2/12 LECTURE 19 : $\mathbb{E}$, VAR & SUMMARY OF DATA.

X r.v. with pmf p_X. $\mu_X = \mathbb{E}[X] = \sum_x x p_X(x)$ $(\mu = \mu_X)$

$\mathbb{E}[f(X)] = \sum_x f(x) p_X(x)$ $f : \mathbb{R} \to \mathbb{R}$

[Expectation/mean]

$\mathrm{VAR}[X] := \mathbb{E}[(X-\mu)^2]$ (Variance of X).

$= \mathbb{E}[X^2 - 2\mu X + \mu^2]$ $\mathrm{VAR}[X] \geq 0$

$= \mathbb{E}[X^2] - \mu^2$ (linearity.)

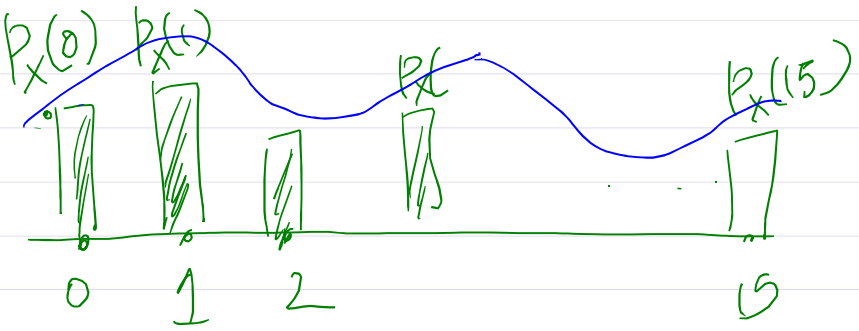
Alternative $\mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 p_X(x)$ (Take $f(x) = (x-\mu)^2$)

$= \sum_x x^2 p_X(x) - \mu^2 = \mathbb{E}[X^2] - \mu^2$.

Eg 19.1 Assume X r.v. with pmf $p_X(x) = 0$ if $x \notin \{0, \dots, 15\}$.

You are given $p_X(0), p_X(1), \dots, p_X(15)$.

Pictorially p_X is an histogram.



Can you summarise the histogram / r.v. ?

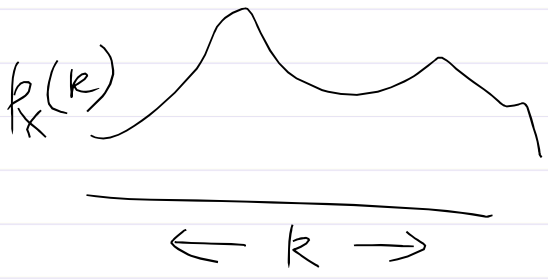
First summary is mean — $\mu_X = \mathbb{E}[X] = \sum_x x p_X(x)$ ← computable from histogram

Eg 19.2 Suppose I have scores of students in Quiz 1.

Define $p_X(k) = \frac{\#\text{ of scores} = k}{\#\text{ of students}}$. $\sum_{k=0}^{15} p_X(k) = 1$, $p_X(x) \geq 0$.

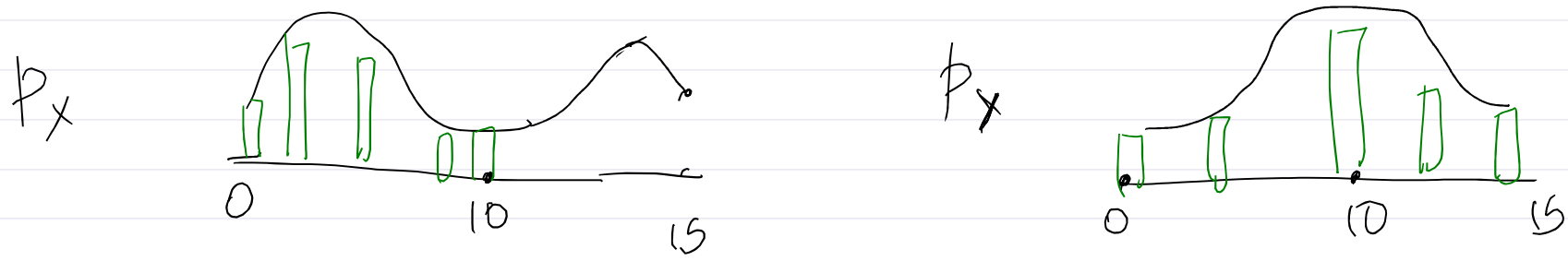
$\Rightarrow p_X$ is a pmf of some r.v. X.

(mean) $\mu_X = \sum_x x p_X(x)$ $(= 9.6\dots)$



Qn*: Find explicitly (Ω, P) & $X : \Omega \to \mathbb{R}$ s.t. $P(X=k) = p_X(k)$.

Qn: Suppose I have two pmfs p_X & p_Y



Assume $\sum_x x p_X(x) = \mu_X = \mu_Y = \sum_y y p_Y(y)$.

$VAR[X] = \sum_x (x-\mu)^2 p_X(x)$ $\Rightarrow$ larger variance $\Rightarrow$ larger "disparity".

LEMMA 19.3 (1) $VAR[X] = 0$ iff $p_X(\mu) = 1$ (i.e., $X \equiv \mu$)

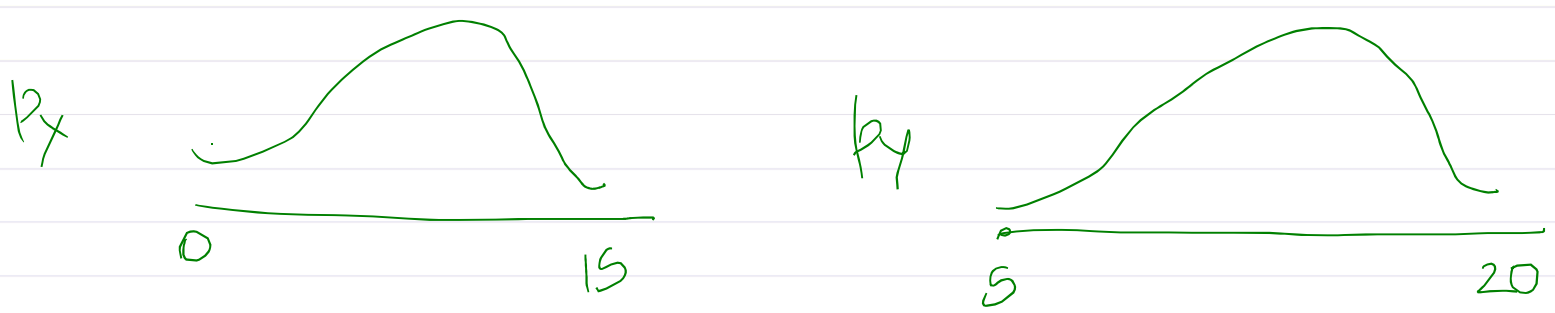
(2) $VAR[aX+b] = a^2 VAR[X]$ $\forall\ a, b \in \mathbb{R}$.

(3) $\mu_{aX+b} = a\mu_X + b$. (recall).

[why not $\sum_x |x-\mu| p_X(x)$??]

Eg 19.4 (Again Eg. of Quiz scores). Suppose I add 5 points to all.

Then the new r.v. $Y = X+5$. $p_Y(x) = p_X(x-5)$, $x \in \mathbb{R}$



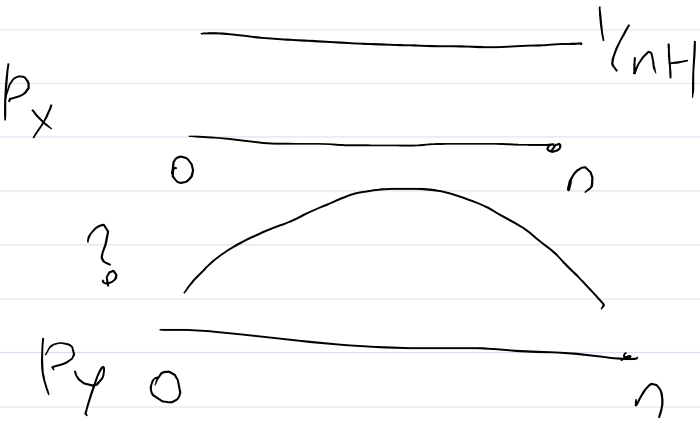
$\mu_Y = \mu_X + 5$, $VAR[Y] = VAR[X]$.

Eg 19.5 (A) $X \overset{d}{=} Unif\{0, 1, 2, \ldots, n\}$. $\mathbb{E}[X] = \frac{n}{2}$

$Y \overset{d}{=} Bin(n, 1/2)$ $\mathbb{E}[Y] = \frac{n}{2}$

Compute $VAR(X)$, $VAR(Y)$.

[& what can you say about disparity ?]



Further summaries — $\mathbb{E}[X]$, $\mathrm{VAR}[X]$, $\mathbb{E}[X^k]$, $k \geq 1$.

For eg. If $X \overset{d}{=} \mathrm{Ber}(p)$ then $X^k \overset{d}{=} \mathrm{Ber}(p)$ $\forall\, k \geq 1$.

$[P_X(1) = p = 1 - P_X(0)]$ $[$i.e., $P_{X^k}(1) = p = 1 - P_{X^k}(0)]$

$\Rightarrow \mathbb{E}[X^k] = p \quad \forall\, k \geq 1$.

Ex 19.7 Let X be a r.v. $f : \mathbb{R} \to \mathbb{R}$ & define $Y = f(X)$.

Then Y is a r.v. with pmf $p_Y(\cdot) = ??$ (write it in terms of p_X).

Ex 19.8 **Sampling with replacement.** Total population $= N$.

There are two types. # of type 1 $= Np \in \mathbb{N}$, $p \in [0,1]$

# of type 2 $= N(1-p) \in \mathbb{N}$.

Suppose you choose a sample of size n from the population at random & with replacement.

Let X = # of samples of type 1 in the 'n' randomly chosen samples.

$$p_X(k) = P(X = k) \underset{(\text{check})}{=} \binom{n}{k} \left(\frac{Np}{N}\right)^k \left(\frac{N(1-p)}{N}\right)^{n-k}$$

$$= \binom{n}{k} p^k (1-p)^{n-k}.$$

So X is a $\mathrm{Bin}(n, p)$ r.v.

Qn. How to find p?

Define $Y = \frac{X}{n}$. $\quad p_Y\left(\frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$ $\quad k = 0, \ldots, n$.

$$\mathbb{E}[Y] = \underset{(\text{linearity})}{\frac{\mathbb{E}[X]}{n}} = p. \quad (\text{mean of } Y)$$

$$\mathrm{VAR}[Y] = \underset{(\text{by LEMMA 19.6})}{\frac{\mathrm{VAR}[X]}{n^2}} \overset{(A)}{=} \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\mathrm{VAR}[Y] = \sum_x (y-p)^2 P_Y(y) = \frac{p(1-p)}{n} \le \frac{1}{4n}$$

$(p(1-p) \le \frac{1}{4})$

$$\mathrm{VAR}[Y] \le \frac{1}{4n}$$

If $n = 2500$, $\mathrm{VAR}[Y] \le 10^{-4}$

[Intuitively, "$|Y-p| \le 10^{-4}$";
if p is 10^{-6} then this is bad.; if "p is small/large", $\frac{p(1-p)}{n}$ is very small]

Ex 19.9 (A) Repeat the above exercise without replacement.

Find P_Y, $\mathbb{E}[Y]$ & $\mathrm{VAR}[Y]$.

$\mu_Y = \mathbb{E}[Y]$; $\sigma_Y = \mathrm{SD}[Y] := \sqrt{\mathrm{VAR}[Y]}$.

LEMMA 19.10 (Chebyshev's Ineq.) X r.v. $\mathbb{P}(|X-\mu| \ge t) \le \frac{\sigma^2}{t^2}$ $\forall\, t>0$.

Proof: $|X-\mu| \ge t \Rightarrow (X-\mu)^2 \ge t^2$ $\forall\, t>0$.

So $\{|X-\mu| \ge t\} \subseteq \{(X-\mu)^2 \ge t^2\}$

$\Rightarrow \mathbb{P}(|X-\mu| \ge t) \le \mathbb{P}((X-\mu)^2 \ge t^2)$

$$\overset{(M.\ \text{Ineq})}{\le} \frac{\mathbb{E}[(X-\mu)^2]}{t^2} = \frac{\sigma_X^2}{t^2}. \quad \square$$

($(X-\mu)^2 \ge 0$)

So Chebyshev's ineq. $\Rightarrow$

$$\mathbb{P}(|X-\mu| \ge k\sigma) \le \frac{1}{k^2} \quad \forall\, k \ge 1$$

Apply to Eg. of sampling with replacement.

$Y = \frac{X}{n}$; $\mu_Y = p$; $\sigma_Y = \sqrt{\frac{p(1-p)}{n}}$

$$P(|Y-p| \geq k\sigma_Y) \leq \frac{1}{k^2}$$

"Prob. of Error in $|Y-p|$ is larger than $\frac{k\sqrt{p(1-p)}}{\sqrt{n}}$ is atmost $\frac{1}{k^2}$"

If $k=10$, then $|Y-p| \leq \frac{k\sqrt{p(1-p)}}{\sqrt{n}}$ with 99% accuracy.

Ex: Do this for sampling without replacement.

Find t s.t. $P(|Y-a| \geq t) \leq \frac{1}{100}$??