Statistics
Revision Series 2017

Stephen King

May 28, 2017

Statistics is the study of the collection, analysis, interpretation, and presentation of data. Data can be classified into two separate sections; Numerical or Categorical

Numerical Data

Numerical Data results from asking questions which can be answered with numbers.

Numerical Data

Numerical Data results from asking questions which can be answered with numbers, eg;

- What is the average height of people in the class?
- How many people in Limerick play rugby?
- What is the temperature in the sea at Kilkee on New Year's Day?

Numerical Data

Numerical Data results from asking questions which can be answered with numbers, eg;

- What is the average height of people in the class?
- How many people in Limerick play rugby?
- What is the temperature in the sea at Kilkee on New Years Day?

Numerical Data can be organised into two subsections; Discrete and Continuous.

Discrete Data

Numerical data which can only take on specific values are called discrete data. Discrete variables assume values that can be counted. A simple example of discrete data is the number of people in your family. The only possible answers to this particular question are Natural numbers (Positive whole numbers).

Discrete Data

Numerical data which can only take on specific value are called discrete data. A simple example of discrete data is the number of people in your family. The only possible answers here are Natural numbers (Positive whole numbers). Other examples of discrete data are:

- Any question regarding numbers of people, animals or objects which cannot be divided up repeatedly.
- Shoe size, dress size...

Even though we can get a shoe size or dress size of 5.5 etc, there are still only a limited amount of different values available. For example, we cannot get a shoe size of 5.4695 or 5.4696 and so on. This is the difference between discrete and continuous variables.

Continuous Data

Numerical Data which be any one of an infinite set of values *within a given range* is called continuous data. An example of continuous data is the time taken to complete the 100m sprint. The answer to this can be any possible time such as 9.79 seconds or 10.962 seconds or 21.355 seconds etc.

Continuous Data

Numerical Data which be any one of an infinite set of values *within a given range* is called continuous data. An example of continuous data is the time taken to complete the 100m sprint. The answer to this can be any possible time such as 9.79 seconds or 10.962 seconds or 21.355 seconds etc. Other examples of continuous data are:

- height or weight of a person.
- average levels of rainfall/windspeed/% cloudcover
- % of RDA of Vitamin C in a food substance.

Categorical Data

Any question which cannot be answered with numbers provides us with Categorical Data.

Categorical Data

Any question which cannot be answered with numbers provides us with Categorical Data. Examples of Categorical Data include:

- What mode of transport is used to get to work?
- Are you satisfied with your local Politician?
- What is your favourite fruit?

Categorical Data

Any question which cannot be answered with numbers provides us with Categorical Data. Examples of Categorical Data include:

- What mode of transport is used to get to work?
- Are you satisfied with you local Politician?
- What is your favourite fruit?

Categorical Data can be subdivided into two sections; Ordinal and Nominal.

Ordinal Data

Ordinal Data is Categorical Data which can be ordered. The most common example of this is exam grades,A,B,C,D,E,F,NG.

Ordinal Data

Ordinal Data is Categorical Data which can be ordered. The most common example of this is exam grades,A,B,C,D,E,F,NG. Other examples of ordinal data include:

- Social Class, (Upper, Middle, Lower)
- Satisfaction levels with a good or service (Extremely Satisfied, Very Satisfied, Satisfied, Quite Unsatisfied, Extremely Unsatisfied.)
- Karate Belt Ranks (White, Yellow....Black.)

Nominal Data

Nominal Data is Categorical Data which cannot be placed into a certain order or hierarchy.

Nominal Data

Nominal Data is Categorical Data which cannot be placed into a certain order or hierarchy. Examples of this include:

- Modes of transport to work or school.
- Favourite food/colour/sports teams.
- Most visited country on holidays.
- What is your occupation?

Surveys and experiments provide us with data about a population.

The *population* is the group that is being studied.

Surveys and experiments provide us with data about a population.

The *population* is the group that is being studied.

A *census* is a survey of the entire population. Most often it is too time consuming and too expensive to carry out a census, so a sample is selected from the population to be surveyed.

Surveys and experiments provide us with data about a population.

The *population* is the group that is being studied.

A *census* is a survey of the entire population. Most often it is too time consuming and too expensive to carry out a census, so a sample is selected from the population to be surveyed.

A *sample* is a group, selected from the population being studied, in order to gain information about that population. There are various sampling methods which can be used, depending on the type of study.

Sampling Methods

- *Simple Random Sample* - a sample of a certain size in which each member of the population has equal chance of being included in the sample.

Sampling Methods

- *Simple Random Sample* - a sample of a certain size in which each member of the population has equal chance of being included in the sample.
- *Stratified Sample* - a population is divided into a minimum of two subgroups. Simple random samples are selected from each subgroup, then combined to form the stratified sample.

Sampling Methods

- *Simple Random Sample* - a sample of a certain size in which each member of the population has equal chance of being included in the sample.
- *Stratified Sample* - a population is divided into a minimum of two subgroups. Simple random samples are selected from each subgroup, then combined to form the stratified sample.
- *Cluster Sample* - The population is divided into sections called clusters. One or more clusters are then randomly selected, and some or all of the members of the selected cluster are included in the sample.

Sampling Methods

- *Quota Sampling* - The researcher fills a quota of people surveyed from a certain subgroup, (teenage boys, middle aged women etc.)

Sampling Methods

- *Quota Sampling* - The researcher fills a quota of people surveyed from a certain subgroup, (teenage boys, middle aged women etc.)
- *Convenience Sampling* - Sample subjects chosen in most convenient way possible.

Bias

It is important to avoid bias as much as possible when undertaking a survey or experiment. To avoid bias it is important that a sample is representative of the population.

Bias

It is important to avoid bias as much as possible when undertaking a survey or experiment. To avoid bias it is important that a sample is representative of the population.

Surveys which require voluntary responses are often biased, since generally only those with strong opinions make the effort to respond. Other forms of bias result from *undercoverage*, *non-response bias* and *response bias* (leading questions, desirability)

Bias

It is important to avoid bias as much as possible when undertaking a survey or experiment. To avoid bias it is important that a sample is representative of the population.

Surveys which require voluntary responses are often biased, since generally only those with strong opinions make the effort to respond. Other forms of bias result from *undercoverage*, *non-response bias* and *response bias* (leading questions, desirability)

Random sampling can help avoid voluntary response bias and undercoverage bias, but is often innefective against response bias.

Frequency table.

Data, when collected in its original form, is called raw data. Often this data is organised into a frequency distribution. A *frequency distribution* is the organisation of raw data in table form, using classes and frequencies.For example the raw data from a sample size of 20 school children examining the number of children in a family can be represented in a frequency table.

Frequency table.

Data, when collected in its original form, is called raw data. Often this data is organised into a frequency distribution. A *frequency*

distribution is the organisation of raw in table form, using classes and frequencies.For example the raw data from a sample size of 20 school children exaiming the number of children in a family can be represented in a frequency table.

Raw Data 1, 3, 4, 3, 5, 2, 4, 2, 3, 1, 3, 2, 4, 5, 3, 2, 4, 1, 2, 3.

Frequency table.

Data, when collected in its original form, is called raw data. Often this data is organised into a frequency distribution. A *frequency distribution* is the organisation of raw in table form, using classes and frequencies.For example the raw data from a sample size of 20 school children exaiming the number of children in a family can be represented in a frequency table.

Raw Data 1, 3, 4, 3, 5, 2, 4, 2, 3, 1, 3, 2, 4, 5, 3, 2, 4, 1, 2, 3.

Frequency Table

No. of Children	1	2	3	4	5
Frequency	3	5	6	4	2

Categorical Frequency Table

A categorical frequency table is used for data that can be placed in specific categories, usually nominal or ordinal data.

Fav.Team	Man.U.	Man.C	Ars.	Liv.	Chel
Frequency	12	8	6	2	4

Grouped Frequency Table

Grouped frequency tables are used when *numerical* data can be bunched into groups. The following is a grouped frequency table illustrating exam results of a group of students;

Result (%)	0-20	21-40	41-60	61-80	81-100
Frequency	2	5	12	18	8

Cumulative Frequency Table

A cumulative frequency table can be calculated from a grouped frequency table by adding the frequency in each class to the total of the frequencies of the preceding classes. This can be seen in the example below.

Result (%)	0-20	21-40	41-60	61-80	81-100
Frequency	2	5	12	18	8
Result (%)	< 20	< 40	< 60	< 80	< 100
Cum. Freq.	2	7	19	37	45

Stem and Leaf Plots

A *stem and leaf plot* is a data plot that uses part of the data number as the stem, and part of the data value as the leaf.

A *stem and leaf plot* is a data plot that uses part of the data number as the stem, and part of the data value as the leaf.
Example Construct a stem and leaf plot for the following data;

25, 31, 20, 32, 13, 14, 43, 02, 57, 23, 36, 32, 33, 32, 44, 32, 52, 44, 51, 45.

The first step will be to group the data according to the first digit (stem).

02 13, 14 20, 23, 25 31,32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57

Stem and Leaf Plot

0	2						
1	3	4					
2	0	3	5				
3	1	2	2	2	2	3	6
4	3	4	4	5			
5	1	2	7				

The stem and leaf plot has an advantage over frequency tables as it retains the raw data, whilst displaying it in a graphical form.

Back to Back

We also need to be able to create and interpret back to back stem and leaf plots. These are used to compare related distributions.

Back to Back

Example Construct a stem and leaf plot summarising the following data in relation to height increase (in cms) of 12 year old boys and girls over the course of one year.

Boys	Girls
08 21 07 14	09 06 23 25
22 12 16 05	16 21 09 19
06 14 16 09	30 16 10 18

This time the stem goes in the middle and we put one leaf either side.

				Boys		Girls				
5	6	7	8	9	0	6	9	9		
2	4	4	6	6	1	0	6	6	8	9
			1	2	2	1	3	5		
					3	0				

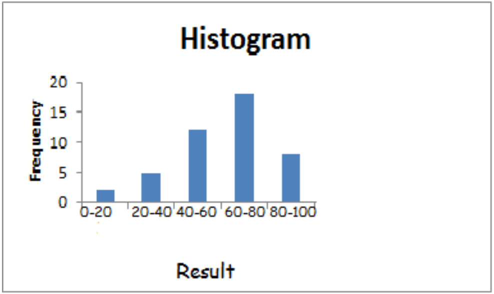
Histograms

A *histogram* is a graphical method of displaying the information of a grouped frequency table. The following grouped frequency table can be represented by a histogram.

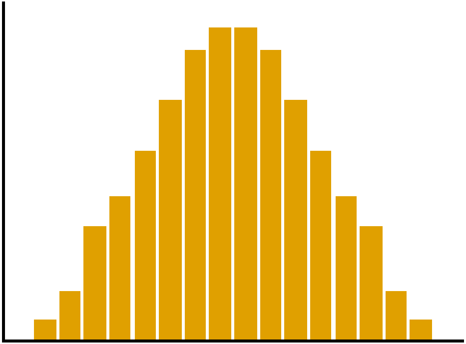
Result (%)	0-20	21-40	41-60	61-80	81-100
Frequency	2	5	12	18	8

Histogram

Result (%)	0-20	21-40	41-60	61-80	81-100
Frequency	2	5	12	18	8

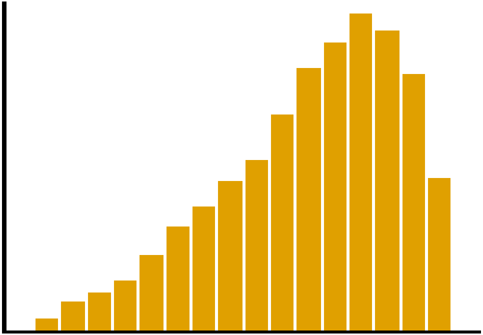


From a histogram, we can see how the data is distributed. This histogram displays data which has a symmetric distribution.



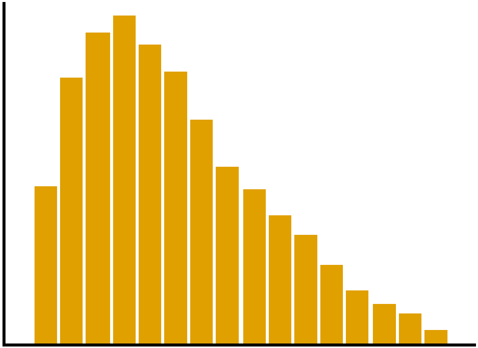
mean = median = mode

In the following histogram we say that the data is left skewed, or negatively skewed.



mean < median < mode

In the following histogram we say that the data is right skewed or positively skewed.



mode < median < mean

Mean

There are various measures of centrality for data, the mean, the mode and the median.

The *mean* is the sum of the data values, divided by the total number of values. The sample mean is represented by $\overline{x}$, and the population mean is represented by μ.

Example: find the mean of the set of values: 84, 12, 27, 15, 40, 18, 33, 33, 14, 4.

$$\overline{x} = \frac{84+12+27+15+40+18+33+33+14+4}{10}$$

$$\overline{x} = \frac{280}{10} = 28$$

Mode

The mode is the value that has the highest frequency (appears most often).

Example: Find the mode of the following values: 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5,

3 is the number which appears most often, hence it is the mode.

Median

The median is the midpoint of the data array. If the data array has an even amount of numbers in it, the median is the midpoint of the two middle numbers.

Example - Find the median of the following data: 684, 764, 656, 702, 856, 1133, 1132, 1303.

Median

The first thing to do is to arrange the numbers in ascending order: 656, 684, 702, 764, 856, 1132, 1133, 1303.

The two middle numbers are 764 and 856, so we must get their midpoint:

$$\text{median} = \frac{764+856}{2}$$
$$\frac{1620}{2} = 810$$

810 is the median.

Frequency tables

We can also get the mean of data displayed in a frequency table.

No. of Children	1	2	3	4	5
Frequency	2	5	5	3	3

We must make sure to multiply each data entry by its frequency before adding them all together. Then we divide by the sum of the frequencies.

$$\overline{x} = \frac{(1)(2)+(2)(5)+(3)(5)+(4)(3)+(5)(3)}{2+5+5+3+3}$$

$$\overline{x} = \frac{54}{18} = 3$$

In this particular sample the average amount of children per family is 3.

Grouped frequency table

We can also get the mean of a data from a grouped frequency table. It is important that we use the *mid-interval values* when calculating the mean.

Find the mean of the data from the following grouped frequency table.

Result (%)	0-20	21-40	41-60	61-80	81-100
Frequency	2	5	12	18	8

Mid-interval Values

Mid-int.	10	30	50	70	90
Result (%)	0-20	20-40	40-60	60-80	80-100
Frequency	2	5	12	18	8

$$\overline{x} = \frac{(10)(2)+(30)(5)+(50)(12)+(70)(18)+(90)(8)}{2+5+12+18+8}$$

$$\overline{x} = \frac{2750}{45} = 61.11\%$$

Measures of centrality tell us a little bit about a list of data. It gives us an insight into what the *average value* of the dataset is. However we usually need to know how spread out the data is, ie. are all of the values bunched closely around the mean, or are they much more spread out. This is where we use our measures of variation;

- Standard deviation.
- Range
- Interquartile range.

Measure of variation can also tell us a lot about the consistency of a variable!

Range

The Range is simply the difference between the highest value and the lowest value.
The range of the dataset 2,3,5,5,8,11,11,14,18,23,28 is:

$$28 - 2 = 26$$

Range $= 26$

Interquartile Range

The interquartile range is the difference between the upper quartile (Q_3) and the lower quartile (Q_1). It is important that the data set is ranked in ascending order.

The lower quartile is the value such that one quarter of the values of the dataset are less than or equal to it.

The upper quartile is the value such that one quarter of the values of the dataset are greater than or equal to it.

Example - Find the upper and lower quartiles of the following data. Hence find the interquartile range.

$$2, 3, 5, 5, 8, 10, 11, 14, 18, 23, 28$$

Example - Find the upper and lower quartiles of the following data. Hence find the interquartile range.

$$2, 3, 5, 5, 8, 10, 11, 14, 18, 23, 28$$

There are eleven values in the dataset. The median is the 6^{th} value which is 10.

Lower Quartile - Since $\frac{1}{4}$ of 11 is 2.75, we will **round up** to three. It is important to note that this only tells us the position of the lower quartile, it is the third entry of the dataset. The lower quartile is **5**.

Upper Quartile - $\frac{3}{4}$ of 11 is 8.25 which we **round up** to 9 .The upper quartile will be the 9^{th} entry of the dataset. The upper quartile is **18**.

Interquartile Range

Therefore the interquartile range is:

$$Q_3 - Q_1 \rightarrow 18 - 5 = 13$$

Stem and leaf Plot

We can also get the median and interquartile range from a stem and leaf plot.

0	2						
1	3	4					
2	0	3	5				
3	1	2	2	2	2	3	6
4	3	4	4	5			
5	2	7					

Stem and Leaf Plot

The above stem and leaf plot has nineteen entries. The lower quartile will be the 5^{th} entry, which is 23, and the upper quartile is the 15^{th} entry, which is 44.
The interquartile range is:

$$44 - 23 = 21$$

The median is the 10^{th} entry, which is 32.

Standard Deviation (σ)

$$\sigma = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n}}$$

Standard deviation measure the average deviation from the mean, how spread out the data are from the mean.

Calculating Standard Deviation

Example - Calculate the standard deviation of the following frequency distribution.

Value	1	2	3	4	5	6
Frequency	7	8	4	4	3	4

A common method when calculating standard deviation is to use a table.

Calculating Standard Deviation

f	x	$\overline{x}$	$x-\overline{x}$	$(x-\overline{x})^2$	$f(x-\overline{x})^2$
7	1	3	2	4	28
8	2	3	1	1	8
4	3	3	0	0	0
4	4	3	1	1	4
3	5	3	2	4	12
4	6	3	3	9	36

We sum up the first column to give us n (Σf).

$$n = 30$$

We then sum up the final column, this gives us $\Sigma(x-\overline{x})^2$.

$$\Sigma(x-\overline{x})^2 = 88$$

Calculating Standard Deviation

We have what we need to find the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{\frac{88}{30}} = 1.71$$

We are allowed to use **stat mode** in our calculators to calculate standard deviation.
Different calculators have slightly different methods. Please make sure you understand how to find standard deviation on your own calculator before leaving tonight!

Centrality

Mean

- Can only be used with numerical data. Uses all the values.
- Affected by outliers.
- Is used in calculating standard deviation

Median

- Finds the middle value (centre) of the data set.
- Used only with numerical data.
- Easy to find.
- Not affected by outliers.

Centrality

Mode

- Can be used for numerical and categorical data.
- Easy to find.
- Not affected by outliers.
- May be more than one mode, or perhaps no mode.

Variation

Range

- Very easy to find.
- Affected by outliers.

Interquartile Range

- Not affected by outliers

Standard Deviation

- Use all the data.
- Affected by outliers.
- Used in Inferential Statistics.

Scatter Plot

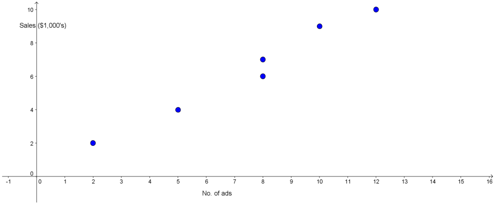
Scatter plots are a visual method of investigating whether or not there is a relationship between two variables. We plot the points from a set of **bivariate** data, (paired data), onto the graph to allow us to tell whether or not a correlation exists.

Example - A manager wishes to find out whether or not there is a relationship between the number of radio ads aired per week and the amount of sales (in thousands of dollars) of a product. The data is displayed below:

No. of ads, x	2	5	8	8	10	12
Sales(1,000's), y	2	4	7	6	9	10

Example

When creating a scatter plot, we put the *independent* variable, otherwise known as the *explanatory* variable, on the x-axis.



We can see from the above scatter plot that there is a strong positive relationship (correlation) between the number of radio ads aired and the amount of sales.

Correlation Coefficient

Scatter plots are a visual way of investigating the correlation between two sets of data. There is also a mathematical method known as the **correlation coefficient**.

- The correlation coefficient measures the strength and direction of a **linear** relationship between two variables.
- The correlation coefficient has a value between -1 and 1.
- The symbol used for sample correlation coefficient is r.
- If r is very close to 1, this indicates a strong positive linear relationship.
- If r is very close to -1, this indicates a strong negative linear relationship.
- If r is closer to 0, this indicates there is no linear relationship between the variables.

Calculating the correlation coefficient

A formula for calculating coefficient is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

This formula is easier to use than it looks. However, we don't need to worry about using formulas to calculate r, we are only required to find it with our calculators, in **stat** mode.
Please make sure you know how to calculate correlation coefficient on your calculators before leaving tonight

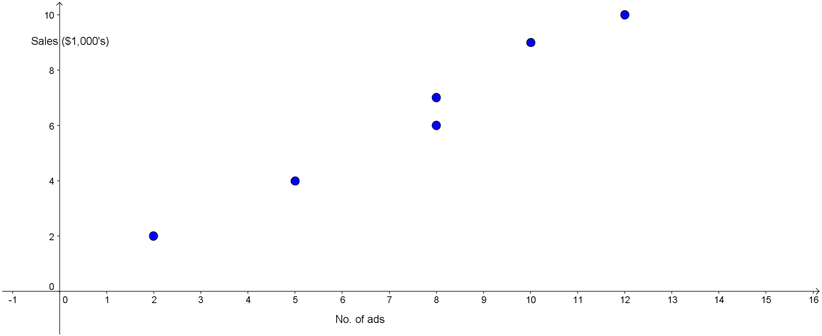
Calculating r

Calculate the correlation coefficient for the following set of paired data.

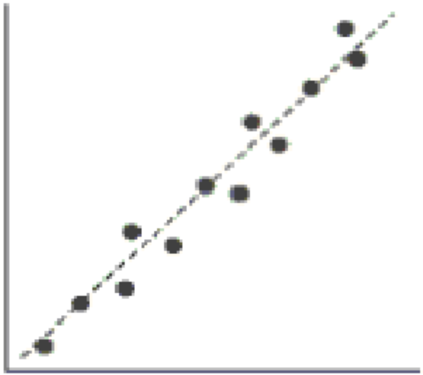
No. of ads, x	2	5	8	8	10	12
Sales(1,000's), y	2	4	7	6	9	10

It is important to remember to clear the memory in your calculator before beginning a new question in stat mode.

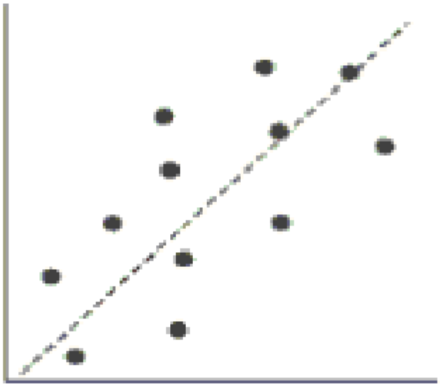
Using Stat mode in the calculator, we get a r value of .978. This suggest a strong positive correlation between the number of radio ads aired and the amount of sales.



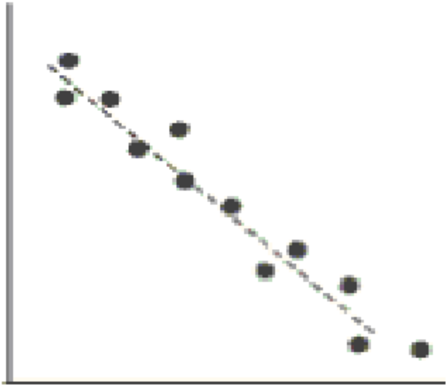
Strong Positive Correlation



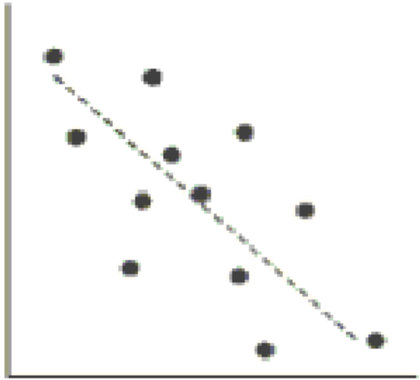
Weak Positive Correlation



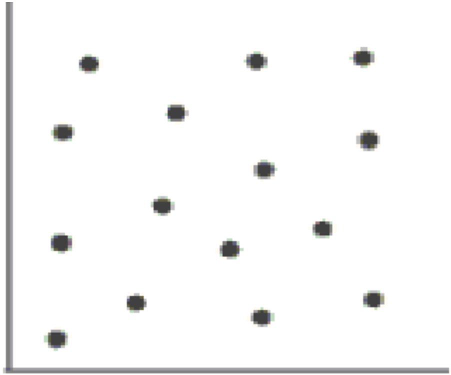
Strong Negative Correlation



Weak Negative Correlation



No Correlation



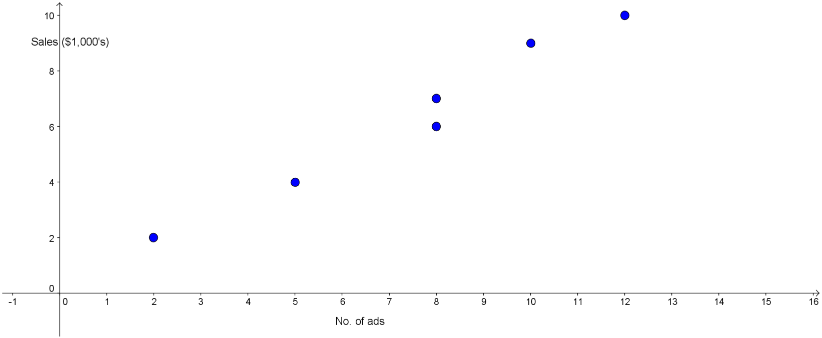
Line of Best Fit

If the scatter plot and correlation coefficient suggest a significant linear relationship, we can find the **line of best fit**, otherwise known as the *regression line*. This line helps us to make predictions about the data.

There is a mathematical method knowns as the *least squares* method for exactly calculating the regression line. On our course, we only need to be able to draw the line of best fit by eye.

Example

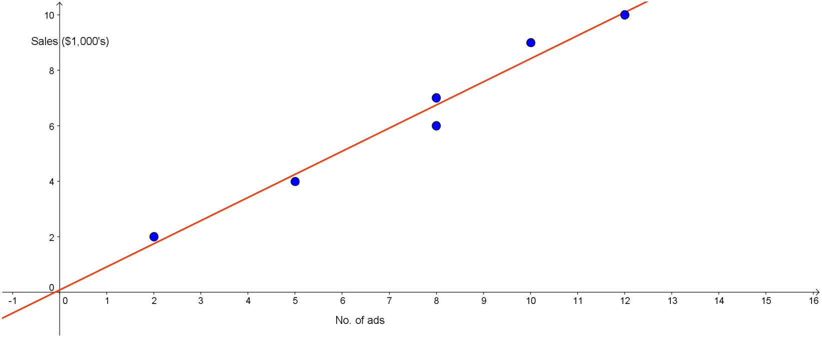
Draw the line of best fit on the scatter plot below.



It is important to try draw the line so it is as close as possible to all of the points.

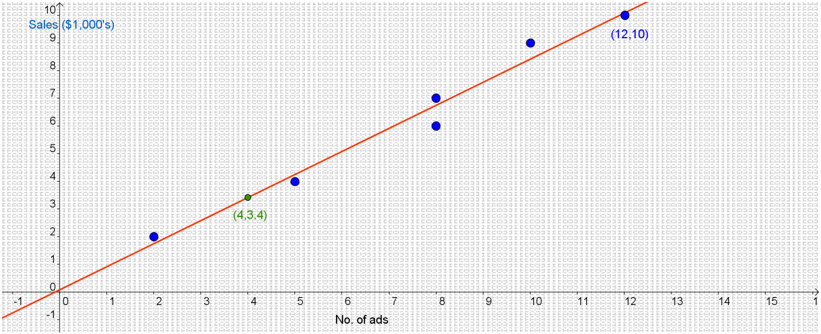
Line of Best Fit

Below is the line of best fit for the data.



We can now visually select two points on the regression line to create the equation of the line of best fit.

Regression Equation



I can see from the diagram above that the point $(12, 10)$ is on the regression line, I have also selected the point $(4, 3.4)$ (in green).

Regression Equation

Two points:

$$(12, 10) \text{ and } (4, 3.4)$$

Slope:

$$m = \frac{3.4-10}{4-12}$$
$$m = \frac{-6.6}{-8} = .825$$

Eqn of line:

$$y - 10 = .825(x - 12)$$
$$y = .825x + .1$$

The equation of the line of best fit is $y = .825x + .1$

$y = .825x + .1$

We can use our regression equation to make predictions.

Example

1. Estimate the amount of sales (in 1,000's) generated when the company buys 10 advertisements.
2. Estimate how many advertisements lead to $15,000$ in sales.

Example

Estimate the amount of sales (in 1,000's) generated when the company buys 10 advertisements.
We use our regression equation,$y = .825x + .1$, where x is the amount of radio adverts, and y is the amount of sales, (in 1,000's).

$$y = .825(10) + .1$$
$$y = 8.35$$

So, we estimate that 10 advertisements generate $8,350$ in sales revenue.

Example

Estimate how many advertisements lead to 15, 000 in sales.

$$15 = .825x + .1$$
$$x = 18.0606$$

The company would need to buy an estimated 18 adverts to yield 15, 000 in sales revenue.

Correlation

Correlation does not imply causality! If we find that there is a correlation between two variables, this does not necessarily mean that a change in one is causing a change in the other. Sometimes there is a third 'hidden' variable, causing the changes.

Confidence Intervals

A confidence interval is a range which contains the value most likely to be the population mean or population proportion.
We use information gained from a sample to make a prediction about the population

Confidence Intervals

The formula for confidence intervals is:

Means and Standard Deviations:

$$\text{Interval: } \bar{x} \pm 1.96(\text{Standard Error})$$

where

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

Proportions:

$$\text{Interval: } \hat{p} \pm 1.96(\text{Standard Error})$$

where

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}}$$

Margin of Error

The margin of error is the maximum likely difference between the sample proportion($\hat{p}$) and the population proportion(p). The margin of error is calculated by:

$$E = 1.96(\text{Standard Error})$$

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}}$$

Where n is the sample size, and for p we will actually use $\hat{p}$, the sample proportion. The greater the sample size, the smaller the margin of error.

Using the margin of error we can calculate confidence intervals (95%) in the form:

$$\hat{p} - E < p < \hat{p} + E$$

Example

Phonetec, a mobile phone manufacturer wants to estimate the proportion(p) of its present stock of phones that have a defect. From a random sample of 500, 40 are found to be have a defect. Calculate the 95% confidence interval for the proportion of all phone with defects.

$$\hat{p} = \frac{40}{500} = .08$$

$$\text{Standard Error} = \sqrt{\frac{.08(1-.08)}{500}} = .01213$$

$$\text{Margin of Error: } E = 1.96 \times .01213 = .0238$$

Therefore the 95% confidence interval is :$\hat{p} - E < p < \hat{p} + E$

$$.08 - .0238 < p < .08 + .0238$$

$$.056 < p < .104$$

Example

A sample of the results of 200 Leaving Cert students has a mean of 375 points. The standard deviation of the sample is 90 points. Find the 95% confidence interval for the true mean in relation to Leaving Cert points.

$$\bar{x} = 375 \ \sigma = 90 \ \ n = 200$$

$$\text{Interval: } 375 \pm 1.96\left(\frac{90}{\sqrt{200}}\right)$$

So the 95% confidence interval is:

$$362.53 \leq \mu \leq 387.47$$

A hypothesis is a conjecture about a property of a population.
Procedure for Hypothesis Testing (95%):

- Identify **null** and **alternate** hypotheses.
- Calculate the confidence interval (95%).
- Make the decision

Decision making: - If the stated hypothetical proportion lies within the confidence interval we **accept** the null hypothesis.
If the stated hypothetical proportion lies outside the confidence interval, then we **reject** the null hypothesis.

The **null** hypotheses (H_0) states that there is no difference between the population proportion and the stated parameter.

The **alternate** hypotheses (H_a) states that there is a difference between population proportion and the stated parameter.

Example:
The CEO of Phonetec claims that only 1% of their phones have a defect. We have already seen that out of a random sample of 500, 40 had a defect. Bases on these findings, can we reject the CEO's claim?

Example

The CEO of Phonetec claims that only 1% of their phones have a defect. We have already seen that out of a random sample of 500, 40 had a defect. Bases on these findings, can we reject the CEO's claim? The null hypothesis is that the proportion of all phones

with a defect is equal to 1%.$(p = .01)$
The alternate hypothesis is that the proportion of all phone with a defect is not equal to 1%. $(p \neq .1)$

We have already calculated the confidence interval for this proportion:

Example

$$H_0 : p = 0.01[\text{claim}]$$
$$H_a : p \neq 0.01$$

Confidence Interval:

$$.056 < p < .104$$

Since .01 does not lie within this confidence interval we reject the null hypothesis. Therefore we are rejecting the claim that the proportion of Phonetec phones with a defect is only 1.%

Sources

Bluman, 2003, Elementary Statistics: A Brief Version, Mcgraw Hill, New York.

Images

- http://www.itjobswatch.co.uk/about.aspx
- http://www.education.com/study-help/article/tweaks-trends-correlation/?page=2