14th January 2025 (Tuesday) →

Nsr Ankified

Median →

Ex: 7, 9, 5, 10, 4, 8, 12 = X

$\tilde{x}$ = median.

First, we need to order the data.

x = 4, 5, 7, 8, 9, 10, 12

Median = $\frac{12-4}{2} + 4 = 4 + 4 = 8$ → Works if n is odd → Does not work (Why? What if data skewed)

If n is even, the median is any number b/w the middle two numbers.

$$\therefore \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , n \text{ is odd} \\ \frac{1}{2}\left[x_{\lfloor \frac{n}{2} \rfloor} + x_{\lfloor \frac{n}{2} \rfloor + 1}\right] & , n \text{ is even.} \end{cases}$$

Ex → Marks = 50, 41, 72, 89, 92, 72, 60, 100

ordered = 41, 50, 60, 72, 72, 89, 92, 100

n → even.

$$\therefore \tilde{x} = \frac{1}{2}\left[x_4 + x_5\right]$$

$$= \frac{1}{2}\left[(2)(72)\right] = 72$$

Now, in a frequency table.

X_i	f_i	C.F.
x_1	f_1	f_1
x_2	f_2	f_1+f_2
⋮		⋮
x_K	f_K	n
	$\sum f_k = n$	

If n → odd

$F_i \geq \frac{n}{2}$

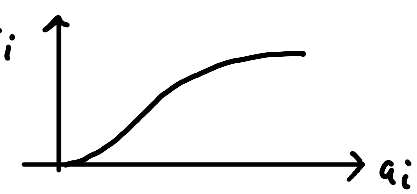
Now, if this is b/w two classes,

$F_i \geq \frac{n}{2} \rightarrow x$ $\Rightarrow \tilde{x} = \frac{x + x'}{2}$

$F_i \geq \frac{n}{2} + 1 \rightarrow x'$

What if the data is continuous?

Ogive, for example, looks like —



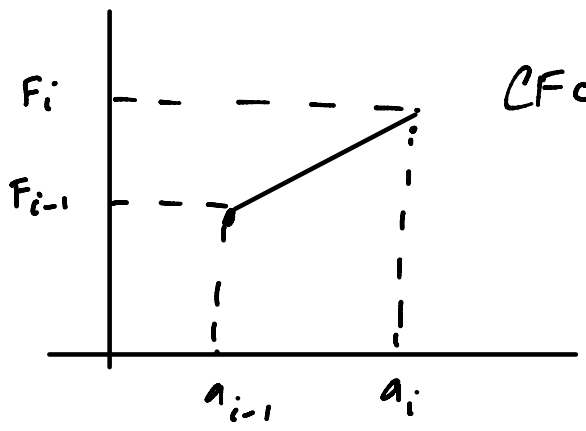
i.e. data is like,

Int	f_i	F
$[a_1, a_2]$	f_1	f_1
$(a_2, a_3]$	f_2	f_2+f_1
⋮	⋮	⋮

Say, we have,

$$F_i \geq \frac{n}{2} \rightarrow (a_i, a_{i+1}]$$

We assume that F_i grows linearly b/w $i-1^{th}$ and i^{th} class.



CF changes linearly – assumption.

$\therefore$ Eqn of line,

$$\frac{y - F_{i-1}}{\underbrace{F_i - F_{i-1}}_{f_i}} = \frac{x - a_{i-1}}{\underbrace{a_i - a_{i-1}}_{h}}$$

We should also substitute, $y = \frac{n}{2}$

$$\therefore \frac{\frac{h}{2} - F_{i-1}}{f_i} = \frac{x - a_{i-1}}{h} \Rightarrow \boxed{x = \frac{h}{f_i}\left[\frac{n}{2} - F_{i-1}\right] + a_{i-1}}$$

How find out the median graphically? Simply, it the point of intersection of ogive with $F_i = \frac{n}{2}$.

Δ Formal defn of median: For any cont.. r.v. X, the median 'm' is defined as that value for which,

$$P(X \geq m) \geq \frac{1}{2} \text{ and } P(X \leq m) \geq \frac{1}{2}$$

Say, we do n expt and on the r.v. values.

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

$n = Odd = 2k+1 \longrightarrow$

Let $m = k+1$

$$\Rightarrow \frac{\#obs \leq x_m}{\#obs} = \frac{\#obs \leq x_{k+1}}{2k+1} = \frac{k+1}{2k+1} \geq \frac{1}{2}$$

Also, $\frac{\#obs \geq x_m}{\#obs} = \frac{k+1}{2k+1} \geq \frac{1}{2}$

$n = Even = 2k$

$\Rightarrow m = \frac{1}{2}\left[x_k + x_{k+1}\right]$

$$\therefore \frac{\#obs \geq m}{\#obs} = \frac{k}{2k} = \frac{1}{2}$$

$\therefore$ Median defined for data satisfies the defn. for probability.

Merits	Demerits
① Rigid Defn	① no exact $\tilde{x}$ for n = even
② Easily understood	② algebraic treatment - not possible
③ Not affected by outliers	③ not based on all obs
	④ Affected by sampling fluctuations

What is the optimisation provided by the median ?

$$\tilde{x} = \text{argmin} \left\{ \sum_i |x_{(i)} - c| \right\}$$

Proof : Say, data is,

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Say, $n = 2k+1$ (odd)

Consider 2 terms,

$$|x_{(1)} - c| + |x_{(n)} - c|$$

When $c < x_{(1)}$,

$$|x_{(1)} - c| + |x_{(n)} - c| = x_{(1)} - c + x_{(n)} - c + x_{(1)} - x_{(1)}$$

$$= (x_{(n)} - x_{(1)}) + 2\underbrace{(x_{(1)} - c)}_{\hookrightarrow +ve}$$

If $c > x_{(n)}$,

$$|x_{(1)} - c| + |x_{(n)} - c|$$

$$= c - x_{(1)} + c - x_{(n)} + x_{(n)} - x_{(n)}$$

$$= (x_{(n)} - x_{(1)}) + 2\underbrace{(c - x_{(n)})}_{\hookrightarrow +ve}$$

Now, $x_{(1)} \leq c \leq x_{(n)}$

$$\therefore |x_{(1)} - c| + |x_{(n)} - c|$$

$$= c - x_{(1)} + x_{(n)} - c$$

$$= x_{(n)} - x_{(1)} \longrightarrow \underline{\text{Minimal}}$$

$\Rightarrow$ for $c \in [x_{(1)}, x_{(n)}]$, we will minimize $|x_{(1)} - c| + |x_{(n)} - c|$

Similarly proceeding,

for $c \in [x_{(k)}, x_{(n-k+1)}]$, we will minimize $|x_{(k)} - c| + |x_{(n-k+1)} - c|$

16th January 2025 (Thursday) →

⓫ Mode →

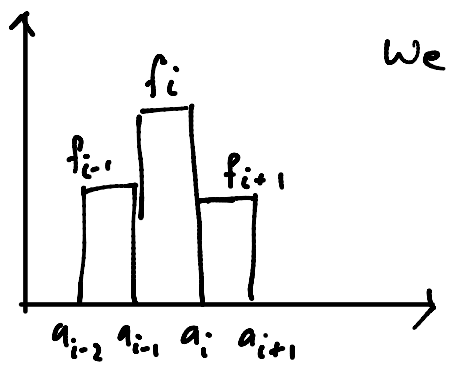
$x = 10, 7, 8, 9, 10, 5, 12, 6, 10, 15, 9, 8, 10, 9, 12$

$\hat{x} = \text{Mode} = 10$

What happens if there is more than one element having highest frequency?

Mode is not defined here — Mode only defined for unimodal distributions.

What do we do with continuous data?

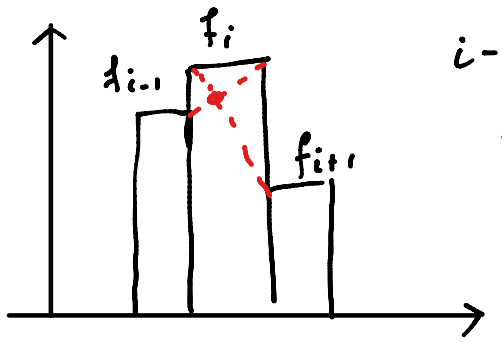


We just take the mode to be the middle point of

$(a_{i-1}, a_i]$ → i.e,

$$\hat{x} = \frac{a_{i-1} + a_i}{2} = a_{i-1} + \frac{h}{2}$$ → Assuming const class size.

What if we have →



Makes sense for the mode to be biased towards $i-1$ class.

∴ We find intersection of these interpolated linear curves. Like joining (a_{i-1}, f_{i-1}) to (a_i, f_i)

$$\frac{y - f_{i-1}}{f_i - f_{i-1}} = \frac{x - a_i}{\underbrace{a_i - a_{i-1}}_{h}}$$

$$\Rightarrow y = f_{i-1} + \frac{(f_i - f_{i-1})(x - a_{i-1})}{h}$$

and the other line, joining (a_{i-1}, f_i) to (a_i, f_{i+1})

$$\frac{y - f_i}{f_{i+1} - f_i} = \frac{x - a_{i-1}}{\underbrace{a_i - a_{i-1}}_{h}} \Rightarrow y = f_i + \frac{(f_{i+1} - f_i)}{h}(x - a_{i-1})$$

Solving these,

$$\hat{x} = a_{i-1} + \frac{(f_i - f_{i-1})h}{2f_i - (f_{i+1} + f_{i-1})}$$

Empirical relation for nearly symmetric distribution →

$$\bar{x} - \hat{x} = 3(\bar{x} - \tilde{x})$$

For perfectly symmetric dist,

$$\tilde{x} = \hat{x} = \bar{x}$$

Merits / Demerits of mode →

Merits	Demerits
→ Easily computed by inspection	→ Ill defined
→ Not affected by outliers	→ Not all observation
→ Unequal class-interval, or open-ended of class intervals	→ Sampling fluctuation.
	→ Not mathematically treatable (Chronic cancer)

✗ argmin f(i)

(*) Remarks →

① mean, median (except even n) rigidly defined.
Mode → only for unimodal data.

② Determination of mode not possible if very few data points are given

③ Mean → value passed by each unit if total volume ($\sum x_i$) were distributed equally amongst all units.
$\tilde{x}, \hat{x}$ → missing data.

We may ignore top 25% and bottom 25% of data in $\bar{x}$ calculation to deal with the outlier effect. → This is called a trimmed mean.
If don't delete, but replace top 25% with value just below it, bottom 25% with value just above ↳ Winsorized mean.

(*) We thus generally take $\bar{x}$ to be the best measure
→ The minimization of sum of squares deviations, accounting all of the data, etc.

Ex: Weights (kg) : 138, 143, 141, 159, 152, 148, 160, 267

$n = 8$, $\bar{x} = 161$ kg → Not good central measure, 7 lie below it

Why? 267 is extreme outlier.

Here, $\tilde{x}$ is appropriate.

Eg: 4 values, {15, 30, 33, 36}

we sample (i.e. take out randomly with replacement)

		$\bar{x}$	$\tilde{x}$
1	15, 30, 33	26	30
2	15, 30, 36	27	30
3	15, 33, 36	28	33
4	30, 33, 36	33	33

→ Median is pretty stable, mean fluctuates (due to 15, outlier)

There are 2 more central tendencies, not generally used but useful in some distributions →

(I) Geometric Mean →

$$x_g = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

$$\Rightarrow \log x_g = \frac{1}{n} \sum_{i=1}^{n} \log x_i \quad \rightarrow \text{A.M. of log of values.}$$

Note: $\left\{ \prod_{i=1}^{n} \left(\frac{x_i}{y_i} \right) \right\}^{1/n} = \frac{\left\{ \prod_{i=1}^{n} x_i \right\}^{1/n}}{\left\{ \prod_{i=1}^{n} y_i \right\}^{1/n}}$ → Average price Relative / Ratio

useful here, the g.m. of ratio is ratio of g.m.s.

Also, say,

$t_1 = 0$, $y = a$

$t_2 = t$, $y = ar^t$

To find value at mid point, we calculate G.M.

$$ar^{t/2} = \sqrt{a \cdot ar^t}$$

(*) If we have any $x_i = 0$ or $x_i < 0$, we hit a problem with GM.

Do we really? We may simply shift entire dataset by fixed value to ensure $x_i \neq 0$ and $x_i \neq 0 \ \forall i$, calculate GM, then shift back.

17th January 2025 (Friday) →

Harmonic Mean →

$$\bar{X}_n = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

→ AM of the reciprocals is the reciprocal of HM.

→ x units per y units (??)

Equal share of x units → HM
Equal share of y units → AM

① Weighed Data → (x_i, ω_i) → Treating the weights as frequencies of the data.

→ AM: $\bar{X}_A = \frac{\sum_i \omega_i x_i}{\sum_i \omega_i}$

→ GM: $\bar{X}_G = \left[\prod_i x_i^{\omega_i}\right]^{1/\sum_i \omega_i}$

→ HM: $\bar{X}_H = \frac{\sum_i \omega_i}{\sum_i \frac{\omega_i}{x_i}}$

→ Generalized for weighed data.