statstutor

community project

encouraging academics to share statistics support resources

stcp-marquier-continousS

The following resources are associated:
The Birthweight dataset.csv, ANOVA in SPSS

Summarising Continuous Variables by group in SPSS

Dependent variable: Continuous

Independent variable: Categorical

Data: The data set 'Birthweight_reduced.sav' contains details of 42 babies and their parents at birth e.g. birthweight, mothers age and whether or not the mother smokes (smoker). Open the SPSS dataset '*birthweight_reduced.sav'*.

Birthweight	Gestation	smoker	motherage	mnocig
5.80	33	0		0
4.20	33	1		7
6.40	34	0		0
4.50	35	1	41	7

Mother smokes = 1

Choosing the right summary statistics

Different summary statistics are appropriate depending on the distribution of the data. Use histograms to check if the data is approximately normally distributed.

Normally distributed data	**Skewed data**
Average: Mean **Measure of spread**: Standard deviation Similar mean and median	**Average**: Median **Measure of spread**: Interquartile range Very different mean and median
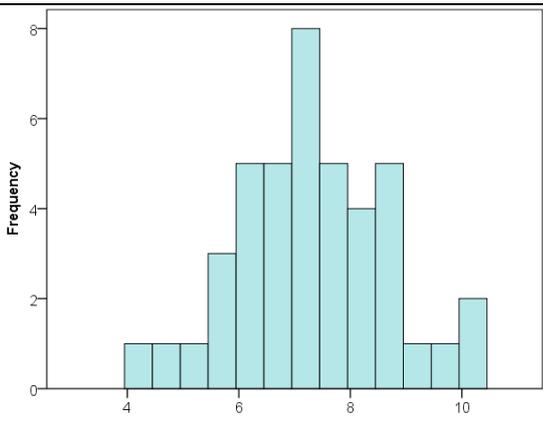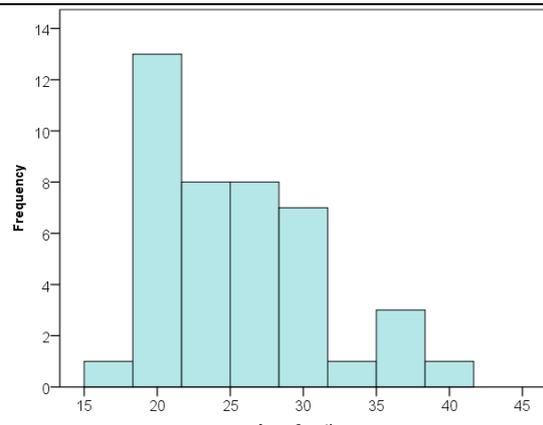
---

Summarising continuous variables by group in SPSS

Research question: Does a mother smoking have an effect on the birthweight of a baby? The dependant variable is Birth weight (lbs) and the independent variable is whether or not the mother smokes (smoker).

Data Visualisation

To display the information graphically, use either histograms or boxplots by group.

To plot histograms representing the distribution of Birthweight for the smoker and non-smoker groups, click on
Graphs → Legacy Dialogs →Histogram
Put *Birthweight* in the 'Variable' box and *smoker* in the 'Columns' box to plot the histograms side by side or in 'Rows' to plot one on top of the other. Ticking 'Display normal curve' will add the curve to the plot to help assess normality. Then click **OK** to obtain the histograms below.
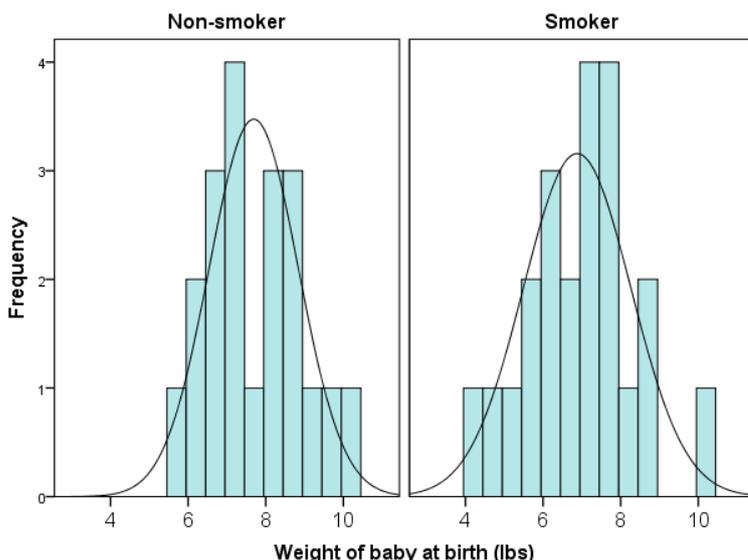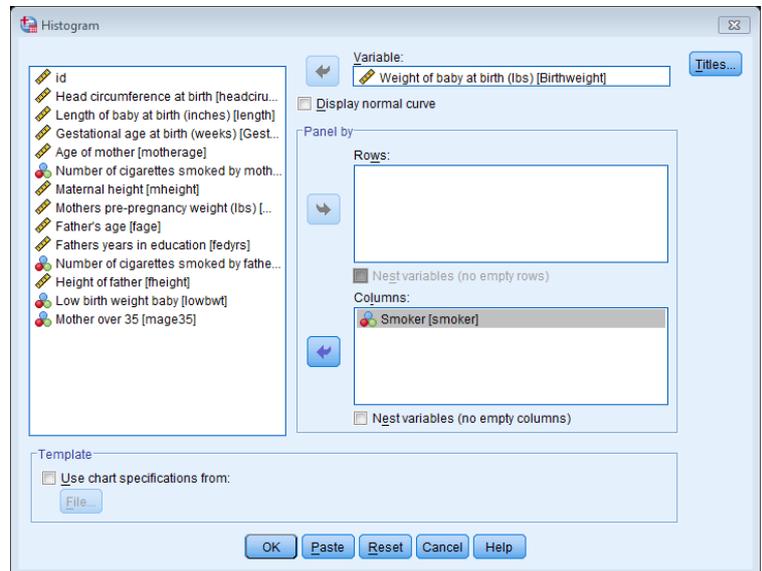
The two histograms show that birthweight is approximately normally distributed for the babies of smokers and non-smokers. So it is appropriate to use means and standard deviations to describe the data and an independent t-test to look for a significant difference between the groups.

However, it's not easy to see if there is a difference between the two groups from the histograms. Boxplots or plots with confidence intervals are usually clearer when looking for differences between groups. Box-plots help visualise the median, upper and lower quartiles of the data, and identify outliers.
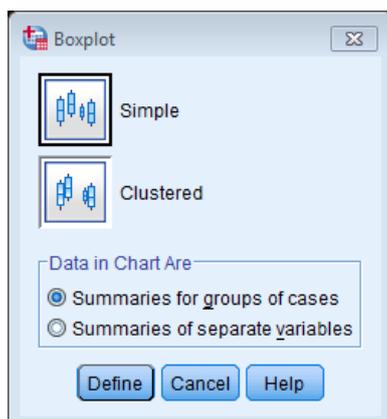
To produce a box-plot, go to *Graphs → Legacy Dialogs → Boxplots*. Then click 'Simple' and 'Summaries of groups of cases'. After this, put the continuous variable of interest, i.e. *Birthweight*, inside the 'Variable' box and move the categorical variable *smoker* in the 'Category Axis' box. Then click **OK** to obtain the following boxplots:

Summarising continuous variables by group in SPSS
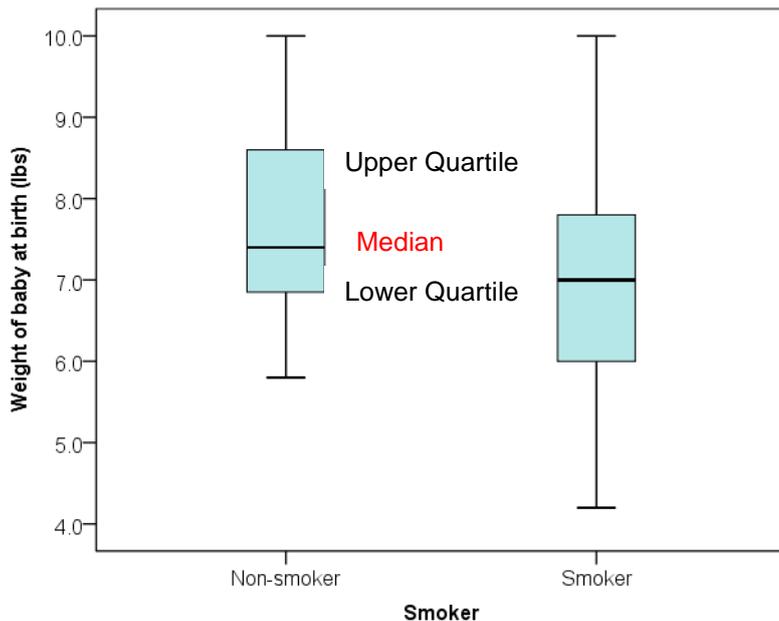
The boxplot represents the spread of birthweight for the smoker and non-smoker groups. The median is represented by the line in the middle of the box. The box limits represent the first quartile and the third quartile, and thus represents the middle 50% of the data called the "Inter Quartile Range". The longer the box, the more spread out the data are. The bottom and top whiskers represent the minimum and maximum values unless there is an outlier when the whiskers represent 1.5 times the interquartile range. If an outlier is present, then it will be represented by a circle with a row identifier for that individual. Extreme outliers are represented by a star.

Non-smoker mothers seem to have given birth to slightly heavier babies than smoker mothers. The smoker box contains a bigger range of values, which suggests a higher variation in the birth weight of babies born from a mother who smokes.

Summary statistics

There are several methods for producing summary statistics by group through the **Analyze** menu:

Descriptive Statistics → Explore
Tables→Custom Tables
Compare Means → Means

The **Explore** option gives a large number of statistics but the output is messy.
Custom Tables is good for summarising lots of variables at once in any style of your choosing but can be tricky to use. The **Compare Means** option gives a nice table with options for most summary statistics and is easy to use so we will use that here.

Go to *Analyze → Compare means → Means*, move *Birthweight* into the 'Dependent List' box and variable *Smoker* in the Independent List. Then you should click on 'Options'.
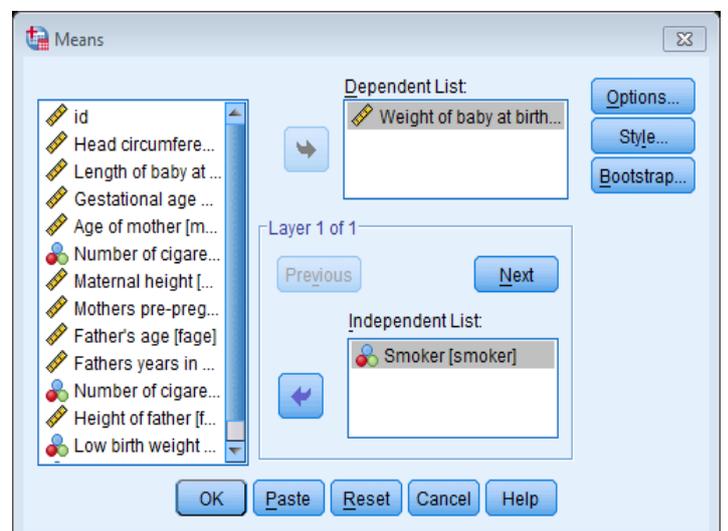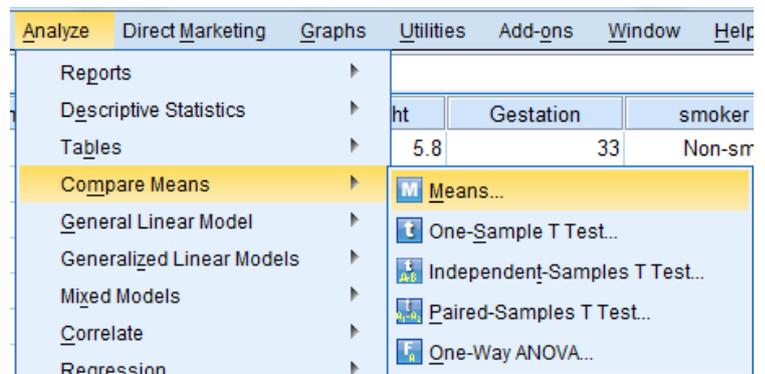
Summarising continuous variables by group in SPSS
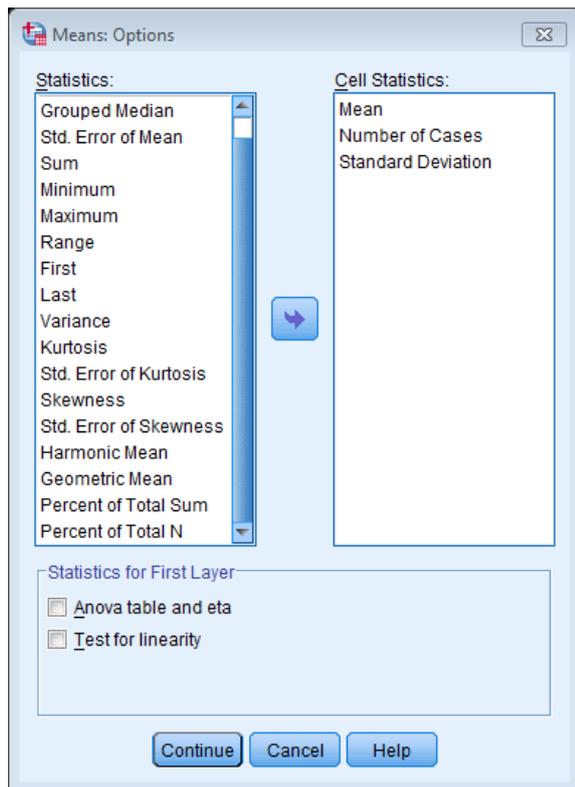
By default the Mean, standard deviation and number of cases are given but there are numerous options on the left hand side e.g. median. Move your chosen statistics to the 'Cell Statistics' box

By clicking **Continue** and **OK**, the following table is obtained:

Report
Weight of baby at birth (lbs)

Smoker	Mean	N	Std. Deviation
Non-smoker	7.690	20	1.1480
Smoker	6.877	22	1.3894
Total	7.264	42	1.3297

When reporting results from SPSS, it is often preferable to reduce the number of decimal places. This can be done by double clicking on the table in the output window, highlighting the cells to be edited, right clicking and opening the **Cell Properties** window. Reduce the number of decimals in the **Format Value** tab.

Report
Variables Weight of baby at birth (lbs)

Smoker	Mean	N	Std. Deviation
Non-smoker	7.7	20	1.1480
Smoker	6.9	22	1.3894
Total	7.3	42	1.3297

Delete
Select Table
Select cells with similar significance
Sort Rows
Create Graph
Table Properties...
Cell Properties...
TableLooks...
Insert Footnote

Format Value

Interpretation of the results

Do the group means and standard deviations look similar or very different?

All samples will vary slightly just by chance so only classify big differences as 'different'. When carrying out tests comparing independent groups (Independent t-tests and ANOVA), the variances (standard deviation squared) need to be similar. A big difference is when one standard deviation is more than twice the other.

The mean birthweight for babies of smokers (Mean = 6.9) is 0.8 lbs lower than the mean for non-smokers (Mean = 7.7). The standard deviations are quite similar so the groups are equally spread out.