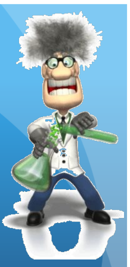
What is statistics?

- Even though you may not have realized it, you probably have made some statistical statements in your everyday conversation or thinking. Statements like "I sleep for about eight hours per night on average" and "You are more likely to pass the exam if you start preparing earlier" are actually statistical in nature.
- Statistics is a discipline which is concerned with:
- designing experiments and other data collection,
- summarizing information to aid understanding,
- drawing conclusions from data, and
- estimating the present or predicting the future.

Why start with statistics?

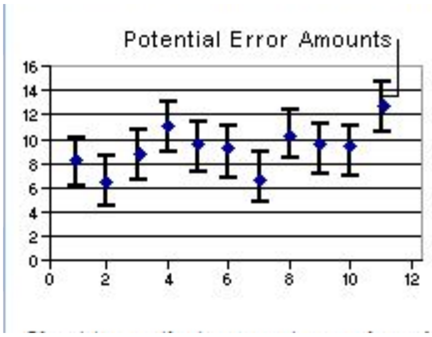


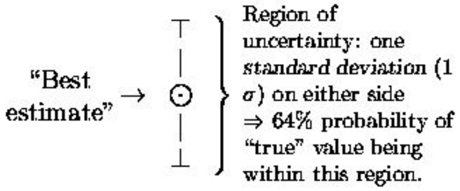
- You will need to design your own experiments and collect your own data
- You need to know whether your data is ‘statistically significant’ and also whether is SPECIFIC, ACCURATE, RELIABLE and VALID.
-

6.1.1 Outline that error bars are a graphical representation of the variability of data.

- The knowledge that any individual measurement you make in a lab will lack perfect precision often leads a researcher to choose to take multiple measurements at some independent variable level.
- Though no one of these measurements are likely to be more precise than any other, this group of values, it is hoped, will *cluster* about the true value you are trying to measure.
- This distribution of data values is often represented by showing a single data point, representing the *mean* value of the data, and *error bars* to represent the overall *distribution* of the data.

- The mean, or average, of a group of values describes a middle point, or central tendency, about which data points vary.
- The mean is a way of summarizing a group of data and stating a best guess at what the true value of the dependent variable value is for that independent variable level.





“Best estimate” →

Region of uncertainty: one standard deviation (1 σ) on either side ⇒ 64% probability of “true” value being within this region.

Standard Deviation (SD) is the measure of spread of the numbers in a set of data from its mean value.

- The error bars shown in a line graph represent a description of how confident you are that the mean represents the true value.
- The more the original data values range above and below the mean, the wider the error bars and less confident you are in a particular value.

Adding error bars in excell

- On the **Format** menu, click **Selected Data Series**.
- On the **X Error Bars** tab or the **Y Error Bars** tab, select the options you want.
- http://cal3.fmc.flinders.edu.au/BTECskills/excel/Errorbars.htm

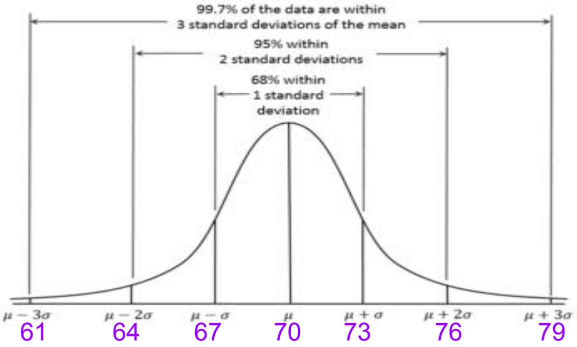
6.1.2 Calculate the mean and standard deviation of a set of values.

- Students should specify the sample standard deviation not the population.
- Students are not expected to know the formulas. They will be expected to use the statistics function of a scientific calculator.

- **Standard deviation** is a simple measure of the variability or dispersion of a data set.
- A low standard deviation indicates that the data points tend to be very close to the same value (the mean),
- while high standard deviation indicates that the data are “spread out” over a large range of values.

- For example, the average height for adult men in the United States is about 70 inches, with a standard deviation of around 3 inches. This means that most men (about 68%, assuming a normal distribution) have a height within 3 inches of the mean (67 inches – 73 inches),
- while almost all men (about 95%) have a height within 6 inches of the mean (64 inches – 76 inches).

- If the standard deviation were zero, then all men would be exactly 70 inches high.
- If the standard deviation were 20 inches, then men would have much more variable heights, with a typical range of about 50 to 90 inches.



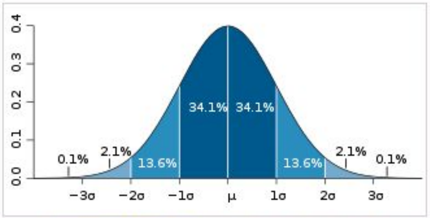
μ (mean) = 70 inches

σ (SD) = 3 inches

Calculate

- x = one value in your set of data
  avg (x) = the mean (average) of all values x in your set of data
  n = the number of values x in your set of data
- For each value x, subtract the overall avg (x) from x, then multiply that result by itself (otherwise known as determining the square of that value).
- Sum up all those squared values. Then divide **that** result by (n-1).
- find the square root of that last number. **That's** the standard deviation of your set of data.

6.1.3 State that the statistic standard deviation is used to summarize the spread of values around the mean, and that within a normal distribution approximately 68% and 95% of the values fall within plus or minus one or two standard deviations respectively.



A plot of a normal distribution (or bell curve). Each colored band has a width of one standard deviation.

6.1.4 Explain how the standard deviation is useful for comparing the means and the spread of data between two or more samples.

- A small standard deviation indicates that the data is clustered closely around the mean value.
- Conversely, a large standard deviation indicates a wider spread around the mean.

6.1.5 Outline the meaning of coefficient of variation.

- The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a dimensionless number.
- Coefficient of variation is the standard deviation expressed as a percentage.

- When the SD and mean come from repeated measurements of a single subject, the resulting coefficient of variation is an important measure of reliability.
- This form of within-subject variation is particularly valuable for sport scientists interested in the variability an individual athlete's performance from competition to competition or from field test to field test. The coefficient of variation of an individual athlete's performance is typically a few percent.

- For example, if the coefficient of variation for a runner performing a 10,000-m time trial is 2.0%, a runner who does the test in 30 minutes has a typical variation from test to test of 0.6 minutes.

6.1.6 Deduce the significance of the difference between two sets of data using calculated values for *t* and the appropriate tables.

- The t-test can be used to measure whether there is a significant difference between the means of two populations.
- For example if you measure the weight of the inhabitants on two islands the t-test formula will work out whether there is a significant difference based on the difference between the means and the degree of variation among them.

- A table of critical t values is used to determine the probability that the difference is simply random chance.
- See example of t test between a sample and a population mean.

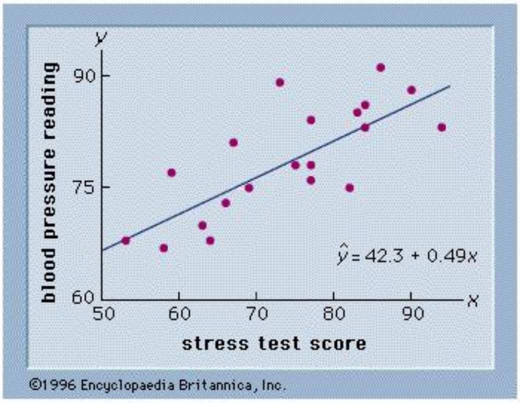
- Independent t test: The most frequently used t test determines whether two sample means differ reliably from each other.
  - Do two groups training at different levels of intensity differ from each other on a measure of cardiorespiratory endurance?
- Dependent t test: The two group of scores are related in some way.
  - Two groups of subjects are matched on one or more characteristics OR
  - One group of subjects is tested twice on the same variable.

- Two tailed t test: it is assumed that the difference between the means could favour either mean.
- One tailed test: can do only one direction.

6.1.7 Explain that the existence of a correlation does not establish that there is a causal relationship between two variables.

- **Correlation** (co-relation) is a term which is used to define the extent of relatedness or relationship between two variables.
- Are the two variables related in such a way that random chance cannot account for the relationship?
- It should be noted that just because you can mathematically determine how related two variables are one cannot use correlation to validate a cause and effect relationship between the two variables.

- Therefore correlation is not sufficient for validity of the relationship. This concept is loosely phrased “**correlation does not imply causation**.”
- They can indicate only how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables.



©1996 Encyclopaedia Britannica, Inc.

- **Pearson product-moment correlation coefficient (r)** is the correlation between two variables (X and Y)
- This calculation provides a measure of the linear relationship between the two variables. Does X and Y increase or decrease together or is it relationship due to random chance.
- The correlation coefficient will be a value between +1.000 and -1.000.
- The closer the number is to 0 the less likely there is a linear relationship with a value of 0 is there is no linear relationship between X and Y.

- The closer the number is to 1 the more likely there is a linear relationship between X and Y with a value of 1.0 indicating a perfect linear relationship. The sign (+ or -) indicates the direction of the relationship. A plus (+) sign tells you that as X increase so does Y. A minus (-) sign tells you as X increases, Y decreases or as X decrease, Y increases.

- **Coefficient of determination (r^2)** is the proportion of the variance of one variable which is predictable from the other variable. In other words this helps determine (in percentage) how much the variation of Y is based on the variation of X. Is the variation in Y related to the linear relationship between X and Y.

- Example
- “If your r = 0.922, then r2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between X and Y (as described by the regression equation).
- The other 15% of the total amount of variation in Y remains unexplained.”