Statistics for ISBM

Michael Stewart

November 2003

Contents

1 Introduction

There are many sayings about statistics - 'Figures don't lie, but liars figure' is one. It is true that you can misuse statistics to prove anything, and this is why a basic grasp of statistics can help us to work out if a claim is valid, or an abuse of mathematics.

2 Types of data

- Quantitative (numerical)

  Continuous *Any value in a range is possible - e.g. height*
  Discrete *Set values only are possible - e.g. blood group*

- Qualitative (categorical)

  Nominal *Several options available - e.g. favourite colour*
  Ordinal *Several ordered options available - e.g. is the pain mild, moderate, severe*
  Ranked *List of options to be placed in order - e.g. rate these restaurants in order of preference*

3 Measures

3.1 Location

A measure of location is intended to give a value in some way representative of a set of numbers. It is commonly referred to as the 'average'. Three measures are used - mean, median, and mode. Each has strengths and weaknesses.

Mean - $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ where i is a reference number, used to keep track of which data point is being looked at. $\sum_{i=1}^{N} x_i$ is the 'sum over all i' - that is, the total of all the data points. Hence the mean is the total of all data points, divided by the number of points that exist. Usually what is meant when someone refers to 'the average'.

Median - if the data points are put in order, the middle one. If there are two in the middle, take the mean of these.
Mode - most commonly occuring value.

The mean is arguably the most representative if all numbers in the set are 'fair'. For instance, if we consider length of hospital stay after an operation, the mean gives a useful measure unless one patient who had pre-existing problems can't go home because of these for a year. This one patient will distort the mean value. In this case, the median is more useful, as it is relatively insensitive to outliers - the one or two values that lie far away from the others and are not truly representative. The mode may not be representative in the same way, but has a use still. If the set of values are shoe sizes, the modal value might be the best size to keep a spare pair of slippers in!

3.2 Spread

The spread indicates how variable a set of numbers are. It can also indicate how good a representation of a set that a measure of location is - the smaller the spread, the more representative the location is of each number in the set individually, as well as for the set as a whole.

Range - highest value subtract the lowest value
Interquartile range (IQR) - if the data points are put in order, the one $\frac{3}{4}$ of the way through, subtract the one $\frac{1}{4}$ of the way through.
Variance (s^2) - $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. However, it can be calculated more easily using the rearranged formula $\frac{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2}{n-1}$. It is a measure of how spread out the data is around the mean.
Standard deviation (s) - $\sqrt{(s^2)}$
Standard error (SE) $SE = \frac{s}{\sqrt{n}}$ This can be understood as follows - the bigger the standard deviation (s), the less reliable the results are (bigger SE); the more samples that are used (bigger n), the more reliable the results should be (smaller SE).

3.3 Populations

The measures of location and spread detailed above are calculated for a *sample* of a larger *population*. The calculated values are then used as an estimate of the values for the population as a whole.

Population mean $\mu = \bar{x}$
Population standard deviation $\sigma = s$

4 Graphical

Histogram area of each box is proportional to the number of data points falling in the range it covers. If each block is of equal width, height is proportional to number of points *(I think this will always be the case for ISBM)*
Box plot the extent of the lines gives the range. The box shows the IQR. The line in the box gives the median
Stem and leaf easy to find the median from it
Scatter allows you to plot two variables against each other
Pie chart size of each segment represents the fraction of the total number of points in that category. Size of angle = 3.6x%

5 Confidence Intervals

Noted above that we use the results from our sample to estimate the values for a population. It is unlikely that the mean of a sample will be exactly identical to the mean for the population as a whole, so confidence intervals are used as a way of setting limits on the values. A 95% confidence interval means that we can be 95% sure that the true value lies within the range given. Put another way, if we draw 100 random samples from a population, and calculated the 95% confidence interval on the mean for each, then 95 would include the true population mean, and 5 would not.
In general, we work them out as:

$\mu = \bar{x} \pm$ *(test statistic) SE*

We use the sample mean as the central point, then the range is given by the standard error (so the less reliable are results bigger SE the larger the range of values we need to include to be 95% sure we have the right one), multiplied by a test statistic.

The value of the test statistic depends on the number of degrees of freedom (in this kind of test, df $= n-1$), and the confidence level we want. If $n > 30$, the df is no longer important (all values are the same), and we need to remember that the test statistic is:

1.96 for 95% confidence
2.58 for 99% confidence

In this case ($n > 30$) the test statistic is called the z-statistic

For $n < 30$ (called the t-statistic), and other confidence levels, refer to a table - *Would be provided if we had to use it in ISBM, but chances are that anything that comes up will be one of the two z-statistics.*

Again, to be more certain that we include the true value (higher confidence), we need a wider range about the sample mean hence a higher test statistic.

For categorical data, can still get confidence limits. If p is the sample proportion (e.g. $p = 0.1$ if 10% of those in a sample had black hair), and π is the population proportion, then:

$\pi = p \pm 1.96SE$ at the 95% significance level, for $n > 30$

SE is given by $\sqrt{\frac{p(1-p)}{n}}$

In this case, we need $0.05 \leq p \leq 0.95$

NB A 95% confidence level is the same as a p-value of 0.05 i.e. a high confidence level is the same as a low chance of error.

6 Hypothesis Testing

H_0 null hypothesis any discrepancy seen is due to chance *e.g. took a sample with more short people than is typical for the population*
H_1 alternative hypothesis a discrepancy is due to a genuine difference *e.g. the population is, on average, shorter than the value previously suggested*

To test the hypothesis, we calculate a test statistic, and compare this with a table of values. If it is larger than the critical value for a given confidence level with the right df, then we reject H_0 in favour of H_1. If it is lower, then we keep H_0. *The larger this test statistic is, the further the sample is from the value given, and the less likely that the difference is due to chance. The difference doesnt translate directly into a probability, hence we need to find the test statistic and refer to tables.*

The test statistic is simply:

$\frac{O-E}{SE}$ Where O is the observed result and E is the expected result.

Clearly, the further apart O and E are, the bigger the test statistic (more likely that there is a significant difference), and the bigger the SE, the smaller the test statistic (the less reliable the results are, the more likely it is that a discrepancy is due to chance)

The hypothesis test relates simply to the confidence interval:

$$\begin{aligned} & \mu && = \bar{x} \pm (test\ statistic)SE \\ \Rightarrow\ & \mu\bar{x} && = (test\ statistic)SE \\ \Rightarrow\ & \frac{\mu\bar{x}}{SE} && = (test\ statistic) \end{aligned}$$

6.1 One Tail Test

A normal hypothesis test has a null hypothesis that a parameter is the same as some pre-determined value. A one-tail test has a null hypothesis that a parameter is equal to, or lies on one side of a pre-determined value. Expressed mathematically, H_0 will always include an equality sign. *e.g. A one-tail test*

could have the null hypothesis test that basketball players have a mean height of at least 2.0m. That gives:
H_0: $x \geq 2.0$m
H_1: $x \leq 2.0$m

The test statistic is calculated in exactly the same way as before. The only difference comes in interpretation, where the probability associated with a given value of the test statistic is halved. *For a two-tail test, $Z = 1.96$ gives $p = 0.05$ (95% confidence level). For a one-tail test, $Z = 1.96$ would give $p = 0.025$ (97.5% confidence level)*

6.2 Errors

Two types of error are possible:

Type I wrongly reject H_0 *i.e. conclude that there is a significant difference when none exists*
Type II wrongly accept H_0 *i.e. conclude that there is no significant difference when one does exist*

Type I is considered more serious.

The probability of a type I error is α, also called the significance level of the test. Need p-value less than α for the result to be significant.

The maximum probability of a type II error is β. An test should be designed to have $\alpha \leq \beta$.

6.3 Power

The power of a test is given by $100(1 - \beta)\%$. It is the likelihood of the test detecting a significant difference, if such a difference exists. *i.e. If the alternative hypothesis is true (H_0 wrong), then a test with $\beta = 0.2$ will pick this up only 80% of the time.*

7 Comparing Data

7.1 Paired Data

Two sets of data are paired if there is known to be an association between them for instance weights before and after treatment with a drug. The same people are used, so the two sets of data are not completely independent.

The method is the same as a one sample test. *i.e. To test for a significant change in weight after treatment, use the same test as for comparing one sample with a given population parameter.* This is because we are simply testing the differences between each entry in the two sets of data, to see if it diffrerent to zero, as opposed to comparing the overall properties of the two sets.

7.2 Unpaired Data

Two sets of data are unpaired if there is no association between them for instance, the weights of two different groups of people.

The calculation is carried out as for a single sample test, but finding the standard error is more involved. If we have two samples, with n_1 points, sd $= s_1$ in the first, and n_2 points, sd $= s_2$ in the second, then the pooled sample variance is given by:

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

Calculate the T- or Z- statistic for the difference between the means of the two samples. Standard error is given by:

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This method works provided s_1 and s_2 are not too different - no more than a factor of 2 between them.

7.3 Qualitative Data

To compare two sets of qualitative data, we use a chi-squared (χ^2) test *(For ISBM other tests do exist)* Like any other test, this assumes (H_0) that two sets of data are independent, and generates a test statistic to test this.

The method generalises to a table with r rows and c columns. For an example, use a 2x2 table:

Outcome	Treated	Not Treated	Total
Good	40	60	100
Bad	20	280	300
Total	60	340	400

First, we need to generate a table of expected frequencies *i.e. the number of people in each box if there is no association between treatment and outcome.*

Considering everyone in the study, $100/400 = \frac{1}{4}$ of people had a good outcome, and $\frac{3}{4}$ had a bad outcome. Assuming that treatment has no effect on outcome, we would expect 45 of those treated ($60 \times \frac{3}{4}$) to have a bad outcome, and 15 ($60 \times \frac{1}{4}$) a good outcome. Extending the same method to those not treated we get a table of expected results:

Outcome	Treated	Not Treated	Total
Good	15	85	100
Bad	45	255	300
Total	60	340	400

Note that the figures in the total row and column are unchanged.

To find the test statistic, we calculate $\chi^2 = \frac{(O-E)^2}{E}$

In the example, we get:

$$\begin{aligned}\chi^2 &= \frac{(40-15)^2}{15} + \frac{(60-85)^2}{85} + \frac{(20-45)^2}{45} + \frac{(280-255)^2}{255} \\ &= \frac{625}{15} + \frac{625}{85} + \frac{625}{45} + \frac{625}{255} \\ &= 63.4\end{aligned}$$

This is then compared to a table of critical values, as for any other hypothesis test. To find the df for a χ^2, calculate (r-1)(c-1) so for a 2x2 table, df = 1.

In the example, the χ^2 calculated is greater than that for the 99% confidence level i.e. can be fairly sure that the outcome is related to whether the person was treated or not.

There are limitations on when we can use the chi-squared test:

Need $n > 30$
Need all $E \geq 1$
Need at least 75% of $E \geq 5$

8 Further Reading

Statistical consideration for research. S Carley. EMJ 2003:258-262

Author

Michael Stewart, March 2003.
Updated November 2003
michael.stewart@physics.org