Comparing two proportions

1. Effect estimates (risk difference, relative risk, odds ratio)
2. 2 × 2 contingency tables

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
11.04.2024

Risk difference

- The **risk difference** *RD* is a measure of the difference in risk, $\pi_1 - \pi_0$, between the exposed and unexposed groups in the population
- It is estimated by the sample difference

$$\widehat{\mathrm{RD}} = p_1 - p_0$$

- Providing that
    - $n_1 \times p_1 \geqslant 10$ and $n_1 \times (1 - p_1) \geqslant 10$ in the exposed group, and
    - $n_0 \times p_0 \geqslant 10$ and $n_0 \times (1 - p_0) \geqslant 10$ in the unexposed group

  we use a **normal approximation** to the sampling distribution of $p_1 - p_0$

- The **standard error** of the sample difference is

$$\begin{aligned}\mathrm{s.e.}(p_1 - p_0) &= \sqrt{\mathrm{s.e.}(p_1)^2 + \mathrm{s.e.}(p_0)^2} \\ &= \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_0(1-\pi_0)}{n_0}},\end{aligned}$$

  where $\mathrm{s.e.}(p_1)$ and $\mathrm{s.e.}(p_0)$ are the standard errors of the proportions in the exposed and unexposed groups respectively

CI for the risk difference

- The **confidence interval** for the risk difference, i.e., for the difference between two proportions $\pi_1 - \pi_0$, is given by

  $$\mathrm{CI} = (p_1 - p_0) \pm z' \times \mathrm{s.e.}(p_1 - p_0),$$

  where z' is the appropriate percentage point of the standard normal distribution

Example: 16.1 in Kirkwood & Sterne

We consider the results from an **influenza vaccine trial** carried out during an epidemic.

Of $n = 460$ adults who took part, $n_1 = 240$ received **influenza vaccination** and $n_0 = 220$ received **placebo vaccination**. Overall $d = 100$ people contracted influenza, of whom $d_1 = 20$ were in the vaccine group and $d_0 = 80$ in the placebo group.

The results are displayed in a 2×2 table.

	Influenza		
	Yes	**No**	**Total**
Vaccine	20 (8.3%)	220 (91.7%)	240 (100%)
Placebo	80 (36.4%)	140 (63.6%)	220 (100%)
Total	100 (21.7%)	360 (78.3%)	460 (100%)

The **overall proportion** of subjects in the sample who got influenza is

$$p = \frac{100}{460} = 0.217 = 21.7\%$$

The percentage getting influenza was much lower in the vaccine group (8.3%) than in the placebo group (36.4%)

The estimated **risk difference** between the vaccine and placebo groups is:

$$\widehat{\mathrm{RD}} = 0.083 - 0.364 = -0.281.$$

Its estimated **standard error** is

$$\begin{aligned}\widehat{\mathrm{s.e.}}(p_1 - p_0) &= \sqrt{\frac{0.083 \times (1 - 0.083)}{240} + \frac{0.364 \times (1 - 0.364)}{220}} \\ &= 0.037.\end{aligned}$$

The approximate 95% **confidence interval** for this reduction is:

$$\begin{aligned} 95\%\ \text{CI} &= (-0.281 - 1.96 \times 0.037, -0.281 + 1.96 \times 0.037) \\ &= (-0.353, -0.208). \end{aligned}$$

This means that we are 95% confident that in the population the vaccine would reduce the risk of contracting influenza by between 20.8% and 35.3%.

Relative Risk

- The **relative risk**, or **risk ratio**, *RR* is the ratio of the two population proportions π_1/π_0
- Estimated by

$$\widehat{\mathrm{RR}} = \frac{p_1}{p_0} = \frac{d_1/n_1}{d_0/n_0},$$

  where p_1 and p_0 are the sample proportions in the exposed and unexposed groups

Properties of the relative risk

- $RR = 1$: the risks are the same in the two groups
- $RR > 1$: the risk of the outcome is *higher* among those exposed to the risk factor
- $RR < 1$: the risk of the outcome is *lower* among those exposed to the risk factor

- The further the relative risk is from 1, the stronger the association between exposure and outcome

CI for the relative risk

- The 95% **confidence interval** for the relative risk is

$$95\%\ \mathsf{CI} = \left(\exp\left\{\log \widehat{\mathrm{RR}} - 1.96 \times \mathrm{s.e.}\left(\log \widehat{\mathrm{RR}}\right)\right\},\right.$$
$$\left.\exp\left\{\log \widehat{\mathrm{RR}} + 1.96 \times \mathrm{s.e.}\left(\log \widehat{\mathrm{RR}}\right)\right\}\right),$$

  where the estimated **standard error** of the natural logarithm of the estimated risk ratio (i.e., the sample ratio) is

$$\widehat{\mathrm{s.e.}}(\log \widehat{\mathrm{RR}}) = \sqrt{1/d_1 - 1/n_1 + 1/d_0 - 1/n_0}$$

Example: 16.2 in Kirkwood & Sterne

	Lung cancer		
	Yes	**No**	**Total**
Smokers (exposed)	39 (0.13%)	29961 (99.87%)	30000 (100%)
Non-smokers (unexposed)	6 (0.01%)	59994 (99.99%)	60000 (100%)
Total	45 (0.05%)	89955 (99.95%)	90000 (100%)

A **cohort study** to investigate the association between smoking and lung cancer. The estimated **risk ratio** is

$$\widehat{\mathrm{RR}} = \frac{0.0013}{0.0001} = 13.$$

The estimated **standard error** of the natural logarithm of the estimated risk ratio is:

$$\widehat{\mathrm{s.e.}}(\log \widehat{\mathrm{RR}}) = \sqrt{1/39 - 1/30000 + 1/6 - 1/60000} = 0.438$$

The 95% **confidence interval** for the risk ratio is therefore:

$$\begin{aligned} 95\%\ \mathsf{CI} &= (\exp\{\log(13) - 1.96 \times 0.438\}, \\ &\qquad \exp\{\log(13) + 1.96 \times 0.438\}) \\ &= (5.5, 30.7). \end{aligned}$$

This means that we are 95% confident that the risk of lung cancer among smokers is between 5.5 and 30.7 times larger than the risk of lung cancer among non-smokers

Odds

- The **odds** of an outcome D is defined as

$$\mathrm{Odds} = \frac{\mathrm{P}(\text{D happens})}{\mathrm{P}(\text{D does not happen})} = \frac{\mathrm{P}(\mathrm{D})}{1 - \mathrm{P}(\mathrm{D})}$$

- The odds is estimated by

$$\widehat{\mathrm{Odds}} = \frac{p}{1-p} = \frac{d/n}{1-d/n} = \frac{d/n}{h/n} = \frac{d}{h},$$

  which is the number of individuals who experience the event divided by the number of individuals who *do not* experience the event

Odds Ratio

- The **odds ratio** is denoted by OR and is the ratio between the odds in the exposed group and the odds in the unexposed group
- It is estimated by

$$\widehat{\mathrm{OR}} = \frac{d_1/h_1}{d_0/h_0} = \frac{d_1 \times h_0}{d_0 \times h_1},$$

  which is also known as the **cross-product ratio** of the 2×2 table

Properties of the odds ratio

- OR is one of the most common effect measures in medical statistics, even though it is less intuitive than RR
- Odds used in for example **logistic regression**
- $OR = 1$ occurs when the odds, and hence the proportions, are the same in the two groups
- The OR is **always further away from 1 than the corresponding** RR,
- For **rare outcomes** the OR is approximately equal to the RR
- **OR(disease) = 1/OR(healthy)** (this is not the case for RR)

Example: 16.4 in Kirkwood & Sterne

Consider a study in which we monitor the risk of **severe nausea** during chemotherapy for breast cancer. A **new drug** is compared with **standard treatment**

	Number with severe nausea	**Number without severe nausea**	**Total**
New drug	88 (88%)	12	100
Standard treatment	71 (71%)	29	100

The estimated **risk** of severe nausea in the group treated with the new drug is

$$p_1 = \frac{88}{100} = 0.880 = 88.0\%,$$

and the estimated **risk** of severe nausea in the group given the standard treatment is

$$p_0 = \frac{71}{100} = 0.710 = 71.0\%.$$

The estimated **relative risk** is

$$\widehat{\mathrm{RR}} = \frac{88/100}{71/100} = 1.239,$$

an apparently moderate increase in the prevalence of nausea. The estimated **odds ratio** is

$$\widehat{\mathrm{OR}} = \frac{88/12}{71/29} = 2.995,$$

a much more dramatic increase

Suppose now that we consider our outcome to be *absence* of nausea. Then the estimated **risk ratio** is:

$$\widehat{\mathrm{RR}} = \frac{12/100}{29/100} = 0.414,$$

which means that the proportion of patients without severe nausea has more than halved. The estimated **odds ratio** is:

$$\widehat{\mathrm{OR}} = \frac{12/88}{29/71} = 0.334,$$

which is exactly the inverse of the odds ratio for nausea $(1/2.995=0.334)$

CI for the odds ratio

- The 95% **confidence interval** for the odds ratio is

  $$95\%\ \mathrm{CI} = \left(\exp\left\{\log\widehat{\mathrm{OR}} - 1.96 \times \mathrm{s.e.}\left(\log\widehat{\mathrm{OR}}\right)\right\},\right.$$
  $$\left.\exp\left\{\log\widehat{\mathrm{OR}} + 1.96 \times \mathrm{s.e.}\left(\log\widehat{\mathrm{OR}}\right)\right\}\right),$$

  where the estimated **standard error** of the natural logarithm of the estimated odds ratio (i.e., the sample ratio) is

  $$\widehat{\mathrm{s.e.}}(\log\widehat{\mathrm{OR}}) = \sqrt{1/d_1 + 1/h_1 + 1/d_0 + 1/h_0},$$

  which is also known as **Woolf's formula**

Example: 16.3 in Kirkwood & Sterne

Consider the survey from Example 15.5 in Kirkwood & Sterne (2003) of $n = 2000$ patients aged 15 to 50 registered with a particular general practice. It showed that $d = 138$ (6.9%) were being treated for asthma.

	Asthma		
	Yes	**No**	**Total**
Women	81	995	1076
Men	57	867	924
Total	138	1862	2000

The estimated **prevalences** of asthma (proportions with asthma) in women and men are:

$$p_1 = \frac{81}{1076} = 0.0753 = 7.53\%$$

and

$$p_0 = \frac{57}{924} = 0.0617 = 6.17\%,$$

respectively. The estimated **risk ratio** is:

$$\widehat{\mathrm{RR}} = \frac{0.0753}{0.0617} = 1.220.$$

The estimated **odds** of asthma in women and men are:

$$\frac{p_1}{h_1} = \frac{81}{995} = 0.0814$$

and

$$\frac{p_0}{h_0} = \frac{57}{867} = 0.0657,$$

respectively. The estimated **odds ratio** is:

$$\widehat{\mathrm{OR}} = \frac{0.0814}{0.0657} = 1.238.$$

The estimated odds ratio of 1.238 indicates that asthma is more common among women than men.

The estimated **standard error** of the natural logarithm of the estimated odds ratio is given by

$$\widehat{\text{s.e.}}(\log \widehat{\text{OR}}) = \sqrt{1/81 + 1/995 + 1/57 + 1/867} = 0.179$$

The 95% **confidence interval** for the odds ratio is therefore:

$$\begin{aligned} 95\%\ \mathrm{CI} &= (\exp\{\log(1.238) - 1.96 \times 0.179\}, \\ &\qquad \exp\{\log(1.238) + 1.96 \times 0.179\}) \\ &= (0.872, 1.758) \end{aligned}$$

This means that with 95% confidence, the odds ratio in the population lies between 0.872 and 1.758