**SPSS - Testing for normality**

Previouslyy we described quantitative data using graphs e.g. histogram, boxplots and we also calculated numerical summaries e.g. mean, median, standard deviation etc. How do we decide which numerical summaries to use for a particular set of data? First we must decide if the data is normally distributed.

* If the data is normally distributed we will use the mean as the measure of centrality and the standard deviation as the measure of variability.
* If the data is not normally distributed we will use the median as a measure of centrality and the interquartile range as a measure of variability.

How do we check if the data is normally distributed? We can compare the values of the mean and median (they should be similar for a normal distribution), we can look at the shape of the histogram (which should be reasonably bell-shaped and symmetric for a normal distribution), we can also use more formal tests of normality and normal probability plots.

**Example**

The amount of money spent on advertising for 18 companies is as follows:

5000 2500 500 10000 2000 250 20000 6000 250

500 750 6000 7000 4000 2000 1000 3000 2000

Enter the data into one column in SPSS called amount. Select **Analyze** from the main menu and then **Descriptive Statistics** and then **Explore**. Select amount in the Dependent List. Click on **Plots** and tick *Histogram*. Tick *Normality Plots with Tests*. Click continue. Click OK.

Is the variable amount normally distributed?

* The histogram is positively skewed.
* There is an outlier in the boxplot (observation 7 with a value of 20000).
* The mean (4041.66) is greater than the median (2250).



Not all the points are close to the line in the normal probability plot which indicates that the data may not be normally distributed.

**Tests of Normality**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnov(a) | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| amount | .216 | 18 | .025 | .739 | 18 | .000 |

a Lilliefors Significance Correction

The ***Shapiro-Wilk test*** is used when the sample size (n) is less than 50. The p-value from the test (called Sig.) is .000 (written as <0.001).

The p-value gives the probability that the data come from a normal distribution – if the probability is very small (<0.05) then we reject the idea that the data are normally distributed and conclude that the data are not normally distributed.

Here the probability is so small (< 0.001), we conclude that the data is not normally distributed.