Lecture 10

Paul Holaway, Abhi Thanvi

July 7th, 2022

Lecture 9 Review

Before we move on to the new material, we will do a quick review of Lecture 9 content. Last time we finished up learning about custom functions and a new method for creating samples. We will do both together in one quick example before moving onto today's lecture material. If you do not remember the specific syntax to use for custom functions and the sampling procedure, refer back to Lecture 8 and 9 notes.

Example 1; Custom Functions and Sampling Review

Let's suppose you have been hired by an independent professional baseball team to do a study on their players performance after the first month. They want you to do an individual study on each player, but in a random order. IE) A random sample containing every player. Remove the player with just 1 AB.

```
baseball <- read.csv("~/Classes/DPISu22/Data Sets/baseball.csv", stringsAsFactors=TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
baseball = baseball %>% filter(GP > 1) #Removing the player with 1 AB.
```

```
PlayerSample = function(data){
  x = nrow(data)
  num = sample(1:x, x, replace = FALSE)
  data = data[num,]
  return(data)
}
```

```
set.seed(143572)
random = PlayerSample(baseball)
random[,2:6] #Printing out 5 columns to save space.
```

```
##    Number GP PA AB  H
## 4       9 29 98 74 22
## 2       3 29 96 73 29
## 12     32 29 91 71 18
## 7      22 27 35 25  2
## 11     29 25 62 54 13
## 8      23 27 76 63 20
## 1       1 29 93 71 20
## 10     27 25 68 55 15
## 5      17 21 39 22  4
## 9      24 27 68 53  9
## 3       4 29 84 77 30
## 6      21 29 77 56 17
```

One thing to note above is the use of the `nrow()` function. This function calculates the number of rows in a data set. Fun Fact: The Number of Rows = Length of a Data Set (`length()`). The reason for using `nrow()` here is because we do not specify a column from the data set in our custom function. While it is possible, it is beyond the programming scope of this course. With `length()` you need to specify a column for a data set, while `nrow()` you do not specify a column. You may use whichever one you prefer as the two methods when done correctly are interchangeable. I personally prefer to use `length()`, but you may use either. Okay, now we can move onto the next portion of lecture content.

Distributions

Today we will be learning one of the most important topics in introductory statistics, probability distributions. A probability distribution (simply referred to as a distribution) is a statistical function that describes all the possible values and likelihoods that a random variable (#NextTime) can take within a given range. There are a ***LOT*** of different distributions, so we will just focus on three to introduce you to the topic. (Technically just two as one is a special case of another.) There are two different types of distributions, discrete and continuous. Discrete distributions can only take on values that are finite (most commonly integers) while continuous distributions can take on any value within a given range. Below is a list of the most commonly used distributions. The ones in italics will be the three we will cover today. You may look at the others in your spare time if you wish.

- **Discrete Distributions**
  - *Bernoulli* (Special Case of Binomial)
  - *Binomial*
  - Geometric
  - Hypergeometric
  - Negative Binomial
  - Poisson
  - Uniform (Discrete Form)
- **Continuous Distributions**
  - *Normal*
  - Beta
  - Chi-Square
  - Exponential
  - Gamma
  - Uniform (Continuous Form)

Likelihoods (probabilities) can be calculated using a "Probability Mass Function" (PMF) if the distribution is discrete or a Probability Density Function" (PDF) if the distribution is continuous.

Bernoulli Distribution

The Bernoulli distribution will be the first of three distributions that we talk about today. It is a special case of the Binomial distribution (we will show that later). The Bernoulli distribution models a *single* outcome in an experiment that asks a yes or no question. In the notation of the distribution, yes = 1 and no = 0. Its PMF is . . .

$$f(x) = p^x(1-p)^{1-x}; x = 0, 1$$

where p is the probability of success. This is the simplest distribution of all. For something to follow a Bernoulli distribution it must have only two outcomes and only one trial.

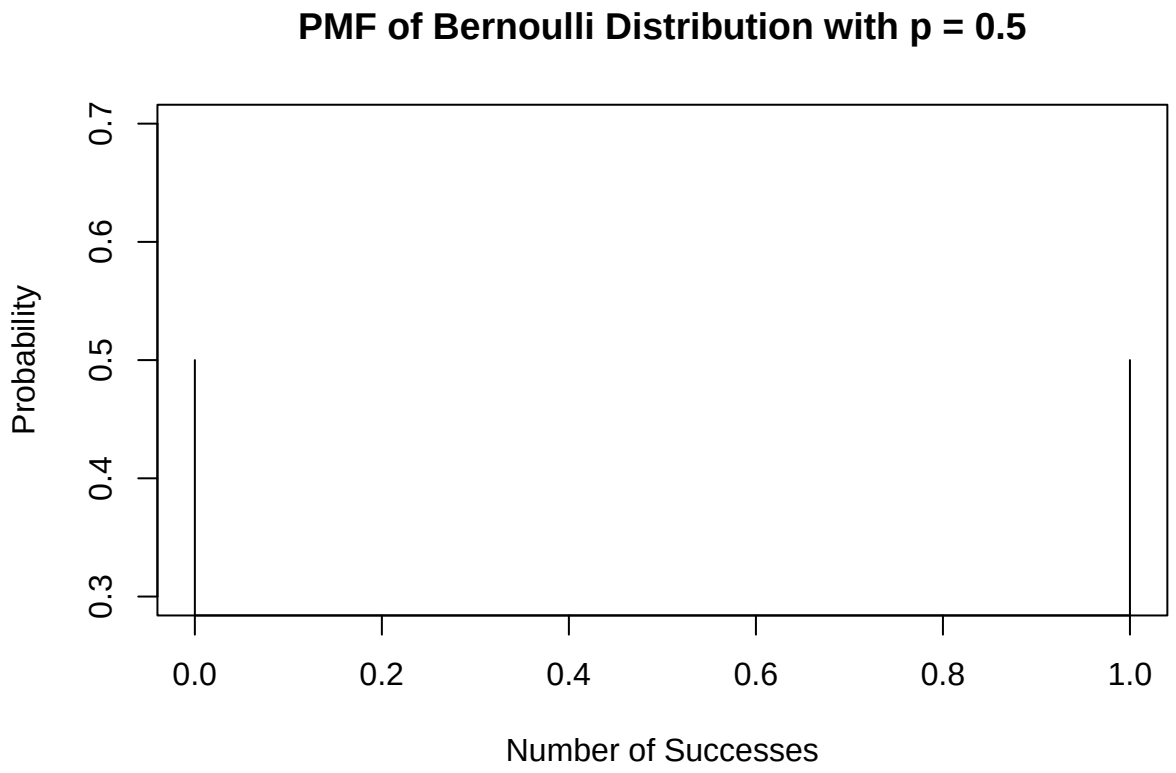
Example 2; Flipping One Coin

Flipping a coin one time is a classic Bernoulli example. You only have two outcomes (head or tail) and since you do one flip, that is one trial. For a fair coin, the probability of flipping a head or tail is the same, one-half. So. . .

$$P(H) = 0.5^1(1-0.5)^{1-1} = 0.5$$

Let me plot this for you. You do not have to know how to do this. It is more advanced.

```
plot(0:1,dbinom(0:1,size = 1, prob = 0.5), type = "h", xlab = "Number of Successes",
     ylab = "Probability", main = "PMF of Bernoulli Distribution with p = 0.5")
```

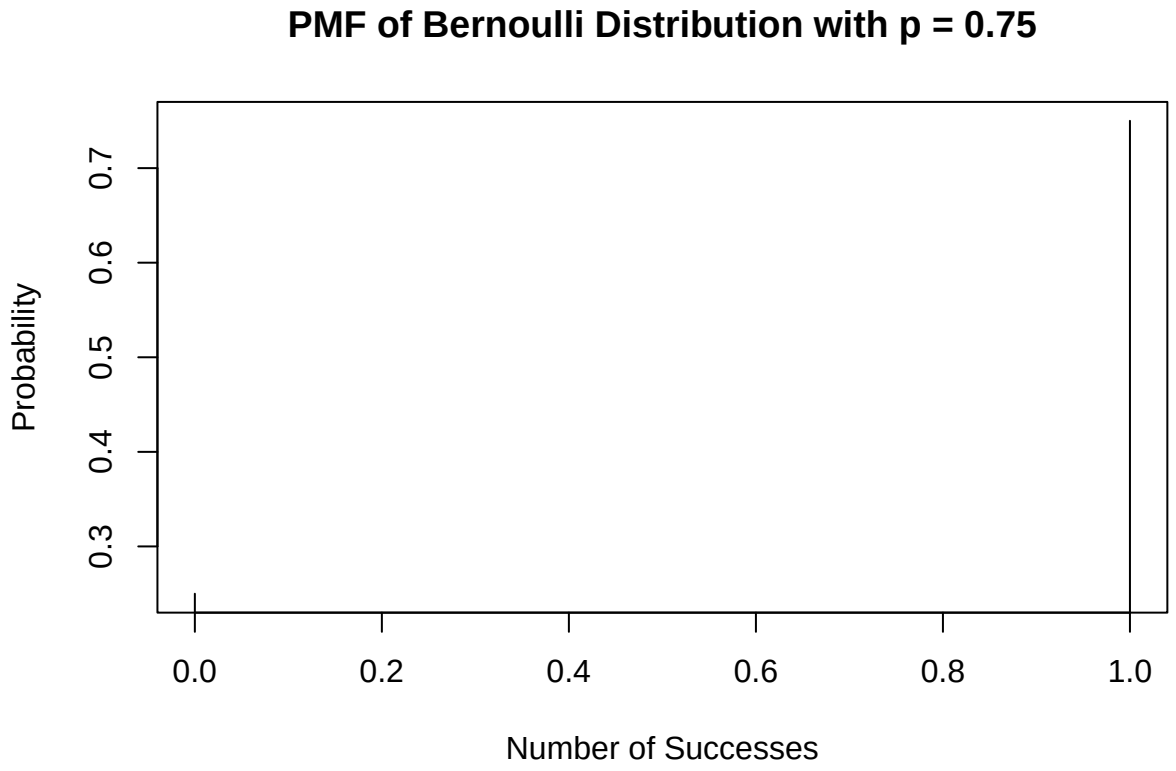


What if we did not have a fair coin and the probability of heads $p = 0.75$.

$$P(H) = 0.75^1(1-0.75)^{1-1} = 0.75$$

Again here is the PMF.

```r
plot(0:1,dbinom(0:1,size = 1, prob = 0.75), type = "h", xlab = "Number of Successes",
     ylab = "Probability", main = "PMF of Bernoulli Distribution with p = 0.75")
```



Okay, that is pretty simple, so let's move onto the next distribution.

Binomial Distribution

The Binomial distribution is the more general case of the Bernoulli distribution. Instead of only having one trial, we have n trials. For something to follow a Binomial distributions it must have only two outcomes, each trial is independent, a fixed number of trials (n), and a probability of success (p). Its PMF is ...

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}; x = 0, 1, 2, \ldots, n$$

Whenever you see a !, that in mathematics is called the factorial sign. The formula is below.

$$n! = 1 * 2 * 3 * 4 * \ldots * n$$

Example 3; Flipping More Than One Coin

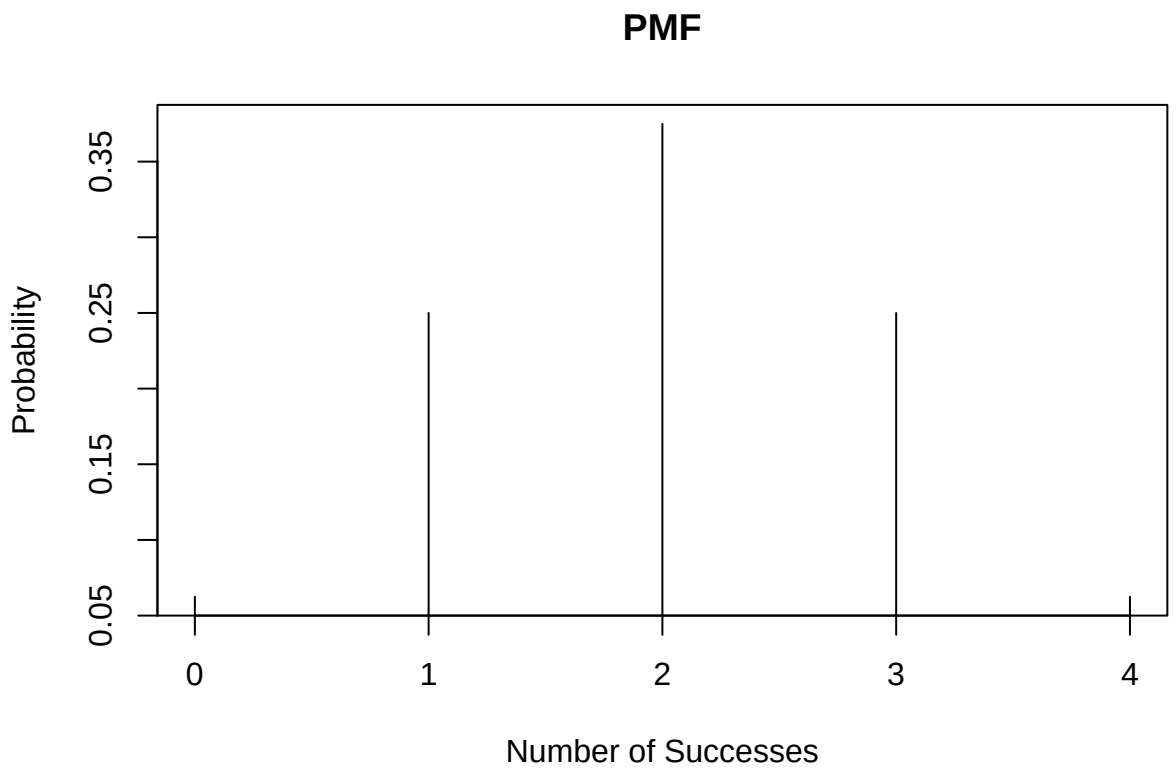
You are out with your friend for the day. Both of you have different things you want to do, so you decide to flip a fair coin to decide who chooses what you do. If you can do four things, what is the probability that you will choose any n number of things to do?

$$P(X = 0) = \frac{4!}{0! * 4!}0.5^0(1-0.5)^{4-0} = 0.0625$$

$$P(X=1) = \frac{4!}{1!*3!}0.5^1(1-0.5)^{4-1} = 0.25$$
$$P(X=2) = \frac{4!}{2!*2!}0.5^2(1-0.5)^{4-2} = 0.375$$
$$P(X=3) = \frac{4!}{3!*1!}0.5^3(1-0.5)^{4-3} = 0.25$$
$$P(X=4) = \frac{4!}{4!*0!}0.5^4(1-0.5)^{4-4} = 0.0625$$

As you can see, all the probabilities add up to one (as they should). Here is the PMF.

```
plot(0:4,dbinom(0:4,size = 4, prob = 0.5), type = "h", xlab = "Number of Successes",
     ylab = "Probability", main = "PMF")
```



Let's do one more example before moving onto normal.

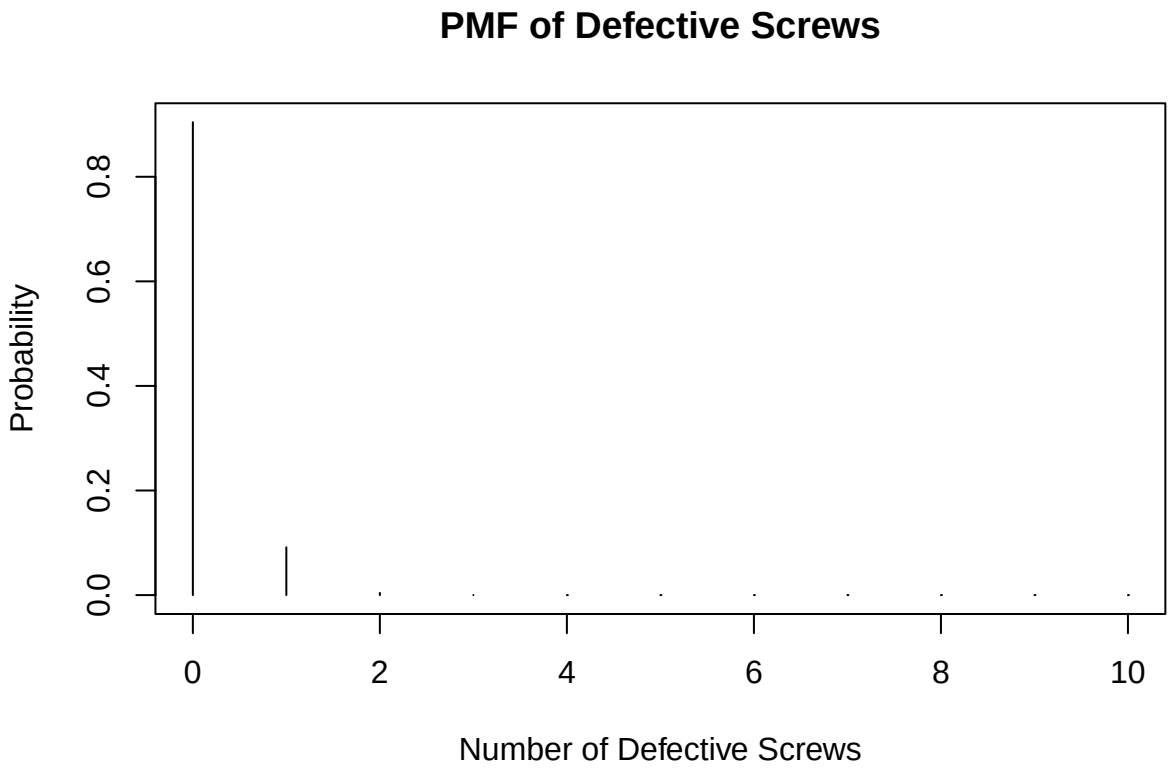
Example 4; Production Effeciency

It is known that screws produced by a certain company will be defective with probability 0.01, independently of one another. The company sells the screws in packages of 10 and offers a money-back guarantee that at most 1 of the 10 screws is defective. What proportion of packages sold must the company replace? Hint: Since we want to know if a screw is defective, then we will count a defective screw as a "success".

$$P(Refund) = P(X \geq 2) = 1 - P(X=0) - P(X=1) = 1 - \frac{10!}{0!*10!}0.01^0(1-0.01)^{10-0} - \frac{10!}{1!*9!}0.01^1(1-0.01)^{10-1} \approx 0.004$$

Here is the PMF for this problem.

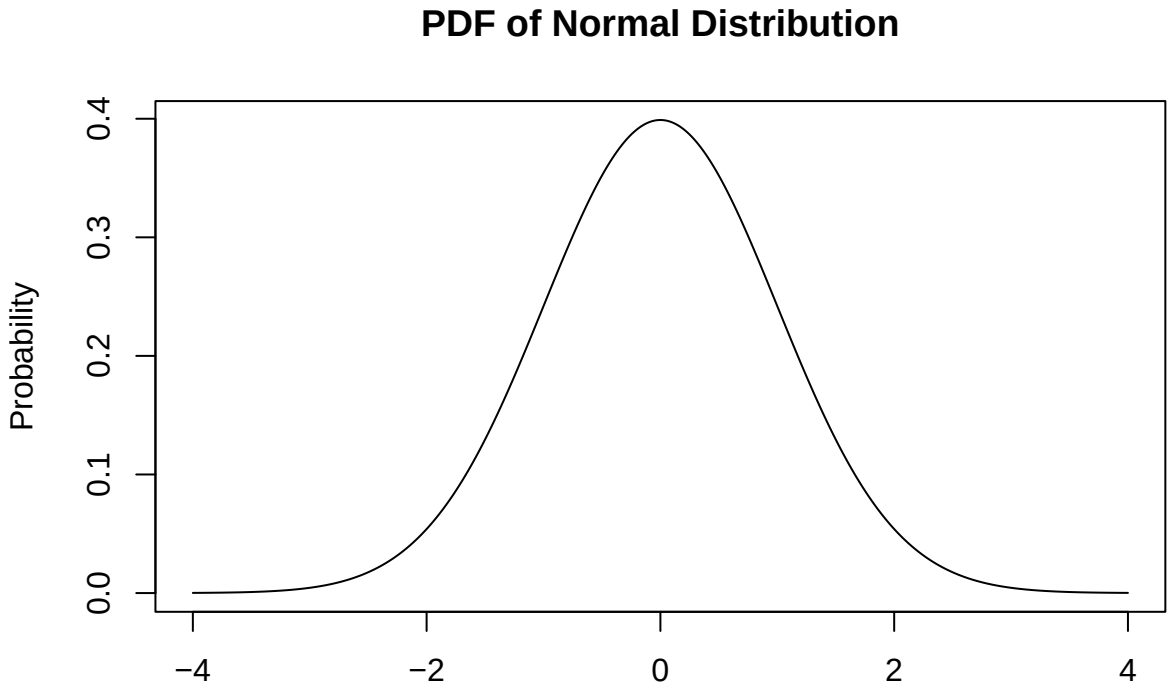
```r
plot(0:10,dbinom(0:10,size = 10, prob = 0.01), type = "h",
     xlab = "Number of Defective Screws", ylab = "Probability",
     main = "PMF of Defective Screws")
```



Normal Distribution

The Normal distribution is the most famous in statistics and one of the most applicable in real world applications. The PDF for the normal distribution is below.

```r
plot(seq(-4,4,0.001),dnorm(seq(-4,4,0.001),mean = 0,sd = 1), type = "l", xlab = "",
     ylab = "Probability", main = "PDF of Normal Distribution")
```



The normal distribution is also referred to as the bell curve by many due to its bell shape. One thing to note is that this PDF above is of a standard normal distribution. A standard normal distribution has a mean of 0 and a standard deviation of 1. In general, a normal distribution does not have to have the mean be 0 and the standard deviation be 1. The PDF for a Normal distribution is ...

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; -\infty < x < \infty$$

Where μ is the mean and σ is the standard deviation. Now calculating probabilities here gets more tricky because continuous variables can take all values within their domains. So you cannot use the tricks from discrete cases. This is when we need to employ an external tool, the **Z-table**. The Z-table helps us calculate normal probabilities because it has (almost) all of them calculated for us. One thing to note is that the Z-table by default reads as $P(Z < x)$, so if you want to find $P(Z > x)$, you would need to do $1 - P(Z < x)$. Let's do a quick example.

Example 5; Reading a Z-Table

Let's suppose that you have some case where $\mu = 0$ and $\sigma^2 = 1$. You want to calculate...

- $P(Z < 0.5)$
- $P(Z > 0.85)$
- $P(0.5 < Z < 0.85)$

Let's use the Z-table. For the first part, you scroll down until you find 0.5 on the left-hand side of the table, and move over to the column .00 because it is just 0.5 we want to look at.

$$P(Z < 0.5) \approx 0.6915$$

Pretty easy, now for the second one. It will work the same way, you just need to do the 1 - part.

$$P(Z > 0.85) = 1 - P(Z < 0.85) \approx 1 - 0.8023 \approx 0.1977$$

Now what about the third one? This is a bit more difficult. However, here is how you can think about it. You know how to find $P(Z > 0.5)$ and $P(Z < 0.85)$. You just want the area in between them, this leads us to the following conclusion...

$$P(a < X < b) = P(X < b) - P(X < a)$$

You are first finding the probability of X less than b, and then subtracting the overlap from X less than a.

$$P(0.5 < Z < 0.85) = P(Z < 0.85) - P(Z < 0.5) \approx 0.8023 - 0.6915 \approx 0.1108$$

Okay, but what if you do not have a mean of 0 and a standard deviation of 1 (most common)? It turns out there is a remedy for this. We can use what is called a Z-score. If our normal distribution is not standard normal, we can calculate the Z-score to find out where that value would be if the distribution was standard normal. The formula to calculate the Z-score is ...

$$Z = \frac{x - \mu}{\sigma}$$

Let's try out an example.

Example 6; Oil Profits

An oil company on average makes $\$40,000$ a day with a standard deviation of $\$10,000$. The owner wants to know the following ...

- $P(X < 30,000)$; The probability that profits will be less than $\$30,000$.
- $P(X > 60,000)$; The probability that profits will be more than $\$60,000$.
- $P(30,000 < X < 60,000)$; The probability that profits will be between $\$30,000$ and $\$60,000$.

First, we need to find the Z-score.

$$Z = \frac{30000 - 40000}{10000} \implies Z = -1.00$$

$$P(X < 30,000) = P(Z < -1.00) \approx 0.1587$$

The second part will run the same.

$$Z = \frac{60000 - 40000}{10000} \implies Z = 2.00$$

$$P(X > 60,000) = P(Z > 2.00) = 1 - P(Z < 2.00) \approx 1 - 0.9772 \approx 0.0228$$

The third part will use the same formula when we have both less than and greater than symbol as before. We would need to find *two* Z-scores here. However, since we already have them, we can just plug them in.

$$P(30,000 < X < 60,000) = P(X < 60,000) - P(X < 30,000) = P(Z < 2.00) - P(Z < -1.00) \approx 0.9772 - 0.1587 \approx 0.8185$$

That is all we will cover for now. Next time we will go into more statistical concepts and do a bit of a deeper dive into these distributions.

End of Lecture 10 Notes