



THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Probability Review 1

Lecture 2

STA 371G

Probability Theory

The Concept of Probability

What is common among the following?

- Outcome of rolling a die
- S&P500 index at the end of January
- Number of iPhone 7s to be sold over the next year
- Number of unique visitors to Amazon.com over the next week
- Lifetime of your MacBook Air

We cannot predict any of these with certainty.

Yet, we can model them using **probability theory** and study the values they might take, associated probabilities etc.

Probability Theory

Definitions

Definition

A **random variable** expresses the outcome of a **random experiment** as a number. It is denoted by an uppercase letter.

- Random experiment \rightarrow Rolling a die
- Random variable $\rightarrow X$: The outcome when a die is rolled
- Random variable \rightarrow

$$Y : \begin{cases} 1, & \text{if outcome is odd number,} \\ 2, & \text{if outcome is even number} \end{cases}$$

Probability Theory

Definitions

Definition

A **discrete random variable** is a random variable with a finite (or countably infinite) range.

A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range.

Examples

- X : Number of stocks on NYSE whose price change today (discrete)
- Y : Average price change of the stocks on NYSE (continuous)

Probability Theory

Exercise

Discrete or continuous?

- Number of iPhone 7s to be sold over the next year
- Lifetime of your MacBook Air
- Number of unique visitors to Amazon.com over the next week
- S&P500 index at the end of January

Probability Theory

Definitions

Definition

Probability is the measure of the likelihood that a particular outcome (or set of outcomes) will be observed.

Probability is a number always between 0 and 1.

0 implies impossibility, 1 implies certainty.

Examples

- X : S&P500 at the end of 2017, $P(X > 2270) = 0.85$
- Y : Lifetime of your MacBook, $P(Y > 15 \text{ years}) = 0.05$

Probability Distributions

So, we have defined a random variable. How do we know what the probabilities are?

For example, what is the probability that your MacBook will break down after 5 years but before 7 years? That is, $P(5 < X < 7) = ?$

Definition

The **probability distribution** of a random variable X is a description of the probabilities associated with the possible values of X .

Discrete random variable \rightarrow Probability Mass Function (p.m.f.)

Continuous random variable \rightarrow Probability Density Function (p.d.f.)

Probability Distributions

Discrete Random Variables

Example

X : The outcome when you roll n -sided fair die.

Since this is a fair die, the corresponding probability mass function:

$$f(x) = \begin{cases} \frac{1}{n} & x = 1, \dots, n, \\ 0 & \text{otherwise.} \end{cases}$$

- $f(2) = P(X = 2)$, which is the probability of observing a “2.” This interpretation will not hold for continuous random variables.
- Sum of probabilities is always 1. ($n \times \frac{1}{n}$).
- This is an example of **Discrete Uniform Distribution**.

Probability Distributions

Continuous Random Variables

Example

Y : Lifetime of your MacBook (in years)

Let's assume Y has a **Continuous Uniform Distribution** with a maximum of 20 years. Its probability distribution is then given by the following probability density function:

$$f(y) = \begin{cases} \frac{1}{20} & 0 \leq y \leq 20, \\ 0 & \text{otherwise.} \end{cases}$$

What is $P(Y = 5) = ?$ or $P(Y = 5.5) = ?$ or $P(Y = 5.551234123) = ?$

They are all 0.

Probability Distributions

Continuous Random Variables

Warning!

For a continuous random variable, $P(Y = a)$ is always zero, regardless of a .

For this reason, for continuous random variables we ask questions like “ $P(a \leq Y \leq b) = ?$ ”

And we take integrals to find such probabilities.

Probability Distributions

Continuous Random Variables

Example

Y : Lifetime of your MacBook (in years)

$$f(y) = \begin{cases} \frac{1}{20} & 0 \leq y \leq 20, \\ 0 & \text{otherwise.} \end{cases}$$

What is $P(5 < Y < 7) = ?$

$$P(5 < Y < 7) = \int_5^7 \frac{1}{20} dy = \frac{y}{20} \Big|_5^7 = \frac{7}{20} - \frac{5}{20} = \frac{1}{10}$$

In general, $P(a \leq Y \leq b) = \int_a^b f(y) dy$.

Probability Distributions

Graphs Go Here

Mean, Variance and Standard Deviation

Definition

Mean or **Expected Value** of a random variable X is a measure of the center of its probability distribution. It is a weighted average of all possible values X can take, where the weights are the corresponding probabilities.

Discrete random variable X

$$\mu_X = E[X] = \sum_x xf(x)$$

Continuous random variable Y

$$\mu_Y = E[Y] = \int_y yf(y)dy$$

Mean, Variance and Standard Deviation

Definition

Variance of a random variable X is a measure of the dispersion, or variability in its distribution. **Standard Deviation** of X is the square root of its variance.

Discrete random variable X

$$\sigma_X^2 = \text{Var}(X) = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f(x)$$

Continuous random variable Y

$$\sigma_Y^2 = \text{Var}(Y) = E[(Y - \mu_Y)^2] = \int_y (y - \mu_Y)^2 f(y) dy$$

Law of Large Numbers

Definition

For a random variable X , the average of X_1, X_2, \dots, X_n gets very close to the expected value of X ($E[X]$) for large n .

Example

A die is rolled $n = 4$ times: $x_1 = 4, x_2 = 6, x_3 = 1, x_4 = 1$. The average is

$$\frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{4 + 6 + 1 + 1}{4} = 3$$

For large n , the average will be around 3.5; because $E[X] = 3.5$.

Law of Large Numbers

R Exercise

Go to R Studio...

```
# Generate a random number in [0,1]
runif(1)

# Generate a random number in [1,7]
runif(1, min=1, max=7)

# Floor it down to simulate a die
floor(runif(1, min=1, max=7))

# Simulate 3 dice
floor(runif(3, min=1, max=7))

# Take the average
mean(floor(runif(3, min=1, max=7)))

# Let's increase the number of dice
mean(floor(runif(10, min=1, max=7)))
```


Normal Distribution

A very common continuous probability distribution. Also known as the **bell curve**.

The mean and variance uniquely defines the distribution: $N(\mu, \sigma^2)$

The curve is symmetric around the mean.

Central Limit Theorem

Add numbers

Central Limit Theorem

R Exercise

X : The price of a house. Assume X is uniform in $[100, 400]$ (\$K).

Y : Average house price in a zip code

We will simulate z number of zip codes, each containing n houses.

```
# Simulating a zip code with 3 houses
runif(3, min=100, max=400)
# Repeat this for 5 zip codes.
house_prices <- t(replicate(5, runif(3, min=100, max=400) ))
# Find the average house price in each zip code
avg_house_prices <- rowMeans(house_prices)
# See what you got
hist(avg_house_prices)
# Increase n and z and try again!
```