Chapter 3
Review of Statistics

■ Solutions to Exercises

1. The central limit theorem suggests that when the sample size (n) is large, the distribution of the sample average ($\bar{Y}$) is approximately $N\left(\mu_Y, \sigma_{\bar{Y}}^2\right)$ with $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$. Given a population $\mu_Y = 100$, $\sigma_Y^2 = 43.0$, we have

   (a) $n = 100$, $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} = \frac{43}{100} = 0.43$, and

$$\Pr(\bar{Y} < 101) = \Pr\left(\frac{\bar{Y} - 100}{\sqrt{0.43}} < \frac{101 - 100}{\sqrt{0.43}}\right) \approx \Phi(1.525) = 0.9364.$$

   (b) $n = 64$, $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{64} = \frac{43}{64} = 0.6719$, and

$$\Pr(101 < \bar{Y} < 103) = \Pr\left(\frac{101 - 100}{\sqrt{0.6719}} < \frac{\bar{Y} - 100}{\sqrt{0.6719}} < \frac{103 - 100}{\sqrt{0.6719}}\right)$$
$$\approx \Phi(3.6599) - \Phi(1.2200) = 0.9999 - 0.8888 = 0.1111.$$

   (c) $n = 165$, $\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} = \frac{43}{165} = 0.2606$, and

$$\Pr(\bar{Y} > 98) = 1 - \Pr(\bar{Y} \le 98) = 1 - \Pr\left(\frac{\bar{Y} - 100}{\sqrt{0.2606}} \le \frac{98 - 100}{\sqrt{0.2606}}\right)$$
$$\approx 1 - \Phi(-3.9178) = \Phi(3.9178) = 1.0000 \text{ (rounded to four decimal places).}$$

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. Denote each voter's preference by Y. $Y = 1$ if the voter prefers the incumbent and $Y = 0$ if the voter prefers the challenger. 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) $\hat{p} = \frac{215}{400} = 0.5375.$

(b) $\widehat{\text{var}(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.5375 \times (1 - 0.5375)}{400} = 6.2148 \times 10^{-4}.$ The standard error is $\text{SE}(\hat{p}) = (\text{var}(\hat{p}))^{\frac{1}{2}} = 0.0249.$

(c) The computed t-statistic is

$$t^{act} = \frac{\hat{p} - \mu_{p,0}}{\text{SE}(\hat{p})} = \frac{0.5375 - 0.5}{0.0249} = 1.506.$$

Because of the large sample size $(n = 400),$ we can use Equation (3.14) in the text to get the p-value for the test $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$:

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-1.506) = 2 \times 0.066 = 0.132.$$

(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

$$p\text{-value} = 1 - \Phi(t^{act}) = 1 - \Phi(1.506) = 1 - 0.934 = 0.066.$$

(e) Part (c) is a two-sided test and the p-value is the area in the tails of the standard normal distribution outside ± (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 $H_0 : p = 0.5$ vs. $H_1 : p > 0.5,$ we cannot reject the null hypothesis at the 5% significance level. The p-value 0.066 is larger than 0.05. Equivalently the calculated t-statistic 1.506 is less than the critical value 1.645 for a one-sided test with a 5% significance level. The test suggests that the survey did not contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey.

4. Using Key Concept 3.7 in the text
   (a) 95% confidence interval for p is

$$\hat{p} \pm 1.96 SE(\hat{p}) = 0.5375 \pm 1.96 \times 0.0249 = (0.4887, 0.5863).$$

   (b) 99% confidence interval for p is

$$\hat{p} \pm 2.57 SE(\hat{p}) = 0.5375 \pm 2.57 \times 0.0249 = (0.4735, 0.6015).$$

   (c) The interval in (b) is wider because of a larger critical value due to a lower significance level.
   (d) Since 0.50 lies inside the 95% confidence interval for p, we cannot reject the null hypothesis at a 5% significance level.

5. (a) (i) The size is given by $\Pr(|\hat{p} - 0.5| > .02)$, where the probability is computed assuming that $p = 0.5$.

$$\begin{aligned}\Pr(|\hat{p} - 0.5| > .02) &= 1 - \Pr(-0.02 \le \hat{p} - 0.5 \le .02) \\ &= 1 - \Pr\left(\frac{-0.02}{\sqrt{.5 \times .5/1055}} \le \frac{\hat{p} - 0.5}{\sqrt{.5 \times .5/1055}} \le \frac{0.02}{\sqrt{.5 \times .5/1055}}\right) \\ &= 1 - \Pr\left(-1.30 \le \frac{\hat{p} - 0.5}{\sqrt{.5 \times .5/1055}} \le 1.30\right) \\ &= 0.19\end{aligned}$$

where the final equality using the central limit theorem approximation

(ii) The power is given by $\Pr(|\hat{p} - 0.5| > .02)$, where the probability is computed assuming that $p = 0.53$.

$$\begin{aligned}\Pr(|\hat{p} - 0.5| > .02) &= 1 - \Pr(-0.02 \le \hat{p} - 0.5 \le .02) \\ &= 1 - \Pr\left(\frac{-0.02}{\sqrt{.53 \times .47/1055}} \le \frac{\hat{p} - 0.5}{\sqrt{.53 \times .47/1055}} \le \frac{0.02}{\sqrt{.53 \times .47/1055}}\right) \\ &= 1 - \Pr\left(\frac{-0.05}{\sqrt{.53 \times .47/1055}} \le \frac{\hat{p} - 0.53}{\sqrt{.53 \times .47/1055}} \le \frac{-0.01}{\sqrt{.53 \times .47/1055}}\right) \\ &= 1 - \Pr\left(-3.25 \le \frac{\hat{p} - 0.53}{\sqrt{.53 \times .47/1055}} \le -0.65\right) \\ &= 0.74\end{aligned}$$

where the final equality using the central limit theorem approximation.

(b) (i) $t = \frac{0.54 - 0.5}{\sqrt{0.54 \times 0.46/1055}} = 2.61, \Pr(|t| > 2.61) = .01,$ so that the null is rejected at the 5% level.

(ii) $\Pr(t > 2.61) = .004,$ so that the null is rejected at the 5% level.

(iii) $0.54 \pm 1.96\sqrt{0.54 \times 0.46/1055} = 0.54 \pm 0.03,$ or 0.51 to 0.57.

(iv) $0.54 \pm 2.58\sqrt{0.54 \times 0.46/1055} = 0.54 \pm 0.04,$ or 0.50 to 0.58.

(v) $0.54 \pm 0.67\sqrt{0.54 \times 0.46/1055} = 0.54 \pm 0.01,$ or 0.53 to 0.55.

(c) (i) The probability is 0.95 is any single survey, there are 20 independent surveys, so the probability if $0.95^{20} = 0.36$

(ii) 95% of the 20 confidence intervals or 19.

(d) The relevant equation is $1.96\times \text{SE}(\hat{p}) < .01$ or $1.96\times\sqrt{p(1-p)/n} < .01.$ Thus n must be chosen so that $n > \frac{1.96^2 p(1-p)}{.01^2}$, so that the answer depends on the value of p. Note that the largest value that $p(1 - p)$ can take on is 0.25 (that is, $p = 0.5$ makes $p(1 - p)$ as large as possible). Thus if $n > \frac{1.96^2\times 0.25}{.01^2} = 9604,$ then the margin of error is less than 0.01 for all values of p.

6. (a) No. Because the p-value is less than 5%, $\mu = 5$ is rejected at the 5% level and is therefore not contained in the 95% confidence interval.
   (b) No. This would require calculation of the t-statistic for $\mu = 5$, which requires $\overline{Y}$ and $\text{SE}(\overline{Y})$. Only one the p-value for $\mu = 5$ is given in the problem.

7. The null hypothesis in that the survey is a random draw from a population with $p = 0.11$. The t-statistic is $t = \frac{\hat{p}-0.11}{SE(\hat{p})}$, where $SE(\hat{p}) = \hat{p}(1-\hat{p})/n.$ (An alternative formula for SE($\hat{p}$) is $0.11\times(1-0.11)/n,$ which is valid under the null hypothesis that $p = 0.11$). The value of the t-statistic is −2.71, which has a p-value of that is less than 0.01. Thus the null hypothesis $p = 0.11$ (the survey is unbiased) can be rejected at the 1% level.

8 $1110 \pm 1.96\left(\frac{123}{\sqrt{1000}}\right)$ or $1110 \pm 7.62.$

9. Denote the life of a light bulb from the new process by Y. The mean of Y is μ and the standard deviation of Y is $\sigma_Y = 200$ hours. $\overline{Y}$ is the sample mean with a sample size $n = 100$. The standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{200}{\sqrt{100}} = 20$ hours. The hypothesis test is $H_0 : \mu = 2000$ vs. $H_1 : \mu > 2000$. The manager will accept the alternative hypothesis if $\overline{Y} > 2100$ hours.
   (a) The size of a test is the probability of erroneously rejecting a null hypothesis when it is valid. The size of the manager's test is

$$\begin{aligned}\text{size} &= \Pr(\overline{Y} > 2100|\mu = 2000) = 1 - \Pr(\overline{Y} \leq 2100|\mu = 2000)\\ &= 1 - \Pr\left(\frac{\overline{Y} - 2000}{20} \leq \frac{2100 - 2000}{20}|\mu = 2000\right)\\ &= 1 - \Phi(5) = 1 - 0.999999713 = 2.87\times 10^{-7}.\end{aligned}$$

   $\Pr(\overline{Y} > 2100|\mu = 2000)$ means the probability that the sample mean is greater than 2100 hours when the new process has a mean of 2000 hours.
   (b) The power of a test is the probability of correctly rejecting a null hypothesis when it is invalid. We calculate first the probability of the manager erroneously accepting the null hypothesis when it is invalid:

$$\begin{aligned}\beta &= \Pr(\overline{Y} \leq 2100|\mu = 2150) = \Pr\left(\frac{\overline{Y} - 2150}{20} \leq \frac{2100 - 2150}{20}|\mu = 2150\right)\\ &= \Phi(-2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.\end{aligned}$$

   The power of the manager's testing is $1 - \beta = 1 - 0.0062 = 0.9938.$

(c) For a test with 5%, the rejection region for the null hypothesis contains those values of the t-statistic exceeding 1.645.

$$t^{act} = \frac{\bar{Y}^{act} - 2000}{20} > 1.645 \Rightarrow \bar{Y}^{act} > 2000 + 1.645 \times 20 = 2032.9.$$

The manager should believe the inventor's claim if the sample mean life of the new product is greater than 2032.9 hours if she wants the size of the test to be 5%.

10. (a) New Jersey sample size $n_1 = 100$, sample average $\bar{Y}_1 = 58$, sample standard deviation $s_1 = 8$. The standard error of $\bar{Y}_1$ is $\text{SE}(\bar{Y}_1) = \frac{s_1}{\sqrt{n_1}} = \frac{8}{\sqrt{100}} = 0.8$. The 95% confidence interval for the mean score of all New Jersey third graders is

$$\mu_1 = \bar{Y}_1 \pm 1.96\text{SE}(\bar{Y}_1) = 58 \pm 1.96 \times 0.8 = (56.432, 59.568).$$

(b) Iowa sample size $n_2 = 200$, sample average $\bar{Y}_2 = 62$, sample standard deviation $s_2 = 11$. The standard error of $\bar{Y}_1 - \bar{Y}_2$ is $\text{SE}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{64}{100} + \frac{121}{200}} = 1.1158$. The 90% confidence interval for the difference in mean score between the two states is

$$\begin{aligned}\mu_1 - \mu_2 &= (\bar{Y}_1 - \bar{Y}_2) \pm 1.64\text{SE}(\bar{Y}_1 - \bar{Y}_2) \\ &= (58 - 62) \pm 1.64 \times 1.1158 = (-5.8299, -2.1701).\end{aligned}$$

(c) The hypothesis tests for the difference in mean scores is

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

From part (b) the standard error of the difference in the two sample means is $\text{SE}(\bar{Y}_1 - \bar{Y}_2) = 1.1158$. The t-statistic for testing the null hypothesis is

$$t^{act} = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{SE}(\bar{Y}_1 - \bar{Y}_2)} = \frac{58 - 62}{1.1158} = -3.5849.$$

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

$$p - \text{value} = 2\Phi(-|t^{act}|) = 2\Phi(-3.5849) = 2 \times 0.00017 = 0.00034.$$

Because of the extremely low p-value, we can reject the null hypothesis with a very high degree of confidence. That is, the population means for Iowa and New Jersey students are different.

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.

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

12. Sample size for men $n_1 = 100,$ sample average $\overline{Y}_1 = 3100,$ sample standard deviation $s_1 = 200.$ Sample size for women $n_2 = 64,$ sample average $\overline{Y}_2 = 2900,$ sample standard deviation $s_2 = 320.$ The standard error of $\overline{Y}_1 - \overline{Y}_2$ is $\text{SE}(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{200^2}{100} + \frac{320^2}{64}} = 44.721.$

(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

$$t^{act} = \frac{\overline{Y}_1 - \overline{Y}_2}{\text{SE}(\overline{Y}_1 - \overline{Y}_2)} = \frac{3100 - 2900}{44.721} = 4.4722.$$

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

$$p\text{-value} = 2\Phi(-|t^{act}|) = 2\Phi(-4.4722) = 2 \times (3.8744 \times 10^{-6}) = 7.7488 \times 10^{-6}.$$

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. To examine whether there is gender discrimination in the compensation policies, we take the following one-sided alternative test

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

With the t-statistic $t^{act} = 4.4722,$ the p-value for the one-sided test is:

$$p\text{-value} = 1 - \Phi(t^{act}) = 1 - \Phi(4.4722) = 1 - 0.999996126 = 3.874 \times 10^{-6}.$$

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 greater than 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.

13. (a) Sample size $n = 420,$ sample average $\overline{Y} = 654.2,$ sample standard deviation $s_Y = 19.5.$ The standard error of $\overline{Y}$ is $\text{SE}(\overline{Y}) = \frac{s_Y}{\sqrt{n}} = \frac{19.5}{\sqrt{420}} = 0.9515.$ The 95% confidence interval for the mean test score in the population is

$$\mu = \overline{Y} \pm 1.96\text{SE}(\overline{Y}) = 654.2 \pm 1.96 \times 0.9515 = (652.34,\ 656.06).$$

(b) The data are: sample size for small classes $n_1 = 238$, sample average $\bar{Y}_1 = 657.4$, sample standard deviation $s_1 = 19.4$; sample size for large classes $n_2 = 182$, sample average $\bar{Y}_2 = 650.0$, sample standard deviation $s_2 = 17.9$. The standard error of $\bar{Y}_1 - \bar{Y}_2$ is $SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = 1.8281$. The hypothesis tests for higher average scores in smaller classes is

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

The t-statistic is

$$t^{act} = \frac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{SE}(\bar{Y}_1 - \bar{Y}_2)} = \frac{657.4 - 650.0}{1.8281} = 4.0479.$$

The p-value for the one-sided test is:

$$p\text{-value} = 1 - \Phi(t^{act}) = 1 - \Phi(4.0479) = 1 - 0.999974147 = 2.5853 \times 10^{-5}.$$

With the small p-value, the null hypothesis can be rejected with a high degree of confidence. There is statistically significant evidence that the districts with smaller classes have higher average test scores.

14. We have the following relations: $1\,in = 0.0254\,m$ (or $1\,m = 39.37\,in$), $1\,lb = 0.4536\,kg$ (or $1\,kg = 2.2046\,lb$). The summary statistics in the metric system are $\bar{X} = 70.5 \times 0.0254 = 1.79\,m$; $\bar{Y} = 158 \times 0.4536 = 71.669\,kg$; $s_X = 1.8 \times 0.0254 = 0.0457\,m$; $s_Y = 14.2 \times 0.4536 = 6.4411\,kg$; $s_{XY} = 21.73 \times 0.0254 \times 0.4536 = 0.2504\,m \times kg$, and $r_{XY} = 0.85$.

15. Let p denote the fraction of the population that preferred Bush.

(a) $\hat{p} = 405/755 = 0.536$; $\mathrm{SE}(\hat{p}) = .0181$; 95% confidence interval is $\hat{p} \pm 1.96\,\mathrm{SE}(\hat{p})$ or $0.536 \pm .036$

(b) $\hat{p} = 378/756 = 0.500$; $\mathrm{SE}(\hat{p}) = .0182$; 95% confidence interval is $\hat{p} \pm 1.96\,\mathrm{SE}(\hat{p})$ or 0.500 ± 0.36

(c) $\hat{p}_{Sep} - \hat{p}_{Oct} = 0.036$; $\mathrm{SE}(\hat{p}_{Sep} - \hat{p}_{Oct}) = \sqrt{\frac{0.536(1-0.536)}{755} + \frac{0.5(1-0.5)}{756}}$ (because the surveys are independent. The 95% confidence interval for the change in p is $(\hat{p}_{Sep} - \hat{p}_{Oct}) \pm 1.96\,\mathrm{SE}(\hat{p}_{Sep} - \hat{p}_{Oct})$ or $0.036 \pm .050$. The confidence interval includes $(p_{Sep} - p_{Oct}) = 0.0$, so there is not statistically significance evidence of a change in voters' preferences.

16. (a) The 95% confidence interval if $\bar{Y} \pm 1.96\,\mathrm{SE}(\bar{Y})$ or $1013 \pm 1.96 \times \frac{108}{\sqrt{453}}$ or 1013 ± 9.95.

(b) The confidence interval in (a) does not include $\mu = 0$, so the null hypothesis that $\mu = 0$ (Florida students have the same average performance as students in the U.S.) can be rejected at the 5% level.

(c) (i) The 95% confidence interval is $\bar{Y}_{prep} - \bar{Y}_{Non-prep} \pm 1.96\,SE(\bar{Y}_{prep} - \bar{Y}_{Non-prep})$ where $\mathrm{SE}(\bar{Y}_{prep} - \bar{Y}_{Non-prep}) = \sqrt{\frac{S_{prep}^2}{n_{prep}} + \frac{S_{non-prep}^2}{n_{non-prep}}} = \sqrt{\frac{95^2}{503} + \frac{108^2}{453}} = 6.61$; the 95% confidence interval is $(1019 - 1013) \pm 12.96$ *or* 6 ± 12.96.

(ii) No. The 95% confidence interval includes $\mu_{prep} - \mu_{non-prep} = 0$.

(d) (i) Let X denote the change in the test score. The 95% confidence interval for μ_X is $\bar{X} \pm 1.96\, SE(\bar{X})$, where $\text{SE}(\bar{X}) = \frac{60}{\sqrt{453}} = 2.82$; thus, the confidence interval is 9 ± 5.52.

(ii) Yes. The 95% confidence interval does not include $\mu_X = 0$.

(iii) Randomly select n students who have taken the test only one time. Randomly select one half of these students and have them take the prep course. Administer the test again to all of the n students. Compare the gain in performance of the prep-course second-time test takers to the non-prep-course second-time test takers.

17. (a) The 95% confidence interval is $\bar{Y}_{m,2004} - \bar{Y}_{m,1992} \pm 1.96\,\text{SE}(\bar{Y}_{m,2004} - \bar{Y}_{m,1992})$ where

$\text{SE}(\bar{Y}_{m,2004} - \bar{Y}_{m,1992}) = \sqrt{\frac{S^2_{m,2004}}{n_{m,2004}} + \frac{S^2_{m,1992}}{n_{m,1992}}} = \sqrt{\frac{10.39^2}{1901} + \frac{8.70^2}{1592}} = 0.32$; the 95% confidence interval is $(21.99 - 20.33) \pm 0.63$ or 1.66 ± 0.63.

(b) The 95% confidence interval is $\bar{Y}_{w,2004} - \bar{Y}_{w,1992} \pm 1.96\,\text{SE}(\bar{Y}_{w,2004} - \bar{Y}_{w,1992})$ where

$\text{SE}(\bar{Y}_{w,2004} - \bar{Y}_{w,1992}) = \sqrt{\frac{S^2_{w,2004}}{n_{w,2004}} + \frac{S^2_{w,1992}}{n_{w,1992}}} = \sqrt{\frac{8.16^2}{1739} + \frac{6.90^2}{1370}} = 0.27$; the 95% confidence interval is $(18.47 - 17.60) \pm 0.53$ or 0.87 ± 0.53.

(c) The 95% confidence interval is

$(\bar{Y}_{m,2004} - \bar{Y}_{m,1992}) - (\bar{Y}_{w,2004} - \bar{Y}_{w,1992}) \pm 1.96\,\text{SE}[(\bar{Y}_{m,2004} - \bar{Y}_{m,1992}) - (\bar{Y}_{w,2004} - \bar{Y}_{w,1992})]$, where

$\text{SE}[(\bar{Y}_{m,2004} - \bar{Y}_{m,1992}) - (\bar{Y}_{w,2004} - \bar{Y}_{w,1992})] = \sqrt{\frac{S^2_{m,2004}}{n_{m,2004}} + \frac{S^2_{m,1992}}{n_{m,1992}} + \frac{S^2_{w,2004}}{n_{w,2004}} + \frac{S^2_{w,2004}}{n_{w,2004}}} = \sqrt{\frac{10.39^2}{1901} + \frac{8.70^2}{1592} + \frac{8.16^2}{1739} + \frac{6.90^2}{1370}} = 0.42.$

The 95% confidence interval is $(21.99 - 20.33) - (18.47 - 17.60) \pm 1.96 \times 0.42$ or 0.79 ± 0.82.

18. $Y_1, \ldots, Y_n$ are i.i.d. with mean μ_Y and variance σ_Y^2. The covariance $\text{cov}(Y_j, Y_i) = 0$, $j \neq i$. The sampling distribution of the sample average $\bar{Y}$ has mean μ_Y and variance $\text{var}(\bar{Y}) = \sigma^2_{\bar{Y}} = \frac{\sigma_Y^2}{n}$.

(a)

$$
\begin{aligned}
E[(Y_i - \bar{Y})^2] &= E\{[(Y_i - \mu_Y) - (\bar{Y} - \mu_Y)]^2\} \\
&= E[(Y_i - \mu_Y)^2 - 2(Y_i - \mu_Y)(\bar{Y} - \mu_Y) + (\bar{Y} - \mu_Y)^2] \\
&= E[(Y_i - \mu_Y)^2] - 2E[(Y_i - \mu_Y)(\bar{Y} - \mu_Y)] + E[(\bar{Y} - \mu_Y)^2] \\
&= \text{var}(Y_i) - 2\,\text{cov}(Y_i, \bar{Y}) + \text{var}(\bar{Y}).
\end{aligned}
$$

(b)

$$\begin{aligned}
\operatorname{cov}(\bar{Y}, Y) &= E[(\bar{Y}-\mu_Y)(Y_i-\mu_Y)] \\
&= E\left[\left(\frac{\sum_{j=1}^{n} Y_j}{n}-\mu_Y\right)(Y_i-\mu_Y)\right] \\
&= E\left[\left(\frac{\sum_{j=1}^{n} (Y_j-\mu_Y)}{n}\right)(Y_i-\mu_Y)\right] \\
&= \frac{1}{n}E[(Y_I-\mu_Y)^2+\frac{1}{n}\sum_{j\neq i}E[(Y_j-\mu_Y)(Y_i-\mu_Y)] \\
&= \frac{1}{n}\sigma_Y^2+\frac{1}{n}\sum_{j\neq i}\operatorname{cov}(Y_j, Y_i) \\
&= \frac{\sigma_Y^2}{n}.
\end{aligned}$$

(c)

$$\begin{aligned}
E\left(s_Y^2\right) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}E[(Y_i-\bar{Y})^2] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\operatorname{var}(Y_i)-2\operatorname{cov}(Y_i,\bar{Y})+\operatorname{var}(\bar{Y})] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\sigma_Y^2-2\times\frac{\sigma_Y^2}{n}+\frac{\sigma_Y^2}{n}\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{n-1}{n}\sigma_Y^2\right) \\
&= \sigma_Y^2.
\end{aligned}$$

19. (a) No. $E(Y_i^2)=\sigma_Y^2+\mu_Y^2$ and $E(Y_iY_j)=\mu_Y^2$ for $i\neq j$. Thus

$$\begin{aligned}
E(\bar{Y}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}E(Y_i^2)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}E(Y_iY_j) \\
&= \mu_Y^2+\frac{1}{n}\sigma_Y^2
\end{aligned}$$

(b) Yes. If $\bar{Y}$ gets arbitrarily close to μ_Y with probability approaching 1 as n gets large, then $\bar{Y}^2$ gets arbitrarily close to μ_Y^2 with probability approaching 1 as n gets large. (As it turns out, this is an example of the "continuous mapping theorem" discussed in Chapter 17.)

20. Using analysis like that in equation (3.29)

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$
$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)\right] - \left(\frac{n}{n-1}\right)(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)$$

because $\bar{X} \xrightarrow{p} \mu_X$ and $\bar{Y} \xrightarrow{p} \mu_Y$ the final term converges in probability to zero.

Let $W_i = (X_i - \mu_x)(Y_i - \mu_Y)$. Note W_i is *iid* with mean σ_{XY} and second moment $E[(X_i - \mu_X)^2 (Y_i - \mu_Y)^2]$. But $E[(X_i - \mu_X)^2 (Y_i - \mu_Y)^2] \le \sqrt{E(X_i - \mu_X)^4}\sqrt{E(Y_i - \mu_Y)^4}$ from the Cauchy-Schwartz inequality. Because *X* and *Y* have finite fourth moments, the second moment of W_i is finite, so that it has finite variance. Thus $\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} E(W_i) = \sigma_{XY}$. Thus, $s_{XY} \xrightarrow{p} \sigma_{XY}$ (because the term $\frac{n}{n-1} \to 1$).

21. Set $n_m = n_w = n$, and use equation (3.19) write the squared SE of $\bar{Y}_m - \bar{Y}_w$ as

$$[SE(\bar{Y}_m - \bar{Y}_w)]^2 = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}(Y_{mi} - \bar{Y}_m)^2}{n} + \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}(Y_{wi} - \bar{Y}_w)^2}{n}$$

$$= \frac{\sum_{i=1}^{n}(Y_{mi} - \bar{Y}_m)^2 + \sum_{i=1}^{n}(Y_{wi} - \bar{Y}_w)^2}{n(n-1)}.$$

Similary, using equation (3.23)

$$[SE_{pooled}(\bar{Y}_m - \bar{Y}_w)]^2 = \frac{\frac{1}{2(n-1)}\left[\sum_{i=1}^{n}(Y_{mi} - \bar{Y}_m)^2 + \frac{1}{(n-1)}\sum_{i=1}^{n}(Y_{wi} - \bar{Y}_w)^2\right]}{2n}$$

$$= \frac{\sum_{i=1}^{n}(Y_{mi} - \bar{Y}_m)^2 + \sum_{i=1}^{n}(Y_{wi} - \bar{Y}_w)^2}{n(n-1)}.$$