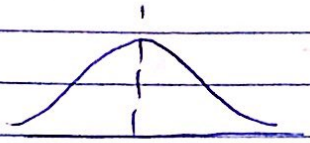
9th October 2018 — Nikhila Kadiyala

3 possible vertical scales for Histograms
- raw area
- % (relative frequency)
- density scale
  - (a) relative frequency
  - (b) total area under Histogram = 100%

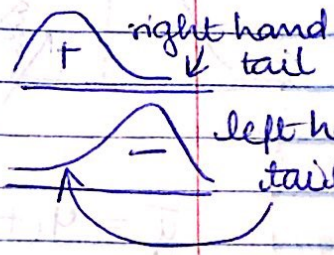
value	raw area	% (relative frequency)
3.3	1	4%
3.4	0	0%
3.5	1	
3.6	2	8%

Convention: all histograms are implicitly on the density scale

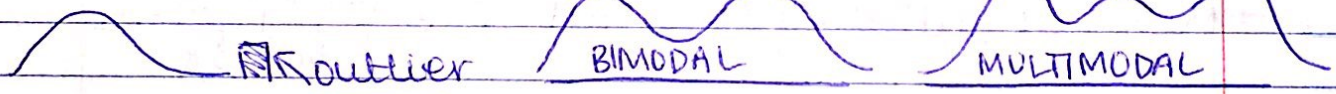


← point of symmetry

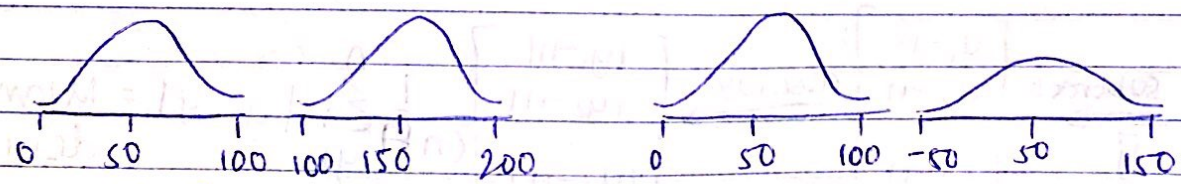
Not symmetric = skewed



- Skewness occurs when there is a barrier beyond/below which you can't go.



Bell Curve – symmetric, unimodal

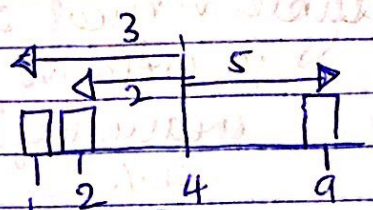


same shape, different center

same shape, same center, different spread

Measures of center:
1. Mean (average)
2. Median (sort and the middle # is median)
3. Mode

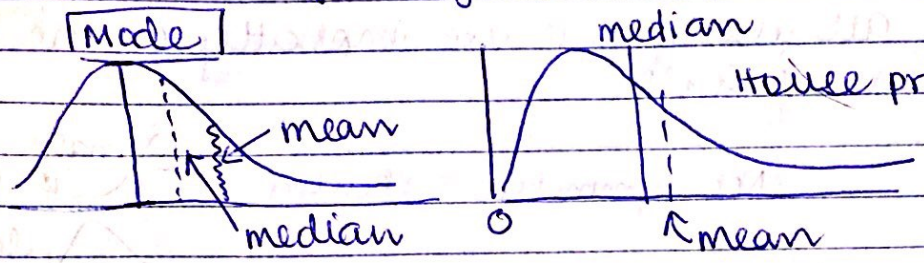
$Y = \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ $n=3$ Mean $\bar{y}=4$



$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ subtract 4 $\rightarrow \begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$ ← deviations of the mean

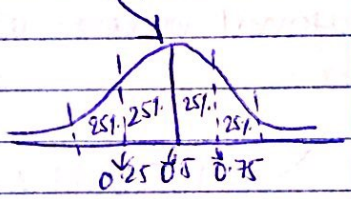
center of gravity (balance point)

mean of $(y_n - \bar{y}) = 0$



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

Median: 50% percentile
0.5 quantile



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

Measures of spread: Typical amount by which each # differs from centre

$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$ subtract $\bar{y}$ $\rightarrow \begin{bmatrix} y_1 - n \\ y_2 - n \\ \vdots \\ y_n - n \end{bmatrix}$ absolute $\rightarrow \begin{bmatrix} |y_1 - n| \\ |y_2 - n| \\ \vdots \\ |y_n - n| \end{bmatrix}$ $\frac{1}{(n-1)}\sum_{y_i}^{n} |y_i - \bar{y}|$ = Mean Absolute deviation

↓ Square

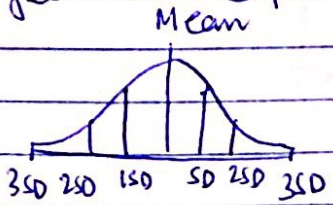
$\begin{bmatrix} (y_1 - n)^2 \\ (y_2 - n)^2 \\ \vdots \\ (y_n - n)^3 \end{bmatrix}$ $\frac{1}{(n-1)}\sum_{y_i}^{n} (y_i - \bar{y})^2$ = variance (s^2)

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

• Graphical interpretation of SD.

EMPIRICAL RULE:

① Start @ mean, go 1 SD either way, you will capture $\frac{2}{3}$ of the data (≈68%.)



② start @ mean, go 2 SDs either way capture most data (95%.)

③ start @ mean, go 3 SDs either way, capture almost all data (99.7%.)