Level 2 Use Statistical Methods to make an Inference

The requirement of this topic is that students make a **Statistical Inference**.

> **–** The process of drawing conclusions about the *population* based on a *sample* taken from the *population*, based on statistical analysis.

It is not sufficient to merely produce some statistics and say something about them. There must be a discussion about what the sample(s) says about the population.

As a means of doing this, students are required to do a **Statistical Investigation**

For the purposes of this year, students are expected to use the *statistical enquiry cycle* (Problem, Plan, Data, Analysis, Conclusion) in their investigation.

The understanding of, and correct use of, terminology is a vital part of this topic. It is vital that students understand all the words used (up to their desired level of achievement). The following words are absolutely key and you must be comfortable with them:

Population – All the members of some group. For example, all people aged 15 and over living in New Zealand on 1 January 2012. (Note the precision with which the population is limited.)

Sometimes a population is either infinite (all possible rolls of a dice) or can only be imagined as complete (all possible essays written by students in Year 9 in NZ).

Population parameter – A number representing a property of a population. Common examples are the population mean, μ and the population standard deviation, σ. Population parameters, although fixed, are usually not known but can be estimated by sample statistics.

Variability – The tendency for a property to have different values for different individuals/groups or to have different values at different times.

Variable – a measureable property of some member of a population, such as height or number of children.

Population distribution – The variation in the values of a variable (measureable result) for every member of the population.

Discrete distribution – values of a variable that can only take on exact number values, usually whole numbers. Examples include number of children or percentage score in a test.

Continuous distribution – values of a variable that can take any value inside a suitable range. That is there is always a potential value between any two other values. Note that for practical purposes we often round continuous distributions. Examples include weight and amount of rain fallen.

Sample – a selection of some members of the population. A **survey** is a measure of one or more variables from a sample (whereas a census is the same process applied to the entire population).

The most common symbol for sample size is *n*.

Sample statistic – A number calculated from a sample, such as a mean or range.

Estimate – a value calculated from a sample, which we can apply to the population, but will always have an error range associated with it.

Precision – How close an estimate is expected to be to the true value of a population parameter (measure). This measure is based on the degree of similarity among estimates of a population parameter, if the same sampling method were repeated over and over again.

Data

Your investigation will use **Data**.

> – A collection of facts, numbers or information; the individual values of which are often the results of an experiment or observations.
>
> (Originally data was the plural of datum, which was a single piece of information. Nowadays we often use “data” to apply to a single item.)

In this unit the data will generally be **Multivariate**.

> – Each item in the population will have measurements for different *variables*.

And will be in the form of a **Data Set**.

> – A table of values from *experiment* or observations. Usually the columns of the table consist of measurements of the same variable for each member, while the rows are for different members or the same member but at different times.

Cleaning data

The process of finding and correcting (or removing) errors in a *data set* in order to improve its quality.

Mistakes in *data* can arise in many ways such as:

- A respondent may interpret a question in a different way from that intended by the writer of the question.
- An experimenter may misread a measuring instrument.
- A data entry person may mistype a value.

Missing Data Points

Often data sets are incomplete.

You must not attempt to even out sample sizes by leaving out data.

Missing data points are not, in themselves, a problem. One of the reasons why statisticians use measures such as mean and median is so that different sample sizes can be directly compared.

However, sometimes missing data introduces, or at least could introduce, a source of *bias*, and should be discussed. For example, if you are testing the effectiveness of a drug, the death of a person taking it cannot be ignored because the results are no longer measureable.

You cannot deal with missing data by setting it a value of zero, even if you suspect the missing data is introducing some bias.

The correct way to deal with missing data is to calculate the statistical measures normally and then to discuss any possible effects in the analysis.

Variation

Your data will not all be the same, or there would be no point studying it. But there are many reasons why different members of a population get different numerical values for some measurement.

Individual variation (or **natural** or **real variation**) occurs because individuals are different.

Occasion-to-occasion variation happens as some variables are not constant even over short time periods. Blood pressure, for example, is always changing and every time it is measured will be different, even for the same person. Even things we think of as fixed, such as a person’s height, is not exactly the same all the time.

Well designed experiments/observations try to reduce this variation by always taking readings at the same time, or in the same situations.

Measurement variation comes because no measuring instrument and no person reading it are perfect.

Well designed experiments/observations try to reduce this as much as possible, by using the most accurate instruments possible and careful measurement techniques. Good experiments also quote the error range in any reading.

Induced variation is that caused by members of the population being in different circumstances. For example, doing exercise will change heart rate compared to periods of rest. Even this variation will depend on the type and length of the exercise.

Experiments or observations intended to measure induced variation will usually try to keep the individual variation as low as possible. For example if measuring the effect of a new method of teaching they might use a test sample of students all approximately the same age and skill.

Another method used commonly for induced variation is the addition of a **control group**. That is a separate group which do not have the different circumstances or procedures. This is virtually always used in medical tests, where a new drug is compared against both an old drug and against a group not taking any drug.

Most often in the real world we are trying to decide on the strength of an induced variation – e.g. does smoking increase the rate of lung cancer? All the other variations merely act to complicate the picture, and they should be discussed as part of your analysis (even to say that they are minor).

Sampling Error

Sampling error arises because every sample is different even if the sample is completely random and unbiased, and so any statistic calculated from a sample will differ from the same statistic calculated from the next sample. As a result, no sample statistic can be assumed to be anything but an *estimate* of the actual population statistic.

The sampling error can be minimised only by taking larger samples. It can never be removed, but can be estimated statistically with *confidence intervals* and the like. That is one can estimate the likely sample error.

Non-sampling Errors

Students need to try to understand the difference between sampling error, which is inevitable as a result of individual and occasion-to-occasion variation, and non-sampling error, which is not.

Non-sampling errors are those that arise from an incorrect selection of samples, for whatever reason. These can often be reduced by good experimental technique, but sometimes the situation does not allow for a proper representative and unbiased sample to be taken.

Our aim with sampling is to get a sample that is, as far as possible, **Representative**.

> – One where all groups of different elements of the population appear in the sample in proportion to their distribution in the population.

It is never entirely possible with a sample to get it truly representative, but there are techniques, discussed on the next page, that improve our chances.

Our sampling needs, most of all, to avoid **Bias**.

> – An influence that leads to results which are different from the true value in a particular direction, e.g. consistently too high or too low, but also too varied or regular.

A **biased sample** is one in which the system used to create the *sample* produces results that are not *representative* of the *population* not just randomly, but reliably in a particular way. Such as error is sometimes called **systematic**.

Any one sample may be turn out to be unrepresentative, as a result of sampling variation. This is not an example of bias if repeated sampling would balance out on average. It is only bias if the method tends to lead to an unrepresentative sample because of some flaw.

Some major examples of bias being introduced are:

- The sampling process lets individuals to select themselves. Individuals with strong opinions or with a stake in the subject will tend to be over-represented, creating bias.
- When the survey questions themselves lead to bias. This can be by way of “leading” questions, poorly worded questions, inappropriate permitted answers, or even if the order questions are asked in tends to set up a specific response.
- Answers given by respondents do not always reflect their true beliefs, for example when people are asked about racial issues they tend to disguise extreme views.
- People who report staying with the regime of a trial but who actually are not, such as failing to follow an exercise programme.
- Drop-outs cannot be measured for results. When people, animals or plants die during a trial it is rarely the ones who were doing best, which tends to bias results towards the high end. On the other hand people can drop out of a drug trial because the problem has gone away, which will tend to bias results towards failure.

Sampling Methods

All sampling must involve **Random Selection**.

– All members of the population have the same probability of being chosen in the sample.

Simple Random Sampling

– The members of a population are numbered off and the selection is made based on a random process in which each member is equally likely to be selected.

An member of the population can be chosen only once: it is **sampling without replacement**.

While simple this process increases the chance of unrepresentative samples by chance, and is not recommended for heavily *stratified* samples. It can also prove difficult to number off the population in any sensible way (e.g. for taking political polls, numbering all voters would be awkward).

Students should not call this merely “random sampling” – the addition of the word “simple” is vital to distinguish this from the other methods. All sampling should be random.

Stratified sampling

– The *population* is split into non-overlapping groups based on some distinct difference. A *simple random sample* is then taken from each group, with the number taken from each group being in proportion to the size of the group in the population.

This technique is recommended when groups in the population have, or are suspected to have, quite different characteristics, especially for small sample sizes.

The name comes from the word *strata*, meaning layer, with a singular of *stratum*. However “stratified” samples need not be actually physically separated into distinct physical groups, e.g. it is very common to stratify by gender and race when trying to get a representative sample of a country.

Systematic sampling

– A method of sampling from a list of the population so that the sample is made up of every k^{th} member on the list, after randomly selecting a starting point from 1 to k.

For example, if selecting students from a school, a stratified sample might take every 20th student after arranging them all in order of age.

This method is recommended when the population is separated into groups with distinct characteristics which can be arranged in such a way that systematic sampling effectively acts as a means of stratifying.

This often increases the representativeness of the sample, and is can be easier to operate than simple random sampling (depending on how hard it is to sort the population).

Cluster sampling

– The population is split into naturally forming groups (clusters). All of the clusters, or a simple random sample of clusters is selected. Either the individuals in these clusters form the sample or simple random samples chosen from each selected cluster form the sample.

For example, the classes of a school for period 1 on a Monday are selected as clusters, and one student from each is randomly selected. That would ensure a spread of ages and academic levels, but would tend to bias the sample towards older students (who tend to be in smaller classes).

Cluster sampling has little to recommend it statistically, but is often used nonetheless in the real world, generally due to issues of time and/or money.

Features of Data

While your answers must focus on the statistical measures of your sample(s) you can also discuss the non-measureable features of the data separately.

This should be done with care, as students are prone to considerably under-estimating the effects of natural variation. Just by sheer chance odd things are going to happen from time to time.

Symmetry and Skew

> – If the smaller values of a distribution tend to be further from the centre of the distribution than the larger values, the distribution is said to have negative skew or be skewed to the left.

Small samples tend to lack good symmetry thanks to natural variation, even when the population is itself symmetrical. Students should therefore not read too much into a lack of symmetry.

However sometimes a lack of symmetry is expected, such as distribution of incomes, and a match of symmetry to an expected result is worth noting.

Outlier

> – A variable’s value significantly different from most other values in a sample.
>
> An outlier may be genuine, indicating an individual of particular interest. Or an outlier may be the result of a mistake, and should be corrected if possible.

Students are far too prone to interpreting any extreme value as an outlier. Natural variation will always throw up extreme values. Mistakes should be removed, but generally unless four or more standard deviations from the mean, or perhaps a quarter of the range away from other values, any outlier should not be removed from a data set.

Students should, in general, do all their calculations including any extreme values, and then mention in their analysis the result of this (for example dragging the mean away from the true value).

Cluster

> – A distinct group of values in a distribution.

In general ignore the presence of light clustering. Clusters will always occur as a result of natural variation, most especially for small sample sizes students deal with (and also for thinly spread data).

However clustering may represent real features and are worth discussing if the effect is very strong. In the middle values of the data clusters may indicate a *bi-modal* or multi-modal distribution. If an obvious cluster occurs towards the high or low ends, it may indicate a bias in the sampling (some distinct group is over-represented).

Modality

> – A unimodal distribution has values clustering mainly about a single central value.
>
> – A bimodal distribution has values clustering about two separated central points. It is common in populations divided in half by some critical difference, such as gender.

Modality is best seen visually, and is one occasion where box-and-whisker graphs are not useful. Note that modality says nothing about symmetry.

Tendency at Extremes

Most distributions cluster around a central value or values, and extreme values are uncommon. (Statisticians talk about *kurtosis* to describe this effect, though not at school level.)

However there are some distributions with a high number of extreme events (e.g. in general the weather tends to work this way –droughts are more common than expected from short dry periods). If your distribution has a large number of extreme values this might be worth mentioning.

Measures of Data

Measures of data fall into two distinct groups. Those that measure the *central tendencies* and those that measure the *spread or variability*.

Discussion of all these measures should include using the graphs of the distribution to stress their meaning.

Measures of centre

These measure the typical value for a variable. Students should more or less stick to the *mean* and the *median*.

Median

– The middle value of a data set when placed in order.

The median is the most stable measure of centre as it is not influenced by the random appearance of extreme values. It is only rarely inappropriate (e.g.for strongly *bi-modal* distributions).

We can use our sample statistics to calculate how likely the population's true median is to fall inside a range. This is a **confidence interval** for the median.

– A 95% confidence interval for the median can be found using: **median** $\pm \frac{\mathbf{1.5 \times \mathit{IQR}}}{\sqrt{\mathbf{n}}}$

This finds the interval (range of values) which contains the true value of the median 95% of the time in the long run. The lower and upper *bounds* of a confidence interval are the *confidence limits*.

Two Samples using Medians

When we compare two samples we cannot just compare the sample statistics, as they will always have some sampling error (variation due to sampling), and so it might be just fluke that one is bigger than the other. We compare the medians using their *confidence intervals*.

If the confidence intervals of two medians do not overlap then we can have confidence that there is a difference between the medians of the two groups. Sometimes this is described as a "*statistically significant*" difference.

Similarly when studying an *induced variation*, we can say that we are confident that the effect of the variation is ***significant*** when the confidence intervals of the medians do not overlap.

Mean

– The centre of mass of the values in a distribution

Calculated by adding the values and then dividing this total by the number of values.

The mean can be influenced by unusually large or unusually small values. Too much reliance should not be placed on a mean without ensuring that is appropriate for the distribution. Generally it should be stressed less heavily than the median, especially for skewed distributions.

While not part of Y12 we can also find **confidence intervals** for our means

– A 95% confidence interval for the mean is can be found using: **mean** $\pm \frac{\mathbf{2}\,\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$

This finds the interval which contains the true value of the mean 95% of the time in the long run. We can compare different samples using the confidence intervals of means as we do for medians, but not generally at Year 12.

Measures of spread (measures of variability, measures of dispersion)

A number that conveys the degree to which values in a distribution differ from each other.

We generally use: *interquartile range* and *range*. The *standard deviation* is another common measure of spread.

Interquartile range

– The width of an interval that contains the middle 50% of the values in the distribution.

Calculated as the difference between the *upper quartile* and *lower quartile*, that is, the values a halfway between the median and the smallest and largest values when the values are ordered by size.

The interquartile range is a stable measure of spread in that it is not influenced by unusually large or unusually small values. The interquartile range is more useful as a measure of spread than the *range* because of this stability.

Range

– The difference between the largest and smallest values in the distribution.

The range is less useful than other measures of spread because it is strongly influenced by the presence of a the single largest and smallest values. It also is dependent on sample size, because as samples grow there is an increased tendency to get very large and very small values, whereas the other measures used are, ideally, mostly independent of sample size.

Other Ranges

Because the range is influenced by the most extreme values and the IQR only includes half the values, statisticians often use ranges in between them. The most commonly used is the “95% range” in which the top and bottom 2.5% range of values are excluded. Note that for the sample size of 30 generally expected in NCEA assessments at Y12 that is all the values excluding only the largest and smallest.

Sometimes measures such as “effective range” are used, judged on the basis of removing outliers. This is highly dependent on personal choices and should be used with extreme care.

Standard deviation

– A measure of spread from the mean for a distribution.

It is calculated by taking the square root of the *average* of the squares of the deviations of the values from their mean. A calculator or software should be used to calculate it (most will offer two ways, and the σ_{x-1} is the correct one for a sample, but it doesn’t matter at Y12).

If most values are close to the mean then the standard deviation is small, but like the mean it is heavily influenced by unusually large or unusually small values.

For a *Normal* distribution 66% of values will lie within ± 1 standard deviation, 95% will lie within ± 2 standard deviations and 99% of of values will lie within ± 3 standard deviations.

The standard deviation of the sample is also useful in generating the *confidence interval* of the mean.

Data displays

An important part of most analyses is the display of the data. Graphs, in particular, are a good way of spotting features such as skew and outliers.

There is no point providing a graph that does not show any features, but merely repeats the data presentation in another form. Students should think carefully about what the graph does to help their analysis (and not just draw the simplest graph they can think of out of laziness).

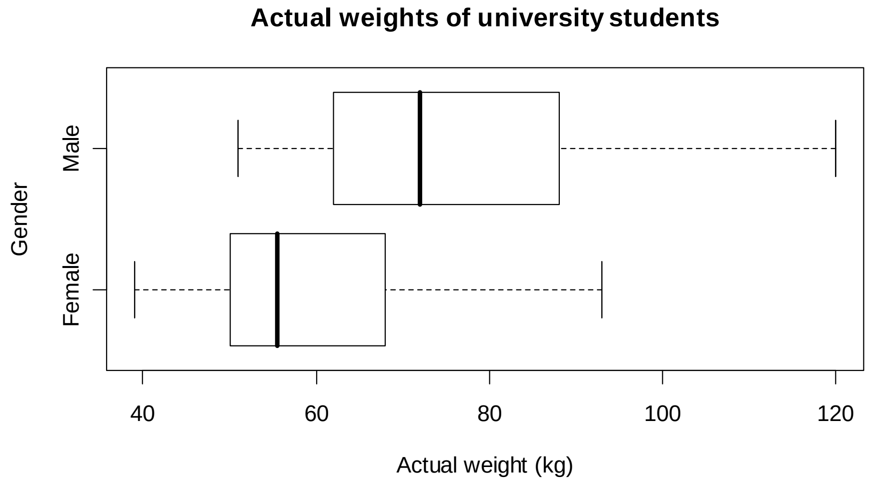
Generally the data will be presented to you in a table. If not then it is important that it is provided in full in your analysis.

Box-and-Whisker Plot (box and whisker diagram, box and whisker graph, box plot)

> – A ‘box’ that extends from the lower quartile to the upper quartile, with a line drawn at the median. One ‘whisker’ is drawn from the upper quartile to the maximum value and the other ‘whisker’ is drawn from the lower quartile to the minimum value.
>
> Sometimes *extreme* values are marked as dots and the whiskers exclude them.
>
> Some box and whisker plots only extend to the 95% limits. If this is the case then it should be clearly marked.
>
> Some plots also mark the mean, usually with a cross.

Box and whisker plots may be drawn horizontally or vertically.

Box and whisker plots are particularly useful for comparing the distribution of a numerical variable by displaying side-by-side box and whisker plots on the same scale. They are also useful when the number of values to be plotted is reasonably large.



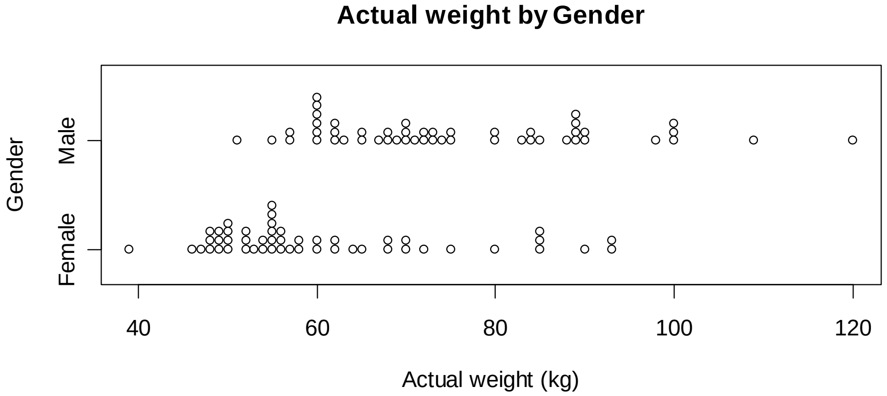
When using a comparative box-and-whisker to make a decision on whether medians are statistically different it is helpful to draw in the confidence intervals across the median lines. That way the amount of overlap, or lack of overlap, is clearly visible.

Year 12 students should consider box-and-whisker plots as the starting point of any comparison of variables between two or more groups.

Dot plot (dot graph, dotplot)

– Each dot represents each value of the variable in the sample. If a value occurs more than once, the dots are placed one above the other so that the height of the column of dots represents the frequency for that value.

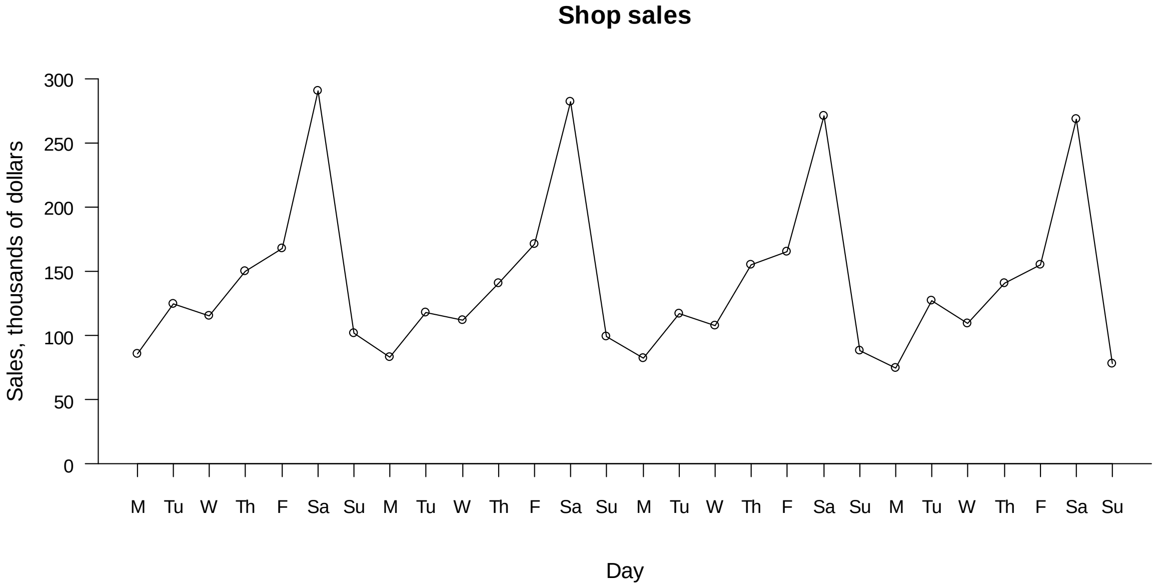
Dot plots are particularly useful for comparing distributions when the number of values is reasonably small.



Because a dot plot merely repeats the data in another form they generally provided little extra information. Students should have a good reason for using a dot plot instead of a box-and-whisker (such as to demonstrate a bi-modal distribution, or peculiar clustering).

Line graph

– A series of points representing individual observations spaced proportionally to their time of measurement are connected by line segments.



Line graphs are useful for showing changes in a variable over time, and particularly for showing trends in data. They should **never** be used for data that does not have a natural order.

Bar graphs and Histograms

– Display the distribution of a variable in rectangles, one drawn for each discrete value (bar graph) or range of values (histograms), where the height of each rectangle is proportional to the frequency of values in that value or interval.

Students should resist the temptation to draw histograms because they are easy. They are, in general, a lower level of analysis, and not suitable on their own for Achievement at Year 12.

Writing a Question

The structure of the assessment will require you to write a “question” that can be answered using the data you are given.

It is important to achieve at Year 12 that the question be at a sufficiently high level.

- **It must involve a comparison**. One group’s values for one *variable* must be compared to another group’s values for that same *variable*.
- **It must use measurable statistics**. It is no good asking which is best or if they are different unless “best” or “different” is well defined by some *measure* (i.e. number value).
- The question should address **a possible difference in the population**. You should be investigating that.
- However, while there must be some chance of there being a result, **the answer can be negative**. You are not punished for finding that some measure is not increased by some activity or that different parts of the population have the same values.

  Do not spend excessive time looking for something exciting in the data to report on.
- **Stick to the population sampled**. Your investigation cannot answer questions about populations other than the one sampled.

  You can speculate in your conclusion, if you want, about the likelihood the results carry over, but not in the actual question posed at the start of your investigation.

Do not expect to use all the data presented to you in the test. In your question you will generally chose one variable which you will compare for different members of the population. That generally means that you can cross out all the data columns except the one you are investigating.

In particular you must **not** do a **bivariate** investigation – that is, you may not compare one *variable* with another *variable* (for example seeing if the height of students is linked to their weight). While a commonly used and very powerful way to analyse data, it is the subject of another Achievement Standard.

Students looking for Achieved should pick a simple, easily understood question that involves comparing one variable for two groups in the population.

Those looking for Merit and Excellence should think about questions that involve looking at one variable across multiple groups in the population, especially if they could result in a *trend*.

Generating Random Numbers

Samples must always be taken truly randomly, which does not include any method which involves personal choice.

Students should know how to find and use the Ran# function on their calculators.

To select a random number between 1 and B, my suggested method is to use **Ran# × B + 1** and **truncate** the result (i.e. ignore decimal places).

To select a random number between A and B, use **Ran# × (B + 1 – A) + A** and **truncate**.

It is important to add the +1 when finding the difference between the A and B, because it is inclusive of both end numbers (e.g. from 5 to 7 is three numbers, 5, 6 and 7, not 7 – 5 = 2).

Others suggest the same basic formulas, but adding + 0.5, and then setting your calculator to round to the nearest whole number.

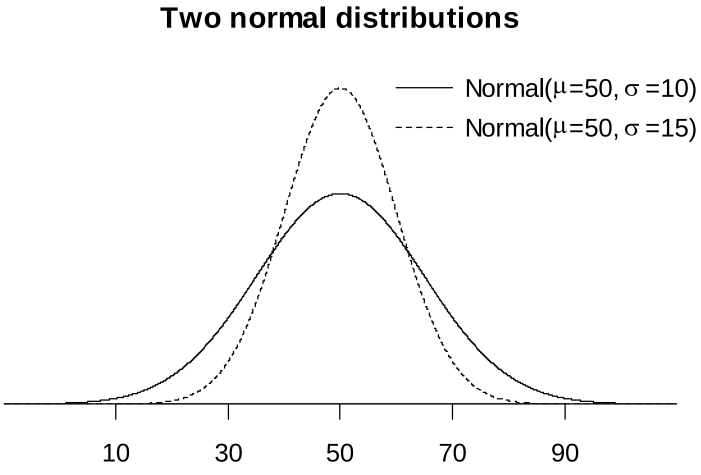
MSExcel has a roundbetween(low, high) function, that automatically gives an integer value from the low number to the high one.

Normal distribution

A family of theoretical distributions that is useful as a model for some continuous random variables, as they commonly appear in nature, and are a useful approximation for many more.

Each member of this family of distributions is identified by specifying the mean μ and standard deviation σ.

The probability density function of a normal distribution is a symmetrical, bell-shaped curve, centred at its mean μ. The graphs of the probability density functions of two normal distributions are shown below, one with μ = 50 and σ = 15 and the other with μ = 50 and σ = 10.


Two normal distributions
Normal(μ=50, σ =10)
Normal(μ=50, σ =15)