Review 2 Answer Key

1.

Tableware.

a) Since there are 57 degrees of freedom, there were 59 different products in the analysis.

b) 84.5% of the variation in retail price is explained by the polishing time.

c) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (59 – 2) = 57 degrees of freedom. We will use a regression slope *t*-interval. For 95% confidence, use $t^*_{57} \approx 2.0025$, or estimate from the table $t^*_{50} \approx 2.009$.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = 2.49244 \pm (2.0025) \times 0.1416 \approx (2.21, 2.78)$$

d) We are 95% confident that the average price increases between \$2.21 and \$2.78 for each additional minute of polishing time.

2.

AP Statistics scores.

a) H_0: The distribution of AP Statistics scores at Ithaca High School is the same as it is nationally.
H_A: The distribution of AP Statistics scores at Ithaca High School is different than it is nationally.

Counted data condition: The data are counts.
Randomization condition: Assume that this group of students is representative of all years at Ithaca High School.
Expected cell frequency condition: The expected counts (shown in the table) are all greater than 5.

Under these conditions, the sampling distribution of the test statistic is χ^2 on 5 – 1 = 4 degrees of freedom. We will use a chi-square goodness-of-fit test.

Score	Observed	Expected	Residual = $(Obs - Exp)$	Standardized Residual = $\frac{(Obs - Exp)}{\sqrt{Exp}}$	Component = $\frac{(Obs - Exp)^2}{Exp}$
5	26	11.155	14.845	4.445	19.756
4	36	22.698	13.302	2.792	7.7955
3	19	24.153	– 5.153	– 1.049	1.0994
2	10	18.527	– 8.527	– 1.981	3.9245
1	6	20.467	– 14.47	– 3.198	10.226
					$\sum \approx 42.801$

$\chi^2 \approx 42.801$. Since the *P*-value is essentially 0, we reject the null hypothesis.

There is strong evidence that the distribution of scores at Ithaca High School is different than the national distribution. Students at IHS get fewer scores of 2 and 1 than expected, and more scores of 4 and 5 than expected.

b) H_0: Gender and AP Statistics score are independent at Ithaca High School.

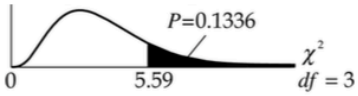
H_A: There is an association between gender and AP Statistics score at Ithaca High School.

Counted data condition: The data are counts.
Randomization condition: Assume this year's students are representative of all years.
Expected cell frequency condition: After combining the cells for scores of 2 and 1, the expected counts are all greater than 5.

	Boys (Obs / Exp)	Girls (Obs / Exp)
5	13 / 13.67	13 / 12.33
4	21 / 18.928	15 / 17.072
3	6 / 9.9897	13 / 9.0103
2 or 1	11 / 8.4124	5 / 7.5876

Under these conditions, the sampling distribution of the test statistic is χ^2 on 3 degrees of freedom. We will use a chi-square test for independence. (This is a test for independence, since we have one group that has been classified according to two variables, gender and score. However, if you said it was a test for homogeneity, since you were comparing two groups, no one would get terribly upset!)

With $\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} \approx 5.59$, the P-value ≈ 0.1336.



Since P-value ≈ 0.1336 is high, we fail to reject the null hypothesis. There is no evidence of an association between gender and score at Ithaca High School. The boys seem to do just as well as the girls.

3.

Twins.

H_0: There is no association between duration of pregnancy and level of prenatal care.

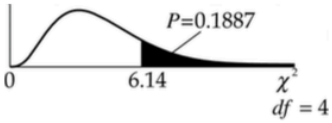
H_A: There is an association between duration of pregnancy and level of prenatal care.

Counted data condition: The data are counts.
Randomization condition: Assume that these pregnancies are representative of all twin births.
Expected cell frequency condition: The expected counts are all greater than 5.

	Preterm (induced or Cesarean) (Obs / Exp)	Preterm (without procedures) (Obs / Exp)	Term or postterm (Obs / Exp)
Intensive	18 /16.676	15 / 15.579	28 / 28.745
Adequate	46 / 42.101	43 / 39.331	65 / 72.568
Inadequate	12 / 17.223	13 / 16.090	38 / 29.687

Under these conditions, the sampling distribution of the test statistic is χ^2 on 4 degrees of freedom. We will use a chi-square test for independence.

$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} \approx 6.14$, and the P-value ≈ 0.1887.



Since the P-value ≈ 0.1887 is high, we fail to reject the null hypothesis. There is no evidence of an association between duration of pregnancy and level of prenatal care in twin births.

4.

Infliximab.

H_0: The remission rates are the same for the three groups.

H_A: The remission rates are different for the three groups.

Counted data condition: The data are counts.
Randomization condition: Assume that these patients are representative of all patients.
Expected cell frequency condition: The expected counts are all greater than 5.

	Placebo (Obs / Exp)	5 mg (Obs / Exp)	10 mg (Obs / Exp)
Remission	23 / 38.418	44 / 39.466	50 / 39.116
No Remission	87 / 71.582	69 / 73.534	62 / 72.884

Under these conditions, the sampling distribution of the test statistic is χ^2 on 2 degrees of freedom. We will use a chi-square test for homogeneity.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} \approx 14.96,$$

and the P-value ≈ 0.0006.

Since the P-value ≈ 0.0006 is low, we reject the null hypothesis. There is strong evidence that the remission rates are different in the three groups. Patients receiving 10 mg of Infliximab have higher remission rates than the other groups. These data indicate that continued treatment with Infliximab is of value to Crohn's disease patients who exhibit a positive initial response to the drug.

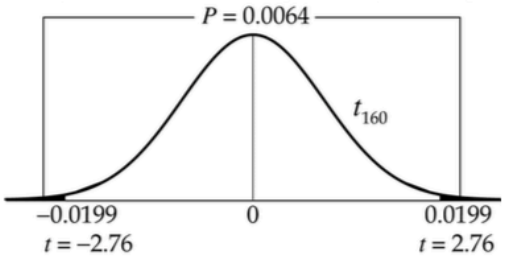
5.

Depression and the Internet.

a) H_0: There is no linear relationship between depression and Internet usage. $(\beta_1 = 0)$
H_A: There is a linear relationship between depression and Internet usage. $(\beta_1 \neq 0)$

Since the conditions for inference are satisfied (given), the sampling distribution of the regression slope can be modeled by a Student's t-model with (162 – 2) = 160 degrees of freedom. We will use a regression slope t-test. The equation of the line of best fit for these data points is: $\hat{Depression}After = 0.565485 + 0.019948(InternetUsage)$.

The value of $t \approx 2.76$. The P-value of 0.0064 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between depression and Internet usage. Those with high levels of Internet usage tend to have high levels of depression. It should be noted, however, that although the evidence is strong, the association is quite weak, with $R^2 = 4.6\%$. The regression analysis only explains 4.6% of the variation in depression level.



b) The study says nothing about causality, merely association. Furthermore, there are almost certainly other factors involved. In fact, if 4.6% of the variation in depression level is related to Internet usage, the other 95.4% of the variation must be related to something else!