

An Introduction to the Analysis of Proportions

These notes are linked to both the [Binomial](#) and [Categorical Data Analysis](#) notes. It also forms an introduction to topics that beyond the scope of these notes such as *Logistic Regression*. A comprehensive discussion of these and related topics can be found in [Collet, D. \(1991\)](#).

The binomial distribution assumes that we know p , the probability of a success. However, often all that we have is an estimate of its true value from a sample. *This is analogous to the statistic - parameter relationship*. Suppose that we have a sample obtained from plasma donors on Merseyside and we wish to use the proportion of smokers in this sample to estimate the proportion of all Merseyside plasma donors that smoked. If we questioned all of the plasma donors we could obtain the true proportion. If a value is an estimate a caret symbol (^) is often used to identify it as such, e.g. $\hat{p}$ is an estimate of p . If $\hat{p}$ is an estimate of p then $1 - \hat{p}$ must be an estimate of q .

- [Confidence limits for a proportions](#)
- [Comparing 2 proportions](#)
- [Odds ratios](#)

How reliable is $\hat{p}$?

We can assess the reliability of $\hat{p}$ as an estimate of p if we calculate its standard error and then its confidence limits.

$$s. e. (\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Using the [plasma donor data](#) as an example.

116/225 were smokers, therefore $\hat{p}$ is 0.515 and its standard error is

$$s. e. (0.515) = \sqrt{\frac{0.515(1 - 0.515)}{225}}$$

which is 0.0333. Because the divisor in the above equation is n , which is large, our estimate of $\hat{p}$ would appear to be precise. 0.0333 is small compared to 0.515. If n had been 25 the standard error would have been three times larger at 0.1.

The standard error can be converted into approximate confidence limits by making use of z values (standard normal deviates). The confidence limits are:

$$\hat{p} \pm z_{\alpha/2} \text{ s.e. } (\hat{p})$$

where $z_{\alpha/2}$ indicates the two-tailed value of z for a particular value of α . Thus:

α	confidence limits (1- α).100	$z_{\alpha/2}$ value
0.10	90%	1.645
0.05	95%	1.960
0.01	99%	2.576

Therefore the 95% confidence limits for the proportion of plasma donors from Merseyside that smoke are:

$$0.515 \pm 1.960 (0.0333) \text{ giving the limits } (0.4497, 0.5803).$$

Comparing two proportions

Is the proportion of male plasma donors that smoke different from the number of female plasma donors that smoke? Two approaches to this question will be described. The first uses confidence intervals, the second uses a z statistic.

If two proportions are the same the difference between them should be 0. However, because of sampling error it is possible to achieve a difference that is not 0 even when the two proportions are identical. We need to test if the difference is large enough to suggest the existence of a real difference in the two proportions.

Method 1 : Confidence intervals

We can calculate a confidence interval for the difference. If this interval includes 0 we have evidence to suggest that there is no real difference.

Let p_A be the true proportion of male smokers and p_B be the true proportion for female smokers (who are plasma donors in Merseyside). The true difference between the two is $(p_A - p_B)$ which we estimate from our sample as $(\hat{p}_A - \hat{p}_B)$.

$$\hat{p}_A = 82 / 140 = 0.586 \text{ and}$$

$$\hat{p}_B = 34 / 85 = 0.400.$$

$$(\hat{p}_A - \hat{p}_B) = 0.586 - 0.400 = 0.186$$

The standard error of the difference is found from

$$\sqrt{\left\{ \frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B} \right\}}$$

which is

$$\begin{aligned} & \sqrt{\left\{ \frac{0.586(0.414)}{140} + \frac{0.400(0.600)}{85} \right\}} \\ &= \sqrt{0.001733 + 0.002823} \\ &= 0.0675 \end{aligned}$$

From this we find approximate 95% confidence limits for the true difference. These are:

$$0.186 \pm 1.96 (0.0675) = (0.0537, 0.3183)$$

Because the interval defined by these limits does not include 0 we can say that the difference between the two proportions is significant at the 5% level and that we have evidence for the statement that proportion of male plasma donors who smoke is greater than the proportion of female plasma donors that smoke.

Method 2. A z test

We can test the same hypotheses by a z test. (H_0 : The two proportions are the same). The equation is:

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\left\{ \hat{p}_{AB}(1-\hat{p}_{AB}) \left[\frac{1}{n_A} + \frac{1}{n_B} \right] \right\}}}$$

$\hat{p}_{AB}$ is a common estimate of the proportion of plasma donors who smoke, i.e. ignoring their gender. This estimate is $(82 + 34) / (140 + 85) = 0.515$. Consequently our estimate of z is:

$$\begin{aligned} & \frac{0.586 - 0.400}{\sqrt{\left\{ 0.515(0.485) \left[\frac{1}{140} + \frac{1}{85} \right] \right\}}} \\ &= 2.707 \end{aligned}$$

From z tables we find the P value for this test to be 0.0068 ($< 1\%$) and we have strong evidence in favour of the H_1 . This suggests that the proportion of male plasma donors who smoke is greater than the proportion of female plasma donors who smoke.

Odds Ratios

The ratio of the probability of a success to the probability of a failure is called the odds. If we have two sets of measures we can calculate the ratio of their odds, this is known as the **odds ratio** and is usually given the symbol .

$$= \frac{p_A / (1 - p_A)}{p_B / (1 - p_B)}$$

Using the plasma donor data again we can find for male and female smokers.

$$\frac{0.586 / 0.414}{0.400 / 0.600} = 2.1234$$

What is the interpretation of the odds ratio?

If $\hat{p}_A = \hat{p}_B$ the two odds would be same and the odds ratio would be 1.0. In the plasma donor example the odds for male smokers is over twice that of females, hence males are more likely to smoke.

Again we can use a confidence interval as the basis for a test of significance. We wish to know if is significantly different from 1. If is significantly different from 1 there is evidence that the two odds are not the same.

$$\text{s.e.} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

where a , b , c & d are found from

	Successes	Failures
Group 1	a	b
Group 2	c	d

for the plasma donors

	Smoke	Non-Smoker
Males	82	58
Females	34	51

$$\text{therefore the s.e.} = \sqrt{\frac{1}{82} + \frac{1}{58} + \frac{1}{34} + \frac{1}{51}} = 0.28$$

and the 95% confidence limits are:

$$2.1234 \pm 1.96 (0.28) = (1.5744, 2.6724)$$

Since the interval defined by these limits does not include 1.0 we have evidence that the odds for male smokers is greater than that for female smokers. Hence male plasma donors are more likely to smoke than female plasma donors.

Summary

The examples in this appendix have demonstrated some different, and usually complementary, approaches to the analysis of proportions. Often they can be used as a more powerful alternative to the traditional chi-squared association analysis test.

Main Menu	Univariate statistics
---------------------------	---------------------------------------