MATH 1700

Instructor: Mehdi Maadooliat

Chapter 3

Department of Mathematical and Statistical Sciences

CHAPTER 3

- Descriptive Analysis and Presentation of Bivariate Data
- Bivariate Data
  - Qualitative vs Qualitative
    - Contingency table
  - Qualitative vs Quantitative
    - Side-by-side Box Plot
  - Quantitative vs Quantitative
    - Scatter diagram
- Linear Correlation
- Lurking Variable
- Linear Regression

BIVARIATE DATA

- **Bivariate data** The values of two different variables that are obtained from the same population element.

- Each of the two variables may be either *qualitative* or *quantitative*. As a result, three combinations of variable types can form bivariate data:

1. Both variables are qualitative (categorical).
2. One variable is qualitative (categorical), and the other is quantitative (numerical).
3. Both variables are quantitative (numerical).

TWO QUALITATIVE VARIABLES

- **Qualitative vs Qualitative:**
  - A **cross-tabulation** or **contingency table** will be used

- **Example: Thirty students from our college were randomly identified and classified according to two variables: gender (M/F) and major (liberal arts (LA), business administration (BA), technology(T))**

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

EXAMPLE 1 – CONSTRUCTING CROSS-TABULATION TABLES

- **We can construct a 2×3 table.**

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

- **Given:**

  M = male
  F = female
  LA = liberal arts
  BA = business admin
  T = technology

- **The entry in each cell is found by determining how many students fit into each category. Adams is male (M) and liberal arts (LA) and is classified in the cell in the first row, first column.**

<table>
<tr><th rowspan="2">Gender</th><th colspan="3">Major</th></tr>
<tr><th>LA</th><th>BA</th><th>T</th></tr>
<tr><td>M</td><td>||||| (5)</td><td>||||| | (6)</td><td>||||| || (7)</td></tr>
<tr><td>F</td><td>||||| | (6)</td><td>|||| (4)</td><td>|| (2)</td></tr>
</table>

EXAMPLE 1 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

Gender	Major					
	LA		BA		T	
M	|||||	(5)	||||| |	(6)	||||| ||	(7)
F	||||| |	(6)	||||	(4)	||	(2)

- **The resulting 2 × 3 contingency (cross-tabulation) table is:**

Gender	Major			
	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

- **Percentages Based on the Grand Total (Entire Sample)**

Gender	Major			
	LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

- **Percentages Based on Row Totals**
- **(Marginal: within Gender)**

Gender	Major			
	LA	BA	T	Row Total
M	28%	33%	39%	100%
F	50%	33%	17%	100%
Col. Total	37%	33%	30%	100%

Percentages Based on Column Totals

(Marginal: within Major)

Gender	Major			
	LA	BA	T	Row Total
M	45%	60%	78%	60%
F	55%	40%	22%	40%
Col. Total	100%	100%	100%	100%

QUANTITATIVE VS QUALITATIVE: SIDE-BY-SIDE COMPARISONS

MARQUETTE UNIVERSITY
Be The Difference.

- **Quantitative vs Qualitative:**
- **Example 2: The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs.**

Design A ($n = 6$)			Design B ($n = 6$)			Design C ($n = 6$)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43

- **5-Number Summary for Each Design**

	Design A	Design B	Design C
High	40	42	43
Q_3	38	38	41
Median	36.5	34.5	40.5
Q_1	34	34	40
Low	32	33	39

Mean and Standard Deviation for Each Design

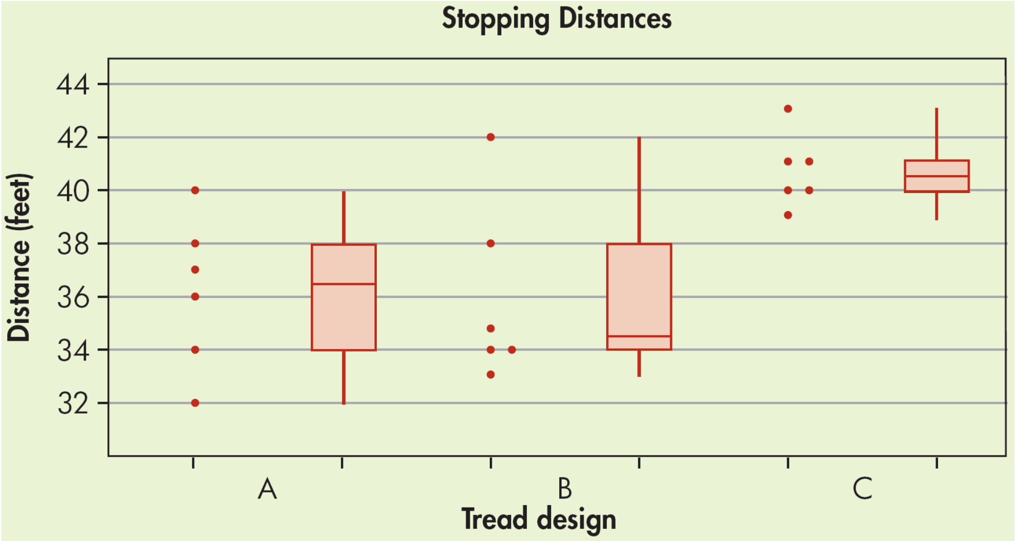
	Design A	Design B	Design C
Mean	36.2	36.0	40.7
Standard deviation	2.9	3.4	1.4

EXAMPLE 2 – CONSTRUCTING SIDE-BY-SIDE COMPARISONS

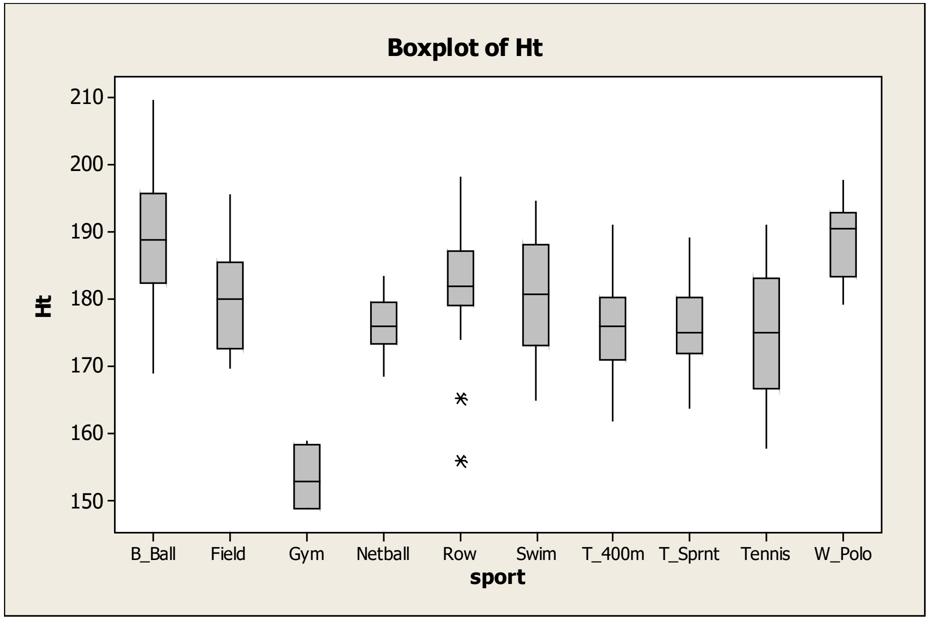
- The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs

Design A (n = 6)			Design B (n = 6)			Design C (n = 6)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43

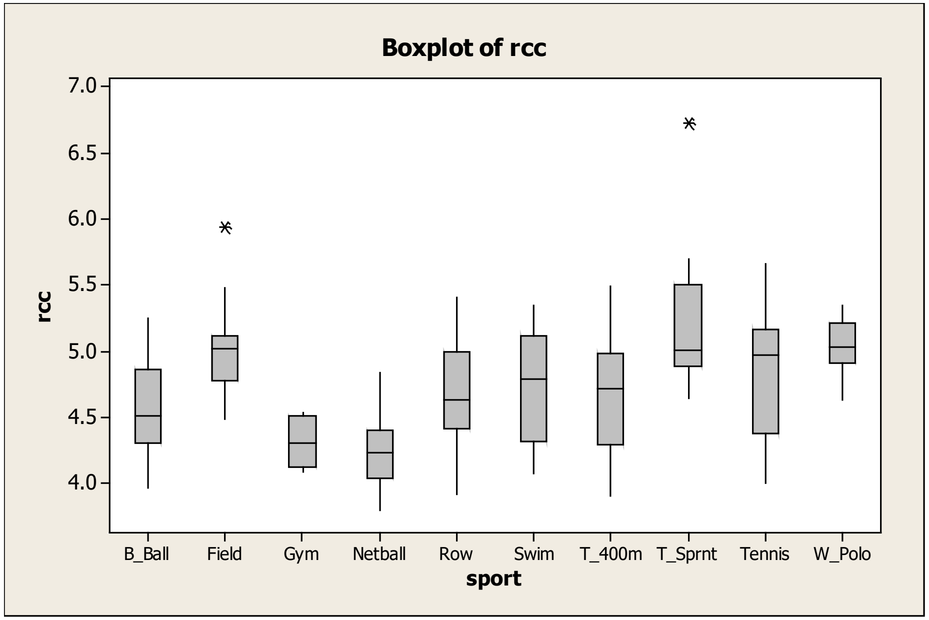
- Side by side Box -and-Whiskers



SIDE-BY-SIDE BOX PLOT FOR AIS DATA (AUSTRALIAN INSTITUTE OF SPORT)



- [JAMM: STAT-Calculator]()



TWO QUANTITATIVE VARIABLES

MARQUETTE UNIVERSITY
Be The Difference.

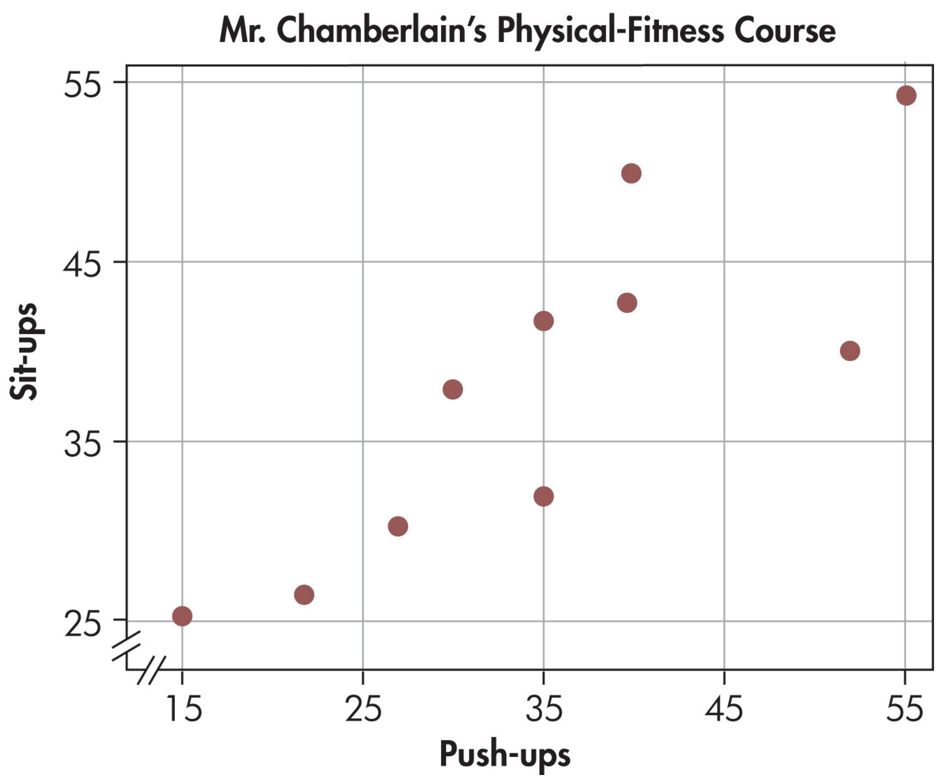
- It is customary to express the data mathematically as **ordered pairs** (x, y), where
  - x is the **input variable** (called the **independent variable**)
  - y is the **output variable** (called the **dependent variable**).
- The data are said to be *ordered* because one value, x, is always written first.
- They are called *paired* because for each x value, there is a corresponding y value from the same source.

- **Scatter diagram** A plot of all the ordered pairs of bivariate data on a coordinate axis system. The input variable, x, is plotted on the horizontal axis, and the output variable, y, is plotted on the vertical axis.

TWO QUANTITATIVE VARIABLES

- **Example: Push-ups vs Sit-ups**

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43



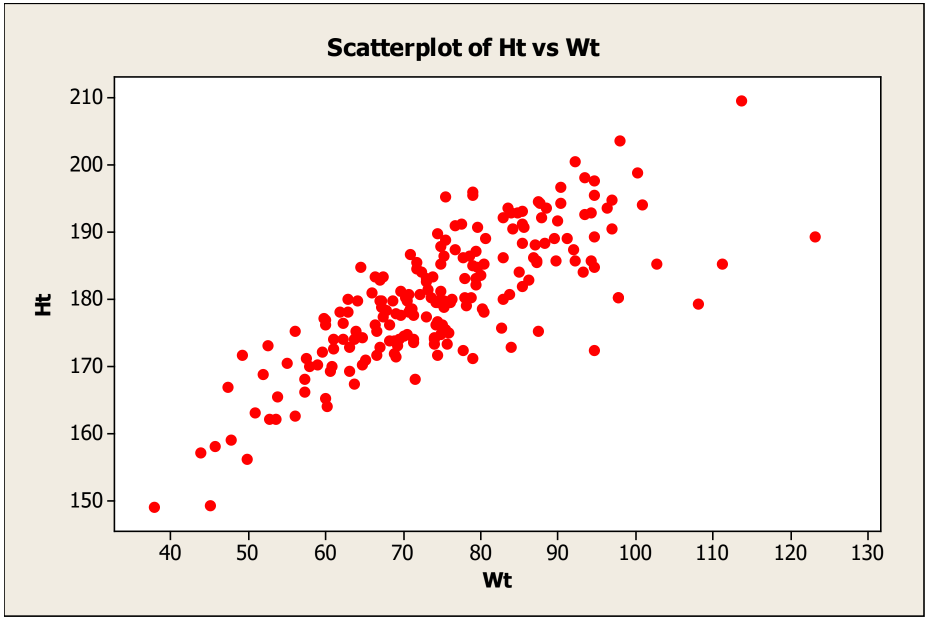
TWO QUANTITATIVE VARIABLES

- There is not always an explanatory-response (dependent-independent) relationship.

- More examples:
  - Height and Weight
  - Income and Age
  - SAT scores on math exam and on verbal exam
  - Amount of time spent studying for an exam and exam score

- [JAMM: STAT-Calculator]

Australian Institute of Sport (AIS.xlsx)

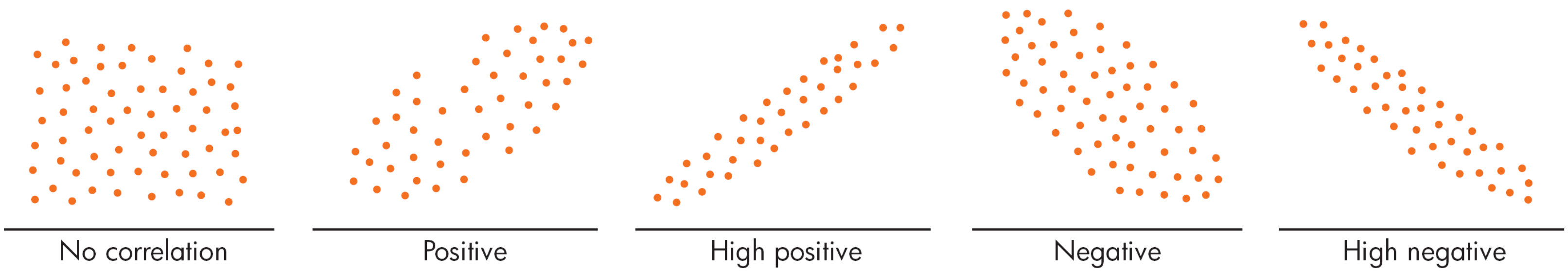


TWO QUANTITATIVE VARIABLES

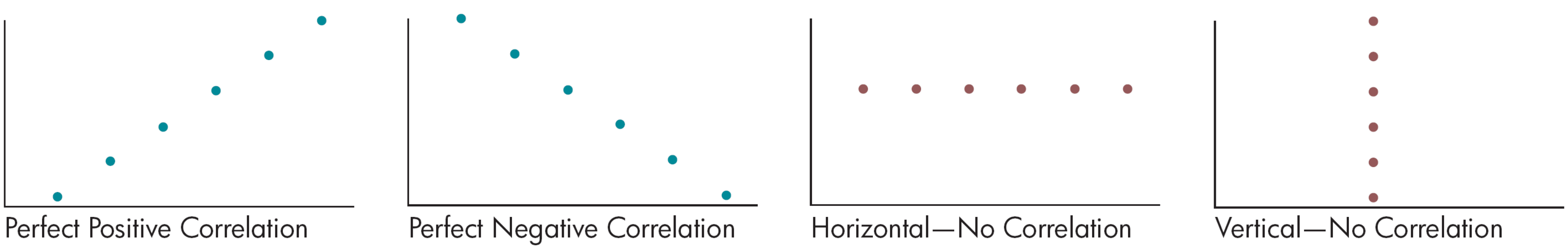
- Why Scatterplots?
- Look for overall pattern and any striking deviations from that pattern.
- Look for outliers, values falling outside the overall pattern of the relationship
- You can describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship.
  - Form: Linear or clusters
  - Direction
    - Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and likewise below-average values also tend to occur together.
    - Two variables are negatively associated when above-average values of one variable accompany below-average values of the other variable, and vice-versa.
  - Strength-how close the points lie to a line

LINEAR CORRELATION

- **Linear Correlation, r, is a measure of the strength of a linear relationship between two variables x and y.**
- $-1 \leq r \leq 1$ **Will discuss the definition in a minute**



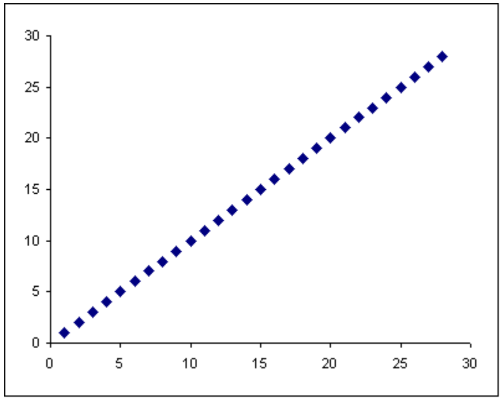
- $r \approx 0$ $r \approx 0.5$ $r \approx 0.8$ $r \approx -0.5$ $r \approx -0.8$



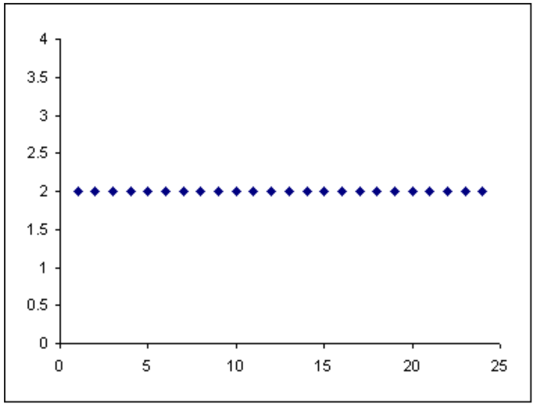
LINEAR CORRELATION (FORMULA 1)

- $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$, **where**
  - s_x **and** s_y**: are the standard deviations of** x**'s and** y**'s**
  - $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- **Examples of extreme cases**

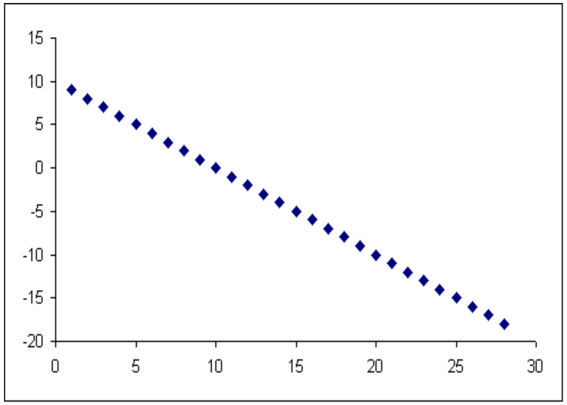
$r = 1$



$r = 0$



$r = -1$



LINEAR CORRELATION (FORMULA 2)

- $r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$ **(*), where**
  - $SS(x) = \sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$
  - $SS(y) = \sum_{i=1}^{n} y_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$
  - $SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_i x_i\right)\left(\sum_i y_i\right)$

OR:

- $SS(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $SS(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SS(xy) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- **(*) is equivalent to the first formula:** $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$
- **Example 5: In Mr. Chamberlain's physical-fitness course, several fitness scores were taken. The following sample is the numbers of push-ups and sit-ups done by 10 randomly selected students:**

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

EXAMPLE 5 – *SOLUTION*

- **Find the linear correlation coefficient for the push-up/sit-up data.**

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\sum x = 351$ sum of x	$\sum x^2 = 13{,}717$ sum of x^2	$\sum y = 380$ sum of y	$\sum y^2 = 15{,}298$ sum of y^2	$\sum xy = 14{,}257$ sum of xy

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 13{,}717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(y) = \sum y^2 - \frac{(\sum y)^2}{n} = 15{,}298 - \frac{(380)^2}{10} = 858.0$$

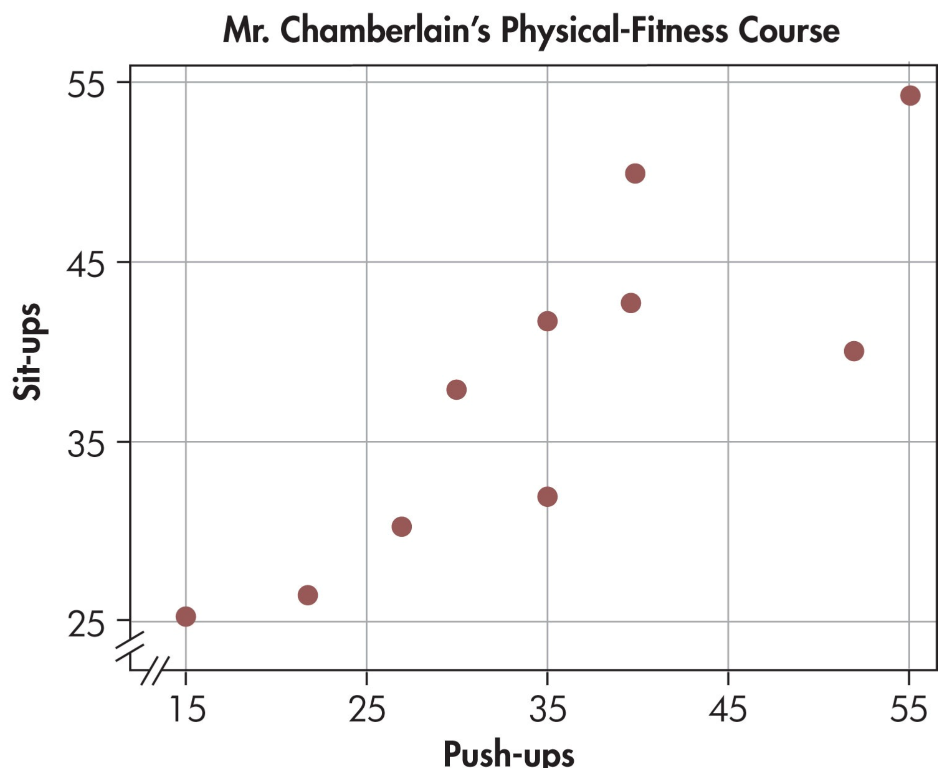
$$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} = 14{,}257 - \frac{(351)(380)}{10} = 919.00$$

EXAMPLE 5 – *SOLUTION*

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

$$= \frac{919.0}{\sqrt{(1396.9)(858.0)}}$$

$$= 0.8394 = \mathbf{0.84}$$



RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

- **Correlation** or r : measures the direction and strength of the **linear** relationship between two numeric variables

- **General Properties**
  - It must be between -1 and 1, or $(-1 \leq r \leq 1)$.
  - If r is negative, the relationship is negative.
  - If $r = -1$, there is a perfect negative linear relationship (extreme case).
  - If r is positive, the relationship is positive.
  - If $r = 1$, there is a perfect positive linear relationship (extreme case).
  - If r is 0, there is no **linear** relationship.
  - r measures the strength of the **linear** relationship.

- Correlation Applet

CAUSATION AND LURKING VARIABLES

- The **cause-and-effect relationship:** Correlation does not necessarily imply causation. Just because two things are highly related does not mean that one causes the other.

- A perceived relationship between a dependent(response) variable and an independent(explanatory) variable that has been misestimated due to the failure to account for a confounding factor (lurking variable) is termed a spurious relationship

- Examples of spurious relationship

LURKING VARIABLE AND SIMPSON'S PARADOX

MARQUETTE UNIVERSITY
Be The Difference.

- **Lurking variable:** A variable that is not included in a study but has an effect on the variables of the study and makes it appear that those variables are related.
- **Simpson's Paradox:** An association or comparison that holds for all of several groups can reverse direction when a lurking variable is present.
- **Example: Kidney stone treatment(** Br Med J (Clin Res Ed) 292 (6524): 879–882 **)**

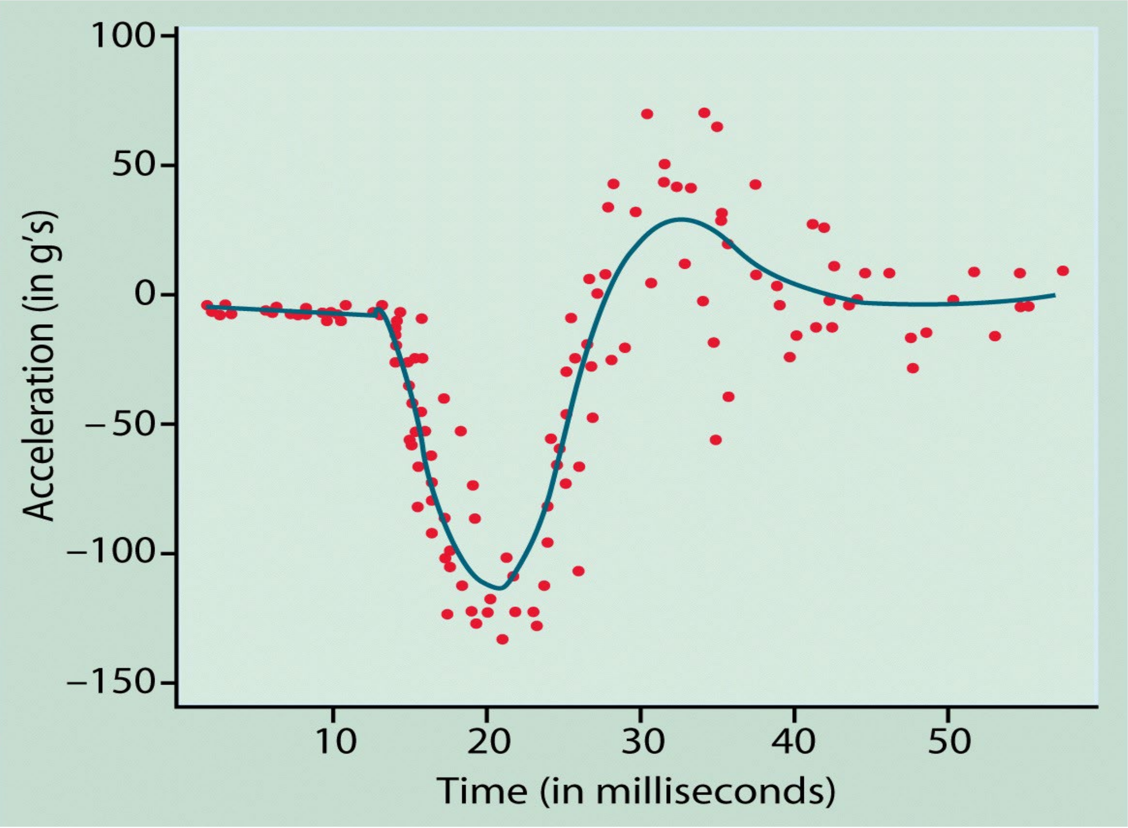
	Treatment A	Treatment B
Small Stones	*Group 1* 93% (81/87)	*Group 2* 87% (234/270)
Large Stones	*Group 3* 73% (192/263)	*Group 4* 69% (55/80)
Both	78% (273/350)	83% (289/350)

- http://en.wikipedia.org/wiki/Simpson's_Paradox

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

It is possible for there to be a strong relationship between two variables and still have $r \approx 0$.

EX.



LINEAR REGRESSION

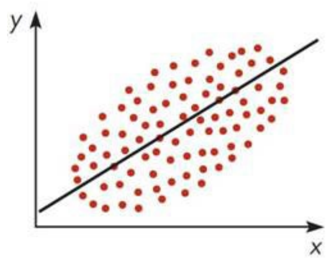
- **Regression analysis** finds the equation of the line that best describes the relationship between two variables.

- Here are some examples of various possible relationships, called *models* or **prediction equations:**
  - Linear (straight-line): $\hat{y} = b_0 + b_1 x$
  - Quadratic: $\hat{y} = a + bx + cx^2$
  - Exponential: $\hat{y} = a(b^x)$
  - Logarithmic: $\hat{y} = a \log_b x$

What we cover in this book

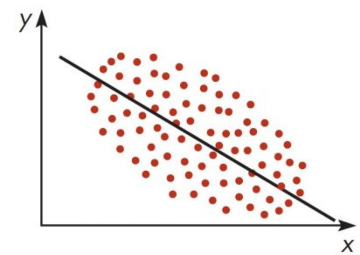
LINEAR REGRESSION

- **Suppose that $\hat{y} = b_0 + b_1 x$ is the equation of a straight line, where $\hat{y}$ (read "y-hat") represents the predicted value of y that corresponds to a particular value of x.**



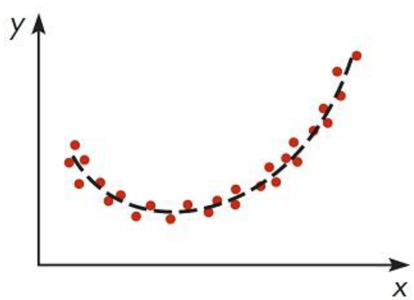
Linear Regression with Positive Slope

Figure 3.17



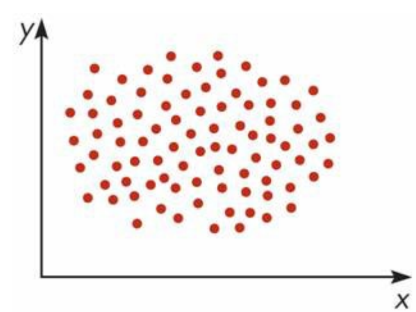
Linear Regression with Negative Slope

Figure 3.18



Curvilinear Regression (Quadratic)

Figure 3.19

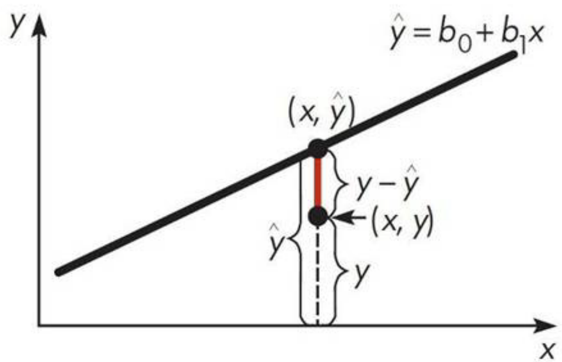


No Relationship

Figure 3.20

LINEAR REGRESSION

- The **least squares criterion** requires that we find the constants b_0 and b_1 such that $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is as small as possible.

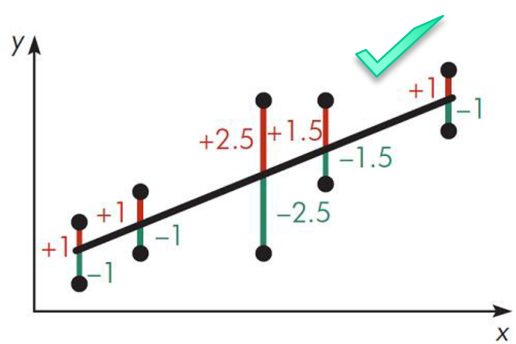


Observed and Predicted Values of *y*

- The length of this distance represents the value $(y_i - \hat{y})$ (shown as the red line segment in the Figure. We call it **residual**).
- Note that $(y_i - \hat{y})$ is **positive** when the point (x, y) is above the line and **negative** when (x, y) is below the line

LINEAR REGRESSION

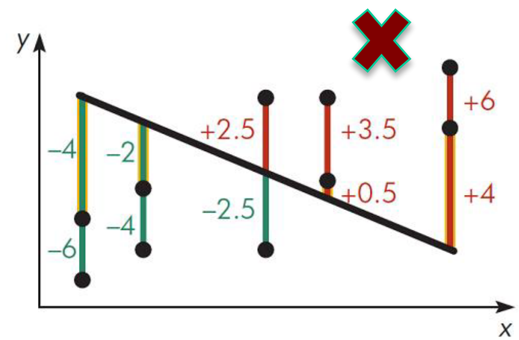
- Figure 3.22 shows a scatter diagram with what appears to be the **line of best fit**, along with 10 individual $(y_i - \hat{y})$ values. (Positive values are shown in red; negative, in green.)



$\Sigma (y - \hat{y})^2 = (-1)^2 + (+1)^2 + \ldots + (+1)^2 = 23.0$

The Line of Best Fit

Figure 3.22



$\Sigma (y - \hat{y})^2 = (-6)^2 + (-4)^2 + \ldots + (+6)^2 = 149.0$

Not the Line of Best Fit

Figure 3.23

- Figure 3.23 shows the same data points as Figure 3.22. The 10 individual values $(y_i - \hat{y})$ are plotted with a line that is definitely not the line of best fit. [Applet]

LINEAR REGRESSION

- **Our job is to find the one line that will make $\sum_{i=1}^{n}(y_i - \hat{y})^2$ the smallest possible value.**

- **The equation of the line of best fit is determined by its slope (b_1) and its y-intercept (b_0).**

- **Slope:** $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS(xy)}{SS(x)}$, **where**
  - $SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_i x_i\right)\left(\sum_i y_i\right)$
  - $SS(x) = \sum_{i=1}^{n} {x_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$

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

LINEAR REGRESSION

- **Example-5: push-up/sit-up data.**

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\sum x = 351$ *sum of x*	$\sum x^2 = 13,717$ *sum of x^2*	$\sum y = 380$ *sum of y*	$\sum y^2 = 15,298$ *sum of y^2*	$\sum xy = 14,257$ *sum of xy*

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 13,717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} = 14,257 - \frac{(351)(380)}{10} = 919.00$$

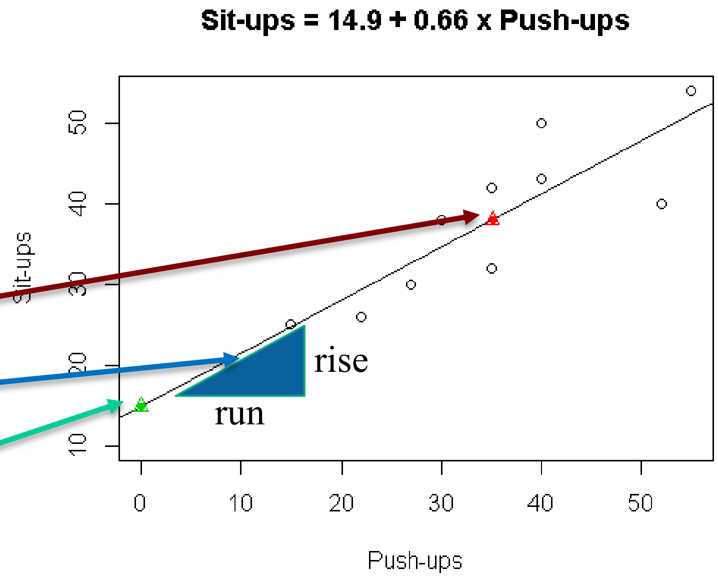
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{351}{10} = 35.1$
- $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{380}{10} = 38$

LINEAR REGRESSION (EXAMPLE)

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

- **Give** $\bar{x} = 35.1$, $\bar{y} = 38$, $SS(x) = 1396.9$ **and** $SS(xy) = 919$

- **We want to find the line of best fit,** $\hat{y} = b_0 + b_1 x$**, where**
- $b_1 = \dfrac{SS(xy)}{SS(x)} = \dfrac{919}{1396.9} = 0.6579 = 0.66$
- $b_0 = \bar{y} - b_1\bar{x} = 38 - 0.6579 \times 35.1 = 14.9077 = 14.9$
- **Therefore**
- $\hat{y} = 14.9 + 0.66x$



- **Important Notes:**
  - **The line goes through** $(\bar{x}, \bar{y})$
  - **The slope:** $b_1 = 0.66$
  - **The y-intercept:** $b_0 = 14.9$

LINEAR REGRESSION

- **Notes**

1. **Remember to keep at least three extra decimal places while doing the calculations to ensure an accurate answer.**

2. **When rounding off the calculated values of b_0 and b_1, always keep at least two significant digits in the final answer.**

3. **The slope, b_1, represents the predicted change in y per unit increase in x.**
   - **In our example, where $b_1 = 0.66$, if a student can do an additional 10 push-ups (x), we predict that he or she would be able to do approximately $7\ (0.66 \times 10)$ additional sit-ups (y).**

4. **The y-intercept is the value of y where the line of best fit intersects the y-axis.**

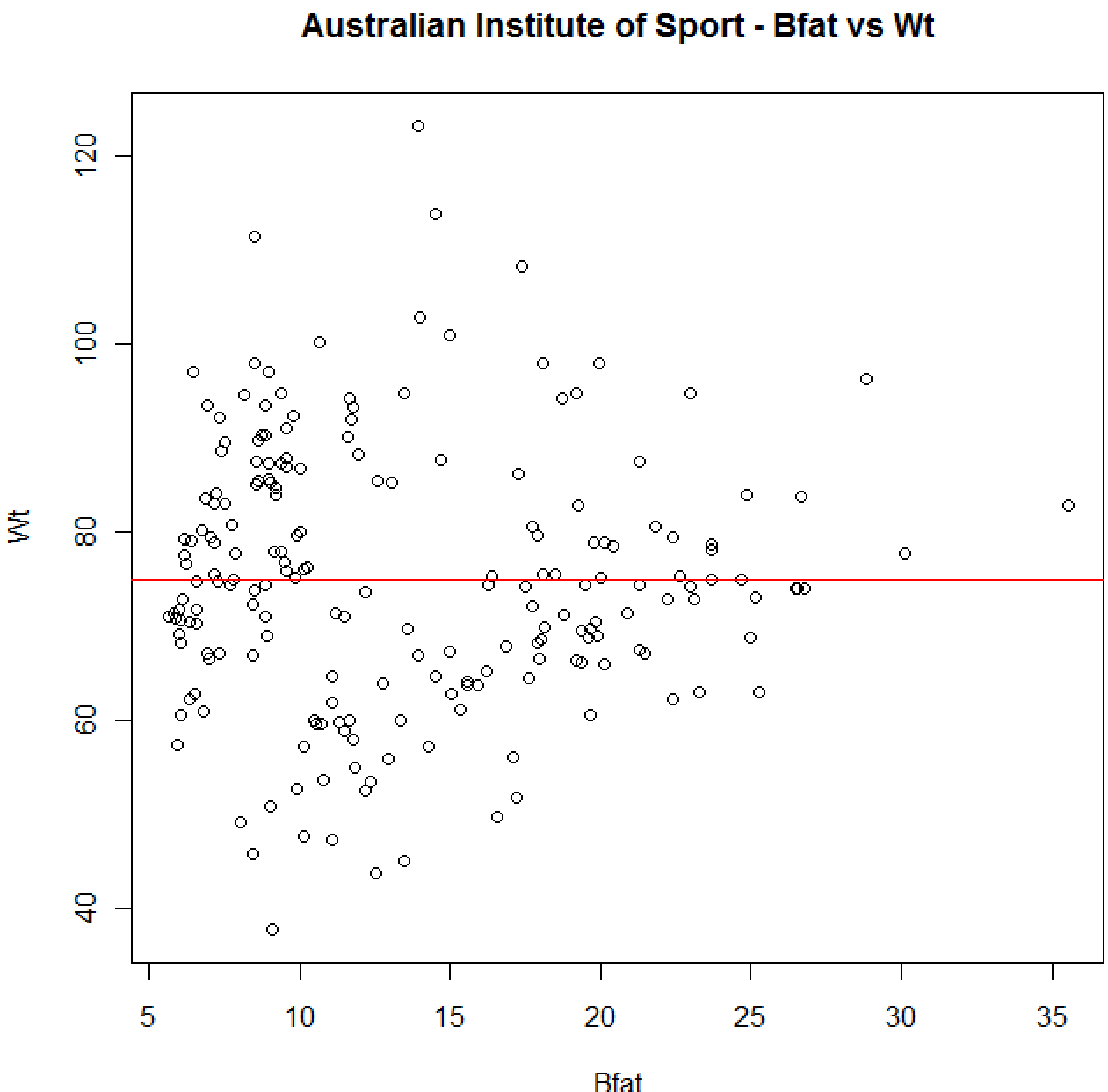
EXAMPLE – AUSTRALIAN INSTITUTE OF SPORT

MARQUETTE UNIVERSITY
Be The Difference.

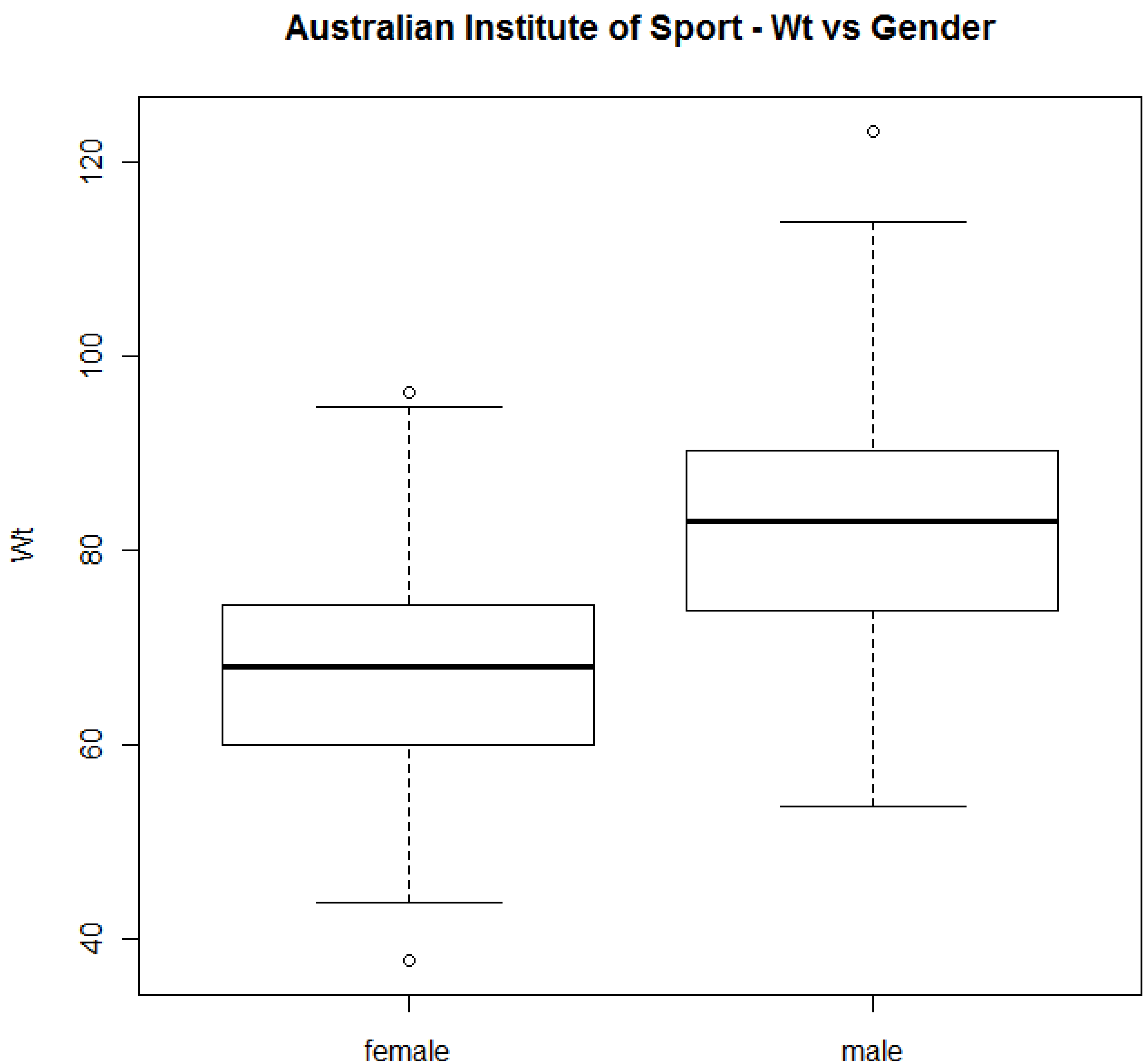
- **Data on 102 male and 100 female athletes collected at the Australian Institute of Sport,** (courtesy of Richard Telford and Ross Cunningham.)

	Gender	Bfat	Wt
1	female	19.75	78.9
2	female	21.30	74.4
3	female	19.88	69.1
4	female	23.66	74.9
5	female	17.64	64.6
	:	:	:
198	male	11.79	93.2
199	male	10.05	80.0
200	male	8.51	73.8
201	male	11.50	71.1
202	male	6.26	76.7

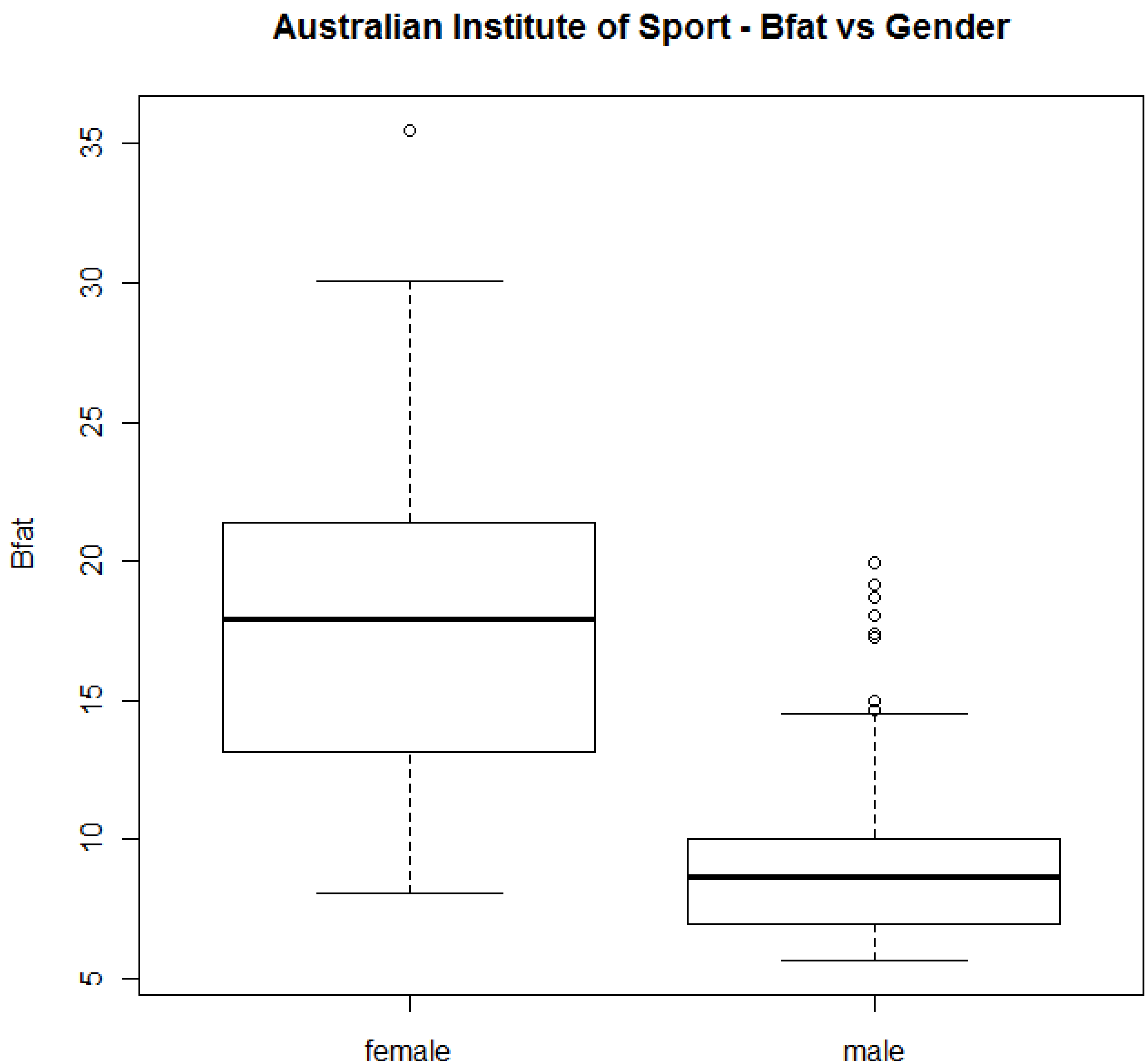
EXAMPLE CONT'D



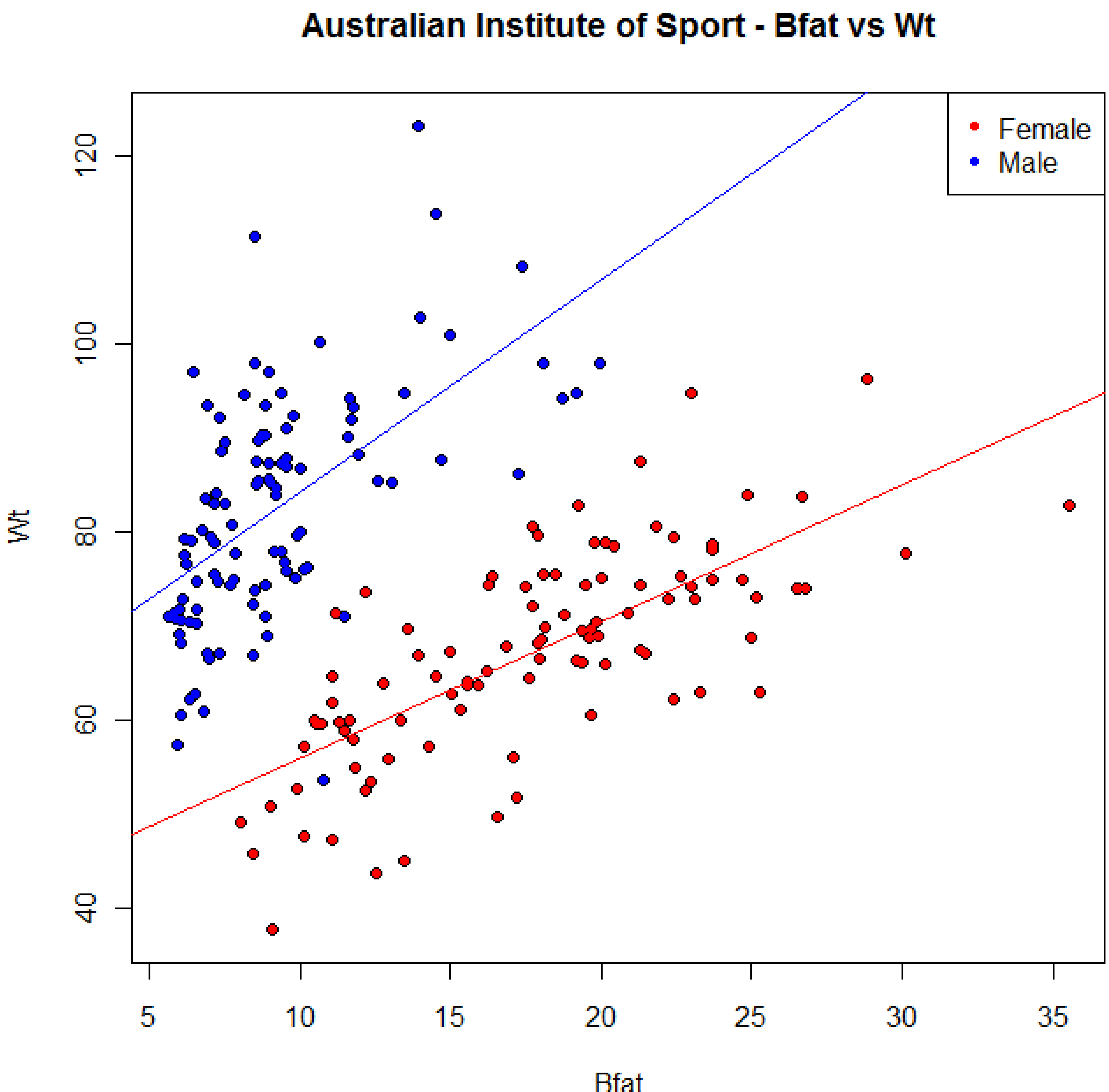
EXAMPLE CONT'D



EXAMPLE CONT'D

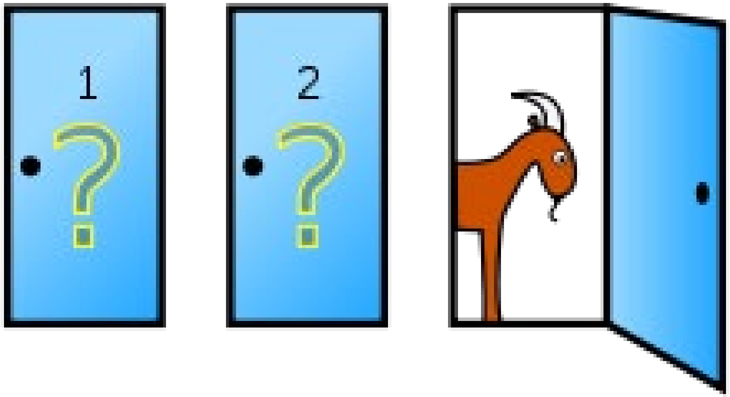


EXAMPLE CONT'D



LET'S MAKE A DEAL (PREPARE FOR CHAPTER 4)

- Let's Make a Deal (Monty Hall problem)
  - http://www.mathwarehouse.com/monty-hall-simulation-online/

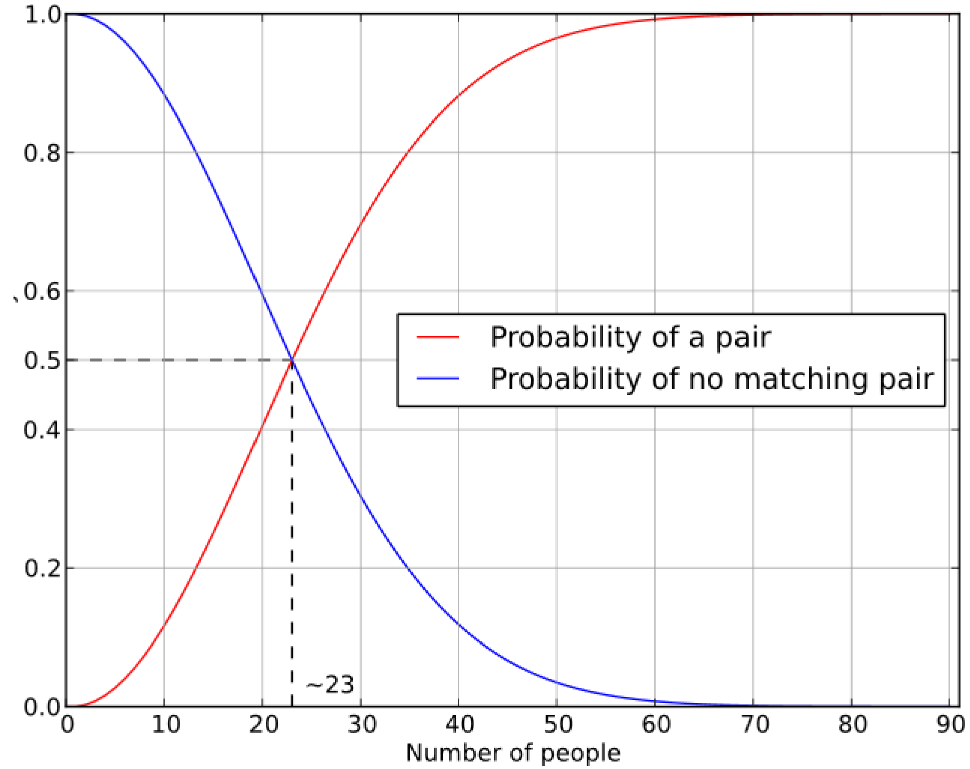


- This is motivation to study probability.

- Should you switch or should you stay with your original choice?

BIRTHDAY PARADOX (PREPARE FOR CHAPTER 4)

- What's the chances that two people in our class have the same birthday?



QUESTIONS?

- **ANY QUESTION?**