RLP-STATSA · RLP Statistics A — Exploring Data, Study Design and Probability

Unit 2: Two-Variable Data and Regression

Practice Workbook. The same problems, in the same order, as the Practice sections of the online lessons.

Name _____ Period _____

Lesson 2.1 — Scatterplots and Correlation

FLUENCY

1. Use technology to find the least-squares line and the correlation coefficient for (3, 9), (4, 12), (9, 20), (10, 20), (12, 24).

2. Use technology to find the least-squares line and the correlation coefficient for (3, 10), (5, 18), (6, 19), (12, 28), (13, 31).

3. Use technology to find the least-squares line and the correlation coefficient for (4, 7), (5, 10), (6, 8), (8, 12), (9, 13), (14, 24).

4. A scatterplot shows that as x increases, y tends to decrease, with points close to a straight line. Describe the direction, form and strength.

5. Which value of r shows the strongest linear association: 0.62, −0.91, 0.05 or 0.88?

6. Every y-value is converted from inches to centimeters. What happens to r?

7. True or false: if r = 0, there is no relationship of any kind between x and y.

8. You swap the x and y variables. What happens to r?

APPLICATION

9. Precipitation (inches) at Kansas City and Springfield, 2019-2023 (NOAA): Kansas City 53.22, 34.10, 41.26, 34.57, 37.40; Springfield 54.49, 49.71, 46.97, 43.28, 43.08. Compute r by hand from z-scores, then check with technology.

10. Over all 52 years (1973-2024), r = 0.452 for Kansas City against Springfield precipitation. Interpret it in context and explain why it is positive.

11. Kansas City precipitation against year, 1973-2024, has r = 0.050. A student says 'Kansas City's rain doesn't vary.' Correct the statement.

CHALLENGE

12. Compute r for the points (−3, 9), (−2, 4), (−1, 1), (0, 0), (1, 1), (2, 4), (3, 9). What does the result show?

REVIEW

13. For the data 3, 15, 15, 20, 21, 31, 32: find the mean, median, range and interquartile range. (Find the quartiles as the medians of the lower and upper halves, leaving out the overall median when the count is odd.)

14. (Unit 1) St. Louis mean annual temperature, 1950-2024 (NOAA), has mean 56.64°F and SD 1.88°F. The 2012 value was 61.2°F: find its z-score.

Lesson 2.2 — Least-Squares Regression

FLUENCY

1. Use technology to find the least-squares line and the correlation coefficient for (1, 7), (2, 11), (7, 12), (11, 16), (12, 18), (13, 16), (14, 17).

2. Use technology to find the least-squares line and the correlation coefficient for (1, 6), (2, 10), (3, 12), (4, 12), (5, 12), (9, 21), (10, 23).

3. Use technology to find the least-squares line and the correlation coefficient for (1, 11), (3, 13), (6, 20), (7, 24), (9, 25), (12, 31), (14, 33).

4. Use technology to find the least-squares line and the correlation coefficient for (4, 22), (8, 32), (9, 39), (10, 40), (11, 44), (12, 44), (13, 51).

5. For Missouri's age-adjusted cancer death rate, 1999-2017 (CDC), with x = years since 1999: r = −0.9897, s_x = 5.627, s_y = 13.376, $\bar{x} = 9$, $\bar{y} = 189.484$. Find the slope and intercept of the least-squares line.

6. Which point does every least-squares line pass through?

7. A line has $r^2 = 0.64$. What is r if the slope is negative?

APPLICATION

8. Interpret the slope (−2.353) and intercept (210.66) of the cancer model in context.

9. Use the cancer model to predict the 2010 rate and find the residual (the actual 2010 rate was 185.6).

10. Regressing Columbia's annual precipitation on St. Louis's (1973-2024) gives predicted COU = 9.897 + 0.766 × STL, with r = 0.777. Predict Columbia's total in a year when St. Louis gets 50 inches, and interpret r^2.

CHALLENGE

11. Regressing St. Louis on Columbia instead gives slope 0.789. Show that this is not 1/0.766, and explain why the two lines differ.

REVIEW

12. (Unit 1) For the data 12, 15, 15, 18, 30, find Q1, Q3 and the upper 1.5 × IQR fence. Is 30 an outlier?

13. (Unit 1) Which is more resistant to an outlier, the mean or the median?

Lesson 2.3 — Residuals, Influence and Causation

FLUENCY

1. The heart-disease model is predicted rate = 282.43 − 6.12x (x = years since 1999). The actual 2005 rate was 236.8. Find the residual.

2. A residual plot shows a random scatter around 0 with no pattern. What does it suggest?

3. A residual plot is U-shaped. What does it suggest?

4. Classify the explanation: ice-cream sales and drownings rise together because both increase in hot weather.

5. Classify: students who choose to take an SAT prep course score higher, but they also tend to be more motivated, so the effects cannot be separated.

6. Which points tend to be most influential in a regression: those with extreme x-values or extreme y-values near $\bar{x}$?

APPLICATION

7. A line fitted to Missouri's unintentional-injury death rate, 1999-2017 (CDC), has r = 0.809. Its residuals are positive in 1999, negative 2000-2001, positive 2002-2008, negative 2009-2010, positive in 2011, negative 2012-2015, and positive 2016-2017. What should the report conclude about the model?

8. Kansas City precipitation against year, 1973-2024, has slope 0.274 inches per decade. Refit without 2001 (53.51 in, a wet year near the middle of the range): the slope becomes 0.242 inches per decade. Refit instead without 1973 (55.31 in, the first year): the slope becomes 0.681 inches per decade. Which point is influential? Explain.

9. A news story reports that Missouri counties with more hospitals have more deaths each year, and suggests hospitals are dangerous. Name the lurking variable and explain the association.

CHALLENGE

10. Missouri's age-adjusted heart-disease death rate fell from 300.5 in 1999 to 191.1 in 2017, with r = −0.948 against year. Explain why 'year' cannot be the cause, and describe what kind of evidence could identify a cause.

REVIEW

11. For the data 10, 15, 27, 29, 32, 34, 36: find the mean, median, range and interquartile range. (Find the quartiles as the medians of the lower and upper halves, leaving out the overall median when the count is odd.)

12. (Unit 1) Convert an SD of 8.40 inches (St. Louis annual precipitation, 1973-2024) to millimeters.

Lesson 2.4 — Exponential Models and the Limits of Extrapolation

FLUENCY

1. A quantity starts at 15,000 and grows by 8% each year. Write a model and find its value after 9 years, to the nearest whole number.

2. A quantity starts at 15,000 and grows by 12% each year. Write a model and find its value after 5 years, to the nearest whole number.

3. A quantity starts at 1,200 and grows by 12% each year. Write a model and find its value after 12 years, to the nearest whole number.

4. Evaluate: $\log_3 81$

5. Evaluate: $\log_2 (1/4)$

6. A log fit gives $\log_{10} y = 2.5 + 0.02x$. Write the exponential model $y = a \cdot b^x$.

7. An exponential model has growth factor 1.03 per year. Find the doubling time.

APPLICATION

8. Missouri's census population (millions): 1970 4.677, 1980 4.917, 1990 5.117, 2000 5.595, 2010 5.989, 2020 6.155. Fit an exponential model by regressing $\log_{10} P$ on t = years since 1970 (P in people). Report the annual growth rate.

9. Compare that rate with the 1910-1970 model (about 0.58% per year) and with the same two fits for the whole United States, and say what the comparison shows.

10. A linear fit to the same 1970-2020 Missouri data gives residuals (millions) of +0.060, −0.016, −0.133, +0.029, +0.106, −0.045. What do they suggest?

CHALLENGE

11. Use your 1970-2020 exponential model to predict 2050, then argue whether the prediction should be published.

REVIEW

12. (Unit 1) A data set has mean 40 and SD 5. Every value is multiplied by 2 and then 10 is added. Find the new mean and SD.

13. (Lesson 2.2) State the formula for the least-squares slope in terms of r and the SDs.

Unit 2 checkpoint

1. Which r indicates the weakest linear association?
   A. −0.85
   B. 0.12
   C. 0.67
   D. −0.40
2. True or false: If r = 0.777, then 77.7% of the variation in y is explained by x.
3. With r = 0.8, s_y = 10 and s_x = 4, the least-squares slope is _____.
4. The model predicts the death rate falls 6.12 per 100,000 per year. Which is the best interpretation?
   A. The rate falls exactly 6.12 each year
   B. The model predicts an average decrease of about 6.1 deaths per 100,000 per year
   C. 6.12 people die each year
   D. The rate is 6.12
5. A point has actual y = 201.8 and predicted y = 215.1; its residual is _____.
6. A residual plot is curved. What should you do?
   A. Report the line
   B. Try a different form, such as exponential via logs
   C. Delete points
   D. Swap x and y
7. True or false: A strong correlation between two variables shows that one causes the other.
8. Which point is most likely to be influential?
   A. One near $\bar{x}$ with a large residual
   B. One with an extreme x-value far from the line
   C. One on the line near $\bar{x}$
   D. Any point with y = 0
9. If $\log_{10} P = 6.51 + 0.0025t$, the annual growth factor is about _____.
10. True or false: A model with r = 0.9894 can safely be used to predict 50 years beyond its data.
11. St. Louis and Columbia rainfall correlate at r = 0.777 because…
    A. St. Louis rain causes Columbia rain
    B. The same weather systems reach both (common response)
    C. Of coincidence
    D. Columbia rain causes St. Louis rain
12. True or false: The least-squares line always passes through $(\bar{x}, \bar{y})$.