TOPIC 9
COMPUTER PRACTICAL
COMPUTER ANALYSIS OF CONTINUOUS VARIABLES:
COMPARING TWO MEANS & ANALYSIS OF VARIANCE

1. Introduction

In the practicals for **Sampling Variability of a Mean** and **Comparing Two means**, you learned how to calculate confidence intervals for the mean of a continuous variable, and how to compare means between two groups using a t-test. You also learned how to compare two means when they are based on separate measurements for the same person (paired data). All the calculations in those sessions were carried out on paper using a calculator.

In the lecture for **Comparing Two means**, you were also introduced to the concept of analysis of variance for comparing means between two or more groups.

In this session, we will learn how to carry out these tasks in Stata.

Using Stata it is very easy to look at the distributions of variables and this is always recommended as the preliminary stage of an analysis.

In this session we are going to cover some of the techniques used to analyse continuous response variables. These include:

- Describing the overall distribution of the response
- Calculating the confidence interval for the mean
- Describing the distribution of the response in sub-groups defined by some other variables
- Comparing the mean in different sub-groups using t-tests
- Comparing the mean in different sub-groups using ANOVA

2. Getting started

- Open Stata
- Open a new Do file
- Write some comments to yourself to remind you what the purpose of the Do file is
- Change the working directory
- Load the bab9 file

3. Describing the distribution of a continuous variable

In Practical 8 we saw that the command

```
summ bweight, detail
```

produces a detailed description of the distribution of bweight. This includes the mean and standard deviation, the median, some percentiles, the four smallest and largest values and the number of non-missing observations. It is useful to accompany this summary with a histogram of the distribution of bweight (which we also looked at in Practical 8), by typing:

```
hist bweight, bin(12) freq
```

and

```
hist bweight, bin(12) freq norm
```

The choice of 12 bins - that is, with birthweight divided into 12 groups - is arbitrary but seems to work OK. The normal distribution, with the same mean and standard deviation as that observed in the dataset, is superimposed in the second plot.
Try repeating this command, leaving the number of bins at the default (that is by omitting the option `bin`(12).
Look at the distribution of gestation age (`gestwks`), in the same way.

What would you say about this distribution?

Neither birthweight nor gestational age have truly symmetric distributions. Furthermore, the distribution of gestational age suggests a sizeable number of premature births.
Type in a suitable command to get the histogram for the distribution of birthweight for babies of more than 32 weeks gestation.

Does this appear to be closer to a Normal distribution?

4. Confidence intervals for the mean

The mean and the confidence interval can be estimated using the command `ci means` (the command `mean` could also be used to give the exact same result, the only difference between the two commands occurs when you apply them to several variables at once and some have missing values, which is not an issue with this dataset). Type:

```
ci means bweight
```

One can ask for confidence intervals for several variables at a time:

```
ci means bweight gestwks matage
```

One can also vary the confidence level using the option `level()`. For example, you can ask for 99% confidence intervals:

```
ci means bweight gestwks matage, level(99)
```

These confidence intervals are based on the t-distribution and, strictly, this assumes that the data follow a normal distribution. We have noted that the distributions are somewhat skewed but this effect is not very large and the sample size is quite large. Therefore, one can trust these confidence intervals on account of the central limit theorem (which we recall states that the distribution of a mean is closer to the normal distribution than the original data, and that this effect grows stronger with increasing numbers of subjects).

5. Distribution of the response variable across subgroups

Suppose we wish to investigate the relationship between sex and birthweight, in other words we want to assess whether boys and girls have similar mean birthweight. We start by looking at the mean birthweight for each sex. These means can be obtained by typing

```
tab sex, summarize(bweight)
```

1) *Does this suggest that birthweight is similar for girls and boys?*

Generally speaking, it only makes sense to compare means for distributions which are roughly normal (or at least symmetrical) and which have the same spread of values (approximately the same SD). The importance of this assumption depends on the number of subjects, because of the central limit theorem. The distributions in the two groups are quite similar to the overall distribution as you can see using the commands:

```
histogram bweight, by(sex,total) normal percent bin(20)

graph box bweight, over(sex,total)
```

The option total asks Stata to include a plot of the distribution of birthweight for both sexes as well as plots for boys and girls. The histogram looks rather messy. Stata's graph commands are very flexible and with practice you can draw very good quality graphs. This is, however, beyond the scope of this course since the commands are often long and complex. For a better looking graph use the following command[1]:

```
histogram bweight, by(sex,total cols(3)) normal percent bin(20)
xlabel(0 2500 5000,format(%9.0f))
```

2) *Do you think the distributions are roughly normal with a similar spread of values? Are the sample sizes still reasonably large in the two groups?*

6. Comparing means

In **Comparing Two means** you used the z-test and the t-test to assess evidence for a difference between two separate groups of subjects. Both tests assume the response variable follows a (roughly) normal distribution in each of the groups being compared, and we have now looked at this for birth weight in this dataset. The t-test also assumes that the standard deviation of the response variable is the same in each group, and takes account of the fact that its value must be estimated, while the z-test allows for different standard deviations but assumes them to be known. The t-test is recommended for small samples where there could be appreciable additional uncertainty due to estimating the standard deviation in the two groups.

In Stata, there is only one command to compare two means: `ttest`. When one is comparing two groups, it has options that allow one to let the standard deviations on the two groups be different, or to require them to be the same[2]. It always uses the t-distribution to go from the test statistic:

$$t= \frac{\textit{difference between the means}}{\textit{SE of the difference}} \qquad (1)$$

to the p-value. Recall that when the number of subjects, and therefore the degrees of freedom, are large enough, p-values under the normal (z) and the t-distributions are virtually the same.

We start with a one-sample test. This is a test of the hypothesis that the true mean of a variable in a data set is equal to a selected value. The command:

```
ttest bweight=3300
```

---

[1] If you are interested... `by(sex, total cols(3))` instructs Stata to draw a plot for boys, girls and both sexes and to arrange these side by side in three columns. `xlabel(0 2500 5000, format(%9.0f))` tells Stata to label the x-axis at 0, 2500 and 5000 grammes and to display the x-values without decimal places.

[2] When the standard deviations are allowed to be different, there a special formula to adjust the degrees of freedom which ensures the test is appropriate in this situation (see next page).

tests whether the true mean birthweight of BAB babies is equal to 3300g, the England and Wales national average for normal births.
The null hypothesis for this test is that the estimated mean birthweight (in the population) is 3300 grammes. This is given by Stata as: `Ho: mean(bweight) = 3300`
Stata also gives three sets of hypotheses and p-values. We want to test the hypothesis that the mean birthweight is not 3300g. We don't mind whether the mean is greater or less than 3300 and therefore we are interested in the central p-value[3] which is given under the hypothesis:
`Ha: mean != 3300`
This is given as Pr(|T| > |t|) = 0.0000 which we would write as P<0.0001

3) *What do you conclude from this test?*

A two-sample test is a test of the null hypothesis that the true mean outcome is the same in each of two groups. To test whether the mean of the birthweight of babies is the same for boys and girls, type:

```
ttest bweight, by(sex)
```

4) *What do you conclude about the difference in birthweight between boys and girls?*

This version of the test assumes that the standard deviations of the response in the two samples are the same "in the population" (or roughly the same in the data). This assumption can be tested statistically by typing:

```
sdtest bweight, by(sex)
```

This test calculates the ratio of the standard deviation of birthweight among males/SD of birthweight among females. The null hypothesis is that the SD are the same in both groups
`ratio = sd(male) / sd(female)          Ho: ratio = 1`
Again, you want the p-value from the central set of results- which tests whether the ratio is not equal to one. If this p-value is very small then it is very likely that the SD of the two samples are not the same.
If the assumption of equal standard deviations was found to be inappropriate, you could use this version of the command:

```
ttest bweight, by(sex) unequal
```

This allows for unequal standard deviations when it calculates the standard error of the difference. An alternative is to carry out a non-parametric or distribution-free test; these will be covered in Topic 20.

5) *Was it reasonable to use the t-test with equal; standard deviations to compare birthweights between boys and girls?*

[3] The other alternatives relate to one-tailed tests, and we do not generally recommend these. You would only use a one-tailed test if you were interested in testing for a particular difference i.e. "mean A is larger than mean B" instead of "mean A is different from mean B". You need a good reason to do this, and should have decided to do this before you look at the data.

7. The analysis of variance for two groups

Analysis of variance (ANOVA) compares the means of an outcome variable across groups defined by another variable. The purpose is to assess how much of the observed difference between the group means is due to differences in the variance of the outcome variable estimates for each group. Again, this test works on the principle that the observed difference between the means may result from chance variation in the values observed in the different groups.

When there are only two means to be compared analysis of variance and the t-test are essentially the same. So why do we need to use analysis of variance? Remember that, using the t-test, you can only compare two means. Using analysis of variance you can compare several means. So in the BAB data we could compare the mean birth weight in all 4 groups of maternal age to see whether the babies' birth weights are associated with the age of the mother.

In **Comparing Two means**, we saw how analysis of variance works by partitioning the variance into the between-group variance, and the within-group variance. If we carry out an analysis of variance of birthweight by sex then the within-group variance will measure how far the individual babies' birthweights deviate from the mean male birthweight for boy babies, and how far they deviate from the mean female birthweight for girl babies. The between-group variance measures the amount of variability in mean birthweight between boys and girls.

Suppose that there is no true difference between boys' and girls' mean birthweights. If this is true, then the girls' and boys' means in our data are different only because their sample means vary about a single underlying (unobserved) population mean. If that is true, then the between-sexes (between-group) mean square (MS) will, on average, be the same as the within group mean square and any tendency to be larger will be due to chance. This is the situation under the null hypothesis.

The F statistic is therefore used to assess evidence for whether the between-group mean square is greater than the within-group mean square. Under the null hypothesis, (and assuming normally distributed data) it has a theoretical basis (the F distribution) from which a p-value can be calculated.

Let's see this in practice. We will use the command oneway. Stata has several commands for carrying out analysis of variance. oneway is the most appropriate for comparing the mean of one variable across groups defined by another variable. Try:

```
oneway bweight sex, tabulate
```

It produces the following output:

```
    Sex of |       Summary of Birthweight
    infant |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
      male |   3211.2791   665.97976         326
    female |    3044.127   628.66035         315
-----------+------------------------------------
     Total |   3129.1373   652.78265         641
```

Between and within groups variance

F= Between groups MS / Within groups MS
In this example there is 10 time as much variation between groups as within groups

The difference is unlikely to be due to chance

```
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups       4476041.4      1   4476041.4     10.66     0.0012
 Within groups       268244081    639   419787.293
------------------------------------------------------------------------
    Total            272720122    640   426125.19

Bartlett's test for equal variances:  chi2(1) =   1.0593  Prob>chi2 = 0.303
```

Bartlett's test checks the vailidity of the assumption that each group has a similar population standard deviation. A large p value indicates that this assumption is very likely to be valid.

The first part of the output does not appear unless you type the option tabulate. It is simply a useful reminder of the means in each group. The second part always appears: it is the analysis of variance itself. Look at the "Within groups" MS. The value of this is 419787.293 which is the within-group variance. The male and female mean birthweights also vary from the mean birthweight for all the babies together and the measure of this variation is given by the "Between groups" MS, which is 4476041.4. You can see there is a big difference between these values, therefore the F statistic is large and the small p-value shows that the difference is unlikely to be due to chance.

Compare the p-value to that obtained using the t-test.

The first column in the results table, SS, gives quantities called Sums of Squares. The mean square (MS) is the SS divided by the appropriate degrees of freedom.

8. Analysis of variance for several groups

It is easy to extend the analysis of variance from two groups to more than two. Try comparing the mean birth weight in the four categories of maternal age group:

```
oneway bweight matagegp, tabulate
```

The output is very similar to the output from the previous test. There are now four means in the top table. Notice that the analysis of variance now shows between groups variance to have 3 degrees of freedom. We have 4 groups of maternal age, so *k*-1 gives 3 degrees of freedom.

6) From the results of this test:

a) What is the null hypothesis being tested in the analysis of variance?

b) What is the result of the test?

9. Non-parametric tests

Distribution-free tests do not assume that the distributions being compared are normal, so are useful alternatives in these situations. They are also called "non-parametric tests" because they do not work through the parameters of the normal or other distributions. Below there are two examples of these kinds of test along with the output you can expect.

Wilcoxon rank sum test

The 'Wilcoxon Rank Sum test' (also called 'Mann-Whitney test'), is a distribution-free alternative to the t-test, and is used to test the hypothesis that the distributions in the two groups have the same median. For this test you will need to use the BABNEW.DTA dataset. The command in order to compare the birthweight in boys and girls is:

```
ranksum bweight, by(sex)
```

The output is as follows:

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         sex |      obs    rank sum    expected
-------------+---------------------------------
        male |      326    113074.5      104646
      female |      315     92686.5      101115
-------------+---------------------------------
    combined |      641      205761      205761

unadjusted variance  5493915.00
adjustment for ties     -302.63
                     ----------
adjusted variance    5493612.37

Ho: bweight(sex==male) = bweight(sex==female)
             z =   3.596
    Prob > |z| =   0.0003
```

The p-value of 0.003 can be interpreted as strong evidence that the median birthweight of males and females is different, however it is not immediately obvious from the output which is higher. To interpret this it would also be useful to obtain the median observations from each group.

```
bysort sex: summarize bweight, detail
```

Wilcoxon signed rank test

The distribution-free equivalent of the t-test **for matched pairs** is the 'Wilcoxon signed rank test', with Stata command `signrank`. Using the dataset NONPARAMETRIC.DTA, which contains 15 pairs of skinfold measurements, with each pair being a skinfold measurement on a single individual by two observers A and B. We want to test for a difference in the two observers, so we should first check the distribution of the outcome variable in each group (using histograms):

```
hist sfa, bin(5)
```

```
hist sfb, bin(5)
```

Experiment with changing the `bin` option. Do you think the distributions are Normal? With so few observations, it can be difficult to judge, so it is often safer to use a nonparametric test.
We can apply the command:

```
signrank sfa=sfb
```

Which gives output of:

```
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       14       109.5          60
    negative |        1        10.5          60
        zero |        0           0           0
-------------+---------------------------------
         all |       15         120         120

unadjusted variance       310.00
adjustment for ties        -0.13
adjustment for zeros        0.00
                        ----------
adjusted variance         309.88

Ho: sfa = sfb
             z =   2.812
    Prob > |z| =   0.0049
```

The p-value of 0.0049 provides strong evidence that there is a difference between the two observers (With A measuring systematically higher than B).
For comparison, run a t-test.

7) What do you conclude?