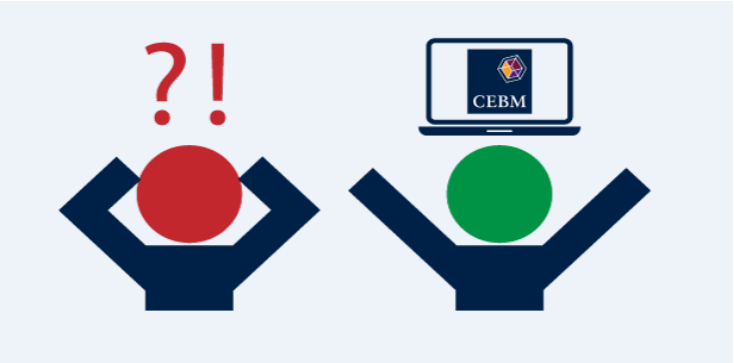
Tip for data extraction for meta-analysis – D7



What if something is missing from categorical risk data?

Kathy Taylor

Unbounded limits of outer categories in categorical risk (quantile or dose-response) data is a common problem in meta-analysis of prognostic studies. It's worth checking to see if the global range of the exposure variable is reported, as this provides the outer bounds of the outer categories but the global range is not often available. Earlier (post D5), I used imputation to deal with the problem of unbounded limits, by setting the range of the exposure of unbounded categories equal to a multiple of the average range of exposure of the inner categories. In this post I'll look at some other problems of missing data in categorical data and how you might deal with them. I'll present five different examples. Bekkering et al provides further examples.

Examples 1 to 4 provide ways to complete missing categorical data so that you can then apply the trend estimation method that I highlighted before (post D4) to estimate a hazard ratio, relative risk or odds ratio. Example 5 provides a method where incomplete categorical case-control data are used to directly estimate an odds ratio.

Example 1

In dose-response data, if the serving size is missing, standard serving sizes may be reported elsewhere and ranges may be inferred from verbal descriptions e.g. "once a week".

Example 2

If an odds ratio is missing, an unadjusted odds ratio may be derived from the number of first events and total number in each group by the following:

$$p = \frac{events}{total}$$

$$odds = \frac{p}{1-p}$$

$$Unadjusted\ odds\ ratio\ = \frac{p_{Intervention}/(1-p_{Intervention})}{p_{Control}/(1-p_{Control})} = \frac{p_{Intervention}\ (1-p_{Control})}{p_{Control}\ (1-p_{Intervention})}$$

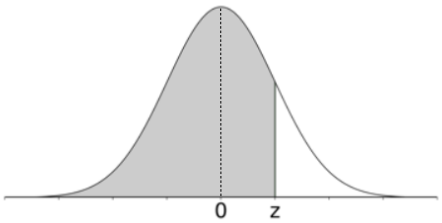
Note that for this post and future posts, when I give equations, for A multiplied by B, instead of $A \times B$, I will write AB, as this is easier to read.

Example 3

A missing confidence interval may be derived from a p value. Altman and Bland show how to estimate a confidence interval for a hazard ratio, odds ratio or relative risk from a p value. The method relies on effect estimates being log-transformed. See post G8 to find out about log-transformations.

They give an example of a relative risk of 0.30 with a reported p-value of 0.034.

First you need to calculate the z value which corresponds to the reported p-value from a table of the standard normal distribution. The table reports the cumulative probability P(Z<z), indicated by the shaded area.



The cumulative probability corresponding to a 2-sided p value of 0.034 is $1 - \frac{0.034}{2} = 0.983$. The corresponding z value is 2.12 which is found in the table or calculated by
abs(norm.s.inv(0.017)) in EXCEL
abs(invnormal(0.017)) in STATA
abs(qnorm(0.017)) in R.

Taking logs of the relative risk, $ln(0.30) = -1.204$
The z value is assumed to be equal to the log of the effect estimate divided by its standard error (SE). Therefore $SE = \frac{-1.204}{2.12} = -0.568$
The negative sign is ignored and the 95% confidence interval of -1.204 is calculated as
$-1.204 \pm 1.96SE$ i.e. -2.317 to -0.091.
Taking exponentials produces the 95% confidence interval of the relative risk, 0.30, as 0.10 to 0.91.

Example 4

In a case-control study, if the numbers of patients in the case and control groups are missing, they can be derived from a reported odds ratio provided the total number of first events in each group are also reported, by using a bit of maths (see below if you're interested) as

$$n_1 = \frac{E_1 T \ + \ E_1 E_2 (OR - 1)}{ORE_2 + E_1}$$

and

$$n_2 = \frac{E_2 TOR \ + \ E_1 E_2 (1 - OR)}{ORE_2 E_1}$$

where:

OR is the reported odds ratio

n_1 and n_2 are the numbers in each group

E_1 and E_2 are the total number of first events in each group.

If the odds ratio is unadjusted the calculations above will be exact, and if the reported odds ratio is adjusted for confounders, the calculation will produce estimates which will be reasonable provided the adjustment does not alter the odds ratio markedly.

Example 5

Perez et al show how simulation can be used to derive hazard ratios when categorical data are reported as the number of cases and controls for each level of a marker, and the overall mean and standard deviation of the marker are also available (possibly reported in a table of baseline characteristics). They publish the R code that they used to run the simulation.

Here's a tip…

You can deal with missing categorical risk data in several ways, by making inferences from similar studies, doing simple calculations, or making direct estimates by using simulation.

As you can see, a number of data extraction methods involve making estimates, and it's important test the impact of these estimates in sensitivity analysis. Therefore, when extracting data, you'll find it useful to record which studies involved estimates. This involves the process of data extraction. In my next blog post, I'll give some tips on how you can improve the process of data extraction.

Where did the equations come from?

(A multiplied by B is represented in equations as AB)

For example 4, the odds ratio (OR) and the total number of first events in each group (E_1 and E_2) have been reported. The numbers in each group (n_1 and n_2) have not been reported.

$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}$ **(equation 1)**

$p_1 = \frac{E_1}{n_1}$ **(equation 2)**

$p_2 = \frac{E_2}{n_2}$ **(equation 3)**

$T = n_1 + n_2$ **(equation 4)**

Substitute equations 2 and 3 into equation 1

$OR = \frac{\frac{E_1}{n_1}\left(1-\frac{E_2}{n_2}\right)}{\frac{E_2}{n_2}\left(1-\frac{E_1}{n_1}\right)}$

Rearrange and multiply both sides by $n_1 \times n_2$

$OR\ (E_2 n_1 - E_1 E_2) = (E_1 n_2 - E_1 E_2)$ **(equation 5)**

Substitute $n_2 = T - n_1$ (from equation 4) into equation 5 and rearrange

$$n_1 = \frac{E_1 T\ +\ E_1 E_2 (OR - 1)}{OR E_2 + E_1}$$

Substitute n_1 into $n_2 = T - n_1$ (from equation 4) and rearrange

$$n_2 = \frac{E_2 TOR\ +\ E_1 E_2 (1 - OR)}{OR E_2 + E_1}$$

Dr Kathy Taylor teaches data extraction in Meta-analysis. This is a short course that is also available as part of our MSc in Evidence-Based Health Care, MSc in EBHC Medical Statistics, and MSc in EBHC Systematic Reviews.

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter @dataextips