this time: standard deviation (SD); the normal curve

Read: DS lites ch. 1-3 (A), 1-6 (B) earlier; LN pp. 1-94

AMS 7 14 Apr 17

next time: controlled experiments & observational studies

(1)

our official note-taker is Kaitlyn Abercrombie

median

$\begin{bmatrix} 9 \\ 2 \\ 1 \end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ $n=3$

median = middle #, after data set has been sorted

$\begin{bmatrix} 9 \\ 1 \\ 4 \\ 2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 2 \\ 4 \\ 9 \end{bmatrix}$ $n=4$

median (3) mean of middle 2 #s (n even)

(n odd)

guant. y

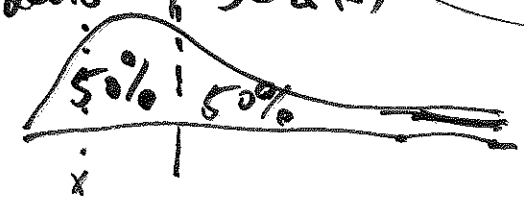
$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ n #

mean $\bar{y}$

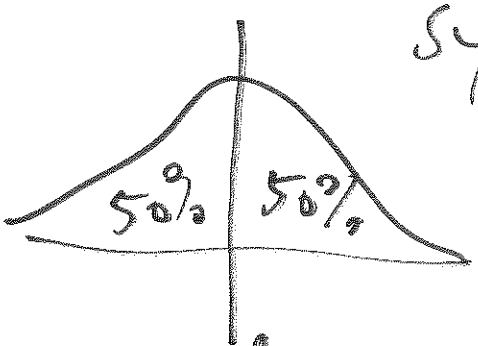
median $\tilde{y}$

graphical interpretation of median

(density scale)

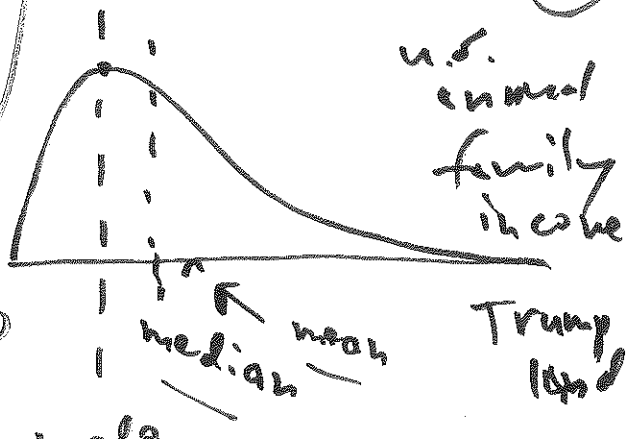


50%/50% point in data



symmetric unimodal

mode = point of symmetry = mean = median

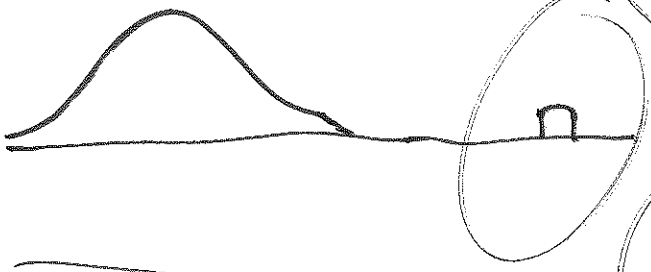


median stays same

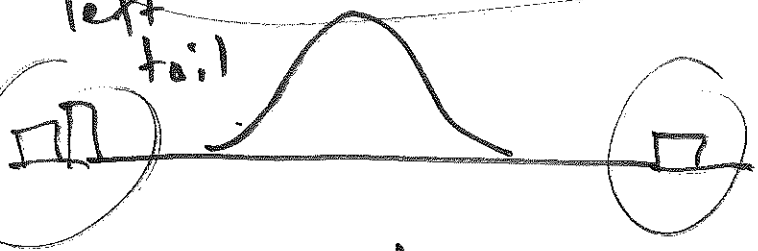
mean moves right

right tail

outlier



left tail

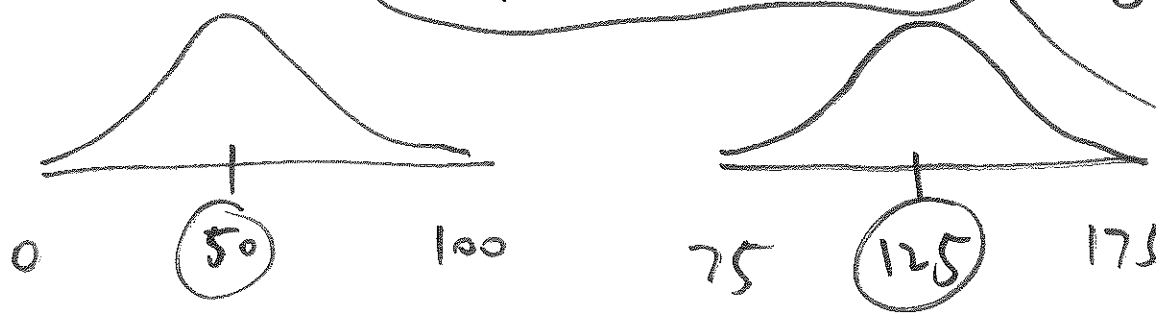


outliers

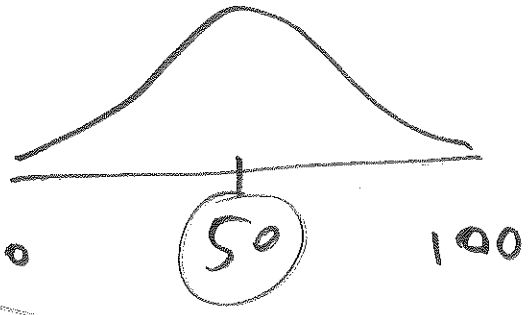
median insensitive to outliers, but mean can be highly sensitive to outliers

different center

same spread



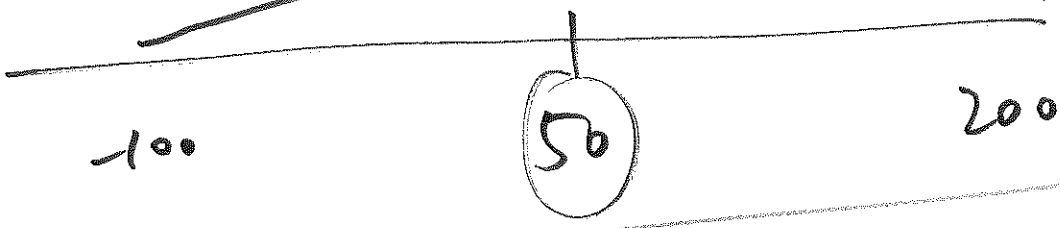
same shape

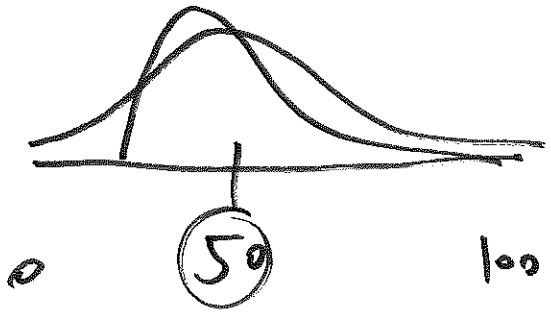


same center

different spread

same shape

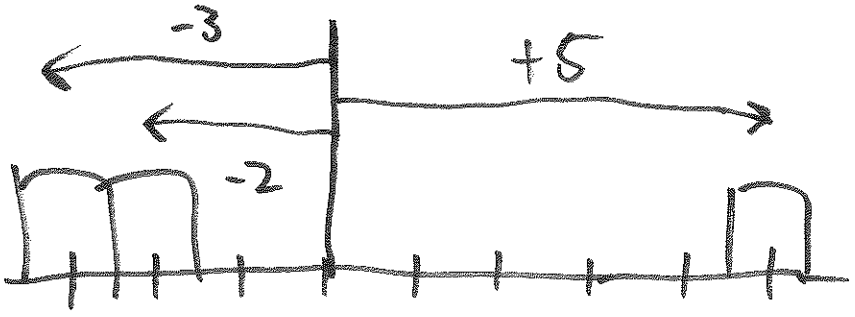




same center

same spread

different shape



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

mean $\bar{y} = 4$

mean $\bar{y} = 4$

subtract $\bar{y} = 4$ from all data values

$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$ ← deviations from the mean

mean 0 | 0

to avoid ⊕ & ⊖ cancellation,

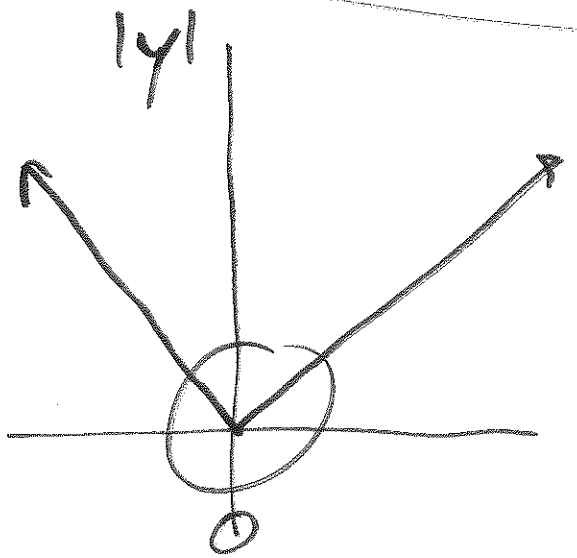
[1] absolute values

$$\begin{bmatrix} |-3| \\ |-2| \\ |+5| \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} \begin{bmatrix} |y_1 - \bar{y}| \\ \vdots \\ |y_n - \bar{y}| \end{bmatrix}$$

MAD = mean absolute deviation → mean 3.3 mean

MAD (Sir Arthur Eddington) ~1910 ⑤

not used much today



|y| is not differentiable at 0

$\begin{bmatrix} \$1 \\ \$2 \\ \$9 \\ \$4 \end{bmatrix}$ subtract 4 $\begin{bmatrix} \$-3 \\ \$-2 \\ \$+5 \end{bmatrix}$ square $\begin{bmatrix} (-3)^2 = +9\ \$^2 \\ (-2)^2 = +4 \\ (+5)^2 = +25 \end{bmatrix}$

mean $4 mean $12.7 $^2

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

mean $\bar{y}$

$$\frac{(-3)^2 + (-2)^2 + (+5)^2}{3-1} = \text{(sample) variance} = s^2$$

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

(sample) standard deviation = $\sqrt{\text{(sample) variance}} = s$

(SD)

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

here

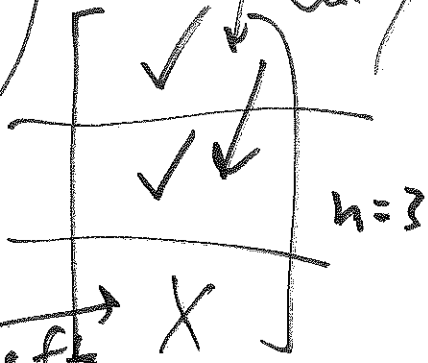
$$s = \sqrt{\frac{38}{2}} = 4.3$$

why $(n-1)$ not n?

$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ subtract mean $\rightarrow$ $\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$

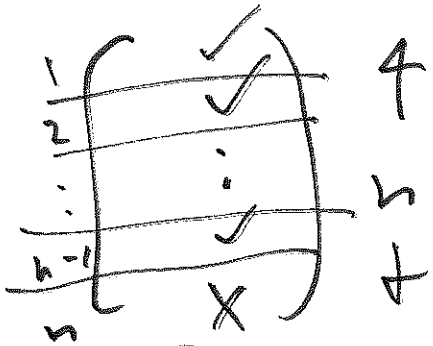
mean $\bar{y} = 4$

free to vary



no. of freedom left

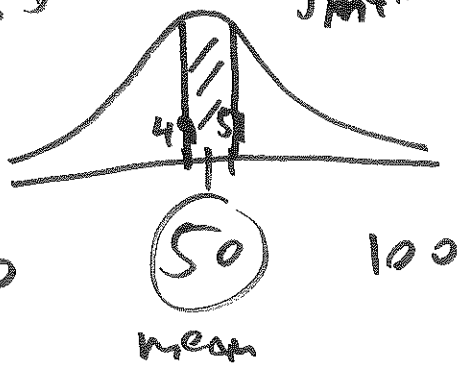
mean 4



mean $\bar{y}$

a data set with n obs. only has $(n-1)$ degrees of freedom for measuring spread

SD 1 ← too small

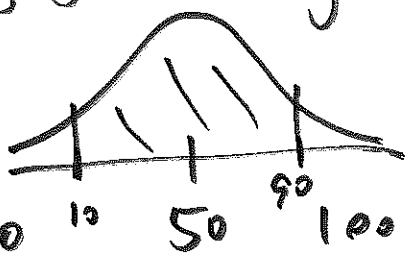


graphical interpretation of SD

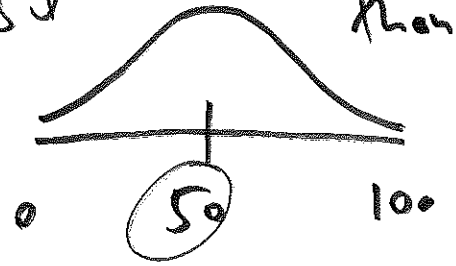
empirical rule

For virtually any data set, if you start at the mean & go $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ SD(s), either way, you will usually capture (about 2/3 ~~68%~~ / most 95% / almost all ~~99.7%~~) of the data in that interval

SD too big



SD a bit smaller than 25



5) ~~0.1~~ too small ~~0.5~~ too big $\approx$ 0.29 cm