**Laboratory Tutorial 1-4: Descriptive Statistics**

In this laboratory tutorial you will:

1. Summarise the demographic properties of a data sample
2. Compute the mean, range and standard deviation for given variables
3. Check for homogeneity of variance between groups within given variables
4. Check distributions of values for given variables for normality

# Preamble

For this tutorial, you need your modified version of the Sleep3ED data file as last used in Lab 1-2. This should contain the new variables that you created i.e. agegp3, sleepdef, BMI and the recoded BMI category variable. If these do not exist, create (or recreate) these before you continue, using the instructions provided in the earlier Labs.

**This is a mandatory tutorial. In order to pass the coursework, you must achieve a score of 50% or higher on the associated Blackboard quiz (Lab Quiz 1-4).**

**\*Note: We strongly suggest you do not start the Lab Quiz for this tutorial before you have all your answers ready.**

# Exercise 1: Demographic properties of the Sleep3ED sample

Normally, when writing up a research report, you would describe the characteristics of your sample at the start of the Methods section. For example, a sub-section called “Sample” or “Participants” may go as follows:

“The sample consisted of 300 students enrolled at Brunel University (52% male, 48% female), ranging in age from 21 to 52 years (mean = 35 years, standard deviation = 12 years).”

This allows the reader to put the results of your analysis into context. Specifically, it allows important questions like the following to be answered:

1. How many people took part in the study in total? Is this large enough to be a reliable estimate of the population as a whole?
   * There are ~180 Level 1 CS/IS students enrolled in DISC. Would a sample of 5 students be sufficient to draw reliable conclusions about the views of the whole cohort?
2. Is the sample representative of the target population?
   * An 80:20 ratio of males to females might be representative of a CS cohort, but not for a social science programme.
   * For a sample of CS undergraduates, would a mean age of 45 be a cause for suspicion about the data?

In the lecture a range of descriptive statistics were introduced. You will now use appropriate SPSS commands to compute the statistics required to complete the blanks[[1]](#footnote-1) in the following sample summary:

**Q1: “Staff working at the University of Melbourne, Australia were invited to complete a questionnaire about sleep behaviour. The sample consisted of \_\_\_ respondents, (\_\_% male, \_\_% female), ranging in age from \_\_ to \_\_ years (mean = \_\_ yrs, standard deviation = \_\_ yrs). “**

**Sample size**

This should be easy to determine just by looking at the number of rows in the data table however it is best to compute descriptive statistics across all relevant variables to be sure, as some cases may contain missing values that render them unusable in all or certain analyses that you plan to do. For the purpose of Q1, you can use the total number of non-empty rows, but as you compute the various descriptive statistics later in the exercise, pay attention to how the number of valid cases varies from one variable to the next.

**Ratio of males to females**

The variable “sex” is a categorical variable, in that each case must belong to one of a finite set of categories and is encoded with the corresponding category number. It makes no sense to compute the average value of “sex”. It is useful, however, to compute the proportion of cases belonging to each category. To do this, we can use the Frequencies command:

Go to: “Analyze 🡪 Descriptive Statistics 🡪Frequencies

Move the variable “sex” into the variables box and click OK. In the resulting output table find the percentage of males and females. You should round these values to the nearest percentage point. The table should also confirm the size of the overall sample.

**Computing the mean, range and standard deviation of ages**

The variable “age” is a continuous variable. Usually, counting the frequency of each occurring value (of which there may be hundreds) is cumbersome and not particularly informative. It is usually better to describe such variables using summary statistics like average (mean, mode, median), range and standard deviation. These statistics can be found in one form or another in many of the SPSS commands. For example, if you click on “Statistics” in the Frequencies command you can select all of the statistics we need here. However, if we just need mean, range and standard deviation it is quicker to use the “Descriptives” command:

Go to: “Analyze 🡪 Descriptive Statistics 🡪Descriptives

Move the variable “age” into the variables box. Click “options” and make sure that all the required statistics (mean, minimum, maximum and standard deviation) are checked. Then click “Continue”, then “OK”. Extract the statistics that you need from the output table.

**Why compute the mean?**

The **mean** is just one *measure of central tendency*, or **average**. It is measured by summing all cases then dividing by the number of valid cases. It is the most commonly used average statistic for continuous data for a number of reasons. It is easy to compute, and it is what most people think of when they hear the word average. However, the mean is also important because it is what most statistical difference tests are using when they measure the significance of differences occurring between groups (e.g. mean height of males vs. females). The mean age tells us, in general terms how old the sample was.

The median and mode are also measures of average. **Median** is the middle value in the distribution when all values are placed in rank order (i.e. the 50th percentile). Median can sometimes be useful as a summary statistic because, unlike the mean, it is unaffected by extreme values (outliers). Finally, the **mode** is simply the most commonly occurring value in the distribution. This is less useful for continuous data where the most common value may only occur a small number of times relative to the overall sample size and there may be a tie for mode – if the ages 25, 29 and 31 are jointly most common, what is the average?

**Understanding standard deviation**

The mean provides us with a way of describing the typical value on a measure. However, in order to compute a parametric statistical test, like t-test, it is also necessary to measure the extent to which values vary around this central value.

There are several ways to measure dispersion or variance. One simple measure is the **range**, composed of the minimum and maximum values within the sample (or the difference between them). Providing that no data entry errors or extreme values (outliers) exist, the minimum and maximum values provide us with a useful measure of how similar or ‘homogeneous’ the sample is with respect to the mean. For instance, if a sample taken from the student population is described as having a mean age of 19 years, a minimum age of 18 years and a maximum age of 21 years it clearly indicates to the reader that this is a sample taken from a typical student population and that no mature students took part in the study.

Frequently, however, outliers will tend to exist within the data, making range less useful. This is especially likely for dependent measures, where we have no control over the values. For example, if all but one of our participants were between 166 cm and 180 cm tall, except for one person who was 200 cm tall, then describing dispersion in terms of the minimum and maximum values is misleading. A more sophisticated measure of dispersion that deals with this problem is called the standard deviation (SD).

This is calculated by:

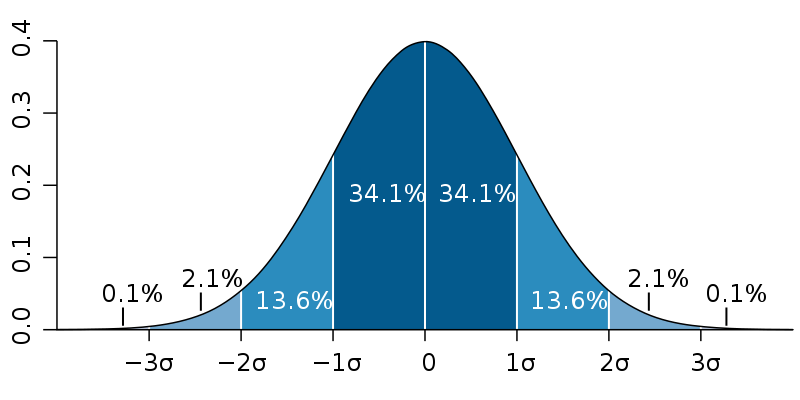
* Computing the difference between each case and the mean
* Squaring each of these values
* Adding them all up
* Dividing by N-1 (or N if the whole population has been sampled)
* Finding the square root of this value

Hence, SD can be thought of as the average difference between case values and the overall mean. Because it is an average of all cases, it is relatively unaffected by the odd extreme value, making it a more robust measure of dispersion than the range.

# Exercise 2: Checking normality

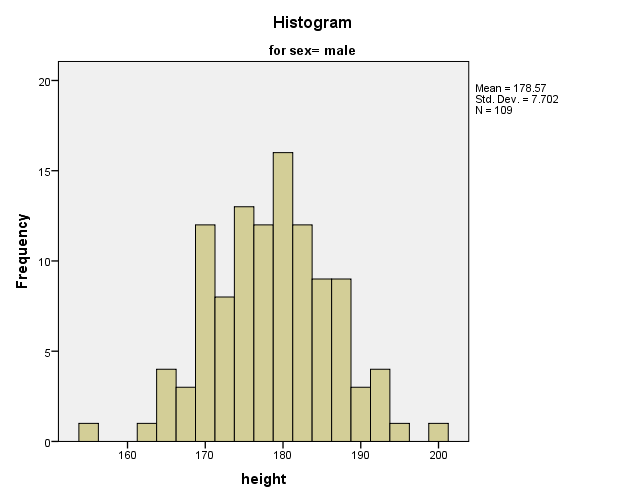
So far, we have discussed how a sample of values for a particular variable will have an average, or measure of central tendency, and a standard deviation or measure of dispersion. When values tend to be spread in an even fashion around the mean, we say that the variable has a normal distribution.

When a variable taken from a sample is normally distributed, we can make estimates about the larger population using the mean and SD values. When plotted as a frequency histogram, a (perfect) normal distribution of values looks like this:

[](http://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg)

Note that the distribution is symmetrical, with the mean, median and mode all occurring at the same value in the middle (50% of cases are above/below the mean). Each interval along the x-axis of the histogram represents one standard deviation. We find that approximately 68% of cases (34% each side) will fall within 1 SD and about 95% (48% each side) of cases occur within 2 SDs.

Why is this useful? Well, if we look at our sleep dataset we find that the distribution height amongst males is roughly normal, with a mean of 178.6 cm and an SD of 7.7 cm. Based on this we can estimate that 95% of all males, assuming this is a representative sample, from the target population (University of Melbourne staff) will be between 163.2 cm and 194 cm tall. Anyone significantly below or above this range would be somewhat unusual and their height value might even be an error.



For instance, we can see that one respondent is recorded as being 155 cm tall. This would place them some 3 SDs below the mean, meaning that they were shorter than 99.8 % of the population. This may or may not be true but it is helpful to know when checking our data.

Further to this, knowing the range in which most values normally fall is useful when it comes to doing statistical tests of difference and correlation. If we estimate the ranges within which a certain percentage of values fall each sample group, we can better judge whether an apparent difference in their means is meaningful (or significant in statistical language) or not. For instance, if two groups differ by 5 points on a measure, is this significant or not? Well, this would depend upon the distribution (specifically the SD) of each group sample. The more the distributions overlap, the less confident we would be that the difference is significant. We will look at difference and correlations tests, along with the notion of statistical significance more closely in lab 1-5.

So, how can we tell if a distribution is normal? One way is to visualise using a frequency histogram as shown above. The simplest way to do this is to use the **Frequencies** command:

* In the main panel, click on Charts
* Select histogram from the “Chart Type” frame
* Click continue and then run the Frequencies procedure as you did before

If you want to visualise the distribution for a particular group or subset of the data (like the histogram shown above), then use the **Explore** function that is also in the Descriptive Statistics sub-menu.

In the real world, we very rarely see a perfectly smooth, symmetrical distribution – the class bell shaped curve. This is particularly true for smaller samples (i.e. N<100). However, providing the distribution looks reasonably normal, then that is normally good enough to use most **parametric** **statistical tests** (e.g. t-test and Pearson’s correlation). To provide additional reassurance, it is also possible to quantify how normal a distribution is. The most common measures – **skewness** and **kurtosis** – were introduced in the Descriptive Statistics lecture.

Both the **Frequencies** and **Descriptives** commands in SPSS will allow you to compute a measure of skewness. A positive skew value means more values in the lower range (peak shifted to the left). A negative skew value refers to the opposite effect. The higher the value, the more exaggerated the skew is. Ideally a variable should have a skewness value of near to zero and certainly less than one. If a distribution is particularly skewed, then one should not use parametric statistical tests. Instead, a **non-parametric** equivalent test should be used. Non-parametric tests are more tolerant of skewed/kurtotic data, because they ignore measures of dispersion, transforming values instead into ranks (ordinal data). We will come back to this later when we look at correlation and difference tests.

Whilst skewness refers to the symmetry, kurtosis on the other hand refers to “peakedness”. A normal distribution is shaped similarly to a bell. A positive kurtosis score means that the curve is too tall (too many values close to the mean), whilst a negative value means that it is too flat (too many values on the tails of the distribution). Clearly, both positive and negative kurtosis means that we can no longer estimate the probability of a particular value occurring based on its distance, in terms of the number of SDs, from the mean. A strong kurtosis value has the same as a strong skewness value when it comes to choosing statistical tests.

Here’s one way to compute skewness and kurtosis statistics using SPSS:

* Go to “Descriptives”, click on “options” and check the option boxes for “Skewness” and “Kurtosis”
* Click “Continue
* Select variables for weight, the number of alcoholic drinks consumed per day (“alcohol”) and the number of caffeine drinks consumed per day (“caffeine”)
* Click OK and compare these variables using the values in the output table to answer the following questions:

**Q2: Which variable has the most normal distribution?**

**Q3: Is “weight” skewed towards the lower or upper end of the value range?**

**Q4: What is the skewness score for “alcohol” (to 1 decimal place)?**

**Q5: Is the distribution of “alcohol” too peaked or too flat?**

# Exercise 3: Comparing groups

Means, ranges and standard deviations can tell us a lot about the overall sample. However, they are also useful when comparing groups within the sample. For instance, two groups might have very similar looking means leading to an initial conclusion that they are very similar in nature. Further analysis, however, might reveal one group to have a much broader distribution or be more skewed than the other. Let’s take an example with the Sleep data. We will see whether being overweight has an effect on the number of hours of sleep people think they need.

* Go to the “Compare Means” command in SPSS
* Move the variable “hourneed” into the “Dependent list”
* Move “BMI\_grp” (or whatever you named it) into the “Independent list”
* Click “options” and move the statistics for mean, number of cases, standard deviation, minimum, maximum, skewness and kurtosis into the “Cell Statistics” list. Click “Continue”.

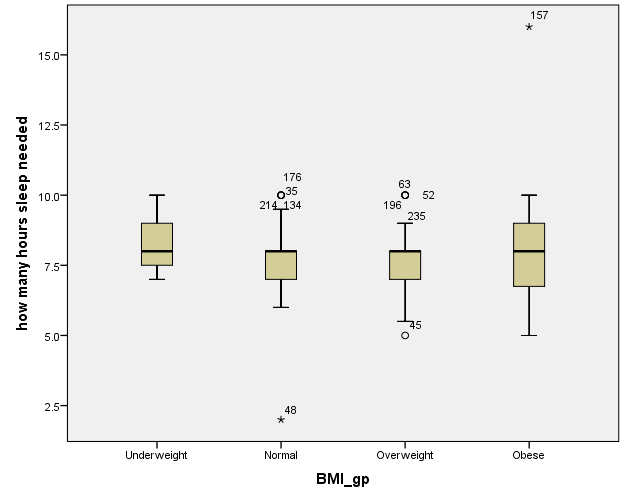
Click OK and study the output. You should be able to see that the normal, overweight and obese groups have very similar means (we will ignore the underweight group because there are only 8 cases). Notice, however, that the standard deviation for the obese group is almost double that of the normal and overweight groups. Notice also that the obese group is positively skewed, whilst the other two groups are skewed slightly in the negative direction. Further exploratory analysis would be necessary to explain why these differences might exist. For now, answer the following descriptive questions:

**Q6: What is standard deviation score for the obese group?**

**Q7: What is the skewness score for the normal BMI group?**

# Exercise 4: Creating a box-plot

Box plots can be very useful for comparing distributions across multiple groups. They are a compact and elegant way of displaying a number of salient features of a distribution including the median average, inter-quartile range, range and outliers. The example below shows a box-plot of ‘hourneed’ (perceived number of hours of sleep needed) between different BMI groups.



Each BMI category has its own box plot. Each plot has a box, coloured beige, which represents the inter-quartile range (25th to 75th percentile). The line in the middle of the box represents the median average. A box has ‘whiskers’ sprouting above and below it. The length of the whiskers represents the values within 1.5 of the inter-quartile range of the top/bottom of box. Finally, outliers are represented as individual points above and below the plot. These are rare values that lie outside of the whiskers and are identified by their case number. Extreme values are shown as asterisks (see case 157 in the figure above).

We can see that whilst their averages are quite similar, the BMI categories vary quite considerably in their distributions of ‘hourneed’ values. The Normal and Overweight groups seem to have a negative skew, with a median high within the box and several outliers high in the range. In contrast, the Underweight group seems relatively normal with no outliers.

You can re-create the above plot by following these steps:

* Select the menu Graphs 🡪 Boxplot
* Choose Simple (Summaries for groups of cases)
* Move hourneed to the ‘Variable’ box
* Move BMI\_gp to the ‘Category Axis’ box
* Click ‘OK’

# Summary

In this tutorial, you learned how to compute summary statistics for both categorical and continuous variables, how to use these techniques to summarise the characteristics of the sample in a report and to explore the distributions of various continuous variables and groups within such variables.

Before you move on to these exercises, make sure you have successfully answered all questions in this tutorial (through the Blackboard Lab Quiz 1-4) and feel generally confident about all aspects of the material covered. If you have any queries, please ask a tutor during your laboratory session. **Don’t forget to** **save your data file** before you close SPSS**.**

# Further Reading

Pallant, J. (2007) SPSS survival manual : a step by step guide to data analysis using SPSS for Windows (Version 15), Chapter 6.

1. Enter whole numbers, not decimals e.g. 32% not 32.1%) [↑](#footnote-ref-1)