British Standards Institution Study Day

How do we make sense of the results?

Martin Bland
Prof. of Health Statistics
University of York
http://martinbland.co.uk

1. Making sense of the results

In this lecture we shall look at the statistical principals involved in the presentation of the results of research studies. We shall illustrate this with a look at how people presented results in a leading health research journal: *The British Medical Journal*, or *BMJ*. The BMJ is published both on paper and on-line, but it now uses the on-line journal as the key publication. For this lecture, we shall look at the four research papers published in the week starting between 30 May 2011 and 5 Jun 2011. You can find these on

http://www.bmj.com/archive/online/2011/05-30

The first is by Anke Steckelberg, Christian Hülfenhaus, Burkhard Haastert, Ingrid Mühlhauser: Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial (*BMJ* 2011; **342**: d3193). In the Abstract under Results we have:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). A "positive attitude" towards colorectal screening prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, −5.9% to −0.3%; P<0.01). The intervention had no effect on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); P=0.87) . . . .'

We can pick some terms out of this passage:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, **99% confidence interval 25.7% to 36.7%**; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, **37.8% to 49.1%**; P<0.001). A "positive attitude" towards colorectal screening prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, **−5.9% to −0.3%**; P<0.01). The intervention had no effect on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); P=0.87) . . . .'.

These are all confidence intervals, one method used to make sense of the results. We can also pick out some other terms:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; **P<0.001**). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; **P<0.001**). A "positive attitude" towards colorectal screening prevailed in both groups but was **significantly lower** in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%,

−5.9% to −0.3%; **P<0.01**). The intervention had no effect on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); **P=0.87**) . . . .'.

These all refer to the results of significance tests, another method used to make sense of research results.

The second paper is by Astrid Guttmann, Michael Schull, Marian Vermeulen, and Therese Stukel: Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada (*BMJ* 2011; **342**: d2983). In the Abstract under Results we read:

'13 934 542 patients were seen and discharged and 617 011 left without being seen. The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 v <1 hour the adjusted odds ratio (95% confidence interval) was 1.79 (1.24 to 2.59) for death and 1.95 (1.79 to 2.13) for admission in high acuity patients . . .'.

We can pick out confidence intervals here, too:

'13 934 542 patients were seen and discharged and 617 011 left without being seen. The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 v <1 hour the adjusted odds ratio **(95% confidence interval)** was 1.79 **(1.24 to 2.59)** for death and 1.95 **(1.79 to 2.13)** for admission in high acuity patients . . .

An odds ratio = 1 for no effect, greater than 1 for a positive effect, so this means that people with a long stay in the emergency department were more likely to have an adverse event. We shall cover odds ratios in more detail in Lecture 2.

The third publication was by Kari Johansson, Erik Hemmingsson, Richard Harlid, Ylva Trolle Lagerros, Fredrik Granath, Stephan Rössner, and Martin Neovius: Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study (*BMJ* 2011; **342**: d3017). In the Abstract under Results we read:

'. . . After the very low energy diet period, apnoea-hypopnoea index was improved by −21 events/hour (95% confidence interval −17 to −25) and weight by −18 kg (−16 to −19; both P<0.001). After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001).'

Again we have confidence intervals:

'. . . After the very low energy diet period, apnoea-hypopnoea index was improved by −21 events/hour **(95% confidence interval −17 to −25)** and weight by −18 kg **(−16 to −19**; both P<0.001). After one year the apnoea-hypopnoea index had improved by −17 events/hour **(−13 to −21)** and body weight by −12 kg **(−10 to −14)** compared with baseline (both P<0.001).'

and significance tests:

'. . . After the very low energy diet period, apnoea-hypopnoea index was improved by −21 events/hour (95% confidence interval −17 to −25) and weight by −18 kg (−16 to −19; **both P<0.001**). After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (**both P<0.001**).'

The fourth paper was by Geeta Kumar, Harshpal Sachdev, Harish Chellani, Andrea Rehman, Vini Singh, Harsh Arora, and Suzanne Filteau: Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months: randomised controlled trial (*BMJ* 2011; **342**: d2975). In the Abstract under Results we read:

'Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group v 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), or referral to the outpatient clinic for moderate morbidity. . .'.

Again we have confidence intervals:

'Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group v 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, **95% confidence interval 0.68 to 1.29**; P=0.68), or referral to the outpatient clinic for moderate morbidity. . .'.

and significance tests:

'Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group v 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; **P=0.68**), or referral to the outpatient clinic for moderate morbidity. . .'.

Like an odds ratio, a rate ratio = 1.0 if there is no difference, greater than 1 for a positive effect, so this means that infants in the vitamin D group were less likely to die or be admitted to hospital than were infants in the placebo group. We shall cover rate ratios and their difference from odds ratios in more detail in Lecture 2.

Confidence intervals and significance tests are two things we can use to help us to interpret the results of research studies. They are two methods of what we call 'statistical inference'. We shall look here at what each means, what they have in common, and what is the difference between them.

2. Samples and populations

In almost all health research studies, the data we have are from a sample drawn from a much larger population.

- sample: people we have in the study,
- population: all the other people like them, including people who will be like them in the future.

We want to use the sample to tell us about the population. The problem is that we have this particular sample. Would another sample give us a different answer? For example:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001) . . .'.

Would another sample give a difference = 31.2 percentage points? Might it give 30.0 percentage points? Or 40 percentage points? Or even −31.2 percentage points? The confidence interval enables us to deal with this problem.

How strong is the evidence that informed choice increases following evidence-based information? Would informed choice increase in another sample? Would it increase in the population? The significance test P value enables us to deal with this problem.

3. Confidence intervals

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7% . . . '

In the sample the difference is 31.2 percentage points. We want to know the difference ***in the population***. We cannot know exactly what it is. The sample difference, 31.2 percentage points, is only an estimate of the difference in the population.

The range of values 25.7% to 36.7% is also an estimate. We estimate that, ***in the population***, the difference in percentage making an informed choice is somewhere between 25.7 to 36.7. In mathematics, the set of values between 25.7 and 36.7 is called an interval. 'Between 25.7 and 36.7 percentage points' is called an interval estimate. We can think of '25.7 and 36.7 percentage points' as a range of values within which we estimate the difference in the population to be.

Why '99%' confidence interval? This is the difficult bit. We choose the interval so that for 99% of the possible samples which we could take, of which this is just one, the interval would include the population reduction. This means that 99% of confidence intervals include their population value. 1% do not.

Why 99%? 99% was a choice made for this particular study. More often, we use 95% rather than 99%.

How is the 99% confidence interval calculated? There are many methods for doing this in different situations. This is what a lot of statistical method is about. We are not going into this for this lecture. We shall leave it until later in the course.

4. Significance tests

The other approach to statistical inference which is often used is significance tests. In the statement

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001) . . . '

'P<0.001' is the result of a significance test. Like the confidence interval, the significance test tells us something about the population from which our sample was drawn. Here, the P value is a indicator of the strength of the evidence which this sample provides that, ***in the population***, the percentage making an informed choice increases when people are given evidence based information. P is a probability, between 0 and 1. It is the proportion of possible samples which would give us the observed value.

- Small P ➔ strong evidence = statistically significant.
- Large P ➔ weak evidence = not statistically significant.

The usual cut-off for a decision about whether there is sufficient evidence is P= 0.05. Sometimes researchers might decide that a different cut-off is better. In this study, the cut-off chosen for a decision that we have evidence for an effect is P= 0.01. The authors do not say why they made this decision, which they should.

What is a P value, exactly? Suppose in this population the information given did not affect informed choice. Would many possible samples of this size produce a difference as big as 31.2 percentage points? The proportion of possible samples which would produce a difference as big as 31.2 percentage points is P. In the example, P<0.001.

The proportion of samples which would have a difference as big as 31.2, or bigger, if the difference in the population were zero, is less than 0.001, or less than 1 in 1000. This is small, so ***either*** we have a sample which is very unusual ***or*** the difference in population is not zero. Here the evidence for a non-zero difference is strong. The difference is said to be statistically significant, meaning that the sample provides evidence for something existing in the population. The word 'significant' comes from a Latin word meaning a sign; we can think of a significant difference as a sign of something in the population. Because P is so small, less than 0.001, we would say that this difference is very highly significant. There is very strong evidence.

Sometimes people say that the P value is the probability that the null hypothesis, e.g. that there is no difference in the population from which our sample was taken, is true. Sometimes people say that the P value is the probability that the observed difference arose by chance. Neither of these is quite correct. The P value is the probability of a sample being as far from what you would expect, if the null hypothesis were true. If the data would be unlikely if the null hypothesis were true, we take this as evidence that it is not.

Steckelberg *et al.* go on to say:

'. . . A "positive attitude" towards colorectal screening prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, −5.9% to −0.3%; P<0.01) . . . '

How do we decide whether there is enough evidence? The usual choices for meaningful P values are:

- P>0.05 ➔ not significant, weak or insufficient evidence.
- P≤0.05 ➔ significant, sufficient evidence.
- P<0.01 ➔ highly significant, strong evidence.
- P<0.001 ➔ very highly significant, very strong evidence.

We can vary these cut-offs sometimes, depending on the reason for doing the test. In this study they used P<0.01 as significant.

Why say 'P<0.001' rather than give the exact P value? Most statistical computer programs give P to only 3 or to 4 decimal places. They do this because for many tests of significance, the calculation of P becomes imprecise when P is very small. We do not say P = 0.000, even when the P value is so small that that is the value to three decimal places (e.g. P = 0.00000223 is P = 0.000 to three decimal places). This is because it is usually possible for the difference to have occurred if the null hypothesis were true, even if very, very unlikely. This is just a convention and most computer programs will display P = 0.000 in this case.

How is the P value calculated? There are many methods for different situations. As for confidence intervals, this is what a lot of statistical method is about. We shall come back to it on another occasion.

5. The most important thing about significance tests:

If P>0.05, we say that the difference is not significant. ***This does not mean that there is no difference!*** It means that we have not found evidence for a difference, which is very different. Absence of evidence is ***not*** the same as evidence of absence.

In general, it is very difficult to prove that something does not exist. At one time, it was thought by Europeans that all swans were white, because all the species of swans known, mute, whooper, Bewick's, etc., were white. Then Europeans reached Australia and found that there the swans were black! So the University of York keeps black swans on our lake to remind us of this: never conclude that something does not exist just because you cannot find it.

Steckelberg *et al*. go on to say:

'The intervention ***had no effect*** on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); P=0.87) .'

This inference is quite wrong. There is no evidence that there is a difference in the population, but we do not have evidence that there is no difference. The 99% confidence interval for the difference is = −6.3 to +5.3 percentage points. This means that a difference of this size would be compatible with the data. ***We cannot conclude that the treatment had no effect.*** At most, we could conclude that take-up is increased by no more than five percentage points. We should give the confidence interval.

6. Significance test or confidence interval?

Significance tests and confidence intervals are two ways to make the link between sample and population. Why have two approaches? There are several reasons:

- historical reasons, including some very argumentative statisticians,
- computation problems, sometimes we cannot find a confidence interval in a straightforward way or without modern computing power,
- sometimes there is no meaningful estimate to find,
- sometimes we are concerned with existence of an effect more than how big it is,
- we cannot always find a confidence interval; but we can almost always do a significance test.

Recommendations to authors for most major health research journals advise confidence intervals be given where possible.