**Data exploration and visualisation plan (Descriptive statistics)**

Juan C. Rivillas, Ph.D. Student ICL MRC

Updated: 09.01.2022

Importing CSV

Export SABE’s data from Excel to a separate CSV file.

*Descriptive statistics*- Summary

1. Displaying data (tables and plots).
   * Use appropriate plots to present the distributions.
   * Outliers, errors, missing data.
2. Summarise the data (Mean, Median, Mode).
   * Summarise the data using whatever statistics you think appropriate.
3. Measure the variability in the data (inter-quartile range, variance, SD, FWHM).
   * Quantify variability, so we can assess whether differences between groups are real
4. Quantify the strength of associations between variables (Pearson correlation coefficient).
   * Quantify the effects of interest, so they can be compared between different groups (ACEs, family SEP).
   * Investigate possible dependencies between the variables.
   * Bivariate is a great step beyond correlation analysis towards multivariate analysis.

Overall and specific tasks

1. **Displaying Data**

Tables (Frequency Tables).

Figures (composition -static: pie charts, components of components, and tree map to show accumulation to total and absolute measures). Bar Charts, histograms, and scatter plots.

Distributions

Use histograms to look at distributions for each column.

Which variables appear to be normally distributed?

What happens when the data are split by categories/different values?

Dealing with errors

To take care of suspicious outliers.

To eliminate any columns that appear to contain systematic errors.

Box plots

Use the Box Plot to compare variable distributions between different attributes (columns).

Which variables have significantly different means between attributes?

Scatter plots

Use Scatter Plots to look at covariation between variables.

Use correlations to suggest other interesting projections of the data.

k-means clustering

Use a k-Means to find clusters in the data set.

How many clusters appear to be present?

Which features can be used to separate clusters?

How do the clusters correspond to type?

1. **Descriptive statistics**

Need to quantify the effects of interest: SEP and ACEs on BHS, AL, subsystems scores, biomarkers.

Need to quantify the dispersion e.g., are the variations between individuals too great to draw conclusions.

Arithmetic mean

* Can be used for both continuous and discrete data – not for categorical data. Biomarkers, medication, and health risk (alcohol and tobacco? -it depends on categories, e.g., number of drinks or cigarros applies, while YES or NOT does not apply.
* Investigating the Body Mass Index for people with different SEP or ACEs.

Median

* Middle observation - such that 50% of data lies below its value.
* Main advantage of median over arithmetic mean is insensitivity to outliers (extreme points). This is also the main disadvantage.
* Use tables and Box (and Whisker) plots.

Mode

* + Data value or category occurring most frequently.
  + Particularly useful for qualitative data and for discrete data – when fractional values don’t make sense
  + e.g. biomarkers, health risks, SEP measures.
  + Use barplots to SEP and ACEs measures, diseases, and health risks.

Variability

* People differ in their characteristics (SEP, ACEs, biomarkers, etc…)

🡪 BMI will vary between people with the same SEP or ACEs.

* to quantify whether the difference in BMI, ldl or cholesterol is due to tobacco or alcohol, other factors, or a combination of the two?
* Use histograms (single var) or scatter plots (two var).

Measures of dispersion aims to characterise degree of spread or variability within data set.

* + E.g., weight and weigh, ldl and HbAc1
  + Use histograms and Box (and Whisker) plots to quantify variability in data: Quartiles: divide the distribution in 4 equal parts.

1st quartile =25th percentile, 2nd = 50th percentile (median), 3rd  = 75th percentile

* + Inter Quartile Range – IQR
    - Inter-quartile range: (3rd – 1st quartile) 25th to 75th percentile.
    - Not sensitive to outliers, but only uses central 50% of observations
  + The variance quantifies amount of variability or spread around the sample mean. The variance of *n* observations, *x1,…,xn*, is defined as



e.g., investigating the cholesterol or blood pressure for people with different SEP or ACEs.

* Standard Deviation (SD)

Notes: Look at the shapes of a distribution **Skewness** measures symmetry around the mean

* Positively skewed 🡪 long right tail.
* Negatively skewed 🡪 long left tail.
* Symmetric 🡪 equal tails. If skewed, median and IQR are preferable

1. **Correlation of two variables**

* If a relationship exists between two variables (e.g., BMI or blood pressure and SEP or ACEs), we say that they have a correlation, or dependence.
* We are often interested in establishing whether a linear relationship exists between two things.
* One method of calculating the correlation is the Pearson Correlation Coefficient.

Text

Description automatically generated

**4. The normal distribution**

* Using the Normal distribution as a model for the population behavior, we can use it to answer questions like
  + - Does this population have a larger/smaller mean than another population? (e.g., low vs high SEP).
    - Does this population have a larger/smaller variance than another population?
* To use *hypothesis testing* to answer these types of questions and potentially show a statistically significant difference between variables of interest.