Basics of Statistical Analysis

Maj Gen (R) Suhaib Ahmed, HI (M)
MBBS; MCPS; FCPS; PhD (London)
Genetics Resource Centre (GRC)

Statistics includes data collection, summarizing, presentation and interpretation. Without analysis of scientific data by statistical methods it could loose most, if not all, of its meaning. In spite of its fundamental importance most medical professionals have a poor understanding of why and how statistics is used. Most books on statistics are themselves to be blamed for this. The authors have been using statistics for so long that they usually forget what it is like not to know statistics at all. Almost all of the books lay emphasis on how to do statistical analysis. Ironically, little stress is laid on "why" there is a need to do statistical analysis?

Population

It literally means the inhabitants of a place for example a village, city, district or a province etc. Let us take the example of a city with a population of 100,000. We can single out all individuals in the city who are less than 15 years of age and call it the "paediatric population". Similarly all individuals in the city, who have a disease, diabetes for example, can be labelled as the "diabetic population". The city could have any number of subpopulations depending on any particular criteria. The "diabetic population" in this example would include all patients of diabetes whether they are symptomatic or not, young or old, rich or poor, educated or not, male or female, black or white, tall or short and so on and so forth. Let us examine a study conducted on diabetics in a hospital of the same city. Blood glucose levels in all diabetics who came to a hospital during one year were studied. The subjects of the study are called the "sample population". The "sample population" in this study is much smaller than the actual "diabetic population". It is generally taken to represent the actual population of patients without realizing that the patients of diabetes who are symptom free or who do not have an access to the hospital for several reasons are not included in the "sample population". A careful sampling can result in minimum of errors. The slight amount of errors due to sample selection can be minimized by statistical analysis. However, no amount of clever analysis can compensate for gross errors in sampling.

In conclusion statistical analysis is required to use the information gained from a "sample of individuals" to make inferences about the "actual population of patients".

Distribution patterns of a population

From the point of view of statistics it is also important to see how the individuals are distributed in a population. The distribution pattern can be normal or abnormal.

Normal distribution

Let us examine the heights of individuals in a village with a population of 1000. Most people would have an average height whereas a small proportion would have very tall or

very short heights. If we plot the heights and the number of people on X and Y-axis respectively we get a bell shaped curve (Fig: 1). This is a normal distribution pattern also called “normal distribution”.

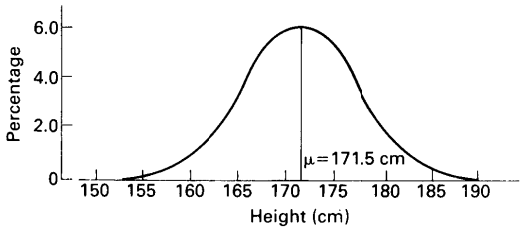

Fig: 1. Normal distribution pattern of population.

The t-distribution

The preceding example is of a very large sample population. If the sample population is small say 30 or less the normal distribution pattern is replaced by another type of distribution called the “t-distribution”. It tends to be more flat (the peak is not as high as in a normal distribution) and it also has long tails at the two extreme ends i.e. right and left.

Abnormal distribution

Imagine a selected population sample that contains predominantly very short or very tall people. The distribution pattern (curve) in such a sample (Fig: 2) would be “skewed” either to the left or to the right. In selected samples, for example Hb of children in a paediatric ward, greater number of children is expected to have low Hb than the normal paediatric population. The distribution pattern in this sample would be “skewed” towards right.

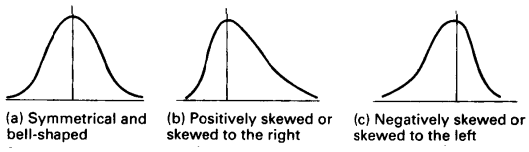

Fig: 2. Abnormal distribution patterns.

Binomial distribution

A scientific observation quite commonly takes only two possibilities e.g. positive or negative, normal or abnormal, male or female. This type of data has a binomial distribution.

Population mean

Every sample, regardless of its distribution pattern, has a mean (average) value. In a normally distributed population it is represented by the arithmetic mean (the centre point on the X-axis in Fig: 1 represents the “population mean”). In an abnormally distributed

population the calculation of mean is more difficult and it is better represented by a geometric mean or the median (midway when ranked in order). Calculation of population mean is of fundamental importance in statistical analysis because in comparing different populations one is actually comparing the means of each population.

Mean of the actual population, for example mean ALT level in hepatitis or mean blood glucose level in diabetics, can be calculated by carrying out analysis on a very large population sample. It can also be calculated by sampling from several different representative places and averaging the individually calculated means.

Standard Error of the mean

Standard Error of the Mean (SEM) is a very useful measurement that shows the difference between the individual sample means and the actual population mean. A sample mean is unlikely to be exactly equal to the actual population mean. A different sample would give a different mean, which is usually due to sampling variations. SEM for a qualitative variable may be calculated by the formula:

$$SEM = \sqrt{p\ (1-p)/n}$$

where p is the frequency of occurrence of the variable and n is the number of observations. Similarly for a quantitative variable the SEM may be calculated by the formula:

$$SEM = s/\sqrt{n}$$

where s=SD of the sample mean and n is the number of observations.

Dispersion (distribution) around the mean

The mean of a population can be made more meaningful if the distribution of the individual values around the mean is also expressed. In fact during comparison of two population means by statistical methods the dispersion around the mean of each population is also taken into consideration. Most commonly the dispersion (distribution) around the mean is expressed as Standard Deviation (SD) or Coefficient of variation (CV). In a normally distributed sample the dispersion around mean that includes approximately 67% of the results is called one standard deviation ($\pm$1 SD). Two standard deviation ($\pm$2 SD) includes 95% of the results while three SD ($\pm$3 SD) includes 99% of the results.

Confidence Interval (CI) of the mean

Confidence Interval (CI) is a range of values in which one is confident that it contains the actual population mean. For example a 95% CI ranging from 21%-27% means that there is 95% chance that the actual population mean lies within this range. In other words we can also say that if the same experiment was repeated 100 times then 95 times the mean of the result would fall within the range 21%-27%. We can calculate CI by multiplying the sample mean with its SEM. However, its calculation is dependent on the sample size. Depending on the requirement one can calculate 90%, 95%, or 99% CI. But most

commonly 95% CI is used. The confidence interval in a large sample (>60) is calculated as follows:

$$95\%\ CI = x \pm (1.96 \times SEM)$$

where x is the sample mean, 1.96 is the 5% point on either side of a normal distribution, SEM is the standard error of the mean.

In a small sample the distribution pattern is different (t-distribution). Therefore calculation of CI on such a sample is also different. In a small sample the degree of freedom is restricted by n-1 (n is the number in the sample). The 5% point on a t-distribution is calculated from the t-table. For example, if the number of sample (n) is 8, the degree of freedom (n-1) is 7 and the 5% point is 2.36. In this case the 95% CI would be:

$$95\%\ CI = x \pm (2.36 \times SEM)$$

Statistics in Practice

Statistics is the science of collecting, summarizing, presenting, and interpreting data, and using them to test hypothesis. Unfortunately, for statistical analysis we tend to concentrate mostly on the testing alone. It is often not realized that the statistical test would produce a correct result only if the data were collected and interpreted correctly.

Data collection

In collecting data one needs to keep in mind the difference between the "sample population" and the "actual population". In a hospital/institution-based study there is a tendency to ignore, consciously or unconsciously, that all patients of a disease would not come to a hospital. For example it is unlikely for a patient to visit a hospital if he/she is not symptomatic, is not educated, is unable to afford the hospital charges, or lives in an area that is far from the hospital. The net result is that the "sample population" examined in a hospital is not expected to represent the "actual population" of patients.

Let us consider a study on the prevalence of a micro-organism resistant to a particular antibiotic. An institution found that 30% of the organisms isolated in the bacterial cultures from infected wounds were resistant to an antibiotic. Extrapolation of this result to the "actual population" could be justified provided it is ensured that samples from all of the infected wounds are referred for culture sensitivity. But it should be kept in mind that culture & sensitivity is requested mostly when the wound infection fails to respond to empirical antibiotic treatment. Therefore it would be inappropriate to extrapolate the results of this study to the "actual population". However, in such studies one can calculate the Standard Error of the Mean (SEM) and then construct a 95% CI that would provide a range of values that includes the mean of the actual population.

Sample size

Another important factor that determines whether the results of a "sample population" can be extrapolated to the "actual population" is the number of individuals tested. The

larger the sample size closer would be its mean to the actual population mean. However adequate sample size means that it should be neither too small nor too large. A common mistake is to examine a small sample size. Let us consider the example of tossing a coin. If one tosses it 10 times it is quite likely that head comes 3 times and the tail comes 7 times. It should not lead to a conclusion that head would come 70% of the times if the tossing is repeated. However, if the same coin is tossed 100 times it is least likely that the head comes up 70% of the times unless there is something wrong with the coin itself! Occasionally one tends to examine a much larger sample size than is actually required. This obviously is waste of time as well as resources. For example, in order to see the frequency of head or tail while tossing a coin, 100 observations should be enough. Extending the observations to 5000 or higher would not alter the results to a significant extent.

Calculation of the correct sample size is of fundamental importance in a study. It can be done through manual calculations or by using computer software designed for this purpose. However, it requires some prior knowledge about the approximate frequency of occurrence of the variable under study.

Expression of variability

Table: 1 shows Hb measurements in two groups of patients each having a mean of 9.1 g/dl. The means of the two groups are similar, however, the distribution of results around the mean is not alike. In practice it is essential that one should express how the results are distributed around the mean. A simple method of expressing variability is to give the range (highest and the lowest values). An even better method is to give Standard Deviation (SD). As seen earlier about 67% of the observations lie within one standard deviation, about 95% lie within two standard deviation and about 99% lie within three standard deviations.

Table: 1. Distribution pattern, and mean values of Hb levels in two groups of patients.

Sr. #:	**Group-1 (Hb: g/dl)**	**Group-2 (Hb: g/dl)**
1	7.3	2.5
2	8.4	4.5
3	8.6	7.2
4	8.7	7.9
5	9.0	8.7
6	9.1	9.6
7	9.6	9.8
8	9.7	10.5
9	10.1	11.7
10	10.2	13.2
11	10.3	14.1
Mean	**9.1**	**9.1**
Range:	**7.3-10.3**	**2.5-14.1**
SD (± 2SD)	**± 0.9**	**± 3.4**

Comparison between groups

The subjects of a scientific study are a population that has its own distribution pattern, mean and a standard deviation (SD). The common objectives of a scientific study are to compare its result with a reference population value or to compare the results of various subgroups of the study population. In such comparisons we are actually trying to compare the means of various populations. The comparison can be done by two different approaches i.e. confidence interval and hypothesis testing.

Comparison by confidence interval:

In comparison by confidence interval we calculate the CI of each group and see if the two ranges are overlapping or not. Table: 2 shows the mean and 95% CI from three study groups. Comparison of Group-1 and Group-2 shows no overlap between the 95% CI therefore the results from the two groups are significantly different. The 95% CI of Group-3 has an overlap with the 95% CI of Group-1 therefore these two results are not significantly different from each other.

Table: 2. Mean and 95% confidence intervals in the three groups of a study.

Groups:	Mean:	95% CI:
Group-1	3.9	3.1-4.9
Group-2	6.4	5.2-7.3
Group-3	3.1	2.1-3.8

A more scientific way of comparing groups by confidence intervals would be to calculate the 95% CI for the difference between the means of the two groups. If the CI does not include zero, for example 1.05 to 3.41, the groups are significantly different from each other. On the other hand if the CI includes zero, for example -1.31 to 1.55, the groups are not different.

Comparison by hypothesis testing:

The comparison by hypothesis testing begins with the “null hypothesis” which states that there is no difference between the groups. If a statistical test proves the null hypothesis correct then there is no difference between the groups. If the null hypothesis is proven wrong then the difference is statistically significant.

P value:

P means probability. Consider the means of the groups shown in Table: 2. If a statistical test for comparing the means of group-1 and group-2 gets a P value of 0.03 it would mean that there is 0.03 (3%) probability that the null hypothesis is true. This would also mean that there is 3% probability (chance) that there is no difference between the groups or there is a 97% probability that the two groups are different from each other. By convention we use a cut off value for P as 0.05. When P value is less than 0.05 we call it statistically significant. In the recent years there has been a move towards quoting the actual P value (P=0.03 or P=0.10) rather than using the < or > signs.

Comparison by confidence interval or hypothesis testing?

The approach for comparison by confidence interval seems so straightforward that it may come as a surprise that why most people do not use it? The confidence intervals show the uncertainty or lack of precision in the estimate of interest and therefore convey more useful information than the P value. However, the presentation of both the actual P value and the CI is desirable, but if one is given the choice the P value may be omitted because it is less important and can also be gauged roughly from the CI.

Parametric and non-parametric tests

Theoretical distributions are described by quantities called parameters, notably the mean and standard deviation, so the statistical methods that use distribution assumptions are called parametric tests. There is another class of statistical tests that do not involve distributional assumptions and are called non-parametric tests. Because these methods are based on analysis of ranks rather than actual data they are also called rank tests. Samples with a skewed distribution are commonly analysed by non-parametric tests. Moreover the rank methods tend to be more suited to hypothesis testing (P value) than the CI methods.