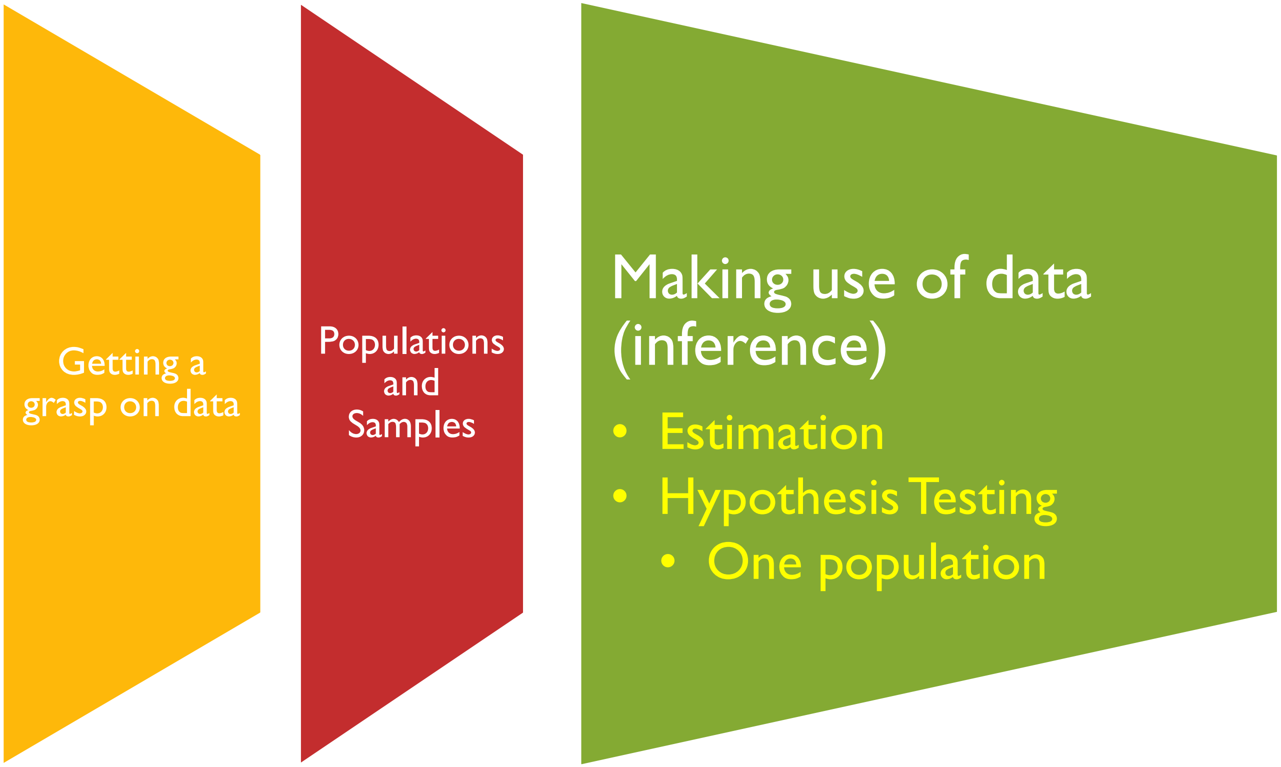
Welcome

What did we do last class?



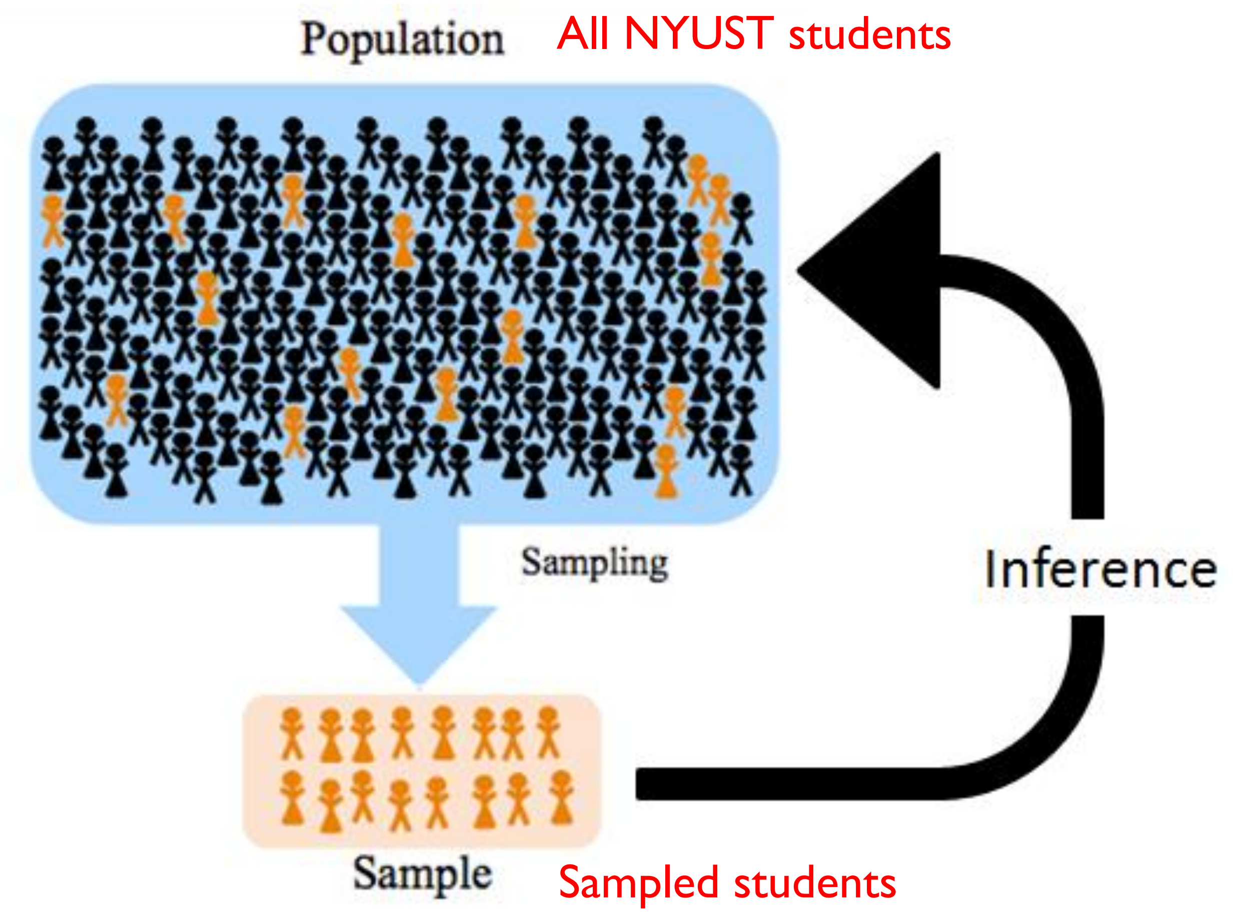
Getting a grasp on data

Populations and Samples

Making use of data (inference)

- Estimation
- Hypothesis Testing
  - One population

Overall, students are very satisfied with the teaching results. Can we infer if the average of students' rating is higher than 4.



All NYUST students

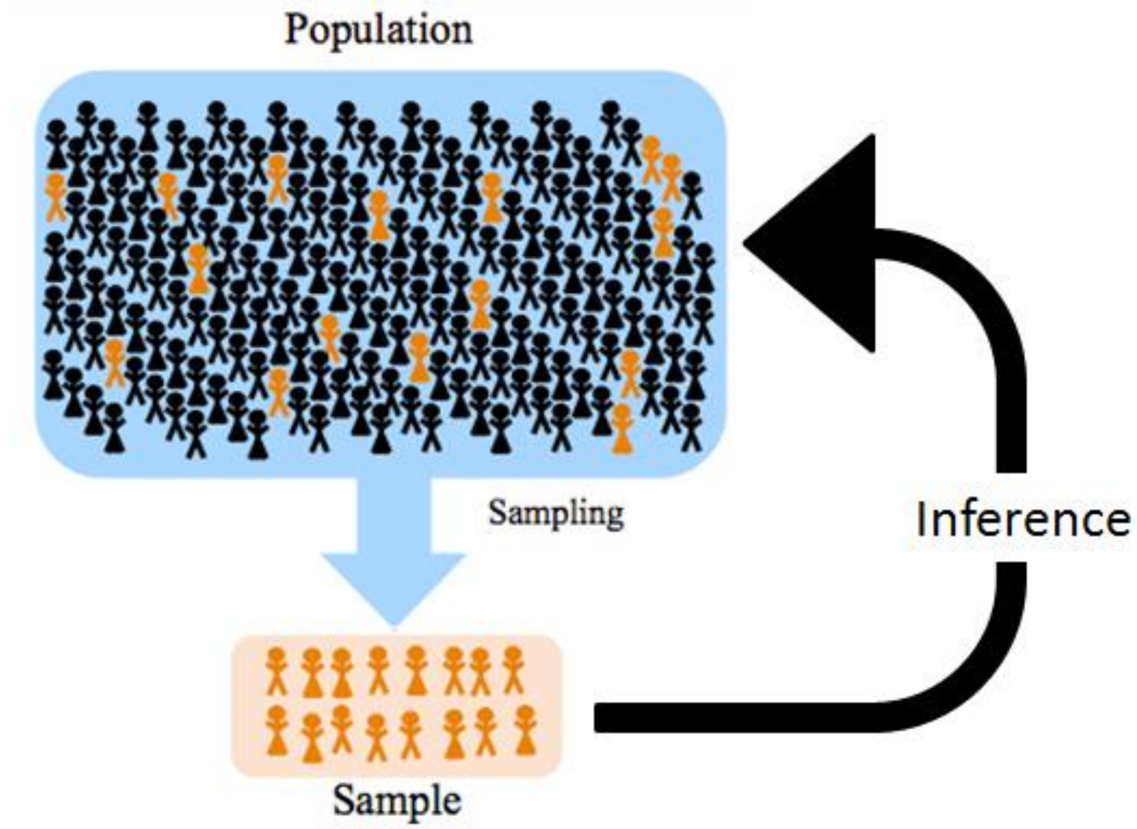
Hypothesis test

Sampled students

One population

- Mean
  - to test whether the sample mean differ from a population mean
- Proportion
  - to test whether the sample proportion differ from a population proportion
- Variance
  - to test whether the sample variance differ from a population variance

One population-Mean

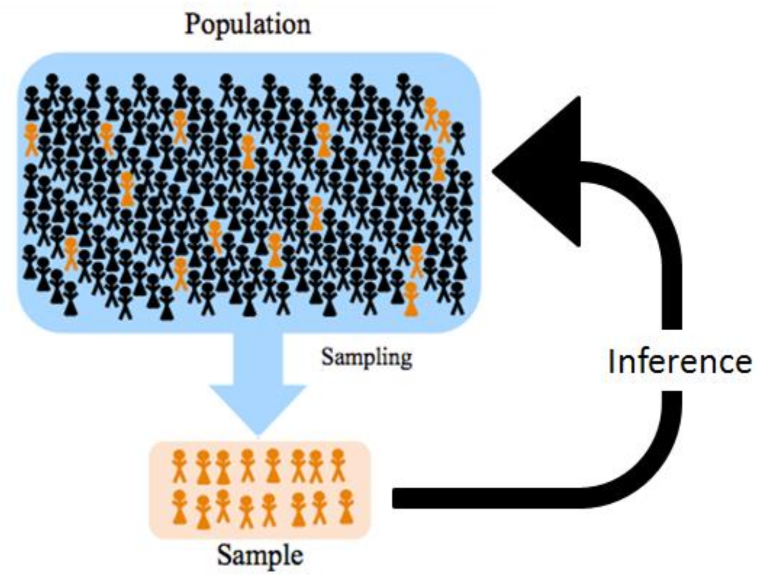


test whether the sample mean differ from a population mean

Now consider the case in which you have a normal distribution data but you do not know the population variance

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

Population Variance



If you can calculate population variance

It means that

- You must know the *true* population mean

- You must know the frequency distribution of X and actual informants in your population (N). Therefore, you can obtain *true* population proportion

- You will have the answer of *true* population variance

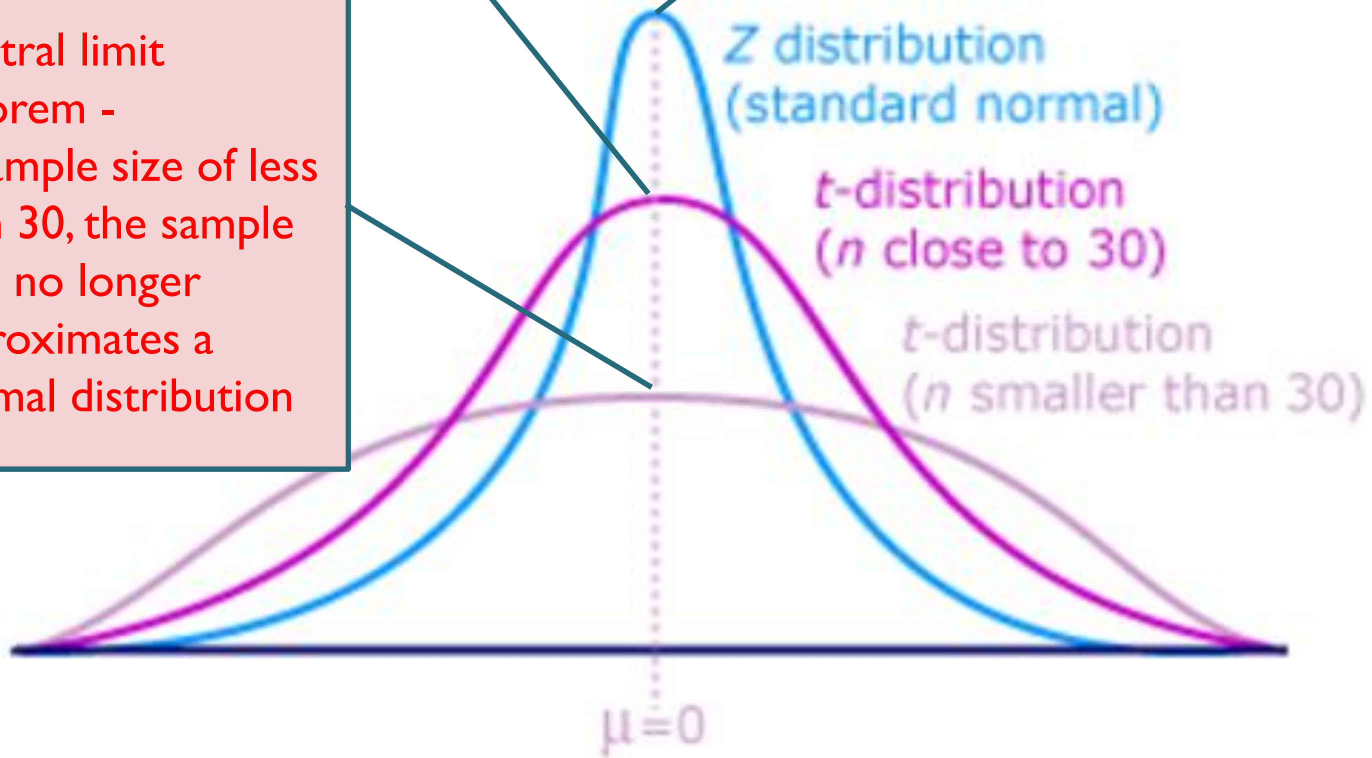
If you can calculate population variance, it means that you have all *true* population parameters and why we need to redo "statistical analysis" to test mean, proportion and variance?

Normal distribution vs t-distribution

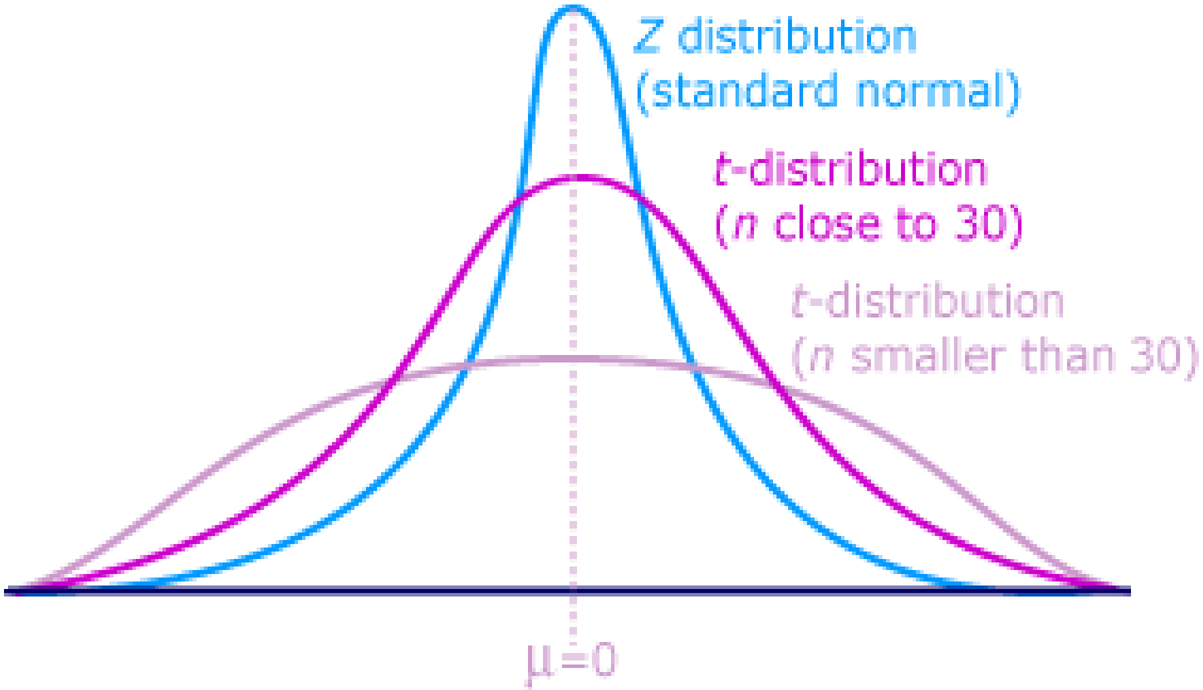
A t distribution is a sampling distribution

Central limit theorem -
A sample size of less than 30, the sample data no longer approximates a normal distribution

A *Z* distribution is a normal distribution



Z distribution and t distribution



We use the t distribution when the population variance is unknown

As *n* increases in size, the shape of the *t*-distribution begins to resemble a normal distribution

The t-distribution, like the z-distribution, is bell-shaped and symmetric about a mean of 0

Z distribution and t distribution

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$ Sample Variance

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$ Population Variance

no

☒

If the population variance is unknown, the estimation of population mean is given by **t-distribution**

yes

☑

If the population variance is known, the estimation of population mean is given by **z-distribution**

Confidence Intervals

- A Confidence Interval is an interval of numbers containing the most plausible values for our Population Parameter.

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Now, we are moving to t-distribution

Use *z*-distribution

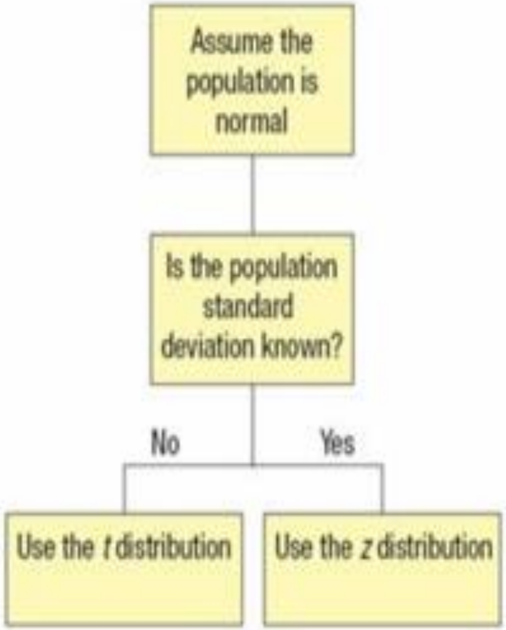
If the population standard deviation is known or the sample is greater than 30.

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Use *t*-distribution

If the population standard deviation is unknown and the sample is less than 30.

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$



Assume the population is normal

Is the population standard deviation known?

No

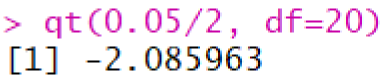
Yes

Use the *t* distribution

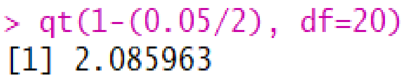
Use the *z* distribution

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

- Sample size = 21
- Significance level α is 0.05
- The degrees of freedom (df) = sample size -1

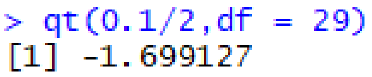


```
> qt(0.05/2, df=20)
[1] -2.085963
```

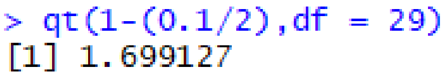


```
> qt(1-(0.05/2), df=20)
[1] 2.085963
```

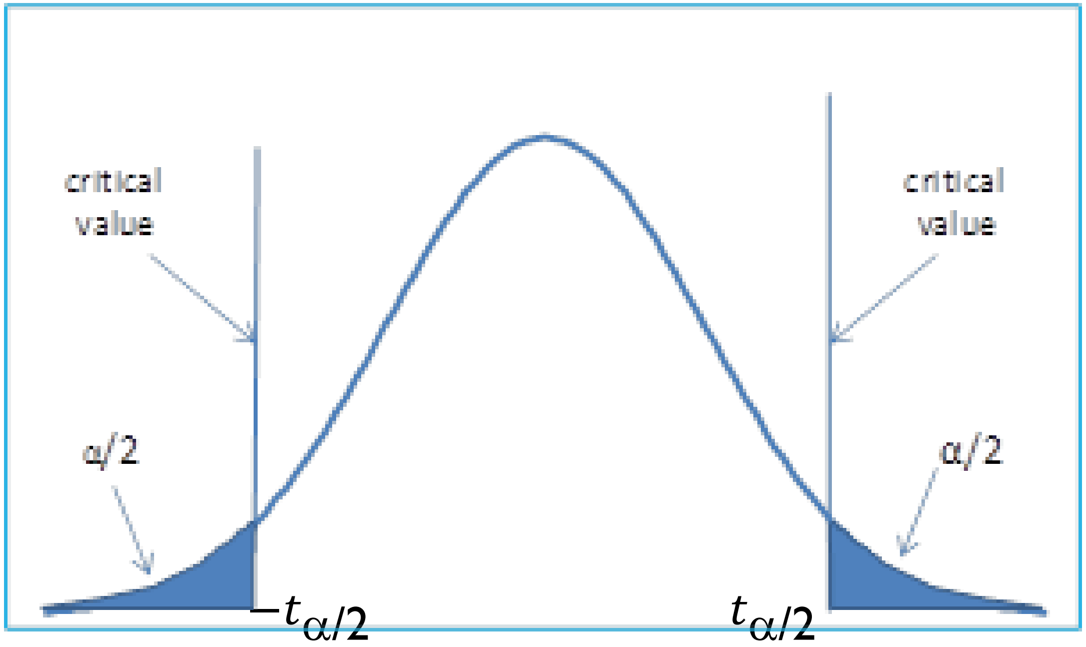
- Sample size = 30
- Significance level α is 0.1
- The degrees of freedom (df) = sample size -1



```
> qt(0.1/2,df = 29)
[1] -1.699127
```

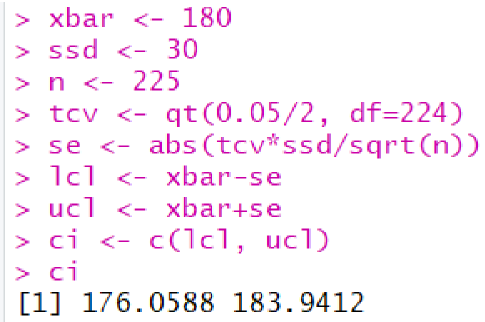


```
> qt(1-(0.1/2),df = 29)
[1] 1.699127
```



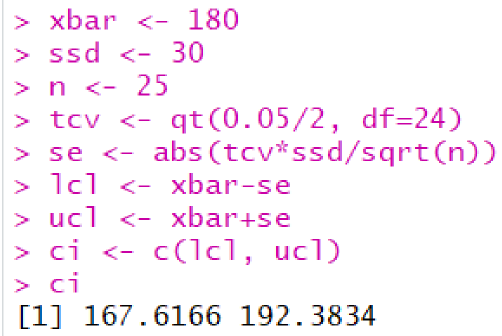
Problem

- Suppose we want to estimate the average weight of NYUST student (male). We draw a random sample of 225 men from the population and weight them.
  - We find that the average in our sample weighs 180 pounds, and the standard deviation of the sample is 30 pounds. What is the 95% confidence interval.



```
> xbar <- 180
> ssd <- 30
> n <- 225
> tcv <- qt(0.05/2, df=224)
> se <- abs(tcv*ssd/sqrt(n))
> lcl <- xbar-se
> ucl <- xbar+se
> ci <- c(lcl, ucl)
> ci
[1] 176.0588 183.9412
```

  - What happen if we only draw a random sample of 25 men. What is the 95% confidence interval.



```
> xbar <- 180
> ssd <- 30
> n <- 25
> tcv <- qt(0.05/2, df=24)
> se <- abs(tcv*ssd/sqrt(n))
> lcl <- xbar-se
> ucl <- xbar+se
> ci <- c(lcl, ucl)
> ci
[1] 167.6166 192.3834
```

Do you aware?

- In terms of t-critical values, what is the difference when the sample size decreases from 225 to 25?

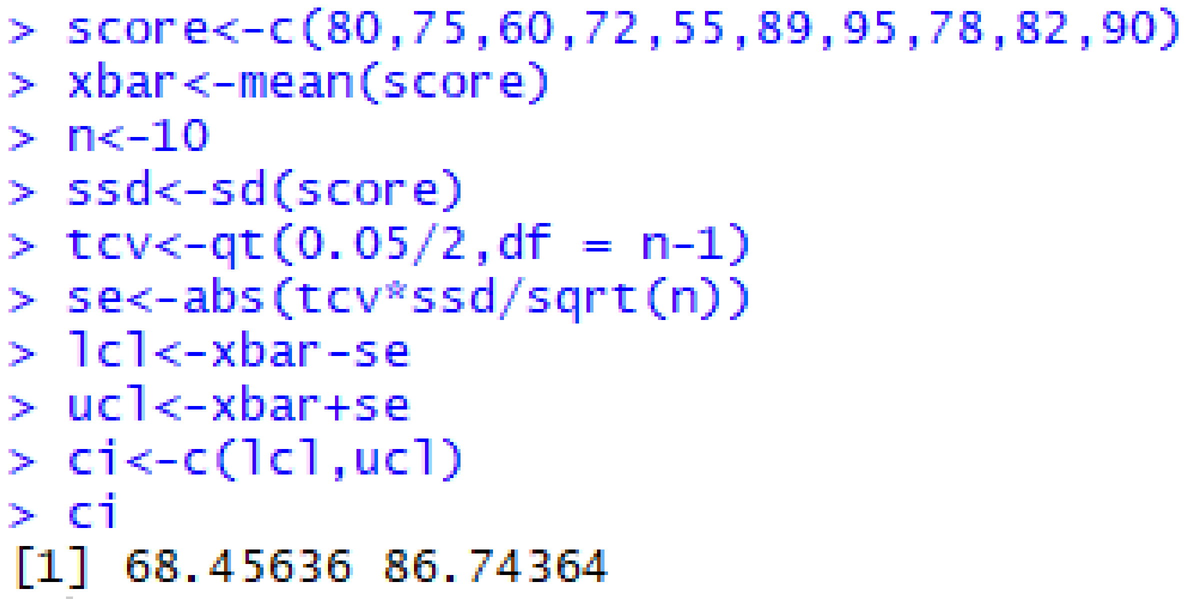
- qnorm(0.05/2) = -1.959964

- qt(0.05/2, df=224) = -1.970611
- qt(0.05/2, df=24) = -2.063899

- qt(0.05/2, df=2000) = -1.961151
- qt(0.05/2, df=8000) = -1.960261

Exercise 1

- A class randomly selected 10 students' statistical scores, and the scores were 80, 75, 60, 72, 55, 89, 95, 78, 82, and 90. What is the 95% confidence interval for the average grade μ of the class?



```
> score<-c(80,75,60,72,55,89,95,78,82,90)
> xbar<-mean(score)
> n<-10
> ssd<-sd(score)
> tcv<-qt(0.05/2,df = n-1)
> se<-abs(tcv*ssd/sqrt(n))
> lcl<-xbar-se
> ucl<-xbar+se
> ci<-c(lcl,ucl)
> ci
[1] 68.45636 86.74364
```

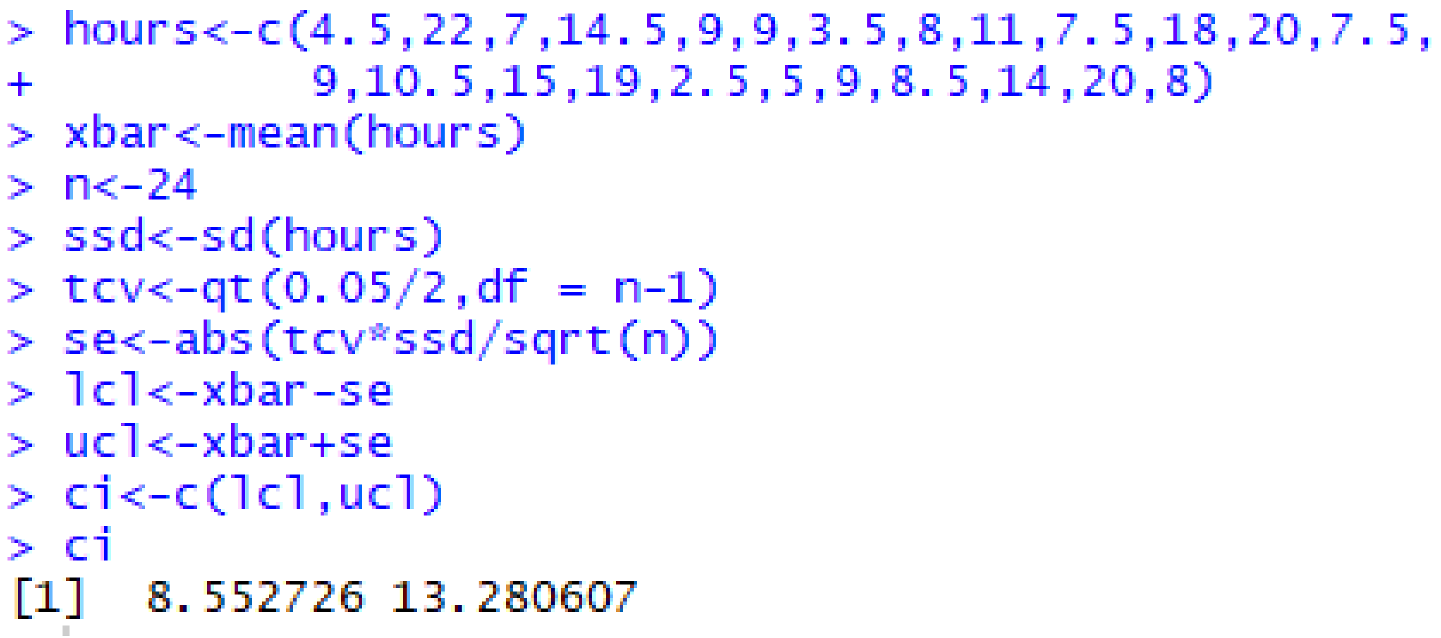
Exercise 2

- A sample of students from an introductory stats class were polled regarding the number of hours they spent studying for the last exam.All students submitted the number of hours on a card.There were 24 individuals in the one section of the course polled.The data was used to make inferences regarding the other students taking the course.There data are below:

- Compute a 95 percent confidence interval.What does this tell us?

4.5	22	7	14.5	9	9	3.5	8	11	7.5	18	20
7.5	9	10.5	15	19	2.5	5	9	8.5	14	20	8

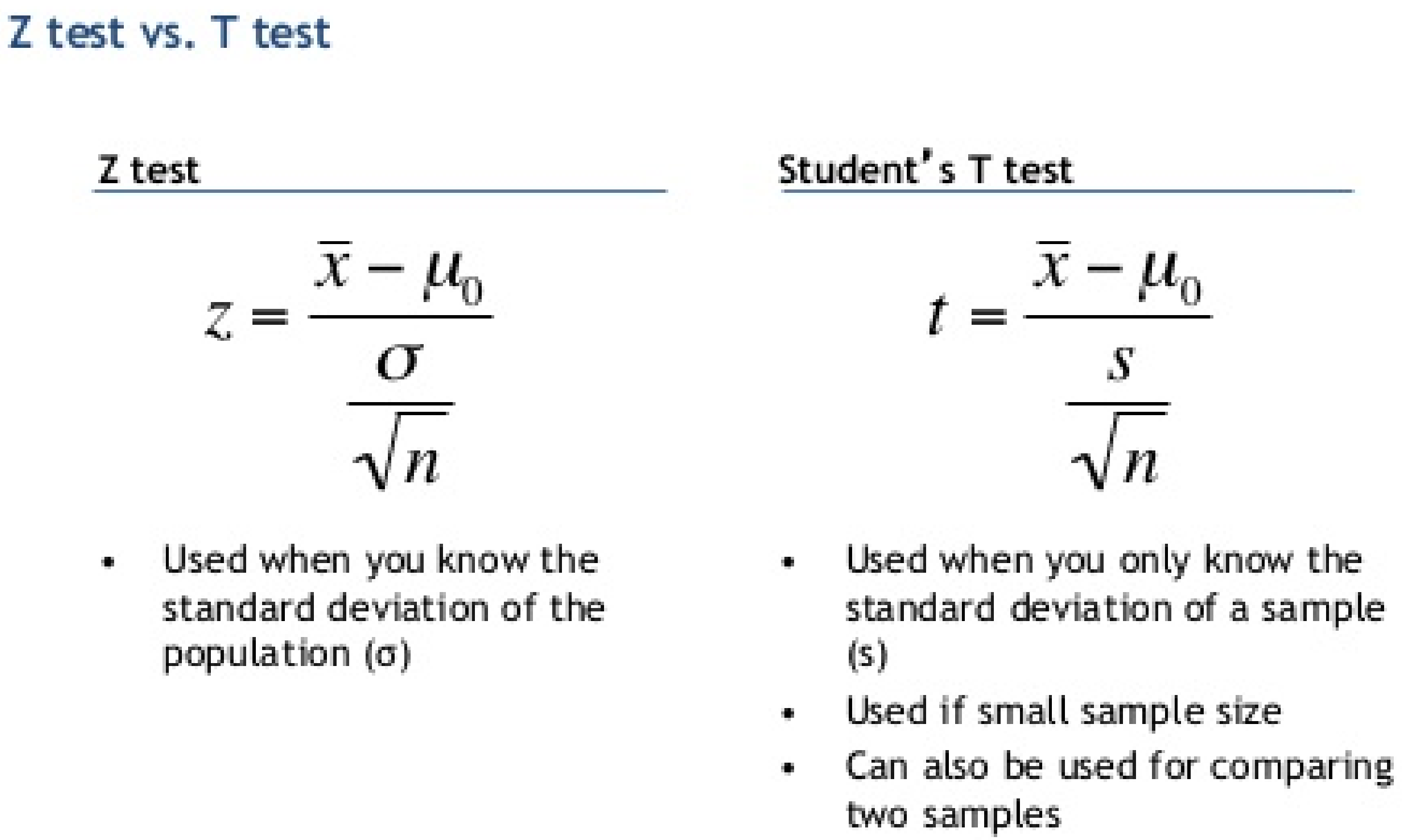
Ans



```
> hours<-c(4.5,22,7,14.5,9,9,3.5,8,11,7.5,18,20,7.5,
+          9,10.5,15,19,2.5,5,9,8.5,14,20,8)
> xbar<-mean(hours)
> n<-24
> ssd<-sd(hours)
> tcv<-qt(0.05/2,df = n-1)
> se<-abs(tcv*ssd/sqrt(n))
> lcl<-xbar-se
> ucl<-xbar+se
> ci<-c(lcl,ucl)
> ci
[1]  8.552726 13.280607
```

I am 95% confident that the average studying hours in the stats class is between 8.55 and 13.28 hours.

Hypothesis test



Z test vs. T test

Z test

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

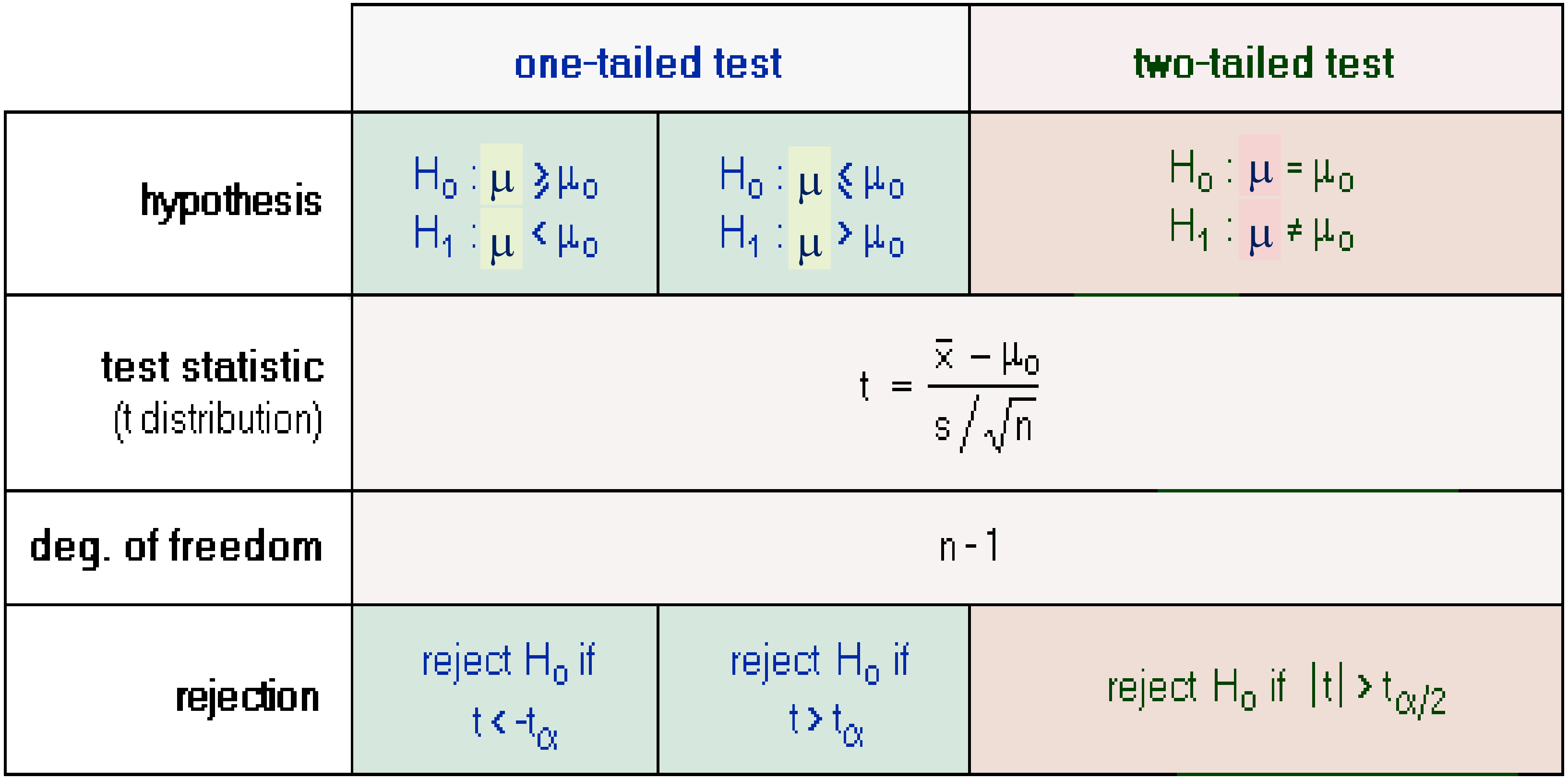
- Used when you know the standard deviation of the population (σ)

Student's T test

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Used when you only know the standard deviation of a sample (s)
- Used if small sample size
- Can also be used for comparing two samples

Inference about a population mean



	one-tailed test		two-tailed test
hypothesis	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$
test statistic (t distribution)	$t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$		
deg. of freedom	n - 1		
rejection	reject H_0 if $t < -t_{\alpha}$	reject H_0 if $t > t_{\alpha}$	reject H_0 if $\lvert t \rvert > t_{\alpha/2}$

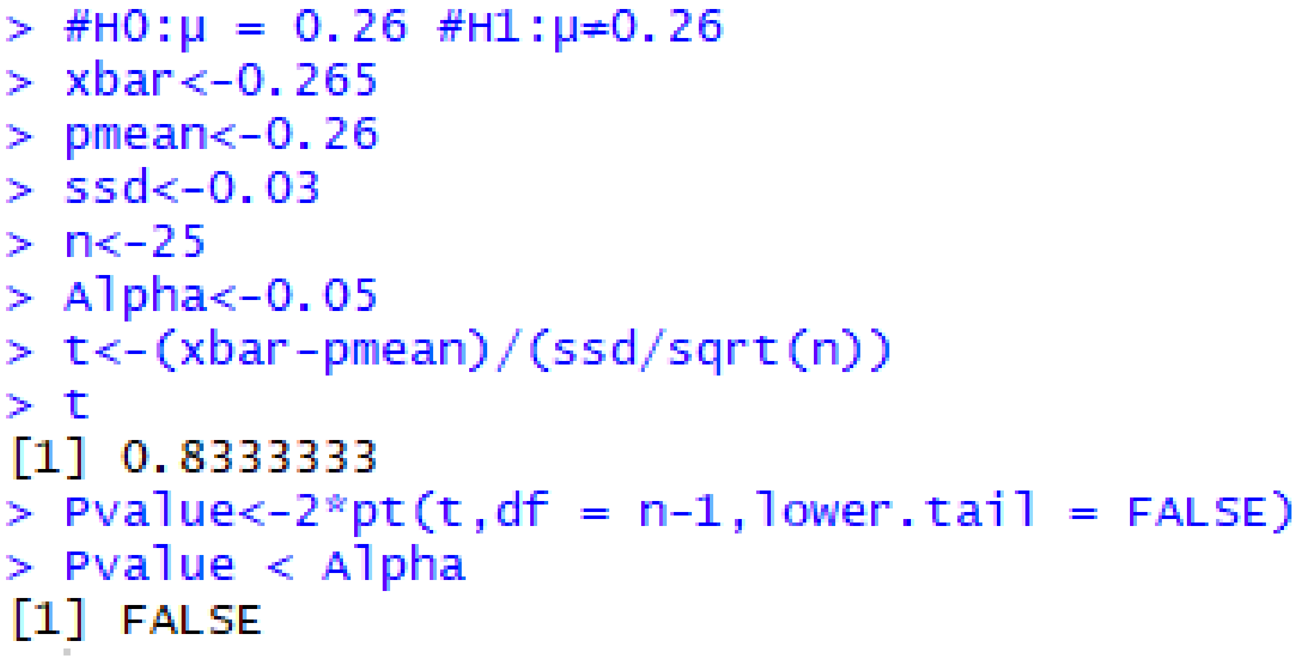
Now you are a manager of a baseball team in MLB

- Let's look at the batting average (AVG) in MLB from the years 1985 and 2013. You randomly recruit 25 players in your team.

- Their batting average is 0.265, and the sample standard deviation of is 0.03.

- Determine whether their batting average is significantly different from the history record of 0.26. Set the significance level at 5%.

Ans



```
> #H0:μ = 0.26 #H1:μ≠0.26
> xbar<-0.265
> pmean<-0.26
> ssd<-0.03
> n<-25
> Alpha<-0.05
> t<-(xbar-pmean)/(ssd/sqrt(n))
> t
[1] 0.8333333
> Pvalue<-2*pt(t,df = n-1,lower.tail = FALSE)
> Pvalue < Alpha
[1] FALSE
```

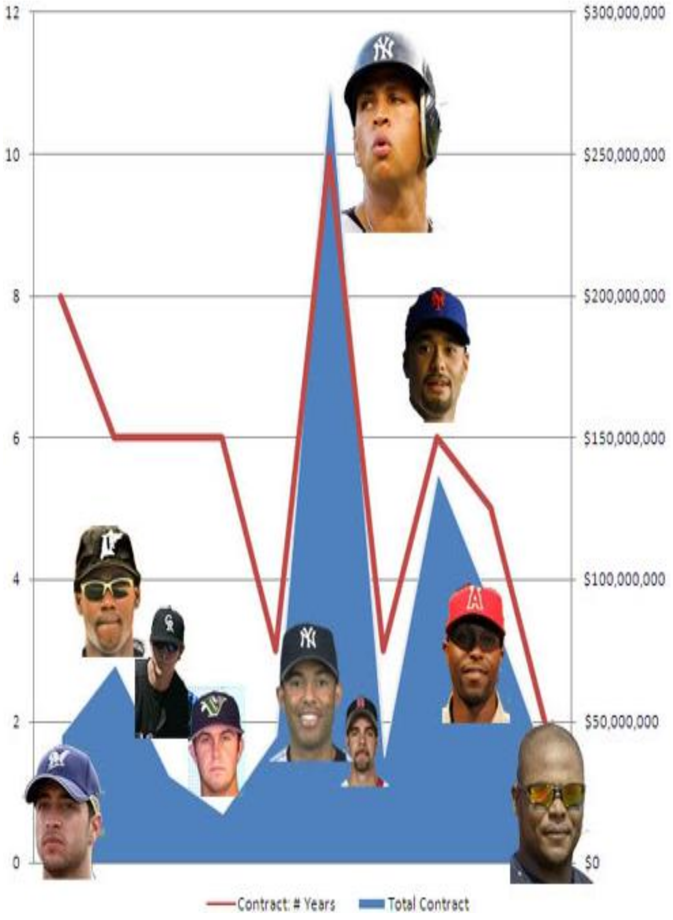
Non-rejection H0. At 5% significance level, we do not have sufficient evidence to suggest that their batting average is different from 0.26. Their batting performance is the same as the history record.

Now, you try to recruit some top players in your team

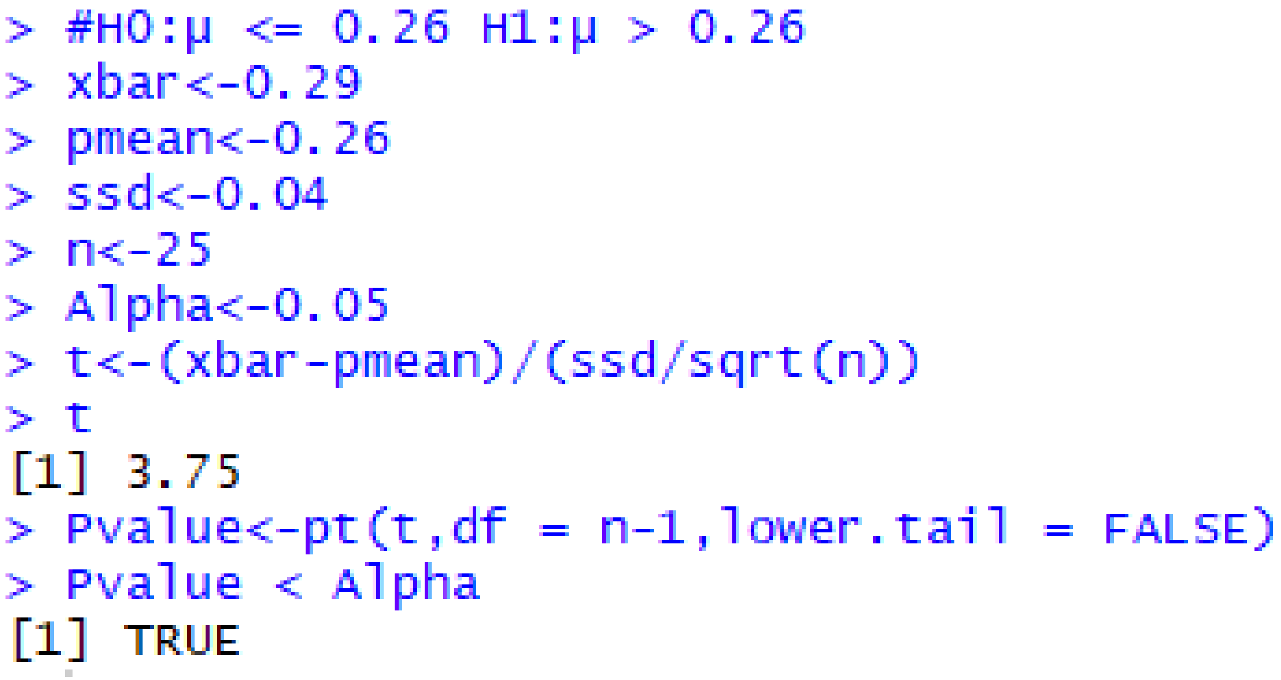
- You are not happy with previous recruitment.

- Thus, your are seeking for some top players to join to your team. The new team (25 players) batting average is now as 0.29 and the sample standard deviation of is 0.04.

- Determine whether their batting average is significantly higher than the 0.26. Set the significance level at 5%. Are you happy with the results?



Ans



```
> #H0:µ <= 0.26 H1:µ > 0.26
> xbar<-0.29
> pmean<-0.26
> ssd<-0.04
> n<-25
> Alpha<-0.05
> t<-(xbar-pmean)/(ssd/sqrt(n))
> t
[1] 3.75
> Pvalue<-pt(t,df = n-1,lower.tail = FALSE)
> Pvalue < Alpha
[1] TRUE
```

Reject H0. At 5% significance level, we do have sufficient evidence to suggest that their batting average is higher than 0.26. Their batting performance is better than the history record, and you should happy with the new recruitment.

However, the budget is limited….

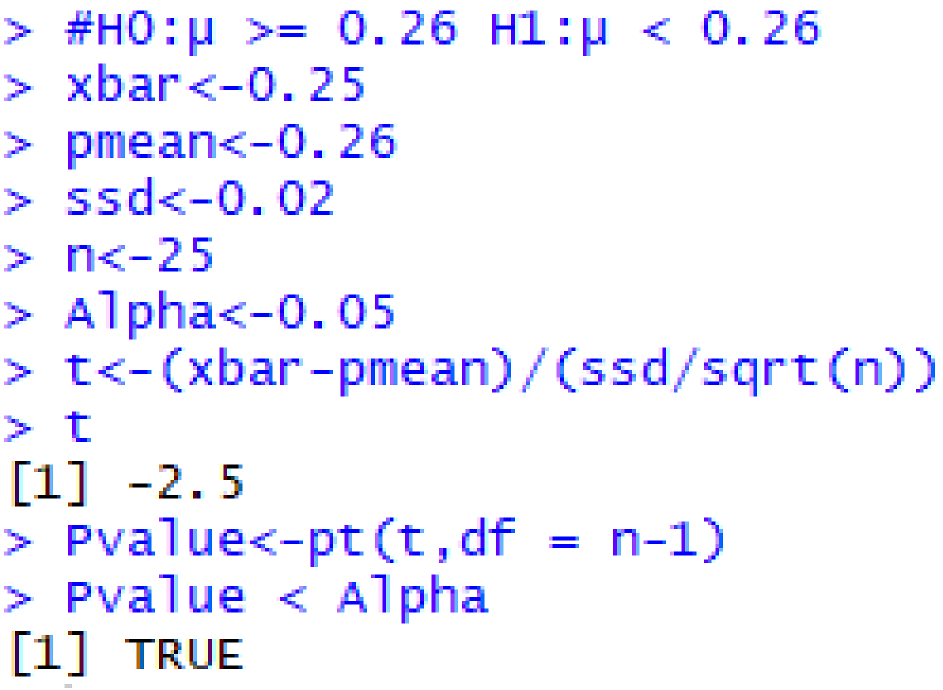
- Working with a limited budget means you can not have all top players in your team. You have to do some adjustments !!

- Your final 25 players their batting average is 0.25 and the sample standard deviation of is 0.02.

- Determine whether their batting average is significantly lower than the 0.26.

- Set the significance level at 5%.

Ans



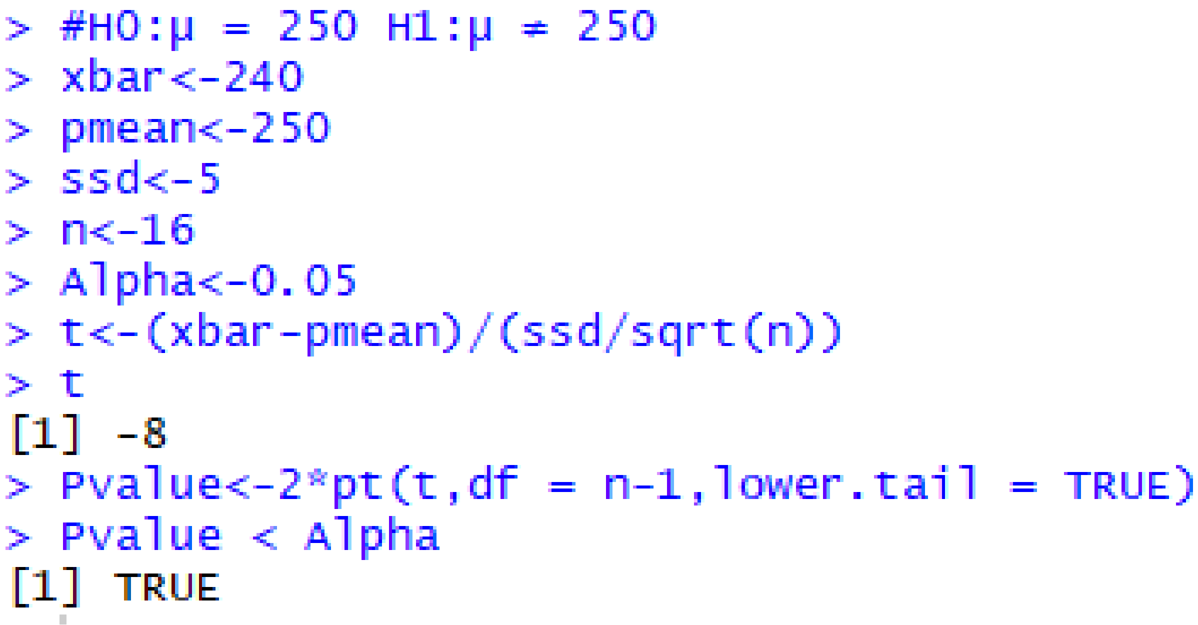
```
> #H0:µ >= 0.26 H1:µ < 0.26
> xbar<-0.25
> pmean<-0.26
> ssd<-0.02
> n<-25
> Alpha<-0.05
> t<-(xbar-pmean)/(ssd/sqrt(n))
> t
[1] -2.5
> Pvalue<-pt(t,df = n-1)
> Pvalue < Alpha
[1] TRUE
```

Reject H0. At 5% significance level, we do have sufficient evidence to suggest that their batting average is lower than 0.26. Their batting performance is lower than the history record, and you must be disappointed with the final result.

Exercise

- The manufacturer claims that the average weight of its products is exactly 250 grams. If the weight of the 16 products is randomly selected by the company, the average weight is 240 grams and the sample standard deviation is 5 grams. Please verify that the manufacturer claims is true (significant level = 0.05)?

Ans



```
> #H0:μ = 250 H1:μ ≠ 250
> xbar<-240
> pmean<-250
> ssd<-5
> n<-16
> Alpha<-0.05
> t<-(xbar-pmean)/(ssd/sqrt(n))
> t
[1] -8
> Pvalue<-2*pt(t,df = n-1,lower.tail = TRUE)
> Pvalue < Alpha
[1] TRUE
```

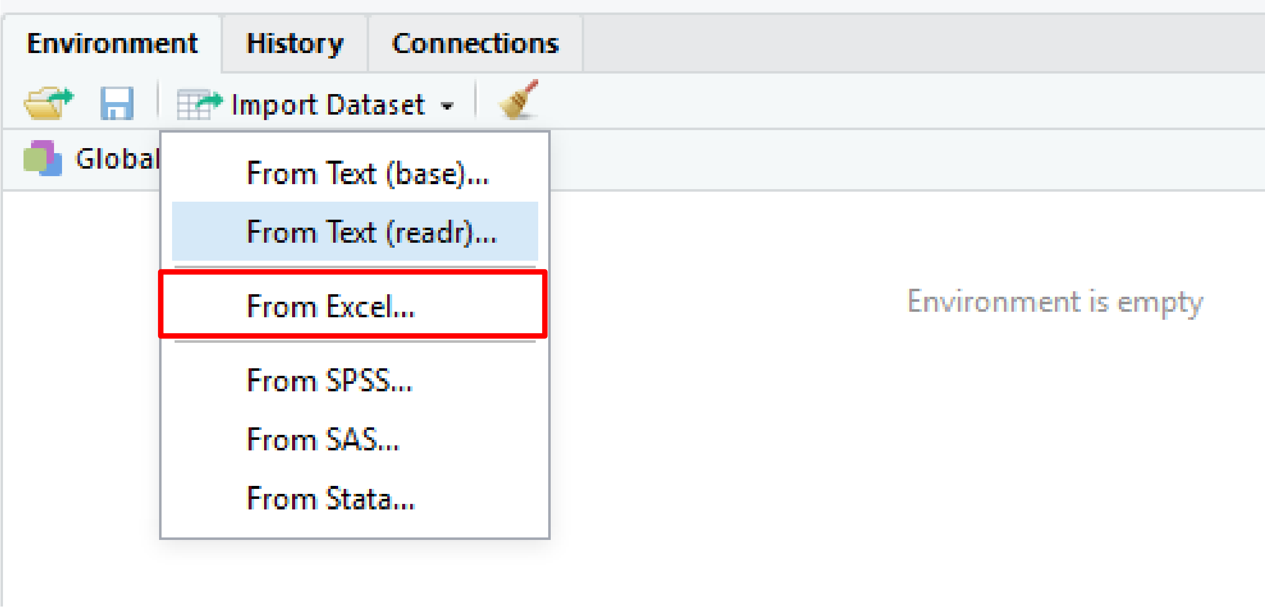
REAL DATA ANALYSIS

Basic R programming

- **getwd()**
  - Get the current path of R studio
- **setwd(“路徑”)**
  - Specify the path to crawl the data to R studio
- **data<-read.csv(“data_name.csv”)**
  - Read data
- **View(data)**
  - View the data
- **str(data)**
  - View the data structure
- **data.frame(data)**
  - Convert data to data frame
- **data$column**
  - Specify a column for the data
- **t.test(x, y = NULL, alternative = c(“two.sided”, “less”, ”great”), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)**
  - Calculate the p value of t value.

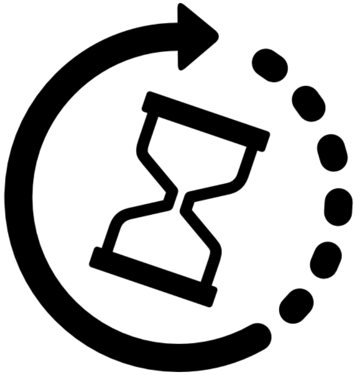
Import data

1. Login to the NYUST elearning system to download Data.rar.
2. Import Excel data into R studio

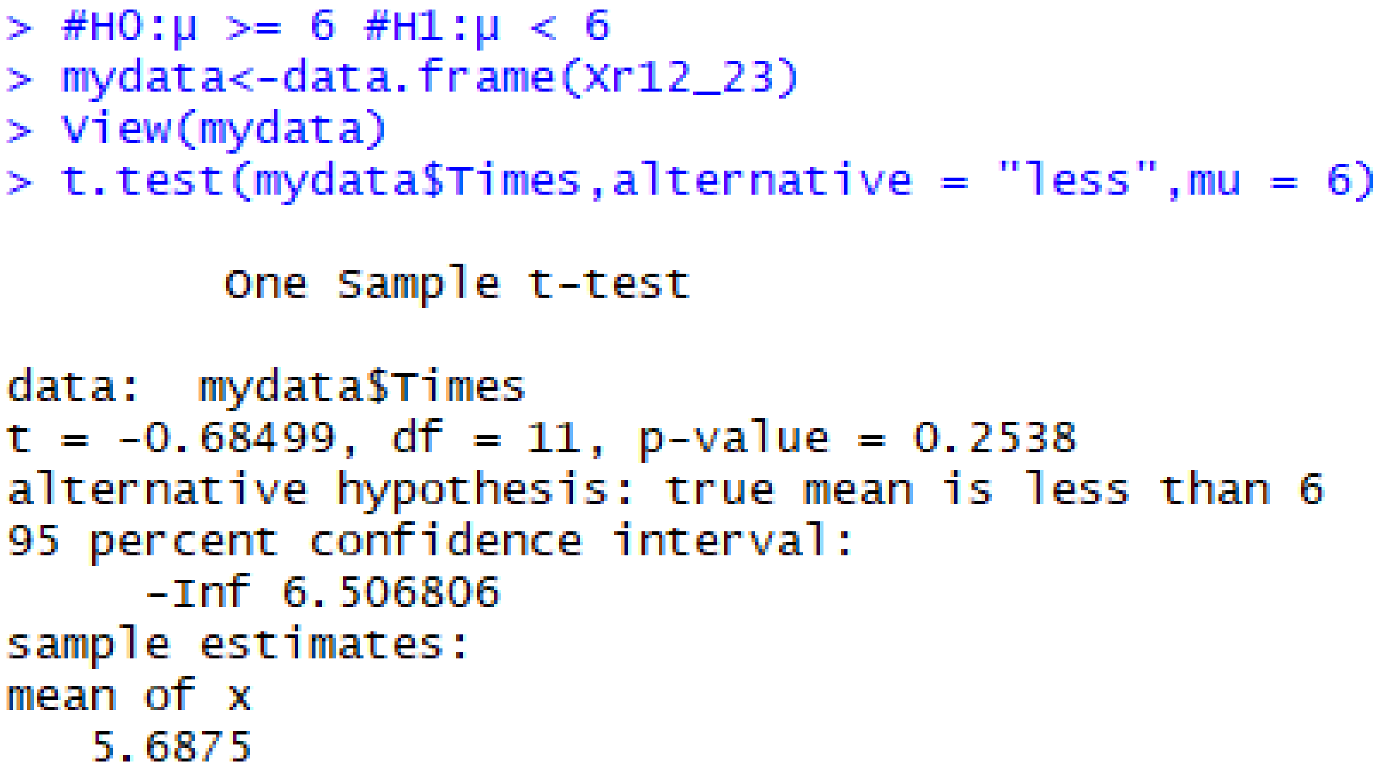


R dataset practices - Exercise 1

- (use Xr12-23)
- A courier service advertises that its average delivery time is less than 6 hours for local deliveries. A random sample of times for several deliveries to an address across town was recorded. These data are shown here. Is this sufficient evidence to support the courier's advertisement, at the 5% level of significance?



Ans

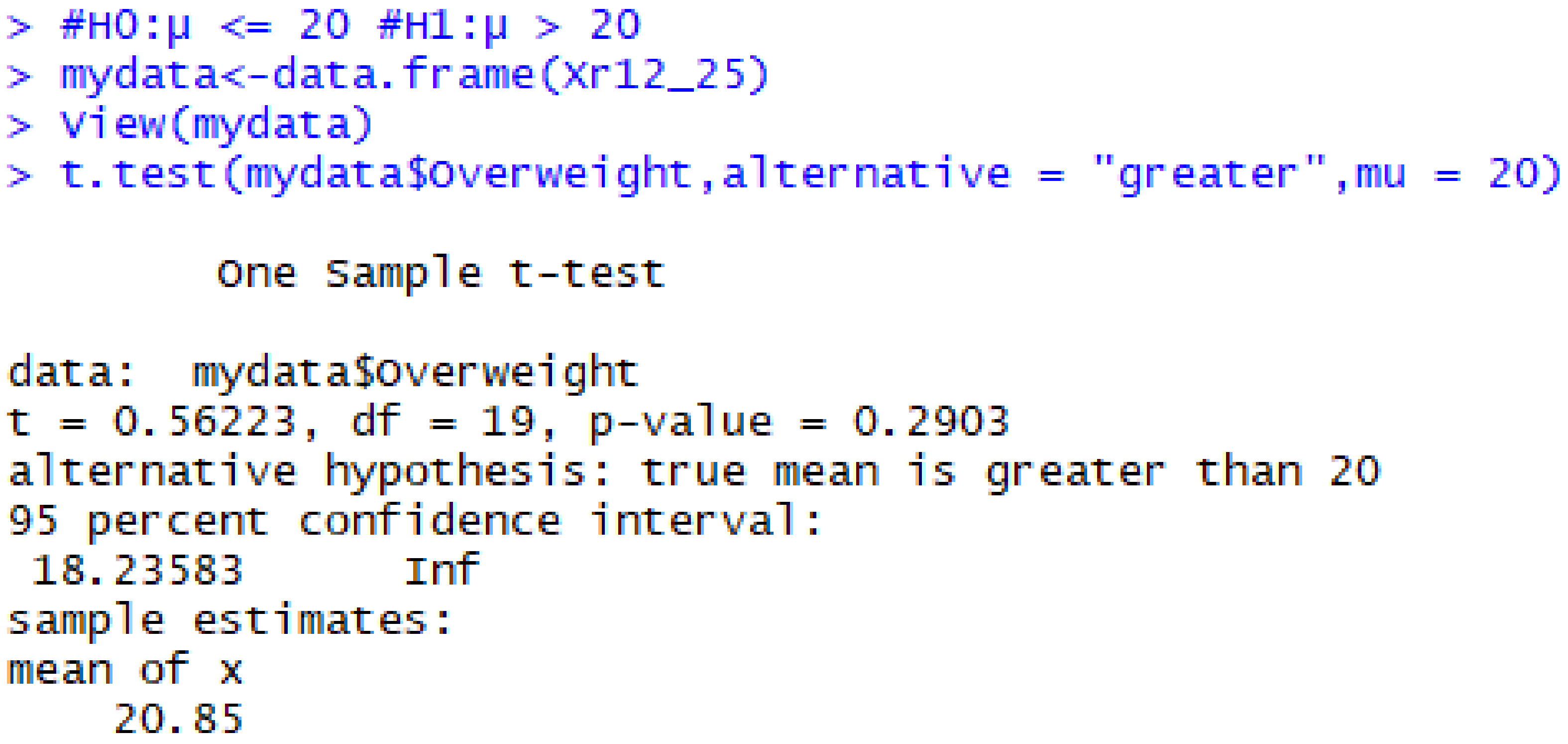


```
> #H0:μ >= 6 #H1:μ < 6
> mydata<-data.frame(Xr12_23)
> View(mydata)
> t.test(mydata$Times,alternative = "less",mu = 6)

        One Sample t-test

data:  mydata$Times
t = -0.68499, df = 11, p-value = 0.2538
alternative hypothesis: true mean is less than 6
95 percent confidence interval:
     -Inf 6.506806
sample estimates:
mean of x 
   5.6875 
```

R dataset practices - Exercise 2

- (use Xr12-25)
- A diet doctor claims that the average North American is more than 20 pounds overweight. To test him claim, a random sample of North American was weighed, and the difference between their actual and idea weights was calculated.
- The data is listed at Xr12-25. Do these data allow us to infer at the 5 % significance level that the doctor's claim is true?

Ans



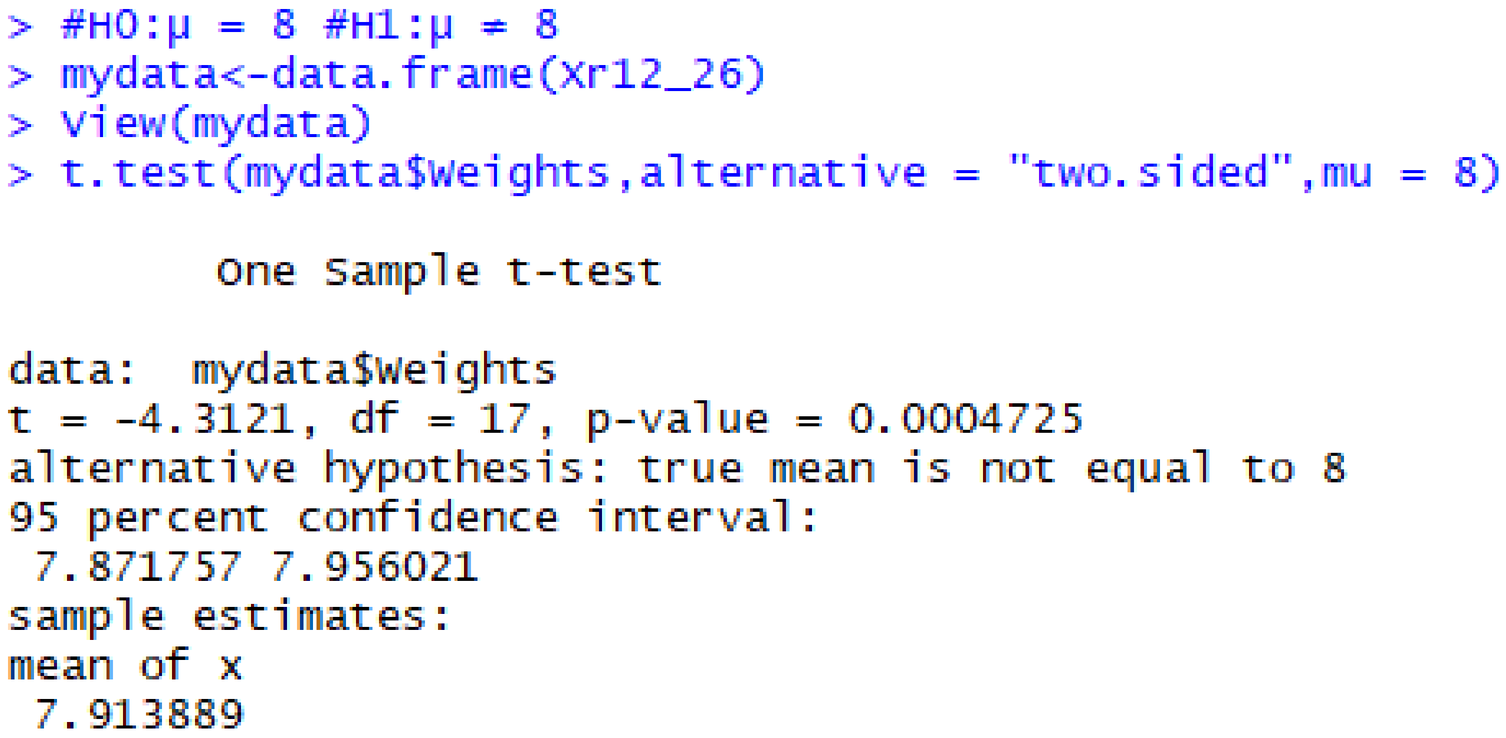
```
> #H0:μ <= 20 #H1:μ > 20
> mydata<-data.frame(Xr12_25)
> View(mydata)
> t.test(mydata$Overweight,alternative = "greater",mu = 20)

        One Sample t-test

data:  mydata$Overweight
t = 0.56223, df = 19, p-value = 0.2903
alternative hypothesis: true mean is greater than 20
95 percent confidence interval:
 18.23583      Inf
sample estimates:
mean of x
    20.85
```

R dataset practices - Exercise 3

- (use Xr12-26)
- A federal agency is responsible for enforcing laws governing weights and measures routinely inspects packages to determine whether the weight of the contents is the same as that advertised on the package.

- A random sample of containers whose packaging states that the contents weight 8 ounces was drawn. The contents were weighted and the results in Xr12-26. Can we conclude at the 1% significance level that on average the containers are mislabeled?

Ans



```
> #H0:μ = 8 #H1:μ ≠ 8
> mydata<-data.frame(Xr12_26)
> View(mydata)
> t.test(mydata$Weights,alternative = "two.sided",mu = 8)

        One Sample t-test

data:  mydata$Weights
t = -4.3121, df = 17, p-value = 0.0004725
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.871757 7.956021
sample estimates:
mean of x 
 7.913889 
```

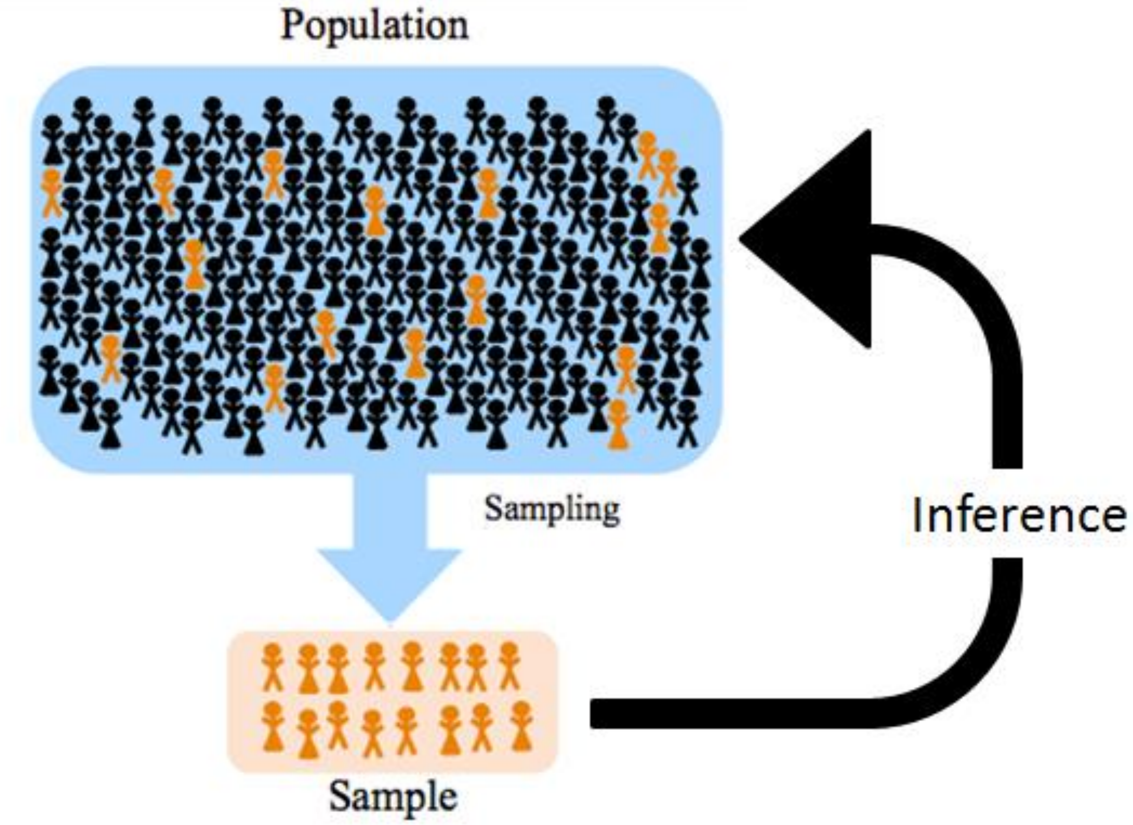
One population

- Mean DONE

- Proportion

- Variance

One population-Proportion



test whether the sample proportion differ from a population proportion

Inference about a population proportion

- Each sample point can result in just two possible outcomes.
  - we call one of these outcomes a success and the other, a failure.

- For example
  - Is the proportion of female students in the NYUST different from .50 ?
    - In this case, we call female students as a success and male students as a failure
    - We have 100 samples and 40 female students and 60 male students. So what is our p-hat?

Hypothesis test

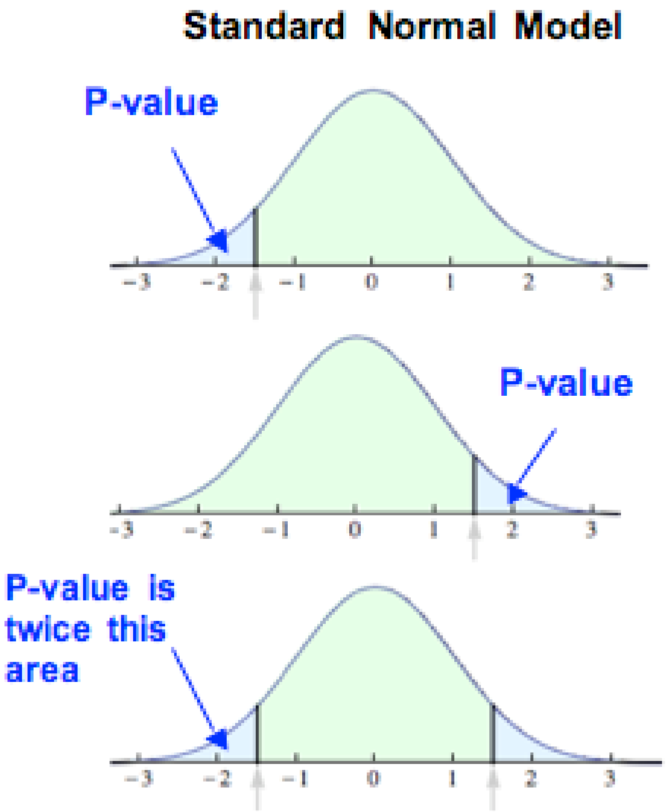
$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Approximate Normal distribution because

$np \geq 5$ & $n(1-p) \geq 5$

The difference between the sample proportion and hypothesized population proportion divided by the standard error of $\hat{p}$

Hypothesis test



$H_a : p < p_o$

Left-tailed P-value

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

$H_a : p > p_o$

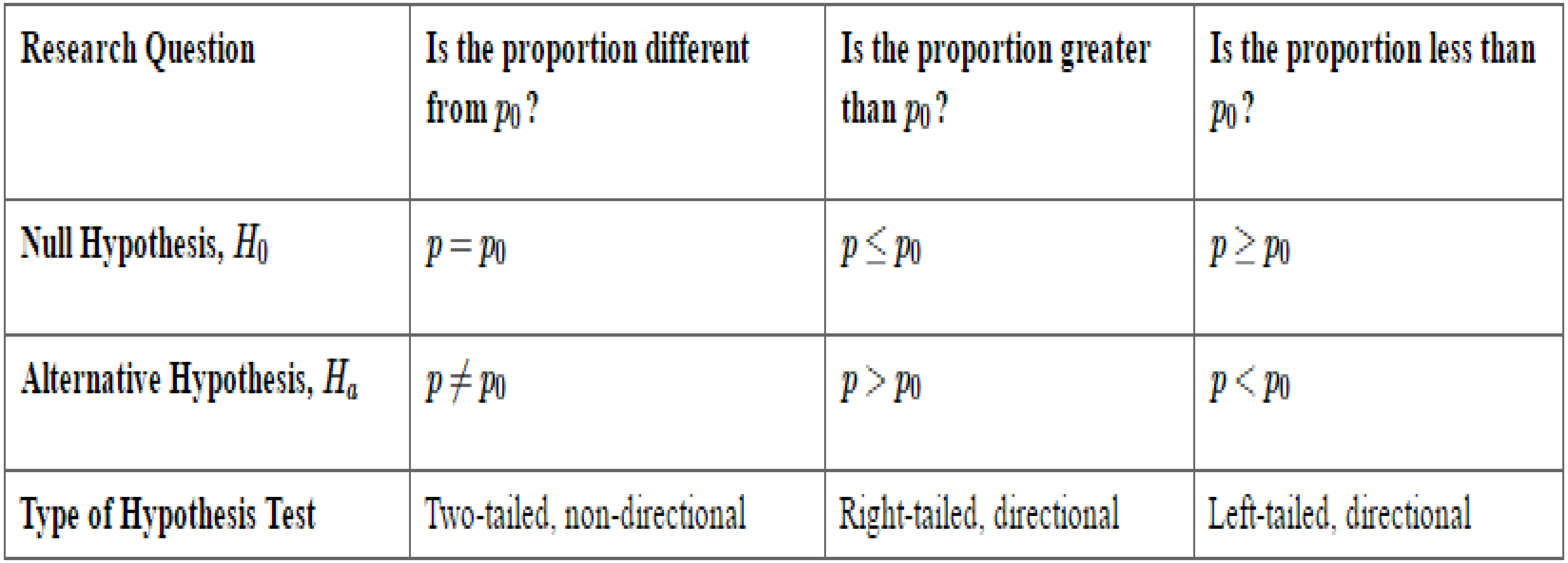
Right-tailed P-value

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

$H_a : p \neq p_o$

Two-tailed P-value

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

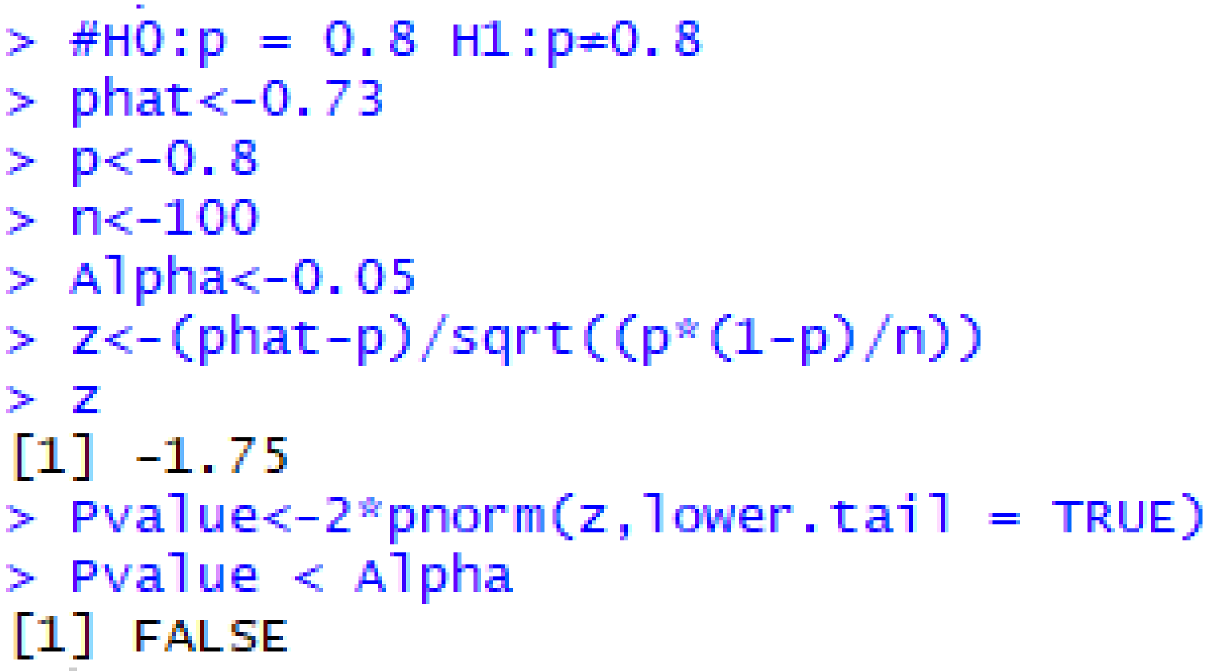


Research Question	Is the proportion different from p_0?	Is the proportion greater than p_0?	Is the proportion less than p_0?
Null Hypothesis, H_0	$p = p_0$	$p \leq p_0$	$p \geq p_0$
Alternative Hypothesis, H_a	$p \neq p_0$	$p > p_0$	$p < p_0$
Type of Hypothesis Test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed, directional

Problem

- The COB Dean claims that 80 percent of COB students are very satisfied with the student services they receive.
- To test this claim, we surveyed 100 students, using simple random sampling. Among the sampled students, 73 percent say they are very satisfied.
- Can we reject the Dean's hypothesis that 80% of the students are very satisfied? Use a 0.05 level of significance.

Ans



```
> #H0:p = 0.8 H1:p≠0.8
> phat<-0.73
> p<-0.8
> n<-100
> Alpha<-0.05
> z<-(phat-p)/sqrt((p*(1-p)/n))
> z
[1] -1.75
> Pvalue<-2*pnorm(z,lower.tail = TRUE)
> Pvalue < Alpha
[1] FALSE
```

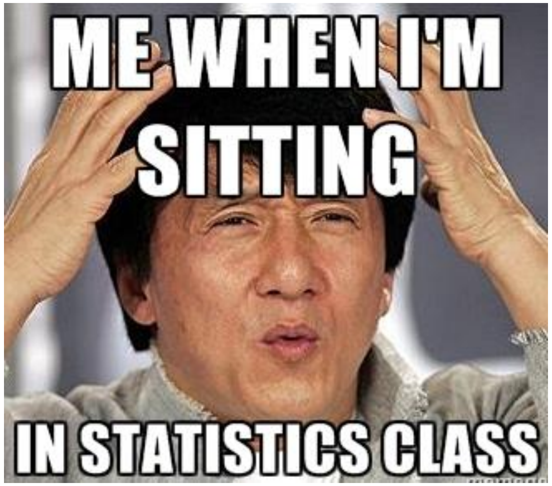
Non-rejection H0. At 5% significance level, we do not have sufficient evidence to reject the statement that 80% of the students are very satisfied with student services they receive.

Problem

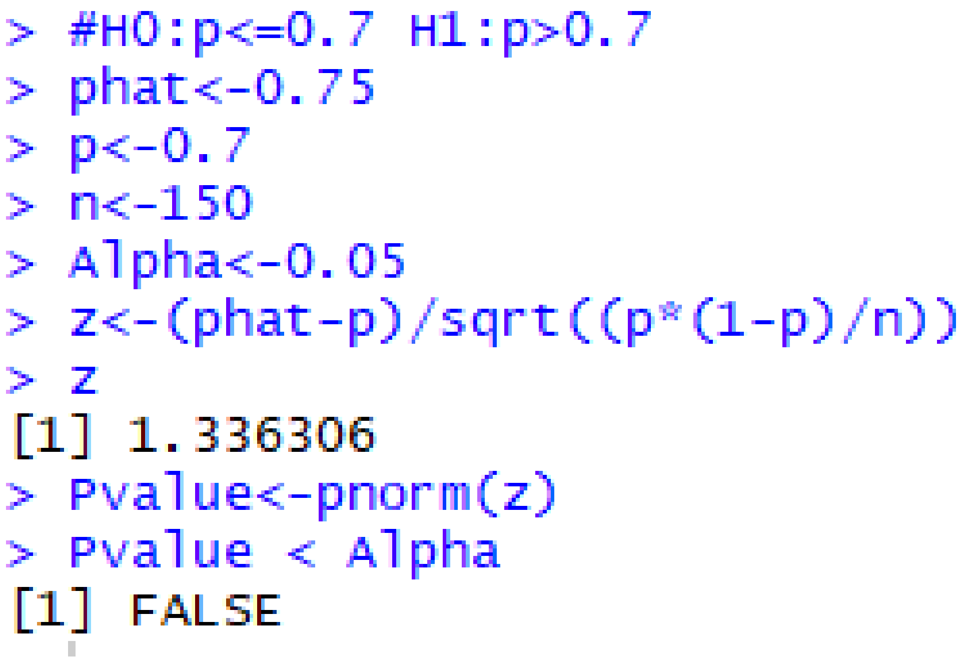
- The Dean claims that *at most* 70 percent of COB students are satisfied with the teaching.

- To test this claim, we surveyed 150 students, using simple random sampling. Among the sampled students, 75 percent say they are very satisfied.

- Can we reject the Dean's hypothesis that 70% of the students are very satisfied? Use a 0.05 level of significance.



Ans



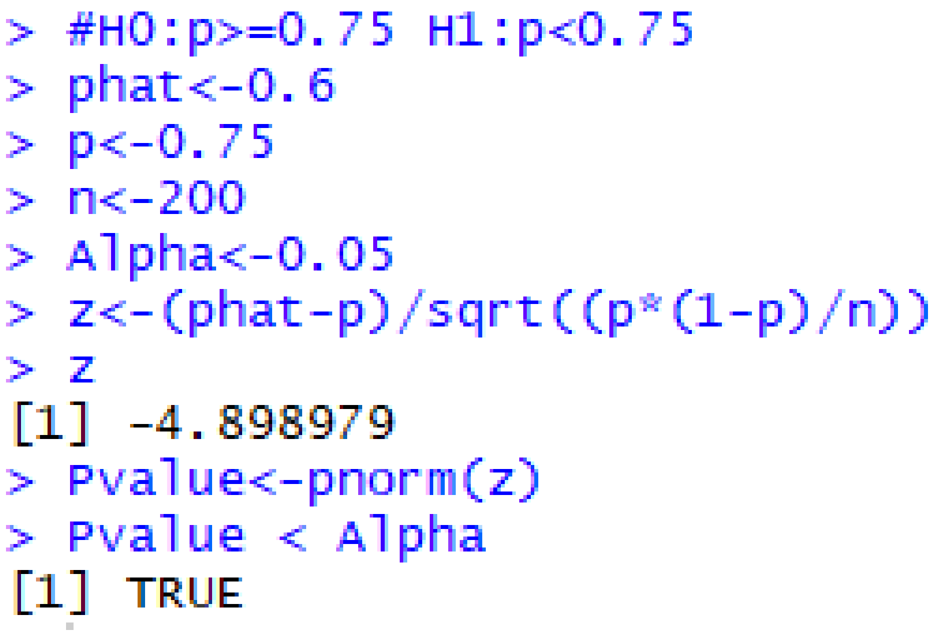
```
> #H0:p<=0.7 H1:p>0.7
> phat<-0.75
> p<-0.7
> n<-150
> Alpha<-0.05
> z<-(phat-p)/sqrt((p*(1-p)/n))
> z
[1] 1.336306
> Pvalue<-pnorm(z)
> Pvalue < Alpha
[1] FALSE
```

Non-rejection H0. At 5% significance level, we do not have sufficient evidence to reject the statement that *at most* 70 percent of COB students are satisfied with the teaching.

Problem

- The COB Dean claims that *at least* 75 percent of COB students are very satisfied with the tuition.
- To test this claim, we surveyed 200 students, using simple random sampling. Among the sampled students, 60 percent say they are very satisfied.
- Can we reject the Dean's hypothesis that 75% of the students are very satisfied? Use a 0.05 level of significance.

Ans



```
> #H0:p>=0.75 H1:p<0.75
> phat<-0.6
> p<-0.75
> n<-200
> Alpha<-0.05
> z<-(phat-p)/sqrt((p*(1-p)/n))
> z
[1] -4.898979
> Pvalue<-pnorm(z)
> Pvalue < Alpha
[1] TRUE
```

Reject H0. At 5% significance level, we do have sufficient evidence to reject the statement that *at least* 75 percent of COB students are very satisfied with the tuition.

REAL DATA ANALYSIS

Basic R programming

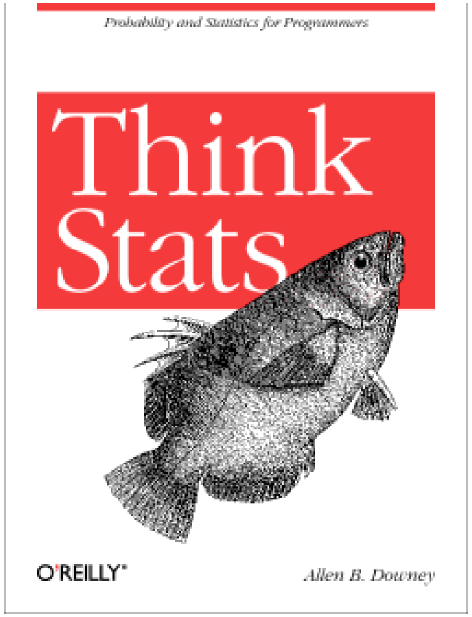
- as.factor(data$clumn)
  - Define a variable as a factor

- table(data)
  - Calculate the number of groups

- prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, correct = TRUE)
  - Use proportion test to calculate p value.

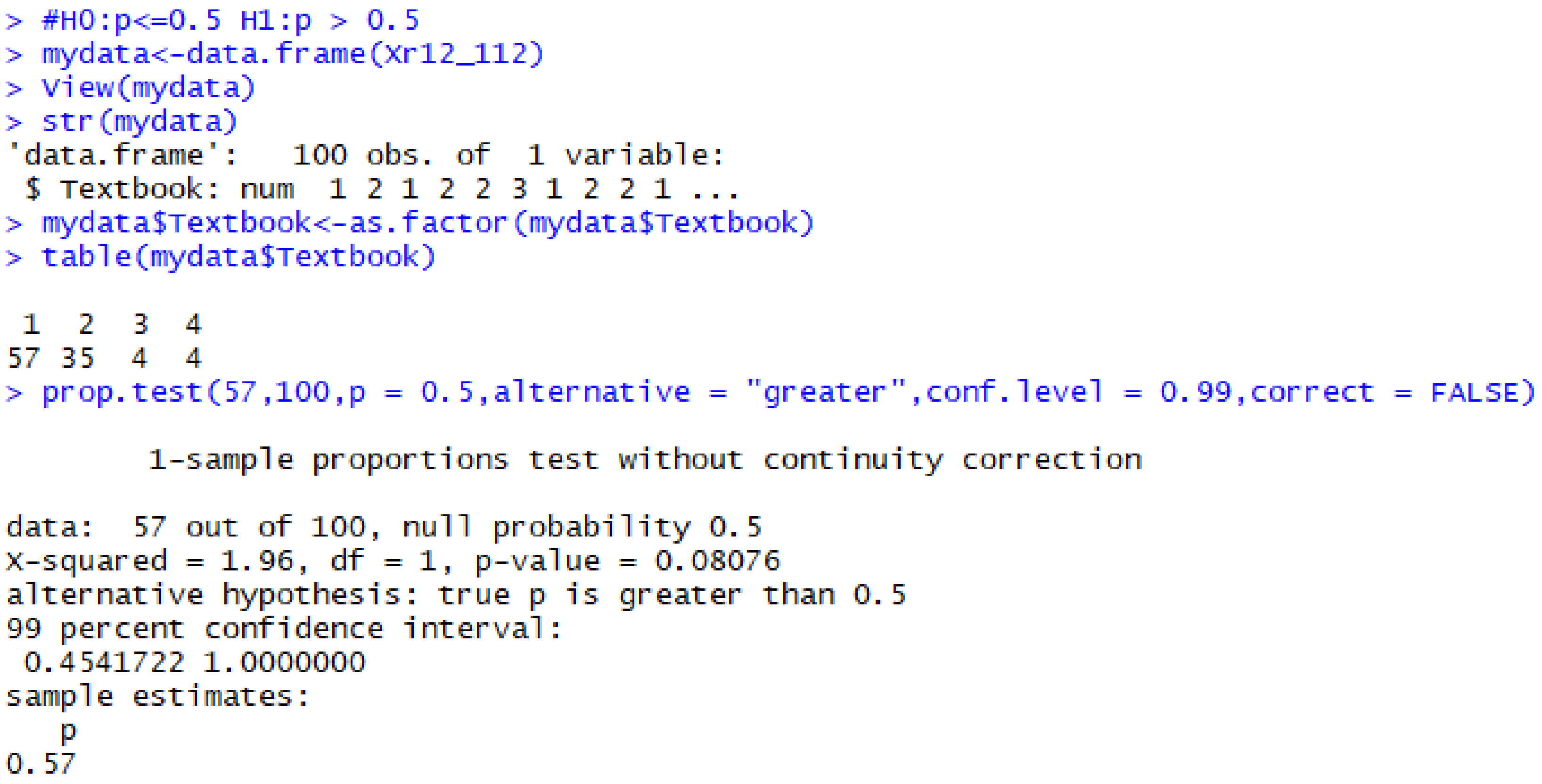
R dataset practices - Exercise 1

- (use Xr12-112)
- A professor of business stats recently adopted a new textbook. At the completion of the course, 100 randomly selected students were asked to access the book.



- The responses are as follows :
  - (1) = Excellent; (2) = Good; (3)= Adequate; (4) = Poor
- The results are stored using the codes. Do the data allow us to conclude that more the 50 percent of all business students would rate the book as excellent at 1% significance level?

Ans



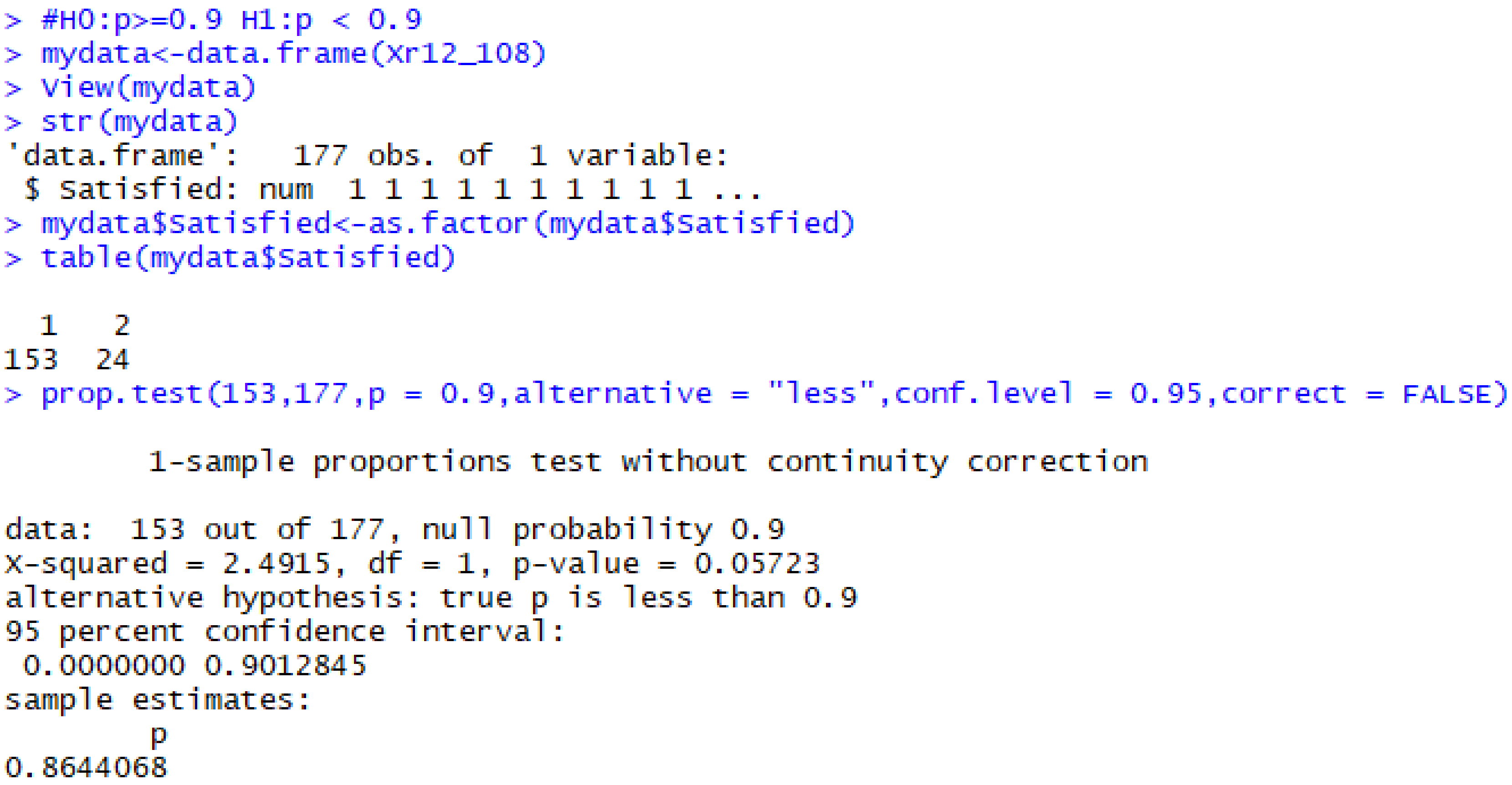
```
> #H0:p<=0.5 H1:p > 0.5
> mydata<-data.frame(Xr12_112)
> View(mydata)
> str(mydata)
'data.frame':	100 obs. of  1 variable:
 $ Textbook: num  1 2 1 2 2 3 1 2 2 1 ...
> mydata$Textbook<-as.factor(mydata$Textbook)
> table(mydata$Textbook)

 1  2  3  4 
57 35  4  4 
> prop.test(57,100,p = 0.5,alternative = "greater",conf.level = 0.99,correct = FALSE)

	1-sample proportions test without continuity correction

data:  57 out of 100, null probability 0.5
X-squared = 1.96, df = 1, p-value = 0.08076
alternative hypothesis: true p is greater than 0.5
99 percent confidence interval:
 0.4541722 1.0000000
sample estimates:
   p 
0.57 
```

R dataset practices - Exercise 2

- (use Xr12-108)
- The results of an annual Claimant Satisfaction Survey of policyholders who have had a claim with State Farm Insurance Company revealed a 90% satisfaction rate for claim service.

- To check the accuracy of this claim, a random sample of State Farm claimants was asked to rate whether they were satisfied with the quality of the service (1= satisfied and 2 = Unsatisfied). Use 5% significance level, can we infer that the satisfaction rate is less than 90%?

Ans

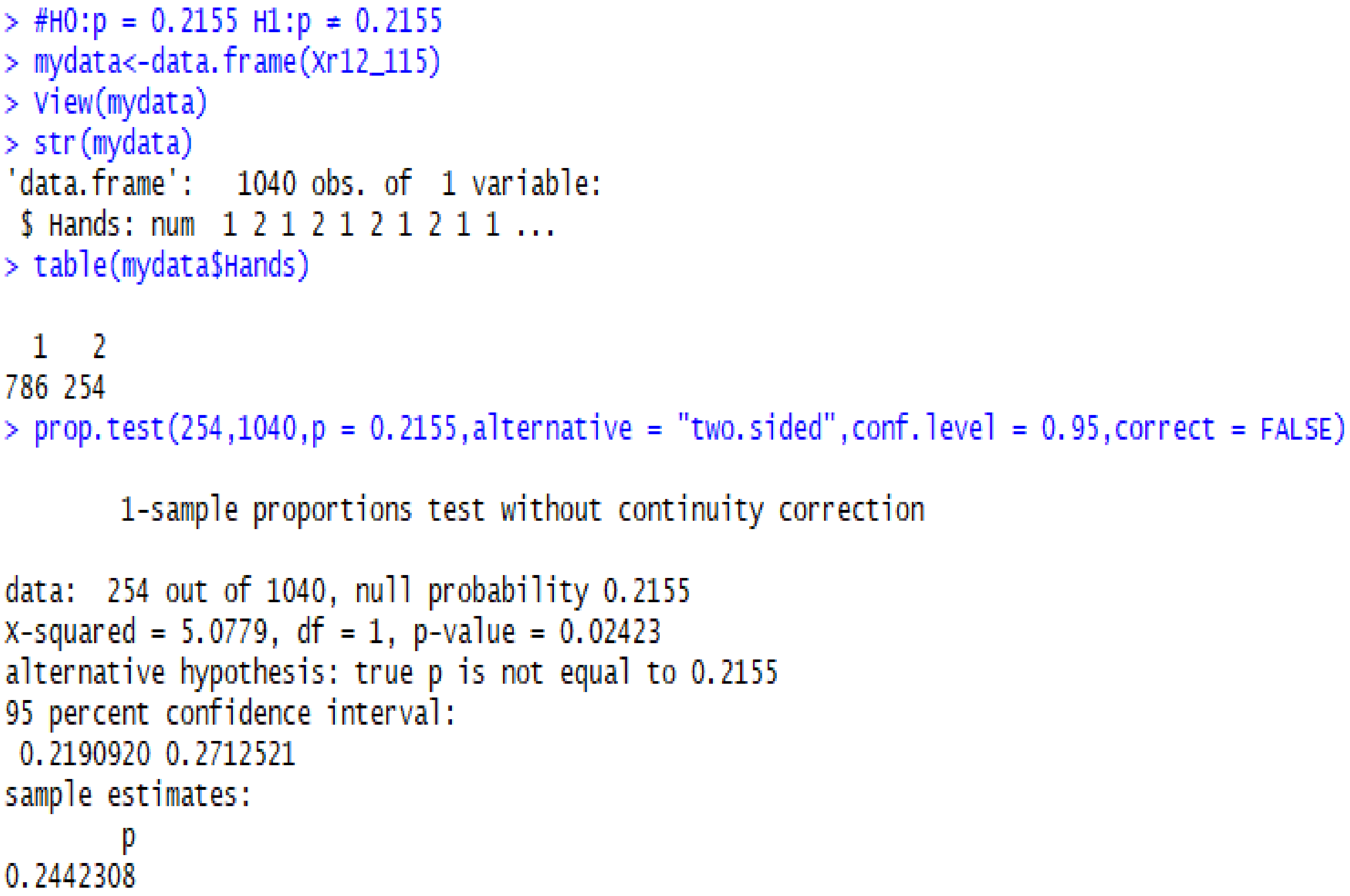


```
> #H0:p>=0.9 H1:p < 0.9
> mydata<-data.frame(Xr12_108)
> View(mydata)
> str(mydata)
'data.frame':	177 obs. of  1 variable:
 $ Satisfied: num  1 1 1 1 1 1 1 1 1 1 ...
> mydata$Satisfied<-as.factor(mydata$Satisfied)
> table(mydata$Satisfied)

  1   2 
153  24 
> prop.test(153,177,p = 0.9,alternative = "less",conf.level = 0.95,correct = FALSE)

	1-sample proportions test without continuity correction

data:  153 out of 177, null probability 0.9
X-squared = 2.4915, df = 1, p-value = 0.05723
alternative hypothesis: true p is less than 0.9
95 percent confidence interval:
 0.0000000 0.9012845
sample estimates:
        p 
0.8644068 
```

R dataset practices - Exercise 3

- (use Xr12-115)
- According to the ACBL bridge hands that contain two 4-card suits, one 3-card suit and one 2-card suits(4-4-3-2) occur with 21.55% probability.
- Suppose that a bridge-playing statistic professor with time on his hands tracked the number of hands over a one-year period and recorded the following hands with 4-4-3-2 (code 2) and others (code 1).
- Test to determine whether the proportion of 4-4-3-2 hands differs from the theoretical probability at 5% significance level.

Ans



```
> #H0:p = 0.2155 H1:p ≠ 0.2155
> mydata<-data.frame(Xr12_115)
> View(mydata)
> str(mydata)
'data.frame':   1040 obs. of  1 variable:
 $ Hands: num  1 2 1 2 1 2 1 2 1 1 ...
> table(mydata$Hands)

  1   2 
786 254 
> prop.test(254,1040,p = 0.2155,alternative = "two.sided",conf.level = 0.95,correct = FALSE)

        1-sample proportions test without continuity correction

data:  254 out of 1040, null probability 0.2155
X-squared = 5.0779, df = 1, p-value = 0.02423
alternative hypothesis: true p is not equal to 0.2155
95 percent confidence interval:
 0.2190920 0.2712521
sample estimates:
        p 
0.2442308 
```

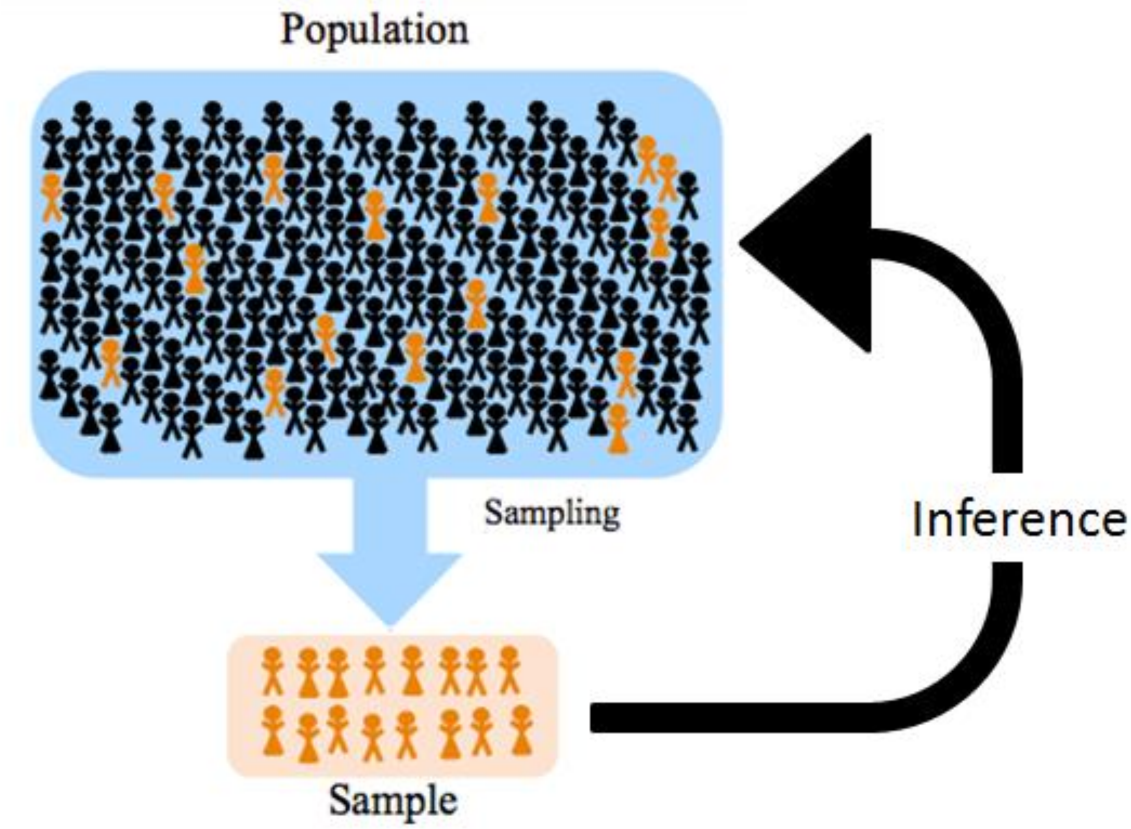
One population

- Mean DONE

- Proportion DONE

- Variance

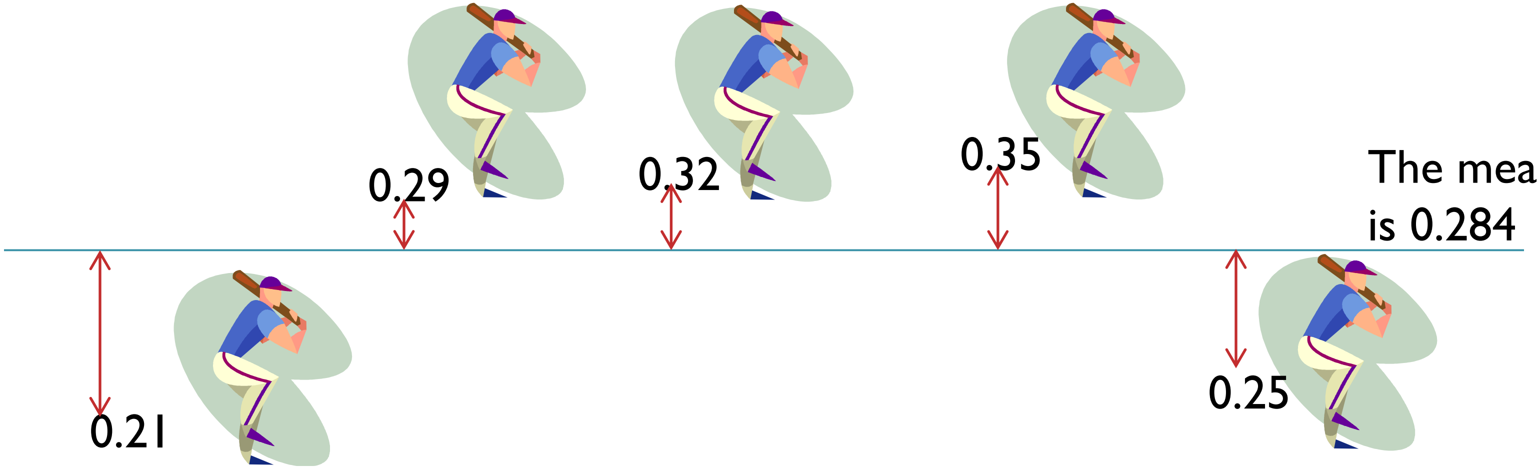
One population-Variance



test whether the sample variance differ from a population variance

Variability - Formula

The batting average is distributed from 0.21-0.35 in the five players



The mean is 0.284

The variance of a **population** is: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

The variance of a **sample** is: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

The variance is the average deviation from the mean, in "squared sum"
The standard deviation is the square root of the variance
The range is the largest observation – smallest observation

One population- Variance

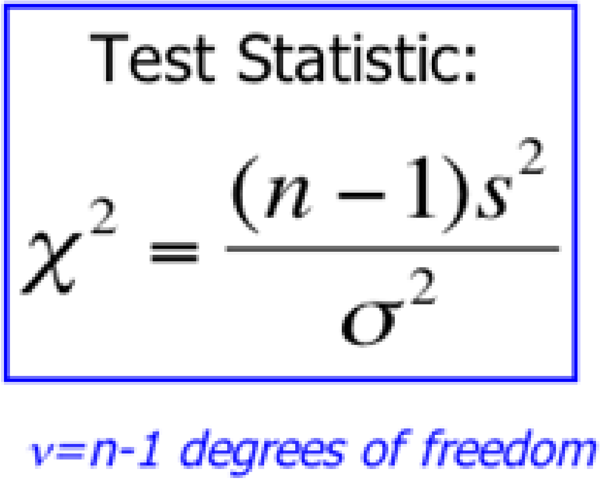
- A critical aspect of production is quality

If a sport shoes is not made to fit its specifications.

To improve the quality of products, we need to ensure there is a little variation

One population- Variance

- The chi-square(X^2) assignment can be used to make a hypothetical verification of one population variation.



Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

v=n-1 degrees of freedom

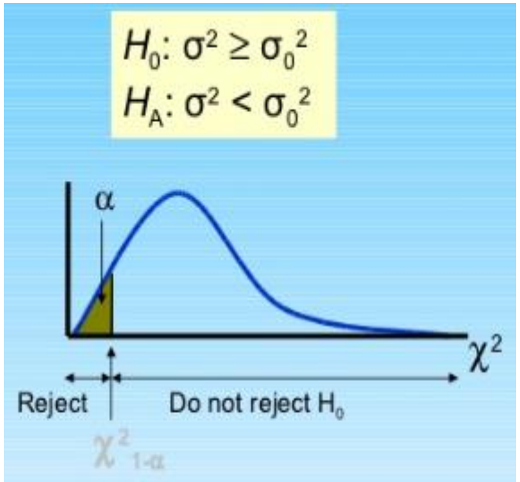
x^2 = standardized chi-square

n = sample size

s^2 = sample variance

σ^2 = hypothesized variance

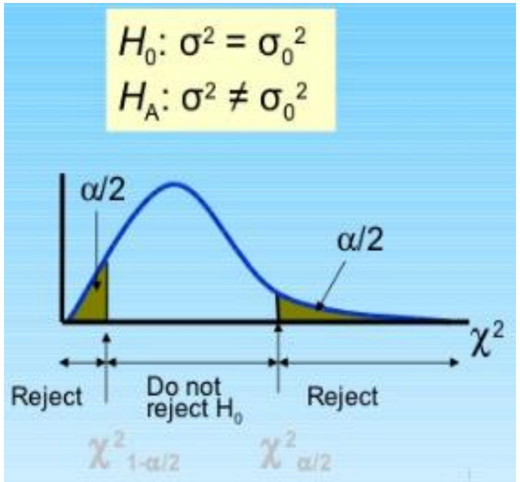
Hypothesis test


H_0: $\sigma^2 \geq \sigma_0^2$
H_A: $\sigma^2 < \sigma_0^2$
α
χ^2
Reject
Do not reject H_0
$\chi^2_{1-\alpha}$

Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

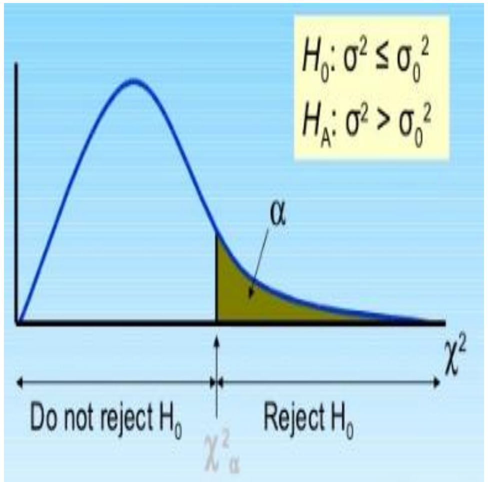
ν=n-1 degrees of freedom



Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

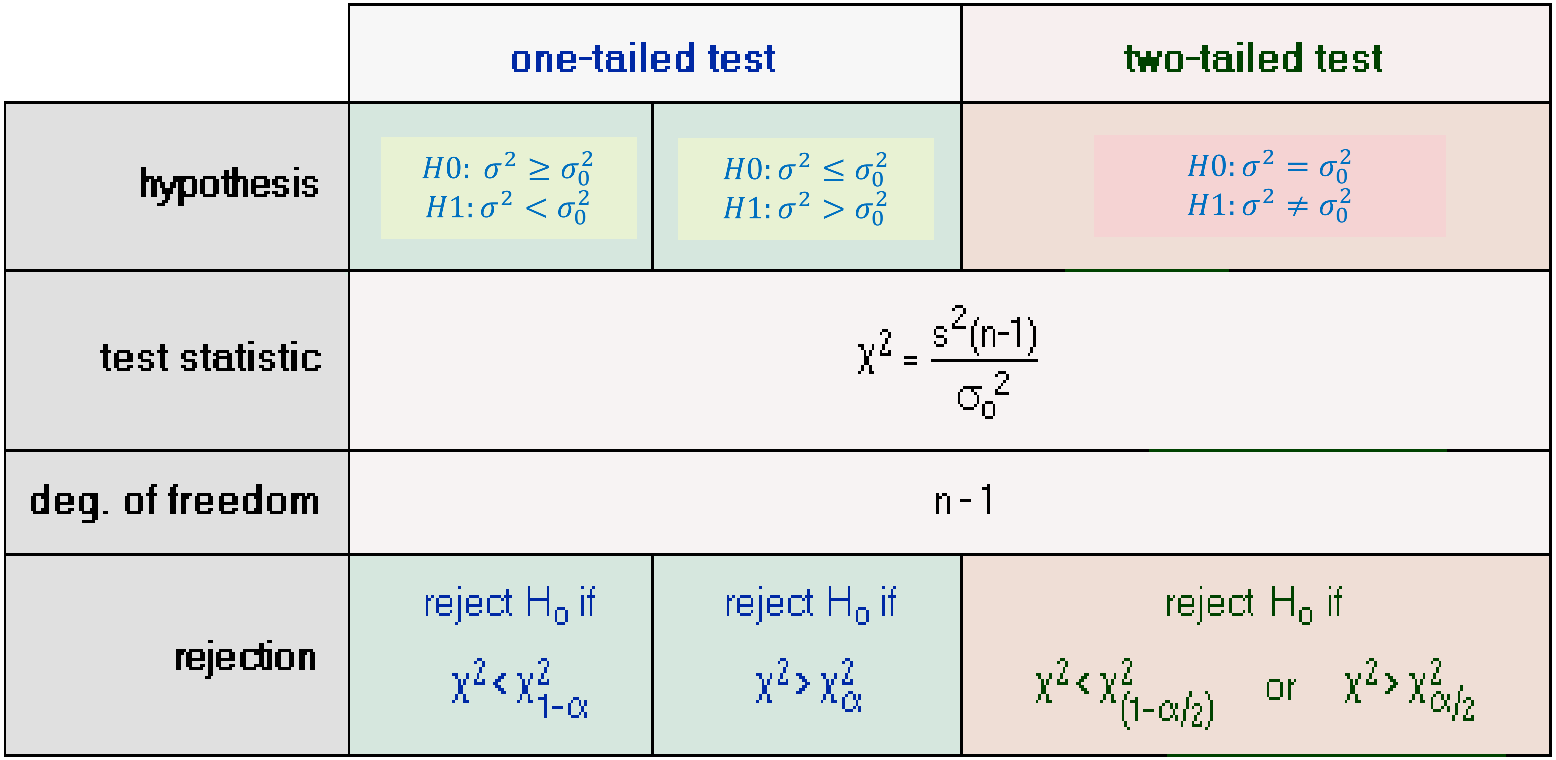
ν=n-1 degrees of freedom


H_0: $\sigma^2 \leq \sigma_0^2$
H_A: $\sigma^2 > \sigma_0^2$
α
χ^2
Do not reject H_0
Reject H_0
χ^2_{α}

Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

ν=n-1 degrees of freedom



	one-tailed test		two-tailed test
hypothesis	$H0: \sigma^2 \geq \sigma_0^2$ $H1: \sigma^2 < \sigma_0^2$	$H0: \sigma^2 \leq \sigma_0^2$ $H1: \sigma^2 > \sigma_0^2$	$H0: \sigma^2 = \sigma_0^2$ $H1: \sigma^2 \neq \sigma_0^2$
test statistic	$\chi^2 = \frac{s^2(n-1)}{\sigma_0^2}$		
deg. of freedom	n - 1		
rejection	reject H_0 if $\chi^2 < \chi^2_{1-\alpha}$	reject H_0 if $\chi^2 > \chi^2_{\alpha}$	reject H_0 if $\chi^2 < \chi^2_{(1-\alpha/2)}$ or $\chi^2 > \chi^2_{\alpha/2}$

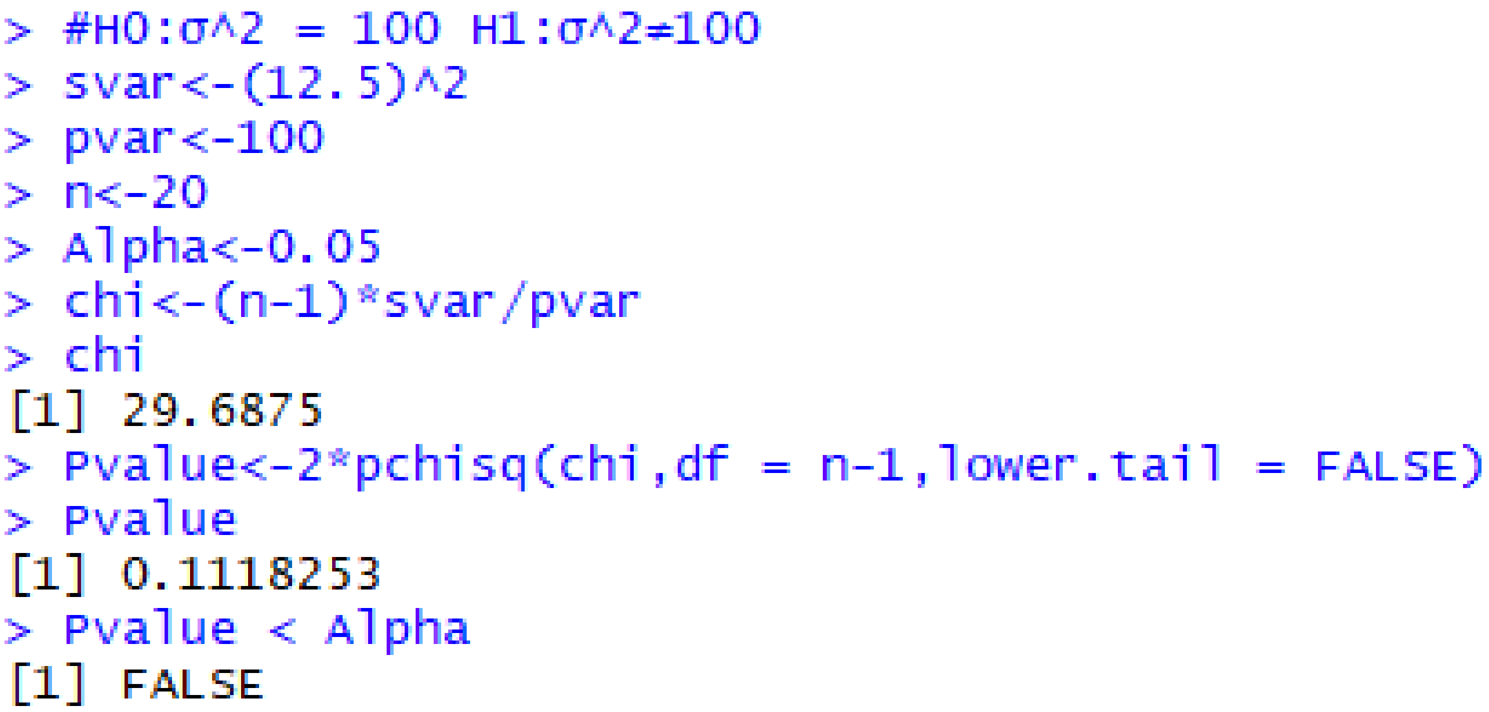
To find the p-value, we use “pchsiq”

Please type help(“pchisq”)

Problem

- You have a random sample of size 20, with a sample standard deviation of 12.5.
- You have good reason to believe that the underlying population is normal.
- Is the population variance different from 100, at the 0.05 significance level?

Ans



```
> #H0:σ^2 = 100 H1:σ^2≠100
> svar<-(12.5)^2
> pvar<-100
> n<-20
> Alpha<-0.05
> chi<-(n-1)*svar/pvar
> chi
[1] 29.6875
> Pvalue<-2*pchisq(chi,df = n-1,lower.tail = FALSE)
> Pvalue
[1] 0.1118253
> Pvalue < Alpha
[1] FALSE
```

Non-Rejection H0. At 5% significance level, we do not have sufficient evidence to suggest that the population variance is different from 100.

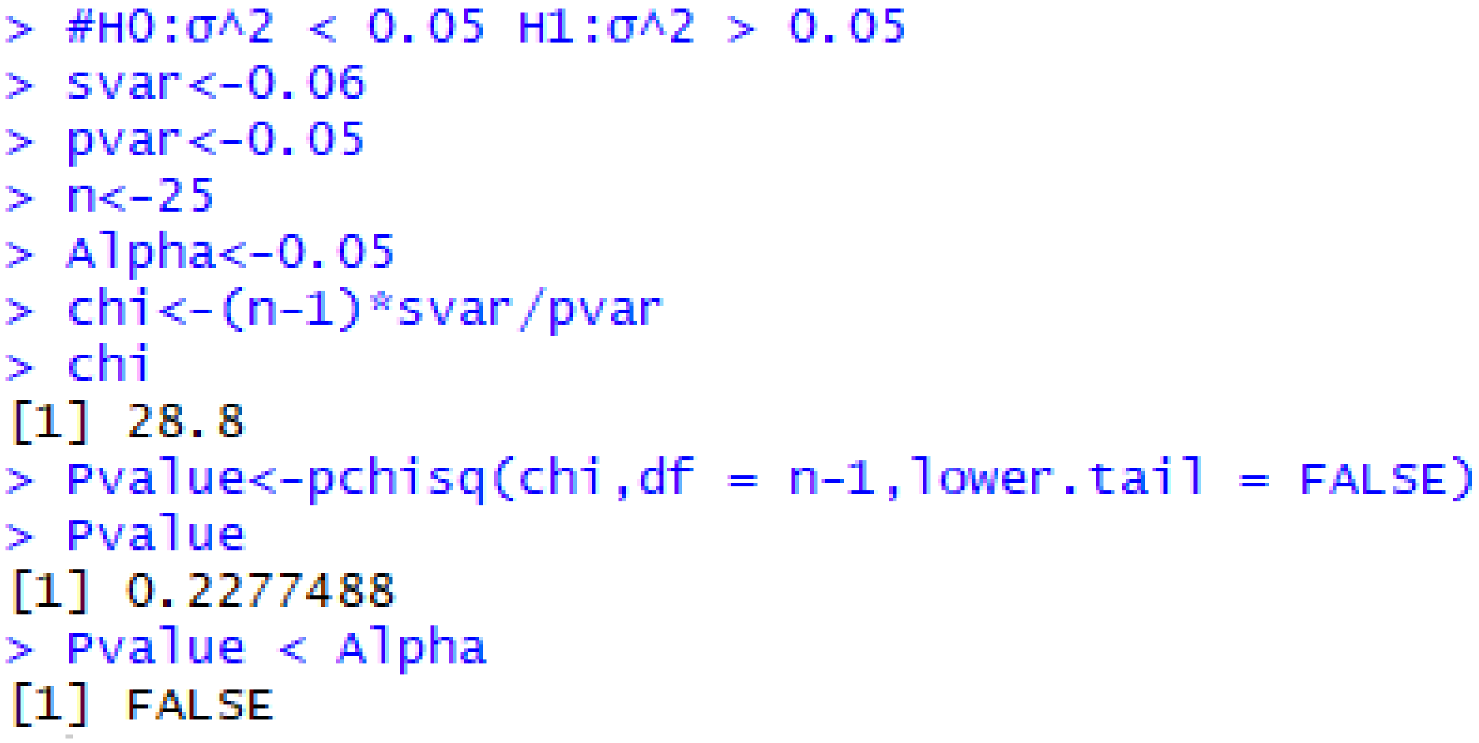
Problem

- You don’t want too much variation from sport shoes to sport shoes. You assume that a population variance of no more than 0.05 inch is acceptable.

- To determine whether the machine is operating within specification, you randomly select 25 shoes. The sample variance, which is 0.06.

- Is the population variance larger than 0.05, at the 0.05 significance level?

Ans



```
> #H0:σ^2 < 0.05 H1:σ^2 > 0.05
> svar<-0.06
> pvar<-0.05
> n<-25
> Alpha<-0.05
> chi<-(n-1)*svar/pvar
> chi
[1] 28.8
> Pvalue<-pchisq(chi,df = n-1,lower.tail = FALSE)
> Pvalue
[1] 0.2277488
> Pvalue < Alpha
[1] FALSE
```

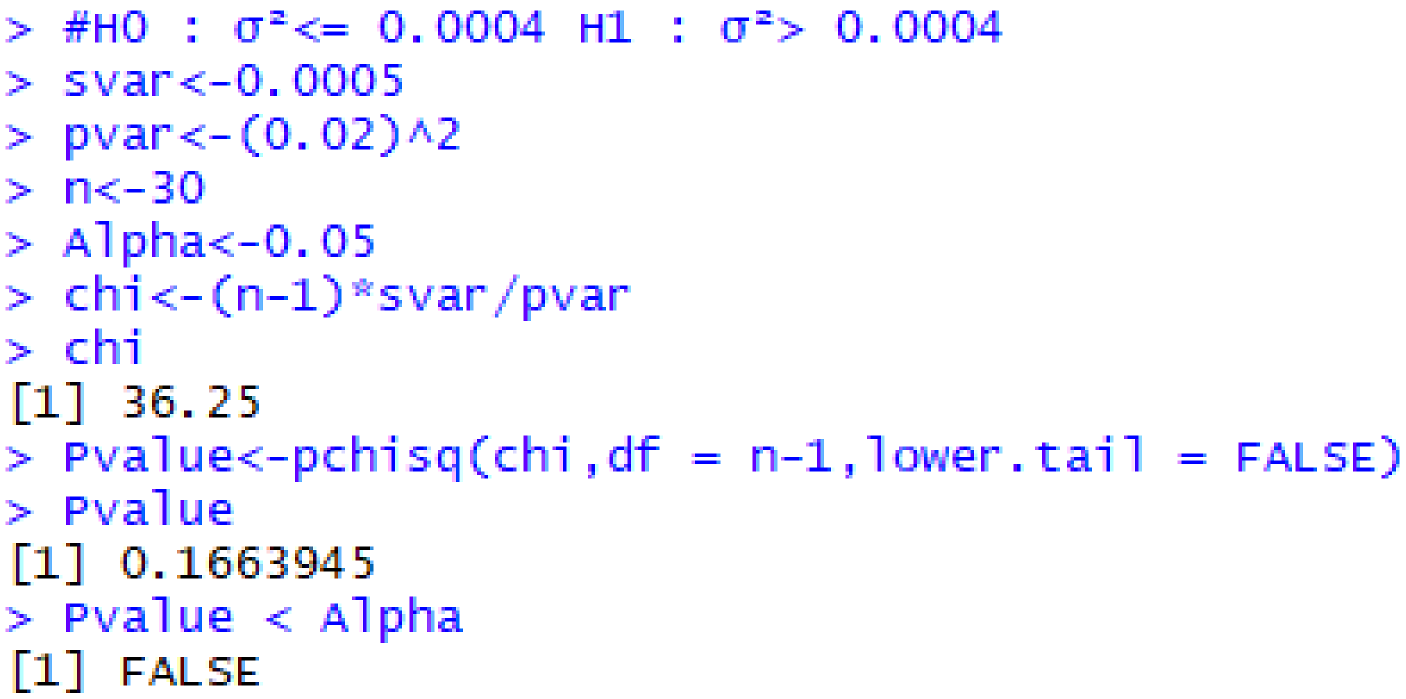
Non-Rejection H0. At 5% significance level, we do not have sufficient evidence to suggest that the population variance is larger than 0.05

Problem

- The manufacturing specification for a part requires a maximum standard deviation of 0.02 inches for the length of the part.

- The sample variation of the 30 parts extracted by the company is S^2 = 0.0005. Is the requirement of this manufacturing specification reached with α = 0.05?

Ans



```
> #H0 : σ²<= 0.0004 H1 : σ²> 0.0004
> svar<-0.0005
> pvar<-(0.02)^2
> n<-30
> Alpha<-0.05
> chi<-(n-1)*svar/pvar
> chi
[1] 36.25
> Pvalue<-pchisq(chi,df = n-1,lower.tail = FALSE)
> Pvalue
[1] 0.1663945
> Pvalue < Alpha
[1] FALSE
```

Non-Rejection H0. At 5% significance level, we do not have sufficient evidence to suggest that the population variance is larger than 0.0004

REAL DATA ANALYSIS

Basic R programming

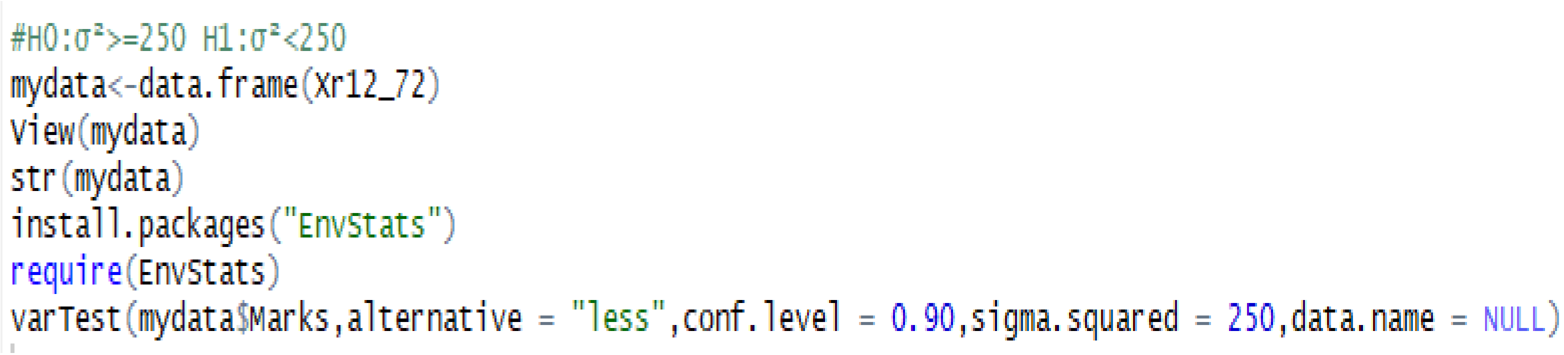
-
  ```
  install.packages("EnvStats")
  require(EnvStats)
  ```
  - Install package which is including varTest method.

-
  ```
  varTest(x, alternative = "two.sided", conf.level = 0.95, sigma.squared = 1, data.name = NULL)
  ```

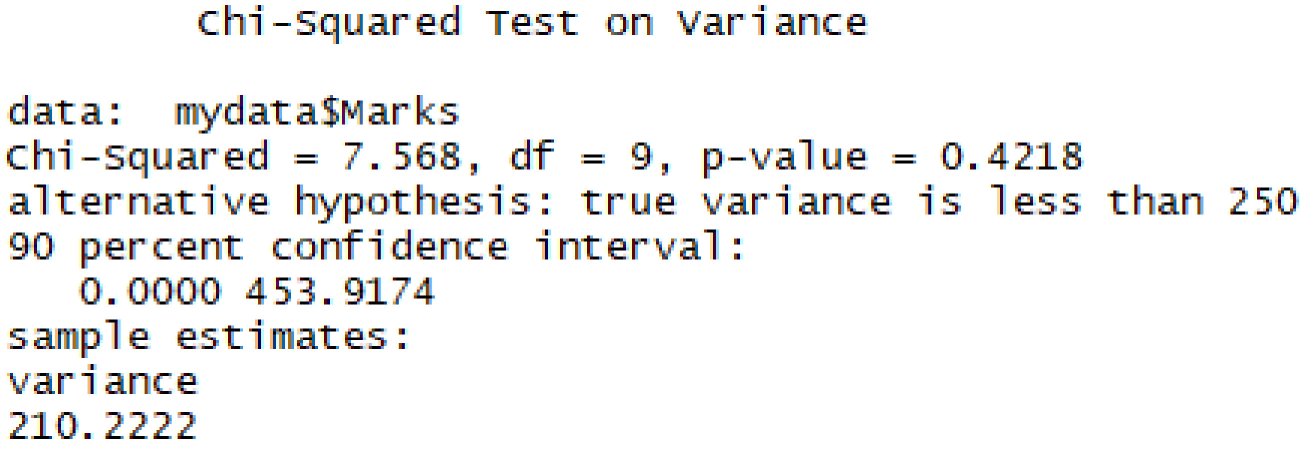
R dataset practices - Exercise 1

- (use Xr12-72)
- After many years of teaching, a statistics professor computed the variance of the marks on her final exam and found the population variance to be 250. She recently made changes to the way in which the final exam is marked and wondered whether this would result in a reduction in the variance.
- A random sample of this year's final exam marks are listed here. Can the professor infer at the 10% significance level that the variance has decreased?

Ans



```
#H0:σ²>=250 H1:σ²<250
mydata<-data.frame(Xr12_72)
View(mydata)
str(mydata)
install.packages("EnvStats")
require(EnvStats)
varTest(mydata$Marks,alternative = "less",conf.level = 0.90,sigma.squared = 250,data.name = NULL)
```

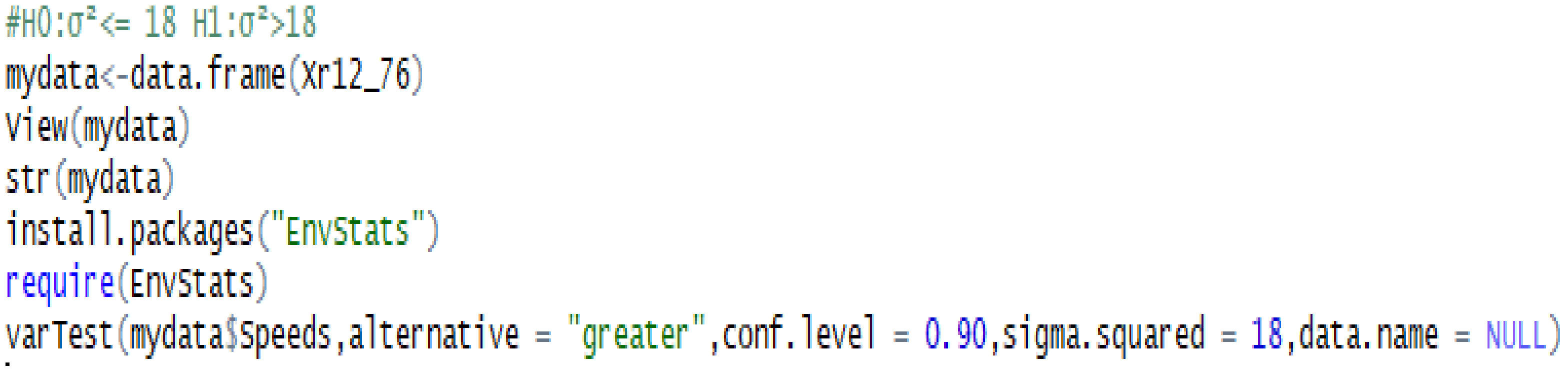


```
        Chi-Squared Test on Variance

data:  mydata$Marks
Chi-Squared = 7.568, df = 9, p-value = 0.4218
alternative hypothesis: true variance is less than 250
90 percent confidence interval:
   0.0000 453.9174
sample estimates:
variance
210.2222
```

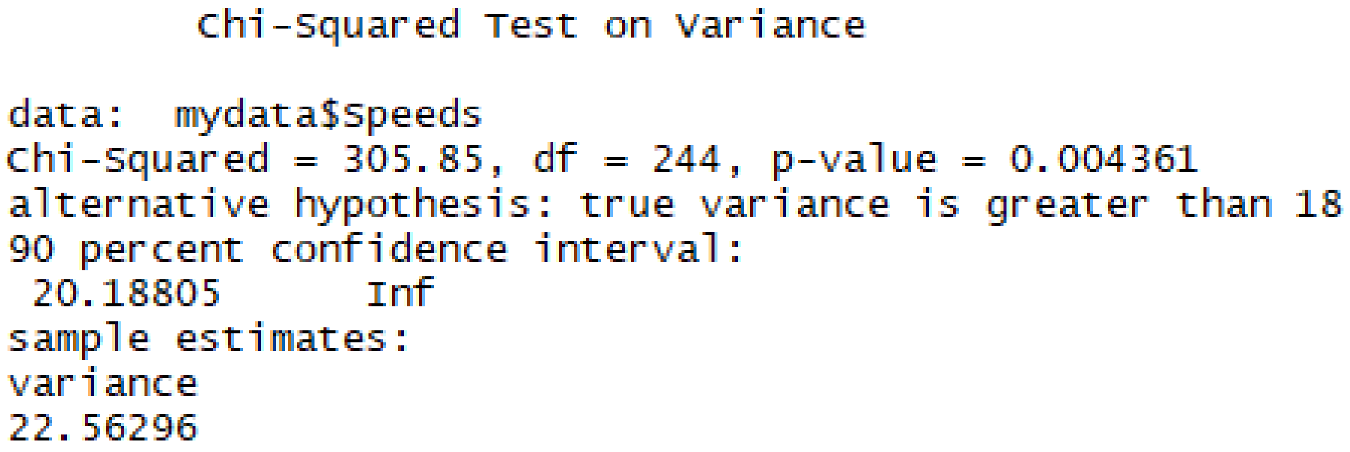
R dataset practices - Exercise 2

- (use Xr12-76)
- Some traffic experts believe that the major cause of highway collisions is the differing speeds of cars. That is, when some cars are driven slowly while others are driven at speeds well in excess of the speed limit, cars tend to congregate in bunches, increasing the probability of accidents. Thus, the greater the variation in speeds, the greater will be the number of collisions that occur.
- Suppose that one expert believes that when the variance exceeds 18 mph, the number of accidents will be unacceptably high. A random sample of the speeds of 245 cars on a highway with one of the highest accident rates in the country is taken. Can we conclude at the 10% significance level that the variance in speeds exceeds 18 mph.

Ans



```
#H0:σ²<= 18 H1:σ²>18
mydata<-data.frame(Xr12_76)
View(mydata)
str(mydata)
install.packages("EnvStats")
require(EnvStats)
varTest(mydata$Speeds,alternative = "greater",conf.level = 0.90,sigma.squared = 18,data.name = NULL)
```

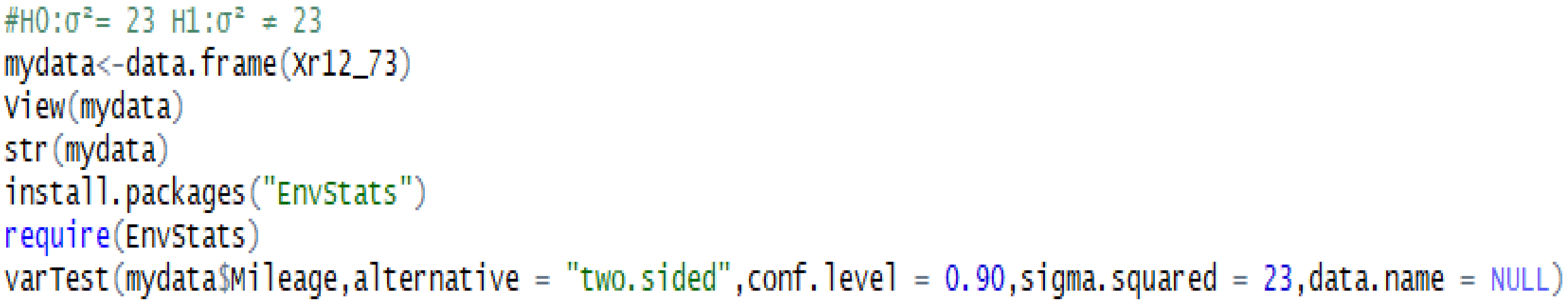


```
        Chi-Squared Test on Variance

data:  mydata$Speeds
Chi-Squared = 305.85, df = 244, p-value = 0.004361
alternative hypothesis: true variance is greater than 18
90 percent confidence interval:
 20.18805      Inf
sample estimates:
variance
22.56296
```

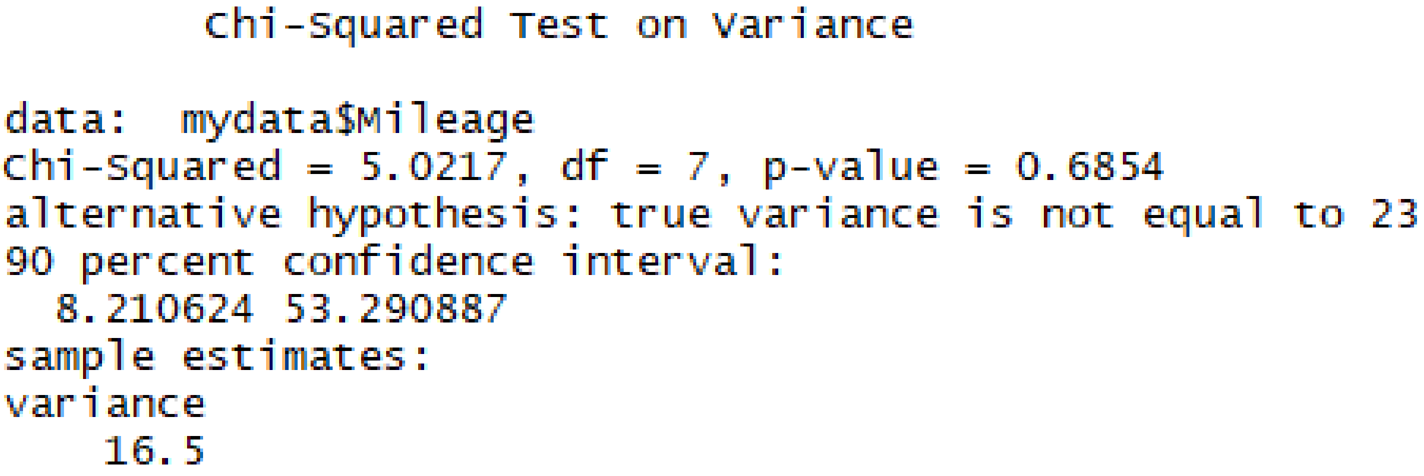
R dataset practices - Exercise 3

- (use Xr12-73)
- With gasoline prices increasing, drivers are more concerned with their cars' gasoline consumption. For the past 5 years a driver has tracked the gas mileage of his car and found that the population variance from fill-up to fill-up was 23. Now that his car is 5 years old, he would like to know whether the variability of gas mileage from his last eight fill-ups; these are listed in the data.
- Conduct a test at a 10% significance level to infer whether the variability has changed.

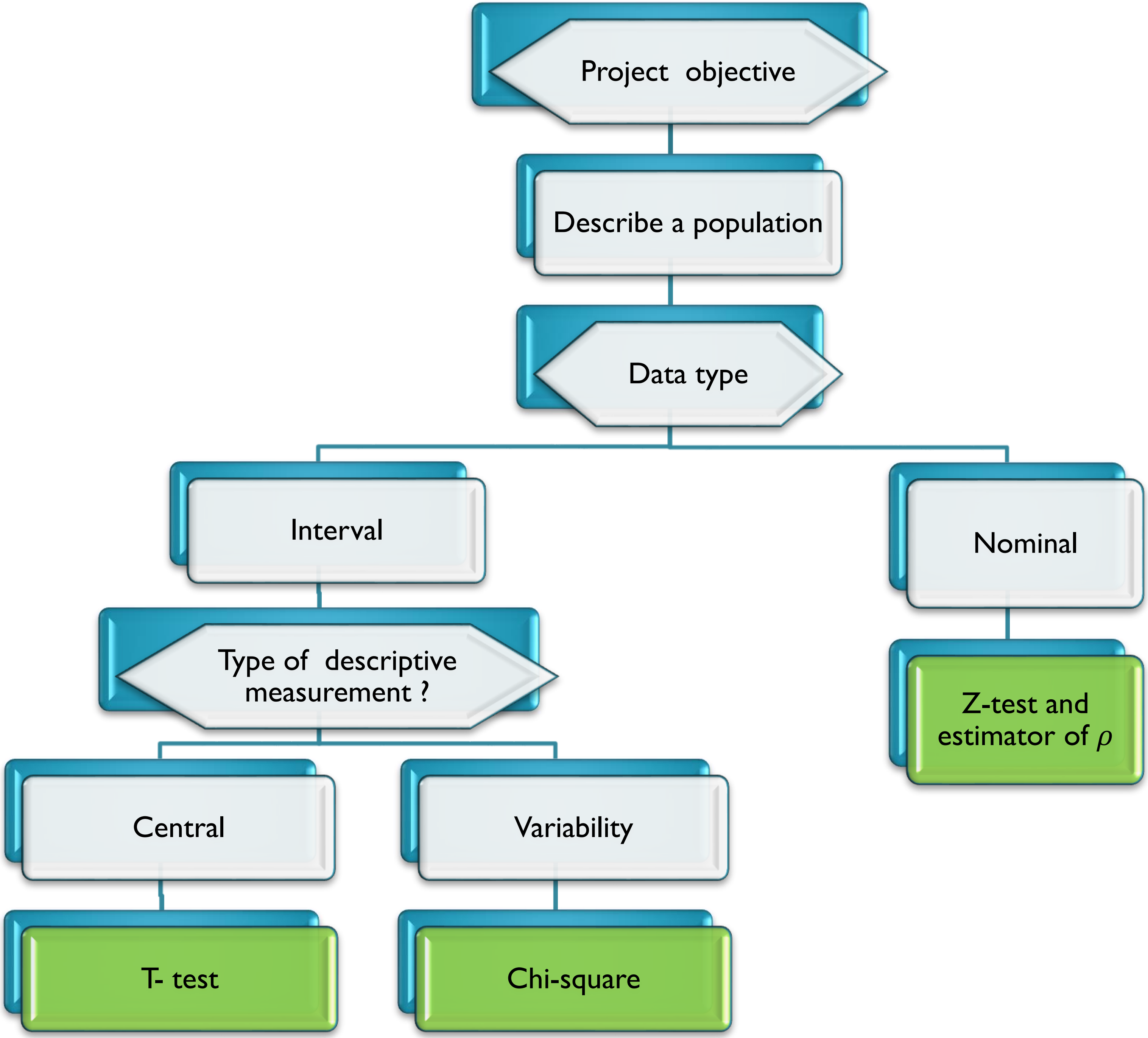
Ans



```
#H0:σ²= 23 H1:σ² ≠ 23
mydata<-data.frame(Xr12_73)
View(mydata)
str(mydata)
install.packages("EnvStats")
require(EnvStats)
varTest(mydata$Mileage,alternative = "two.sided",conf.level = 0.90,sigma.squared = 23,data.name = NULL)
```



```
        Chi-Squared Test on Variance

data:  mydata$Mileage
Chi-Squared = 5.0217, df = 7, p-value = 0.6854
alternative hypothesis: true variance is not equal to 23
90 percent confidence interval:
  8.210624 53.290887
sample estimates:
variance
    16.5
```



Project objective

Describe a population

Data type

Interval

Nominal

Type of descriptive measurement ?

Z-test and estimator of ρ

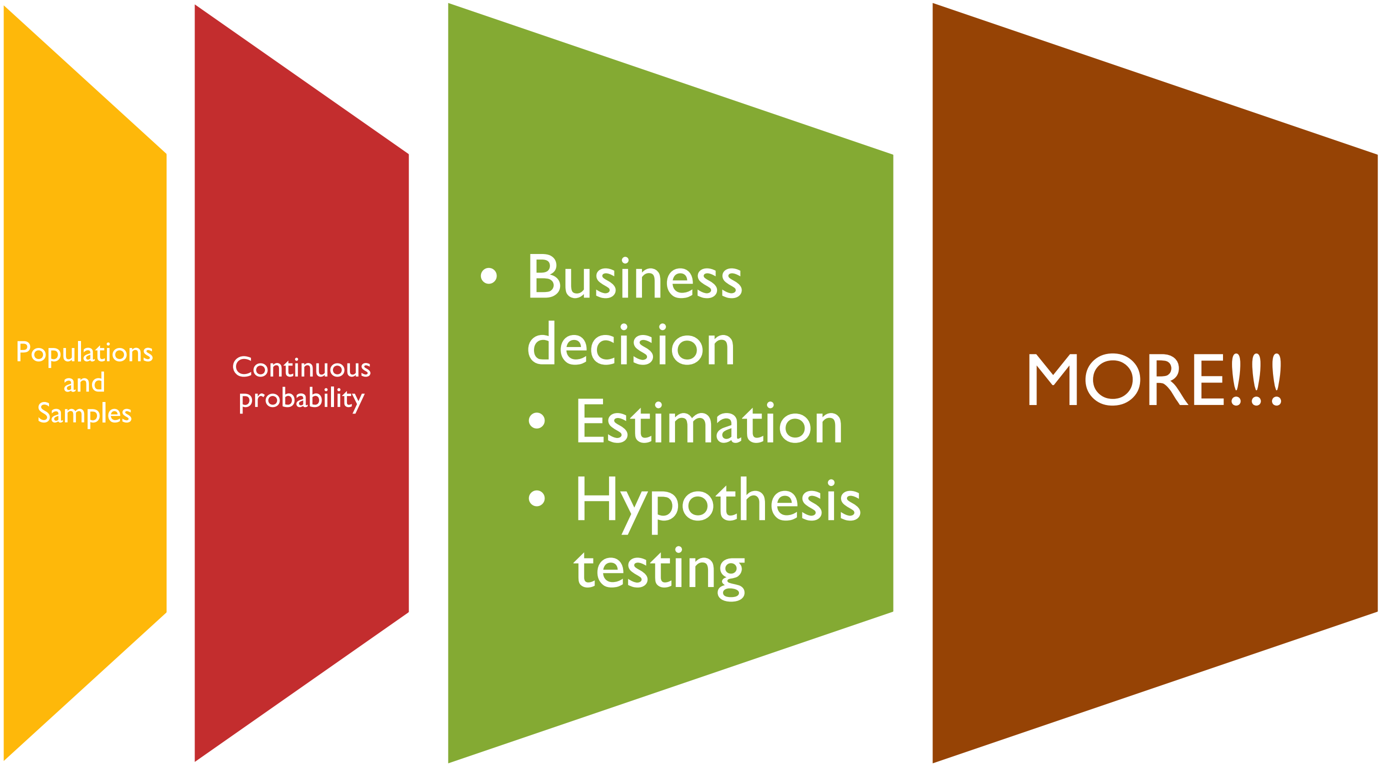
Central

Variability

T- test

Chi-square

Where are we and where are we going ?



Populations and Samples

Continuous probability

- Business decision
  - Estimation
  - Hypothesis testing

MORE!!!