Lecture 6

Histograms
Bar Charts (Review)
Types of Data

All values in a column should be both the same type and be comparable to each other in some way

● **Numerical** — Each value is from a numerical scale
  ○ Numerical measurements are ordered
  ○ Differences are meaningful

● **Categorical** — Each value is from a fixed inventory
  ○ May or may not have an ordering
  ○ Categories are the same or different
Bar Charts of Counts

*Distributions:*

- The distribution of a variable (a column) describes the frequency of its different values
- The **group** method counts the number of rows for each value in a column

Bar charts can display the distribution of categorical values

- Proportion of how many US residents are male or female
- Count of how many top movies were released by each studio

*(Demo)*
Binning Numerical Values

Binning is counting the number of numerical values that lie within ranges, called bins.

- Bins are defined by their lower bounds (inclusive)
- The upper bound is the lower bound of the next bin

188, 170, 189, 163, 183, 171, 185, 168, 173, ...

The [185,190) bin
Histogram

Chart to display the distribution of numerical values using bins

(Demo)
Clicker question

What row are you sitting in?

- A) 1-2
- B) 3-4
- C) 5-6
- D) 7-8
- E) 9+
Clicker question

What row are you sitting in?

- A) 1
- B) 2-3
- C) 4-5
- D) 6-8
- E) 9+
The Density Scale
Histogram Axes

By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%

- The horizontal axis is a number line (e.g., years)
- The vertical axis is a rate (e.g., percent per year)
- The area of a bar is a percentage of the whole

(Demo)
How to Calculate Height

The [20, 40) bin contains 59 out of 200 movies

- “59 out of 200” is 29.5%
- The bin is 40 - 20 = 20 years wide

29.5 percent

Height of bar = \[ \frac{29.5 \text{ percent}}{20 \text{ years}} \]

= 1.475 percent per year
Height Measures Density

\[
\text{Height} = \frac{\% \text{ in bin}}{\text{width of bin}}
\]

- The height measures the percent of data in the bin \textit{relative to the amount of space in the bin}.
- So height measures crowdedness, or \textit{density}.

(Demo)
Area Measures Percent

\[
\text{Area} = \% \text{ in bin} = \text{Height} \times \text{width of bin}
\]

- “How many individuals in the bin?” Use area.
- “How crowded is the bin?” Use height.
Discussion Question

What's the height of each bar in these two histograms?

```python
actress.hist(1, bins=[0,15,25,85])
actress.hist(1, bins=[0,15,35,85])
```

What are the vertical axis units?

<table>
<thead>
<tr>
<th>Name</th>
<th>2016 Income (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jennifer Lawrence</td>
<td>61.7</td>
</tr>
<tr>
<td>Scarlett Johansson</td>
<td>57.5</td>
</tr>
<tr>
<td>Angelina Jolie</td>
<td>40</td>
</tr>
<tr>
<td>Jennifer Aniston</td>
<td>24.75</td>
</tr>
<tr>
<td>Anne Hathaway</td>
<td>24</td>
</tr>
<tr>
<td>Melissa McCarthy</td>
<td>24</td>
</tr>
<tr>
<td>Bingbing Fan</td>
<td>20</td>
</tr>
<tr>
<td>Sandra Bullock</td>
<td>20</td>
</tr>
<tr>
<td>Cara Delevingne</td>
<td>15</td>
</tr>
<tr>
<td>Reese Witherspoon</td>
<td>15</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>15</td>
</tr>
<tr>
<td>Kristen Stewart</td>
<td>12</td>
</tr>
<tr>
<td>Amanda Seyfried</td>
<td>10.5</td>
</tr>
<tr>
<td>Tina Fey</td>
<td>10.5</td>
</tr>
<tr>
<td>Julia Roberts</td>
<td>10</td>
</tr>
<tr>
<td>Emma Stone</td>
<td>10</td>
</tr>
<tr>
<td>Natalie Portman</td>
<td>8.5</td>
</tr>
<tr>
<td>Margot Robbie</td>
<td>8</td>
</tr>
<tr>
<td>Meryl Streep</td>
<td>6</td>
</tr>
<tr>
<td>Meryl Streep</td>
<td></td>
</tr>
<tr>
<td>Mila Kunis</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Clicker question

What are the vertical axis units?

- Counts
- %
- % per millions 
- % per $
Chart Types
Bar Chart Versus Histogram

Bar Chart
- 1 categorical axis & 1 numerical axis
- Bars have arbitrary (but equal) widths and spacings
- For distributions: height (or length) of bars are proportional to the percent of individuals

Histogram
- Horizontal axis is numerical, hence to scale with no gaps
- Height measures density; areas are proportional to the percent of individuals
Overlaid Graphs

For visually comparing two populations

(Demo)