Inferential statistics - 1

B. K. Dassanayake, MBBS (Peradeniya)

Department of Surgery, Faculty of Medicine, University of Peradeniya.

Key words: Inferal statistics.

The previous section of this series dealt with how to describe the data that we collect using various tools. The data that we have collected can be thought of as a 'Sample'. Thus, descriptive statistics stops with describing the Sample. Inferential Statistics attempts to go a step further, and use the sample data to predict certain things about an entire Population.

In summary, Inferential mainly attempts to do two things:

1. Make population predictions using a sample
2. Compare groups values to see if groups are different from each other

We will focus on these two areas separately. However, before venturing further, a few definitions need to be laid out:

Hypotheses testing

The most important difference between inferential and descriptive statistics is that, although we are fully certain about the values we calculate regarding the sample we collect, we can never be fully certain about the things we infer from this sample about the population. As noted in the first section of this series, all we can do is measure the probability of our inferences being true. This is known as testing a hypothesis.

Suppose we wish to see if there is a difference in IQ between males and females. We have collected IQ values, and are planning to use an appropriate test to compare the male and female values. Inferential statistical tests usually end with giving us a 'p-value'. The following process can be followed in all cases where a p value needs to be interpreted:

Step 1. What is our Question: "Is there a difference in IQ values between Males and Females?"

Step 2: Rewrite this Question as a Negative Statement: "There is No Difference in IQ values between Males and Females"

This statement is known as the Null Hypothesis

Step 3: What is the Probability of The Null Hypothesis being True?

The p-value we obtain is thus the probability of the Null Hypothesis being true.

i.e. If the P value is 0.01, this means that The statement "There is No Difference in IQ values between Males and Females" only has a 1% chance of being true.

For most tests, we set the P value at 0.05. This means we accept a difference if the Null Hypothesis only has less than a 5% chance of being true.

Type I and Type II error

By setting the P value at 0.05, we are deciding that we will accept that the Null Hypothesis if the chance of it being true is more than 5%. In other words, if that probability is less that 5%, then the difference between the IQ values is assumed to be significant. However, we are only 95% certain that there is indeed a difference; there is room for error. This is known as Type I error.

Type I Error/Alpha Error: claiming a significant difference between two groups when there actually isn't one. We attempt to keep the probability of this happening low, and as in the above example, we usually decide to set it at 0.05.

Type II Error/ Beta Error: The probability of claiming that there is no difference between two groups when there actually is one. Thus obviously Beta error will be dependent on how powerful the statistical test we use is at finding a difference between the groups. If we use a less powerful test, we will conclude that there is no difference between groups, although there is. Thus, (1-?) is known as The Power of the Test.

Now let us refocus on Inferential Statistics.

1. Making population predictions using a sample

It is usually impossible to measure a certain parameter in an entire population. For example, it would be very cumbersome to calculate the average height of a Medical Student by measuring heights of every single Medical Student in the world. Thus we have to use data gathered from a Sample to predict what the Population value actually is.

Eg:

1. We may have collected data on 120 oesophageal lengths in Sri Lankan males. This is our Sample, and we can calculate Descriptive Statistics (Length, Standard Deviation etc) for our Sample. We can go a step further and use our Sample to answer the question 'what is the mean length of the oesophagus in a Sri Lankan Male?' (i.e, the Population)

Correspondence: B. K. Dassanayake, Department of Surgery, Faculty of Medicine, University of Peradeniya.
E-mail: thraless@gmail.com

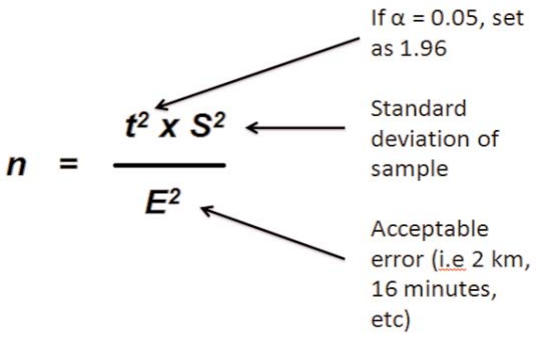
2. Using a Sample of 100 females between 40 and 60 from whom we collect fasting blood sugar values, we can attempt to answer the question What proportion of Sri Lankan women between 40 and 60 are diabetic? This is different from simply describing what proportion of our Sample was diabetic.

Its easy to see that a problem arises immediately: exactly how large should the sample be? This depends on how accurate we need our result to be.

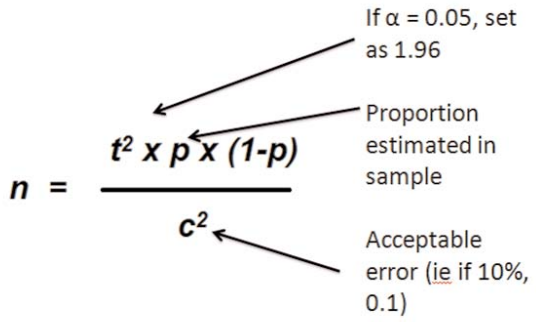
There are formulae which allow us to calculate a sample size depending on how accurate a result we need for our prediction. The formulae vary depending on whether the variable in question is Interval/Ratio data or Categorical Data.

A simple sample size formula for interval/ratio data

Two factors affect the sample size in this formula: First, the standard deviation of the variable in question, as calculated from a small sample of data we collect. (the higher the variability, the larger the sample size). The second factor is the level of accuracy we need. The formula is as follows. (n = sample size)



As mentioned earlier, ? is the Alpha-Error, set at 0.05. The acceptable error is how much we are willing to add as 'plus or minus' to the value we estimate. For example, if the study is being done to assess the average distance a patient travels to get to hospital, the error could be in kilometers. And thus if the sample mean is 10 km, we can predict a population mean of 10 ± 2 km.



A simple sample size formula for categorical data

In this case the following formula can be used:

Here, the value p is the proportion we estimate from a small sample in the study. For example, if our question is to find out what proportion of Sri Lankans are diabetic, and we have collected data from 20 Sri Lankans out of whom 2 are diabetic, that gives p in this formula a value of 0.1. Here, acceptable error is calculated as a percentage (i.e - 34% of Sri Lankans are diabetic ± 8%)

Problems in using the formulae

The most important problem is that the formulae require us to know something beforehand: in case of interval/ratio data, the standard deviation; and in case of categorical data, the proportion. Thus, before we begin to collect data for our study, it is important to first perform a Pilot Study to gather this information from a small sample. The values calculated from the Pilot Study data can then be used to calculate a sample size for the planned study.

The second drawback is that both these formulae assume that there are no other factors affecting the variable in question. For example, age and BMI will have an effect on Diabetes; and thus the study would make more sense if we calculate the proportion of diabetics in separate age and BMI categories rather than overall. This requires sample sizes to be calculated for each of those sub-categories separately. Alternately, more complex formulae may be used to stratify sample size calculations. Many such formulae and sample size calculators may be found online.

This concludes the first aspect of Inferential Statistics. The next section will deal with how to compare groups to see if they are different from each other.

References (also to be used as further reading)

1. Thattil, R.O. and L.H.P. Gunaratne, (2000). Basic Statistics for Advanced Level Students. Godage & Bros. Colombo, Sri Lanka.
2. Thattil, R.O. (1999). Handbook on Survey Design and Analysis. PGIA Publication.
3. Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, Oxford University Press.