BSCMA1004

STATISTICS II NOTES

WEEK 3 NOTES

IITM B.S Degree

PREPARED BY

Vehaan Handa, IIT Madras

Stats II – Week III Notes

DATE

★ Expected Value of RV

Suppose X is a drv with range T_X and pmf f_X. Exp. value $E[X]$:

$$E[X] = \sum_{t \in T_X} t f_X(t) = \sum_{t \in T_X} t P(X=t)$$

- mean, average value
- $E[X]$ may/may not be in range of X
- Same units as X.

1. $X \sim \text{Bernoulli}(p)$

$$E[X] = 0(1-p) + 1(p) = p$$

2. $X \sim \text{Unif}\{1,2,3,4,5,6\}$

$$E[X] = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

3. Val. of lottery ticket is $\{200, 20, 0\}$ (1/1000, 27/1000, 972/1000)

$$E[X] = 200.\frac{1}{1000} + 20.\frac{27}{1000} + 0.\frac{972}{1000} = 0.56$$

4. $X \sim \text{Unif}\{a, a+1, \ldots, b\}$.

$$E[X] = a.\frac{1}{b-a+1} + (a+1).\frac{1}{b-a+1} + \ldots + b.\frac{1}{b-a+1}$$

Identity: $a + (a+1) + \ldots + b = (b-a+1)\frac{a+b}{2}$

$$E[X] = \frac{a+b}{2}$$

1. $X \sim \text{Geometric}(p)$

$$E[X] = \sum_{t=1}^{\infty} t(1-p)^{t-1} p$$
$$= p + 2(1-p)p + 3(1-p)^2 p + \dots$$

2. $X \sim \text{Poisson}(\lambda)$

$$E[X] = \sum_{t=0}^{\infty} t e^{-\lambda} \frac{\lambda^t}{t!}$$

3. $X \sim \text{Binomial}(n, p)$

$$E[X] = \sum_{t=0}^{n} t \binom{n}{t} p^t (1-p)^{n-t}$$

★ How to simplify sum?

1. Difference Equation (DE): $a_{t+1} - r a_t = b_t \ (r \neq 1)$ → another sequence

$$\sum_{t=1}^{n} a_t = \frac{a_1 - r a_n}{1-r} + \frac{1}{1-r} \sum_{t=1}^{n-1} b_t$$

2. Geometric Progression (GP): $a_{t+1} - r a_t = 0 \ (r \neq 1)$

$$\sum_{t=1}^{n} a_t = \frac{a_1 - r a_n}{1-r} \xrightarrow[n \to \infty]{|r|<1} \frac{a_1}{1-r}$$

3. Exponential Function:

$$\sum_{t=0}^{\infty} e^{-\lambda} \frac{\lambda^t}{t!} = 1$$

DATE

4. Binomial : $\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n$

So, expected value of:

1. $X \sim Geom(p)$ [DE : $a_i = p$, $r = 1-p$ $b_t = r^t p$]

$a_{t+1} = (t+1)(1-p)^t p$

$a_{t+1} - (1-p)a_t = (1-p)^t p$

$$\boxed{E[X] = \frac{1}{p}}$$

2. $X \sim Poisson(\lambda)$ (Using ③)

$$\boxed{E[X] = \lambda}$$

3. $X \sim Binomial(n,p)$

$$\boxed{E[X] = np}$$

★ Properties of E[X]

Constant RV and Positive RV

$$E[X] = \sum_{t \in T_X} t P(X=t)$$

1. Consider constant c as RV X with $P(X=c) = 1$

$$E[c] = c$$

2. Suppose X takes only non negative values. Then

$$E[X] \geq 0$$

DATE

★ Expected value of function of RV's:

Suppose $X_1, \ldots, X_n$ have joint PMF $f_{X_1 \ldots X_n}$ with range of X_i denoted T_{X_1}. Let $g: T_{X_1} \times T_{X_2} \times \ldots \times T_{X_n} \to \mathbb{R}$ be a function, let $Y = g(X_1, X_2 \ldots, X_n)$ have range T_Y, PMF f_Y. Then:→

$$E[g(X_1 \ldots, X_n)] = \sum_{t \in T_Y} t f_Y(t) = \sum_{t_i \in T_{X_1}} g(t_1, \ldots, t_n) \cdot f_{X_1 \ldots X_n}(t_1, \ldots, t_n)$$

- To find $E[Y]$, you don't always need f_Y. Joint pmf of $X_1, \ldots, X_n$ can be used directly.

Eg. $X \sim \{-2, -1, 0, 1, 2\}$ (1/5 1/5 1/5 1/5 1/5) $\quad g(X) = X^2 \sim \{0, 1, 4\}$ (1/5 2/5 2/5)

$$E[g(X)] = 0 \cdot \frac{1}{5} + 1 \cdot \frac{2}{5} + 4 \cdot \frac{2}{5} = \underline{\underline{2}}$$

$$E[g(X)] = (-2)^2 \cdot \frac{1}{5} + (-1)^2 \cdot \frac{1}{5} + (0)^2 \cdot \frac{1}{5} + (1)^2 \cdot \frac{1}{5} + (2)^2 \cdot \frac{1}{5} = \underline{\underline{2}}$$

★ Linearity of Expected Value

① $E[cX] = cE[X]$ for RV X and constt. c.

② $E[X+Y] = E[X] + E[Y]$ for any two RV's X and Y.

■ $$E[aX + bY] = aE[X] + bE[Y]$$

- $X \sim f_X$, $Y \sim f_Y$, joint pmf f_{XY} is not given. Can we compute $E[X+Y]$? Yes.

$$E[X+Y] = E[X] + E[Y] = \sum_{t \in T_X} t f_X(t) + \sum_{t \in T_Y} t f_Y(t)$$

Note: Exp. value of $E[g(X)+h(Y)]$ can be computed with marginal PMF and does not depend on joint PMF.

Eg. Expected value of Binomial (n,p).

$$E[Y] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np$$

- Difficult to simplify.
- Alternative: Linearity of Exp. value.

Let $X_1, \ldots, X_n$ be iid Bernoulli (p). Then $E[X_i] = p$ and $Y = X_1 + X_2 + \ldots + X_n \sim$ Binomial (n,p)

$$E[Y] = E[X_1] + \ldots + E[X_n] = np.$$

★ Zero Mean RV's

A RV X with $E[X] = 0$ is said to be a zero mean RV.

Translation of RV: $X+c$, $c \in$ Constt, is "translated" version of X.

- Range of $X+c$ is $\{t+c : t \in T_X\}$, translated vn of T_X.
- $P(X+c = t+c) = P(X=t)$. PMF translated too.

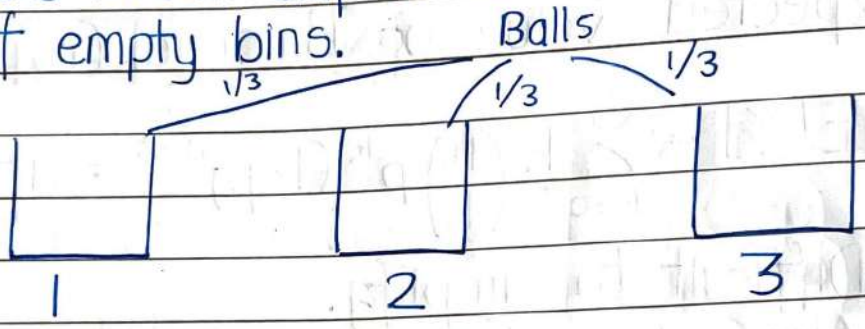
DATE

IMP! $E[X]$ is a constant.

$Y = X - E[X]$ is a translated version of X and $E[Y] = 0$.

So $\underbrace{X - E[X]}_{\text{"centering"}}$ is zero mean RV.

Eg. 10 balls are thrown unif. at random in 3 bins.
Exp. no. of empty bins.



$$X_i = \begin{cases} 1, & \text{if bin } i \text{ is empty} \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, 2, 3$$

$$P(X_i = 1) = \frac{2^{10}}{3^{10}} = \left(\frac{2}{3}\right)^{10} \qquad P(X_i = 0) = 1 - \left(\frac{2}{3}\right)^{10}$$

Y = empty bins $= X_1 + X_2 + X_3$.

$$E(Y) = E(X_1) + E(X_2) + E(X_3) = 3 \cdot \left(\frac{2}{3}\right)^{10}$$

DATE

★ Variance and Standard Deviation:

Variance of rv X, denoted $Var(X)$ is:→

$$Var(X) = E[(X - E(X))^2]$$

Standard Deviation X, $SD(X) = +\sqrt{Var(X)}$

- $Var(X) = \sum_{t \in T_X} (t - E[X])^2 P(X = t)$
- Variance non negative, standard deviation well defined.
- Units of $SD(X)$ are same as units of X.

Eg. $X \sim Unif\{1,2,3,4,5,6\}$ $\quad E[X] = 3.5$

$$Var(X) = (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} + (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6}$$

$$= \frac{35}{12} = 2.916$$

$$SD(X) = 1.7078.$$

★ Properties

1. $Var(aX) = a^2 Var(X)$
2. $SD(aX) = |a| SD(X)$
3. $Var(X+a) = Var(X)$
4. $SD(X+a) = SD(X)$

★ Other formula for variance

1. $Var(X) = E[X^2] - (E(X))^2$

$E(X) \rightarrow$ first moment
$E(X^2) \rightarrow$ second moment
$Var(X) \rightarrow$ second central moment

DATE

If X and Y are independent rv's

1. $E[XY] = E[X]E[Y]$
2. $Var(X+Y) = Var(X) + Var(Y) = Var(X-Y)$

Eg. Variance of sum of two dice?

$$Var(X+Y) = Var(X) + Var(Y)$$
$$= 2\sqrt{\frac{35}{12}} = \sqrt{\frac{35}{3}}$$

Distribution	Exp. Value	Variance
Bernoulli(p)	p	$p(1-p)$
Binomial (n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Unif$\{1,2,\ldots,n\}$	$(n+1)/2$	$(n^2-1)/12$

★ Standardised Random Variables

A rv X is said to be standardised if: →

$$E[X] = 0, \quad Var[X] = 1$$

Let X be a rv. Then $Y = \frac{X - E[X]}{SD(X)}$ is a standardised random variable.

DATE

- There are rv's s.t $E[X]$ goes to ∞ or $-\infty$
  - $X \sim \{1, 2, 4, \ldots, 2^{n-1}, \ldots\}$
  - $E[X^2]$ will go to ∞ and $Var(X)$ not defined. (ill defined)
- There are rv's s.t $E[X]$ is not well defined.
  - $X \sim \{1, -2, 4, \ldots, (-2)^{n-1}, \ldots\}$
  - $E(X^2)$ will go to ∞ in this case and $Var(X)$ ill defined.
- Sometimes $E[X]$ is finite, but $E[X^2]$ goes to ∞.
  - $Var(X)$ goes to ∞.

★ Covariance and Correlation

Consider joint PMFS:

f_{XY}	$X=0$	$X=1$	f_Y
$Y=0$	1/4	1/4	1/2
$Y=1$	1/4	1/4	1/2
f_X	1/2	1/2	

f_{XY}	0	1	f_Y
0	0	1/2	1/2
1	1/2	0	1/2
f_X	1/3	1/2	

Same marginal PMFs in both cases. Mean, variance are identical. But:→

- X and Y are independent in one.
- Val. of X determines val. of Y in other.

Covariance: Suppose X and Y are rv's on same prob. space. Covariance of X and Y, Cov(X,Y) is:→.

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$

DATE

- $Cov(X,Y)$ is positive :→
  - $(X-E[X])(Y-E[Y])$ tends to be +ve.
  - When X is above/below avg, Y tends to be correspondingly above/below its avg.

- $Cov(X,Y)$ is negative :→
  - $(X-E[X])(Y-E[Y])$ tends to be -ve
  - When X is above/below avg, Y tends to be correspondingly below/above its avg.

- $Cov(X,Y) = 0$.
  - X and Y are uncorrelated.

Eg. 1. X : height (person) Y : weight (person)
- Higher X tends to result in higher Y.
- Expect $Cov(X,Y)$ to be +ve

2. X : Monsoon rainfall Y : Farmer Debt
- Higher X tends to result in lower Y.
- Expect $Cov(X,Y)$ to be -ve.

	X = -1	X = 0	X = 1	f_Y	
Y = -1	1/15	2/15	2/15	1/3	E(Y) = 0
Y = 0	2/15	1/15	2/15	1/3	E(X) = 0
Y = 1	2/15	2/15	1/15	1/3	
f_X	1/3	1/3	1/3		

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = 1.\frac{1}{15} + (-1).\frac{2}{15} + (-1).\frac{2}{15} + (1).\frac{1}{15}$$

$$= \frac{-2}{15}$$

classmate

PAGE

DATE

Properties

1. $Cov(X,X) = Var(X)$
2. $Cov(X,Y) = E[XY] - E[X]E[Y]$
3. Covariance is symmetric, $Cov(X,Y) = Cov(Y,X)$
4. Covariance is a linear quantity.
   (i) $Cov(X, aY+bZ) = a\,Cov(X,Y) + b\,Cov(X,Z)$
   (ii) $Cov(aX+bY, Z) = a\,Cov(X,Z) + b\,Cov(Y,Z)$
5. If X and Y are independent, they are uncorrelated. i.e $Cov(X,Y) = 0$. If X and Y are uncorrelated, may be dependent.

☆ Correlation Coefficient

Correlation Coefficient / correlation of two RV's X and Y is defined :→.

$$\boxed{\rho(X,Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)}} \qquad -1 \leq \rho \leq 1$$

$$-SD(X)\,SD(Y) \leq Cov(X,Y) \leq SD(X)\,SD(Y)$$

Simplify $E\left[\left(\frac{X-E(X)}{SD(X)} + \frac{Y-E(Y)}{SD(Y)}\right)^2\right] \geq 0$ (lower bound)

$E\left[\left(\frac{X-E(X)}{SD(X)} - \frac{Y-E(Y)}{SD(Y)}\right)^2\right] \geq 0$ (upper bound)

If $\rho(X,Y) = 1$ or -1,

There exists $a \neq 0$ and b so that $Y = aX + b$ with probability 1. Y is a linear fn of X.

$\mu \rightarrow$ Mean $E[X]$
$\sigma^2 \rightarrow$ Variance $Var(X)$
$\sigma \rightarrow$ Standard Deviation $SD(X)$

For multiple RV's :→
- $\mu_X \rightarrow$ mean $E[X]$
- $\sigma_X^2 \rightarrow$ variance $Var(X)$
- $\sigma_X \rightarrow$ standard deviation $SD(X)$

Suppose avg marks = 50/100. What fraction of students will have marks ≥ 50?
- Cannot be 0.

What fraction has marks ≥ 80?
- Cannot be 1
- Cannot be 0.9 (★)

So we can estimate bounds.

★ Standard units in statistics

Consider rv X with mean μ and variance σ^2. In an experiment, X may take a value close to or away from μ.

$X - \mu$ measures dist. of X from mean μ. Could be positive/negative.

DATE

Standard Units: The number of standard deviations that a realisation of a random variable is away from the mean.

We expect $X-\mu$ to fall b/w $-c\sigma$ and $c\sigma$ for small c.
i.e X to fall between $\mu - c\sigma$ and $\mu + c\sigma$

Eg. Throw pair of die. X = sum of two numbers

- $\mu = 7 \quad \sigma \approx 2.42$
- $P(|X-\mu| \leq \sigma) = P(4.58 \leq X \leq 9.42) = P(X \in \{5,6,7,8,9\}) = 2/3$
- So $P(|X-\mu| > \sigma) = 1/3$
- $P(|X-\mu| > 2\sigma) = P(X \in \{2,12\}) = 2/36 \approx 0.056$

Eg. $X \sim \text{Unif}\{1,2,\ldots,100\}$

- $\mu = 50.5, \quad \sigma \approx 28.9$
- $P(|X-\mu| > \sigma) = 1 - 58/100 = 42/100$
- $P(|X-\mu| > 2\sigma) = 0$

★ Markov's Inequality

Let X be a drv taking non negative values with finite mean μ. Then :→

$$\boxed{P(X \geq c) \leq \frac{\mu}{c}}$$

Proof:

1. $$\mu = \sum_{t \in T_X} t P(X=t) = \sum_{t<c} t P(X=t) + \sum_{t \geq c} t P(X=t)$$

Since first sum is non negative,

$$\mu \geq \sum_{t \geq c} t P(X=t) \geq \sum_{t \geq c} c P(X=t) = c P(X \geq c)$$

DATE

★ Chebyshev's Inequality

Let X be a drv with a finite mean μ and finite variance σ^2. Then:→

$$\boxed{P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}}$$

Proof: Apply MI to $(X-\mu)^2$

Other forms:

- $P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$, $P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{1}{k^2}$
- $P(\mu-k\sigma < X < \mu+k\sigma) \geq 1-\frac{1}{k^2}$
- $P(X \geq \mu+k\sigma) + P(X \leq \mu-k\sigma) \leq \frac{1}{k^2}$
- $P(X \geq \mu+k\sigma) \leq \frac{1}{k^2}$, $P(X \leq \mu-k\sigma) \leq \frac{1}{k^2}$

★ Actual vs Chebyshev

Let X be a rv with finite mean μ and variance σ^2

$$P(|X-\mu| \geq 2\sigma) \leq \frac{1}{4}$$

① $X \sim \text{Binomial}(10, 1/2)$, $\mu = 5$, $\sigma = \sqrt{2.5} = 1.58$.

$$P(|X-5| \geq 2\sigma) = P(X \in \{0,1,9,10\}) \approx 0.021$$

$(\leq 1/4)$

classmate

PAGE

DATE

② $X \sim \text{Geometric}(1/4)$, $\mu = 4$ $\sigma \approx 3.46$

$P(|X-4| \geq 2\sigma) = P(X \in \{11, 12, \ldots\}) \approx 0.056$

- Mean μ, through Markov's Inequality, bounds the probability that a non negative random variable takes values much larger than mean.

- Mean μ and SD σ, through Chebyshev's inequality, bound the probability that X is away from μ by $k\sigma$.

Eg. Suppose [redacted] no. of accidents decreases by 10000/day across the country. Is it significant? If SD of no. of accidents is known, we can find how high 10000 is in terms of SD's to answer this.