this time: measures of center & spread

read: DD book A ch. 1, 2
book B ch 1-3

AMS 7
9 Oct 18

1

next time: normal curve

lecture notes: LN pp. 1 - 22

today LN pp. 22+

new comers deadline for HW 1: next wed (8 days from now) 11:59pm

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 $\leftrightarrow$ area of histogram bars

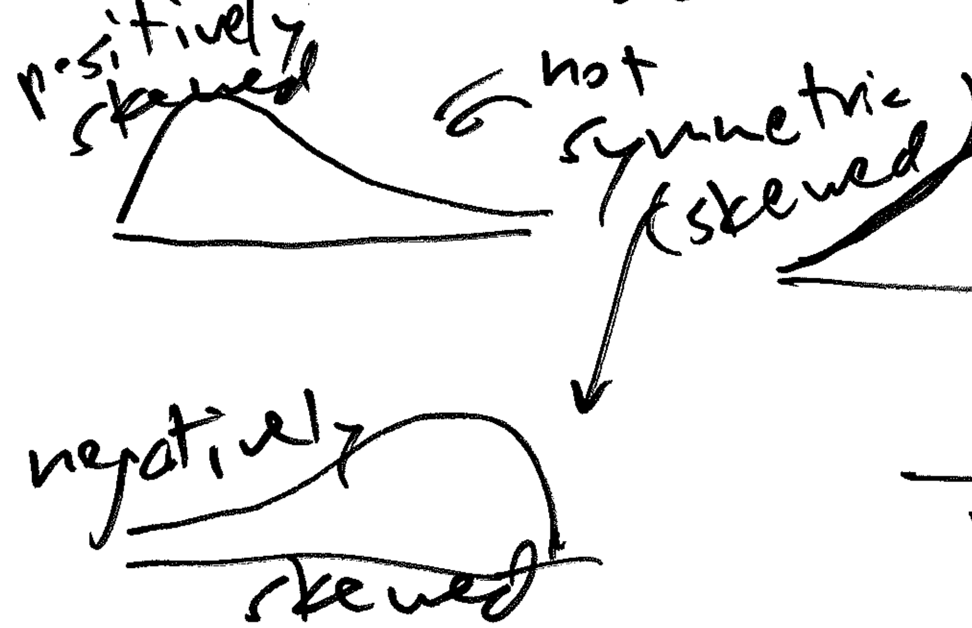
curve

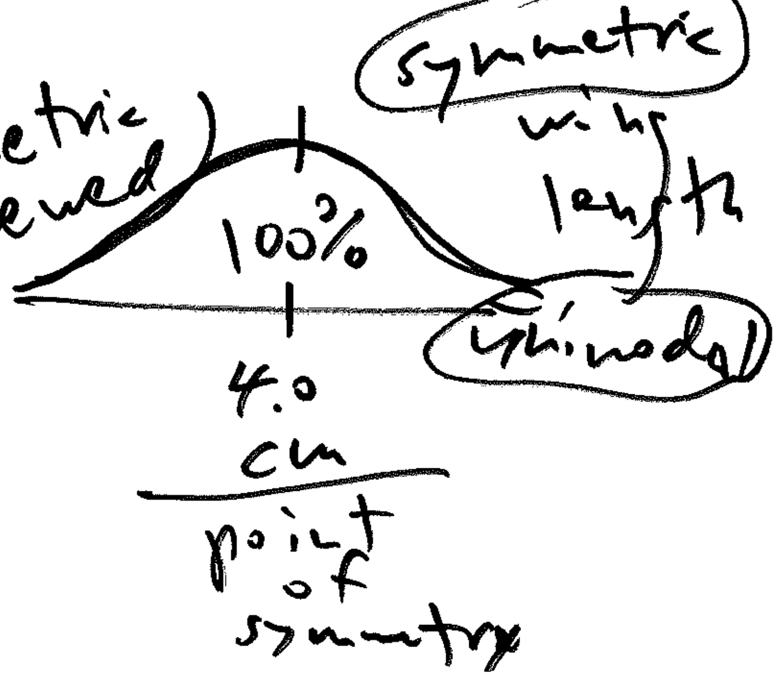
(b) total area under hist. = 100%

butterfly data

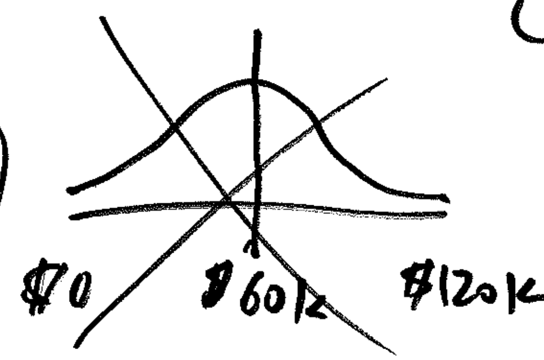
convention: all hist. from now on are implicitly on density scale

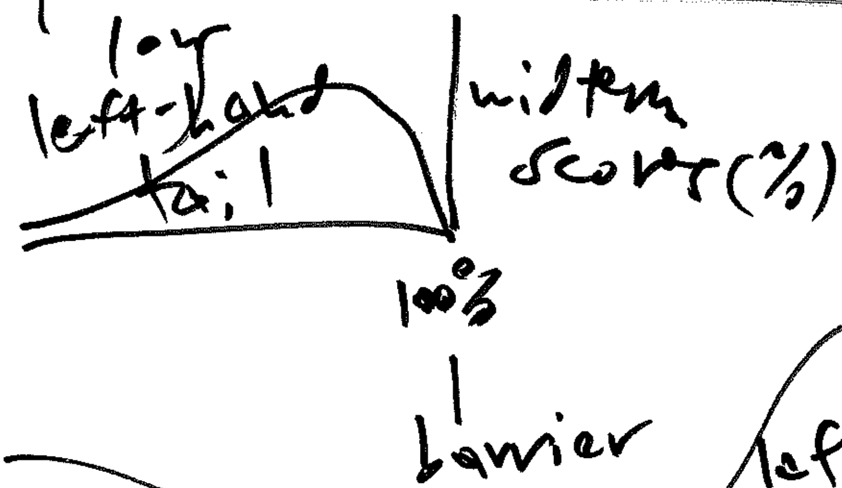
value	raw freq.	% (relative freq.)
3.3	1	$1/24 \doteq 4\%$
3.4	0	0%
3.5	1	4%
3.6	2	8% $\left(\frac{2}{24}\right)$
⋮		⋮
4.5	1	⋮
	$n = 24$	100%

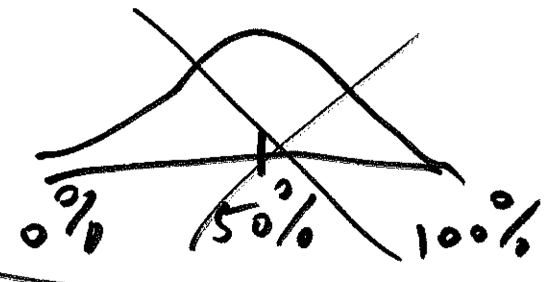


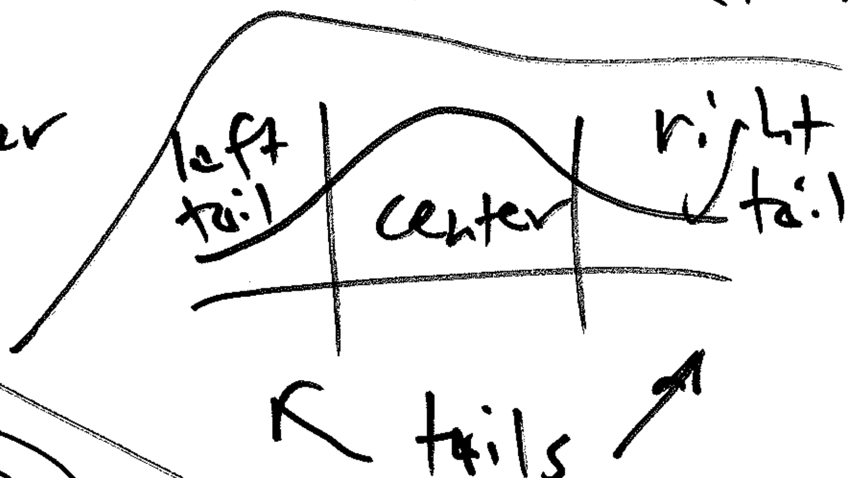


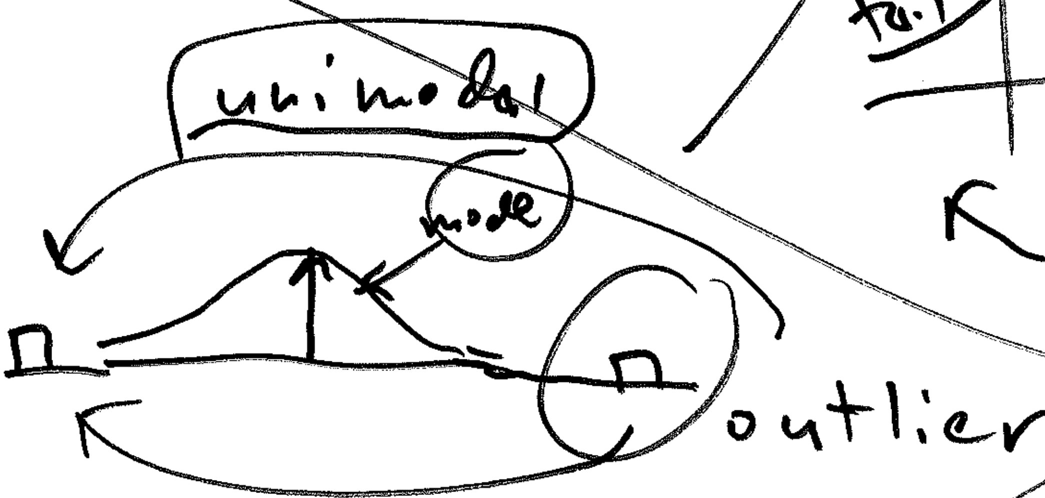

Lawier
long right-hand tail
$0
Bill Gates
U.S. family income in 2017

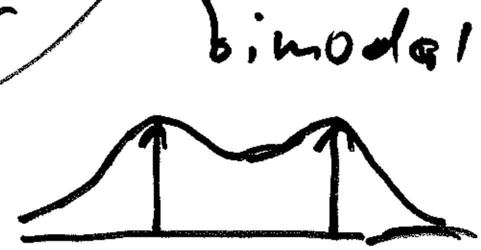

$0
$60k
$120k

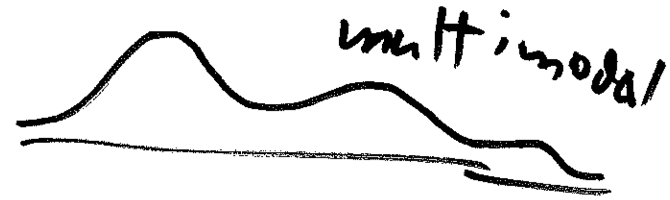

long left-hand tail
midterm scores (%)
100%
Lawier

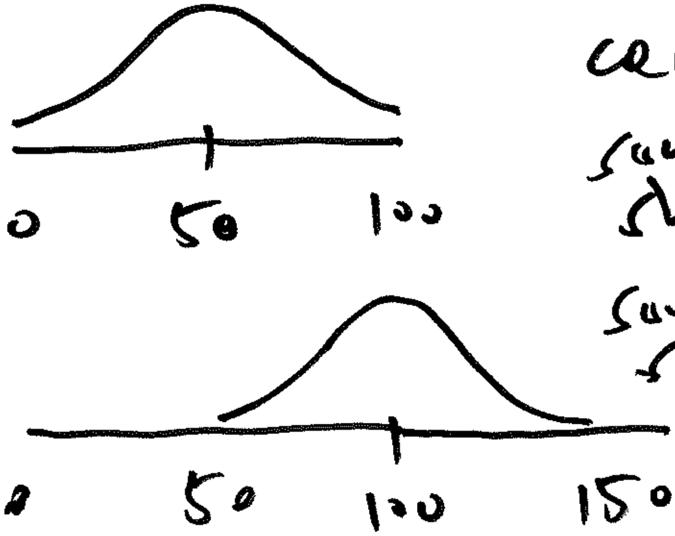

0%
50%
100%


left tail
center
right tail
tails

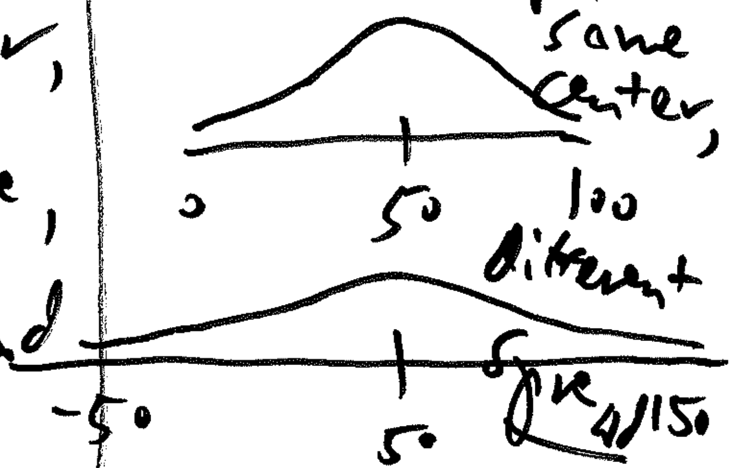

unimodal
mode
outlier


bimodal

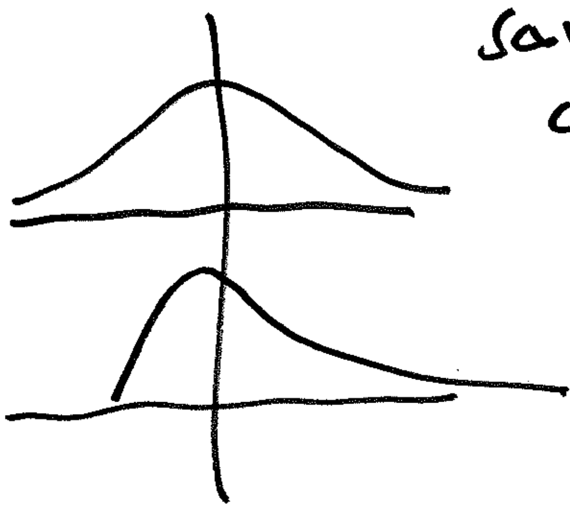

multimodal



different center, same shape, same spread



same shape, same center, different spread



same center, same spread, different **shape**

qualitatively

measures of center

L-15

quant. cont. ratio

y wing length (cm)

$$\begin{array}{l} y_1 \\ y_2 \\ \vdots \\ y_{24} \end{array} \begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \quad n = 24$$

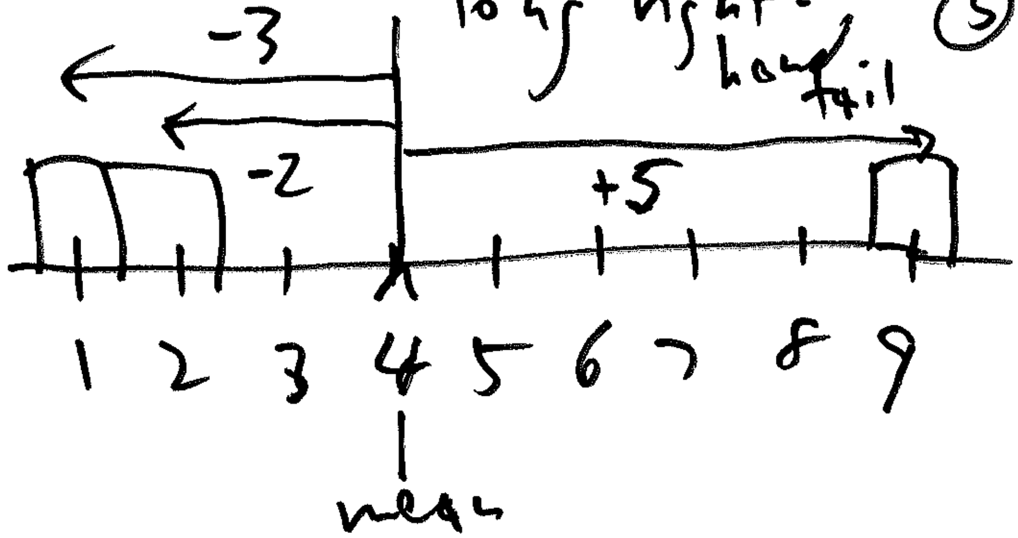
1. mean (average)
2.

(3. mode)

mean $\bar{y} = \dfrac{4.4 + \dots + 3.9}{24} \doteq 4.0\text{ cm}$

$$y = \begin{array}{c} y_1 \\ \vdots \\ y_n \end{array}\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \quad n=3$$

mean $\bar{y} = 4$



$$\underset{\text{mean } 4}{\begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix}} \xrightarrow[n=3]{\text{subtract } 4} \underset{\text{mean } 0}{\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}}$$

deviations from the mean

$$\underset{\text{mean } \bar{y}}{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}} \xrightarrow[\bar{y}]{\text{subtract}} \underset{\text{mean } 0}{\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}}$$

graphical interpretation of mean:

center of gravity

= balance point

Symmetric unimodal

point of symmetry = mean = mode

$$\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \xrightarrow[n=24]{\text{sort}} \begin{bmatrix} 3.3 \\ 3.5 \\ \vdots \\ 4.0 \\ 4.0 \\ \vdots \\ 4.4 \\ 4.5 \end{bmatrix}$$

median

middle obs after sorting

$$\frac{4.0 + 4.0}{2} = 4.0$$

$n = 3$

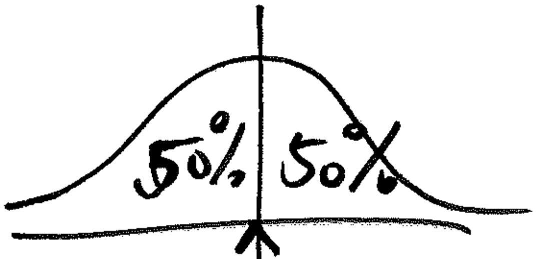
$$\begin{bmatrix} 2 \\ 1 \\ 9 \end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \leftarrow \underline{\text{median}}$$

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

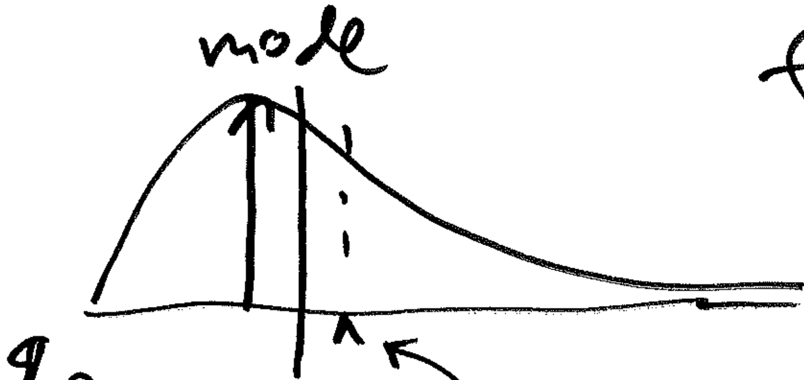
---

graphical interpretation of median

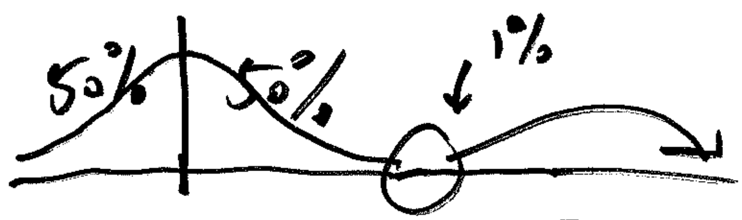
50%/50% point in data in relative frequency terms



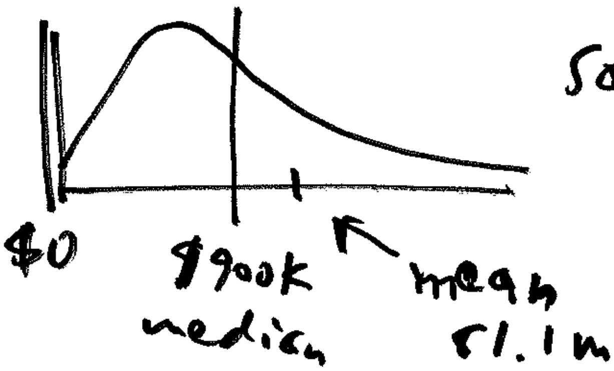
point of symmetry = mean = mode = median



family income



mean = median



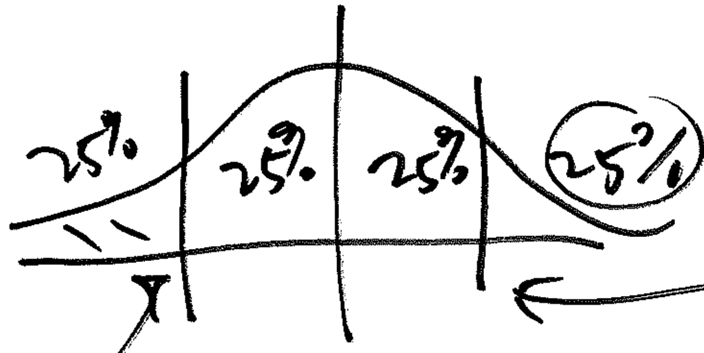
sale price of SF houses

σ lower case sigma

upper case Σ

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

index of summation



75th percentile = 0.75 quantile

median = 50th percentile = 0.5 quantile

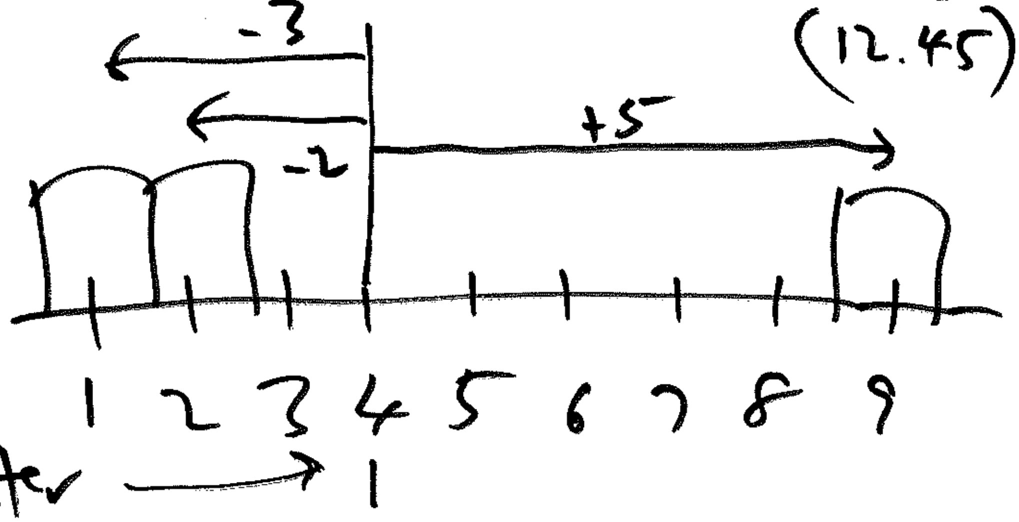
25th percentile = 0.25 quantile

influence of outliers on mean:

mean is pulled by the tail

measures of spread

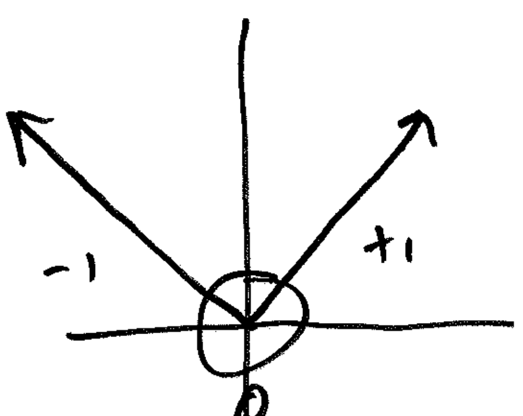
typical amount by which each # differs from center



$$\underset{\text{mean } 4}{\overset{\text{sample}}{\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}}} \xrightarrow[4]{\text{subtract}} \underset{\text{mean } 0}{\begin{bmatrix} \$-3 \\ \$-2 \\ \$+5 \end{bmatrix}} \xrightarrow{\text{absolute value}} \underset{\text{mean } \$\frac{10}{3} \doteq 3.3}{\begin{bmatrix} \$3 \\ \$2 \\ \$5 \end{bmatrix}}$$

$$\underset{\text{mean } \bar{y}}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \xrightarrow[\bar{y}]{\text{subtract}} \underset{\text{mean } 0}{\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}} \xrightarrow[\text{value}]{\text{abs.}} \underset{\text{mean}}{\begin{bmatrix} |y_1 - \bar{y}| \\ \vdots \\ |y_n - \bar{y}| \end{bmatrix}}$$

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



not diff. at 0

$$\underset{\text{mean } \$4}{\begin{bmatrix} \$1 \\ \$2 \\ \$9 \end{bmatrix}} \to \begin{bmatrix} \$-2 \\ \$-3 \\ \$+5 \end{bmatrix} \xrightarrow{\text{square}} \underset{\text{mean } \$^2 \frac{38}{3} \doteq 12.7}{\begin{bmatrix} +4\ \$^2 \\ +9\ \$^2 \\ +25\ \$^2 \end{bmatrix}}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \to \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \to \underset{\text{mean } (*)}{\begin{bmatrix} (y_1 - \bar{y})^2 \\ \vdots \\ (y_n - \bar{y})^2 \end{bmatrix}}$$

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

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

$SD = \sqrt{\frac{38 \ \$^2}{2}} = \$4.4$

(100 yrs old)

← free

✓

✓

✗

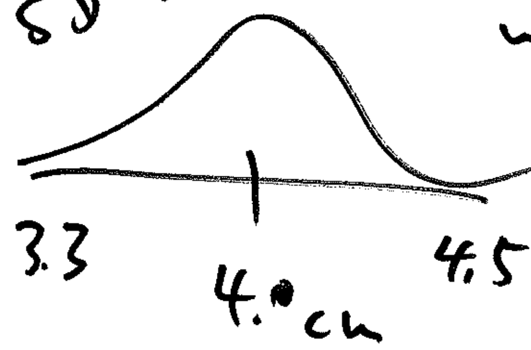
$n = 3$

not free

mean 4

the data set has $n = 3$ observations in it, but only $(n-1) = 2$ __degrees of freedom__ for measuring spread graphical interpretation of SD

SD?



wing length

empirical rule

mean

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

mean $\bar{y} = 4.0$ cm

SD $s = 0.29$

start at mean,

go $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ SD(s) either way: you will capture $\begin{pmatrix} \text{about } \frac{2}{3} \\ \text{most} \\ \text{almost all} \end{pmatrix}$ $\begin{pmatrix} 68\% \\ 95\% \\ 99.7\% \end{pmatrix}$ of the data

(normal curve)

0.5 too big
0.3 about right
0.1 too small