STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Review of Probability and Statistics

Part 2: Joint Distribution and Random Sample

Two Random Variables and Joint Distribution

- We can generalize the single random variable to two random variables
- The distribution of two random variables is called joint distribution
- Example
  - Two random variables X and Y are discrete, and their joint density function is

			y	
	$p(x, y)$	500	1000	5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

Two Random Variables and Joint Distribution

- Question
  1. P(X = Y) = ?
  2. P(X $>$ 500) = ?

	$p(x, y)$	y = 500	1000	5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

Two Random Variables and Joint Distribution

- Marginal distribution is the distribution of single random variable obtained from the joint distribution
- In this example, what is the marginal distribution for X and Y?

	$p(x, y)$	y = 500	1000	5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

Two Random Variables and Joint Distribution

- If the two random variables X and Y are continuous, then their joint distribution is defined by a two-dimensional smooth function
- For example

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

Two Random Variables and Joint Distribution

- The probability under joint density function (pdf) is integration
- For example

$$P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x + y^2)\, dx\, dy$$

Two Random Variables and Joint Distribution

- Marginal distribution of continuous random variables under joint pdf
- We need to integrate along one random variable

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

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

Independent Random Variables

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

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

or $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (5.1)

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

Conditional Distribution

- If random variable pair (X, Y) has a joint distribution, then the value of one variable will affect the distribution of another

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

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

If X and Y are discrete, replacing pdf's by pmf's in this definition gives the **conditional probability mass function of Y when $X = x$.**

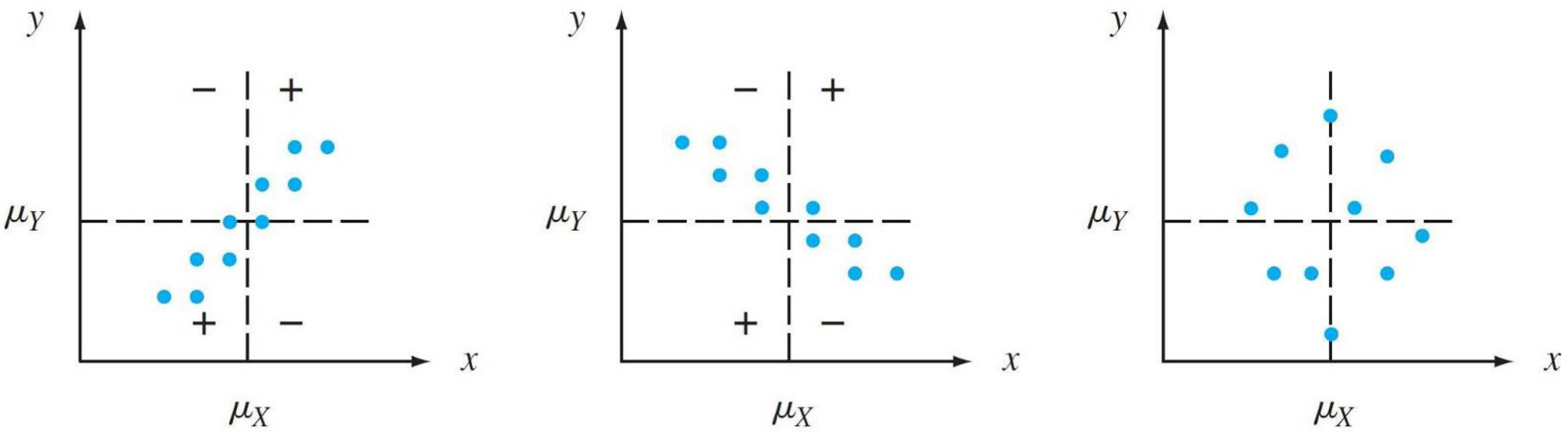
Covariance of Two Random Variables

- Covariance measures the linear relationship between two random variables
- How one variable's change affects the other one

The **covariance** between two rv's X and Y is

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

Covariance of Two Random Variables



(a) Positive covariance (b) Negative covariance (c) Covariance near zero

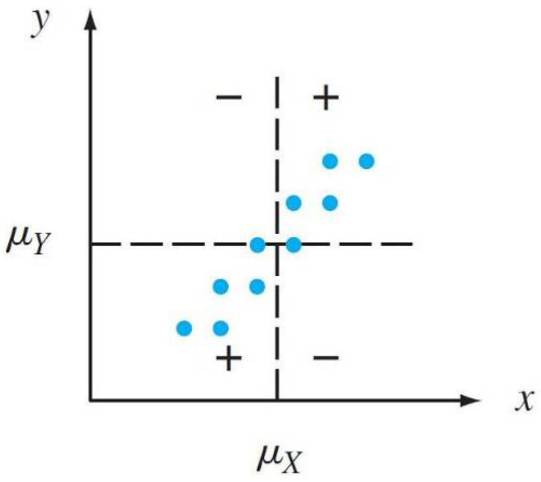
Correlation Coefficient of Two Random Variables

- The range of covariance is $(-\infty, \ +\infty)$ and the specific value depends on the scale of random variables
- The covariance of different random variables are not comparable
- Correlation coefficient (ρ) is the covariance normalized by standard deviation
- $\rho \in [-1, 1]$

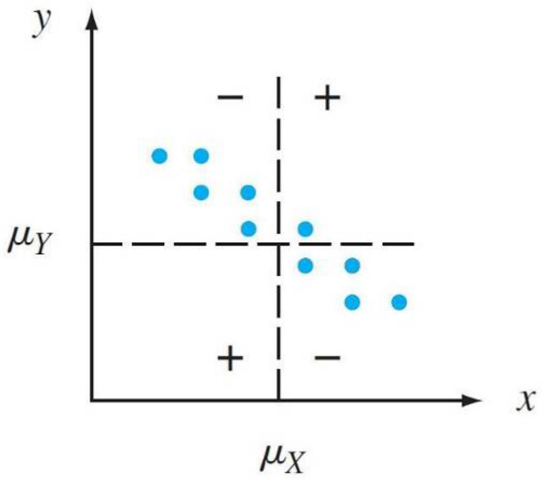
The **correlation coefficient** of X and Y, denoted by Corr(X, Y), $\rho_{X,Y}$, or just ρ, is defined by

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

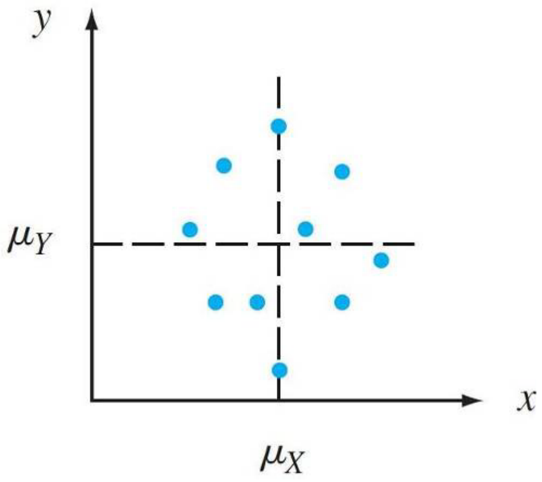
Correlation Coefficient of Two Random Variables



(a) $\rho > 0$



(b) $\rho < 0$



(c) $\rho \approx 0$

Random Sample

- Random sample is a subset we randomly select from the population
- In the random sample, we view each observation as a random variable and denote the sample by $X_1, X_2, \dots, X_n$
- X_i's are independent and identically distributed (i.i.d.)
- Random samples obtained by different sampling process are different
- Any sample statistics (mean $\bar{X}$, std S) are also random

Central limit theorem

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. If sample size n is reasonably large (>30), then the sample mean $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

Central Limit Theorem