Chapter 8

Estimation and Confidence Intervals: Additional Topics

- This chapter simply follows the methods in Chapter 7 for forming confidence intervals

- The text is a bit disorganized here so hopefully we can simplify

Figure 8.1:
8.1 Proportions of Successes

- The same ideas can be used to find point and interval estimates of proportions.
- Suppose, for example, that we have \( n \) observations, with \( X \) successes, so that

\[
\hat{p} = f = X/n
\]  

(8.1)

- Then the number of successes \( X \) has

\[
E(X) = np; \quad Var(X) = np(1-p)
\]  

while the proportion of successes \( p \) has

\[
E(\hat{p}) = p; \quad Var[p] = \frac{p(1-p)}{n}.
\]  

(8.3)

- We can estimate the variance of the proportion by

\[
s^2_p = \frac{p(1-p)}{n}.
\]  

(8.4)

- Then we simply use the normal approximation to the binomial:

\[
Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}.
\]  

(8.5)

- The 100(1 - \( \alpha \)) % confidence interval for \( \hat{p} \) when \( n \) is large is:

\[
P\{\hat{p} - Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}\} = (1 - \alpha)
\]  

(8.6)

- Notice that this expression involves the unknown \( p \), so we are going to replace \( p \) by its estimator \( f \) to give:

\[
P\{p - Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < p + Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\} = (1 - \alpha)
\]  

(8.7)

Keep in mind that this approximation holds only for large samples.
8.1.1 Example of Confidence Interval for Population Proportion \( p \)

- A sample of 100 voters are asked about a municipal bond issue; 64 favour it.
- What is a 95 percent confidence interval for the proportion who favour it in the entire municipal population?

- Here \( n = 100, X = 64 \), and \( \hat{p} = .64 \).
- So the 95 percent confidence interval is \((.55, .73)\). (Can you get these numbers?)
- Interpret this interval.

8.2 Confidence Interval for the Difference Between Means

- It looks like there are a whole bunch of cases to remember here, but actually we are just reapplying the same principles over and over
- For instance, assume that we have two independent samples of size \( n_X \) and \( n_Y \).

\[
X_i = X_1 \ldots X_{n_X} \quad Y_j = Y_1 \ldots Y_{n_Y}
\]

- Let the means and variances be \( \mu_X, \mu_Y \) and \( \sigma_X^2, \sigma_Y^2 \).

8.2.1 Known Variances

- Let us assume that the samples are both normal, it should be obvious after some thinking that the \((1 - \alpha)\)% for known variances is

\[
\bar{X} - \bar{Y} \pm Z_{\frac{\alpha}{2}} \sigma_{\bar{X} - \bar{Y}}
\]

\[
\sigma_{\bar{X} - \bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}
\]

since the variables are independent (the variance of the sum is the sum of the variances)

\[
V[X_i - Y_j] = V[X_i] + V[Y_j] \text{ for any } i, j
\]
8.2.2 Unknown Variances with Variances Presumed Different

- If the variances are unknown (and assumed different) which is usually the case we want to estimate \( \sigma_{\bar{X} - \bar{Y}} \) by \( s_{\bar{X} - \bar{Y}} \):

\[
X - Y \pm t_v s_{\bar{X} - \bar{Y}}
\]

\[
s_{\bar{X} - \bar{Y}} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}
\]

- In my own work I have used

\[
v = n_X + n_Y - 2
\]

whereas NCT uses a terribly complicate expression:

\[
v = \frac{\left[ \frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \right]^2}{\frac{s_X^2}{(n_X-1)} + \frac{s_Y^2}{(n_Y-1)}}
\]

- Of course as \( n_X \) and \( n_Y \) gets larger \( t \Rightarrow Z \) (standard normal) and it makes no difference which we use for \( v \)

8.2.3 Unknown Variances that are Assumed to be Equal: \( \sigma_X^2 = \sigma_Y^2 = \sigma^2 \)

- Notice then that

\[
\sigma_{\bar{X} - \bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}
\]

we can estimate \( \sigma \) by using a **pooled estimator** \( s_p \)

\[
s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}
\]

and the estimate for

\[
\sigma_{\bar{X} - \bar{Y}} = \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}
\]

is

\[
s_{\bar{X} - \bar{Y}} = s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}
\]

which can be substituted in (8.9) with \( v = n_X + n_Y - 2 \)
8.3 Confidence Intervals for the Differences in Population Proportions

- This should be straightforward as another example:
- Recall from Chapter 7, consider two independent populations.

1. Population 1: \( X_1 = \) number of successes, \( n_1 = \) number in sample 1, so \( \hat{p}_1 = \frac{x_1}{n_1} \).
   Recall
   \[
   E[p_1] = p_1 \\
   V[\hat{p}_1] = \frac{p_1(1-p_1)}{n_1} \\
   \hat{p}_1 \sim N \left( \left( p_1, \frac{p_1(1-p_1)}{n_1} \right) \right) \text{ asymptotically}
   \]

2. Population 2: \( X_2 = \) number of successes, \( n_2 = \) number in sample 2, so \( \hat{p}_2 = \frac{x_2}{n_2} \).
   \[
   E[\hat{p}_2] = p_2 \\
   V[\hat{p}_2] = \frac{p_2(1-p_2)}{n_2} \\
   \hat{p}_2 \sim N \left( \left( p_2, \frac{p_2(1-p_2)}{n_2} \right) \right) \text{ asymptotically}
   \]

3. Form Difference of Sample Proportion: \( \hat{p}_1 - \hat{p}_2 \)

   If \( n_1 \) and \( n_2 \) are large, i.e. \( n_1p_1(1-p_1) \geq 9 \) and \( n_2p_2(1-p_2) \geq 9 \) then:
   \[
   \hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}) \text{ approximately.}
   \]

4. So the confidence interval is
   \[
   \hat{p}_1 - \hat{p}_2 \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}
   \]
8.4 Sample Size Determination

8.4.1 Sample Size Determination for Population Mean from known variance

- Let $B_{(1-\alpha)}$ sampling error bound (called margin of error in text) in any confidence interval for $\hat{\theta}$

\[ \hat{\theta} \pm B_{(1-\alpha)} \]

- $B_{(1-\alpha)}$ is a function of the sample size, so that we can for a given size for $B$ calculate a single sample size

- For instance, for the confidence interval of the population mean from known variance we have

\[ B_{(1-\alpha)} = \frac{Z_{\alpha/2} \sigma}{\sqrt{n}} \]

which we could rearrange to solve for $n$

\[ \sqrt{n} = \frac{Z_{\alpha/2} \sigma}{B_{(1-\alpha)}} \Rightarrow n = \frac{Z_{\alpha/2}^2 \sigma^2}{B_{(1-\alpha)}^2} \]

- Example

- Suppose we have metal rods in an industrial process with a $\sigma = 1.8$mm and want a 99% confidence interval that extends no further than 0.50 on either side of sample mean ($B_{(0.99)} = .50$). What is the required sample size?

- Answer ($Z_{\alpha/2} = Z_{.005} = 2.576$):

\[ \frac{Z_{\alpha/2}^2 \sigma^2}{B_{(1-\alpha)}^2} = \frac{(2.576)^2 \times (1.8)^2}{(0.5)^2} \approx 86 \]

8.4.2 Sample Size Determination for Population Proportion

- In this case

\[ B_{(1-\alpha)} = Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \]

- Notice that since $B_{(1-\alpha)}$ involves $p$ the sample proportion (which obviously depends on $n$) we cannot simply calculate this directly

- Instead we can set an upper limit for $n$ by assuming a sample proportion of 0.5 (this gives the largest value of $n$ and hence is the most conservative) so that

\[ B_{(1-\alpha)} = Z_{\alpha/2} \sqrt{\frac{0.5(0.5)}{n}} = Z_{\alpha/2} \sqrt{\frac{0.25}{n}} = 0.5Z_{\alpha/2} \sqrt{\frac{1}{n}} \]
which leads to
\[ \sqrt{n} = \frac{0.5Z^2}{B_{(1-\alpha)}} \Rightarrow n = \frac{.25Z^2}{B^2_{(1-\alpha)}} \]

- Suppose a polling firm wants to have no more than 3% band around the population proportion confidence interval at the 95% level, how large a polling sample should be obtained

- Answer \((Z^2 = Z_{.025} = 1.96)\)

\[ n = \frac{.25Z^2}{B^2_{(1-\alpha)}} = \frac{.25 \times 1.96^2}{.03^2} \approx 1067 \text{ people} \]