3.1 Association between Two Categorical Variables

Response and Explanatory variables

- **Response variable** (Dependent Variable) - the outcome variable on which comparisons are made
- **Explanatory variable** (Independent variable) - defines the groups to be compared with respect to values on the response variable

  Example: Response/Explanatory
  - Blood alcohol level/# of beers consumed
  - Grade on test/Amount of study time
  - Yield of corn per bushel/Amount of rainfall

Association

- The main purpose of data analysis with two variables is to investigate whether there is an association and to describe that association
- An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable

contingency table:

- Displays two categorical variables
- The rows list the categories of one variable
- The columns list the categories of the other variable
- Entries in the table are frequencies

Example : Food type and Pesticide status

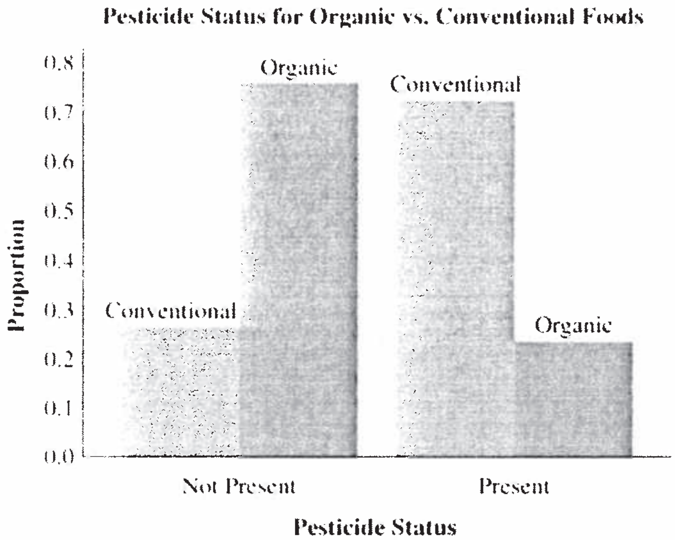
	Pesticide Status		
Food Type	Present	Not Present	**Total**
Organic	29	98	**127**
Conventional	19485	7086	**26571**
Total	**19514**	**7184**	**26698**

Calculate proportions and conditional proportions

These treat pesticide status as the response variable. The sample size n in a row shows the total on which the conditional proportions in that row were based.

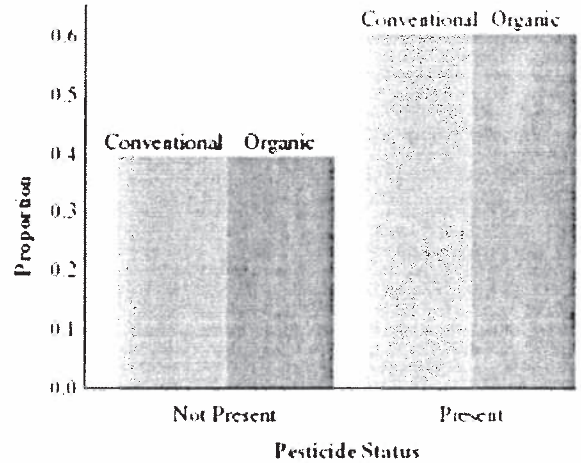
	Pesticide Status			
Food Type	Present	Not Present	**Total**	n
Organic	0.23	0.77	**1.000**	127
Conventional	0.73	0.27	**1.000**	26571

- What proportion of organic foods contain pesticides?
- What proportion of conventionally grown foods contain pesticides?



- Use side by side bar charts to show conditional proportions
- Allows for easy comparison of the explanatory variable with respect to the response variable

❖ If there was no association between organic and conventional foods, then the proportions for the response variable categories would be the same for each food type

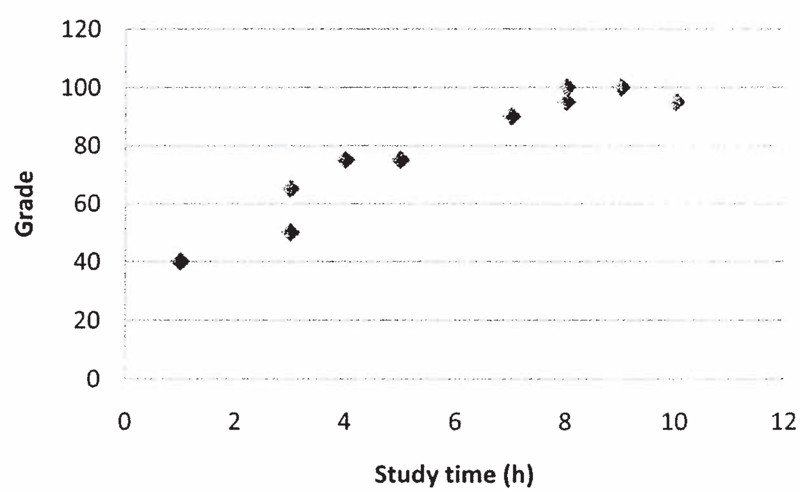


3.2 Explore the Association between Two Quantitative Variables

- Graphical display of relationship between two quantitative variables:
  - Horizontal Axis: *Explanatory variable*, x
  - Vertical Axis: *Response variable*, y

Example:

Study time (h) per week (X)	Grade on test (Y)
8	95
10	95
5	75
3	65
8	100
3	50
4	75
7	90
9	100
1	40



Interpreting Scatterplots

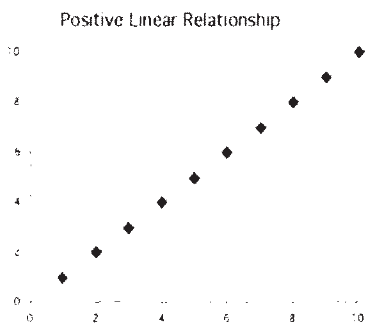
➢ We can describe the overall pattern of a scatterplot by the trend, direction, and strength of the relationship between the two variables
  - Trend: linear, curved, clusters, no pattern
  - Direction: positive, negative, no direction
  - Strength: how closely the points fit the trend

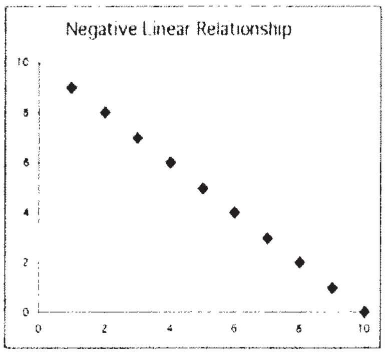
➢ Also look for outliers from the overall trend

Interpreting Scatterplots: Direction/Association

Two quantitative variables x and y are

- *Positively associated* when
  - High values of x tend to occur with high values of y
  - Low values of x tend to occur with low values of y
- *Negatively associated* when high values of one variable tend to pair with low values of the other variable





Positive association: As x goes up, y tends to go up

Negative association: As x goes up y tends to go down

Ex: Would you expect a positive association, a negative association or no association between the

(i) age of the car and the mileage on the odometer - Positive association
(ii) age of the car and the resale value - Negative association
(iii) age of the car and the total amount that has been spent on repairs- Positive association
(iv) weight of the car and the number of miles it travel on a gallon of gas - Negative association

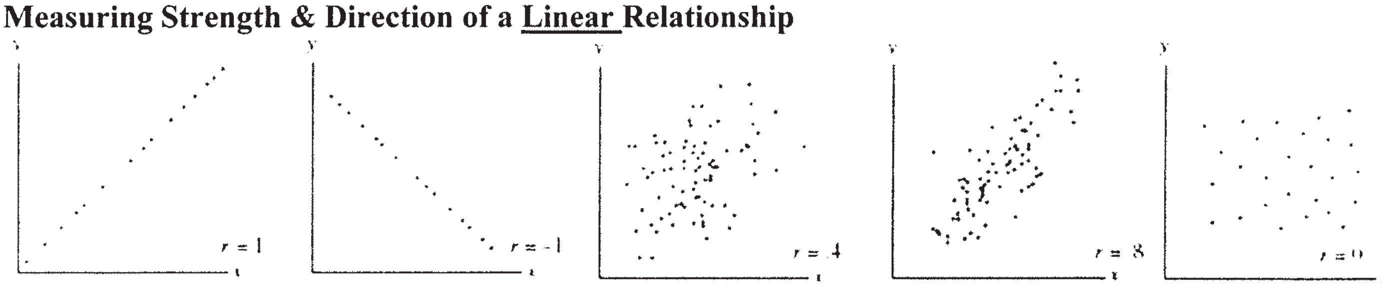
Linear Correlation, r

Measures the strength and direction of the linear association between x and y. Correlation coefficient, r is defined by

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = \frac{1}{n-1}\sum Z_x Z_y$$

- A positive r value indicates a positive association
- A negative r value indicates a negative association
- An r value close to +1 or -1 indicates a strong linear association
- An r value close to 0 indicates a weak association

Measuring Strength & Direction of a Linear Relationship



Properties of Correlation

- Always falls between -1 and +1
- Sign of correlation denotes direction
  - (-) indicates negative linear association
  - (+) indicates positive linear association
- Correlation has a unitless measure - does not depend on the variables' units
- Two variables have the same correlation no matter which is treated as the response variable

3.3 Regression Analysis

Regression Line

A regression line is a straight line that describes how the response variable (y) changes as the explanatory variable (x) changes. It predicts the value of the response variable (y) for a given level of the explanatory variable (x). Regression line has the form

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

The y-intercept of the regression line is denoted by a & the slope of the regression line is denoted by b

Ex: Regression Equation: $\hat{y} = 61.4 + 2.4x$

Here $\hat{y}$ is the predicted height and x is the length of a femur (thighbone), measured in centimeters

Use the regression equation to predict the height of a person whose femur length was 50 centimeters

$$\hat{y} = 61.4 + 2.4(50) = 181.4$$

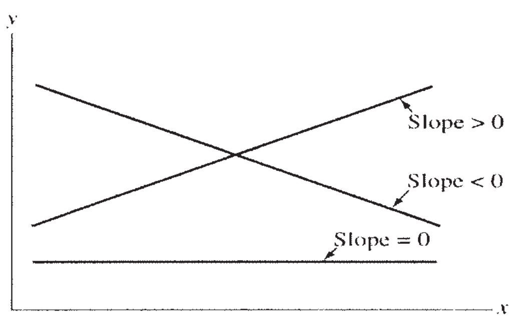
Interpreting the y-Intercept

- The predicted value for y when $x = 0$
- Helps in plotting the line
- May not have any interpretative value if no observations had x values near 0

Interpreting the Slope

Slope: measures the change in the predicted variable (y) for a 1 unit increase in the explanatory variable in (x)

Example: A 1 cm increase in femur length results in a 2.4 cm increase in predicted height



Regression Line

- ➢ At a given value of x, the equation: $\hat{y} = a + bx$
  - Predicts a single value of the response variable
  - But… we should not expect all subjects at that value of x to have the same value of y
    - Variability occurs in the y values!
- ➢ The regression line connects the estimated *means* of y at the various x values
- ➢ In summary, $\hat{y} = a + bx$ describes the relationship between x and the *estimated means* of y at the various values of x

Residuals

- Measures the size of the prediction errors, the vertical distance between the point and the regression line
  - Each observation has a residual
  - Calculation for each residual: $(y - \hat{y})$
- A large residual indicates an unusual observation

"Least Squares Method" Yields the Regression Line

- Residual sum of squares: $\sum(residual)^2 = \sum(y - \hat{y})^2$

- The least squares regression line is the line that minimizes the vertical distance between the points and their predictions, i.e., it minimizes the residual sum of squares
- Note: the sum of the residuals about the regression line will always be zero

Regression Formulas for y-Intercept and Slope

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

Slope: $b = r(\frac{s_y}{s_x})$ **Y-Intercept:** $a = \bar{y} - b(\bar{x})$

- Regression line always passes through $(\bar{x}, \bar{y})$

The Slope and the Correlation

Correlation:

- Describes the strength of the linear association between 2 variables
- Does not change when the units of measurement change
- Does not depend upon which variable is the response and which is the explanatory

Slope:

- Numerical value depends on the units used to measure the variables
- Does not tell us whether the association is strong or weak
- The two variables must be identified as response and explanatory variables
- The regression equation can be used to predict values of the response variable for given values of the explanatory variable

The Squared Correlation(r^2)

- r^2 measures the proportion of the variation in the y-values that is accounted for by the linear relationship of y with x
- A correlation of .9 means that $.9^2 = .81 = 81\%$
  81% of the variation in the y-values can be explained by the explanatory variable, x

Some Cautions in Analyzing Association

Extrapolation: Using a regression line to predict y-values for x-values outside the observed range of the data

- Riskier the farther we move from the range of the given x-values
- There is no guarantee that the relationship given by the regression equation holds outside the range of sampled x-values

Outliers and Influential Points

- Construct a scatterplot
  - Search for data points that are well outside of the trend that the remainder of the data points follow
- A *regression outlier* is an observation that lies far away from the trend that the rest of the data follows
- An observation is influential if
  - Its x value is relatively low or high compared to the remainder of the data
  - The observation is a regression outlier

Influential observations tend to pull the regression line toward that data point and away from the rest of the data

Correlation does not Imply Causation

- A strong correlation between x and y means that there is a strong linear association that exists between the two variables
- A strong correlation between x and y, does not mean that x *causes* y
- Data are available for all fires in Chicago last year on x = number of firefighters at the fires and y = cost of damages due to fire
  - Would you expect the correlation to be negative, zero, or positive?
  - If the correlation is positive, does this mean that having more firefighters at a fire causes the damages to be worse? Yes or No
  - Identify a third variable that could be considered a common cause of x and y:
    a) Distance from the fire station, b) Intensity of the fire, c) Size of the fire

A ***lurking variable*** is a variable, usually unobserved, that influences the association between the variables of primary interest

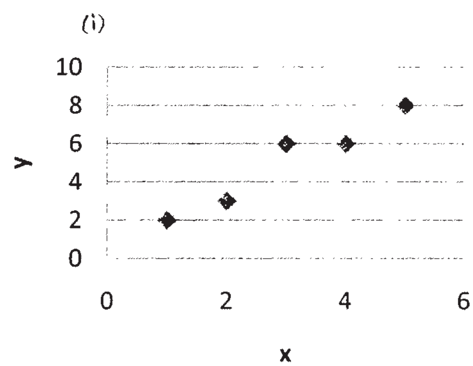
- Reading level and shoe size – lurking variable=age
- Childhood obesity rate and GDP-lurking variable=time

- When two explanatory variables are both associated with a response variable but are also associated with each other, there is said to be *confounding*
- *Lurking variables* are not measured in the study but have the potential for *confounding*

Example :Consider the following data
(i) draw the scatter plot and predict r
(ii) calculate correlation coefficient
(iii) find the regression line

x	y
1	2
3	6
4	6
5	8
2	3

(i)



x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$Z_x = \frac{x-\bar{x}}{s_x}$	$Z_x = \frac{y-\bar{y}}{s_y}$	Z_xZ_y
1	2	$(1-3)^2 = 4$	$(2-5)^2 = 9$	-1.266	-1.224	1.550
3	6	$(3-3)^2 = 0$	$(6-5)^2 = 1$	0.000	0.408	0.000
4	6	$(4-3)^2 = 1$	$(6-5)^2 = 1$	0.633	0.408	0.258
5	8	$(5-3)^2 = 4$	$(8-5)^2 = 9$	1.266	1.224	1.550
2	3	$(2-3)^2 = 1$	$(3-5)^2 = 4$	-0.633	-0.816	0.517
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma(x-\bar{x})^2 = 10$	$\Sigma(y-\bar{y})^2 = 24$			3.875

$\Sigma Z_x Z_y$

$$\bar{x} = \frac{\Sigma x}{5} = \frac{15}{5} = 3$$

$$\bar{y} = \frac{\Sigma y}{5} = \frac{25}{5} = 5$$

$$S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{10}{5-1}} = 1.58$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{24}{5-1}} = 2.45$$

(ii) $r = \frac{1}{n-1} \Sigma Z_x Z_y = \frac{1}{5-1} 3.875 = 0.957$

(iii) Regression Line $\Rightarrow \hat{y} = a + bx$.

$b = r\left(\frac{S_y}{S_x}\right) = 0.957\left(\frac{2.45}{1.58}\right) = 1.484$.

$a = \bar{y} - b(\bar{x}) = 5 - 1.484(3) = 0.56$

$\Rightarrow \hat{y} = 0.56 + 1.484x$

MATH 2300 - Chapter 3 Problems

Answer true or false.

1) A side-by-side bar graph is a graphical display for two categorical variables. 1) ____________

2) If the absolute value of the correlation is approximately one, then the points lie close to a line that slopes upward or downward. 2) ____________

Complete the contingency table and use it to solve the problem.

3) The partially completed contingency table gives the frequencies of the data on age (in years) and sex from the residents of a retirement home. 3) ________

	Age (yrs)			
	60-69	70-79	Over 79	Total
Male	11	9	5	
Female	9	2	4	
Total				

What is the proportion of male residents in the age group 60-69?

A) 0.55 B) 0.44 C) 0.50 D) 0.35 E) 0.60

Select the most appropriate answer.

4) If a positive association exists between two quantitative variables, 4) ________

A) the movement of x does not affect the movement of y.

B) y tends to decrease as x increases.

C) y tends to decrease as x decreases.

D) y tends to increase as x decreases.

E) none of these.

Provide an appropriate response.

5) Almost all of the acidity of soda pop comes from the phosphoric acid which is added to give them a sharper flavor. Is there an association between the pH of the soda and the amount of phosphoric acid (in grams)? The correlation between pH and phosphoric acid is -0.991. Describe the association. 5) ________

A) Weak linear association in a negative direction

B) Very strong linear association in a negative direction

C) No evidence of association

D) Strong linear association in a positive direction

E) Weak linear association in a positive direction

6) For the 14 teams in baseball's American league , the correlation with number of wins in the 2007 regular season is 0.51 for shutouts, 0.61 for hits made, -.70 for runs allowed and -0.56 for homeruns allowed. (mlb.mlb.com/stats/) Which variable has the strongest linear association with number of wins? 6) ________

A) homeruns allowed B) hits made

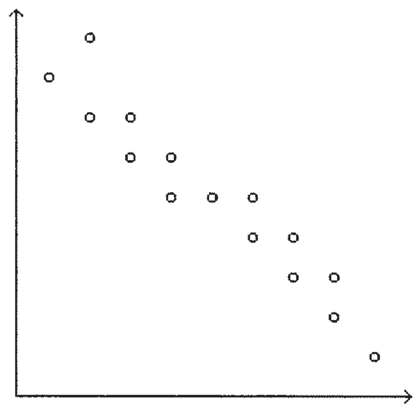
C) shutouts D) runs allowed

7) In 2007, the number of wins had a mean of 81.79 with a standard deviation of 10.89 for the teams of baseball's American league. The equation that predicts the number of wins (y) using the number of runs allowed (x) is $\hat{y} = 159.62 - 0.10x$. What is the predicted number of wins for a team that allowed 800 runs? Round your answer to the nearest integer. 7) ______

A) 168 B) 82 C) 80 D) 160

Determine the type of association apparent in the following scatterplot.

8) 8) ______



A) Linear association, moderately strong association

B) Linear association, very strong association

C) Negative association, linear association

D) Negative association, moderately strong association

E) Negative association, linear association, very strong association

Provide an appropriate response.

9) Based on findings from the Health and Nutrition Examination Survey conducted by the National Center for Health Statistics from April 1971 to June 1974, the regression equation predicting the average weight of a male aged 18–24 (y) based on his height (x) is given by $\hat{y} = -172.63 + 4.842x$. (www.cdc.gov/nchs/data/ad/ad014acc.pdf) Interpret the slope of the regression line. 9) ______

A) for every unit increase in weight, the predicted height increases by 4.842 pounds

B) for every unit increase in height, the predicted weight increases by 4.842 pounds

C) for every unit increase in weight, the predicted height decreases by 4.842 pounds

D) for every unit increase in height, the predicted weight decreases by 4.842 pounds

10) The regression equation relating dexterity scores (x) and productivity scores (y) for the employees of a company is $\hat{y} = 5.50 + 1.91x$. Ten pairs of data were used to obtain the equation. The same data yield $r = 0.986$ and $\bar{y} = 56.3$. What is the best predicted productivity score for a person whose dexterity score is 20? 10) ______

A) 56.30 B) 43.7 C) 111.91 D) 38.20 E) 58.20

11) A regression line for predicting Interenet usage (%) for 39 countries is $\hat{y} = -3.61 + 1.55x$, where x is the per capita GDP, in thousands of dollars, and y is Internet usage. What is the residual for a country with a per capita GDP of $28,000 and actual Internet use of 38 percent? 11) ______

A) -4.5 B) -1.79 C) 5.4 D) 1.79 E) -5.4

12) Which statement is true about residuals? 12) _______

A) Residuals measure the size of prediction errors.

B) Not all observations have residuals.

C) The larger the absolute value of a residual, the closer the predicted value is to the actual value.

D) In a scatterplot, the residual for an observation is the horizontal distance between the point and the regression line.

E) None of these

13) Which of the following is <u>not</u> a property of r? 13) _______

A) The closer r is to zero, the weaker the linear relationship between x and y.

B) r measures the strength of any kind of relationship between x and y.

C) r is always between -1 and 1.

D) r does not depend on the units of y or x.

E) r does not depend on which variable is treated as the response variable.

Provide an appropriate response.

14) A random sample of records of electricity usage of homes in the month of July gives the amount of electricity used and size (in square feet) of 135 homes. A simple linear regression was performed to predict the amount of electricity used (in kilowatt-hours) based on size. The resulting model is $\widehat{usage}$ = 1229 + 0.02 size. The residual for a family living in a house that is 2290 square feet is negative. Interpret. 14) _______

A) They are using less electricity than other houses of the same size based on the regression equation.

B) Their house is smaller than predicted by the regression equation.

C) They are using more electricity than predicted by the regression equation.

D) Their house is bigger than predicted by the regression equation.

E) They are using less electricity than predicted by the regression equation.

15) The relationship between the number of games won by a minor league baseball team and the average attendance at their home games is analyzed. A regression to predict the average attendance from the number of games won has an $r^2 = 0.256$. Assume that a linear model is appropriate. What is the correlation between the average attendance and the number of games won? 15) _______

A) 0.863 B) 0.256 C) 0.744 D) 0.07 E) 0.506

Select the most appropriate answer.

16) Among the possible lines that can go through data points in a scatterplot, the regression line results from the least squares method and has the smallest value for the ____________________. 16) _______

A) residual sum of squares

B) slope

C) correlation

D) residual sum

Fill in the missing information.

17) 17) _______

$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = a + bx$
2.6	1	?	110	?	$\hat{y} = -110 + 40x$

A) $\bar{y} = 104$; r = -0.36 B) $\bar{y} = -6$; r = 0.36 C) $\bar{y} = -70$; r = 0.02 D) $\bar{y} = -6$; r = 0.02