

Statistical Concepts Lecture

Bill Sellers

October 12, 1998

1 Summary

This lecture will attempt to introduce (hopefully re-introduce) the basic concepts behind the interpretation and use of statistics. It makes no assumptions about previous knowledge. In fact, given the level of misunderstanding about statistics, the cleaner the slate the better.

The topics I will cover are as follows

1. What are statistics?
2. What are they for?
3. Common problems

What I won't do is give you a cookbook of techniques to cover every eventuality. It's not really possible in any case, and it's easy enough to look up what you need (or ask someone if you are stuck). In addition, there are often several perfectly good alternatives.

2 What are statistics?

Statistics are (by definition) estimates of parameters. Good statistics are good (by which I mean reasonably accurate and unbiased) estimates of parameters, and really good statistics allow the calculation of just how good an estimate of the parameter they are.

This is not a helpful definition until you know what a parameter is, and why you might be interested in estimating it.

Parameters are what we generally think of when we talk about statistics. The parameters often have the same name as the parameter, but often the mathematical symbol used for the parameter is Arabic (such as m being the parameter, the arithmetic mean of a population and μ being the statistic, the arithmetic mean of a sample). More about populations and samples later.

So, parameters are numbers that we can calculate that tell us something about a population. Thus values such as the maximum, minimum, range, arithmetic mean, variance, standard deviation, skew, mode, median, inter-quartile range are all parameters. There are an infinite number of parameters — you can make up your own ones as long as they are mathematically based on measurements taken from members of the population you are interested in. The ones I've mentioned are popular because they are generally useful, but for a particular application you may be interested in, say, the value of the 69th percentile divided by the value of the 68th, and you could calculate it.

Useful measures generally fall into two categories.

1. Measures of central tendency (including various means, median, and mode).
2. Measures of distribution (including variance, standard deviation, skew, range, percentiles and quartiles).

This begs a fundamental question. Why are we interested in these values? Let me illustrate this with a few examples.

We know that not everyone gets paid the same amount of money. Like many people, I am interested in how my salary compares with other people's salaries. I would like to know whether I am better off than the 'average' person. An average is a measure of central tendency. I want to know the value of a 'middling' salary, so I can either feel smug knowing that I earn more than that, or I can complain to my union if I earn less.

However, I am not just interested in the middle value. I am also interested in the spread. If the middle value turns out to be, say, \$15,000, and it turns out that almost everyone earns between \$12,000 and \$25,000 then I shall be extremely cross if I earn \$10,000 and very happy if I earn \$30,000. If, on the other hand, most people fall into the \$8,000 to \$35,000 range then I will be rather less worried about my \$10,000 and less smug about \$30,000.

Obviously, one good way of presenting this information is visually with a graph.

(Draw graphs on blackboard showing a narrow distribution and a wide distribution)

These show the two parameters I am interested in — the value of the central tendency and the distribution. For salaries I might also be interested in the range. Especially the upper range — just how much do the richest people earn?

Now we can begin to look at the difference between parameters and statistics. If I have access to the Inland Revenue's computer database, I could find out the exact salary of everyone who works in the UK. I could decide what values I am interested in. For example, I might choose the arithmetic mean as my measure of central tendency and the standard deviation as my measure of distribution and I could then calculate these values. These would be two parameters that describe this population.

However, it is more likely that I wouldn't have access to this database. Generally speaking, it is very rare to have access to all the data for a population. It is much more common to have data for a sample. For example I could have sent out a questionnaire asking for this information, or I might

have access to the database for Edinburgh only. So this is where statistics come in. It turns out that if I measure the arithmetic mean for the sample, then this is a good estimate of the arithmetic mean of the population — assuming that it's a good sample, but more of that later. The standard deviation of the sample is not the best estimate of the standard deviation of the population. The standard deviation of samples, it turns out, are slightly smaller than the standard deviations of populations so a correction factor, which depends on the size of the sample needs to be applied. This is generally the case — and this adjustment for the size of the sample is often referred to as the number of degrees of freedom.

When dealing with samples — as we almost always do — we often want to know how good an estimate of a population parameter we have. Obviously if you calculate a mean salary from asking a couple of people you met at the pub last Friday night, you are not going to expect it to be a very good estimate. If you asked (and got an answer from) everyone at the pub, you would have a better estimate. However, this throws up the first problem with samples. Is the sample a good sample? Generally, what we mean by a good sample, is a genuinely random selection of members of the population. Is, 'people at the pub on a Friday night' a good selection. I suspect not. It might be convenient, but it is certainly a biased selection. This is quite a common problem. We know that it is a biased sample, but it might be the only sample we can easily get and so we would have to make a value judgement on how good an estimate we have. The problem with this is that someone else might not think very much of our judgement and disagree.

This sort of sample is probably good enough for a pilot study. It doesn't give us a conclusive answer but it does give us a feel for the value. If the population we were interested in was 'pub goers in Edinburgh' rather than 'the adult population of the UK' (because we are doing the study for a brewery, for example), then it is probably a good sample. The

only thing we would have to worry about then is that people may well not have given us honest answers, and quite a few people wouldn't have bothered to reply to our questions. This latter can be a real problem if it later turned out that people on low incomes were rather less likely to want to tell us about it than people on high incomes. We would then find that our estimate of the population mean is higher than we would otherwise expect. Getting an unbiased sample can be very much harder than you might think, and it is important to get a good sample.

Once a population has been well characterised, then rather than random samples, one can use balanced samples. This is what opinion poll companies do. Rather than ask a very large number of people, they ask a smaller number of people but they make sure that this sample contains people from a mix of racial groups, jobs, sexes, ages, regions with the numbers of people representing each of these groups chosen so that the overall balance matches that of the total population. Obviously, for this to work, you need to know in advance the makeup of the total population and which factors are likely to affect the outcome. Obviously, this is very rarely the case.

Once we have a good sample, and have calculated the value of the statistic we can then often precisely calculate how good an estimate of the population parameter it is. For example, if we calculate an arithmetic mean of a sample, we know that this is our best estimate of the population mean. If we use a statistics program to do this calculation, it will also be able to tell us the 95% confidence limits of this value. That is to say, that we can be 95% certain that the population mean that we are trying to estimate is within these confidence limits. Put another way, there is a 1 in 20 chance that the population mean is outside these limits. In biology, because it is a rather imprecise science, 95% is considered acceptably reliable most of the time.

3 What are they for?

3.1 Differences

In experimental science, a very common question is to ask whether a particular treatment has any effect. In this case treatment means any sort of experimental manipulation to the subjects. The general protocol is that the subjects are randomly divided into two (or more than two if several treatments are being investigated) groups and one is given the treatment and the other isn't (the control). Some sort of measure is then taken and the measurements of the two groups compared. We are interested in two bits of information. Is there a detectable effect? How big is the effect? Statistics will only help you answer the first question. This is very important because the second question may be much more important. A very sensitive test may show that there is a statistically significant effect on life expectancy of drinking beer. But when you actually look at the data, you might find that the difference works out as a reduction of half a day over 60 years, which is rather less important than all those days we've lost with hangovers!

The problem we have with 'detectable effect' is that all biological entities vary slightly. Treating one animal/organ/cell never has quite the same effect as treating another one — even when every effort has been made to make them as identical as possible. This means that we would expect every subject to have a slightly different response. What this means in practice is that we calculate some statistics for the two different samples and use them to estimate parameters for the populations they have come from (the populations being, for example, 'all normal mice', and 'all mice receiving growth factor A'). If it turns out that the parameters estimated with their confidence limits overlap, then we have no evidence that the treatment had any effect at all. If they don't overlap, then we have reasonable evidence that we have a real effect.

The statistic that we chose is up to us. We can

easily estimate the mean or the median if we are interested in central tendency, and we can estimate the variance or standard deviation if we are interested in distribution. There are lots of other parameters we could estimate if we wanted, but those are the common ones. The choice depends on what the experimenter is interested in. If you are designing a drug to cure flatulence, then you are more likely to be interested in detecting effects on the mean amount of gas produced, than on any changes in the distribution.

Again, for a pilot experiment, or when the effects are very large and reliable, this sort of statistical analysis may not be necessary. If you add a chemical to a chick embryo and it grows an extra limb, then you can conclude that this effect is due to the drug as long as you can repeat it a few times. Growing extra limbs is extremely rare (considerably less than 5%) in normal circumstances, so even a single occurrence is interesting. However, for quantitative effects (size changes, pH changes or whatever), some sort of statistical analysis is almost always necessary.

Also in this category are ‘goodness of fit’ tests. This is where you have some theoretical ideas about the population, and you want to see whether your sample matches the theory. Your theory can predict values for various parameters and there are a number of tests that can be used to see whether your sample data matches.

3.2 Relationships

The examples so far have been so-called univariate statistics. They have involved measuring one thing on each subject. Obviously we can measure more than one thing — height and weight for example. When we measure two things, we enter the realm of bivariate statistics. Often, we are then interested in the relationship between these two measurements.

In this case, we want to know whether there is a numerical link between a quantity of one thing and a quantity of another. Again, if we have access to

the data for the whole population (unlikely!) then we can calculate various parameters, such as the correlation coefficient, which give us an indication of such an association. More likely, we will once again have a sample and we will calculate the correlation coefficient of the sample and use that as an estimate for the population and once again we will be able to put confidence limits on our estimation.

In the case of correlation coefficients, these vary from -1 for a completely inverse relationship to +1 for a perfect relationship. What we are interested in (what we mean by statistically significant in this case) is that the range of our estimate for the population value does not include 0 (no relationship).

Similarly, we can use the extra information from the extra variable we have measured to increase the power of our tests to identify an effect of the treatment. Often, to do this, the measurements are treated separately and a univariate test is simply repeated twice. However, if the two measured variables (height and weight are good examples here) are strongly related to each other (tall people are generally heavier than short people), then a multivariate test is more appropriate.

4 Groups

Science needs to create groupings. Sometimes these groupings are real — such as whether an animal is male or female — and sometimes perhaps more for convenience such as grouping animals by sub-species. Once groupings exist, it is often necessary to decide whether something belongs in that grouping or not. Usually, groups are identified by associations of features, but often a specific feature is not diagnostic and a group of features need to be considered together to place an individual in a group. An example of this is deciding the sex of a skeleton. We know the skeleton had a sex when it was alive, and nowadays we can identify the sex reliably from DNA probes. However, we can also attempt to identify the sex from differences in the shapes and sizes of the bones which we know vary

between males and females. Unfortunately, none of the specific features can be guaranteed to occur, and the most reliable way to assess sex morphologically is to measure all the features that have some bearing on the sex, such as the angle of the femoral head and the size of the sciatic notch and incorporate all this information in an equation that calculates a single value for the individual, and depending on this value, the skeleton can be assigned, with some known reliability as either male or female. Producing the equation in the first place is a matter of taking these measurements on a number of skeletons of known sex and performing a discriminant analysis.

4.1 Structure

The last thing I shall talk about are so-called exploratory statistics. This is where the calculation of the statistic is an end in itself. The values calculated will hopefully give us an insight into the structure of the data we have measured. A common use of this branch of statistics is in taxonomy where people are investigating the differences between groups of closely related animals. A technique known as principal component analysis (for example, there are others) allows the identification of which of the measurements that have been taken are important for segregation, and plotting these principal components allows the groupings to be visualised, and allows verification that supposed differences are real.

5 Common Problems

5.1 How do I choose a test?

The first thing to think about is which of the four categories you fit into. Are you trying to show a relationship between variables or a difference between two or more groups. Group membership and structure are usually obvious (and difficult — they don't get covered in most standard text books). In experimental work where you have treatments and

controls, you are probably looking at differences (although if you are trying to work out some sort of dose-response you might be looking at relationships). If you are doing survey or observational work, you could be interested in either or both depending on what research question you are trying to answer.

Then you need to decide how you are modelling your data. This may sound strange, but all statistical tests depend on making assumptions about how your sample should relate to the population you are interested in. Almost all tests assume that you are using a random, independent, unbiased sample. Random means that (ideally) you rolled dice or tossed a coin to decide who should be in the sample. Independent means that the other samples you have taken had no effect on this one. Unbiased means that you are not systematically selecting specific individuals (your random selection procedure is uniformly likely to pick any individual — this can be quite insidious). Often tests will also make assumptions about the approximate distribution of the data. The commonest distribution assumed is the so called 'normal' distribution. This is the classic 'bell curve'. Fortunately, much biological data fits this distribution fairly well. Although one can use 'goodness of fit' tests to check that data matches this sort of distribution, generally speaking it's good enough to just look at a plot and see if it's bell curved. Most tests are actually reasonably robust and give good answers even if the data isn't perfectly normally distributed.

For cases where data is clearly not nicely distributed. There are distribution-free tests. These are the 'non-parametric' tests. They rely on ranking the measurements rather than using their measured values so they throw away quite a lot of statistical power — which means that you might need to use twice as many animals or collect field data for twice as long. Also, there are situations where you want to use a specific measure of central tendency. Comparing means requires parametric tests and comparing medians requires non-parametric. However, with some sets

of data that are clearly not bell-shaped, they can be a good choice. Generally you can find a non-parametric partner to most common parametric statistics. So the median is the non-parametric equivalent of the mean, Kruskal-Wallis is non-parametric ANOVA, Mann-Whitney is the non-parametric t-test.

5.1.1 Some examples

In the following examples, identify which group or groups (Relationships, Differences, Groups, Structure) the research question belongs to:

Evaluating the effect of an anti-baldness drug by giving half a group of volunteers a placebo and the other half the drug.

Looking at the effect of alcohol on driving performance by asking volunteers to drink varying amounts and then measuring their success in a virtual reality driving simulator.

Identifying whether changes in shape of hand bones can be used to assess manual dexterity in extinct primates by measuring the hand bones in a variety of living primates whose manual dexterity is already known.

Investigating the differences between the phylogenetic tree generated by looking at mitochondrial DNA sequences with that generated by cladistic analysis of anatomical traits.

Looking at the effect of a neural growth factor by measuring the dendrite lengths of nerves grown with and without the factor.

Looking at the effect of a neural growth factor by measuring the dendrite lengths of nerves grown in a progressively increasing concentration of growth factor.

Investigating whether women are better than men at anatomical science by analysing the results obtained on this course over the last 5 years.

Investigating whether a growth factor enhances collective duct branching by measuring the inter-branch interval on treated and control kidneys.

5.2 Outliers

Outliers are values that just don't fit. Going back to our first example where you are asking everyone in the pub what they earn. And just by an unlucky chance, Paul McCartney happens to be there. Suddenly, out of a group of 40 people all with earnings between the \$12,000 and \$30,000 (which would give you quite a reasonable estimate of the population average), you have someone with a yearly earning of \$5,000,000. Quite clearly, you will get a better estimate of the parameter if you just ignore this value — and that is exactly the correct thing to do. It's not cheating. You clearly have a bad sample if you include this value, and you clearly would have a better sample without it. Of course, you would have to report that you excluded an outlier, and if in a repeat experiment you happened to bump into Sean Connery, then you might have to decide that actually these high earners are representative and shouldn't have been excluded.

5.3 Bias

You have to be really careful about biased data. An undetected bias in your sample will make your conclusions potentially meaningless. With experimental data, this depends on good experimental design. With observational or survey data it is a good idea to validate your protocol. Generally speaking a protocol in this case is a compromise between accuracy and the amount of data that can be collected in unit time (for example, you can record the behaviour of two monkeys at once to double the amount of data collected), so they can often be validated by doing a comparison between the data collected by a slow but thorough method with that obtained by the quick and dirty methodology.

5.4 Hypotheses

Commonly used statistical techniques owe their origin to Fisher. They rely on the concept of testing whether a null hypothesis can be rejected. In this context, they can only identify when things are

different, and not when they are the same. Thus a t-test can be used to say, with a known degree of certainty, that two samples come from populations with a different mean. It cannot be used to say that they come from populations with the same mean — only that no difference was detected. This is an important limitation.

Bayesian statistics (after another mathematician, Bayes), works out the probabilities of chains of events occurring, and can be used to calculate the probability that two samples are from the same population. However, Bayesian statistics cannot (currently) be applied in a cook-book fashion. A full probabilistic analysis needs to be performed for the chain of events which can be very difficult to do. This can obviously be very useful in certain situations and is enjoying a resurgence in popularity.

5.5 Multiple tests

The validity of significance levels assume that only one test was performed on the data to test a specific null hypothesis. The practice of correlating everything with everything, or performing every known statistical test on the data until something significant turns up is invalid. If 95% significance turns up randomly 1 time in 20, if you perform 20 tests, it's very likely to turn up. This sort of data mining is a valid exploratory technique — suitable for a pilot study and for formulating hypotheses. You can't then use the same data to test those hypotheses.

5.6 Graphs

9 times out of 10, statistics won't miraculously reveal anything in your data that you couldn't have seen on a graph. And if you can't see it on a graph, it's not likely to be terribly important in any case. This means that you should always try and visualise your data — preferably in several different ways. It's very easy with most computer programs and almost always worth doing.

5.7 Computers

There is very little point (most of the time) doing statistical calculations by hand nowadays. The best way is to get to grips with a computer package. I would recommend SPSS, but other people like various other programs such as SAS, StatView, MINITAB and others. Excel is not a statistics program. It's quite good for plotting graphs, but is not suitable for anything but the simplest statistics. Saying that, the simplest statistics are often all that is required. Means, t-tests and linear regression are all that most people ever need!

5.8 Interpretation

Interpretation is probably the most important part of statistical analysis. Statistics become very easy when you have a clear research questions. You pose your questions, devise your protocol, collect your results, perform any necessary statistical analysis, and interpret the results in light of the research questions that you initially posed. Remember, that just because something is statistically significant, it doesn't mean that it is at all important.

5.9 Projects

Here are some mistakes that occur in projects:

1. Proving hypotheses (rather than rejecting null hypotheses)
2. Comparing everything with everything until you find a significant result
3. Not graphing the data
4. Not drawing error bars
5. Not saying what the error bars are (standard error, 95% confidence limits or whatever)
6. Putting in lots of tedious working which clearly shows that you don't really know what you are talking about

7. Using Excel for statistics (OK, I'm being a little unfair here)
8. Not linking the statistics to the research questions

6 Help

There are hundreds of books on statistics. Which you choose depends on personal taste. I would recommend one that is tied to the computer programme that you use since it will tell you which test to use and how to do it. There is no need to do any of this by hand any more! There are plenty in the library, and rather more available in various bookshops.

There are also some computer based tutorials. "Statistics for the terrified" is available on the University computers and there are other resources that you can find on the web. The full text of this lecture is on the web, for example.

If all else fails, you can always ask someone to help. This is better phrased as, "can you check that I've done the right thing" and usually gets you the help you need.

7 Conclusion

Hopefully you now have a better understanding of the basic concepts of statistics. Most of you will have to use statistics in your projects — assuming they go well and you actually have some data. About 90% of those will use t-tests to compare means, and maybe 10% will need to use regression and analysis of variance. All of you will have to draw various graphs, and all of you will have to read papers and be able to critically assess the statistics that other people have done and be suitably sceptical about what you read.