# Lecture 11

Paul Holaway, Abhi Thanvi

July 11th, 2022

## Lecture 10 Review

Before we move on to the new material, we will do a quick review of Lecture 10 content. Last time we learned about three different distributions, Bernoulli, Binomial, and Normal. We then learned how to calculate probabilities using them. We will do a few quick examples today before moving on.

### Example 1; Distribution Review

Hogan is a prisoner of war in a WWII prisoner of war camp. Unfortunately for Hogan, the Kommandant and guards are the best, and the probability of escape is 0.005 (not great for Hogan). Hogan stays and helps other prisoners escape every four months because that is how long it takes to dig a new escape tunnel.

**Question 1:**

If it is January 1942 now (the next escape attempt is scheduled for April) what is the probability six of the escape attempts will be successful? (WWII ended on September 2, 1945)

$$P(Escape = 6) = \frac{11!}{6! * 5!} 0.005^6 (1 - 0.005)^{11-6} \approx 7.04 * 10^{-12}$$

Wow, that's *really* bad.

**Question 2:**

If it is January 1942 now (the next escape attempt is scheduled for April) what is the probability at least one of the escape attempts will be successful? (WWII ended on September 2, 1945)

$$P(Escape \geq 1) = 1 - P(Escape = 0) = 1 - \frac{11!}{0! * 11!} 0.005^0 (1 - 0.005)^{11-0} \approx 0.0536$$

**Question 3:** The Kommandant is trying to figure out when the next escape will happen to maximize the chances of him stopping it. He knows it is every fourth month, but the specific time is never the same. The average number of days into the fourth month when the escape attempt happens is 16 with a standard deviation of 4 days. The Kommandant is going to be out of camp (Hogan does not know this) the 12th through 15th. What is the probability that the escape will happen while he is gone?

$$Z = \frac{15 - 16}{4} = -0.25$$

$$Z = \frac{12 - 16}{4} = -1.00$$

$P(12 < Escape < 15) = P(Escape < 15) - P(Escape < 12) = P(Z < -0.25) - P(Z < -1.00) \approx 0.4013 - 0.1587 \approx 0.2426$

Okay, now we can move onto the next portion of lecture content.

# Central Limit Theorem and Random Variables

Today we are going to learn one of the most important concepts in statistics, the Central Limit Theorem. If you ever take any other statistics courses, you will either talk about it, or use it in either a homework or project. It is so important that almost any general statistics textbook will have a section on it. After that, we will go into random variables and some things about them.

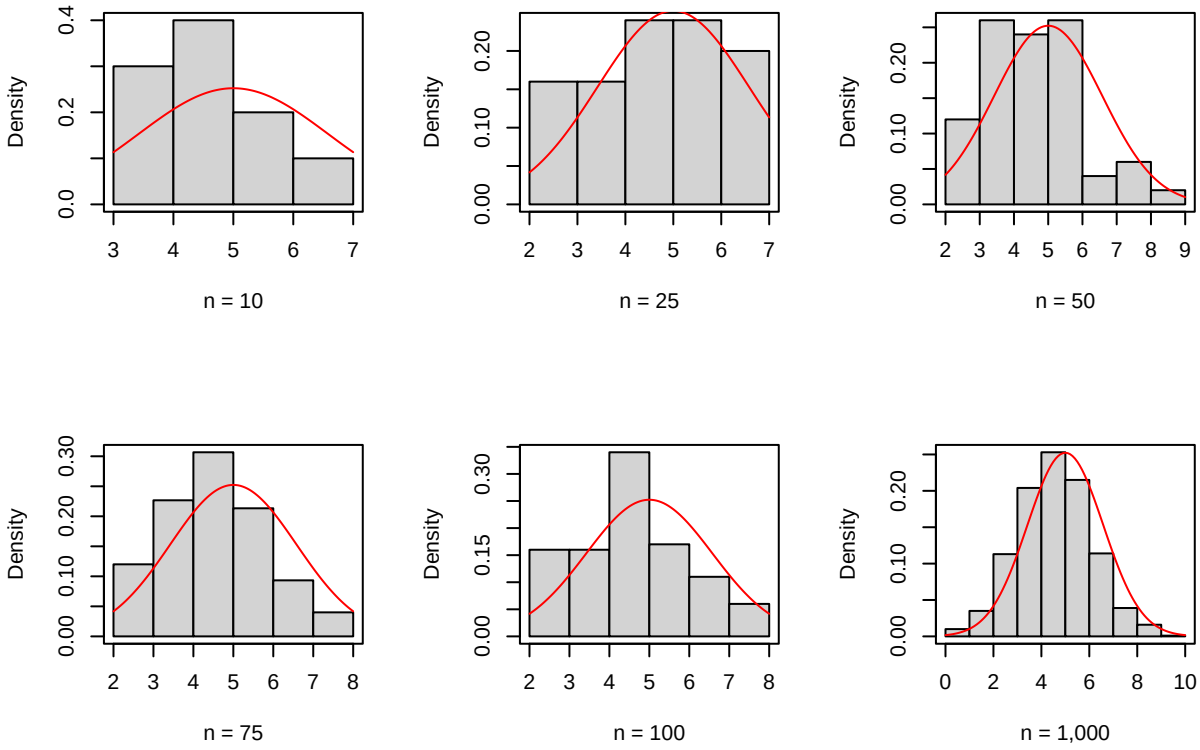
## Central Limit Theorem

The central limit theorem is one of the most remarkable and important theorems in probability and statistics. It states that the sum of a large number of independent random variables has a distribution that is approximately normal. In more general terms, the larger number of observations you have, the closer the distribution gets to becoming approximately normal. French mathematician Abraham DeMoivre noticed that when $n$ was large, his binomial distribution looked very close to a normal one. He then started to use the normal distribution to approximate binomial probabilities when $n$ was large. He found this to be close enough to what it actually would be. This result was later extended on by Laplace and others, which then later became the Central Limit Theorem (CLT). We will not be asking you questions that directly use CLT because they are far beyond the scope of this course. Instead, you will be indirectly using CLT in your work. However, I will illustrate (in a condensed example) what DeMoivre saw.

### Example 2; CLT at Work

We will look at a binomial distribution using $p = 0.5$, but will increase $n$ as we go along. You do **NOT** need to know how this code works. This is just an illustrative example. The normal curve is added in red.

```
set.seed(314439)
par(mfrow=c(2,3))
hist(rbinom(10,size = 10,prob = 0.5), xlab = "n = 10", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
hist(rbinom(25,size = 10,prob = 0.5), xlab = "n = 25", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
hist(rbinom(50,size = 10,prob = 0.5), xlab = "n = 50", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
hist(rbinom(75,size = 10,prob = 0.5), xlab = "n = 75", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
hist(rbinom(100,size = 10,prob = 0.5), xlab = "n = 100", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
hist(rbinom(1000,size = 10,prob = 0.5), xlab = "n = 1,000", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 5, sd = sqrt(2.5)), add = TRUE, col = "red")
```



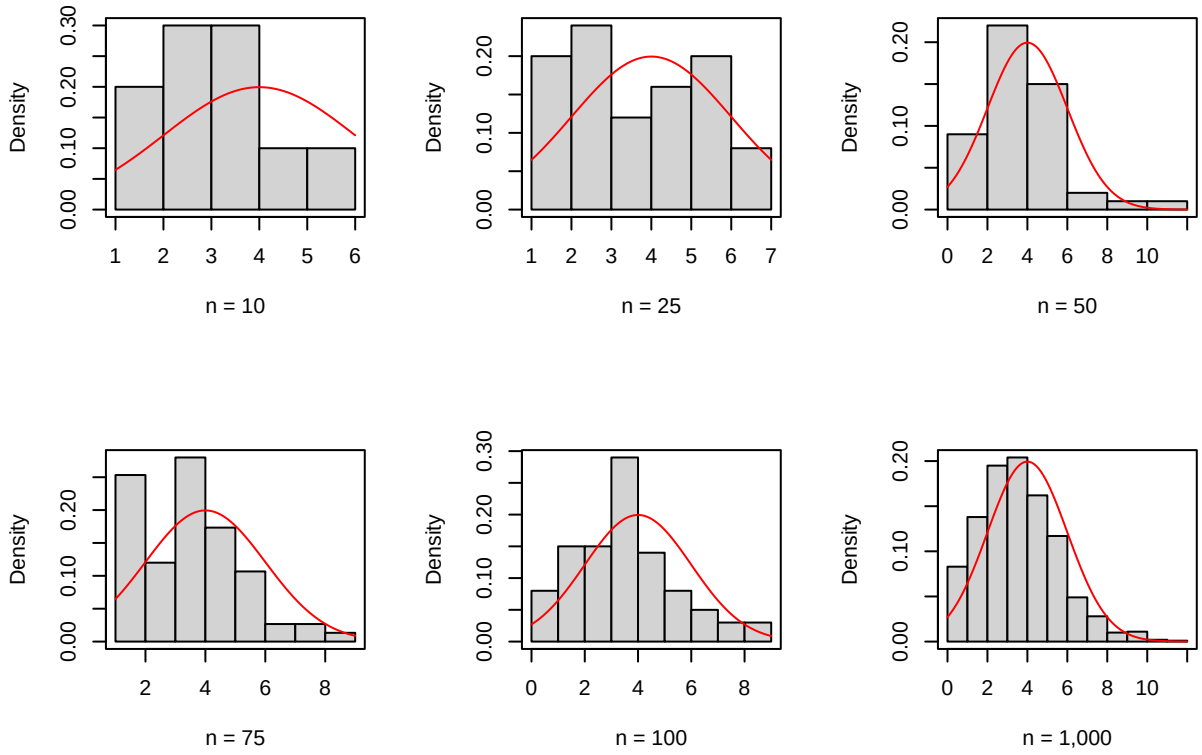

As you can see, as you increase the value of $n$ (the number of observations) the binomial distribution looks to be more and more normal. You can now see where DeMoivre is coming from. However, the next thought is, "Does this work on other distributions?" The answer is yes.

**Example 3; CLT at Work, Again**

Here, we will look at another distribution, the Poisson distribution. It is one of the other discrete distributions that has many applications. Again, this example is for illustrative purposes only, so if you do not fully understand the Poisson distribution or the code, that is perfectly fine.

```
set.seed(314439)
par(mfrow=c(2,3))
hist(rpois(10,lambda = 4), xlab = "n = 10", main = "", probability = TRUE)
box()
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
hist(rpois(25,lambda = 4), xlab = "n = 25", main = "", probability = TRUE)
box()
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
hist(rpois(50,lambda = 4), xlab = "n = 50", main = "", probability = TRUE)
box()
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
hist(rpois(75,lambda = 4), xlab = "n = 75", main = "", probability = TRUE)
box()
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
hist(rpois(100,lambda = 4), xlab = "n = 100", main = "", probability = TRUE)
box()
```

```
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
hist(rpois(1000,lambda = 4), xlab = "n = 1,000", main = "", probability = TRUE)
box()
curve(dnorm(x,mean = 4, sd = 2), add = TRUE, col = "red")
```



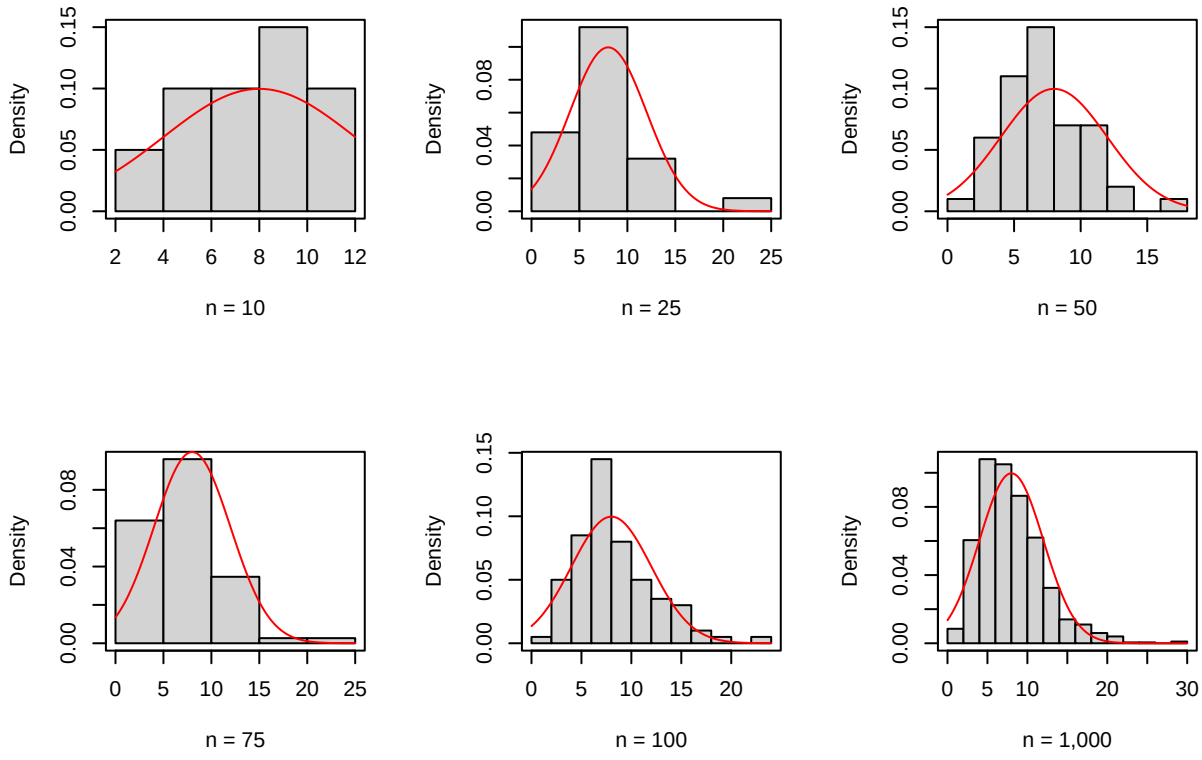

A bit rougher, but you can still see CLT at work, and it does a great job. Now what about continuous distributions?

**Example 4; CLT Still At Work**

Here, we will look at another distribution, the Gamma distribution. It is one of the continuous distributions that has many applications. Again, this example is for illustrative purposes only, so if you do not fully understand the Gamma distribution or the code, that is perfectly fine.

```
set.seed(314439)
par(mfrow=c(2,3))
hist(rgamma(10,shape = 4, scale = 2), xlab = "n = 10", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
hist(rgamma(25,shape = 4, scale = 2), xlab = "n = 25", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
hist(rgamma(50,shape = 4, scale = 2), xlab = "n = 50", main = "",
```

```
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
hist(rgamma(75,shape = 4, scale = 2), xlab = "n = 75", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
hist(rgamma(100,shape = 4, scale = 2), xlab = "n = 100", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
hist(rgamma(1000,shape = 4, scale = 2), xlab = "n = 1,000", main = "",
     probability = TRUE)
box()
curve(dnorm(x,mean = 8, sd = 4), add = TRUE, col = "red")
```




Once again, CLT does its thing.

## Random Variables

Now let's move onto the next part of today's lecture, random variables. A **Random Variable** is a function that assigns a unique numerical value to each outcome in a sample space. Each random variable follows a probability distribution (PMF/PDF), however, the true distribution is usually unknown. We generally use known distributions (the three we learned last time for example) if the random variable fits the criteria to use that distribution. This process is *very* theoretical and way beyond the scope of this course, so we will not

be looking into that. We will just be looking at random variables that we have already determined follow a certain distribution and look at properties of those distributions.

## Expected Value

The expected value of a distribution is pretty self-explanatory. It is the value of a random variable that we expect to get if we sample from that distribution. Fun Fact: This is another way of saying the average value of a distribution, or the mean. In mathematical notation, $\mathbb{E}(X) = \mu$.

## Variance

The variance is also pretty self-explanatory. It is how much the value of the random variable tends to vary. One thing to note is that the variance is the standard deviation squared $Var(X) = SD(X)^2$. The difference between variance and standard deviation is not large. Both measure variability in the distributions. However, the standard deviation is more intuitive to understand as its units will be expressed in the same way as the original variable.

## Important Formulas

Now comes the important part, how to actually calculate these values. Well how you calculate them depends on if the distribution is continuous or discrete.

- **Discrete**

$$\mu_x = \mathbb{E}(X) = \sum_{\forall x} xP(x)$$

$$\sigma_x^2 = Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X)^2) = \sum_{\forall x} x^2 P(x) - \mu_x^2$$

- **Continuous**

For the record, the continuous case requires you to know calculus to evaluate. We **DO NOT** expect you to know or learn calculus for this course. The formulas below are just for show.

$$\mu_x = \mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$\sigma_x^2 = Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X)^2) = \int_{-\infty}^{\infty} x^2 f(x)dx - (\int_{-\infty}^{\infty} xf(x)dx)^2$$

Yuck!! Be glad you do not need to evaluate these. There are actually (pretty accurate) estimators created for both the mean and variance for a normal distribution. Those are below.

- **Normal Distribution**

$$\mu_x = \mathbb{E}(X) = \frac{1}{n}\sum_{\forall x} x$$

$$\sigma_x^2 = \frac{1}{n-1}\sum_{\forall x}(x-\mu_x)^2$$

If you wish to calculate the standard deviation...

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{\forall x}(x-\mu_x)^2}$$

Okay, but what about the Binomial distribution? Is there a convenient formula that I can mindlessly plug in numbers to? Yes, yes there is.

- **Binomial/Bernoulli if** $n = 1$

$$\mu_x = np$$
$$\sigma_x^2 = np(1-p)$$

Okay, so far we have had an info dump lecture. Now let's move onto examples.

**Example 5; Every One Wins**

Let's suppose you are playing a game with a friend called Every One Wins. You win the game if you roll a one, but lose if you roll anything else. What would the expected number and standard deviation of ones be if you rolled seven times?
**Answer:**

$$\mathbb{E}(1's) = 7 * \frac{1}{6} = \frac{7}{6}$$
$$SD(1's) = \sqrt{Var(1's)} = \sqrt{7 * \frac{1}{6} * \frac{5}{6}} = \sqrt{\frac{35}{36}} = \frac{\sqrt{35}}{6} \approx 0.9860$$

**Example 6; Plankton!**

Every month Plankton tries to steal the Krabby Patty formula. Mr. Krabs is trying to save money (naturally) on his security measures so he "pays" Sandy to help him out. Sandy's first step is to determine the average number and standard deviation of times Plankton will attempt to steal the formula per month. Using the chart below, help her calculate this.

| Times Per Month | Probability |
|---|---|
| 1 | 0.50 |
| 2 | 0.20 |
| 3 | 0.15 |
| 4 | 0.10 |
| 5 | 0.05 |

$$\mathbb{E}(Attempts) = (1)(0.5) + (2)(0.2) + (3)(0.15) + (4)(0.1) + (5)(0.05) = 2$$
$$SD(Attempts) = \sqrt{\mathbb{E}(Attempts^2) - (\mathbb{E}(Attempts)^2)}$$
$$\mathbb{E}(Attempts^2) = (1^2)(0.5) + (2^2)(0.2) + (3^2)(0.15) + (4^2)(0.1) + (5^2)(0.05) = 5.5$$
$$SD(Attempts) = \sqrt{5.5 - 2^2} = \sqrt{5.5 - 4} = \sqrt{1.5} \approx 1.2247$$

**Example 7; Let R Do It For Us**

Let's re-look at the `hello` data set.

```
hello <- read.csv("~/Desktop/DPISu22/Data Sets/hello.csv", stringsAsFactors=TRUE)
```

Let's suppose you wanted to figure out the average shoe size and the standard deviation of the shoe size for students in STAT107. We *can* use the formulas listed above, but do you really want to look through and manually calculate 192 rows? Of course not. This is where `RStudio` comes in. The nice thing is we already know how to calculate the average (`mean()`), but what about the standard deviation? That is also super simple using the `sd()` function. Recall the syntax for the `mean()` function is `mean(data$column)`. The syntax for the `sd()` function is the same, `sd(data$column)`.

```
mean(hello$Shoe.Size)
```

```
## [1] 9.434896
```

```
sd(hello$Shoe.Size)
```

```
## [1] 3.074499
```

There we go, hopefully this all makes sense. Just remember to identify whether your variable is binomial or normal so you use the correct formula. Over the next two lectures, we will learn more reasons why calculating the mean and standard deviation for a random variable is useful.

## End of Lecture 11 Notes