# w05

Yekta Amirkhalili

June 5th, 2023

# Week 5 (Session 4) - June 5, 2023

###**material to cover:**

**0.** …**Descriptive Statistics**… Describe the results based on Statistics calculated on a sample set from the population.

1. Central Tendency or Location: Mean, Median, Mode, Quantiles,
2. Measures of Scale or Dispersion: Range, Variance, Standard Deviation, CV
3. Functional Form: Skewness, Kurtosis
4. Graphical Form: Histograms, BoxPlot

**1.** …**Sampling Distributions**…

## Standard Normal

$$P(Z > Z_\alpha) = \alpha$$

$$E(Z) = 0,$$
$$Var(Z) = 1$$

.

#### 1.1 t-distribution

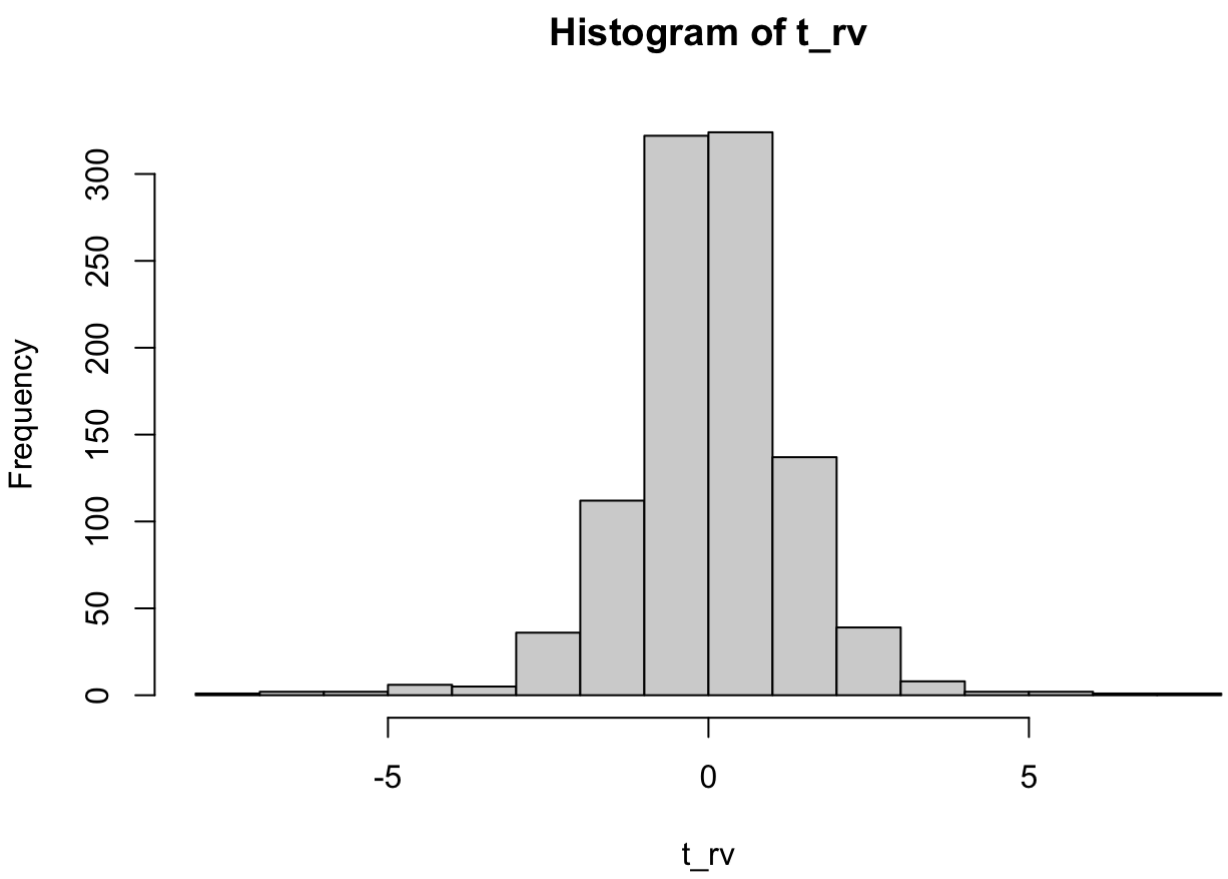
Let $Z$ be a standard normal RV and let $Y$ be independent of $Z$ and a $\chi^2$ RV with $n$ degrees of freedom. Random Variable $t_n$ follows a t-student distribution:

$$t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

```
set.seed(0)
#t random variables
t_rv <- rt(1000, 5)

#print first 10
print(t_rv[1:10])
```

```
##  [1]  1.59612424 -0.59734723  0.51893251 -0.33018072 -0.22250611  0.55853573
##  [7] -0.39149113 -0.01357225  0.75881946  0.80487760
```

```
#draw
hist(t_rv)
```

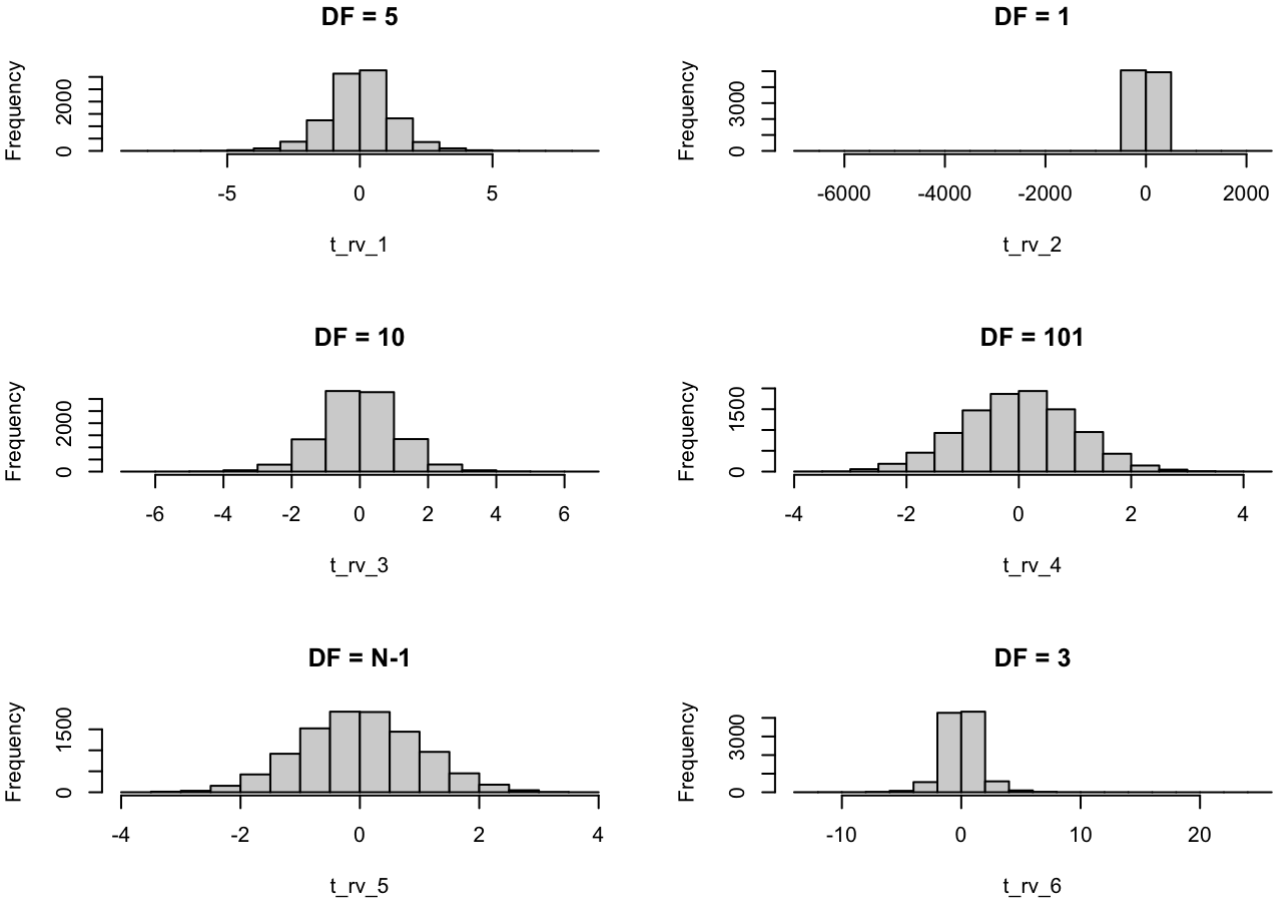



$$E(t_n) = \begin{cases} 0 & n > 1 \\ \infty - \infty & n = 1 \end{cases}$$

$$Var(t_n) = \begin{cases} \frac{n}{n-2} & n > 2 \\ \infty & n \leq 2 \end{cases}$$

Question: What if $n = 1$ or $n = 2$?

```
set.seed(1)
N <- 10000

df_1 <- 5
df_2 <- 1
df_3 <- 10
df_4 <- 101
df_5 <- N - 1

t_rv_1 <- rt(N, df_1)
t_rv_2 <- rt(N, df_2)
t_rv_3 <- rt(N, df_3)
t_rv_4 <- rt(N, df_4)
t_rv_5 <- rt(N, df_5)
t_rv_6 <- rt(N, 3)

par(mfrow = c(3,2))
hist(t_rv_1, main = 'DF = 5')
hist(t_rv_2, main = 'DF = 1')
hist(t_rv_3, main = 'DF = 10')
hist(t_rv_4, main = 'DF = 101')
hist(t_rv_5, main = 'DF = N-1')
hist(t_rv_6, main = 'DF = 3')
```




- large n value, $t_n \rightarrow N(0,1)$.

**1.2 F-distribution** Let $U$ and $V$ be independent $\chi^2$ RVs with $n$ and $m$ degrees of freedom, respectively. F RV follows the F-distribution:

$$F_{n,m} = \frac{\frac{\chi^2_n}{n}}{\frac{\chi^2_m}{m}}$$

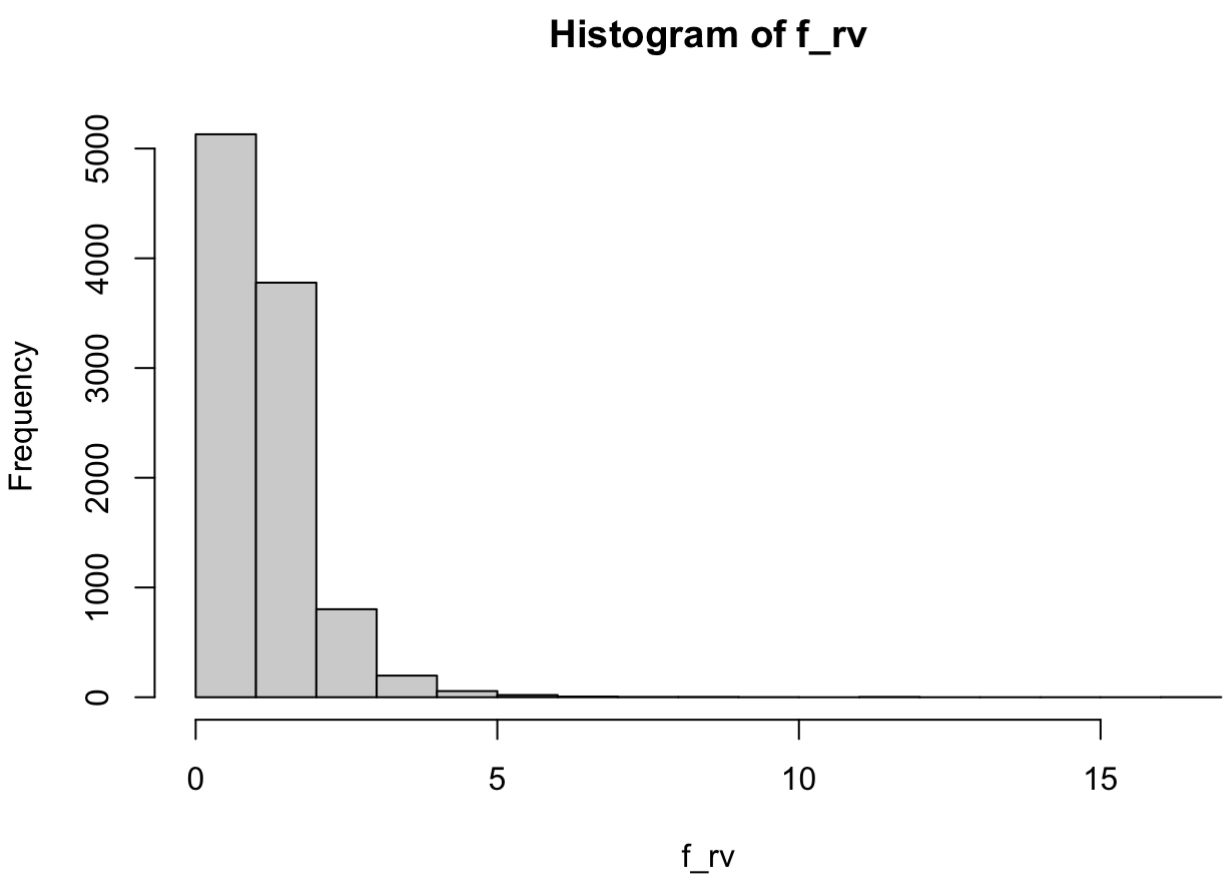
```
set.seed(2)

N <- 10000
n <- 10
m <- 15

f_rv <- rf(N, n, m)

xchi1 <- rchisq(10000, 10)
xchi2 <- rchisq(10000, 15)

f_rv_0 <- (xchi1/n)/(xchi2/m)

print(mean(f_rv))
```

```
## [1] 1.158077
```

```
print(mean(f_rv_0))
```

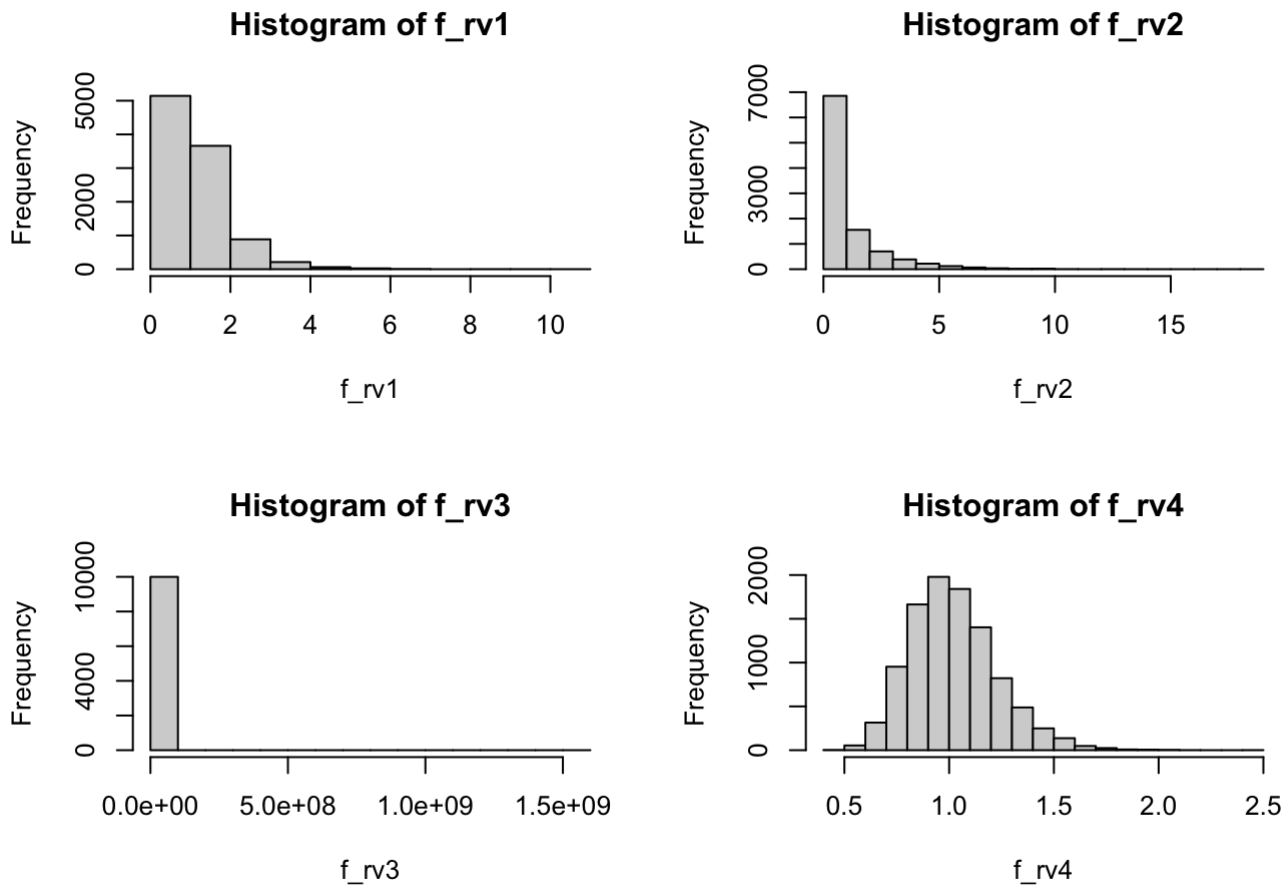
```
## [1] 1.139963
```

```
hist(f_rv)
```




$$E(F_{n,m}) = \frac{m}{m-2} \quad m > 2$$

```
set.seed(3)

N <- 10000
n1 <- 10
m1 <- 15

n2 <- 1
m2 <- 100

n3 <- 100
m3 <- 1

n4 <- 100
m4 <- 100


f_rv1 <- rf(N, n1, m1)
f_rv2 <- rf(N, n2, m2)
f_rv3 <- rf(N, n3, m3)
f_rv4 <- rf(N, n4, m4)

par(mfrow = c(2,2))
hist(f_rv1)
hist(f_rv2)
hist(f_rv3)
hist(f_rv4)
```




**2. Inferenetial Statistics** From the Descriptive results, figure out what the population statistics are. This is when you need to estimate things based on sample statistics.

**2.1 Point Estimates** Different ways to measure how well you're estimating something.

**2.1.1 Bias** Bias is how far away your estimated value is from the actual value and is formulated:

$$b = E(\hat{\theta}) - \theta$$

**2.1.2 Mean Squared Error** Bias isn't always the best way to evaluate your estimate. (Why?)

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

or,

$$MSE(\hat{\theta}) = b^2 + Var(\hat{\theta})$$

Ok, but how do we even estimate?

**2.2 Point Estimate Methods 2.2.1 Method of moments** No real formulation, it is just based on parameters of population.

**2.2.2 Maximum Likelihood Method** An optimization problem where the objective is to maximize the likelihood function. What is a likelihood function? It is how likely you it is (the probability of) that the distribution you think the data has based on the samples is actually the real distribution? Basically, how likely are you to have guessed the correct value for whatever parameter you're estimating?

**Additional: Point Estimates and P-value** A p-value is the smallest critical value ($\alpha$) with which you can reject a null hypothesis. This $\alpha$ value is given to you, the only thing you calculate is the p-value and then compare it to the critical value and make a decision.

$$p - value = P(Z \geq Z_0)$$

If $p - value \leq \alpha$, reject the null hypothesis. Else, you cannot reject the null hypothesis.

Ok, but what does any of this mean?! Let's look at a simple example with standard normal distribution. Here is all the information you have:

- Data Follows $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Let's say I want to estimate the population mean based on sample mean. Let's say the actual population mean is $\mu$, and the value I have calculated based on samples is $\overline{X}$ and the value I have estimated the actual population mean to be is $\mu_0$. Basically, I need to know how far off am I from the actual population mean value. The issue: I don't know the population mean value! I can however break down $Z$:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

Based on the area under the curve for a standard normal distribution, what do I know? Let's say I am working with $\alpha = 0.05$ which means a critical value of 5%. We will learn what this means later on. Just remember this:

$$P(Z > Z_\alpha) = \alpha$$

Plug in that value and we have:

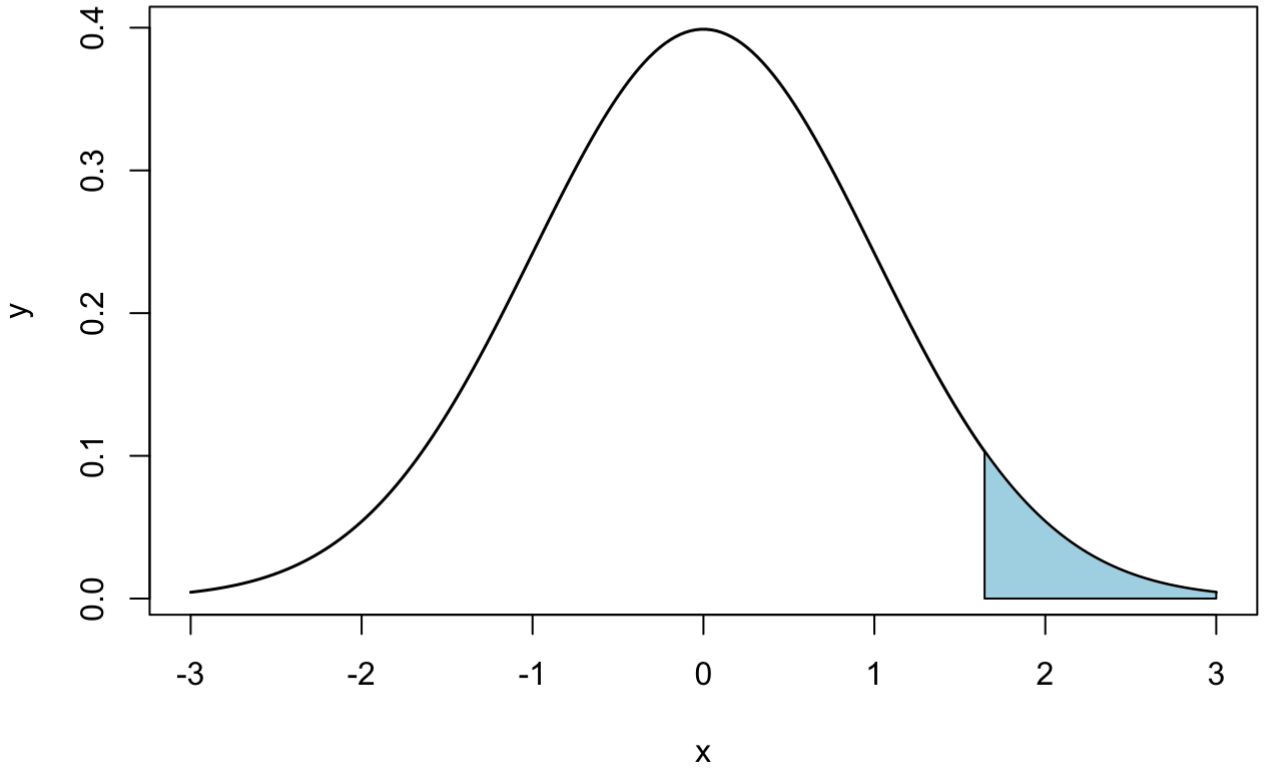
$$P(Z > Z_{0.05}) = 0.05$$

What is the value of $Z_{0.05}$? If it is difficult to think of 0.05, we can look at it the other way because we know: $Z_{1-\alpha} = -Z_\alpha$. So, we are basically looking for the $Z$ value of the 95% quantile!

```
quant_95 <- qnorm(0.95, mean = 0, sd = 1)
quant_95
```

```
## [1] 1.644854
```

```
x <- seq(-3,3,length=200)
y <- dnorm(x)

plot(x, y, type = "l", lwd = 1.5)

x_ <- seq(quant_95,3,length=200)
y_ <- dnorm(x_)
polygon(c(quant_95,x_,3), c(0, y_, 0), col = 'lightblue')
```




But what is the value of the blue part?

```
value <- pnorm(3, 0, 1) - pnorm(quant_95, 0, 1)
value
```

```
## [1] 0.0486501
```

And that is the $\alpha$ value! Perfect! Now we know what we mean when we say $P(Z > Z_\alpha)$. It means the probability (area under the curve) of falling in that blue section. How do we use this though?

Let's go back to our Hypothesis. I said the actual mean was $\mu$ and I estimated some value $\mu_0$. My hypothesis is that these values are equal. I write this mathematically as:

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

We just work with the null hypothesis (first one). This is an equal hypothesis, which means your estimation can be either greater than or less than the actual value. Right? So, you are essentially looking at both sides of the actual $\mu$. Remember:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$Z\frac{\sigma}{\sqrt{n}} = \overline{X} - \mu_0$$

We can also say based on previous calculations that:

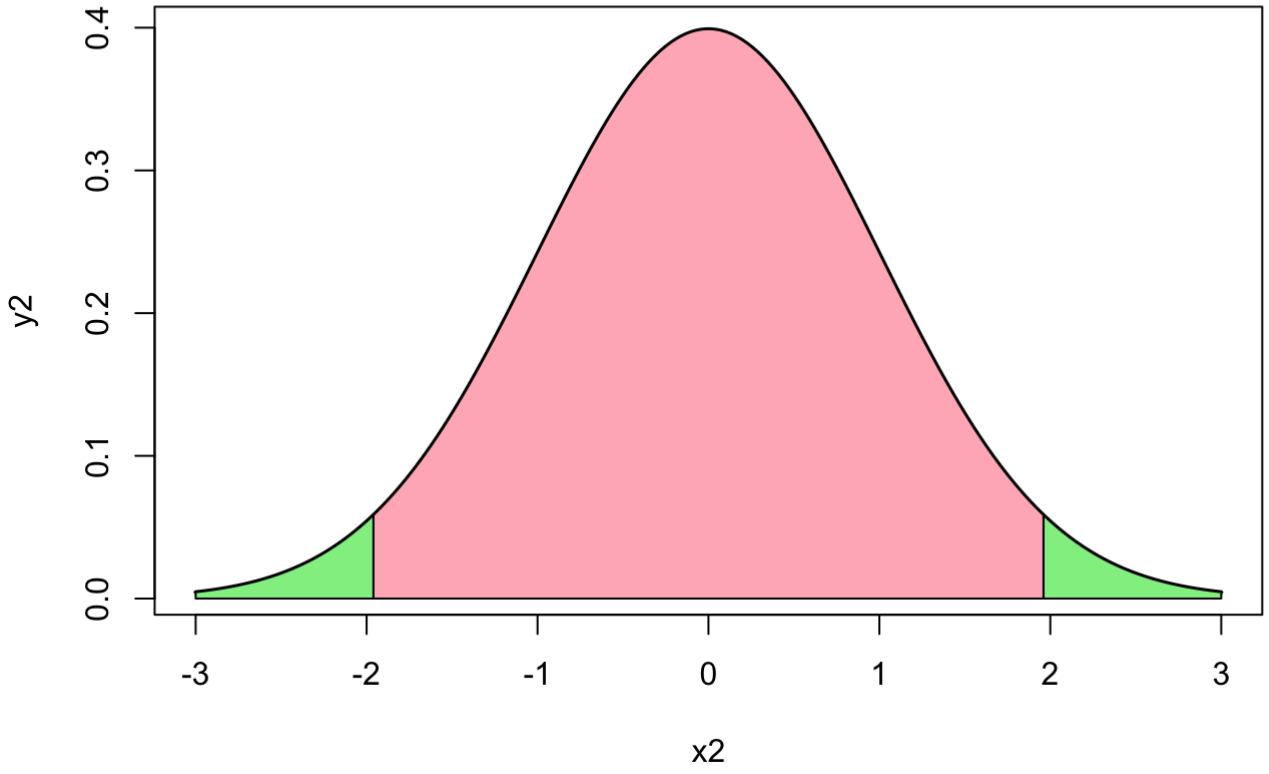
$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

Where $2Z_{\alpha/2} = Z_{\alpha}$. So, we can now write this for $\overline{X} - \mu_0$:

$$-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu_0 \leq Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This is called a confidence Interval. Let's visualize this for the previous example:

```
sample_size <- 200
sample_mean <- 1
sample_std <- 2
alpha_new <- (0.05)/2

#we now need to calculate the Z value for alpha = 0.025
z_alphaHalf <- qnorm(1 - alpha_new, mean = 0, sd = 1)

x2 <- seq(-3, 3, length = 200)
y2 <- dnorm(x2)

plot(x2, y2, type = "l", lwd = 2)

x2_ <- seq(z_alphaHalf, 3, length = 100)
y2_ <- dnorm(x2_)
polygon(c(z_alphaHalf, x2_, 3), c(0, y2_, 0), col = 'lightgreen')

x3_ <- seq(-3, -z_alphaHalf, length = 100)
y3_ <- dnorm(x3_)
polygon(c(-3, x3_, -z_alphaHalf), c(0, y3_, 0), col = 'lightgreen')

x4_ <- seq(-z_alphaHalf, z_alphaHalf, length = 200)
y4_ <- dnorm(x4_)
polygon(c(-z_alphaHalf, x4_, z_alphaHalf), c(0,y4_,0), col = 'lightpink')
```




Let's actually see these values:

```
lower_bound <- sample_mean - (1 * z_alphaHalf * sample_std / sqrt(sample_size))
upper_bound <- z_alphaHalf * sample_std / sqrt(sample_size) + sample_mean

print(paste('Accept values between: ', lower_bound, ' and ', upper_bound))
```

```
## [1] "Accept values between:  0.722819235130065  and  1.27718076486994"
```

Which means if your estimate is $\mu_0 = 1$ then you are in the acceptance area (pink colored). Which means, you cannot reject the null hypothesis and we say that your estimated mean is statistically significantly close to the actual mean.