Introducing Statistics for Bioscientists

Keith Gregson

February 14, 2007

Contents

1 Why statistics?

Students often ask "Why do we do statistics?" They usually do so because:

- they are convinced that statistics is hard

- they don't perceive any use for it

- they don't like the way statistics is used by the media and politicians (with good reason)

- and they don't like anything which involves numbers

Usually they don't get an answer, because the person they ask also believes a combination of the above!

Consider the following scenarios

1. Two researchers have discovered separate processes which they believe will improve the texture of bread. Both processes would involve similar expense in modifying the baking process which is currently in use. When asked by the production manager about the prospects of their research "A" says that he thinks his process will improve the bread while "B" says that he is 95% confident that his process is indeed an improvement. If you were the production manager, responsible for the expense of changing the process, who would you take most notice of?

2. If you were given the choice of two medical treatments, both of which may produce unpleasant side effects. One of the treatments affects 1 in 1000 people, while the other affects 1 in 100000. Which would you opt for?

This is why we do statistics; to give an objective value to things which are not certain. "Maybe" and "possibly" are not good enough to help us make important decisions, and beware the character who is certain!

Sometimes chance doesn't really matter - "Heads we go to the pictures tonight, tails we go to the bar", but sometimes it does!

Variation - the difference between individuals

I have often heard so called "Hard Scientists"(you know who I mean) tell us that

> "If we can't measure something accurately enough, then we should build a better meter or do the measurement more carefully."

The same scientists are able to measure the distance to the Moon or Mars to within half a cat's whisker, thus making it difficult to argue. However they don't usually get their feet wet or their hands dirty, and nor do the Moon or Mars behave in any random way. Living things do! So while the *hot-shots* spend their time trying to solve the three-body problem[1] (most biological systems

---

[1] the analytical solution of three objects moving under gravity - should be simple but isn't!

have millions) we have to make do with what we can extract from experiments which are never repeatable: and hence which give different results each time, irrespective of how much care we take. Thus it is necessary to accept and to deal with variation.

Because our subject material has its own inherent variation, we know that the results of an experiment will show some differences, whatever we do to the treated group. The problem that we have therefore, is to identify the variation due to the treatment as opposed to that of the inherent variation.

In order to do this we usually compare our experimental results with a group of samples which have not had the treatment (the *control* group). We then evaluate the probability that our experimental results came from the same population as the control group.

In order to calculate the probability we often have to evaluate fairly complicated statistics (numerical values such as mean, correlation coefficient *etc.*) which we can then compare with tables that have been produced for us by theoretical statisticians/mathematicians. In effect they have simulated similar experiments many times in order to calculate the probabilities of all possible results. From these tables we are able to predict the probability that our experimental results were achieved by chance. If this probability is small (significant) we can be justified in assuming our experiment had some kind of effect.

Probability - the predicted likelihood of an event occurring

In order to gain some understanding of probability we shall begin by examining the simple process of tossing a coin. What is the likelihood of a coin coming down heads, assuming it's got a head on one side and a tail on the other? Yes! 50% or 0.5 (which I prefer). We can reach this conclusion because there is an equal chance of a head or a tail and the probability of any result (i.e. a head or a tail) is 1. For the purist (or pedant!) I have ignored the possibility of the coin landing on its edge, or never coming down, etc. If any of these happen simply toss the coin again or use another - preferably not yours!

Suppose that you have just spun the coin three times and seen three heads. Now what is the chance of: a head? a tail? *"The law of averages"* tells "Joe Public" that we should be getting a tail by now (so as to balance things out!), so the chance of getting a tail must be bigger. BUNKUM! or cobblers, or even rubbish, whichever you prefer. The law of averages (if it exists at all) does not work in this way. The probability of tossing a head with a fair coin is 0.5, however many heads or tails have appeared before. After a long (think of a number and then double it!) sequence of tosses we would expect to see roughly the same number of heads as tails, but we would not expect to see head, tail, head, tail, etc. otherwise the results are not random and we would be able to predict the outcome.

Lesson 1: **beware the law of averages**, it's not a law - simply an expectation. In this case 'that we would expect to see roughly the same number of heads and tails', but we can not predict the order in which they will occur.

We can extend the same argument to the National (C)lottery. Here the probabilities are slightly different (approximately 1 in 14 million [2] of winning, and therefore $1 - 1/(14$ million$)$ of not winning.

If you stand in the shop queue listening to people intent on buying their lottery tickets you are almost certain to hear "I've been using the same six numbers ever since the lottery started, it must be my turn soon (by the law of averages)". They dare not miss now because their combination must come up shortly - NONSENSE! The probability of any one combination coming up in any one lottery draw is always the same (1 in 13983816). End of lesson 1, and you've saved a pound a week already.

We can extend our understanding of probability by looking at how to combine probabilities.

For example we could consider the likelihood of two coin tosses producing two heads. This is simply the probability that the first coin produces a head (0.5), times the probability that the second coin produces a head (0.5). Thus the result is 0.25. An alternative method of evaluating this is to calculate the ratio of the number of ways in which we can obtain the required result divided by the total of all possible results (1/4).

A more interesting question is, "What is the probability of obtaining a head and a tail when both coins have been tossed?" In this case we don't specify the order, a head and a tail is just as good as a tail and a head. Thus two of the four possibilities meet our criterion and therefore the probability is 0.5 This type of situation, where the result may consist of several different permutations of results is known as a combination. In this case the combination is the set of results (HT, TH) which can make up the result of one head and one tail.

Unfortunately the word combination is misused in many ways, for example if you were told that the combination of a safe was 213 you would not expect to open it by typing 123 or 321 or any of the other possibilities (permutations). Thus the type of lock on this type of safe should properly be called a permutation lock. Combination locks would be much easier to open!

We could look at other examples, many text books do, and explain how statistics had its birth in various card games, but I don't intend to do this. I gave up playing bridge because of the endless arguments about probabilities which erupted after the games. To my mind, games should be fun, and while I enjoyed the mental athletics at the time, I couldn't subscribe to the serious chat afterwards, much better to have a beer.

However, when it comes to performing experiments and analyses which result in applications of drugs, food manufacturing, *etc.* where serious consequences may result, we have an obligation to get it right. Thus we need to evaluate our

---

[2]The probability of winning the N.L. (predicting 6 numbers from 49) is the combination of the probabilities of six successive events: the probability of drawing one of your six selections from the 49 balls, the probability of drawing one of your remaining 5 from 48 ... and the probability of drawing your last selection from the remaining 44. Here is the calculation:

$$\frac{6}{49} \times \frac{5}{48} \times \frac{4}{47} \times \frac{3}{46} \times \frac{2}{45} \times \frac{1}{44} = \frac{1}{13983816} \approx 0.00000007$$

results and to perform relevant statistical analyses in order to apply our science as safely as possible.

Making use of Statistics

The process that we go through when trying to evaluate the results of our experiments/research, comprises the following steps.

1. - Understand the system.

2. - Design an experiment/survey/observation, including the means of analysis.

3. - Carry out the experiment.

4. - Calculate the probability of achieving the observed results.

5. - Consider the odds to decide whether our results could be due to chance alone. If the probability is significantly small we can assume that the results were due to our experimental treatment, otherwise our treatment can not be assumed to have an effect. There is no shame in failure, provided it has been achieved with rigour and honesty. It is often as important to know that a treatment does not have an effect as it is to know that it does.

The above is the answer to "Why do we do statistics?" It gives us a quantitative (probability of success) evaluation of the situation rather than a qualitative ("I think it's a good idea"). Your hunches may be good, but they will be better backed up with a good statistical analysis.

2 Basic Statistics

Mean

The mean of population (μ) is calculated using the formula:-

$$\mu = \Sigma x/n$$

n.b. Usually, we are unable to calculate the value of μ because we don't have access to all (n) the individuals in the population, and so we have to make do with an estimate ($\bar{x}$) calculated from a sample containing n values

$$\bar{x} = \Sigma x/n$$

where n is now the sample size.

Variance

The variance of a population with a mean μ and containing n observations is

$$\sigma^2 = \Sigma(x - \mu)^2/n$$

However, we do not usually know μ and so must use an estimate $\bar{x}$ instead. When we calculate the variance this way we have to adjust the formula a little bit. (because it gives answers which are slightly too small, there's some mathematical jiggery pokery here, but we can prove it if necessary!) The best *estimate* of the variance (s^2) that we can calculate from a *sample* of size n turns out to be given by the formula:-

$$s^2 = \Sigma(x - \bar{x})^2/(n - 1)$$

Note that the denominator is not n. We use $n - 1$ degrees of freedom, arguing that one degree of freedom has been lost in estimating the mean. Note also that it is therefore impossible to estimate the variance from a sample of 1!

Further it will be useful to note that $(n - 1)s^2 = \Sigma(x - \bar{x})^2$

Standard Deviation

The standard deviation of a population (σ) is the square root of the variance (σ^2).

The estimate of the standard deviation (s) is the square root of the variance estimate (s^2) and is a measure of how much the sample values vary about the sample mean.

Standard Error (of ...)

Usually when we deal with the standard error we are concerned with the standard error of a derived value - e.g. the mean of a sample, or the difference between two sample means as in the t-test.

The standard error of a sample mean is given by:

$$SE_{mean} = \frac{s}{\sqrt{n}}$$

where n is the sample size.

Sometimes the term is used on its own to mean the standard error *of the sample*, which is the same as the standard deviation.

3   The Statistical Method

1. State the experimental hypothesis - e.g. "Whizzo" petrol additive increases the mpg.

2. Define the critical probability value, usually 5%, 1%, or 0.1% but can be any relevant value, depending upon the importance of the conclusion.

3. Define the null hypothesis as "**all possible alternatives** to the experimental hypothesis"- e.g. "Whizzo" petrol additive either decreases mpg or makes no difference.

4. Observe the results

5. Perform the relevant test to calculate the probability of making the observation if the null hypothesis is true.

6. Reject the null hypothesis if this probability is less than the critical value.

4 The Normal Frequency Distribution

The normal distribution is fundamental to much of modern statistical analysis. We need not concern ourselves with the underlying mathematical theory. Suffice it to say that many naturally occurring phenomena may be described in terms of the normal distribution. In addition, measurements which intrinsically are not *normal* may be combined by taking the mean of several values to give a close approximation to normality. Thus the assumption, which is often made, that a population follows a normal distribution is generally a reasonable one.

The normal frequency distribution is a bell shaped curve which is symmetrical about the mean (μ). Many values are observed close to the mean, while fewer occur as we move away from it.

roughly 15% values are larger

roughly 2% are larger

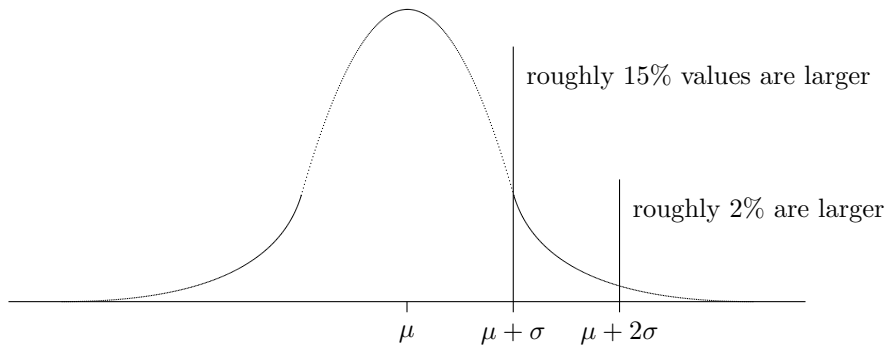$\mu \qquad \mu + \sigma \qquad \mu + 2\sigma$

Figure 1: The Normal Distribution

The y axis represents the probability of observing the corresponding value of x and thus the area under the curve is unity. The areas under the curve to the right of the vertical lines represent the probabilities of recording values of x greater than $x = \mu + \sigma$ and $x = \mu + 2\sigma$ respectively.

The flatness (or otherwise) of the curve will depend upon the variation of the data. If the variation is small the curve will be sharply peaked, if the variation is large the curve will be reminiscent of a Lincolnshire Alp! Two properties of importance therefore are the variance and the mean.

The mathematical equation of the curve is

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

where μ is the mean and σ the standard deviation. The standard normal distribution is a normal curve with a mean of zero and standard deviation unity. The equation for the standard normal curve is therefore

$$y = \frac{1}{\sqrt{2\pi}} \; e^{-z^2/2} \text{ where } z = \frac{x - \mu}{\sigma}$$

The probability of obtaining values greater than z are obtained from standard normal distribution tables or computer programs.

9

Application

A bottling process is set up to dispense 501cc of liquid, with a standard deviation of 0.5cc, into marked 500cc bottles. Each bottle is capable of holding 502.5cc of liquid before overflowing.

In a batch of a thousand filled bottles:

1. How many bottles will suffer spillage?

2. How many bottles will contain short measure?

Question 1 requires us to find the probability of dispensing quantities greater than502.5cc. This is the same as calculating the probability of obtaining values greater than 3 standard deviations above the mean $((502.5 - 501)/0.5)$ from the standard normal curve. The computer informs me that this probability is 0.00135 and hence we should expect a spillage once in every thousand bottles filled.

Question 2 requires us to calculate the probability of obtaining quantities less than 500cc. This is the same as asking what is the probability of obtaining a value more than 2 $((501 - 500)/0.5)$ standard deviations nelow the mean of a standard normal distribution. Since the normal distribution is symmetrical, this is equivalent to finding the probability of obtaining values greater than $+2$. This probability is 0.0227 and thus we should expect 23 bottles in each batch to be under-filled.

Who would be interested in the answers to the above and why?

5 The t-test for the difference between two means

A common statistical task is the comparison of means of two samples to establish if there is a difference. Perhaps we have a number of plots, n_1 of which have been treated with a new fertilizer, and n_2 are untreated plots. In this case we would like to know if the fertilizer has any effect on the yield. Typical results would be recordings of yield from random plants within each plot.

The test statistic to use in this case is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

where $\bar{x}_i$ is the estimate of the mean for sample$_i$

n_i is the no. of observations for sample$_i$

s^2 is the estimate of variance of the populations

$= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

with $n_1 + n_2 - 2$ degrees of freedom

this is the weighted mean of the two sample variances

and s_1^2 and s_2^2 are the variance estimates for each of the samples.

This looks a complicated statistic but, mathematics apart, the composition of the statistic is easy to see. Basically it comprises the difference between the two means (the top line) divided by the pooled estimate of its standard error (the bottom line).

$$t = \frac{\text{difference between the means}}{\text{standard error of the difference}}$$

This latter is a measure of the difference we would expect due to random effects if the samples were from the same population.

Thus, the bigger the difference between the means - the bigger the value of t, and the bigger the standard error - the smaller the value of t. A large value of t leads us to believe that the population means are not the same.

Values showing the probability of observing values greater than t for specific values of n (degrees of freedom) are to be found in tables (or via the computer!). This tells us what the probability of obtaining a value greater than or equal to t would be if the two samples came from the same population. Thus a t value of 3.365 with 5 degrees of freedom indicates a probability of 1%, i.e. there is a chance of 1 in 100 that these two samples came from the same population.

As with most statistical tests we make the hypothesis that there is no difference between the population means - *the null hypothesis*. The probability of obtaining the observed result is then a good measure of this assumption i.e. if the probability were 1 in a million we would reject the null hypothesis and conclude that the measured difference was "real" and therefore due to the experimental treatment. However if the tables showed a probability of 1 in 10 we would find it difficult to disbelieve the null hypothesis, and we would conclude that there was no difference between the samples.

5.1 Performing a t-test

Assume that we have observed data for two samples A and B. We would perform a t-test to verify one of three possible experimental hypotheses :

1. The average of sample A is larger than that of sample B :

   1. Calculate the means $\bar{A}$ and $\bar{B}$.
   2. If $\bar{A} \leq \bar{B}$ accept the null hypothesis, obviously! Otherwise -
   3. Calculate the t-value $= (\bar{A} - \bar{B})/SE(\bar{A} - \bar{B})$
   4. Look up the probability p via the computer or in tables. p is the probability of obtaining a larger value of t by chance when the null hypothesis is true.
   5. If p is less than the critical value (typically 5%, 1% or 0.1%) accept the experimental hypothesis, otherwise accept the null hypothesis.

2. The average of sample A is smaller than that of sample B : reverse the sample names and proceed as above.

3. The average values of samples A and B differ : In this case we must consider both the above cases since we are not concerned which of the samples is larger. Hence we must include both possibilities in our calculation. This is done as follows:

   1. Calculate the t value for the case that generates a positive value.
   2. Look up the probability p corresponding to t which is the probability of observing a value of t greater than the observed value by chance if the null hypothesis is true. We must also take into account the possibility of the observed value being less than $-t$ which will also be p. Thus the overall probability of observing a value which is either larger than t or smaller than $-t$ will be $2p$.
   3. If $2p$ is less than the critical value (typically 5%, 1% or 0.1%) accept the experimental hypothesis, otherwise accept the null hypothesis.

A note of caution to advanced t-testers

In the above we have *assumed that the variance is the same in both samples* and that *the samples are independent of each other*, this is usually the case.

However, there are adapted versions of the t-test which should be used when the sample variances are different (e.g. the size of doughnuts in New York and London) or when "paired" measurements are taken on individuals before and after treatment (e.g. weights during a dieting regime)

6   The χ^2 Test for Frequencies

Sometimes the observations from our experiments can not be transformed or expressed within the framework of the normal distribution. One such case is that of recording frequencies of events or occurrences, for example the number ofmales/females in litters, or the number of heads/tails in sequences of coin tosses. In these cases we resort to a different test known as the χ^2 test. This test is not based upon any underlying data distribution, it is one of many so called *non-parametric* tests.

Consider the situation in which seeds with an "assured" germination rate of 90% only 210 of 250 seeds did in fact germinate. Would we be justified in complaining to the seed company?

The appropriate non-parametric test in this case uses the $\chi^2 statistic$ which is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where n is the number of possible results, O_i are the observed frequencies and E_i are the expected (predicted) frequencies.

In our case we have two possible results and their expected frequencies are 0.9×250 and 0.1×250 respectively so we have:

$$\chi^2 = \frac{(210 - 225)^2}{225} + \frac{(40 - 25)^2}{25} = 10$$

There is a large family of χ^2 distributions based upon a single parameter called *degrees of freedom*. Its value in this case is 1 and so we could look up the value 10 with 1 degree of freedom and find that the probability of taking a sample from a population whose germination rate really was 90% and getting a χ^2 value greater than or equal to 10 would be 0.002. This probability is so small that we do indeed have some justification for complaint.

6.1   Degrees of Freedom

The concept of *degrees of freedom* is a difficult one. In this case it can be thought of as a measure of the flexibility which the system has to generate results. For example in the above problem a complete set of results can be generated by counting the germinated plants, since if we know the total number of seeds and the number which germinated we can calculate the number which did not germinate without having to count them. Indeed it is highly unlikely that we dug up the seeds to count the ones which did not germinate! Thus, of the two possible outcomes for each seed, we need only count the occurrences of one of them. The value for *degrees of freedom* is therefore 1.

In general in a situation with n possible outcomes we need only count $n-1$ of them, since the remaining one must be consistent with the overall total. Hence the degrees of freedom in such a trial is $n - 1$.

6.2 Contingency Tables

When each member of a sample is classified according to two properties the observed frequencies can be expressed in a contingency table. As an example consider the following survey in which we have classified individual people according to eye and hair colour:-

		Hair Colour			
		Blonde	Brown	Black	Totals
Eye	Blue	80	90	40	210
Colour	Brown	20	210	10	240
Totals		100	300	50	450

We may wish to know whether eye and hair colour are related or not. In order to do this we perform the usual statisticians trick of assuming that they are not (the null hypothesis - H0) and then see if the data is consistent with this assumption. Thus we need to see if the observed values differ from the expected (which we calculate making the assumption that there is no difference). We are obviously in a situation where the χ^2 test may prove useful.

Now, if we are to use the χ^2, how do we calculate the expected values? In this case we have no underlying theory to help so we must rely on our assumptions and a bit of common sense. Let us look at blue eyed blondes (who's biased?) as an example:-

Looking at our table we can calculate

1. the probability of being blonde $= \frac{100}{450}$. Remember that we are assuming that eye colour does not influence hair colour and that we found 100 blondes in our sample of 450.

2. similarly the probability of being blue eyed $= \frac{210}{450}$

Therefore the expected number of blue eyed blondes is:-

$$\frac{100}{450} \times \frac{210}{450} \times 450 = 46.66$$

In order to ease the calculation you can rearrange this as:-

$$\frac{\text{total blonde} \times \text{total blue eyed}}{\text{grand total}}$$

Basically, if we assume that the two factors are independent, we use the row and column totals because they are the best values that we have available. In the same way we can find the expected values for brown eyed blondes.

$$\frac{\text{total blonde} \times \text{total brown eyed}}{\text{grand total}}$$

etc.

Thus we can construct a table as follows:-

		Hair Colour			
		Blonde	Brown	Black	Totals
Eye	Blue	80 (46.7)	90 (140)	40 (23.3)	210
Colour	Brown	20 (53.3)	210 (160)	10 (26.7)	240
Totals		100	300	50	450

where the figures in brackets are the expected frequencies.

We calculate the χ^2 statistic using the six observed and expected frequencies from the table.

$$
\begin{aligned}
\chi^2 &= \frac{(80 - 46.7)^2}{46.7} + \ldots \\
&= 100.45
\end{aligned}
$$

The value for *degrees of freedom* in the case of contingency tables is calculated as follows:

$$df = (rows - 1) \times (cols - 1)$$

We thus look up the value 100.45 with 2 degrees of freedom. Tables show that this value is much larger than $\chi^2(2)$ at 0.001. We therefore conclude (with confidence in excess of 99.9%) that the null hypothesis must be rejected and that hair and eye-colour are therefore related in some way.