Module 7
Random Variables
& Probability Distributions

Module Learning Outcomes

- Construct a probability distribution table with a given discrete random variable X.
- Calculate and interpret the expected value of a discrete random variable X.
- Calculate and interpret the standard deviation of a discrete random variable X.
- Verify a situation is a proper Binomial experiment by checking 4 conditions.
- Define the Binomial random variable, given the situation.
- Find Binomial probabilities using the Binomial equation and R.
- Calculate and provide an interpretation of the expected value and standard deviation of Binomial random variables.
- Verify the three Poisson conditions.
- Define the Poisson random variable, given the situation.
- MLO: Find Poisson probabilities using the Poisson equation and R.
- Make use of Z-scores to compare values from different group of subjects.
- Use the 68-95-99.7 Rule or Empirical rule to find basic percentages of a group of subjects.
- Perform (forward) Normal probability calculations using the Z-table (and R).
- Perform backward Normal probability calculations using the Z-table (and R).
- Find Exponential probabilities using the Exponential equation and R.

7.1 General Random Variables

- There are two types of random variables – **discrete random variables** and **continuous random variables**.
- The main difference is that the former can only have handful (or countable) possible values and the latter can have infinite number of possible values.
- Note: The proper definition of discrete random variables is different from above but it is outside of the scope of this course.

7.2 Variables vs. Random Variables

- All categorical variables can be considered discrete random variables (as long as each class of the categorical variable can be represented in or coded as numbers).
- However, not all numerical variables are continuous random variables.
- In fact, when responses of numerical variables are counts or frequencies, then it will be considered as discrete random variable. It is because responses in counts are “discrete” in nature.
- The rest of numerical variables are typically continuous random variables.
- Note: Do not stress over the definition here. It is rather obvious when you see an example.

7.3 Discrete Probability Distributions Table

- A **discrete probability distribution table** of a discrete random variable (X) has two main things: (1) all possible values of the random variable and (2) their corresponding probabilities (or $f(x)$).
- It generally has the following form:

X					
$f(x)$					

- MLO: Construct a probability distribution table with a given discrete random variable X.

7.4 Expected Value of Discrete Random Variables

- The **expected value** of a random variable X is like an average and it can be calculated by:

$$\mu = E(X) = \sum x * f(x)$$

- The symbol or notation μ is “mu” and it is a Greek alphabet and it is universally accepted as the mean of the population (or when dealing with probability).
- Another notation $E(X)$ is more descriptive. The “E” part stands for the “expected value” and “(X)” part means “of the random variable X”. But these two notations are equivalent.
- There are three key parts in the **interpretation** (or the meaning) of the expected value: 1) it is an average (cf. single value) of 2) of the random variable (cf. variable) from an infinite trials

<u>of the experiments</u> (cf. large number of subjects in the population).
Cross reference to the original definition of parameters (Modules 3 and 4).

- In short, it is usually presented as “average of the random variable (X), among infinite trials of the same experiment”. Or it is the “long run average”.
- MLO: Calculate and interpret the expected value of a discrete random variable X.

7.5 Standard Deviation of Discrete Random Variables

- The **standard deviation** of a random variable X has the following formula.

$$\sigma = SD(X) = \sqrt{\sum (x-\mu)^2 * f(x)}$$

- The symbol or notation σ is “sigma” and it is a Greek alphabet and it is universally accepted as the standard deviation of the population (or when dealing with probability).
- Another notation $SD(X)$ is more descriptive. The “SD” part stands for the “standard deviation” and “(X)” part means “of the random variable X”. But these two notations are equivalent.
- The standard deviation measures how far apart the values of the random variable are from the expected value. (Just like how far apart the observed values of a regular variable are from the mean.)
- The **interpretation** of standard deviation is very similar to expected value, except that the “average” part is replaced by “the typical difference from the expected value”.
- In other words, it is usually presented as the “typical difference of the random variable (X) from the expected value, among infinite trials of the same experiment”
- MLO: Calculate and interpret the standard deviation of a discrete random variable X.

7.6 Applications

- The expected value used here is synonymous with **Expected Monetary Value** (or EMV) in the business world when the random variable is about money.
- For example, the amount of return (X) is based on the outcome of an economy. If the economy is good, the amount of return is \$3 million, i.e. \$3 million gain. In contrast, if the economy is bad, the amount of return is negative \$2 million, i.e. \$2 million loss. The probabilities of the economy being good and bad are 0.6 and 0.4 respectively.
- After the calculation, the EMV of the amount of return is positive \$1 million, it means that the amount of return is expected to be \$1 million per investment, **on average**, <u>when the investment is invested infinite number of times</u>.
- Note #1: In reality, the investment cannot be repeated infinitely. Hence, the above number is only a very good approximation.
- Note #2: How good is the approximation? It is only as good as the estimation of the probability values (0.6 vs. 0.4) as well as the stated amount of return values (\$3 million vs. minus \$2 million).
- Note #3: A positive (or negative) EMV means there is an expected profit (or loss).

Binomial Random Variables and Probability Distributions

- Binomial random variables are a special case of the general discrete random variables.
- In short, you will see some with "finite" outcomes, probability distribution table, expected value and standard deviation.
- Note: Binomial probability distribution is used for probability calculation when there is one categorical variable.

7.7 Binomial Experiments

- To find Binomial probability, we need to verify if an experiment is indeed a **Binomial experiment**.
- To check if an experiment is a Binomial experiment, we have to check four conditions. Below is a table show the four conditions and their corresponding expectations.

	Conditions	**Expectation (To Do)**
1	The process or experiment can be broken down into n **identical stages** or **trials**.	a) Provide a description of the individual trial or stage. b) Identify the value of *n* – the number of trials.
2	In each trial, outcomes can be classified in to two **complementary** events and they are defined as **Success** and **Failure**.	Define in detail the Success of each trial. Note: Definition of Failure is not necessary because of their complementary relation.
3	All trials are **independent** to each other. In other words, the outcome from one trial will not affect the outcome of another trial, and vice versa.	a) Provide the meaning of independence, in the context. b) Check, in the context, if trials are indeed independent to each other.
4	In each trial, the probability of success is a **constant** and the **same** in all $\boldsymbol{n}$ trials.	Identify the value of $\boldsymbol{p}$, the probability of getting Success in each trial.

- If one or more of the above conditions are not met, then we cannot use the techniques below to find Binomial probabilities.
- Instead, you will have to use the basic probabilities (like conditional probabilities).
- MLO: Verify a situation is a proper Binomial experiment by checking 4 conditions.

7.8 Binomial Random Variables and Binomial Distributions

- A **Binomial random variable** in any Binomial experiment is defined as the number of Successes, with the sample of size n.
- Like the way we have with general random variable, we use the letter *X* here to represent the Binomial random variable, and *X* has a probability distribution (table) called the **Binomial probability distribution**.

X					
$f(x)$					

- Note: There are always $(n+1)$ possible outcomes for Binomial random variables.
- Note: The sum of all probabilities, or $f(x)$, is always one, except maybe due to rounding errors.
- MLO: Define the Binomial random variable, given the situation.

7.9 Binomial Probability Calculations

- The probability of a Binomial random variable is given by:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

  where $x = 0, 1, 2, \cdots, n$, $n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$, and $0! = 1$.
- On the right hand side of the equation, there are n, p, and x. The first two items are directly from the Binomial Experiment verification. And the last one will be directly from questions.
- Besides equation, we could also find Binomial probabilities using R.
- There are two main ones: dbinom(x, n, p) and pbinom(x, n, p, cumulative).
- MLO: Find Binomial probabilities using the Binomial equation and R.

7.10 Binomial Expected Value and Standard Deviation

- Once we have verified that X is a Binomial random variable from a proper Binomial experiment, its expected value and standard deviation are defined respectively:

$$\mu = E(X) = n \times p \qquad \sigma = SD(X) = \sqrt{n \times p \times (1-p)}$$

- Please note that the interpretation of them is the same as before.
- MLO: Calculate and provide an interpretation of the expected value and standard deviation of Binomial random variables.

Poisson Random Variables and Probability Distributions

- Poisson random variables are also a special case of the general discrete random variables.
- Similarities: Loosing speaking, a Poisson random variable also has two complementary outcomes: Success and Failure.
- Differences: A Binomial variable can have $n+1$ possible values. In contrast, a Poisson random variable can have infinite possible values.
- Note: In practice, there are only "countably infinite" values.

7.11 Poisson Random Variable and Assumptions

- The Poisson Random variable (X) is defined as the number of Success within a unit.
- The "unit" could be 1) a period of time, 2) a well-defined space, or 3) both time and space.
- There are three main Poisson assumptions or conditions.
- 1) The number of Successes happens in one unit is independent of the number of Success happening in another unit.
- 2) The probability of Successes is a constant in all units of the same size. And it grows with the same proportions as the unit size. In other words, then the unit goes double in size, the probability of Success will be doubled.
- 3) Successes are assumed to occur one at a time.
- MLO: Verify the three Poisson conditions.
- MLO: Define the Poisson random variable, given the situation.

7.12 Poisson Probability Calculations

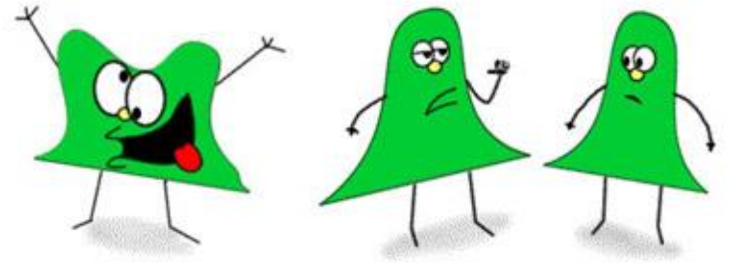
- The probability of a Poisson random variable is given by:

$$f(x) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad or \; f(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

  where $x = 0, 1, \ldots, n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$, and $0! = 1$.
- Either μ or λ are called the **mean parameter**.
- Besides equation, we could also find Poisson probabilities using R.
- There are two main ones: dpois(x, μ) and ppois(x, μ, cumulative).
- MLO: Find Poisson probabilities using the Poisson equation and R.

7.13 Poisson Approximation to Binomial

- When n is large and np is reasonably small like smaller than 7 (rather arbitrary), the Poisson probability is very close to the Binomial probability.
- It would be useful when you have no access to software, especially when n is large, the term $n!$ will be huge and the calculation will be smaller.



"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

<u>Normal Random Variables and Probability Distributions</u>

- Normal random variables are one special case, but an important one, of the continuous random variable family.
- By definition, there are infinite number of outcomes, where it will never happen in reality.
- In other words, the "definition" in reality is "countably infinite" number of outcomes.
- Normal distributions are so versatile but it is mainly used for <u>one numerical variable</u> situation.
- It is, however, being used in multi-dimensional situations later in the program.
- The probability of a Normal random variable (X) is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

7.14 Normal Shape or Bell Shape

- In Module #3, we learned that **Normal shape** (or **Normal distribution)** was used to describe histograms. Here we will elaborate more about this important shape.
- The Normal shape is by far the most popular or well-known shape/distribution and we know a great deal about this rich family of distributions.
- All family members look about the same (i.e. all has Normal shape or a Normal-shape mountain), but they differ in two key components.
- 1) **Centre** (represented by the **mean** and it is denoted as $\boldsymbol{\mu}$): it represents the "middle" of the base of the mountain.
- 2) **Spread** (represented by the **standard deviation** and it is denoted as $\boldsymbol{\sigma}$): it represents how big or long the "range" (the English word range that refers to the mountain) of the base of the mountain.
- Once we know about the mean and standard deviation, together with the shape, we can do calculations about it.
- One important note we need to recall: just as the sum of all bars in a histogram is always one (or 100%), the <u>area underneath the Normal curve is also one</u>.

7.15 Z-Score

- Before doing any Normal calculations, we first need to understand **Z-scores**.
- Z-scores are also called the **Standard Normal Scores** and it is defined as:

$$Z = \frac{Value - Mean}{SD} = \frac{X - \mu}{\sigma}$$

- This process is called the **standardization**, or the X-value is being **standardized**.
- Note: The letter "Z" is a reserved letter in statistics.
- Z-scores tell us two main things: (1) the location in relation to the mean (positive value means on the right side of the mean), and (2) the standard distance in relation to the mean (bigger magnitude means farther away from the mean).
- Please note that Z-scores do not have units, and that is the reason why it is used to compare values from two different group of subjects.
- The probability of the Standard Normal random variable (Z) is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

- MLO: Make use of Z-scores to compare values from different group of subjects.

7.16 Z-Table

- The Z-table contains 2 kinds of entries: 1) the **cumulative area** (it refers to the area to the left side of the given Z-score), and 2) the Z-score.
- Areas are all given in decimal forms (four decimal places). Make sure you are able to convert decimals to percentage, and vice versa.
  For example, an entry in the Z-table has 0.8869 or 88.69%. A 0.5% cumulative area means the left-hand area is 0.0050.
- To find areas using Z-table, the second decimal point of the Z-score is used to determine which column to use and the remaining part determines which row to use in the Z-table, i.e.

$$Z = 1.65 = 1.6|5 = \underbrace{1.6}_{row} \mid \underbrace{5}_{column}$$

- Note: When the given Z-score cannot be found in the table because it is either too small (or too big), it is always wise to sketch the diagram and shade the area of interest. In this case, when the area is too small (or it is covering the whole area), then the answer should be close to zero (or close to one).

7.17 General Concept in Normal Calculations

- Once we know that a numerical variable follows a Normal shape and has the values of mean and standard deviation, we are able to do Normal probability calculations.
- There are two kinds (or directions) of Normal calculations: Forward vs. Backward.

- In the **forward Normal calculations**, you are given a value of X-variable and asked to find an answer in the form of percentage (or probability).
- In the **backward Normal calculations**, you are given a value in percentage and asked to find answers that are about the X-variable.

7.18 Forward Normal Calculations

- Suppose a numerical variable follows a Normal distribution with an average of μ and a standard deviation of σ, then we can find the area by doing the following two steps.
- (1) Convert a given X-value to a Z-score by $Z = \frac{X-\mu}{\sigma}$.
- (2) Look up the Z-table for the corresponding area.
- The schematic looks like:

Given a value of X	Step 1	Z-score	Step 2	Answer
X	$\xrightarrow[using\ Z=\frac{X-\mu}{\sigma}]{}$	Z	$\xrightarrow[using\ Z-table]{}$	Area

- Alternately, we could use Excel function "NORM.DIST()" to compute Normal probabilities.
- In R, you can use pnorm(x, μ, σ). Note the default always refers to the left-hand area.
- MLO: Perform (forward) Normal probability calculations using the Z-table (and R).

7.19 Backward Normal Probability Calculations

- Sometimes, an area is provided, and we would like to find some value of the numerical variable.
- This is essentially a reverse process of the previous method.
- (1) Use the area provided to get the corresponding Z-score in the Z-table.

Note: Make sure the proportion or area used to find the Z-score is a cumulative area.

- (2) Convert the Z-score back to a value of the numerical variable by $X = \mu + Z \times \sigma$.
- The schematic looks like:

Given an area	Step 1	Z-score	Step 2	Value of X
Area	$\xrightarrow[using\ Z-table]{}$	Z	$\xrightarrow[using\ X=\mu+Z\times\sigma]{}$	X

- Alternately, we could use Excel function "NORM.INV()" to compute find X.
- In R, you can use qnorm(cumulative probability, μ, σ).
- MLO: Perform backward Normal probability calculations using the Z-table (and R).

Exponential Random Variables and Probability Distributions

- Exponential random variables are another special case of the continuous random variable family.
- Exponential random variables are used to model the time between two events (or two Successes).
- They are highly related to Poisson random variables in Queuing Theory.

7.20 Exponential Probability Calculations

- The probability of a Poisson random variable is given by:

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \; or \; f(x) = \lambda e^{-\lambda x}$$

  where $X \, \epsilon \, \mathcal{R}^{+}$ (i.e. all positive real number), μ is called the **mean parameter** while λ is the **rate parameter**.
- It represents "how many times something occurs in a period of time" vs. "at what rate something occurs in a period of time.
- In fact, one is a reciprocal of another.
- Example: If the average number of emails I get per hour is 5 (i.e. $\mu = 5$), then the rate at which I am receiving emails is 12 minutes (or 1/5 of an hour) per email (i.e. $\lambda = 1/5$).
- Unlike Normal distribution, which is impossible to find the area under the function using integration (or Calculus), it is much easier with Exponential functions.
- The **right-hand area** of an Exponential random variable (X) is given by:

$$P(X > k) = \int_{k}^{\infty} \frac{1}{\mu} e^{-\frac{x}{\mu}} dx = e^{-\frac{k}{\mu}}$$

- One interesting property: Exponential distribution is also called the "memoryless function". The lifetime of a lightbulb typically follows an Exponential distribution. The remaining lifetime of a lightbulb is the same at time = 0 and at time = 100 hours.
- In R, you can use pexp(x, λ). Note the parameter is R is always the rate parameter (λ) and the default always refers to the left-hand area.
- MLO: Find Exponential probabilities using the Exponential equation and R.