Exercise 6: Descriptive Statistics

Launch Stata, open a new do-file and save as *Stata_Exercise6.do*. Add appropriate comments at the beginning of the do-file. Add commands to change the current directory to the Exercise 6 folder and load bl_combined2.dta. Remember to keep saving the do-file as you go along. Run through these exercises referring to chapter 7 in the module notes.

6.1 Summarising distribution of a continuous variable

For each of the variables *bmi*, *egfr*, *lvef* and *bl_creat*:

- Use the `summarize` command to obtain the mean, SD, median and interquartile range.
- Which of these statistics would you use to summarize the distribution of each variable?
- Produce a histogram with an overlaid normal distribution.
- Produce a box plot.

6.2 Describing the association between continuous variables

For the variables *bmi*, *sbp*, *egfr*, *lvef* and *bl_creat*:

- Obtain the pairwise correlation coefficients.
- Produce a scatterplot matrix.
- Produce a scatter plot of (i) *egfr* against *bl_creat* and (ii) *sbp* against *bmi*.

6.3 Describing the association between a continuous and a categorical variable

Explore the association between (i) *bmicat* and *sbp*, (ii) *diab* and *bmi*

- Use the `table` and/or `tabstat` commands to produce tables showing the number of observations, mean and standard deviation of the continuous variable in each category of the categorical variable.
- Copy and paste the tables in to an Excel worksheet.
- Produce box plots of the distribution of the continuous variable over the categorical variable.
- Carry out a two-sample t-test to investigate whether mean *sbp* varies by *overwt*.

6.4 Describing the association between two categorical variables

Explore the associations between (i) *agegroup* and *pep*, (ii) *bmicat* and *diabetes* and (iii) *bmicat* and *pep* (note pep is the primary endpoint of heart failure hospitalisation or death from cardiovascular causes).

- For each of the above produce a twoway table with row percentages and a p-value from a chi-square test.
- How are the variables associated? Are the associations as you might expect?
- Are there any missing values? Produce a table that also includes the missing values.

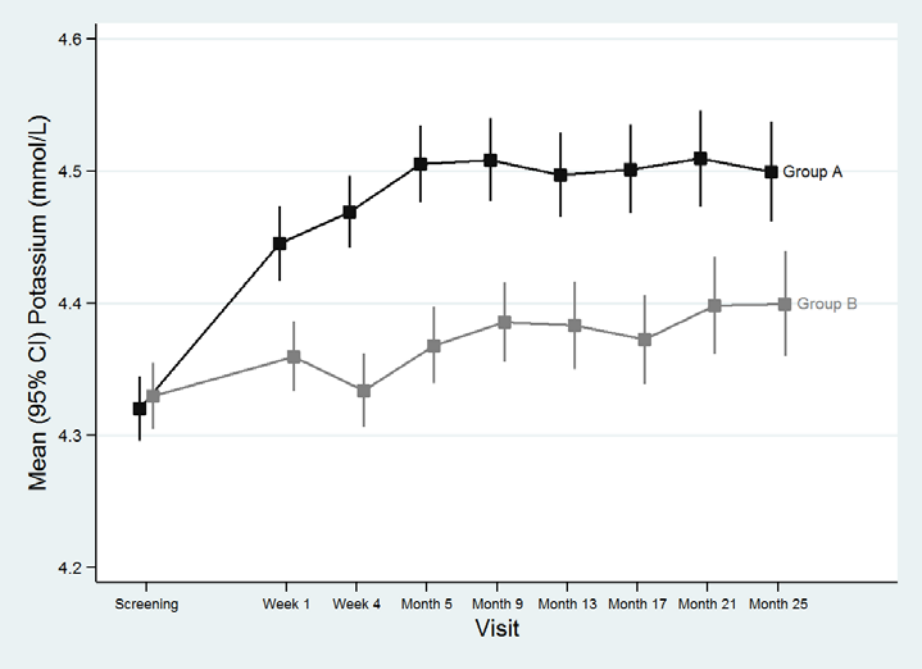
- For *bmicat* calculate the odds of *pep* in each category. How do the odds relate to the row percentages seen in the twoway table?
- Obtain odds ratios and a graph showing the odds and 95% confidence intervals.

6.5 Overlaid Twoway Graphs

(You may find it helpful to refer to the Introduction to Stata Graphics notes, available on Moodle.)

Open meanpot.dta. You will need to do a little bit of data processing.

- Using the *Graphics > Twoway Graph* GUI try to reproduce the figure below.
- The figure consists of four overlaid plots – two scatter plots and two connected plots.



- Once you've completed the task using the GUI copy the command syntax into your do-file.
- Use the triple forward slash (///) to enable the command to be laid out clearly over several lines.
- What do you think about the x-axis (visit) scale? How might you improve this figure?