Practical for Session 2 Summarising Data

18/10/2022

1   Hand Calculations

This section gives you the chance to do some calculations for yourself and see how the concepts we saw in the lecture work in real life. Once upon a time, these calculations would have been done by hand: I'm sure you could do them in your head, but getting stata to do them for you will be quicker. However, we are going to go through the steps that you would have to perform if you were calculating them by hand, so that you can see how it works. In practice, you would simply ask stata to churn the results out rather than calculating them this way.

> To start Stata, click on Start Button ⇒ All Programs ⇒ Stata 14 ⇒ StataIC 14 (64 bit)
>
>
> This should work for most of you, but depending on your faculty and whether you are staff or a student, the exact route may vary. Let me know if you have any difficulty.

> Solutions for all practicals can be found at
> `http://personalpages.manchester.ac.uk/staff/mark.lunt/stats_course.html`

Load in the PImax dataset with the commands

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/pimax.dta
```

(Note that there are two separate commands above, and they need to be entered separately, each on its own line.)

This dataset contains a single variable, pimax, which is the maximal static inspiratory pressure, measured in cm H_2O, of 25 cystic fibrosis patients.

Sort the data into ascending order with the command

```
sort pimax
```

1.1   Now look at the data in the spreadsheet view using the command `browse` or the

browse button, and look up the median value   ......

1.2   From the same browse view, what are the 25th and 75th centiles (the values of the 7th and 18th observation respectively).

   Lower quartile   ......

   Upper quartile   ......

1.3   Now we will calculate the mean. First put the data back in order with the command

```
sort id
```

Then we can generate a running sum with the command

```
gen sum = sum(pimax)
```

If you look at the data with `browse`, you will see the variable `sum` contains a running sum, i.e. the sum of all observations from the top of the dataset to that observation inclusive. Hence, `sum` in the last observation contains the sum of `pimax` in all of the observations.

Finally, we have to divide by the number of observations. We can do this with two commands: first

```
gen n = _n
```

will put the number of the observation into a variable called `n`, then

```
gen mean = sum/n
```

will generate a running mean: the mean of all observations from the top of the dataset to the current observation. Again, you can browse the data.

What is the mean PImax in this sample ?   . . . . . .

1.4   Finally, we are going to calculate the standard deviation. First, we need to create a variable containing the mean value for each observation:

```
drop mean
egen mean = mean(pimax)
```

If you `browse` the data, you will see that the variable `mean` now contains the same value for each observation, the overall mean. This is an important difference between `gen` and `egen`. (From now on, I will assume that you are `browse`ing the data each time you change it and will not explicitly mention it).

Next we calculate the difference between each observation and this mean (i.e. $x_i - \bar{x}$):

```
gen diff = pimax - mean
```

Square the difference to give $(x_i - \bar{x})^2$:

```
gen diff2 = diff * diff
```

and add them all up $(\Sigma (x_i - \bar{x})^2)$:

```
gen diff2_sum = sum(diff2)
```

Dividing by the observation number will give a "running variance", with the value of the variance in for the entire sample in the last observation $(\frac{\Sigma(x_i - \bar{x})^2}{n})$

```
gen variance = diff2_sum / n
```

Finally, take the square root of the variance to get the standard deviation $(\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}})$:

```
generate sd = sqrt(variance)
```

Again, the value for the entire sample will be in the last observation.
What is the standard deviation of PImax in this sample ?[a]   . . . . . .

2   Summarising Data in Stata

Read the stata file `htwt.dta` into stata by typing

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/htwt.dta, clear
```

This file includes two BMI values: `bmi` which was based on measured data and `bmirep` which was based on reported data.

2.1   Examine the distribution of measured BMI scores by displaying a histogram of the data, using the command

```
histogram bmi
```

Is the data normally distributed, or does it show some skewness ?

....

3 Summarise

3.1   Calculate summary measures of the measured BMI using the command
        `summarize bmi, det`

3.2   Write down the mean BMI.   ......

3.3   How does the mean compare to the median ?

....

....

....

3.4   What are the lower and upper quartiles of the data ?

....

....

....

4 Summarise by group.

4.1   Display the data separately for each sex, using the commands
        `sort sex`
        `by sex:  summ bmi, det`

4.2   The distributions for the two sexes can be neatly compared graphically using boxplots. The command to do this is `graph box bmi, by(sex)`. Write down a short description of what you see.

....

....

....

....

5   tabstat

This command can be used to produce tables of summary statistics: it is similar to `summarize`, but the output it produces is far more controllable. The basic syntax is `tabstat varlist, statistics(statname [statname ...]` The option `statistics` can be given one or more of the options in table 1.1.

For example,

`tabstat bmi, statistics(mean sd)`

would give the mean and SD of BMI. There is also a `by()` option, which enables you to obtain the statistics for different subgroups:

`tabstat bmi, statistics(mean sd) by(sex)`

will give the mean and SD of BMI for men and women separately.

5.1   Use `tabstat` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately (If you have forgotten the names of the variables to use for this, try typing `describe`.

....

....

6   table

An alternative to `tabstat` is `table`. This is more flexible in some ways and less flexible in others: in particular, it can only produce a maximum of 5 statistics per group. The syntax is

statname	definition
mean	mean
count	count of nonmissing observations
n	same as count
sum	sum
max	maximum
min	minimum
range	range = max - min
sd	standard deviation
var	variance
cv	coefficient of variation (sd/mean)
semean	standard error of mean = sd/sqrt(n)
skewness	skewness
kurtosis	kurtosis
median	median (same as p50)
p1	1st percentile
p5	5th percentile
p10	10th percentile
p25	25th percentile
p50	50th percentile (same as median)
p75	75th percentile
p90	90th percentile
p95	95th percentile
p99	99th percentile
iqr	interquartile range = p75 - p25
q	equivalent to specifying "p25 p50 p75"

Table 1.1: Statistics available in `tabstat`

```
table groupvars, contents(contents_list)
```

where **contents_list** consists of pairs of statistic names and variable names. So, to get the mean and SD of BMI for men and women using `table`, you would type

```
table sex, contents(mean bmi sd bmi)
```

6.1 Use `table` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately

....

....

7 Further exercises

7.1   What is the average age of the subjects ?   . . . . . .

7.2   Draw a histogram of the ages, using the command `histogram age`. Do the ages follow a normal distribution ?

. . . .

. . . .

7.3   How old are the youngest and oldest males and females in the study ?

	Youngest	Oldest
Males	. . . . . .	. . . . . .
Females	. . . . . .	. . . . . .

7.4   What is the mean of the self-reported BMI. Is this greater or less than the mean of the BMI as measured by the nurse ?

. . . .

. . . .

7.5   Create a variable for the difference between measured BMI and self-reported BMI:

```
gen bmidiff = bmi - bmirep
```

Write down its mean value, standard deviation and the number of subjects for whom both BMI measures are available.

Mean   . . . . . .

Standard Deviation   . . . . . .

Both available on . . . . subjects

7.6   Produce histograms of height in men and women, using the commands

```
histogram nurseht, by(sex)
```

and

```
histogram nursewt, by(sex)
```

Add a superimposed normal distribution to the graphs with the `normal` option to the `histogram` command.

7.7   Start a Word document. Select the Graph window in stata and use Ctrl+C to copy the graph. Switch to the Word document and use Ctrl+V to paste the graph.