10.1 Review and Preview

*Focus of this chapter is to analyze paired sample data:

→ Determine whether a correlation, or association, between 2 variables exists & whether the correlation is linear

→ For linear correlations, we can identify an equation of a straight line that best fits the data

→ We can use that equation to predict the value of one variable given the value of the other variable

10.2 Correlation

Bivariate data:

- Ex: (4, 2)
- Ex: paired sample data consisting of durations (x) and time intervals after eruptions (y) of the Old Faithful geyser

Part I: Basic Concepts of Correlation

Correlation:

Scatterplot:

- *Examining scatterplots for patterns:*
  - Uphill direction: as one variable increases, the other also increases
  - Downhill direction: as one variable increases, the other decreases
  - Outliers
  - As the pattern becomes closer to a straight line, the relationship between x and y becomes stronger

- Ex #1: Using the paired duration and time interval after eruption data from the following table, to graph a scatter plot. Then describe the pattern of the relationship between the duration and time intervals.

Duration	240	120	178	234	235	269	255	220
Interval after	92	65	72	94	83	94	101	87

Linear correlation coefficient:

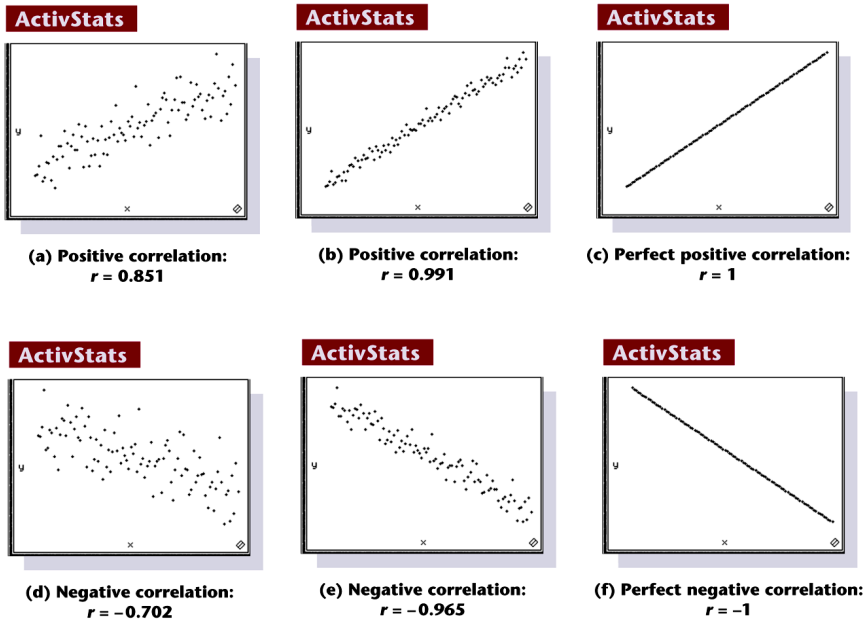
- Notation: r
- Its value is computed using the following formula:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

- **Important Notation:**

n	the number of pairs of data present
Σ	the addition of the items indicated
Σx	the sum of all x values
Σx^2	each x value should be squared and then add those squares
$(\Sigma x)^2$	the x values should be added and then find the sum squared
Σxy	each x value should be multiplied by its corresponding y value and then add those products
r	the linear correlation coefficient for a sample
ρ	Greek letter rho used to represent the linear correlation coefficient for a population

- It is a sample statistic that is used to measure the strength of the linear correlation between x and y. If we had every pair of population values for x and y, the result of r would be a population parameter, represented by ρ (Greek rho)



- **Requirements:** Given any collection of sample paired data, the linear coefficient r can always be computed, but the following requirements should be satisfied when testing hypothesis or making other inferences about r.
    1. The sample of paired (x, y) data is a random sample of independent quantitative data
    2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern
    3. Any outliers must be removed if they are known to be errors. The effects of any outliers should be considered by calculating r with and without the outliers included.
    4. The pairs of (x, y) data must have a bivariate normal distribution (for any fixed value of x, the corresponding values of y have a bell-shaped distribution and for any fixed value of y, the corresponding x values also have a bell-shaped distribution)

- Round the linear correlation coefficient to _____ so that its value can be directly compared to critical values in table A-6

- *To use a graphing calculator to find the linear correlation coefficient r:*
  - Press STAT then 1: Edit
  - Enter the lists of values into L1 and L2
  - Press STAT then select CALC
  - Choose 4: LinReg
  - The Xlist should be L1 and the Ylist should be L2 (leave freqlist & store blank)
  - Put the cursor on CALCULATE, press ENTER
  - Scroll down and you'll see a value given for r
  - If r does not appear, go to MODE to make sure the Stat Diagnostics are set to ON

  - Ex #1: Using the following simple random sample of data given, find the value of the linear correlation coefficient r:

    --Requirements:

x	3	1	3	5
y	5	8	6	4

r = _____

- *Interpreting the linear correlation coefficient r:*
  - The value of r must always fall between -1 and +1

  - General Interpretation:
    - If r is close to zero, we conclude that there is NO linear correlation between x and y

    - If r is close to -1 or +1, we conclude that there IS a linear correlation between x and y

  - *Using Table A-6*:
    - Correlation→ If the computed linear correlation coefficient lies in the left tail beyond the leftmost critical value or if it lies in the right tail beyond the rightmost critical value, conclude that _____
      _____

    - No Correlation→ If the computed linear correlation coefficient r lies between the two critical values, conclude that _____
      _____

- *Properties of the Linear Correlation Coefficient r:*
    - The value of r is always between -1 and + 1
    - The value of r does not change if all values of either variable are converted to a different scale
    - The value of r is not affected by the choice of x or y. Interchange all x and y values and the value of r will not change
    - r measures _____ It is not designed to measure the strength of a relationship that is not linear.

Ex #1: Use the previous example, and a significance level of 0.05 and table A-6 to determine if there is sufficient evidence to support the claim of a linear correlation:

♦ Ex #2: Using the paired duration and time interval after eruption data from the following table, find the value of the linear correlation coefficient r. Then refer to table A-6 to determine whether there is a linear correlation between the duration times and the time intervals after eruptions. In table A-6, use the critical value for $\alpha = 0.05$.

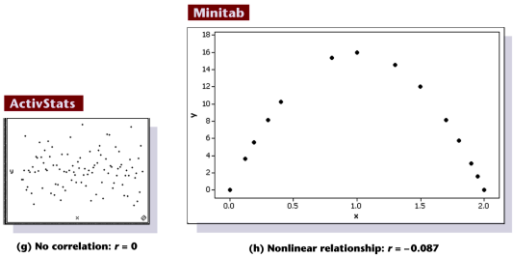
Duration	240	120	178	234	235	269	255	220
Interval after	92	65	72	94	83	94	101	87

→ *Common Errors involving Correlation:*

- A common error is to conclude that correlation implies causality.
  --From the previous example, we can conclude that there is a correlation between the duration times and the interval after eruption times, but we cannot conclude that longer duration times CAUSE longer interval after eruption times
  --outside air temperature may affect the duration of an eruption and the time interval after an eruption (outside air temp. would be called a lurking variable)

- Data based on averages suppress individual variation and may inflate the correlation coefficient
  --One study produced a 0.4 linear correlation coefficient for bi-variate data relating income and education among individuals, but the linear correlation coefficient was 0.7 when regional averages were used.

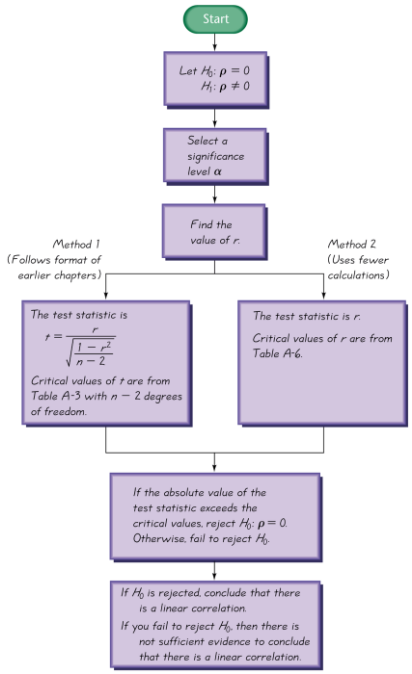
- A relationship may exist between x and y even when there is no linear correlation (could be a curved or exponential pattern)



Part II: Formal Hypothesis Test

Hypothesis Test for Correlation:

- H_0: $\rho = 0$ (there is _____)
- H_1: $\rho \neq 0$ (there is _____)



- *Critical Value Method (#2): Test statistic is r*
  - Test statistic: r (use graphing calculator)
  - Critical values: refer to table A-6
  - Conclusion: If $|r| >$ critical value from table A-6, reject H_0 and conclude that there is a linear correlation
  - Conclusion: If $|r| \leq$ critical value, fail to reject H_0: there is not sufficient evidence to conclude that there is a linear correlation

- Example #1: Using the previous example about duration and interval after eruption times, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption.

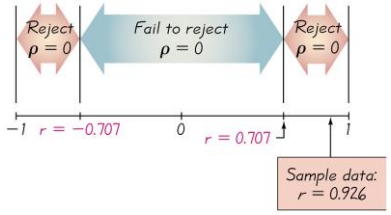
  - H_0:

  - H_1:

  - No significance level was specified, so use α = _____

  - Previously found that r = _____

  - Using table A-6 with $\alpha = 0.05$ and n = 8, the critical values of r are _____



  Since 0.926 falls in the critical region, we _____

  - Conclusion:

- Example #2: Use the following data to test the claim that there is a linear correlation between shoe print lengths and heights of males. Use a significance level of 0.01 to test the claim.

Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

H_0:

H_1:

α = _____

The test statistic is r = _____

Using table A-6 with α = 0.01 and n = 5, the critical values of r are _____

Since 0.591 does not fall in the critical region, we _____

Conclusion:

10.3 Regression

Part I: Basic Concepts of Regression

Deterministic: relationship between 2 variables meaning that

Ex: the total cost y of an item with a list price of x and a sales tax of 5% can be found using the deterministic equation: $y = 1.05x$

Probabilistic: relationship between 2 variables meaning that

Ex: a child's height is not determined completely by the height of the father (or mother)

Regression Equation:

$b_0 =$
$b_1 =$

where b_0 and b_1 are sample statistics that are used to estimate the population parameters β_0 and β_1

- *Regression line:*
  - Also called: the line of best fit or the least-squares line
- x is called _____
- y is called _____
- From algebra, you may remember the general equation of a line is

- *Requirements:*
    1. The sample of paired (x, y) data is a random sample of quantitative data
    2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern
    3. Any outliers must be removed if they are known to be errors (consider the effects of any outliers that are not known errors)

- *Requirements for Formal Regression Analysis:*
    1. For each fixed value of x, the corresponding y values have a bell-shaped distribution.
    2. For the different fixed of x, the distributions of the corresponding y values have the same variance.
    3. For the different fixed values of x, the distributions of the corresponding y values have means that lie along the same straight line
    4. The y values are independent

- *Notation for Regression Equation:*

	Population Parameter	Sample Statistic
y intercept of regression equation		
slope of regression equation		
equation of the regression line		

- Finding the slope b_1 and y intercept b_0 in the regression equation (round to 3 significant digits):

--Slope: $b_1 = \dfrac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$

--Y intercept: $b_0 = \bar{y} - b_1\bar{x}$

*Equations done manually and using technology may differ slightly, due to rounding in manual calculations (we are going to use the graphing calculator this year)

- *Using the graphing calculator to find the equation of the regression line:*
  - Press STAT then 1: Edit
  - Enter the lists of values into L1 and L2
  - Press STAT then select TESTS
  - Choose 4: LinReg
  - The Xlist should be L1 and the Ylist should be L2 (leave freqlist & store blank)
  - Put the cursor on CALCULATE, press ENTER
  - you'll see the general form for the equation of a line y = ax + b
  - then you'll see the values of a and b for the given data

Ex#1: In section 10-2, we found the linear correlation coefficient of r = -0.956. Use the given sample data to find the regression equation.

x	3	1	3	5
y	5	8	6	4

Equation of the regression line:

Ex#2: Using the duration/interval after eruption times, we found that the linear correlation coefficient is r = 0.926. Use the following sample data to find the equation of the regression line.

Duration	240	120	178	234	235	269	255	220
Interval after	92	65	72	94	83	94	101	87

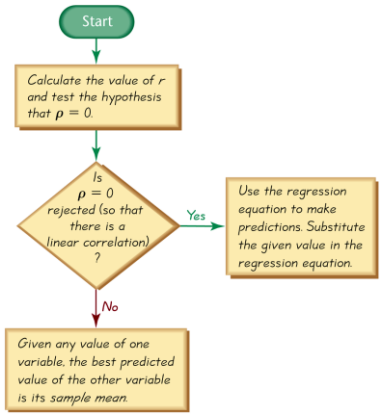
Equation of the regression line:

- *Using the Regression Equation for Predictions:* Regression equations are often useful for predicting the value of one variable, given some particular value of the other variable.

  Requirements:

  1. Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.
  2. Use the regression equation for predictions only if the linear correlation coefficient r indicates that there is a linear correlation between the two variables
  3. Use the regression line for predictions only if the data does not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called extrapolation, and could result in bad predictions)
  4. If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its sample mean.

- *In predicting a value of y based on some given value of x…*
  1. If there is NOT a linear correlation, _____
  2. If there IS a linear correlation, _____
     _____



* Think of r as a measure of how well the regression line fits the sample data

--Regression equations obtained from sample data with r very close to -1 or +1 are likely to be much better models for the population data than regression equations from data sets with values of r that are not so close to -1 or +1.

- Ex #1: Using the duration times and time intervals after eruptions of Old Faithful, we found that there is a linear correlation between the two variables, and we also found that the regression equation is $\hat{y} = 34.783 + 0.234x$. Assuming that the current eruption has a duration of x = 180 sec, find the best predicted value of y, the time interval after this eruption.

  Note: If there was not a linear correlation, we would use $\bar{y}$ as the best prediction of y:

- Ex #2: There is obviously no linear correlation between hat sizes and IQ scores of adults. Given that an individual has a hat size of 7, find the best predicted value of this person's IQ score.

Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7