

# **PRINCIPLES OF ECONOMETRICS**

**5<sup>TH</sup> EDITION**

**ANSWERS TO ODD-NUMBERED**  
**EXERCISES IN APPENDIX C**

**EXERCISE C.1**

- (a) A linear estimator is one that can be written in the form  $\sum a_i Y_i$  where  $a_i$  is a constant.

$Y^*$  is a linear estimator where  $a_i = 1/2$  for  $i = 1, 2$  and  $a_i = 0$  for  $i = 3, 4, \dots, N$ .

- (b) The expected value of an unbiased estimator is equal to the true population mean.

$$E(Y^*) = E\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{2}E(Y_1) + \frac{1}{2}E(Y_2) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

- (c) The variance of  $Y^*$  is given by

$$\text{var}(Y^*) = \text{var}\left(\frac{1}{2}Y_1 + \frac{1}{2}Y_2\right) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{\sigma^2}{2}$$

- (d) The sample mean is a better estimator because it uses more information. The variance of the sample mean is  $\sigma^2/N$  which is smaller than  $\sigma^2/2$  when  $N > 2$ , thus making it a better estimator than  $Y^*$ . In general, increasing sample information reduces sampling variation.

**EXERCISE C.3**

The probability that, in a 9 hour day, more than 20,000 pieces will be sold is the same as the probability that average hourly sales of fried chicken is greater than  $20,000/9 \approx 2,222$  pieces.

$$P[\bar{X} > 2222] = P\left[Z > \frac{2222 - 2000}{500/\sqrt{9}}\right] = 0.091$$

**EXERCISE C.5**

- (a) We set up the hypotheses  $H_0 : \mu \leq 170$  versus  $H_1 : \mu > 170$ . The alternative is  $H_1 : \mu > 170$  because we want to establish whether the mean monthly account balance is more than 170.

The test statistic, given  $H_0$  is true, is  $t = \frac{\bar{X} - 170}{\hat{\sigma}/\sqrt{N}} \sim t_{(399)}$ . The rejection region is  $t \geq 1.649$ .

The value of the test statistic is  $t = 2.462$ . Since  $t = 2.462 > 1.649$ , we reject  $H_0$  and conclude that the new accounting system is cost effective.

- (b)  $p = P[t_{(399)} \geq 2.462] = 1 - P[t_{(399)} < 2.462] = 0.007$

**EXERCISE C.7**

- (a) To test whether current hiring procedures are effective, we test the hypothesis that  $H_0 : \mu \leq 450$  against  $H_1 : \mu > 450$ . The manager is interested in workers who can process

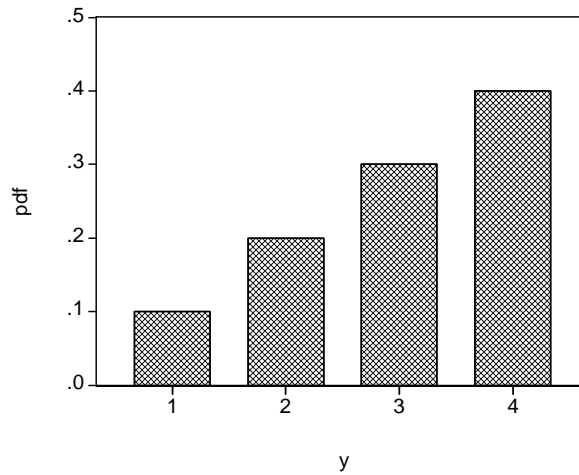
at least 450 pieces per day. The test statistic, when  $H_0$  is true, is  $t = \frac{\bar{X} - 450}{\hat{\sigma}/\sqrt{N}} \sim t_{(49)}$ . The

value of the test statistic is  $t = 1.861$ . Using a 5% significance level at 49 degrees of freedom, the rejection region is  $t > 1.677$ . Since  $1.861 > 1.677$ , we reject  $H_0$  and conclude that the current hiring procedures are effective.

- (b) A type I error occurs when we reject the null hypothesis but it is actually true. In this example, a type I error occurs when we wrongly reject the hypothesis that the hiring procedures are effective. This would be a costly error to make because we would be dismissing a cost effective practice.
- (c)  $p\text{-value} = (1 - P(t_{(49)} < 1.861)) = (1 - 0.9656) = 0.0344$

**EXERCISE C.9**

- (a) A sketch of the *pdf* is shown below.



- (b)  $E(Y) = 3$
- (c)  $\text{var}(Y) = 1$
- (d)  $E(\bar{Y}) = 3$  and  $\text{var}(\bar{Y}) = \frac{1}{3}$

**EXERCISE C.11**

- (a) The sample variance and sample standard deviation are as follows

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1} = \frac{33296.003}{967} = 34.43227$$

$$\hat{\sigma} = \sqrt{34.43227} = 5.8679$$

- (b) The required sample moments are

$$\tilde{\mu}_2 = \left( \frac{N-1}{N} \right) \hat{\sigma}^2 = \frac{967}{968} \times 34.43227 = 34.3967$$

$$\tilde{\mu}_3 = \frac{\sum_{i=1}^N (X_i - \bar{X})^3}{N} = \frac{-137910.04}{968} = -142.46905$$

$$\tilde{\mu}_4 = \frac{\sum_{i=1}^N (X_i - \bar{X})^4}{N} = \frac{6392783.3}{968} = 6604.11498$$

- (c) The skewness and kurtosis are given by

$$S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3} = \frac{-142.46905}{(34.3967)^{3/2}} = -0.7062$$

$$K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4} = \frac{6604.11498}{(34.3967)^2} = 5.5819$$

To be compatible with the normal distribution, we require  $S \approx 0$  and  $K \approx 3$ . The above values appear to be too different from 0 and 3 to be compatible with the normal distribution. We can check on this assertion by using the Jarque-Bera test.

- (d) Calculating the value of the Jarque-Bera test statistic, we obtain

$$JB = \frac{N}{6} \left( S^2 + \frac{(K-3)^2}{4} \right) = \frac{968}{6} \left( (-0.7062)^2 + \frac{(5.5819-3)^2}{4} \right) = 349.33$$

Because  $JB = 349.33$  exceeds the 1% critical value  $\chi_{0.99,2}^2 = 9.210$ , we reject a null hypothesis that the data are normally distributed.

### EXERCISE C.13

- (a) The likelihood function is

$$L(\lambda | y_1, \dots, y_N) = \prod_{i=1}^N P(Y = y_i) = \prod_{i=1}^N \left( \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right)$$

Thus, the log-likelihood function is

$$\begin{aligned} \ln L(\lambda | y_1, \dots, y_N) &= \ln \left[ \prod_{i=1}^N \left( \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right) \right] = \sum_{i=1}^N \ln \left( \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right) \\ &= \sum_{i=1}^N (-\lambda + y_i \ln \lambda - \ln y_i!) \\ &= \ln \lambda \sum_{i=1}^N y_i - N\lambda - \sum_{i=1}^N \ln(y_i!) \end{aligned}$$

- (b) Differentiating the log-likelihood with respect to  $\lambda$ , we obtain

$$\frac{d \ln L}{d \lambda} = \frac{\sum_{i=1}^N y_i}{\tilde{\lambda}} - N = 0 \Rightarrow \tilde{\lambda} = \frac{\sum_{i=1}^N y_i}{N}$$

- (c) The second derivative of the log-likelihood function is  $\frac{d^2 \ln L}{d \lambda^2} = -\frac{\sum_{i=1}^N y_i}{\lambda^2}$ . Since all  $y_i \geq 0$ , and  $\lambda^2 > 0$  is always positive, the second derivative is negative providing at least one  $y_i > 0$ .

- (d) The mean and variance of  $\tilde{\lambda}$  are given by

$$E(\tilde{\lambda}) = \frac{E\left(\sum_{i=1}^N y_i\right)}{N} = \frac{\sum_{i=1}^N E(y_i)}{N} = \frac{\sum_{i=1}^N \lambda}{N} = \frac{N\lambda}{N} = \lambda$$

$$\text{var}(\tilde{\lambda}) = \frac{\text{var}\left(\sum_{i=1}^N y_i\right)}{N^2} = \frac{\sum_{i=1}^N \text{var}(y_i)}{N^2} = \frac{\sum_{i=1}^N \lambda}{N^2} = \frac{N\lambda}{N^2} = \frac{\lambda}{N}$$

- (e) The information measure is given by

$$I(\lambda) = -\left\{E\left[\frac{d^2 \ln L(\lambda)}{d\lambda^2}\right]\right\} = -E\left(-\frac{\sum_{i=1}^N y_i}{\lambda^2}\right) = \frac{\sum_{i=1}^N E(y_i)}{\lambda^2} = \frac{N\lambda}{\lambda^2} = \frac{N}{\lambda}$$

### EXERCISE C.15

- (a) From Exercises C.12 and C.13, the maximum likelihood estimate is  $\tilde{\lambda} = 5/3$  and the information measure is

$$I(\tilde{\lambda}) = \frac{N}{\tilde{\lambda}} = \frac{3}{5/3} = \frac{9}{5}$$

- (b) The likelihood ratio test statistic is given by

$$LR = 2[\ln L(\tilde{\lambda}) - \ln L(1)]$$

Noting that

$$\ln L(\lambda) = \ln(0.25) - 3\lambda + 5 \ln \lambda$$

we find  $LR = 1.108$ . Since  $LR = 1.108 < \chi_{(0.95, 1)}^2 = 3.841$ , we do not reject  $H_0: \lambda = 1$ .

- (c) Using the Wald statistic, we have

$$W = (\tilde{\lambda} - 1)^2 \times \left[ -\frac{d^2 \ln L(\lambda)}{d\lambda^2} \right]_{\lambda=\tilde{\lambda}} = (\tilde{\lambda} - 1)^2 \times \frac{5}{\tilde{\lambda}^2} = 0.8$$

Since  $W = 0.8 < \chi_{(0.95, 1)}^2 = 3.841$ , we do not reject  $H_0: \lambda = 1$ .

- (d) Using a modified version of the Wald test, we have

$$W = (\tilde{\lambda} - 1)^2 \times I(\tilde{\lambda}) = \left(\frac{5}{3} - 1\right)^2 \times \frac{3}{5/3} = 0.8$$

Since  $W = 0.8 < \chi_{(0.95, 1)}^2 = 3.841$ , we do not reject  $H_0: \lambda = 1$ .

- (e) The score function, evaluated at  $\lambda = 1$ , is

$$s(1) = \frac{d \ln L}{d\lambda} \Big|_{\lambda=1} = \left(-3 + \frac{5}{\lambda}\right) \Big|_{\lambda=1} = 2$$

- (f) When  $\lambda = 1$ ,  $I(\lambda) = 3/\lambda = 3$ .

- (g) The LM statistic for hypothesis  $H_0 : \lambda = 1$  versus the alternative  $H_1 : \lambda \neq 1$  is

$$LM = [s(1)]^2 [I(1)]^{-1} = 2^2 \times \frac{1}{3} = 1.333$$

Since  $LM = 1.333 < \chi_{(0.95, 1)}^2 = 3.841$ , we do not reject  $H_0 : \lambda = 1$ .

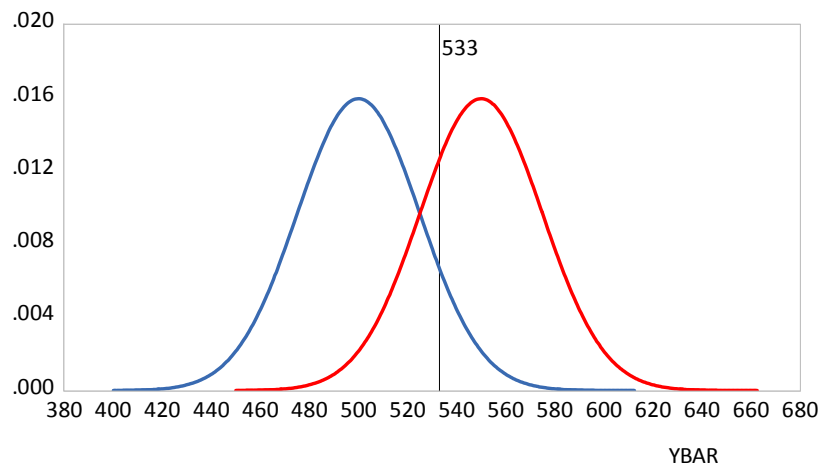
### EXERCISE C.17

- (a) When  $Y \sim N(500, 100^2)$  and  $N = 25$ ,  $\bar{Y} \sim N(500, 100^2/25)$ , and

$$P(\bar{Y} \geq 510) = P\left(Z \geq \frac{510 - 500}{100/\sqrt{25}}\right) = 0.3085$$

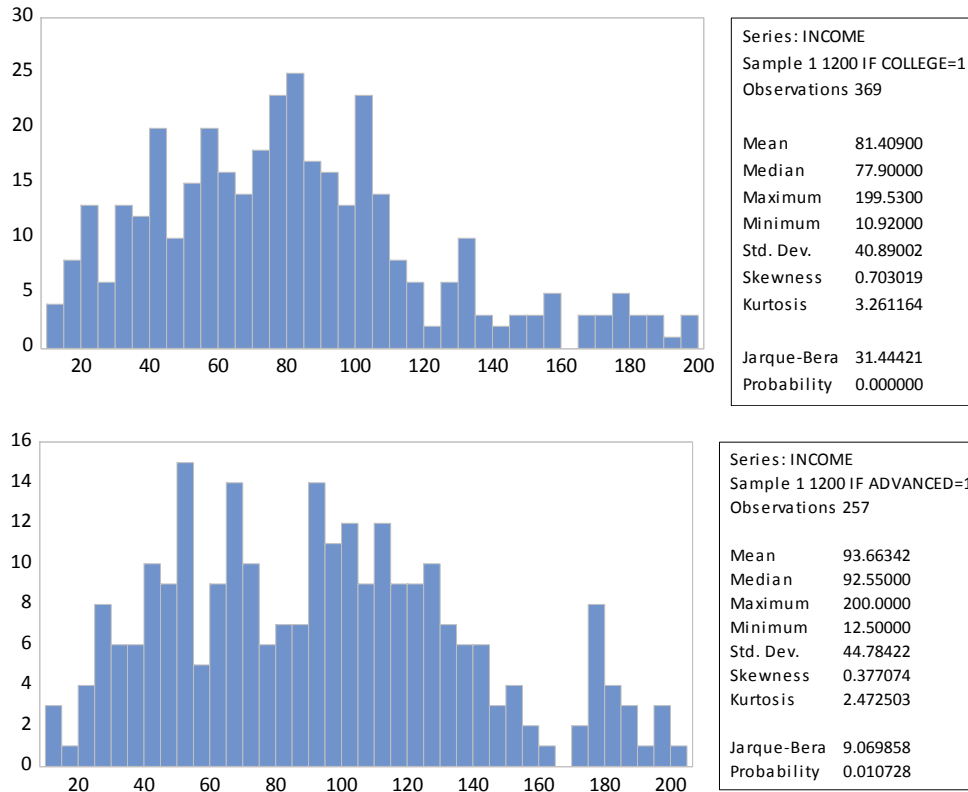
The average score of 510 is not convincing. There is still a 30% chance of getting a score that high through sampling error.

- (b) In this case,  $P(\bar{Y} \geq 533) = 0.0495 \approx 0.05$ . The statement is correct. We can conclude that smaller classes raise average test scores, with a sampling error probability that this conclusion is incorrect.
- (c) When  $Y \sim N(550, 100^2)$  and  $N = 25$ ,  $\bar{Y} \sim N(550, 100^2/25)$ , and  $P(\bar{Y} \geq 533) = 0.8023$ . We want this probability to be large. It is the probability of rejecting the incorrect hypothesis  $H_0 : \mu = 500$  when  $\mu = 550$  and a 5% significance level is used.
- (d) From part (c),  $P(\bar{Y} \leq 533) = 1 - P(\bar{Y} \geq 533) = 0.1977$ . This value is the probability of a Type II error when  $\mu = 550$  and the probability of a Type I error is 0.05.
- (e) In the diagram below the probability of a Type I error is the area under the blue curve to the right of 533; the probability of a Type II error is the area under the red curve to the left of 533. If the threshold is pushed to the right, the probability of a Type I error decreases, but the probability of a Type II error increases. If the threshold is pushed to the left, the probability of a Type II error decreases, but the probability of a Type I error increases.



**EXERCISE C.19**

- (a) The college-degree histogram is more skewed to the right than the advanced-degree histogram. If we take a central point on the income axis, say 100, then the proportion of households with incomes greater than 100 is much greater for the advanced-degree households. The advanced-degree histogram is located further to the right.



- (b) The proportion of advanced-degree households with incomes more than \$10,000 is

$$111/257 = 0.4319$$

The proportion of college-degree households with incomes more than \$10,000 is

$$103/369 = 0.2791$$

- (c) The value of the test statistic for testing  $H_0 : \mu_{ADV} \leq 90$  against the alternative  $H_1 : \mu_{ADV} > 90$  is

$$t = \frac{\bar{y}_{ADV} - 90}{\hat{\sigma}_{ADV} / \sqrt{N_{ADV}}} = \frac{93.66342 - 90}{44.78422 / \sqrt{257}} = 1.311$$

The corresponding  $p$ -value is  $P(t_{(256)} > 1.311) = 0.0955$ . The 5% critical value is  $t_{(0.95, 256)} = 1.651$ . We do not reject  $H_0 : \mu_{ADV} \leq 90$ . There is insufficient evidence to conclude that mean income for advanced-degree households is greater than \$9,000.

- (d) The test statistic value for testing  $H_0 : \mu_{COLL} \leq 90$  against the alternative  $H_1 : \mu_{COLL} > 90$  is

$$t = \frac{\bar{y}_{COLL} - 90}{\hat{\sigma}_{COLL} / \sqrt{N_{COLL}}} = \frac{81.409 - 90}{40.89 / \sqrt{369}} = -4.036$$

The corresponding  $p$ -value is  $P(t_{(368)} > -4.036) = 1.000$ . The 5% critical value is  $t_{(0.95, 368)} = 1.649$ . We do not reject  $H_0 : \mu_{COLL} \leq 90$ . There is insufficient evidence to conclude that mean income for college-degree households is greater than \$9,000. (Because  $\bar{y}_{COLL} < 90$ , this conclusion can be reached without computing the value of the  $t$ -statistic.)

- (e) A 95% interval estimate for the mean  $\mu_{ADV}$  is given by

$$\bar{y}_{ADV} \pm t_{(0.975, 256)} \hat{\sigma}_{ADV} / \sqrt{N_{ADV}} = 93.66342 \pm 1.9693 \times 44.78422 / \sqrt{257} = (88.162, 99.165)$$

A 95% interval estimate for the mean  $\mu_{COLL}$  is given by

$$\bar{y}_{COLL} \pm t_{(0.975, 368)} \hat{\sigma}_{COLL} / \sqrt{N_{COLL}} = 81.409 \pm 1.9664 \times 40.89 / \sqrt{369} = (77.223, 85.595)$$

- (f) The choice of test statistic for testing  $H_0 : \mu_{ADV} \leq \mu_{COLL}$  against the alternative  $H_1 : \mu_{ADV} > \mu_{COLL}$  depends on whether the population variances  $\sigma_{ADV}^2$  and  $\sigma_{COLL}^2$  are assumed to be equal. The estimates  $\hat{\sigma}_{ADV} = 44.784$  and  $\hat{\sigma}_{COLL} = 40.89$  are similar, but it is not obvious whether they could be two estimates of the same population variance. The  $F$ -test for equality of variances in Section C.7.3 of *POE5* does not reject  $H_0 : \sigma_{ADV}^2 = \sigma_{COLL}^2$  at a 5% significance level ( $p$ -value = 0.111). Thus, we use the  $t$ -test that assumes the population variances are equal. The pooled variance estimate is

$$\hat{\sigma}_P^2 = \frac{(N_{ADV} - 1)\hat{\sigma}_{ADV}^2 + (N_{COLL} - 1)\hat{\sigma}_{COLL}^2}{N_{ADV} + N_{COLL} - 2} = \frac{256 \times 44.78422^2 + 368 \times 40.89^2}{256 + 368} = 1808.8677$$

The value of the  $t$ -statistic for testing  $H_0 : \mu_{ADV} \leq \mu_{COLL}$  against  $H_1 : \mu_{ADV} > \mu_{COLL}$  is

$$t = \frac{\bar{y}_{ADV} - \bar{y}_{COLL}}{\hat{\sigma}_P \sqrt{\frac{1}{N_{ADV}} + \frac{1}{N_{COLL}}}} = \frac{93.66342 - 81.409}{42.5308 \sqrt{\frac{1}{257} + \frac{1}{369}}} = 3.546$$

The 5% critical value is  $t_{(0.95, 624)} = 1.647$ . Since  $3.546 > 1.647$ , we reject  $H_0 : \mu_{ADV} \leq \mu_{COLL}$ ; The mean income of advanced-degree households is greater than the mean income of college-degree households.

## EXERCISE C.21

- (a) The sample proportion of students enrolled in regular sized classes who score 500 points or more is  $\hat{p} = 0.35922$ . A 95% interval estimate for the population proportion is

$$\hat{p} \pm t_{(0.975, 411)} \sqrt{\hat{p}(1 - \hat{p}) / N} = (0.3128, 0.4057)$$



Because this interval contains the value 0.4, for a two-tail test and a 5% significance level, we cannot reject  $H_0 : p = 0.4$ .

- (b) The value of the test statistic for testing  $H_0 : p \leq 0.4$  against the alternative  $H_0 : p > 0.4$  is  $t = -1.725$ . The 5% critical value is  $t_{(0.95, 411)} = 1.649$ . Because  $-1.725 < 1.649$ , we do not reject  $H_0 : p \leq 0.4$ . There is no evidence to suggest that, in regular sized classes, the population proportion of students who score 500 points or more is greater than 0.4.
- (c) The value of the test statistic for testing  $H_0 : p = 0.4$  against the alternative  $H_0 : p < 0.4$  is  $t = -1.725$ . The 5% critical value is  $t_{(0.05, 411)} = -1.649$ . Because  $-1.725 < -1.649$ , we reject  $H_0 : p = 0.4$ . We conclude that, in regular sized classes, the population proportion of students who score 500 points or more is less than 0.4.
- (d) The sample proportion of students enrolled in small classes who score 500 points or more is  $\hat{p} = 0.38017$ . A 95% interval estimate for the population proportion is

$$\hat{p} \pm t_{(0.975, 362)} \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = (0.3301, 0.4303)$$

Because this interval contains the value 0.4, for a two-tail test and a 5% significance level, we cannot reject  $H_0 : p = 0.4$ .

The value of the test statistic for testing  $H_0 : p \leq 0.4$  against the alternative  $H_0 : p > 0.4$  is  $t = -0.778$ . The 5% critical value is  $t_{(0.95, 362)} = 1.649$ . Because  $-0.778 < 1.649$ , we do not reject  $H_0 : p \leq 0.4$ . There is no evidence to suggest that, in small classes, the population proportion of students who score 500 points or more is greater than 0.4.

The value of the test statistic for testing  $H_0 : p = 0.4$  against the alternative  $H_0 : p < 0.4$  is  $t = -0.778$ . The 5% critical value is  $t_{(0.05, 362)} = -1.649$ . Because  $-0.778 > -1.649$ , we fail to reject  $H_0 : p = 0.4$ . There is insufficient evidence to conclude that, in small classes, the population proportion of students who score 500 points or more is less than 0.4.

## EXERCISE C.23

### Using data file *cps5*

- (a) Using subscript  $F12$  for females with 12 years of education, and subscript  $F16$  for females with 16 years of education, the mean wages are  $\bar{y}_{F12} = 15.65354$  and  $\bar{y}_{F16} = 25.16512$ . These values suggest a large payoff to the extra years of education.
- (b) A 95% interval estimate for  $\mu_{F12}$ , the population mean wage of females with 12 years of education is  $\bar{y}_{F12} \pm t_{(0.975, 1049)} \hat{\sigma}_{F12} / \sqrt{N_{F12}} = (15.171, 16.137)$ . A 95% interval estimate for  $\mu_{F16}$ , the population mean wage of females with 16 years of education is  $\bar{y}_{F16} \pm t_{(0.975, 1160)} \hat{\sigma}_{F16} / \sqrt{N_{F16}} = (24.395, 25.935)$ . The intervals do not overlap.

- (c) Using subscript  $M12$  for males with 12 years of education, and subscript  $M16$  for males with 16 years of education, the mean wages are  $\bar{y}_{M12} = 19.02089$  and  $\bar{y}_{M16} = 30.6343$ . Again, there is a large payoff to the extra years of education. The payoff for males is  $\bar{y}_{M16} - \bar{y}_{M12} = 11.61341$ . The payoff for females is  $\bar{y}_{F16} - \bar{y}_{F12} = 9.51158$ . There is a larger payoff for males than to females for the extra years of education.
- (d) A 95% interval estimate for  $\mu_{M12}$ , the population mean wage of males with 12 years of education is  $\bar{y}_{M12} \pm t_{(0.975, 1588)} \hat{\sigma}_{M12} / \sqrt{N_{M12}} = (18.542, 19.499)$ . A 95% interval estimate for  $\mu_{M16}$ , the population mean wage of males with 16 years of education is  $(29.737, 31.531)$ . In both cases, there is no overlap with the comparable female intervals.

- (e) An estimate of the gender difference in payoff to extra education is

$$\hat{\theta} = (\bar{y}_{F16} - \bar{y}_{F12}) - (\bar{y}_{M16} - \bar{y}_{M12}) = 9.51158 - 11.61341 = -2.10183$$

- (f) To calculate this interval estimate, we need an estimate of the variance of  $\hat{\theta}$ . Since the different sample means will be independent, the variance of  $\hat{\theta}$  is equal to the sum of the variances of its components. Thus, we have

$$\begin{aligned} \widehat{\text{var}}(\hat{\theta}) &= \widehat{\text{var}}(\bar{y}_{F16}) + \widehat{\text{var}}(\bar{y}_{F12}) + \widehat{\text{var}}(\bar{y}_{M16}) + \widehat{\text{var}}(\bar{y}_{M12}) \\ &= \frac{13.37096^2}{1161} + \frac{7.976889^2}{1050} + \frac{16.85173^2}{1359} + \frac{9.728424^2}{1589} = 0.48311 \end{aligned}$$

A 95% interval estimate for  $\theta$  is given by

$$\hat{\theta} \pm Z_{(0.975)} \text{se}(\hat{\theta}) = -2.10183 \pm 1.96 \times \sqrt{0.48311} = (-3.947, -0.256)$$

Because this interval does not contain zero, and all values in the interval are negative, we can conclude, at a 5% significance level, that the extra 4 years of education benefits males more than females.

## EXERCISE C.23

Using data file *cps5\_small*

- (a) Using subscript  $F12$  for females with 12 years of education, and subscript  $F16$  for females with 16 years of education, the mean wages are  $\bar{y}_{F12} = 14.38919$  and  $\bar{y}_{F16} = 26.28851$ . These values suggest a large payoff to the extra years of education.
- (b) A 95% interval estimate for  $\mu_{F12}$ , the population mean wage of females with 12 years of education is  $\bar{y}_{F12} \pm t_{(0.975, 123)} \hat{\sigma}_{F12} / \sqrt{N_{F12}} = (13.345, 15.433)$ . A 95% interval estimate for  $\mu_{F16}$ , the population mean wage of females with 16 years of education is  $\bar{y}_{F16} \pm t_{(0.975, 147)} \hat{\sigma}_{F16} / \sqrt{N_{F16}} = (24.106, 28.471)$ . The intervals do not overlap.

- (c) Using subscript  $M12$  for males with 12 years of education, and subscript  $M16$  for males with 16 years of education, the mean wages are  $\bar{y}_{M12} = 19.28104$  and  $\bar{y}_{M16} = 32.18359$ . Again, there is a large payoff to the extra years of education. The payoff for males is  $\bar{y}_{M16} - \bar{y}_{M12} = 12.90255$ . The payoff for females is  $\bar{y}_{F16} - \bar{y}_{F12} = 11.89932$ . There is a larger payoff for males than to females for the extra years of education.
- (d) A 95% interval estimate for  $\mu_{M12}$ , the population mean wage of males with 12 years of education is  $\bar{y}_{M12} \pm t_{(0.975, 182)} \hat{\sigma}_{M12} / \sqrt{N_{M12}} = (18.040, 20.522)$ . A 95% interval estimate for  $\mu_{M16}$ , the population mean wage of males with 16 years of education is  $(28.686, 35.681)$ . In both cases, there is no overlap with the comparable female intervals.
- (e) An estimate of the gender difference in payoff to extra education is

$$\hat{\theta} = (\bar{y}_{F16} - \bar{y}_{F12}) - (\bar{y}_{M16} - \bar{y}_{M12}) = 11.89932 - 12.90255 = -1.00323$$

- (f) To calculate this interval estimate, we need an estimate of the variance of  $\hat{\theta}$ . Since the different sample means will be independent, the variance of  $\hat{\theta}$  is equal to the sum of the variances of its components. Thus, we have

$$\begin{aligned} \widehat{\text{var}}(\hat{\theta}) &= \widehat{\text{var}}(\bar{y}_{F16}) + \widehat{\text{var}}(\bar{y}_{F12}) + \widehat{\text{var}}(\bar{y}_{M16}) + \widehat{\text{var}}(\bar{y}_{M12}) \\ &= \frac{13.43445^2}{148} + \frac{5.87459^2}{124} + \frac{22.11877^2}{156} + \frac{8.50972^2}{183} = 5.029668 \end{aligned}$$

A 95% interval estimate for  $\theta$  is given by

$$\hat{\theta} \pm Z_{(0.975)} \text{se}(\hat{\theta}) = -1.00323 \pm 1.96 \times \sqrt{5.029668} = (-5.399, 3.392)$$

Because this interval contains zero – it includes both positive and negative values – we are unable to conclude, at a 5% significance level, that the extra 4 years of education benefits males more than females.

### EXERCISE C.25

- (a) The summary statistics follow. The statistics p50 and p90 are the 50<sup>th</sup> and 90<sup>th</sup> percentiles.

variable	$N$	mean	variance	Std. dev.	p50	p90
<i>TOTEXP</i>	1200	8.511769	85.97714	9.272386	5.7	18

- (b) There are 602 observations on households with total expenditures less than or equal to the median. The sample proportion of income spent on food by these households is 0.44418256. A 95% interval estimate of the true proportion is

$$\hat{p} \pm 1.96 \text{se}(\hat{p}) = 0.4442 \pm 1.96(0.020251) \cong [0.40, 0.48]$$

- (c) There are 128 observations on household with total expenditures greater than or equal to the 90<sup>th</sup> percentile value, which is given in (a). The sample proportion of income spent on food by these households is 0.24611172. The standard error is  $se(\hat{p}) = 0.03807279$ . A 95% interval estimate of the true proportion is  $[0.17, 0.32]$ .
- (d) We estimate with 95% confidence that the proportion of total expenditures going to food for households with less than the median total expenditures is between 0.40 and 0.48. For those households with total expenditures in the top 10% the proportion we estimate with 95% confidence that the proportion spent on food is between 0.17 and 0.32. The budget share of food is estimated to be smaller for those households with higher total expenditures.
- (e) An asymptotically valid  $t$ -test statistic is

$$t = (\hat{p} - 0.4) / se(\hat{p}) \stackrel{a}{\sim} t_{(N-1)}$$

Using the full sample of 1200 observations  $\hat{p} = 0.39044033$  and  $se(\hat{p}) = 0.01408299$ . The resulting  $t = -0.67880952$ . If we use critical values from the standard normal distribution,  $\pm 1.96$ , then we fail to reject the null hypothesis that the sample proportion of total expenditures spent on food is 0.4.

Using the 602 observations for households with total expenditures less than or equal to the median,  $\hat{p} = 0.44418256$  and  $se(\hat{p}) = 0.0202511$ . The resulting  $t = 2.1817363$ . If we use critical values from the standard normal distribution,  $\pm 1.96$ , then we reject the null hypothesis that the sample proportion of total expenditures spent on food is 0.4.

There are 128 observations on household with total expenditures greater than or equal to the 90<sup>th</sup> percentile value. For these observations  $\hat{p} = 0.24611172$  and  $se(\hat{p}) = 0.03807279$ . The resulting  $t = -4.0419496$ . If we use critical values from the standard normal distribution,  $\pm 1.96$ , then we reject the null hypothesis that the sample proportion of total expenditures spent on food is 0.4.

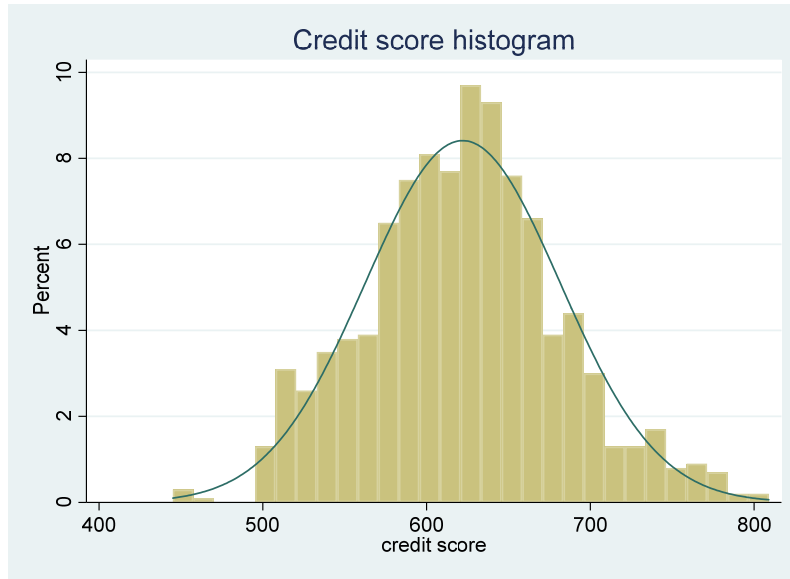
### EXERCISE C.27

- (a) The histogram of *CREDIT* is below. We have superimposed a normal distribution with the sample mean and variance. The sample values are

variable	$N$	mean	variance	skewness	kurtosis
<i>CREDIT</i>	1000	622.069	3543.574	0.1792352	3.061757

The Jarque-Bera test statistic is  $JB = \frac{N}{6} \left( S^2 + \frac{(K-3)^2}{4} \right)$ , where  $S$  is the skewness of the

data and  $K$  is the kurtosis. The calculated Jarque-Bera test statistic is  $JB = 5.5131198$ . The test critical value is  $\chi^2_{(0.95,2)} = 5.991$  so we fail to reject the null hypothesis of normality.



- (b) If the null hypothesis of equal variances is true, then

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(N_1-1, N_2-1)}$$

For population 1, where *DELINQUENT* = 1,  $N_1 = 199$  and the sample variance of *CREDIT* is 3813.9112. For population 2, where *DELINQUENT* = 0,  $N_2 = 801$  and the sample variance of *CREDIT* is 3470.6202. The  $F$ -statistic value is 1.0989134. For a two tail test the 0.025 and 0.975 percentiles of the  $F_{(198, 800)}$  distribution are 0.79617444 and 1.2374978, respectively. Thus, we cannot reject the null hypothesis of equal variances at the 5% level for these two populations.

- (c) If we assume the population variances are equal then we can use the test labelled Case 1. The test statistic is

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - c}{\sqrt{\hat{\sigma}_p^2 \left( \frac{1}{N_1} + \frac{1}{N_2} \right)}} \sim t_{(N_1 + N_2 - 2)}$$

where  $c = 0$  and

$$\hat{\sigma}_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}$$

The sample mean of population 1, where *DELINQUENT* = 1, is 616.26131 and the sample mean of population 1, where *DELINQUENT* = 0, is 623.51186. The estimated value of the pooled variance is 3538.728, and the denominator of the  $t$ -statistic is 4.7117343. The calculated  $t$ -statistic value is  $-1.538829$ . Using the standard normal critical values  $\pm 1.96$ , we cannot reject the null hypothesis that the means of the two populations are equal.

- (d) The test statistic is

$$V = \frac{(N-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi_{(N-1)}^2$$

where

$$\hat{\sigma}^2 = \sum (Y_i - \bar{Y})^2 / (N-1)$$

The sample variance of population 1, where *DELINQUENT* = 0, is 3813.9112 and  $N_1 = 199$ . The value of the test statistic is  $V = 209.76511$ . The critical values for this two tail test are  $\chi_{(0.025, 198)}^2 = 160.92463$  and  $\chi_{(0.975, 198)}^2 = 238.86122$ . Therefore we fail to reject the null hypothesis at the 5% level.

### EXERCISE C.29

- (a) The percentage of men who had completed at least 16 years of education by 1993 is 7.3794%.  
The percentage of the men's mothers who had at least 16 years of education is 7.6632%.  
The percentage of fathers who had at least 16 years of education is 17.1239%.
- (b) The sample mean number of years of schooling completed by men with fathers who had 16 or more years of education is  $\bar{y}_1 = 14.1326$ . The sample standard deviation is  $\hat{\sigma}_1 = 2.009447$ .  
The number of men in this subsample is  $N_1 = 181$ .

The sample mean number of years of schooling completed by men with fathers who had less than 16 years of education is  $\bar{y}_0 = 12.39954$ . The sample standard deviation is  $\hat{\sigma}_0 = 1.643745$ . The number of men in this subsample is  $N_0 = 876$ .

Since we wish to establish that  $\mu_1 > \mu_0$ , we set up the hypotheses as  $H_0: \mu_1 \leq \mu_0$  versus the alternative  $H_1: \mu_1 > \mu_0$ . To establish whether it is reasonable to use a pooled variance estimate, we compute  $F = \hat{\sigma}_1^2 / \hat{\sigma}_0^2 = 1.494$ . The 5% critical value for the  $F$ -test is  $F_{(0.975, 180, 875)} = 1.245$ . We proceed under the assumption the population variances corresponding to the two subsamples are not equal. Some preliminary calculations are

$$\widehat{\text{var}}(\bar{y}_1) = \frac{\hat{\sigma}_1^2}{N_1} = \frac{2.009447^2}{181} = 0.0223087 \quad \widehat{\text{var}}(\bar{y}_0) = \frac{\hat{\sigma}_0^2}{N_0} = \frac{1.643745^2}{876} = 0.00308436$$

Then, the value of the test statistic is  $t = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_0^2}{N_0}}} = 10.876$ .

The degrees of freedom are

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_0^2}{N_0}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{\hat{\sigma}_0^2}{N_0}\right)^2}{N_0 - 1}} = \frac{(0.0223087 + 0.00308436)^2}{\frac{(0.0223087)^2}{180} + \frac{(0.00308436)^2}{875}} = 232$$

Because  $t = 10.876 > t_{(0.95, 232)} = 1.651$ , we reject  $H_0 : \mu_1 \leq \mu_0$  at the 5% significance level. We can conclude that men with more educated fathers have mean years of education that is greater than the mean number of years of education for those with less educated fathers.

Note from the output below, that the same conclusion would be reached if a pooled variance  $t$ -test is used. Also, the  $p$ -value in the output is for a two-tail test. It needs to be halved for a one-tail test (but, of course, half of zero is zero).

Test for Equality of Means of EDUC			
Categorized by values of PF16			
Sample: 1 1057			
Included observations: 1057			
Method	df	Value	Probability
t-test	1055	-12.40064	0.0000
Satterthwaite-Welch t-test*	232.3002	-10.87563	0.0000
Category Statistics			
PF16	Count	Mean	Std. Err. of Mean
0	876	12.39954	1.643745
1	181	14.13260	2.009447
All	1057	12.69631	1.831313

- (c) To investigate whether more highly educated men, those with more than 12 years of schooling, tend to marry more highly educated women, those with more than 12 years of schooling, we consider the sub-sample of observations where  $FATHEDUC > 12$ . Then, we examine whether the proportion of these observations where  $MOTHEDEC > 12$  is significantly greater than 0.5. The size of the sub-sample where  $FATHEDUC > 12$  is  $N = 292$ . The proportion of these observations where  $MOTHEDEC > 12$  is  $\hat{p} = 0.510274$ . Let  $p$  be the population proportion of more highly educated men that marry more highly educated women. Then, the null and alternative hypotheses are  $H_0 : p \leq 0.5$  and  $H_1 : p > 0.5$ . The value of the test statistic is

$$t = \frac{\hat{p} - 0.5}{\sqrt{\hat{p}(1 - \hat{p})/N}} = \frac{0.510274 - 0.5}{\sqrt{0.510274 \times 0.489726/292}} = 0.351$$

The 5% critical value is  $t_{(0.95, 291)} = 1.650$ . There is insufficient evidence to conclude that more highly educated men tend to marry more highly educated women. Since this is a one-tail test, the  $p$ -value in the output below needs to be halved.

Hypothesis Testing for FMGT12		
Sample: 1 1057 IF FATHEDUC>12		
Included observations: 292		
Test of Hypothesis: Mean = 0.500000		
Sample Mean = 0.510274		
Sample Std. Dev. = 0.500753		
Method	Value	Probability
t-statistic	0.350596	0.7261