Facts to know about STAT 200, 220, and 234:

- All three are intro classes — no prior stats expected or required.
- Only STAT 220 and 234 qualify as prerequisites for more advanced STAT courses.
- You do not need to take STAT 200 to prepare for either STAT 220 or 234.
- You do not need to take STAT 220 to prepare for STAT 234.
- You cannot take STAT 200 "after" STAT 220 or 234.
- No student can have credit for both STAT 220 and 234.
- See Linda Collins in E107 if you have questions.

Today's topics (textbook sections 1.1–1.2):

- Categorical variables
  - Bar graphs & pie charts
- Quantitative variables
  - Histograms — shape, mode, skew
  - Time plots
  - Mean and median

NYC public school data

Class size:

Math test scores:
Cases & variables

Class roster:

<table>
<thead>
<tr>
<th>CNET ID</th>
<th>Year</th>
<th>State</th>
<th>Age</th>
<th>Section #</th>
</tr>
</thead>
<tbody>
<tr>
<td>a123</td>
<td>3</td>
<td>VT</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>xyz4</td>
<td>1</td>
<td>ME</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>abc7</td>
<td>2</td>
<td>FL</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In a study,
- We collect data from cases (also called: individuals, units, observations, data points, ...).
  A case might be: a person, a company, a university, a tissue sample, a lake, ....
- The variables are the characteristics of each case that we record information about:
  age, # of students, type of institution, disease status, antibody count, ...

Two types of variables

- Quantitative variables:
  Record a numerical value for each case
- Categorical variables:
  Place each individual or case into one of several categories

Quantitative variables

Record a numerical value for each case.
- How many students in Grade 3 were tested at each school?
- How many Grade 3 students’ test outcome was Level 1?
- What is the unemployment rate each year?
- [not all numbers are quantitative!]
  What section of Stat220 is each student enrolled in?
  What is the first digit of your phone number?

Categorical variable

Place each individual or case into one of several categories
- For each student at school #M015, did he/she take the test?
- Which section of Stat 220 are you enrolled in?
- Which division does each department at U of C belong to?
- Sometimes categories might be coded as numbers:

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Received flu vaccine?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1=yes,2=no,3=delayed)</td>
</tr>
<tr>
<td>18506</td>
<td>1</td>
</tr>
<tr>
<td>84721</td>
<td>2</td>
</tr>
<tr>
<td>93821</td>
<td>1</td>
</tr>
<tr>
<td>17374</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Two types of variables

- Quantitative variables:
  Record a quantitatively meaningful numerical value for each case.
  “How many” / “how much”? Can you add, subtract, or average these numbers?

- Categorical variables:
  Place each individual or case into one of several categories.
  Note that the category label might be a number!

Graphs for categorical variables

Data gathered in 2010 from students in grade 3 at school 01M015:

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Math test level</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC123</td>
<td>Level 3</td>
</tr>
<tr>
<td>XYZ456</td>
<td>Level 1</td>
</tr>
<tr>
<td>TUV789</td>
<td>Level 4</td>
</tr>
<tr>
<td>AAA111</td>
<td>Level 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test result</th>
<th># of students</th>
<th>% of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>6</td>
<td>23.1%</td>
</tr>
<tr>
<td>Level 2</td>
<td>12</td>
<td>46.2%</td>
</tr>
<tr>
<td>Level 3</td>
<td>6</td>
<td>23.1%</td>
</tr>
<tr>
<td>Level 4</td>
<td>2</td>
<td>7.7%</td>
</tr>
</tbody>
</table>

Pie charts

- The size of each “slice” is proportional to the percent of cases that fall into that category.

Bar graphs

- Can use a bar graph to display the proportion of cases that are placed into each of the different categories.
- The height of each bar is proportional to the percent of cases that fall into that category.
Bar graphs

- The bars can also be horizontal
- Can reorder the categories to show largest to smallest

2018 grade 3 test results in school M015

Bar graphs are a good tool for comparing data

Grade 3 students tested in Manhattan district 01 in 2010

Bar graphs

Grade 3 students tested in Manhattan district 01 in 2010
Bar graphs

Grade 3 students tested in Manhattan district 01 in 2010

Bar graphs

I Can also use a bar graph to display any quantitative attribute of the categories
I Not every bar graph can be turned into a pie chart!
(Does it make sense to add the numbers & get a total?)

Bar graphs

Grade 3 students in 2010 with Level 4 results
Dialect survey data

What is your "general" term for the rubber-sole shoes worn in gym class, for athletic activities, etc.

- tennis shoes
- sneakers
- cheer
- gymshoes

Based on data from the National Dialect Survey by Beth New & Scott Golder

Dialect survey data

In each state, how many people participated in the dialect survey?

<table>
<thead>
<tr>
<th>State</th>
<th># of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>247</td>
</tr>
<tr>
<td>AR</td>
<td>145</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Histograms

In each state, how many people participated in the dialect survey?
Histograms

In each state, how many people participated in the dialect survey?

<table>
<thead>
<tr>
<th>State</th>
<th># of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>247</td>
</tr>
<tr>
<td>AR</td>
<td>145</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Not enough bins:

![Histogram example with too few bins](image1)

Too many bins:

![Histogram example with too many bins](image2)

Histograms

Choose the number of bins by trial & error!

- Goal: to summarize the general trend of the data
- Not enough bins → may lose information
- Too many bins → too much detail, can’t see general trend
- In general, use more bins for larger data sets

Histograms

Is this a bar plot or a histogram?

Orange County - Household Income

[www.doctorhousingbubble.com](http://www.doctorhousingbubble.com)
Histograms

In a histogram:

- The measured variable is quantitative.
- Values for this variable are shown on the horizontal axis.
- Each bin covers a range of values for the variable.
- No gaps between the bins.
- Generally, all the bins have the same width.
- (Next class: density plots & bins with different widths)
- The size of the bar is proportional to the # of observations that fall into that bin.

Size: height or area?

- The y-axis should start at zero for barplots & histograms, otherwise the graph can be misleading!

http://www.appstate.edu/~goodmanjm/rcoe/hwr/math/graphs/graphs.html

Distribution shape

A mode of your data is a variable value where there is a strong peak.

“sneakers” histogram: bimodal, not symmetric.

“rilly” histogram: unimodal, symmetric.

Three unimodal histograms:

Left-skewed

Symmetric

Right-skewed
Distribution shape

This data has a right-skewed distribution:

![Histogram showing right-skewed distribution]

Time plots

What are two problems with this time plot?

**Figure 1: Projected Afghan National Army Growth**

<table>
<thead>
<tr>
<th>Date</th>
<th># cancelled flights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 1</td>
<td>438</td>
</tr>
<tr>
<td>Jan 2</td>
<td>942</td>
</tr>
<tr>
<td>Jan 3</td>
<td>2122</td>
</tr>
<tr>
<td>Jan 4</td>
<td>652</td>
</tr>
<tr>
<td>Jan 5</td>
<td>1999</td>
</tr>
<tr>
<td>Jan 6</td>
<td>3811</td>
</tr>
<tr>
<td>Jan 7</td>
<td>899</td>
</tr>
</tbody>
</table>

![Time plot showing # flights cancelled vs. Date]

![Time plot showing # of states vs. # of participants]

![Time plot showing Army size vs. Date]

Source: Afghan Ministry of Defence; Combined Security Transition Command-Afghanistan
Time plots

- Time is on the horizontal axis.
- Measured variable is on the vertical axis.
- Space between points on the horizontal axis should be proportional to the amount of time between points.
- Usually, y-axis should start at zero so that the graph is not misleading.

Mean & median

The mean is the average value.
The median is the middle value.

# of survey participants in the Northeast:

<table>
<thead>
<tr>
<th>State</th>
<th>CT</th>
<th>MA</th>
<th>ME</th>
<th>NH</th>
<th>NJ</th>
<th>NY</th>
<th>PA</th>
<th>RI</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>651</td>
<td>1276</td>
<td>140</td>
<td>191</td>
<td>1221</td>
<td>2749</td>
<td>1747</td>
<td>142</td>
<td>114</td>
</tr>
</tbody>
</table>

- Mean = \[
\frac{651 + 1276 + 140 + 191 + 1221 + 2749 + 1747 + 142 + 114}{9} = 914.5556
\]

- Median: 114 140 142 191 651 1221 1276 1747 2749

# of survey participants in the Rocky Mountains:

<table>
<thead>
<tr>
<th>State</th>
<th>CO</th>
<th>ID</th>
<th>MT</th>
<th>NV</th>
<th>UT</th>
<th>WY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>322</td>
<td>158</td>
<td>294</td>
<td>93</td>
<td>261</td>
<td>70</td>
</tr>
</tbody>
</table>

- Mean = \[
\frac{322 + 158 + 294 + 93 + 261 + 70}{6} = 199.6667
\]

- Median: 70 93 158 261 294 322

\[
\text{Median} = \frac{158 + 261}{2} = 209.5
\]

Why is the mean higher than the median in this data set?