AP Statistics – Unit 2
Exploring Two-Variable Data

2.1 Introducing Statistics: Relationships Between Variables

Often, data will involve several variables that may be related to one another. Here we want to begin discussing relationships between two variables.

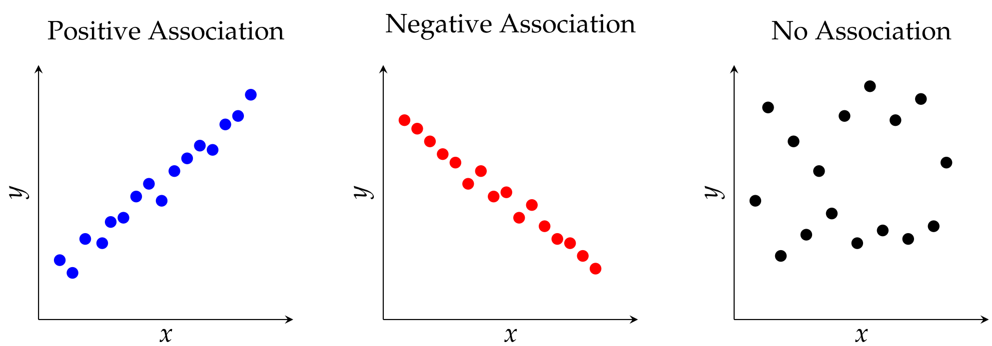
Key Terms

When dealing with two-variable data, we denote an **explanatory variable** x and a **response variable** y.

Key Terms

Two variables may be related in the following ways:

- **Positive association**: As x increases, y increases too.
- **Negative association**: As x increases, y decreases.
- **No association**: Changes in x do not systematically affect y.



Types of Association Between Two Variables

2.2 Constructing and Interpreting Scatterplots

Scatterplots display two quantitative variables measured on the same individuals.
To describe a scatterplot:

- **Direction**: Positive, negative, or no association
- **Form**: Linear, curved, clusters, etc.
- **Strength**: Strong, moderate, or weak association
- **Outliers**: Points that fall outside the general pattern

Interpretation: "There is a strong/moderate/weak positive/negative linear relationship between (explanatory variable) and (response variable)."

Example 0.1. Example: Describe the association shown in a scatterplot of study hours (x) vs. exam scores (y).
Solution: Positive, linear, moderately strong association with one outlier.

2.3 Correlation

The **correlation coefficient** r measures the strength and direction of a linear relationship between two quantitative variables.

- r is between -1 and 1
- $r > 0$: Positive association, $r < 0$: Negative association
- The closer r is to ± 1, the stronger the linear relationship

Important: Correlation **does not imply causation**.

Interpretation: "The correlation r shows/confirms that there is a strong/moderate/weak linear relationship between (explanatory variable) and (response variable)."

Example 0.2. Example: A correlation of $r = 0.85$ suggests a strong positive linear relationship.

2.4 Least Squares Regression Lines (LSRL)

The **LSRL** predicts values of the response variable y given an explanatory variable x.
Equation:

$$\hat{y} = a + bx$$

where:

- $b = r \times \frac{s_y}{s_x}$ (slope)
- $a = \bar{y} - b\bar{x}$ (y-intercept)

Interpretation of Slope: For each additional unit increase in x, the predicted y increases/decreases by b units.
Interpretation of Intercept: The predicted y when $x = 0$.

2.5 Residuals

A **residual** is the difference between an observed y and the predicted $\hat{y}$:

$$\text{Residual} = y - \hat{y}$$

- Positive residual: Actual y is above the predicted $\hat{y}$
- Negative residual: Actual y is below the predicted $\hat{y}$

Residual plots help assess whether a linear model is appropriate.

2.6 Assessing the Fit: r^2 and Standard Deviation of Residuals

Coefficient of Determination (r^2) tells the percent of the variation in y explained by the model.
Standard Deviation of Residuals (s) measures the typical distance between the observed y-values and the predicted $\hat{y}$-values.

Example 0.3. Example: If $r^2 = 0.72$, then 72% of the variation in y is explained by the linear model relating x to y.

Practice Problems

1. Suppose the correlation between hours of exercise and weight loss is $r = -0.64$.
   - (a) Describe the direction and strength.
   - (b) Is weight loss caused by exercise based on this information alone?

   Solution:

   - (a) The relationship is moderately strong and negative.
   - (b) No, correlation does not imply causation.

2. A least squares regression line for predicting exam score (y) from study hours (x) is $\hat{y} = 50 + 5x$.
   - (a) Interpret the slope.
   - (b) Predict the score for someone who studies 6 hours.

   Solution:

   - (a) Each additional hour of study is associated with a predicted increase of 5 points.
   - (b) $\hat{y} = 50 + 5(6) = 80$.

3. Given the residual plot below shows a clear curved pattern, what conclusion can you draw?

   Solution: The relationship between x and y is not linear; a different model may be more appropriate.