Global PaedSurg Research Training Fellowship



**Session 6: 25 April 2019**

**Data Cleaning and Analysis**

**By Dr. Emily Smith and Tessa Concepcion**

# Clean data set

* Think about data you have collected.
* What is your research question?
* What are you trying to get your data to “say”?
* Which statistical tests will best help you answer your research question? This will vary depending on the research question?
* Contact the research team/ statistician to discuss how to analyse your data.

# Data Analysis Process



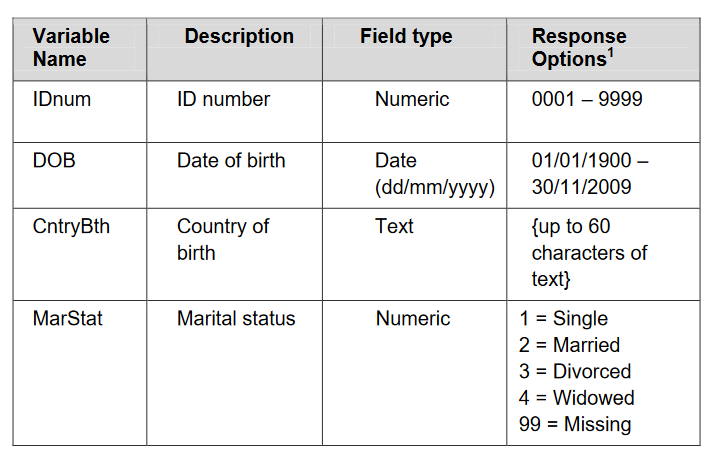
* Managing data – via Excel or REDCap
* Analysing and Interpreting large data sets by using statistical software

# Step 1: Creating an analysis plan

* Formulate your plan according to your research objective and setting
* Collect the data
* Now that you have collected the data, consider how do you assess its accuracy and precision. Example – Global PaedSurg validation process.

# Step 2: Managing Data

* Create a data dictionary.
* A data dictionary should include at minimum:
  + Variable names
  + Variable descriptions – what the variable means
  + Variable types
  + Response options (the answers) and any codes used
* Some data dictionaries include the column from the questionnaire where the variable can be found. Microsoft Excel is a great resource for a data dictionary. An example is attached below

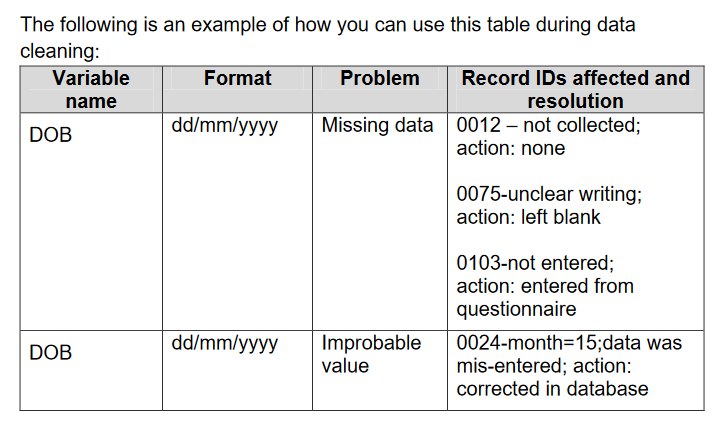


*NB: Make sure you add a column describing how to code the missing values*

# Step 3: Cleaning Data

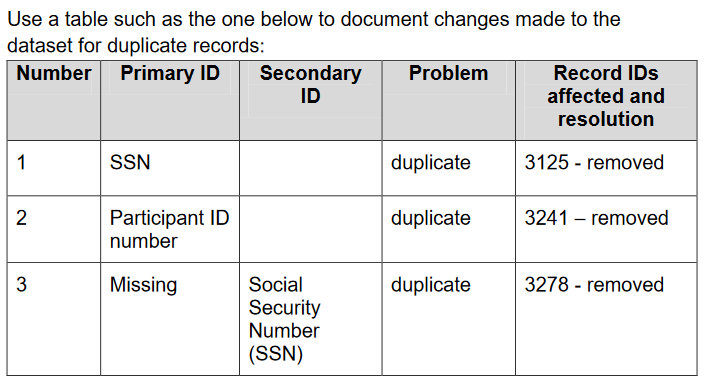
* Before you begin, make a copy of the original dataset!!
* Few databases are free of errors and missing values
* Reviewing dataset to identify errors before analysis is important
* This is an iterative process
* Document, document, document any changes you make! – note down any changes you make such as
  + Changes to the datasheet
  + Decisions about how to assess certain fields
* Documentation will ensure that you make consistent decisions and will provide a reference for those who may have questions about your analysis

The following is an example of how to use this table when documenting any changes.



Check for duplicate records

* Identify how many records are in the dataset. Use your statistical software to check the record count
* Determine if the number of records matches the number of questionnaires/ entries
* If the records are more than the number of questionnaires/ entries, run a frequency listing to look for multiple records with the same identifying information (such as ID number or name). If the data is anonymous you can still assess for duplicates – the statistical software can identify record entries that are the same i.e same weight, gestational age and age at diagnosis.
* If there are two records with the same ID number or name, select the records and examine them to determine if they are identical (a duplicate record) or whether and ID number or name was entered incorrectly
* Use a table such as the one below to document changes made to the dataset for duplicate records



# Step 4: Detecting and correcting missing, miscoded or out of range values

* Few datasets are 100% complete or accurate
* Usually there are a few weird or missing values
* Sometimes it occurs randomly or in patterns

## Types of missing Data

* MCAR (Missing completely at random): Missing data are independent of variables and occur at random.
* MAR (Missing at Random): missingness is related to a particular variable, but not related to the value of the variable that has missing data (accidentally omitting an answer on a questionnaire)
* MNAR (Missing Not at Random): Missing for a reason for example some individuals may be excluded

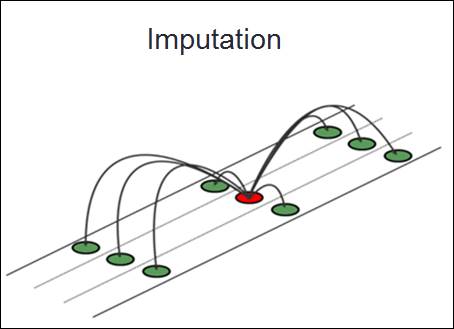
The best way to identify missing variables is to run frequency tables, it shows you your variable distribution and totals.

## Handling Missing Data

### Complete case analysis (complete participant analysis)

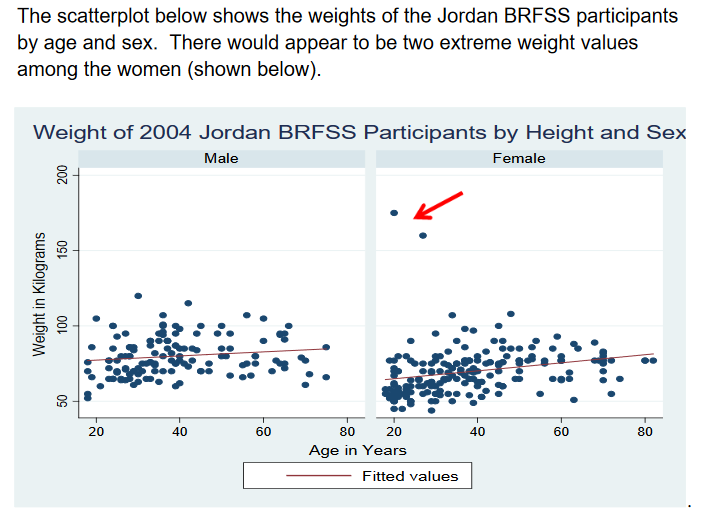
* Uses whole dataset as it is (+/- remove everyone with missing data on one or more variables)
* Can reduce precision
* Unbiased in a wide range of circumstances

### Imputing

Statistical estimation of what the missing variable would be based on other participants with similar variables.

* Default value imputing
* Mean imputing
* Regression imputation
* Multiple imputation
* Inverse probability weighting

### Identify outliers

Making a scatterplot illustrates the value of one variable on the X axis and the value of the other on the Y axis.

# Data analysis

* Now that you have cleaned your data.
* Remember you should have a copy of the raw data that’s untouched as well as the cleaned dataset.
* The cleaned dataset will be used for analysis

## Types of Statistical Analysis

Descriptive Statistics

Describes a phenomenon, such as how many? How Much?

* Frequencies
* Basic measurements

Can be presented in a table, including the raw number and the percentage of the total which the raw number represents. Use of bar charts and pie charts for categorical data. For distribution, Box and whisker Plots and histograms can be used. Continuous variables can be made into categories such as age, using various categories with different ranges.

### Inferential Statistic

Infers about a phenomenon, such as proving or disproving theories, associations between phenomena, if sample relates to larger population i.e. diet and health. And determine if findings are significant.

* Hypothesis testing
* Confidence intervals
* Significance testing
* Prediction
* Correlation

## Analysis of Continuous Data

### Correlation

When to use it?

* When you want to know about the association or relationship between two continuous variables.
* C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\N3L1XTC8\MC900014156[1].wmfExample: food intake and weight; drug dosage and blood pressure; air temperature and metabolic rate, etc.

What does it tell you?

* If a linear relationship exists between two variables, and how strong that relationship is.

What do the results look like?

* The correlation coefficient = Pearson’s *r*
* Ranges from -1 to +1

### T-tests

What does a t-test tell you?

* If there is a statistically significant difference between mean score or value between two groups (either same group of people before and after, or two different groups of people.

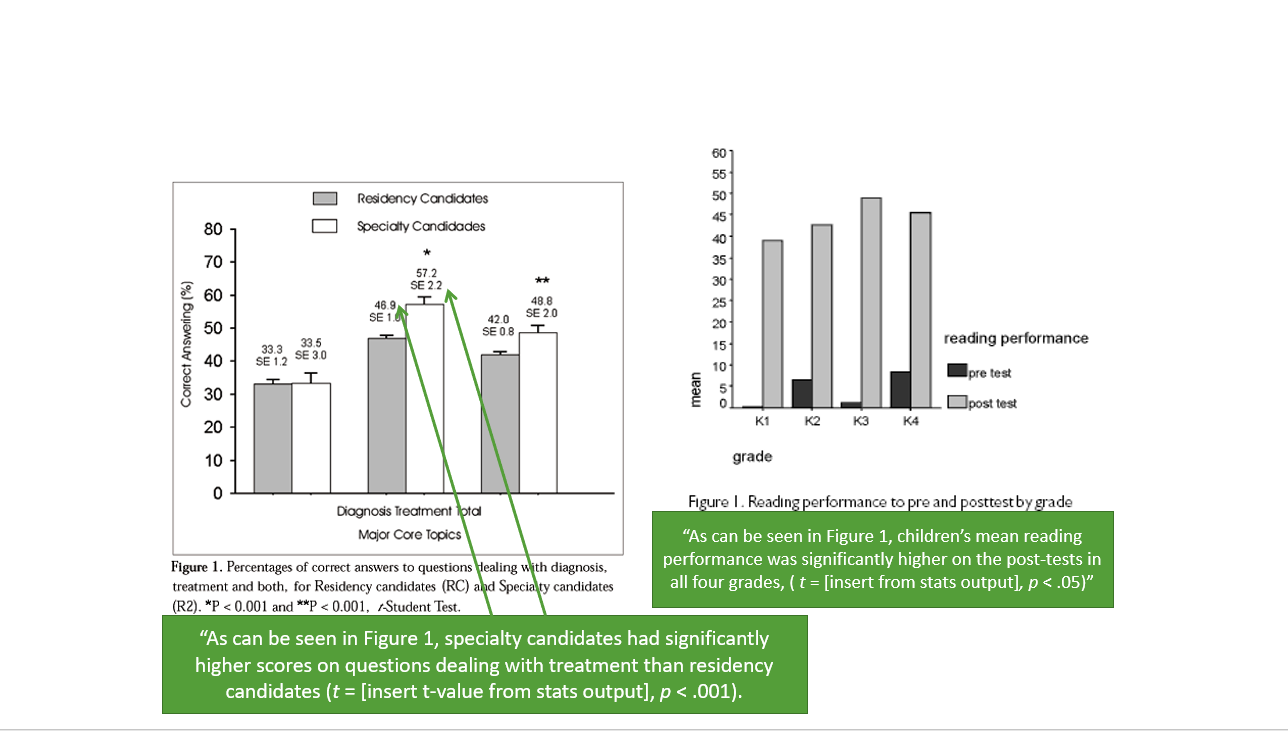
*C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\F0382YT3\MC900014158[1].wmf*

What does the result look like?

* Students’ *t*

How do you interpret it?

* By looking at the corresponding p-value
  + If p < 0,05 then means are significantly different from each other
  + If p > 0,05 then means are not significantly difference from each other

How to report it is illustrated in the diagram below

## Analysis of Categorical (Nominal) Data

### Chi-squared

When to use it?

* When you want to know if there is an association between two categorical (nominal) variables (i.e., between an exposure and outcome)
  + Example: Smoking (yes/no) and lung cancer (yes/no)
  + Example: Obesity (yes/no) and diabetes (yes/no)

What does a chi-square test tell you?

* If the observed frequencies of occurrence in each group are significantly different from expected frequencies (i.e., a difference of proportions)

C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\N3L1XTC8\MC900434529[1].wmf

What do the results look like?

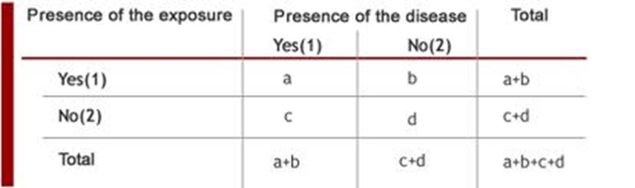
* Chi-square test statistics = *X*2

How do you interpret it?

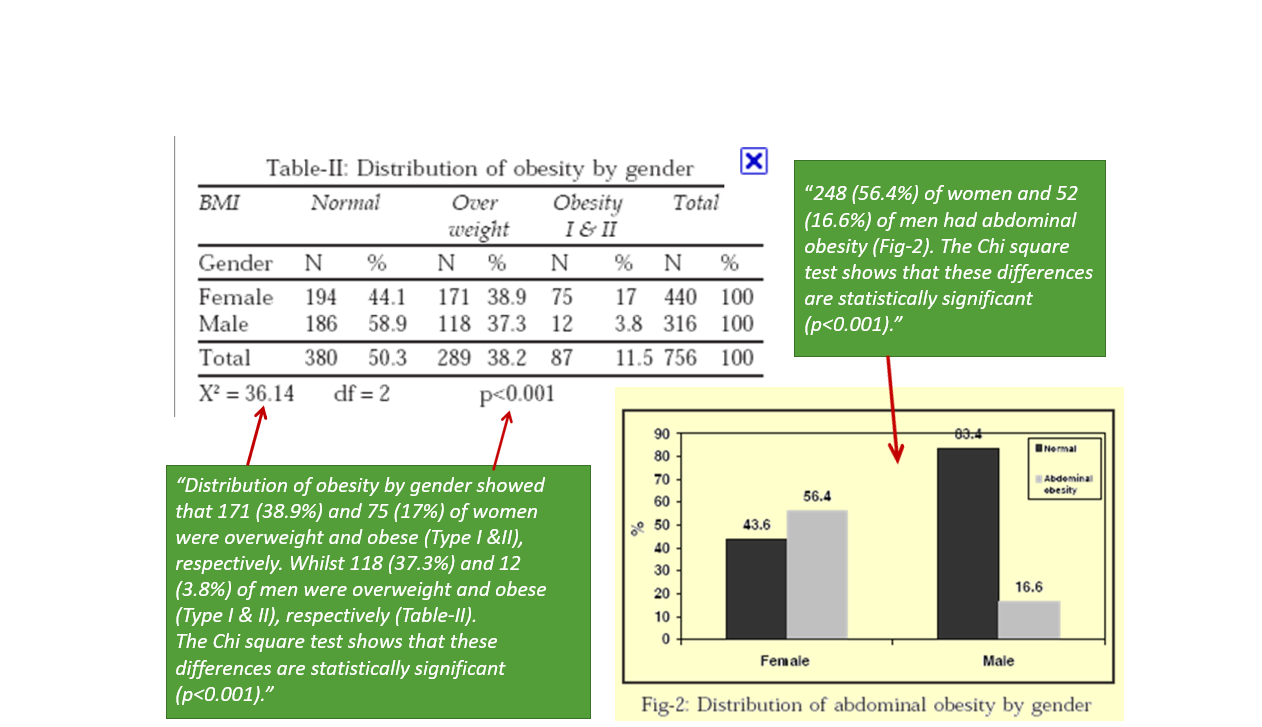
* Usually, the higher the chi-square statistic, the greater likelihood the finding is significant, but you must look at the corresponding p-value to determine significance.

*Tip: Chi square requires that there be 5 or more in each cell of a 2x2 table and 5 or more in 80% of cells in larger tables. No cells can have a zero count.*

An example is in the table below:



How to report is illustrated in the diagram below (P value):



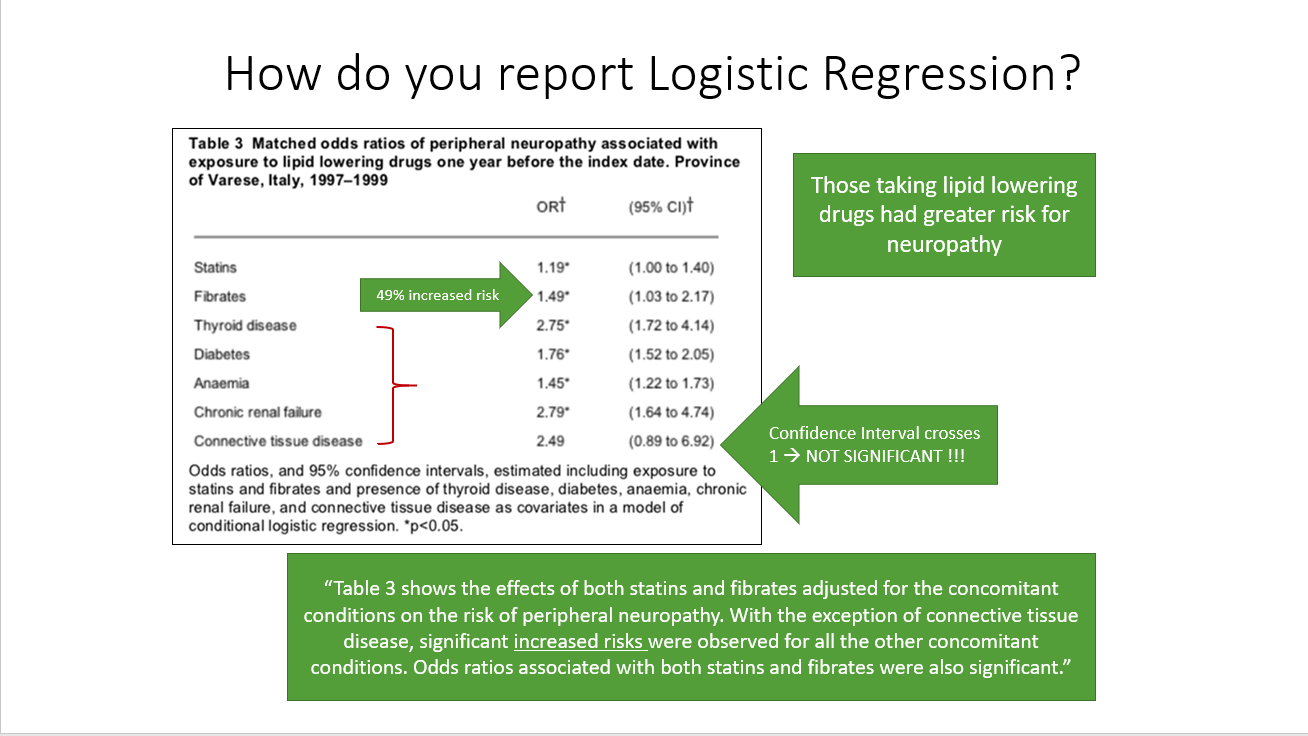
### Logical regression

When to use it?

* When you want to measure the strength and direction of the association between two variables, where the dependent or outcome variable is categorical (e.g., yes/no)
* When you want to predict the likelihood of an outcome while controlling for confounders.

How do you interpret the results?

* Significance can be inferred using by looking at confidence intervals:
* If the confidence interval does not cross 1 (e.g., 0.04 – 0.08 or 1.50 – 3.49), then the result is significant.
* If OR > 1 🡪 The outcome is that many times MORE likely to occur
  + The independent variable may be a RISK FACTOR
  + 2.0 = twice as likely
* If OR < 1 🡪 The outcome is that many times LESS likely to occur
  + The independent variable may be a PROTECTIVE FACTOR
  + 0.50 = 50% less likely to experience the event

How to report, is as illustrated in the image below

## Summary of Statistical Tests

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic Test** | **Type of Data Needed** | **Test Statistic** | **Example** |
| **Correlation** | Two continuous variables | Pearson’s r | Are blood pressure and weight correlated? |
| **T-tests/ANOVA** | Means from a continuous variable taken from two or more groups | Student’s *t* | Do normal weight (group 1) patients have lower blood pressure than obese patients (group 2)? |
| **Chi-square** | Two categorical variables | Chi-square *X*2 | Are obese individuals (obese vs. not obese) significantly more likely to have a stroke (stroke vs. no stroke)? |
| **Logistic Regression** | A dichotomous variable as the outcome | Odds Ratios (OR) & 95% Confidence Intervals (CI) | Does obesity predict stroke (stroke vs. no stroke) when controlling for other variables? |

# References

*Essential Medical Statistics*. Kirkwood & Sterne, 2nd Edition. 2003

http://ocw.tufts.edu/Content/1/lecturenotes/193325

http://stattrek.com/AP-Statistics-1/Association.aspx?Tutorial=AP

http://udel.edu/~mcdonald/statcentral.html

*Background to Statistics for Non-Statisticians*. Powerpoint Lecture. Dr. Craig Jackson , Prof. Occupational Health Psychology , Faculty of Education, Law & Social Sciences, BCU. *ww.hcc.uce.ac.uk/****craigjackson****/Basic%20****Statistics****.ppt.*