Lecture 34

Categorical Responses: proportions and odds

For the rest of the semester, we will discuss the analysis of categorical response data. The simplest case is when there are two responses (yes/no or success/failure) and that is what we will mainly concentrate on. For the most basic analysis: predicting a binary response from a binary predictor (such as is a voter Democrat or Republican based on the voter's sex: male or female) there are many possible analyses. We will review four possibilities and discuss the strength and weakness of each in this and the following lecture.

Data involving a binary response and a binary predictor can be summarized in a $\mathbf{2 \times 2}$ table. For discussing possible analyses we will use the following summary data of handedness in twins born between 1900 and 1910 in Denmark.

Sex	Right handed	Not right handed
Male	1174	119
female	1137	79

Two proportions

For the above table, one might ask whether the proportion of male twins who are left-handed is equal to the proportion of female twins who are left handed and one might want to be able to give a confidence interval for the difference. For the data above, the proportion of male twins who are left handed is $\boldsymbol{p_1 = 119/(119 + 1174) = 0.092}$ while the proportion of female twins who are left handed is $\boldsymbol{p_2 = 79/(79 + 1137) = 0.065}$. Proportions are approximately normally distributed when the sample size is large and $\boldsymbol{np}$ and $\boldsymbol{n(1-p)}$ are both large (> 10) which is the case with these data. So what we need to know is the standard error. For a test, the null hypothesis is that the proportions are equal so one uses the pooled proportion in calculating the standard error. For confidence intervals, there is no such assumption, so the standard error is calculated differently. The formulas are as follows:

For a test: $\boldsymbol{Z = \dfrac{p_1 - p_2}{\text{S.E.}}}$ where S.E. $= \boldsymbol{\sqrt{p_c(1-p_c)(\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$ and $\boldsymbol{p_c = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}}$

For a confidence interval: $\boldsymbol{p_1 - p_2 \pm z_{\alpha/2}}$S.E. where S.E. $= \boldsymbol{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$

The statistic has a standard normal distribution (there are no degrees of freedom: the mean determines the variance for a proportion). For the example above: $\boldsymbol{p_c = \frac{1293 \times 0.092 + 1216 \times 0.065}{1293 + 1216}} =$

$\mathbf{0.079}$ and S.E. $= \sqrt{\mathbf{0.079 \times 0.921 \times (1/1293 + 1/1216)}} = \mathbf{0.0107}$. Thus $\boldsymbol{Z = 2.51}$ and the p-value for whether handedness is male and female twins are different or not is 0.012.

The confidence interval has a standard error that is not based on pooling the proportions. Thus the confidence interval has standard error: S.E. $= \sqrt{\mathbf{0.092 \times 0.908/1293 + 0.065 \times 0.935/1216}} = \mathbf{0.0107}$ (the same in this case: in fact, the difference in the calculation of the formulas to calculate standard errors are often negligible) and the confidence interval is $\mathbf{(0.0061, 0.0480)}$.

These tests and confidence intervals provide a straightforward answer to whether two proportions are equal or not. The requirements are that the sample sizes are large enough that $\boldsymbol{np}$ and $\boldsymbol{n(1-p)}$ are both greater than 10 in both populations. Further, the proportions should be meaningful population proportions. Here, one might suspect that the proportions of left handedness in twins in Denmark might generalize to individuals elsewhere - not necessarily twins and not necessarily Danes.

Odds Ratios

We will begin by making the straightforward analysis of proportions more complicated. After we have gone though the math, we will tell why the more complicated approach of analyzing odds ratios is actually critical to many experimental designs, especially in medicine.

Odds are ratios of proportions of successes to their complements: proportions of failures. That is:

$$\boldsymbol{w = \frac{p}{1-p}}$$

So if the odds are 50:50 or 1:1 then $\boldsymbol{p = 1/2}$. If the odds are 2:1 then $\boldsymbol{p = 2/3}$. In general, if the odds are $\boldsymbol{a : b}$ then $\boldsymbol{p = a/(a+b)}$. For very small values of $\boldsymbol{p}$ the odds are just about equal to $\boldsymbol{p}$. For the data above, the odds a male twin is left handed is $\mathbf{0.092/(1-0.092) = 0.10}$ which is very close to the probability a male twin is left handed. The odds a female twin is left-handed is $\mathbf{0.065/(1-0.065) = 0.0695}$ which is very close to the probability a female twin is left handed.

The question of whether the proportion of left-handed individuals is the same for males and females turns into a question of whether the odds ratio is 1. The odds ratio is $\boldsymbol{w_1/w_2}$ and it varies from zero to infinity. This begs for a log transform. So we discuss the statistics of the log odds ratio:

$$\text{log odds ratio} = \mathbf{log}\left(\frac{\boldsymbol{w_1}}{\boldsymbol{w_2}}\right)$$

For a test, the standard error is calculated as:

$$\sqrt{\frac{\mathbf{1}}{\boldsymbol{p_c(1-p_c)}}(\frac{\mathbf{1}}{\boldsymbol{n_1}} + \frac{\mathbf{1}}{\boldsymbol{n_2}})}$$

whereas for a confidence interval it is calculated as

$$\sqrt{\frac{\mathbf{1}}{\boldsymbol{n_1 \times p_1(1-p_1)}} + \frac{\mathbf{1}}{\boldsymbol{n_2 \times p_2(1-p_2)}}}.$$

Again, the distribution is standard normal.

The corresponding test statistic to the test for equal proportions for the log odds ratio is

$$\boldsymbol{Z = \frac{\log(\frac{0.092/(1-0.092)}{0.065/(1-0.065)})}{\sqrt{\frac{1}{0.079\times 0.921}(\frac{1}{1293}+\frac{1}{1216})}} = 2.54}$$

with 2-sided p-value 0.011.

Why would one ever want to consider log odds ratios instead of the more straightforward population proportions test? The reason is that odds ratios are meaningful when the desired proportions are not. Since the proportions of interest are meaningful above, consider the following example instead:

	Lung cancer	No lung cancer
Smokers	40	20
Non-smokers	60	80

The above data are fabricated but they represent a retrospective study of lung cancer with 100 lung cancer victims matched to 100 controls. Since the number of lung cancer cases and the number of controls is fixed, the proportion of smokers who got lung cancer in this study has no population equivalent. That is, the proportion of smokers who got lung cancer in this study is a meaningless statistic. However, the beauty of the odds ratio is that it is the same no matter whether the columns or the rows are meaningful. That is

$$\frac{\text{the odds a person who smokes gets cancer}}{\text{the odds a person who does not smoke gets cancer}} = \frac{\text{the odds a lung cancer victim is a smoker}}{\text{the odds a non-lung cancer victim is a smoker}}$$

$$= \boldsymbol{\frac{40 \times 80}{60 \times 20}}$$

References and Readings

[1] Olga Basso, Jorn Olsen, Niels V. Holm, Axel Skytthe, James W. Vaupel, and Kaare Christenson. Handedness and mortality: A follow-up study of Danish twins born between 1900 and 1910. *Epidemiology*, 11:576–580, 2000.

Exercises for Lecture 34

1. –

2. –