Class 5

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
Agenda:

Recap Chapter 3.2 - 3.3

Lecture Chapter 4.1 - 4.2

Review Chapter 1 - 2.5
(Exam 1 Chapters)
Recap Chapter 3.2 - 3.3
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 $y$.

**positive relationship:** as $x$ increases so does $y$

**negative relationship:** as $x$ increases $y$ decreases

\[-1 \leq r \leq 1\]

![Graph showing different correlation strengths: No correlation, Positive, High positive, Negative, High negative.](Figure from Johnson & Kuby, 2012.)

Rowe, D.B.
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>

\[ SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 = 13717 - \frac{(351)^2}{10} = 1396.9 \]

\[ SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} y_i \right)^2 = 15298 - \frac{(380)^2}{10} = 858.0 \]

\[ SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) = 14257 - \frac{(351)(380)}{10} = 919.0 \]

\[ r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} = \frac{919.0}{\sqrt{(1396.9)(858.0)}} = 0.84 \]

Questions?

Figure from Johnson & Kuby, 2012.

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

3.3 Linear Regression

We try different lines until we find the “best” one, \( \hat{y} = b_0 + b_1x \)

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

3.3 Linear Regression

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

Plotted points.

The line goes through (x̄, ȳ).

The slope is \( b_1 = 0.8 \).

The y-intercept \( b_0 = 0.5 \).

Two points (2.5, 2.5) and (0, 5).

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

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

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

\[
b_0 = \bar{y} - b_1 \bar{x}
\]
3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Our data!

Height vs. Weight

\[
b_1 = \frac{SS(xy)}{SS(x)} = \frac{2456.7}{472.4} = 5.2
\]

\[
b_0 = (143.2) - (5.2)(66.7) = -203.7
\]

units of lbs/in

point-slope formula
Chapter 4: Probability

Daniel B. Rowe, Ph.D.

Department of Mathematics, Statistics, and Computer Science
4: Probability
4.1 Probability of Events

Let’s talk about experiments, events, and probabilities.

An experiment is a process by which a measurement is taken or observations is made.

i.e. flip coin or roll die
4: Probability
4.1 Probability of Events

An outcome is the result of an experiment. i.e. Heads, or 3

Coin: $O_1=H$, $O_2=T$
Die: $O_1=1$, $O_2=2$, $O_3=3$, $O_4=4$, $O_5=5$, $O_6=6$

Sample space is a listing of possible outcomes.
$S=\{O_1,O_2\}$ or $S=\{O_1,O_2,O_3,O_4,O_5,O_6\}$

Coin: $S=\{H,T\}$
Die: Coin: $S=\{1,2,3,4,5,6\}$
4: Probability
4.1 Probability of Events

An event $A$ is an outcome or a combination of outcomes. i.e. $A=$ *even number when rolling a die* =\{2,4,6\}

The probability of an event $A$ is written $P(A)$.

i.e. $P(A) = P(\text{even number when rolling a die})$

\[ = P(\{2,4,6\}) \\
= \frac{3}{6} \]
4: Probability
4.1 Probability of Events - Properties

Property 1
In words:

“A probability is always a numerical value between 0 and 1.”

In algebra:

\[ 0 \leq P(A) \leq 1 \]

If the event \( A \) can never occur, then \( P(A) = 0 \).
If the event \( A \) is sure to occur, then \( P(A) = 1 \).
4: Probability
4.1 Probability of Events - Properties

Property 2

In words:

“The sum of probabilities for all outcomes of an experiment is equal to exactly 1.”

In algebra: \( \sum_{i=1}^{n} P(O_i) = 1 \) where \( i = 1, ..., n \) and \( O_i \) are nonoverlapping events that include all possibilities.

The book uses \( A \).
4: Probability

4.1 Probability of Events

Now that we talked about events and probabilities, how do we get probabilities of events?

Probability of a event: The relative frequency with which that event can be expected to occur.

There are three different approaches to probability.
(1) Empirical (AKA experimental)
(2) Theoretical (AKA classical or equally likely)
(3) Subjective (expression of belief, not discuss)
4: Probability
4.1 Probability of Events

Empirical (Observed) Probability: $P'(A)$

In words:

empirical probability of $A = \frac{\text{number of times } A \text{ occurred}}{\text{number of trials}}$

In algebra:

$$P'(A) = \frac{n(A)}{n}$$

The “’” in $P'(A)$ means an empirical probability.
4: Probability

4.1 Probability of Events – Law of large numbers

Had computer flip a single coin 1000 times.

Flip # on $x$ axis

$P'(H)$ on $y$ axis.

This shows convergence to true value of $1/2$. 

$P'(H) = \frac{\text{# of heads}}{\text{# coin flips}}$
4: Probability
4.1 Probability of Events

In the empirical method you actually have to perform the experiment of flipping the coin.

The empirical approach may be off in the short run.

Suppose you get on a streak and out of 10 flips all 10 are heads?

By the empirical method we would say that $P'(H) = 1$. 
4: Probability
4.1 Probability of Events

**Theoretical (Expected) Probability: \( P(A) \)**

In words:
theoretical probability of \( A = \frac{\text{number of times } A \text{ occurs in sample space}}{\text{number of elements in the sample space}} \)

In algebra:

\[
P(A) = \frac{n(A)}{n(S)}
\]
4: Probability
4.1 Probability of Events
So let's examine what could potentially happen when we flip a coin twice.

Before flip.
4: Probability
4.1 Probability of Events
So let's flip a coin twice.

Flip once.

Sample space: listing of outcomes for 1 flip

\[ S = \{H, T\} \]

\[ P(H) = \frac{\text{# times } H \text{ occurs in } S}{\text{# elements in } S} \]
4: Probability
4.1 Probability of Events
So let's flip a coin twice.

Flip twice.

Sample space: listing of outcomes for 2 flips

\[ S = \{HH, HT, TH, TT\} \]

\[ P(HH) = \frac{\# \text{ times } HH \text{ occurs in } S}{\# \text{ elements in } S} \]
4: Probability
4.1 Probability of Events

So let's flip a coin twice.

Can flip three times.

Sample space: listing of outcomes for 2 flips

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

$$P(HHH) = \frac{\text{# times } HHH \text{ occurs in } S}{\text{# elements in } S}$$
4: Probability
4.1 Probability of Events

In the theoretical method you do not have to perform the experiment of flipping the coin.

If each of the events are equally likely, then the theoretical approach is correct from the start.

If the events are not equally likely, then the theoretical method is not correct and we should use a different approach.
4: Probability
4.1 Probability of Events – Probabilities as odds

If the odds in favor of an event $A$ are $a$ to $b$ (or $a:b$), then

1. The odds against event $A$ are $b$ to $a$ (or $b:a$).

2. The probability of event $A$ is $P(A) = \frac{a}{a + b}$.

3. The probability that event $A$ will not occur is $P(\text{not } A) = \frac{b}{a + b}$.
4: Probability
4.1 Probability of Events

The odds against event $A$ are $b$ to $a$ (or $b:a$). The probability of event $A$ is $P(A) = \frac{a}{a + b}$.

Therefore, if we are at the race track and we define

$A = $ our horse wins the race.

If the odds against $A$ are 100 to 1 (100:1),

then the probability of $A$ is $P(A) = \frac{1}{100 + 1} = \frac{1}{101}$.
4: Probability
4.2 Conditional Probability of Events

We use conditional probability in our daily lives and sometimes do not realize it.

What is the probability that the Professor will put an exam question on topic $x$?

What is the probability that the Professor will put an exam question on topic $x$ given that he covered topic $x$ in class?
4: Probability
4.2 Conditional Probability of Events

What is the probability that the Professor will put an exam question on topic \( x \)?

What is the probability that the Professor will put an exam question on topic \( x \) given that he covered topic \( x \) in class?

Let \( A = \) Professor will put an exam question on topic \( x \)
\( B = \) he covered topic \( x \) in class

\[ P(A) \text{ vs. } P(A|B) \]
Conditional probability an event will occur: A conditional probability is the relative frequency with which an event can be expected to occur under the condition that that additional preexisting information is known about some other event.

\[ P(A \mid B), \text{ the “|” is spoken as “given” or “knowing”} \]
4: Probability

4.2 Conditional Probability of Events

**Example:** Roll two die.
Let \( A \) be that 10 is the sum of the two die.

\[
P(A) = \frac{n(A)}{n(S)}
\]

Figure from Johnson & Kuby, 2012.
4: Probability
4.2 Conditional Probability of Events

Example: Roll two die.
Let $A$ be that 10 is the sum of the two die.

$$P(A) = \frac{3}{36}.$$
4: Probability
4.2 Conditional Probability of Events

Example: Roll two die.

Let $B$ that the first die is a 4.

$P(B) = \frac{6}{36}.$

Figure from Johnson & Kuby, 2012.
4: Probability
4.2 Conditional Probability of Events

**Example**: Roll two die.
Let $A$ be that 10 is the sum of the two die.
Let $B$ that the first die is a 4.

$P(A|B) = \frac{1}{6}$. 

Figure from Johnson & Kuby, 2012.
4: Probability

Questions?

Homework: Chapter 4 # 3, 11, 12, 13, 31, 51, 57
Finish Reading Chapter 4
Review Chapters 1 – 2.5
(Exam 1 Chapters)

Just the highlights!
1. Summation Notation
\[ \sum_{i=1}^{n} f(x_i) = f(x_1) + f(x_2) + \ldots + f(x_n) \]

2. Factorials
\[ n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \]

3. Computations
\[ x = 20, \ y = 14, \ s = 16, \ w = -2, \ m = 15, \ n = 10 \]
Compute \[ x + y \cdot \sqrt[n]{s} = 25.6 \]

4. Simple Linear Equations
\[ 2 - 2x = 3x + 3 \quad x = -1/5 \]
1: Statistics
1.1 Americans Here’s Looking at you

Statistics is all around us!

How much time between Internet usage?

Figure from Johnson & Kuby, 2012.

Rowe, D.B.
1: Statistics
1.1 What is Statistics?

Population: A collection, or set, of individuals, objects, or events whose properties are to be analyzed.

Sample: Subset of the population.

Variable: A characteristic of interest about each individual element of a population or sample.

Data value: The value of the variable associated with one element of a population or sample.

Parameter: A numerical value summarizing all the data of an entire population.

Statistic: A numerical value summarizing the sample data.
1: Statistics
1.1 What is Statistics?

**Data:** The set of values collected from the variable from each of the elements that belong to the sample.

- **Qualitative**
  - Nominal (names)
  - Ordinal (ordered)

- **Quantitative**
  - Discrete (gap)
  - Continuous (continuum)
1: Statistics
1.1 What is Statistics?

**Qualitative variable:** A variable that describes or categorizes an element of a population.

**Nominal variable:** A qualitative variable that characterizes an element of a population. No ordering. No arithmetic.

**Ordinal variable:** A qualitative variable that incorporates an ordered position, or ranking.

**Quantitative variable:** A variable that quantifies an element of a population.

**Discrete variable:** A quantitative variable that can assume a countable number of values. Gap between successive values.

**Continuous variable:** A quantitative variable that can assume an uncountable number of values. Continuum of values.
2: Descriptive Analysis and Single Variable Data
2.1 Graphs - Qualitative Data

Circle (pie) graphs and bar graphs:
Circle is parts to whole as angle.
Bar graph is amount in each category as rectangular areas.

Figures from Johnson & Kuby, 2012.
2: Descriptive Analysis and Single Variable Data

2.2 Frequency Distributions and Histograms

Statistics Exam Scores [TA02-06]

<table>
<thead>
<tr>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>72</td>
</tr>
<tr>
<td>47</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>77</td>
</tr>
<tr>
<td>82</td>
</tr>
<tr>
<td>95</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>72</td>
</tr>
<tr>
<td>95</td>
</tr>
<tr>
<td>74</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>86</td>
</tr>
<tr>
<td>88</td>
</tr>
<tr>
<td>72</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td>78</td>
</tr>
<tr>
<td>94</td>
</tr>
<tr>
<td>86</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>92</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>91</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>76</td>
</tr>
<tr>
<td>78</td>
</tr>
<tr>
<td>67</td>
</tr>
<tr>
<td>66</td>
</tr>
<tr>
<td>68</td>
</tr>
<tr>
<td>98</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>77</td>
</tr>
<tr>
<td>86</td>
</tr>
<tr>
<td>77</td>
</tr>
<tr>
<td>84</td>
</tr>
<tr>
<td>83</td>
</tr>
<tr>
<td>82</td>
</tr>
<tr>
<td>63</td>
</tr>
<tr>
<td>68</td>
</tr>
<tr>
<td>97</td>
</tr>
</tbody>
</table>

Figures from Johnson & Kuby, 2012.
2: Descriptive Analysis and Single Variable Data

2.3 Measures of Central Tendency

Sample Mean: Usual average, p. 63

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

Sample Median: Middle value, p. 64

- \( n \) odd, \( \tilde{x} = \frac{n+1}{2} \) value
- \( n \) even, avg \( \frac{n}{2} \) & \( \frac{n}{2} + 1 \) values

Sample Mode: Most often, p. 66

\( \hat{x} = \) most often

Measures of central tendency characterize center of distribution.

Measures of dispersion characterize the variability in the data.
2: Descriptive Analysis and Single Variable Data

2.4 Measures of Dispersion

Range: \( H - L \), p. 74

Deviation from mean: value minus sample mean, p. 74

\[ i^{th\ deviation\ from\ mean} = x_i - \bar{x} \]

Sample Variance: avg squared dev using \( n-1 \) in den, p. 76

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

Sample Standard Deviation: \( s = \sqrt{s^2} \)
2: Descriptive Analysis and Single Variable Data
2.3, 2.4 Measures of Central Tendency and Dispersion

Example: Data values: 1, 2, 2, 3, 4

\[
\bar{x} = 2.4 \quad \hat{x} = 2 \quad \tilde{x} = 2
\]

\[
s^2 = 1.3 \quad s = 1.1
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{x} = \text{most often value} \quad \tilde{x} = \text{middle value}
\]

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad s = \sqrt{s^2}
\]
2: Descriptive Analysis and Single Variable Data
2.5 Measures of Position

Measures of Position: Quartiles - ranked data into quarters

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

Figure from Johnson & Kuby, 2012.
2: Descriptive Analysis and Single Variable Data
2.5 Measures of Position

Measures of Position: percentiles - rank data into 100ths

\[ L = \text{lowest value} \]
\[ H = \text{highest value} \]
\[ P_k = \text{value where } k\% \text{ are smaller} \]

\[
\frac{nk}{100} \quad \text{integer, } A \\
\text{not Integer, } B
\]

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

\[ p_k \text{ is value in next largest position, } B+1 \text{ value} \]

Figure from Johnson & Kuby, 2012.
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 $x_1, x_2, ..., x_n$.  

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

Questions?

Homework: Chapter 1 # 7, 9, 11, 41, 49a
vocabulary on page 27.
Chapter 2 # 8, 35, 75, 97, 105,
115, 123c-d, 129, 137