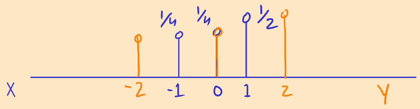
Covariance



$Var(X) < Var(Y)$

$Cov(X,Y)$ how related are X and Y

$$Cov(X,Y) = \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big]$$

$$Var(X) = Cov(X,X)$$

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

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$-1 \le Corr(X,Y) \le 1$$

$$= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$X \perp\!\!\!\perp Y$

$Pr[X=1 \mid Y=2]$

$= Pr[X=1]$

$Cov(X,Y) = 0$

	-2	0	2	
-1	1/16	1/16	1/8	1/4
0	1/16	1/16	1/8	1/4
1	1/8	1/8	1/4	1/2
	1/4	1/4	1/2	

$Pr[X=1 \mid Y=2] = 1$

$Cov(X,Y) =$ high

	-2	0	2	
2	1/4	0	0	1/4
0	0	1/4	0	1/4
-2	0	0	1/2	1/2
	1/4	1/4	1/2	

$Cov(X,Y) =$ med

	-2	0	2	
-1	1/8	1/8	0	1/4
0	1/8	1/8	0	1/4
1	0	0	1/2	1/2
	1/4	1/4	1/2	

10 Variance, Covariance, Concentration Inequalities, and WLLN

Computer Science Mentors 70

Apr 12 - Apr 16
Spring 2021

Amogh Gupta, Sylvia Jin, Aekus Bhathal, Abinav Routhu, Debayan Bandyopadhyay

Roast us here: `https://tinyurl.com/csm70-feedback20`

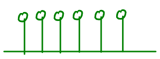
1 Variance

1. **Introduction to Variance**

We are making paper boats out of square sheets of paper. Suppose the size of the sail of the boat is same as the length of the side of the paper. We plan to go to "The Random Sheet Shop" to buy the sheet. This shop have 6 sheets of varying side lengths.

(a) The shop has sheets of side length 1 unit to 6 units. Find the mean and variance for the size of the sail of the boat we could make out of these sheets if the shop gives us a sheet uniformly at random.

X be the length of a sail.



$\mathbb{E}[X] = \frac{1}{6}\cdot(1+2+\cdots+6) = \frac{21}{6} = \boxed{\frac{7}{2}}$

X	X^2
1	1
2	4
3	9
4	16
5	25
6	36

2 3 1 5
5 5 2 3 6
2 3 2

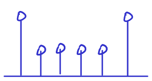
$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$= \frac{1}{6}(1+4+9+16+25+36) - \frac{49}{4} = \frac{91}{6} - \frac{49}{4} = \boxed{\frac{35}{12}}$

(b) Now suppose we go to a shop which lists sizes a little differently and gives a sheets with a different probability distribution. The shop lists the sheet as difference from mean but never tells you the mean of side but rather tells the mean of the area of the sheet is $\frac{101}{6}$. Find the mean and variance for the size of the sail of the boat we could make out of these sheets if the shop gives us a sheet with distribution $(-\frac{5}{2} : \frac{1}{3}, -\frac{3}{2} : \frac{1}{12}, -\frac{1}{2} : \frac{1}{12}, +\frac{1}{2} : \frac{1}{12}, +\frac{3}{2} : \frac{1}{12}, +\frac{5}{2} : \frac{1}{3})$

$\mathbb{E}[X^2]$

25/4 9/4 1/4 1/4 9/4 25/4



2 3 1 1
1 1 6 1 6 6
6 3 5 6

$X - \mathbb{E}[X]$

$\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] = \cdots = \frac{55}{12}$

$= \frac{25}{4}\cdot\frac{1}{3} + \frac{9}{4}\cdot\frac{1}{12} + \cdots =$

$\mathbb{E}[X]$

	X	$\Pr[(X-\mathbb{E}[X])]$
-2.5	1	1/3
-1.5	2	1/12
-0.5	3	1/12
0.5	4	1/12
1.5	5	1/12
2.5	6	1/3

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

$\Rightarrow \mathbb{E}[X] = \sqrt{\mathbb{E}[X^2] - \mathrm{Var}(X)}$

$= \sqrt{\frac{101}{6} - \frac{55}{12}}$

$= \frac{7}{2}$

(c) Give an intuitive explanation for the difference between the variance and mean in part(a) and part (b)

2. **Chaotic Santa**

(Fall '17 Disc) A delivery guy is out delivering n packages to n customers, where $n \in \mathbb{N}$, $n > 1$. Not only does he hand a random package to each customer, he opens the package before delivering it with probability $\frac{1}{2}$.

1 307
2 88
⋮
n 3

(a) What is the expected number of customers that get their own package unopened?

Let total # of customers with own unopened packages be $X = \sum_{i=1}^{n} X_i$, where $X_i = \begin{cases} 1 & \text{person } i \text{ gets own unopened pack.} \\ 0 & \text{o/w} \end{cases}$

$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{2n} = \frac{n}{2n} = \boxed{\frac{1}{2}}$

$\mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3]$

$\mathbb{E}[X_i] = \Pr[X_i = 1]\cdot 1 + \cancel{\Pr[X_i = 0]\cdot 0}$

$= \frac{1}{n}\cdot\frac{1}{2} = \frac{1}{2n}$

$(X_1+X_2+X_3+X_4)^2$
$= X_1^2 + X_2^2 + X_3^2 + X_4^2$
$+ 2X_1X_2 + 2X_1X_3 + 2X_2X_3$
$+ 2X_1X_4 + 2X_2X_4 + 2X_3X_4$

X_iX_j is indicator

$\mathbb{E}[X_iX_j] = \Pr[X_iX_j = 1]$
$= \Pr[X_i = 1 \cap X_j = 1]$
$= \Pr[X_i = 1]\Pr[X_j = 1 \mid X_i = 1]$
$= \frac{1}{2n}\cdot\left(\frac{1}{2}\cdot\frac{1}{n-1}\right) = \frac{1}{4n(n-1)}$

(b) What is the variance for the random variable above?

$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{3}{4} - \frac{1}{4}$

$= \boxed{\frac{1}{2}}$

$\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{1\le i<j\le n} X_iX_j$

$\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = \sum \mathbb{E}[X_i^2] = \sum \mathbb{E}[X_i] = \frac{n}{2n} = \frac{1}{2}$

$\mathbb{E}[X^2] = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right]$

$= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] + \mathbb{E}\left[2\sum_{1\le i<j\le n} X_iX_j\right]$

$= \frac{1}{2} + 2\sum_{1\le i<j\le n}\mathbb{E}[X_iX_j] = \frac{1}{2} + 2\cdot\frac{1}{4n(n-1)}\cdot\binom{n}{2} = \frac{1}{2} + 2\cdot\frac{1}{4n(n-1)}\cdot\frac{n(n-1)}{2} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$

X_iX_j	$\Pr$
0	
1	

$X_j \backslash X_i$	0	1
0	0	0
1	0	1

X_i	X_i^2
0	0
1	1

3. **Jensen's for Special Polynomials**

Jensen's inequality says that for any convex function f, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ (You don't need to know this for CS 70). In this problem, we will prove that Jensen's inequality holds for a subclass of convex functions called "special polynomials" (this is a made up name). We define a special polynomial as any function that can be written as

$$f(x) = a_n x^{(2^{n-1})} + a_{n-1} x^{(2^{n-2})} + \ldots + a_1 x$$

for some $n \in \mathbb{N}$ and $\forall 1 \leq i \leq n$, $a_i \geq 0$

(a) Prove that $\mathbb{E}[X^2] \geq E[X]^2$ (*Hint: use the definition of variance*)

(b) Use part (a) to prove that $\mathbb{E}[X^{(2^k)}] \geq E[X]^{(2^k)}$ for some $k \in \mathbb{N}$

(c) Use part (b) and properties of expectation to prove that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$

2 Covariance and Correlation

1. **One D -> Two D**

   Recall from long, long ago, the binary truth operators OR, AND, XOR, and so forth. In this problem we analyze these operators as probability distributions.

   $c \times (x \oplus y)$

   (a) Suppose X, Y are discrete random variables taking on values in $\{0, 1, 2, 3\}$ and let $\mathbb{P}(X = x, Y = y) = c(x \oplus y)$. $\oplus$ is the XOR operator, and can be expressed in terms of elementary logic operators as $P \oplus Q = (P \vee Q) \wedge \neg(P \wedge Q)$.

   Since X and Y are not constrained to 0 or 1, in this case $\oplus$ is applied bitwise (e.g. $3 \oplus 2 = 11_2 \oplus 10_2 = 01_2 = 1$).

   $11 = 3$

   Find c, and express the joint distribution of X and Y with a probability table, and find the ~~marginal~~ distributions for X and Y.

   univariate

   marg. of Y

Y \ X	0	1	2	3	
0	0	c	2c	3c	1/4
1	c	0	3c	2c	1/4
2	2c	3c	0	c	1/4
3	3c	2c	c	0	1/4
	1/4	1/4	1/4	1/4	← marg. of X

   $\Pr[X = 1] \geq c > 0$ $\quad \Pr[X=1 \mid Y=1] = 0$ $\quad \therefore X \not\perp Y$

   $\sum_x \Pr[X=x \mid Y=0] = \frac{6c}{\square} = 1 \Rightarrow c \neq 0$

   $\Pr[X=1] = 1/4$

   $24c = 1 \Rightarrow c = \frac{1}{24}$

   $\Rightarrow 6c = \frac{1}{4}$

   (b) Using the same definitions from part (a), find the covariance and the correlation of X, Y [5cm]

   (c) Find $\mathbb{E}[X + Y]$ and $Var[X + Y]$

(b) $Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \frac{3}{2}\cdot\frac{3}{2} = \frac{3}{2} - \frac{9}{4} = \boxed{-\frac{3}{4}}$

$\mathbb{E}[XY] = \sum xy \Pr[X=x, Y=y] = \frac{3}{2}$

$Var(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \frac{5}{4} = Var(Y)$

$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-3/4}{5/4} = \boxed{\frac{-3}{5}}$

(c) $\mathbb{E}[X+Y] = \frac{3}{2} + \frac{3}{2} = 3$

$Var(X+Y) = Cov(X+Y, X+Y)$
$= Cov(X,X) + Cov(X,Y) + Cov(Y,X) + Cov(Y,Y)$
$= Var(X) + Var(Y) + 2Cov(X,Y)$
$= \frac{5}{4} + \frac{5}{4} - \frac{6}{4}$
$= \boxed{1}$

bilinearity of Cov means
1. $Cov(aX, bY) = ab\,Cov(X,Y)$
2. $Cov(X+Y, Z+W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W)$

3 Concentration Inequalities

1. Let X be the sum of 20 i.i.d. Poisson random variables $X_1, \ldots, X_{20}$ with $\mathrm{E}(X_i) = 1$. Find an upper bound of $\Pr[X \geq 26]$ using,

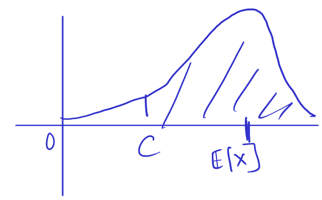
$X \geq 0$

$\mathbb{E}[X] < \infty$

$\Pr[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$

$\mathbb{E}[X] = 100$

$c = 1$



(a) Markov's inequality:

$\Pr[X \geq 26] \leq \frac{20}{26} \approx 0.769$

$\mathbb{E}[X] = \sum_{i=1}^{20} \mathbb{E}[X_i] = 20 \cdot 1 = 20$

(b) Chebyshev's inequality:

$\mathrm{Var}(X) = \sum_{i=1}^{20} \mathrm{Var}(X_i) = 20 \cdot 1 = 20$

$\Pr[X \geq 26] = \Pr[X - \underset{\mathbb{E}[X]}{20} \geq 6]$

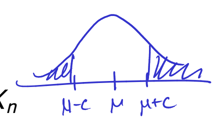
$\leq \Pr[|X - \mu| \geq 6] \leq \frac{20}{36} \approx 0.556$

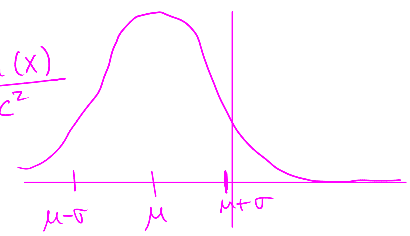
X

$\mathbb{E}[X] = \mu < \infty$

$\mathrm{Var}(X) = \sigma^2 < \infty$

$\Pr[|X - \mu| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$





2. Suppose we have a sequence of iid random variables $X_1, X_2, \ldots, X_n$
Let $A_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$ be the sample mean.
~~Show that the true mean of $X_i = \mu$ is within the interval $[\cancel{\mu}^{A_n} - 4.5\frac{\sigma}{\sqrt{n}}, \cancel{\mu}^{A_n} + 4.5\frac{\sigma}{\sqrt{n}}]$ with 95% probability.~~

Show that the interval $[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]$ contains μ with 95% prob.

$\mathbb{E}[A_n] = \frac{1}{n}\mathbb{E}\left[\sum^n X_i\right] = \mathbb{E}(X_i) = \mu$

$[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}] \ni \mu \iff |A_n - \mu| \leq 4.5\frac{\sigma}{\sqrt{n}}$

$\underline{\mathrm{Var}(X_i) = \sigma^2}$

$\mathrm{Var}(A_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$\Pr\left[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\mathrm{Var}(A_n)}{(4.5)^2\sigma^2/n} = \frac{\sigma^2/n}{(20.25)\sigma^2/n} = \frac{1}{20.25} \leq \frac{1}{20}$

$-\Pr\left[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}\right] \geq \frac{-1}{20}$

$\Pr\left([A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}] \ni \mu\right) = \Pr\left(|A_n - \mu| \leq 4.5\frac{\sigma}{\sqrt{n}}\right) = 1 - \Pr\left[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}\right] \geq 1 - \frac{1}{20} = 0.95$

4 Weak Law of Large Numbers

Introduction to LLN

Leanne has a weighted coin that shows up heads with probability $\frac{4}{5}$ and tails with probability $\frac{1}{5}$. Leanne flips the coin 100 times, and computes X, the average number of coins that show up heads.

(a) What is $E[X]$?

(b) What is Var(X)?

(c) Suppose Leanne flips n coins instead of 100. What does the LLN tell us about X?