0.1 Difference in proportions

We can also construct a confidence interval for the difference between two sample proportions, $\pi_1 - \pi_2$. The point estimate is the difference in sample proportions for tho both groups , $\hat{p}_1 - \hat{p}_2$.

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

- Both samples are simple random samples.
- The samples are independent.
- Each sample includes at least 10 successes and 10 failures.
- The samples comprises less than 10% of their respective populations.

Confidence Interval

Compute the standard Error

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\frac{\hat{p}_1 \times (1-\hat{p}_1)}{n_1}\right] + \left[\frac{\hat{p}_2 \times (1-\hat{p}_2)}{n_2}\right]}$$

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\frac{40 \times 60}{400}\right] + \left[\frac{30 \times 70}{300}\right]} = \sqrt{\left[\frac{2400}{400}\right] + \left[\frac{2100}{300}\right]}$$

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{6+7} = 3.6\%$$

Standard Error for Difference of Proportions

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\frac{\hat{p}_1 \times (1-\hat{p}_1)}{n_1}\right] + \left[\frac{\hat{p}_2 \times (1-\hat{p}_2)}{n_2}\right]}$$

- $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of groups 1 and 2 respectively.
- n_1 and n_2 are the sample sizes of groups 1 and 2 respectively.

N.B. This formula will be provided in the exam paper. Also, there is no accounting for small samples.

Confidence Interval for the Difference Between Two Proportions

- A confidence interval gives us some idea of the range of values which an unknown population parameter (such as the mean or variance) is likely to take based on a given set of sample data.
- Many occasions arise where we have to compare the proportions of two different populations.
- For example, a firm may want to compare the proportions of defective items produced by different machines; medical researchers may want to compare the proportions of men and women who suffer heart attacks etc.
- A confidence interval for the difference between two proportions would specify a range of values within which the difference between the two true population proportions may lie, for such examples.
- The procedure for obtaining such an interval is based on the sample proportions, p1 and p2, from their respective overall populations.

0.2 Testing the Difference Between Two Population Proportions

- When we wish to test the hypothesis that the proportions in two populations are not different, the two sample proportions are pooled as a basis for determining the standard error of the difference between proportions.
- Note that this differs from the procedure used previously on statistical estimation, in which the assumption of no difference was not made.
- Further, the present procedure is conceptually similar to that presented in Section 11.1, in which the two sample variances are pooled as the basis for computing the standard error of the difference between means.

0.2.1 Hypothesis Tests of Differences between Proportions

This procedure is used to compare two proportions from two different populations. For two tailed tests, the null hypothesis states that the population proportion $\pi_1 - \pi_2$ has a specified value, with the alternative stating that $\pi_1 - \pi_2$ does not have this value.

Specifying the Null and Alternative Hypothesis

$$H_0 : \pi_1 = \pi_2 \qquad\qquad H_0 : \pi_1 - \pi_2 = 0$$

$$H_1 : \pi_1 \neq \pi_2 \qquad\qquad H_1 : \pi_1 - \pi_2 \neq 0$$

- Expected Value of differences under null hypothesis: $\pi_1 - \pi_2 = 0$

- Significance level = 0.01

$$SE(p_1 - p_2) = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Calculate Pooled Proportion Estimate

$$\bar{p} = \frac{29 + 62}{1110 + 1553}$$

- Test Statistic

$$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{SE(\pi_1 - \pi_2)}$$

The formula for the estimated standard error is:

$$S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\bar{p}(100-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\bar{p}$ is a aggregate proportion (proportion of successes from overall sample, regardless of which group they are in).

0.2.2 Pooled Estimate for Population Proportion

The pooled estimate of the population proportion, based on the proportions obtained in two independent samples.

0.3 Worked Example

Proportions : Remarks

- Small Sample sizes will not be considered for the case of sample proportions or Difference of proportions. Small samples will considered for "sample mean" cases only.
- The computed p-values is compared to the pre-specified significance level of 5%. Since the p-value (< 0.0001) is less than the significance level of 0.05, the effect is statistically significant.
- Since the effect is significant, the null hypothesis is rejected. The conclusion is that the probability of graduating from high school is greater for students who have participated in the early childhood intervention program than for students who have not.
- The results could be described in a report as: The proportion of students from the early-intervention group who graduated from high school was 0.86 whereas the proportion from the control group who graduated was only 0.52. The difference in proportions is significant, with $p < 0.0001$.

0.4 Summary of Inference Procedures

Point Estimates:

$$\hat{p}_1 = \frac{x_1}{n_1}$$

$$\hat{p}_2 = \frac{x_2}{n_2}$$

Hypotheses:

$$\begin{aligned} H_0 : &\quad \pi_1 \leq \pi_2 \\ H_1 : &\quad \pi_1 > \pi_2 \end{aligned}$$

- The population proportion for group 1 does not exceed the corresponding value for group 2.
- The population proportion for group 1 does exceed (is greater than) the corresponding value for group 2.

$$\begin{aligned} H_0 : &\quad \pi_1 - \pi_2 \leq 0 \\ H_1 : &\quad \pi_1 - \pi_2 > 0 \end{aligned}$$

Critical Vale

- $\alpha = 0.05$
- One-tailed Procedure (refer back to H_1) k=1
- Large sample ($x_1 + x_2 > 30$)

Descision is $|TS| > CV$?
Comclusion: We can reject the null hypothesis, We can reasonably conclude that....

0.5 Example

In the past, 18% of shoppers have bought a particular brand of breakfast cereal. After an advertising campaign, a random sample of 220 shoppers is taken and 55 of the sample have bought this brand of cereal.

Write down the null and the alternative hypothesis for this problem, and state whether it is a one tailed or two tailed test

The conventional treatment for a disease has been shown to be effective in 80% of all cases. A new drug is being promoted by a pharmaceutical company; the Department of Health wishes to test whether the new treatment is more effective than the conventional treatment.

Write down the null and the alternative hypothesis for this problem, and state whether it is a one tailed or two tailed test

0.6 Worked Example : Difference of Two Proportions

Two time-sharing systems are compared according to their response time to an editing command.

- The mean response time of 100 requests submitted to system 1 was measured to be 600 milliseconds with a known standard deviation of 20 milliseconds.
- The mean response time of 100 requests on system 2 was 592 milliseconds with a known standard deviation of 23 milliseconds.

Using a significance level of 5%, test the hypothesis that system 2 provides a faster response time than system 1.

Clearly state your null and alternative hypotheses and your conclusion.