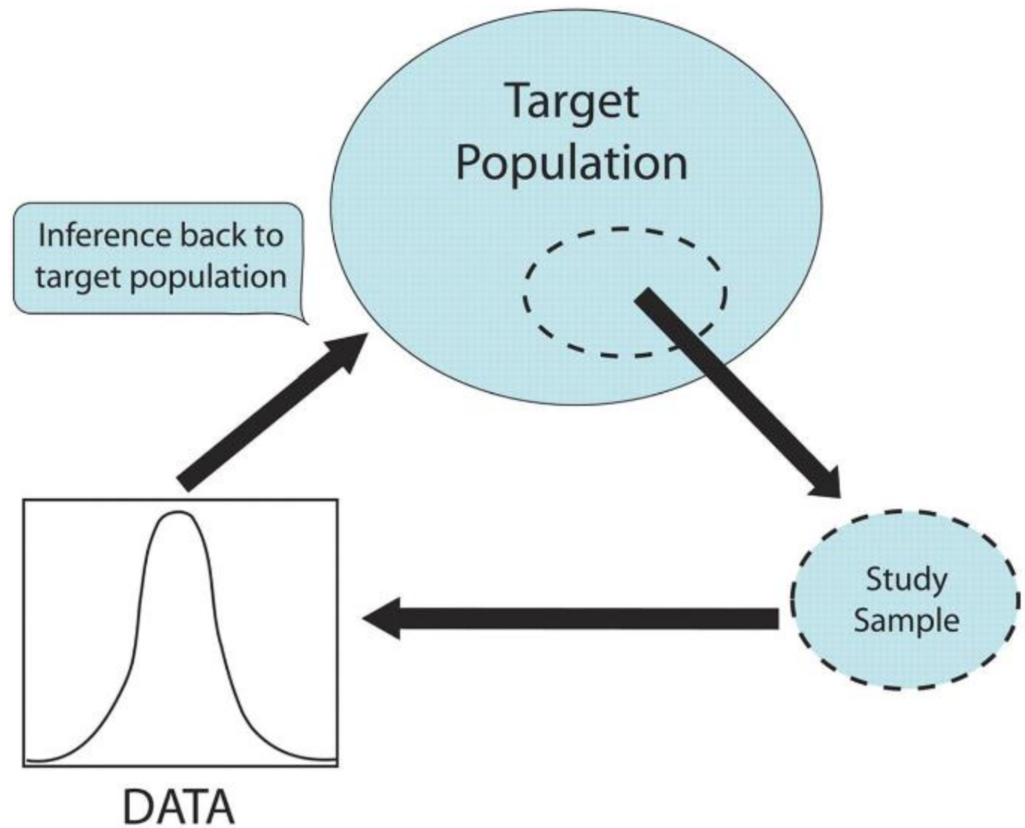
Welcome


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

Getting a grasp on data

Populations and Samples

Making use of data (inference)

- Estimation

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(μ) is 1.5 hours
- Population standard deviation(σ) 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 2.5 hours ?
  - What is the probability that the commute time is *more* than 2 hours ?

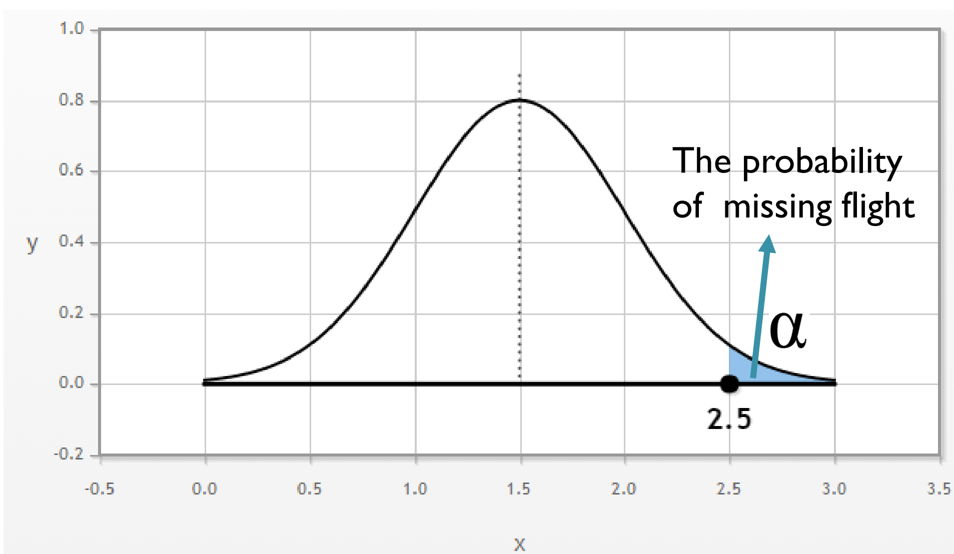
So what is the result ? What is your decision?

Case 1 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



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

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

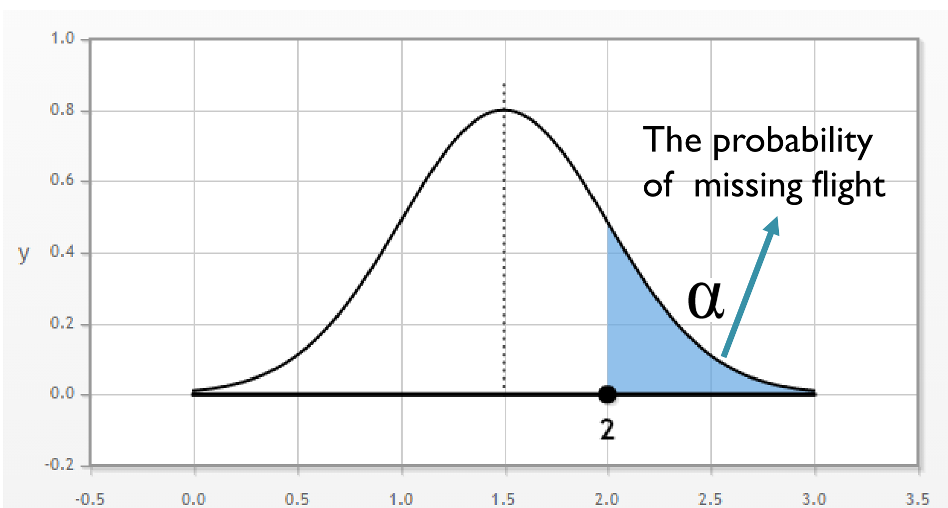
So what is the result ? What is your decision?

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



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

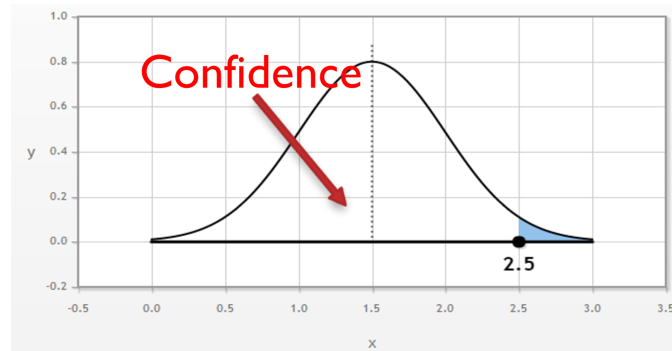
$$P(Z > 1) = 0.1587$$

Which one is correct ?

- You make a decision to leave the university around 1:30 pm. What is the probability that the commute time is more than 2.5 hours ?

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

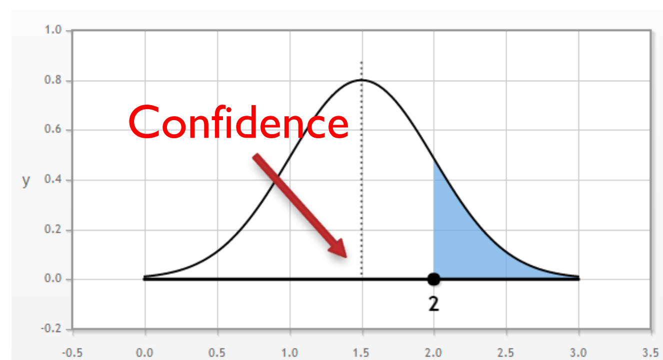
$$P(Z > 2) = 0.0228$$



- You make a decision to leave the university around 2:00 pm. What is the probability that the commute time is more than 2 hours ?

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

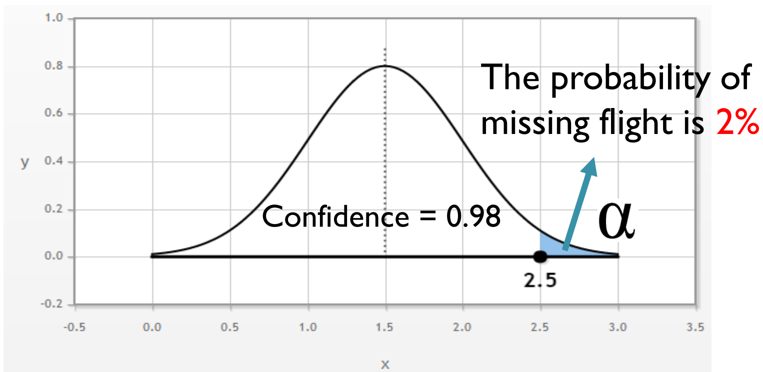
$$P(Z > 1) = 0.1587$$

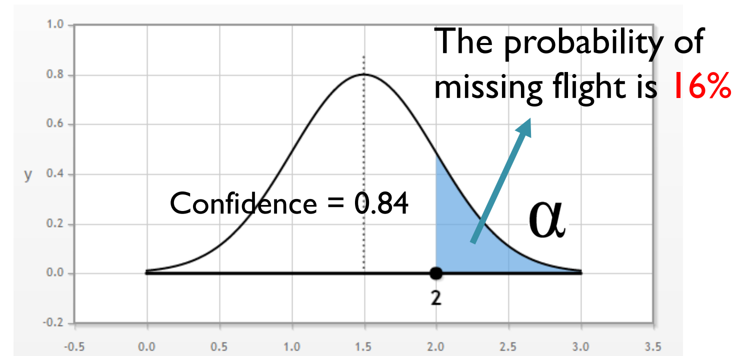


How do you make a decision

- Should I leave at
  - 1:30 pm and take 2% chance to make wrong decision
  - 2:00 pm and take 16% chance to make wrong decision

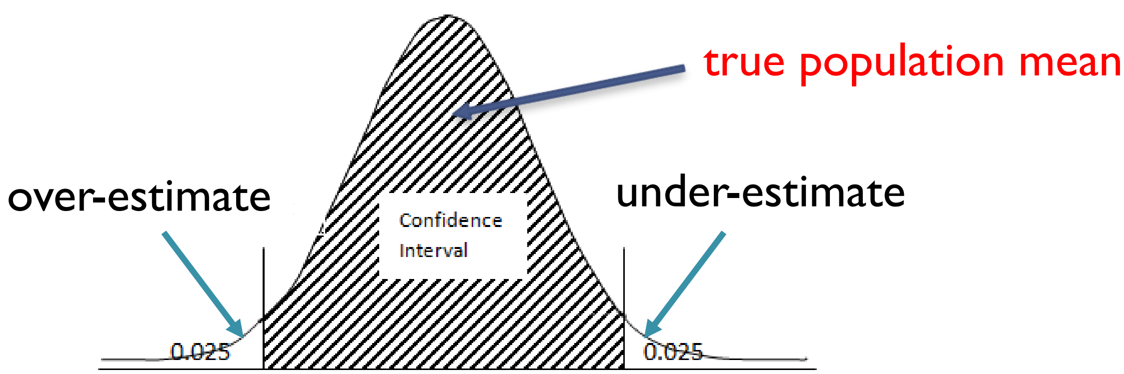
The answer is





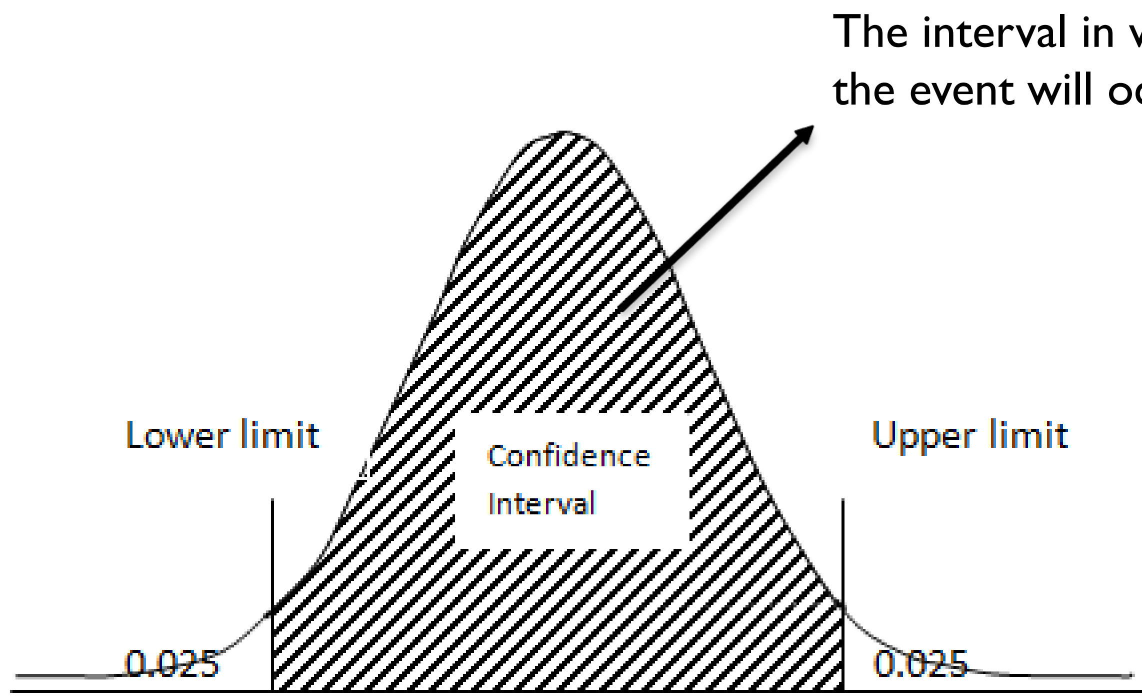
Confidence

- A Confidence Interval is an interval of numbers containing the true population mean
- When we are doing estimation, we have chances to over-estimate or under-estimate the true population mean
  - Therefore, the confidence interval is always two-tailed and the width is changing according to confidence level



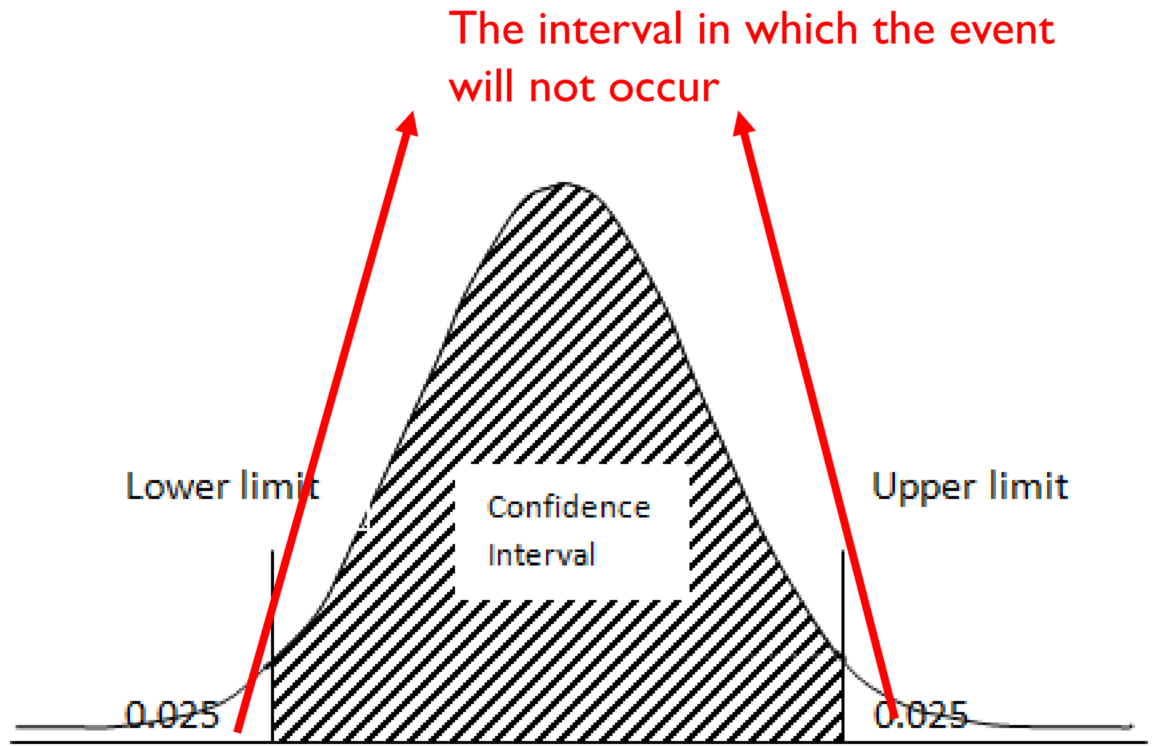
Confidence level

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



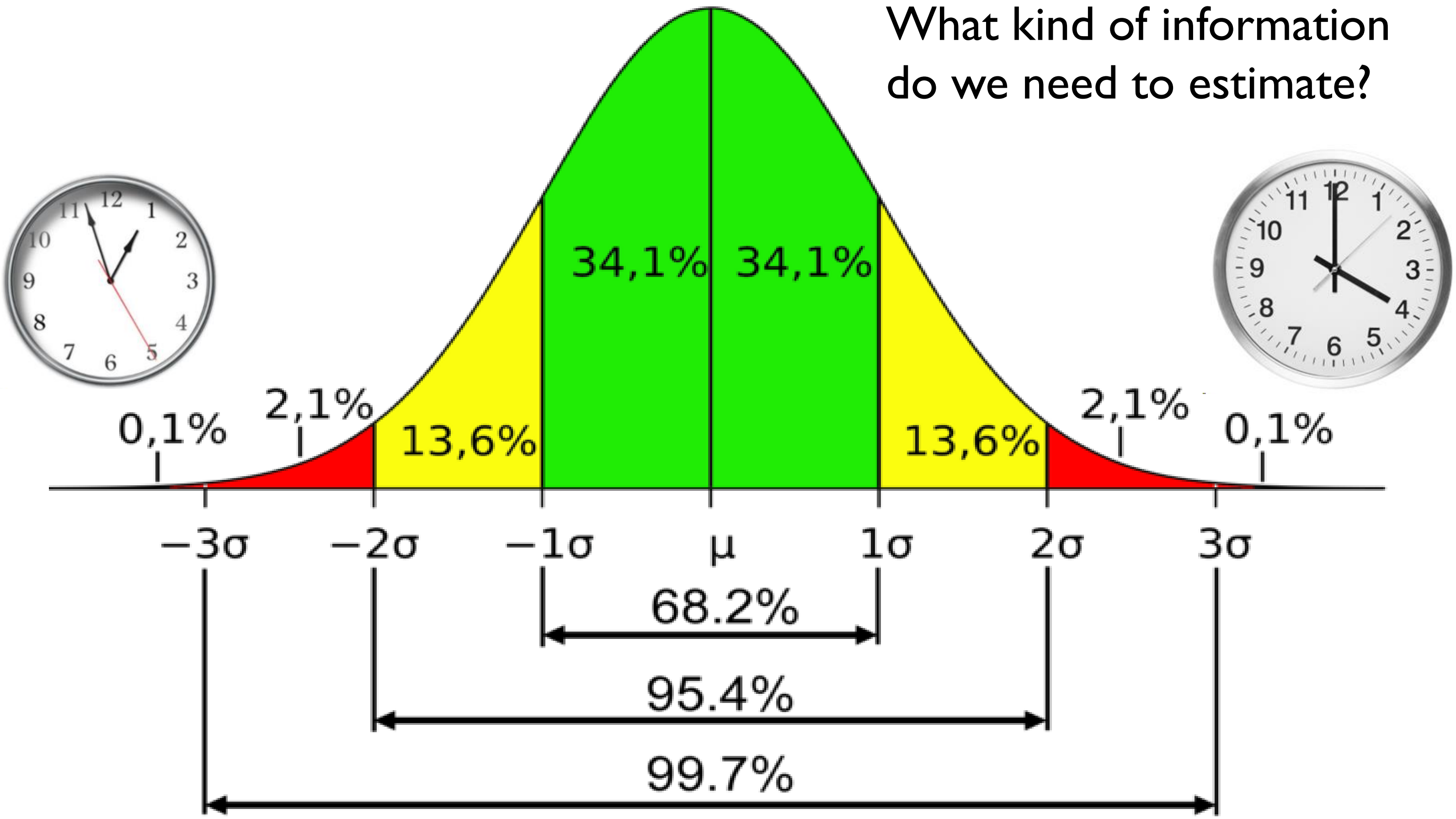
Confidence level

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



Normal distribution- Empirical rule

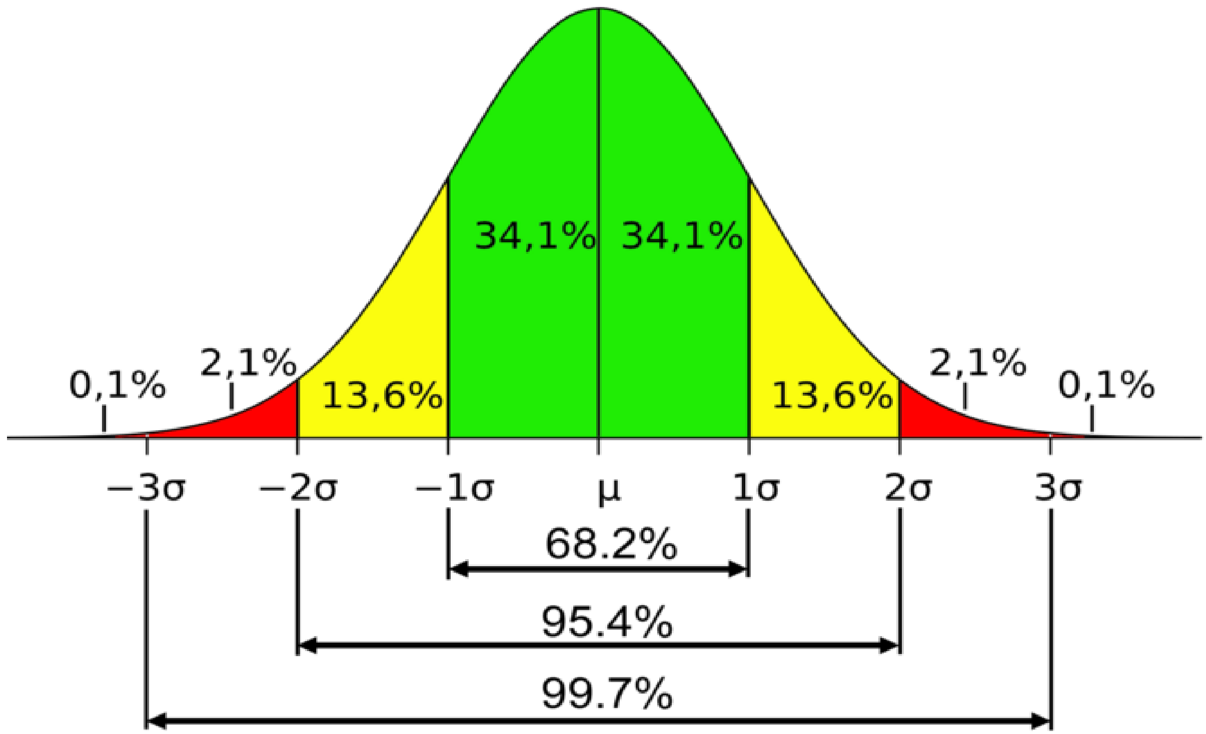
What kind of information do we need to estimate?



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

1.5hours

Confidence level

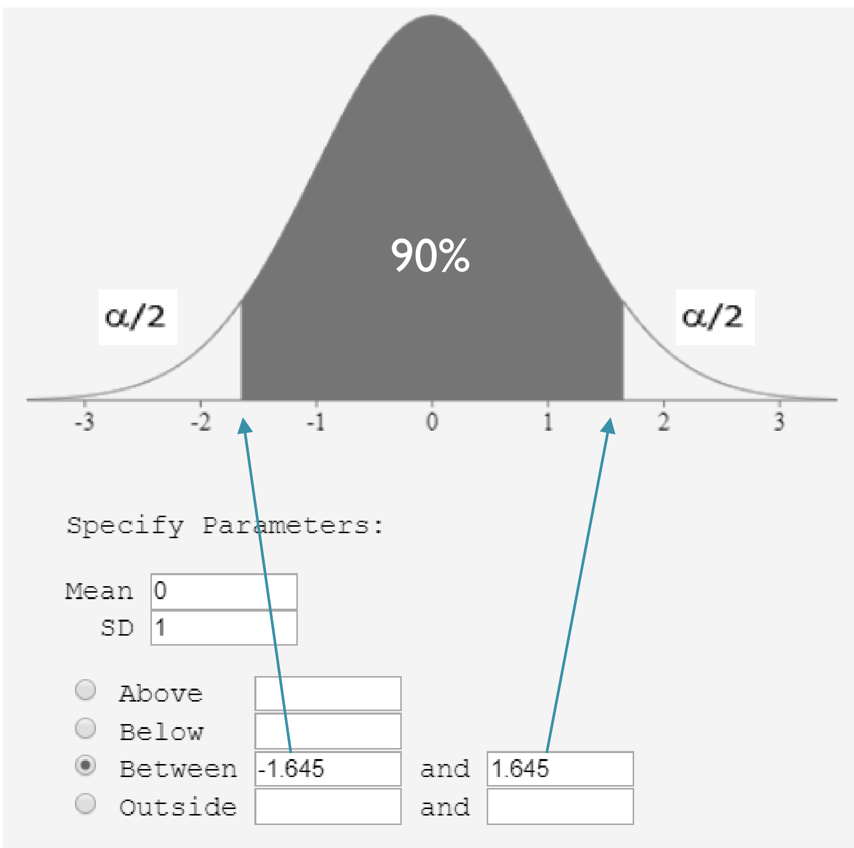


- P ($\mu - 1\sigma \le \mu \le \mu + 1\sigma$) = 0.682
- P ($\mu - 2\sigma \le \mu \le \mu + 2\sigma$) = 0.954
- P ($\mu - 3\sigma \le \mu \le \mu + 3\sigma$) = 0.997

```
> 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

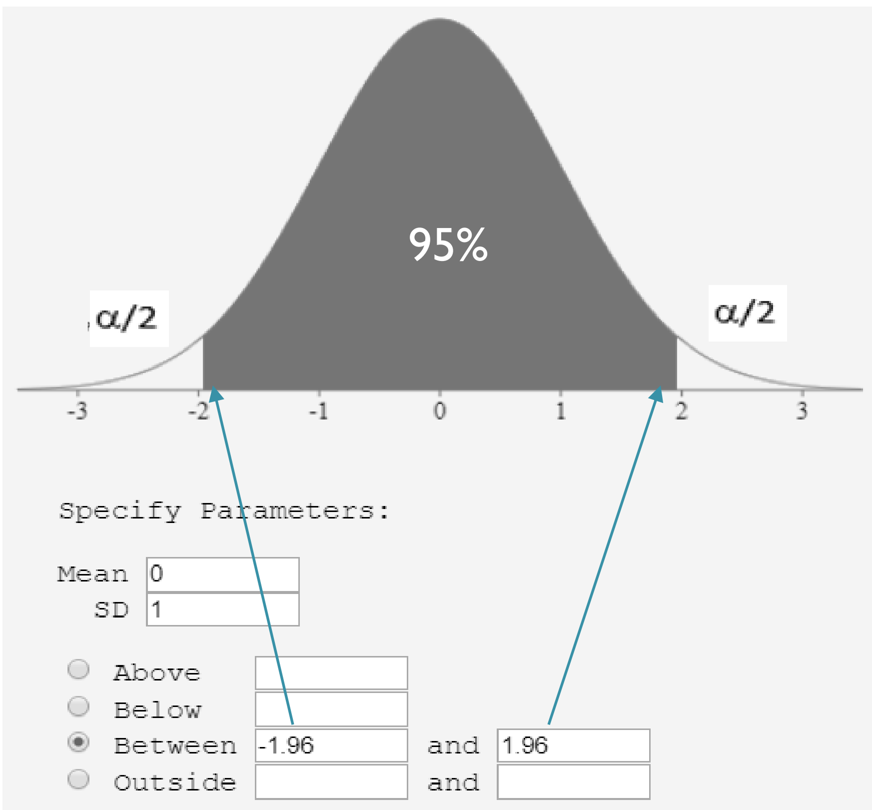


Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

```
> 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

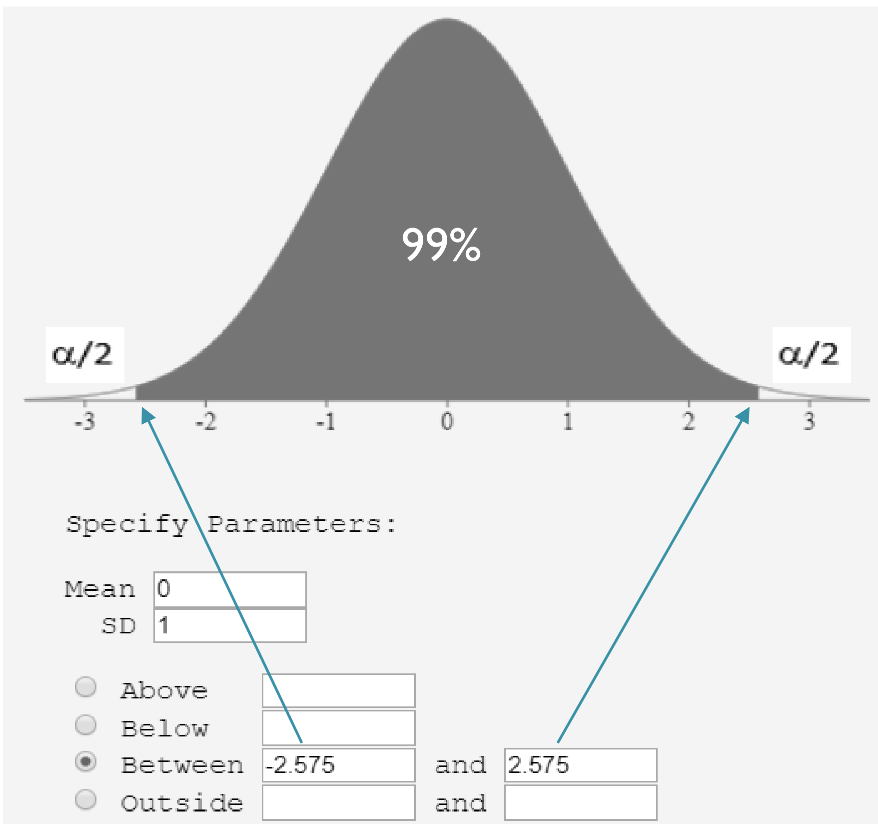


Level of Confidence	z_c
90%	1.645
95%	1.96
99%	2.575

```
> 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

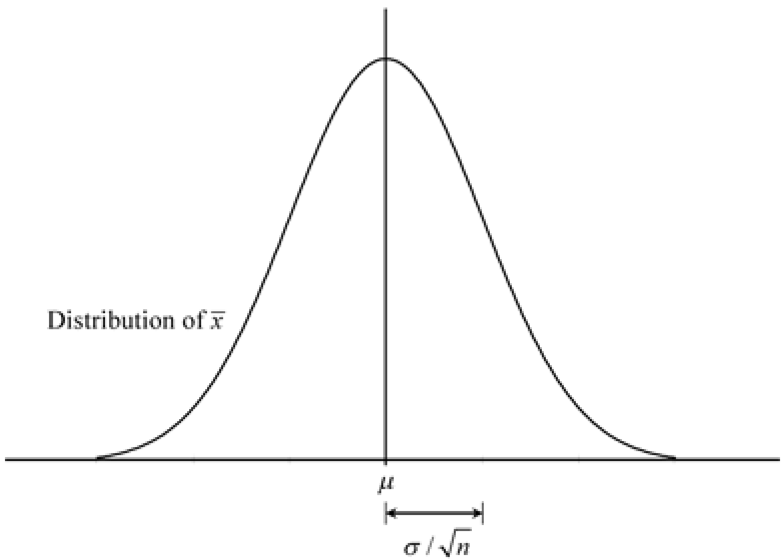
P $(\mu - 2.575\sigma \leq \mu \leq \mu + 2.575\sigma)$ = 0.99



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 ($\mu - 1.645\sigma \le \mu \le \mu + 1.645\sigma$) = 0.90

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

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

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

We use sample mean to estimate population mean

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

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

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

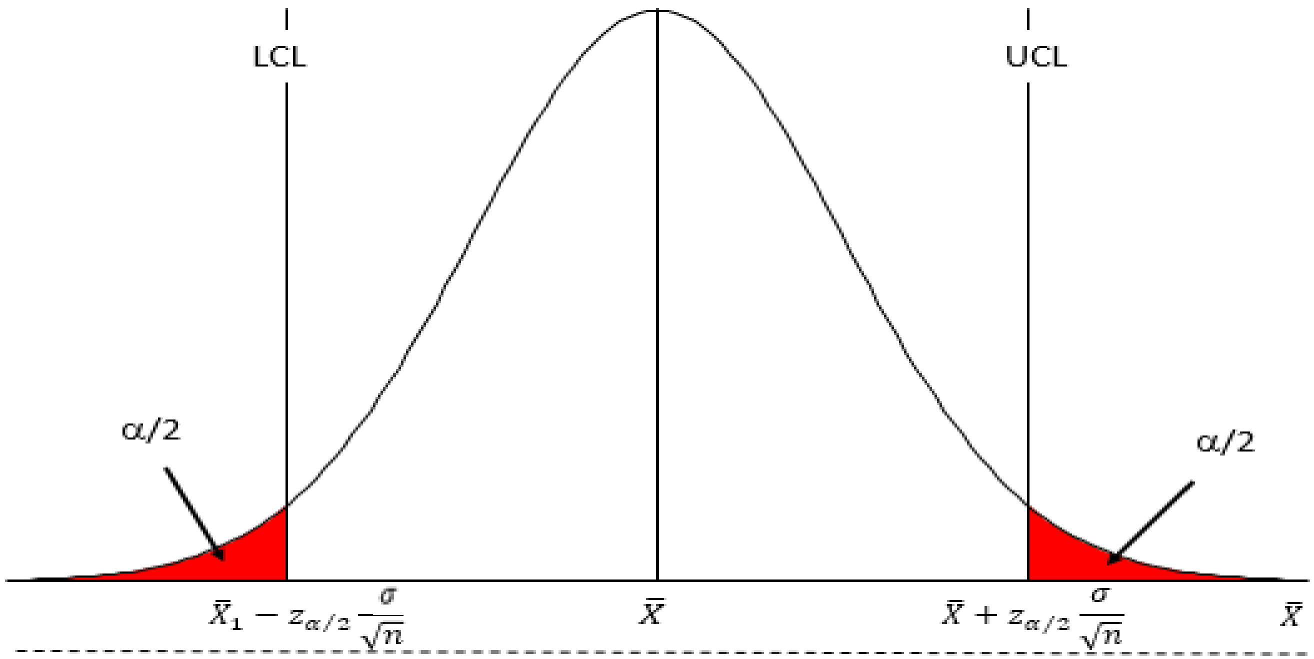
$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

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

Four basic elements in confidence interval

- Sample mean($\bar{x}$)
- Population variance(σ^2)
- Population standard deviation(σ)
- Confidence level
- Sample size(n)

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
```

Additional exercise- confidence level

- Suppose the insurance company randomly select 36 insured persons. The average age of 36 insured persons is 39.5 years old. The known standard deviation of the population is 7.2 years old. What is the 95% confidence interval of the population mean μ?

```
> xbar<-39.5
> psd<-7.2
> n<-36
> se<-abs(qnorm(0.05/2)*psd/sqrt(n))
> lcl<-xbar-se
> ucl<-xbar+se
> ci<-c(lcl,ucl)
> ci
[1] 37.14804 41.85196
```

Additional exercise- confidence level

- Suppose the factory randomly draws 16 canned peaches. The average weight of 16 canned peaches is 5.5. We know that the standard deviation of population is 0.065. Assuming that the peach canned weight follows the Normal distribution, What is the 99% confidence interval of the population mean μ?

```
> xbar<-5.5
> psd<-0.065
> n<-16
> se<-abs(qnorm(0.01/2)*psd/sqrt(n))
> lcl<-xbar-se
> ucl<-xbar+se
> ci<-c(lcl,ucl)
> ci
[1] 5.458143 5.541857
```

What is the error of confidence interval

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

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)
```

Additional exercise-sample size

- A random survey of 36 people driving to work, the average age and standard deviation of the car is 2.6 and 0.3. How many samples do we need to have a 95% confidence level so that the error in the estimate of the population mean does not exceed 0.05?

```
> mu<-2.6
> psd<-0.3
> B<-0.05
> ((qnorm(0.05/2)*psd)/B)^2
[1] 138.2925
```

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

Where are we and where are we going ?

Getting a grasp on data

Populations and Samples

Making use of data (inference)

- Estimation
- Hypothesis Testing

Hypotheses

- Null hypothesis.
  - denoted by H0, is usually the hypothesis that sample observations result purely from chance.
  - We regard the claim of people (enterprises, institutions) as H0, ex. Toothpaste manufacturers claim that their market share is over 40%(H0:p>0.4)

- Alternative hypothesis.
  - denoted by H1, is the hypothesis that sample observations are influenced by some non-random cause.
  - Verification claim ex. Is the rating of the TV station higher than 0.45?(H1:p≥0.45)

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?

EXAMPLES

- The mini shovel produced by the car squad claims to run at least 18.2 kilometers per liter of gasoline. This is very attractive for many small car drivers. Suppose Mr. Zhang wants to buy one, but it is not really so fuel efficient. So he inquired about the friend who bought the car to check whether the car can run at least 18.2 kilometers per liter of gasoline. Please help him set up a hypothesis for hypothesis testing.

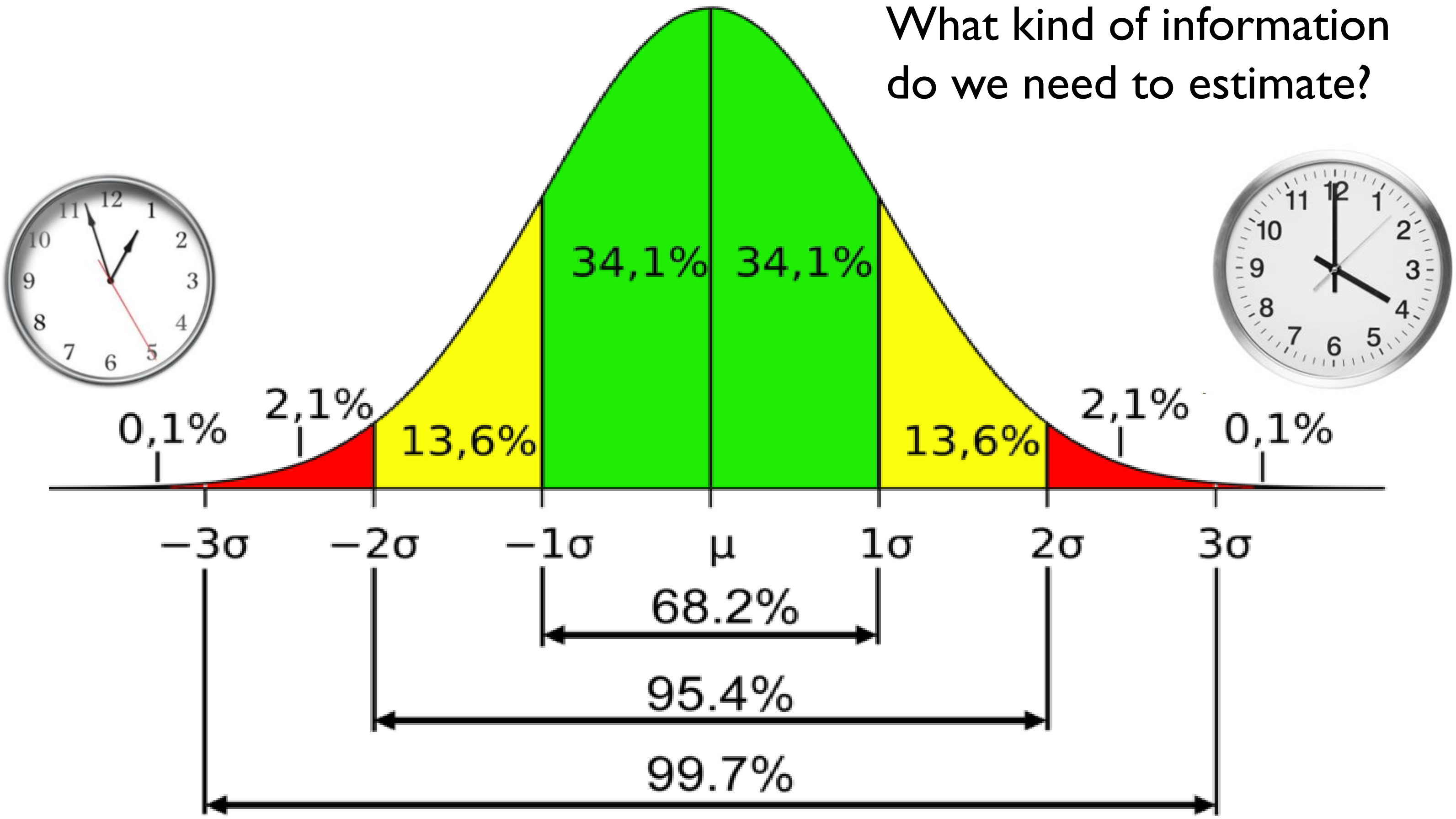
- The latest LCD version of the notebook has a standard length of 10 inches. Too long or too short does not work. The computer company purchased 49,000 display boards, and the quality control department was ordered to do the verification. How should the quality management department assume?

How long does it to get to the airport

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

Normal distribution- Empirical rule

What kind of information do we need to estimate?



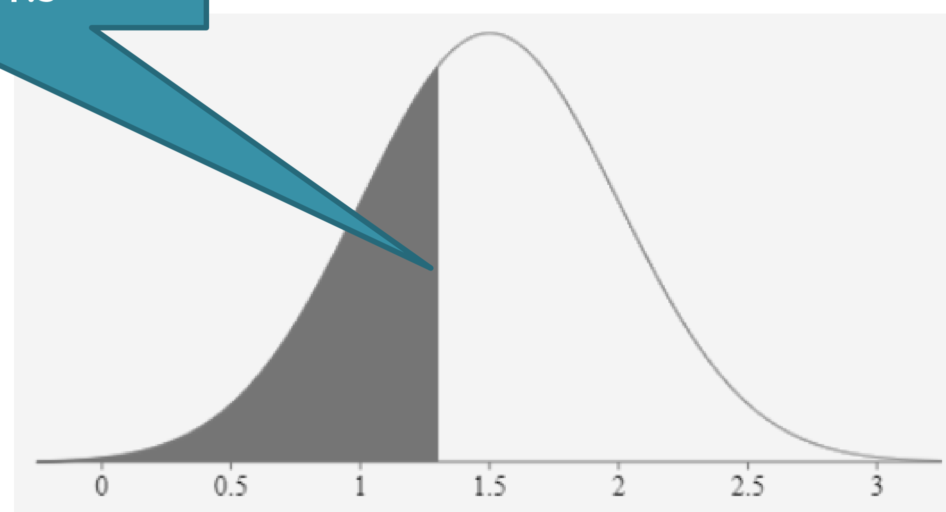
0 hour **0.5 hour** **1 hour** **2hours** **2.5hours** **3 hours**

1.5hours

Left-tailed test

- I state 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 1.3hours.
- 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

Again !

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

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

How to reject the null hypothesis

Recall you memory about

$$\alpha$$

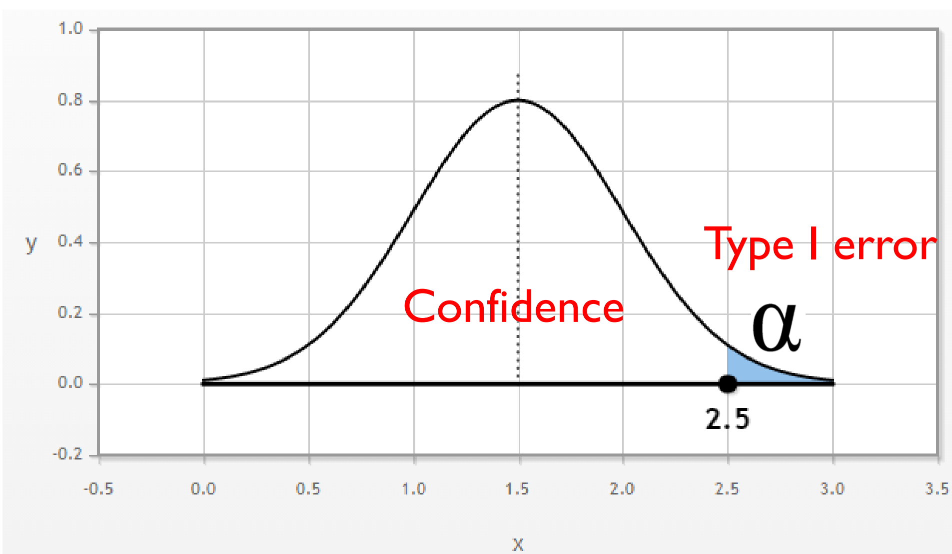
What is your decision?

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



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

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

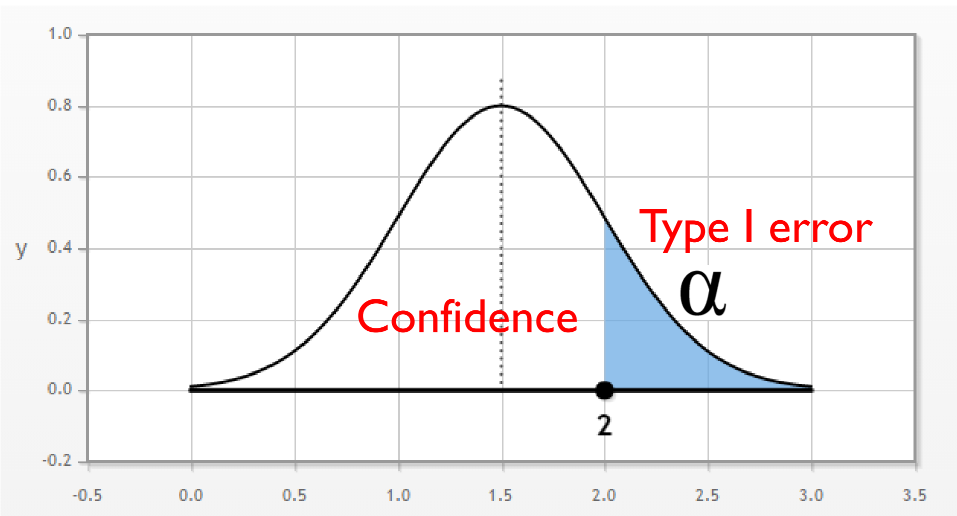
What is your decision?

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



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

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

Type one and type two error

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

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

The probability of getting type I error will be the α level

Type one and type two error

Figure 1		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

- 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
- Some commonly used significance level include 0.1, 0.05, 0.01.

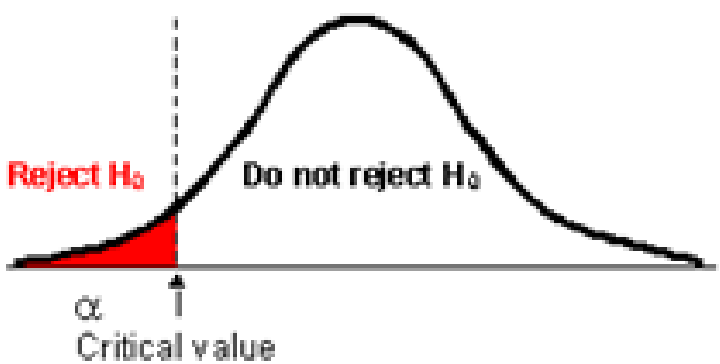
How to reject the null hypothesis

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

Left-tailed test

H0 $\mu \geq$ 1.5 hours

H1 $\mu <$ 1.5 hours



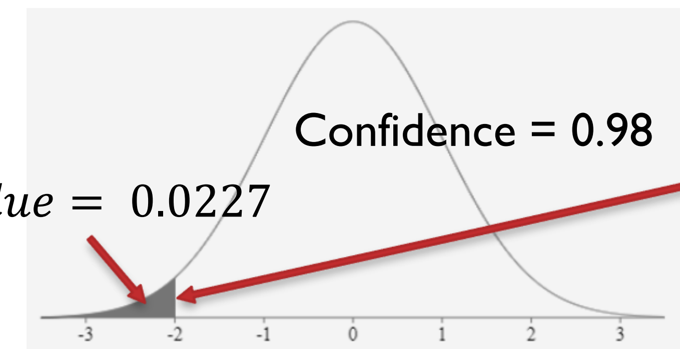
1. We try to prove the commute time is less than 1.5 hours
2. When we can have sufficient evidence to reject H0 ?
3. Critical value is the threshold

Hypothesis test- left-tailed test

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.

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

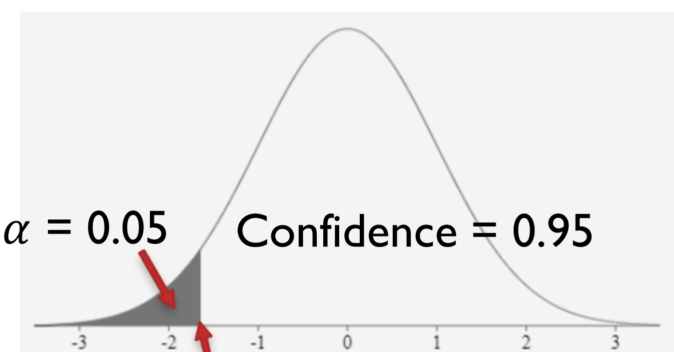
$p\ value =\ 0.0227$



z = $\frac{1.3 - 1.5}{0.5/\sqrt{25}} = -2$

```
> pnorm(-2)
[1] 0.02275013
```

α = 0.05



$critical\ value$ = -1.645

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

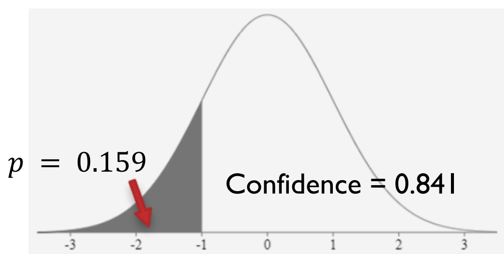
$p\ (\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < crtical\ value\ ) = \alpha$

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

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

Case 1
Sample mean is 1.4 hours

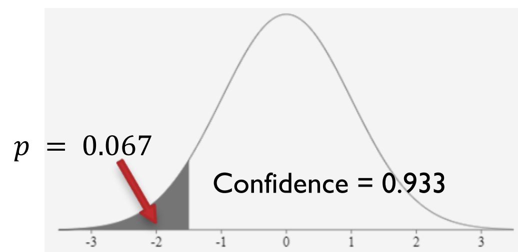
$$z = \frac{1.4 - 1.5}{0.5/\sqrt{25}} = -1$$




```
> pnorm(-1)
[1] 0.1586553
```

Case 2
Sample mean is 1.35 hours

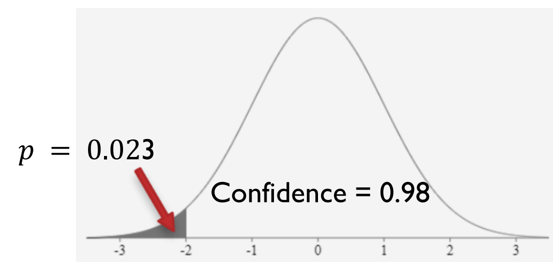
$$z = \frac{1.35 - 1.5}{0.5/\sqrt{25}} = -1.5$$



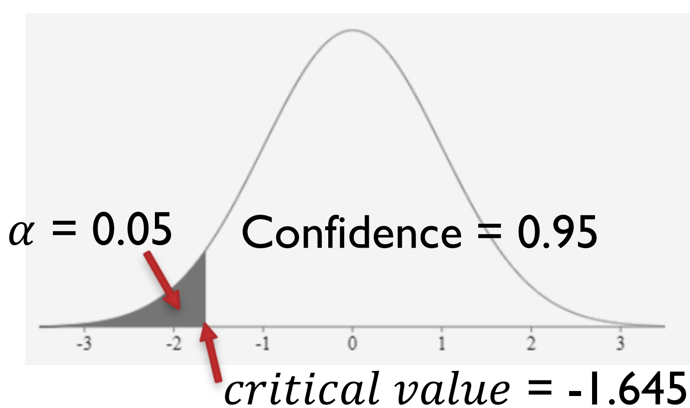

```
> pnorm(-1.5)
[1] 0.0668072
```

Case 3
Sample mean is 1.3 hours

$$z = \frac{1.3 - 1.5}{0.5/\sqrt{25}} = -2$$




```
> pnorm(-2)
[1] 0.02275013
```



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

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

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 ?

第一步

- 設定虛無和對立假設
  - 虛無假設通常來自某個人、事或物的宣稱與假定
  - 對立假設通常來自研究者想驗證的事件
  - 對立假設必定挑戰虛無假設

- 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 ?
  - 題目請你推論是否能驗證雲科學生每周花少於3.5小時寫作業
  - 因此，H1為 H1: $mu < 3.5$
  - 然而對立假設必定挑戰虛無假設
  - 因此，H0為 H0: $mu \geq 3.5$

  - 你在找證據希望能證實學生真的花少於3.5小時在作業上

第二步

- 收集題目給定的型一誤差
- 並記住型一誤差的定義是當H0為真，而你卻誤判拒絕了 H0
- 型一誤差通常是給定的條件，所以不需要自己判斷，商業統計常用的為 1%, 5% 和10%
- Test to determine at the 5 % significance level

- 因此本題的Alpha為 0.05

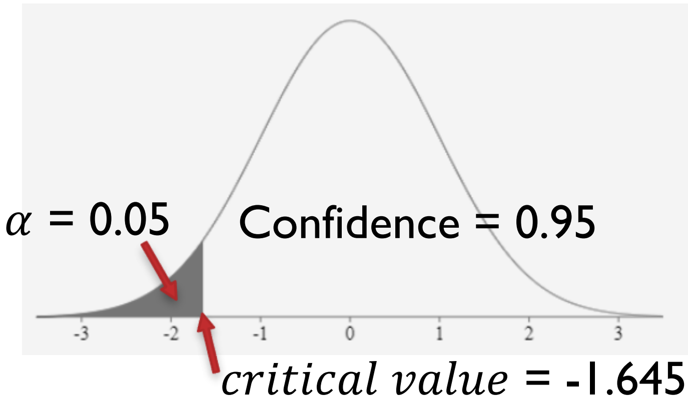
第三步

- 收集題目中所給的資訊
- 25 sample NYUST students
- The sample mean is 2 hours
- σ = 0.6.

- 並將所收集到的樣本平均數轉換成Z值，以便能進行假設檢定，公式請參考抽樣分配。

```
> z <- (xbar - pmean)/(psd/sqrt(n))
> z
[1] -12.5
```

第四步



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

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

當我們設定好Alpha為0.05 時可以分成以下兩種方式來作判別

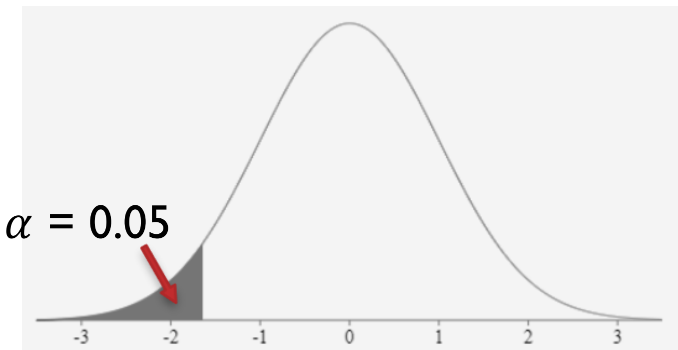
1. p-value (此為通用統計學方法，且實務上多數採用這方式)
2. Z-critical value (此為教科書方法，實務上不太可能採用這方式)

再問你自己一次，你在做什麼題目?

你在找證據希望能証實學生真的花少於3.5小時在作業上，那學生到底要少於3.5小時多少? 才能讓你有充足的證據說 "學生真的花少於3.5小時在作業上"?

第五步- p-value

- 當你轉換成Z值，你才有辦法找機率

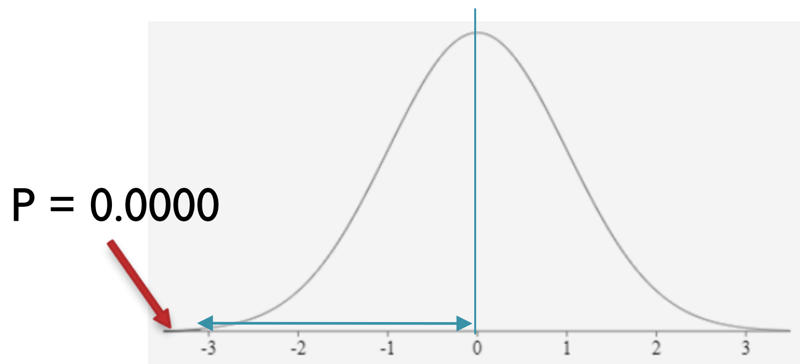



```
> z <- (xbar - pmean)/(psd/sqrt(n))
> z
[1] -12.5
> Pvalue<-pnorm(z)
> Pvalue
[1] 3.732564e-36
```

本題中，你找到的機率非常小，並**不是**表示你找到的犯錯錯誤機率很小，而是表示你樣本平均數與Mu有足夠的差異(距離)

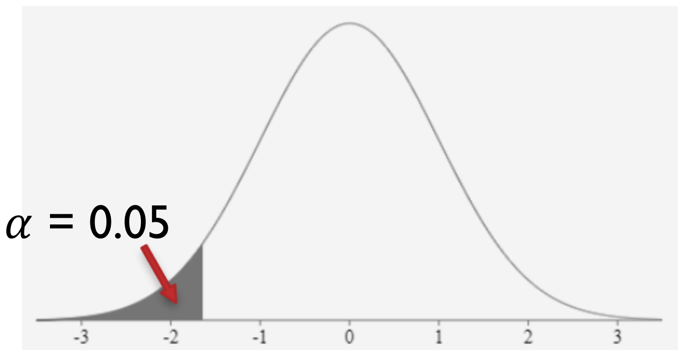
反之，若你找到的機率非常大，並**不是**表示你找到的犯錯錯誤機率很大，而是表示你樣本平均數與Mu沒有足夠的差異(距離)

第五步- p-value



本題中，你找到的機率非常小，並**不是**表示你找到的犯錯錯誤機率很小，而是表示你樣本平均數與Mu有足夠的差異(距離)，當足夠的差異呈現出來時(p< Alpha)，你則有足夠的證據下結論，學生真的花少於3.5小時在作業上，因此你拒絕H0。

第五步- Z-critical value



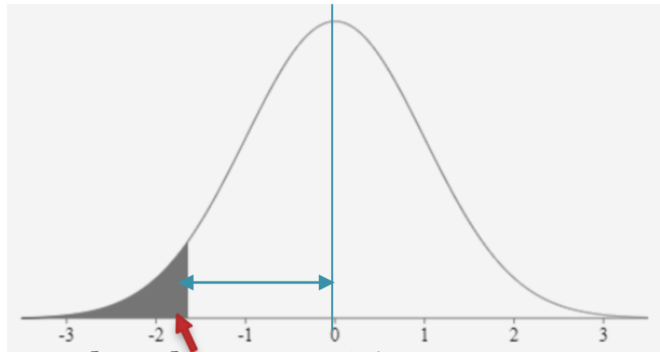

```
> CV<-qnorm(0.05)
> CV
[1] -1.644854
```


```
> z <- (xbar - pmean)/(psd/sqrt(n))
> z
[1] -12.5
```

本題中，你找到的Z 值為-12.5 ，根據 Alpha，你可以反推左尾檢測是的臨界值是 -1.645

這也說明，你樣本平均數要與Mu有足夠的差異(距離)，當足夠的差異呈現出來時(Z< 臨界值= -1.645)，你則有足夠的證據下結論，學生真的花少於3.5小時在作業上，因此你拒絕H0。

第五步- Z-critical value



$critical\ value$ = -1.645

z = -12.5

本題中，你找到的Z值遠比臨界值還小。當足夠的差異呈現出來時(Z< 臨界值= -1.645)，你則有足夠的證據下結論，學生真的花少於3.5小時在作業上，因此你拒絕H0。

第六步

當我們設定好Alpha為0.05 時可以分成以下兩種方式來作判別

1. p-value (此為通用統計學方法，且實務上多數採用這方式)
2. Z-critical value (此為教科書方法，實務上不太可能採用這方式)

但不論你用哪一種方法，最後的答案都是一樣的。

Answer:

When α = 0.05, we have sufficient evidence to reject H0. Therefore, we conclude that the mean amount of doing homework by NYUST students is less than 3.5 hours

R programing

```
#H0 mu ≥ 3.5 H1: mu < 3.5
xbar <- 2
pmean <- 3.5
psd <- 0.6
n <- 25
Alpha<- 0.05
z <- (xbar - pmean)/(psd/sqrt(n))
z
CV<- qnorm(0.05)
CV
Pvalue<- pnorm(z)
Pvalue
Pvalue < Alpha
z < CV
```

```
> #H0 mu ≥ 3.5 H1: mu < 3.5
> xbar <- 2
> pmean <- 3.5
> psd <- 0.6
> n <- 25
> Alpha<- 0.05
> z <- (xbar - pmean)/(psd/sqrt(n))
> z
[1] -12.5
> CV<- qnorm(0.05)
> CV
[1] -1.644854
> Pvalue<- pnorm(z)
> Pvalue
[1] 3.732564e-36
> Pvalue < Alpha
[1] TRUE
> z < CV
[1] TRUE
```

p< Alpha 的判斷方式

Z< 臨界值的判斷方式

Additional exercise- Hypothesis test

- Manufacturers claim that the average coffee pot produced is more than 3 pounds.

- A total of 36 cans were randomly taken to measure the weight, resulting in an average weight of 2.97. Assuming the standard deviation of population is 0.18 pounds, the manufacturer's claim is verified? (significant level $\alpha = 0.01$)

R programing

```
#H0: μ ≥ 3 H1: μ < 3
xbar<- 2.97
pmean<- 3
psd<- 0.18
n<- 36
Alpha<- 0.01
z<- (xbar-pmean)/(psd/sqrt(n))
z
CV<- qnorm(0.01)
CV
Pvalue<- pnorm(z)
Pvalue
Pvalue < Alpha
z < CV
```

```
> #H0: μ ≥ 3 H1: μ < 3
> xbar<- 2.97
> pmean<- 3
> psd<- 0.18
> n<- 36
> Alpha<- 0.01
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] -1
> CV<- qnorm(0.01)
> CV
[1] -2.326348
> Pvalue<- pnorm(z)
> Pvalue
[1] 0.1586553
> Pvalue < Alpha
[1] FALSE
> z < CV
[1] FALSE
```

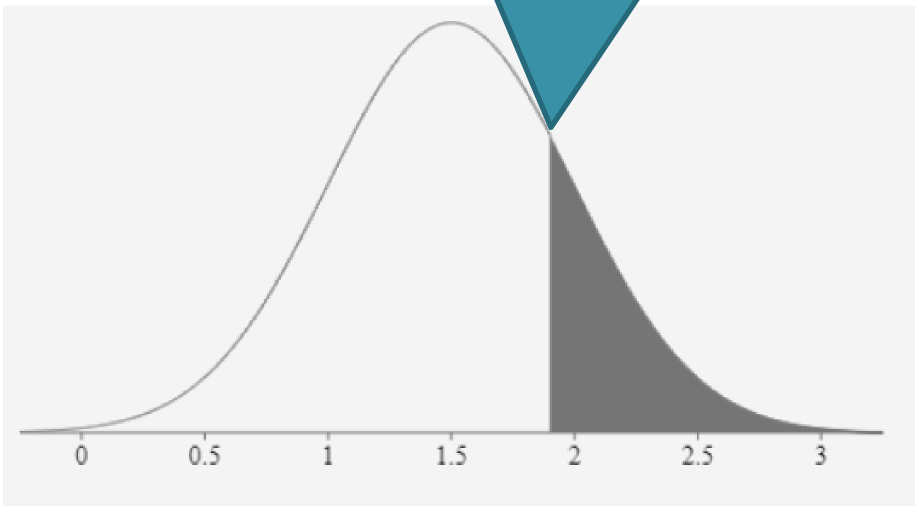
How long does it to get to the airport

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

Right-tailed test

- I state 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 1.75 hours.
- The alternative hypothesis might be that the commute time is larger than 1.5 hours.

You are testing if sample mean is actually larger than 1.5



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

How to reject the null hypothesis

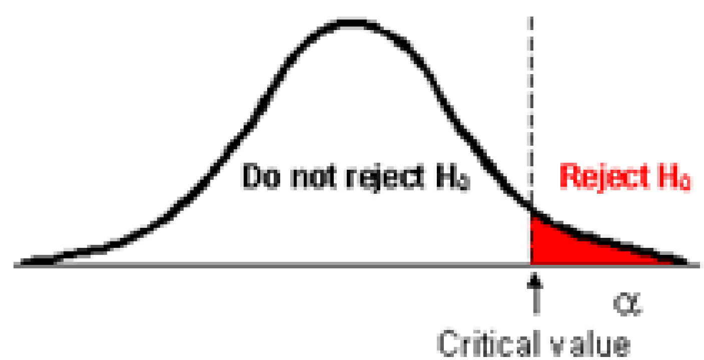
If we establish directional hypotheses, then the rejection region is allocated to right tail of the probability distribution

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

Right-tailed test

H0 $\mu \leq 1.5$ hours

H1 $\mu > 1.5$ hours



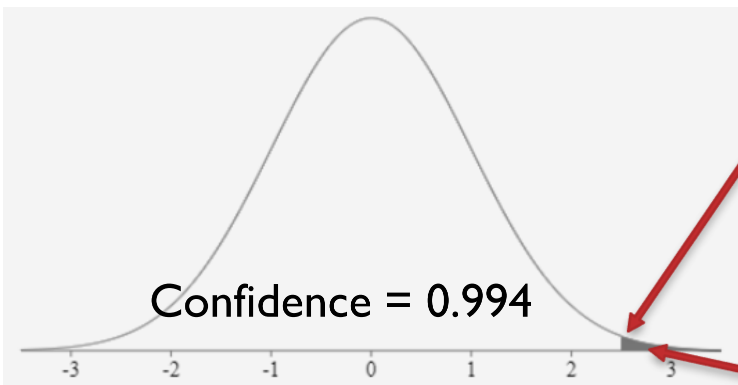
Hypothesis test-right-tailed test

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.

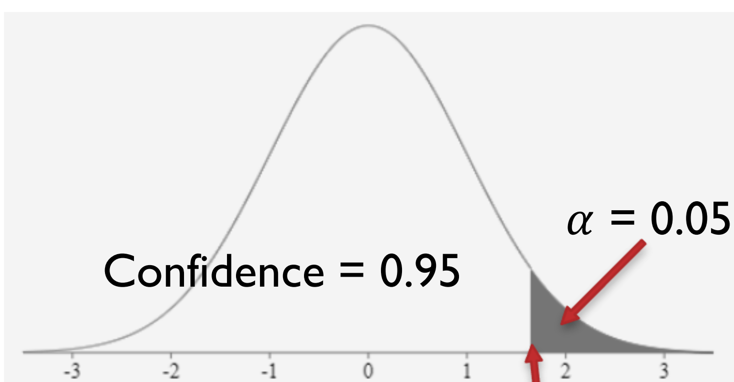
H0 $\mu \leq 1.5$ hours
H1 $\mu > 1.5$ hours

$$z = \frac{1.75 - 1.5}{0.5/\sqrt{25}} = 2.5$$

```
> pnorm(2.5, lower.tail = FALSE)
[1] 0.006209665
```



$p\ value = 0.006$



critical value = 1.645

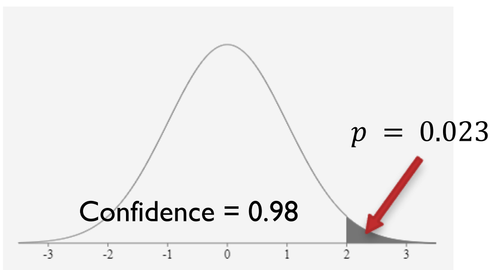
P (rejecting Ho given that H0 is true) = α

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

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

Case 1
Sample mean is 1.7 hours

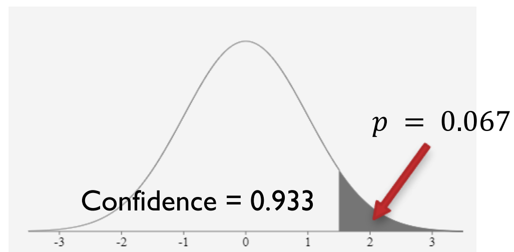
$$z = \frac{1.7 - 1.5}{0.5/\sqrt{25}} = 2$$




```
> pnorm(2,lower.tail = FALSE)
[1] 0.02275013
```

Case 2
Sample mean is 1.65 hours

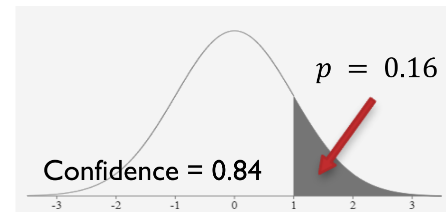
$$z = \frac{1.65 - 1.5}{0.5/\sqrt{25}} = 1.5$$



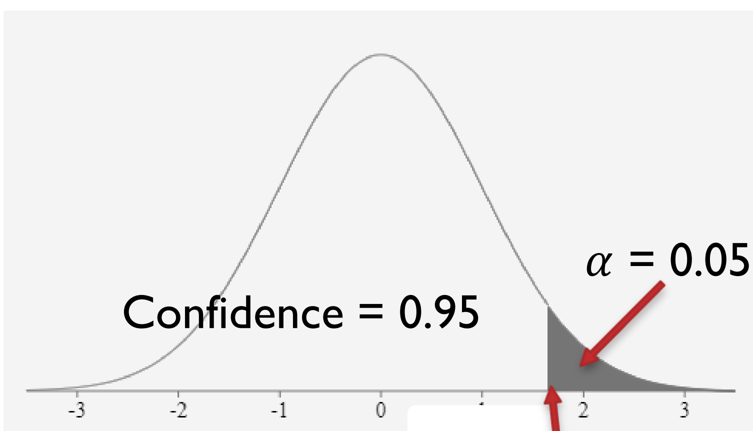

```
> pnorm(1.5,lower.tail = FALSE)
[1] 0.0668072
```

Case 3
Sample mean is 1.6 hours

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




```
> pnorm(1,lower.tail = FALSE)
[1] 0.1586553
```



critical value = 1.645

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

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

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 $\sigma = 10$. 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

```
#H0 : μ ≤ 50 H1 : μ > 50
xbar<- 60
pmean<- 50
psd<- 10
n<- 36
Alpha<- 0.05
z<- (xbar - pmean)/(psd/sqrt(n))
z
CV<-qnorm(1-0.05)
CV
Pvalue<-pnorm(z,lower.tail = FALSE)
Pvalue
Pvalue < Alpha
z > CV
```

```
> #H0 : μ ≤ 50 H1 : μ > 50
> xbar<- 60
> pmean<- 50
> psd<- 10
> n<- 36
> Alpha<- 0.05
> z<- (xbar - pmean)/(psd/sqrt(n))
> z
[1] 6
> CV<-qnorm(1-0.05)
> CV
[1] 1.644854
> Pvalue<-pnorm(z,lower.tail = FALSE)
> Pvalue
[1] 9.865876e-10
> Pvalue < Alpha
[1] TRUE
> z > CV
[1] TRUE
```

Additional exercise- Hypothesis test

- The average number of students in a primary school in the past was 120 cm. Today, 100 students were randomly selected from the new students, the average height was 123 cm and $\sigma^2 = 25$. Is the height of new students higher than before?($\alpha = 0.05$)

R programing

```
#H0 : μ ≤ 120, H1 : μ > 120
xbar<- 123
pmean<- 120
psd<- 5
n<- 100
Alpha<- 0.05
z<- (xbar-pmean)/(psd/sqrt(n))
z
CV<- qnorm(1-0.05)
CV
Pvalue<- pnorm(z,lower.tail = FALSE)
Pvalue
Pvalue < Alpha
z > CV
```

```
> #H0 : μ ≤ 120, H1 : μ > 120
> xbar<-123
> pmean<-120
> psd<-5
> n<-100
> Alpha<-0.05
> z<-(xbar-pmean)/(psd/sqrt(n))
> z
[1] 6
> CV<-qnorm(1-0.05)
> CV
[1] 1.644854
> Pvalue<-pnorm(z,lower.tail = FALSE)
> Pvalue
[1] 9.865876e-10
> Pvalue < Alpha
[1] TRUE
> z > CV
[1] TRUE
```

Two-tailed test

- I start 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

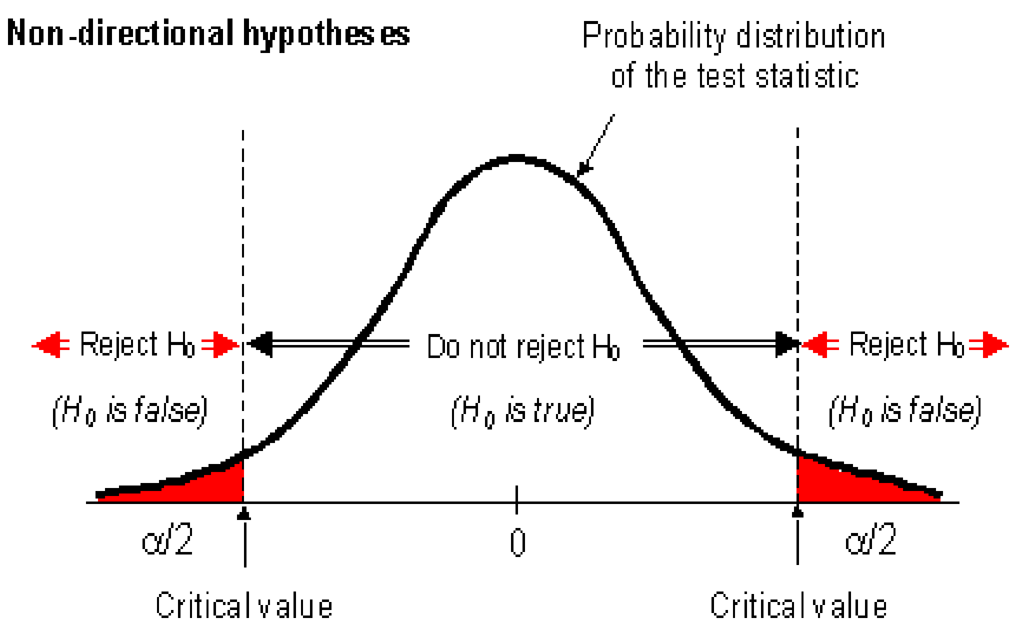
- H0 $\mu = 1.5$ hours
- H1 $\mu \neq 1.5$ hours

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

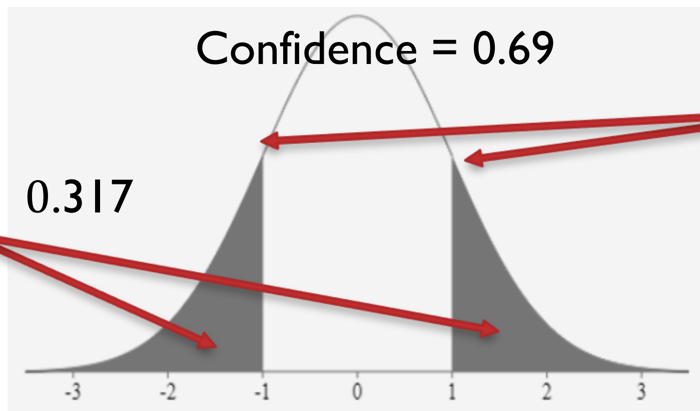
Hypothesis test -3

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

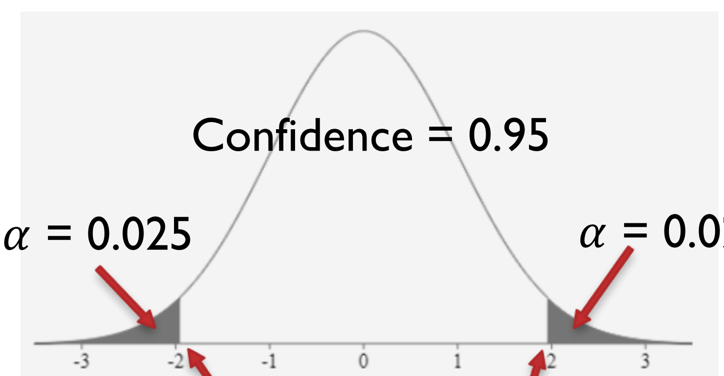
H1 $\mu \neq$ 1.5 hours

$p\ value =$ 0.317



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

```
> 2*pnorm(1, lower.tail = FALSE)
[1] 0.3173105
```



α = 0.025

α = 0.025

$critical\ value$ = -1.96

$critical\ value$ = 1.96

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

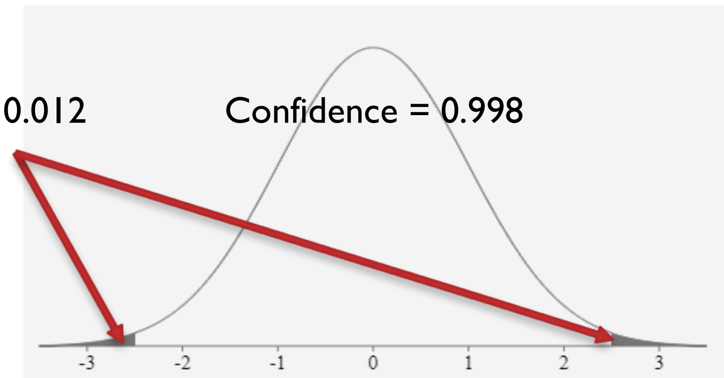
```
> qnorm(0.05/2)
[1] -1.959964
> qnorm(1-(0.05/2))
[1] 1.959964
```

H0 μ = 1.5 hours

H1 $\mu \neq$ 1.5 hours

Case 1
Sample mean is
1.25 hours

$$z = \frac{1.25 - 1.5}{0.5/\sqrt{25}} = -2.5$$

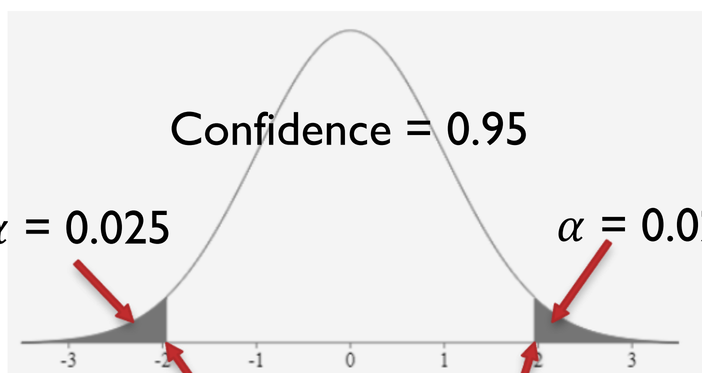



```
> 2*pnorm(-2.5,lower.tail = TRUE)
[1] 0.01241933
```

Case 2
Sample mean
is 1.4 hours

$$z = \frac{1.65 - 1.5}{0.5/\sqrt{25}} = 1.5$$


```
> 2*pnorm(1.5,lower.tail = FALSE)
[1] 0.1336144
```



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

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

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

```
#H0 μ= 0.5 H1 μ≠  0.5
xbar <- 0.495
pmean <- 0.5
psd <- 0.05
n <- 16
Alpha<- 0.05
z <- (xbar - pmean)/(psd/sqrt(n))
z
CVL<- qnorm(0.05/2)
CVU<- qnorm(1-(0.05/2))
c(CVL,CVU)
(z < CVL) | (z > CVU)

Pvalue<- 2*pnorm(z,lower.tail = TRUE)
Pvalue
Pvalue <  Alpha
```

```
> #H0 μ= 0.5 H1 μ=  0.5
> xbar <- 0.495
> pmean <- 0.5
> psd <- 0.05
> n <- 16
> Alpha<- 0.05
> z <- (xbar - pmean)/(psd/sqrt(n))
> z
[1] -0.4
> CVL<- qnorm(0.05/2)
> CVU<- qnorm(1-(0.05/2))
> c(CVL,CVU)
[1] -1.959964  1.959964
> (z < CVL) | (z > CVU)
[1] FALSE
> Pvalue<- 2*pnorm(z,lower.tail = TRUE)
> Pvalue
[1] 0.6891565
> Pvalue <  Alpha
[1] FALSE
```

Additional exercise- Hypothesis test

- The manufacturer claims that the average strength of the fishing line is 8 kg and the standard deviation is 0.5 kg.
- 50 fishing lines were randomly selected and tested for an average strength of 8.5 kg. Please verify the manufacturer's claim at a significant level of 0.01.

R programing

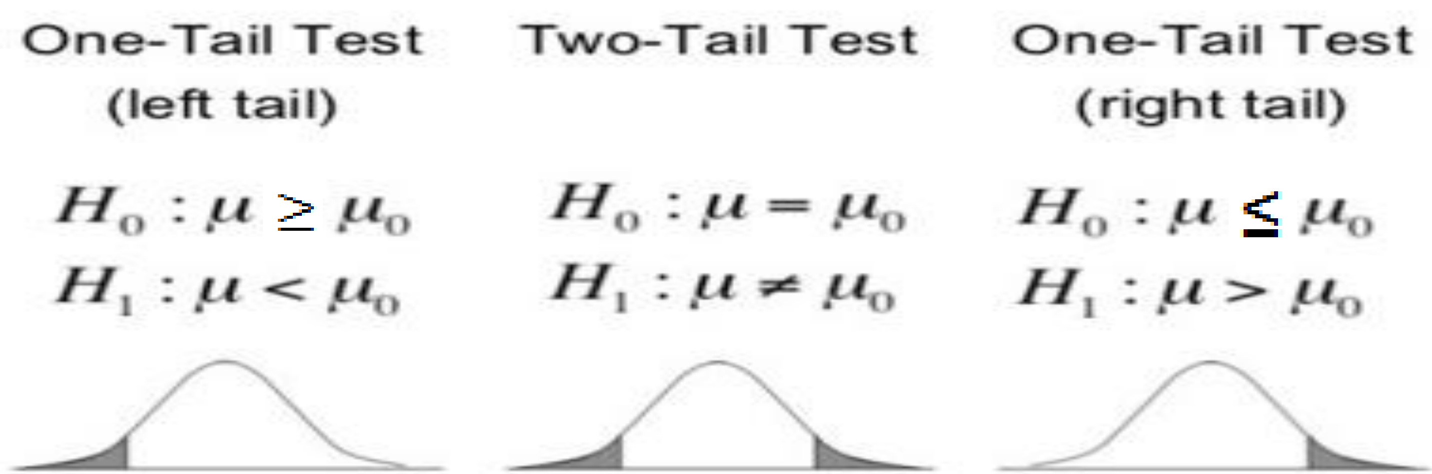
```
#H0 : μ = 8, H1 : μ≠8
xbar<- 8.5
pmean<- 8
psd<- 0.5
n<- 50
Alpha<-0.01
z<- (xbar-pmean)/(psd/sqrt(n))
z
CVL<- qnorm(0.01/2)
CVU<- qnorm(1-(0.01/2))
c(CVL,CVU)
(z < CVL) | (z > CVU)

Pvalue<- 2*pnorm(z,lower.tail = FALSE)
Pvalue
Pvalue <  Alpha
```

```
> #H0 : μ = 8, H1 : μ≠8
> xbar<- 8.5
> pmean<- 8
> psd<- 0.5
> n<- 50
> Alpha<-0.01
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] 7.071068
> CVL<-qnorm(0.01/2)
> CVU<-qnorm(1-(0.01/2))
> c(CVL,CVU)
[1] -2.575829  2.575829
> (z < CVL)|(z > CVU)
[1] TRUE
> Pvalue<-2*pnorm(z,lower.tail = FALSE)
> Pvalue
[1] 1.53746e-12
> Pvalue < Alpha
[1] TRUE
```

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.



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).

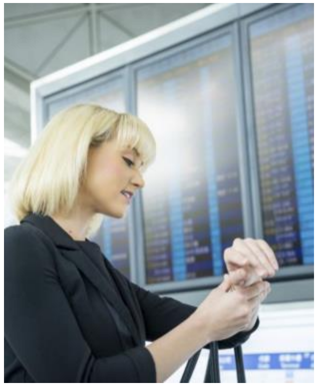
Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the result of the hypothesis test.

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 $p = 1 - \beta$

Where are we and where are we going ?

Getting a grasp on data

Populations and Samples

Making use of data (inference)

- Estimation
- Hypothesis Testing
  - One population
  - Two population

Additional exercise

- Suppose that a doctor claims that those who are 17 years old have an average body temperature that is higher than the commonly accepted average human temperature of 98.6 degrees Fahrenheit. A simple random statistical sample of 25 people, each of age 17, is selected. The average temperature of the sample is found to be 98.3 degrees. Further, suppose that we know that the population standard deviation of everyone who is 17 years old is 0.6 degrees. Is the doctor's claim correct or not ($\alpha = 0.01$)?

Additional exercise

- According to market research, it is stipulated that the bulbs produced can be used for 1200 hours. After formal production, 64 bulbs are tested, and the average is 1194 hours. The population variance is 36 hours. The test is marked with a significant level is 5%. Is the bulb manufactured by the factory compliant?

Additional exercise

- The ice shop is scheduled to open a branch in a certain location. According to experience, the location of the ice shop must be a large number of people, and the average hourly at least 100 people can be profitable. Assume that the planners of the ice shop observed 49 hours at the scheduled location, and the average pedestrian per hour was 106 and the population standard deviation of 10.5. Can they open an ice shop at this location($\alpha=0.01$)?

Additional exercise

- Tire manufacturers claim to produce at least 60,000 kilometers of tires. It is known that the mileage that the tire can travel is a normal distribution, and the standard deviation of population is 26,000 kilometers. Today, 16 tires are tested with an average mileage of 59,000 kilometers. Is the manufacturer's claim correct under 10% of the significant level?

Additional exercise

- A brand mobile phone claimed that its average weight was 78 grams. Today, 10 of the brand's mobile phones were randomly selected, with an average weight of 80 grams with $\sigma = 4$ grams. Please verify that the manufacturer's claim is true at a significant level. (assuming the population conforms to the normal distribution and the significant level is 0.05)

ADDITIONAL EXERCISE ANSWER

Additional exercise

- Suppose that a doctor claims that those who are 17 years old have an average body temperature that is higher than the commonly accepted average human temperature of 98.6 degrees Fahrenheit. A simple random statistical sample of 25 people, each of age 17, is selected. The average temperature of the sample is found to be 98.3 degrees. Further, suppose that we know that the population standard deviation of everyone who is 17 years old is 0.6 degrees. Is the doctor's claim correct or not ($\alpha = 0.01$)?

Ans

```
#H0: μ ≥ 98.6 H1: μ < 98.6
xbar<- 98.3
pmean<- 98.6
psd<- 0.6
n<- 25
Alpha<- 0.01
z<- (xbar-pmean)/(psd/sqrt(n))
z
CV<- qnorm(0.01)
CV
Pvalue<- pnorm(z)
Pvalue
Pvalue < Alpha
z < CV
```

```
> #H0: μ ≥ 98.6 H1: μ < 98.6
> xbar<- 98.3
> pmean<- 98.6
> psd<- 0.6
> n<- 25
> Alpha<- 0.01
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] -2.5
> CV<- qnorm(0.01)
> CV
[1] -2.326348
> Pvalue<- pnorm(z)
> Pvalue
[1] 0.006209665
> Pvalue < Alpha
[1] TRUE
> z < CV
[1] TRUE
```

Additional exercise

- According to market research, it is stipulated that the bulbs produced can be used for 1200 hours. After formal production, 64 bulbs are tested, and the average is 1194 hours. The population variance is 36 hours. The test is marked with a significant level is 5%. Is the bulb manufactured by the factory compliant?

Ans

```
#H0 : μ = 1200, H1 : μ≠ 1200
xbar<- 1194
pmean<- 1200
psd<- sqrt(36)
n<- 64
Alpha<-0.05
z<- (xbar-pmean)/(psd/sqrt(n))
z
CVL<- qnorm(0.05/2)
CVU<- qnorm(1-(0.05/2))
c(CVL,CVU)
(z < CVL) | (z > CVU)

Pvalue<- 2*pnorm(z,lower.tail = TRUE)
Pvalue
Pvalue <  Alpha
```

```
> #H0 : μ = 1200, H1 : μ≠ 1200
> xbar<- 1194
> pmean<- 1200
> psd<- sqrt(36)
> n<- 64
> Alpha<-0.05
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] -8
> CVL<- qnorm(0.05/2)
> CVU<- qnorm(1-(0.05/2))
> c(CVL,CVU)
[1] -1.959964  1.959964
> (z < CVL) | (z > CVU)
[1] TRUE
> Pvalue<- 2*pnorm(z,lower.tail = TRUE)
> Pvalue
[1] 1.244192e-15
> Pvalue <  Alpha
[1] TRUE
```

Additional exercise

- The ice shop is scheduled to open a branch in a certain location. According to experience, the location of the ice shop must be a large number of people, and the average hourly at least 100 people can be profitable. Assume that the planners of the ice shop observed 49 hours at the scheduled location, and the average pedestrian per hour was 106 and the population standard deviation of 10.5. Can they open an ice shop at this location($\alpha=0.01$)?

Ans

```
#H0 : μ ≤ 100, H1 : μ > 100
xbar<- 106
pmean<- 100
psd<- 10.5
n<- 49
Alpha<- 0.01
z<- (xbar-pmean)/(psd/sqrt(n))
z
CV<- qnorm(1-0.01)
CV
Pvalue<- pnorm(z,lower.tail = FALSE)
Pvalue
Pvalue < Alpha
z > CV
```

```
> #H0 : μ ≤ 100, H1 : μ > 100
> xbar<- 106
> pmean<- 100
> psd<- 10.5
> n<- 49
> Alpha<- 0.01
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] 4
> CV<- qnorm(1-0.01)
> CV
[1] 2.326348
> Pvalue<- pnorm(z,lower.tail = FALSE)
> Pvalue
[1] 3.167124e-05
> Pvalue < Alpha
[1] TRUE
> z > CV
[1] TRUE
```

Additional exercise

- Tire manufacturers claim to produce at least 60,000 kilometers of tires. It is known that the mileage that the tire can travel is a normal distribution, and the standard deviation of population is 26,000 kilometers. Today, 16 tires are tested with an average mileage of 59,000 kilometers. Is the manufacturer's claim correct under 10% of the significant level?

Ans

```
#H0 : μ ≥ 60000, H1 : μ < 60000
xbar<- 59000
pmean<- 60000
psd<- 26000
n<- 16
Alpha<- 0.1
z<- (xbar-pmean)/(psd/sqrt(n))
z
CV<- qnorm(0.1)
CV
Pvalue<- pnorm(z,lower.tail = TRUE)
Pvalue
Pvalue
Pvalue < Alpha
z < CV
```

```
> #H0 : μ ≥ 60000, H1 : μ < 60000
> xbar<- 59000
> pmean<- 60000
> psd<- 26000
> n<- 16
> Alpha<- 0.1
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] -0.1538462
> CV<- qnorm(0.1)
> CV
[1] -1.281552
> Pvalue<- pnorm(z,lower.tail = TRUE)
> Pvalue
[1] 0.4388655
> Pvalue
[1] 0.4388655
> Pvalue < Alpha
[1] FALSE
> z < CV
[1] FALSE
```

Additional exercise

- A brand mobile phone claimed that its average weight was 78 grams. Today, 10 of the brand's mobile phones were randomly selected, with an average weight of 80 grams with $\sigma = 4$ grams. Please verify that the manufacturer's claim is true at a significant level. (assuming the population conforms to the normal distribution and the significant level is 0.05)

Ans

```
#H0 : μ = 78, H1 : μ ≠ 78
xbar<- 80
pmean<- 78
psd<- 4
n<- 10
Alpha<-0.05
z<- (xbar-pmean)/(psd/sqrt(n))
z
CVL<- qnorm(0.05/2)
CVU<- qnorm(1-(0.05/2))
c(CVL,CVU)
(z < CVL) | (z > CVU)

Pvalue<- 2*pnorm(z,lower.tail = FALSE)
Pvalue
Pvalue <  Alpha
```

```
> #H0 : μ = 78, H1 : μ ≠ 78
> xbar<- 80
> pmean<- 78
> psd<- 4
> n<- 10
> Alpha<-0.05
> z<- (xbar-pmean)/(psd/sqrt(n))
> z
[1] 1.581139
> CVL<- qnorm(0.05/2)
> CVU<- qnorm(1-(0.05/2))
> c(CVL,CVU)
[1] -1.959964  1.959964
> (z < CVL) | (z > CVU)
[1] FALSE
> Pvalue<- 2*pnorm(z,lower.tail = FALSE)
> Pvalue
[1] 0.1138463
> Pvalue <  Alpha
[1] FALSE
```