Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:
• Presentation of the course
• Introduction to data
Course Home

Instructor’s webpage:

math.ucsd.edu/~eaamari/teaching_2018winter_math11.html

• Lecture slides
• Homework sets
• Course syllabus
• Provisional course calendar
• Link to Piazza
• Office hours times and locations (see syllabus)
Homework

- Homework is due weekly on Tuesday’s lecture.
- Late assignments will not be accepted.
- Your worst homework grade will be dropped.
- Randomly selected problems on the assignment will be graded.

Tacit homework: Read the textbook!

- Homework handed back on Discussion sections.
- No homework re-grading will be allowed after the section ends. This means that if you come back after you went out the room, your grade is fixed and your homework will not be regraded. Complaints/reclamation during the section will be considered with concern.
Computer Labs

In addition, you will be given eight weekly computer labs.

Check out and follow conscientiously the associated website:

http://www.math.ucsd.edu/math11/W18.html

- These labs will be due on **Friday at 9pm** each week without an exam. Labs should be submitted on TritonEd by their respective due dates.
- Lab assignments can be submitted up to 1 hour late at the cost of a 1-point penalty. Assignments submitted later than this will not be accepted, excepting the first assignment which can be submitted up to a week late for a 1-point penalty.
- Questions about labs should be directed to the head lab TA: Selene Xu (yux033@ucsd.edu)
- Lab Office Hours held weekly, on Friday, 11am-1pm AP&M B349
How the Course is Graded

The one following formula giving you the better result will be used:

<table>
<thead>
<tr>
<th>Formula 1</th>
<th>Formula 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>15% Homework</td>
<td>15% Homework</td>
</tr>
<tr>
<td>20% Computer labs</td>
<td>20% Computer labs</td>
</tr>
<tr>
<td>15% Midterm Exam 1</td>
<td>15% Best Midterm Exam</td>
</tr>
<tr>
<td>15% Midterm Exam 2</td>
<td>50% Final Exam</td>
</tr>
<tr>
<td>35% Final Exam</td>
<td></td>
</tr>
</tbody>
</table>

- Your worst homework grade will be dropped for computing your final *Homework* score.
- No makeup exams.
- The grading scheme will be curved and scaled to the best student in class.
<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>January 8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Chapters 3-4-6</td>
<td></td>
<td></td>
<td>Chapters 6-7</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>15 Martin Luther King, Jr. Holiday</td>
<td>16 Chapters 8-9</td>
<td>17</td>
<td>18 Chapters 13-14</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>HW 1 due</td>
<td></td>
<td></td>
<td></td>
<td>Lab 1 due</td>
</tr>
<tr>
<td>Week 3</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Chapter 14</td>
<td></td>
<td></td>
<td>Chapter 15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HW 2 due</td>
<td></td>
<td></td>
<td></td>
<td>Lab 2 due</td>
</tr>
<tr>
<td>Week 4</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>February 1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Midterm 1</td>
<td></td>
<td></td>
<td>Chapter 16</td>
<td></td>
</tr>
<tr>
<td>Week 5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Chapter 15</td>
<td></td>
<td></td>
<td>Chapters 15-5</td>
<td>Lab 4 due</td>
</tr>
<tr>
<td></td>
<td>HW 3 due</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 6</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Chapters 11-16-17</td>
<td>HW 4 due</td>
<td></td>
<td>Chapters 17-18</td>
<td>Lab 5 due</td>
</tr>
<tr>
<td></td>
<td>HW 4 due</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 7</td>
<td>19 Presidents' Day</td>
<td>20 Chapter 19</td>
<td>21</td>
<td>22 Chapters 20-21</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lab 6 due</td>
</tr>
<tr>
<td>Week 8</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>March 1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Midterm 2</td>
<td></td>
<td></td>
<td>Chapters 20-21</td>
<td></td>
</tr>
<tr>
<td>Week 9</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Chapters 22-23</td>
<td></td>
<td></td>
<td>Chapters 22-23</td>
<td>Lab 7 due</td>
</tr>
<tr>
<td></td>
<td>HW 6 due</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 10</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Chapter 25</td>
<td></td>
<td></td>
<td>Chapters 25-24</td>
<td>Lab 8 due</td>
</tr>
<tr>
<td></td>
<td>HW 7 due</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Week</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Final Exam</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3pm-6pm</td>
<td></td>
</tr>
</tbody>
</table>
Components you Need

**Textbook:** Stats, Data & Models, 4th Edition.
by De Veaux, Velleman & Bock

**Software:** Minitab 18
→ Free
Download instructions on TritonEd.
Used for labs

**Calculators:**
- Used on exams and homework
- Need not be graphics, nor have statistical functions
- Cannot be your phone or computer (for exams)
The Piazza forum for our class where questions can be posted and answered.

https://piazza.com/ucsd/winter2018/math11lectureb/home
Before Carrying On...

Any questions so far?
Types of Data

- Categorical/Qualitative: Data that fall into categories or labels; often text ideas; tend NOT to have units. Examples: Gender, marital status, area code.

- Numeric/Quantitative: Numerical data that have units; it usually makes sense to do math operations on these data. Examples: Age, salary, # text messages sent last month.

Be Careful About Data Types:
- Some variables encoded with numbers are not numeric.
  - 0/1 for TRUE/FALSE
  - ZIP codes (92093)
- Not all numeric variables look like numbers.
  - Dates (Friday the 13th, 2017)
  - GPS coordinates (40° 26' 46" N 79° 58' 56" W)
Types of Data

• **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code
Types of Data

- **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code

- **Numeric/Quantitative**
  Numerical data that have units; it usually makes sense to do math operations on these data
  Examples: Age, salary, # text messages sent last month

Be Careful About Data Types:
- 0/1 for TRUE/FALSE
- ZIP codes (92093)
- Not all numeric variables look like numbers.
- Dates (Friday the 13th, 2017)
- GPS coordinates (40°26′46″ N 79°58′56″ W)
Types of Data

- **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code

- **Numeric/Quantitative**
  Numerical data that have units; it usually makes sense to do math operations on these data
  Examples: Age, salary, # text messages sent last month

**Be Careful About Data Types:**
Some variables encoded with numbers are not numeric.
  - 0/1 for TRUE/FALSE
  - ZIP codes (92093)
Not all numeric variables look like numbers.
  - Dates (Friday the 13th, 2017)
  - GPS coordinates (40°26’ 46” N 79°58’ 56” W)
Numerical Data: Histograms

<table>
<thead>
<tr>
<th>ChipotleCost</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.50</td>
</tr>
<tr>
<td>7.29</td>
</tr>
<tr>
<td>8.25</td>
</tr>
<tr>
<td>6.34</td>
</tr>
<tr>
<td>12.31</td>
</tr>
<tr>
<td>11.39</td>
</tr>
<tr>
<td>14.23</td>
</tr>
<tr>
<td>12.65</td>
</tr>
<tr>
<td>12.21</td>
</tr>
<tr>
<td>7.53</td>
</tr>
<tr>
<td>14.21</td>
</tr>
<tr>
<td>6.79</td>
</tr>
<tr>
<td>8.76</td>
</tr>
<tr>
<td>9.32</td>
</tr>
<tr>
<td>10.53</td>
</tr>
</tbody>
</table>

No purchases in the span [13,14)

3 purchases where 6 ≤ cost < 7
The choice of bin size influences crudely the histogram plot.

By default, Minitab tries its best to display an informative histogram.
We say an histogram has a **mode** when it is peaked somewhere.
Numerical Data Vocabulary: Symmetry

An histogram is **symmetric** if both sides of mode look the same.

(a) Symmetric  
(b) Symmetric  
(c) Not symmetric
The tails of an histogram are the parts away from the center.
Numerical Data Vocabulary: Skewness

When an histogram is not symmetric, we can describe further its asymmetry by saying it is

- **Skewed left:** if the left tail is longer than the right tail.
- **Skewed right:** if the right tail is longer than the left tail.

![Histogram examples](image)

(a) Skewed left  
(b) Skewed right

Skewed left/right = the left/right tail stretches out longer.
An outlier is an observation that appears extreme relative to the rest of the data. (= Not conventional)

Examples:
- Extreme values in precision measurements for astrophysics
- Trolls’ answers in online questionnaires
Sometimes outliers are informative, sometimes just annoying.
Describe this histogram.

Heights of NBA players from the 2008–9 season

Unimodal, skewed left, no outlier.
Describe this histogram.

Heights of NBA players from the 2008–9 season

Unimodal, skewed left, no outlier.
Practice

Describe these histograms.

*Population of France - Provisional estimate at 1 January 2017*

1. Birth deficit due to 1914-1918 war (depleted cohorts)
2. Depleted cohorts at reproductive age
3. Birth deficit due to 1939-1945 war
4. Baby boom
5. End of baby boom

(G. Pison, Population & Societies, n° 542, INED, March 2017)
Practice

Describe these histograms.

Multimodal, skewed right (up), no outlier.
The Center of a Data Set/Histogram/Distribution

Idea 1: The center of a distribution should be the data value “in the middle of the list of data” –there should be the same number of data values on each side of “the center”.  

MEDIAN
The Center of a Data Set/Histogram/Distribution

**Idea 1:** The center of a distribution should be the data value “in the middle of the list of data” – there should be the same number of data values on each side of “the center”.

```
4, 6, 0, -2, 45  →  -2, 0, 4, 6, 45
4, 6, 0, 3, -2, 45  →  -2, 0, 3, 4, 6, 45
```

**MEDIAN**

- Put the data in order, choose the middle number.
- If there is no “middle number”, average the two in the middle of the ordered list.

The median is 3.5
Idea 1: The center of a distribution should be the data value “in the middle of the list of data” – there should be the same number of data values on each side of “the center”.

**Median**

\[
\begin{align*}
4, 6, 0, -2, 45 & \rightarrow -2, 0, 4, 6, 45 \\
4, 6, 0, 3, -2, 45 & \rightarrow -2, 0, 3, 4, 6, 45
\end{align*}
\]

Put the data in order, choose the middle number.

If there is no “middle number”, average the two in the middle of the ordered list.

The median is 3.5

Idea 2: The center of a distribution must take into account the data values themselves, not just the order they are in.

**Mean**
The Center of a Data Set/Histogram/Distribution

**Idea 1**: The center of a distribution should be the data value “in the middle of the list of data” – there should be the same number of data values on each side of “the center”.

\[
\begin{align*}
4, 6, 0, -2, 45 & \quad \rightarrow \quad -2, 0, 4, 6, 45 \\
4, 6, 0, 3, -2, 45 & \quad \rightarrow \quad -2, 0, 3, 4, 6, 45
\end{align*}
\]

The median is 3.5

**MEDIAN**

Put the data in order, choose the middle number.

If there is no “middle number”, average the two in the middle of the ordered list.

**Idea 2**: The center of a distribution must take into account the data values themselves, not just the order they are in.

**MEAN**

4, 5, 6 and 4, 5, 990 have the same median

\[
\begin{align*}
\frac{4 + 5 + 6}{3} &= 5 \\
\frac{4 + 5 + 990}{3} &= 333
\end{align*}
\]

Add the data values, divide by how many there are.
Which Center Idea Do I Use?

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>{2, 3, 4, 5, 6, 7}</code></td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td><code>{2, 3, 4, 5, 6, 70}</code></td>
<td>4.5</td>
<td>15</td>
</tr>
<tr>
<td><code>{2, 3, 4, 5, 6, 700}</code></td>
<td>4.5</td>
<td>120</td>
</tr>
</tbody>
</table>

Moral: The median is resistant to outliers and skew. For this reason, it is called "robust."
Which Center Idea Do I Use?

\{2, 3, 4, 5, 6, 7\} \quad \{2, 3, 4, 5, 6, 70\} \quad \{2, 3, 4, 5, 6, 700\}

Median: 4.5  
Mean: 4.5  
Median: 4.5  
Mean: 15  
Median: 4.5  
Mean: 120

**Moral:** The median is resistant to outliers and skew. For this reason, it is called “robust”. 
Which Center Idea Do I Use?

\{2, 3, 4, 5, 6, 7\} \hspace{1cm} \{2, 3, 4, 5, 6, 70\} \hspace{1cm} \{2, 3, 4, 5, 6, 700\}

- **Median**: 4.5
- **Mean**: 4.5

- **Median**: 4.5
- **Mean**: 15

- **Median**: 4.5
- **Mean**: 120

**Moral**: The median is resistant to outliers and skew. For this reason, it is called “robust”.

**Median**: Use for asymmetric distributions or data with outliers.
**Mean**: Use for symmetric distributions without outliers.
The Spread of a Distribution

Most values in here

Most values in here
Three Ideas for Measuring Spread

The Range:
(maximum value in data set) – (minimum value in data set)

- **Pros**: Easy to calculate, gives a sense of the total span of the data
- **Cons**: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):
(upper quartile) – (lower quartile)

- **Pros**: Resistant to skew and outliers, easy to communicate
- **Cons**: Takes work to calculate, no universal definition, not widely known
Three Ideas for Measuring Spread

The Range:
(maximum value in data set) – (minimum value in data set)

• **Pros:** Easy to calculate, gives a sense of the total span of the data
• **Cons:** Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):
(upper quartile) – (lower quartile)
Three Ideas for Measuring Spread

**The Range:**

\[(\text{maximum value in data set}) - (\text{minimum value in data set})\]

- **Pros:** Easy to calculate, gives a sense of the total span of the data
- **Cons:** Summarizes the data using only two of the data points (the extremes); not resistant to outliers

**The Interquartile Range (IQR):**

\[(\text{upper quartile}) - (\text{lower quartile})\]

- **Pros:** Resistant to skew and outliers, easy to communicate
- **Cons:** Takes work to calculate, no universal definition, not widely known
Three Ideas for Measuring Spread

The Range:
(maximum value in data set) – (minimum value in data set)

- **Pros**: Easy to calculate, gives a sense of the total span of the data
- **Cons**: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):
(upper quartile) – (lower quartile)

- **Pros**: Resistant to skew and outliers, easy to communicate
- **Cons**: Takes work to calculate, no universal definition, not widely known

Range: \(19 - 2 = 17\)
IQR: \(13 - 3 = 10\)

\(\{2, 2, 3, 3, 5, 5, 5, 7, 11, 13, 13, 13, 13, 17, 19\}\)

Upper quartile (Q3): Find the median of the data to the right of Q2 (here: 13)
Lower quartile (Q1): Find the median of the data to the left of Q2 (here: 3)
Middle quartile (Q2, Median): 7
Bringing It All Together: The Five Number Summary

The five number summary for the magnitudes of 221 earthquakes

<table>
<thead>
<tr>
<th>Max</th>
<th>9.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3</td>
<td>7.6</td>
</tr>
<tr>
<td>Median</td>
<td>7.2</td>
</tr>
<tr>
<td>Q1</td>
<td>6.7</td>
</tr>
<tr>
<td>Min</td>
<td>4.0</td>
</tr>
</tbody>
</table>

The five number summary for the magnitudes of 221 earthquakes
Visualizing the Five-Number Summary: The Boxplot

- Upper fence (1.5 IQRs above Q3)
- Upper whisker (largest value inside fences)
- Q3
- Lower whisker (smallest value inside fences)
- Lower fence (1.5 IQRs below Q1)

Outliers: Data outside the fences; *extreme outliers* are often labelled with a star and are more than 3 IQRs away from the lower/upper quartiles!
Three Ideas for Measuring Spread

The sample **standard deviation** of \( x = (x_1, \ldots, x_n) \) is defined as:

\[
s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]
Three Ideas for Measuring Spread

The sample **standard deviation** of \( x = (x_1, \ldots, x_n) \) is defined as:

\[
s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

- Sum of squares: to get positive values
- Square root: to undo squaring action and have \( s \) with same units as the \( x_i \)'s.
- \((n - 1)\) instead of \( n \): explained later in the course.
Three Ideas for Measuring Spread

The sample **standard deviation** of \( x = (x_1, \ldots, x_n) \) is defined as:

\[
s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

- Sum of squares: to get positive values
- Square root: to undo squaring action and have \( s \) with same units as the \( x_i \)'s.
- \((n - 1)\) instead of \( n \): explained later in the course.

- **Pros**: Very popular notion of spread; easy to handle mathematically
- **Cons**: Sensitive to outliers and skew; hard to explain to non-statisticians.
Explaining Your Data: SOCS

**Shape**: Modality (uniform, unimodal, bimodal, multimodal)
Symmetry (symmetric, skewed left, skewed right)

**Outliers**: Where are they? Do they have meaning?

**Center**: Mean for unimodal, symmetric data with no outliers, median else

**Spread**: Standard deviation (SD) for unimodal, symmetric data with no outliers, IQR else
Practice

Describe **Shape**, **Outliers**, **Center**, and **Spread** of this dataset.

![Histogram and boxplot](image)

- **Max**: 9.1
- **Q3**: 7.6
- **Median**: 7.2
- **Q1**: 6.7
- **Min**: 4.0
Describe Shape, Outliers, Center, and Spread of this dataset.

Answer: Unimodal and slightly skewed left, no outliers, width median $Q_2 = 7.2$ and spread $IQR = 7.6 - 6.7 = 0.9$. 
Comparing Distributions

When Comparing Histograms:

• Make sure the scales are the same (both axes!)
• Discuss SOCS (Shape, Outliers, Center, Spread)
• Use the same measures of center and spread for both
• Also create five-number summaries
• Write your conclusions so they do not generalize beyond the sample
Comparing Distributions

When Comparing Histograms:

- Make sure the scales are the same (both axes!)
- Discuss SOCS (Shape, Outliers, Center, Spread)
- Use the same measures of center and spread for both
- Also create five-number summaries
- Write your conclusions so they do not generalize beyond the sample
Comparing Boxplots: Tons of Info in One Place

Massachusetts Wind Speed Boxplots for Each Month in 2011

What to think about:

• Which groups have the highest medians, the greatest IQRs
• Where the middle 50% of data is for each group
• What things look like when outliers are minimized from our view
• The context-dependent nature of outliers
Facebook Friends VS Gender

Descriptive Statistics: Friends

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gender</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>Range</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>Female</td>
<td>447.3</td>
<td>326.8</td>
<td>0.0</td>
<td>200.0</td>
<td>400.0</td>
<td>640.0</td>
<td>1410.0</td>
<td>1410.0</td>
<td>440.0</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>385.9</td>
<td>282.9</td>
<td>0.0</td>
<td>200.0</td>
<td>310.0</td>
<td>579.0</td>
<td>1200.0</td>
<td>1200.0</td>
<td>379.0</td>
</tr>
</tbody>
</table>

Histogram of Friends

Boxplot of Friends
Taking Outliers Seriously

• Try researching a particular outlier. Does it really belong?
• Some outliers are errors.
• Run analysis with and without your outliers.
• In the end, state what you are doing with outliers and why.
Taking Outliers Seriously

- Try researching a particular outlier. Does it really belong?
- Some outliers are errors.
- Run analysis with and without your outliers.
- In the end, state what you are doing with outliers and why.

**Practice:** Give an example of a *quantitative variable* and an *outlier* in that data.
Taking Outliers Seriously

- Try researching a particular outlier. Does it really belong?
- Some outliers are errors.
- Run analysis with and without your outliers.
- In the end, state what you are doing with outliers and why.

**Practice:** Give an example of a **quantitative variable** and an **outlier in that data**.
An answer: **Number of different types of cheese in a country, France**
192 students were asked:

*How many songs are in your digital music library?*

---

**Histogram of Songs in Library**

![Histogram of Songs in Library](image_url)
Re-expressing Highly Skewed Data

192 students were asked:

How many songs are in your digital music library?

Examples where data are thought of on a re-expressed scale:

- Earthquake magnitudes
  \[ M = \log_e 10\left(\frac{A}{A_0}\right) \]
- pH of substance
  \[ pH = -\log(a_{H^+}) \]
- Grades on a bad test
  \[ G = 10\sqrt{\text{Original grade}} \]
Re-expressing Highly Skewed Data

We display the new (made-up) variable

\[ \log_{10}(\text{Song Count}). \]

People with \( 1.8 \leq \log_{10}(\text{Song Count}) < 2.1 \)

People with \( 10^{1.8} \leq \text{Song Count} < 10^{2.1} \)
Why Re-express Skewed Data?

• To make it visually more appealing
• To create a more commonly-shaped histogram (the value of this is apparent later)
• To get the “lens of analysis correct”
Why Re-express Skewed Data?

- To make it visually more appealing
- To create a more commonly-shaped histogram (the value of this is apparent later)
- To get the “lens of analysis correct”

“How many songs?” (1, 2, 3 ...)

“What order of magnitude?” (1, 10, 100, ...)

Discuss with a classmate: Will the mean be less than, about equal to, or greater than the median in each distribution?
Discuss with a classmate: Will the mean be less than, about equal to, or greater than the median in each distribution?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Skew</td>
<td>4.5161</td>
<td>4.4005</td>
</tr>
<tr>
<td>Medium Skew</td>
<td>5.0217</td>
<td>4.6575</td>
</tr>
<tr>
<td>High Skew</td>
<td>6.089</td>
<td>5.334</td>
</tr>
</tbody>
</table>

As the graph skews to the right, the mean becomes larger than the median: the mean is pulled right by the larger values in the data set.
What you Should Do After the Lecture

• Buy textbook
• Start Homework 1 (Due next Tuesday in class)
  Print it, and make sure to write down your full name, the PID
  and your section.
• Install Minitab
• Read the course syllabus and look at the calendar