Class 4

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
Agenda:

Recap Chapter 2.5, 3.1

Lecture Chapter 3.2, 3.3
Recap Chapter 2.5
2. Descriptive Analysis and Single Variable Data
2.5 Measures of Position

**Measures of Position: Quartiles** - ranked data into quarters

$L = $ lowest value
$H = $ highest value
$Q_2 = $ median
$Q_1 = 25\%$ smaller
$Q_3 = 75\%$ smaller
$\text{IQR} = Q_3 - Q_1$

Rowe, D.B.
2: Descriptive Analysis and Single Variable Data

2.5 Measures of Position

Measures of Position: percentiles - rank data into 100ths

$L = \text{lowest value}$

$H = \text{highest value}$

$P_k = \text{value where } k\% \text{ are smaller}$

Rank data

\[
\frac{nk}{100}
\]

integer, \( A \)

not Integer, \( B \)

\( p_k \) halfway between value and next one average of \( A^{th} \) and \((A+1)^{th} \) values

\( p_k \) is value in next largest position, \( B \) value

Figure from Johnson & Kuby, 2012.

Rowe, D.B.
2: Descriptive Analysis and Single Variable Data
2.5 Measures of Position

**Standard score, or z-score:** The position a particular value of $x$ has relative to the mean, measured in standard deviations.

$$Z_i = \frac{i^{th}\text{ value} - \text{mean}}{\text{std. dev.}} = \frac{x_i - \bar{x}}{s}$$

There can be $n$ of these because we have $x_1, x_2, \ldots, x_n$. 

2: Descriptive Analysis and Single Variable Data

Questions?

Homework: Read Chapter 2.5-2.7
Chapter 2 # 115, 123c-d, 129, 137
Recap Chapter 3.1
3: Descriptive Analysis and Bivariate Data
3.1 Bivariate Data: two qualitative

Cross-tabulation tables or contingency tables

Example:
Construct a 2×3 table.

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>M</td>
<td>LA</td>
<td>Feeney</td>
<td>M</td>
<td>T</td>
<td>McGowan</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Argento</td>
<td>F</td>
<td>BA</td>
<td>Flanigan</td>
<td>M</td>
<td>LA</td>
<td>Mowers</td>
<td>F</td>
<td>BA</td>
</tr>
<tr>
<td>Baker</td>
<td>M</td>
<td>LA</td>
<td>Hodge</td>
<td>F</td>
<td>LA</td>
<td>Ort</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>Bennett</td>
<td>F</td>
<td>LA</td>
<td>Holmes</td>
<td>M</td>
<td>T</td>
<td>Palmer</td>
<td>F</td>
<td>LA</td>
</tr>
<tr>
<td>Brand</td>
<td>M</td>
<td>T</td>
<td>Jopson</td>
<td>F</td>
<td>T</td>
<td>Pullen</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>Brock</td>
<td>M</td>
<td>BA</td>
<td>Kee</td>
<td>M</td>
<td>BA</td>
<td>Rattan</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Chun</td>
<td>F</td>
<td>LA</td>
<td>Kleeberg</td>
<td>M</td>
<td>LA</td>
<td>Sherman</td>
<td>F</td>
<td>LA</td>
</tr>
<tr>
<td>Crain</td>
<td>M</td>
<td>T</td>
<td>Light</td>
<td>M</td>
<td>BA</td>
<td>Small</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>Cross</td>
<td>F</td>
<td>BA</td>
<td>Linton</td>
<td>F</td>
<td>LA</td>
<td>Tate</td>
<td>M</td>
<td>BA</td>
</tr>
<tr>
<td>Ellis</td>
<td>F</td>
<td>BA</td>
<td>Lopez</td>
<td>M</td>
<td>T</td>
<td>Yamamoto</td>
<td>M</td>
<td>LA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LA</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>BA</td>
</tr>
<tr>
<td>M</td>
<td>6</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>T</td>
</tr>
<tr>
<td>M</td>
<td>7</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Row Total</td>
</tr>
<tr>
<td>M</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Col. Total</td>
</tr>
<tr>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

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

Figures from Johnson & Kuby, 2012.
3: Descriptive Analysis and Bivariate Data
3.1 Bivariate Data: one qualitative and one quantitative

Example:

<table>
<thead>
<tr>
<th>Design A (n = 6)</th>
<th>Design B (n = 6)</th>
<th>Design C (n = 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37 36 38</td>
<td>33 35 38</td>
<td>40 39 40</td>
</tr>
<tr>
<td>34 40 32</td>
<td>34 42 34</td>
<td>41 41 43</td>
</tr>
</tbody>
</table>

Figures from Johnson & Kuby, 2012.

Vertical box-and-whiskers

Rowe, D.B.
3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: two quantitative, Scatter Diagram

Example: Push-ups

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Push-ups, $x$</td>
<td>27</td>
<td>22</td>
<td>15</td>
<td>35</td>
<td>30</td>
<td>52</td>
<td>35</td>
<td>55</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Sit-ups, $y$</td>
<td>30</td>
<td>26</td>
<td>25</td>
<td>42</td>
<td>38</td>
<td>40</td>
<td>32</td>
<td>54</td>
<td>50</td>
<td>43</td>
</tr>
</tbody>
</table>

Input variable: independent variable, $x$.
Output variable: dependent variable, $y$.

Scatter Diagram: A plot of all the ordered pairs of bivariate data on a coordinate axis system.

$(x, y)$ ordered pairs.

Figures from Johnson & Kuby, 2012.
3: Descriptive Analysis and Bivariate Data

Questions?

Homework: Read Chapter 3
   Chapter 3 # 3, 7, 15
3: Descriptive Analysis and Bivariate Data
3.1 Bivariate Data: two quantitative, Scatter Diagram

Our data.

Gender, height, weight.

Use height vs. weight (no gender).
3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: Scatter Diagram

Our data.

Height vs. Weight

Find yourself. If not here then removed or you did not respond.

$n=80$ responses
Chapter 3: Descriptive Analysis and Presentation of Bivariate Data continued

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
3: Descriptive Analysis and Bivariate Data

3.2 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 its computation in a minute.

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

Figure from Johnson & Kuby, 2012.

Rowe, D.B.
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Values closer to +1 or -1 mean a stronger relationship.

Values near 0 mean a weak association.

Positive values mean a positive relationship. Positive relationship means as $x$ increases so does $y$.

Negative values mean a negative relationship. Negative relationship means as $x$ increases $y$ decreases
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

$r = 1$

Perfect Positive Correlation

$r = -1$

Perfect Negative Correlation

Horizontal—No Correlation

Vertical—No Correlation

Figure from Johnson & Kuby, 2012.
Computing the linear correlation coefficient $r$.

1. \[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \]

\[ s_x = \text{std} \ x's \quad s_y = \text{std} \ y's \]

2. \[ r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} \]

\[ SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \]

\[ SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} y_i \right)^2 \]

\[ SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) \]

1. and 2. are equivalent.

Rowe, D.B.
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example:

Rowe, D.B.

\[
\sum_{i=1}^{n} x_i 
\sum_{i=1}^{n} x_i^2 
\sum_{i=1}^{n} y_i 
\sum_{i=1}^{n} y_i^2 
\sum_{i=1}^{n} x_i y_i 
\]

<table>
<thead>
<tr>
<th>Student</th>
<th>Push-ups, x</th>
<th>( x^2 )</th>
<th>Sit-ups, y</th>
<th>( y^2 )</th>
<th>( xy )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>729</td>
<td>30</td>
<td>900</td>
<td>810</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>484</td>
<td>26</td>
<td>676</td>
<td>572</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>225</td>
<td>25</td>
<td>625</td>
<td>375</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>1,225</td>
<td>42</td>
<td>1,764</td>
<td>1,470</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>900</td>
<td>38</td>
<td>1,444</td>
<td>1,140</td>
</tr>
<tr>
<td>6</td>
<td>52</td>
<td>2,704</td>
<td>40</td>
<td>1,600</td>
<td>2,080</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>1,225</td>
<td>32</td>
<td>1,024</td>
<td>1,120</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>3,025</td>
<td>54</td>
<td>2,916</td>
<td>2,970</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>1,600</td>
<td>50</td>
<td>2,500</td>
<td>2,000</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>1,600</td>
<td>43</td>
<td>1,849</td>
<td>1,720</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sum x &= 351 \\
\sum x^2 &= 13,717 \\
\sum y &= 380 \\
\sum y^2 &= 15,298 \\
\sum xy &= 14,257
\end{align*}
\]

Figure from Johnson & Kuby, 2012.
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example:

<table>
<thead>
<tr>
<th>Push-ups, x</th>
<th>27</th>
<th>22</th>
<th>15</th>
<th>35</th>
<th>30</th>
<th>52</th>
<th>35</th>
<th>55</th>
<th>40</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sit-ups, y</td>
<td>30</td>
<td>26</td>
<td>25</td>
<td>42</td>
<td>38</td>
<td>40</td>
<td>32</td>
<td>54</td>
<td>50</td>
<td>43</td>
</tr>
</tbody>
</table>

\[
SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 = 
\]

\[
SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} y_i \right)^2 = 
\]

\[
SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) = 
\]

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

\[
\sum_{i=1}^{n} x_i = 351 \\
\sum_{i=1}^{n} x_i^2 = 13717 \\
\sum_{i=1}^{n} y_i = 380 \\
\sum_{i=1}^{n} y_i^2 = 15298 \\
\sum_{i=1}^{n} x_i y_i = 14257
\]

Figures from Johnson & Kuby, 2012.
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example: Our data!

\[
SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2
\]

\[
SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)
\]

\[
\sum_{i=1}^{n} x_i = 5350 \quad SS(x) = 1582.8
\]

\[
\sum_{i=1}^{n} x_i^2 = 359364
\]

\[
\sum_{i=1}^{n} y_i = 11503 \quad SS(y) = 4871.1
\]

\[
\sum_{i=1}^{n} y_i^2 = 1702699
\]

\[
\sum_{i=1}^{n} x_i y_i = 775934 \quad SS(xy) = 6670.9
\]

\[
r = \frac{SS(xy)}{\sqrt{SS(x) SS(y)}} = 0.76
\]
3: Descriptive Analysis and Bivariate Data
3.2 Linear Correlation

Understanding Linear Correlation
Skip for now. Read on own.

Causation and Lurking Variables
Correlation does not necessarily imply causation. Just because two things are highly related does not mean that one causes the other.

Soda sales go up, flu incidence goes down. Does soda cause flu to go down?
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

Regression analysis finds the equation of a line that “best” describes the relationship between the two variables ($x$ and $y$).

What do we mean by “best?”

How is “bestness” determined?

Least squares regression.
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

Let’s say that we are given points as in figure.

Imagine that there is an underlying line

\[ y = \beta_0 + \beta_1 x \]

that the data fits to (or comes from).

\( \beta_0 \) is \( y \)-intercept and \( \beta_1 \) is slope.

The points are considered to be

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

We want to find the “best” fit line to the data.

Rowe, D.B.
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

We can try different lines until we find the “best” one.

Let’s call the “best” one \( \hat{y} = b_0 + b_1 x \).

Imagine that there is an underlying line

\[ y = \beta_0 + \beta_1 x \]

that the data fits to (or comes from).

\( \beta_0 \) is \( y \)-intercept and \( \beta_1 \) is slope.

The points are considered to be

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n \]

\( b_0 \) is estimated \( y \)-intercept and \( b_1 \) is estimated slope.
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

What is criteria for bestness? → Sum of squared distances.

The points are considered to be

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

The “best” line value at \( x_i \) is

\[ \hat{y}_i = b_0 + b_1 x_i \]

These vertical distances \( \varepsilon_i \) are called residuals.

\[ \varepsilon_i = y_i - b_0 - b_1 x_i \]

\( i = 1, \ldots, n \)
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

What is criteria for bestness? → Sum of squared distances.

\[ \hat{y} = b_0 + b_1 x \]

We move around the line until the sum of the squared residuals

\[ \sum_{i=1}^{n} \varepsilon_i^2 \]

is made a minimum.

Least squares line.

This is a measure of misfit and a criterion for the “best“ line.
We don't actually have to move the line around.

We can find the “best” fit line that minimizes the sum of the squared residuals by using Equations 3.5-3.7a.

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

or

$$b_1 = \frac{SS(xy)}{SS(x)}$$

then $$b_0 = \bar{y} - b_1 \bar{x}$$ because line goes through $$(\bar{x}, \bar{y})$$.
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Using (1,1),(3,2),(2,3),(4,4) \( x = 2.5, \ y = 2.5 \)

\[ b_1 = \frac{SS(xy)}{SS(x)} \]

\[ SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) \]

\[ \sum_{i=1}^{n} x_i = \quad , \quad \sum_{i=1}^{n} y_i = \quad , \quad \sum_{i=1}^{n} x_i y_i = \]

\[ SS(xy) = \]
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

Example: Using \((1,1),(3,2),(2,3),(4,4)\)

\[
\bar{x} = 2.5, \quad \bar{y} = 2.5
\]

\[
b_1 = \frac{SS(xy)}{SS(x)}
\]

\[
SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2
\]

\[
\sum_{i=1}^{n} x_i = \quad , \quad \sum_{i=1}^{n} x_i^2 =
\]

\[
SS(x) =
\]
3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

**Example:** Using \((1,1),(3,2),(2,3),(4,4)\)

\[
\bar{x} = 2.5, \quad \bar{y} = 2.5
\]

\[
b_1 = \frac{SS(xy)}{SS(x)}
\]

\[
b_0 = \bar{y} - b_1 \bar{x} =
\]

point-slope formula
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

\((x, y)\) pairs: (1,1), (3,2), (2,3), (4,4)

The line goes through \((\bar{x}, \bar{y})\).

The slope is \(b_1\) =

The y-intercept \(b_0\) =

Two points
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Our data!

Height vs. Weight

\[ b_1 = \frac{SS(xy)}{SS(x)} = \frac{6708.8}{1582.7} = 4.2 \]

\[ b_0 = (143.8) - (4.2)(66.9) = -138.1 \]

units of lbs/in

point-slope formula
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Our data!

Height vs. Weight

\[
\hat{y} = -138.1 + 4.2x
\]

So if a new student adds the class and is \(x=70\) in tall, the best guess for his/her weight is about \(y=157.0\) lbs.
3: Descriptive Analysis and Bivariate Data

Questions?

Homework: Chapter 3 # 33, 44, 53, 59, 75
Read 4.1 and 4.2