# Critical appraisal for emergency medicine 2: Statistics

S Goodacre

Correspondence to:
Professor S Goodacre, Medical Care Research Unit, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK; s.goodacre@sheffield.ac.uk

Accepted 6 January 2008

Critical appraisal of the statistical aspects of an article can be taxing for anyone without expert knowledge. It is tempting to seek out statistical "rules" that can be used to identify flaws in a study, but few situations are sufficiently cut and dried to allow such a crude approach. It is worth bearing in mind that statistics is a specialist field. The idea of a clinician dabbling in statistics should alarm us as much as the idea of a statistician dabbling in clinical medicine.

Statistics should help readers, not baffle them. The findings of nearly all quantitative medical research are subject to a degree of uncertainty arising from random error, as described in the first article in this series. Appropriate use of statistics should help the reader to understand how the findings of a study may be influenced by this uncertainty. The most useful way of appraising the statistical aspects of a study is therefore to try to work out what the statistics actually say. Example 1 shows how some statistics can actually mean very little.

Of course, the least helpful statistical analysis is none at all. In this situation the authors have ignored the potential influence of chance. It is always worth asking of even the simplest article, where are the statistics? An example of a study where some statistics would be very helpful is shown in Example 2.

There are broadly two ways in which statistics can be used to address uncertainty, depending on the way the research question is asked:

1. Hypothesis testing (the p value)
2. Estimation (the confidence interval)

There is no reason why both approaches cannot be used in the same analysis. Indeed, they often complement each other. However, a number of journals prefer confidence intervals to be reported than p values.

## HYPOTHESIS TESTING (THE p VALUE)

The research question is phrased in the form of a hypothesis and the data are collected to determine whether the hypothesis is true. For example, if we were evaluating the effectiveness of a drug compared with placebo, we would ask: Is this drug more effective than placebo? The study would compare the effect of the drug with placebo and determine the probability that this effect could have arisen by chance.

The objectives should state a hypothesis. The opposite of the stated hypothesis (that the stated hypothesis is false) is known as the null hypothesis. The null hypothesis might be that there is no difference between active treatment and placebo, or that there is no difference between our sample and the rest of the population.

The p value can be expressed in two different ways depending on how precise/pedantic you wish to be with your definition:

- Put simply, the p value is the probability that the null hypothesis is true. This makes sense and is easy to understand. The smaller the p value is, the less likely the null hypothesis is to be true, therefore the more likely we are to reject the null hypothesis and accept our alternative hypothesis.
- Strictly speaking, the null hypothesis must be either true or false. It cannot be "probably true" and whether it is true or not cannot be changed by the findings of a research study. Therefore, a more accurate definition is that the p value is the probability of observing results at least as extreme as those we have, given that the null hypothesis is true. The smaller the p value, the more unlikely it is that we would observe these results given the null hypothesis being true. Hence, the more likely we are to reject the null hypothesis.

Don't worry if you don't understand the more complicated definition; it is perfectly possible to survive statistics without making any mistakes by using the simple definition.

## ESTIMATION (THE CONFIDENCE INTERVAL)

Often a research question can be phrased as a measurement and the data collected to provide an estimate of the measurement. For example, evaluating the effectiveness of a drug compared with placebo would involve asking: What is the effectiveness of this drug compared with placebo? The study would estimate the relative or absolute risk reduction associated with using the drug and use confidence intervals to indicate uncertainty around this estimate.

The 95% confidence interval is a range of values around an estimate that have a 95% probability of encompassing the "true" value of that estimate. Put simply, the true value probably lies within the confidence interval. The confidence interval will tell you how precise an estimate is. The wider the confidence interval, the less precise is the estimate.

## CONFIDENCE INTERVALS OR p VALUES?

The hypothesis testing and estimation approaches both have advantages and disadvantages:

- Hypothesis testing should require the researcher to decide before the study starts what difference is considered to be clinically significant. It therefore has the advantage of requiring a definition of "success" against which the treatment or test may be judged.



*Emerg Med J* 2008;**25**:362–364. doi:10.1136/emj.2007.057315

**Example 1: What do these statistics mean?**

A study of a new clinical prediction score reported that the score was a highly significant predictor of adverse outcome ($p<0.001$). Does this mean that this is a useful score?
No, all it means is that the association between the score and adverse outcome was very unlikely to be due to chance alone. All we can say from this is that the score probably predicts outcome better than rolling dice.

- Both p values and confidence intervals can tell you whether a result is statistically significant or not. The result is significant if the p value is less than 0.05 or if the confidence interval does not encompass the value for no effect (eg, a relative risk of 1 or an absolute risk reduction of 0).
- Confidence intervals provide information about the potential magnitude of an effect, regardless of whether or not the result is statistically significant; p values only tell you how statistically significant the result is.
- Confidence intervals can be used to estimate the likelihood of a type II statistical error (see below).
- Too many p values in an article suggest the possibility of multiple hypothesis testing and an increased risk of type I statistical errors (see below).

The list above suggests that confidence intervals have rather more advantages than p values. At a very simplistic level it is reasonable to interpret this as "confidence intervals are good, p values are bad". This is not a bad rule of thumb for judging the appropriateness of statistical analysis in articles. Poor articles are often loaded with p values and bereft of confidence intervals.

## TYPE I AND TYPE II STATISTICAL ERRORS

Any hypothesis test can produce an erroneous conclusion because of random error. These errors are usually termed statistical errors and classified as type I or II. Table 1 shows how type I and type II statistical errors are defined.

The probability of producing a type I or type II statistical error for any hypothesis test depends on the sample size and the level at which statistical significance is set. The latter is the p value below which we consider the results to be so improbable, given the null hypothesis, that we will reject the null hypothesis and accept our alternative. By convention it is normally set at 0.05.

The level at which statistical significance is set is called alpha ($\alpha$). The p value is the probability for a specific hypothesis test that a significant (ie, positive) result is a false positive. The value at which we set $\alpha$ is therefore the maximum probability of a false positive that we are prepared to accept.

Having set $\alpha$ to determine what probability of a false positive result we are prepared to accept, the probability of a false negative result (defined as beta ($\beta$)) is determined by the sample size. The larger the sample size, the smaller $\beta$ will be. The power of a study is defined as $1-\beta$. By convention, a study should aim to recruit a sufficient sample size for the power to be 80% or 90%.

Study power is determined by:

1. the level at which $\alpha$ is set;
2. the sample size;
3. the variability of the outcome measure, as defined by its standard deviation; and
4. the minimum clinically significant difference we wish to detect.

**Example 2: Where are the statistics?**

A study of a new technique for providing emergency department sedation reported that it had been used in 30 consecutive cases without any serious complications. The authors conclude that this shows that the technique is safe.
This study would benefit from some simple statistical analysis such as a confidence interval around the estimate of a zero adverse event rate. There is a very simple rule for calculating this: the 3/n rule. This states that the upper limit of the 95% confidence interval is roughly 3 divided by the number of cases in the series. Thus, the upper limit in this case is $3/30 = 0.1$. So it is within the bounds of statistical probability that the serious adverse event rate could be as high as 10%.

Since (1) is typically set by convention and (3) is beyond our control, researchers should adjust the sample size to detect the minimum clinically significant difference. In the real world, they may be tempted to adjust the minimum clinically significant difference to fit the sample size they think they can achieve.

### Type I (false positive) errors

For any individual test, the probability of a false positive result is reflected in the p value. However, if more than one test has been performed, this is no longer true. With $\alpha$ set at 0.05, the probability of obtaining a false positive result if there is no true difference is $1-(0.95\times0.95)$ for two tests, $1-(0.95\times0.95\times0.95)$ for three tests, etc. If you do enough tests you will ultimately get a positive result due to chance, even if there is no true difference. Isolated positive results among a series of tests should therefore be viewed with suspicion, particularly if there is no scientific rationale why that test should be positive and not the others, or why that test has been done and not any others. This is illustrated in Example 3.

Multiple hypothesis testing is a common flaw in poorly planned studies. If researchers collect their data without any clear objective and then analyse the data to look for any statistically significant results, they will almost certainly find some. Unfortunately they are likely to be meaningless and false positive.

Multiple hypothesis testing is easy to spot when authors report all the hypothesis tests they have performed. However, they will often only report the positive (significant) results. It is worth asking the following questions about any positive result, particularly from observational data, to identify the possibility of multiple hypothesis testing:

- Is there a clear rationale for the hypothesis tests? This should be explained in the background.
- Does the hypothesis test flow from the study objectives or does it only appear in the results?

**Table 1** Type I and II statistical errors

| | Alternative hypothesis is true | Null hypothesis is true |
|---|---|---|
| Experiment shows significant result | True positive<br>No error | False positive<br>Type I error |
| Experiment shows no significant result | False negative<br>Type II error | True negative<br>No error |

### Example 3: Multiple hypothesis testing

A new drug for treating ventricular fibrillation has been evaluated in a randomised trial. Overall the drug had no significant effect on mortality, but the authors have undertaken a subgroup analysis and identified a significant result in men aged less than 60 who suffered a witnessed arrest. Should we start using this drug for these selected patients?

Subgroup analyses should always be treated with caution because they inevitably involve undertaking multiple hypothesis tests. Using the 5% threshold for statistical significance, there is a 1 in 20 chance that each hypothesis test will be positive due to chance alone. The more hypothesis tests undertaken, the more likely a false positive result will arise due to chance. It is reasonable to assume that at least eight hypothesis tests have been undertaken in this study and there may be many more that the authors have not reported. We should therefore ignore this positive result because there is a substantial probability that it has arisen by chance.

Subgroup analysis should be planned in advance and supported by a strong a priori rationale. In certain circumstance it may be appropriate to use a lower threshold for statistical significance (eg, 0.01).

- Does the methods section include a plan for analysis that sets out what tests will be performed, does it include the positive test, and does it follow from the objectives?
- Do the reported hypothesis tests suggest other more obvious associations that should be explored but have not been reported? For example, it is slightly suspicious if a significant p value has been reported for an association between educational status and outcome, but no association between age or gender and outcome have been explored.

### Type II (false negative) errors

If a study produces a negative result it is worth checking the following factors to identify whether it could be a type II (false negative) error:

- Look for confidence intervals. If they are wide this suggests that estimates are imprecise and a false negative result more likely.
- Look at the extreme ends of the confidence interval. If important differences are possible within the confidence interval, then this study has not ruled out the possibility of an important effect, even though the study is negative.
- Look at the power calculation. Were sensible values for $\alpha$ and $\beta$ used and what was the minimum clinically significant difference? Is this really the smallest difference that would be worthwhile detecting? Has it been justified?

## ISSUES THAT MAY BE BEST LEFT TO AN EXPERT

If you have a little statistical knowledge, it may be tempting to use it to uncover "fundamental flaws" in the statistical analysis. This is a dangerous game to play. The following issues require a fair amount of statistical expertise and careful judgement:

- Parametric versus non-parametric tests: it may be useful to identify whether data are Normally distributed and whether the statistical tests make assumptions about distribution. However, there are many circumstances in which skewed data can be perfectly acceptably analysed using parametric tests.
- Regression, multivariate analysis and any kind of statistical modelling: these are very useful statistical techniques that can be appropriately used to enhance analysis. However, they can often be used to obscure or mislead. Always seek expert statistical advice to determine whether they have been used appropriately.

## SUMMARY

It is possible to get horribly bogged down when appraising the statistical aspects of a paper. The key is to remember that the statistics are supposed to be there to help you. In particular, they should help you to interpret the potential effect of random error upon the findings. In this respect, confidence intervals are often more useful than p values.

**Competing interests:** None.

# Critical appraisal for emergency medicine 2: Statistics

S Goodacre

*Emerg Med J* 2008 25: 362-364
doi: 10.1136/emj.2007.057315