Applied Methods in DigEcon

Applied Lecture 1:
Intro, Basic Statistical Concepts & the R Environment

Michael E. Kummer

Spring 2024

Introduction

Instructor

Michael Kummer

Assoc. Prof at Nova SBE, previously UEA and GaTech

PhD in **ICT, Search and Market Outcomes** at the **Department of Economics** from University of Mannheim and ZEW-Mannheim, Germany

- Research Interests: **Social Networks, Peer Production in Online Settings** and **Firm Behavior in E-Commerce**

Email: michael.kummer@novasbe.pt

Office Hours: upon request/ Please also come and see me before/after class.

Credit: Applied slides based on materials by: Rodrigo Belo and Miguel Godinho de Matos

Statistics: Random Variables and their Probability Distributions

Purpose: learn this vocabulary

Random Variables

A **random variable** is variable whose value is subject to variations due to chance.

Examples of random variables:

- Number of heads appearing in 10 flips of a coin
- Number of sunny days in a calendar year
- Number of people showing up for a flight
- . . .

What is chance?

- Chance does not need to be really random
- In most cases it can be a phenomenon that we are unable or not interested in understanding or predicting

Discrete Random Variables

A **discrete random variable** is one that takes on only a finite or countably infinite number of values.

Examples:

- Number of heads appearing in 10 flips of a coin
- Number of sunny days in a calendar year
- Number of people showing up for a flight

A Bernoulli random variable is the simplest example of a discrete random variable.

Example (1 flip of a coin)

- $P(X = 1) = 1/2$ (read as "the probability that X equals one is one-half")
- Probabilities must sum to 1, so: $P(X = 0) = 1/2$

Discrete Random Variables: pdf

The **probability density function (pdf)** of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, j = 1, 2, ..., k$$

with $f(x) = 0$ for any x not equal to x_j.

The sum of all probabilities must be **1**:

$$\sum_{j=1}^{k} f(x_j) = 1$$

Discrete Random Variables: pdf

Example

- X is the **number of free throws** made by a basketball player out of two attempts.
- X can take the values $\{0, 1, 2\}$
- Assume the pdf of X is $f(0) = 0.2$, $f(1) = 0.44$, and $f(2) = 0.36$
- Q: What is the probability that a player misses both throws?
- Q: What is the probability that a player scores at least once?
- Q: What is the probability that a player scores three times?

Continuous Random Variables

A variable X is a **continuous random variable** if it takes any real value with zero probability.

"A continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers" Wooldridge, Appendix B

Examples:

- Commute time on a rainy day
- Max. temperature on a given day
- Unemployment rate
- . . .

Continuous Random Variables: pdf

The **probability density function (pdf)** of a continuous random variable is a continuous function.

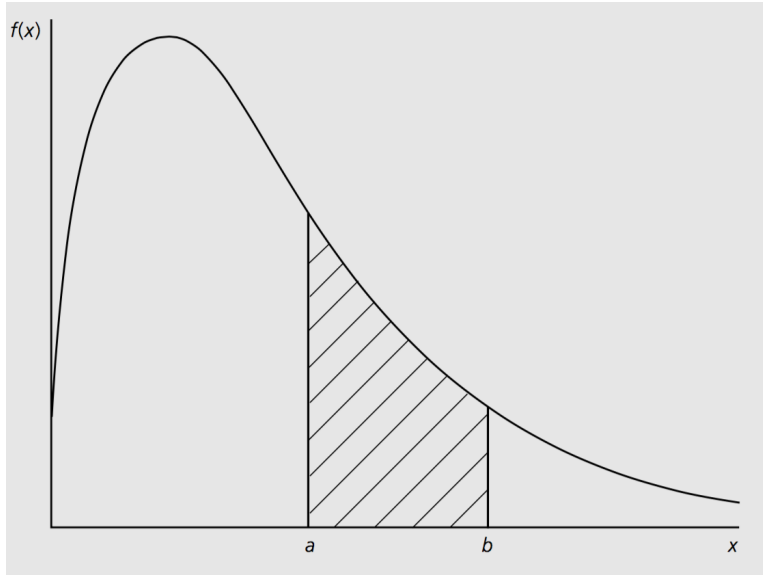
Because the probability of obtaining any real value is zero, we use the **pdf** to compute events that involve a range of values

Example:

- X is the **time between two buses** showing up in the same stop
- Q: What is the probability that the next bus will arrive in more than 5 min. but less than 10 min.?

Continuous Random Variables: pdf

$P(a < X < b)$ corresponds to the shaded area below:



The **integral** of $f(x)$ over all its support must be **1**:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Density Function (cdf)

The **cumulative density function (cdf)** describes the probability that a random variable takes a variable smaller than a given number:

$$F(a) = P(X \leq a)$$

Properties:

- $F(-\infty) = 0$
- $F(\infty) = 1$
- For any number c, $P(X > c) = 1 - F(c)$
- For any numbers a and b, $P(a < X \leq b) = F(b) - F(a)$

Joint Distributions

We are usually interested in the occurrence of events involving **more than one random variable**

- Example: What is the probability that it will rain tomorrow **and** it will be more than 10 degrees Celcius?

Let X and Y be discrete random variables. Then, (X, Y) have a **joint distribution**, which is fully described by the joint probability density function of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

Joint Distributions and Independence

Two random variables, X and Y are **independent**, if and only if:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

In the discrete case we have:

$$P(X = x, Y = y) = P(X = x) P(Y = y)$$

Joint Distributions and Independence

Example: Free Throw Shooting

- A basketball player is shooting **two free throws**
- X is a Bernoulli random variable equal to 1 if **first throw** is a success
- Y is a Bernoulli random variable equal to 1 if **second throw** is a success
- Assume the player succeeds in 80% of the free throws
  - $P(X = 1) = P(Y = 1) = 0.8$

- Q: What is the probability of the player making **both free throws**?
  - If X and Y are independent, then:

$$P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = 0.8^2 = 0.64$$

Note: if the throws are not independent then these calculations are not valid

Conditional Distributions

We are usually interested in **how one random variable is related to one or more other random variables**

The **conditional distribution** of Y given X tells us the distribution of Y conditional on us having information about X:

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

Discrete case:

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

Examples

- Height of a person given that we know their age
- Height of a person given that we know their age and gender
- Probability of blue eyes given that is Portuguese

Conditional Distributions

Example: Free Throw Shooting

- A basketball player is shooting **two free throws**
- $P(X = 1) = 0.8$
- Conditional density:
  - $f_{Y|X}(1|1) = 0.85$, $f_{Y|X}(0|1) = 0.15$
  - $f_{Y|X}(1|0) = 0.70$, $f_{Y|X}(0|0) = 0.30$

- Q: What is the probability of the player making **both free throws**?

$$P(X = 1, Y = 1) = P(Y = 1|X = 1)P(Y = 1) = 0.85 * 0.80 = 0.68$$

Features of Probability Distributions

We are usually interested in a **few features of the distributions** of random variables

The features of interest can be put in **three categories**:

- Measures of **central tendency**
- Measures of **variability or spread**
- Measures of **association between two random variables**

Central Tendency Measures: Expected Value

The **expected value** of a random variable X, **E(X)**, is a weighed average of all possible values of X.

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j)$$

Example:

- X can take the values -1, 0, 2 with probabilities 1/8, 1/2, and 3/8, respectively
- Q: What is the expected value of X?

$$E(X) = (-1) \cdot (1/8) + 0 \cdot (1/2) + 2 \cdot (3/8) = 5/8$$

Measures of Central Tendency: The Expected Value

If X is a continuous random variable, then $E(X)$ is defined as an **integral**:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Properties of the Expected Value

Given a random variable, X, and a function $g(\cdot)$, we can create a new random variable: $g(X)$

- Example: $g(X) = -X$

The expected value of a function of X is the **weighted average** of the function of X over the density of X:

$$E[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j)$$

Continuous case:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx$$

Properties of the Expected Value

Property E.1

For any constant c, $E(c) = c$

Property E.2

For any constants a and b, $E(aX + b) = aE(X) + b$

Property E.3

If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then:

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$$

Measures of Variability: Variance

The **variance** of a random variable X is a measure of its dispersion around the expected value, $\mu \equiv E(X)$:

$$\sigma^2 \equiv Var(X) \equiv E[(X - \mu)^2]$$

$$\sigma^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$$

Properties of the Variance

Property VAR.1

$Var(X) = 0$ if and only if X is a constant, i.e., there is a constant c such that $P(X = c) = 1$. In this case, E(X)=c.

Property VAR.2

For any constants a and b, Var(aX + b) = a^2 Var(X).

Note that b does not affect the variance. This means that adding a constant to a random variable does not change it's variance.

Measures of Variability: Standard Deviation

The **standard deviation** of a random variable, $sd(X)$, is simply the square root of the variance:

$$\sigma \equiv sd(X) \equiv \sqrt{Var(X)}$$

Property SD.1

For any constant c, $sd(c) = 0$

Property SD.2

For any constants a and b, sd(aX + b) = |a| sd(X).

Example: Standardizing a Random Variable

Let's **create a new random variable** from random variable X:

$$Z \equiv \frac{X - \mu}{\sigma}$$

Expected value of Z:

First rewrite Z as $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$

$$E(Z) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

Variance of Z:

$$Var(Z) = Var(\frac{1}{\sigma}X - \frac{\mu}{\sigma}) = \frac{1}{\sigma^2}Var(X) = \frac{\sigma^2}{\sigma^2} = 1$$

Measures of Association: Covariance

It is useful to have summary measures of **how two random variables vary with one another**

The **covariance** between two random variable X and Y, is defined as:

$$\sigma_{XY} \equiv Cov(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)]$$

If, $\sigma_{XY} > 0$, then, on average, when X is above its mean, Y is also above its mean

Properties of Covariance

Property COV.1:

If X and Y are independent, then $Cov(X, Y) = 0$

Property COV.2:

For any constants a_1, b_1, a_2, b_2,

$$Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X, Y)$$

Note: $Cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = Var(X)$

Variance of Sums of Random Variables

Property VAR.3

For constants a and b,

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Note: If X and Y are uncorrelated, i.e., $Cov(X, Y) = 0$, then:

$$Var(X + Y) = Var(X) + Var(Y)$$

and

$$Var(X - Y) = Var(X) + Var(Y)$$

Conditional Expectation

Usually, in social sciences we want to **explain a variable**, Y, **in terms of another variable**, say X.

One way to summarize this information is to calculate the **conditional expectation**, that for the discrete case is:

$$E(Y|X = x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x)$$

Example:

- How does the **expected hourly wage** ($WAGE$) change with the years of formal **education** ($EDUC$)?
  - What the meaning of $E(WAGE|EDUC = 12)$?
  - and $E(WAGE|EDUC = 16)$?

Conditional Expectation

Example:

Suppose that the expected value of *WAGE* given *EDUC* is the linear function:

$$E(WAGE|EDUC) = 1.05 + 0.45EDUC$$

- What is the expected hourly wage for a person with 9 years of education?
  - $E(WAGE|EDUC = 9) = 1.05 + 0.45 * 9 = 5.1$ Eur/Hour

Properties of Conditional Expectation

Property CE.1

$E[c(X)|X] = c(X)$ for any function $c(X)$

Property CE.2

For any functions $a(X)$ and $b(X)$,

$$E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X)$$

Property CE.3

If X and Y are independent, then $E(Y|X) = E(Y)$

Credit Today: My slides today based on slides by:

Miguel Godinho de Matos

Assistant Professor of Information Systems and Management at Católica-Lisbon,

PhD from Carnegie Mellon University

- Research: **Social network analysis** and **consumer behavior**

Rodrigo Belo

Assistant Professor at University of Rotterdam, Post-Doctoral Researcher at Carnegie Mellon University and at Católica-Lisbon.

PhD from Carnegie Mellon University

- Research Interests: **Social Networks** and **Technology on Educational Settings**