Lecture 8

Histograms
Announcements

- Homework 3 out today, due next Friday 3/5.
Let's check if you are registered for PollEverywhere!

I am registered!

I'm not registered.
Area Principle

Areas should be proportional to the values they represent

In 2013,

30% of accidental deaths of males were due to automobile accidents

20% of accidental deaths of females were due to automobile accidents

Example from Tian Zheng
From Tufte, p. 70: “If the area of the dollar is accurately to reflect its purchasing power, then the 1978 dollar should be about twice as big as that shown.”
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
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
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)
For this histogram

The heights are proportional to the counts in each bin.

The areas are proportional to the counts in each bin.

Both heights and areas are proportional to the counts in each bin.
For this histogram

Heights are proportional to the counts in each bin

Areas are proportional to the counts in each bin

Both heights and areas are proportional to the counts in each bin
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)
For this histogram

Heights are proportional to counts in each bin

Areas are proportional to counts in each bin

Both heights and areas are proportional to counts in each bin
How to Calculate Height

The [0, 20) bin contains 80 out of 200 movies

- “80 out of 200” is 40%
- The bin is $20 - 0 = 20$ years wide

\[
\text{Height of bar} = \frac{40 \text{ percent}}{20 \text{ years}}
\]

\[
= 2 \text{ percent per year}
\]
Height Measures Density

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

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

(Demo)
Area Measures Percent

**Area** = % in bin = Height $\times$ 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?

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>Mila Kunis</td>
<td>4.5</td>
</tr>
</tbody>
</table>
What would be the units of the vertical axis?

- Counts
- %
- % per million
- % 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)