MAKING THE DAYS COUNT

iAspire Student Support

An Introduction to Statistics

VOLUME 1, ISSUE 9

By Dr Donnalee Taylor

SPECIAL POINTS OF INTEREST:

- Experimental Research
- Hypothesis
- Sampling Methods
- Statistical Methods
- Analysing Differences
- Definitions

Key Words

How to, statistics, hypothesis, data, analyzing, definitions

INSIDE THIS ISSUE:

Dr Donnalee Taylor

College of Public Health, Medical and Veterinary Science

James Cook University
Townsville, QLD
Australia

Welcome

'An Introduction to Statistics' was created to be an easy guide to statistics. The condensed information format provides an accessible reference to get you comfortable, confident and excited about statistics.

This guide has been set up in a nonlinear way so you can start anywhere and use it as a reference tool. I have intentionally used simple wording to reduce the use of advanced statistical jargon and have tried to provide plausible examples wherever possible along with some comic relief.

Enjoy your statistical journey. Have fun and don't let any mean numbers pick on you ;o)

Lets Get Started

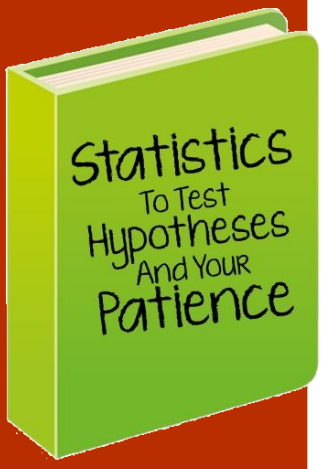
Statistics is the study of a collection of data, the organization of data, the analysis of data and the interpretation and presentation of data. Statistics deals with all aspects of data including: planning data collection (hypothesis), designing surveys or experiments, analysing your findings and presenting your findings. Statistics is a set of tools for the organization and analysis of data. All research experiments start with a Hypothesis.

A hypothesis is the expression of a testable theory. A typical hypothesis states a theory such as: 'if I do this, then this will happen' (i.e. 'If a boy is raised by wolves, he will exhibit wolf-like behaviour.' Or 'If a first year university student attends all of their lectures and labs, they will be able to achieve high grades.') The purpose of a hypothesis is to help create testable parameters through which information can be gained.

A hypothesis states a specific informational goal to be explored. While a hypothesis is usually used for studying the sciences it can be applied to other types of research and assessment projects.

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Writing a Hypothesis



First comes a question you can research then you make a hypothesis. Read about the subject broadly in published manuscripts (primary resources). Creating a hypothesis will require some background research into the area in which you wish to research. Learn about what type of research has been conducted in your areas of interest, what has been successful and what has not been successful? What are the models used for the research? What are the potential models that can be used for this research? Where will the research be conducted? Find the holes in existing published data by paying close attention to statements like 'it is unknown …' or 'future studies …'. The better your background knowledge the stronger and more accurate your hypothesis will be. A broad background into your research area will provide you with information to support or help prove your hypothesis.

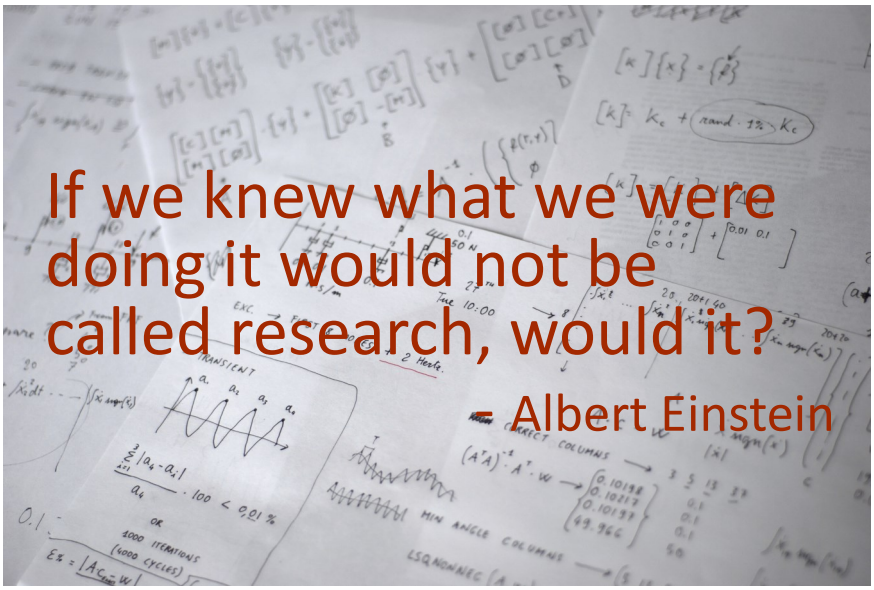
A hypothesis is written before you begin an experiment and not after. Be sure never to alter your data if the hypothesis does not match. Whether your hypothesis is proven right or wrong, it is a significant finding. The results may not be what you thought you would find but that's when research gets interesting and exciting and you have the opportunity to learn and discover some amazing things.

Keep in mind the research you have already read and where the holes are in the published information to help you format a well-constructed hypothesis. To write a solid hypothesis you will need to understand what the variables are for the project. The factors in scientific experiments in search of cause and effect relationships often are referred to as descriptive independent and dependent variables.

"Research is formalized curiosity. It is poking and prying with purpose."

- Zora N Hurston

(Anthropologist 1891-1960)



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Research Variables

INDEPENDENT VARIABLES

An independent variable is the variable that is changed by a researcher or occurs naturally in the environment. As the researcher changes the independent variable they observe what happens to the dependent variable. Usually there is one independent variable in an experiment.

DEPENDENT VARIABLES

The dependent variable responds to the changes made to the independent variable. There are often more than one dependent variable in an experiment or a set of observations. In an experiment a dependent variable's score depends on or is determined by another variable.

CONTROLLED VARIABLES

Controlled variables are quantities that a researcher wants to remain constant and may need to be observed as carefully as the dependent variables. Controlled variables are also referred to as '***constant factors***'.

Example: Opening a water faucet (independent variable – what I change), measure the amount of flowing water (dependent variable – what I observe), the faucet and the water pressure (controlled variables or constant factors – what I keep the same).

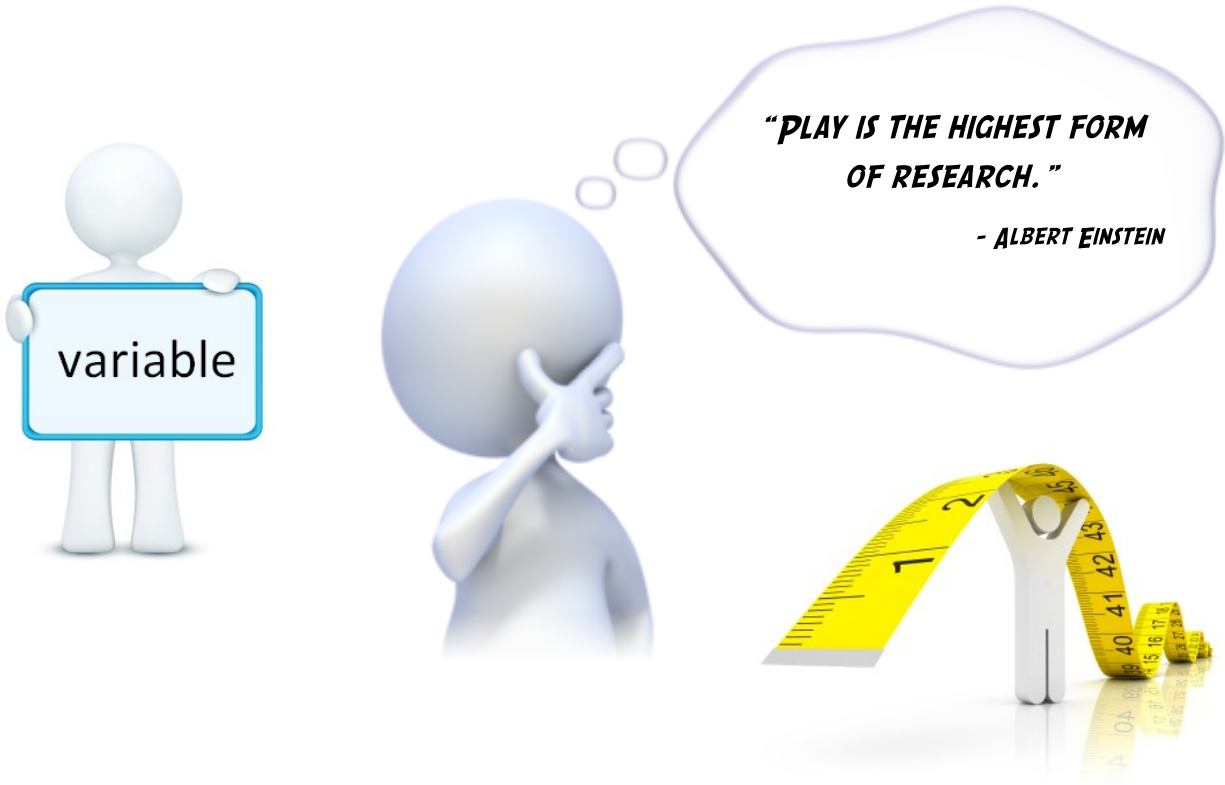
PROBABILITY

The word probability is often used to express a subjective judgment about the likelihood a particular event will or will not occur. For example 'It will probably rain tomorrow'. Sometimes a number between 0 and 1 is assigned to predictions to represent the *degree of confidence* the event will occur. For example, 'the likelihood it will rain tomorrow is 90% (0.9)'. In this example 100% (1.00) represents certainty and a rating of 0 would indicate complete certainty no rain will fall tomorrow.

The probability or a particular outcome or set of outcomes is referred to as a p-value. Often p-values are written as (capital 'P' or lower case 'p') P=(0.05) or P<(0.05) which indicates a significant result or P>(0.05) which indicates a non-significant result.

POPULATION

It is very difficult to research every individual within a population therefore researchers will study a representative sample or a subset of the population as a whole. Researchers will then generalize their findings about the population sample.



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Sampling Methods

Good research is what can set you apart from other graduates and researchers

Sample size is the most important factor in controlling ***margin of error***. The sample size is in the denominator of the ***standard error***, which means as your sample size increases, the standard error goes down and so will the margin of error. Even though a larger sample size provides more information in your analysis and greater precision, using a large sample size is not always feasible or ethical. Sampling is implemented to reduce case numbers to a manageable size. Bigger isn't always better. Sampling is done by ***random sampling*** or ***matching***.

Random sampling does not mean mostly-random, sort-of random or random-enough. A sample is either random or not. True random selection of each member of the population has an equal chance of being selected, like drawing numbers or names from a hat. Be prepared in random selection that you may get a selection that seems odd, weird or even fixed. You can always use a computer to do the selection for you.

Stratified Sampling involves splitting the population into categories and then taking a random sample from each category. The size of the sample is proportional to the size of the category. For example: A company carrying out a survey on employee satisfaction. A stratified sample might select employees proportionally from each department, and level of management. In a small mixed gender group it may be appropriate to ensure that males and females are proportionally represented.

The method of ***matching*** is used to gain accurate and precise results of a study so the results may be applicable to a larger population. After a population is examined and a sample has been chosen, a researcher may consider variables or extrinsic factors that might affect the study. When researchers are aware of extrinsic variables before they conduct a study they may apply *precision matching* or *frequency distribution* matching methods.

Precision matching there is an experimental group that is matched with a control group. Both groups have the same characteristics.

Frequency distribution tends to be more manageable and efficient than precision matching. Rather than one to one precision matching, frequency distribution allows the comparison of an experimental and control group through relevant variables. In saying this, it is very difficult to find an exact match to anyone. Matching produces valid conclusions but there are obvious difficulties associated with this process and that is why researchers tend to reject matching methods in favour of random sampling.

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Descriptive Statistics - Revealing Patterns

Data is either numeric in its origin or transformed by researchers into numbers. Using statistics serves two purposes; one, description and two, predictions. ***Descriptions*** include the group characteristics (variables); recording data for each variable which then can be used to reveal the distribution of data for each of these variables. ***Predictions*** are based on the concept of generalizing or simplifying data so patterns may be revealed through the analysis of the collected data (Jackson et al. 1994). Data can be generalized to provide predictions of how things may occur similar to a probabilistic approach. A researcher cannot be certain that the same thing will happen in other contexts rather, a researcher can only reasonably expect the same thing to happen.

Everyone uses predictions in their daily life. For example, students work on and submit assessment pieces for a grade. Students make the prediction that once an assessment piece is completed and submitted it will be graded based on all the other times this has occurred in the same way. Statistics performs a similar function providing the precise probabilities which are determined in terms of the percentage of chance that an outcome will occur along with a complete range of error. This type of prediction is the primary goal of inferential statistics.

Descriptive statistics 'describes' data; I know ground breaking stuff. Descriptive statistics commonly include frequency counts, ranges (high and low scores or values), means, modes, median scores, standard deviations and standard errors (see definitions section for more details). It is essential to understand ***variables*** and ***distributions*** within descriptive statistics.

Variables

Research observations are recorded in the form of numerical data (Levin 1991; Runyon 1976). Numbers have a variable nature, meaning that quantities vary according to certain factors. For example: when analysing student assessment grades, the scores will vary based on numerous reasons such as student subject knowledge, writing ability and so -on. In statistics these reasons are referred to as variables.

Variables are divided into three basic categories: Nominal variables, Ordinal variables and Interval variables.

NOMINAL OR NAMED VARIABLES

Nominal variables classify data into labelled categories for counting the frequencies of occurrences within the categories (Runyon 1991). A researcher comparing assessment grades between male and female students would compile data using two categories based on sex ('male' and 'female') which would be a nominal variable. Note the categories are not quantified (maleness or femaleness) but rather the category data is quantified (i.e. 15 males and 15 females).

ORDINAL OR ORDERED VARIABLES

Ordinal variables rank data into terms of degree. A researcher that wants to analyse student assessments that have been given a letter grade (ordinal variable) would rank them (i.e. grade A = 1, Grade B = 2, etc.). Grade 'A' ranks higher than Grade 'B' however, the distance between A and B is not defined.

INTERVAL VARIABLES

Interval variables are score data in which the order of data and the precise numeric distance between data points is known (Runyon 1991). A researcher analysing the actual percentage scores of an assessment (i.e. a score of 95 'A' ranks higher than a score of 85 'B', which ranks higher than a score of 70 'C') will know the order of these data points but will also know the distance between them (i.e. there are 10 points between the first two, 15 points between the second two and 25 points between the first and last points).

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Distributions

"Research is creating new knowledge."
- Neil Armstrong
(astronaut 1930-2012)

A distribution is a graphic representation of data. A line formed by connecting data points is called a ***frequency distribution*** and can take many shapes (Jackson et al. 1994). A bell-shaped curve is important as it characterizes the data distribution as 'normal'. This theoretical 'normal' ideal is a mathematical construct with mathematical properties that are helpful in describing the attributes of data distribution (Levin 1991). It is worth noting that actual data points seldom have a perfectly 'normal' distribution but rather approach a normal curve.

The closer the data resembles a normal curve the more probable the distribution maintains the same mathematical properties as the normal curve. This can be used in describing the characteristics of a frequency distribution with greater certainty. Not all frequency distributions approach a normal curve and may be skewed. A skewed frequency distribution no longer has the normal curve characteristics that apply.

NORMAL AND SKEWED DISTRIBUTION

Normal distribution occurs if the distribution of a population is completely normal. A graphed representation of this type of distribution will look like a bell curve; symmetrical and most of the scores clustered toward the middle (see Figure 1).

Skewed distribution simply means the distribution of a population is not normal. The scores might cluster toward the right or the left side of the curve or there might be two or more clusters of scores making the distribution look like hills (Figure 2).

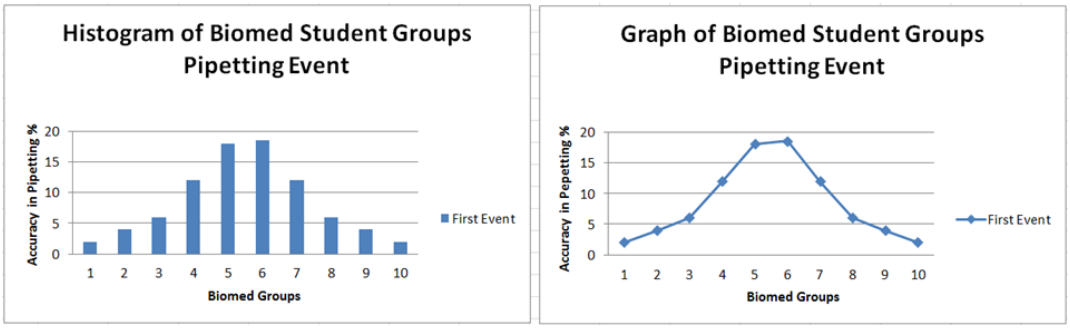
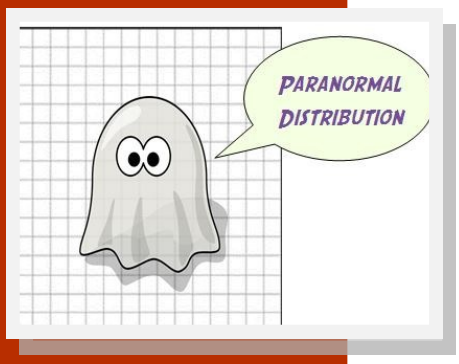


Figure 1: **Diagrams with normal distribution.**



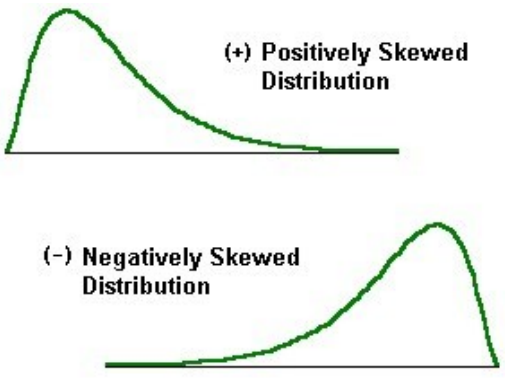


Figure 2: **Positively and Negatively skewed distributions**

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Statistical Methods

Statistics is a methodology that scientists and mathematicians use that concentrates on data analysis. Statistics appears in almost all areas of science, technology, research, marketing, etc. Statistics tools are used whenever data is obtained for the purpose of finding and discerning results (analysing and interpreting results) to explain variations, to draw conclusions from and/or to predict future data. The world of statistics consists of *populations* (individual persons or objects) and *samples* (a set of individuals from the population).

INDIVIDUAL VARIABLES

Note: Statisticians refer to any measured quantity or characteristic as an individual variable. Data collected on a variable will vary from animal to animal or person to person (therefore it is called an independent variable).

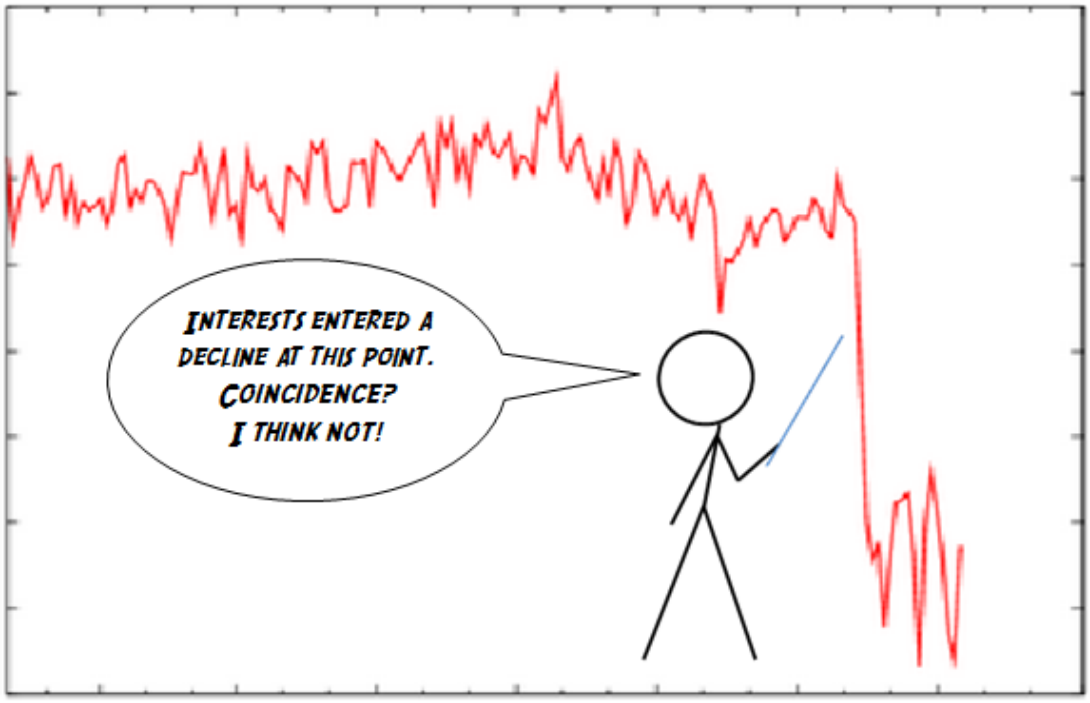
The two major types of variables are:

Qualitative: A qualitative variable (or categorical variable) classifies an individual based on categories. For example: male/female; red hair, blonde, brunette, other etc.

Quantitative variable: A quantitative variable measures or counts a quantifiable characteristic, such as: **height**/weight/number of degrees you have/number of hours you slept last night/etc. A quantitative variable value represents a count (quantity) or measurement with numerical meaning. Which means you can add, subtract, multiply or divide the values of a quantitative variable and the results will make numerical sense. A light-bulb moment: qualitative variables, can take on a numerical value but only as a placeholder rather than a quantity.

Because qualitative- and quantitative-variables represent different types of data, each type has its own set of statistics.

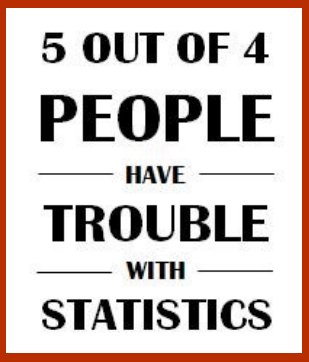
Qualitative variables are mainly summarized by two statistics: ***frequency*** (number in each category) and ***relative frequency*** (the percentage) in each category. As an example 500 people classified by gender, 200 males and 300 females (the total in each gender category is referred to as the ***frequency***). Now using the percentage of data in a specific category (the ***relative frequency***) can be calculated by dividing the frequency by the sample total and multiplying by 100. For example: 200/500*100 = 40% males and 300/500*100 = 60% females. Rather than presenting this information as a *percent* it can be presented as a *sample proportion* which is calculated the same as above without multiplying by 100 (i.e. 200/500 = 0.40). Written as, the sample proportion of males is 0.40 and the sample proportion of females is 0.60.



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Statistics for Qualitative Variables

"He who does not research has nothing to teach."
- Proverb



Statistics for Qualitative Variables may seem limiting but a variety of analysis can be performed using frequencies and relative frequencies to answer a range of possible questions you may be exploring. These are the analysis that uses proportions to ***estimate***, ***compare*** and ***look for relationships between*** the groups of qualitative data.

Estimating a Proportion is done using relative frequencies (see above) to make estimates about a single population proportion. According to the Australian Bureau of Statistics, women comprise half of Australia's total population (50.2% in 2010). However, women comprise less than one-third (30.1%) of all parliamentarians in Australia's parliaments. You are interested in what percentage of Australian females think that women and men should participate equally in the decision-making process of parliament. Using a sample of the Australian female population and not the entire population, these results may vary from sample to sample. The amount of variability would be measured by a ***margin of error*** (the amount that you add and subtract from your sample statistic) will give you an estimated percentage of how female Australians feel about women and men participating equally in the parliament decision-making process. For example: you are estimating the proportion of females in the population and you want to be 95% confident in your results. You sample 1,002 individuals and find that 65% support the issue. The margin of error for this survey turns out to be plus or minus 3 percentage points. This means you can expect the sample proportion of 65% to change by as much as 3 percentage points either way if you took a different sample of 1,002 individuals (i.e. the actual population proportion is somewhere between 65-3=62% and 65+3=68%).

Comparing Proportions is used every day by researchers, media and the average layman. For example: what proportion of biomedical students reading this booklet pass the statistical components of their assignments compared to the non-readers? What percentages of biomedical men watch college rugby league football versus women?

Perhaps you have collected data on a random sample of 1000 JCU biomedical students. First compare the proportion of males to females in the random sample of 1000 biomed students. The proportion of females is 0.52 females and the proportion of males is 0.48.

Looking for relationships between qualitative variables or whether two qualitative variables are related (i.e. is gender related to political affiliation?). To answer this question you will need to put the sample data into a two-way table (Table 1); using the columns and rows to represent the two variables, and analyse the data by using a Chi-square test.

Table 1: An example of data in a table for a Chi-square test.

Gender and Political Affiliation for 56,735 Australian Voters

Gender	Liberal	Labour	Greens	Other
Male	#	#	#	#
Females	#	#	#	#

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Analysing Differences between Groups

t-Tests

A ***t-Test*** is used to determine if the scores of **two groups** differ on a single variable. A t-Test is designed to test for the differences in mean scores. For example: a t-Test can be used to determine whether writing ability differs among students in two classrooms. A t-Test is appropriate when looking at paired data such as, analysing scores of two groups of participants on a particular variable (age group, affiliation) OR for analysing scores of a single group of participants on two variables (products, affiliation, before and after).

Checking the Conditions Step one of ANOVA is checking to be sure all necessary conditions are met before diving into the data analysis. The conditions of using ANOVA are just an extension of the conditions for a t-test (comparing two means). These conditions all need to hold in order for ANOVA to be conducted:

- The populations are independent (i.e. their outcomes don't affect each other).
- The populations each have a normal distribution.
- The variance of the normal distributions are equal.

Matched Pairs t-Test can be used to determine if the scores of the same participants in a study differ under different conditions. For example: used to determine if students write better essays after taking a writing class than they did before taking a writing class.

Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is one of the most commonly used statistical techniques. Analysis of variance is all about examining the amount of variability in a $\mathcal{Y}$ (response) variable and trying to understand where the variability is coming from. For example it can be used to compare several populations in different groups (denoted by an $\mathcal{X}$ variable) such as political affiliation, age group, different brands of product, etc. ANOVA is suitable for experimental data where you apply certain treatments ($\mathcal{X}$) to subjects and measure a response ($\mathcal{Y}$).

The ANOVA is a statistical test which makes a single, overall decision as to whether a significant defence is present among three or more sample means (Levin & James 1991).

An Analysis of Variance test is similar to a t-Test. However, the ANOVA can also test multiple groups to see if they differ on one or more variables and can be used to test between-groups and within-groups differences. There are two types of ANOVA's: One-Way ANOVA and Multiple ANOVA (MANOVA)

One-Way ANOVA is used to test a group or groups to determine if there are differences in a single set of scores. For example: a one-way ANOVA could be used to determine whether freshmen, sophomores, juniors and seniors differed in their writing abilities.

Multiple ANOVA (MANOVA) is used to test a group or groups to determine if there are differences in two or more variables. For example: a MANOVA could be used to determine whether freshmen, sophomores, juniors and seniors differed in writing abilities and whether those differences were reflected by gender. In this example the researcher could determine 1) whether writing ability differed across class levels, 2) whether writing ability differed across gender and 3) whether there was an interaction between class level and gender. Multivariate analysis of variance (MANOVA) is simply an ANOVA with several dependent variables. An ANOVA tests the difference in means between two or more groups or the effects on individual variables, while MANOVA tests for the difference in two or more vectors of means or patterns. Two common multivariate tests are Wilk's λ (most commonly used) and Pillai's Trace (robust to violations of assumptions).

Another example is a study where two different textbooks are used and a researcher is interested in the students' improvements in biomedicine and chemistry (improvements in biomedicine and chemistry are the two dependent variables). The objective in using a MANOVA is to determine if the response variables (student improvements) are altered by the observer's manipulation of the independent variables (text books).

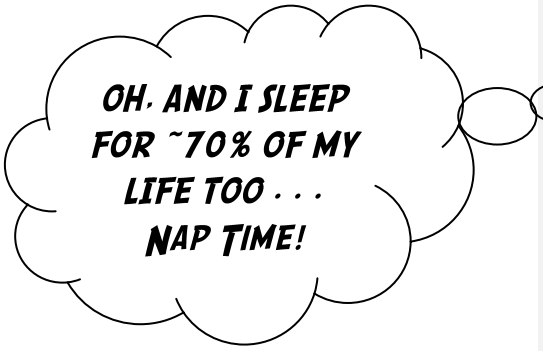
iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Setting up the Hypothesis

"Without data you're just another person with an opinion."
- W. Edwards Deming
Engineer & Statistician

Setting up the Hypothesis to be tested in an ANOVA is step two after checking to be sure all necessary conditions are met before diving into the data analysis. You are testing to see whether or not all of the population means can be deemed equal to each other. *The null hypothesis (H_o)* for ANOVA is that all the population means are equal. The *alternative hypothesis (H_a)* must be the opposite of what is in the null hypothesis. For example: the opposite of having all of the population means deemed equal to each other is a minimum of two of these means that are not equal. If two of these means are not equal it is enough to disprove H_O. Therefore the alternative hypothesis (H_a) is that at least two of the population means are different. The alternative hypothesis in a t-test may be that one mean is greater than, less than, or not equal to the other, in an ANOVA you don't consider the alternative other than [1]. After you reach the conclusion that the null hypothesis (H_o) is rejected in ANOVA, you can proceed to figure out how the means are different, which ones are bigger than others and soon using *multiple comparisons*.

Now you have conducted ANOVA to see whether a group of populations have the same mean, and you have rejected H_o. You conclude that at least two of those populations have different means. You can continue on to find out how many and which means are different by conducting *multiple comparison tests*. The two most common multiple comparison procedures are: Fisher's paired differences (known as Fisher's test or Fisher's LSD) and Tukey's simultaneous confidence intervals (known as Tukey's test).



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Analysing Relationships Among Variables

Statistical relationships between variables rely on the notion of correlation and regression. These two concepts aim to describe the ways in which variables relate to one another:

Correlation tests are used to determine how strongly the scores of two variables are associated or correlated with each other. In statistics a correlation measures the strength and direction of the linear relationship between two quantitative variables that represent counts or measurements. Note it would be incorrect to write 'a correlation exists between eye colour and hair colour'. These variables may be related but they are not quantitative (measurable) variables. In this case you could write 'there may be an association between eye colour and hair colour'. An example of a strong positive correlation would be the correlation between age and job experience (usually, the longer people are alive the more job experience they will have). There is a strong negative correlation between TV viewing and class grades (students who spend more time watching TV tend to have lower grades).

Correlation is a number between -1.0 and +1.0. The positive indicates a perfect positive relationship (i.e. as one increased the other increased in sync). The negative indicates a perfect negative relationship between variables (i.e. if one variable increases the other decreases in sync). A zero indicates that there is no linear relationship at all between the variables. Note: not all correlations will be these exact numbers (-1.0, +1.0, 0) so the closer to -1.0 and +1.0 the stronger the relationship is (Figure 3) and the closer to 0 the weaker the relationship is (in a graph a zero relationship is represented by a horizontal line).

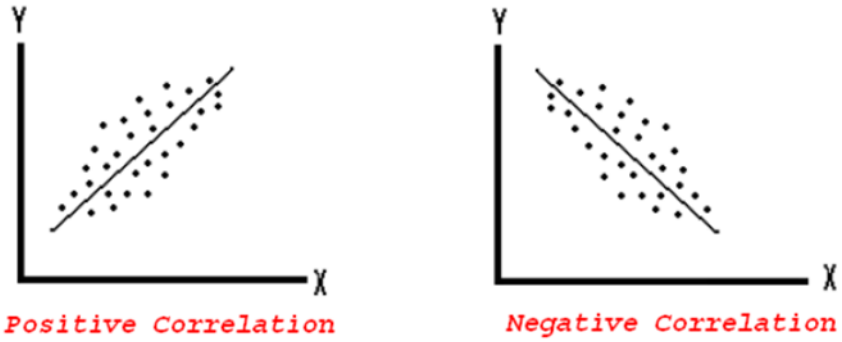
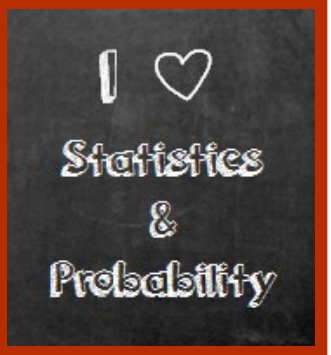


Figure 3: Graph examples of a positive and a negative correlation.

Image Source: http://education-portal.com/cimages/multimages/16/Pos-neg_correlation.PNG

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Ethics in Statistics



Other than animal and human ethics approvals to do research, researchers have ethics in statistics to adhere to. This includes no misrepresentation of data and that the truth is reported. Researchers should steer clear of data manipulation and hiding data to project only what one desires and not what the numbers are actually indicating.

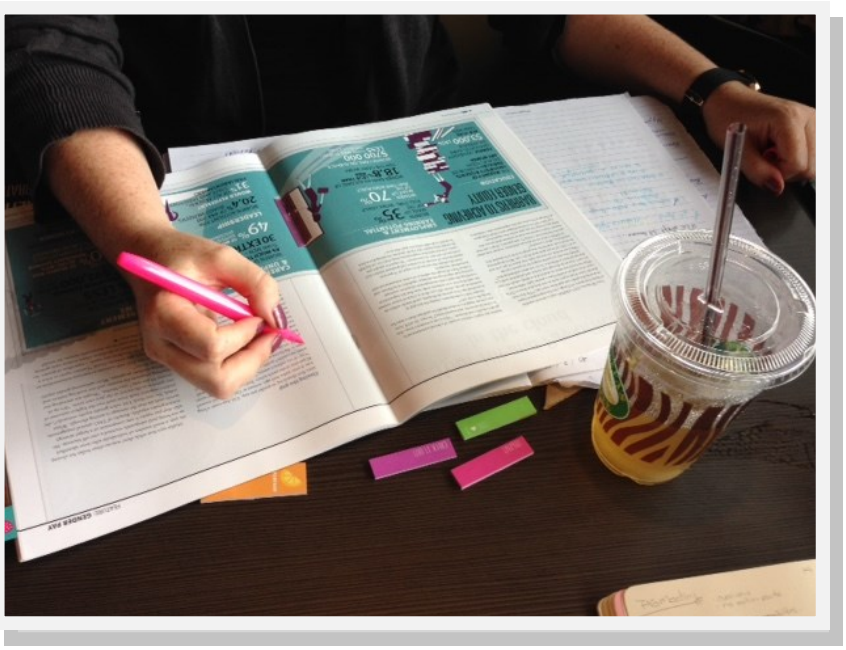
'Lies, damned lies and statistics' is a phrase attributed to the power associated with figures and the representation of these figures. The phrase popularized by Mark Twain is attributed to the 19th century British Prime Minister Benjamin Disraeli (1804-1881). The phrase is commonly used to doubt statistics given to support government positions.

Regrettably it does happen that data is misrepresented which then calls all conclusions based on statistical analysis into refute. Areas to be cautious and diligent in are data collection ensuring it is not biased by posing the wrong questions which stimulate strong emotions (surveys) or the removal of outliers in data and not reporting it.

The old saying 'numbers don't lie' may be true but the interpretation and representation of these numbers can be misleading. An example of this could be a company only publishing the numbers and figures from a customer survey that will reflect well on the company. Without delving too deeply in the philosophy of statistics it is fair to say the interpretation of data is very important as this is what is published and made available to others. This is extremely important in medical research and drug research where a high cost is associated with the misuses of statistics.

Measures against potential misuses of statistics include testing for reliability and validity, testing for statistical significance and critically reading statistics. Most of all researcher vigilance, integrity, veracity and ethical conduct are paramount.

Being a statistician means never having to say you're certain



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Most Popular Graphs in Statistics

Other than understanding your data the other goal of statistics is to present your data in a meaningful way. Listing data on a page or trying to describe it is difficult and even more difficult to understand. The old saying 'a picture is worth a thousand words' is true. Something that may take paragraphs to describe in words may be best represented in a graph. Appropriately chosen graphs convey information quickly and easily and can highlight subtle features of the data. Graphs can show relationships that are otherwise obscure in a list of numbers. Graphs can also provide a convenient way to compare results.

Different types of data (qualitative, quantitative and paired data) require different types of graphs. The following are the most common graphs used in statistics.

Bar Graph contains a bar for each category of a set of qualitative data. The bars are arranged in order of frequency, so that more important categories are emphasized. Vertical bar graphs are the most preferred style in science. Bar graphs allow for the comparison of important data values (Figure 4).

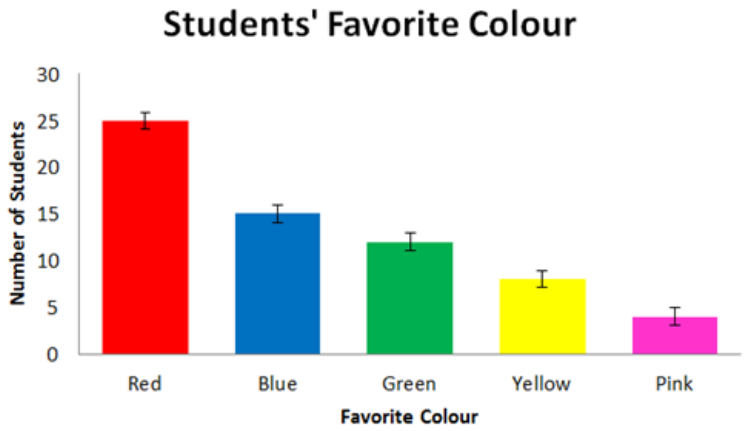


Figure 4: Example of a bar graph.

Pie Chart displays qualitative data in the form of a pie or circle graph. Each slice of pie represents a different category (Figure 5).

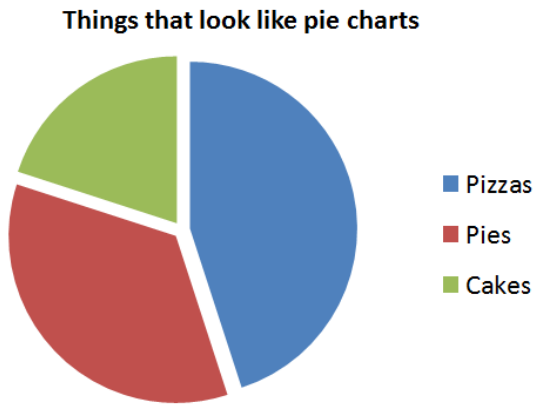


Figure 5: Example of a pie chart.

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Most Popular Graphs in Statistics

"Statistics is the grammar of science."

- Karl Pearson

Histogram is another kind of graph that uses bars to display quantitative data. Ranges of values, called classes, are listed at the bottom, and the classes with greater frequencies have taller bars (Figure 6). Great for showing continuous data (weight, height, how much time, etc.).

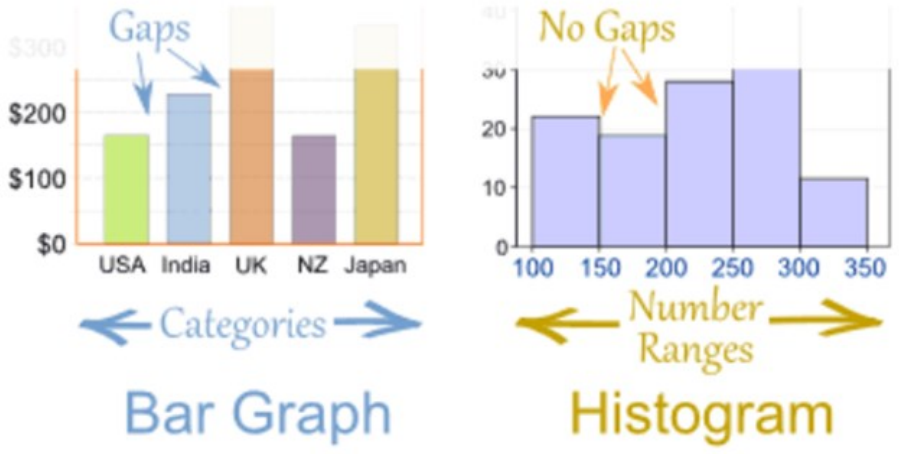
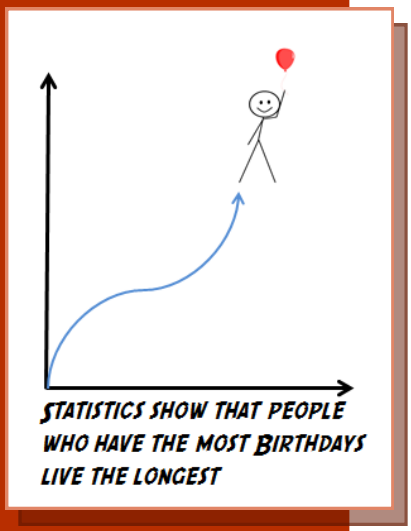


Figure 6: **Example of a Histogram and the difference between a histogram and a bar graph.**
Image Source: http://www.mathsisfun.com/data/histograms.html

Line Graph shows information that is connected in some way, for example magazine sales over time.



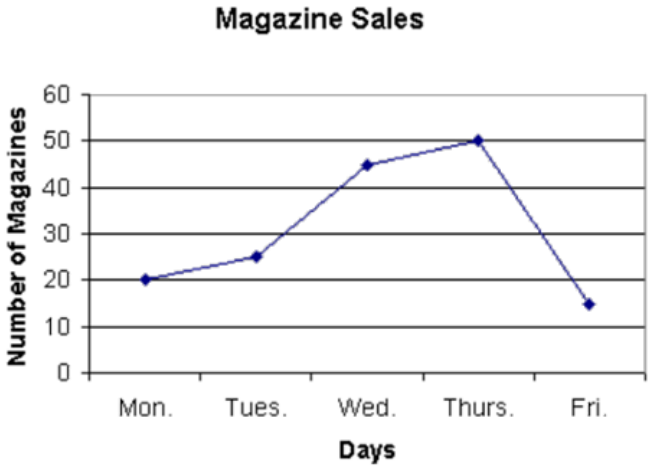


Figure 7: **Example of a line graph**
Image Source: http://www.studyzone.org/testprep/math4/d/linegr4.gif

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Scatterplot displays data that is paired by using a horizontal axis (the x axis) and a vertical axis (the y axis). The statistical tools of correlation and regression are then used to show trends on the scatterplot (Figure 8).

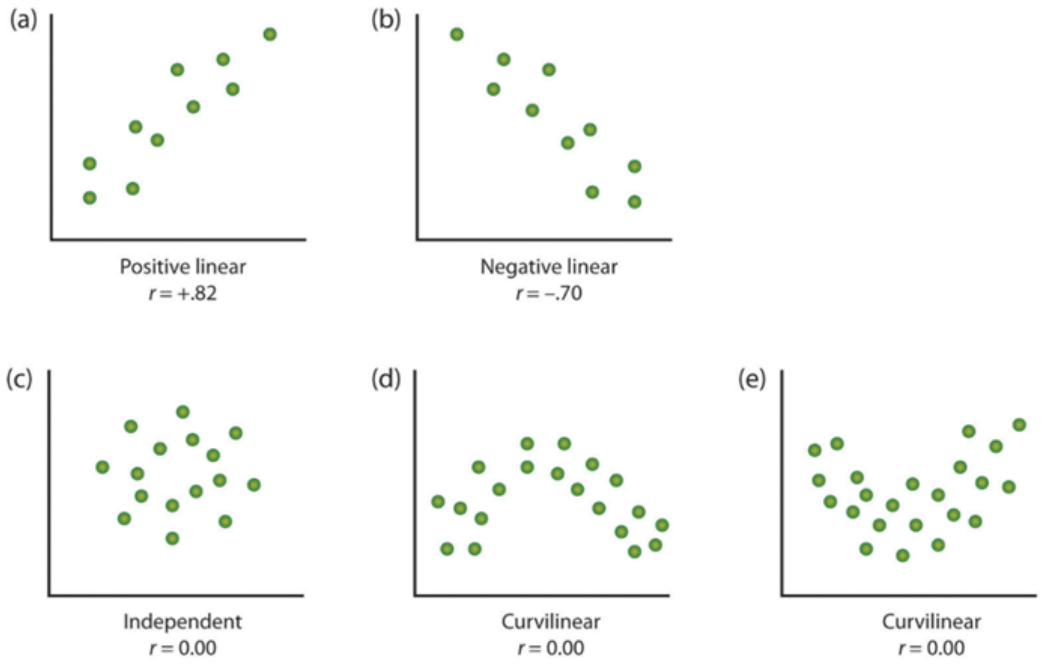


Figure 8: **Examples of scatterplots.**
Image Source: http://images.flatworldknowledge.com/stangor/stangor-fig02_008.jpg

Time-Series Graphs displays data at different points in time used for paired data. The horizontal axis shows the time and the vertical axis is for the data values. Time-series graphs can be used to show trends as time progresses (Figure 9).

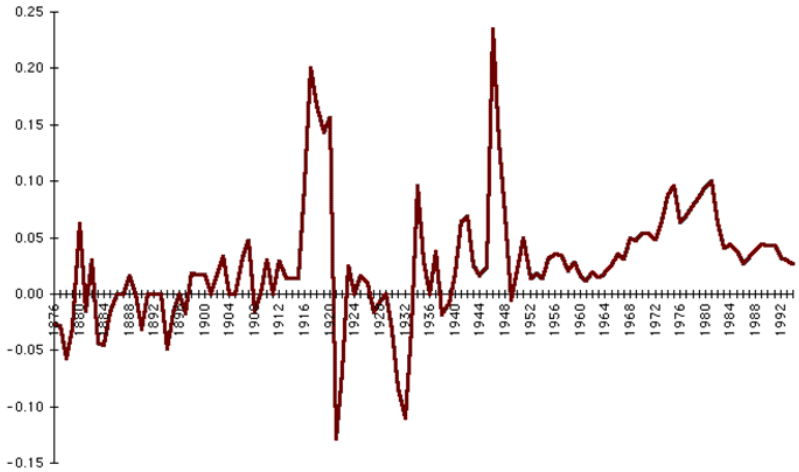


Figure 9: **Example of a time-series graph of inflation rate in the United States between 1876 and 1992.**
Image Source: http://www.uri.edu/artsci/newecn/Classes/Art/INT1/Eco/D_A/Gif/Inf_long.gif

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

"Everything should be made as simple as possible, but not simpler."

- Albert Einstein

Most Important Elements of a Graph

1. A main title for the graph clearly states the relationship between the axes (i.e. Figure 10: 2013 Queensland Regional Bottled Water Sales). Be sure to reference your figure or table in the text of your assignment (Figure/Table #). Note: in scientific writing all figures have a title below the figure and all tables are titled above the table (refer to Figures 1-10 and Table 1).
2. The axes are appropriately scaled and labelled (Figure 10).
3. Independent variable is on the X axis and the dependent variable is on the y axis.
4. The data is plotted accurately.
5. An appropriate key or legend is part of the graph (Figure 10).
6. Appropriate colour and texture is used for the graph so if it is printed in black and white the different bars are distinguishable (Figure 10).
7. Remove background colour in the graph and grid lines (Figure 10).
8. Standard errors bars or Standard deviation bars are present and noted (Figure 10).

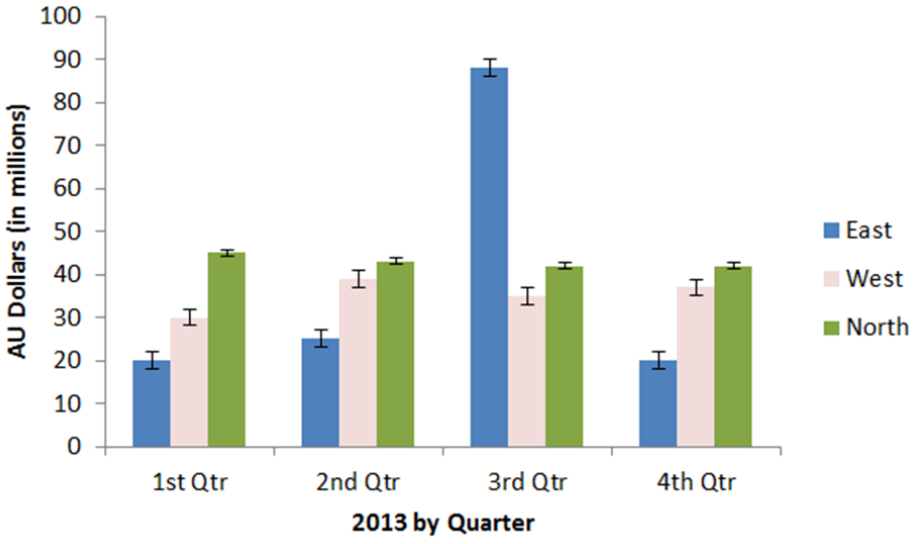


Figure 10: Queensland Regional Bottled Water Sales in 2013 are a guestimate only.

Fun Facts

The word statistics is believed to be derived from the Latin word 'Status' or the Italian word 'Statista'. These words mean 'Political State'.

In Hamlet (1602) Shakespeare used the word Statist.

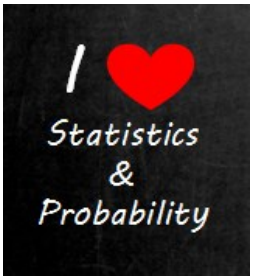
iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

Definitions

Correlation: In a statistical world correlation is only used to discuss the relationship between two quantitative (numerical) variables and NOT two qualitative (categorical) variables. Correlation measures how closely the relationship between two quantitative variables, such as height and weight, follows a straight line and tells you the direction of that line as well (the strength and direction of their linear relationship. If one increases, what does the other do?)

Deviation: Deviation scores are the distance between each data point and the mean.

Mean Average is the sum of the sample values divided by the sample size. Excel Function: =AVERAGE (data_range).



Mean Deviation: Is the average of the absolute values of the deviation cores (i.e. mean deviation is the average distance between the mean and the data points).

Median is the middle value after the sample values have been sorted into order by magnitude. If there are an even number of values in the sample, the average of the two middle values is used. Excel Function: =MEDIAN(data_range).

Mode is the most common value in the sample. Excel Function: =MODE(data_range).

Outlier is a value that "lies outside" (much larger than or smaller than) most of the other values in a data set.

p-Value: is the calculated probability of finding the observed or extreme result. A threshold value for p is called the significance level of the test which is usually 5% or 1% widely used as statistical significance and interpreted as $P \leq 0.05$ and conversely $P > 0.05$ indicates a non-significant result.

Range is the difference between the largest and smallest values in the sample. In other words it is the distance between the lowest data point and the highest data point. Excel Function: =MAX(data_range) –MIN (data_range).

Standard Deviation is the square root of the variance. This calculation is useful because it allows for the same flexibility as variance regarding further calculations and yet also expresses variation in the same units as the original measurements. Is another measure of variation (variance).

Standard Error is the square root of the variance divided by the sample size. Often preferred as a measure of process variation. This method of calculating the standard deviation is known as the Root Mean Square Error (RMSE) method. Excel Function: =STDEV(data_range).

Variance is an estimate of the variation or dispersion of the process from which the sample was drawn. The sample statistic 's2' is an unbiased estimator of the population parameter. Excel Funciton: =VAR(data_range).

Fun Fact

In 1908 the Student's t-Test was introduced by William S Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. Gosset devised the test as an inexpensive means to monitor the quality of stout. Gosset's t-test was submitted along with his other mathematical works under the pseudonym "Student" thus Student's t-Test.

iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016

References

Field A. 2013. Discovering Statistics using IBM SPSS Statistics. 4th ed. SAGE Publications Ltd.

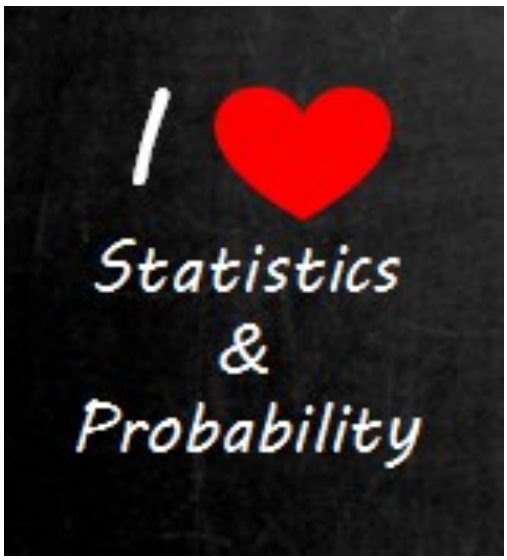
Jackson S. et al. 1994-2012. Statistic: An Introduction. Writing@CSU. Colorado State University. Available at: http://writing.colostate.edu/guides/guide.cfm?guideid=67.

Levine J. & James A.F. 1991. Elementary statistics in social research, 5th ed. New York: Harper Collins.

Quinn G.P. & Keough M.J. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press.

Ramsey D. 2007. Intermediate Statistics for Dummies. Wiley Publishing, Inc. Indianapolis, Indiana

Runyon R.P. & Haber A. 1976. Fundamentals of behavioural statistics, 3rd ed. Reading, MA: Addison-Wesley Publishing Company.



iAspire Student Support v1(9) by Dr Donnalee Taylor is licensed under a Creative Commons Attribution 4.0 International License. 2016