Chapter 5. Joint Probability Distributions and Random Sample

Chapter 5: Joint Probability Distributions and Random Sample

- **5.1. Jointly Distributed Random Variables**
- **5.2. Expected Values, Covariance, and Correlation**
- **5. 3. Statistics and Their Distributions**
- **5.4. The Distribution of the Sample Mean**
- **5.5. The Distribution of a Linear Combination**

Introduction

In chapter 3 and 4, we studied probability models for a single random variable.

However, many problem involve several random variables simultaneously

We first discuss probability models for the joint behavior of several random variable, putting special emphasis on the case in which the variables are independent of one another.

We then study expected value of functions of several random variables, including covariance and correlation

5.1. Jointly Distributed Random Variables

- **The Joint Probability Mass Function for Two Discrete Random Variables**

Let X and Y be two discrete random variables defined on the sample space $\boldsymbol{S}$ of an experiment. The joint probability mass function $p(x,y)$ is defined for each pair of numbers (x,y) by

$$p(x, y) = P(X = x \ \ and \ \ Y = y)$$

5.1. Jointly Distributed Random Variables

- Let A be any set consisting of pairs of (x,y) values. Then the probability $P[(X,Y)\in A]$ is obtained by summing the joint pmf over pairs in A:

$$p[(X,Y)\in A]=\sum_{(x,y)\in A}\sum p(x,y)$$

- **Two requirements for a pmf**

$$p(x,y)\geq 0 \qquad \sum_{x}\sum_{y}p(x,y)=1$$

Example 5.1

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 each type of policy, a deductible amount must be specified. 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* = the deductible amount on the auto policy**, ***Y* = the deductible amount on the homeowner's policy**

Suppose the joint pmf is given the accompanying **Joint Probability Table**

	$p(x,y)$	y: 0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

Then $p(100,100)$=?

$P(Y \geq 100)$=?

5.1. Jointly Distributed Random Variables

- **Example 5.1 (Cont')**

	$p(x,y)$	y = 0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

p(100,100) =P(X=100 and Y=100) = 0.10

P(Y ≥ 100) = p(100,100) + p(250,100) + p(100,200) + p(250,200) = 0.75

5.1. Jointly Distributed Random Variables

- **The marginal probability mass function**

The marginal probability mass functions of X and Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x, y); \quad p_Y(y) = \sum_x p(x, y)$$

p_Y

p_X

	Y_1	Y_2	…	Y_{m-1}	Y_m
X_1	$p_{1,1}$	$p_{1,2}$		$p_{1,m-1}$	$p_{1,m}$
X_2	$p_{2,1}$	$p_{2,2}$		$p_{2,m-1}$	$p_{2,m}$
…					
X_{n-1}	$p_{n-1,m}$	$p_{n-1,m}$		$p_{n-1,m}$	$p_{n-1,m}$
X_n	$p_{n,m}$	$p_{n,m}$		$p_{n,m}$	$p_{n,m}$

- **Example 5.2 (Ex. 51. Cont')**

The possible X values are x=100 and x=250, so computing row totals in the joint probability table yields

	$p(x,y)$	y = 0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

$p_X(100)=p(100,0)+p(100,100)+p(100,200)=0.5$

$p_X(250)=p(250,0)+p(250,100)+p(250,200)=0.5$

$$p_x(x)=\begin{cases}0.5, x=100,250\\ 0, otherwise\end{cases}$$

5.1. Jointly Distributed Random Variables

- **Example 5.2 (Cont')**

$p(x,y)$		y: 0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

$p_y(0)=p(100,0)+p(250,0)=0.2+0.05=0.25$

$p_y(100)=p(100,100)+p(250,100)=0.1+0.15=0.25$

$p_y(200)=p(100,200)+p(250,200)=0.2+0.3=0.5$

$$p_Y(y)=\begin{cases}0.25, y=0,100\\ 0.5, y=200\\ 0, otherwise\end{cases}$$

$P(Y \geq 100) = p(100,100) + p(250,100) + p(100,200) + p(250,200)$
$= p_Y(100)+p_Y(200) = 0.75$

5.1. Jointly Distributed Random Variables

- The Joint Probability Density Function for Two Continuous Random Variables

Let X and Y be two continuous random variables. Then f(x,y) is the joint probability density function for X and Y if for any two-dimensional set A

$$P[(X,Y)\in A]=\iint_A f(x,y)dxdy$$

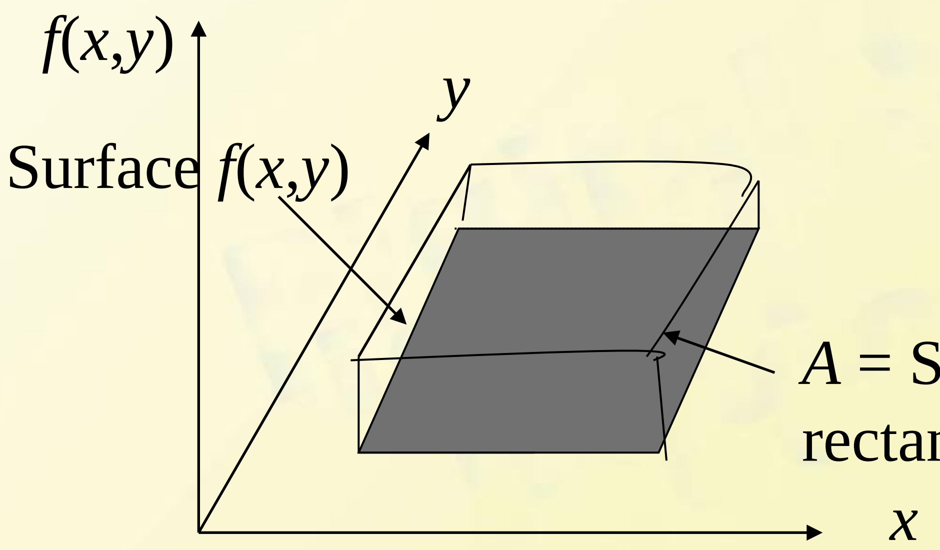
Two requirements for a joint pdf

1. $f(x,y)\geq 0$; for all pairs (x,y) in R^2
2. $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy=1$

5.1. Jointly Distributed Random Variables

- In particular, if A is the two-dimensional rectangle $\{(x,y): a \le x \le b,\ c \le y \le d\}$, then

$$P[(X,Y)\in A] = P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x,y)\,dy\,dx$$



$P[(X,Y)\in A]$ =volume under density surface above A

- **Example 5.3**

A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use, Y = the proportion of time that the walk-up window is in use. Let the **joint pdf of (X,Y)** be

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

1. **Verify that $f(x,y)$ is a joint probability density function;**
2. **Determine the probability** $P(0\le X\le \frac{1}{4}, 0\le Y\le \frac{1}{4})$

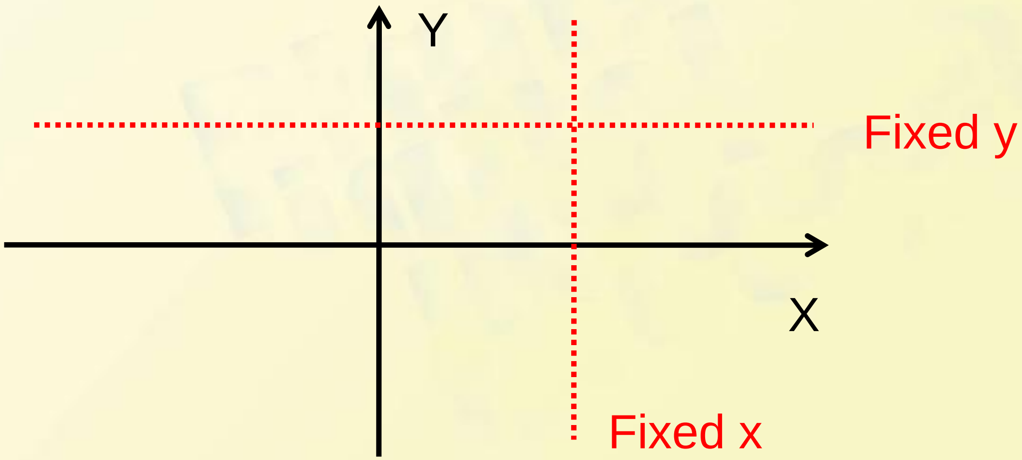
5.1. Jointly Distributed Random Variables

- **Marginal Probability density function**

The marginal probability density functions of *X* and *Y*, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

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

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



5.1. Jointly Distributed Random Variables

- **Example 5.4 (Ex. 5.3 Cont')**

The marginal pdf of X, which gives the probability distribution of busy time for the drive-up facility without reference to the walk-up window, is

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_0^1 \frac{6}{5}(x+y^2)dy = \frac{6}{5}x + \frac{2}{5}$$

for x in (0,1); and 0 for otherwise.

The marginal pdf of Y is

$$f_Y(y) = \begin{cases} \frac{6}{5}y^2 + \frac{3}{5} & 0 \le y \le 1 \\ 0 & otherwise \end{cases}$$

Then

$$P(\frac{1}{4} \le Y \le \frac{3}{4}) = \int_{1/4}^{3/4} f_Y(y)dy = 0.4625$$

Example 5.5

A nut company markets cans of deluxe mixed nuts containing almonds, cashews, and peanuts. Suppose the net weight of each can is exactly 1 lb, but the weight contribution of each type of nut is random. Because the three weights sum to 1, a joint probability model for any two gives all necessary information about the weight of the third type. Let X = the weight of almonds in a selected can and Y = the weight of cashews. The joint pdf for (X,Y) is

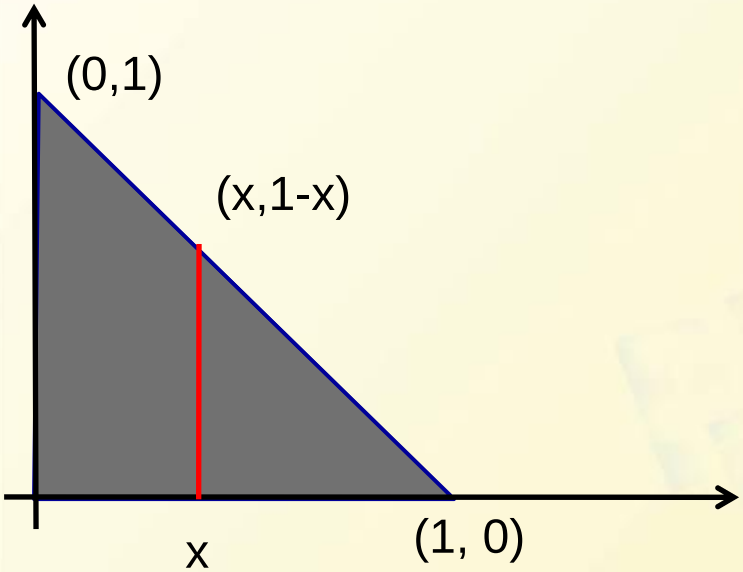
$$f(x,y)=\begin{cases}24xy & 0\le x\le 1, 0\le y\le 1, x+y\le 1\\ 0 & otherwise\end{cases}$$

1. Verify that $f(x,y)$ is a joint probability density function;
2. Let A ={(x,y):0≤x≤1,0≤y≤1, and x+y≤0.5}, P(X,Y) ∈A=?
3. $f_x(x)$ =? And $f_y(y)$=?

5.1. Jointly Distributed Random Variables

- **Solution:**

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



1: f(x,y) ≥ 0

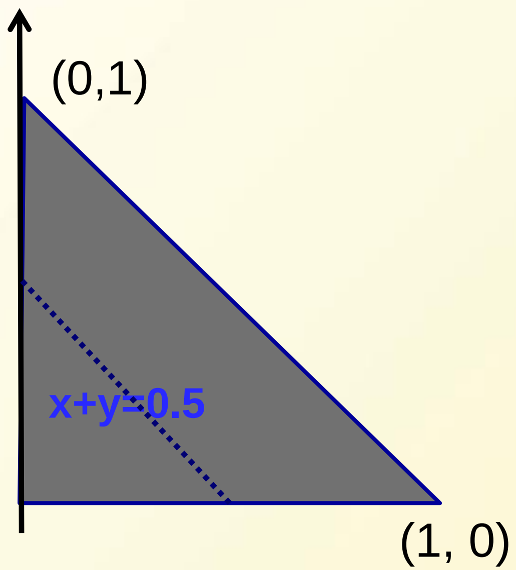
$$2: \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dydx = \iint_D f(x,y)dydx$$

$$= \int_0^1 \{\int_0^{1-x}(24xy)dy\}dx$$

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

- **Example 5.5 (Cont')**

Let the two type of nuts together make up at most 50% of the can, then A={(x,y); 0≤x ≤1; 0 ≤ y ≤ 1, x+y ≤ 0.5}



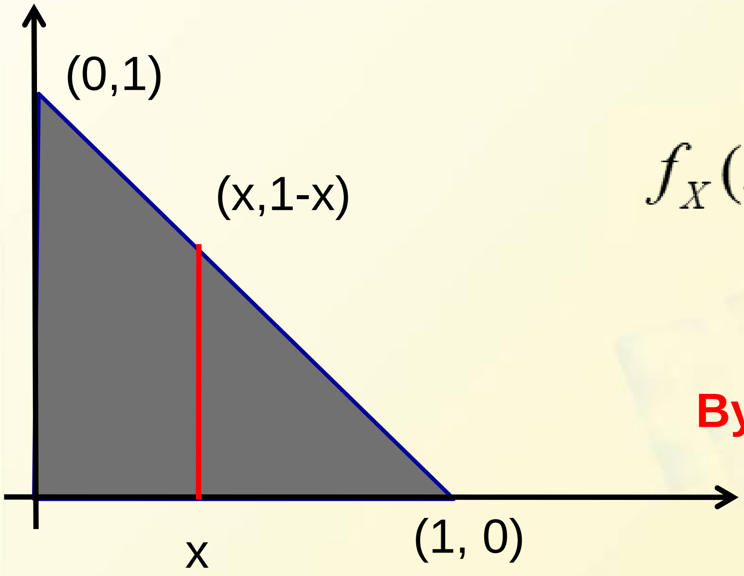
$$P((X,Y)\in A)=\iint_A f(x,y)\,dy\,dx$$

$$=\int_0^{0.5}\left\{\int_0^{0.5-x}(24xy)\,dy\right\}dx$$

$$=0.0625$$

5.1. Jointly Distributed Random Variables

- **Example 5.5 (Cont')**



The marginal pdf for almonds is obtained by holding X fixed at x and integrating f(x,y) along the vertical line through x:

$$f_X(x) = \begin{cases} \int_0^{1-x} (24xy)dy = 12x(1-x)^2, 0 \le x \le 1 \\ 0, otherwise \end{cases}$$

By symmetry

$$f_Y(y) = \int_0^{1-y} (24xy)dx = 12y(1-y)^2$$

- **Independent Random Variables**

Two random variables X and Y are said to be independent if for every pair of x and y values,

$$p(x, y) = p_X(x) \cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are continuous}$$

Otherwise, X and Y are said to be dependent.

Namely, two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's.

- **Example 5.6**

In the insurance situation of Example 5.1 and 5.2

$$p(100,100) = 0.1 \neq (0.5)(0.25) = p_X(100)\,p_Y(100)$$

$p(x,y)$		y: 0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

So, X and Y are not independent.

5.1. Jointly Distributed Random Variables

- **Example 5.7 (Ex. 5.5 Cont')**

  Because f(x,y) has the form of a product, X and Y would appear to be independent. However, although

$$f_X(x) = \int_0^{1-x} (24xy)dy = 12x(1-x)^2$$

$$f_Y(y) = \int_0^{1-y} (24xy)dx = 12y(1-y)^2$$

By symmetry

$$f_x(\frac{3}{4}) = f_y(\frac{3}{4}) = \frac{9}{16}, f(\frac{3}{4},\frac{3}{4}) = 0 \neq \frac{9}{16} \cdot \frac{9}{16}$$

So, X and Y are not independent.

5.1. Jointly Distributed Random Variables

- **Example 5.8**

Suppose that the lifetimes of two components are **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. Then the joint pdf is

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \begin{cases} \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2} & x_1 > 0, x_2 > 0 \\ 0 & otherwise \end{cases}$$

Let λ_1 =1/1000 and λ_2=1/1200. So that the expected lifetimes are 1000 and 1200 hours, respectively. The probability that both component lifetimes are at least 1500 hours is

$$P(1500 \le X_1, 1500 \le X_2) = P(1500 \le X_1) P(1500 \le X_2)$$

5.1. Jointly Distributed Random Variables

▪ More than Two Random Variables

If $X_1, X_2, \ldots, X_n$ are all discrete rv's, the joint pmf of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If the variables are continuous, the joint pdf of $X_1, X_2, \ldots, X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$P(a_1 \le X_1 \le b_1, \ldots, a_n \le X_n \le b_b) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_n \ldots dx_1$$

5.1. Jointly Distributed Random Variables

- **Multinomial Experiment**

An experiment consisting of ***n* independent and identical trials**, in which each trial can result in any one of ***r* possible outcomes**.

Let p_i=P(Outcome *i* on any particular trial), and define random variables by X_i=the number of trials resulting in outcome *i* (i=1,…,r). The joint pmf of $X_1,\ldots,X_r$ is called the multinomial distribution. The joint pmf of $X_1,\ldots X_r$ can be shown to be

$$p(x_1,\ldots,x_r)=\begin{cases}\dfrac{n!}{(x_1!)(x_2!)\ldots(x_r!)}p_1^{x1}\ldots p_r^{xr}, x_i=0,1\ldots with x_1+x_2\ldots+x_r=n\\ 0\end{cases}$$

Note: the case r=2 gives the binomial distribution.

5.1. Jointly Distributed Random Variables

- Independent

Definition:

The random variables X_1, X_2, …X_n are said to be independent if for every subset X_{i1}, X_{i2},…, X_{ik} of the variable (each pair, each triple, and so on), the joint pmf or pdf of the subset is equal to the product of the marginal pmf's or pdf's.

- **Example 5.10**

When a certain method is used to collect a fixed volume of rock samples in a region, there are four resulting rock types. Let X_1, X_2, and X_3 denote the proportion by volume of rock types 1, 2 and 3 in a randomly selected sample. If the joint pdf of X_1,X_2 and X_3 is

$$f(x_1, x2, x3) = \begin{cases} kx_1x_2(1-x_3), 0 \le x_1 \le 1, 0 \le x_2 \le 1, 0 \le x_3 \le 1, x_1 + x_2 + x_3 \le 1 \\ 0, otherwise \end{cases}$$

(1) Determine *k*

(2) $P(X_1+X_2 \leq 0.5)=?$

Solution:

(1) The k is determined by

$$\iiint_{D_1} f(x_1, x2, x3) = 1, D_1 : -\infty \le x_i \le \infty, i = 1, 2, 3 \Rightarrow$$

k=144.

(2) The probability that rocks of types 1 and 2 together account for at most 50% of the sample is

$$\iiint_{D_2} f(x_1, x2, x3) = 0.6066, D_2 : X_1 + X_2 \le 0.5$$

- **Example 5.11**

  If $X_1, \ldots, X_n$ represent the lifetime of ***n* components**, the components operate **independently of one another**, and each lifetime is exponentially distributed with parameter λ, then

  (A) Joint pdf is?

  (B) If there n components constitute a system that will fail as soon as a single component fails, then the probability that the system lasts past *t* time is ?

5.1. Jointly Distributed Random Variables

- **Solution:**

(A)

$$f(x_1, x_2, \ldots x_n) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})\ldots(\lambda e^{-\lambda x_n})$$

$$= \begin{cases} \lambda^n e^{-\lambda \sum xi}, x_1 \geq 0; x_2 \geq 0; \ldots, x_n \geq 0; \\ 0, otherwise \end{cases}$$

(B)

$$P(X_1 > t, X_2 > t, \ldots, X_n > t) = \int_t^\infty \ldots \int_t^\infty f(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n$$

$$= (\int_t^\infty \lambda e^{-\lambda x_1} dx_1)\ldots(\int_t^\infty \lambda e^{-\lambda x_n} dx_n) = e^{-n\lambda t}$$

therefore,

$$P(systemlifetime \leq t) = 1 - e^{-n\lambda t},\ fort \geq 0$$

5.1. Jointly Distributed Random Variables

Conditional Distribution

Let X and Y be two continuous rv's with joint pdf f(x,y) and marginal X pdf $f_X(x)$. Then for any X values x for which $f_X(x)>0$, the conditional probability density function of Y given that X=x is

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

If X and Y are discrete, then

$$f_{Y|X}(y\,|\,x)=\frac{p(x,y)}{p_X(x)},-\infty<y<\infty$$

is the conditional probability mass function of Y when X=x.

- **Example 5.12 (Ex.5.3 and Ex.5.4 Cont')**

X= the proportion of time that a bank's drive-up facility is busy and Y=the analogous proportion for the walk-up window.

(A) The conditional pdf of Y given that X=0.8 is ?

(B) The probability that the walk-up facility is busy at most half the time given that X=0.8 is ?

Solution:

(A)

$$f_{Y|X}(y\,|\,0.8)=\frac{f(0.8,y)}{f_X(0.8)}=\frac{1.2(0.8+y^2)}{1.2(0.8)+0.4}=\frac{1}{34}(24+30y^2),0<y<1$$

(B)

$$f_{Y|X}(y\le 0.5\,|\,X=0.8)=\int_{-\infty}^{0.5} f_{Y|X}(y\,|\,0.8)dy=\int_{-\infty}^{0.5}\frac{1}{34}(24+30y^2)dy=0.39$$

- The Expected Value of a function h(x,y)

Let X and Y be jointly distribution rv's with **pmf** ***p*(*x*,*y*)** or **pdf** ***f*(*x*,*y*)** according to whether the variables are discrete or continuous. Then the expected value of a function ***h*(*X*,*Y*),** **denoted by E[h(X,Y)] or μh(X,Y) ,** is given by

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) \cdot p(x,y), X \& Y : discrete \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) dxdy, X \& Y : continuous \end{cases}$$

- **Example 5.13**

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?

Solution:

Let X and Y denote the seat numbers of the first and second individuals, respectively. Possible (X,Y) pairs are {(1,2),(1,3),…(5,4),and the joint pmf of (X, Y) is }

$$p(x,y)=\begin{cases}\dfrac{1}{20} & x=1,\dots,5;\ y=1,\dots,5;\ x\neq y\\ 0 & \text{otherwise}\end{cases}$$

The number of seats separating the two individuals is

h(X,Y)=|X-Y|-1

Example 5.13 (Cont')

The accompanying table gives h(x,y) for each possible (x,y) pair.

$h(x,y)$		x = 1	2	3	4	5
	1	--	0	1	2	3
	2	0	--	0	1	2
y	3	1	0	--	0	1
	4	2	1	0	--	0
	5	3	2	1	0	--

$$E[h(X,Y)] = \sum_{(x,y)} \sum h(x,y) \cdot p(x,y)$$

$$= \sum_{\substack{x=1 \\ x \neq y}}^{5} \sum_{y=1}^{5} (|x-y|-1) \cdot \frac{1}{20} = 1$$

- **Example 5.14**

In Example 5.5, the joint pdf of the amount X of almonds and amount Y of cashews in a 1-lb can of nuts was

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

If 1 lb of almonds costs the company \$1.00, 1 lb of cashews costs \$1.50, and 1 lb of peanuts costs \$0.50, then

(A) The total cost of the contents of a can is ?

(B) The expected total cost is ?

Solution:

(A) $h(X,Y)=(1)X+(1.5)Y+(0.5)(1\text{-}X\text{-}Y)=0.5+0.5X+Y$

(B) The expected total cost is

$$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\cdot f(x,y)\,dxdy$$

$$= \int_0^1\int_0^{1-x} (0.5+0.5x+y)\cdot 24xy\,dydx = \$1.10$$

Note: The method of computing $E[h(X_1,\ldots, X_n)]$, the expected value of a function $h(X_1, \ldots, X_n)$ of n random variables is similar to that for two random variables.

- Covariance

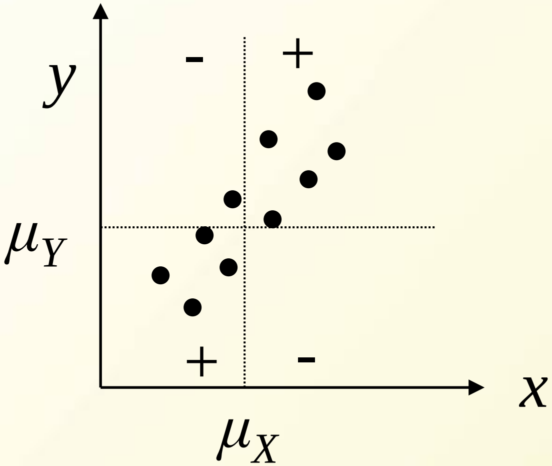
When two rv X and Y are **not independent**, it is frequently of interest to assess **how strongly they are related to one another.**

The Covariance between two rv's X and Y is

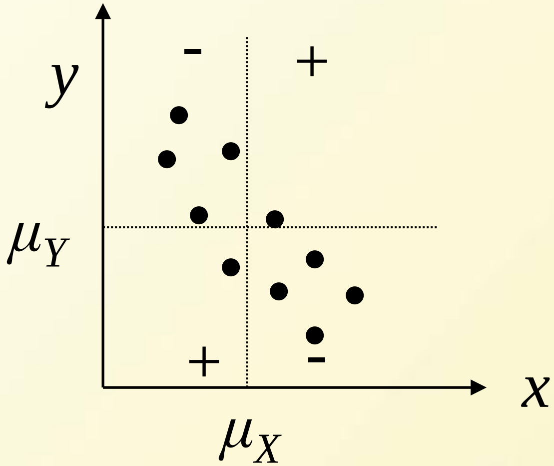
$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, p(x, y) & X, Y \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f(x, y)\, dx\, dy & X, Y \text{ continuous} \end{cases}$$

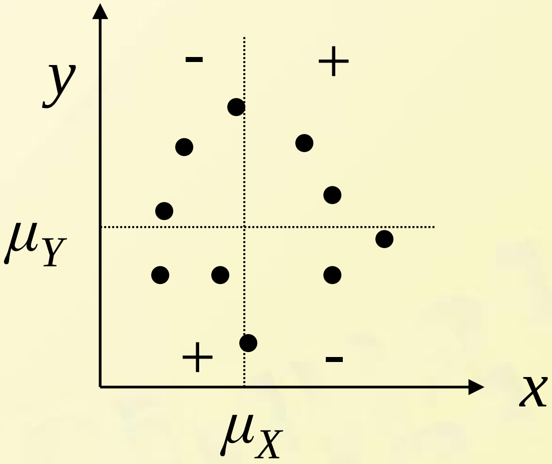
- Illustrates the different possibilities.



(a) positive covariance



(b) negative covariance;



(c) covariance near zero

Here: P(x, y) =1/10

For a strong positive relation ship, Cov(X,Y) should be quite positive.
For a strong negative relation ship, Cov(X,Y) should be quite negative.
If X and Y are not strongly related , Cov(X,Y) near 0

Example 5.15

The joint and marginal pmf's for X = automobile policy deductible amount and Y = homeowner policy deductible amount in Example 5.1 were

	$p(x,y)$	y: 0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

x	100	250
$p_X(x)$	.5	.5

y	0	100	200
$p_Y(y)$	.25	.25	.5

Find cov(X,Y)?

Solution:

From which $\mu_X = \sum x p_X(x) = 175$ and $\mu_Y = 125$.

Therefore

$$\mathrm{Cov}(X,Y) = \sum_{(x,y)}\sum (x-175)(y-125)\,p(x,y)$$

$$= (100-175)(0-125)(0.2) + \ldots + (250-175)(200-125)(0.3)$$

$$= 1875$$

5.2 Expected Values, Covariance, and Correlation

- Proposition

$$\mathrm{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$$

Note: $$\mathrm{Cov}(X, X) = E(X^2) - \mu^2{}_X = V(X)$$

- Correlation

The correlation coefficient of X and Y, denoted by Corr(X,Y), $\rho_{X,Y}$ or just ρ, is defined by

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

The normalized version of Cov(X,Y)

Example 5.17

It is easily verified that in the insurance problem of Example 5.15, $\sigma_X = 75$ and $\sigma_Y = 82.92$. This gives

$$\rho = 1875/(75)(82.92)=0.301$$

5.2 Expected Values, Covariance, and Correlation

- Proposition

1. If a and c are either both positive or both negative

$$\mathrm{Corr}(aX+b,\, cY+d) = \mathrm{Corr}(X,Y)$$

2. For any two rv's X and Y, $-1 \leq \mathrm{Corr}(X,Y) \leq 1$.

Statement 1 says that the correlation coefficient is not affected by a linear change in the units of measurement

According to Statement 2, the strongest possible positive relationship is evidenced by $\rho=+1$, *whereas* strongest possible negative relationship responds to $\rho=-1$

Proposition

1. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.

(when ρ = 0, X and Y are said to be uncorrelated.)

2. $\rho = 1$ or -1 **iff(if and only if)** $Y = aX+b$ for some numbers a and b with $a \neq 0$.