3.2 In this question, you are asked to consider a Bernoulli random variable $Y$, with a success probability $Pr(Y = 1) = p$. You are told that you have $n$ draws from this distribution and that $\hat{p}$ is the fraction of successes (i.e., the percentage of 1’s).

a. The first part of the question asks you to show that $\hat{p} = \bar{Y}$. From our definition on $\bar{Y}$, we know that:

$$\bar{Y} = \frac{1}{n} \left[ \sum_{i=1}^{n} Y_i \right]$$

$$= \frac{1}{n} \text{[number of successes]}$$

$$= \frac{\text{number of successes}}{n}$$

$$= \hat{p}$$

b. The second part asks you to show that $\hat{p}$ is an unbiased estimator of $p$. We know that, since the $Y_i$ are Bernoulli random variables that

$$E(Y_i) = 1 \cdot Pr(Y_i = 1) + 0 \cdot Pr(Y_i = 0) = Pr(Y_i = 1) = p.$$  (1)

Using this information, and our result from part (a), we have that

$$E(\hat{p}) = E(\bar{Y})$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(Y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} p$$

$$= \frac{1}{n} [n \cdot p]$$

$$= p$$

c. Finally, you are asked to show that $\text{var}(\hat{p}) = p(1 - p)/n$. We know from chapter 2 (equation 2.7), that

$$\sigma_Y^2 = \text{Var}(Y_i) = p(1 - p).$$  (2)

Using this information, and the fact that the $Y_i$’s are independent, we then know from equation (2.45) in the text, that:

$$\text{Var}(\hat{p}) = \text{Var}(\bar{Y})$$

$$= \frac{\sigma_Y^2}{n}$$

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

3.3 You are told that, in a survey of likely voters, 215 responded that they would vote for the incumbent and 185 would likely vote for the challenger. Let $p$ denote the fraction of all likely voters who preferred the incumbent at the time of the survey and $\hat{p}$ denote the fraction of the survey respondents who preferred the incumbent.

a. The first part of the question asks you to estimate $p$. We know from question 3.2 that an unbiased estimator of $p$ is $\bar{Y}$, so we could use as our estimator:

$$\hat{p} = \bar{Y}$$

$$= \frac{215}{400}$$

$$= 0.5375$$
b. From part (c) of question 3.2, we know that:
\[ V\text{ar}(\hat{p}) = \frac{p(1-p)}{n}. \]
This suggests the estimator for the variance of
\[ \hat{V}\text{ar}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.5375(1-0.5375)}{400} = 0.00062148. \]
The corresponding standard error estimate would be
\[ SE(\hat{p}) = \sqrt{\hat{V}\text{ar}(\hat{p})} = 0.02493. \]

3.10 You are told that a new standardized test is given to 100 randomly selected third grade students in New Jersey. The sample average score is \( \bar{Y} \) on the test is 58 and the sample standard deviation is \( s_y = 8 \).

a. The first part of the question asks you to construct a 95% confidence interval for the mean score of all New Jersey third graders. Using the information above, we know that:
\[ SE(\bar{Y}) = \frac{s_y}{\sqrt{n}} = \frac{8}{\sqrt{100}} = 0.8 \]
The corresponding 95% confidence interval is then given by:
\[ \{ \bar{Y} \pm 1.96SE(\bar{Y}) \} = \{ 58 \pm 1.96(0.8) \} = \{ 56.432, 59.568 \} \]

b. The same test is given to 200 third graders in Iowa, with a sample average of 62 points and a standard deviation of 11 points. You are asked to construct a 90% confidence interval for the difference in the mean test scores between Iowa and New Jersey. Using equation (3.19) from the text, we have that:
\[ SE(\bar{Y}_{Iowa} - \bar{Y}_{NJ}) = \sqrt{\frac{s_{\text{Iowa}}^2}{n_{\text{Iowa}}} + \frac{s_{\text{NJ}}^2}{n_{\text{NJ}}}} \]
\[ = \sqrt{\frac{11^2}{200} + \frac{8^2}{100}} \]
\[ = \sqrt{0.605 + 0.64} \]
\[ = 1.1158 \]
The corresponding confidence interval is then given by:
\[ \{ \bar{Y}_{\text{Iowa}} - \bar{Y}_{NJ} \pm 1.64SE(\bar{Y}_{\text{Iowa}} - \bar{Y}_{NJ}) \} = \{ 62 - 58 \pm 1.64(1.1158) \} \]
\[ = \{ 4 \pm 1.83 \} \]
\[ = \{ 2.17, 5.83 \} \]
e. Finally, you are asked to assess the degree of confidence you have in the proposition that the population means for Iowa and New Jersey are different. More specifically, you are asked what the standard error of the differences in the two sample means (which we have already calculated as 1.1158) and the \( p \)-value of the test of no differences in means versus some difference. That latter is given by:

\[
\begin{align*}
p - \text{value} & = 2\Phi \left( -\frac{\bar{Y}_{\text{Iowa}} - \bar{Y}_{\text{NJ}} - 0}{SE(\bar{Y}_{\text{Iowa}} - \bar{Y}_{\text{NJ}})} \right) \\
& = 2\Phi \left( -\frac{4}{1.1158} \right) \\
& = 2\Phi (-3.585) \\
& \approx 2(0.00017) \\
& = 0.00034.
\end{align*}
\]

Clearly, we would reject the null hypothesis of no differences between the two groups of students.

3.12 In this question, you are asked to consider evidence on the issue of gender discrimination in a firm. You are given data on the salaries of 100 men and 64 women with similar job descriptions.

a. The first question asks what the data suggest in terms of the wage differences between men and women. The corresponding null and alternative hypotheses would be:

\[
\begin{align*}
H_0 & : E(Y_{\text{men}}) - E(Y_{\text{women}}) = 0 \\
H_1 & : E(Y_{\text{men}}) - E(Y_{\text{women}}) \neq 0
\end{align*}
\]

The data provided in the table indicate that:

\[
\begin{align*}
\bar{Y}_{\text{act men}} &= 3100 \\
\bar{Y}_{\text{act women}} &= 2900 \\
s_{\text{men}} &= 200 \\
s_{\text{women}} &= 320
\end{align*}
\]

Using this information:

\[
\begin{align*}
SE(\bar{Y}_{\text{men}} - \bar{Y}_{\text{women}}) & = \sqrt{\frac{s_{\text{men}}^2}{n_{\text{men}}} + \frac{s_{\text{women}}^2}{n_{\text{women}}}} \\
& = \sqrt{\frac{200^2}{100} + \frac{320^2}{64}} \\
& = \sqrt{400 + 1600} \\
& = 44.72
\end{align*}
\]

This in turn can be used to calculate the appropriate t-statistic as:

\[
\begin{align*}
t^{\text{act}} & = \frac{\bar{Y}_{\text{men}} - \bar{Y}_{\text{women}} - 0}{SE(\bar{Y}_{\text{men}} - \bar{Y}_{\text{women}})} \\
& = \frac{200}{44.72} \\
& = 4.47
\end{align*}
\]

Finally, we can compute the corresponding \( p \)-value as:

\[
\begin{align*}
p - \text{value} & = 2\Phi (-|t^{\text{act}}|) \\
& = 2\Phi (-4.47) \\
& \approx 2(0) \\
& = 0.
\end{align*}
\]
b. In the second part of the question, you are asked to assess whether the data suggest that the firm is guilty of gender discrimination in its compensation policies. With the extremely small p-value, the null hypothesis can be rejected with a high degree of confidence. There is overwhelming statistical evidence that mean earnings for men are different from the mean earnings for women. However, by itself, this does not imply gender discrimination by the firm. Gender discrimination means that two workers, identical in every way but gender, are paid different wages. The data description suggests that some care has been taken to make sure that workers with similar jobs are being compared. But, it is also important to control for characteristics of the workers that may affect their productivity (education, years of experience, etc.). If these characteristics are systematically different between men and women, then they may be responsible for the difference in mean wages. (If this is true, it raises an interesting and important question of why women tend to have less education or less experience than men, but that is a question about something other than gender discrimination by this firm.) Since these characteristics are not controlled for in the statistical analysis, it is premature to reach a conclusion about gender discrimination.

3.15 This question focuses on the outcomes of two polls around the time of the 2004 elections. The first poll was in September of 2004, with 405 of 755 likely voters preferring Bush, whereas the second poll found that 378 our of 756 favoring Bush.

a. The first question asks that you construct of 95% confidence bound for the fraction of likely voters favoring Bush in September 2004. [As an aside, what you are doing here is actually constructing a confidence interval on the estimator of this fraction for a similar sample of 755 likely voters.] Using the data, we know that:

$$\hat{Y}_{Sept} = \frac{405}{755} = 0.536$$

As estimate of variance of this estimator would be

$$Var(\hat{Y}_{Sept}) = \frac{\hat{Y}_{Sept}(1 - \hat{Y}_{Sept})}{n} = \frac{0.536(1 - 0.536)}{755} = 0.00032941.$$

Using this,

$$SE(\hat{Y}_{Sept}) = \sqrt{Var(\hat{Y}_{Sept})} = 0.01815.$$ (7)

The corresponding confidence interval would be given by:

$$\{\hat{Y}_{Sept} \pm 1.96SE(\hat{Y}_{Sept})\} = \{0.536 \pm 1.96(0.01815)\} = \{0.500, 0.571\}$$ (8)

b. The second question asks you to repeat these steps for the October 2004 poll. In this case

$$\hat{Y}_{Oct} = \frac{378}{756} = 0.5$$

As estimate of variance of this estimator would be

$$Var(\hat{Y}_{Oct}) = \frac{\hat{Y}_{Oct}(1 - \hat{Y}_{Oct})}{n} = \frac{0.5(1 - 0.5)}{756} = 0.00033069.$$

Using this,

$$SE(\hat{Y}_{Oct}) = \sqrt{Var(\hat{Y}_{Oct})} = 0.01818.$$ (10)

The corresponding confidence interval would be given by:

$$\{\hat{Y}_{Oct} \pm 1.96SE(\hat{Y}_{Oct})\} = \{0.5 \pm 1.96(0.01818)\} = \{0.464, 0.536\}$$ (11)
c. The last question asks whether or not there was a significant change in voters’ opinions across the two surveys. This is just a test of the difference between means between the two samples. That is, we have the null hypothesis

\[ H_0 : E(Y_{Sept}) - E(Y_{Oct}) = 0 \]  

versus the alternative hypothesis:

\[ H_1 : E(Y_{Sept}) - E(Y_{Oct}) \neq 0 \]  

To construct the corresponding t-statistic, we will need sample variances for both polls. That is, we need:

\[ s^2_{Sept} = \bar{Y}_{Sept}(1 - \bar{Y}_{Sept}) = 0.536(1 - 0.536) = 0.2487 \]  

and

\[ s^2_{Oct} = \bar{Y}_{Oct}(1 - \bar{Y}_{Oct}) = 0.5(1 - 0.5) = 0.25 \]  

This can then be used to form

\[ SE(\bar{Y}_{Sept} - \bar{Y}_{Oct}) = \sqrt{s^2_{Sept}/n_{Sept} + s^2_{Oct}/n_{Oct}} \]

\[ = \sqrt{0.2487/755 + 0.25/756} \]

\[ = \sqrt{0.00032941 + 0.00033069} \]

\[ = 0.0257 \]

This in turn can be used to calculate the appropriate t-statistic as:

\[ t^{act} = \frac{\bar{Y}_{Sept} - \bar{Y}_{Oct} - 0}{SE(\bar{Y}_{Sept} - \bar{Y}_{Oct})} \]

\[ = \frac{0.036}{0.027} \]

\[ = 1.33 \]

Finally, we can compute the corresponding p-value as:

\[ p - value = 2\Phi(-|t^{act}|) \]

\[ = 2\Phi(-1.33) \]

\[ \approx 2(0.0918) \]

\[ = 0.1836. \]

With this high a p-value, we should be reluctant to reject the null hypothesis that there has been no change in voter’s opinions.