Topic 2

Descriptive Statistics for Bivariate Data

The Association Between
Two Categorical Variables

Response and Explanatory Variables

Response variable (Dependent Variable, y)

- The **outcome** variable on which comparisons are made.

Explanatory variable (Independent variable, x)

- (categorical) the groups to be compared.
- (quantitative) the change in different numerical values to be compared.

Example: Response/Explanatory

- Survival status/Smoking status
-
- GPA/Number of hours a week spent studying

Association Between Two Variables

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 Tables

A 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

Response variable

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	**127**
Conven.	19,485	7,086	**26,571**
Total	**19,514**	**7,184**	**26,698**

Calculate Proportions and Conditional Proportions

	Pesticide Status		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	127
Conventional	19,485	7,086	26,571
Total	**19,514**	**7,184**	**26,698**

These proportions are called **conditional proportions** because their formation is **conditional** on (in this example) food type.

	Pesticide Status			
Food Type	**Present**	**Not Present**	**Total**	***n***
Organic	0.23	0.77	**1.00**	127
Conventional	0.73	0.27	**1.00**	26,571

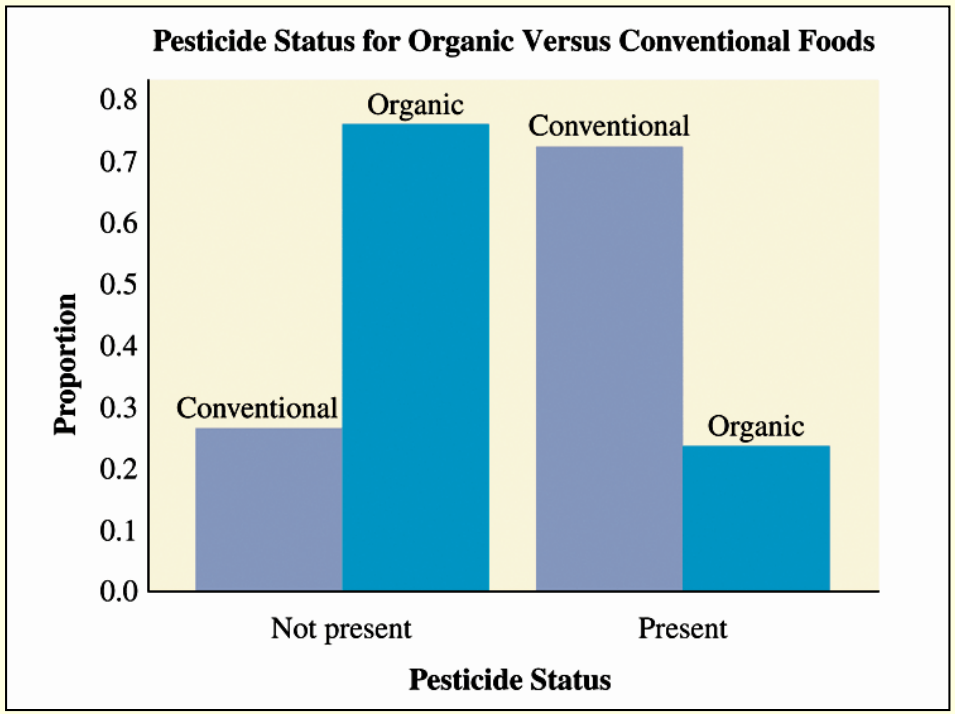
Questions:

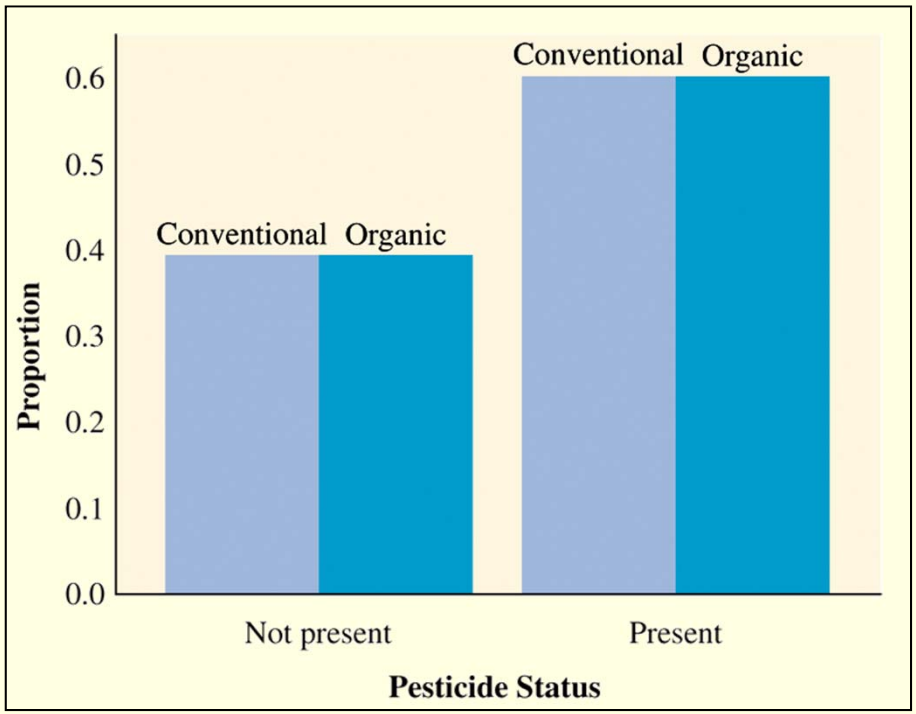
1. What proportion of organic foods contain pesticides?
2. What proportion of conventionally grown foods contain pesticides?
3. What proportion of all sampled items contain pesticides?

Calculate Proportions and Conditional Proportions

Using side by side bar charts to show conditional proportions allows for easy comparison of the explanatory variable with respect to the response variable.

Side by side bar charts below show the Conditional Proportions on Pesticide Status, Given the Food Type.





Topic 2

Descriptive Statistics for Bivariate Data

The Association Between
Two Quantitative Variables

Scatterplot

Graphical display of the relationship between two quantitative variables:

- Horizontal Axis: *Explanatory variable*, x
- Vertical Axis: *Response variable*, y

We examine a scatterplot to study association.

How do values on the response variable change as values of the explanatory variable change?

You can describe the overall pattern of a scatterplot by

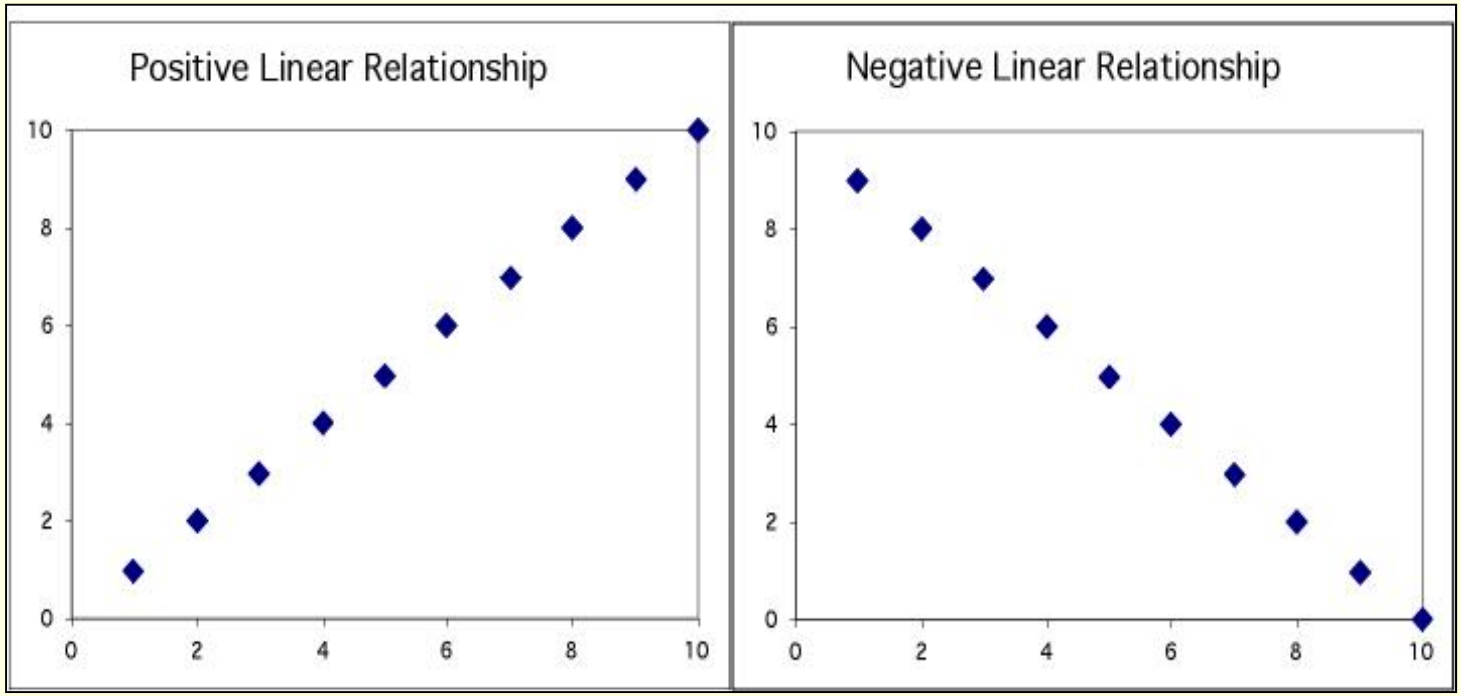
- **Trend**: linear, curved, clusters, no pattern
- **Direction**: positive, negative, no direction
- **Strength**: how closely the points fit the trend

Also look for **outlier**s from the overall trend.

Scatterplot

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.



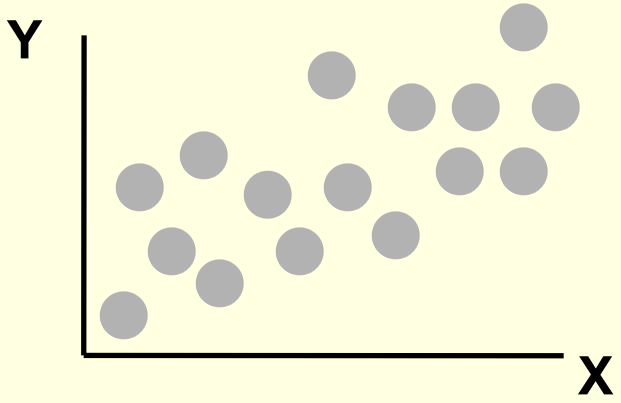
Example: 100 cars on the lot of a used-car dealership

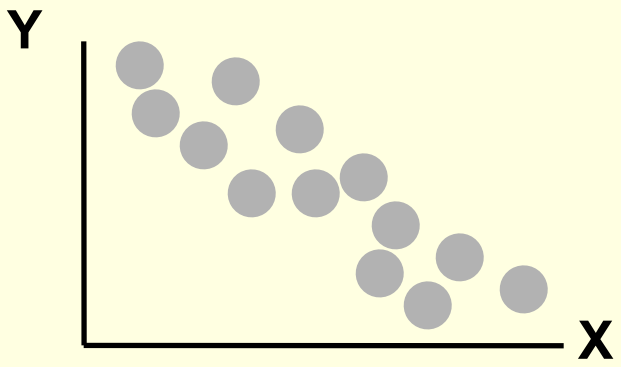
Question: Would you expect a positive association, a negative association or no association between the age of the car and the mileage on the odometer?

- Positive association
- Negative association
- No association

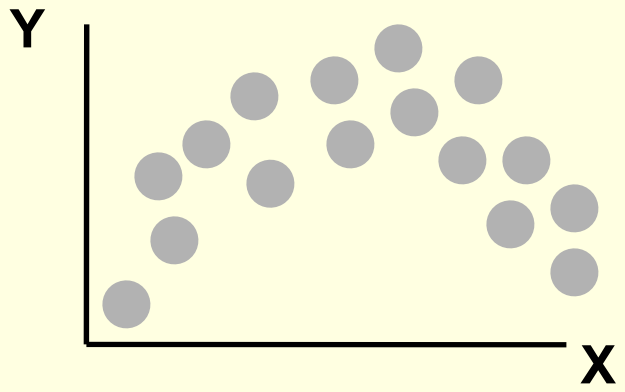
Types of Relationships (trend & direction)

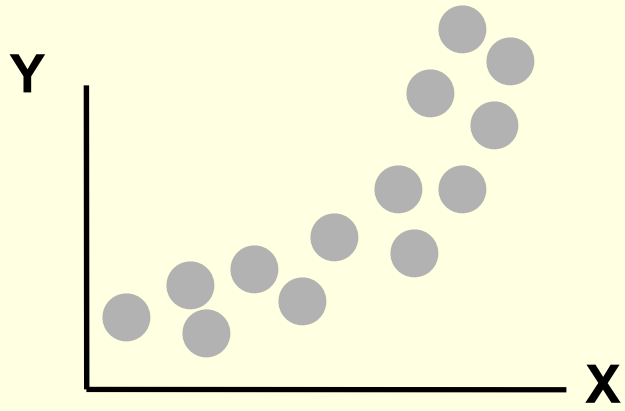
Linear relationships



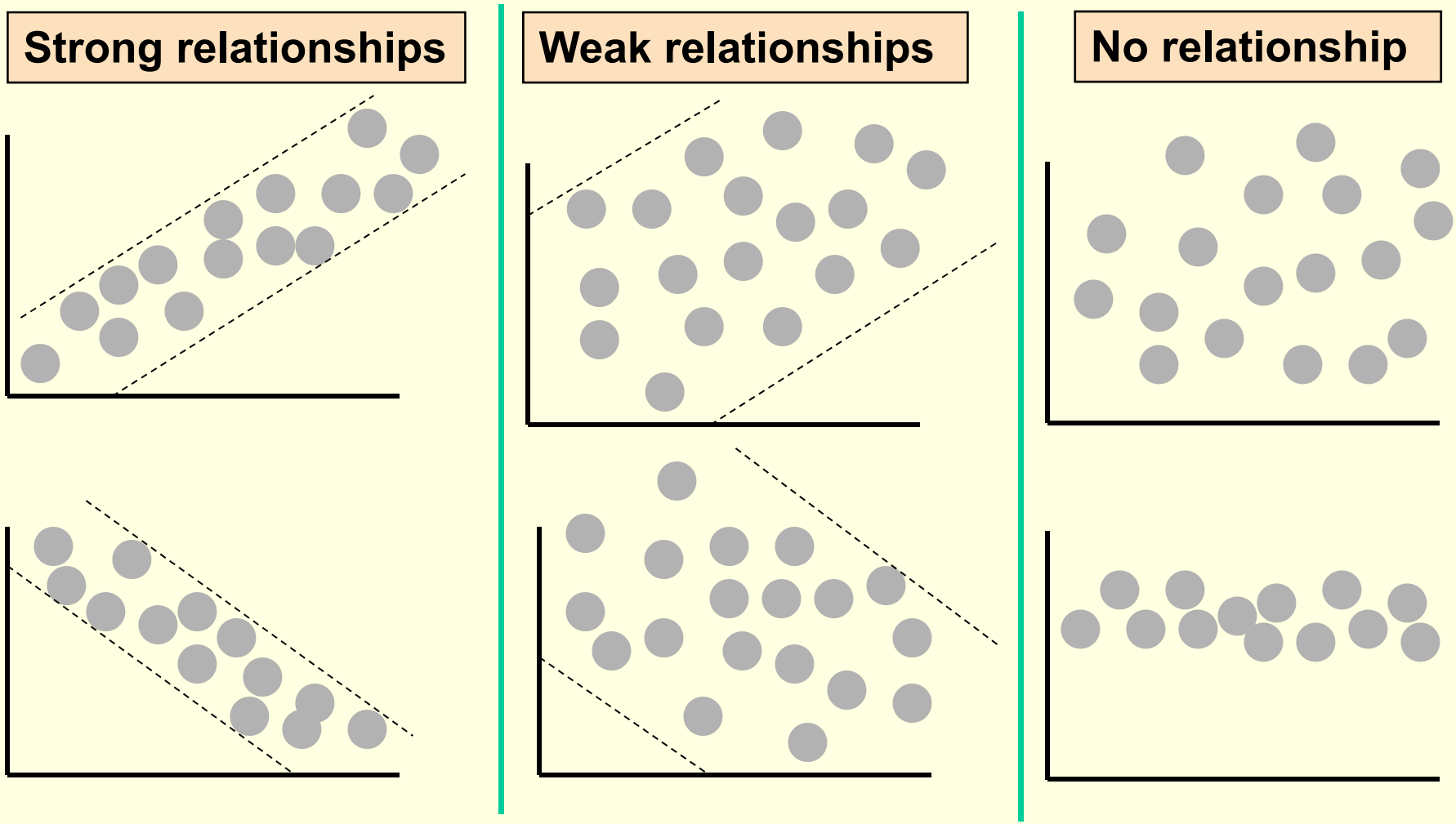


Curvilinear relationships





Types of Relationships (strength)



The Scatterplot: Looking for a Trend

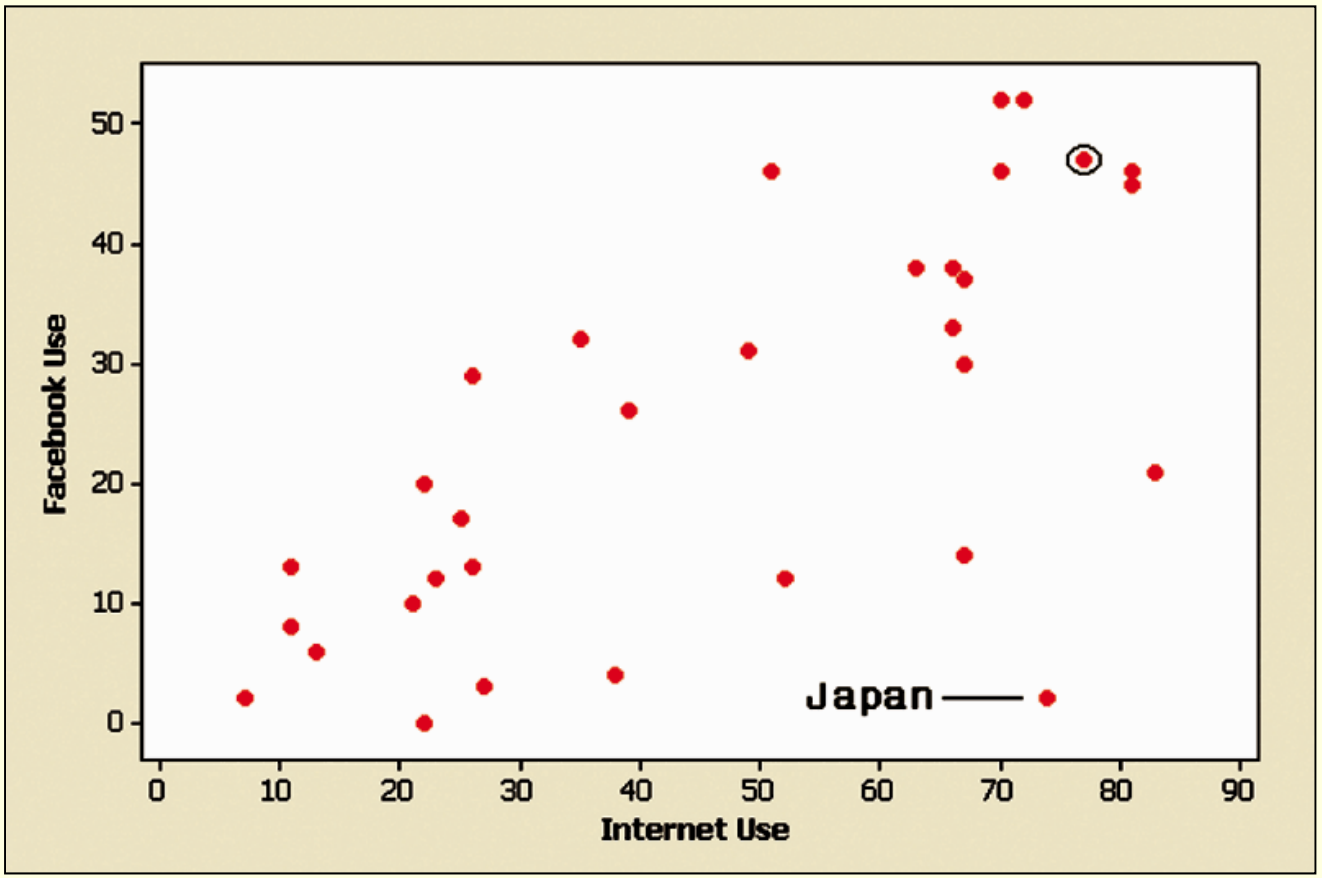


Figure 3.5 MINITAB Scatterplot for Internet Use and Facebook Use for 33 Countries

Question:

Is there any point that you would identify as standing out in some way? Which country does it represent, and how is it unusual in the context of these variables?

Example: Internet and Facebook Penetration Rates

Table 3.4 Internet and Facebook Penetration Rates For 33 Countries

Country	Internet Penetration	Facebook Penetration
Argentina	49.40%	30.53%
Australia	80.60%	46.01%
Belgium	67.30%	36.98%
Brazil	37.76%	4.39%
Canada	72.30%	52.08%
Chile	50.90%	46.14%
China	22.40%	0.05%
Colombia	38.80%	25.90%
Egypt	12.90%	5.68%
France	65.70%	32.91%
Germany	67.00%	14.07%
Hong Kong	69.50%	52.33%
India	7.10%	1.52%
Indonesia	10.50%	13.49%
Italy	48.80%	30.62%
Japan	73.80%	2.00%
Malaysia	62.80%	37.77%
Mexico	24.90%	16.80%
Netherlands	82.90%	20.54%
Peru	26.20%	13.34%
Philippines	21.50%	19.68%
Poland	52.00%	11.79%
Russia	27.00%	2.99%
Saudi Arabia	22.70%	11.65%
South Africa	10.50%	7.83%
Spain	66.80%	30.24%
Sweden	80.70%	44.72%
Taiwan	66.10%	38.21%
Thailand	20.50%	10.29%
Turkey	35.00%	31.91%
USA	77.33%	46.98%
UK	70.18%	45.97%
Venezuela	25.50%	28.64%

Example: Internet and Facebook Penetration Rates

Using MINITAB, we obtain the following numerical measures of center and spread:

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Max
Internet Use	33	47.00	24.40	7.00	24.00	49.00	68.50	83.00
Facebook Use	33	24.73	16.49	0.00	11.00	26.00	38.00	52.00

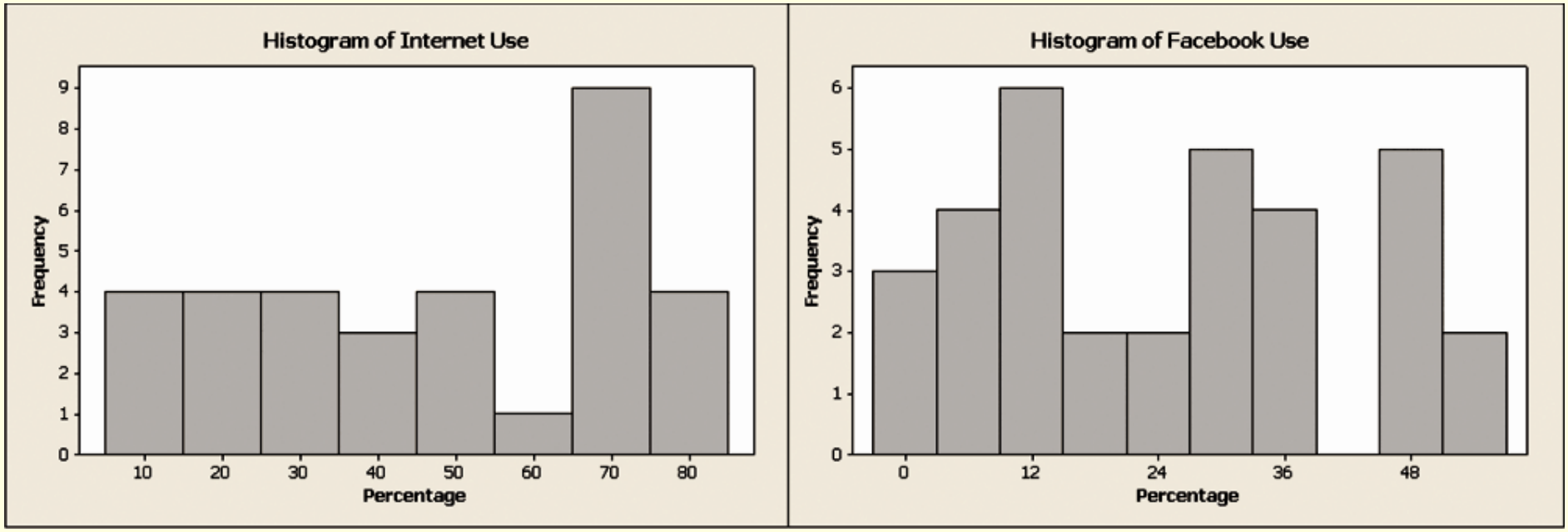
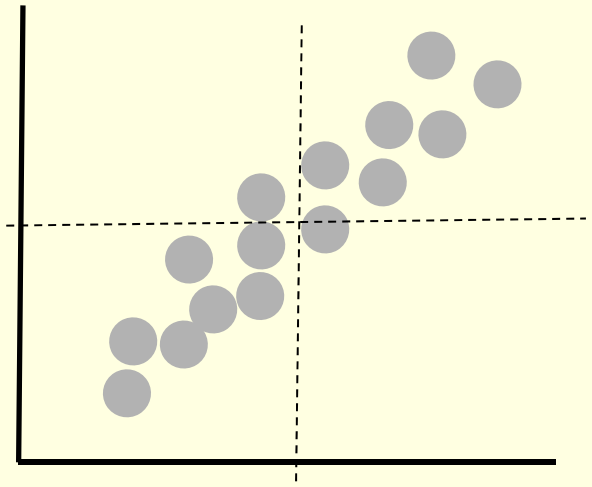


Figure 3.4 MINITAB histograms of Internet use and Facebook use for the 33 countries.
Question: Which nations, if any, might be outliers in terms of Internet use? Facebook use? Which graphical display would more clearly identify potential outliers?

Summarizing the Strength of Association: The Correlation, *r*

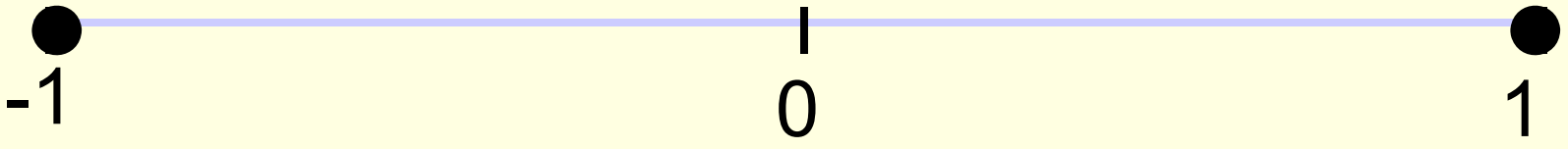
$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$



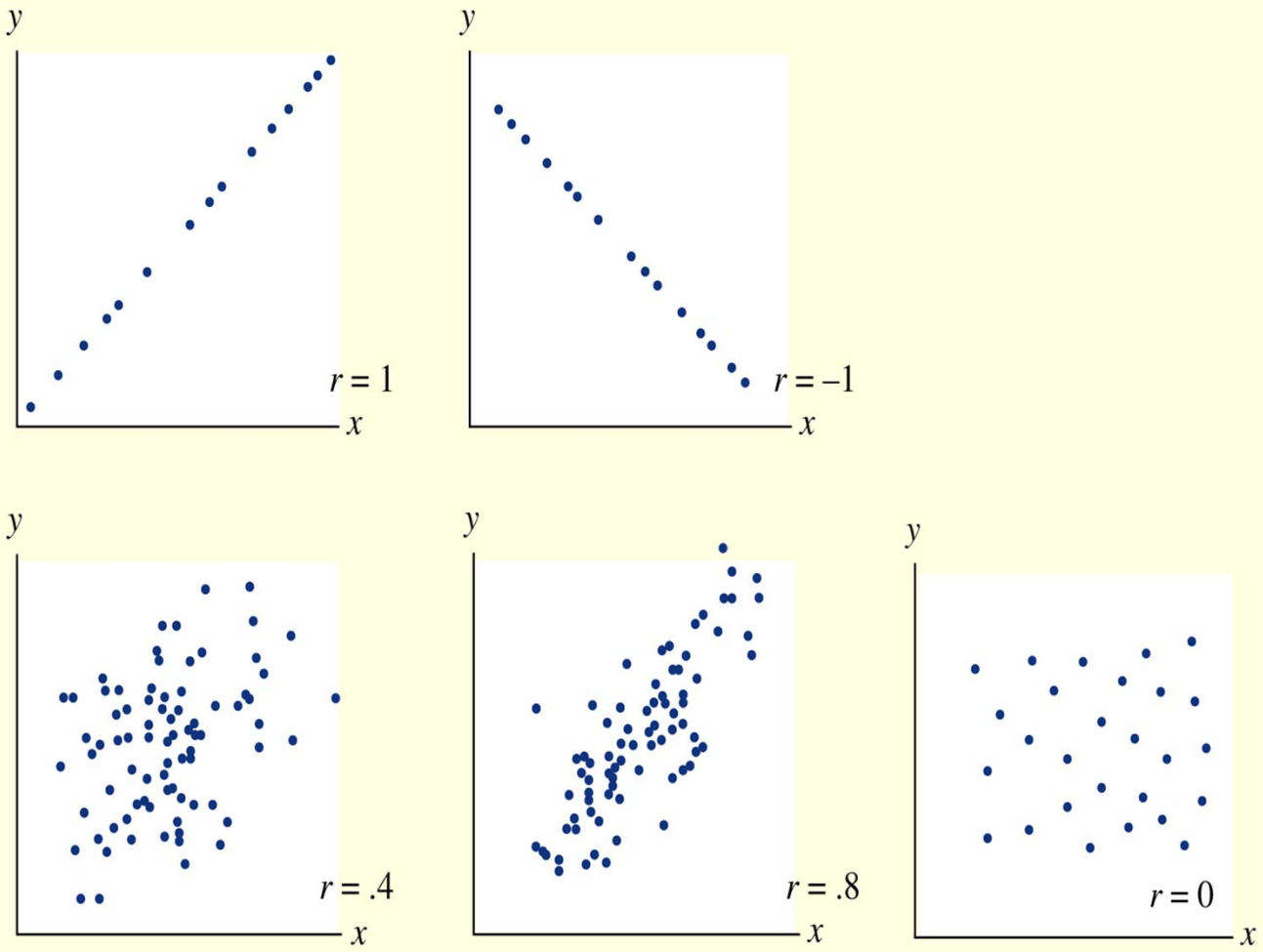
Summarizing the Strength of Association: The Correlation, *r*

The **Correlation** measures the strength and direction of the *linear association* between x and 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.



Correlation Coefficient: Measuring Strength and 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 unit-less measure, it does not depend on the variables' units.
- Two variables have the same correlation no matter which is treated as the response variable.
- Correlation is ***not*** resistant to outliers.
- Correlation ***only*** measures strength of a **linear relationship.**

Calculating the Correlation Coefficient

Example: Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe.

	x	y
Country	**Per Capita GDP**	**Life Expectancy**
Austria	21.4	77.48
Belgium	23.2	77.53
Finland	20.0	77.32
France	22.7	78.63
Germany	20.8	77.17
Ireland	18.6	76.39
Italy	21.5	78.51
Netherlands	22.0	78.15
Switzerland	23.8	78.99
United Kingdom	21.2	77.37

Calculating the Correlation Coefficient

Example: Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe.

x	Y			
21.4	77.48	-0.078	-0.345	0.027
23.2	77.53	1.097	-0.282	-0.309
20.0	77.32	-0.992	-0.546	0.542
22.7	78.63	0.770	1.102	0.849
20.8	77.17	-0.470	-0.735	0.345
18.6	76.39	-1.906	-1.716	3.271
21.5	78.51	-0.013	0.951	-0.012
22.0	78.15	0.313	0.498	0.156
23.8	78.99	1.489	1.555	2.315
21.2	77.37	-0.209	-0.483	0.101
= 21.52	= 77.754			sum = 7.285
s_x =1.532	s_y =0.795			

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \left(\frac{1}{10-1}\right)(7.285)$$

$$= 0.809$$

Topic 2

Descriptive Statistics for Bivariate Data

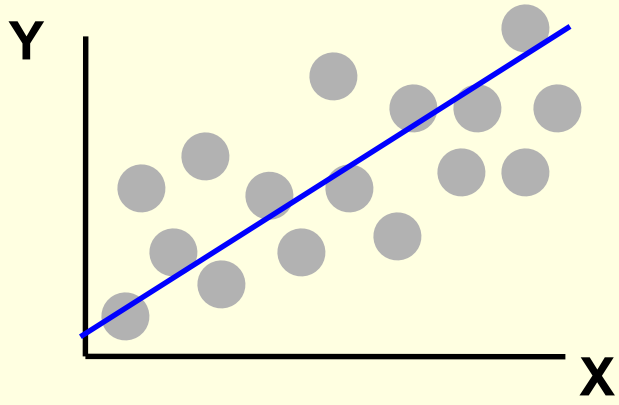
Predicting the Outcome of a Variable

(Linear Regression)

Regression Line

The first step of a *regression analysis* is to identify the response and explanatory variables.

- We use *y* to denote the *response variable.*
- We use *x* to denote the *explanatory variable.*



Regression Line: An Equation for Predicting the Response Outcome

The **regression line** predicts the value for the *response variable y* as a straight-line function of the *value x of the explanatory variable*.

Let $\hat{y}$ denote the **predicted value** of *y.* The equation for the regression line has the form

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

In this formula, *a* denotes the ***y*-intercept** and *b* denotes the **slope.**

Example: Height Based on Human Remains

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

$\hat{y}$ 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

***y*-Intercept:**

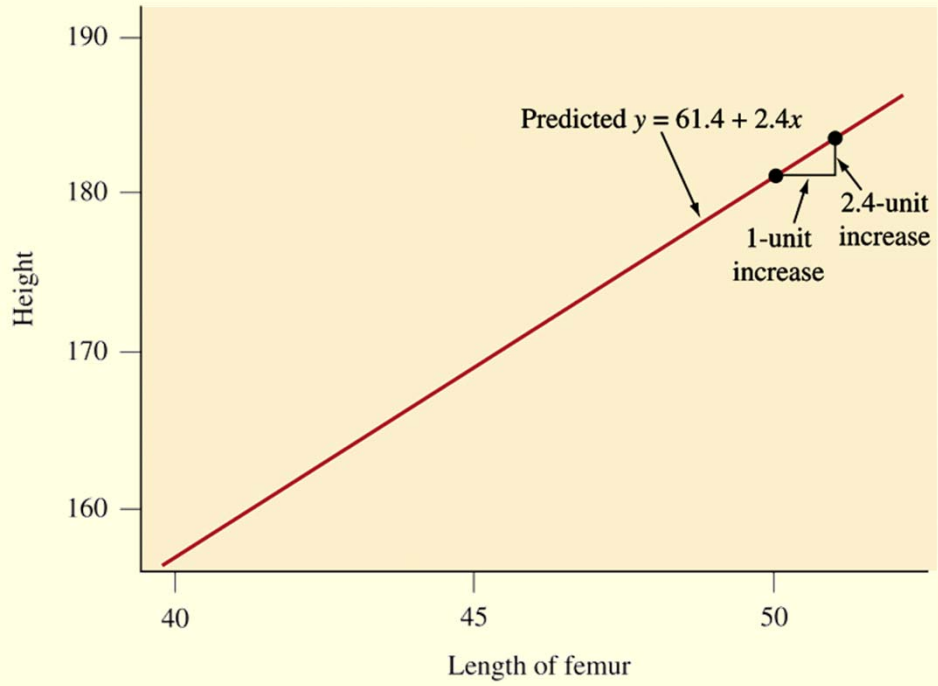
- The predicted value for *y* when $x = 0$
- This fact helps in plotting the line
- May not have any interpretative value if no observations had *x* values near 0

It does not make sense for femur length to be 0 cm, so the *y*-intercept for the equation $\hat{y} = 61.4 + 2.4x$ is ***not*** a relevant predicted height.

Interpreting the Slope

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

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



Slope Values

Positive: **The relationship is positive**

Negative: **The relationship is negative**

Zero: **There is no relationship**

Question Would you expect a positive or negative slope when y = annual income and x = number of years of education?

Residuals Measure the Size of Prediction Errors

Residuals measure 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.
- The smaller the absolute value of a residual, the closer the predicted value is to the actual value, so the better is the prediction.

The Method of Least Squares Yields the Regression Line

Residual sum of squares:

$$\sum (residuals)^2 = \sum (y - \hat{y})^2$$

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

Slope: $$b = r\left(\frac{s_y}{s_x}\right)$$

***y*-Intercept:** $$a = \overline{y} - b(\overline{x})$$

Notice that the slope *b* is directly related to the correlation *r*, and the *y*-intercept depends on the slope.

Calculating the slope and *y*-intercept for the regression line

Using the baseball data to illustrate the calculations.

Table 3.5 Team Batting Average and Team Scoring (Mean Number of Runs per Game) for American League Teams in 2010[5]

Team	Batting Average	Team Scoring
NY Yankees	0.267	5.30
Boston	0.268	5.05
Tampa Bay	0.247	4.95
Texas	0.276	4.86
Minnesota	0.273	4.82
Toronto	0.248	4.66
Chicago Sox	0.268	4.64
Detroit	0.268	4.64
LA Angels	0.248	4.20
Kansas City	0.274	4.17
Oakland	0.256	4.09
Cleveland	0.248	3.99
Baltimore	0.259	3.78
Seattle	0.236	3.17

$\bar{x} = .2597$

$\bar{y} = 4.45$

$s_x = 0.01257$

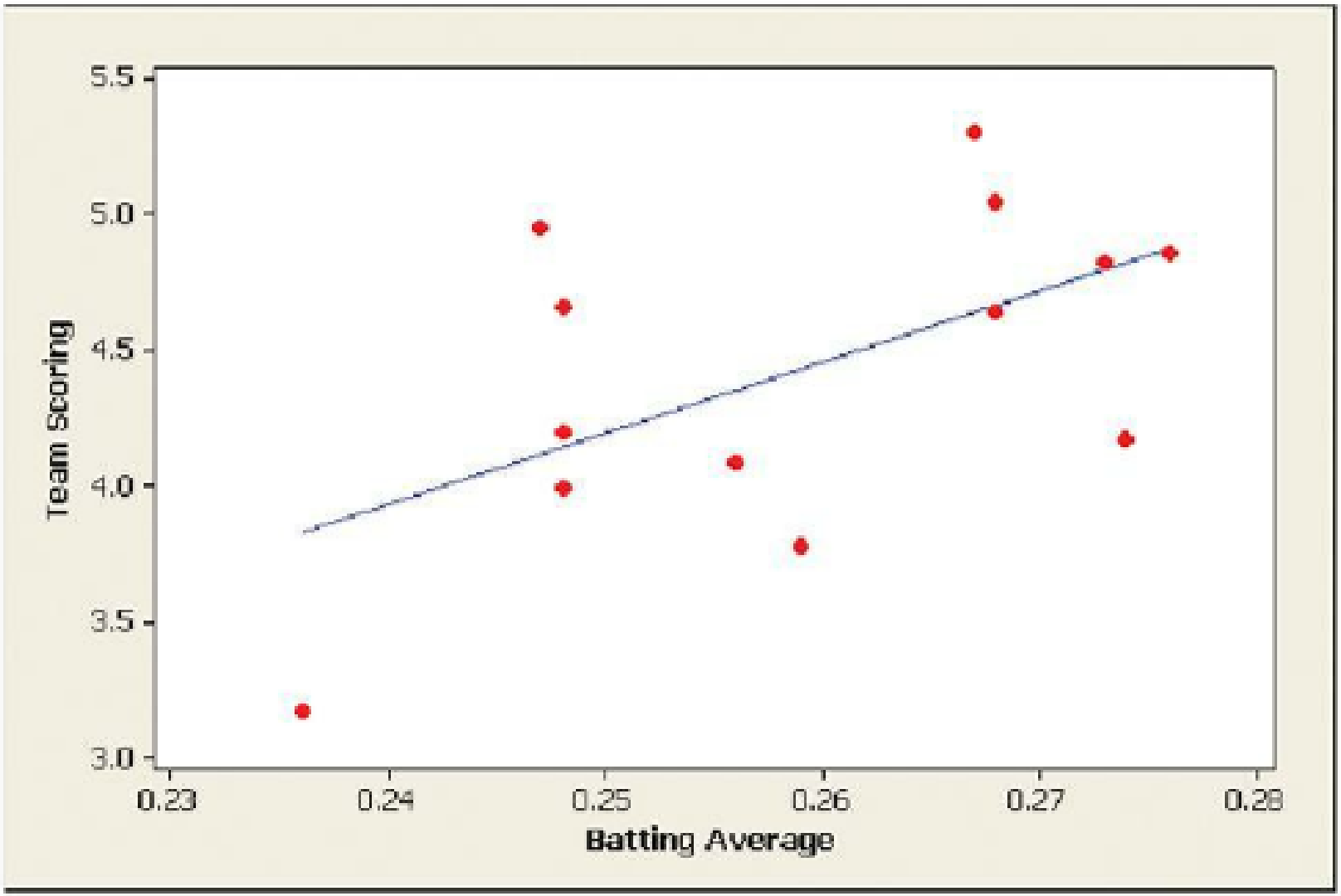
$s_y = 0.577$

$r = 0.568$

$$b = r\left(\frac{s_y}{s_x}\right) = 0.568\left(\frac{0.577}{0.01257}\right) = 26.07$$

$$a = \bar{y} - b(\bar{x}) = 4.45 - 26.07(0.2597) = -2.32$$

The regression line to predict team scoring from batting average is $= -2.32 + 26.1x$



▲ **Figure 3.13 MINITAB Output for Scatterplot of Team Batting Average and Team Scoring, with Regression Line Superimposed. Question** How can you find the prediction error that results when you use the regression line to predict team scoring for a team?

The Slope and the Correlation

Correlation:

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

The Slope and the Correlation

Slope:

- The two variables must be identified as response and explanatory variables.
- Numerical value depends on the units used to measure the variables.
- Does not tell us whether the association is strong or weak.
- 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)

The typical way to interpret the **squared correlation** r^2 is as the proportion of the variation in the *y*-values that is accounted for by the linear relationship of *y* with *x.*

When a strong linear association exists, the regression equation predictions tend to be much better than the predictions using only $\bar{y}$.

We measure the *proportional reduction in error* and call it, r^2. It is also called the coefficient of determination.

A correlation of 0.9 means that

$$.9^2 = .81 = 81\%$$

- 81% of the variation in the *y*-values can be explained by the explanatory variable, *x.*

The Squared Correlation (r^2)

Another way to describe the strength of association refers to **how close** predictions for y tend to be to observed y values.

The variables are **strongly associated** if you can predict y **much better** by substituting x values into the prediction equation than by merely using the sample mean $\bar{y}$ and ignoring x.

Consider the prediction error: the difference between the observed and predicted values of y.

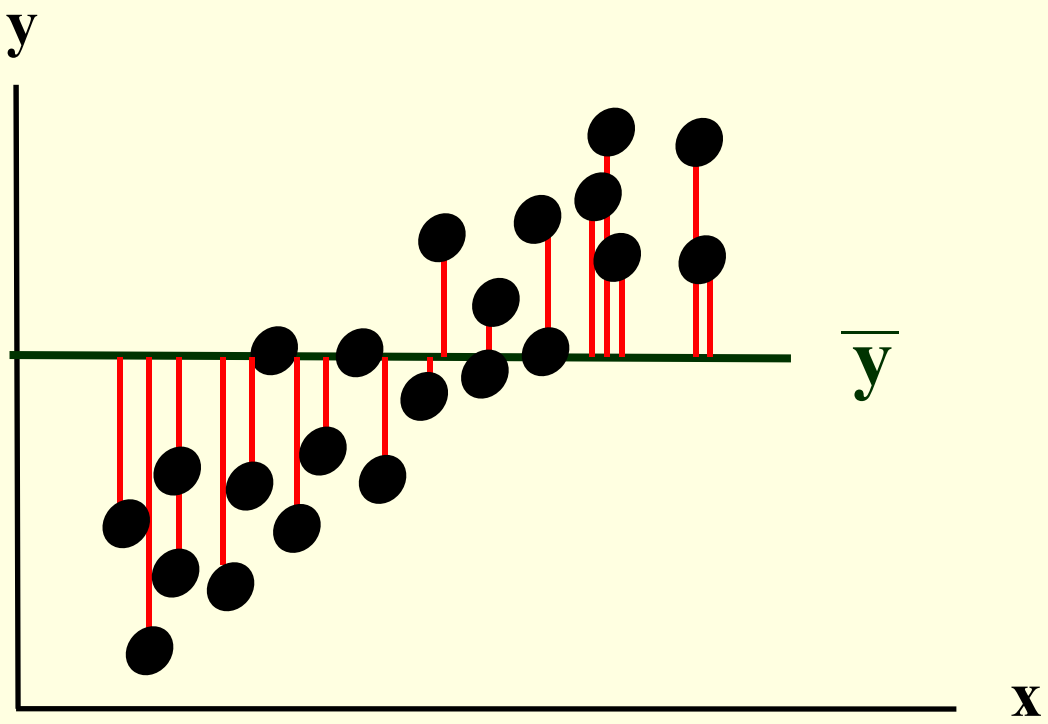
- Using the regression line to make a prediction, each error is: $y - \hat{y}$

- Using only the sample mean, $\bar{y}$, to make a prediction, each error is: $y - \bar{y}$

The Squared Correlation (r^2)

When we predict y using $\overline{y}$ (that is, ignoring x), the error summary equals: $\sum(y-\overline{y})^2$

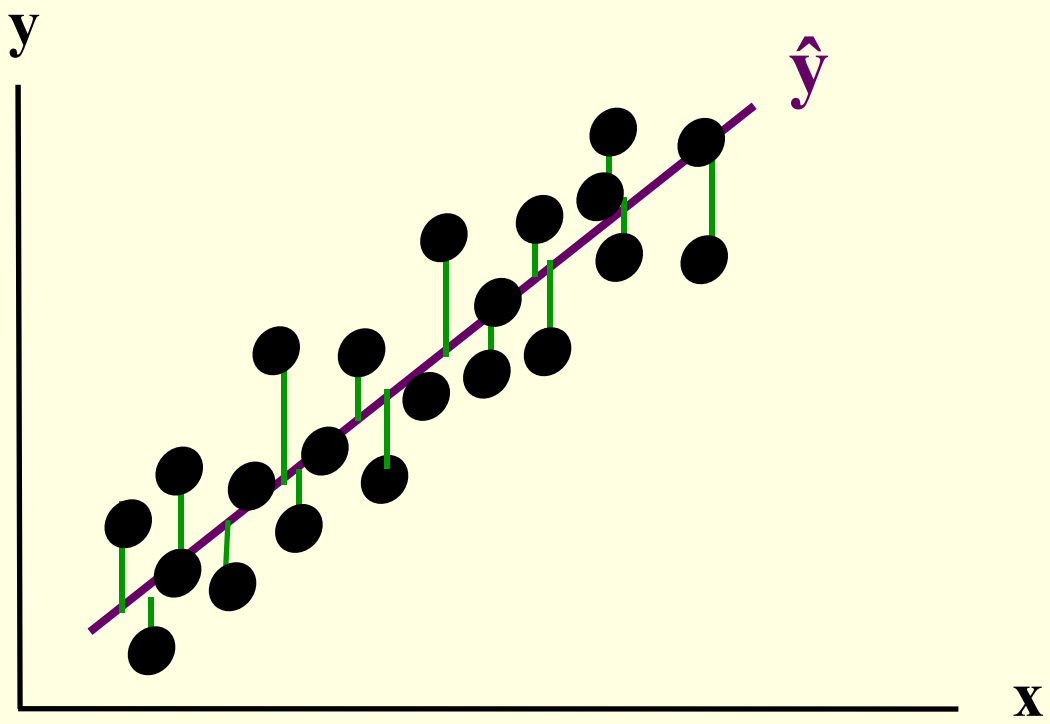
This is called the *total sum of squares*.



The Squared Correlation (r^2)

When we predict y using x with the regression equation, the error summary is: $\sum(y-\hat{y})^2$

This is called the *residual sum of squares*.



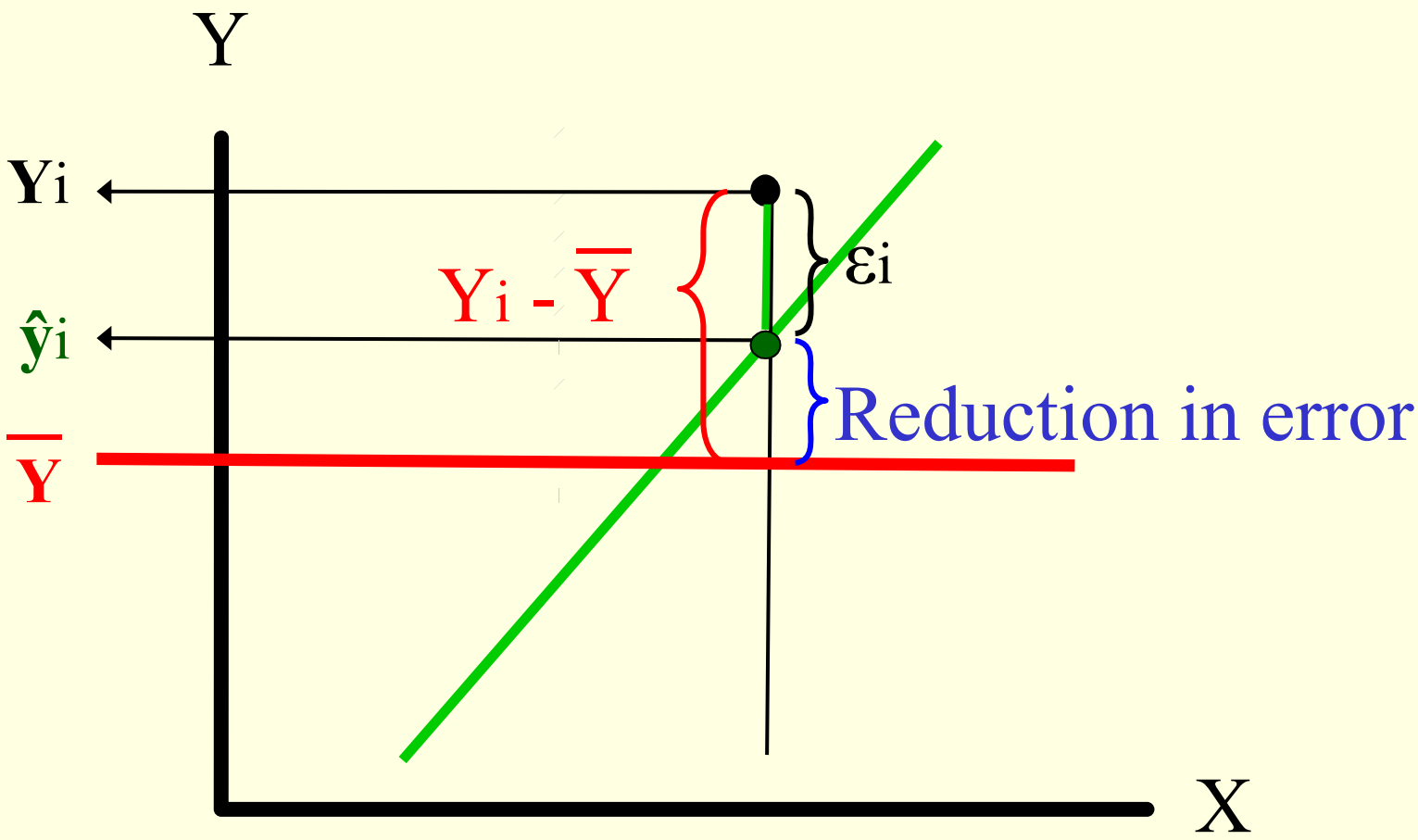
The Squared Correlation (r^2)

When a **strong** linear association exists, the regression equation predictions tend to be **much better** than the predictions using $\overline{y}$

We measure the ***proportional reduction in error*** and call it, ***r*²**

We use the notation r^2 for this measure because it equals the square of the correlation *r.*

$$r^2 = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$


Y
Yi
ŷi
Ȳ
Yi - Ȳ
εi
Reduction in error
X

Example:
What Does r^2 Tell Us in the Strength Study?

For the female athlete strength study:

- x: number of 60-pund bench presses
- y: maximum bench press
- The correlation value was found to be $r = 0.80$

We can calculate r^2 from r: $(0.80)^2=0.64$

For predicting maximum bench press, the regression equation has 64% less error than $\bar{y}$ has

64% of the variation in Y is explained by the regression equation using X.

Correlation r and Its Square r^2

Both r and r^2 describe the strength of association

'r' falls between -1 and +1

- It gives us the direction and the strength of the linear relationship between x and y.

'r^2' falls between 0 and 1

- It summarizes the reduction in sum of squared errors in predicting y using the regression line instead of using $\overline{y}$

Topic 2

Descriptive Statistics for Bivariate Data

Cautions in Analyzing Associations

Extrapolation Is Dangerous

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.

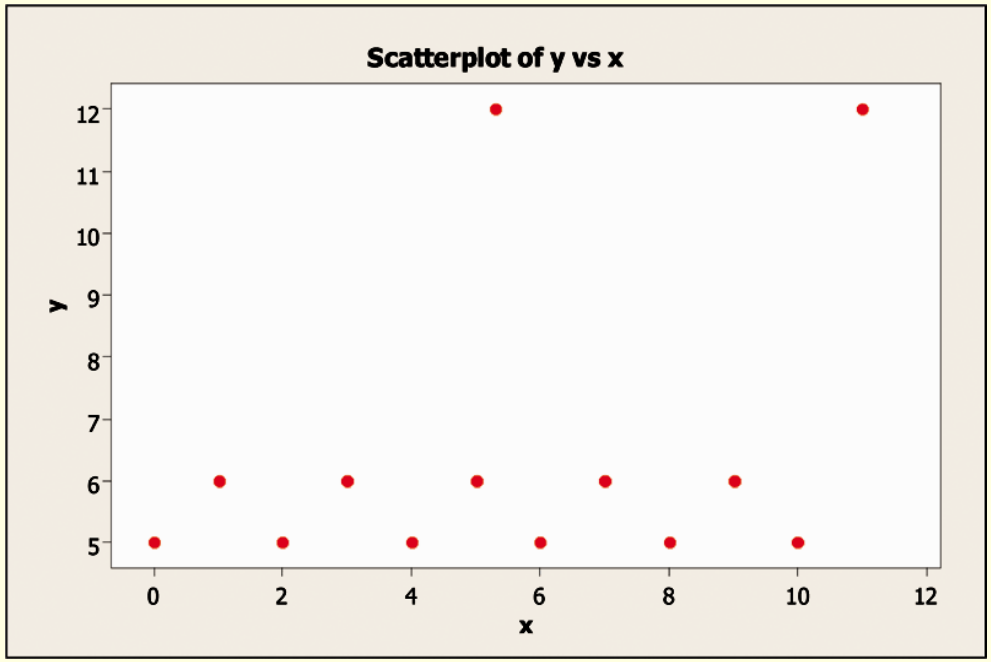
- There is no guarantee that the relationship given by the regression equation holds outside the range of sampled *x*-values.

Outliers and Influential Points

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

- the observation is a regression outlier.
- its *x* value is relatively low or high compared to the remainder of the data.



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

Simpson's Paradox

Simpson's Paradox:

When the direction of an association between two variables changes after we include a third variable and analyze the data at separate levels of that third variable.

Simpson's Paradox
Example: Smoking and Health

Is Smoking Actually Beneficial to Your Health?

Table 3.7 Smoking Status and 20-Year Survival in Women

	Survival Status		
Smoker	**Dead**	**Alive**	**Total**
Yes	139	443	582
No	230	502	732
Total	**369**	**945**	**1,314**

Probability of Death of **Smoker** = 139/582 = 24%

Probability of Death of **Nonsmoker** = 230/732 = 31%

smoking improves your chances of living !!!???

Simpson's Paradox
Example: Smoking and Health

5/(5 + 174) =
0.028 or
2.8%

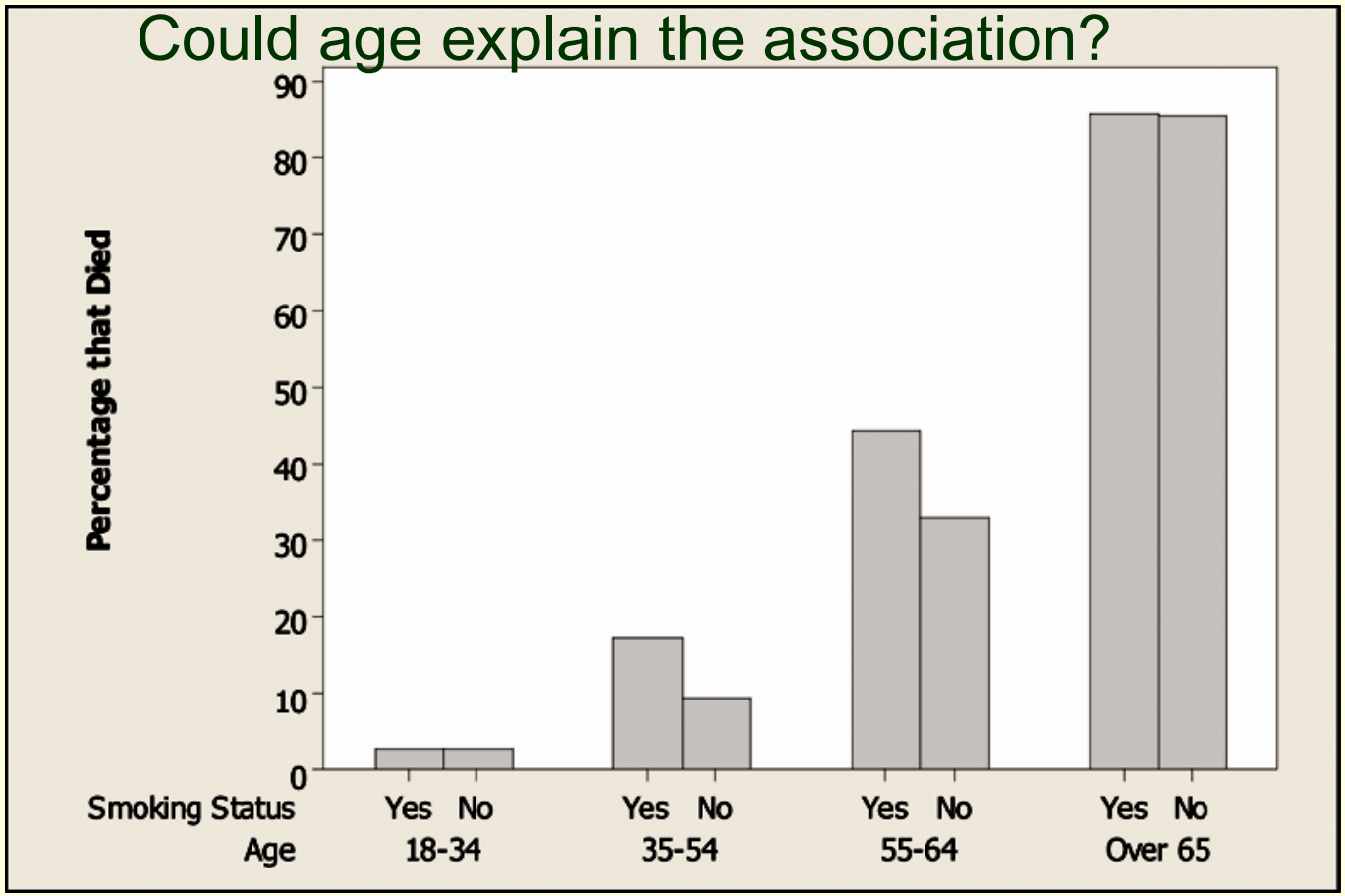
	Age Group							
	18–34 Survival?		35–54 Survival?		55–64 Survival?		65+ Survival?	
Smoker	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28

Table 3.8 Smoking Status and 20-Year Survival, for Four Age Groups

6/(6 + 213) =
0.027 or
2.7%

	Age Group			
Smoker	**18–34**	**35–54**	**55–64**	**65+**
Yes	2.8%	17.2%	44.3%	85.7%
No	2.7%	9.5%	33.1%	85.5%

Table 3.9 Conditional Percentages of Deaths for Smokers and Nonsmokers, by Age



Correlation Does Not Imply Causation

In a regression analysis, suppose that as *x* goes up, *y* also tends to go up (or down). Can we conclude that there's a causal connection, with changes in *x* causing changes in *y*?

- 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* to change.

Example: Correlation Does Not Imply Causation

Data are available for all fires in Chicago last year on x = number of firefighters at the fire and *y* = cost of damages due to the fire.

1. Would you expect the correlation to be negative, zero, or positive?

2. If the correlation is positive, does this mean that having more firefighters at a fire causes the damages to be worse? Yes or No?

3. Identify a third variable that could be considered a common cause of *x* and *y*:

Lurking Variables & Confounding

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

- Ice cream sales and drowning
  – lurking variable = temperature
- Reading level and shoe size
  – lurking variable = age

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.*