This measure of time: center & spread

next time: normal curve

new course deadline for HW 1: 11/5
next Wed (8 days from now)

3 possible vertical scales for histograms

1. raw frequency: plot the counts
2. relative frequency: plot the %
3. density scale: when hist are plotted on density scale:
   (a) rel. freq $\frac{\text{area of histogram}}{\text{bars}}$
   (b) total area under hist $= 100\%$
butterfly data

Convention:
all hist. from now on are implicitly on density scale.

<table>
<thead>
<tr>
<th>value</th>
<th>raw freq.</th>
<th>% (relative freq.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>1</td>
<td>1/24 = 4%</td>
</tr>
<tr>
<td>3.4</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>3.5</td>
<td>1</td>
<td>4%</td>
</tr>
<tr>
<td>3.6</td>
<td>2</td>
<td>8% (2/24)</td>
</tr>
<tr>
<td>4.5</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

n = 24 100%

relative freq. raw freq.

4

3

2

1

0

3.3 3.5 3.6

positively skewed

(not symmetric, skewed)

symmetric

with

mean

length

unimodal

4.0 cm

point of symmetry

negatively skewed

3

2

1

0

density scale
Lower income in 2017

$0 $120k

Low-income tail

Unimodal

Outlier

Multimodal
measures of center

1. mean (average)
2. 2
3. mode

4. height (cm)
   4.4
   3.6
   1.0
   3.9
   h = 24
   Y = $\frac{4.4 + \ldots + 2.8}{24}$
   = 4.0 cm

Qualitatively

same shape, different center,
same shape, different spread
same center, same spread,
same center, different shape
different center,
same shape,
same spread,
\[ y_n = \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}, \quad n = 3 \]

Mean \( \bar{y} = 4 \)

\[ \left[ \begin{array}{c} 1 \\ 2 \\ 9 \\ y_1 \\ \vdots \\ y_n \end{array} \right], \quad \text{mean} \bar{y} = 4 \]

\[ \left[ \begin{array}{c} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_n - \bar{y} \end{array} \right], \quad \text{mean} = 0 \]

Deviations from the mean:

\[ \text{Graphical interpretation of mean:} \]

Center of gravity = balance point

Symmetrical distribution

Point of symmetry = mean = mode
A graphical interpretation of median

50% 50% point in data in relative frequency terms
Mean = mode = median

Mode

Median

Mean

Sale price of 50% houses

Upper cases

Lower cases

Index of summation

\[ \bar{y} = \frac{y_1 + \ldots + y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i \]
25\% 25\% 25\%
\[ \text{median} \]

\[ = 0.5 \text{ Quantile} \]

75\% Percentile

\[ = 0.75 \text{ Quantile} \]

25th Percentile

\[ = 0.25 \text{ Quantile} \]

Influence of outliers on mean:

Mean is pulled by the tail

Measures of Spread

Typical amount by which each # differs from center
Sample: \[
\begin{bmatrix}
1 \\
2 \\
9 \\
\end{bmatrix}
\]
mean $4$

Subtract \[
\frac{\begin{bmatrix}
5 \\
2 \\
8 \\
\end{bmatrix}}{4}
\]
mean $0$

Subtract \[
\begin{bmatrix}
\frac{y_1 - 7}{y} \\
\frac{y_2 - 7}{y} \\
\frac{y_n - 7}{y}
\end{bmatrix}
\]
abs. value $\begin{bmatrix}
1 - \frac{y_1}{7} \\
1 - \frac{y_2}{7} \\
1 - \frac{y_n}{7}
\end{bmatrix}$
mean \(\frac{13}{3} \approx 3.3\)

\[
\frac{1}{n} \sum_{i=1}^{n} |y_i - \bar{y}| = \text{(MAE)} \text{ mean absolute deviation}
\]
\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \text{(sample variance)} \]

\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \text{(sample standard deviation)} \]

\[ s_d = \sqrt{\frac{6^2}{38}} = 4.4 \]

The data set has \( n = 3 \) observations, but only \((n-1) = 2\) degrees of freedom for measuring spread.
Graphical interpretation of SD

Empirical rule:
Start at mean:
\( \begin{align*} \mu = 4.0 \text{ cm} \end{align*} \)

Go (\( \pm \frac{1}{2} \)) SD (5) either way: you will capture (about \( \frac{\Sigma}{3} \)) \( 68\% \) of the data.

0.5 too big, 0.3 about right, 0.1 too small.

Total length:
\( 3.3 \) \( 4.0 \) \( 4.5 \)

\( n = 24 \)

Sample mean \( \overline{y} = 4.0 \text{ cm} \)

Sample standard deviation \( s = 0.29 \)