Chapter 8. Statistical inference for proportions

8.1 Single Proportions

This chapter focusses on inference for proportions, and therefore data on categorical variables instead of inference for means or spread. Also this form of inference is based on approximately Normal distributions.

To estimate the unknown proportion p, you use the *sample proportion* of successes $\hat{p}$ = X/n. In the case of the population being at least 20 times larger than the sample, X (the count of successes) has a binomial (B(*n,p*)) distribution. When n is large, both X and $\hat{p}$ are approximately Normal, and this makes it possible to calculate significance tests and confidence intervals, despite the discrete nature of binomial distributions.

Single proportion confidence intervals

When n is sufficiently large $\hat{p}$ has mean $\mu_{\hat{p}}$=p and standard deviation $\sigma_{\hat{p}}$ = $\sqrt{p(1-p)/n}$. As p in the formula is unknown, we need to replace it by the estimate $\hat{p}$.

The standard error of $\hat{p}$ is therefore:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

As the margin of error must be m = z*$SE_{\hat{p}}$, then the C level confidence interval can be calculated by using: $\hat{p}$ ± m. You can use this interval for 90%, 95% or 99% confidence in the cases that both the failures and successes are each 15 or more.

Additional notes when using the CI:

- When reporting results, be sure to round up the values to the amount of digits that are meaningful to the purpose.
- Remember that the margin of error is based only on the random sampling error. If participants participate dishonestly then that affects the results, and this error is not included in the margin of error.

In the case of the number of successes and failures being less than 15, you can use the *plus four confidence interval*. The plus four rule assumes that the sample has 2 more successes and 2 more failures. As a result, the *plus four estimate* is written as below.

$$\tilde{p} = \frac{X+2}{n+4}$$

The standard error then becomes: $SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$, which can then be filled into the confidence interval formula.

Single proportion significance tests

In a significance test, the H_o specifies a value for p (so p_o), which is what we use as a substitute in the z test formula for a single proportion, as shown below. This is for the null hypothesis: H_o: $p = p_o$.

$$z = \frac{\hat{p} - po}{\sqrt{\frac{po(1-po)}{n}}}$$

The p-values can then be found according to:

- H_a: $p > p_o$ is $P(Z \geq z)$
- H_a: $p < p_o$ is $P(Z \leq z)$
- H_a: $p \neq p_o$ is $2P(Z \geq |z|)$

This test can be used as long as the expected number of successes (np_o) and the expected number of failures ($n(1-p_o)$) are each larger than 10.

As it is rare for there to be a specific p_o that we want to test, significance tests for single proportions are not used often. Whenever possible, comparative studies are always favoured.

Deciding on a sample size

To get a specific margin of error you need to guess the value of $\hat{p}$. The guessed value is known as p* and can be found by a) using a sample estimate from previous similar studies or a pilot study, or b) by using p*=0.5 as the margin of error is the largest at this value. It is a safe choice as it gives an *n* slightly larger than needed. The sample size needed can then be calculated by filling in the formula:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

When using a p* value of 0.5 an easier formula can be used:

$$n = \frac{1}{4}\left(\frac{z^*}{m}\right)^2$$

When p is smaller than 0.3 or greater than 0.7, using the p* = 0.5 will give a much larger n than necessary.

Additional notes about margins of error:

- Margins of error are always the same for $\hat{p}$ and 1 - $\hat{p}$.
- Margins of error are highest at p* = 0.5
- All formulas concerned with margin of error only apply to that one specific type of error. Other error types are not counted for here.

8.2: Two proportion comparisons

In two proportion comparisons two populations are involved. These you name Population 1 and Population 2, and their two population proportions of successes are p_1 and p_2, the sample sizes n_1 and n_2. Here are the notations needed in this part of the chapter:

Population	Population Proportion	Sample Size	Count of Successes	Sample Proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

We need the difference between the two populations to compare the proportions:

$$D = \hat{p}_1 - \hat{p}_2.$$

This distribution is approximately Normal when n is large enough. As it is easier to work with positive numbers, in general we should choose the higher proportion as the first population. The mean of D can be calculated by using the addition rule for means:

$$\mu_D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2.$$

The sum of the variances is

$$\sigma_D^2 = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2 = \frac{p1(1-p1)}{n1} + \frac{p2(1-p2)}{n2}.$$

So D is approximately Normal when both n's are large enough, with mean $\mu_D = p_1 - p_2$ and standard deviation:

$$\sigma_D = \sqrt{\frac{p1(1-p1)}{n1} + \frac{p2(1-p2)}{n2}}.$$

Confidence intervals for difference in proportions

The standard error of D is:

$$SE_D = \sqrt{\frac{\hat{p}1\,(1-\hat{p}1)}{n1} + \frac{\hat{p}2(1-\hat{p}2)}{n2}}.$$

The margin of error is therefore: m = z*SE_D, which gives an approximate C confidence interval for p_1 - p_2 being: D ± m. You can use this interval for 90%, 95% or 99% confidence in the cases that both the failures and successes are each 10 or more.

The *plus four confidence interval* can also be used for the difference in proportions. The plus four rule in this case assumes that the data has 2 more successes and 2 more failures, but they are now divided equally over the two samples. As a result, the *plus four estimates* are written as below. X_1 n_1

$$\tilde{p}_1 = \frac{X1+1}{n1+2} \text{ and } \tilde{p}_2 = \frac{X2+1}{n2+2}$$

The estimated difference becomes:

$$D = \tilde{p}_1 - \tilde{p}_2$$

and so the standard error is:

$$\sigma_{\tilde{D}} = \sqrt{\frac{p1(1-p1)}{n1+2} + \frac{p2(1-p2)}{n2+2}},$$

which can then be filled into the confidence interval formula.

Significance tests for difference in proportions

The null hypothesis is a significance test for the difference in proportions is: H_o: $p_1 = p_2$. To test this you need to compute the z statistic:

$$z = \frac{\hat{p}1-\hat{p}2}{SEDp}$$

where the pooled standard error is:

$$SE_{Dp} = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n1}+\frac{1}{n2})}$$

In which $\hat{p}$ is the number of successes in both samples divided by the sum of the sample sizes (See below). This is called the *pooled estimate of p*. p represents the common value of p_1 and p_2.

$$\hat{p} = \frac{X1+X2}{n1+n2}$$

The p-values can then be found according to:

- H_a: $p_1 > p_2$ is $P(Z \geq z)$
- H_a: $p_1 < p_2$ is $P(Z \leq z)$
- H_a: $p_1 \neq p_2$ is $2P(Z \geq |z|)$

This z test can be used when the number of failures and number of successes in each sample is at least 5.

Relative Risk

You can also summarize the comparison between two proportions as relative risk (RR):

$$RR = p_1/p_2$$

When both proportions are equal you will get a relative risk of 1.