Quiz 1 (January 30, 2019 (Wednesday)).

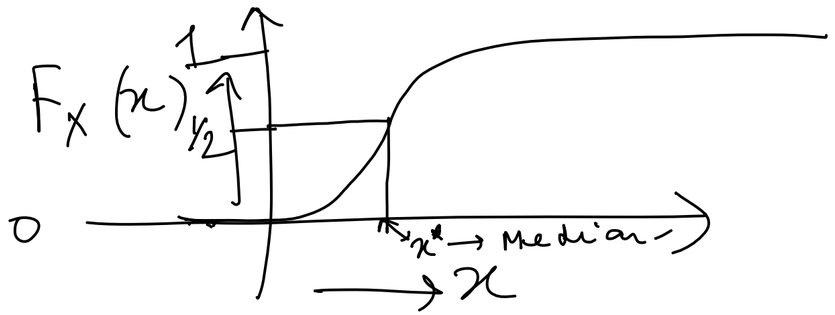
Time :- 6.35 p.m. to 7.15 p.m. Venue : L18, L19, L20.

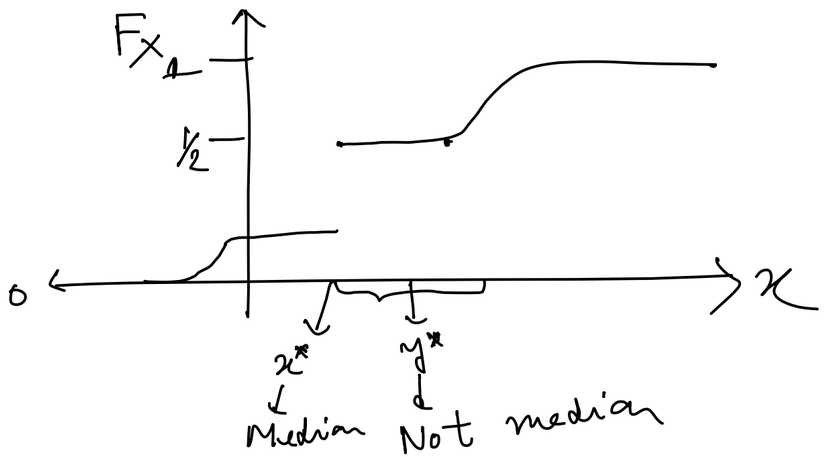
X : Random variable.

F_X : CDF of X.

$$\text{Med}_F = \inf_x \left\{ x : F(x) \geq \frac{1}{2} \right\} \longrightarrow \text{General defn.}$$

If F is cont^{n}, we then have $\text{Med}_F = F^{-1}\left(\frac{1}{2}\right)$.





Advantage of using median:-

$X = \{x_1, \ldots, x_n\}$ → data

Want to compute median (X).

↓ Middle most observation.

Suppose $X^* = \{x_{(1)}, \ldots, x_{(n)}\}$

↓ The smallest ↓ → Ordered data

$x_{(i)}$ → i-th smallest, $i = 1, \ldots, n$.

$$\text{Med}(X) = x_{\left(\frac{n+1}{2}\right)} \quad \text{if } n \text{ is odd.}$$

$$= \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \quad \text{if } n \text{ is even.}$$

$A = \{3, 1, 2, 4\}$, $A^* = \{1, 2, 3, 4\}$ → ordered

$\text{Med}(A) = \frac{2+3}{2} = 2.5.$

Suppose, there is one contamination

$A_{con} = \{3, 2, 400, 1\}$, $A^*_{con} = \{1, 2, 3, 400\}$.

$$\text{Med}(A^*_{con}) = \frac{2+3}{2} = 2.5.$$

In other words, this numerical example indicates that median does not get contaminated by single obs.

However,

$$\text{Mean}(A) = \frac{1+2+3+4}{4} = 2.5$$

$$\text{but Mean}(A_{con}) = \frac{1+2+3+400}{4} = 101.5.$$

That mean mean is "highly" affected by a single contamination.

Comment:- Overall, median is more "Robust" against the influential observations.

Quantile:- X : Random variable

F : CDF of X

$$\text{Quantile}_F(X) = \inf_x \{x : F(x) \geq \alpha\}, \quad \alpha \in (0,1).$$

If $\alpha = \frac{1}{2}$, then Quantile = Median.

Some well-known distributions (discrete & cont^{n}):-

Discrete Distributions:-

1. Discrete uniform :-

Suppose $X = \{x_1, \ldots, x_n\}$.

X : Random variable

P.M.F :- $P[X = x_i] = \frac{1}{n} \quad \forall \; i = 1, \ldots, n.$

$P[\cdot]$ is a proper p.m.f. (Try to convince yourself).

$$E(X) = \sum_{i=1}^{n} x_i \times \left(\frac{1}{n}\right) = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{(average)}.$$

$$E(X^K) = \sum_{i=1}^{n} x_i^K \times \left(\frac{1}{n}\right) = \frac{1}{n} \sum_{i=1}^{n} x_i^K$$

$$M_X(t) = \sum_{i=1}^{n} e^{tx_i} \times \left(\frac{1}{n}\right) = \frac{1}{n} \sum_{i=1}^{n} e^{tx_i}$$

$M_X(t)$ or MGF

$$Var(X) = E(X^2) - \{E(X)\}^2$$ (check $Var(X) = E[X - E(X)]^2$)

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

Check that $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

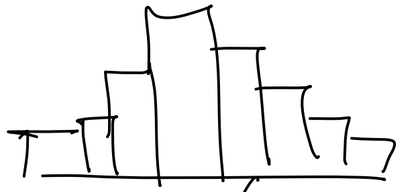
Remark:- Under Discrete uniform distribution, we are having the sample version of different measure in descriptive statistics.

2. **Degenerate dist^{n}:-**

X : Random variable.

$P[X = c] = 1$ and $P[X = x] = 0 \ \forall \ x \neq c$.

(c: some point)



Histogram of the data obtained from N(0,1)

$$E(X) = cx1 = c$$

$$E(X^2) = c^2 x 1 = c^2$$

$$M_X(t) = e^{tc} x 1 = e^{te}.$$

$$Var(X) = E(X^2) - \{E(X)\}^2 = c^2 - c^2 = 0$$

$$F_X(x) = P[X \leq x] = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases}$$

3. Two points distribution (e.g. Bernoulli dist^{n})

mass points

X :	x_1	x_2
$P[X = x]$:	p	$p_2 = (1-p)$

x_1 and x_2

$E(X) = x_1 \times (p) + x_2 (1-p)$ ——— ①

$E(X^2) = x_1^2 \times (p) + x_2^2 (1-p)$ ——— ②

$Var(X) = E(X^2) - \{E(X)\}^2 = ② - ①^2$ check $= p(1-p)(x_1 - x_2)^2$

$M_X(t) = E(e^{tX}) = e^{tx_1} \times (p) + e^{tx_2} \times (1-p)$.

Remark:- $E(X) = \left.\frac{d M_X(t)}{dt}\right|_{t=0} = x_1 p + x_2 (1-p)$

↓

This is another way of deriving $E(X)$

or any other moments.

Note:- If $x_1 = 0$ & $x_2 = 1$, it is called Bernoulli dist^n.

$P[X = x] = p^x (1-p)^{1-x}$, $x = 0$ or 1, $p \in (0,1)$.

Binomial dist^{n}:

Formulation:- n indep. Bernoulli trials with success prob $= p$.

X: No. of succes in n trials.

$\mathcal{X}$: Range of X

$$\mathcal{X} = \{0, 1, \ldots, n\}.$$

p-m.f. $\swarrow$

$$p(x) = P[X = x] = \begin{cases} 0 & \text{if } x \notin \mathcal{X} \\ \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in \mathcal{X}. \end{cases}$$

$X \sim \text{Bin}(n, p)$

$\mathcal{X} = \{0, \ldots, n\}$.

$$P[X = x] = \begin{cases} 0 & \text{if } x \notin \mathcal{X} \\ \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in \mathcal{X} \end{cases}$$

Note $P[X = x] \geq 0 \quad \forall \; x = 0, \ldots n$

And, $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (1 + 1 - p)^n = 1$.

Hence, $P[X = x]$ is a p.m.f.