STATS Lecture #4: Measures of Center + Spread

- 3 possible vertical scale histograms
  1. Raw-frequency plot the counts
  2. Relative-frequency plot the %
  3. Density scale

<table>
<thead>
<tr>
<th>Value</th>
<th>Raw Freq</th>
<th>( \text{Frequency} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>1</td>
<td>( \frac{1}{24} = 0.04 )</td>
</tr>
<tr>
<td>3.4</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>3.5</td>
<td>1</td>
<td>0.41</td>
</tr>
<tr>
<td>3.6</td>
<td>2</td>
<td>0.83</td>
</tr>
<tr>
<td>4.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( n = 24 )</td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

- When histograms are plotted on density scale:
  a) rel freq \( \rightarrow \) area of hist. bars (curve)
  b) total area under hist. = 100%.

- Convention: all histograms from now on, are implicitly on the density scale.

- Positively skewed
- Not symmetric
- Point of symmetry negatively skewed
long right hand tail

U.S. family income in 2017

Bill Gates

left

tail

Midterm scores (\%)

100%

0%

50%

100%

unimodal

mode

left
tail

center
	right
tail

bimodal

multimodal

different center, same shape, same spread

same shape, same center, different spread

same center, same spread, different shape

qualitatively
Measures of center:

Quant. Cond. Ratio

\[ \text{mean} = \frac{4.4 + \ldots + 3.8}{24} = \frac{24}{4.0 \text{ cm}} \]

\[ n = 24 \]

\[ y_1 = \begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ \vdots \\ 3.8 \end{bmatrix} \]

\[ y_n = 3.8 \]

\[ \text{mean} = 4 \]

\[ \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \]

Long right-hand tail

\[ \text{mean} = 4 \]

\[ \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \rightarrow \text{subtract} \begin{bmatrix} -3 \\ -2 \\ +6 \end{bmatrix} \]

\[ \begin{bmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ y_n - \overline{y} \end{bmatrix} \]

Graphical interpretation of the mean: Center

of Gravity = Balance Point

Symmetric

Unimodal

\[ \text{point of symmetry} = \text{mean} = \text{mode} \]

\[ \begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \]

\[ \begin{bmatrix} 3.3 \\ 3.5 \\ 4.0 \\ 4.0 \\ 4.0 \end{bmatrix} \]

\[ \text{middle} = \frac{4.0 + 4.0}{2} = 4.0 \]

\[ \text{median} \]
\[
\begin{align*}
\text{median} & = [\frac{1}{3}, \frac{2}{3}] \\
\text{mean} & = \frac{\sum y_i}{n} \\
\text{mode} & = \text{peak point of symmetry}
\end{align*}
\]
Influence of outliers on the mean: Mean is pulled by the tail.

Measures of spread:

- Typical amount by which each # differs from center.

\[
\begin{align*}
\text{sample:} & \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \\
\text{mean:} & \quad 4 \\
\text{mean:} & \quad 0 \\
\text{mean:} & \quad 3.3 \\
\frac{1}{n} \sum_{i=1}^{n} |y_i - \bar{y}| = (MAD) \text{ mean absolute deviation}
\end{align*}
\]

\[
\begin{align*}
\text{mean:} & \quad 4 \\
\text{mean:} & \quad 0 \\
\text{mean:} & \quad 12.7 \\
\text{mean:} & \quad \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} = (sample) \text{ standard deviation (sd)}
\end{align*}
\]

\[
\begin{align*}
\text{mean:} & \quad 0 \\
\text{mean:} & \quad 12.7 \\
\text{mean:} & \quad \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2} = (sample) \text{ variance}
\end{align*}
\]
The data set has $n = 3$ observations in it, but only $(n-1) = 2$ degrees of freedom for measuring spread.

Empirical rule: start at mean, go $1$ SD either way, you will capture about $\frac{2}{3}$ of the data.

- $2$ SD $= 95\%$.
- $3$ SD $= almost all $ $99.7\%.$ (from normal curve)

ex) $0.5$ too big
$0.1$ too small
$0.3$ about right