Biostatistics (MATH11230)
Estimation and inference for the association measures

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Estimation and inference for measures of association

$\hookrightarrow$ Depending on the study design, we already know, even if only informally, how to estimate, when possible, the relative risk and the odds ratio.

$\hookrightarrow$ But, of course, as statisticians/data scientists, we should also determine the level of uncertainty associated with the estimators we propose, expressed via calculation of confidence intervals.

$\hookrightarrow$ We can then assess whether an observed association of D and E in a sample reflects a population in which D and E are truly associated or it may be simply due to chance.

$\hookrightarrow$ We further restrict our attention to cohort and case-control studies as these are also the most commonly used designs for studying the association between exposure and disease.

Estimation and inference for measures of association

Estimate of the odds ratio

$\hookrightarrow$ In what follows, let us consider a generic 2×2 contingency table whose entries are given by

	D	not D	Totals
E	a	b	a + b
not E	c	d	c + d
Totals	a + c	b + d	a + b + c + d

$\hookrightarrow$ For cohort designs, we can estimate $\text{OR}_{D|E}$, where

$$\text{OR}_{D|E} = \frac{\Pr(D \mid E)}{\Pr(\text{not } D \mid E)} \bigg/ \frac{\Pr(D \mid \text{not } E)}{\Pr(\text{not } D \mid \text{not } E)},$$

by simply estimating the involved probabilities.

$\hookrightarrow$ For example, and letting $p_1 = \Pr(D \mid E)$ and $p_2 = \Pr(D \mid \text{not } E)$, we have

$$\widehat{p}_1 = \frac{a}{a + b}, \qquad \widehat{p}_2 = \frac{c}{c + d}.$$

Estimation and inference for measures of association
Estimate of the odds ratio

$\hookrightarrow$ We thus have

$$
\begin{aligned}
\widehat{\text{OR}}_{D|E} &= \frac{\widehat{p}_1}{1 - \widehat{p}_1} \Big/ \frac{\widehat{p}_2}{1 - \widehat{p}_2} \\
&= \frac{a/(a+b)}{1 - (a/(a+b))} \Big/ \frac{c/(c+d)}{1 - (c/(c+d))} \\
&= \frac{a}{b} \Big/ \frac{c}{d} \\
&= \frac{ad}{bc}.
\end{aligned}
$$

Estimation and inference for measures of association
Estimate of the odds ratio

$\hookrightarrow$ For case-control data, we can rely only on the probabilities that condition on the disease status and thus we estimate

$$\text{OR}_{E|D} = \frac{\Pr(E \mid D)}{\Pr(\text{not } E \mid D)} \bigg/ \frac{\Pr(E \mid \text{not } D)}{\Pr(\text{not } E \mid \text{not } D)}.$$

$\hookrightarrow$ The corresponding estimate is given by

$$\widehat{\text{OR}}_{E|D} = \frac{a/(a+c)}{1-(a/(a+c))} \bigg/ \frac{b/(b+d)}{1-(b/(b+d))}$$

$$= \frac{ad}{bc}.$$

$\hookrightarrow$ The estimate of the OR thus does not depend on the study design employed. This should be no surprise as we have already learned that the OR is symmetric in the roles of D and E.
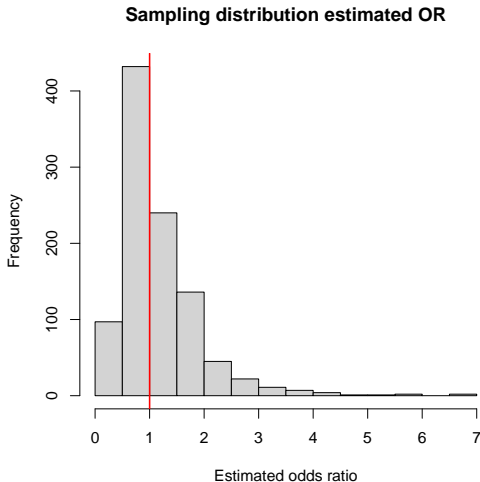
Estimation and inference for measures of association
Sampling distribution of the odds ratio

↪ The estimate $\widehat{\text{OR}} = ad/bc$ is a random variable and, as such, has an associated sampling distribution describing the probabilities that the estimate will fall in specified ranges of values for a given sample size.

↪ Stated in an equivalent way, the sampling distribution of $\widehat{\text{OR}}$ describes the frequency with which ad/bc takes on certain values after repeated samples under identical sampling conditions.

↪ To illustrate the sampling distribution of $\widehat{\text{OR}}$, let us consider a cohort study for which we set $p_1 = \Pr(D \mid E) = p_2 = \Pr(D \mid \text{not } E) = 0.2$, thus implying that the true odds ratio, OR, is one.

↪ Let us now simulate 1000 repeated versions of this cohort study and compute the associated odds ratio (see the R code in the Supplementary Materials for this week).

Estimation and inference for measures of association

Sampling distribution of the odds ratio

Sampling distribution estimated OR
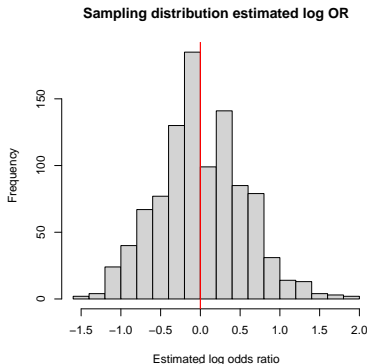
Estimation and inference for measures of association
Sampling distribution of the odds ratio

$\hookrightarrow$ It is evident that the sampling distribution of $\widehat{OR}$ is skewed to the right.

$\hookrightarrow$ For the specific random seed that we have used, the smallest value of the OR estimate was 0.2 and the maximum was 6.77.

$\hookrightarrow$ The mean and median of the 1000 $\widehat{OR}$s was 1.19 and 1, respectively.

$\hookrightarrow$ Note that in this specific case that the true OR is one, the estimate for the OR cannot be much smaller than one but it could be much larger with nonneligible probability.

$\hookrightarrow$ It does not seem reasonable to approximate the sampling distribution of $\widehat{OR}$ by a normal distribution (which is symmetric), unless the sample size is extremely large.

$\hookrightarrow$ This complicate things as approximating sampling distributions by normal distributions is the basis of the construction of confidence intervals (thanks to the central limit theorem) for many estimation problems.

Estimation and inference for measures of association
Sampling distribution of the odds ratio

↪ However, the sampling distribution of the estimated log odds ratio is fairly symmetric and bell-shaped.

Sampling distribution estimated log OR

Estimated log odds ratio

↪ The mean and median across the 1000 log ORs is now 0.023 and 0, respectively.

↪ We will therefore use the normal approximation for the sampling distribution of the estimated log odds ratio.

Estimation and inference for measures of association

Confidence interval for the odds ratio

$\hookrightarrow$ As Jewell (2003, p 79) remarks, the general shape of the sampling distributions of the $\widehat{OR}$ and $\log \widehat{OR}$ would be the same if the data were generated from a case-control study.

$\hookrightarrow$ Because the the mean of $\log(\widehat{OR})$ is close to the true $\log(OR)$ when the sample size is large, it follows that $\log(\widehat{OR}) - \log(OR)$ has an approximately normal sampling distribution with expectation equal to zero.

$\hookrightarrow$ We are now left with the task of estimating the variance of this normal sampling distribution. Let us assume a cohort design.

$\hookrightarrow$ Remember that we have used the notation $p_1 = \Pr(D \mid E)$ and $p_2 = \Pr(D \mid \text{not } E)$. The estimate of the log odds ratio is then given by

$$\log(\widehat{OR}) = \log\left(\frac{\widehat{p}_1}{1 - \widehat{p}_1}\right) - \log\left(\frac{\widehat{p}_2}{1 - \widehat{p}_2}\right).$$

$\hookrightarrow$ We then have

$$\text{var}\{\log(\widehat{OR})\} = \text{var}\left\{\log\left(\frac{\widehat{p}_1}{1 - \widehat{p}_1}\right) - \log\left(\frac{\widehat{p}_2}{1 - \widehat{p}_2}\right)\right\}.$$

Estimation and inference for measures of association

Confidence interval for the odds ratio

$\hookrightarrow$ Because the exposed and unexposed are samples are independent of each other, we can write

$$\text{var}\{\log(\widehat{\text{OR}})\} = \text{var}\left\{\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right\} + \text{var}\left\{\log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right)\right\}.$$

$\hookrightarrow$ Let us now deal with the first term in the right hand side of the above expression and let us write

$$f(\widehat{p}_1) = \log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right).$$

$\hookrightarrow$ We do not know how to calculate directly the variance of $f(\widehat{p}_1)$. However, we do know how to calculate the variance of $\widehat{p}_1$ (more later) and so in order to make the calculations feasible, let us use a Taylor expansion, of order 1, at p_1, so that

$$f(\widehat{p}_1) \approx f(p_1) + (\widehat{p}_1 - p_1)f'(p_1).$$

$\hookrightarrow$ Further, we have that

$$f'(p_1) = \frac{1}{p_1(1-p_1)},$$

and hence

$$f(\widehat{p}_1) \approx \log\left(\frac{p_1}{1-p_1}\right) + (\widehat{p}_1 - p_1)\frac{1}{p_1(1-p_1)}.$$

Estimation and inference for measures of association
Confidence interval for the odds ratio

$\hookrightarrow$ Therefore,

$$\text{var}\left\{\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right\} \approx \text{var}\left\{\log\left(\frac{p_1}{1-p_1}\right) + (\widehat{p}_1 - p_1)\frac{1}{p_1(1-p_1)}\right\}.$$

$\hookrightarrow$ Because p_1 is a constant (unknown but constant),

$$\text{var}\left\{\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right\} \approx \text{var}(\widehat{p}_1)\frac{1}{[p_1(1-p_1)]^2}.$$

Estimation and inference for measures of association
Confidence interval for the odds ratio

↪ The question now: what is the variance of $\widehat{p}_1$?

↪ We have that
$$\text{var}(\widehat{p}_1) = \text{var}\left(\frac{a}{a+b}\right) = \frac{1}{(a+b)^2}\text{var}(a),$$
where $a+b$, the number of exposed individuals, is fixed by design.

↪ In a cohort design, the a entry in the contingency table, denoting the number of exposed individuals with D, follows a binomial distribution with size $a+b$ (total number of exposed individuals) and probability of success p_1 (probability of disease given one is exposed).

↪ Therefore, we have that $\text{var}(a) = (a+b)p_1(1-p_1)$, thus leading to
$$\text{var}(\widehat{p}_1) = \frac{1}{(a+b)^2}(a+b)p_1(1-p_1) = \frac{p_1(1-p_1)}{(a+b)}.$$

Estimation and inference for measures of association
Confidence interval for the odds ratio

$\hookrightarrow$ We thus have

$$\text{var}\left\{\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right\} \approx \frac{p_1(1-p_1)}{(a+b)}\frac{1}{[p_1(1-p_1)]^2}$$
$$\frac{1}{(a+b)p_1(1-p_1)}.$$

$\hookrightarrow$ We are almost there but note that the expression of the variance we want depends on p_1, which is unknown. As it is usually done, we estimate this variance by plugging in $\widehat{p}_1$ for p_1, giving

$$\widehat{\text{var}}\left\{\log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right)\right\} \approx \frac{1}{(a+b)\widehat{p}_1(1-\widehat{p}_1)} = \frac{1}{a} + \frac{1}{b}.$$

$\hookrightarrow$ The same calculations lead to

$$\widehat{\text{var}}\left\{\log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right)\right\} \approx \frac{1}{c} + \frac{1}{d}.$$

Estimation and inference for measures of association
Confidence interval for the odds ratio

$\hookrightarrow$ We finally have the expression for the (estimated) sampling variance of $\log \widehat{\text{OR}}$

$$\widehat{\text{var}}(\log \widehat{\text{OR}}) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

$\hookrightarrow$ Under a case-control design we would arrive at the same variance estimate for $\log \widehat{\text{OR}}$.

$\hookrightarrow$ We thus have

$$\frac{\log \widehat{\text{OR}} - \log \text{OR}}{\sqrt{\widehat{\text{var}}(\log \widehat{\text{OR}})}} \dot{\sim} N(0, 1),$$

so that a two-sided $100(1 - \alpha)\%$ confidence limits to $\log \text{OR}$ are given by

$$\log \widehat{\text{OR}} \pm z_\alpha \sqrt{\widehat{\text{var}}(\log \widehat{\text{OR}})},$$

where z_α is the $1 - \alpha/2$th percentile of the standard normal distribution.

Estimation and inference for measures of association
Confidence interval for the odds ratio

↪ We have obtained a confidence interval for $\log$ OR. In order to obtain a CI for OR, our final goal, we just need to remember the following fact

> If g is a monotonic increasing function, then
>
> $$\Pr(L < X < U) = 1 - \alpha \Rightarrow \Pr(g(L) < g(X) < g(U)) = 1 - \alpha.$$

↪ Therefore, we only need to exponentiate the limits of the $\log$ OR to obtain a CI for OR.

↪ If the CI for the odds ratio does not contain the value one (or if the CI for the log odds ratio does not contain the value zero), then this is equivalent to rejecting the null hypothesis $H_0 : \text{OR} = 1$, that is, it is equivalent to rejecting the null hypothesis that there is no association between E and D.

Estimation and inference for measures of association
Confidence interval for the odds ratio: example

Sudden unexplained deaths in apparently normal babies under one year of age are known as sudden infant deaths, or cot deaths. In the UK, they are the leading cause of death in babies aged between one month and one year.

The causes of Sudden Infant Death Syndrome (SIDS) are not known. In 1990 a case-control study was published (*) that suggested that babies who were put down to sleep on their front and who were too heavily wrapped were more likely to die of SIDS.

Following this study, the 'Back to Sleep' campaign was launched in several countries to encourage parents to place their babies to sleep on their back and to avoid overheating and smoky environments. In subsequent years, deaths from SIDS dropped by over 50%. The table in the next slide shows data from the 1990 study.

(*) Flemming, P.J., Gilbert, R., Azaz, Y. et al. (1990). Interaction between bedding and sleeping position in the sudden infant death syndrome: a population based case-control study. *British Medical Journal*, **301**, 85–89.

Estimation and inference for measures of association
Confidence interval for the odds ratio: example

Position baby last placed down to sleep	Cases	Controls
On its front	62	76
In another position	5	55

A total of 67 babies who died of SIDS were included. The controls are live babies of similar ages and from the same localities as the babies that died. The exposure is last placing the baby down to sleep on its front.

1. Estimate the odds ratio of SIDS associated with the front sleeping position.

2. Calculate a 95% confidence interval for the odds ratio.

3. Interpret the results.

Estimation and inference for measures of association
Confidence interval for the odds ratio: example

↪ We know that the estimate of the $\widehat{\text{OR}}$ is given by ad/bc, and so we have
$\widehat{\text{OR}} = (62 \times 55)/(76 \times 5) = 8.974$ (3 dp).

↪ We thus have that $\log \widehat{\text{OR}} = \log 8.974 = 2.194$.

↪ We know that

$$\widehat{\text{var}}(\log \widehat{\text{OR}}) \approx \frac{1}{62} + \frac{1}{76} + \frac{1}{5} + \frac{1}{55} = 0.247 (\text{3dp}).$$

↪ Given that $z_{0.975} = 1.96$, the 95% CI for log OR is

$$(2.194 - 1.96 \times \sqrt{0.247}, 2.194 + 1.96 \times \sqrt{0.247}) = (1.220, 3.168).$$

↪ This leads to the following 95% CI for OR

$$(e^{1.220}, e^{3.168}) = (3.387, 23.760).$$

Estimation and inference for measures of association
Confidence interval for the odds ratio: example

↪ The estimated odds ratio is 8.974. This means that the odds of SIDS for babies placed to sleep on their front are roughly 9 times the odds of SIDS for babies laid down to sleep in other positions.

↪ The 95% CI for the OR is $(3.387, 23.760)$. We are 95% confident that in the population the odds of SIDS for babies placed to sleep on their front are, roughly, 3 to 24 higher than the odds of SIDS for babies laid down to sleep in other positions.

↪ The CI limits lie well above 1. Therefore, the data indicate that there exists a positive association between death from SIDS and putting the baby down to sleep on its front.

↪ SIDS seems to be a rare condition; see more here:

    https://www.nhs.uk/conditions/sudden-infant-death-syndrome-sids/

and so we can approximate the relative risk by the odds ratio (remember this is a case-control study and hence we cannot estimate the RR directly).

Estimation and inference for measures of association
Confidence interval for the relative risk

$\hookrightarrow$ The relative risk can be estimated in a cohort study by simply estimating $p_1 = \Pr(D \mid E)$ and $p_2 = \Pr(D \mid \text{not } E)$, i.e.,

$$\widehat{RR} = \frac{\widehat{p}_1}{\widehat{p}_2} = \frac{a/(a+b)}{c/(c+d)}.$$

$\hookrightarrow$ The sampling distribution of $\widehat{RR}$ is also skewed to the right but, again, the $\log$ transformation helps (see the plots in the Supplementary Materials file).

$\hookrightarrow$ To centre the confidence interval, we estimate $\log RR$ by $\log \widehat{RR}$.

$\hookrightarrow$ To estimate the variance of the approximate normal sampling distribution we follow a similar reasoning as we did for the odds ratio

$$\text{var}(\log \widehat{RR}) = \text{var}\left\{ \log\left(\frac{\widehat{p}_1}{\widehat{p}_2}\right) \right\} = \text{var}(\log \widehat{p}_1 - \log \widehat{p}_2) = \text{var}(\log \widehat{p}_1) + \text{var}(\log \widehat{p}_2).$$

Estimation and inference for measures of association
Confidence interval for the relative risk

$\hookrightarrow$ As we did for the OR, here we will also use a Taylor expansion of $f(\widehat{p}_1) = \log \widehat{p}_1$, of order 1, at p_1, so that

$$\log \widehat{p}_1 \approx f(p_1) + (\widehat{p}_1 - p_1)f'(p_1)$$
$$= \log p_1 + (\widehat{p}_1 - p_1)\frac{1}{p_1},$$

and therefore

$$\text{var}(\log \widehat{p}_1) \approx \frac{1}{p_1^2}\text{var}(\widehat{p}_1).$$

$\hookrightarrow$ Because we already know that $\text{var}(\widehat{p}_1) = p_1(1 - p_1)/(a + b)$, we finally have

$$\text{var}(\log \widehat{p}_1) \approx \frac{1 - p_1}{p_1(a + b)}.$$

Estimation and inference for measures of association

Confidence interval for the relative risk

$\hookrightarrow$ Again, we plug in $\widehat{p}_1$ for p_1, obtaining the following estimate for the variance of the estimate of the log relative risk

$$\begin{aligned}\widehat{\text{var}}(\log \widehat{p}_1) &\approx \frac{1 - \widehat{p}_1}{\widehat{p}_1(a+b)} \\ &= \frac{b/(a+b)}{(a/(a+b))(a+b)} \\ &= \frac{b}{a(a+b)}\end{aligned}$$

$\hookrightarrow$ The same calculations lead to

$$\widehat{\text{var}}(\log \widehat{p}_2) \approx \frac{d}{c(c+d)}$$

$\hookrightarrow$ Finally,

$$\widehat{\text{var}}(\log \widehat{\text{RR}}) \approx \frac{b}{a(a+b)} + \frac{d}{c(c+d)}.$$

Confidence interval for the relative risk

$\hookrightarrow$ As before and as usual,

$$\frac{\log \widehat{\text{RR}} - \log \text{RR}}{\sqrt{\widehat{\text{var}}(\log \widehat{\text{RR}})}} \dot\sim N(0, 1),$$

and now we can determine the confidence limits exactly as we did for the log odds ratio. Exponentiating the limits gives us the CI for the relative risk.