**Analysis of continuous variables II practical**

In this practical session we will learn how to use Stata to explore means of continuous variables and conduct one sample t-tests. We will use the dataset that you have been using throughout the course so far. We will look at associations between two variables in this dataset and whether there is suggestion of confounding or effect modification by other variables.

There is a video which shows you how to use the Stata commands. You might choose to watch the video first and then work through the practical exercise by yourself. Or you might instead choose to work through the practical alongside watching the video.

Remember!

* Open up the lecture notes/slides which match this session
* Make sure you save your commands in a do file
* Some of the Stata commands you will use in this session are ones you will have seen before, so you should open do files used in previous sessions.

**A. Exploring continuous variables**

***A1: Calculating means and 95% confidence intervals (CIs) of variables***

In this first section we will explore some of the continuous variables in our dataset by calculating their range, mean and 95% CIs. There are at least two Stata commands we can use to show the mean of a continuous variable. In this practical we will use the Stata commands –summarize- and –mean-. The summarize command gives the number of observations, mean, standard deviation, plus minimum and maximum values. The mean command reports the number of observations, mean, standard error and the 95% CI.

*Q. Describe the distribution of BMI (variable name bmi) in our sample based on its minimum and maximum values, mean and 95% CIs.*

*A. To find out the minimum and maximum values of BMI in our sample we can use the summarize command in Stata.*

. summarize bmi

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

bmi | 2,930 27.53164 4.492647 14.80984 52.47276

*There are 2,930 participants with an observed BMI measure in our sample. The mean BMI is 27.5 kg/m2 (standard deviation = 4.5). We see that the minimum BMI is 14.8kg/m2 and the maximum is 52.5kg/m2. The summarize command does not give us the 95% confidence intervals. To get these we need to use the mean command.*

It is good practice to include the units for the values. For example, the mean BMI is 27.5**kg/m2**

. mean bmi

Mean estimation Number of obs = 2,930

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 27.53164 .0829982 27.3689 27.69438

--------------------------------------------------------------

*The output from the mean command shows that again the mean BMI in our sample is 27.5kg/m2. This time we get the standard error in the output rather than the standard deviation. The 95% CI ranges from 27.4 – 27.7. The 95% CI gives 95% coverage of the population mean BMI that are compatible with the study data.*

. summarize dbp

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

dbp | 2,692 77.32448 11.89693 45 133.5

*Q. Now let’s do the same for diastolic blood pressure (variable name dbp).*

*A. We first use the summarize command to find the mean and range of DBP.*

*There are 2,692 participants with an observed DBP value in our sample. The mean DBP is 77.3mmHg (standard deviation = 11.9). The minimum DBP is 45mmHg and maximum is 133.5mmHg.*

*We use the mean command to find the 95% CIs.*

. mean dbp

Mean estimation Number of obs = 2,692

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

dbp | 77.32448 .2292964 76.87486 77.77409

--------------------------------------------------------------

*The 95% CIs ranges from 76.9mmHg to 77.8mmHg, showing 95% coverage of the population mean DBP that are compatible with the study data.*

***A2: Calculating means and 95% CIs for different sub-groups***

Sometimes we might want to show the mean values of a continuous variable for different groups of our population. For example, showing the mean BMI of men and women separately. You can do this in Stata by combining the mean commands with the ‘if’ option.

*Q: Show the mean BMI and 95% CIs for men and women in this sample separately. What do you conclude?*

*A: It is a good idea to first check the coding of the variable you are planning to use in the ‘if’ statement first using the codebook command.*

. codebook sex

---------------------------------------------------------------------------sex Sex: 1=men,2=women

---------------------------------------------------------------------------

type: numeric (byte)

label: sex

range: [1,2] units: 1

unique values: 2 missing .: 0/3,129

tabulation: Freq. Numeric Label

1,368 1 Men

1,761 2 Women

*This shows us that men are coded as 1 and women as 2. We can now use this information to calculate the mean and 95% CIs for men and women separately.*

. mean bmi if sex==1

Mean estimation Number of obs = 1,292

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 27.48871 .1069373 27.27892 27.6985

--------------------------------------------------------------

*Here is the Stata output for men. The mean BMI of men is 27.5kg/m2 and the 95% CIs range from 27.3kg/m2 – 27.7kg/m2.*

*We repeat this to find the mean BMI and 95% CIs for women.*

. mean bmi if sex==2

Mean estimation Number of obs = 1,638

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 27.5655 .1221986 27.32582 27.80518

--------------------------------------------------------------

*The mean BMI for women is 27.6kg/m2 and the 95% CI is 27.3kg/m2- 27.8kg/m2. BMIs are very similar for men and women in our sample.*

We can use this method to explore the means of any number of different sub-groups in our sample in a descriptive manner. In the above exercise we just reported the means and 95% CIs for two sub-groups (men and women), but we could have looked at differences by a variable with more than two categories, such as physical activity, ethnicity or social class.

*Q. Report the means and 95% CIs for BMI for the different physical activity categories (variable physact). What do you conclude?*

*A. Again we should first check the coding of the physical activity variable (physact).*

. codebook physact

---------------------------------------------------------------------------

physact 3 categories of physical activity (1=low, 3=high)

---------------------------------------------------------------------------

type: numeric (byte)

label: adt30gp

range: [1,3] units: 1

unique values: 3 missing .: 1/3,129

tabulation: Freq. Numeric Label

1,321 1 Group 1 -low

1,051 2 Group 2 - medium

756 3 Group 3 - high

1 .

*We can see that there are three categories of physical activity. 1 = low, 2 = medium and 3 = high.*

*Let’s now report the mean BMIs and 95% CIs for these three groups.*

. mean bmi if physact==1

Mean estimation Number of obs = 1,215

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 28.33287 .1385862 28.06098 28.60477

--------------------------------------------------------------

*The mean BMI of people with low physical activity is 28.3kg/m2  (95% CI: 28.1, 28.6).*

. mean bmi if physact==2

Mean estimation Number of obs = 1,000

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 27.25954 .133911 26.99676 27.52232

--------------------------------------------------------------

*The mean BMI of people reporting medium physical activity is 26.6kg/m2 (95% CI: 27.0, 27.5).*

. mean bmi if physact==3

Mean estimation Number of obs = 714

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 26.56143 .1488129 26.26927 26.85359

--------------------------------------------------------------

*The mean BMI of people reporting high physical activity is 26.6kg/m2 (95% CI: 26.3, 26.9).*

*The mean BMIs of people decreases with increased physical activity. You will learn how to formally test such differences in practical xx.*

**B. One sample t-tests**

***B1: Comparing means in our sample to population means***

We can use a one sample t-test when we want to compare the estimates from our sample with those from the general population. To do this in Stata we use the ttest command as follows:

ttest <variable name>==<population value>

*Q. We know from another study that the mean BMI of* ***adults aged 60+*** *in the whole population is 28.5kg/m2. Are adults aged 60+ different from the general population in terms of BMI?*

*A: Using the Stata command we can perform a one-sided t-test to assess whether the mean BMI of adults aged 60+ in our sample is different from the population mean.*

. ttest bmi==28.5 if age>=60

One-sample t test

---------------------------------------------------------------------------

Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+-----------------------------------------------------------------

bmi | 1,409 27.64003 .1134224 4.257496 27.41753 27.86252

---------------------------------------------------------------------------

mean = mean(bmi) t = -7.5821

Ho: mean = 28.5 degrees of freedom = 1408

Ha: mean < 28.5 Ha: mean != 28.5 Ha: mean > 28.5

Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

*The output shows that we had 1,409 adults age 60+ with BMI values in our sample. The mean BMI of these adults was 27.6kg/m2 (95% CI: 27.4, 27.9). Under the output it shows our null hypothesis is that the mean = 28.5 (ie that there is no difference between our sample and the population). The Student t statistic is 7.58.*

*In the bottom left on the output (Ha: mean <28.5) – this shows the one-tailed p-value testing the null hypothesis against the alternative that the mean in our sample is less than 28.5kg/m2. The P value associated with this test is very low (<0.001) and suggests low compatibility with the null hypothesis.*

*On the bottom right of the output (Ha: mean >28.5) – this shows the one-tailed p-value testing the null hypothesis against the alternative that the mean in our sample is greater than 28.5kg/m2. The p value associated with this very high, suggesting high compatibility with the null hypothesis.*

*In the bottom centre (Ha: mean != 28.5) is the two-tailed p value testing the null against an alternative hypothesis that the mean does not equal 28.5. Again the p value is very low <0.001, suggesting low compatibility with the null hypothesis.*

In Stata language != means “does not equal”

*We can conclude that our sample mean BMI is different from the general population mean. The mean BMI in our sample is lower than the population mean.*

*Q. What about if the general population mean BMI was 27.5kg/m2? Do our conclusions change?*

*A: Using the same process as in the previous exercise:*

. ttest bmi==27.5 if age>=60

One-sample t test

---------------------------------------------------------------------------

Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+-----------------------------------------------------------------

bmi | 1,409 27.64003 .1134224 4.257496 27.41753 27.86252

---------------------------------------------------------------------------

mean = mean(bmi) t = 1.2346

Ho: mean = 27.5 degrees of freedom = 1408

Ha: mean < 27.5 Ha: mean != 27.5 Ha: mean > 27.5

Pr(T < t) = 0.8914 Pr(|T| > |t|) = 0.2172 Pr(T > t) = 0.1086

*In contrast to the previous results, the Students t statistic is low (t=1.23). Looking at the results testing the three alternative hypotheses at the bottom of the output we can see that the p values for all three are relatively high. These findings suggest that our reasonably high compatibility with the null hypothesis that the mean BMI in our sample is the same as the population mean.*

***B2: One sample t-tests for sub-groups***

We can also conduct t-tests for specific sub-groups in our sample. For instance, we might want to know whether the mean BMI of people aged 60+ is higher in those who report low physical activity compared to the whole sample.

*Q: Find out whether the mean BMI of people aged 60+ who report low physical activity is higher than the mean for the whole sample. You can identify those who reported low physical activity using the variable physact. You also need to first calculate the mean for those aged 60+.*

*A. Firstly calculating the mean for people aged 60+*

. mean bmi if age>=60

Mean estimation Number of obs = 1,409

--------------------------------------------------------------

| Mean Std. Err. [95% Conf. Interval]

-------------+------------------------------------------------

bmi | 27.64003 .1134224 27.41753 27.86252

--------------------------------------------------------------

*The mean BMI of adults aged 60+ in our sample is 27.6kg/m2 (95% CI: 27.4, 27.9).*

*We can now use this mean in a one sample t-test to assess whether the mean BMIs of people who report low physical activity is different from the mean BMI for the whole sample.*

. ttest bmi=27.6 if age>=60 & physact==1

One-sample t test

---------------------------------------------------------------------------

Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+-----------------------------------------------------------------

bmi | 693 28.15418 .1717148 4.520374 27.81704 28.49133

---------------------------------------------------------------------------

mean = mean(bmi) t = 3.2273

Ho: mean = 27.6 degrees of freedom = 692

Ha: mean < 27.6 Ha: mean != 27.6 Ha: mean > 27.6

Pr(T < t) = 0.9993 Pr(|T| > |t|) = 0.0013 Pr(T > t) = 0.0007

*The output shows that the mean BMI for people aged 60+ who reported low physical activity is 28.2kg/m2 (95% CI: 27.8, 28.5). The Student t statistic is 3.2. Looking at the results of tests of the three alternative hypotheses at the bottom of the output – on the bottom left, the p value from the test of the alternative hypothesis that the mean BMI of those who report low physical activity is* lower *than for the whole sample is high. This suggests high compatibility with the null hypothesis. In the bottom middle, the alternative hypothesis that the mean is different from 27.6 is supported by the data (p=0.001). On the bottom right, the p value is also low (p=0.001) suggesting high compatibility with the alternative hypothesis that the mean BMI for those reporting low physical activity is higher than the whole sample mean.*

**Formative exercise**

If you would like to look at more examples of conducting one sample t-tests you might like to work through the optional exercise below.

Q: Do participants in non-manual social classes aged 60+ have a different mean BMI to the whole sample aged 60+? Note you will need to create a new binary social class variable combining social classes I, II, IIIN as ‘non-manual’ and social classes IIIM, IV, V as ‘manual’.

Q: Do participants in non-manual social classes aged 60+ have a different mean DBP to the whole sample aged 60+?