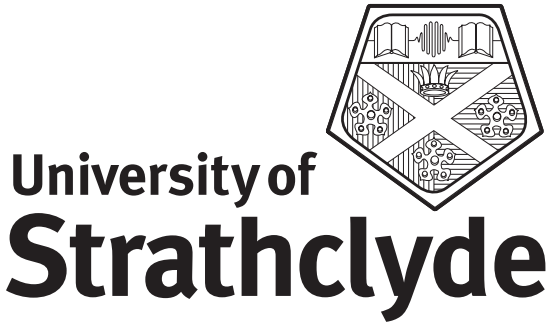


MM104 / MM107
Statistics and Data Presentation Semester 2

Project 2: Descriptive Statistics

Ainsley Miller
ainsley.miller@strath.ac.uk

Lecture Overview

In this lecture we will

- discuss numerical data.
- discuss measures of location and spread.
- discuss skewness and what descriptive statistics to report when your data is skewed symmetric.
- learn how to use Minitab to create histograms and boxplots.
- discuss outliers and how to calculate them.
- learn about confidence intervals and how to interpret the results in the context of the question.

Types of Data

There are two types of data

1. Qualitative data (also called categorical data)
2. Quantitative data (also called numerical data)

Quantitative Data

Quantitative data is also called numerical data.

Common examples of quantitative data are:

- Number of children in a family
- Height
- Weight
- Blood pressure
- Number of school pupils in each local authority

Quantitative Data

Quantitative data can be split into two sub-categories

- Discrete numerical data
- Continuous numerical data

Discrete numerical data is counting data. This is data which can only take a finite amount of values e.g. number of children in a family, number of cars in a household.

Continuous numerical data is data which has been measured using a some apparatus. Since this data is being measured using a device there are infinite possibilities (dependent on the quality of the device). Examples of this data are weights and height.

Location

The location tells us what to expect from our data.

There are 3 ways to measure location

- Mode
- Mean
- Median

Spread

The spread (variability) tells how much sample values depart from expectations and about the diversity of values within our sample.

There are many ways to measure variability

- Range
- Upper and Lower Quartiles
- Interquartile Range (IQR)
- Standard Deviation
- Variance

Skewness Statistic

The skewness statistic measures how skewed a data set is.

- If $-0.5 \leq$ Skewness Statistic ≤ 0.5 the data is roughly symmetric.
- If the Skewness Statistic > 0.5 the data is right tail skewed.
- If the Skewness Statistic < -0.5 the data is left tail skewed.

Which Descriptive Statistics to use ?

If the data is symmetric we quote the mean and standard deviation.
If the data is not symmetric we quote the median and the IQR.

Graphical Summaries - Histogram

- A diagram that uses rectangles to represent frequency.
- Similar to a bar chart, but a histogram groups numbers into ranges.
- Used when trying to graphically display the distribution of numeric data.

How to Create a Histogram in Minitab

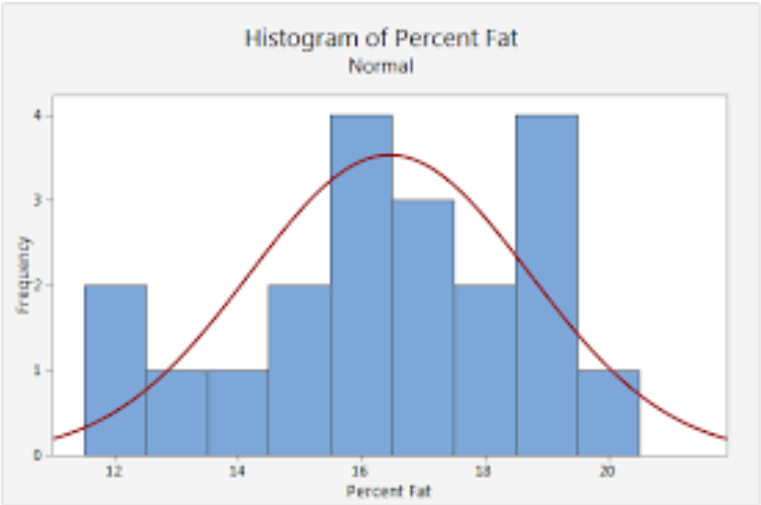
Graph > Histogram

But which one to choose ? Your job as statisticians is to make this choice.

- Simple - Plot a single column of numeric data.
- Simple with Fit: Plot a single column of numeric data and fits a curve of best fit to the data.
- Single Y Variable With Groups: Plot a single column of numeric data, with groups that are defined by the values in a column of categorical data.

Examples of Histograms

Simple Histogram with fit:



Is there a difference ?

For this week's project we are working with Paired Treatment data. Let's consider an example.

Suppose we are carrying out a study for a weight-loss drug. At the beginning of the study we weigh all the participants. The participants then take the weight-loss drug for a number of weeks/months and we then weigh them again at the end of the study. We are of course interested to find out if the participants have lost any weight.

Calculating the Difference

To calculate the difference we would simply do Start Weight - End Weight, for each participant.

But how do we do this in Minitab ?

> Calc > Calculator.
> Then select where you want to store this variable and then in the Expression box type in the equation above using the appropriate columns of your data set.

Visualising the Difference

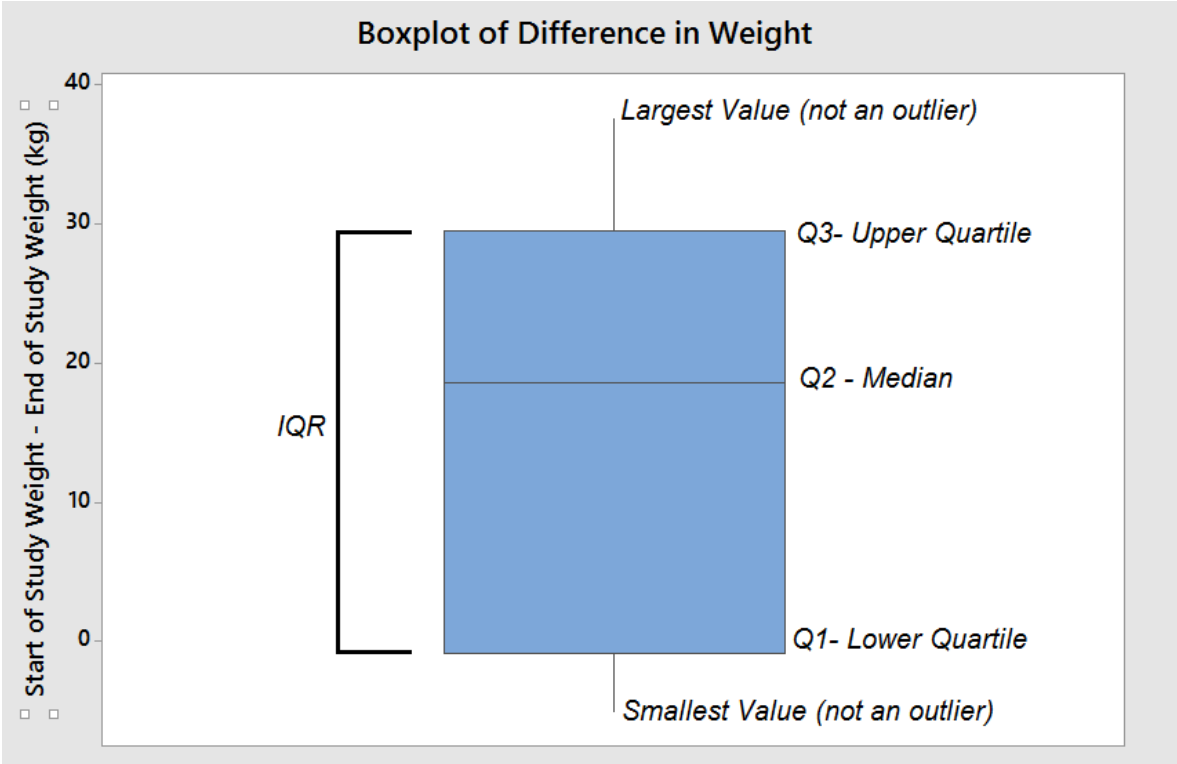
We can visualise this difference in weights using a **box plot**. A box plot shows a visual representation of the median and quartiles of a set of data. A box plot also highlights any outliers in the data set. Outliers are illustrated with stars (*).

But how do we do this in Minitab ?

Graph > Boxplot > Simple

Example of a Boxplot

Simple Boxplot. There are no stars (*) in this boxplot, therefore there are no outliers.



Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

A data point is said to be an outlier if

- it is less than $Q_1 - (1.5 \times IQR)$
- or greater than $Q_3 + (1.5 \times IQR)$

There is a lot of discussion in statistics about whether or not to remove outliers, it depends on the data set.

Confidence Intervals

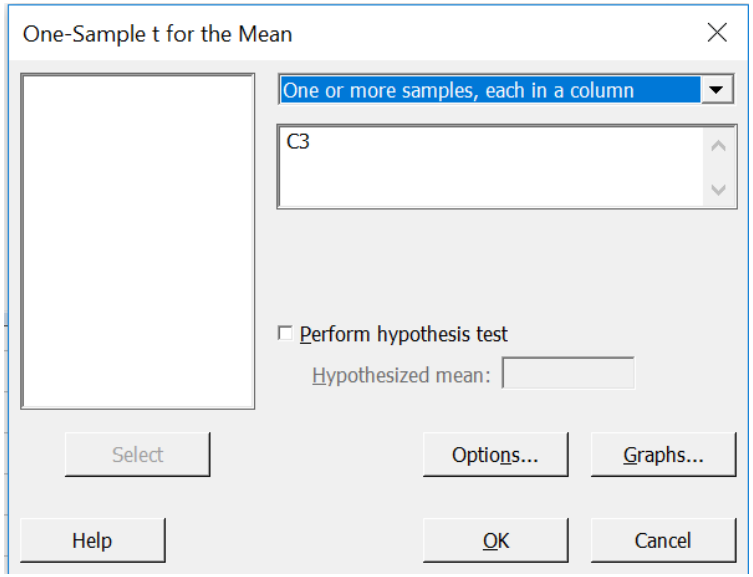
A confidence interval (CI) is the best estimate of the range of a population value that we can come up with given the sample value.

In Statistics we always use a 95 % confidence interval.

Back to our Example

But how do we do find the confidence interval for the difference in weight loss in Minitab ?

Stat > Basic Statistics > 1 Sample t. Here C3 is the column containing the differences.



Interpreting The Result

One-Sample T: C3

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
10	16.82	15.46	4.89	(5.76, 27.89)

μ: mean of C3

Interpretation:
The treatment is associated with a mean weightloss of 16.82 kg. The 95 % CI is 5.76 to 27.89 kg and this interval contains the true mean increase with 95 % confidence.