Statistics for the Surgeon

Dr. Ajeesh Sankaran,

KIMS Al Shifa Hospital, Kerala

Men may construe things after their fashion,

Clean from the purpose of the things themselves

- Cicero in William Shakespeare's *Julius Caesar*

Basic statistical concepts and techniques have been discussed already in a prior ISSH Article of the Month. This article is focused on the practical aspects of statistics. Mostly it deals with understanding why statistical testing is necessary. Then we move on to how to perform these analyses with the help of some (free!) resources detailed at the end.

This article may seem rather stretched out and lengthy, but that is not an attempt to break the record for the longest ISSH Article. Instead, it is an honest effort to cover all the aspects of statistics that would be essential to see any study through. Since surgeons are not exactly famed for their mathematical prowess, this involves fairly long tracts of dreary prose. So interspersed are 'Time Outs' in dark boxes, which are short vignettes intended to illustrate a point that needs deep thought. A quick revision of the earlier statistics article is definitely recommended.

1. Understanding Probability

We start with the most abused of all examples in statistics – the toss of a coin. When a fair coin is tossed, it is nearly impossible to predict *exactly* which face it would land on. A large number of factors are at play here, which cannot be solved for exactly. However, we know that only two outcomes are possible - Heads [H] or Tails [T]. We can also understand that both these outcomes should be equally likely. In other words, the coin should land H or T about half the number of times. This is the traditional sense that probability is understood in, that H has a probability of 1/2. Probability of the event 'H' can be denoted by p[H]. We can then straight away express the coin toss situation as follows:

$$p[H] = ½ = p[T]$$

The above expression makes it clear that i) there are two events and two events only, ii) both are equally likely.

Let us move on to another battered example – throw of a die. The usual die is a cube, with six faces numbered from 1 through 6. Using the same arguments as above, we can conclude that each number has a 1/6 probability of turning up. We can express this as

$$p[1] = p[2] = p[3] = p[4] = p[5] = p[6] = 1/6$$

Probability of an event is not always so easy to interpret. For e.g., many weather services provide an estimate of the likelihood of rain/ snow over a time period (please search Google Weather and it may provide information like "60% chance of precipitation tonight"). What do these percentage figures mean? '50% rain' or '60% rain' obviously has no meaning. One way to look at it is in the context of repeated events. If we toss a coin 100 times, we expect 'H' about 50 times. So, when we say "50% chance of snow tonight", we can understand that of 100 nights with similar weather conditions, 50 nights can be expected to be snowy.

This view of probability in the context of repeated events or measurements is an extremely useful one in statistics. When we expect 50 Heads out of a 100 coin tosses, we also have to accept that often we may not get *exactly* 50 heads. So, when we get 55 heads out of 100 throws, can we confidently conclude that the coin is fair or unfair? The answer to this question is the crux of all inferential statistics.

2. The Need for Statistical Testing

After any data is compiled, it is obviously useful to visualise it in the form of graphs/ charts etc or summarise it in the form of a small set of numbers. As the amount of data increases, it becomes more and more difficult for the human mind to grasp any meaning out of the data set. When the same set is reduced to a Mean and Standard Deviation, for example, we have immediately gained an easy understanding of the entire set. Such 'Descriptive Statistics' are basically summary figures of two aspects of data – what is an average or representative value of this data and how far are the other data points from this representative value. The first category of values are the *measures of central tendency* and the second category *measures of dispersion*. Put together they attempt to give a snapshot of all the data that we have. These also allow us to compare two or more related data sets. For e.g., we can compare the Mean heights of Boys and Girls of a certain age and conclude which group is taller.

While the utility of descriptive statistics is easily appreciated, the need for the other major area of 'Inferential Statistics' is not so easily understood. Inferential Statistics deals with Hypothesis Testing, basically telling us whether the hypothesis that we have based any study upon is to be accepted or not. Going back to the earlier example of 100 coin tosses, what should our conclusion be about the 'fairness' of the coin if we land 55 Heads? In fact, we can calculate that the probability of 55 Heads in 100 tosses with a fair coin is 4.8% (see Appendix 5). This is sufficiently low to raise our suspicions. With 60 Heads in 100, the probability is only 1%. So, while 60 Heads is not impossible, we would be justified in thinking that there is some problem with the coin that makes it more likely to land Heads.

THE IMPROBABLE WORLD

The game of roulette at the Monte Carlo casino is played on a wheel with alternating black and red slots. Bets are placed on which slot the ball would land in after the wheel is set in motion. On August 18, 1913, the ball fell into black slots 26 times in a row. Such an event has a probability of only **1 in 136 million**. Gamblers bet against black in increasing amounts believing that a red was due. Their losses totaled millions of francs in 1913 terms! This is a popular illustration of the Gambler's Ruin or Fallacy. Basically, we are instinctively programmed to believe that somehow the probability of getting a red should be higher after a string of blacks. However, it is important to remind ourselves that each wheel spin is independent of all others before it and, hence, the probability is the same for every throw.

Cut to August 6, 1945, Hiroshima. Tsutomu Yamaguchi, an engineer with Mitsubishi, was preparing to return home after a business trip when the first nuclear bomb in history fell 3kms away. Having survived with radiation burns and temporary blindness, he returned to his hometown in, where else, Nagasaki. The resolute Yamaguchi returned to work on August 9, in time to be hit by the last nuclear bomb in history. Yamaguchi agonized through radiation sickness for weeks but survived to be officially honoured by the Japanese government in 2009 as a survivor of both bomb blasts. Moral of the story: Improbable events are not impossible.

3. Understanding Hypothesis Testing

For this section, let us design an experiment with Red pills and Blue pills. We have before us a very large number of pills which are either Red or Blue. This constitutes the *population* of the study. Out of this, we begin by choosing 10 pills at random. This represents our *sample*. We observe that we have 6 Red and 4 Blue pills.

3.1. The Null Hypothesis

We assume that there are an equal number of Red and Blue pills in the study population. This means that there is an equal probability of choosing either one of them. We can represent this by

$$p[R] = p[B] = \tfrac{1}{2}$$

The Null Hypothesis is just this assumption that there is no difference between these groups, in this case meaning an equal probability of Red and Blue pills. In clinical studies, the Null Hypothesis would postulate that there is no difference between any of the groups involved. All our testing is then directed to either accept or reject this Null Hypothesis. Designating the Null Hypothesis by $\{\mathbf{H_0}\}$, for our Red/Blue pill study we have

$$\{\mathbf{H_0}\} \equiv p[R] = p[B]\ (=1/2)$$

3.2. The Alternate Hypothesis

Usually, the opposite of the Null Hypothesis, the Alternate Hypothesis mostly postulates that there is a difference between the groups. If we designate this by **{H₁}**, we can express this for our study as

$$\{\mathbf{H_1}\} \equiv p[R] \neq p[B]$$

This expression tells us simply that the Alternate Hypothesis states that the probability of picking a Red pill is different from that of a Blue pill, implying that they are not equal in number in our population.

A word of caution: setting up the Null and Alternate Hypotheses may look like overstating the obvious. But having them clearly defined is a necessary and important step for all statistical analysis. In large or multi-faceted studies, this can be a source of much confusion, ending up with coming to improper conclusions regarding the results of the study.

3.3. Hypothesis Testing

In our study, the Null Hypothesis basically states that there an equal number of Red and Blue pills in the population. Our sample of 10 pills contained 6 Red and 4 Blue pills. Hypothesis testing gives us the probability of finding such data, **given that the Null Hypothesis is true**. This probability is what is designated as the '*p – value*'. When this probability is sufficiently low, we would be justified in rejecting the Null Hypothesis.

For our example, the Null Hypothesis requires p[R] = ½. Given this value, it is possible to calculate the probability of getting exactly 6 Reds out of 10. Let's represent that by p[6R, 4B]. Some mathematics leads us to,

$$p[6R, 4B] = 0.2051$$

So even if there were an equal number of Red and Blue pills in the population, there is a good 20.5% chance of drawing 6 Red ones in 10. Table 1 depicts the entire spectrum of possibilities in drawing 10 pills and their associated probabilities, *given that the Null Hypothesis is true*.

0 Reds, 10 Blues	p[0R, 10B]	0.0009
1 Red, 9 Blues	p[1R, 9B]	0.0098
2 Reds, 8 Blues	p[2R, 8B]	0.0439
3 Reds, 7 Blues	p[3R, 7B]	0.1172
4 Reds, 6 Blues	p[4R, 6B]	0.2051
5 Reds, 5 Blues	p[5R, 5B]	0.2461
6 Reds, 4 Blues	p[6R, 4B]	0.2051
7 Reds, 3 Blues	p[7R, 3B]	0.1172

8 Reds, 2 Blues	p[8R, 2B]	0.0439
9 Reds, 1 Blue	p[9R, 1B]	0.0098
10 Reds, 0 Blue	p[10R, 0B]	0.0009

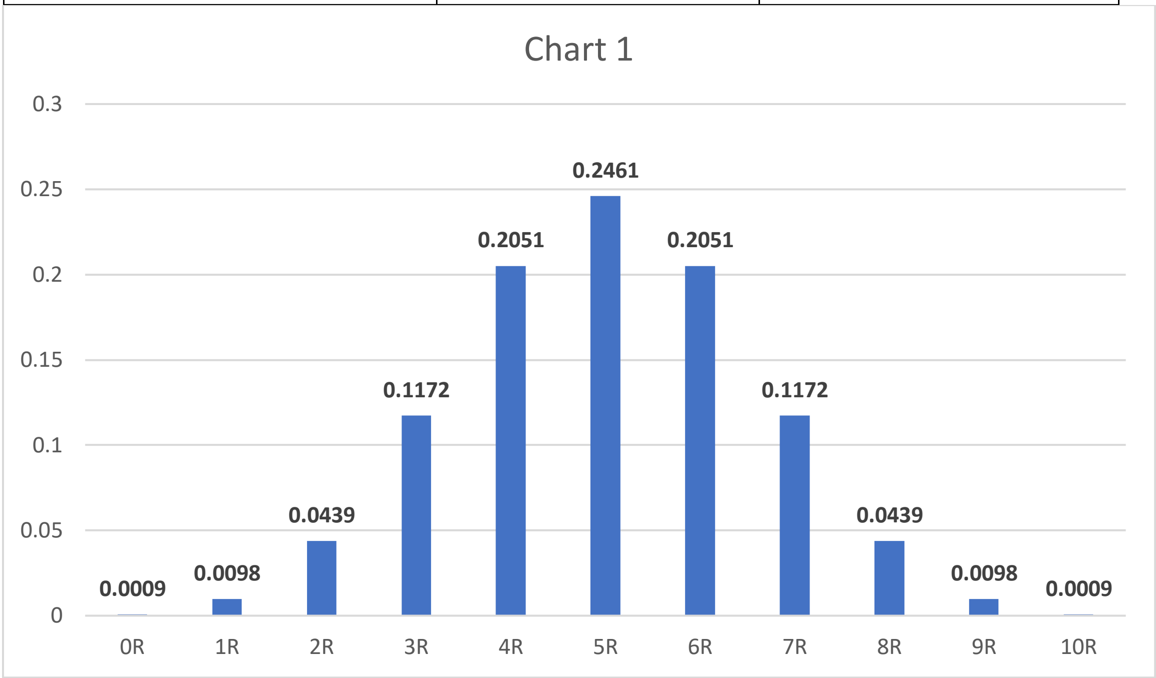


Figure 1 depicts these probabilities in a bar chart. While the symmetry is visually striking, a sobering feature is the probability of drawing 5 Reds. Even when there are an equal number of Red pills and Blue pills in the population, there is less than 25% chance of getting *exactly* 5 Reds out of 10 pills! This non intuitive aspect of the nature of probability underscores why hypothesis testing is essential.

3.4 The 0.05

We have already concluded that the Null Hypothesis is to be rejected when the p-value is sufficiently low. But *how* low is 'sufficiently low'?

Let us set up a thought experiment, with no more than a coin and generous imagination. We toss the coin and find Tails. We continue tossing and find Tails *every time*. Most people would begin to feel there is something wrong with the coin by the 4th or 5th consecutive Tails. The probability for 4 consecutive Tails with a fair coin is 6.25% and for 5 Tails is 3.25%. So, 5% or p=0.05 is certainly a value that most would consider quite unlikely. Ronald Fisher, considered the most important statistician of the 20th century, introduced the p=0.05 as a cut off value in his 1925 publication *Statistical Methods for Research Workers.* It is safe to say that the stature of Fisher was enough to deeply entrench this value in research publications. Apart from this, there is very little rationale for this specific value of 0.05. As one statistician put it, "*Surely God loves the 0.06 as much as the 0.05*". However, in the real world, good luck pushing past journal editors any article with the cut off placed at any other value!

Applying the p=0.05 cut off to Table 1 [and Figure 1], we see that the Null Hypothesis is to be accepted for Red pills between 3 and 7. For the other values, we would have to conclude that the Alternate Hypothesis is to be accepted. In other words, we would infer that the Red pills and Blue pills are not equal in the population. If we find 0, 1 or 2 Red pills out of 10, we would conclude there are more Blue pills. With 8,9 or 10 Reds, we conclude there are more Red pills in the population.

0 Reds, 10 Blues	p[0R, 10B]	0.0009	**{H_1}** **p[B]>p[R]**
1 Red, 9 Blues	p[1R, 9B]	0.0098	
2 Reds, 8 Blues	p[2R, 8B]	0.0439	
3 Reds, 7 Blues	p[3R, 7B]	0.1172	**{H_0}** **p[R] = p[B]**
4 Reds, 6 Blues	p[4R, 6B]	0.2051	
5 Reds, 5 Blues	p[5R, 5B]	0.2461	
6 Reds, 4 Blues	p[6R, 4B]	0.2051	
7 Reds, 3 Blues	p[7R, 3B]	0.1172	
8 Reds, 2 Blues	p[8R, 2B]	0.0439	**{H_1}** **p[R]>p[B]**
9 Reds, 1 Blue	p[9R, 1B]	0.0098	
10 Reds, 0 Blue	p[10R, 0B]	0.0009	

Table 2.

3.5 Probability Distributions

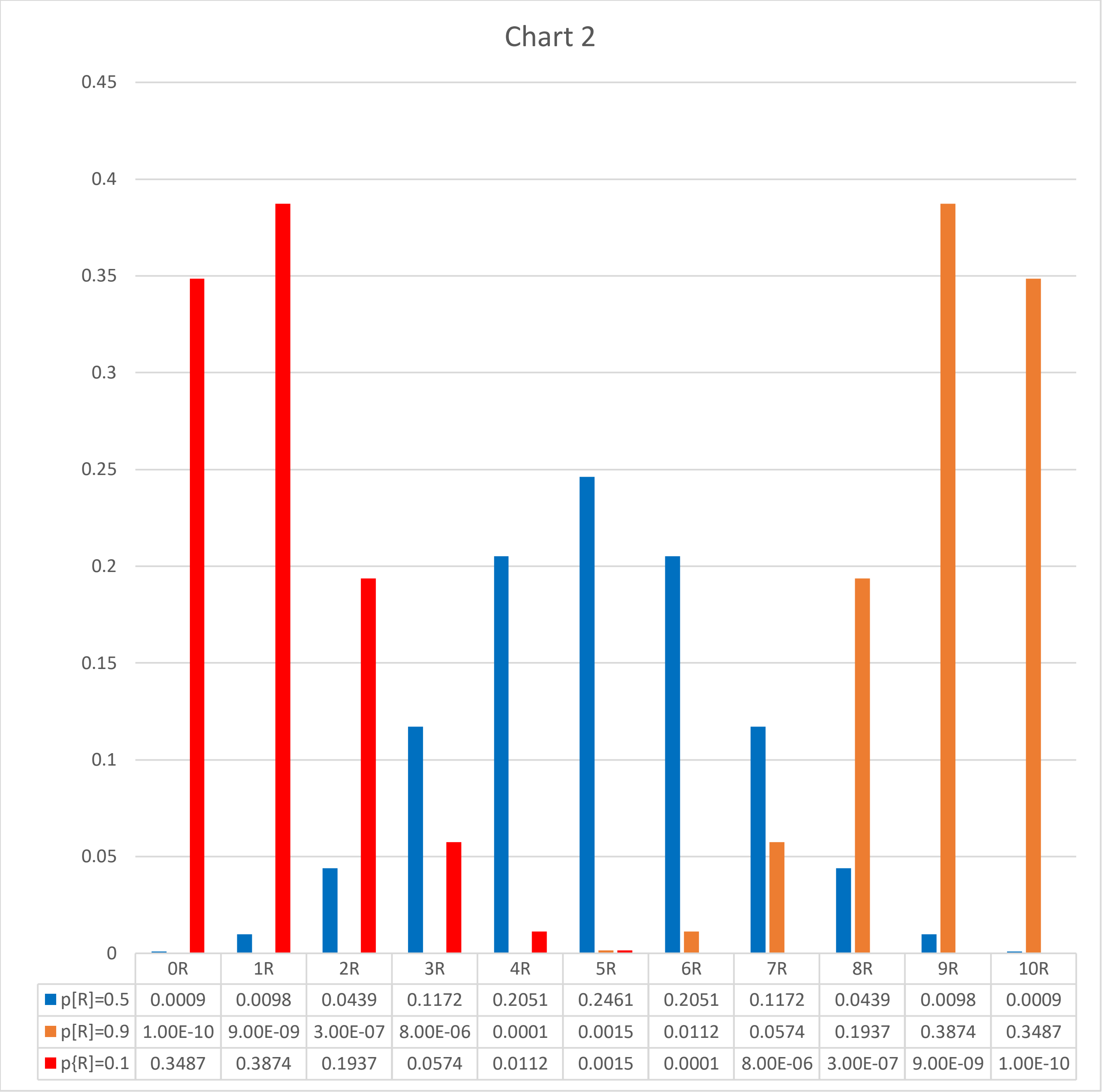
Revisiting Chart 1, we remind ourselves that the chart shows *all* possibilities and the probability associated with *each* one of them. For the situation of equal probabilities and a sample of 10 pills, there is nothing more to know. As expected, all the probabilities add up to 1(apart from rounding error). The chart then is said to constitute a probability distribution. Specifically, Chart 1 is an example of a Binomial Distribution.

Chart 2 shows three such distributions for a sample of 10 pills. They differ in the probabilities p[R] and p[B]. The three sets of probabilities are

$$p[R] = p[B] = o.5$$

$$p[R] = 0.9,\ p[B] = o.1$$

$$p[R] = 0.1,\ p[B] = o.9$$



	0R	1R	2R	3R	4R	5R	6R	7R	8R	9R	10R
🟦 p[R]=0.5	0.0009	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0009
🟧 p[R]=0.9	1.00E-10	9.00E-09	3.00E-07	8.00E-06	0.0001	0.0015	0.0112	0.0574	0.1937	0.3874	0.3487
🟥 p{R]=0.1	0.3487	0.3874	0.1937	0.0574	0.0112	0.0015	0.0001	8.00E-06	3.00E-07	9.00E-09	1.00E-10

Figure 2.

This chart once again underscores how important it is to be wary while drawing conclusions from our findings. Even when the population of pills contains 90% Reds, the chances of drawing exactly that proportion out of a sample of 10 pills is just 38.74%.

3.6 Sample Sizes

In the above discussion we varied the relative number of the pills and their probabilities, but what if we vary the sample size? We are in no way limited to 10 and can easily extend the same analysis to larger samples. We assume the same Null Hypothesis that the Red and Blue pills are equal in number in the population. To reiterate,

$$\mathbf{\{H_0\}} \equiv p[R] = p[B] \; (=1/2)$$

$$\mathbf{\{H_1\}} \equiv p[R] \neq p[B]$$

(By now, the above statements should make clear sense. If they do not, it would not be a bad idea to restart Section 3!)

Let us now consider a sample of 20 pills. We construct a probability distribution in the same way as Chart 1, by basically enumerating all the possible outcomes and calculating their probabilities.

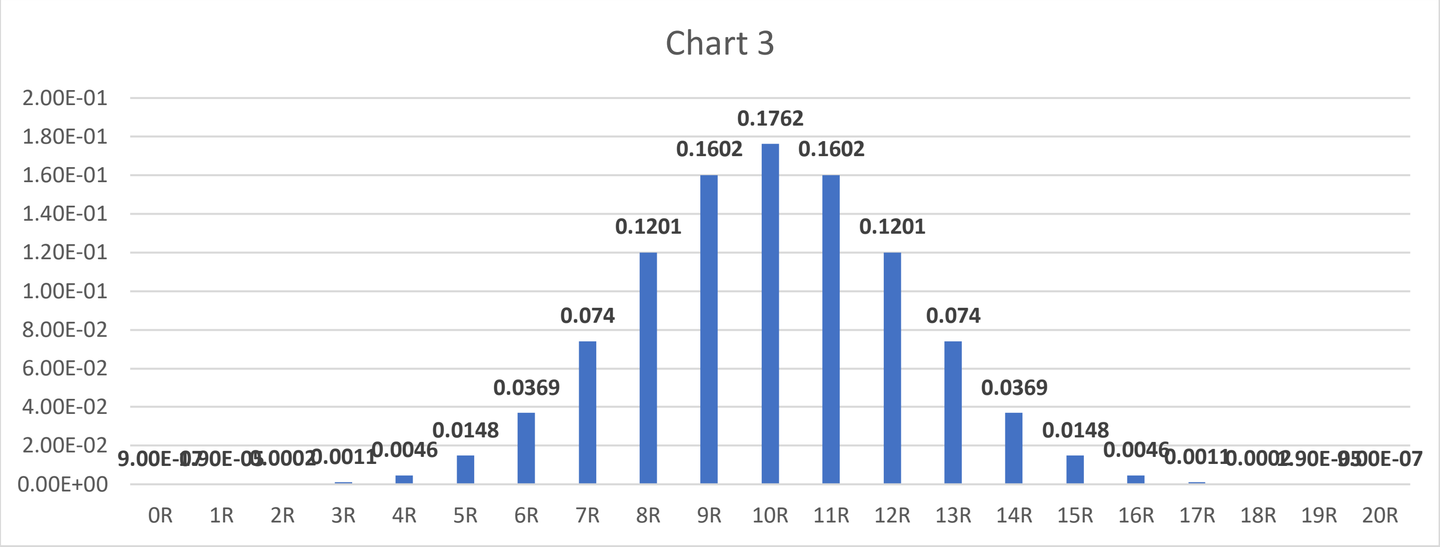


Figure 3.

This distribution looks, of course, very similar to Chart 1. But the interesting aspect is when we apply the 0.05 cut-off value to accept or reject **{H₀}**. The above chart shows that we would have to accept **{H₀}** if our study finds 7 to 13 Red pills, since the probabilities of these are above 0.05. For all others, probabilities are lower than 0.05 and we would have to reject **{H₀}**. Recall from Table 2, that the Null Hypothesis for a 10 sample would have to be accepted for finding 3 to 7 Red pills. When the sample size is increased to 20, the range of acceptance of the Null Hypothesis seems to have narrowed.

Building on that theme, let us look at greater sample sizes and cut-off values for **{H₀}** acceptance/ rejection. Table 3 shows these values for sample sizes 10, 20, 50, 100 and 1000.

Sample Size	Accept {H₀}	Reject {H₀}
10	3 Reds to 7 Reds	0 to 2 Reds + 8 to 10 Reds
20	7 Reds to 13 Reds	0 to 6 Reds + 14 to 20 Reds
50	21 Reds to 29 Reds	0 to 20 Reds + 30 to 50 Reds
100	46 Reds to 54 Reds	0 to 45 Reds + 55 to 100 Reds
1000	472 Reds to 528 Reds	0 to 471 Reds + 529 to 1000 Reds

Clearly, there is a tendency for the range of acceptance of the Null Hypothesis to get narrower as the sample size gets larger. In other words, we are more likely to reject the Null Hypothesis as the sample size increases. In research studies with a small sample, therefore, a large difference would be necessary to reject the Null Hypothesis. Even if an intervention did produce a difference, the magnitude of this difference would have to be large for us to be able to prove it statistically. A larger sample size makes it easier to demonstrate this difference. However, there are numerous practical difficulties in increasing the sample size. Calculating an adequate sample size is, therefore, of great importance and will be dealt with later.

3.7 Putting it all together

A lot seems to have happened in this section. To summarise, we conjured up a world populated by pills of the Red and Blue varieties. We assumed that they were equal in number and designated it as the Null Hypothesis. This leads to the assumption that there is an equal probability of picking the Red or Blue pills out of this population. We picked a sample of 10 pills and found 6 Reds. Based on some concealed mathematics, we realised that this is quite possible. Finally, we demonstrated that with smaller sample sizes even large differences may not be 'statistically significant'. This significance was arrived upon the basis of the p-value, which is the probability of finding the observations that we did, given that the H_0 is true. It is important to understand that the p-value is **not** the probability of the H_0 being true. Hypotheses do not have probabilities attached to them, but data sets do (more in Appendix 4).

This is all very well, but how does all this translate to clinical research? When we start a research project, we have no information about the population involved. Till now we have basically discussed the behaviour of a sample from a *known* population. In actual research, this procedure has to be reversed. Our sample provides some estimates, that we intend to extrapolate to the population. When we found 6 Reds out of 10, we had to conclude that we cannot confidently declare there are more Reds in the population. Had we found 600 out of 1000, however, we would be confident that there have to be more Reds than Blues in the population [refer Table 3- 600 Reds is in the Rejection area for $\{H_0\}$]. This entire process is the core of statistical analysis.

In a nutshell, say a new procedure is performed in 10 patients with 6 'Good' and 4 'Bad' outcomes can we conclude the procedure is useful? No. Certainly, not yet. Sadly, for most clinical studies outcomes are not so simply defined. Usual outcome measures are *continuous* data, where a large number of values are possible. Obviously handling them statistically involves more complex mathematics. But the principles and ideas set up above are more or less still applicable.

GETTING IT DEAD RIGHT

In a study by Bennett *et al*,* the researchers performed fMRI of a subject at the time of presenting it with some pictures of humans in social situations. They found active signals in the brain cavity and dorsal column of this subject. However, there was a twist to this study, which is quoted in the Methods section of their paper – "One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female but given the post-mortem state of the subject this was not thought to be a critical variable".

This article was not entirely in jest. Image acquisition in MRI involves interpreting signals for significance. The researchers illustrated how false positive signals are almost guaranteed when a very large sample is analyzed, since they had studied 130,000 voxels. As is customary in fMRI circles, the p-value had been set at 0.001 or 50 times stricter than the usual 0.05! They concluded that statistical methods need to be improved to avoid such errors. The group was awarded the 2012 Ig Nobel Prize in Neuroscience 'for demonstrating that brain researchers, by using complicated instruments and simple statistics, can see meaningful brain activity anywhere'.

4. The Normal Distribution

In the prior sections, we were introduced to probability distributions. To put it simply, they plots the probability associated with all possible outcomes. For discrete values the plots are similar to bar charts, as above. However, for *continuous* data they are smooth curves – the most famous and well known being **the bell curve**.

While the term 'bell curve' refers to the shape of the curve, mathematically it represents **the normal distribution**. The form that this curve takes is defined by a specific equation, similar to the equations for a line or a circle that many may be familiar with from school. For any value of a variable, the equation gives an associated frequency of its occurrence. Figure 4 shows a typical normal distribution, with Mean of 2 and Standard Deviation of 0.5. An immediate striking feature of note is its symmetry, with half the values to the left of the value '2' and the other half to the right. By definition, 2 is also the Median. Since the peak of the curve is also at 2, that makes it the Mode as well. Therefore, the elegant symmetry of the normal distribution means that its Mean, Median and Mode are all equal.

The distribution is also defined by its standard deviation, for Figure 4 this being 0.5. Let us now focus on the portion of the graph one SD to either side of the Mean *i.e.,* between 1.5 and 2.5. It is obvious that this portion of the graph contains most of the *total* area under the graph. You can go ahead and count the small squares contained

*Bennett CM, Baird AA, Miller MB, Wolford GL. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. J Serendipitous Unexpected Results 2009; 1: 1–5.

in this portion as a measure of area. Then let us sit back and widen our focus area to two SDs on either side of the Mean, between 1 and 3. Here it appears that almost the *entire* area under the graph is encapsulated within this portion. There is very little remaining in the portions of the graph to the left of 1 and to the right of 3.

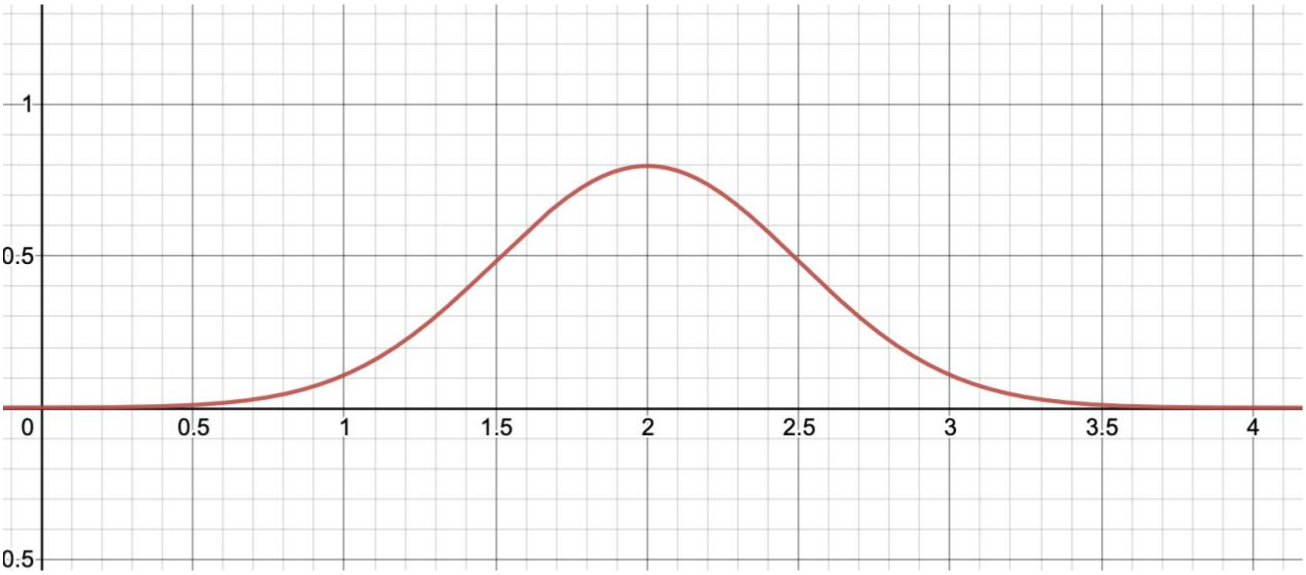
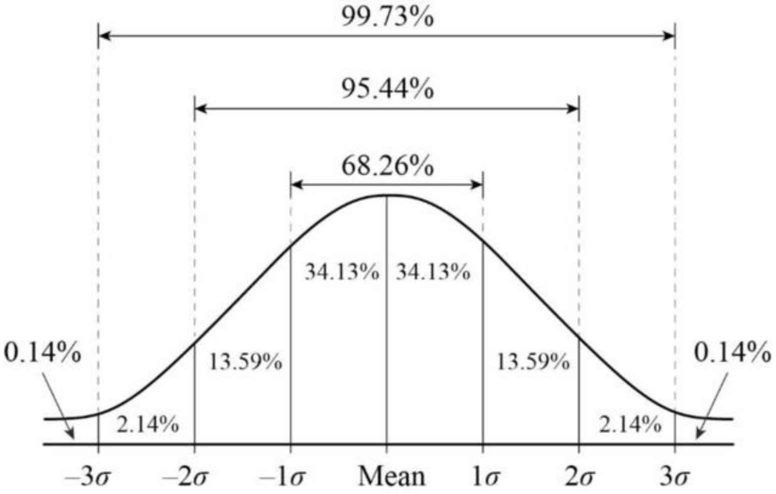


Figure 4. Normal distribution with Mean 2 and SD of 0.5. Counting the small squares under the graph is an approximate method to estimate the area under the graph.

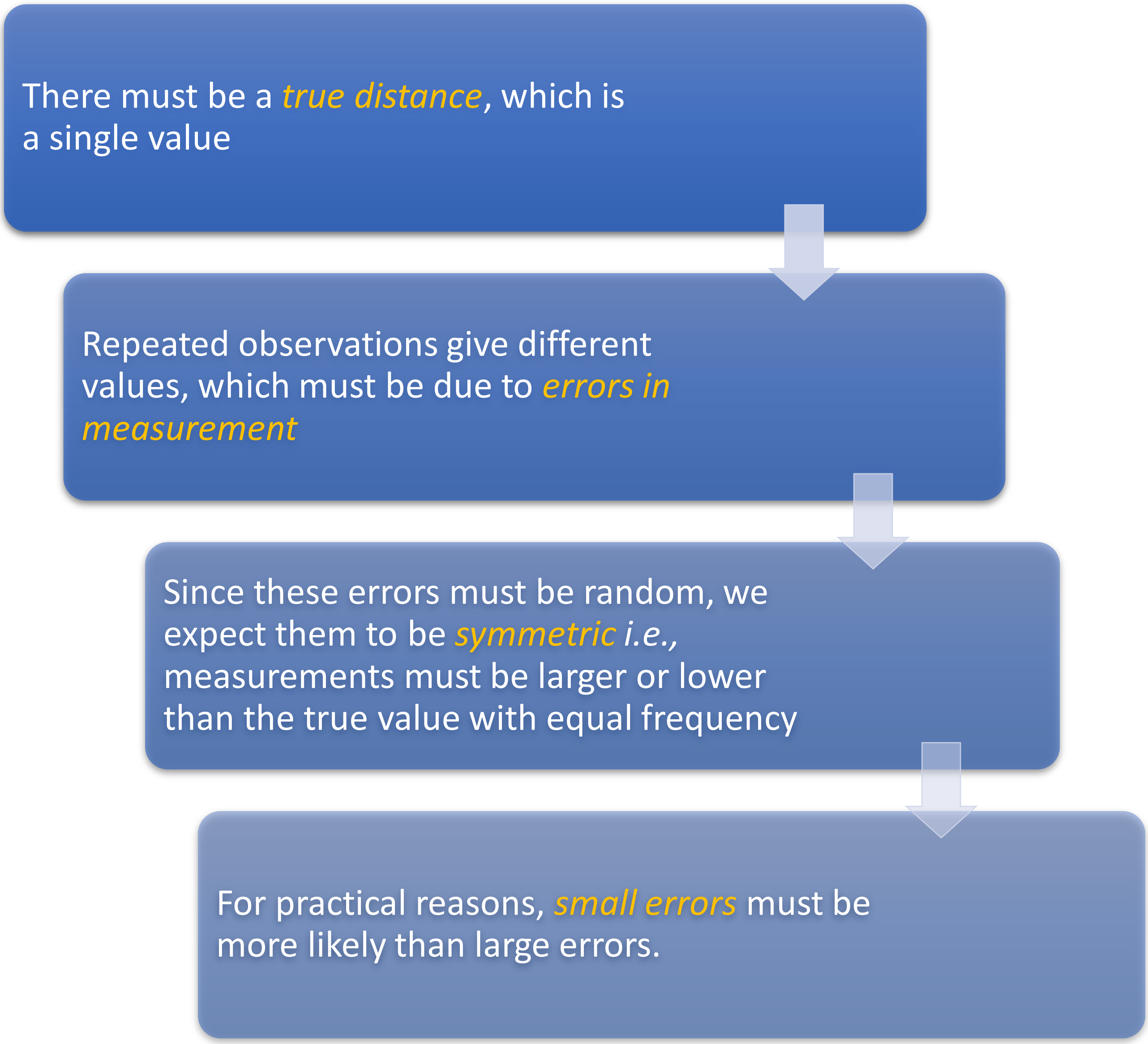
If you actually had the patience and perseverance to count the small squares in Figure 4, it would be possible to calculate the areas under different portions of the graph. Most would find ~70% of the squares in the interval of Mean ± SD, and ~95% in the interval Mean ± 2SD. Fortunately, with the equation of the normal distribution, mathematicians can calculate exactly these areas. These are illustrated in Fig 5 below. The area contained in each segment is provided in percentage values.



4.1 Why so Normal?

Carl Freidrich Gauss, considered the greatest mathematician in history, is often credited for introducing the Normal Distribution. Like so often in science, others like de Moivre and Laplace also made significant contributions prior to Gauss. Needless to say, the

Normal Distribution and its properties are extremely difficult to comprehend *non-mathematically*. A bit of historical perspective might go a long way in understanding its importance in statistics. We pick up the thread with Galileo in 1632, trying to make sense of measuring astronomical distances. His reasoning went as follows:



There must be a *true distance*, which is a single value

Repeated observations give different values, which must be due to *errors in measurement*

Since these errors must be random, we expect them to be *symmetric i.e.,* measurements must be larger or lower than the true value with equal frequency

For practical reasons, *small errors* must be more likely than large errors.

The last couple of assumptions basically lead us to conclude that most of our measurements must be close to the true value.

In the early 18th century, the equation for the normal distribution first makes an appearance from the work of Abraham de Moivre. Being a lifelong monetarily challenged mathematician, he supported himself as a consultant to gamblers at a local den. This led him to work on the Binomial distribution, which we have seen in our Red/Blue Pill example. He worked out that for large numbers, the Binomial Distribution is approximated by the Normal Distribution [please compare the shapes of Figures 3 and 4]. However, in the statistical sense, the Normal Distribution appears in history from a stupendous feat of applied mathematics from Carl Gauss.

On 1 January of 1801, Guiseppe Piazzi of Palermo noticed a new celestial object which he considered a dwarf planet and named it Ceres. He was able to observe and make positional measurements for only 6 weeks, before it was lost in the Sun's glare. These measurements were insufficient to predict where Ceres would reappear in a about a year or so. Enter Gauss, then 24 years old and not yet 'The Great Gauss' as he was later named. Expanding on basically the same ideas as Galileo, with a generous dose of mathematical muscle applied to Piazzi's measurements, Gauss predicted the possible location of Ceres' reappearance. Many astronomers also made predictions of different locations of the sky [Level V evidence!]. On 31 December that year, Gauss was proved right. He had used the, now standard, least squares method on errors to find the most probable orbit. These errors were proved to be normally distributed, which he then used to make the prediction. It is instructive to note how this episode closely resembles any modern research – a set of planetary observations [the sample] was used to make valid predictions on the orbit [the population]. A further development was the Central Limit Theorem, formalised by Laplace in 1810.

The credit for bringing the Normal Distribution to the life sciences, from just a theory of errors, goes to Adolphe Quetelet. In 1846, he contended that the recently published Chest Girth measurements of 5738 Scottish soldiers were normally distributed. Many prominent scientists of the era picked up this assertion and a slew of measurements appeared to prove that many biological characteristics were near normally distributed. Quetelet is now mostly remembered for the Quetelet index, better known by the term Body Mass Index [BMI].

So why do so many characteristics, like height or weight, appear to be approximately normally distributed? Observational or measurement errors are mostly random errors with a large number of factors affecting their magnitude. Similarly, many biological characters are affected by a large number of factors which are all not known. So, both these situations can be expected to have similar mathematical behaviour. It is no surprise that the Normal Distribution turns up in the world around us so frequently. It is defined to fit in such situations.

4.2 The Central Limit Theorem

The importance of the Normal Distribution comes not only from its supposed ubiquitous presence, but also from its gratifying appearance in the Central Limit Theorem [CLT]. To help understand the Theorem, we go back to the Red and Blue pills. As at the beginning, we take 10 pills, then note the number or Reds. We replace these pills, then take another sample of 10 pills. We continue with this process, each time noting the number of Reds. Let us say, having nothing better to do, we end up taking many thousands of samples. What can we now say about the number of Reds that were picked up in each sample?

The Central Limit Theorem tells us that the number of Red pills *per sample* should be near normally distributed. If we plot the number of times we picked up 0 Reds, 1 Red, 2 Reds…. etc, that graph would look like a bell curve. A pause is needed here, where we rewind to Figure 1. We had discussed that Figure 1 represents a Binomial Distribution. It is important to remember that Figure 1 characterises the probabilities of ***one*** sample of 10 pills. The CLT tells us about the behaviour of ***many*** such samples.

When the results of many samples are plotted, we should get a normal distribution. This distribution of repeated samples is called the '*sampling distribution of the means'*. The CLT goes on to say that larger the size of these samples, the closer the distribution gets to Normal.

The final important aspect of the CLT is that the initial distribution does not matter! For each sample the distribution can be of any variety [Normal or any other], the *sampling distribution of the means* still tends to a normal distribution. Since the Normal Distribution has mathematical features that allow easy analysis, this single unifying feature leads to methods of testing hypotheses for a wide variety of situations. We know that the Normal Distribution is characterised by a Mean and SD. This *sampling distribution of the means* has a Mean equal to the population Mean and SD equal to the population SD/√n [n = size of each sample].

At this point many may be saying to themselves, "We don't really perform research this way; we do not take repeated samples and try to get at the true value of a population". Very true! When we complete a study, we have just *one* sample to analyse. The point of the CLT is not to make us do the same study again and again. However, the CLT is the basis for developing tests, like the t-test, and has a rightful central role in inferential statistics.

4.3 Working with the Normal Distribution – the z score

We had already seen from Figures 5 and 6 that the area under the normal curve can be calculated with knowledge of the Mean and SD. The Z score measures the location of each point in units of SD. In simpler words, it measures how many SDs any point is from the Mean. Mathematically we can represent this by,

$$Z = (X - \mu) / \sigma$$

where X is our point of interest, μ Mean and σ Standard Deviation.

A word about notation. Since statistics deals with a sample and population, symbols have to be carefully chosen to avoid confusion. For *e.g.*, from any study we measure a mean which is the *sample* Mean. This sample Mean is what we use as an estimate of the *population* Mean, which is a different entity. Hence, it makes sense to have separate symbols for *sample* and *population* measures. The convention is to use Greek and/or Upper-Case letters for population attributes, and Roman counterparts for the sample. Table 4 shows these symbols.

Attribute	**Population**	**Sample**
Number of elements	N	n
Mean	μ	$\bar{x}$ [x-bar]
Variance	μ^2	s^2
Standard Deviation	μ	s
Proportions	P, Q	p, q
Coefficient of Correlation	ρ	r

Table 4.

To further understand the Z-score, let us take a look at Figure 6. This is the same as Figure 4 with an added point of interest. What is the Z-score of this point on the graph? On the x-axis, the point is at 2.7. Remember that this Figure had a Mean of 2 and SD of 0.5. Hence the Z-score is,

$$Z = (2.7-2)/0.5 \Rightarrow Z = 0.7/0.5 = \mathbf{1.4}$$

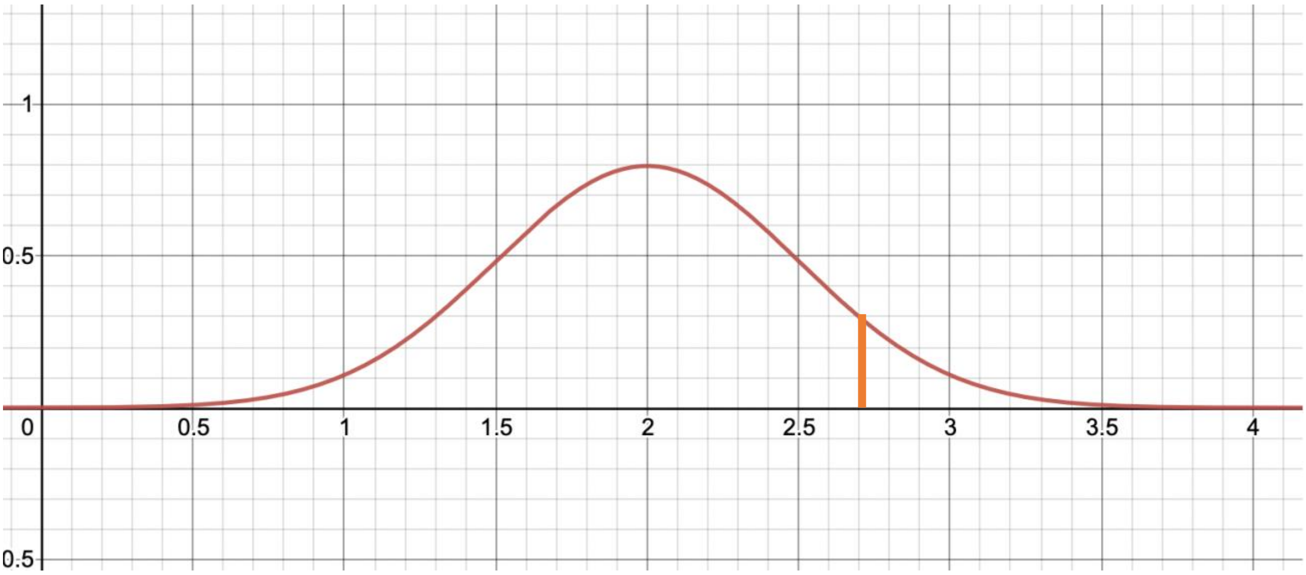


Figure 6. The point at 2.7 has a Z score of 1.4.

Calculating the Z-score allows us to calculate the area to the left or right of this point. That can be done from the equation for this distribution, with a sufficiently advanced knowledge of calculus. Thankfully for lesser mortals, standard tables are available that allow estimation of the area based on the z-score. For a z-score of 1.4, these tables tell us that ~92% of the area lies to the left of this point and ~8% to the right. We can also interpret that any random point has a 92% chance of being less than 2.7 and 8% chance of being greater than 2.7. Figure 5 showed that 95.44% of the area under the Normal Distribution lies in the range $\mu + 2\sigma$ and $\mu - 2\sigma$. That can be taken as the area between the two points with Z-scores +2 and -2. Furthermore, to delineate an area of exactly 95% around the Mean, the points would be those with z-scores +1.96 and -1.96.

To understand this better, let us assume that the distribution in Figures 5 and 7 represents the pinch strength of an adult population in kilograms. From the above z-score calculation, we can conclude that someone with 2.7kgs of pinch strength has better strength than 92% of the population. We can also invert the question. For *e.g.*, what should be the grip strength to be in the top 1%? In other words, at what point is 99% of area to the left? From the standardised tables, that z-score value turns out to be 2.33. We can now plug it in to the z-score formula to get our answer:

$$Z = (X - \mu)/\sigma \Rightarrow \quad 2.33 = (X - 2)/0.5 \quad \Rightarrow \mathbf{X = 3.165} \text{ kgs}$$

So, all those with pinch strengths greater than 3.165 kgs can be considered to be the top 1%. In that sense, 3.165 kgs represents the *cut off value* for the top 1%. Similar cut off values are used in almost all statistical tests to decide significance.

HIGH WATER

In April 1997, the city of Grand Forks in USA received a flood forecast warning from their National Weather Service [NWS]. The city braced for a record 49 feet flood level in the Red River, repairing levees up to 51 feet in height. However, the river crested at 54 feet, inundating an area up to 3 miles inland and necessitating the hurried evacuation of 50,000 people. Estimated losses from the flood for the region was $3.5 billion. Later reviews revealed that the NWS forecasts have an error margin of ±9 feet. Assuming that the error is normally distributed, this leads to a 35% probability of a flood level greater than 51 feet. Providing a single value forecast of 49 feet gives the impression of *certainty*, which caused the officials to fail to prepare for the eventual disaster. An old joke seems apt: the statistician drowned in a river that was 3 feet deep on *average*.

4.4 Applying a statistical test – a drive through example

This section seeks to get a ringside view of the steps involved in any hypothesis testing. For that purpose, we will apply a classic test on the data from a classic paper in hand surgery. The paper being 'Ulnar variance in carpal instability' by Czitrom, Linscheid and Dobyns from 1987.* The test performed is the most commonly employed of all statistical tests (possibly of all time) – the *t-test*.

The t-test has its origins in brewing quality stout beer. William Gosset, a chemist working with the Guinness Brewery in Ireland around the turn of the 20th century, needed to monitor the quality of the unique Guinness stout from small samples from large batches. He developed the 'hypothesis test statistic', which was later shortened

* Czitrom AA, Dobyns JH, Linscheid RL. Ulnar variance in carpal instability. J Hand Surg Am. 1987 Mar;12(2):205-8. doi: 10.1016/s0363-5023(87)80272-1. PMID: 3559071.

to the *t-statistic*, for this purpose. Since Guinness was not too keen on divulging its processes, Gosset published his articles under the pseudonym of Student, leading the test to be christened *Student's t-test*. Later giants like Pearson and Fisher expanded on Gosset's work to place the t-test on a firm mathematical basis. It is fascinating today to learn that Gosset wrote to Fisher "you are the only man that's ever likely to use them"!

The t-test was used in the above article to study the hypothesis that ulnar variance affects patterns of carpal instability. The data of interest to us is summarised in Table 5. The Mean and SDs of the ulnar variance of three groups are given – normal controls, acute scapholunate dissociations and old scapholunate injuries. Note that the entire data is not available, but for the t-test this is not necessary.

Variance [in mm]	**Normal**	**Scapholunate Dissociations**	**Old Scapholunate dissociations with arthrosis**
Mean	**-0.38**	**-1.36**	**-0.49**
SD	**1.48**	**1.60**	**1.31**
n	**65**	**78**	**53**

Table 5.

The t-statistic is a value similar to the z-score, calculated from the data, given by

$$t = [\overline{x1} - \overline{x2}] / s_t$$

where the numerator is the difference between two groups' Means, and s_t is a combination of the SDs of the groups. The exact formulae for calculating s_t involves the SDs and the sample sizes of the groups and are not detailed here.

The point of calculating the t-statistic is that it follows a unique distribution called the *t distribution*. The exact shape depends on a value called *degrees of freedom [**df**]*, which depends on the sample sizes. The **df** for a two-sample test is given by (n_1 -1) + (n_2 – 1), where n_1 and n_2 are the number of elements in each group. Hence, for a comparison of the Normal and Scapholunate dissociation groups from Table 5, the **df** would be (65 – 1) + (78 – 1) = 141. For the same comparison the t-statistic can be calculated to be 3.773.

Similar to the z-score, this t value allows calculating the probability of 3.773 at **df** of 141. This is the p-value and for this example is 0.000237. This being much lower than the usual cut off of 0.05, means that the difference between the groups is statistically significant. In addition, the t value also has cut offs for 95% area similar to the z-score. While for the z-score this is always +1.96 and -1.96, for the t-statistic this depends on the **df**. For a **df** of 141, these cut offs are +1.977 and -1.977. Our value of 3.773 is outside this range and therefore indicates, once again, of a statistically significant difference.

Proceeding to the comparison of Normal wrists vs Old scapholunate dissociations, the above procedure is carried out again. The **df** is (65 – 1) + (53 – 1) = 116. The t-statistic can be calculated to be 0.423. The cut off values for t at 116 **df** are -1.981 and +1.981. Since our t-statistic lies *within* this range we have to conclude that the Means do not have a statistically significant difference. The p-value turns out to be 0.6734, which is much larger than 0.05 and confirming that the groups are not statistically different.

For both the above comparisons, the Null Hypothesis would be that the groups are not different, essentially contending that the Means are not different and part of the same distribution. In the first instance, Normal vs Scapholunate, the t-test tells us that this *data has a probability* of only 0.000237 (0.0237%), if the Null Hypothesis is true. Being less than our pre-decided cut off of 0.05 (or 5%), we have to reject the Null Hypothesis and conclude that the Means of these groups come from *different* distributions. In the 2nd comparison, the data has a probability of 0.6734 (or 67.34%). We are unable to reject the Null Hypothesis in this case.

The idea of this example is not to encourage painful calculation of cut offs and p-values. But rather to illustrate the hidden workings of a statistical test. When raw data gets converted magically to a p-value, it is easy to lose track of what the test actually means. Knowing, at least broadly, how this is done should help choose the right test for the data to be analysed.

5. The Right Test

Not too long ago performing a statistical test required knowledge of its inner workings, along with protracted calculations. Then would be needed the appropriate set of tables to finally come to a p-value. However, the world has leaped forward enough to have even statisticians not needing to know exact formulae for each statistic. On the other hand, there are a multitude of tests and their variations available to choose from. All that is required at present is to choose the proper test and the rest can be carried out with suitable means (Appendix 2).

5.1 Types of Data

Data is collected in terms of **Variables**, which is a simple way to place data into its appropriate class. Any data can be divided into two broad classes: **Categorical** and **Numerical**. **Categorical** data just consists of categories into which elements have been arranged. Gender is a common example, with two or more categories. The MRC grading of muscle strength consists of 6 categories (0 to 5 grades). Here Gender and MRC Grade are the variables, which can only be described in categories. **Numerical** data consists of variables that can be *measured*. Height and Weight being prime examples. They can be measured to any degree of accuracy that we please to achieve. Contrast that with the variable Cancer Stage, which is described in categories I through IV with elements being *counted* into each of them. So Categorical data comes in classes and needs to be counted, while Numerical data can be directly measured.

We can further sub-divide data in to *scales* as in Figure 7 (more details in the earlier Statistics Article). The division of Categorical data into **Nominal** or **Ordinal** is simply based on whether the categories have a definite Order. So, Gender would be Nominal, while MRC Grading is Ordinal. It is also useful to divide Numerical data into **Discrete** or **Continuous** data. Discrete data consists of whole numbers *only*, while Continuous data can take fractional and irrational values.

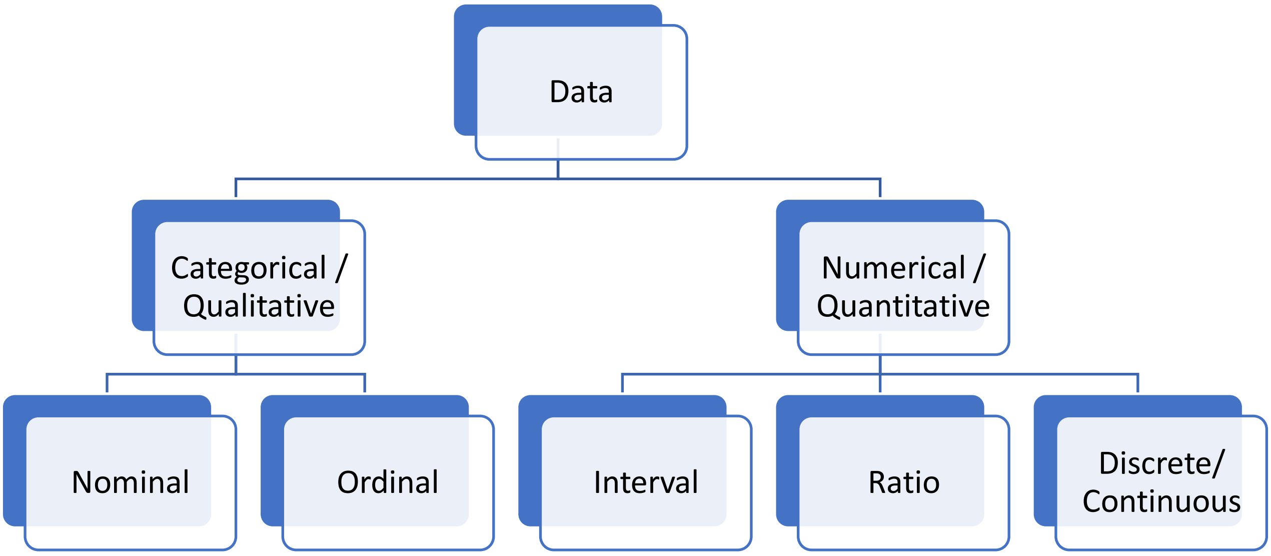


Figure 7.

5.2 Parametric vs Non-parametric tests

In Section 4, we had discussed the Normal Distribution and the Central Limit Theorem. We concluded with an example of the t-test. This test is based on an assumption of the underlying data being normally distributed. The t-statistic is a *parameter* that was calculated based on such as assumption. Similarly, other tests that rely on such an underlying assumption also have parameters calculated as part of the test procedure. These family of tests are termed **parametric** tests, implying the assumption of normality of data.

It is obvious that not all data can be blindly assumed to be normally distributed. In many situations, data may be severely affected by extreme values and the distribution skewed in one direction. For an example, length of hospital stay after routine surgery clusters around a usual value, say 2-3 days. However, some patients may develop complications and require a prolonged stay. Such data is likely to be skewed with a long tail towards increasing stay. Another example is the APGAR score. Here the tail is towards lower scores, since most newborns have a score of 7 or above. In addition, since the Normal Distribution is a continuous distribution, discrete variables cannot be considered normally distributed by default. The VAS score is such an example as it is a discrete measure, often rated from 1 to 10.

Non-parametric tests were developed to remove this assumption of normality of data. Hence, they are also aptly called **distribution-free** tests. These tests mostly involve ranking the observations and then performing calculations on these rankings. Examples include the *Mann-Whitney Test*, *Wilcoxon Signed-Rank test* etc. For most research situations, there is an alternative non-parametric test available that must be considered in lieu of traditional parametric tests. This immediately raises the question, “Why not just always use the non-parametric tests? And be done with the old-fashioned parametric ones”.

The answer lies in that there is a trade-off involved. As is often in life, the wide applicability of the non-parametric tests comes with its own set of problems. Most importantly, they are less likely to find a significant difference between groups than parametric tests. In statistical parlance, they are less *powerful*. The moral of the story is to apply non-parametric tests only when the normality of the data is questionable. To help with that question statisticians have, of course, developed even more tests. These *tests of normality* can be applied to the data before proceeding towards hypothesis testing. For example, the Shapiro- Wilk test is a commonly employed test for normality which has a Null Hypothesis that the data is normally distributed. Much like any other test, the output is a p-value which can be used to accept / reject this Null Hypothesis.

5.3 Independent and Dependent Variables

We have already seen in 5.1 regarding types of data. Variables can also be considered Independent or Dependent, depending on the design of the study. The variables which are chosen/controlled by the investigator are the Independent variables. Other variables respond to changes in these variables and are termed the Dependent variables. It is easier to consider these as Input variables and Output variables. For example, in our example of the t-test the Input variable is the pattern of carpal instability. The division into Scapholunate Dissociation and Old Scapholunate Injury is Categorical Nominal scale data. The Output is the measurement of ulnar variance, which is Numeric Ratio scale data. Clearly understanding these differences is important in determining which tests to employ.

5.4 Tests for matched or paired observations

Study designs can involve observations made before and after intervention or exposure. Such data is called ‘matched or paired data’. It is obvious that we are actually interested in the *difference* between these paired observations. While it is possible to work with such data as two separate populations (like the t-test example in 4.4), specific tests have been designed for these situations for a better analysis. They are summarised based on the scale of data of the outcome in Table 6.

The paired t-test is possibly the most common of these, performed in much the same way as our example in 4.4. However, it is important to note the wide applicability of the Wilcoxon Signed-Rank test. It is excluded only for Nominal scale data and is a non-parametric option to be considered in almost all clinical situations. The test basically involves first tabulating the difference between each pair of observations. These differences are then *ranked* i.e., arranged from lowest to highest. These ranks are then used for hypothesis testing.

Type of Output Variable	Test
Nominal	McNemar Test
Ordinal	Wilcoxon Signed-Rank test
Quantitative [Discrete/ Non normal]	Wilcoxon Signed-Rank test
Quantitative [Normal – 2 observations]	Paired t-test
Quantitative [More than 2 observations]	Repeated measures ANOVA

Table 6.

5.5 Tests for independent observations

We have already seen the t-test being applied to two independent groups – Normal controls and Scapholunate Dissociation. For more groups, the t-test is not valid. The **An**alysis **o**f **Va**riance [**ANOVA**] test was developed as an extension of the t-test for multiple groups.

i) The t-test family of tests

We had seen that the t-test is based on the assumption of underlying normality of data in two groups. The procedure involved calculating a t-statistic, the formula for which involved the difference between the two means in the numerator. It is apparent that this would not be possible for *three* means. We could, if we did not know better, perform a pairwise comparison. That means testing two means at a time, which for three groups ends up with three t-tests. But for more groups this idea ends up getting ugly pretty soon. With four groups we would need 6 t-tests and for five groups no less than 10 t-tests.

In 1919, Ronald Fisher joined the Rothamsted Experimental Station, one of the oldest agricultural research stations in the world. Faced with a vast amount of data collected since 1842, he came up with the ANOVA test for multiple groups. In short, the test involves comparing the variation *between* groups and *within* groups. Similar to the t-statistic, the test calculates the F-statistic which is used for hypothesis testing. The F-distribution, like the t-distribution, depends upon the degrees of freedom and the cut off values are calculated similar to the t-test.

Note that the multiple groups mentioned here are in the *Output* variable. We can also have a situation with multiple *Input* or *Independent* variables. As an example, say we design a study comparing three different treatment protocols for Kienbock's disease. One of the output variables could be grip strength, which can be assumed to be normally distributed. Adding a twist, we also decide to study the effect of gender on the outcomes. So now we have two Categorical [Nominal] Input variables, namely Male and Female. We can perform an ANOVA for the Males and *another* ANOVA test for the

Females. However, it is possible to expand the ANOVA to study the effect of Gender as well. This is called *Two- way ANOVA*, different from the *One-way ANOVA* we had discussed in the last paragraph. To cut a long story short, it is indeed important to *exactly* understand what is being measured and what are the comparisons made. This pretty much is all that is necessary to choose the right test.

ii) Chi-square test

The chi-square [χ^2] test is applied for data with *both* the Input and Outcome variables in Categorical scale. Basically, data consists of the counts or the frequency of each category. The simplest instance is the 2x2 Contingency table, illustrated by an example. A retrospective study by Grandizio *et al*,* compared the incidence of Trigger Digits [TD] after Carpal Tunnel Release [CTR] in Diabetics vs Non-Diabetics. While they studied a number of risk factors also, their primary data is summarised in Table 7.

	TD in 1 year post CTR	No TD in 1 year	Total
Diabetics	21	193	**214**
Non-Diabetics	42	961	**1003**
Total	**63**	**1154**	**1217**

Table 7.

Here the Input variable is diabetic status in post CTR patients and the Outcome we are interested in is incidence of post op triggering. Both are Categorical [Nominal] scale data and Table 7 just *counts* the number of patients in each group. From the table, we note that 63 out of a total of 1217 patients developed post CTR triggering [about 5.18%]. However, in Diabetics about 10% developed triggering [21 out of 214]. The chi-square test measures this difference between the *observed* frequency and an *expected* frequency based on the total values. Table 8 adds what the expected frequencies would be for data in Table 7.

	TD in 1 year post CTR **Expected**	No TD in 1 year post CTR **Expected**	Total
Diabetics	21 **11.07**	193 **202.93**	**214**
Non-Diabetics	42 **51.92**	961 **951.08**	**1003**
Total	**63**	**1154**	**1217**

Table 8.

* Grandizio LC, Beck JD, Rutter MR, Graham J, Klena JC. The incidence of trigger digit after carpal tunnel release in diabetic and nondiabetic patients. J Hand Surg Am. 2014 Feb;39(2):280-5. PMID: 24360881.

The chi-square goes on to, unsurprisingly, calculate the χ^2 statistic. Also unsurprisingly, the χ^2 distribution then allows calculation of cut off values and decide if the difference between observed and expected frequencies is significant or not. For the data in Tables 7 and 8, the χ^2 statistic is 11.37, with a p-value of 0.0007. We conclude that Diabetics are more likely to develop TD within 1 year of CTR.

iii) Correlation and Regression

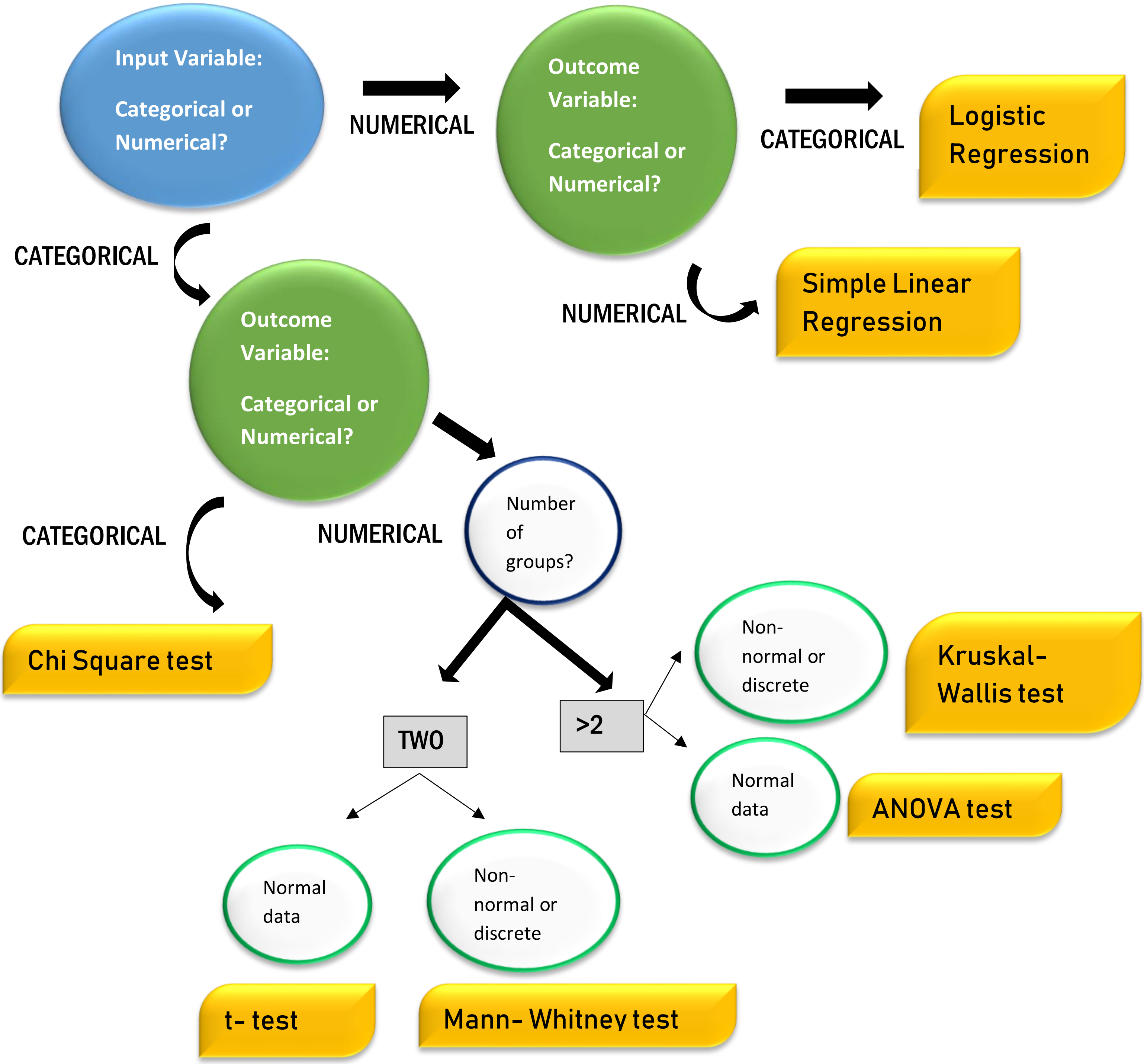
We are left with the broad class of situations where both the Input and Outcome variable are Numerical. This may sound confusing, so we proceed to an example. Meislin *et al* studied the difference in measurements of elbow ROM performed by goniometry vs smartphone photography. Now the Input variable here is the method of measurement (goniometer or photography) which is Nominal scale and the Output is Numerical (the ROM). Indeed, the authors perform a t-test between these groups to come up with a p-value of 0.90 for the Left elbow and 0.88 for the Right elbow, concluding both the methods are similar. However, what if we wished to *predict* the ROM measured via photography from a goniometer measurement? That is performed by Correlation and Regression analysis.

In correlation, one set of values are compared to another to study how they vary with each other. Mathematically this translates to a *correlation coefficient*, that ranges between -1 and +1. Values around zero, positive or negative, signify that the sets do not vary together. Values closer to +1 signify that they both increase/ decrease together, and values close to -1 imply that when one set increases the other decreases. In the above Elbow ROM study, the correlation coefficients were 0.85 on the Left and 0.76 on the Right. We conclude that the ROM measured by both the methods are strongly correlated with each other. Regression analysis take this further, by using the correlation coefficient to set up an equation connecting the sets of data. Feeding one value, like the goniometer measured ROM, then gives the most likely value of photography measured ROM.

iv) Decision making

Armed with the understanding of the previous sections, we are now ready for a step-by-step decision tree for choosing the right test. As a disclaimer, this tree is intended only as an illustration of the steps involved. Every study may not fit exactly into these patterns, and each should be properly examined to come up with a choice of statistical test.

The Input variable is often the starting point as it is the easiest to identify, being usually the basis of the study. We go on to recognise the Output variable(s), which is / are the data that is measured. Each combination of Input and Output variable then has a few options to choose from, as shown below.



A short decision tree for choice of test

A GENDER BENDER

Urban legend goes that in 1973 the University of California in Berkeley [UC Berkeley] was sued for gender discrimination in their admission process. While no record of such a lawsuit exists, the authorities were certainly worried that they may be sued. To their credit, they invited statisticians to analyze the data. The initial data, that had raised eyebrows, looked like this –

	Applications	Admitted (%)
Men	8442	**44**
Women	4321	**35**
Total	12763	41

Significance testing gave a p-value of $\approx 10^{-26}$ for this difference of 9%! The researchers then decided to break down the data department wise, with that of the largest 6 departments given below.

Departments	MEN		WOMEN	
	Applied	Admitted (%)	Applied	Admitted (%)
A	*825*	62	108	**82**
B	*560*	63	25	**68**
C	325	**37**	*593*	34
D	417	33	375	**35**
E	191	**28**	*393*	24
F	373	6	341	**7**

It appeared that women applied more to competitive departments with a lesser acceptance rate, while men applied more to departments with higher acceptance rates. This difference in proportions skews the total *ungrouped* data downwards for women. The tendency for ungrouped data to lead to different conclusions than grouped data is called **The Simpson's Paradox**. The moral of this story? Look closely!

6. Error, Power and Sample Size

While the above sections covered the analysis of data mostly after a study has been conducted, some key concepts are necessary for actually *designing* a project. It is critical to have a prior understanding of the nature of data *expected* to be collected. The statistician's input is doubtless most decisive before the project commences.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of "

- Ronald Fisher

6.1 Types of Errors

We have seen that hypothesis testing involves setting up a Null Hypothesis [H_0] and then testing its validity. Due the very nature of this set up there are some errors that are **unavoidable.** The best that we can hope for is attain a minimum state for these errors. It is important to realise that, apart from these, there are *avoidable* errors that creep in due to the study design or practical difficulties (for e.g. a small sample size). These errors are too numerous to enlist, but adhering to basic principles should help minimising them as well.

Table 9 shows the usual setup of hypothesis testing, comparing the truth of the Null Hypothesis [True or False] with the result of the study [Accept or Reject].

	Null Hypothesis [H_0]		
Result		True	False
	Accept	CORRECT	Type II (β)
	Reject	Type I (α)	CORRECT

Table 9.

We see that there are two possible errors – rejecting a true H_0 and accepting a false H_0. For easier understanding let us compare these with our justice system. The Null Hypothesis is similar to the *presumption of innocence* ('innocent until proven guilty'). All accused are presumed innocent until enough evidence proves their guilt 'beyond reasonable doubt'. In this analogy, a Type I error is when we find an innocent person guilty. The H_0 that a person is innocent, has been rejected in spite of it being true. So a Type ***i*** *error **i**ncarcerates **i**nnocents*. On the other hand, when a guilty person is found innocent a Type II error has occurred. The H_0 was false, but we failed to reject it. The investigative team was not *powerful* enough to get the job done! As we shall see, this analogy is a useful definition of statistical Power.

The Type I error rate, designated as α, is directly decided by the level of significance chosen for the study. When the level of significance is chosen at a p-value of 0.05, we have set up to reject the Null Hypothesis when the probability of the data is less than

5%. But ‘less than 5%’ is not zero. When the p-value for a study has been calculated to be 0.04 (or 4%), we reject the Null Hypothesis *in spite* of the 4% chance of being wrong. So the level of significance is also the chances of rejecting a true Null Hypothesis, which is the definition of Type I error. Hence, for most studies the Type I error rate or α would be 0.05.

In an ideal world, we would certainly want to minimise all error rates. For the α error, this means setting the level of significance lower, say 1% or 0.01. This of course means that many more studies would end up with non-significant results. Bad news if you are trying to get your research published!

6.2 Type II error and Power

We had defined Type II error to be when we accept a false H_o, usually designated as β. From Table 9, the rate of *correctly* rejecting a false H_o would be $1-\beta$. This is defined to be Power. In other words, it is also the ability to detect a difference that is actually present. Evidently, this is an important measure which we would like to maximise. However, unlike α, Power and β are dependent upon multiple factors and so are more difficult to control. Some of these are –

i) Effect Size : When comparing groups it is not just sufficient to detect a difference, it is also important to know the *magnitude* of that difference. For e.g., when measuring finger ROM after tendon repair, we may find a difference in ROM after using different repair techniques or rehabilitation protocols. But finding a 2° difference in mean ROM is not the same as ending up with a 10° difference, although hypothesis testing may find a significant difference in both the cases. Patients and caregivers would not feel that a 2° difference is of any practical use. This concept is captured as Effect Size (ES). Obviously, when the difference between groups is larger it would be easier to find that difference. From our understanding of Power, that means that Power should be higher for higher Effect Sizes. It is now considered good form to report Effect Sizes as well, when reporting on differences in means of groups.

Calculating the ES is not just a simple matter of calculating the difference between means. The ES needs to be comparable between measures, which may be in different units. We would not be able to compare a 5° difference in ROM with a 2kgs difference in Grip Strength. To overcome this, the ES is standardised by dividing the difference of the means with a measure of the standard deviation of the groups. There are different ways described to calculate this standard deviation measure, leading to different measures of ES. In a nutshell, there are more than 50 (!) such measures known with names like *Cohen's d*, *Glass' Δ*, *Standardised Response Mean* etc.

ii) α error : The Type II error rate and Power are actually dependent on the Type I error rate as well. This may seem surprising at first, but it easy to see why. Let us say we attempt to decrease the α by decreasing the level of significance to 1% from the usual 5%. This makes it more difficult to reject the H_0. Since Power is a measure of correctly rejecting the H_0, it also decreases (or Type II error increases).

iii) Sample Size : Instinctively, we feel that a larger sample size should help us reach the right decision. This is, thankfully, true mathematically as well. A larger sample size should help us correctly reject a wrong H_0, increasing Power. The next section deals with the interplay of Power, ES and sample size calculation.

6.3 Sample Size calculation

We have come, at long last, to what should be one of the initial steps of a study. However, sample size calculation involves understanding a broad range of concepts. In fact, as a final step, we require to understand Confidence intervals [CI]. Refreshing up on Sections 4.2 and 4.3 would be useful.

We know that our sample mean ($\bar{x}$) is only an estimate of the true population mean (μ). Such an estimate is called a **point estimate**. Now we know from the Central Limit Theorem (CLT) that the sampling distribution of the means has a mean equal to the μ, and a standard deviation equal to $\sigma/\sqrt{n}$. This knowledge allows us to make an **interval estimate** of μ from $\bar{x}$. With 95% confidence we can estimate that the μ lies within an interval 1.96SD on either side of $\bar{x}$.This is called the 95% Confidence Interval for the mean [95%CI].

$$95\%\ \text{CI} \rightarrow [\bar{x} + 1.96\sigma/\sqrt{n},\ \bar{x} - 1.96\sigma/\sqrt{n}]$$

The general formula for a CI would be $\pm z.\sigma/\sqrt{n}$, where z is the z-score for the desired level of precision. For 95%, this would be 1.96 as seen earlier.

Now let us set up to perform a study that plans to estimate a mean with 95% confidence. The interval on either side of the mean can be taken as the *error* that we are ready to tolerate. Let us call it E.

$$E = 2 \text{ x } z.\ \sigma/\sqrt{n}$$

The additional 2 is since we have to take the range on *both* sides of the mean, each side being $z.\ \sigma/\sqrt{n}$.

Rearranging,

$$\sqrt{n} = 2z \text{ x } \sigma/\ E \Rightarrow \quad n = 4z^2.\sigma^2/\ E^2$$

Let us take an example and work out a needed sample size. We plan to perform a study on the functional outcome after carpal tunnel release, with grip strength as a measure. Some assumptions are required for all sample size calculations. Firstly, we want to our grip strength to be correct within a range of 3kgs because a wider interval would be clinically less useful. So this becomes our margin of error (E). From historical data, we find that the standard deviation of grip strength in the population is about 6kgs, which would be our estimate for σ. The z is 1.96 when we set our level of significance at 0.05. Plugging them all in, $n = 4\text{x}1.96^2\text{x}6^2/3^2 = 61.47 \Rightarrow$ **62**

Since we are calculating a *minimum* sample size we must *round up* to the nearest whole number, as people do not come in fractions. We conclude that a sample size of 62 is required to measure the grip strength to within a 3kgs range with 95% confidence.

But this is not the end of the story, as usual. The above was a simplistic analysis illustrating the general principles. However, in practice, different formulae are required for different situations. As another example, the sample size calculation for a comparison of two means and Numerical scale data is given by –

$$n = 2[\frac{z_{1-\frac{\alpha}{2}}+z_{1-\beta}}{ES}]^2$$

where n is the minimum size of *each* group, α and β are the rates of Type I and II errors and ES is Effect Size. $Z_{1-\frac{\alpha}{2}}$ and $z_{1-\beta}$ represent the z-scores for these error levels. We know that the usual α is set to be 0.05, while most agree that β must be no more than 0.2 or 20%. This choice of β translates to a Power of 80%. Under these standard circumstances the above formula can be simplified to,

$$n = 2[\frac{1.96+0.84}{ES}]^2 = \frac{15.7}{ES^2}$$

Recall that the ES is a complex measure calculated as a ratio of the difference of means and standard deviation. As an example, say we perform the above study on grip strength after carpal tunnel release as a comparison of endoscopic and open techniques. We assume that a 3kgs difference would be clinically important, with the estimate for σ being 6kgs. Then a crude estimate for ES is 3kgs/6kgs = 0.5. From the above formula, we get

$$n = 15.7 / 0.5^2 = 62.8 \sim \mathbf{63}$$

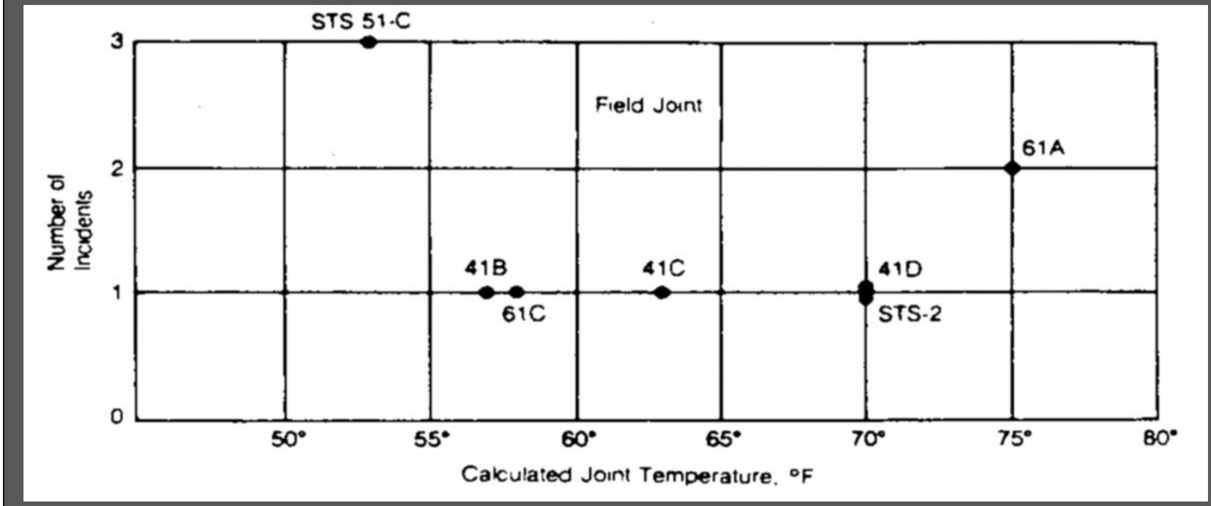
We conclude that to detect a difference of 3kgs between groups, with 80% power, we need 63 patients in *each* group.

Let us step back and digest all this number crunching. In the first example, the sample size calculation was performed in the context of estimating a mean grip strength. In the second, calculations were performed for *hypothesis testing* for the difference between two groups. Sample size calculations are performed always in the specific framework of the study. In summary, it is not important to know all these formulae so much as to understand the context of the calculation. As long as this is understood, the calculation can be performed via some of the resources detailed in the Appendices.

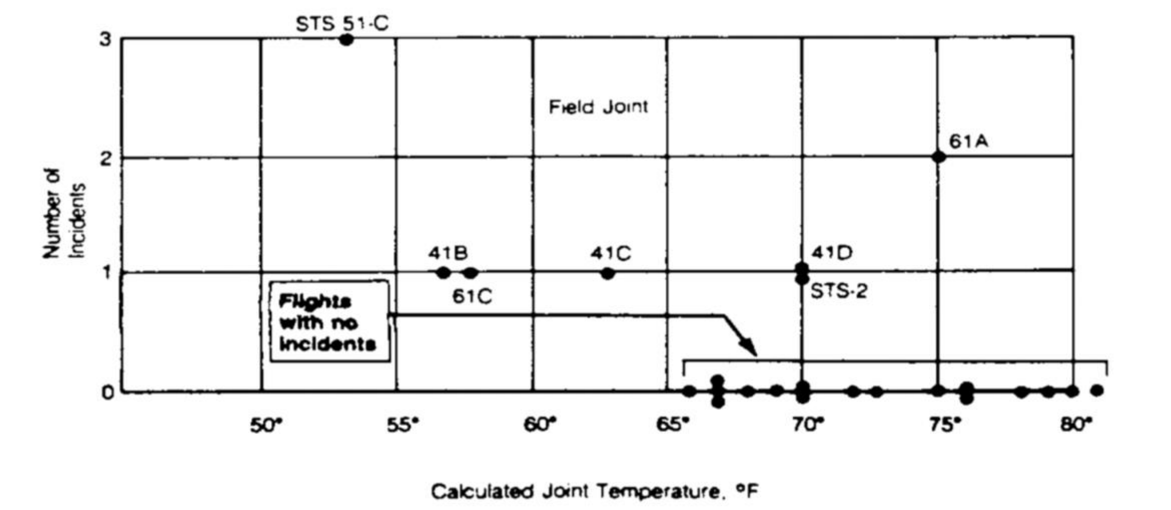
STS – 51 – L

Mission STS-51-L was the tenth flight of the Space Shuttle *Challenger*. On January 28, 1986, *Challenger* took off from the Kennedy Space Center with 7 crew members on board. The launch was a much-publicized event as one of the crew members was a high school teacher, Christa McAuliffe, for inspiring interest in science. About 17% of the US population watched the launch live, including children at school, when the shuttle exploded at an altitude of 15kms about 73 seconds after lift-off, killing all crew members. To address the concerns of a grief stricken nation the Rogers Commission was set up under William P Rogers, ex Secretary of State and ex Attorney General of the USA. The commission included Neil Armstrong and the Nobel Prize winning physicist Richard Feynman.

The Commission zeroed in on faulty rubber seals called O-rings in the booster. During the investigation, it was revealed that there were prior concerns about the O-rings failing at low temperatures during launch. These O-ring incidents had been identified and plotted vs temperature. Based on data similar to the figure below, the eventual conclusion was that temperature did not affect O-ring performance.



However, the Commission noted that the analysis was faulty since it did not include the flights where there had been ***no incidents***. When these were included, the picture changed considerably.



It was obvious that the flights without problems were all at higher launch temperatures. A correlation analysis showed no correlation for the first data set but the second figure displayed negative correlation, implying higher incidents at lower temperatures and *vice versa*. In the figure above, the lowest temperature without any incidents is 66^{o} F. The *Challenger* was launched at temperatures of around 30^{o}F. It turned out that the go ahead for the launch was decided upon by NASA managers *in spite* of warnings from engineers.

Amidst rising concerns that the O-ring anomaly may be buried to avoid showing NASA in poor light, the colourful and venerated Richard Feynman made sure this could not happen. During a televised hearing, he placed material from an O-ring in ice water and demonstrated that it became stiff. He was also irked by NASA insisting that the risk of catastrophic malfunction in the Space Shuttle was 'necessarily 1 in 10^5'. Feynman noted that this meant that NASA could launch a shuttle every day for 274 years while suffering, on average, only one loss. He reckoned this should be about 1 in 100 with 1 in 10^5 sounding wildly fantastical. He anonymously polled the engineers themselves, who turned in estimates ranging from 1 in 50 to 1 in 200. When the Space Shuttle program was eventually cancelled in 2011, it had suffered two losses (*Challenger* in 1986 & *Columbia* in 2003) in 135 flights.

Feynman wrote in his report, "For a successful technology, reality must take precedence over public relations, for nature cannot be fooled".

Epilogue

In 1710, John Arbuthnot, a Scottish polymath and satirist, published his study on the observed ratio of male and female births. He used the baptism records of London for the years 1629 to 1710 to note that there were more male births in *every* one of those 82 years. The ratio ranged from 1.156 in 1661 to 1.011 in 1703. He argued that if the natural ratio of births was indeed equal, we should expect to find more *female* births in half the years and more male births in the rest half. So, the probability of finding more male births in a year must be ½ and of that happening for 82 years in a row must be $(1/2)^{82}$, which is in the range of 10^{-25}! He concluded that the continued birth of more males that females was an 'Argument for Divine Providence'. While that inference has its doubters, we can celebrate the study for the first significance test in history. The probability that Arbuthnot had calculated was basically a p-value with the Null Hypothesis that male and female births should be equal.

Three centuries later, we seem to have come a long way. Talk of p-values and 'statistical significance' pervades academia. However, all is not well in the world of research. We have already seen the dead salmon fMRI study as an example of how analysis can go wrong, when disconnected from reality. In a controversial 2005 article, John Ioannidis of Stanford University made the claim that 'most published research findings are false'.[*] While the article has been critiqued for exaggerating the problem, most statisticians would agree that false positive results are far more than understood by the scientific community. This is also related to what has been termed the 'reproducibility crisis' in science. Since, many initially 'significant' results may be false positives, repeat studies fail to come to the same conclusion. For e.g., the Reproducibility Project in Psychology repeated 100 studies from reputed journals. Originally, 97 of these studies claimed significant results, which reduced to 35 (36.1%) on the repeat attempt. Similarly, another study estimated that about $28 billion worth of preclinical research was non-reproducible. Much of the blame is laid on the flawed understanding of p-values and excessive importance attached to 'statistical significance'. The American Statistical Association [ASA] came up with a 'Statement on Statistical Significance and P-values' in 2016 to clear the air on the issue. This is a must read for researchers and is discussed in Appendix 4.

Let us conclude with the words of Ron Wasserstein, ASA Executive Director, explaining the need for the above statement – "The p-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p<0.05 era'". Which takes us back to Cicero at the beginning.

[*] I Ioannidis JP. Why most published research findings are false. PLoS Med. 2005 Aug;2(8):e124. PMID: 16060722; PMCID: PMC1182327.

Appendix 1 – Further Reading

- Textbooks: *Statistics* by Freedman, Pisani and Purves is a classic textbook in this field but may be too technical for a beginner. *Biostatistics: The Bare Essentials* by Norman and Streiner is not only a witty read, but also an easier introduction to the sort of statistics that is most likely to come up in medical research. *Biostatistics* by Daniel and Cross is a more rigorous approach for those unfortunately mathematically inclined.

- Casual Reads: Many books meant for the non-technical reader are excellent primers to the field of statistics. These do not focus on the tests or techniques, but on principles that drive research. Some are:

  - *Statistics without Tears* by Derek Rowntree is a comprehensive, all round introduction for non-mathematicians.
  - *The Art of Statistics* by David Spiegelhalter outlines the basic principles and has an excellent coverage of the dangers of bad statistics.
  - *Freakonomics* and its sequel *Superfreakonomics* by Steven Levitt and Stephen Dubner deals with the use of data in solving everyday issues. Well received for its humour and unconventional style.
  - *The Signal and The Noise* by Nate Silver focuses on prediction and forecasting, with common pitfalls made in interpreting data.

Appendix 2 – Statistical Resources

It is a daunting task for *statisticians* to remember the formulae for each distribution and statistic. Surgeons, therefore, stand no chance in calculating these by themselves. Mercifully, the Internet has solved this issue for us and many such less fortunate practitioners who need to know statistics. We start with some, mostly free, software and then move on to fully free websites that get the job done.

- Microsoft Excel: the classic spreadsheet software is a surprising statistics powerhouse. Most everyday statistics can be carried out with Excel, if you know where to look. Appendix 3 shows where to look.

- JASP: an open-source platform supported by the University of Amsterdam, it is "free, friendly and flexible". JASP allows all the basic tests to be done easily, with a simple and clean interface. Free textbooks are also available online to learn how to use JASP. Probably the best choice for a beginner in statistical testing.

- PAST: acronym for PAleontological STatistics, PAST has been developed over two decades by the University of Oslo for palaeontologists. The current iteration is a no-nonsense comprehensive package that not just executes hypothesis testing but permits complex modelling and some involved data manipulation. It also comes with a much necessary manual for its myriad features.

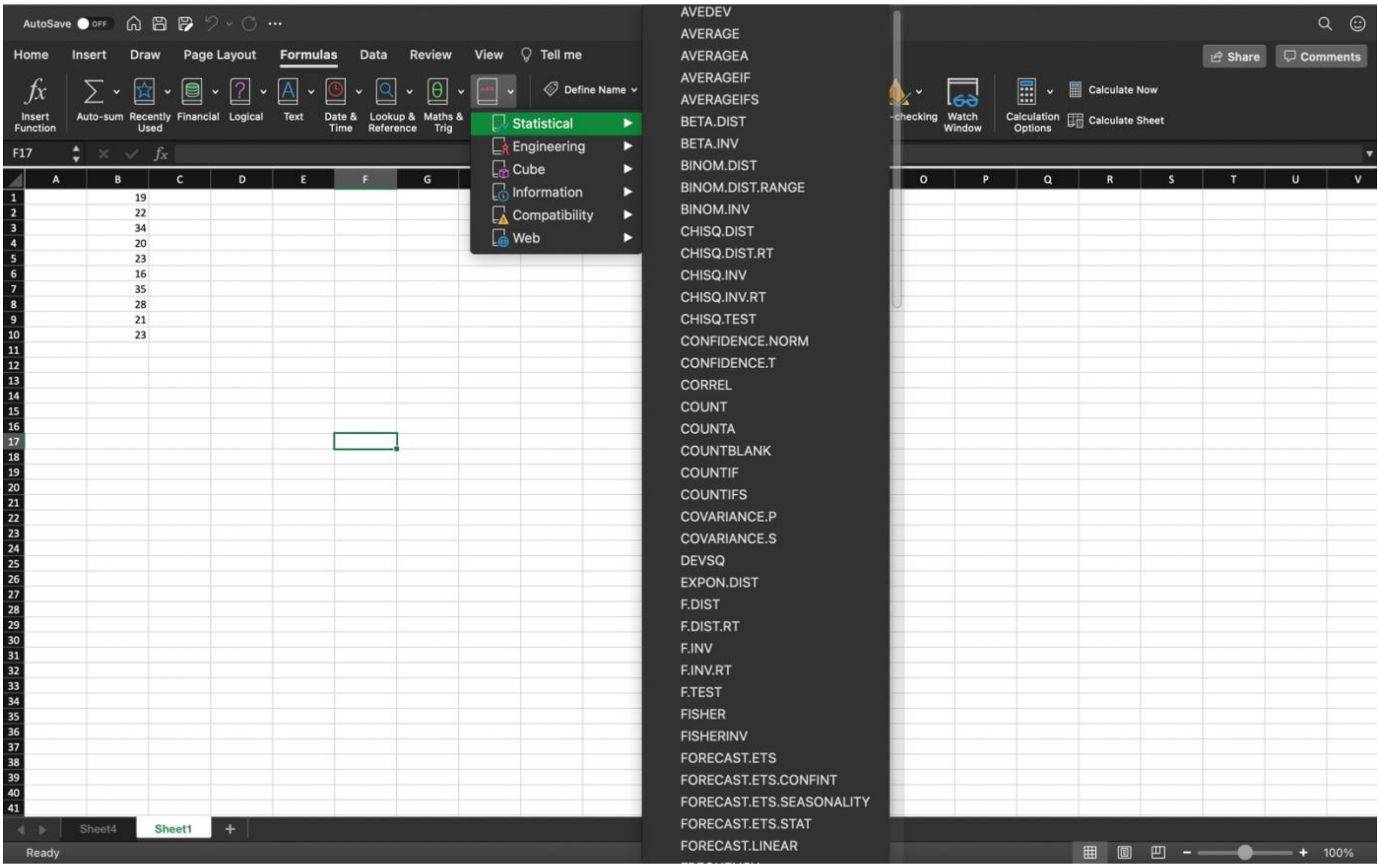
- SOFA: another open-source package that allows creating and exporting attractive charts, along with basic statistics.

- Websites:

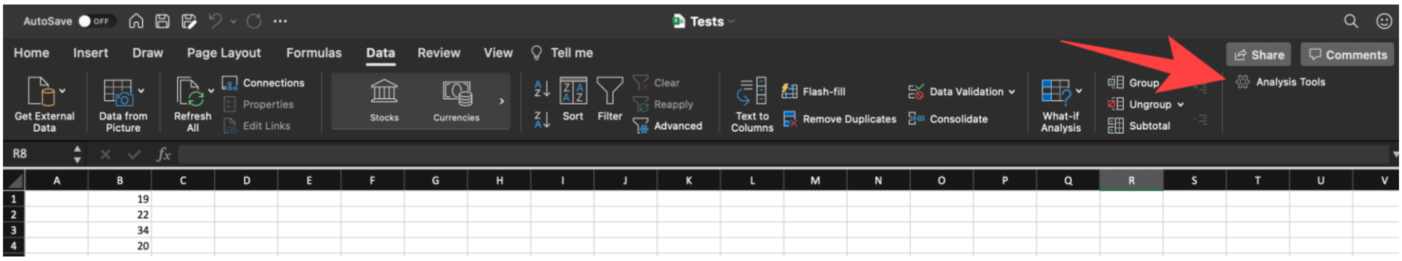
    - Vassarstats.net – free site that allows online statistical calculation. For clinical research, performs all the common tests likely to come up in clinical research. The interface is spartan and it takes time to get used to data entry. Also linked to an online version of the book *Concepts & Applications of Inferential Statistics* by Richard Lowry, erstwhile of Vassar College, New York, USA.
    - SISA [Simple Interactive Statistical Analysis] - simple interface that allows all the basic tests. Not as complete as Vassarstats but has extra features of sample size and power calculation.
    - OpenEpi – as simple as it gets interface, with all the basic tests.
    - Statibot – interactive site that works out the right test to perform from entered data and does some of them as well. It is a beautiful means to get introduced to hypothesis testing for the absolute beginner, as it works out an analysis from minimal presumptions.

Appendix 3 – Excel-lent!

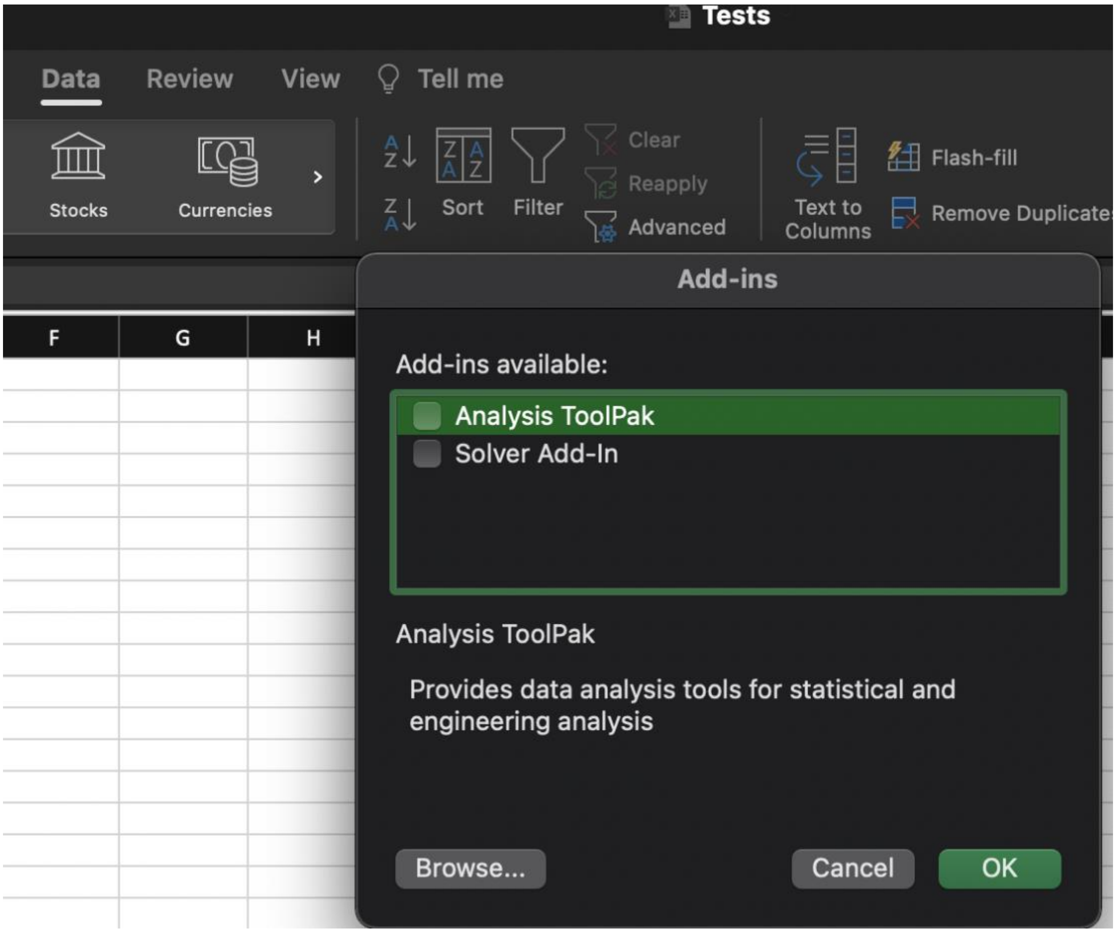
MS Excel is a simple resource to get through basic testing, especially since most of us already use Excel for creating and storing our research data. While there are many of these utilities available under the **Formulas** menu (picture below), there is a neater approach with the **Analysis ToolPak**.



The **Analysis ToolPak** is an add-in that needs to be, well, added in.



Under the **Data** menu, click on **Analysis Tools**. This opens up the Add-ins menu, select **Analysis ToolPak** and click 'OK'. This adds the add-in.



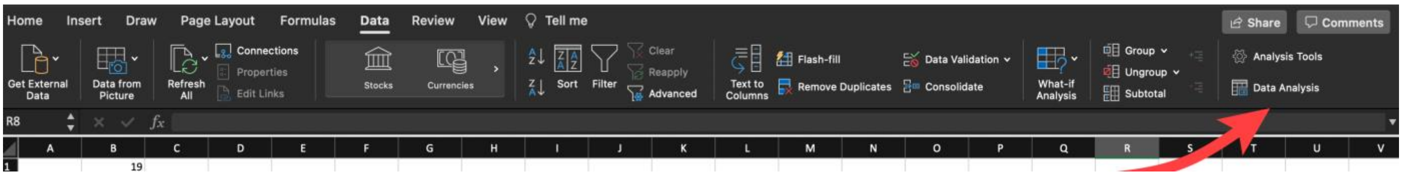
Add-ins available:

- □ Analysis ToolPak
- □ Solver Add-In

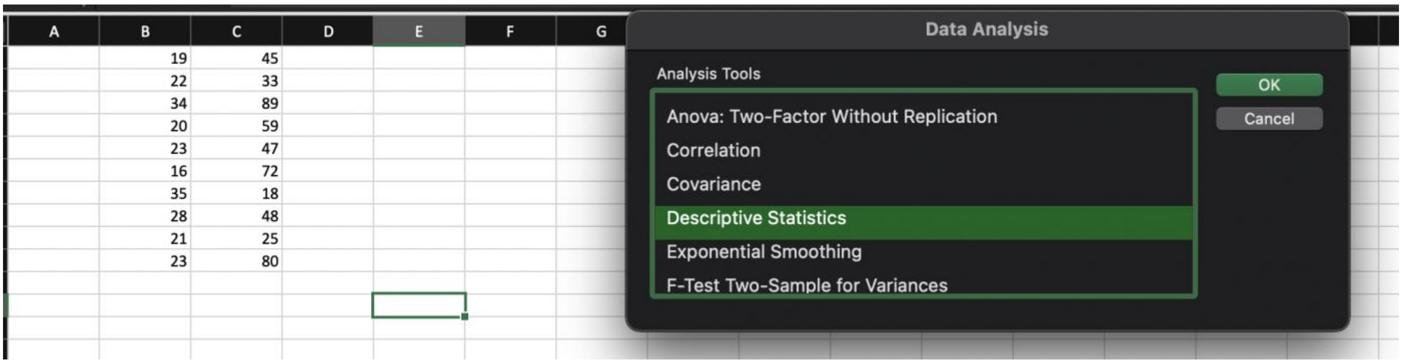
Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

After this is done, the pack appears in the **Data** menu as the **Data Analysis** tab.



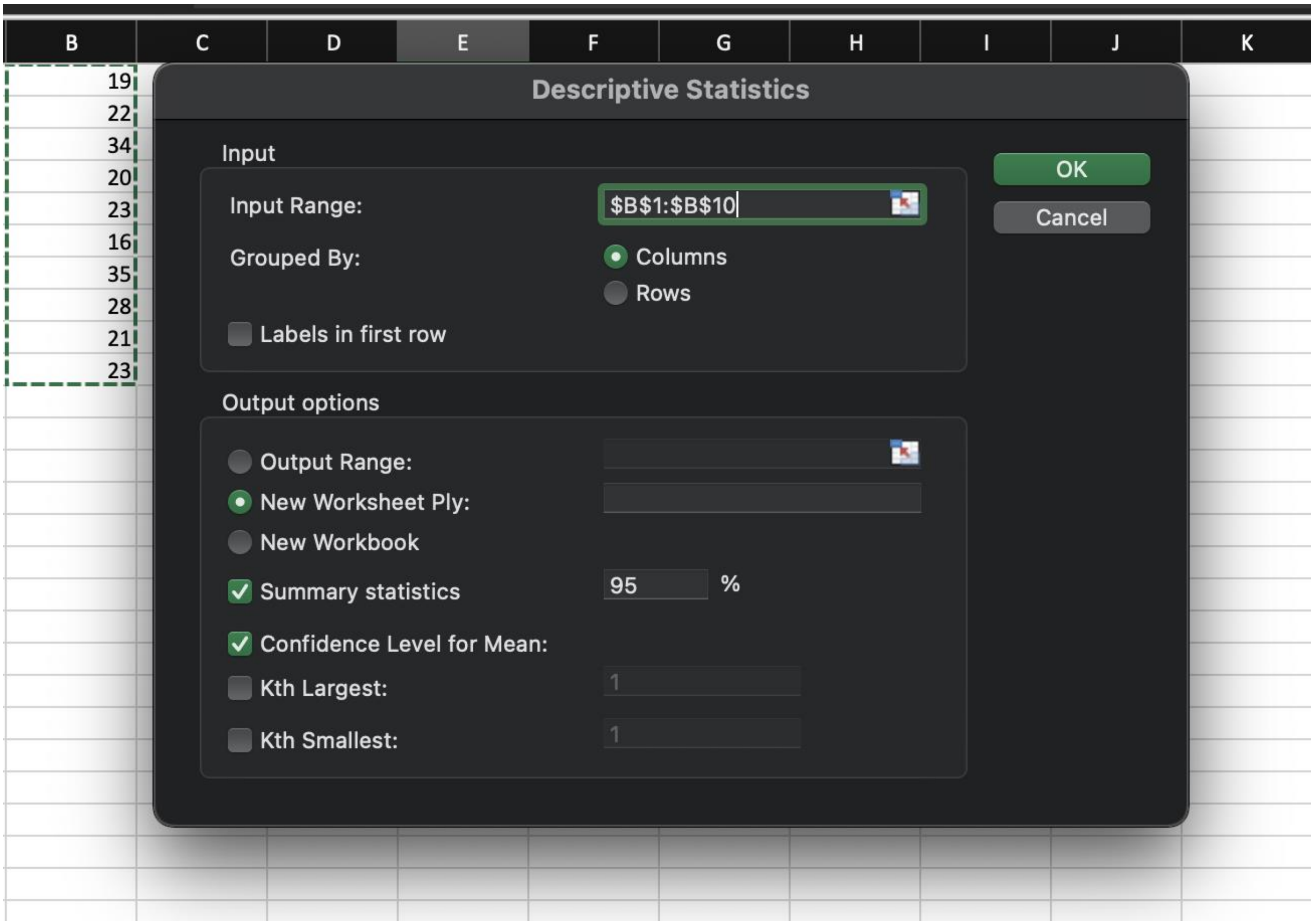
On clicking this opens the window,



A	B	C	D
	19	45	
	22	33	
	34	89	
	20	59	
	23	47	
	16	72	
	35	18	
	28	48	
	21	25	
	23	80	

The choices include Descriptive Statistics, t-tests, ANOVA, Correlation etc.

Clicking on **Descriptive Statistics** leads to another window with options to add the data in a column or row for analysis, along with options regarding the output.



19
22
34
20
23
16
35
28
21
23

Descriptive Statistics

Input

Input Range: $B$1:$B$10

Grouped By: ● Columns ○ Rows

□ Labels in first row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

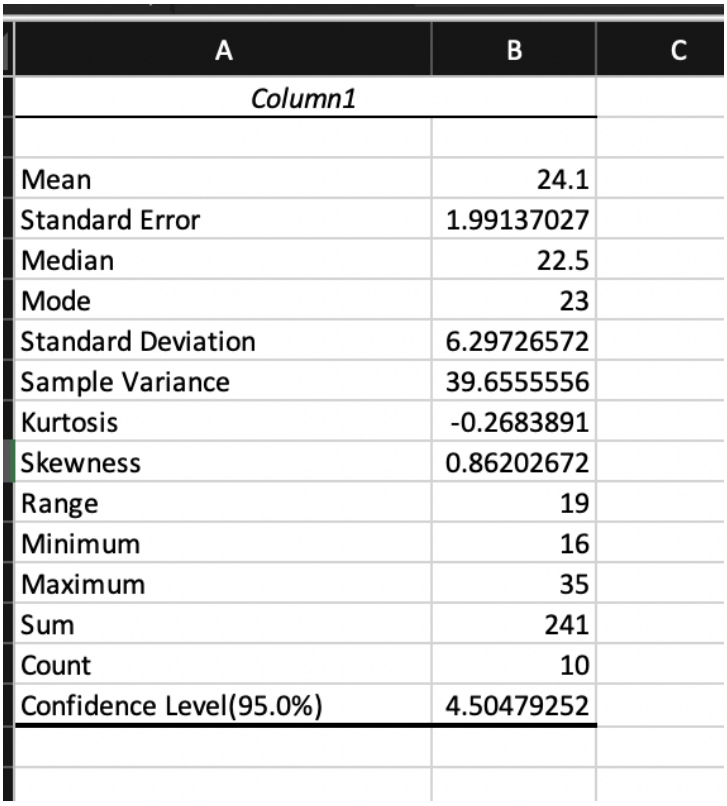
☑ Summary statistics 95 %

☑ Confidence Level for Mean:

□ Kth Largest: 1

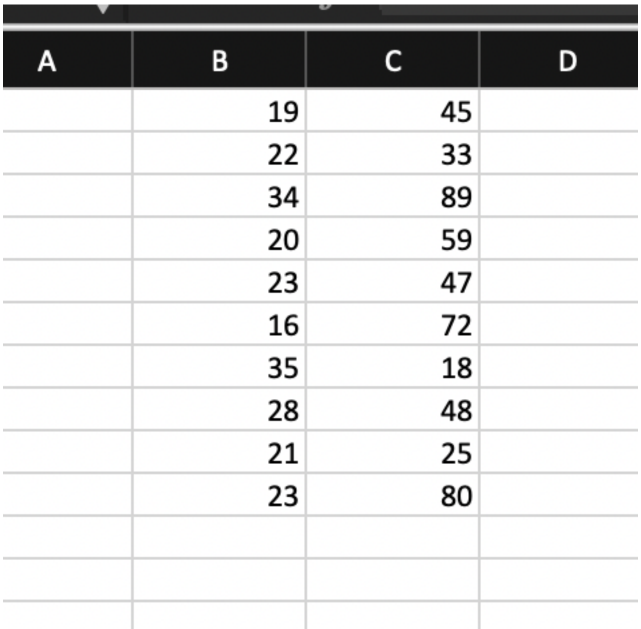
□ Kth Smallest: 1

In the above example the data has been selected in the column **B** and we have chosen the options for ‘Summary Statistics’ and 95% confidence level for the mean. The output is,



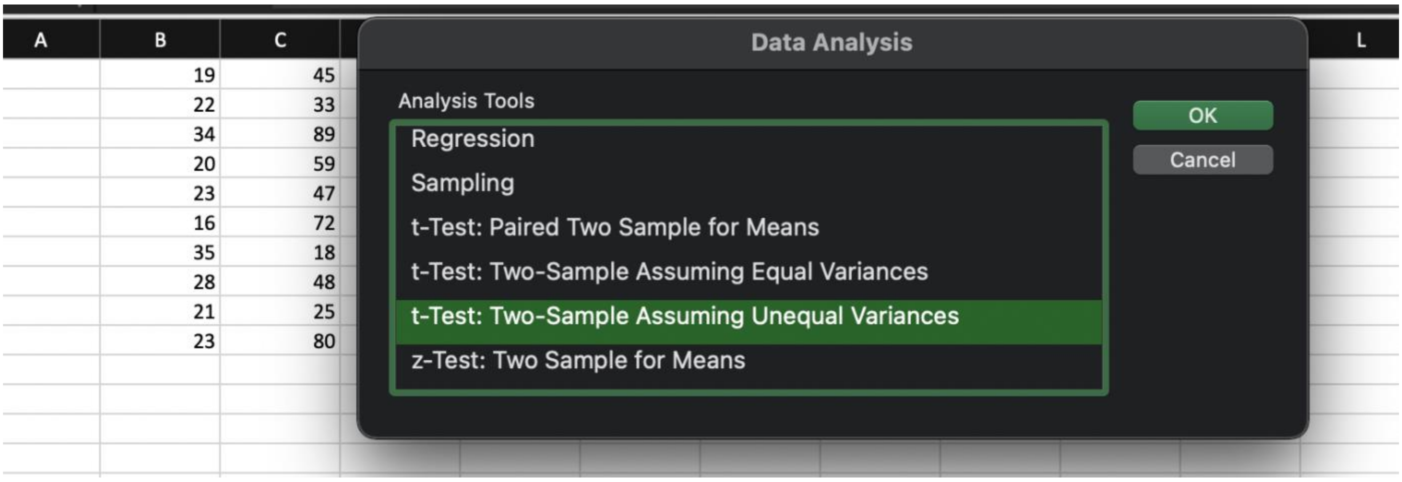
Column1	
Mean	24.1
Standard Error	1.99137027
Median	22.5
Mode	23
Standard Deviation	6.29726572
Sample Variance	39.6555556
Kurtosis	-0.2683891
Skewness	0.86202672
Range	19
Minimum	16
Maximum	35
Sum	241
Count	10
Confidence Level(95.0%)	4.50479252

Moving on to a testing example, say we need to perform a t-test on the following data.



A	B	C	D
	19	45	
	22	33	
	34	89	
	20	59	
	23	47	
	16	72	
	35	18	
	28	48	
	21	25	
	23	80	

Open **Data Analysis** window and choose among the **t-test** options.

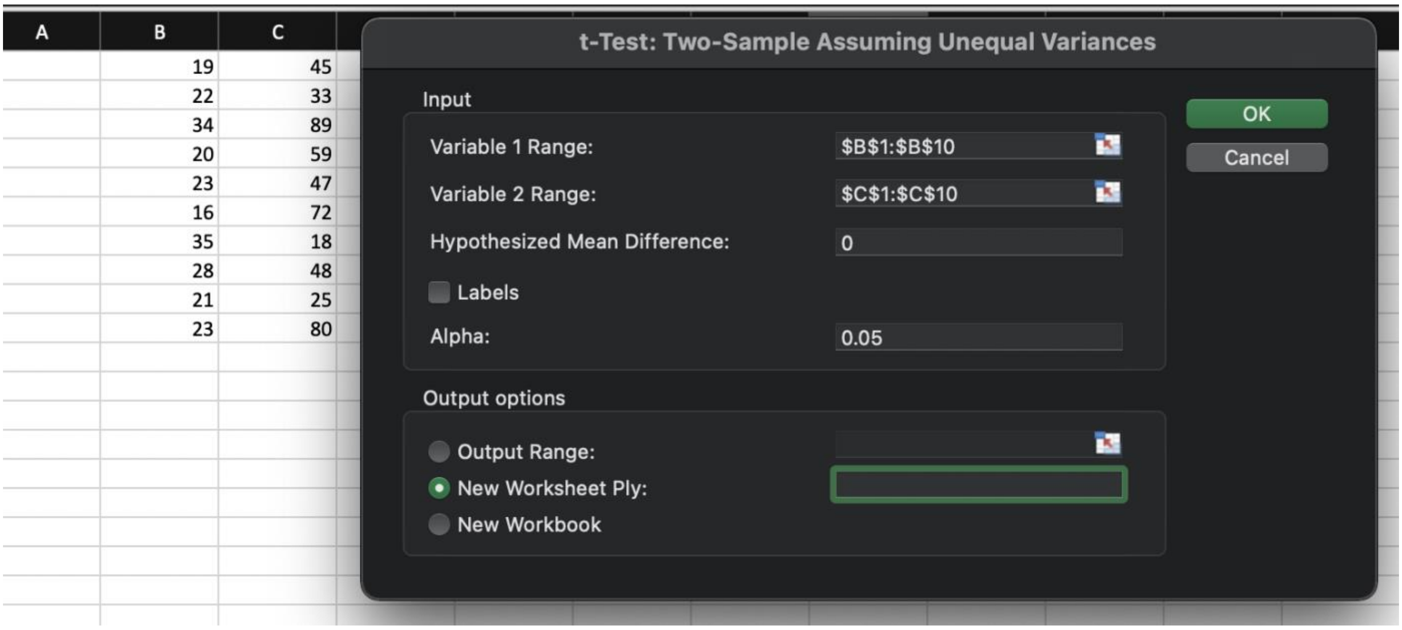


A
B
C
L
Data Analysis
Analysis Tools
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel

A	B	C
	19	45
	22	33
	34	89
	20	59
	23	47
	16	72
	35	18
	28	48
	21	25
	23	80

It is always safe to choose the ’Unequal Variances’ option, unless there are strong reasons to assume that two samples have equal variances. If it is a paired or matched sample, choose ‘Paired Two Sample for means’ option.

The next step is to select the two columns, here **B** & **C**, in the range for inputs. Since we are usually working with a Null Hypothesis that the groups are similar with similar means, the entry for ‘Hypothesized Mean Difference’ is 0. The α is by default set at 0.05.



A	B	C
	19	45
	22	33
	34	89
	20	59
	23	47
	16	72
	35	18
	28	48
	21	25
	23	80

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $B$1:$B$10

Variable 2 Range: $C$1:$C$10

Hypothesized Mean Difference: 0

□ Labels

Alpha: 0.05

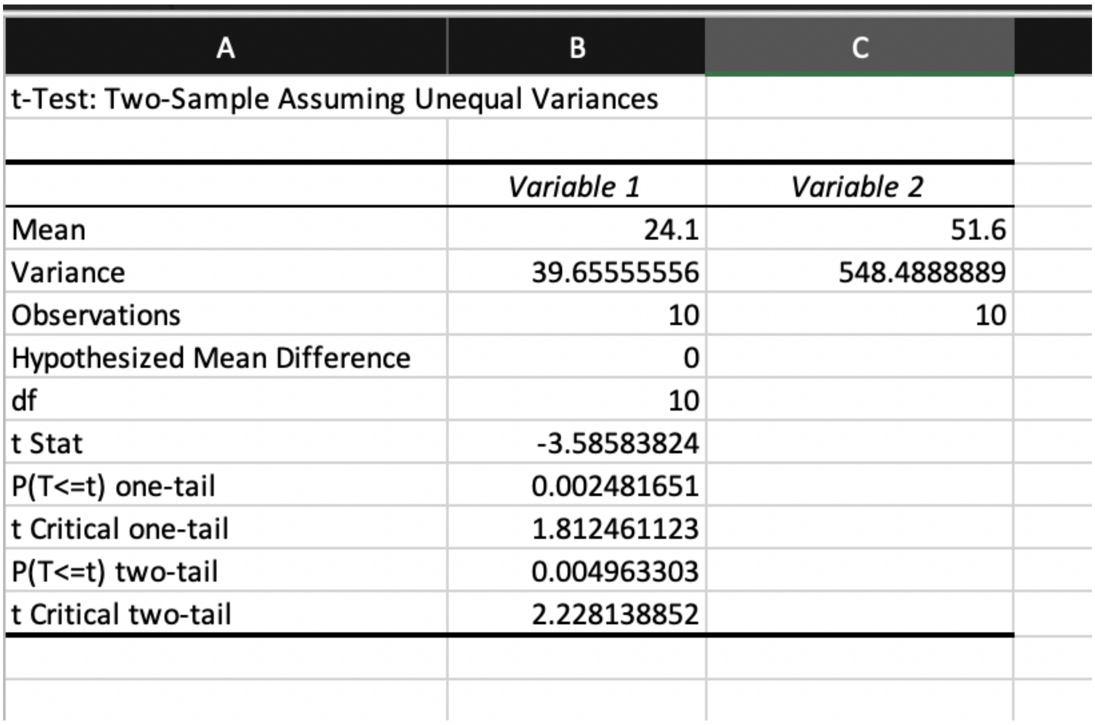
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

The output is,



A	B	C
t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	24.1	51.6
Variance	39.65555556	548.4888889
Observations	10	10
Hypothesized Mean Difference	0	
df	10	
t Stat	-3.58583824	
P(T<=t) one-tail	0.002481651	
t Critical one-tail	1.812461123	
P(T<=t) two-tail	0.004963303	
t Critical two-tail	2.228138852	

We learn that the t-statistic has a value of. -3.585 and that a two tailed p-value is 0.0049, which is statistically significant.

Other tests can also be performed similarly and all it takes is a bit of fiddling to get them right!

Appendix 4 – The American Statistical Association Statement on Statistical Significance and P-values

This section lists the six principles set out in the ASA statement, with a short discussion clarifying their import. It is highly recommended to read the original statement.

Principle 1: **P-values can indicate how incompatible the data are with a specified statistical model**

P-values summarize the incompatibility of the data with a proposed model, usually this being the Null Hypothesis. Hence, P-values tell us the probability of our findings in the setting of the Null Hypothesis and the lower this value, the greater the incompatibility of the data with the Null Hypothesis.

Principle 2: **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone**

This is a tricky, yet critical point. From principle 1, the p-value is the probability of the data *given* the Null Hypothesis is true. It is wrong to say that it is the probability of the Null Hypothesis being true *given* the data that we have. The second (wrong) assumption is often called the *prosecutor's fallacy*, possibly because such arguments have occurred in courts regarding DNA matches from crime scenes. Saying that 'there is a 1 in a million chance of getting such a DNA match if the subject is innocent' cannot be interpreted as 'there is a 1 in a million chance of the subject being innocent'.

Principle 3: **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold**

The threshold for statistical significance at 0.05 or any other value is quite arbitrary and cannot be considered the goal of research. A hypothesis does not suddenly become true on one side of the divide and false on the other side. The use of 'statistical significance' as the only claim to 'scientific truth' is a distortion of the scientific process.

Principle 4: **Proper inference requires full reporting and transparency**

'Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.' Selectively reporting only those p-values which pass the 'significance' threshold, called data dredging/ significance chasing/p-hacking, leads to a false excess of significant results.

Principle 5: **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result**

'Statistical significance is not equivalent to scientific, human, or economic significance'. If large enough samples are studied, even a tiny difference can be statistically significant. This difference may not be of any importance in making clinical or policy decisions. On the other hand, large differences may not pass the threshold of significance in small studies but may actually hold importance in decision making.

Principle 6: **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis**

A p-value without the context of the study design and assumptions, other evidence and arguments etc provides limited information. Data analysis should not end with the calculation of a p-value.

From the Conclusion, *'No single index should substitute for scientific reasoning'*.

Appendix 5 – Only for the Math Enthusiast

This section is for those who feel cheated by not knowing exactly how some of those probabilities and values in this article were calculated. And for the small minority who have been so inspired by all this that they are determined to learn more.

Section2

Calculating the probability of landing exactly 55 heads out of 100 tosses requires the Binomial distribution. The general formula for the probability of *r* events out of total *n* events is,

$$\binom{n}{r} p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where $\binom{n}{r}$ is the number of ways of choosing r events out of n, n! being the products of all whole numbers up to n; p is the probability of the event we are interested in and q the probability of that event *not* happening. The Binomial Distribution is valid where there are only two possibilities, like Heads/ Tails, Red/ Blue pills. When p and q are equally likely *i.e.* p=q=o.5, then the above can be simplified to,

$$\frac{n!}{r!\,(n-r)!}(0.5)^n$$

So, getting 55 Heads out of 100 tosses has a probability of

$$\frac{100!}{55!45!}(0.5)^{100} = 0.04847$$

And getting 60 Heads out of 100 tosses,

$$\frac{100!}{60!40!}(0.5)^{100} = 0.0108$$

Similarly, in Sec 3.3 and Table 1 the above formula is used to calculate the probabilities. For *e.g.,* the probability of drawing 6 Reds out of 10 (assuming they have an equal chance at in each pick),

$$\frac{10!}{6!4!}(0.5)^{10} = 0.2051$$

Section 4

The equation of the normal curve is given by,

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

This equation is the general form for a normal distribution with mean μ and standard deviation σ.

The 'Standard Normal Distribution' is the normal curve with mean 0 and standard deviation of 1. This has the much simpler equation,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-1/2(x)^2}$$

The total area under the curve is 1. The area under *any segment* can then be calculated as the definite integral of the above expression. To calculate the area between -1 σ and +1 σ, for the Standard Normal Distribution, is simply the area between -1 and +1, given by

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 0.68269$$

This means that 68.27% of the total area lies between ±1 σ.

Section 6.3

Sample Size formulae

i) Estimation of a proportion

In the context of a study that seeks to estimate a Categorical variable with proportion p with total width of error E,

$n = 4{Z_\alpha}^2\, p(1-p)/E^2$, where Z_α is 1.96 for the usual 95% confidence level

ii) Estimation of a mean

$$n = 4{Z_\alpha}^2\, \sigma^2/E^2$$

iii) Hypothesis testing of Continuous outcome sample mean vs known population mean

$H_0 \rightarrow \mu = \mu_0$, and $H_A \rightarrow \mu \neq \mu_0$, where μ_0 is the known population mean [historical control].

$n = [\frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})}{ES}]^2$, where ES is Effect Size given by

$$ES = (\mu - \mu_0)/\sigma$$

For the usual α of 0.05 and β of 0.2, this reduces to $\frac{7.84}{ES^2}$

iv) Hypothesis testing of two Continuous outcome sample means

$H_0 \rightarrow \mu_1 = \mu_2$, and $H_A \rightarrow \mu_1 \neq \mu_2$

$$n = 2[\frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})}{ES}]^2$$

For the usual α of 0.05 and β of 0.2, this reduces to $\frac{15.7}{ES^2}$