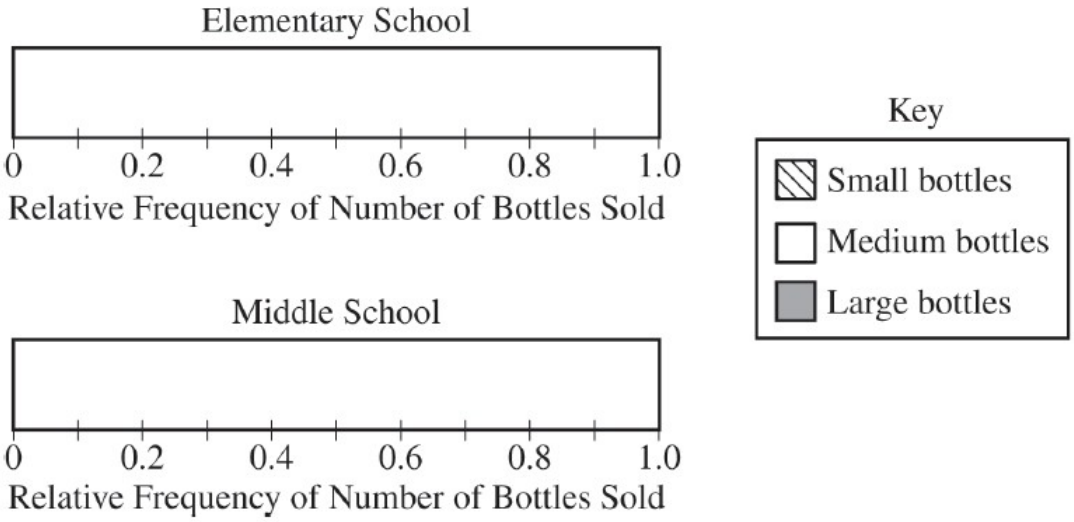
• A local elementary school decided to sell bottles printed with the school district's logo as a fund-raiser. The students in the elementary school were asked to sell bottles in three different sizes (small, medium, and large). The relative frequencies of the number of bottles sold for each size by the elementary school were 0.5 for small bottles, 0.3 for medium bottles, and 0.2 for large bottles.

A local middle school also decided to sell bottles as a fund-raiser, using the same three sizes (small, medium, and large). The middle school students sold three times the number of bottles that the elementary school students sold. For the middle school students, the proportion of bottles sold was equal for all three sizes.

(A) Complete the segmented bar graphs representing the relative frequencies of the number of bottles sold for each size by students at each school.

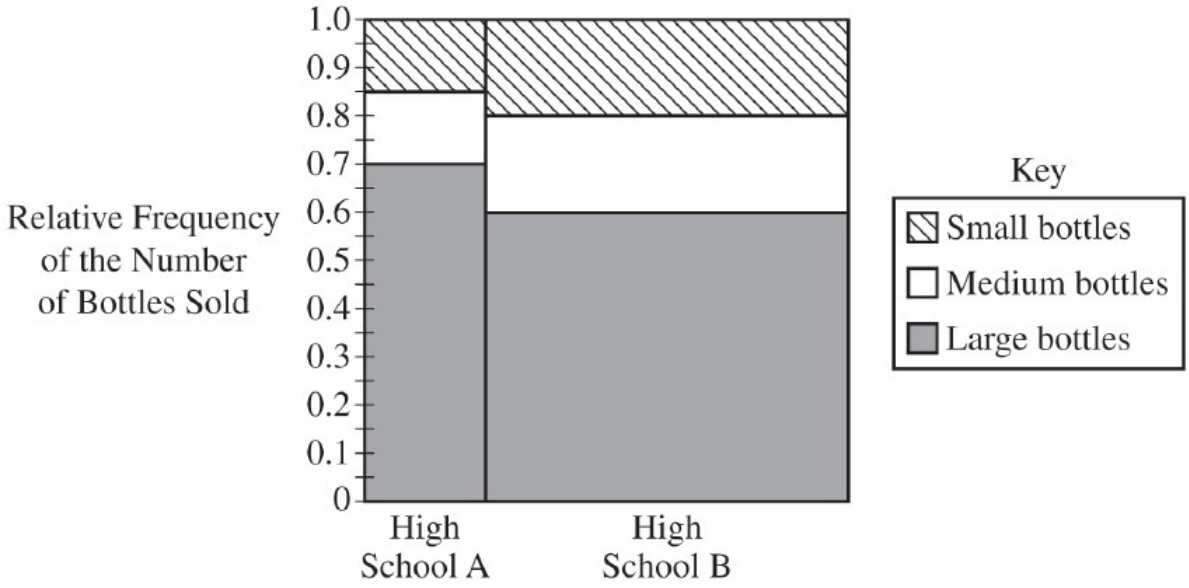


(B) An administrator at the elementary school concluded that the elementary school students sold more small bottles than the middle school students did. Is the elementary school administrator's conclusion correct? Explain your response.

Two high schools are also selling the bottles and are competing to see which one sold more large bottles.

(C) A mosaic plot for the distribution of the number of bottles sold by each of the high schools is shown here.

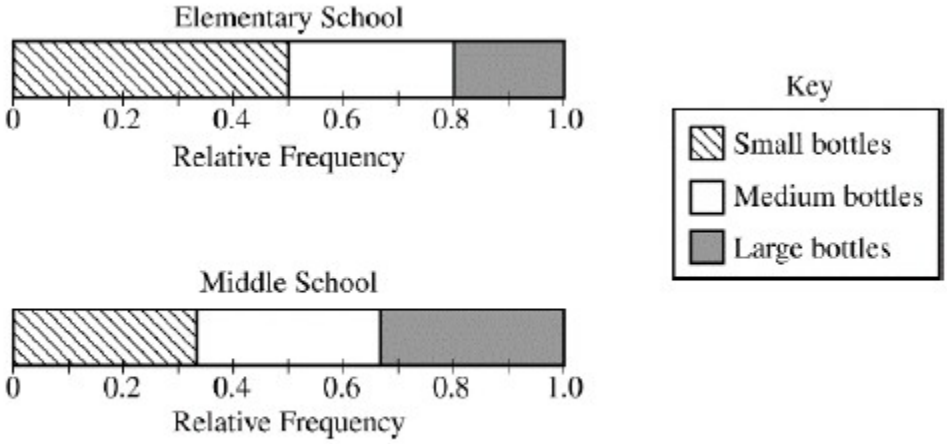
Distribution of the Number of Bottles Sold by High School



(i) Which of the two high schools sold a greater proportion of large bottles? Justify your answer.

(ii) Which of the two high schools sold a greater number of large bottles? Justify your answer.

Solution:
(A)



(B) No, the segment for small bottles for the elementary school is wider than the segment for small bottles for the middle school; however, the middle school students sold three times as many bottles as the elementary school students. So, if the elementary school sold x number of bottles, the middle school sold $3x$ number of bottles.

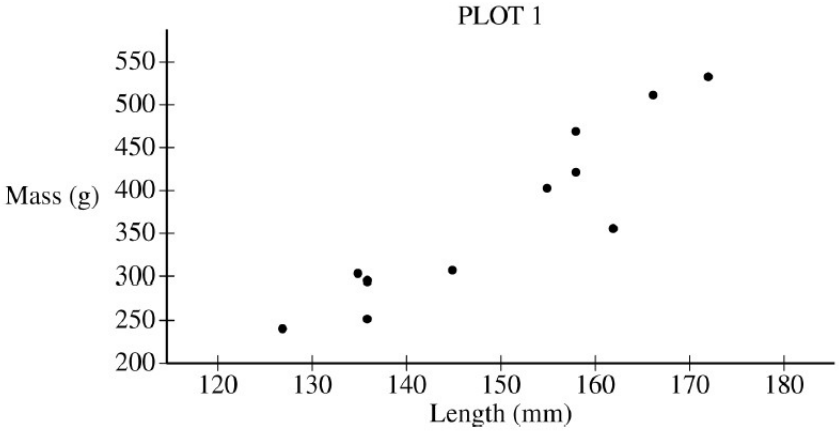
For example, if the elementary students sold 100 bottles total, then they sold 0.5(100) or 50 small bottles. However, because the middle school students sold three times the total number of bottles as the elementary students, they would have sold 300 bottles total and $(0.\overline{3})(300) = 100$ small bottles. Because $100 > 50$, the middle school sold more small bottles, and the elementary school's administrator is not correct.

(C)

(i) The mosaic plot shows that the proportion of large bottles sold by High School A was 0.7 and the proportion of large bottles sold by High School B was 0.6. High School A sold a greater proportion of large bottles because $0.7 > 0.6$.

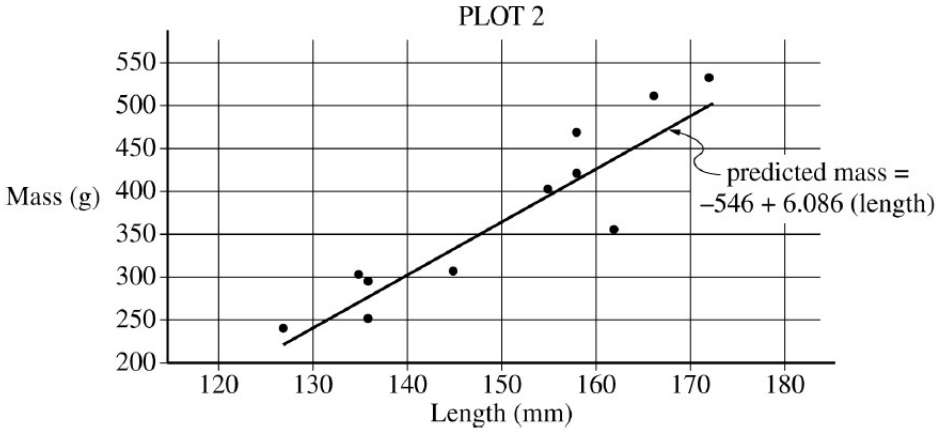
(ii) The number of bottles sold is represented by the area of the shaded region. The area of the rectangle representing large bottles sold by High School B is clearly larger than the area of the rectangle representing large bottles sold by High School A. Therefore, High School B sold more large bottles than High School A.

• A biologist gathered data on the length, in millimeters mm, and the mass, in grams g, for 11 bullfrogs. The data are shown in Plot 1.



(A) Based on the scatterplot, describe the relationship between mass and length, in context.

From the data, the biologist calculated the least-squares regression line for predicting mass from length. The least-squares regression line is shown in Plot 2.



(B) Identify and interpret the slope of the least-squares regression line in context.

(C) Interpret the coefficient of determination of the least-squares regression line, $r^2 \approx 0.819$, in context.

(D) From Plot 2, consider the residuals of the 11 bullfrogs.

(i) Based on the plot, approximately what is the length and mass of the bullfrog with the largest absolute value residual?

(ii) Does the least-squares regression line overestimate or underestimate the mass of the bullfrog identified in part (d-i)? Explain your answer.

Solutions:
(A) The scatterplot reveals a strong, positive, roughly linear association between the mass and length of bullfrogs. There are no points that seriously deviate from the straight-line pattern of the points in the plot.

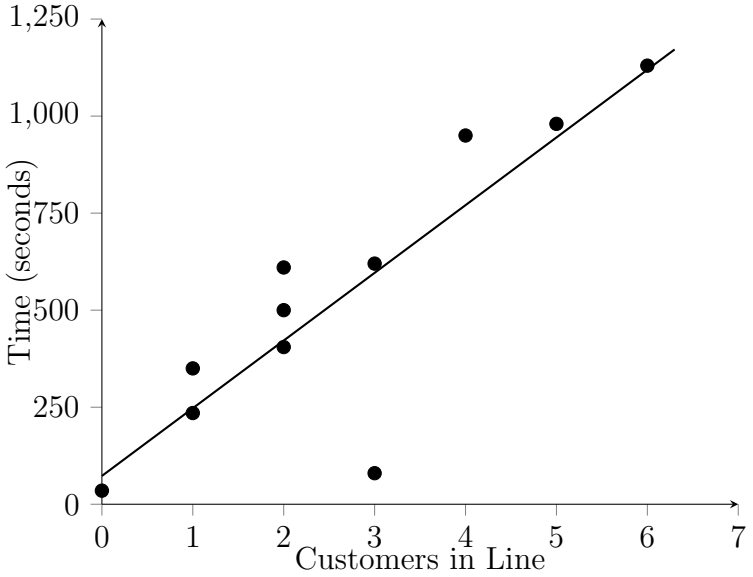
(B) The value of the slope of the least-squares regression line is 6.086. This value indicates that the predicted mass of a bullfrog increases by 6.086 grams for each additional millimeter of length.

(C) The coefficient of determination is $r^2 \simeq 0.819$. This value indicates that 81.9% of the variation in bullfrog mass can be explained by variation in bullfrog length as described by the least-squares line.

(D)
(i) The largest residual in absolute value belongs to the bullfrog with length 162 millimeters and mass 356 grams.

(ii) The least-squares regression line overestimates the mass of the bullfrog with length 162 millimeters. Plot 2 shows that the point for the bullfrog with length 162 millimeters is below the least-squares regression line.

• The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%		R-Sq (adj) = 70.37%

(A) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

(B) Identify and interpret in context the coefficient of determination, r^2.

(C) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

Solution

(A) The estimate of the intercept is 72.95. It is estimated that the average time to finish checkout if there are no other customers in line is 72.95 seconds.

(B) The coefficient of determination is $r^2 = 73.33\%$. This value indicates that 73.33% of the variability in the times it takes customers to finish checkout, including time waiting in line, can be explained by knowing how many customers are in line in front of the selected customer.

(C) The outlier is the point with $x = 3$ and y close to 0. This point is considered an outlier because the combination of x and y values differs from the pattern of the rest of the data. Specifically, the value of y (time to finish checkout) is much lower than would be expected when there are $x = 3$ customers in line in front of the selected customer, given the remaining data.

• Researchers studying a pack of gray wolves in North America collected data on the length x, in meters, from nose to tip of tail, and the weight y, in kilograms, of the wolves. A scatterplot of weight versus length revealed a relationship between the two variables described as positive, linear, and strong.

(A) For the situation described above, explain what is meant by each of the following words.

(i) Positive:

(ii) Linear:

(iii) Strong:

The data collected from the wolves were used to create the least-squares equation $\hat{y} = -16.46 + 35.02x$.

(B) Interpret the meaning of the slope of the least-squares regression line in context.

(C) One wolf in the pack with a length of 1.4 meters had a residual of -9.67 kilograms. What was the weight of the wolf?

The primary goals of this question were to assess a student's ability to (1) explain statistical terms used when describing the relationship between two variables; (2) interpret the slope of a linear regression equation; and (3) calculate a value of y when given a regression equation, a value of x, and a residual.

Solution

(A) In the context of a scatterplot in which y represents weight and x represents length, the following are defined.

A positive relationship means that wolves with higher values of length also tend to have higher weights.

A linear relationship means that as length increases by one meter, weight tends to change by a constant amount, on average.

A strong relationship means that the data points fall close to a line (or curve).

(B) The slope of 35.02 indicates that two wolves that differ by one meter in length are predicted to differ by 35.02 kilograms in weight, with the longer wolf having the greater weight.

(C) In general, a residual is equal to actual weight minus predicted weight, or equivalently,

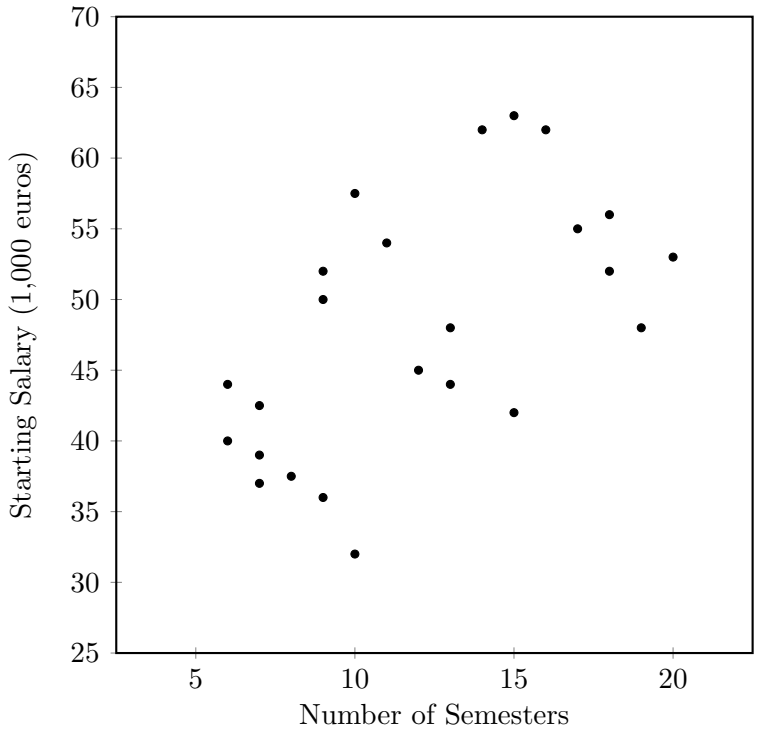
$$\text{actual weight} = \text{predicted weight} + \text{residual}.$$

For the wolf with length 1.4 meters and residual of -9.67, the predicted weight is

$$-16.46 + 35.02(1.4) = 32.568 \text{ kilograms}.$$

Therefore, the actual weight of the wolf is $32.568 + (-9.67) = 22.898$ kilograms.

• A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program. Information was collected on the number of semesters each person in the sample needed to complete the program and the starting salary, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.



(A) Does the scatterplot support the newspaper report about number of semesters and starting salary? Justify your answer.

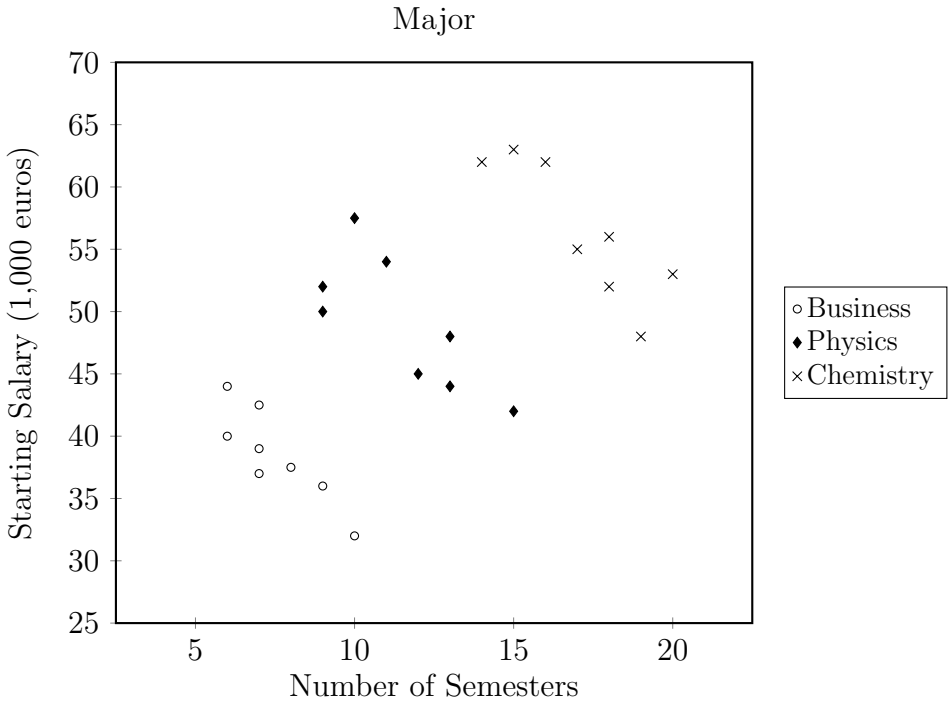
The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003

S = 7.37702 R-Sq = 33.5% R-Sq(adj) = 30.5%

(B) Identify the slope of the least-squares regression line, and interpret the slope in context.

An independent researcher received the data from the newspaper and conducted a new analysis by separating the data into three groups based on the major of each person. A revised scatterplot identifying the major of each person is shown below.



(C) Based on the people in the sample, describe the association between starting salary and number of semesters for the business majors.

(D) Based on the people in the sample, compare the median starting salaries for the three majors.

(E) Based on the analysis conducted by the independent researcher, how could the newspaper report be modified to give a better description of the relationship between the number of semesters and the starting salary for the people in the sample?

Solution

(A) The scatterplot supports the newspaper report about number of semesters needed to complete an academic program and starting salary because it shows a positive association between these two variables.

(B) The slope is 1.1594. For each additional semester needed to complete an academic program, the predicted starting salary increases by €1,159.40.

(C)For the business majors alone, there is a strong, negative, linear association between number of semesters and starting salary. Business majors who need a greater number of semesters to complete an academic program tend to have lower starting salaries.

(D) Business majors have the lowest median starting salary at around €38,000, followed by physics majors at around €48,000, and then chemistry majors with the highest median starting salary at around €55,000.

E The newspaper report should be modified to account for major. Overall, majors that take longer to complete tend to have higher starting salaries, with chemistry the highest, physics the next highest, and business the lowest. However, within a major, students who take a greater number of semesters tend to have lower starting salaries.

Problems adapted from the College Board released tests.