3.2. Each random draw Y_i from the Bernoulli distribution takes a value of either zero or one with probability $\Pr(Y_i = 1) = p$ and $\Pr(Y_i = 0) = 1 - p$. The random variable Y_i has mean

$$E(Y_i) = 0 \times \Pr(Y = 0) + 1 \times \Pr(Y = 1) = p,$$

and variance

$$\begin{aligned} \text{var}(Y_i) &= E[(Y_i - \mu_Y)^2] \\ &= (0 - p)^2 \times \Pr(Y_i = 0) + (1 - p)^2 \times \Pr(Y_i = 1) \\ &= p^2(1 - p) + (1 - p)^2 p = p(1 - p). \end{aligned}$$

(a) The fraction of successes is

$$\hat{p} = \frac{\#(\text{success})}{n} = \frac{\#(Y_i = 1)}{n} = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}.$$

(b)

$$E(\hat{p}) = E\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{1}{n}\sum_{i=1}^n p = p.$$

(c)

$$\text{var}(\hat{p}) = \text{var}\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n \text{var}(Y_i) = \frac{1}{n^2}\sum_{i=1}^n p(1 - p) = \frac{p(1 - p)}{n}.$$

The second equality uses the fact that $Y_1, \ldots, Y_n$ are i.i.d. draws and $\text{cov}(Y_i, Y_j) = 0$, for $i \neq j$.

3.3. Denote each voter's preference by Y, with $Y = 1$ if the voter prefers the democratic party and $Y = 0$ if the voter prefers the republican party. Y is a Bernoulli random variable with probability $\Pr(Y = 1) = p$ and $\Pr(Y = 0) = 1 - p$. From the solution to Exercise 3.2, Y has mean p and variance $p(1 - p)$.

(a) $\hat{p} = \frac{270}{500} = 0.54$

(b) The estimated variance of $\hat{p}$ is $\widehat{var(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.54(1-0.54)}{500} = 0.0004968$. The standard error is $\text{SE}(\hat{p}) = \sqrt{0.0004968} = 0.022289$

(c) The computed t-statistic is

$$t = \frac{0.54 - 0.5}{0.022289} = 1.7946$$

Because of the large sample size (n=500), we can use Equation (3.14) in the text to compute the p-value for the test $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$:

$$p\text{-value} = 2\Phi(-|t|) = 2\Phi(-1.7946) = 2 \times 0.0363 = 0.0726$$

(d) Using Equation (3.17) in the text, the p-value for the test $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$ is

$$p\text{-value} = 1 - \Phi(t) = 1 - \Phi(1.7946) = 1 - 0.9637 = 0.0363$$

(e) Part (c) is a two-sided test and the p-value is the area in the tails of the standard normal distribution outside $\pm$ (calculated t-statistic). Part (d) is a one-sided test and the p-value is the area under the standard normal distribution to the right of the calculated t-statistic.

(f) For the test H_0: $p = 0.5$ versus H_1: $p > 0.5$, we can reject the null hypothesis at the 5% significance level. The p-value 0.0363 is smaller than 0.05. Equivalently the calculated t-statistic 1.7946 is larger than the critical value 1.645 for a one-sided test with a 5% significance level. The test suggests that the survey contained statistically significant evidence that the democratic candidate was ahead of the republican candidate at the time of the survey.

3.4. Using Key Concept 3.7 in the text

(a) 95% confidence interval for *p* is

$$\hat{p} \pm 1.96SE(\hat{p}) = 0.54 \pm 1.96 \times 0.022289 = (0.4963, 0.5837)$$

(b) 99% confidence interval for *p* is

$$\hat{p} \pm 2.57SE(\hat{p}) = 0.54 \pm 2.57 \times 0.022289 = (0.4827, 0.5973)$$

(c) Mechanically, the interval in (b) is wider because of a larger critical value (2.57 versus 1.96). Substantively, a 99% confidence interval is wider than a 95% confidence because a 99% confidence interval must contain the true value of *p* in 99% of all possible samples, while a 95% confidence interval must contain the true value of *p* in only 95% of all possible samples.

(d) Since 0.50 lies inside the 95% confidence interval for *p*, we cannot reject the null hypothesis at a 5% significance level.

3.6. (a) No. Because the *p*-value is less than 0.1 (=10%), $\mu = 10$ is rejected at the 10% level and is therefore not contained in the 90% confidence interval.

(b) No. This would require calculation of the *t*-statistic for $\mu = 8$, which requires $\overline{Y}$ and SE $(\overline{Y})$. Only the *p*-value for test that $\mu = 10$ is given in the problem.

3.11. Assume that n is an even number. Then $\tilde{Y}$ is constructed by applying a weight of $\frac{1}{2}$ to the $\frac{n}{2}$ "odd" observations and a weight of $\frac{3}{2}$ to the remaining $\frac{n}{2}$ observations.

$$\begin{aligned} E(\tilde{Y}) &= \frac{1}{n}\left(\frac{1}{2}E(Y_1) + \frac{3}{2}E(Y_2) + \cdots \frac{1}{2}E(Y_{n-1}) + \frac{3}{2}E(Y_n)\right) \\ &= \frac{1}{n}\left(\frac{1}{2}\cdot\frac{n}{2}\cdot\mu_Y + \frac{3}{2}\cdot\frac{n}{2}\cdot\mu_Y\right) = \mu_Y \\ \text{var}(\tilde{Y}) &= \frac{1}{n^2}\left(\frac{1}{4}\text{var}(Y_1) + \frac{9}{4}\text{var}(Y_2) + \cdots \frac{1}{4}\text{var}(Y_{n-1}) + \frac{9}{4}\text{var}(Y_n)\right) \\ &= \frac{1}{n^2}\left(\frac{1}{4}\cdot\frac{n}{2}\cdot\sigma_Y^2 + \frac{9}{4}\cdot\frac{n}{2}\cdot\sigma_Y^2\right) = 1.25\frac{\sigma_Y^2}{n}. \end{aligned}$$

3.12. Sample size for men $n_1 = 120$ sample average $\overline{Y}_1 = 8200$ sample standard deviation $s_1 = 450$. Sample size for women $n_2 = 150$ sample average $\overline{Y}_2 = 7900$ sample standard deviation $s_2 = 520$ The standard error of $\overline{Y}_1 - \overline{Y}_2$ is:

$$SE(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{450^2}{120} + \frac{520^2}{150}} = 59.0776$$

(a) The hypothesis test for the difference in mean monthly salaries is

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

The *t*-statistic for testing the null hypothesis is

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)} = \frac{8200 - 7900}{59.0776} = 5.0781$$

Use Equation (3.14) in the text to get the *p*-value:

$$p\text{-value} = 2\Phi(-|t|) = 2\Phi(-5.0781) \approx 0$$

The extremely low level of *p*-value implies that the difference in the monthly salaries for men and women is statistically significant. We can reject the null hypothesis with a high degree of confidence.

(b) From part (a), there is overwhelming statistical evidence that mean earnings for men *differ* from mean earnings for women, and a related calculation shows overwhelming evidence that mean earning for men are *greater* that mean earnings for women. However, by itself, this does not imply sex discrimination by the firm. Sex 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 sex discrimination by this firm.) Since these characteristics are not controlled for in the statistical analysis, it is premature to reach a conclusion about sex discrimination.