Chapter 5. Joint Probability Distributions and Random Samples

Math 3670 Spring 2025

Georgia Institute of Technology

Section 1.
Jointly Distributed Random Variables

Two Discrete Random Variables

Definition

Let X and Y be two discrete RVs defined on the sample space $\mathcal{S}$ of an experiment.

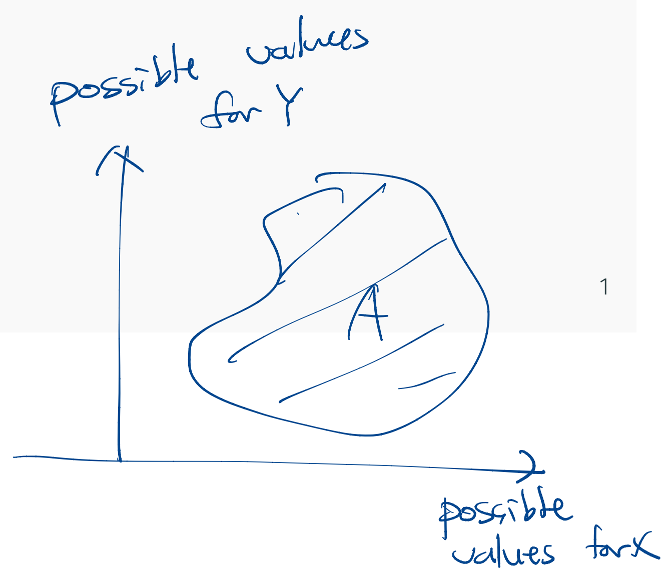
The joint probability mass function $p(x, y)$ is defined by

$$p(x, y) = \mathbb{P}(X = x, Y = y)$$

AND

The joint PMF satisfies

1. $p(x, y) \geq 0$
2. $\sum_{x,y} p(x, y) = 1$
3. $\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$



Two Discrete Random Variables

Example

A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are 0, \$100, and \$200.

Suppose an individual with both types of policy is selected at random from the agency's files. Let X be deductible amount on the auto policy and Y deductible amount on the homeowner's policy.

$X \backslash Y$	0	100	200
100	0.2	0.1	0.2
250	0.05	0.15	0.3

$$p(100,100) = \mathbb{P}(X=100,\ Y=100) = 0.1$$

$$p(250,0) = \mathbb{P}(X=250,\ Y=0) = 0.05.$$

$$p_X(100) = \mathbb{P}(X=100) = \mathbb{P}(X=100, Y=0) + \mathbb{P}(X=100, Y=100) + \mathbb{P}(X=100, Y=200)$$

$$= 0.2 + 0.1 + 0.2 = 0.5.$$

Two Discrete Random Variables

Definition

For a given joint PMF $p(x, y)$ of random variables X and Y, the marginal probability mass function of X is given by

Marginal PMFs

$$p_X(x) := \mathbb{P}(X = x) = \sum_{\text{all } y's} p(x, y)$$

$$p_Y(y) = \mathbb{P}(Y = y) = \sum_{\text{all } x's} p(x, y)$$

Knowing Joint PMF $\Longrightarrow$ Can get p_X, p_Y

$\not\Longleftarrow$ In general

Two Discrete Random Variables

Example

A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are 0, \$100, and \$200.

Suppose an individual with both types of policy is selected at random from the agency's files. Let X be deductible amount on the auto policy and Y deductible amount on the homeowner's policy.

	0	100	200
100	0.2	0.1	0.2
250	0.05	0.15	0.3

Two Continuous Random Variables

Definition

Let X and Y be two continuous RVs.

The joint probability density function $f(x, y)$ is defined by

$$f(x, y) \neq \cancel{\mathbb{P}(X = x, Y = y)}$$

The joint PDF satisfies

1. $f(x, y) \geq 0$
2. $\int f(x, y) =$ $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)\, dx\, dy = 1$
3. $\mathbb{P}((X, Y) \in A) =$

$$\mathbb{P}((X, Y) \in A) = \iint_A f(x,y)\, dx\, dy$$

Two Continuous Random Variables

Example

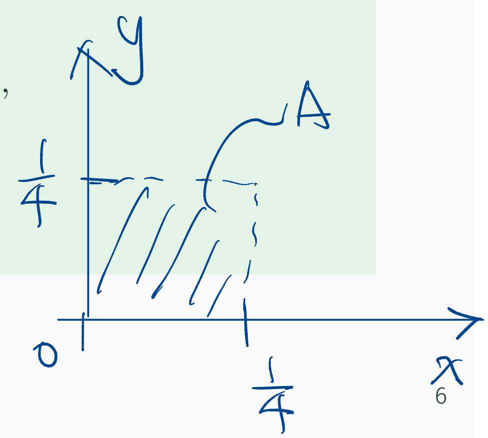
A bank operates both a drive-up facility and a walk-up window.

On a randomly selected day, let X be the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y the proportion of time that the walk-up window is in use.

The joint PDF is given by

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2), & 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4})$.



$$= \mathbb{P}((X,Y) \in A)$$

$$= \iint_A f(x,y)\,dx\,dy$$

$$= \int_0^{\frac{1}{4}}\int_0^{\frac{1}{4}} \frac{6}{5}(x+y^2)\,dx\,dy = \int_0^{\frac{1}{4}} \frac{6}{5}\left[\frac{1}{2}x^2 + y^2 \cdot x\right]_0^{\frac{1}{4}} dy$$

$$= \int_0^{\frac{1}{4}} \frac{6}{5}\left(\frac{1}{32} + \frac{y^2}{4}\right) dy = \frac{6}{5}\left[\frac{1}{32}\cdot y + \frac{y^3}{12}\right]_0^{\frac{1}{4}}$$

$$= \cdots$$

Two Continuous Random Variables

Example

A bank operates both a drive-up facility and a walk-up window.

On a randomly selected day, let X be the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y the proportion of time that the walk-up window is in use.

The joint PDF is given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2), & 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

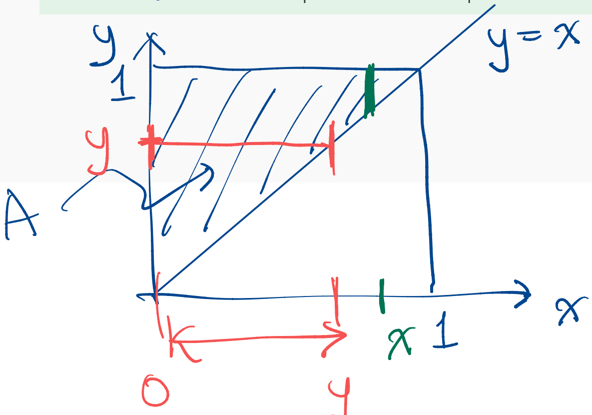
Find $\cancel{\mathbb{P}(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4})}$

$$\mathbb{P}(X \le Y)$$

$$= \mathbb{P}((X, Y) \in A)$$

$$= \int_0^1 \int_0^y \frac{6}{5}(x + y^2)\, dx\, dy$$

$$= \int_0^1 \int_x^1 \qquad dy\, dx$$



Joint PMF : $p(x, y) = \mathbb{P}(X = x, Y = y)$

Joint PDF : $f(x,y)$ such that

(i) $f(x,y) \geq 0$

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$

(iii) $P((X, Y) \in A) = \iint_A f(x,y)\, dx\, dy$.

Two Continuous Random Variables

Definition

For a given joint PDF $f(x, y)$ of random variables X and Y, the marginal probability density function of X is given by

$$f_X(x) := \int_{-\infty}^{\infty} f(x,y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

Two Continuous Random Variables

Example

A bank operates both a drive-up facility and a walk-up window.

On a randomly selected day, let X be the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y the proportion of time that the walk-up window is in use.

The joint PDF is given by

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2), & 0 \le x, y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

$0 \leq x \leq 1$
$0 \leq y \leq 1$

Find the marginal PDFs.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

$$= \int_0^1 \frac{6}{5}(x+y^2)\, dy$$

$$= \frac{6}{5}\left[xy + \frac{1}{3}y^3\right]_0^1 = \frac{6}{5}\left(x+\frac{1}{3}\right)$$

$$f_X(x) = \begin{cases} \frac{6}{5}\left(x+\frac{1}{3}\right), & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Two Continuous Random Variables

Example

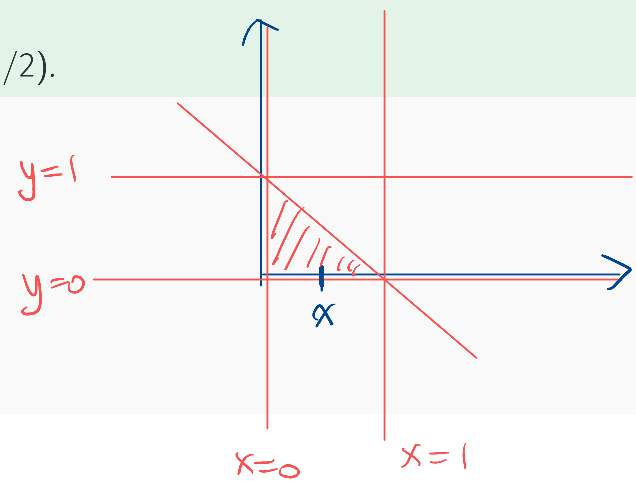
The joint PDF is given by

region Inequalities look at equalities gives bdry

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal PDFs and $\mathbb{P}(X+Y \le 1/2)$.

f_X f_Y



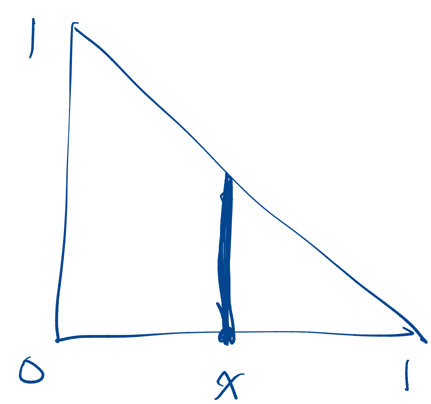
fixed $0 < x < 1$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

$$= \int_{\boxed{0}}^{\boxed{1-x}} 24xy \quad dy$$

$$= \left[12xy^2\right]_0^{1-x}$$

$$= 12x(1-x)^2.$$



$$f_Y(y) = 12y(1-y)^2$$

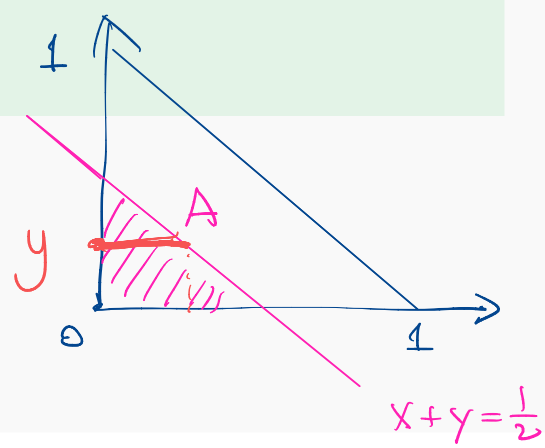
Two Continuous Random Variables

Example

The joint PDF is given by

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal PDFs and $\mathbb{P}(X+Y \le 1/2)$.



$$\mathbb{P}(\underbrace{X+Y \le \tfrac{1}{2}})$$

$$= \mathbb{P}((X,Y) \in A)$$

$$= \iint_A f(x,y)\,dx\,dy$$

$$= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-y} 24xy\,dx\,dy$$

Recall Two events A, B are indep. iff $\mathbb{P}(A \wedge B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Independent Random Variables

Definition

Two random variables X and Y are said to be independent if

for any intervals I, J in $\mathbb{R}$, $\{X \in I\}$, $\{Y \in J\}$ indep.

① If X, Y discrete,

X, Y indep $\iff p(x, y) = p_X(x) \cdot p_Y(y)$

② If X, Y continuous,

X, Y indep $\iff f(x, y) = f_X(x) \cdot f_Y(y)$

Independent Random Variables

Example

The joint PDF is given by

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

$$\left.\begin{aligned} f_X(x) &= \underline{12\, x(1-x)^2} \\ f_Y(y) &= \underline{12\, y(1-y)^2} \end{aligned}\right\}$$ previous example.

$$24xy = f(x,y) \neq f_X(x) \cdot f_Y(y)$$

$$= 12x(1-x)^2 \cdot 12 \cdot y(1-y)^2.$$

$\Rightarrow$ X, Y Not Indep.

Independent Random Variables

Example

Suppose that the lifetimes of two components are $\boxed{\text{independent}}$ of one another and that the first lifetime, X_1, has an exponential distribution with parameter λ_1, whereas the second, X_2, has an exponential distribution with parameter λ_2.

Find the joint PDF.

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$$

$$= \begin{cases} \lambda_1 e^{-\lambda_1 x_1} \cdot \lambda_2 e^{-\lambda_2 x_2}, & x_1, x_2 \geq 0 \\ 0, & o.w. \end{cases}$$

Independent Random Variables

Definition

The random variables $X_1, X_2, \cdots, X_n$ are said to be independent if

$\{X_1 \in I_1\}, \{X_2 \in I_2\}, \cdots, \{X_n \in I_n\}$ indep.

for any choice of $I_1, I_2, \cdots, I_n$.

Conditional Distributions

Definition

Let X and Y be two continuous RVs with joint PDF $f(x, y)$.

Then for any x for which $f_X(x) > 0$, the conditional probability density function of Y given that $X = x$ is

A New RV : $Y \mid X = x$ with $\underbrace{\text{PDF}}_{\text{Condi.}}$

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

Conditional Distributions

Example

A bank operates both a drive-up facility and a walk-up window.

On a randomly selected day, let X be the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y the proportion of time that the walk-up window is in use.

The joint PDF is given by

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2), & 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the conditional PDF of Y given $X = 0.8$.

Compute $\mathbb{P}(Y \le 0.5 | X = 0.8)$.

$$f_{Y|X}(y \mid 0.8) = \frac{f(0.8, y)}{f_X(0.8)} = \boxed{\frac{\cancel{\frac{6}{5}}(0.8+y^2)}{\cancel{\frac{6}{5}}(0.8+\frac{1}{3})}}$$

$0 \le y \le 1$.

$$f_X(x) = \frac{6}{5}\left(x + \frac{1}{3}\right)$$

$$f_X(0.8) = \frac{6}{5}\left(\frac{4}{5} + \frac{1}{3}\right) = \frac{104}{75}$$

$$\cancel{\frac{P(Y \le 0.5, X = 0.8)}{P(X = 0.8)}} = 0$$

$$P(Y \le 0.5 \mid X = 0.8) =$$

$$= P\left((Y \mid X = 0.8) \le 0.5\right)$$

$$= \int_0^{0.5} f_{Y|X}(y \mid 0.8)\, dy$$

$$= \int_0^{0.5} \frac{(0.8 + y^2)}{(0.8 + \frac{1}{3})}\, dy$$

$$E[Y \mid X = 0.8] = \int_0^1 y \cdot f_{Y|X}(y \mid 0.8)\, dy$$

Exercise

(5.1-12) Two components of a minicomputer have the following joint PDF for their useful lifetimes X and Y:

$$f(x,y) = \begin{cases} xe^{-x(y+1)}, & x \geq 0, y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. What is the probability that the lifetime X of the first component exceeds 3?
2. What are the marginal PDFs of X and Y? Are the two lifetimes independent? Explain.
3. What is the probability that the lifetime of at least one component exceeds 3?

Section 2.
Expected Values, Covariance, and Correlation

Expectation of a Function of Two Random Variables

Proposition

Let X and Y be jointly distributed RVs with PMF $p(x, y)$ or PDF $f(x, y)$ according to whether the variables are discrete or continuous.

Let $h(x, y)$ be a function of two variables, then we can define a new random variable $Z = h(X, Y)$.

The expectation of Z is

$$\mathbb{E}[Z] = \mathbb{E}[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x,y)\, p(x,y) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\, f(x,y)\, dx\, dy \end{cases}$$

Example

$h(x,y) = x$

$h(x,y) = y$

$h(x,y) = x \cdot y$

$\vdots$

$$E[X] = \underline{\iint x \cdot f(x,y)\, dx\, dy} = \int x \Big(\underbrace{\int f(x,y)\, dy}_{= f_X(x)} \Big) dx$$

$$= \int x \cdot f_X(x)\, dx$$

$$E[X \cdot Y] = \iint x \cdot y\, f(x,y)\, dx\, dy$$

Expectation of a Function of Two Random Variables

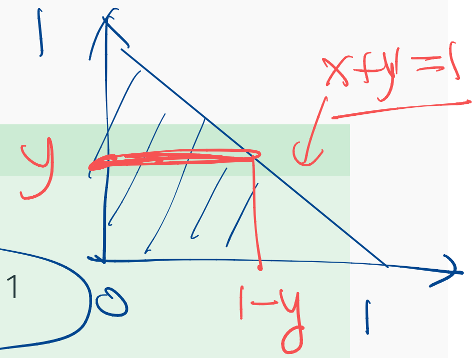
Example

Five friends have purchased tickets to a certain concert.

If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five?

Let X and Y denote the seat numbers of the first and second individuals, respectively.

Expectation of a Function of Two Random Variables



Example

The joint PDF is given by

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}[XY]$.

$$\mathbb{E}[XY] = \iint xy \cdot f(x,y)\, dx\, dy$$

$$= \int_0^1 \int_0^{1-y} 24 x^2 y^2 \, dx\, dy$$

$$= \int_0^1 \Big[8x^3 y^2 \Big]_0^{1-y} dy$$

$$= \int_0^1 8(1-y)^3 \cdot y^2 \, dy$$

$$= \int_0^1 8(1-t)^2 t^3 \, dt \qquad (1-y=t)$$

Covariance

$\mu_X = \mathbb{E}[X]$, $\mu_Y = \mathbb{E}[Y]$

Definition

The covariance between two RVs X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

$$= \iint (x - \mu_X)(y - \mu_Y)\, f(x, y)\, dx\, dy$$

$$= \sum_y \sum_x (x - \mu_X)(y - \mu_Y)\, p(x, y)$$

Covariance

Example

The joint PDF is given by

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the covariance of X, Y.

previous example.

$$\mathrm{Cov}(X,Y) = \underline{\mathbb{E}[X \cdot Y]} - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

$$\mathbb{E}[X] = \iint x \cdot f(x,y)\,dx\,dy$$

$$= \int_0^1 x \cdot \underbrace{12x(1-x)^2}_{f_X(x)}\,dx$$

$$= 12\int_0^1 x^2(1-2x+x^2)\,dx$$

$$= 12 \cdot \left[\frac{1}{3} - 2\cdot\frac{1}{4} + \frac{1}{5}\right] = \frac{12}{30} = \frac{2}{5}$$

$$= \mathbb{E}[Y] \quad \text{(by symmetry)}$$

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

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

Covariance

Proposition

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

4/8/2025

Recall X, Y joint PDF (or PMF)

- $Cov(X,Y) = \mathbb{E}[(X-\mu_X)\cdot(Y-\mu_Y)] = \iint (x-\mu_X)(y-\mu_Y) f(x,y)\, dx\, dy$

where $\mu_X = \mathbb{E}[X]$, $\mu_Y = \mathbb{E}[Y]$

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

Correlation Coefficient

Definition

The correlation coefficient of X and Y is defined by

$$\mathrm{Corr}(X,Y) = \rho_{X,Y} = \rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

(rho)

- $Corr(X,Y) = Corr(Y,X)$
- $Corr(X,X) = 1$.
- $-1 \leq Corr(X,Y) \leq 1$.

In general, $(\mathbb{E}[X\cdot Y])^2 \leq \mathbb{E}[X^2]\cdot\mathbb{E}[Y^2]$

$\Rightarrow \quad Cov(X,Y)^2 \leq Var(X)\cdot Var(Y)$

$\Rightarrow \quad Corr(X,Y)^2 \leq 1$

$$Var(aX+b) = a^2 Var(X)$$
$$Var(cY+d) = c^2 Var(Y)$$

$$Cov(aX+b, cY+d) = a \cdot c\, Cov(X, Y)$$

Correlation Coefficient

Properties

1. For constants a, b, c, d,

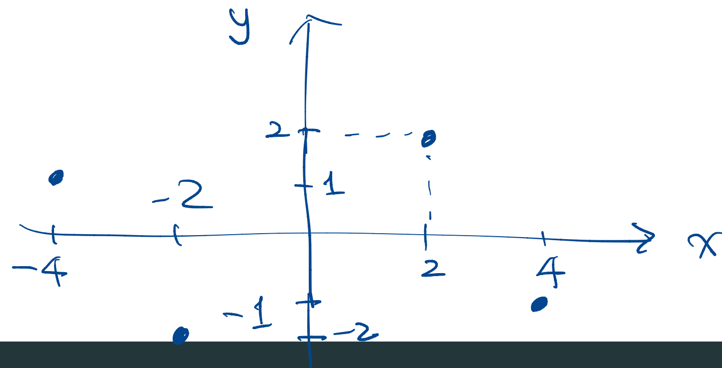
$$\mathrm{Corr}(aX+b, c\cancel{X}Y+d) = \frac{ac}{\sqrt{a^2c^2}} \; Corr(X, Y)$$

2. $-1 \le Corr(X, Y) \le 1$ $\quad (Cov(X,Y)=0)$
3. If X and Y are independent, then $\mathrm{Corr}(X, Y) = 0$. The converse does not hold in general.
4. If $\mathrm{Corr}(X, Y) = 1, -1$, then $Y = aX + b$ for some a, b.

- X, Y indep $\Rightarrow Cov(X, Y) = 0$
$\Rightarrow Corr(X, Y) = 0.$

- $Corr(aX+b, cY+d)$

$$= \begin{cases} Corr(X, Y), & \text{when } ac \geq 0 \\ -Corr(X, Y), & \text{when } a \cdot c < 0 \end{cases}$$



Correlation Coefficient

Example

The joint PMF is given by

$$p(x,y) = \begin{cases} \frac{1}{4}, & (x,y) = (-4,1), (4,-1), (2,2), (-2,-2) \\ 0 & \text{otherwise.} \end{cases}$$

Find the covariance and the correlation coefficient.

$X = 4, 2, -2, -4$ equally likely

$Y = 2 \quad 1 \quad -1 \quad -2$ "

$E[X] = 0 = E[Y]$

$$\mathrm{Cov}(X,Y) = E[X \cdot Y] = \frac{1}{4}\left((-4)\cdot 1 + 4(-1) + 2\cdot 2 + (-2)\cdot(-2)\right)$$

$$= 0 = \mathrm{Corr}(X,Y)$$

Exercise

(5.2-24) Six individuals, including A and B, take seats around a circular table in a completely random fashion.

Suppose the seats are numbered $1, 2, \cdots, 6$.

Let X be A's seat number and Y B's seat number.

If A sends a written message around the table to B in the direction in which they are closest,

how many individuals (including A and B) would you expect to handle the message?

X, Y with joint PDF or PMF

$f_{Y|X}(y|x)$ =

a New RV $Y \mid X = x$ with PDF $\frac{f(x,y)}{f_X(x)}$

PMF $\frac{p(x,y)}{p_X(x)}$ = $p_{Y|X}(y|x)$

$$\mathbb{E}[h(Y) \mid X=x] = \int h(y) \cdot f_{Y|X}(y|x)\, dy$$

Conditional distribution

Definition

The conditional expectation of Y given $X = x$ is defined by

$$\mathbb{E}[Y|X = x] = \sum_y y f_{Y|X}(y|x).$$

The conditional variance of Y given $X = x$ is defined by

$$\begin{aligned} \mathrm{Var}(Y|X = x) &= \mathbb{E}[(Y - \mathbb{E}[Y|X = x])^2|X = x] \\ &= \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2. \end{aligned}$$

$\mathbb{E}[Y \mid X = x] \Leftarrow$ a number

- For each choice of x, $\mathbb{E}[Y|X=x]$ might give a diff #.
- $h(x) = \mathbb{E}[Y \mid X = x]$ a function of x

Contional expectation as a function and a random variable

One can consider $\mathbb{E}[Y|X = x]$ as a function of x.

Say $\boxed{h(x) = \mathbb{E}[Y|X = x]}$ $\underbrace{x}$: real #

We define a random variable $\mathbb{E}[Y|X] = h(\underbrace{X}_{\text{RV}})$.

New RV $h(X) = \mathbb{E}[Y \mid X]$

X, Y

$\Rightarrow$ New RV $Y \mid X = x$ for each x

$\Rightarrow$ $h(x) = \mathbb{E}[Y \mid X = x]$

$\Rightarrow$ From h, define a new RV $h(X)$

$h(X) = \mathbb{E}[Y \mid X]$

Contional expectation as a function and a random variable

$$\mathbb{E}[h(X)] = \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$$

Theorem

1. $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$
2. $\mathsf{Var}(Y) = \mathbb{E}[\mathsf{Var}(Y|X)] + \mathsf{Var}(\mathbb{E}[Y|X])$

Exercise

$$\mathbb{E}[Y] = \mathbb{E}[\,\mathbb{E}[Y \mid X]\,]$$

A miner is trapped in a mine containing 3 doors.

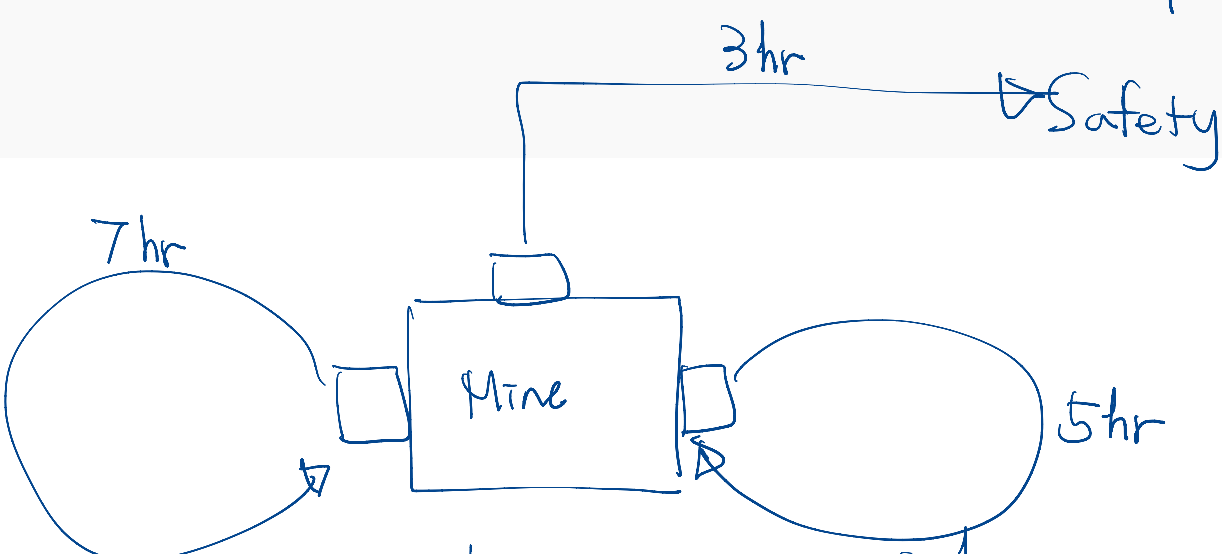
The first door leads to a tunnel that will take him to safety after 3 hours of travel.

The second door leads to a tunnel that will return him to the mine after 5 hours of travel.

The third door leads to a tunnel that will return him to the mine after 7 hours.

If we assume that the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until he reaches safety?

$= Y$



$$X = 1^{st}\text{ choice} = \begin{cases} 1 \\ 2 \\ 3 \end{cases}$$

If $\quad X = 1 \qquad E[Y|X=1] = \underline{3}$

$X = 2 \qquad E[Y|X=2] = \underline{E[Y] + 5}$

$X = 3 \qquad E[Y|X=3] = \underline{E[Y] + 7}$

$$E(E[Y|X]) = \frac{1}{3}\cdot 3 + \frac{1}{3}(E[Y]+5) + \frac{1}{3}(E[Y]+7)$$

$$= E[Y]$$

$$\frac{1}{3}E[Y] = \frac{1}{3}(3+5+7)$$

$$E[Y] = 15.$$

$\blacksquare$

Section 4.
The Distribution of the Sample Mean

joint PDF

$$f(x_1, x_2, \cdots, x_n) = \underbrace{f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n)}_{\text{equal functions}}$$

Sample Mean

Definition

The RVs $X_1, X_2, \cdots, X_n$ are said to form a (simple) random sample of size n if

1. they are independent RVs, and
2. every X_i has the same probability distribution.

] indep. identically distributed = i.i.d.

The sample mean is defined by

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The sample total is defined by

$$T = X_1 + X_2 + \cdots + X_n$$

Simple case $n = 2$.

$$T = X_1 + X_2$$

$$E[T] = E[X_1 + X_2] = E[X_1] + E[X_2]$$

$$Var(T) = E[(X_1+X_2)^2] - (E[X_1+X_2])^2$$

$$= Var(X_1) + Var(X_2)$$

↑ because X_1, X_2 indep.

If not indep. $Var(T) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$

$\mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_n) = \sigma^2$

Sample Mean

$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \mu$

Proposition

i.i.d.

Let $X_1, \cdots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then,

$$\mathbb{E}[\bar{X}] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \cdot n\sigma^2 = \sigma^2/n$$

$$\mathbb{E}[T] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n \cdot \mu$$

$$\mathrm{Var}(T) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n \cdot \sigma^2$$

$$T = X_1 + X_2 + \cdots + X_n$$

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{T}{n}$$

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{T}{n}\right] = \mathbb{E}\left[\left(\frac{1}{n}\right) \cdot T\right] = \frac{1}{n}\mathbb{E}[T]$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\left(\frac{1}{n}\right) T\right) = \frac{1}{n^2}\mathrm{Var}(T)$$

Sample Mean

Example

In a notched tensile fatigue test on a titanium specimen, the expected number of cycles to first acoustic emission (used to indicate crack initiation) is $\mu = 28,000$, and the standard deviation of the number of cycles is $\sigma = 5000$.

Let $X_1, X_2, \cdots, X_{25}$ be a random sample of size 25, where each X_i is the number of cycles on a different randomly selected specimen.

$$\bar{X} = \frac{X_1 + \cdots + X_{25}}{25}$$

$$E[\bar{X}] = \mu = 28,000$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{(5000)^2}{25} = 1000,000$$

$$\sqrt{Var(\bar{X})} = \sqrt{1000000} = 1000$$

$X_1, \cdots, X_n$ indep. same distribution

The Case of a Normal Population Distribution

Proposition indep. $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$

Let $X_1, \cdots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$T \sim N(n\mu, n\sigma^2)$$

$E[\bar{X}] = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n}$

$X_1, \cdots, X_n \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

The Case of a Normal Population Distribution

Smaller variance.

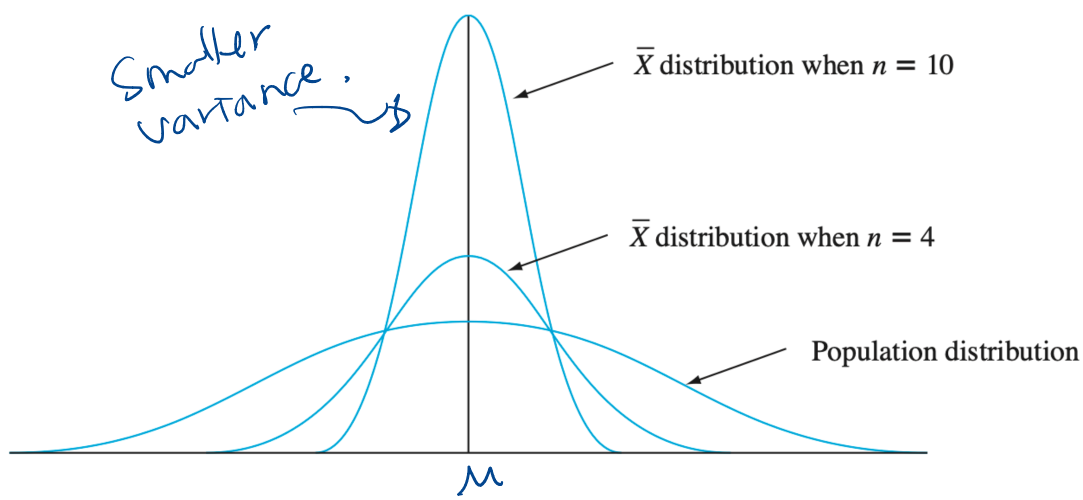


Figure 5.14 A normal population distribution and $\bar{X}$ sampling distributions

The Case of a Normal Population Distribution

Example

Indep.

The time that it takes a randomly selected rat of a certain subspecies to find its way through a maze is $\boxed{\text{a normally}}$ distributed RV with $\mu = 1.5$ min and $\sigma = .35$ min.

Suppose five rats are selected.

Let $X_1, \ldots, X_5$ denote their times in the maze.

$X_1, X_2, \cdots, X_5 \sim N(\mu, \sigma^2)$

Assuming the X_i's to be a random sample from this normal distribution,

what is the probability that the total time is between 6 and 8 min?

$$T = X_1 + \cdots + X_5 \sim N(5\cdot(1.5), 5\cdot(0.35)^2)$$

$$N(7.5, 0.6125)$$

$$P(6 < T < 8)$$

$$Z \sim N(0,1)$$

$$= P\left(\frac{6-7.5}{\sqrt{0.6125}} < \boxed{\frac{T-7.5}{\sqrt{0.6125}}} < \frac{8-7.5}{\sqrt{0.6125}}\right)$$

$$= \Phi(\quad) - \Phi(\quad) = \text{use table!!}$$

The Central Limit Theorem

(CLT)

Theorem i.i.d

finite finite

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean μ and variance σ^2.

If n is sufficiently large, $\bar{X}$ and T have approximately normal distributions.

Rule of Thumb: If $n \geq 30$, the Central Limit Theorem can be used.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\boxed{\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0, 1)}$$

The Central Limit Theorem

Example

A certain consumer organization customarily reports the number of major defects for each new automobile that it tests.

Suppose the number of such defects for a certain model is a random variable with mean value 3.2 and standard deviation 2.4.

Among 100 randomly selected cars of this model, how likely is it that the sample average number of major defects exceeds 4?

$X_1, X_2, \cdots, X_{100}$: i.i.d.

$$\bar{X} \approx N\left(3.2, \frac{(2.4)^2}{100}\right)$$

$$\frac{\bar{X} - 3.2}{0.24} \approx N(0,1)$$

$$P(\bar{X} > 4) = P\left(\frac{\bar{X} - 3.2}{0.24} > \frac{4 - 3.2}{0.24}\right)$$

$$\approx P(Z > 3.33)$$

by CLT

$$= 1 - \Phi(3.33)$$

$X_1, X_2, \cdots, X_n$: random sample of size n
(i.i.d.)

(i) have the same distribution

$$\left(\begin{array}{l} \mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_n] = \mu \\ Var(X_1) = Var(X_2) = \cdots = Var(X_n) = \sigma^2 \end{array} \right)$$

(ii) Independent.

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{T}{n}$$

$$T = X_1 + \cdots + X_n$$

$$\left[\begin{array}{ll} \mathbb{E}[\bar{X}] = \mu, & Var(\bar{X}) = \frac{\sigma^2}{n} \\ \mathbb{E}[T] = n\mu, & Var(T) = n \cdot \sigma^2 \end{array} \right]$$

Why?

In general, for general X_1, X_2 (not neccessarily i.i.d.)

$$\mathbb{E}[aX_1 + bX_2] = a\mathbb{E}[X_1] + b\mathbb{E}[X_2]$$

$$Var(X_1 + X_2) = \mathbb{E}[(X_1 + X_2)^2] - (\mathbb{E}[X_1 + X_2])^2$$

$$= \mathbb{E}[X_1^2 + 2X_1 \cdot X_2 + X_2^2]$$

$$- \left((\mathbb{E}[X_1])^2 + 2\mathbb{E}[X_1]\mathbb{E}[X_2] + (\mathbb{E}[X_2])^2 \right)$$

$$= \left(\mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 \right) + 2\left(\mathbb{E}[X_1 \cdot X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \right)$$

$$+ \left(\mathbb{E}[X_2^2] - (\mathbb{E}[X_2])^2 \right)$$

$$= Var(X_1) + 2\boxed{Cov(X_1, X_2)} + Var(X_2)$$

⟨Central Limit Theorem⟩

Suppose $X_1, X_2, \cdots, X_n$ i.i.d.

Assume $\mu = \mathbb{E}[X_1] < \infty$, $\sigma^2 = Var(X_1) < \infty$

Then,

$\bar{X}$ is approximately normal

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\boxed{\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0, 1)}$$

$n \to \infty$

?

"approximately" or "Converges to"

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq x\right) \xrightarrow[\text{as } n \to \infty]{} \Phi(x)$$

<u>Example</u> $X \sim Bin(n, p)$

$X = X_1 + X_2 + \cdots + X_n$

$X_1, \cdots, X_n \sim Ber(p)$ indep

$\frac{X}{n} = \bar{X} \Rightarrow$ normal by CLT.

$$\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X}{n} - \mu}{\sqrt{\sigma^2/n}} \Rightarrow N(0, 1)$$

$\mu = p$

$\sigma^2 = p(1-p)$

$$= \frac{X - np}{\sqrt{np(1-p)}}$$

The Central Limit Theorem

Normal approximation to Binomial

If $X \sim \text{Bin}(n, p)$ and n is large enough,

X and X/n have approximately normal distribution.

By 2:33

Exercise

(5.4-56) A binary communication channel transmits a sequence of "bits" (0s and 1s). Suppose that for any particular bit transmitted, there is a 10% chance of a transmission error (a 0 becoming a 1 or a 1 becoming a 0).

Assume that bit errors occur independently of one another.

1. Consider transmitting 1000 bits. What is the approximate probability that at most 125 transmission errors occur?
2. Suppose the same 1000-bit message is sent two different times independently of one another.
   What is the approximate probability that the number of errors in the first transmission is within 50 of the number of errors in the second?

① $X = \#$ of errors in 1000 bits $\sim Bin(1000, 0.1)$

$X \approx N(100, 90)$

$P(X \leq 125) \approx P\left(Z \leq \frac{125-100}{\sqrt{90}}\right)$

with half-unit correction:

$P(X \leq 125) = P(X \leq 125.5) \approx P\left(Z \leq \frac{125.5-100}{\sqrt{90}}\right)$

② $X_1 \sim Bin(1000, 0.1)$

$X_2 \sim Bin(1000, 0.1)$ } Indep.

$$\mathbb{P}(-50 \le X_1 - X_2 \le 50)$$

$$\left(\frac{X_1-100}{\sqrt{90}} - \frac{X_2-100}{\sqrt{90}} = \frac{X_1-X_2}{\sqrt{90}} \right)$$

$$= \mathbb{P}\left(-\frac{50}{\sqrt{90}} \le \underbrace{\left(\frac{X_1-100}{\sqrt{90}}\right)}_{\approx Z_1} - \underbrace{\left(\frac{X_2-100}{\sqrt{90}}\right)}_{\approx Z_2} \le \frac{50}{\sqrt{90}}\right)$$

$Z_1, Z_2 \sim N(0,1)$

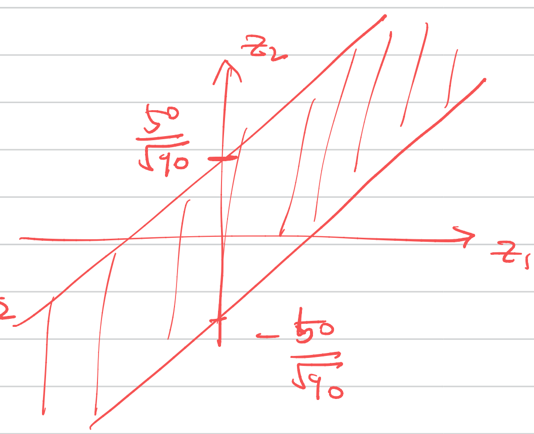
Indep

$$\approx \mathbb{P}\left(-\frac{50}{\sqrt{90}} \le Z_1 - Z_2 \le \frac{50}{\sqrt{90}}\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\frac{50}{\sqrt{90}}+z_2}^{\frac{50}{\sqrt{90}}+z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z_2^2}{2}} \, dz_1 \, dz_2$$

$$\mathbb{P}((Z_1, Z_2) \in A)$$

$$= \iint_A f(z_1, z_2) \, dz_1 \, dz_2$$



Section 5.
The Distribution of a Linear Combination

Linear Combination

Definition

Given a collection of n random variables $X1, \cdots, X_n$ and constants $a_1, \cdots, a_n$,

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

is called a linear combination of the X_i's.

Examples

① If $a_1 = a_2 = \cdots = a_n = 1$

$Y = X_1 + \cdots + X_n = T$

② If $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$

$Y = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n = \overline{X}$.

Linear Combination

Proposition

For a collection of n random variables $X1, \cdots, X_n$ and constants $a_1, \cdots, a_n$, consider

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Then,

$$\mathbb{E}[Y] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2] + \cdots + a_n \mathbb{E}[X_n]$$

$$\mathrm{Var}(Y) =$$

In particular, if they are independent,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2 \, \mathrm{Var}(X_i)$$

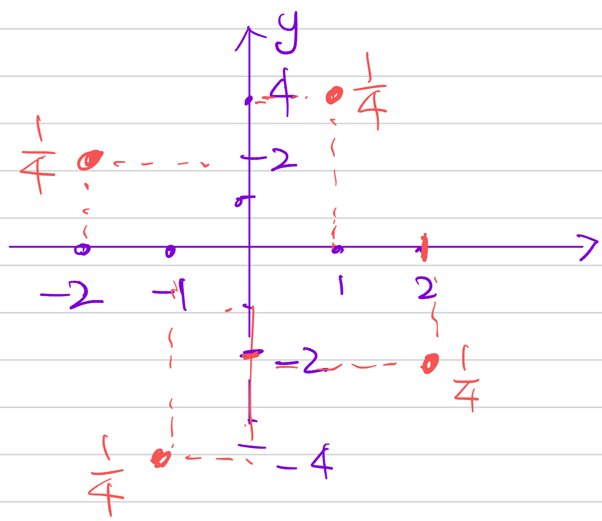
$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(a_i X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(a_i X_i, a_j X_j)$$

$\binom{n}{2}$ terms

$$= \sum_{i=1}^{n} a_i^2 \, \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \, \mathrm{Cov}(X_i, X_j)$$

⟨Counter Example⟩

$\underset{\displaystyle E[XY] - E[X]E[Y]}{\underset{\parallel}{\underline{\mathrm{Cov}(X,Y)}}} = 0$ but Not Indep.



$E[X \cdot Y] = \frac{1}{4} \cdot (4 - 4 + 4 - 4)$

$= 0$

$\therefore \mathrm{Cov}(X, Y) = 0$

$P(X = 2, Y = -2) = \frac{1}{4}$

$P(X = 2) = \frac{1}{4}$

$P(Y = -2) = \frac{1}{4}$

$X \sim \mathrm{Unif}(-1, 1)$ $\qquad \underline{Y = |X|} \sim \mathrm{Unif}(0, 1)$

$P(Y \le t) = P(|X| \le t)$

$= P(-t \le X \le t)$

$= \int_{-t}^{t} \frac{1}{2} \, dx$

$= \frac{2t}{2} = t$

$\mathrm{Cov}(X, Y)$

$= \underline{E[XY]} - \underset{0}{\underset{\parallel}{\underline{E[X]}}} \, \underline{E[Y]}$

$= \int_{-1}^{1} x|x| \cdot \frac{1}{2} \, dx = 0$

$$\mathbb{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$
$$= a_1\,\mathbb{E}[X_1] + a_2\,\mathbb{E}[X_2] + \cdots + a_n\,\mathbb{E}[X_n]$$

$$\mathrm{Var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n)$$
$$= a_1^2\,\mathrm{Var}(X_1) + a_2^2\,\mathrm{Var}(X_2) + \cdots + a_n^2\,\mathrm{Var}(X_n)$$
$$+ 2\big(\mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3) + \cdots + \mathrm{Cov}(X_1, X_n)$$
$$+ \mathrm{Cov}(X_2, X_3) + \mathrm{Cov}(X_2, X_4) + \cdots\big)$$

Linear Combination

Example

A certain automobile manufacturer equips a particular model with either a six-cylinder engine or a four-cylinder engine.

Let X_1 and X_2 be fuel efficiencies for $\boxed{\text{independently}}$ and randomly selected six-cylinder and four-cylinder cars, respectively, with

$$\mu_1 = 22, \qquad \mu_2 = 26, \qquad \sigma_1 = 1.2, \qquad , \sigma_2 = 1.5.$$

Find $\mathbb{E}[X_1 - X_2]$ and $\mathrm{Var}(X_1 - X_2)$.

$$\mathbb{E}[X_1 - X_2] = \mathbb{E}[X_1] - \mathbb{E}[X_2] = 22 - 26 = -4$$

$$\mathrm{Var}(X_1 - X_2) = 1^2\,\mathrm{Var}(X_1) + (-1)^2\,\mathrm{Var}(X_2)$$
$$= (1.2)^2 + (1.5)^2$$
$$= 1.44 + 2.25$$
$$= 3.69.$$

Linear Combination

Proposition

If $X_1, X_2, \cdots, X_n$ are independent, normally distributed RVs (with possibly different means and/or variances), then any linear combination also has a normal distribution.

In particular, the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

$$X_i \sim N(\mu_i, \sigma_i^2) \qquad i = 1, \cdots, n$$

$$\Rightarrow \quad a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim N(\mu, \sigma^2)$$

$$\mu = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

$$\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 .$$

Example

- $X_1, \cdots, X_n$ i.i.d. $N(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- $X_1, \cdots, X_n$ i.i.d. $\mu < \infty$, $\sigma^2 < \infty$

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{by CLT}$$

Exercise

(5.5-62) Manufacture of a certain component requires three different machining operations.

Machining time for each operation has a normal distribution, and the three times are independent of one another.

The mean values are 15, 30, and 20 min, respectively, and the standard deviations are 1, 2, and 1.5 min, respectively.

What is the probability that it takes at most 1 hour of machining time to produce a randomly selected component?