likelihood Estimation (MLE).

Amount and Pattern of Missing Data

3) Missing Not at Random (MNAR)

- Missingness is related to the value of variable with missing data and usually to other variables as well.
- e.g., Participants with very high incomes may choose not to disclose their income due to privacy concerns. Also, individuals with extremely high weights may skip questions about their weight due to embarrassment.

➡ How to handle MNAR: Proper handling requires specialized statistical models, expertise, and often needs additional resources.

Remedies for Missing Data

- Missing data is a serious problem in data analysis as it can affect the generalizability of the results, and so it needs careful handling. There are several potential remedies for missing data. The decision of how to handle missing data is crucial and should be made in the context of the given problem.
- Approaches to dealing with the problem of missing values can be classified:
  1. Deletion methods.
  2. Estimation (Imputation) methods.

Remedies for Missing Data

Deletion methods: The simplest approach to deal with missing data is to delete those cases with missing data and run the data analyses with only the complete cases. Three methods in the deletion category:

1) Listwise Deletion (also called complete case analysis)

Is simply the analysis of those cases for which there are no missing data. The use of listwise deletion is based on an implicit assumption of MCAR. Listwise deletion might be acceptable when MCAR is a realistic assumption and when the percent of missing values is low (<5%) in a large sample of participants.

2) Pairwise Deletion (also called available case analysis)

It involves omitting cases from the analysis on a variable-by-variable basis. In this approach, a case is deleted from the calculation of a statistic only when the specific variables in the analysis have any missing values.

3) Variable Deletion

It involves totally eliminating a variable from consideration in the analyses.

Remedies for Missing Data

Estimation (Imputation) methods

- The next approach is to estimate the missing values and use these estimates in the data analysis.
- Process of estimating missing data based on valid values of other variables or cases in sample.
- Estimation still introduces bias into the analysis, but it is useful when the sample size is small and the deletion of missing data will create further problems.
- There are several useful estimation procedures to replace missing data, such as using prior knowledge, substituting with mean or median values, estimating with regression models, expectation maximization, and using multiple imputations.

Remedies for Missing Data

- **Prior knowledge** is used when a researcher replaces a missing value with a value from an educated guess.

- **Substituting missing data with mean or median values** is another simple approach for estimating missing data. If the distribution of values is skewed or if the measurement scale is ordinal, the median rather than the mean is used as the replacement value.

Outliers

- Any data value that is an unexpected value or outside the expected range of values for that variable is considered an outlier.
- An outlier is a case with such an extreme value on one variable (a univariate outlier) or such a strange combination of scores on two or more variables (multivariate outlier) that it dis torts statistics.
- There are four reasons for the presence of an outlier. **First** is incorrect data entry. **Second** is failure to specify missing-value codes. **Third** is that the outlier is not a member of the population from which you intended to sample. **Fourth** is that the case is from the intended population but the distribution for the variable in the population has more extreme values than a normal distribution.

Detecting Univariate And Multivariate Outliers

Visual Check Using Graphs and Plots

- A review of visual displays of data, such as graphs and plots, is the easiest way of identifying outliers.
- Several graphical displays useful for identifying the outliers include the scatterplot, histogram, and box plot.

Using Standardized Scores

Another way of identifying the outliers is to examine the corresponding standardized score (i.e., z-scores) of the raw data.

https://youtu.be/uH8N5CJ3FEo?si=5jy0f05qOD2laoKF

Remedies for Outliers

Check to see if a transcription error is the cause of the outlier. If transcription is not the problem, other remedies may be implemented.

1) Deletion

- The simplest way to deal with outliers is to delete those cases that are unusual and then run the data analysis with normal cases.

2) Transform the Data

- When outliers are present in the data set, the distribution of the data will likely be skewed.
- There are four commonly used transformation approaches: log transformation, square root transformation, reciprocal transformation, and reverse transformation.

Remedies for Outliers

3) Changing the Score

- If transforming the data fails to correct the problems associated with outliers, you can consider changing the data value(s).
- However, this approach should be used cautiously to avoid any appearance of unethical manipulation of data.
- Diminishes the biasing effect on the data analysis results.

❖ One of the following methods can be used to change the score by changing the data value to one unit above the largest data value and then exclude the outlier. e.g. Suppose you have the following test scores for a group of students 10, 12, 14, 15, 16, 18, 90. Data after handling 10, 12, 14, 15, 16, 18, 19.

Data Should Be Normally Distributed (Normality)

- **The normality assumption can be checked in several ways:**

1. **First**, it can be checked visually through histograms or P-P plots. In a histogram, you will want to see a symmetrical, bell-curved distribution if the data are normally distributed. A P-P plot is the cumulative probability distribution of a variable against the cumulative probability normal distribution.
2. **Second**, you can use the measures of skewness and kurtosis.
3. **Finally**, you can perform statistical tests: the Kolmogorov–Smirnov test or the Shapiro–Wilk test.

https://youtu.be/AVketBmpUTE?si=Yh3udrvzVKX-sW84

https://youtu.be/_oHEZd_OAY0?si=WE47Qub0xKlp2YnK

References

- Saneii, S. H., Doosti, H. (2024). Practical Biostatistics for Medical and Health Sciences. Singapore: Springer Nature Singapore.