EBM notebook

Statistical approaches to uncertainty: p values and confidence intervals unpacked

Introduction

What is statistical uncertainty?

Statistical uncertainty is the uncertainty (present even in a representative sample) associated with the use of sample data to make statements about the wider population.

Why do we need measures of uncertainty?

It usually is not feasible to include all individuals from a target population in a single study. For example, in a randomised controlled trial (RCT) of a new treatment for hypertension, it would not be possible to include all individuals with hypertension. Instead, a sample (a small subset of this population) is allocated to receive either the new or the standard treatment.

What are the measures of uncertainty?

Either hypothesis tests (with the calculation of p values) or confidence intervals (CIs) can quantify the amount of statistical uncertainty present in a study, though CIs are usually preferred.

In a previous Statistics Note, we defined terms such as experimental event risk (EER), control event risk (CER), absolute risk reduction (ARR), and number needed to treat (NNT).[1] Estimates such as ARR and NNT alone, however, do not describe the uncertainty around the results. P values and CIs provide additional information to help us determine whether the results are both clinically and statistically significant (table 1).

Table 1. Outcomes in the RCT comparing streptomycin with bed rest alone in the treatment of tuberculosis

Intervention	Survival	Death	Risk of death	ARR	95% CI	P value
Streptomycin (n = 55)	51	4	4/55 = 7.3% (EER)	25.9% − 7.3% = 19.6%	5.7% to 33.6%	0.006
Bed rest (n = 52)	38	14	14/52 = 25.9% (CER)			

The ARR (the difference in risk) is estimated to be 19.6% with a 95% CI of 5.7% to 33.6%. The p value of 0.006 means that an ARR of 19.6% or more would occur in only 6 in 1000 trials if streptomycin was equally as effective as bed rest. Since the p value is less than 0.05, the results are statistically significant (ie, it is unlikely that streptomycin is ineffective in preventing death). The 95% CI suggests that the likely true benefit of streptomycin could be as small as 5.7% or as large as 33.6%, but is very unlikely to be 0% or less. Our best estimate for the ARR is 19.6% and hence the NNT is 6 (95% CI 3 to 18). This means that we might have to treat as many as 18 people with streptomycin or as few as 3 to prevent 1 additional person dying of tuberculosis.

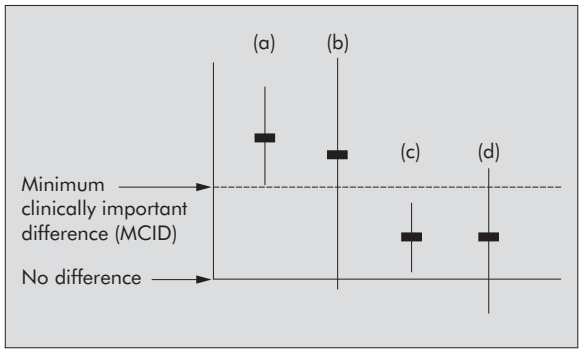

Figure 1 Clinical significance and statistical significance.

WHY IS THE 5% LEVEL (P<0.05) USED TO INDICATE STATISTICAL SIGNIFICANCE?

Conventionally, a p value of <0.05 is taken to indicate statistical significance. This 5% level is, however, an arbitrary *minimum* and p values should be much smaller, as in the above study (p = 0.006), before they can be considered to provide strong evidence against the null hypothesis. Hence reporting the exact p value (eg, p = 0.027) is more helpful than simply stating that the result is significant at the 5% level (or 1% level, as above).

IF AN EFFECT IS STATISTICALLY SIGNIFICANT, DOES THIS MEAN IT IS CLINICALLY SIGNIFICANT?

A *statistically* significant difference is not necessarily one that is of *clinical* significance. In the above example, the statistically significant effect (p = 0.006) is also clinically significant as even a modest improvement in survival is important. For many effects, however, the benefit needs to be somewhat greater than zero for it to be of clinical significance (ie, of sufficient benefit to be worth the effort of treatment). In figure 1, while both studies (a) and (c) show a statistically significant result, with the CIs not overlapping the "no difference" value, only (a) has a result that is consistent (in terms of the CI) with at least a minimum clinically important difference (MCID). Studies (b) and (d) are not statistically significant, as their CIs overlap the values of no difference.

ARE P-VALUES AND CONFIDENCE INTERVALS RELATED?

While the 2 approaches to dealing with the problem of uncertainty are somewhat different, p values and CIs generally provide consistent results. If the effect is statistically significant (at the 5% level), then the 95% CI will not include the value of ''no difference'', and vice versa. While CIs are preferable to p values in summarising study results, both approaches are commonly used.

Comparison of the use of p values and confidence intervals in statistical inference

P values and hypothesis tests	Confidence intervals
What are they used for?	
p values are used to assess whether a sample estimate is significantly different from a hypothesised value (such as zero—ie, no treatment effect). Hypothesis tests assess the likelihood of the estimate under the *null hypothesis* of no difference between 2 population values (or no treatment effect). Conventionally, if $p<0.05$, the null hypothesis is rejected.	Confidence intervals (CIs) present a range of values around the sample estimate within which there is reasonable confidence that the true, but unknown, population value lies. The 95% CI (the range of values within which there is 95% probability that the true value lies) is most commonly used. It corresponds with the typical 5% significance level used in hypothesis tests.
What do they tell us?	
The p value is the probability that the observed effect (or more extreme ones) would have occurred by chance *if in truth there is no effect.* However, it doesn't tell us anything about the size of the true effect and, moreover, since hypothesis tests are 2 tailed (we are interested in differences in either direction) it doesn't even tell us the direction of this effect. Thus, in the above example, the p value of 0.006 indicates that an effect of 19.6% or more, in favour of either streptomycin or bed rest, would occur in only 6 in 1000 trials if in truth there is no effect.	The CI provides a range of values whose limits are, with specified probability (typically 95%), the smallest and the largest true population values consistent with the sample data. A CI can thus effectively function as a hypothesis test for an infinite number of values: if the CI includes any 1 of these values then the sample estimate is not statistically significantly different from it. The 95% CI is of particular relevance to evidence-based practice (EBP), providing valuable information such as whether the interval includes or excludes clinically significant values.
When can they be used?	
There are many different types of hypothesis test, each suitable for a particular type of data. For example, parametric tests (such as t-tests) are only suitable for large samples and for data that are drawn from a population which is approximately normally distributed.	A CI can, and should, be calculated for most measures of effect, such as differences between means (such as scores or weights), and differences in proportions, EER and CER, ARR, NNT, risk ratios (RR), and odds ratios (OR).
How are they calculated?	
The observed effect together with a measure of its variability (such as the standard error, SE) is used to calculate a ''test statistic'' (eg, t, z, χ^2). For example, a t statistic is calculated by dividing the observed effect by its SE. The value of the test statistic is used (from statistical tables) to determine the p value. Thus, in the example above (where SE(ARR) = 7.1%), the z statistic (assuming that the ARR is approximately normally distributed) for the test of whether the risk of death differs between those allocated to streptomycin and those allocated to bed rest is calculated as $z = 19.6/7.1 = 2.76$. This has an associated p value of 0.006.	To calculate a CI around a sample estimate, only 3 pieces of information are needed: the sample size (n), the sample standard deviation (SD), and the ''z score,'' which varies depending on the degree of confidence wanted (95%, 99% etc). For a 95% CI, $z = 1.96$, and for a 99% CI, $z = 2.58$. A 95% CI is calculated as: Sample estimate $\pm$ 1.96 standard errors (SE) of the measure (note: $SE = SD/\sqrt{n}$). Thus, in the above example (where SE(ARR) = 7.1%), the 95% CI for the true ARR is calculated as: 19.6% $\pm$ 1.96 (7.1%) = 5.7% to 33.6%.

WHY DOES SAMPLE SIZE HAVE TO BE CONSIDERED WHEN INTERPRETING THE SIZE OF THE P VALUE AND THE WIDTH OF THE CI?

The larger the sample, the less the uncertainty, the narrower the CI, and hence the smaller the observed effect that can be declared statistically significant ($p<0.05$). Thus, if a sample is very large, even a very small difference (which may be of no clinical relevance) may be statistically significant (see (c) in figure 1). The width of a CI is affected by both the sample size (n) and the sample SD. The larger the sample (and the smaller its variability), the greater the accuracy of the sample estimate and thus the narrower the CI. A wide CI can thus reflect either a small sample or one with large variability (see (b) in figure 1).

CAN THE CONCLUSIONS FROM A HYPOTHESIS TEST BE IN ERROR?

Since hypothesis tests are based on estimates of probability, their conclusions can be in error. There are 2 types of error: rejecting the null hypothesis when it is true (type I error; the probability of this error is 5% if the 5% significance level is used) and failing to reject the null hypothesis when it is false (type II error; the probability of this error is 1 − Power) (figure 2). Power and sample size will be discussed in more detail in a future Statistics Note.

	Truth	
Statistical decision	**Null hypothesis true**	**Null hypothesis false**
Reject null hypothesis	Type I error	Correct (power)
Do not reject null hypothesis	Correct	Type II error

Figure 2 Type I and II errors.

HELEN DOLL, BSc, DIP APP STATS, MSc
Department of Public Health, University of Oxford, Oxford, UK

STUART CARNEY, MB, ChB, MPH, MRCPsych
Department of Psychiatry, University of Oxford, Oxford, UK

1 Carney S, Doll H. *Evidence-Based Medicine* 2005;**10**:102–3.