Principles of Empirical Analysis: Session 2

Aapo Stenhammar

January 19, 2022

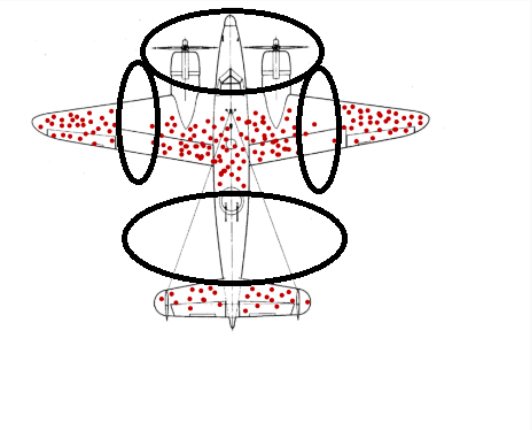
Outline for today's session

1. Comments on the 1st problem set
2. Tips and help for the 2nd problem set

Problem set 1: Question A

Key to this problem is to understand the selection of data

- In the data only planes that come back from mission are observed!



If a plane was hit to a place were there are no red dots it did not return

- Based on this more armor should be added to these locations

Problem set 1: Question B (1/2)

The two data sets were in a different format

- In death data one row was a the level of sex-year while in the population data one row was year

The goal in 1 b.3) was to reformat the data sets to similar format and then merge them

- I reformatted the population data using reshape long command

After formatting the data one needed to calculate the number of deaths and population by 10-year age bins using the egen command

- This was done in first exercise session!

Problem set 1: Question B (2/2)

In calculating the death rate one needed to notice that the population count referred to 31 Dec while the deaths were accumulated over the entire year

- Thus one needed to add the deaths to the denominator

Finally the graphs in 1 b.4) - 1 b.6) could be done using the twoway connected command

Today's data source

Today we use a sample from the Finnish Census (in Finnish väestönlaskentapaneeli) provided freely for education purposes by Statistics Finland

- To find the data click here

The data set follows 839 individuals over years 1950-2010 at 10-year frequency. Variables include:

- Information on which income decile the person is (tuloluokka)
- Whether the individual has graduated from high school (yo)
- The age of father (isa_ika)

Renaming and relabeling data

Very often you need to rename variables. For example today we want to name the variables in English

To rename variables type *rename oldname newname*

Often it is easier to keep track of our data if we assign labels to specific values of our variables

To assign labels to sex variable type:

label define sex 1 "male", add
label define sex 2 "female", add
label values sex sex

Plotting distributions

To plot the cumulative distribution of a variable you can use the distplot command

- Before using it you need to download it (ssc install distplot)

To plot the density of a variable you can use the kdensity command

To split observations in to different quantile groups of a variable you can use the xtile command

- This can for example be used to calculate percentiles of a variable (see Figure 3)

Figure 1: CDF of father's age at birth

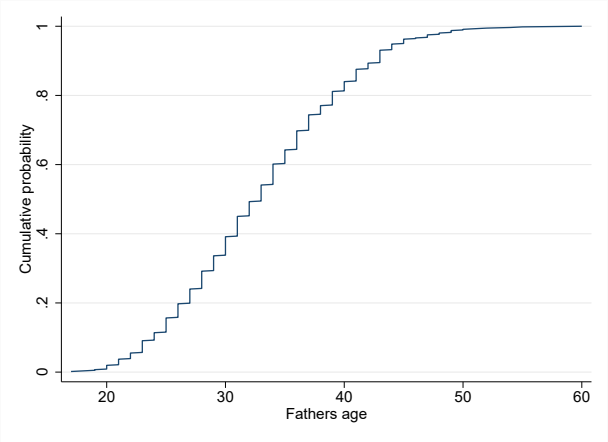


Figure 2: PDF of father's age at birth

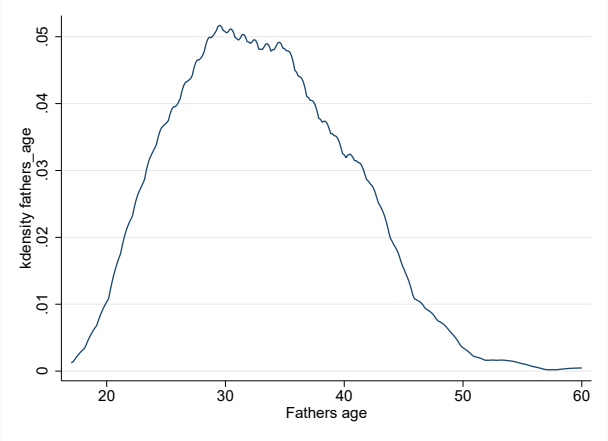
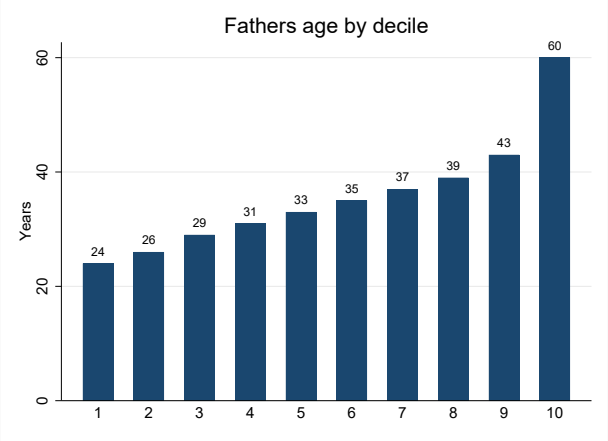


Figure 3: Percentiles of father's age at birth



Calculating summary statistics by group

To calculate the mean of a certain group we can use the egen newvar = mean(var), by(group) command

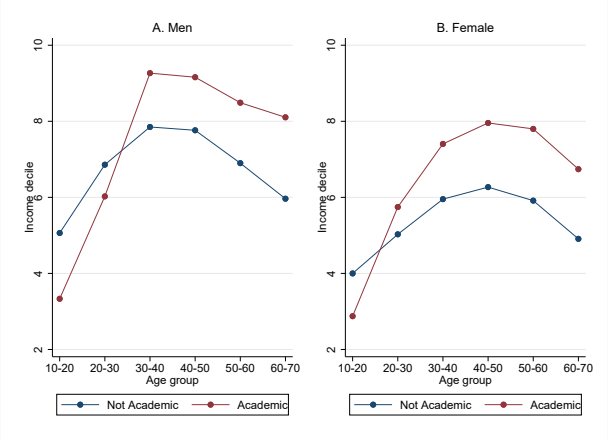
- If you would rather want to calculate the minimum or maximum value in a group you can instead of mean specify min or max

Second option is to collapse your data by specifying collapse (mean) var, by(group)

- If at some point you want to return to the original data format you should use the preserve and restore options

Tip: you can combine graphs using graph combine

Figure 4: Development of income



Calculating changes by group

Often you want to calculate the change in a specific variable over time and by group

One way to do this is to use the egen command to specify the value of the variable over time

egen newvar_xx = mean(var) if year == xx, by(group)

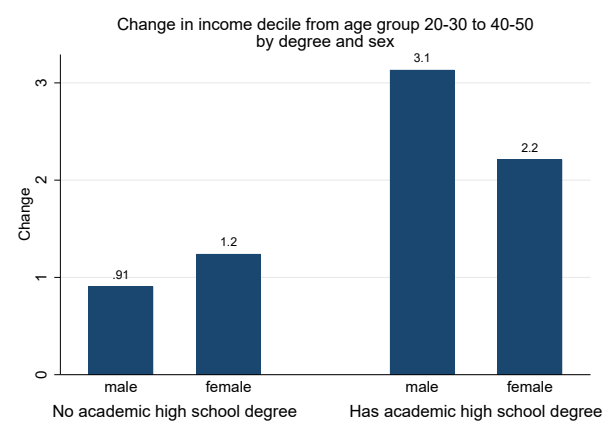
and then collapse the data

collapse (mean) newvar_*, by(group)

After this you can simply use the gen command to calculate the change over time!

Tip: bar graphs can be made in Stata using the graph bar command

Figure 5


Change in income decile from age group 20-30 to 40-50
by degree and sex
Change
0
1
2
3
.91
1.2
3.1
2.2
male
female
male
female
No academic high school degree
Has academic high school degree