6.3 Other Methods for Comparing Proportions

6.3.1 Odds Ratios

An alternative way of analyzing two groups in terms of how likely some outcome is to occur is through an *odds ratio*. We saw how to calculate the odds of an outcome in Section 1.4.3. For example, the probability of a mangrove snail found on the trunk being yellow was estimated to be 0.1259. The odds of the snail being yellow are then

$$\frac{0.1259}{1-0.1259} = 0.144 \text{ to } 1.$$

Similarly, the probability of a mangrove snail found on the foliage being yellow was 0.3728. The odds of the snail on the foliage being yellow are

$$\frac{0.3728}{1-0.3728} = 0.594 \text{ to } 1.$$

Thus the ratio of the odds of being yellow between the foliage and trunk is

$$\frac{0.594}{0.144} = 4.13.$$

That is, the odds of a foliage snail being yellow are 4.13 times the odds of a trunk snail being yellow. This suggests that snails are more likely to be yellow on the foliage than on the trunk, the same relationship we saw when looking at proportions.

Nicotine Inhalers

A common use for odds ratios is to assess the effect of the presence of some condition on certain outcomes. As an example, Bolliger *et al.* [4] describe a randomized double-blind experiment on the effectiveness of oral nicotine inhalers in reducing smoking. This involved 400 volunteers who had smoked at least 15 cigarettes for at least 3 years, and who had tried to reduce their smoking but had failed to do so. The subjects were given an oral inhaler to use as needed, for up to 18 months, and were encouraged to limit their smoking as much as possible. Nicotine inhalers were randomly assigned to half of the subjects while the other half received a placebo.

After 4 months the researchers recorded which subjects had sustained a reduction of at least 50% in the number of cigarettes smoked each day. Table 6.3.1 gives a two-way table of these results, with 26% of the nicotine group achieving a smoking reduction compared to only 9% for the placebo group. The odds for a reduction in the nicotine group are 0.26/0.74 = 0.3514 to 1, while in the placebo group the odds are 0.09/0.91 = 0.0989 to 1. This gives an odds ratio of

$$\text{OR} = \frac{.3514}{.0989} = 3.55.$$

That is, the odds of sustaining a reduction in smoking after 4 months are 3.55 times higher if someone is using a nicotine inhaler[2].

Table 6.3.1: Sustained reductions after 4 months of inhaler use

	Nicotine	Placebo
Reduction	52	18
No Reduction	148	182
Total	200	200

6.3.2 Confidence Intervals

Finding an odds ratio of 3.55 seems to suggest that there is evidence that nicotine inhalers are beneficial in assisting the sustained reduction of smoking. However, it should be clear by now that we are not happy with an estimate by itself. We need some measure of precision. Could it be that there is really no effect and the ratio of 3.55 was just due to sampling variability?

We can determine a confidence interval for the true odds ratio in a similar way to those we have already calculated for means and proportions. The main difference arises from the fact that odds ratios can never be negative but they can be arbitrarily large. It is no surprise then that the sampling distribution for odds ratios is going to be skewed to the right, so our methods based on the Normal

[2] Notice that in this sentence an odds ratio has been used but only one treatment has been mentioned. This is typically what you will find when reading research articles and you should always ask yourself what is the underlying group for the odds ratio. Typically it will be a control group of some form, such as with the placebo treatment used in this study.

distribution are not going to be appropriate. However, it turns out that if you take the logarithm[3] of the odds ratio then you get a statistic where the sampling distribution can be approximated by the Normal. Here we find

$$\log(\text{OR}) = \log(3.55) = 1.267.$$

All we need now is the standard error of this statistic. The formula for this is given by

$$\text{se}(\log(\text{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

where the counts a, b, c, and d are the four entries in the two-way table. From Table 6.3.1, we have $a = 52$, $b = 18$, $c = 148$, and $d = 182$, giving

$$\text{se}(\log(\text{OR})) = \sqrt{\frac{1}{52} + \frac{1}{18} + \frac{1}{148} + \frac{1}{182}} = 0.2950.$$

Now a 95% confidence interval for $\log(\text{OR})$ is

$$1.267 \pm 1.96 \times 0.2950 = 1.267 \pm 0.5782,$$

giving the range $(0.6888, 1.845)$ for $\log(\text{OR})$. This is not what we want though, since we are interested in the odds ratio itself, rather than its logarithm. We can obtain the confidence interval for the odds ratio by raising e to the power of each endpoint. This gives

$$(e^{0.6888}, e^{1.845}) = (1.991, 6.328).$$

Thus we are 95% confident that the odds of a sustained reduction in smoking is between 1.99 and 6.33 times higher when using a nicotine inhaler.

If we were testing a null hypothesis that the nicotine inhaler had no effect on the reduction of smoking then we would expect an odds ratio of 1. This would mean that the odds were the same for both groups. Since 1 is outside the confidence interval we have found, we have evidence against this null hypothesis, suggesting that nicotine inhalers are effective.

6.3.3 Logistic Regression

One reason that odds ratios are often used, instead of proportions, is because of their link to logistic regression. In Section 2.4 we saw that it was not possible to

[3] Unless otherwise stated, all logarithms are natural logarithms. We write "log" for "$\log_e$".

model the relationship between a categorical response variable and a quantitative predictor using standard linear regression. The alternative was to use *logistic regression* instead, and this was achieved by fitting a straight line to the logarithm of the odds.

The example we used in Section 2.4 looked at the relationship between height and p, the probability that a person is male. Logistic regression gave the straight line

$$\log\left(\frac{p}{1-p}\right) = -30.513 + 0.176x,$$

where x was height in centimetres. The "logarithm of the odds" is affectionately called the "log odds". We rearranged this for p, giving the logistic curve shown in Figure 2.4.1. However, we can also use it directly to calculate odds and odds ratios.

As an example, suppose we want to estimate how much more likely it is for a person to be male if they are 170 cm tall rather than 160 cm tall. We could estimate the individual odds by substituting 170 and 160 into the straight line equation and taking exponentials of each, and then finding the ratio. However, it is slightly easier to remember that the log of a ratio is the *difference* of the logs, and then use this to work out $\log(\text{OR})$ directly. Here we have

$$\begin{aligned}\log(\text{OR}) &= (-30.513 + 0.176 \times 170) - (-30.513 + 0.176 \times 160)\\ &= 0.176 \times 10 = 1.76.\end{aligned}$$

Thus the odds ratio is $e^{1.76}$ = 5.81, so the odds of a person who is 170 cm tall being male are 5.81 times the odds of a person who is only 160 cm tall.

Note that the intercepts cancel out when doing this calculation so all that matters is the slope. This value, 0.176, can thus be interpreted as the rate of increase in the log odds for each unit increase in the explanatory variable. This is analogous to the interpretation of slope for standard linear regression, to be discussed further in the next chapter.

Assuming that log odds can be described by a straight line means that the rate of increase is constant across all values of the explanatory variable. Thus the odds ratio of being male between 180 cm and 170 cm will be the same as the odds ratio between 170 cm and 160 cm, 5.81 from above. Of course, this may not always be a realistic assumption in practice.

6.3.4 Adjusted Odds Ratios

Logistic regression was used to estimate the log odds of being male for a particular height, x, using the straight line

$$\log\left(\frac{p}{1-p}\right) = -30.513 + 0.176x.$$

However, regression can also be carried out when there is more than one explanatory variable. It turns out that a better estimate of the log odds of being male can be calculated by

$$\log\left(\frac{p}{1-p}\right) = -93.5 + 0.232x_1 + 1.883x_2,$$

where x_1 is height and x_2 is shoe length, both in centimetres. For example, if a person is 170 cm tall but has a shoe which is 29 cm long, then

$$\log\left(\frac{p}{1-p}\right) = -93.5 + 0.232 \times 170 + 1.883 \times 29 = 0.547,$$

so the odds of being male are $e^{0.547}$ = 1.73 to 1, suggesting they are more likely to be male. Knowing only that they were 170 cm tall, the original relationship gives a log odds of -0.593 and so the odds are 0.55 to 1. From these odds we might have have thought that they were female.

We can do similar calculations with odds ratios. Previously we found that the odds of being male were 5.81 times higher for someone 170 cm tall than for someone 160 cm. Calculating the same odds ratio from the new model uses

$$\begin{aligned}\log(\text{OR}) &= (-93.5 + 0.232 \times 170 + 1.883x_2) \\ &\quad -(-93.5 + 0.232 \times 160 + 1.883x_2) \\ &= 0.232 \times 10 = 2.32.\end{aligned}$$

This gives an odds ratio of $e^{2.32}$ = 10.2, much higher than the previous estimate of 5.81. Note that it did not matter what the shoe length, x_2, actually was, since those terms cancelled out. Why then do we get a different odds ratio?

The reason is that adding shoe length has helped explain more of the variability in the sex observations than could be explained by height alone. This has allowed us to be more precise about the effect of height on probable sex, in this case resulting in a higher odds ratio. We call this an *adjusted odds ratio*.

Fake Tans and Indicator Variables

Beckmann *et al.* [3] carried out a study which illustrates a common use of adjusted odds ratios. A telephone survey, with 2005 participants, was carried out to explore the use of fake tanning lotions and its relationship with other factors and outcomes. One of the questions asked was "Over the last summer, did you get sunburn which was sore and tender the next day?" Table 6.3.2 shows a two-way table of responses split by whether they said "Yes" to this sunburn question and whether they had used a fake tanning lotion.

Table 6.3.2: Sustained reductions after 4 months of inhaler use

	Sunburn		
Fake Tanning	Yes	No	Total
Yes	46	129	175
No	302	1528	1830
Total	348	1657	2005

From this table, the odds of sunburn for people who used a fake tanning lotion are

$$\frac{46/175}{129/175} = 0.3566 \text{ to } 1,$$

while for those who did not use a fake tanning lotion, the odds of sunburn are

$$\frac{302/1830}{1528/1830} = 0.1976 \text{ to } 1.$$

This gives an odds ratio of

$$\text{OR} = \frac{0.3566}{0.1976} = 1.80,$$

so the odds of being sunburnt are 1.80 times higher if a fake tanning lotion was used.

The log odds ratio is $\log(\text{OR}) = 0.588$ with standard error $\text{se}(\log(\text{OR})) = 0.183$, giving a 95% confidence interval for $\log(\text{OR})$ of $(0.229, 0.947)$. Taking exponentials, a 95% confidence interval for the odds ratio is

$$(e^{0.229}, e^{0.947}) = (1.27, 2.58).$$

Since 1 is not in this range, there appears to be evidence then that fake tanning lotions are associated with higher rates of sunburn.

The calculation of the odds ratio and its confidence interval are straightforward, but caution must be used in such survey settings. The researchers were careful to also ask “How often do you wear SPF 15+ or higher sunscreen?” The results from this question showed that those who used a fake tanning lotion were more likely to regularly use sunscreen than those who did not. This means that the odds ratio of 1.80 will not truly reflect the effect of fake tan use on the odds of sunburn, since the use of sunscreen will tend to decrease these odds. We say that sunscreen use *confounds* the effect of fake tanning lotion use on sunburn rates.

Logistic regression provides a way of isolating the effect of the use of fake tanning lotion on the odds of sunburn by modelling the relationship with a number of factors simultaneously. The explanatory variables in this case are categorical. These are handled in regression using *indicator variables*. We might define a variable

$$x_1 = \begin{cases} 1, & \text{if fake tanning lotion was used} \\ 0, & \text{if fake tanning lotion was not used} \end{cases}$$

We can then find a linear relationship of the form

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

using logistic regression. The log of the odds ratio between those who used fake tanning lotions and those who didn’t is then

$$(b_0 + b_1 \times 1) - (b_0 + b_1 \times 0) = b_1,$$

so the odds ratio is simply e^{b_1}. Logistic regression for this single variable gives

$$\log\left(\frac{p}{1-p}\right) = -1.621 + 0.590 x_1.$$

Thus the odds ratio is $e^{0.590} = 1.80$, the same as before. Logistic regression can also be used to give confidence intervals for these estimates.

However, with logistic regression we can also add more indicator variables to the model, including factors such as sex, skin type, sunscreen use, hat wearing, protective clothing, and other sun-protection practices. We can then use the coefficient of the fake tanning lotion variable to estimate the odds ratio, but now it will have been adjusted to take into account the relationships between sunburn and the other possible factors. In this case the authors reported an adjusted odds ratio of 2.07, higher than when sunscreen use was not separated. The 95% confidence interval (1.17, 3.69), so we can conclude that, even after taking other factors into account, there is evidence that the odds of being sunburnt are higher for those who used fake tanning lotions than those who did not.

6.3.5 Relative Risk

A simpler comparison between proportions is given by *relative risk*. This is the ratio of the two probabilities of a certain outcome between two groups, rather than the ratio of the odds. For example, the proportion of people using a fake tanning lotion who were sunburnt was 46/175 = 0.2629 while for those not using a tanning lotion the proportion was 302/1830 = 0.1650. This gives a relative risk of

$$\text{RR} = \frac{0.2629}{0.1650} = 1.59.$$

That is, people using a fake tanning lotion are 1.59 times more likely to be sunburnt.

As the name suggests, relative risks are popular in studies of factors affecting the risk of diseases and accidents. For example, Åkerstedt *et al.* [1] followed a sample of 47 860 individuals in Sweden over a 20-year period. In that time, 166 suffered fatal accidents at work. Of the total individuals, 5659 were classified as having difficulties in sleeping, and of these 5659, 32 had fatal accidents. The remaining 134 fatal accidents involved the 42 201 people who were not exposed to sleeping difficulties. The increase in the risk of having a fatal accident can be measured by

$$\text{RR} = \frac{32/5659}{134/42201} = 1.78.$$

This relative risk suggests that people with sleeping difficulties are 1.78 times more likely to suffer a fatal accident at work.

Relative risks are not as commonly used as odds ratios because they do not have the same simple statistical theory or rich relationship with logistic regression.