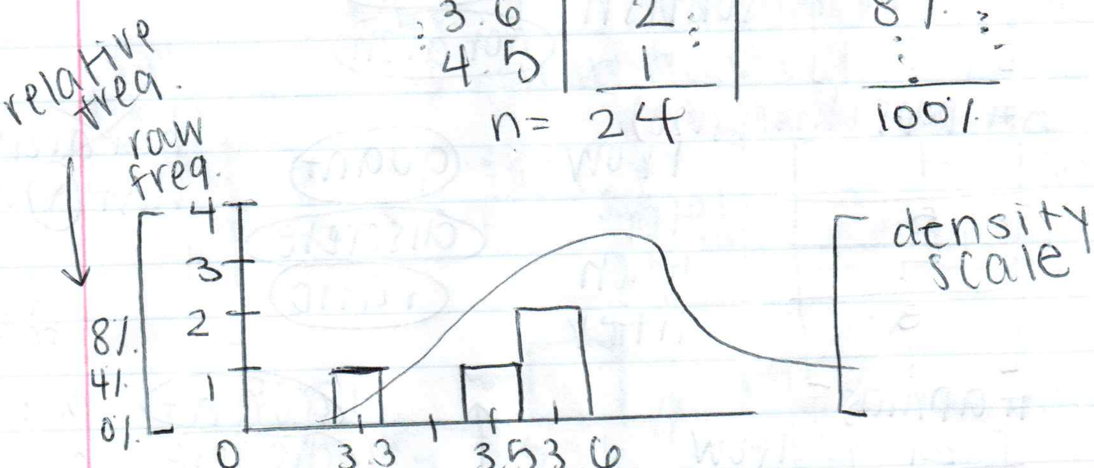
10/9/18 Wave Moretto

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

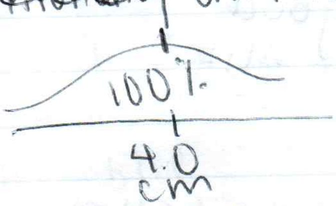
butterfly data

value	raw freq	% (relative frequency)
3.3	1	1/24 = 4%
3.4	0	0%
3.5	1	4%
3.6	2	8%
⋮	⋮	⋮
4.5	1	
n = 24		100%



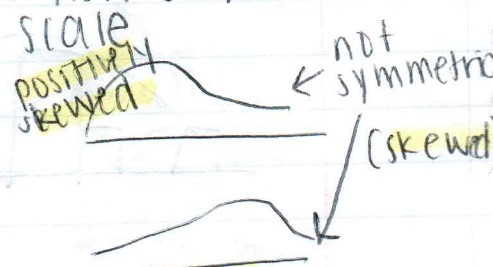
- When hist are plotted on density scale:
  a) rel. freq ⟷ area of hist. bars (curve)
  b) total area under hist. = 100%
- Convention: all hist. from now on, are implicitly ~~[illegible]~~ on the density scale

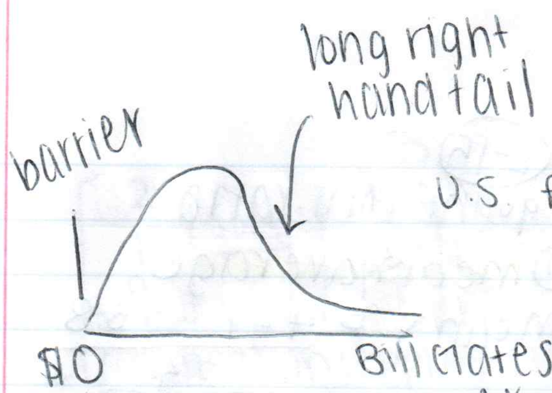
ex)



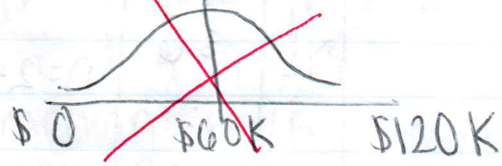
wing length

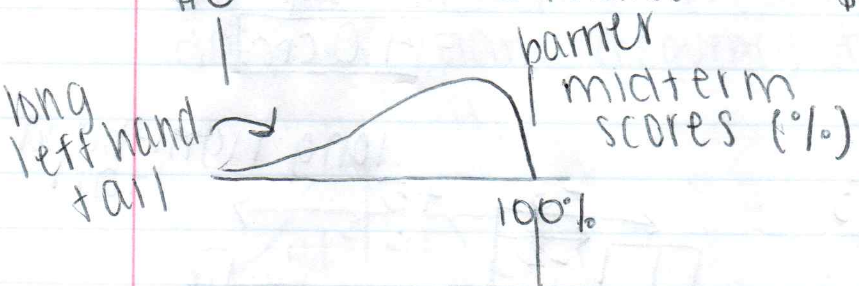
point of symmetry

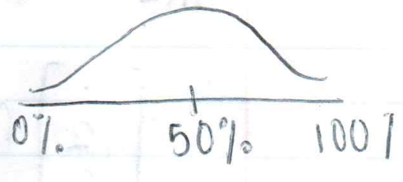




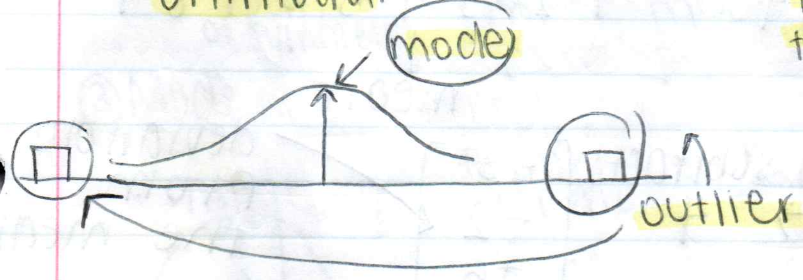
U.S. family income in 2017

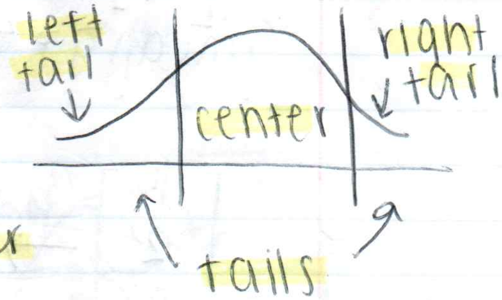




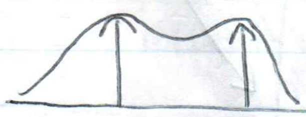


unimodal

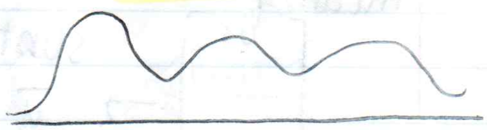


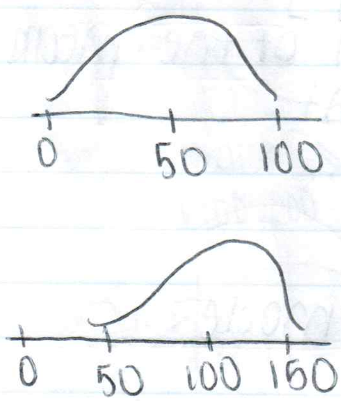


bimodal

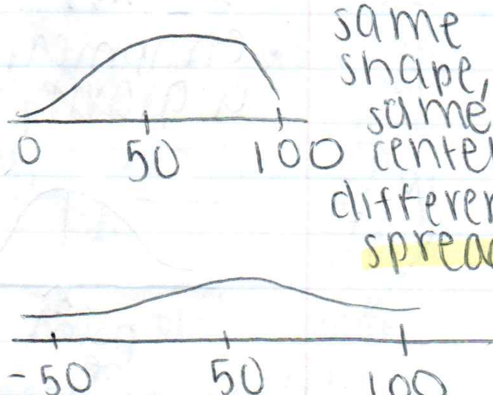


multimodal

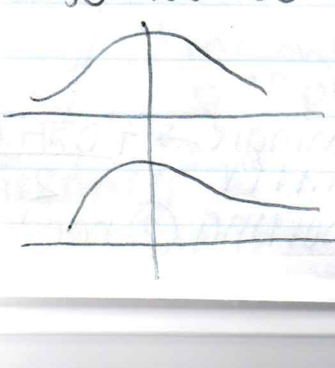




different center, same shape, same spread



same shape, same center, different spread



same center
same spread
different shape → qualitatively

Measures of center: (L-15)

$$\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array}\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.8 \end{bmatrix} \quad n = 24$$

quant. cont. ratio

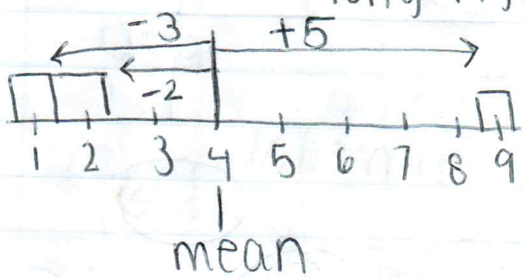
① mean/average

$$\text{mean } \bar{y} = \frac{4.4 + \dots + 3.8}{24} = \boxed{4.0\text{ cm}}$$

$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \quad n = 3$$

mean $\bar{y} = 4$

long right-hand tail



$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \xrightarrow{\text{subtract } 4} \begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$$

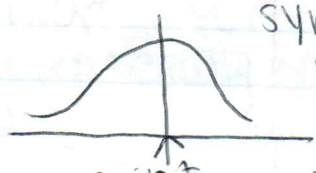
mean 4 — mean 0

deviations from the mean

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \xrightarrow{\text{subtract } \bar{y}} \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_n - \bar{y} \end{bmatrix}$$

- Graphical interpretation of the mean: center of gravity = balance point

symmetric unimodal



point of symmetry = mean = mode

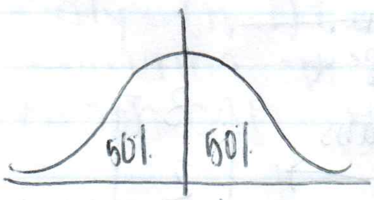
$$\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix} 3.3 \\ 3.5 \\ 4.0 \\ 4.0 \\ 4.4 \\ 4.5 \end{bmatrix}$$

middle after sorting $\rightarrow \frac{4.0 + 4.0}{2} = 4.0$ ② median

$$n=3$$

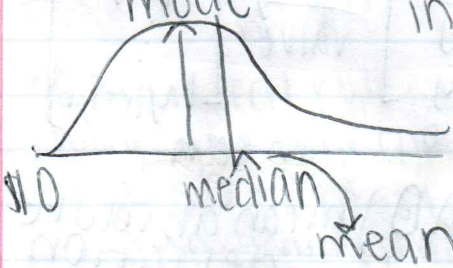
$$\begin{bmatrix}2\\1\\9\end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix}1\\2\\9\end{bmatrix} \leftarrow \text{median} \qquad \begin{bmatrix}2\\1\\3\\9\end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix}1\\2\\3\\9\end{bmatrix} \rightarrow \text{median} = 2.5$$

- Graphical interpretation of median: 50/50 point in data in relative frequency terms

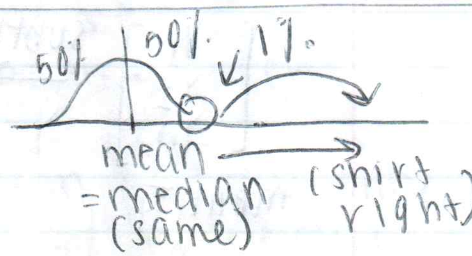


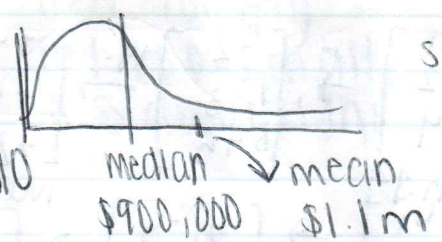
point of symmetry = mean = mode = median

③ Mode



family income





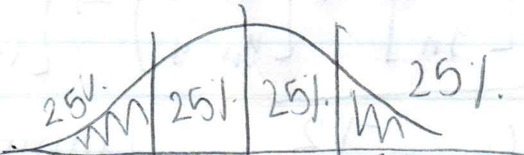
sale price of SC housing

σ lower case sigma

upper case sigma

$$\bar{y} = \frac{y_1 + \dots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

index of summation



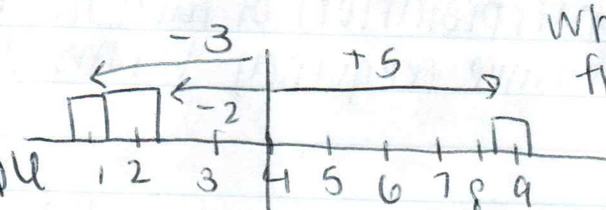
median = 50th percentile = 0.5 quantile

75th percentile = 0.75 quantile

25th percentile = 0.25 quantile

- Influence of outliers on the mean
- Mean is pulled by the tail

Measures of spread: - typical amount by which each # differs from center



sample

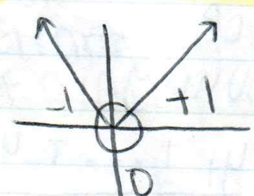
$$\begin{bmatrix}1\\2\\9\end{bmatrix} \xrightarrow[4]{\text{subtract}} \begin{bmatrix}-3\\-2\\+5\end{bmatrix} \xrightarrow[\text{values}]{\text{abs}} \begin{bmatrix}3\\2\\5\end{bmatrix}$$

mean 4   mean 0   mean = 3.3

$$\begin{bmatrix}y_1\\y_2\\y_n\end{bmatrix} \xrightarrow[\bar{y}]{\text{subtract}} \begin{bmatrix}y_1-\bar{y}\\y_2-\bar{y}\\y_n-\bar{y}\end{bmatrix} \xrightarrow[\text{value}]{\text{abs}} \begin{bmatrix}|y-\bar{y}|\\ \vdots \\ |y_n-\bar{y}|\end{bmatrix}$$

mean $\bar{y}$ mean 0   mean

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} |y_i-\bar{y}| = \text{(MAD) mean absolute deviation}}$$



not diff at 0

$$\begin{bmatrix}\$1\\\$2\\\$9\end{bmatrix} \rightarrow \begin{bmatrix}\$-2\\\$-3\\\$+5\end{bmatrix} \xrightarrow{\text{sq.}} \begin{bmatrix}+4\ \$^2\\+9\ \$^2\\+25\ \$^2\end{bmatrix}$$

mean \$4   mean: 12.7

$$\begin{bmatrix}y_1\\ \\y_n\end{bmatrix} \rightarrow \begin{bmatrix}y_1-\bar{y}\\ \\y_n-\bar{y}\end{bmatrix} \rightarrow \begin{bmatrix}(y_1-\bar{y})^2\\ \\(y_n-\bar{y})^2\end{bmatrix}$$

mean:

$$s = \boxed{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \text{(sample) standard deviation (sd)}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \text{(sample) variance}$$

$SD = \sqrt{\dfrac{\$^2 38}{2}} = \$4.4$ (100 yrs old)

✓
✓
X

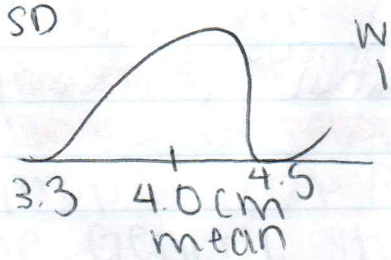
free

not free

mean

n=3

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



wing length

$\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix}$ n=24

mean $\bar{y}$=4.0cm

SD s=?

- Empirical rule: start at mean go 1 SD either way: you will capture about $\frac{2}{3}$/68% of the data
  - 2 SD → most → 95%
  - 3 SD → almost all → 99.7% (from normal curve)

ex) 0.5 too big

0.1 too small

0.3 about right