AMS 7
12 April 18

this time: center & spread } normal curve

read: DD ch. A: 1-3

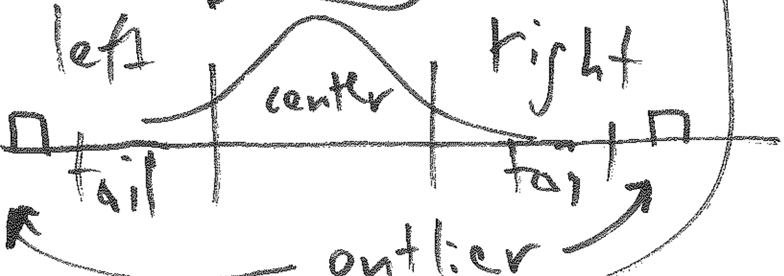
next time: experimental design

DD B: ch. 1-6

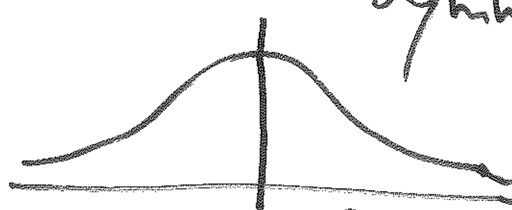
lecture notes ①

LN pp. 1-94

tails



symmetric



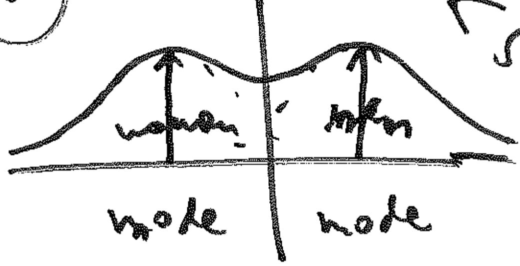
skewed

= asymmetric

long right-hand tail

2

bimodal



← symmetric →

height

1

unimodal symmetric

= mode

= mean

(multimodal)

≥ 2

numerical measures of center

(1) mode when unimodal

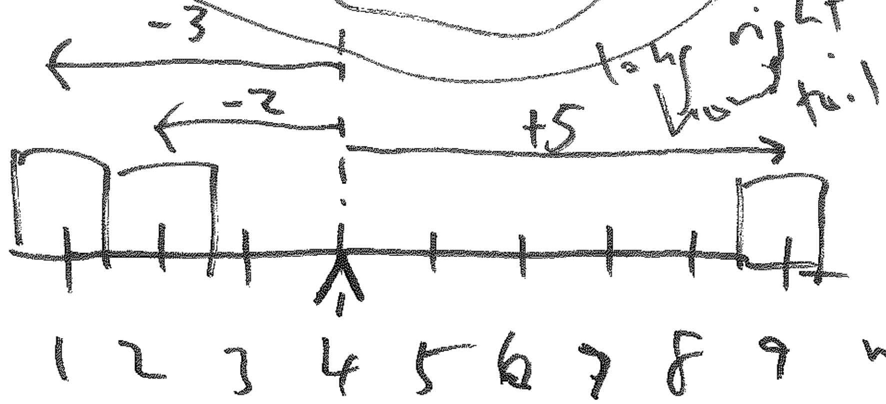
(2) mean

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

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad n$$

mean $\bar{y}$

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



mean $\bar{y} = 4$

outlier

mean = center of gravity

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

mean 0

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

deviations (from the mean)

mean 0

(3) median $\xrightarrow{\text{sort}} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

median = center (middle) data value, after sorting from smallest to largest

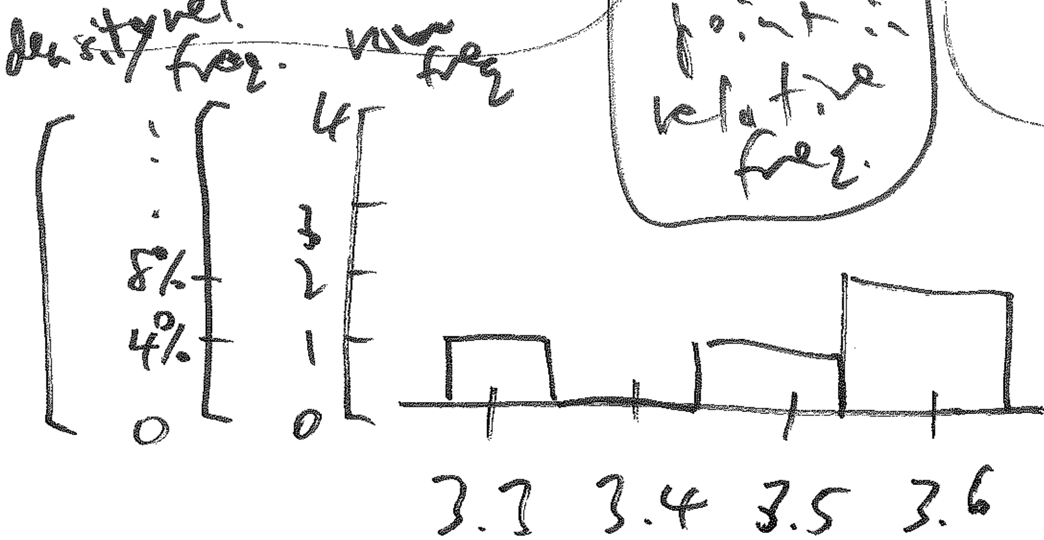
50/50 point in relative freq.

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

mean 4 median 2

$n=4$ (even) $\begin{bmatrix} 9 \\ 1 \\ 4 \\ 2 \end{bmatrix} \to \begin{bmatrix} 1 \\ 2 \\ 4 \\ 9 \end{bmatrix}$

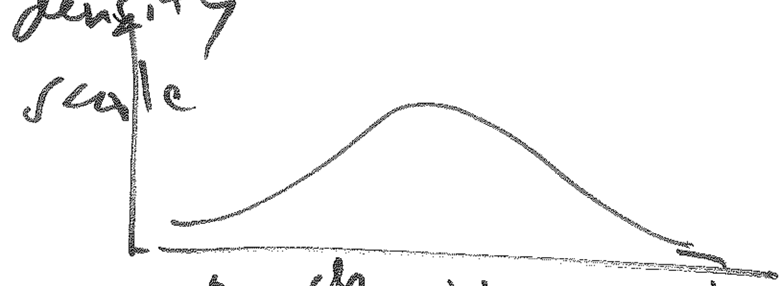
median = mean of 2 middle #s



raw freq hist = identical to rel. freq hist. = density scale hist.

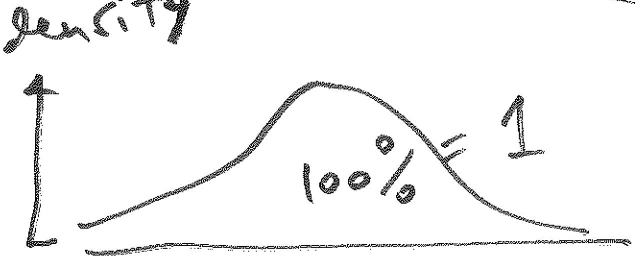
value	raw freq.	relative freq.
3.3	1	$(1/24) \cdot 100\% = 4\%$
3.4	0	$(0/24) \cdot 100\% = 0\%$
3.5	1	1 (4%)
3.6	2	$(2/24) \cdot 100\% = 8\%$
⋮	⋮	
4.5	1	
	$n = 24$	

magic
of
density
scale

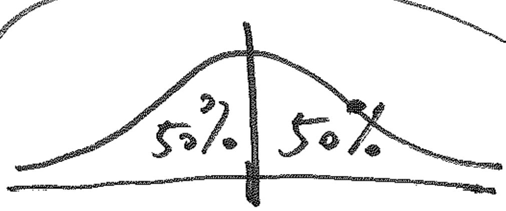


on density scale

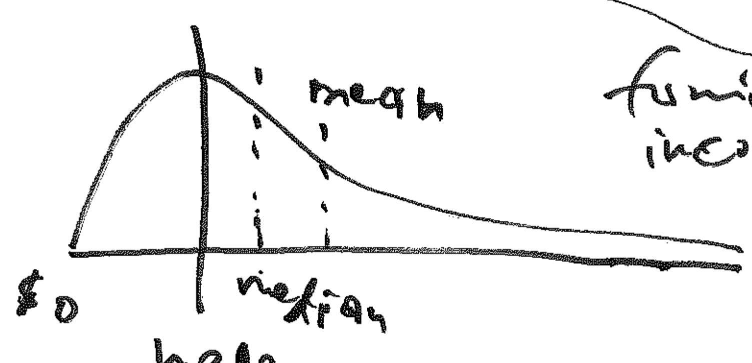
relative frequency = area under hist (curve)



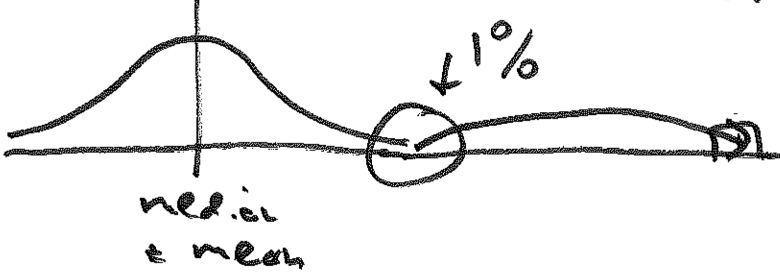
convention: all hist. sketches in this course are implicitly drawn on density scale



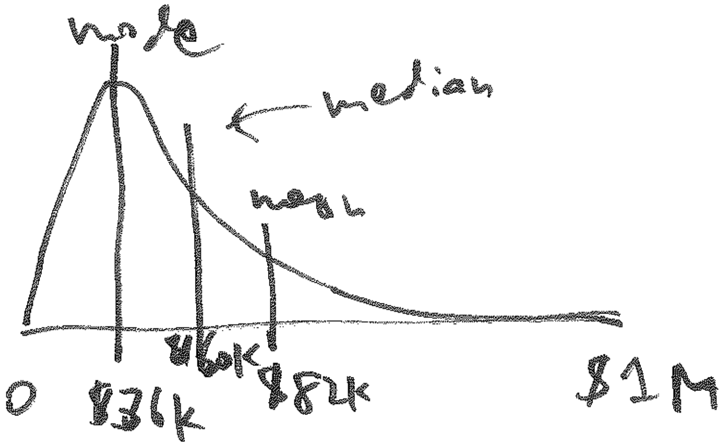
point of symmetry
= mode
= mean
= median

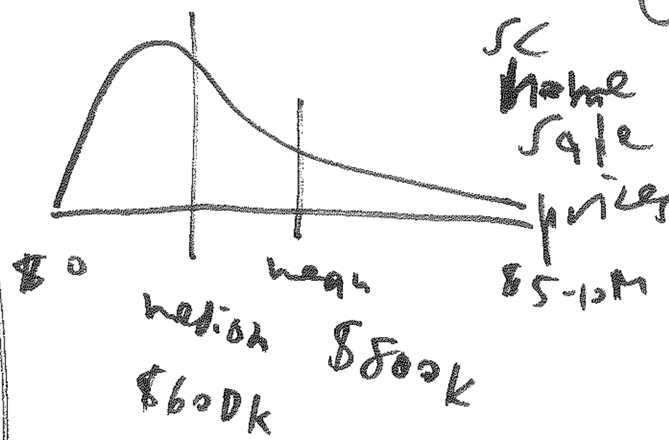


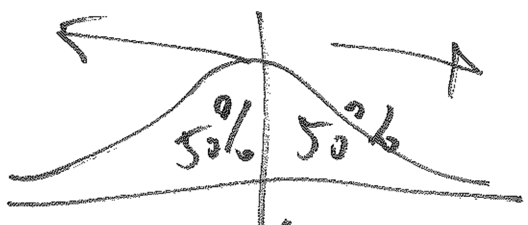
family income



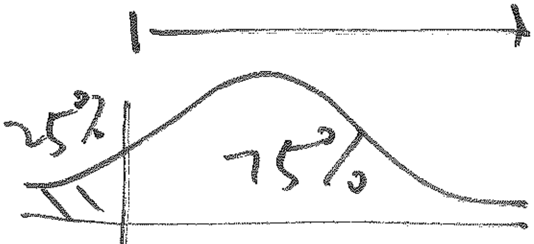
median same, mean shifted to right







median = 50^{th} percentile

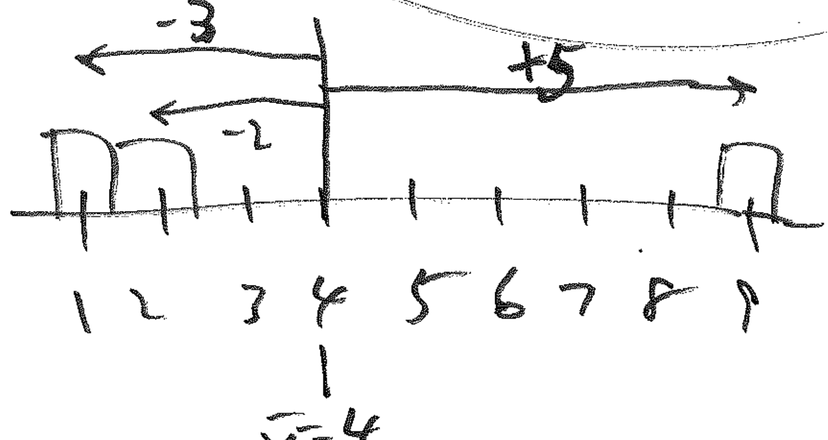


25^{th} percentile = 25^{th} quantile

7

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

mean $\bar{y} = \$4$



$\boxed{\text{measures of spread}}$

$$\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix} \leftarrow \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$$

mean 0

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

mean 0

⑥

one idea to avoid +/− cancellation

$$\begin{pmatrix} -3 \\ -2 \\ +5 \end{pmatrix} \xrightarrow{\text{take absolute values}} \begin{pmatrix} +3 \\ +2 \\ +5 \end{pmatrix}$$

mean $\frac{10}{3} = 3.3$

$$\begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \xrightarrow{|\;|} \begin{pmatrix} |y_1 - \bar{y}| \\ |y_2 - \bar{y}| \\ \vdots \\ |y_n - \bar{y}| \end{pmatrix}$$

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

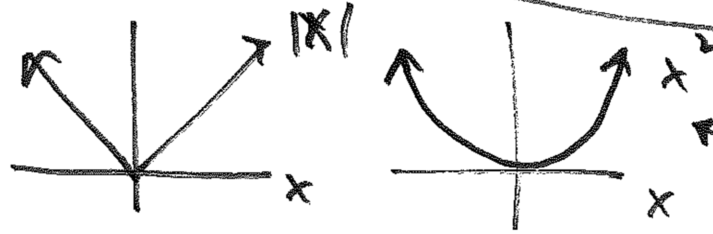
another way

$$\begin{pmatrix} \$ \\ -3 \\ \$ \\ -2 \\ \$ \\ +5 \end{pmatrix} \xrightarrow{\text{square}} \begin{pmatrix} (-3)^2 = +9 \; \$^2 \\ (-2)^2 = +4 \; \$^2 \\ (+5)^2 = +25 \; \$^2 \end{pmatrix}$$

mean $\frac{38}{3} = 12.7 \; \$^2$ $\quad \sqrt{\$^2 \, 12.7} = \$3.7$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \xrightarrow[\bar{y}]{\text{subtract}} \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \xrightarrow{(\;)^2} \begin{bmatrix} (y_1 - \bar{y})^2 \\ \vdots \\ (y_n - \bar{y})^2 \end{bmatrix}$$

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



matches bell curve

7

(intuitive)

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

(better)

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

sample variance

$$\begin{bmatrix} +9 \\ +4 \\ +25 \end{bmatrix}$$

sum 38

sample variance $\frac{38}{2} = \cancel{8^2}\ 19$

SD $\sqrt{\cancel{8^2}\ 19} = \cancel{8}\ 4.3$

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

sample standard deviation (SD)

sample variance $= \left(\text{sample SD}\right)^2$

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

mean 4 mean 0

$$\begin{bmatrix} \cancel{y_1} \\ y_2 \\ \cancel{y_3} \end{bmatrix}$$

mean $\bar{y}$ ← constrain mean to

free to vary

$n=3$

$$\begin{bmatrix} \checkmark \\ \checkmark \\ \end{bmatrix}$$

mean 4

a dataset with n obs. actually only has $(n-1)$ independent pieces of information about spread

degrees of freedom

empirical rule

graphical interpretation of SD

For almost all datasets,

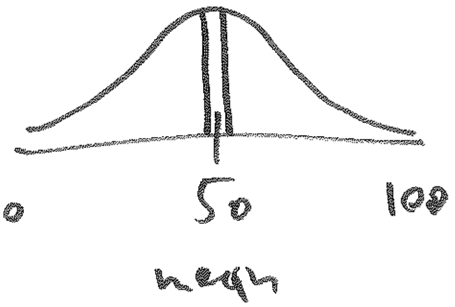
if you start at the mean

and go $\begin{Bmatrix} 1 \text{ SD} \\ 2 \text{ SDs} \\ 3 \text{ SDs} \end{Bmatrix}$ either way,

you will usually encompass (capture)

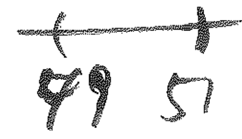
about $\begin{Bmatrix} 2/3 \\ \text{most} \\ \text{almost all} \end{Bmatrix} = \begin{Bmatrix} 68\% \\ 95\% \\ 99.7\% \end{Bmatrix}$ of the data

rough theoretically



SD 1 way too small

50 ± 1

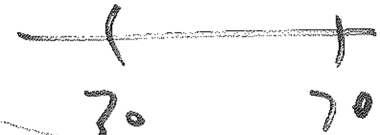


SD 50 way too big

50 ± 50 = all of data

SD 10 too small

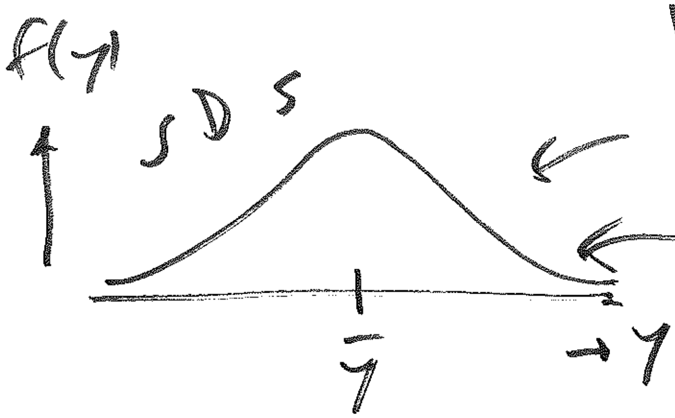
50 ± 2(10)



SD 20 : about right

butterfly data: $s \doteq 0.2$ cm

actual: $s \doteq 0.29$ cm



bell curve = Normal curve (1800) = Gaussian

$$f(y) = \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{1}{2s^2}(y-\bar{y})^2\right]$$

distribution

de Moivre (1720)