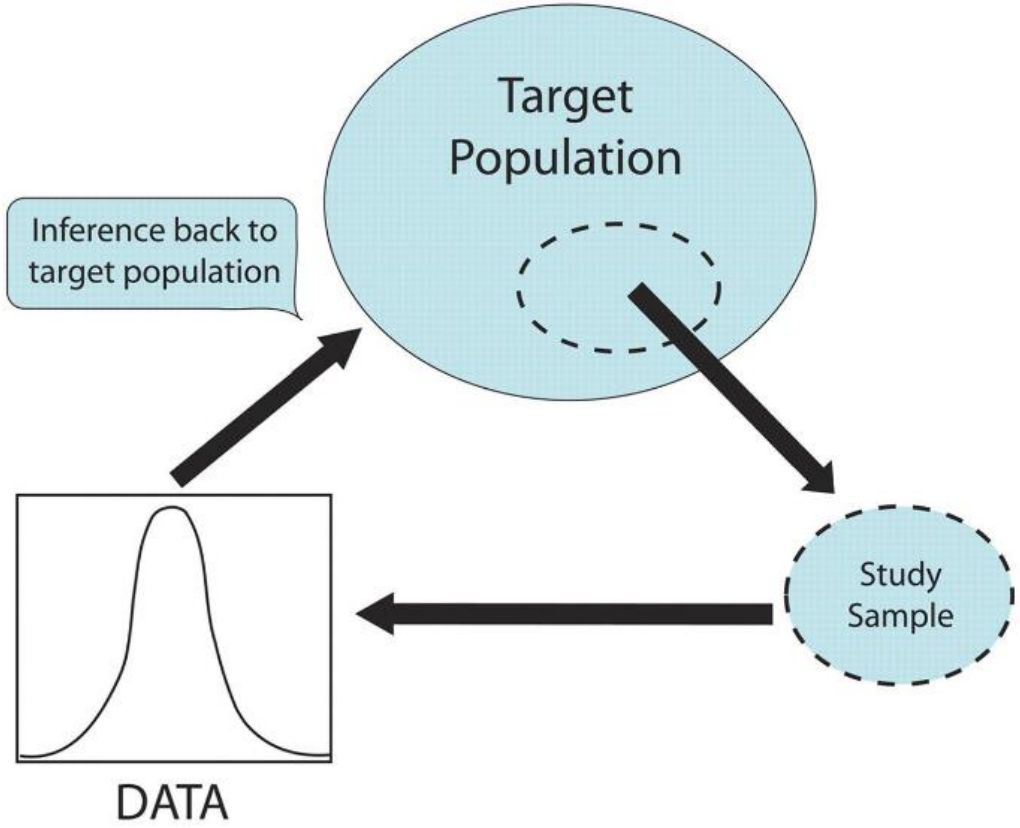
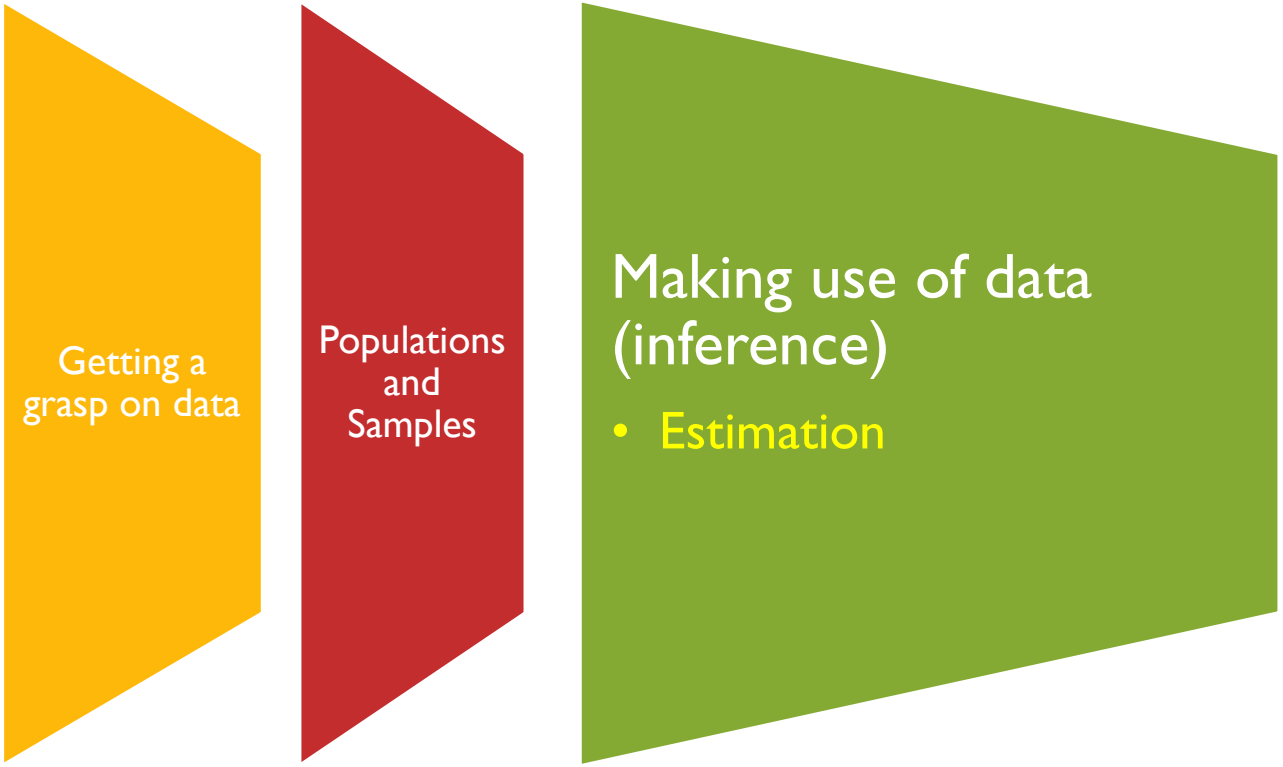
Welcome


Target Population
Inference back to target population
Study Sample
DATA

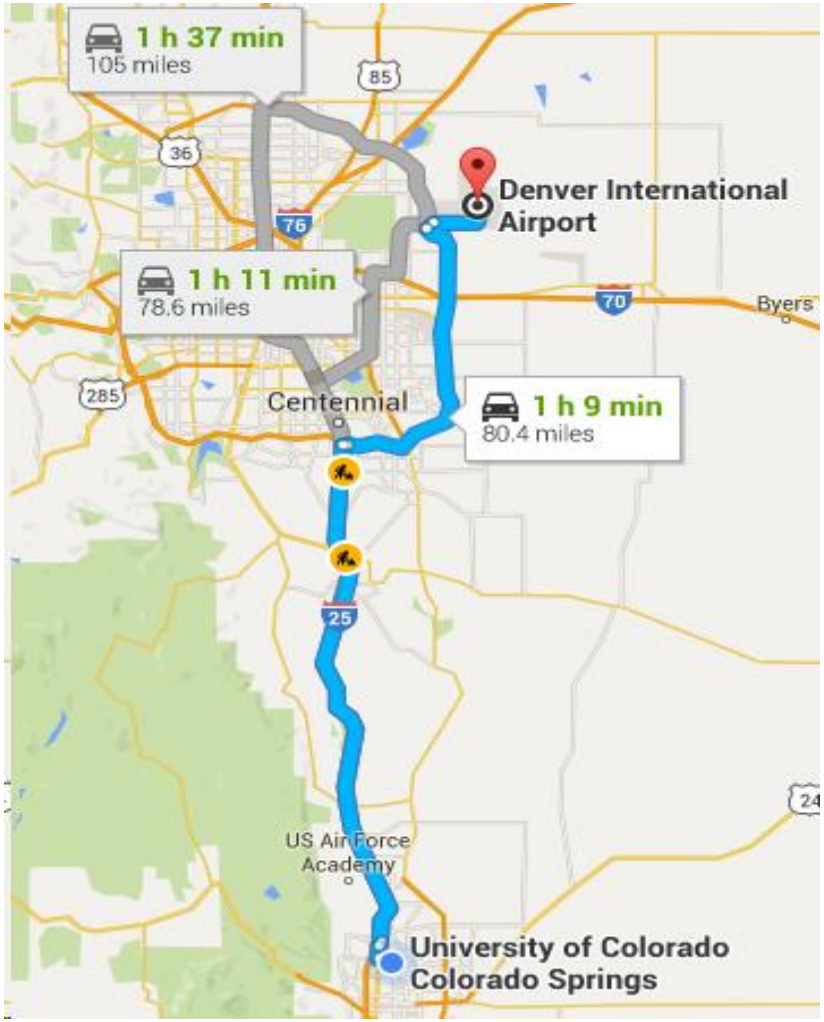
What did we do last class?

- Probability density functions
- Normal distribution (z distribution)
- Sampling distribution
  - Mean
  - Proportion

What have we accomplished and



How long does it to get to the airport



You must arrive at the airport before 4:00 pm when your flight departs

If the commute time is normally distributed

- Population mean is 1.5 hours
- Population standard deviation is 0.5

- Right now is 1:00 pm here and you must arrive at the airport before 4:00 pm when your flight departs

  - What is the probability that the commute time is *more* than 2 hours ?
  - What is the probability that the commute time is *more* than 2.5 hours ?

So what is the result ? What is your decision?

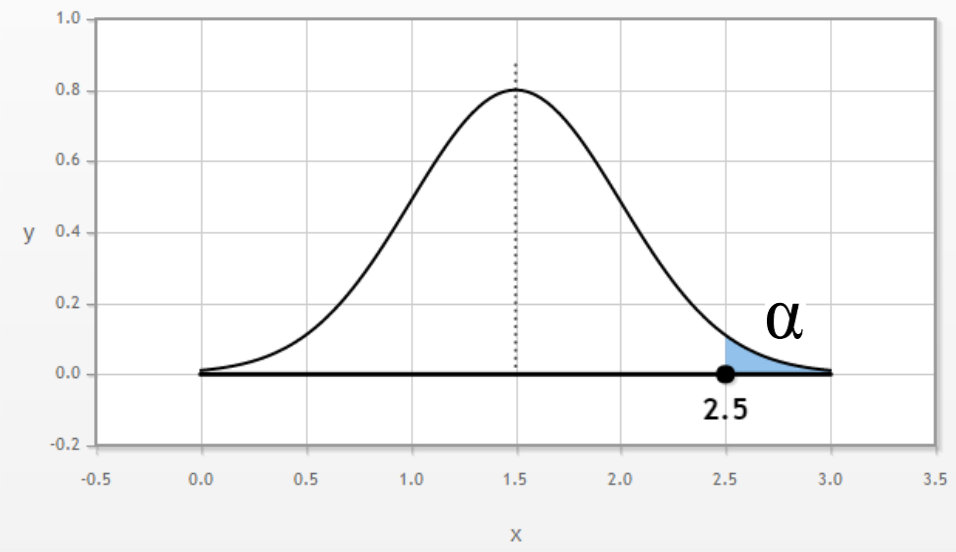
Case I You don't have a girlfriend Right now is 1:00 pm

You must arrive the airport before 4:00 pm

Based upon the following information :

You can leave the university around 1:30 pm, and you will only have 2.28% chance to miss your flight

我女朋友呢!



$$P\left(\frac{X-\mu}{\sigma} > \frac{2.5-1.5}{0.5}\right)$$

$$P(Z > 2) = 0.0228$$

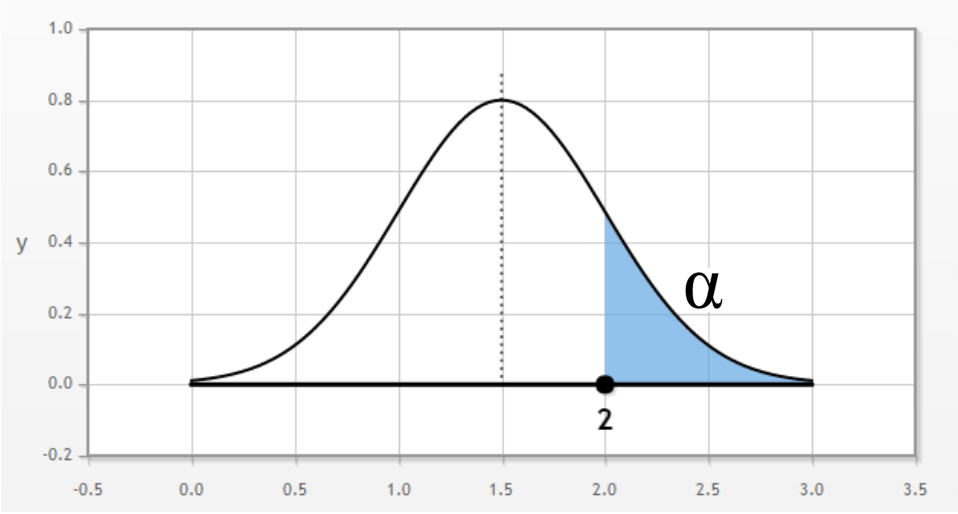
So what is the result ? What is your decision?

Case 2 You have a girlfriend Right now is 1:00 pm

You must arrive the airport before 4:00 pm

Based upon the following information :

You can leave the university around 2:00 pm, but you will have 15.87% chance to miss your flight

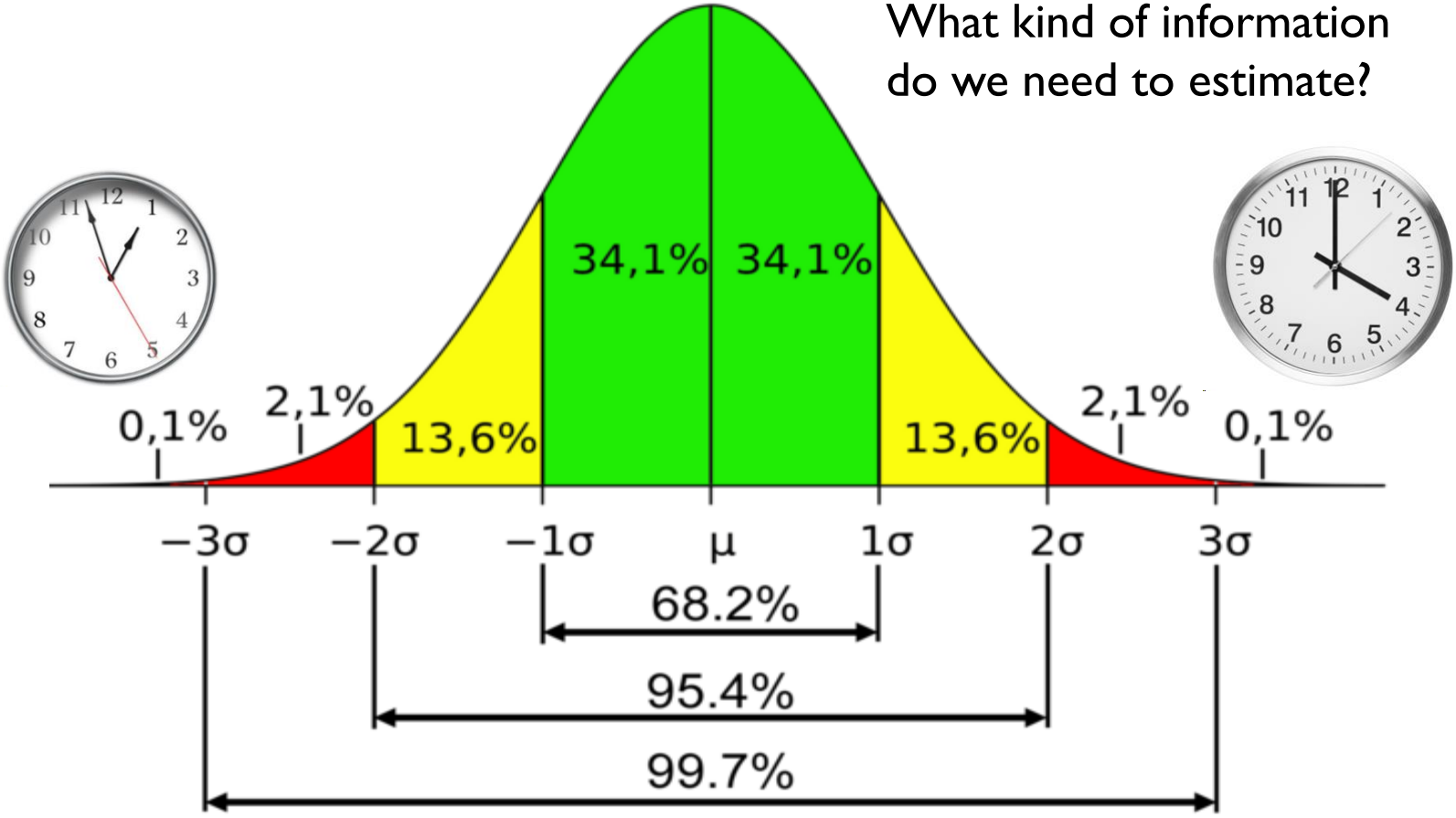


$$P\left(\frac{X-\mu}{\sigma} > \frac{2-1.5}{0.5}\right)$$

$$P\,(Z > 1) = 0.1587$$

Normal distribution- Empirical rule

What kind of information do we need to estimate?



0 hour **0.5 hour** **1 hour** **1.5hours** **2hours** **2.5hours** **3 hours**

Confidence level

A Confidence Interval is an interval of numbers containing the true population mean

It is always two-tailed and the width is changing according to confidence level

$P(\mu - 1\sigma \leq \mu \leq \mu + 1\sigma) = 0.682$

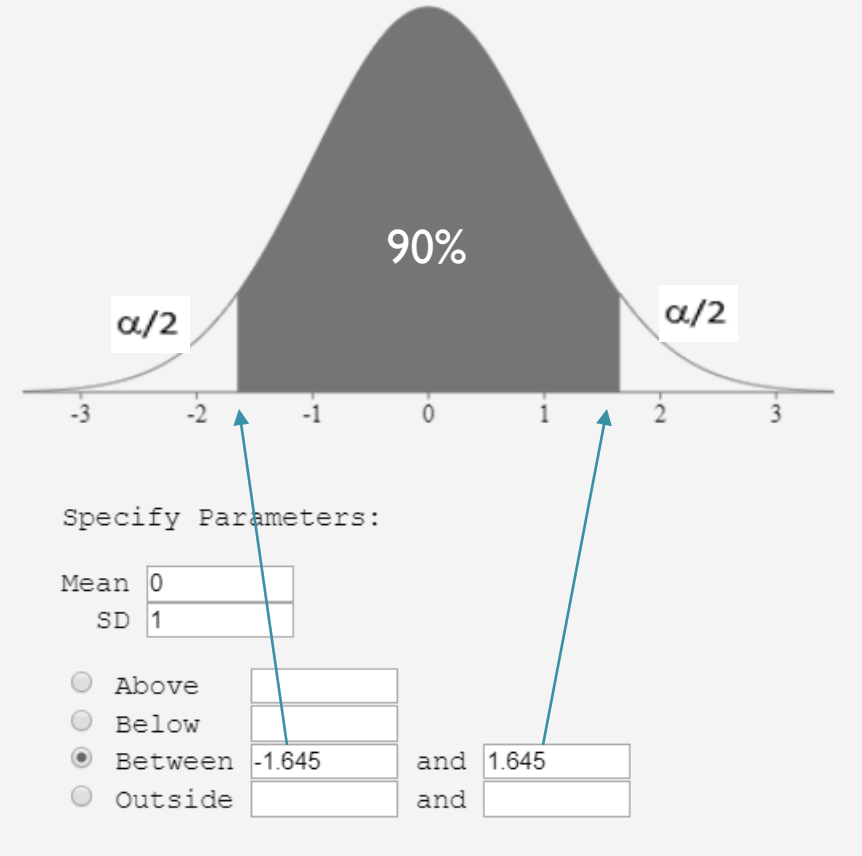
$P(\mu - 2\sigma \leq \mu \leq \mu + 2\sigma) = 0.954$

$P(\mu - 3\sigma \leq \mu \leq \mu + 3\sigma) = 0.997$

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
>
> pnorm(2)-pnorm(-2)
[1] 0.9544997
>
> pnorm(3)-pnorm(-3)
[1] 0.9973002
```

In business, we like to use 90%, 95%, 99% confidence level

P $(\mu - 1.645\sigma \leq \mu \leq \mu + 1.645\sigma)$ = 0.90

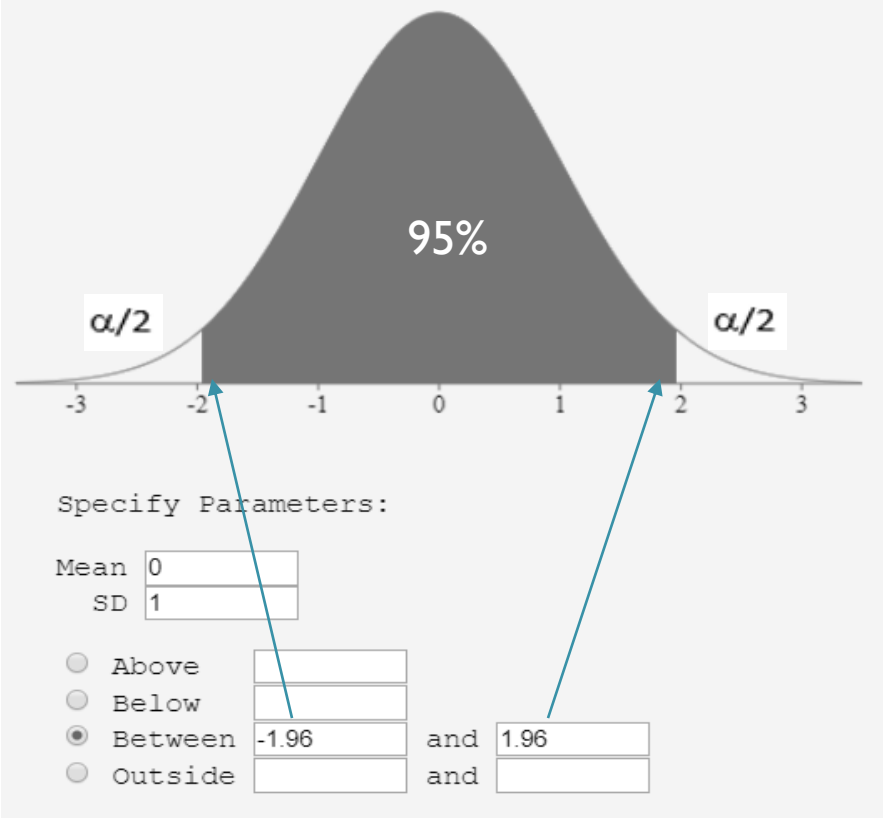


Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

```
> qnorm(0.1/2)
[1] -1.644854
>
> qnorm(0.05/2)
[1] -1.959964
>
> qnorm(0.01/2)
[1] -2.575829
>
```

In business, we like to use 90%, 95%, 99% confidence level

P $(\mu - 1.96\sigma \leq \mu \leq \mu + 1.96\sigma)$ = 0.95

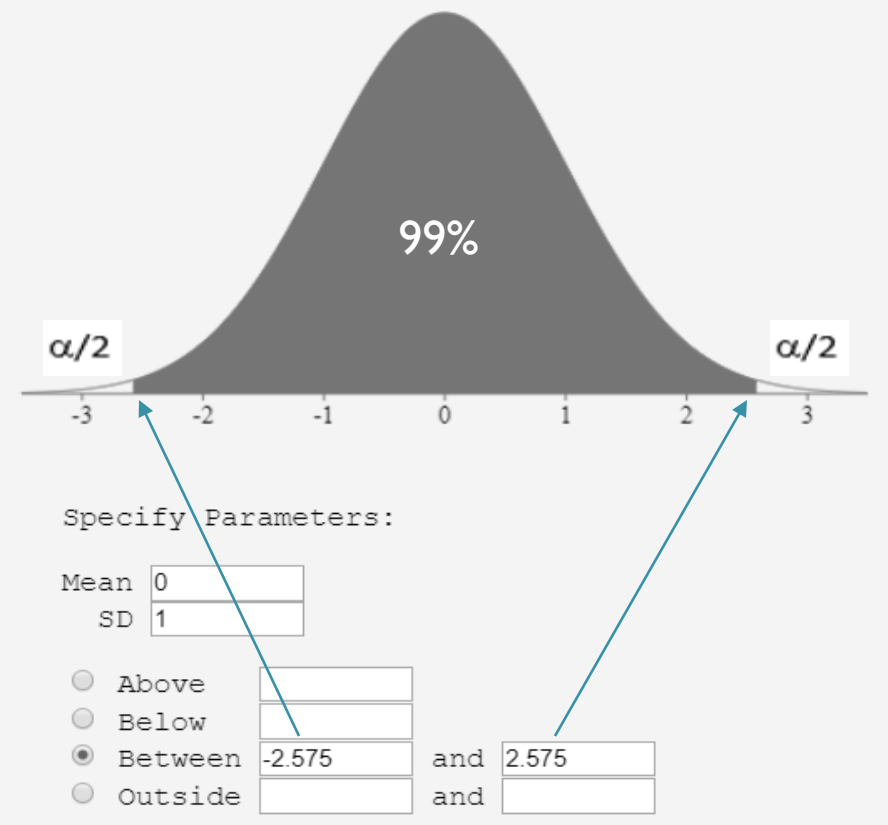


Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

```
> qnorm(0.1/2)
[1] -1.644854
>
> qnorm(0.05/2)
[1] -1.959964
>
> qnorm(0.01/2)
[1] -2.575829
>
```

In business, we like to use 90%, 95%, 99% confidence level

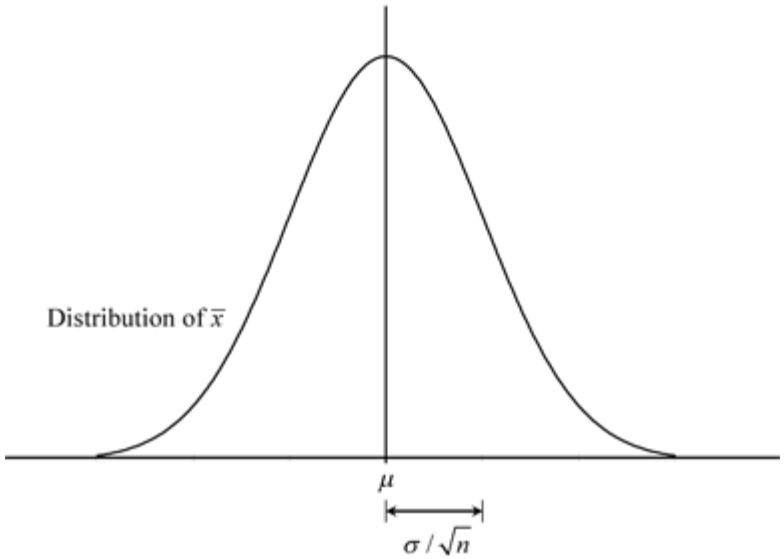
P $(\mu - 2.575\sigma \leq \mu \leq \mu + 2.575\sigma) = 0.99$



Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

```
> qnorm(0.1/2)
[1] -1.644854
>
> qnorm(0.05/2)
[1] -1.959964
>
> qnorm(0.01/2)
[1] -2.575829
>
```

Did you still remember sampling distribution of mean



$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

What is wrong ?

$P(\mu - 1.645\sigma \leq \mu \leq \mu + 1.645\sigma) = 0.90$

$P(\mu - 1.96\sigma \leq \mu \leq \mu + 1.96\sigma) = 0.95$

Why we need to re-estimate the true population mean if we have knew it???

$P(\mu - 2.575\sigma \leq \mu \leq \mu + 2.575\sigma) = 0.99$

We use sample mean to estimate population mean

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$P\left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.90$$

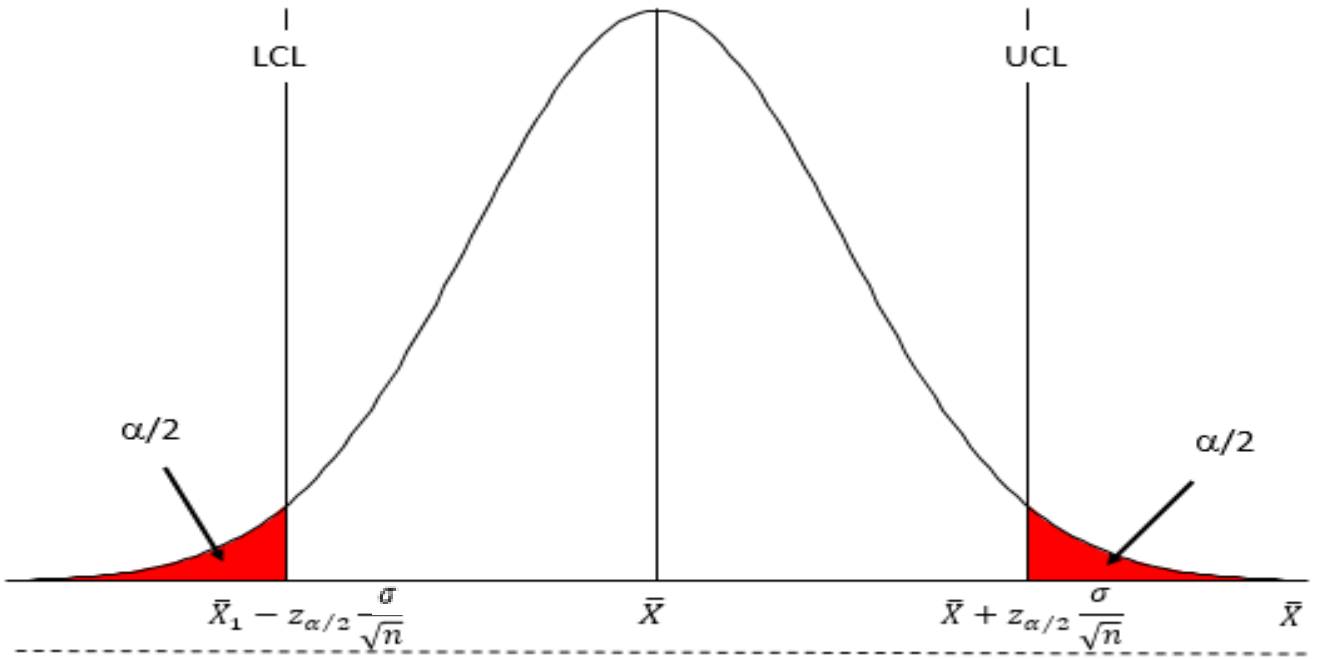
$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{x} - 2.575\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.575\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

Four basic elements in confidence interval

- Sample mean
- Population variance
- Confidence level
- Sample size

Lower & Upper confidence limit



Lower confidence limit (LCL)

Upper confidence limit (UCL)

GENERAL FORMULA

$$\bar{x} \pm (z \text{ critical value})\frac{\sigma}{\sqrt{n}}$$

Problem

- The commuting time between the university and the airport is normally distributed. A random sample of 25 was drawn from a normal distribution with a standard deviation σ of 0.5. The sample mean is 1.5 hours.

- Determine the 90%, 95% and 99% confidence interval estimate of the population mean.

- Determine the 95% confidence interval with a sample size of 100.

R programing

```
xbar <- 1.5
psd <- 0.5
n <- 25
se <- abs(qnorm(0.1/2)*psd/sqrt(n))
lcl <- xbar-se
ucl <- xbar+se
ci <- c(lcl, ucl)
ci
```

Problem

- A group of 16 foot surgery patients had a mean weight of 240 pounds. The standard deviation σ was 25 pounds.

- Find a confidence interval for a sample for the true mean weight of all foot surgery patients. Find a 90% confidence interval.

```
> xbar <- 240
> psd <- 25
> n <- 16
> se <- abs(qnorm(0.10/2)*psd/sqrt(n))
> lcl <- xbar-se
> ucl <- xbar+se
> ci <- c(lcl, ucl)
> ci
[1] 229.7197 250.2803
```

What is the error of confidence interval

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \left(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

If sample mean is equal population mean

The bound on the error (B) of estimation can be rewritten as

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Sample size

- When we want to estimate population mean within a given bound of error with a certain level of confidence.

- We can calculate the sample size needed by solving the equation

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2$$

Problem

- We would like to estimate a population mean to within 10 units. The confidence level has been set at 95% and σ = 200. Determine the sample size.

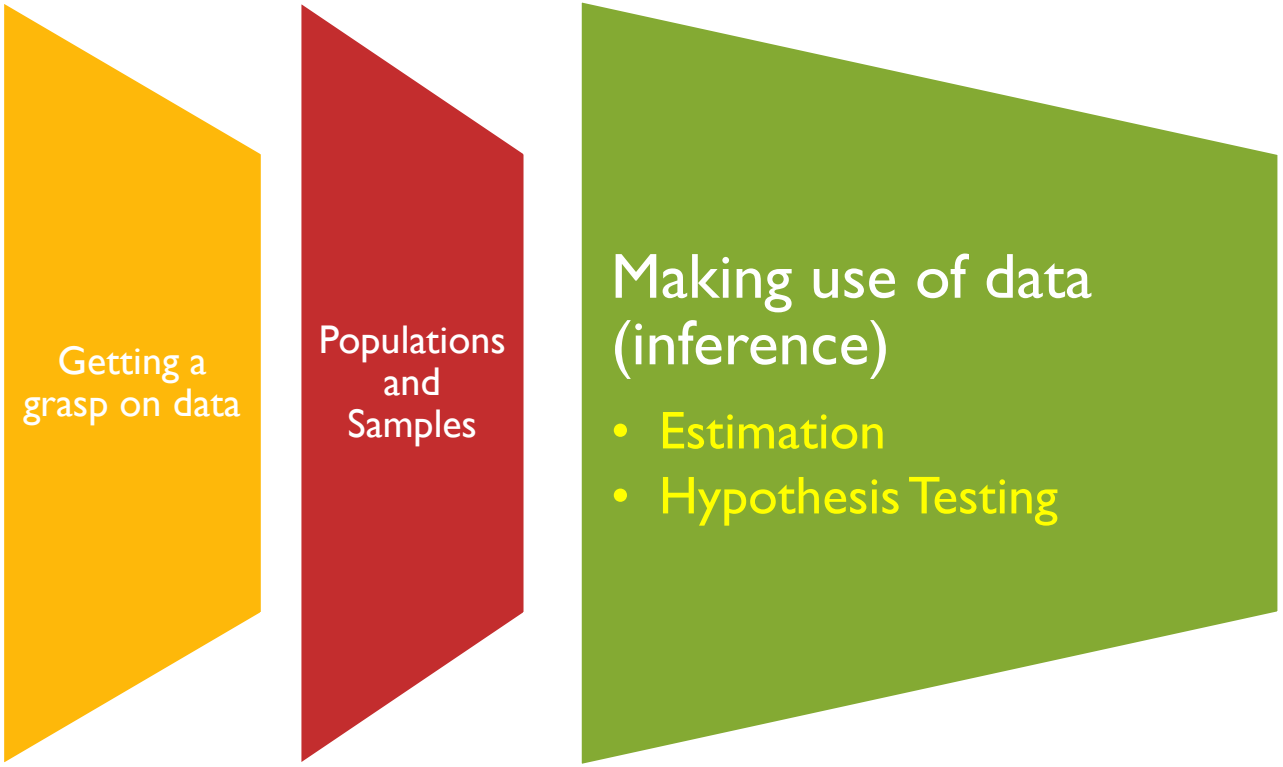
- We would like to estimate a population mean to within 10 units. The confidence level has been set at 95% and σ = 100. Determine the sample size.

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2$$

R programing

```
psd <- 200
b <- 10
n <- (qnorm(0.05/2)*psd/b)^2
round(n)
```

Where are we and where are we going ?



Hypotheses

- Null hypothesis.
  - denoted by H0, is usually the hypothesis that sample observations result purely from chance.

- Alternative hypothesis.
  - denoted by H1, is the hypothesis that sample observations are influenced by some non-random cause.

Your are testing H1 and try to find the statistical evidence to reject H0

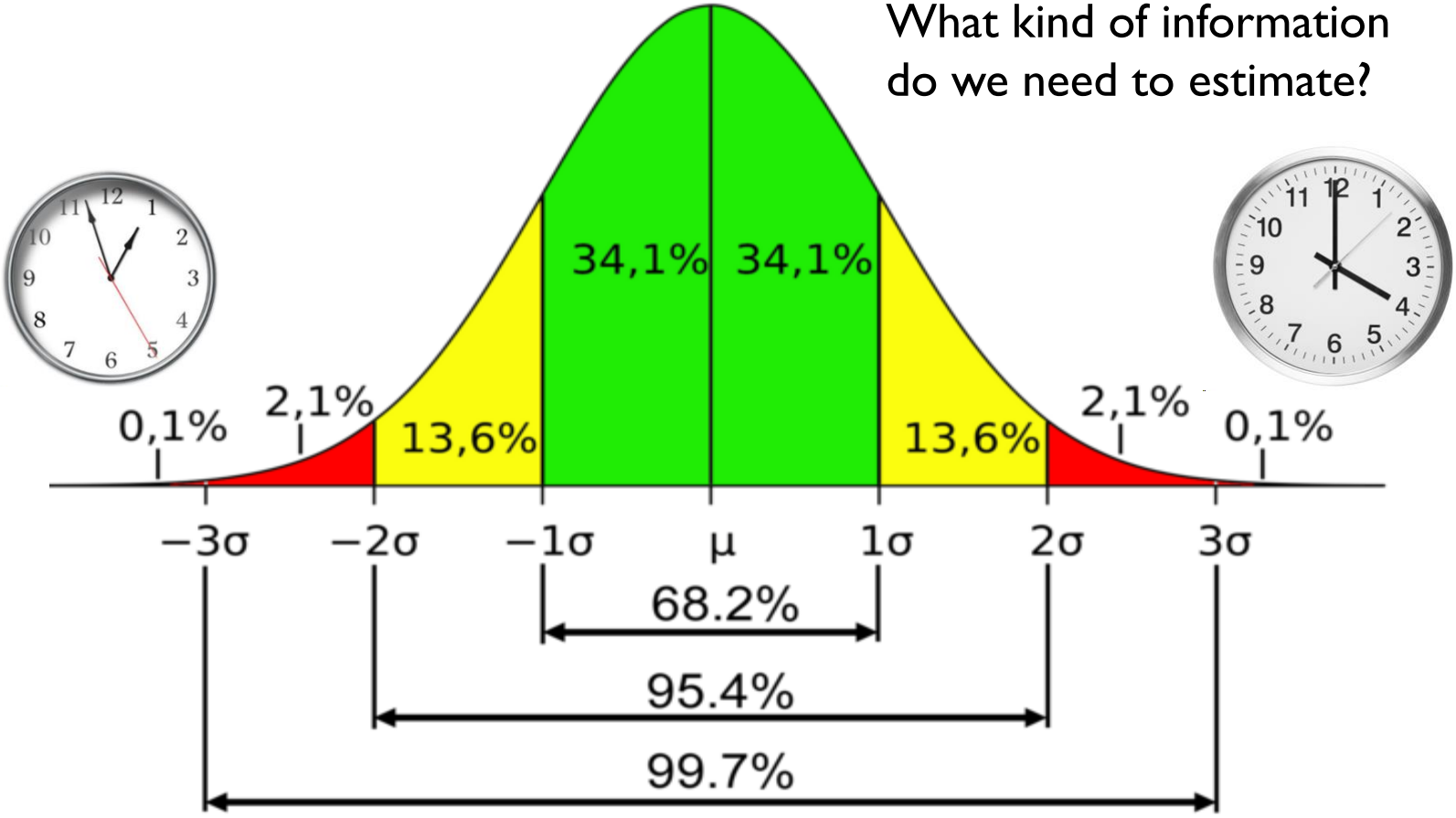
EXAMPLES

- The General Manager tells an investigative reporter that at least 85% of its customers are "completely satisfied" with their overall purchase performance. What hypotheses will be used by the reporter to test the claim?

- A student counsellor claims that first year Science students spend an average 3 hours per week doing exercises in each subject. What hypotheses will be used by a lecturer to test the claim?

Normal distribution- Empirical rule

What kind of information do we need to estimate?

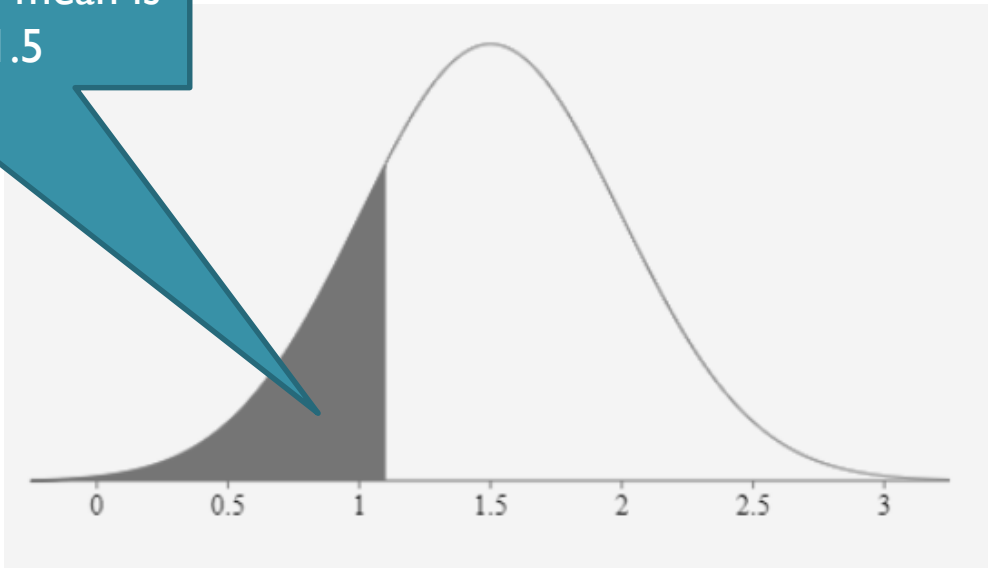


0 hour **0.5 hour** **1 hour** **1.5hours** **2hours** **2.5hours** **3 hours**

Left-tailed test

- I believe that the commute time between the university and the airport is larger than or equal 1.5 hours.
- Suppose that our random sample of n = 25 students and their average commute time is 1.1hours.
- The alternative hypothesis might be that the commute time is less than 1.5 hours.

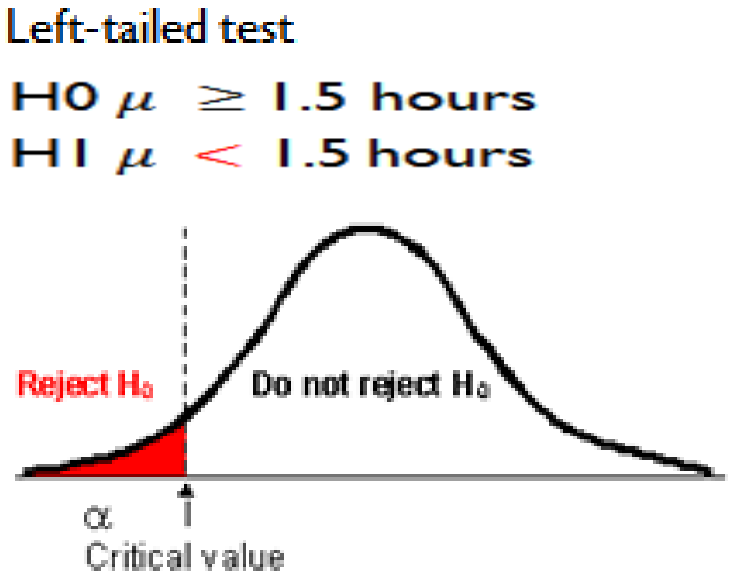
You are testing if sample mean is actually less than 1.5



- H0 $\mu \geq$ 1.5 hours
- H1 $\mu <$ 1.5 hours

How to reject the null hypothesis

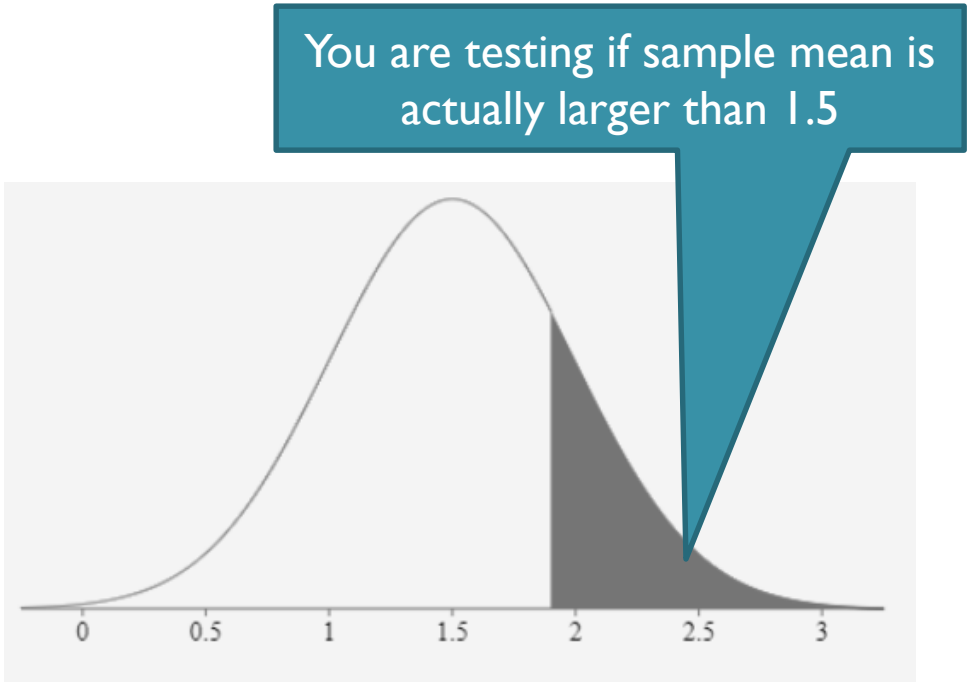
If we establish directional hypotheses, then the rejection region is allocated to left tail of the probability distribution



1. We try to prove the commute time is greater than 1.5 hours
2. When we can have sufficient evidence to reject H0 ?
3. Critical value is the threshold

Right-tailed test

- I believe that the commute time between the university and the airport to be less than or equal 1.5 hours.
- Suppose that our random sample is n = 25 students and their average commute time is 1.9 hours.
- The alternative hypothesis might be that the commute time is larger than 1.5 hours.

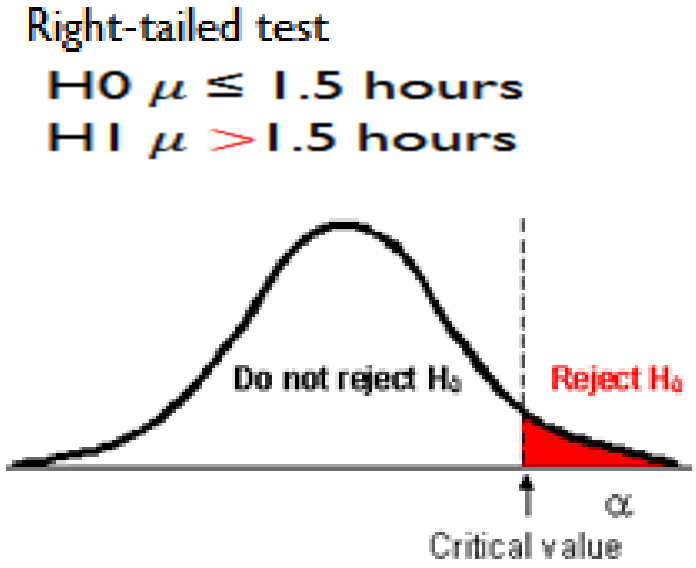


- H0 $\mu \leq$ 1.5 hours
- H1 $\mu >$ 1.5 hours

How to reject the null hypothesis

If we establish directional hypotheses, then the rejection region is allocated to right tail of the probability distribution

1. We try to prove the commute time is greater than 1.5 hours
2. When we can have sufficient evidence to reject H0 ?
3. Critical value is the threshold



Two-tailed test

- I believe that the commute time between the university and the airport is 1.5 hours.
- Suppose that our random sample is n = 25 students and their average commuting time is 1.6, which is not equal to 1.5 hours.
- The alternative hypothesis might be that the commute time is different from 1.5 hours.

You are testing if sample mean is actually larger than 1.5

You are testing if sample mean is actually smaller than 1.5

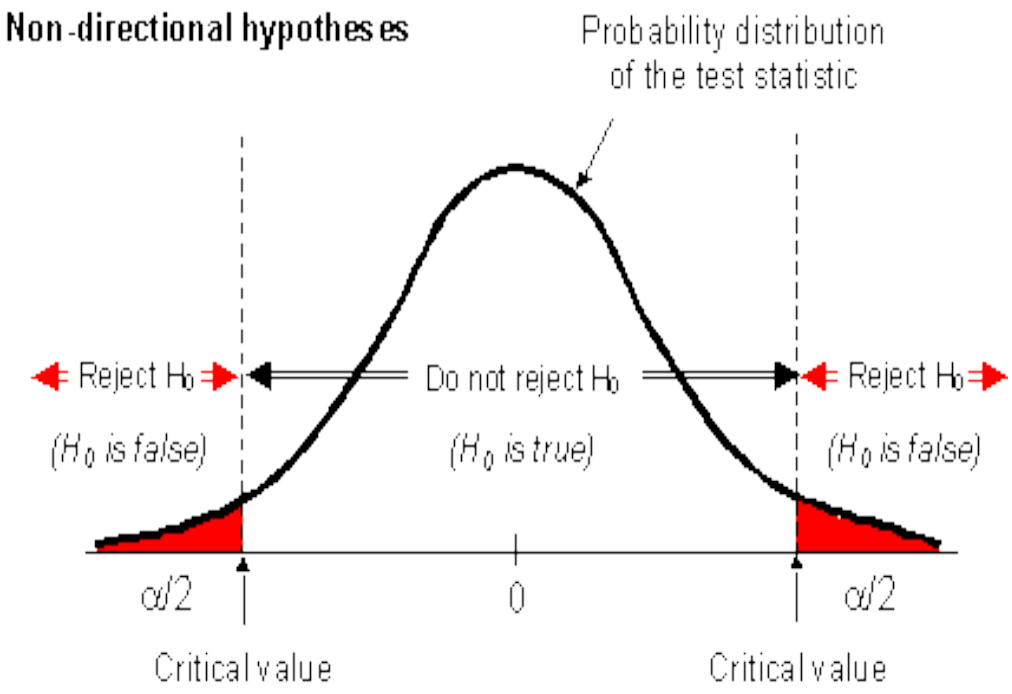
- H0 μ = 1.5 hours
- H1 $\mu \neq$ 1.5 hours

How to reject the null hypothesis

The rejection region associated with two tailed test

H0 μ = 1.5 hours

H1 $\mu \neq$ 1.5 hours

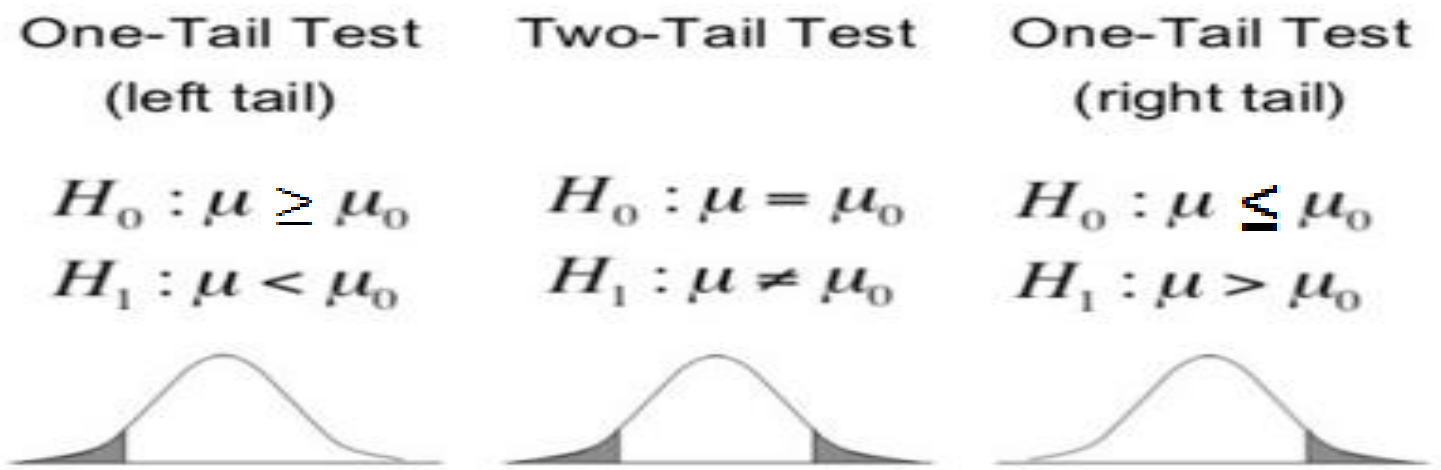


1. We try to prove the commute time is not equal to 1.5 hours
2. When we can have sufficient evidence to reject H0 ?
3. Critical value is the threshold

One-tailed test or two-tailed test

- A one-tailed test (if one mean is greater or less than another mean, but not both)
  - A direction must be chosen prior to testing.

- A two-tailed test (if two means are different from one another)
  - A direction does not have to be specified prior to testing.



Recall you memory about

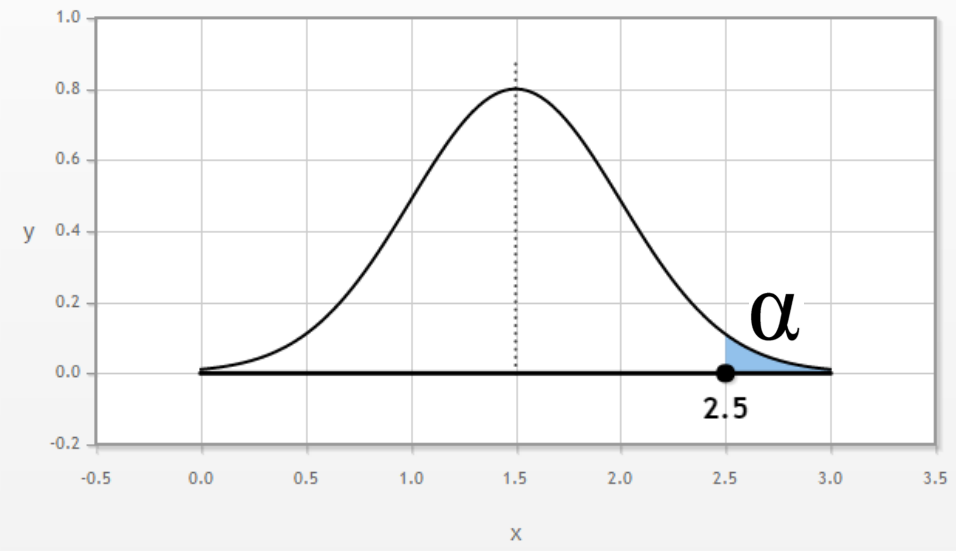
What is your decision?

Case 1 Right now is 1:00 pm

You must arrive the airport before 4:00 pm

Based upon the following information :

You can leave the university around 1:30 pm, and you will only have 2.28% chance to miss your flight



$$P\left(\frac{X-\mu}{\sigma} > \frac{2.5-1.5}{0.5}\right)$$

$$P\left(Z > 2\right) = 0.0228$$

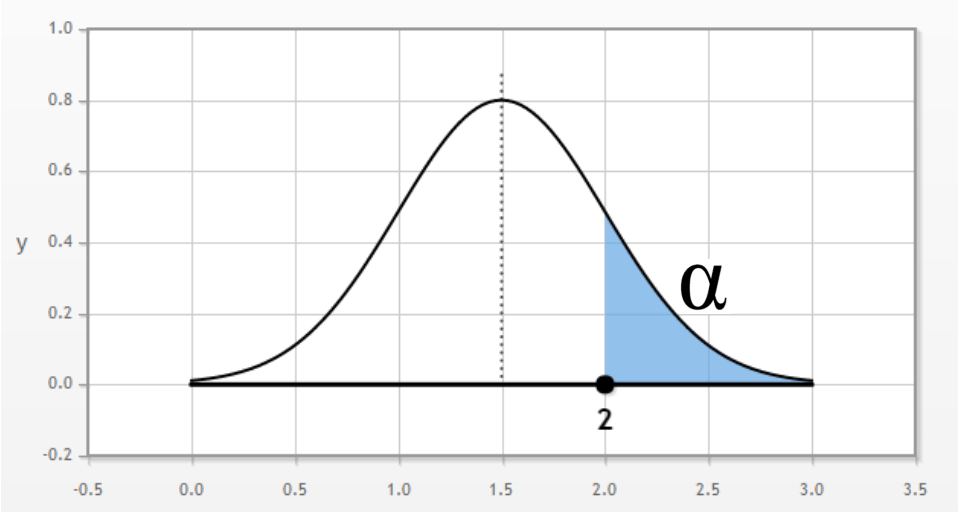
What is your decision?

Case 2 Right now is 1:00 pm

You must arrive the airport before 4:00 pm

Based upon the following information :

You can leave the university around 2:00 pm, but you will have 15.87% chance to miss your flight



$$P\left(\frac{X-\mu}{\sigma} > \frac{2-1.5}{0.5}\right)$$

$$P\left(Z > 1\right) = 0.1587$$

Type one and type two error

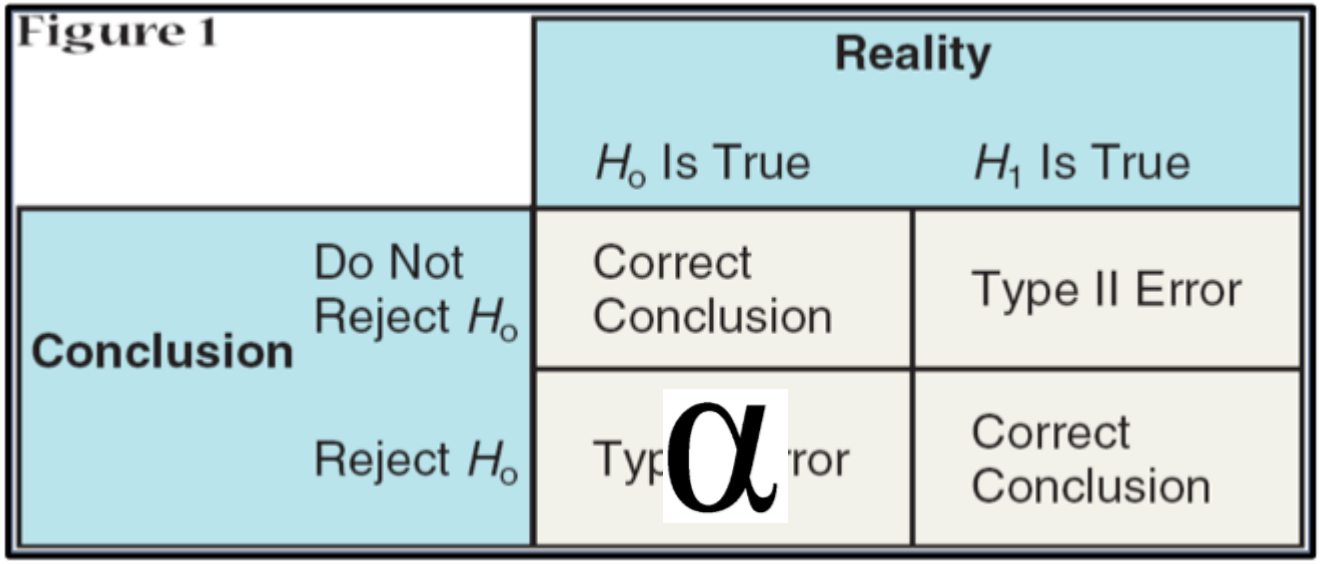


Figure 1		Reality	
		H_o Is True	H_1 Is True
Conclusion	Do Not Reject H_o	Correct Conclusion	Type II Error
	Reject H_o	Type I Error (α)	Correct Conclusion

Type one and type two error

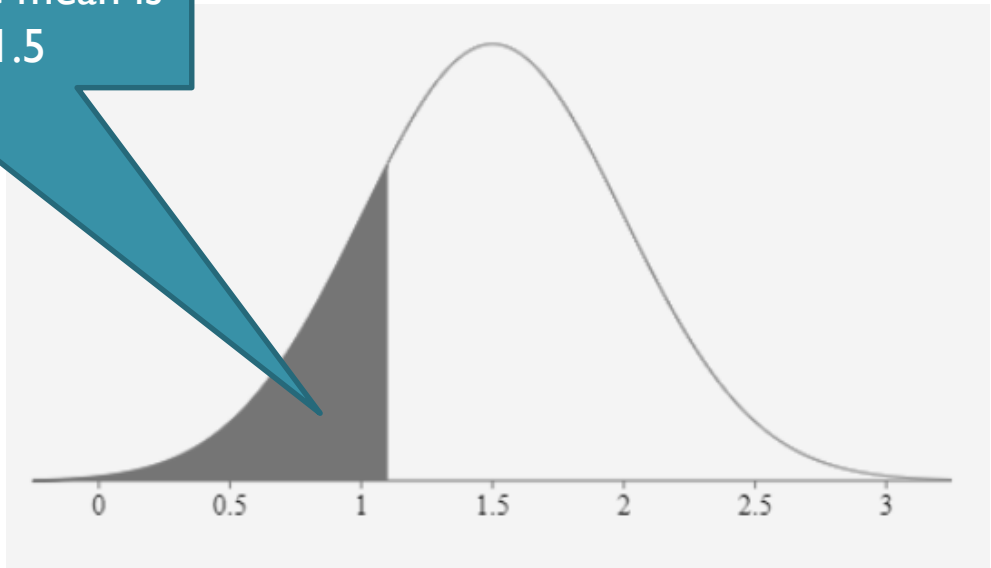
- Because a Type one error is defined as rejecting a true H0, and the probability of committing a Type one error is alpha

- *P* (rejecting Ho given that H0 is true) = α

Hypothesis test- 1

- I believe that the commute time between the university and the airport is larger than or equal 1.5 hours.
- Suppose that our random sample of n = 25 students and their average commute time is 1.1hours.
- The alternative hypothesis might be that the commute time is less than 1.5 hours.

You are testing if sample mean is actually less than 1.5



- H0 $\mu \geq$ 1.5 hours
- H1 $\mu <$ 1.5 hours

Hypothesis test- 1

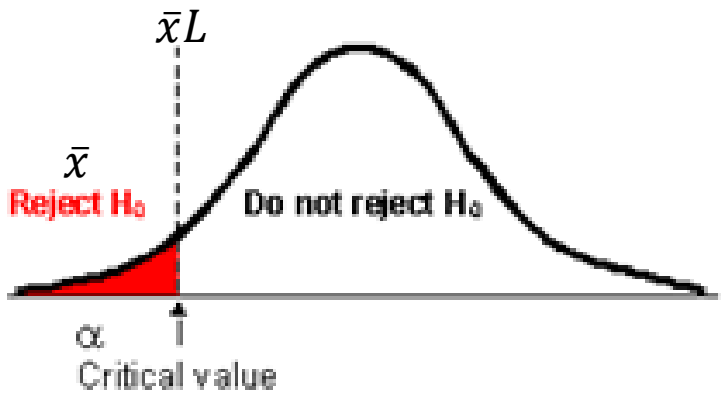
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Test to determine at the 5 % significance level whether there is enough statistical evidence to infer that the commute time is less than 1.5 hours.

Left-tailed test

H0 $\mu \geq$ 1.5 hours

H1 $\mu <$ 1.5 hours



P (rejecting Ho given that H0 is true) = α

$$p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{\bar{x}L-\mu}{\sigma/\sqrt{n}}\right) = \alpha$$

$$p\left(\frac{1.1-1.5}{0.5/\sqrt{25}} < z_\alpha\right) = 0.05$$

```
> qnorm(0.05)
[1] -1.644854
```

$$z = \frac{1.1 - 1.5}{0.5/\sqrt{25}} = -4 < -1.645$$

Problem

- A random sample of 25 sample NYUST students enrolled in a business statistics course was drawn. Each student was asked how many hours he or she spent doing homework in statistics.

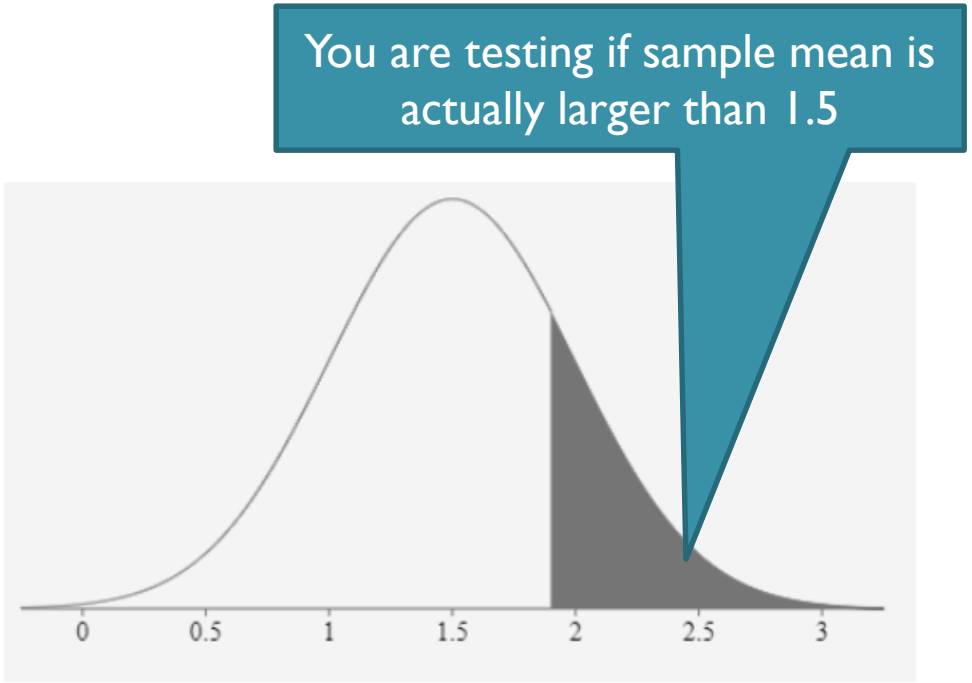
- The sample mean is 2 hours with σ = 0.6. Test to determine at the 5 % significance level whether there is enough statistical evidence to infer that the mean amount of doing homework by NYUST students is less than 3.5 hours ?

R programing

```
xbar <- 2
pmean <- 3.5
psd <- 0.6
n <- 25
z <- (xbar - pmean)/(psd/sqrt(n))
z
qnorm(0.05)
```

Hypothesis test- 2

- I believe that the commute time between the university and the airport to be less than or equal 1.5 hours.
- Suppose that our random sample is n = 25 students and their average commute time is 1.9 hours.
- The alternative hypothesis might be that the commute time is larger than 1.5 hours.

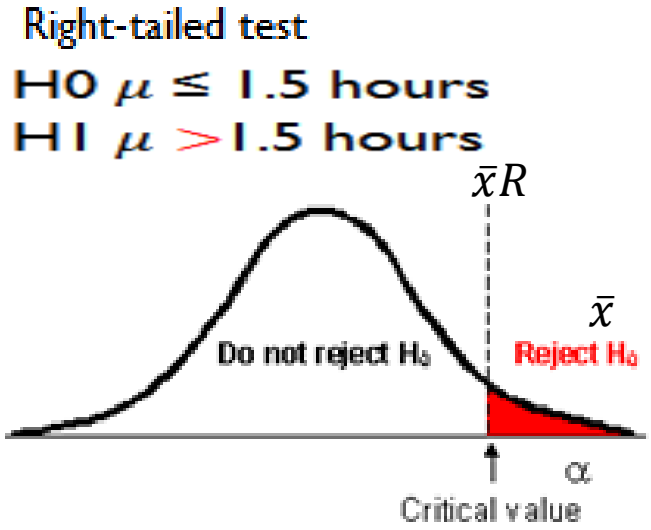


- H0 $\mu \leq$ 1.5 hours
- H1 $\mu >$ 1.5 hours

Hypothesis test- 2

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Test to determine at the 5 % significance level whether there is enough statistical evidence to infer that the commute time is greater than 1.5 hours.



P (rejecting Ho given that H0 is true) = α

$$p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} > \frac{\bar{x}R-\mu}{\sigma/\sqrt{n}}\right) = \alpha$$

$$p\left(\frac{1.9-1.5}{0.5/\sqrt{25}} > z_\alpha\right) = 0.05$$

```
> qnorm(1-0.05)
[1] 1.644854
>
```

$$z = \frac{1.9 - 1.5}{0.5/\sqrt{25}} = 4 > 1.645$$

Problem

- A random sample of 36 sample NYUST students was collected. Each student was asked how many minutes of sports he or she watched daily.

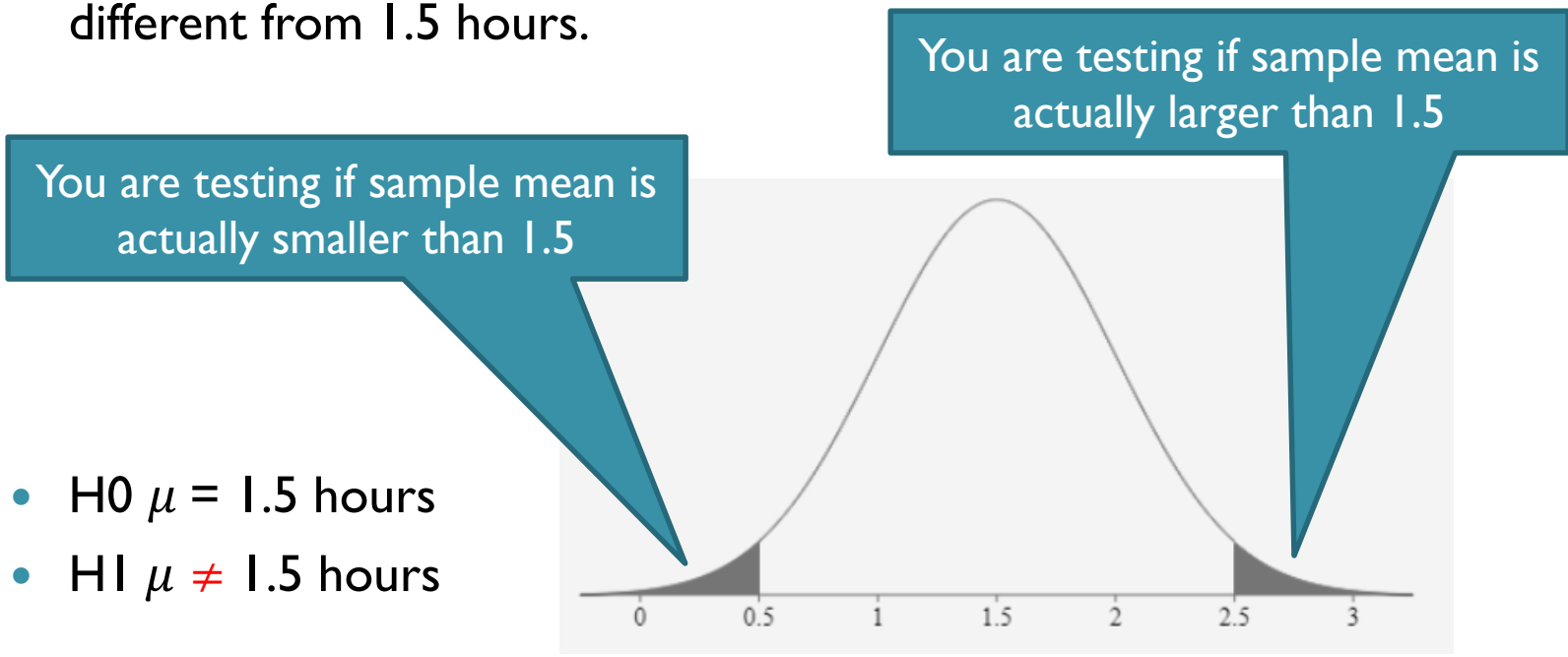
- Sample mean is 60 mins with σ = 10. Test to determine at the 5 % significance level whether there is enough statistical evidence to infer that the mean amount of sport TV watched daily by NYUST students is greater than 50 mins?

R programing

```
xbar = 60
pmean = 50
psd = 10
n = 36
z = (xbar - pmean)/(psd/sqrt(n))
z
qnorm(1-0.05)
```

Hypothesis test -3

- I believe that the commute time between the university and the airport is 1.5 hours.
- Suppose that our random sample is n = 25 students and their average commuting time is 1.6, which is not equal to 1.5 hours.
- The alternative hypothesis might be that the commute time is different from 1.5 hours.



- H0 μ = 1.5 hours
- H1 $\mu \neq$ 1.5 hours

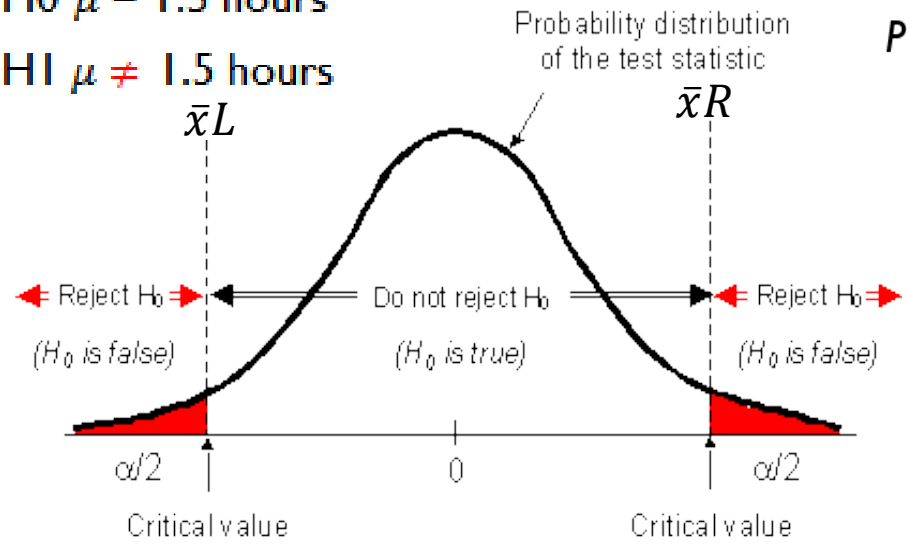
Hypothesis test -3

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Test to determine at the 5 % significance level whether there is enough statistical evidence to infer that the commute time is different from 1.5 hours.

H0 μ = 1.5 hours

H1 $\mu \neq$ 1.5 hours



P (rejecting Ho given that H0 is true) = α

$$p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} > \frac{\bar{x}R-\mu}{\sigma/\sqrt{n}}\right) = \alpha/2$$

$$p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{\bar{x}L-\mu}{\sigma/\sqrt{n}}\right) = \alpha/2$$

$$p\left(\frac{1.6-1.5}{0.5/\sqrt{25}} > z_\alpha\right) = 0.25$$

$$p\left(\frac{1.6-1.5}{0.5/\sqrt{25}} < z_\alpha\right) = 0.25$$

```
> qnorm(0.05/2)
[1] -1.959964
> qnorm(1-(0.05/2))
[1] 1.959964
```

$$-1.96 < z = \frac{1.6-1.5}{0.5/\sqrt{25}} = 1 < 1.96$$

Problem

- A machine that produce ball bearings is set that the average diameter is 0.5 inch. A sample of 16 ball bearings was measured, with the results shown here.

- The sample mean is 0.495 with with σ = 0.05. Can we conclude at the 5% significance level that the mean diameter is not 0.5 inch ?

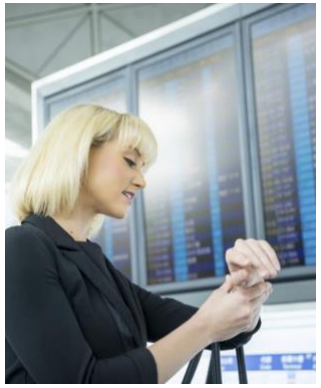
R programing

```
xbar <- 0.495
pmean <- 0.5
psd <- 0.05
n <- 16
z <- (xbar - pmean)/(psd/sqrt(n))
z
qnorm(0.05/2)
qnorm(1-(0.05/2))
```

Type one and type two error

- Suppose that the reality is that the null hypothesis is true – the true mean is the communing time larger than 1.5.

H0 $\mu \geq$ 1.5 hours
H1 $\mu <$ 1.5 hours



	when H0 is true	when H1 is true
Do not Reject H0	correct decision $p = 1 - \alpha$	Type II error $p = \beta$
Reject H0	Type I error $p = \alpha$	correct decision $p = 1 - \beta$

Hypothesis testing

Step 1 State the null and alternative hypotheses.
Step 2 Decide on the significance level, α.
Step 3 Compute the value of the test statistic.
Step 4 Determine the critical value(s).
Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.
Step 6 Interpret the result of the hypothesis test.

The common critical values in Z test

significance level (α)	Tail	Critical value
0.01	Two-tailed	± 2.575
	Right-tailed	2.33
	Left-tailed	-2.33
0.05	Two-tailed	±1.96
	Right-tailed	1.645
	Left-tailed	-1.645
0.1	Two-tailed	±1.645
	Right-tailed	1.28
	Left-tailed	- 1.28

Where are we and where are we going ?