**AP Statistics:**

**Chapter 12 Introduction**

Many people believe that students learn better if they sit closer to the front of the classroom. Does sitting closer *cause* higher achievement, or do better students simply choose to sit in the front? To investigate, an AP Statistics teacher randomly assigned students to seat locations in his classroom for a particular chapter and recorded the test score for each student at the end of the chapter. The explanatory variable in this experiment is which row the student was assigned (Row 1 is closest to the front and Row 7 is the farthest away). Do these data provide *convincing* evidence that sitting closer causes students to get higher grades?

Row 1: 76, 77, 94, 99

Row 2: 83, 85, 74, 79

Row 3: 90, 88, 68, 78

Row 4: 94, 72, 101, 70, 79

Row 5: 76, 65, 90, 67, 96

Row 6: 88, 79, 90, 83

Row 7: 79, 76, 77, 63



Predictor Coef SE Coef T P

Constant 85.706 4.239 20.22 0.000

Row -1.1171 0.9472 -1.18 0.248

S = 10.0673 R-Sq = 4.7% R-Sq(adj) = 1.3%

1. Describe the association shown in the scatterplot.
2. Using the computer output, determine the equation of the least-squares regression line.
3. Calculate the value of the correlation.
4. Calculate and interpret the residual for the student who sat in Row 1 and scored 76
5. Interpret the slope of the least-squares regression line.
6. Interpret the standard deviation of the residuals.
7. Interpret the value of ****.
8. Explain why it was important to randomly assign the students to seats rather than letting each student choose his or her own seat.
9. Does the negative slope provide convincing evidence that sitting closer causes higher achievement, or is it plausible that the association is due to the chance variation in the random assignment? Let’s do a simulation to find out!

 **12.1 Sampling Distribution of *b***

What is the difference between a sample regression line and population (true) regression line?

*.*

What is the sampling distribution of *b*? What shape, center, and spread does it have?

The sampling distribution will have these properties when the following regression model is valid.

* *Means are in a line*
* *SD of y values is constant for each x value*
* *Y values are normally distributed for each x value*

Suppose that the *true* regression line for the seating chart study is  = 87 – 1*x* with = 10 and that the regression model is valid. What is the probability that someone sitting in the second row will get at least 90 on the test?

Read 741–744

What are the conditions we must check to make sure that the regression model is valid and inference for regression is appropriate? How do you check them?

Verify that the conditions for inference are satisfied for the seating chart experiment.

For the seating chart experiment, state and interpret the standard error of the slope.

**HW: page 759 (1, 3, 5, 7a)**

**Confidence Intervals for **

Read 747–750

What is the formula for constructing a confidence interval for a slope? Where do you get the value of *t*\*? How many degrees of freedom should you use?

Alternate Example: *Fresh flowers?*

|  |  |
| --- | --- |
| Sugar(tbs.) | Freshness(hours) |
| 0 | 168 |
| 0 | 180 |
| 0 | 192 |
| 1 | 192 |
| 1 | 204 |
| 1 | 204 |
| 2 | 204 |
| 2 | 210 |
| 2 | 210 |
| 3 | 222 |
| 3 | 228 |
| 3 | 234 |

For their second-semester project, two AP Statistics students decided to investigate the effect of sugar on the life of cut flowers. They went to the local grocery store and randomly selected 12 carnations. All the carnations seemed equally healthy when they were selected. When the students got home, they prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared and placed in the same location, the students randomly assigned one flower to each vase and observed how many hours each flower continued to look fresh. Here are the data and computer output.

Predictor Coef SE Coef T P

Constant 181.200 3.635 49.84 0.000

Sugar (tbs) 15.200 1.943 7.82 0.000

S = 7.52596 R-Sq = 86.0% R-Sq(adj) = 84.5%

Construct and interpret a 99% confidence interval for the slope of the true regression line.

***State:*** *We want to estimate the slope  of the true regression line relating hours of freshness y to amount of sugar x at the 99% confidence level.*

***Plan:*** *If conditions are met, we will use a t interval for the slope to estimate .*

* *Linear The scatterplot shows a linear pattern, and there is no obvious leftover curvature in the residual plot, even though all the residuals are negative for x = 2.*
* *Independent All the flowers were in different vases, so knowing the hours of freshness for one flower shouldn’t provide additional information about the hours of freshness for other flowers.*
* *Normal The histogram of residuals does not show any skewness or outliers.*
* *Equal variance Although the variability of the residuals is not constant at each value of x, there is no systematic pattern such as increasing variation as x increases.*
* *Random Flowers were randomly assigned to the treatments.*

***Do:*** *Using**df = 12 − 2 = 10 the t critical value is t\* = 3.169. Thus, the 99% confidence interval is 15.2  3.169(1.943) = 15.2  6.16 = (9.04, 21.36).*

***Conclude:*** *We are 99% confident that the interval from 9.04 to 21.36 captures the slope of the true regression line relating hours of freshness y to amount of sugar x.*

**HW: page 759 (2, 4, 6, 8, 9, 11)**

 **12.1 Significance Tests for**

Read 751–754

What is the standardized test statistic for a significance test for the slope? Is this formula on the formula sheet? What degrees of freedom should you use?

What are the two explanations for the positive association in the crying and IQ example?

* *There really is an association between crying and later IQ*
* *There really is no association and the apparent association was due to sampling variability (remind them about the rossmanchance applet—when sampling from a population with no association we are bound to see some association just by chance)*

|  |  |
| --- | --- |
| **Time (minutes)** | **Tip (dollars)** |
| 23 | 5.00 |
| 39 | 2.75 |
| 44 | 7.75 |
| 55 | 5.00 |
| 61 | 7.00 |
| 65 | 8.88 |
| 67 | 9.01 |
| 70 | 5.00 |
| 74 | 7.29 |
| 85 | 7.50 |
| 90 | 6.00 |
| 99 | 6.50 |

Alternate Example: Do customers who stay longer at buffets give larger tips? Charlotte, an AP statistics student who worked at an Asian buffet, decided to investigate this question for her second semester project. While she was doing her job as a hostess, she obtained a random sample of receipts, which included the length of time (in minutes) the party was in the restaurant and the amount of the tip (in dollars). Do these data provide convincing evidence that customers who stay longer give larger tips?

(a) Here is a scatterplot of the data with the least-squares regression line added. Describe what this graph tells you about the relationship between the two variables.

More Minitab output from a linear regression analysis on these data is shown below.

Predictor Coef SE Coef T P

Constant 4.535 1.657 2.74 0.021

Time (minutes) 0.03013 0.02448 1.23 0.247

S = 1.77931 R-Sq = 13.2% R-Sq(adj) = 4.5%

 

(b) What is the equation of the least-squares regression line for predicting the amount of the tip from the length of the stay? Define any variables you use.

(c) Interpret the slope and *y* intercept of the least-squares regression line in context.

(d) Carry out an appropriate test to answer Charlotte’s question.

***State:*** *We want to perform a test of :  = 0 vs.:  > 0 where  is the true slope of the population regression line relating length of stay x to tip amount y. We will use  = 0.05.*

***Plan:*** *If the conditions are met we will do a t test for the slope .*

*Linear The scatterplot shows a weak, positive, linear relationship between length of stay and tip amount. The residual plot looks randomly scattered about the residual = 0 line.*

*Independent Knowing one tip amount shouldn’t provide additional information about other tip amounts. Since we are sampling without replacement, we must assume that there are more than 10(12) = 120 receipts in the population from which these receipts were selected.*

*Normal The Normal probability plot looks roughly linear.*

*Equal variance The residual plot shows a fairly equal amount of scatter around the residual = 0 line.*

*Random The receipts were randomly selected.*

*With no obvious violations, we proceed to inference.*

***Do:***  *From the computer output:*

*Test statistic t = 1.23*

*P-value Since the P-value in the computer output is for a two-sided test, we must cut it in half for a one-sided test. P = 0.247/2 = 0.1235 using df = 12 – 2 = 10.*

***Conclude:*** *Since the P-value is larger than 0.05, we fail to reject . We do not have convincing evidence that parties who stay longer at buffets leave larger tips.*

 *Note: Computer printout always gives 2-sided P-values.*

Read 756–757

Can you use your calculator to conduct a test for the slope? What about a confidence interval?

**HW: page 761 (13, 15, 19)**