Critical Reasoning 13 - Probability Distributions

with guest editor Sintu Tonjeni

Carl Friedrich Gauss (*1777 - 1855*) *Mathematician and Philosopher after whom the Normal* (*or Gaussian*) *Distribution is Named.*

In Critical Reasoning 10 we took a preliminary look at some of the basic properties of probabilities, including some mathematical techniques for calculating them, either singly or in combination with another probability. In the world of research, whether in the physical or social sciences however, we are not so much interested in the probabilities of individual outcomes as the distribution of possible outcomes of a random experiment, survey or statistical inference procedure. A **probability distribution** then assigns a probability to each measurable subset of outcomes of such a procedure. (Wikipedia: Probability distribution) If possible we want to gauge and to quantify the likelihood of events, so examined.

Why Do Probability Distributions Matter?

One reason is because when we naturally look at events we find patterns in what we see. Almost habitually, we make decisions based on how we perceive such events as they happen or as they happen over again. This is the basis of intuition. However, intuition does not always serve us well. As we saw in Critical Reasoning 04, concerning fallacies and Critical Reasoning 06, concerning heuristics, prejudices, biases and just plain faulty thinking lead to unwarranted and erroneous conclusions, none more so than those associated with probabilities. Studying probability distributions therefore allows us to be more mindful of the way we use our fallible intuition, by encouraging us to come up with logically consistent, intuitive hypotheses and by challenging our biases with empirical evidence. This is not simply an abstract or academic exercise. In fact, it goes to the heart of how, for example, we estimate risk, perceive the trustworthiness of persons and corporations, to how impartial judges are expected to weigh the preponderance of evidence before them.

Why Are Probability Distributions Now Included in the Humanities Program?

Unfortunately, this is where many highly capable humanities students are either tempted to or actually do drop out from their academic program, either because they feel intimidated by the dreaded, compulsory “Stats Modules” or because they fail to see the relevance of such material for their intended career path. The short answer to this question then is that universities around the world are no longer prepared to confer degrees upon students who are mere repositories of information but who lack the numerate capacity to actually generate knowledge through scientific

research or, at the very least to *understand* the process. We cannot promise to inspire you in this and the remaining study units regarding probability, but we have at least endeavoured to be clear.

Types of Variables

Before we can represent a probability distribution we first have to decide whether the variable(s) we want to represent is (are) either discrete or continuous. A **discrete variable** can only take on a finite number of countable values, such as the number of students attending a lecture: 1, 2, 3, 4… We cannot have halves or quarters of a student present – such cases are impossible.[1] **Categorical**, also known as **qualitative variables** are all discrete and can take on one of a limited, and usually fixed, number of possible values, assigning each individual or unit of observation to a particular group or nominal category on the basis of some qualitative property. (Wikipedia: Categorical variable) If, for example, your first name is Amy and you have black hair then those are two of the many "valves" that the discrete variables, "first name" and "hair colour" can take on. Such variables are also mutually exclusive – you cannot have these different attributes at the same time. If again your first name is Amy and you have black hair, your first name cannot also be Becky with blond hair. Nor, obviously can you be 0.4 Amy and 0.6 Becky *etc.* There are three types of categorical variables: binary, nominal, and ordinal variables. **Binary variables** can take one of two mutually exclusive values *e.g.* positive *vs.* negative, true *vs.* false, diagnosed *vs.* undiagnosed *etc.* **Nominal variables** are used to name, label or categorize particular attributes that are being measured. They can take on values that represent logically separate concepts that cannot be meaningfully ordered *e.g.* the names 'Amy' or 'Becky' or … the nationalities 'South African' or 'Namibian' or …, the species *Homo sapiens* or *Pan troglodytes* or … *etc.* **Ordinal variables** are categorical variables that can be ranked or ordered, even though we may not know the distances or time intervals between such categories *e.g.* 'Amy 1st position', 'Becky 2nd position' and 'Charlie 3rd position' in a race, competition or alphabetically.

Random Variables & Observations

In the interests of expediency, in this context we have used the term "variable" rather loosely to mean either "random variable" or "observation", whereas there is a subtle but important distinction. *E.g.* There may be X students attending a lecture which can be observed to take on a value of 0, 1, 2, 3 *etc.* The value that X takes on in this event, might be denoted x. This little x is a variable.

Big X, meanwhile, is also a variable. It's actually the random variable in this context because it stands for the unknown, prior event, whose probability we are trying to measure with the probability distribution. This makes a meaningful difference when expressing probabilities in terms of events like $Pr(\{X = x\})$.

If the value of a discrete variable may be obtained by counting, then the value of a **continuous variable** may be obtained by measuring an uncountable set of values along a continuum. Continuous variables, such as the distance you walk to class or the time you took to do your hair, can take on any value, including fractional values within a certain range. Maybe you walked 1.1 km to class and took 9.7 minutes to do your hair but you could not have walked a million km to class or taken

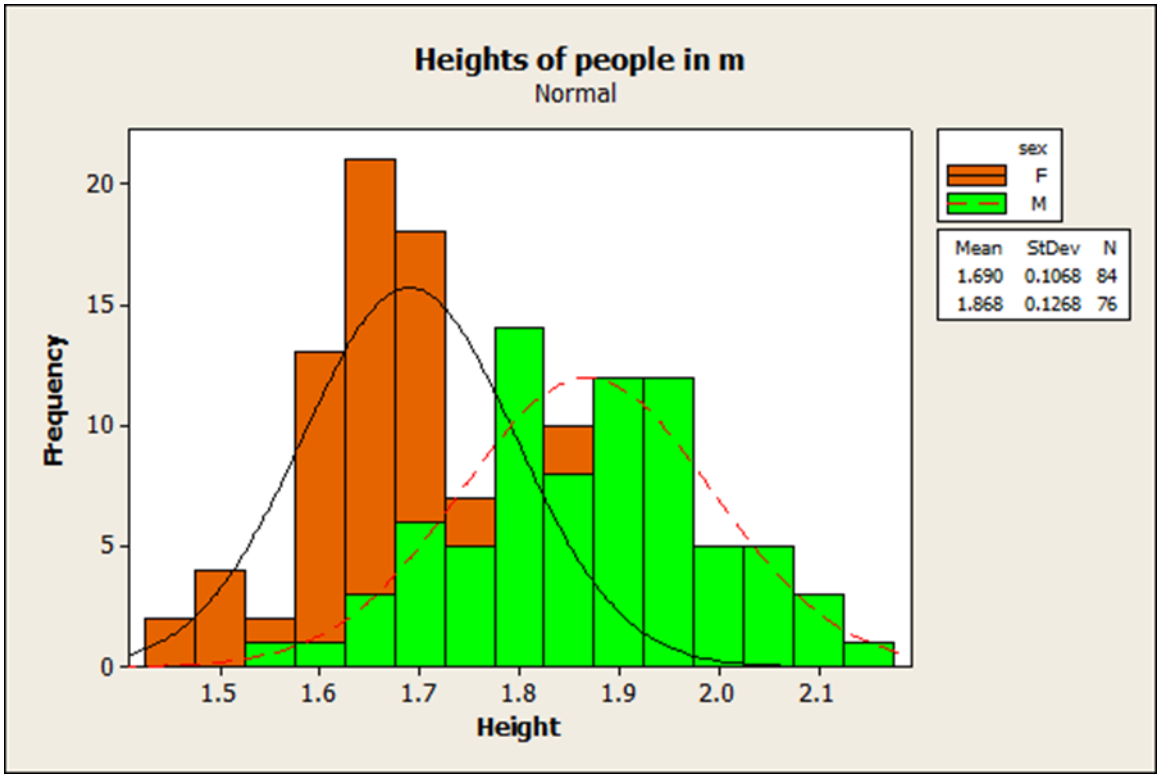
[1] Even a person with multiple identities counts as one being, not several fractions.

negative 3 minutes to do your hair. Continuous variables can be measured according to various scales. **Interval scales** comprise of units of equal size which allow us to both order and quantify the difference between values *e.g.* the Celsius scale. O °C however is not really zero because it is not the coldest temperature that can be. 0 °K is however *is* absolute zero. **Ratio scales** are quantitative scales with a defined zero value. Therefore they allow us not only to order and quantify the difference between values but also to calculate ratios, such as for height or weight.

From Histograms to Distribution Curves

We have all drawn histograms at school. They make it easier to take in, at a glance, the way the data we have collected are distributed. If we have a manageable number of discrete values, we can construct a histogram directly, otherwise we have to divide the range in to a series of intervals and then count how many values fall within each interval. If we have chosen a variable such as the height of a population of students, we will probably end up with a histogram that is clustered about the middle and tails off on either side. If we join the tops of each of our rectangles we will obtain a jagged line approximating a **distribution curve** which is a graph of the frequencies of different values of a variable in a statistical distribution. (Merriam-Webster Dictionary)

For now, take a look at the following histogram created by Mr. Anderson[2] as part an activity with his 7th Grade maths class.



Clearly this was a very tall **cohort** (or group of people with a shared defining characteristic) for Grade 7, or maybe they were his Grade 12 class. What we can see is two sets of data for height presented

[2] See his blog "Human Histogram" at https://banderson02.wordpress.com/2014/05/12/human-histogram/

on the same set of axes – one for females, coded orange, and one for males, coded green. This represents a **bimodal distribution** of data (having two obvious relative modes, or data peaks).

Please note how all the axes are correctly labelled as well as numbered and that there is a key explaining the colour coding. And (this is often omitted even by senior students) there is a title that meaningfully describes the graph! Also Mr. Anderson has provided a few descriptive statistics at right that, at a glance, help in the interpretation of the data.

If Mr. Anderson had simply joined the tops of each set of rectangles in his histogram, he would have ended up with two very chunky "curves" (if one could even call them curves). Instead, he probably used a statistics program to smoothen the curves and then superimposed them on the diagram. What is going on mathematically from histograms to distribution curves is that to obtain ever smoother curves would require one to include ever more and more little rectangles representing ever more data. In other words the number of such cases (n) included would have to become very large. On the other hand, to accommodate so many little rectangles in the same space would require that they be drawn ever more closely. In other words, the width of each of their bases would have to become very small, as small as you like proportional to their number, but not zero. You can do this manually, which is very laborious, or you can use a program like Excel™ to draw them for you. Either which way you will have constructed, what is informally known as a "bell curve" because of its characteristic shape. This figure will crop up time and again in the remaining discussion, so it is important to understand its geometric origins, although it can also be derived algebraically via a somewhat complicated formula.

Probability Distributions

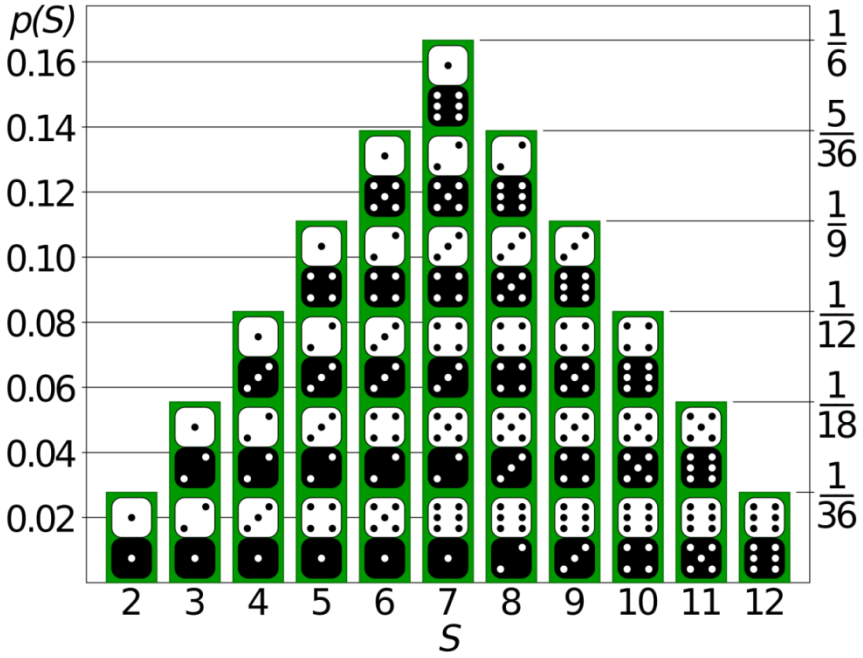
It is not just physical data that are distributed; sometimes we may want to visualise the way that certain abstract information is distributed, in particular probabilities. And although probabilities are abstractions, they can represent very real states of affairs: What is the probability of my passing a given statistics module, knowing that historically just such a proportion has succeeded? What is the probability of my avoiding lung cancer, given that I gave up smoking a pack a day ten years ago?

By way of example, consider the much simpler question: What is the probability of throwing a particular number using only two dice? We know from common sense that all possible throws have to be whole numbers ranging from a minimum of 2 ("snake eyes") to a maximum of 12 (double six). We also know that that there are more than one way of throwing certain numbers like 4. You could throw a 1 and a 3 or a 2 and a 2, or indeed a 3 and a 1 if you are using different coloured dice. So clearly, throwing a 4 is more likely than throwing a 2 on any round. More precisely throwing a 4 with two different coloured dice is exactly 3 times more likely than throwing a 2 because there are 3 out of 36 ways of throwing a 4 and only 1 out of 36 ways of throwing a 2.

Here, it worth noting a few conventions with regard to probability distributions and probabilities. Probability distributions are functions that take events as inputs and deliver probabilities as outputs. Recall, from Critical Reasoning 10, that these probabilities are always numbers between 0 and 1. When an event that has no chance of happening, we say that it has a probability of zero, while an event that is certain to happen has probability 1. Probabilities in between impossibility and certainty

are, by scientific convention, expressed as decimal fractions like 0.3 instead of the more familiar 30%. Recall also that, if you add up all the possible probabilities, you always get the probability of 1.

Tim Stellmach has constructed the following graphic to help visualise the probability distribution for throws on two dice of different colours. The graph, cleverly depicted using the facets of different coloured dice, represents a probability mass function (PMF) which, (recall from Critical Reasoning 10), gives the probability that a discrete random variable is exactly equal to some value. The horizontal axis represents the event that a certain number is thrown. It indexes events that the random variable S takes on the value s. As such it is the input of the PMF. The vertical axes (as a decimal on the left and a proper fraction on the right) are the probabilities that correspond to each of these events. More precisely, we take the event $\{S = s\}$ and find it's probability, $P(\{S = s\})$. For short, we simply to write $p(S)$.



A Representation of the Probability Mass Function of the Sum of Two Regular Dice by Tim Stellmach (*Wikipedia: Dice Distribution* (*bar*))

There are a several features of Stellmach's graphic that stand out: Clearly 7 is the most probable of all possible throws, which is why, no doubt, many people consider it a "lucky" number. Of course "luck" has nothing to do with it. There are simply more ways of throwing a 7 on two different coloured dice than for any other number. Secondly, the figure is neatly symmetrical about the middle, tapering off on either side. Thirdly, the PMF allows one to calculate the probabilities of events represented by intervals, such as the probability of throwing a number greater than 9 *i.e.* $p(S > 9)$. The event $\{S > 9\}$ is the event that the dice are thrown to get a number greater than 9. If a throw gives a number greater than 9 then either a ten, eleven or twelve was thrown. Using our notation, the event $\{S > 9\}$ occurs when either $\{S = 10\}$, $\{S = 11\}$ or $\{S = 12\}$ occurs. We can thus re-express the event $\{S > 9\}$ as occurring if and only if the event $\{S = 10\}$ occurs, or $\{S = 11\}$ occurs, or $\{S = 12\}$ occurs.

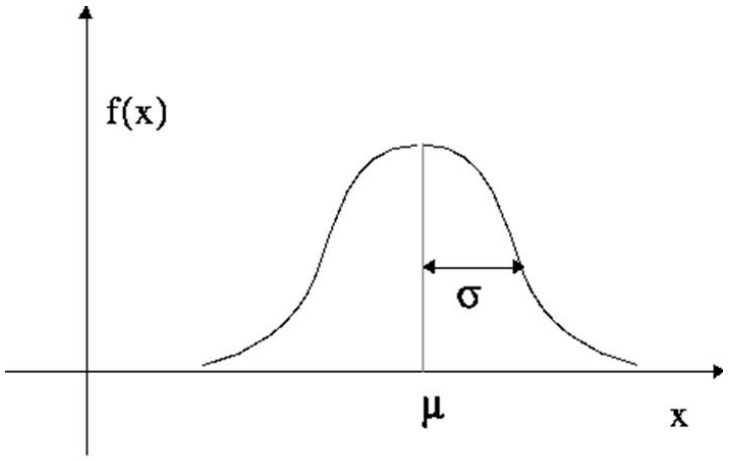
We can consult the chart to look up the probability of throwing a ten. According to the chart the value of $p(S = 10)$ is $\frac{1}{12}$. Similarly we find that $p(S = 11)$ is $\frac{1}{18}$ and $p(S = 12)$ is $\frac{1}{36}$. Using the additive rule from Critical Reasoning 10, we then add these probabilities to get:

$$p(S > 9) = p(S = 10) + p(S = 11) + p(S = 12) = \frac{1}{12} + \frac{1}{18} + \frac{1}{36} = \frac{1}{6}$$

For some continuous random variables such as height, mass or time however, we will have to consider the **normal (Gaussian) distribution** which is a type of continuous probability distribution for a random variable. A random variable with a Gaussian distribution is said to be normally distributed. Such distributions have unique properties that are useful in analytic studies as we shall see below. The general form of a normal distribution is given by its **probability density function** (**PDF**). The PDF is a function of a continuous random variable that specifies the probability of a random variable falling within a particular range of values. As mentioned, such curves can be constructed geometrically; however it is much less time consuming to simply hand the task over to a spreadsheet program like Excel™, which calculates the function algebraically. For any given random variable x, then, the PDF is given by:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

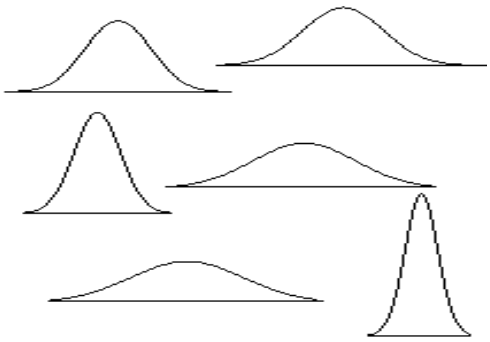
Please do not be alarmed, you will not be required to memorise or even understand this formula. The reason for reproducing it is to make the point that for any variable x, the value of the function depends on only two parameters, the mean (μ) and the standard deviation (σ) shown below; the rest are constants. Both μ and σ are defined and discussed under the sections on 'Measures of Central Tendency' and 'Measures of Spread or Dispersion', below.



The Value of the Probability Density Function above, for a Particular x Depends only on the Mean (μ) and the Standard Deviation (σ)

When μ is set at 0 and σ at 1 we get the **standard normal distribution**, which is a special case of the normal distribution. If the values of μ and σ are tweaked, as they have been below, we obtain variations on the normal distribution; one resembling a speed bump, another more like a traffic cone. What all these curves below have in common however is that they are continuous (smooth),

symmetrical about the mean and approach the x axis **asymptotically** *i.e.* as a limit; they get ever closer to the x axis but don't ever touch it or intersect.



Variations on the Normal Distribution Curve

Populations and Samples

Before we continue it is important that we draw a clear distinction between populations and samples. A **population** is a complete set of items, beings or events that share some property in common, be they: girls over 14; motorists in Cape Town; solar eclipses since the year 1970; Catholics across the globe *i.e. every single one of them*. Populations, understandably, tend to be impracticably large, unless that is, they represent a very precisely defined group, such as: "students in Mr. Anderson's Grade 7 Maths class of 2012." Suppose now, a drug company wants to ensure that their cough lozenges for teenagers are safe for use by all humans 13 and over (that is the defined population). They would have to enrol a sample, test it on them in a clinical trial and submit their results for approval. A **sample** then is a subset of the population, drawn at random or by some other statistical procedure from the population. If the drug company had enrolled all and only those students in Mr. Anderson's Grade 7 Maths class of 2012 in their clinical trial, they would now represent a sample of the defined population. So clearly, the terms "sample" and "population" are relative.

By convention, numbers summarizing data for populations are referred to as **parameters** and are represented by Greek letters. Those summarizing data for samples, on the other hand, are referred to as **statistics** and are represented by Roman letters. This is important in statistics because sometimes a formula for one measure looks slightly different depending upon whether it is describing a population or a sample. Then of course, it is important to know the context in which these terms are used. Is someone using them in the vernacular or are they trying to convey something quite precisely?

Measures of Central Tendency

Most data sets, such as for height, weight, intelligence *etc.*, if plotted as a histogram, will tend to cluster around a central point. That is what people loosely call "the average", however there are several ways, statistically speaking, of being "average" or "the average". The terms "mean," "median" and "mode" all describe "averages" in one sense or another. We shall examine each briefly in turn. To begin, we must decide whether the distribution of variables is discrete (like dice throws) or continuous (like height).

For discrete distributions the mean, μ is calculated by summing up every single value of a variable, x multiplied by its probability, $P(x)$. Thus:

$$\mu = \sum xP(x)$$

A similar formula applies to continuous probability distributions; however it is very seldom that the population mean can be calculated in this way or even precisely known because of the impossibly large number of individuals that have to be taken into account - one can not measure or know *everybody's* height, *everybody's* mass, *everybody's* mathematical aptitude *etc.* One exception to this is IQ scores, where the mean is simply *defined* as 100. For the most part we will be dealing with data sets that are samples of populations.

Instead, we can calculate the arithmetic sample mean ($\bar{x}$) of data set as the sum of the sampled values divided by a count of their number (n). Thus:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

This formula for the **arithmetic mean** is probably familiar to most of us and is the one that we will use almost exclusively in the Social Sciences when we refer loosely to "the mean". Note, that it is important to distinguish the sample mean ($\bar{x}$) from the population mean (μ) because when we come to calculate other statistics for samples it is the former that we have to use rather the latter.

There are besides, a couple of other measures of the mean, namely the geometric and harmonic means. We will only discuss the **geometric mean** further, which is used to find the mean of quantities involving rates. A **rate** is simply one quantity measured against another, in different units (often per unit time). In in Physical Anthropology, for example, it is not uncommon to compare growth rates of children and adolescents. As the word "geometric" implies, we have to multiply the rates rather than add them. Thus:

$$\overline{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

If you are unfamiliar with the big Π operator above, it is analogous to the more familiar big Σ. This one is telling us to multiply all the values from 1 to n. So for example, if we take a sample of the growth rates of 5 boys from Mr. Anderson's 7th Grade class over a period of a year we might obtain the following data:

Growth rates of sample: 3; 4; 3; 5; 4 cm per year

We can then calculate the geometric mean for rate of growth by multiplying the five numbers together and then taking the 5th root ($\sqrt[5]{\ }$) of the product.

Besides the mean, the terms "median" and "mode" also convey something of what we informally mean by "average". The **median** is the middle most value of a data set. If for example, we were considering height, for an odd number of students lined up shortest to tallest; the median is simply the height of the student in the middle. For an even number of students, the median is the sum of heights of two middle most students divided by 2.

The **mode**, on the other hand, is the value that appears most often in a set of data values. (Think of the Afrikaans word for fashion: *mode*.) If, for example, there are three students of 143 cm in height and there is no larger group of students who are the same height as each other, then the mode of the heights of all the students is simply 143 cm, because that is the most frequent height.

In an ideal world of symmetrically distributed data, the mean, median and mode would coincide, *i.e.* they would lie one on top of the other when depicted graphically. Suppose however that Mr. Anderson's 7th Grade class of 2012 had an unusually high number of students on the basketball team, say 10 out of a total of 24, then because taller students tend to **self-select** for sports such as basketball, the distribution of heights would be skewed to the right. In such a distribution the mean, median and mode would not coincide but be dragged apart.

Self-Selection

In the case of Basketball we are not surprised to find tall players over represented, because tallness is an advantage in such a game. It is not that playing Basketball causes tallness. Rather, those who are already unusually tall tend to self-select for participation in games like basketball, whereas those of a slighter build might self-select as gymnasts or jockeys. It is important therefore to exclude self-selection as a source of bias in certain investigations.

Measures of Spread or Dispersion

Some data values are remarkably uniform, such as the mass of electrons; while others, such as the mass of humans are remarkably dispersed or spread out. Measures of spread or **dispersion**, then all quantify this tendency one way or the other, increasing from 0 for absolute conformity upwards as data become more diverse. Fortunately, some measures of spread or dispersion are measured in the same units as the data which they describe. For these measures then, if the data are, for example, measured in kg then so too will be the units of measure of dispersion describing the data. For other such measures of dispersion, however it may not make sense to describe them in the same units. (Wikipedia: Statistical dispersion)

In High School everybody learned about **range**, the interval from the lowest to highest value of any data set, and about **interquartile range**, as the interval between upper and lower quartiles $Q_3 - Q_1$ of any distribution, as measures of dispersion. Below we introduce the **standard deviation** (and its square, **variance**), as additional measures of dispersion. Before we define them though, it is important to decide whether the data before us represent a *population*, in which case we use the lower case Greek letter sigma σ for population standard deviation and σ^2 for population variance. If however a sample *is drawn from a population*, we just use a regular s for sample standard deviation and s^2 for sample variance.

Standard deviation is a measure of the extent to which the data points of group deviate from the mean of the group. As such, it is measured in the same units as the data it describes. Consider the following population of 8 values from the example at "Wikipedia: Standard deviation."

2; 4; 4; 4; 5; 5; 7; 9

The first step is to calculate the mean for the population (μ). This requires that we add all and then divide their sum by a count of their number (n = 8). Thus,

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = 5$$

Next we could calculate how much each number differs from the mean, then sum those differences and divide by the total. But this won't work because some of the numbers differ from the mean by a positive amount, two don't differ at all, and the rest differ by a negative amount. So some

differences are going to end up cancelling each other out, which we don't want. The easiest way out is to simply instruct a program, like Excel™, to truncate the sign before each difference. The mathematical function needed to do this is the absolute value function "Abs x" or "$|x|$" for short. However the absolute vale function doesn't simply truncate a sign, it performs a two-step procedure. First it squares the number, then it delivers the positive square root of the square. Thus:

$$|x| = \sqrt[+]{x^2}$$

So we will have to replicate this two-step process with our data, as follows:

$(2 - 5)^2 = (-3)^2 = 9$ $(5 - 5)^2 = (0)^2 = 0$

$(4 - 5)^2 = (-1)^2 = 1$ $(5 - 5)^2 = (0)^2 = 0$

$(4 - 5)^2 = (-1)^2 = 1$ $(7 - 5)^2 = (2)^2 = 4$

$(4 - 5)^2 = (-1)^2 = 1$ $(9 - 5)^2 = (4)^2 = 16$

Next we calculate the mean of these squared differences and take the positive square root of the result. Even a modestly good calculator will be able to do this on one line, like this:

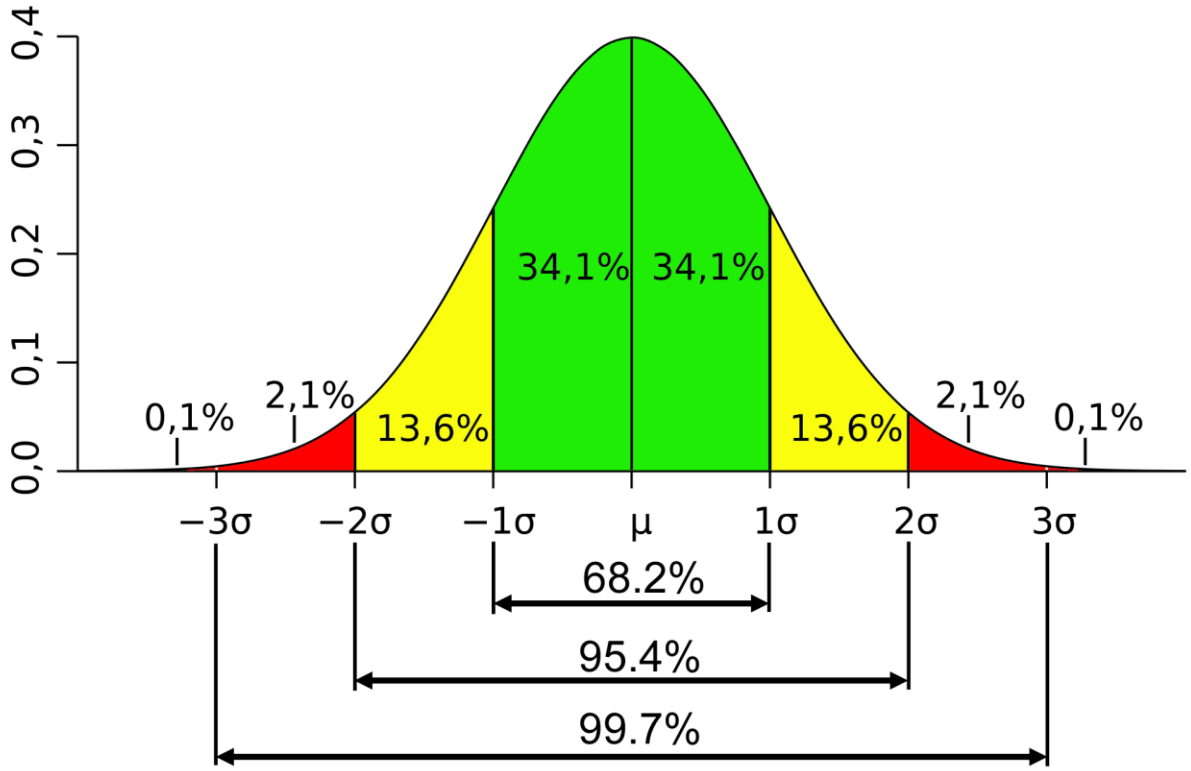
$$\sqrt[+]{\frac{9+1+1+1+0+0+4+16}{8}} = 2$$

This number then is the population standard deviation (σ) of our data and is one important measure of its spread. In general the formula for the population standard deviation of a finite data set, in which each value has the same probability, is given by:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

where n is the number of our population and μ is the mean as calculated before. Some may find this notation daunting but actually it is quite clear. Working from inside, outward, this formula tells us that for all the numbers 1 through n, we should subtract the mean, μ, from the i^{th} x and square the result. Then we should sum up all n of these results and divide the total by our n. Finally, we should take the (positive) square root of this number. This is just what we did in our example, so if you understood the example above step-by-step, you already understand the formula.

It is important to be able to visualise just what information the standard deviations convey, as in the following plot of the standard normal distribution (or bell curve), where each coloured band has a width of 1 standard deviation and the total area under the graph from -∞ to +∞ also equal to 1. Note that the colours below are just for contrast.

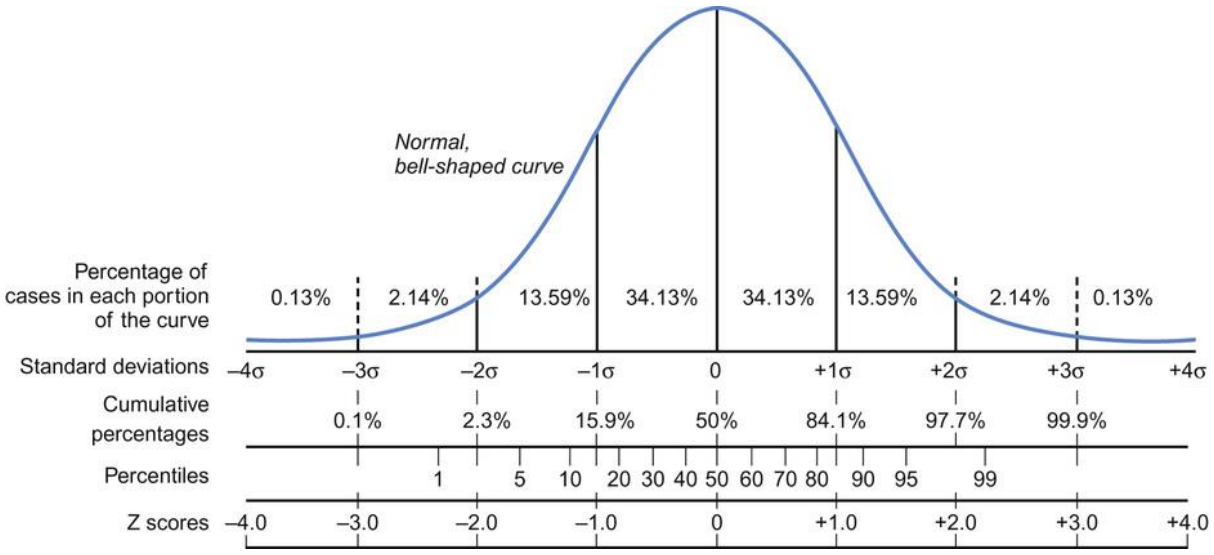


As we can see, about ⅔ of the population falls within 1 standard deviation either side of the mean (μ), while about 95% fall within 2 standard deviations and very nearly all but 0,2% within 3 standard deviations. Because this figure is symmetrical and has a total area of 1, we can, for example, figure out the proportion that fall, say up to and including 1 standard deviation to the right: 0.50 left to μ + 0.341 = 0.841. Similarly, we can calculate only the lesser portion to the right of 1 standard deviation: 1.00 - 0.841 = 0.159. This property and such calculations will be very useful in the following section.

Standardising Scores

Consider the following example: Mike takes a (reliable) test for some sort of aptitude, say lateral thinking. You are informed that he scored 114 points with a mean score of μ = 100 and a standard deviation of σ = 15 points. What should you do with such information? What does it mean in this context? What is the probability of someone scoring 114 on this test? Is Mike an expert lateral thinker?

First we must consider the standard normal function on p. 6 with a mean, μ = 0 and standard deviation, σ = 1. This is also known as the ***z*-distribution**, which is used in in testing hypotheses about means or proportions of samples drawn from populations under certain conditions. The standard score or **z-score** is a measure of how many standard deviations (positive or negative) a datum or raw score x is above the mean. In the graphic below the standard normal curve is shown in blue with the various z-scores marked off at the very bottom corresponding to the standard deviations -4σ to +4σ marked off above.



The process of converting raw scores to z-scores is known as **standardising** or **normalisation** and allows us to compare our data with a standardised scheme for interpreting the distribution of probabilities. There are however a number of provisos. Firstly, the population distribution must be normal or approximately normal. Secondly, we must know the population parameters μ an σ, although in practice we almost never do. Nevertheless, supposing we have the requisite population parameters, we can convert a raw score, x into z-score as follows:

$$z = \frac{x - \mu}{\sigma}$$

This z score (positive or negative) tells us the distance between the raw score, x and the mean, μ in units of standard deviations, σ. If we substitute the information for Mike's lateral thinking score, we get:

$$z = \frac{114-100}{15} = 0.93 \text{ standard deviations}$$

which is just less than +1 standard deviation away from the mean. So clearly Mike's score lies above the mean for lateral thinking, but not quite as exceptionally as he might have hoped.

That is not to say that +1 or indeed (+2 or -3 *etc.*) standard deviations are "special" or critical intervals above or below which a score is observed to fall. Instead, as continuous curves, we should regard whole number standard deviations as statistical landmarks that allow us to estimate (and visualise) into what proportion a score falls. See both of the normal curves above. As a rule of thumb however, we should be guided by our intended audience. If a particular medical journal reports results as percentiles, then such intervals should correspond to our critical values. When we come to Critical Reasoning 15 concerning statistical hypothesis testing however, we will follow the more familiar scientific convention of using critical values corresponding to preselected confidence levels of 66.7%, 95%, 99.9% and so on.

Using z-tables

The area under the z-function to our standardised valve can be calculated and represents the probability of finding such a score in that range. For convenience these calculations are usually published at the end of most statistics textbooks or in a separate manual. They are available here courtesy of the University of Florida. There are two separate tables on separate pages for converting z-scores to probabilities. The first is for negative z-scores that lie to the left of, and including, the mean. The second is for positive z-scores that lie to the right of, and including, the mean.

Once you have selected the appropriate table for the z-score you want to convert, say Mike's z-score of 0.93, run your eye (or finger) down the first column till you come to the row beginning with 0.9. Next, run your eye (or finger) across that row until you come to the column for the second digit, in this case 0.03. The p-value corresponding to the z-score you are looking up will be found at the intersection of the row and column for the first and second digits, respectively. In the case of Mike's z-score this corresponds to a p-value of 0.8238. This means that approximately 82% of the population have a score equal to or less than that of Mike's, *i.e.* the greater proportion.

If we want to work out the lesser proportion *i.e.* those whose scores were equal or better than, or fell to the right of Mike's score, we simply subtract the p-value for the greater proportion from 1 because that is the total area under the graph. Thus: 1 - 0.8238 = 0.1762 which is the proportion of the population who would score equal to or better than Mike's score.

Examples

The following two examples are slightly modified from the ones in the study guide for the Psychological Research module PCY3704, presented by UNISA. Besides being straightforward, questions very much like these have been included in almost every past paper for this module that we have consulted, so they are quite likely to come up again, and not just in UNISA Psychology exams:

1. Zola has an I.Q. score of 120 (standardised with mean of 100 and a standard deviation of 15 on a normal distribution). She also scores an 8 on a 9 point test for mathematical aptitude (standardised with a mean of 5 and a standard deviation of 1.5, also on a normal distribution). If I.Q. were a predictor of mathematical aptitude how would her scores compare?

It is difficult to compare these scores at a glance but transforming them both into z-scores allows us to compare them in terms of standard deviations from the mean. For I.Q. our transformation into a z-score looks like this:

$$z = \frac{120-100}{15} = 1.33 \text{ standard deviations}$$

while for mathematical aptitude, the transformation is as follows:

$$z = \frac{8-5}{1.5} = 2.00 \text{ standard deviations}$$

Therefore Zola's mathematical aptitude is higher than we would expect based on her I.Q. alone.

2. Phoebe, meanwhile scores 56% for an exam in Cognitive Psychology and wants to know where she stands in relation to the rest of the class, where the mean score was 52% and the standard deviation was 4. How did she fare?

Strictly speaking we have to treat Phoebe's class as a population, rather than sample from a larger population of students; otherwise we are not entitled to use the z-formula as is. Suppose that we do, then we can transform Phoebe's mark of 56% into a z-score, as follows:

$$z = \frac{56-52}{4} = 1.00 \text{ standard deviations}$$

So Phoebe's mark is one standard deviation above the mean for her class. If we look up this value on the z-table for the greater proportion, we find that her score corresponds to a p-value of 0.8413, which means that Phoebe fared better on this test than 84% of her classmates.

Sampling

When conducting research it is often necessary to draw a sample, or subset of individuals, from a given population in the hope that such a sample will be representative, or a good estimate, of the population for some characteristic of interest. The long term goal of such research is to draw inferences, from the results obtained by means of sampling, back to the population. If, for example, one obtains permission from number of residents in an old age home to check their existing blood samples for calcium levels and finds that they are unusually low, one might make an inference back to the population of elderly residents that they require some form of supplementation.

When drawing samples we try to randomise the process as far as possible in the hope that our sample will be independent and **representative**, accurately reflecting the characteristics and proportions of the of the parent population. Secondly, we try to draw as many samples as are practicable and cost effective, because smaller samples are less likely to be representative than larger samples. Unfortunately, there is no way of knowing beforehand which sample will be representative and which will not. Furthermore, any two random samples may look quite different because, after all, they are likely to consist of different individuals. The upshot of all this is that any statistic derived from a sample, such as the sample mean, will vary from one sample to the next. So what chance do we take that our sample statistics reflect the population parameters?

Self-Selection - A Related Problem

The process of sampling is also beset by a problem related to that of self-selection. If, for example one elderly resident's calcium levels are unusually low, then it is very likely that there will be another elderly resident in the same sample whose calcium levels are also unusually low. Ideally, we would like our samples to be independent; however in a later study unit we will encounter tests such as the t_d test, specifically designed to accommodate dependent samples.

Fortunately, the **sampling distribution of a statistic** (such as the mean) leads to a very useful estimate of the size of the error we may make in in estimating the mean of a population μ from the mean of our sample $\overline{x}$. If we had unlimited time and resources we could draw multiple samples of a given size in every possible way, with replacement, until we had exhausted all possibilities. If we were to work out a statistic, such as the sample mean $\overline{x}$ for each sample and then plot the

distribution of all the possible sample means, we would arrive at a sampling distribution of that statistic, which together with our individual samples have some useful properties.

This is a rather counterintuitive way of going about the business of random sampling but it is so important that it bares repetition, this time concentrating just on the mean: If we were to draw every possible sample of a given size, with replacement, from a population and calculate the arithmetic mean $\overline{x}$ for each sample, we would end up with a **sampling distribution of the mean**. This distribution has its own mean $\mu_{\overline{x}}$ which is identical to the population mean μ because, in a roundabout way, we would have sampled every individual or element in every possible way that there is to sample a given number at a time for a population of a certain size. Therefore we have:

$$\mu_{\overline{x}} = \mu$$

which is an important insight because it states that: the mean of the sampling distribution of the mean is the same as the parameter of the population mean. That we would have to go to such unusual and protracted lengths to obtain it is all par for the course is mathematics. Nobody balks at infinite sums or series or infinitesimal volumes, at least, not any more. Strangeness is no enemy of logic.

The Central Limit Theorem (CLT)[3]

The CLT is a mathematical theorem that is fundamental to our understanding of much of what remains to be explained in this study unit and beyond. Indeed, it is so important that it is hard to imagine the subject of Statistics without it. It states:

> Given a population with a finite mean μ and a finite, non-zero standard deviation σ, the sampling distribution of the mean (made using samples that were independently selected from the population), approaches a normal distribution with a mean of μ and a standard deviation of $\sigma/\sqrt{n}$ as the sample size, n, increases.

It is not necessary that you be able to prove the central limit theorem or even describe it in every detail, only that you know what it states, and what it entails for our purposes, namely that:

- Regardless of the shape, mean or standard deviation of the parent population, the distribution of the sampling means approaches a normal distribution as n increases. (In fact, it approaches very close to normal with an n of as low as 30.)
- The distribution of the sample means is described by the mean ($\mu_{\overline{x}} = \mu$) and its standard deviation is given by $\sigma/\sqrt{n}$.

While this quantity $\sigma/\sqrt{n}$ is literally the **standard deviation of the sampling means**, it is better known as the **standard error** because it is an estimate of the size of the error we are likely to make if we use the mean of the sampling means $\mu_{\overline{x}}$ as an estimate of the population mean μ. Because the standard error is so frequently used in statistics and in scientific reports it has its own symbol: $\sigma_{\overline{x}}$ The σ indicates that we are dealing with a population parameter rather than a sample statistic. The

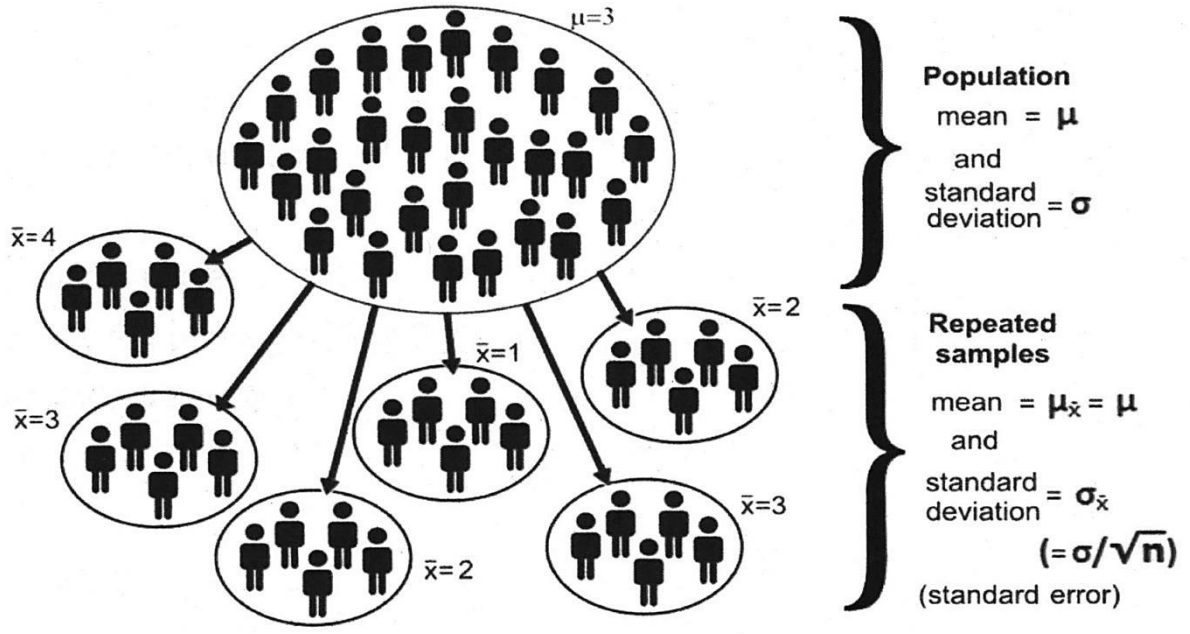
[3] This section is particularly well explained for the non-specialist by UNISA's Professors Kruger & Janeke (2012). We have been guided by their example.

little x bar subscript, $\overline{x}$ meanwhile, signifies that we are referring to a population of sample means as opposed to sigma, σ without a subscript which refers to the standard deviation of the population. Thus:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

According to this formula one should be able to diminish the standard error by reducing σ and increasing n; however because the standard deviation, σ is a population parameter we seldom have control or even knowledge of the parameters of large populations, unless like I.Q. we rig them that way. What we do have control of is the number, n that we include in our samples; but because the formula takes the square root of the number n, we have to include 4x as many the number in our sample just to mitigate the standard error by half. However see the note in parentheses in the first bullet above. Also, it is not necessary or practicable to take ever larger and larger samples when, a) we do not need to have confidence in some finding beyond a certain level of significance and, b) we could be could be sampling "smarter" rather than just more numerously. (See Critical Reasoning 15 on Statistical Hypothesis Testing.)

For now let us revise graphically what we have tried to explain verbally. Professors Kruger & Janeke of UNISA have included the following diagram in their study guide for the 3rd level, undergraduate course in "Psychological Research", presented by the Department of Psychology.



Samples and sample means in relation to a population ©UNISA 2012

The relatively large set of figures in the upper oval represents the parent population. This population is described by its parameters, the mean μ and standard deviation σ, right of curly bracket. The smaller ovals below represent, only a fraction of, all the repeated samples to be drawn from parent population, five at a time (n = 5). Note that not every sample has the same sample mean, $\overline{x}$ but if we were to take the mean of all the possible sample means, $\mu_{\overline{x}}$ it would be numerically equal to the population mean, μ. Furthermore, the standard deviation of all the sampling means, $\sigma_{\overline{x}}$ would be

given by the quantity $\sigma/\sqrt{n}$, according to the Central Limit Theorem. This then is what is otherwise known as standard error. See right of the curly bracket, bottom of the diagram.

Examples

The following example, which requires knowledge of both the z-distribution and the CLT, is selected form Stefan Waner and Steven R. Costenoble's (1998) *Sampling Distributions & The Central Limit Theorem Miscellaneous on-line Topics for Finite Mathematics.* (Available here)

3. A lightbulb manufacturer claims that the lifespan of its lightbulbs has a mean of 54 months and a standard deviation of 6 months. Your consumer advocacy group tests 50 of them. Assuming the manufacturer's claims are true, what is the probability that it finds a mean lifetime of less than 52 months?

Solution: The quantity we seek is the probability that the mean lifetime, $\overline{x}$ of bulbs manufactured by this company is 52 months or less. In symbols, $p(\overline{x} \leq 52)$. According to the CLT, $\overline{x}$ will have an approximately normal distribution with a mean of μ = 54 months and standard error of:

$$\sigma_{\overline{x}} = \frac{6}{\sqrt{50}} \approx 0.85 \text{ months}$$

To find the probability, we have to convert these values into z-scores. Therefore we let:

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{52-54}{0.85} \approx -2.35 \text{ standard deviations}$$

Now since this is the lesser proportion, we consult the first z-table to find the corresponding p-value, which is:

p = 0.0094

So the probability of this happening is 0.0094 or 0.94%. Alternatively, we can be 100% - 0.94% = 99.06% certain that this won't be the case (if the manufacturer's claim is true!)

This last example emphasises another point or two so please take a look:

4. Suppose Mr. Anderson's school gives you consent (in the interests of research) to examine the permanent records of some of his Grade 7 students. You take a sample of 16 cards from among 25 of his students'. You are surprised to discover that the mean I.Q. of these students is exactly 110 points. Given that I.Q. tests are so designed as to produce a mean test score of 100 and a standard deviation of 15 points, what is the probability of the sample of Mr. Anderson's students scoring better than the mean score that they did? Put your result into perspective. Do you suppose it was a matter of chance?

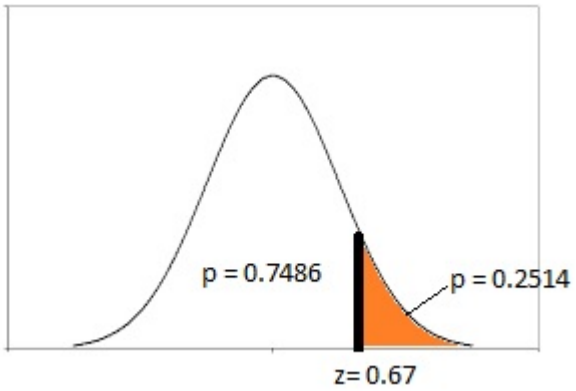
Solution: The quantity that we seek is $p(\overline{x} > 110)$. However, this time we do not, indeed must not, use the standard error when we *already know* the population mean μ = 100 and standard deviation σ = 15. What we still need to do is to convert our raw score into a z-value, therefore:

$$z = \frac{\overline{x} - \mu}{\sigma} = \frac{110-100}{15} = 0.67 \text{ standard deviations}$$

Since this represents the larger portion, we must consult the second z-table here to find the corresponding p-value, which is:

p = 0.7486

But this is the probability that people's I.Q. scores will fall within the interval *up to and including* 110. What we were asked is the probability of the sample having a mean score better than 110. Therefore we have to look at the area under the curve to the right of the z-score of 0.67. The easiest way to calculate this is to simply subtract our p-value from 1, because remember, the entire area under the normal distribution is 1. And so:



A rough sketch of the quantities involved is often helpful to visualise the problem.

1 - 0.7486 = 0.2514

which is the probability of the mean of a sample of students scoring above 110 I.Q. points.

This is modestly improbable, because after all, a score of 110 I.Q. points is already above "average" but not above one standard deviation to the right of 115 points. Just being in school puts one at a distinct advantage in the way intelligence tests are structured. Moreover, schools actively recruit candidates with higher intelligence based on performance, though in the public system, they are not permitted to turn a student away based *solely* on their I.Q. Therefore the mean value of 110 points obtained is probably not due to chance (alone).

Conclusion

Although the point of the above examples is to demonstrate how distributions can be standardised as well as the consequences of the CLT, you will not be required to do any complex calculations in an undergraduate Social Science and Humanities examination. Means and standard deviations will usually be supplied because they are time consuming to calculate. Most examination papers will also come with a detachable list of formulae. What you will be asked in most Research Methodology modules is what a result means in real terms, what confidence you should ascribe to it and how it was derived.

A key skill in this area is learning how to read, and judge the merit of a scientific article. What statistical evidence is being presented? What does it mean? How does it support the conclusion? And are the appropriate tests being used? Although we have only met one such test so far, there are several more to follow.

Finally, most Social Science and Humanities programs will require you to carry out some project towards your major that involves the collection, assimilation and interpretation of data, such as responses to questionnaires or basic anthropometry. This is where statistical skills are paramount. It is one thing to enter your data into a spreadsheet program. It is quite another to know what

functions to select and what the output actually means, and how you should include it in your project report or "mini-thesis".

Task

By way of revision, it is useful to make sure you understood the meaning of the following table of quantities used in this study unit. One is included that you may have to look up. Why are some written with Greek letters while others are written using the Roman alphabet? Why do some have subscripts (or superscripts,) while others don't? In which case, what do such scripts denote?

μ	$\mu_{\overline{x}}$	$\overline{x}$
σ	$\sigma_{\overline{x}}$	s
σ^2	$\sigma_{\overline{x}}{}^2$	s^2
z	p	n

Feedback

The quantities written in Greek letters refer to population parameters that represent some feature of a population such as its mean (μ) or standard deviation (σ). Mostly such quantities are unknown but can be estimated. Quantities written using the Roman alphabet refer to statistics which are measurable features of finite samples such as the mean ($\overline{x}$) or standard deviation (s). Variance, σ^2 and s^2, recall, is simply the square of the standard deviation, or the value displayed on your calculator <u>just before</u> you hit the final square root button when calculating standard deviation. Several mathematical theorems are stated in terms of variance rather than standard deviation, therefore we have retained the notation.

The variables: $\mu_{\overline{x}}$, $\sigma_{\overline{x}}$ and $\sigma_{\overline{x}}{}^2$ are derived by sampling a population in the peculiar fashion described above; counting every possible sample of a population, with replacement, so many (n) at a time. Because the entire population is sampled in this way, they are regarded as parameters; however the x bar subscripts ($\overline{x}$) indicate that they are, nevertheless, samples. $\mu_{\overline{x}}$ represents the sampling mean; $\sigma_{\overline{x}}$ the standard error of the sampling mean or standard error, whereas $\sigma_{\overline{x}}{}^2$ is simply the square of this quantity, though we shall not be concerned with it here.

z is a score (positive or negative) measured in standard deviations, that a raw score is above the mean of a normal distribution. p meanwhile, is just a probability, expressed as a decimal between 0 for impossible to 1 for certainty. Each z-score is associated with a particular probability that has been calculated for us in the link provided. On the other hand, not every p-value is associated with a z-score. We saw several examples of this in Critical Reasoning 10. Finally, n of course is just the number of a population or the number drawn as a sample. Since we are unlikely ever to be unclear about which, we don't have to bother with a separate nu (ν) and n.

If you were unable to identify one or more of the variables above, please do not rely solely on the this feedback. Rather, go back into the text and find the definition in context.

The next Critical Reasoning Study Unit will cover the Logic of Relations.

References:

KRUGER, P. & JANEKE, H. C. (2012) *Psychological Research - Study Guide for PYC3704*. UNISA Department of Psychology

Z- tables originally provided by the University of Florida. Available on our server at: http://philosophy.org.za/uploads_other/Ztables.pdf