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

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

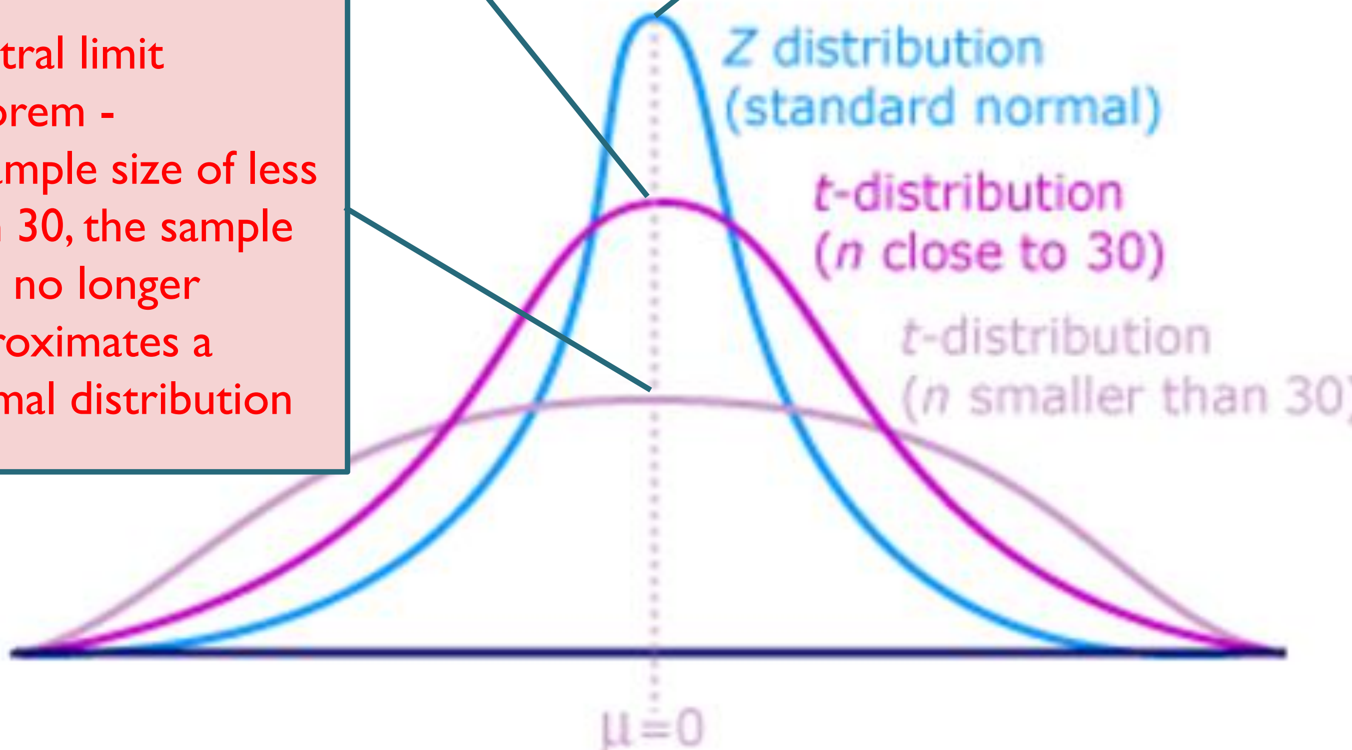
Now consider the case in which you have a normal distribution data but you do not know the population variance

Normal distribution vs t-distribution

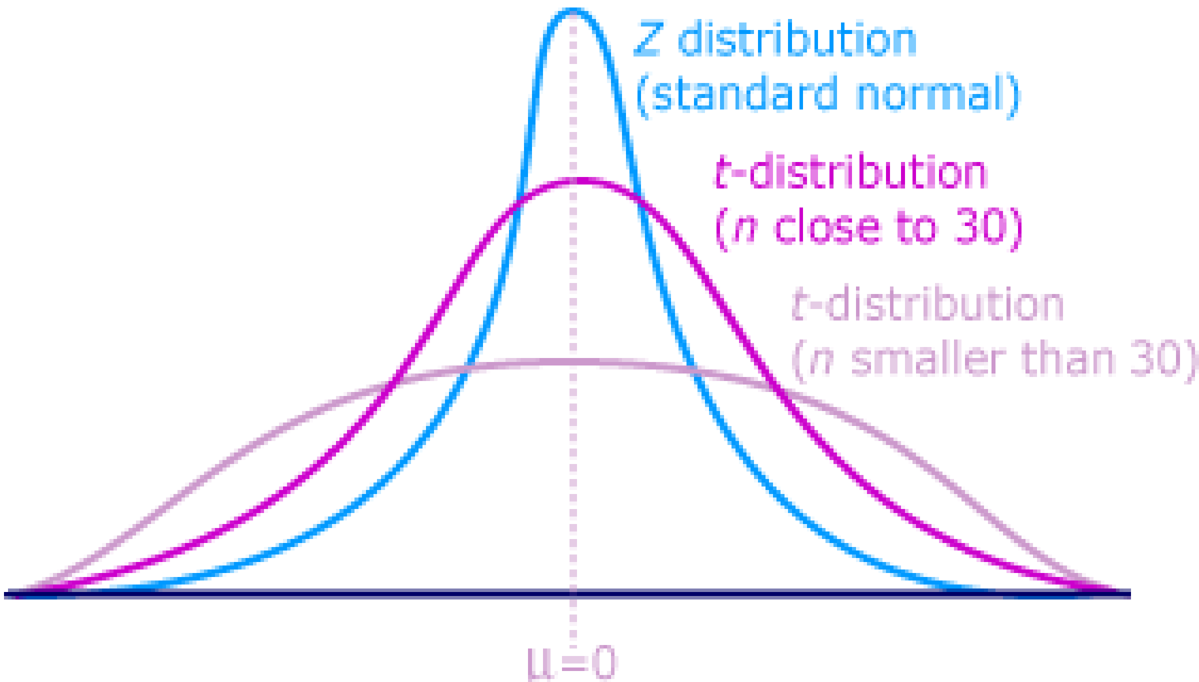
A t distribution is a sampling distribution

A *Z* distribution is a normal distribution

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



Z distribution and t distribution



We use the t distribution when the population standard deviation is unknown

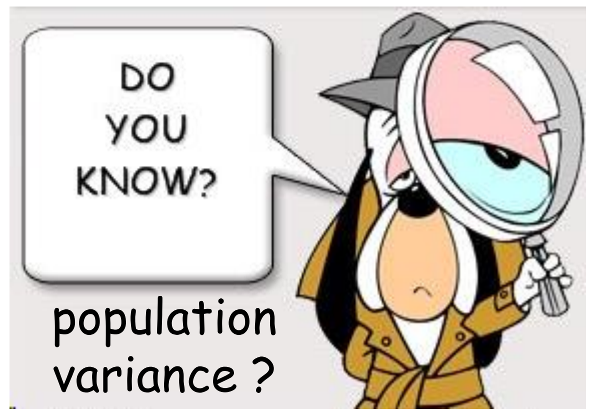
As *n* increases in size, the shape of the *t*-distribution begins to resemble a normal distribution

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**

Now, we are moving to t-distribution

Use z-distribution

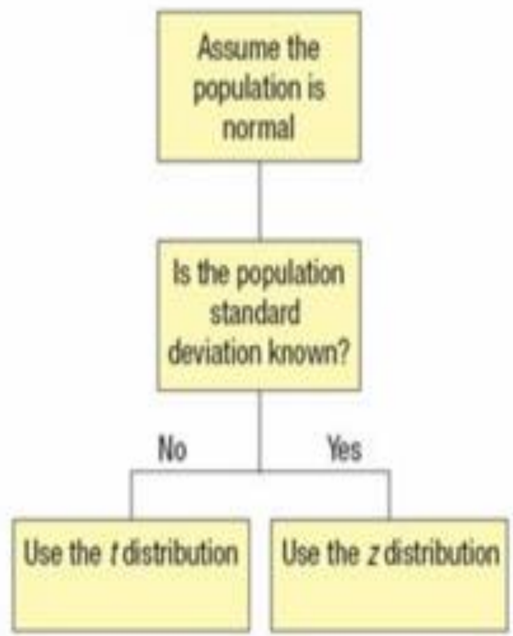
If the population standard deviation is known or the sample is greater than 30.

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

Use t-distribution

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

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



Now, we are moving to t-distribution

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

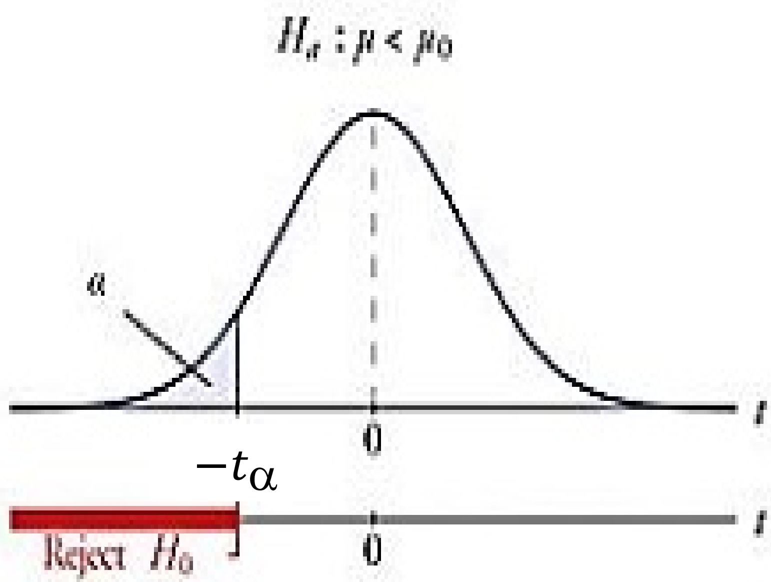
Inference about a population mean

	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 $\lvert t \rvert > t_{\alpha/2}$

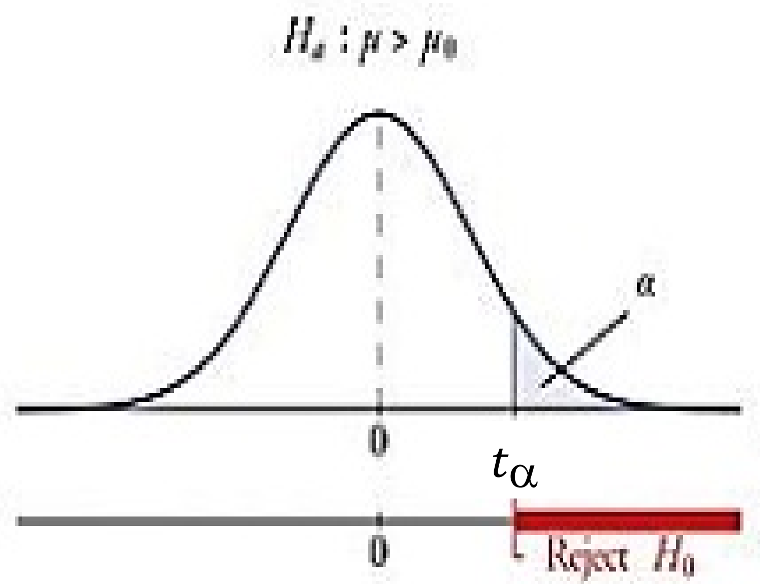
For a left-tailed test and a right-tailed test

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

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




```
> qt(0.05, df=20)
[1] -1.724718
```




```
> qt(1-0.05, df=20)
[1] 1.724718
```

For a two-tailed test

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

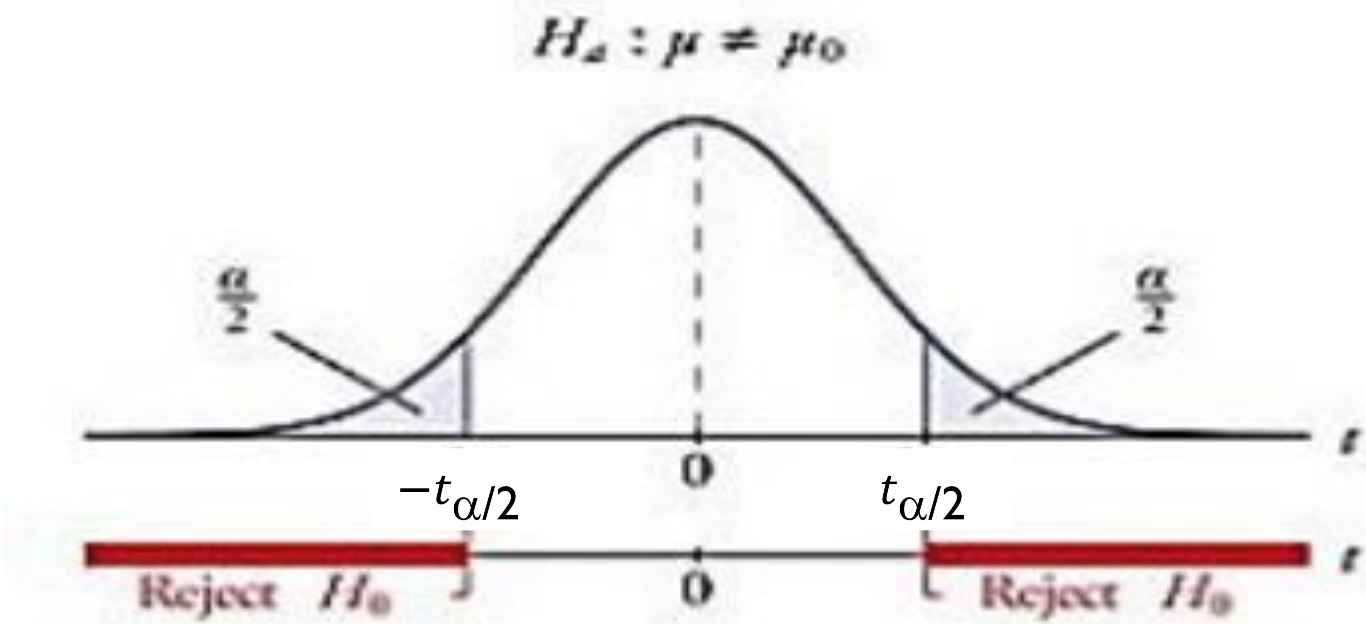
- Sample size = 30
- Significance level α is 0.1
- The degrees of freedom (df) = sample size -1

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

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

```
> qt(0.05/2, df=29)
[1] -2.04523
```

```
> qt(1-(0.05/2), df=29)
[1] 2.04523
```



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

- 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
```

Now you are a manager of a baseball team in MLB

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

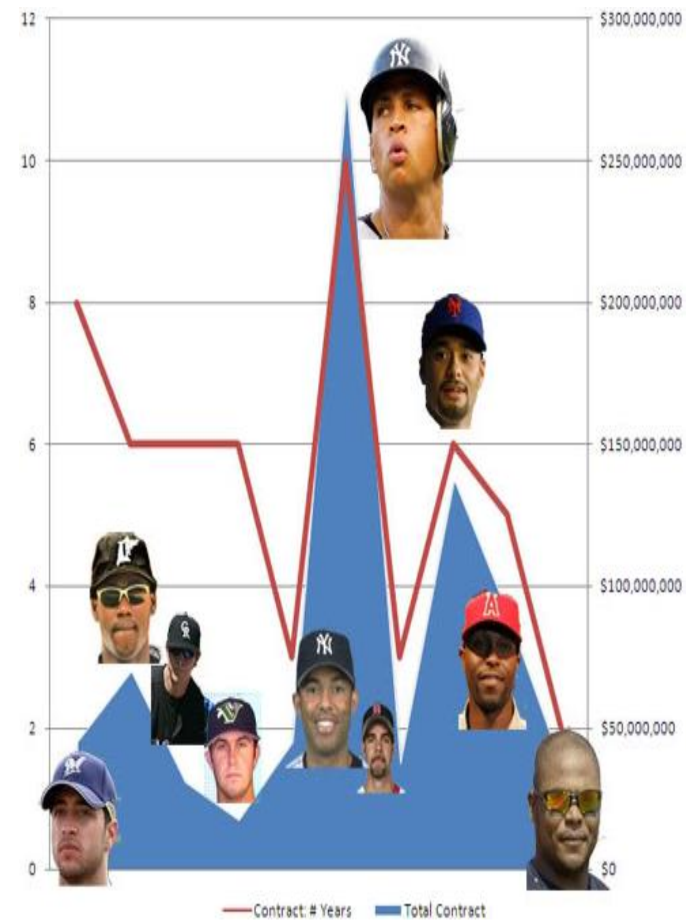
- Let's look at the batting average (AVG) in MLB from the years 1985 and 2013. You randomly recruit 25 players in your team.
- Their batting average is 0.265, and the sample standard deviation of is 0.03.
- Determine whether their batting average is significantly different from the 0.26. Set the significance level at 5%.

```
> xbar <- 0.265
> pmean <- 0.26
> ssd <- 0.03
> n <- 25
> t <- (xbar - pmean)/(ssd/sqrt(n))
> t
[1] 0.8333333
> qt((0.05/2), df=n-1)
[1] -2.063899
```

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

- Their (25 players) batting average is 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%.

```
> xbar <- 0.29
> pmean <- 0.26
> ssd <- 0.04
> n <- 25
> t <- (xbar - pmean)/(ssd/sqrt(n))
> t
[1] 3.75
> qt((1-0.05), df=n-1)
[1] 1.710882
```



However, the budget is limited….

- 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%.


```
> xbar <- 0.25
> pmean <- 0.26
> ssd <- 0.02
> n <- 25
> t <- (xbar - pmean)/(ssd/sqrt(n))
> t
[1] -2.5
> qt((0.05), df=n-1)
[1] -1.710882
```

One population

- Mean 

- Proportion 

- Variance 

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 ?

Population Proportion

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

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 $\hat{p}$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{B}\right)^2$$

Problem $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

- A survey done by a national research center, 747 out of 1168 college students said they have drink beer before the legal drinking age.

- Let's construct a 95% confidence interval for the proportion of college students in the population who have drink beer before the legal drinking age.

```
> 747/1168
[1] 0.6395548
> phat <- 0.6395548
> n <- 1168
> se <- abs(qnorm(0.05/2)*sqrt((phat*(1-phat))/n))
> lcl <- phat-se
> ucl <- phat+se
> ci <- c(lcl, ucl)
> ci
[1] 0.6120198 0.6670898
```

Problem $n = (\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{B})^2$

- Consider p, the true proportion of voters who favor a particular political candidate. A pollster is interested in finding 95 % confidence interval of $\hat{p}$

- The confidence interval will be no wider than the interval 0.03.

- Find the sample size n at the alternative $\hat{p}$ = 0.55.

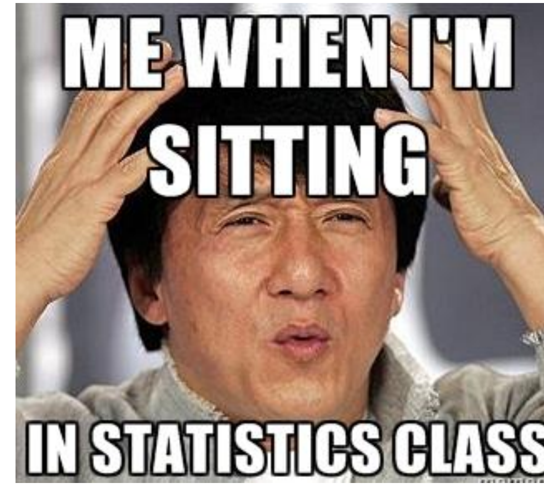
```
> phat <- 0.55
> b <- 0.03
> n <- ((qnorm(0.05/2)*sqrt(phat*(1-phat)))/b)^2
> round(n)
[1] 1056
> args(round)
function (x, digits = 0)
NULL
> round(n,1)
[1] 1056.4
> ?"ceiling"
> ceiling(n)
[1] 1057
```

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

- 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.

```
> phat <- 0.73
> p <- 0.8
> n <- 100
> z <- (phat-p)/sqrt((p*(1-p)/n))
> z
[1] -1.75
> qnorm(0.025)
[1] -1.959964
```

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



- 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.

```
> phat <- 0.75
> p <- 0.7
> n <- 150
> z <- (phat-p)/sqrt((p*(1-p)/n))
> z
[1] 1.336306
> qnorm(0.05)
[1] -1.644854
```

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

- 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.

```
> phat <- 0.60
> p <- 0.75
> n <- 200
> z <- (phat-p)/sqrt((p*(1-p)/n))
> z
[1] -4.898979
> qnorm(0.05)
[1] -1.644854
```

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

One population

- Mean 

- Proportion 

- Variance 

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

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

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} = \frac{\sum (x_i-\mu)^2}{\sigma^2}$$

$$\sum Z_i^2 = \frac{\sum (x_i-\bar{x})^2}{\sigma^2}$$

Test Statistic:

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

v=n-1 degrees of freedom

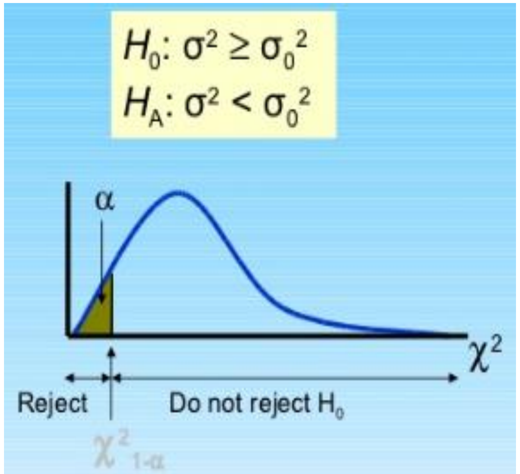
x^2 = standardized chi-square

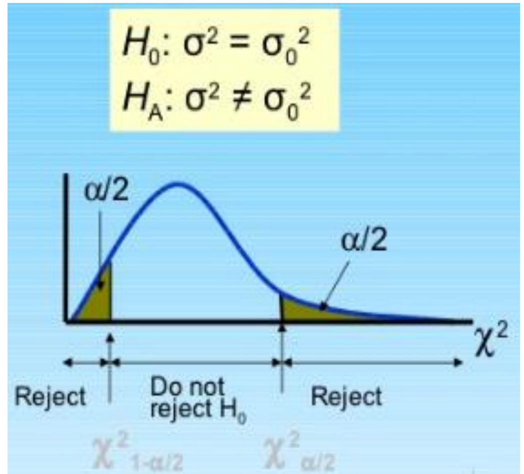
n = sample size

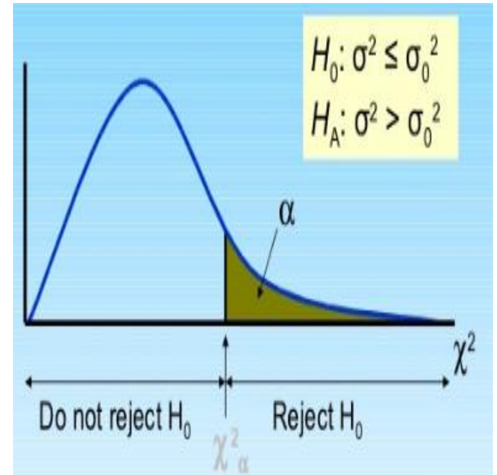
s^2 = sample variance

σ^2 = hypothesized variance

Hypotheses testing







Test Statistic:

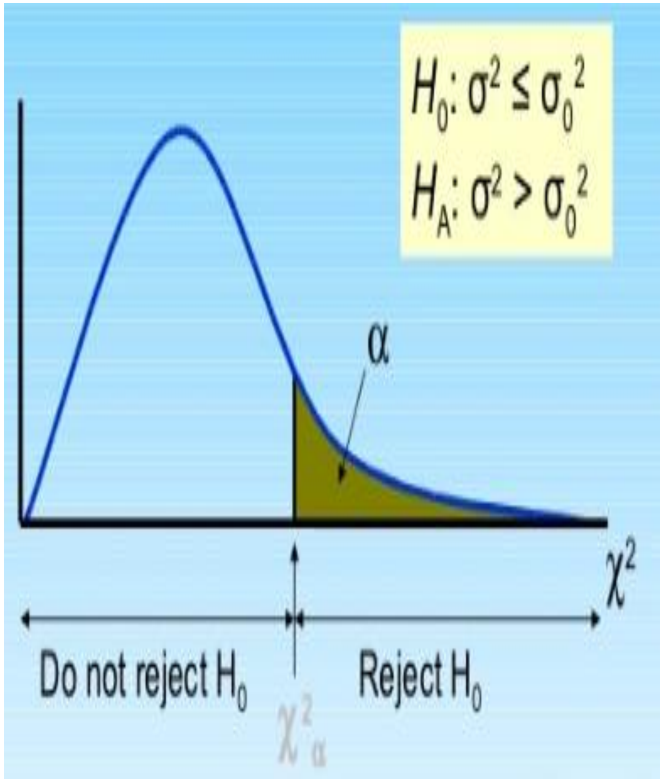
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

ν=n-1 degrees of freedom

Finding Critical Values

- Look at your textbook appendix
- Find the significance level α
- Calculate the number of degrees of freedom (n-1)
- Look up degrees of freedom and α in the chi-square table.

For a right-tailed test



Sample size is 30. What is the degree of freedom?

Significance level α is 0.05

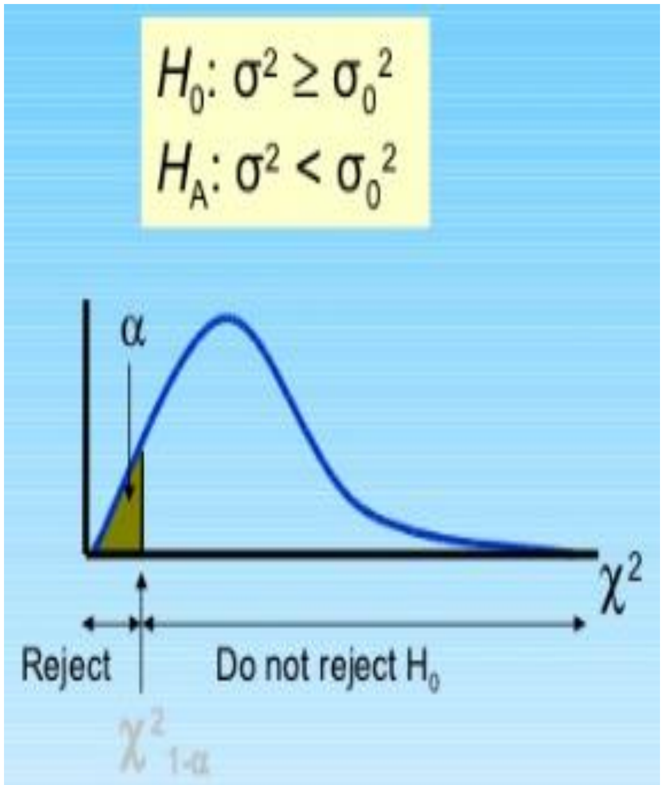
For a right-tailed test, find the column corresponding to α

What is the critical value?

Reject the null hypothesis if the test statistic is greater than the critical value.

```
> qchisq(0.05, df=29, lower.tail=FALSE)
[1] 42.55697
```

For a left-tailed test



Sample size is 25. What is the degree of freedom?

Significance level α is 0.05

For a left-tailed test, find the column corresponding to 1 - α

What is the critical value?

Reject the null hypothesis if the test statistic is less than the critical value.

```
> qchisq(0.05, df=24, lower.tail=TRUE)
[1] 13.84843
```

For a two-tailed test

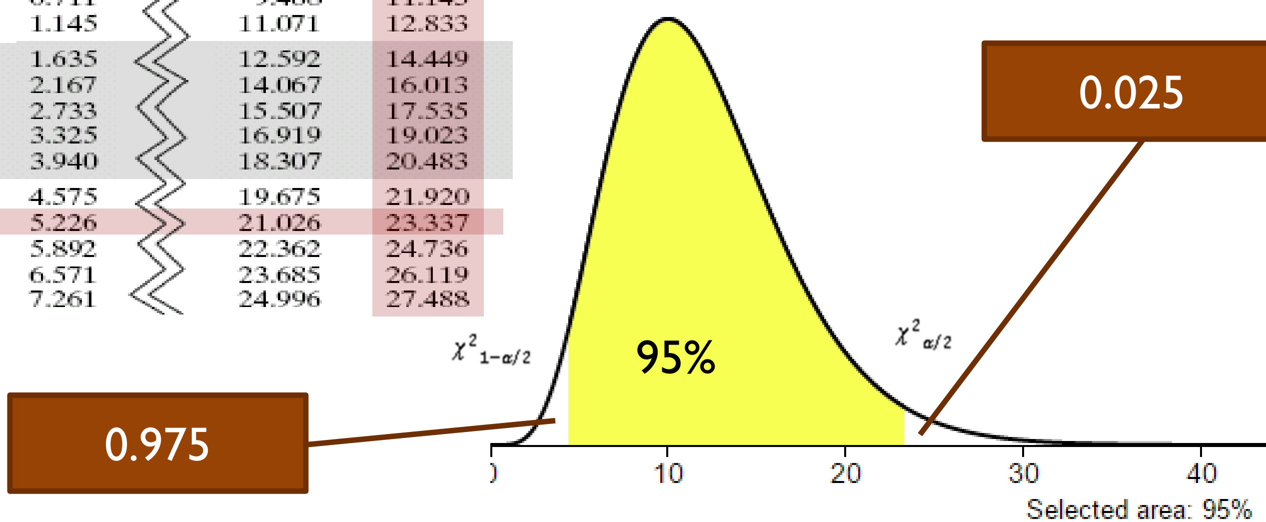
- Sample size = 13
- Significance level is 0.05
- 2.5% in each tail

```
> qchisq(0.025, df=12, lower.tail=TRUE)
[1] 4.403789
> qchisq(0.025, df=12, lower.tail=FALSE)
[1] 23.33666
```

Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of Freedom	0.975	0.95	0.05	0.025
1	0.001	0.004	3.841	5.024
2	0.051	0.103	5.991	7.378
3	0.216	0.352	7.815	9.348
4	0.484	0.711	9.488	11.143
5	0.831	1.145	11.071	12.833
6	1.237	1.635	12.592	14.449
7	1.690	2.167	14.067	16.013
8	2.180	2.733	15.507	17.535
9	2.700	3.325	16.919	19.023
10	3.247	3.940	18.307	20.483
11	3.816	4.575	19.675	21.920
12	4.404	5.226	21.026	23.337
13	5.009	5.892	22.362	24.736
14	5.629	6.571	23.685	26.119
15	6.262	7.261	24.996	27.488



	one-tailed test		two-tailed test
hypothesis	$H0\colon \sigma^2 \geq \sigma_0^2$ $H1\colon \sigma^2 < \sigma_0^2$	$H0\colon \sigma^2 \leq \sigma_0^2$ $H1\colon \sigma^2 > \sigma_0^2$	$H0\colon \sigma^2 = \sigma_0^2$ $H1\colon \sigma^2 \neq \sigma_0^2$
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 $\chi^2 > \chi^2_{\alpha/2}$

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?

```
> svar<-(12.5)^2
> pvar<-100
> n<-20
> chi<-(n-1)*svar/pvar
> chi
[1] 29.6875
> qchisq(0.025, df=19, lower.tail=TRUE)
[1] 8.906516
> qchisq(0.025, df=19, lower.tail=FALSE)
[1] 32.85233
```

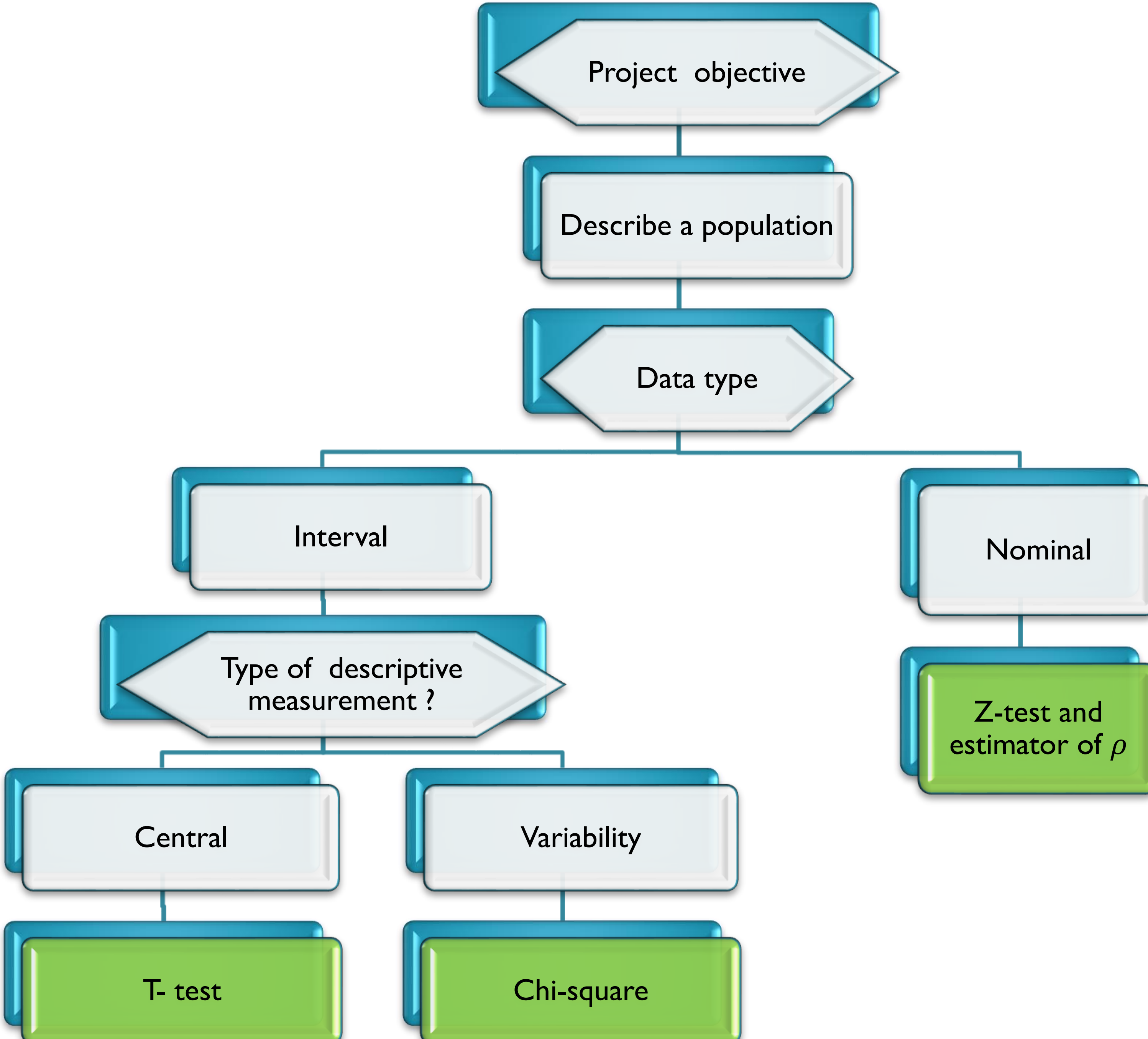
Problem

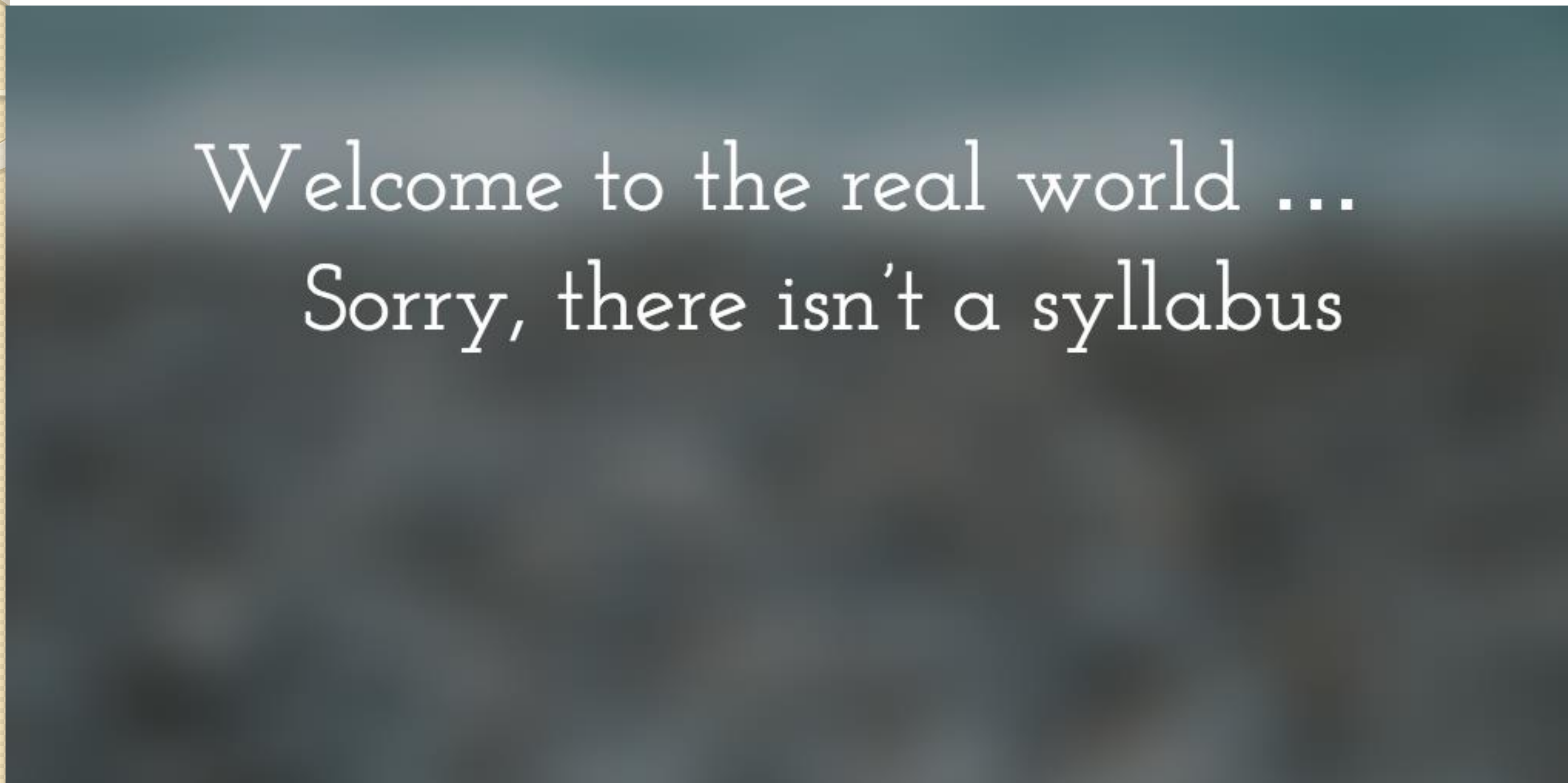
- You don't want too much variation from sport shoes to sport shoes. You assume that a population variance of no more than 0.05 inch is acceptable.

- 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?

```
> svar<-0.06
> pvar<-0.05
> n<-25
> chi<-(n-1)*svar/pvar
> chi
[1] 28.8
> qchisq(0.05, df=24, lower.tail=TRUE)
[1] 13.84843
> qchisq(0.05, df=24, lower.tail=FALSE)
[1] 36.41503
```


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


Welcome to the real world …
Sorry, there isn’t a syllabus

R practices in this section - 1

- Example 1-1 (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 12 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?

Your results should look like this

$H_0: \mu = 6$

$H_0: \mu < 6$

a Rejection region: $t < -t_{\alpha,n-1} = -t_{.05,11} = -1.796$

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5.69 - 6}{1.58/\sqrt{12}} = -.68$, p-value = .2554. There is not enough evidence to support the courier's advertisement.

```
mydata<- data.frame(Xr12_23)
View(mydata)
#> t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired =
FALSE, var.equal = FALSE, conf.level = 0.95)

t.test(mydata$Times, alternative="less", mu=6)
```

R practices in this section - 2

- Example 2-1 (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%?

$H_0 : p = .90$

$H_1 : p < .90$

$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{.8644 - .90}{\sqrt{.90(1-.90)/177}} = -1.58$, p-value = $P(Z < -1.58) = .0571$. There is not enough evidence to infer that the satisfaction rate is less than 90%.

```
mydata<- data.frame(Xr12_108)
View(mydata)
str(mydata)
mydata$fSatisfied<- as.factor(mydata$Satisfied)
table(mydata$fSatisfied)

#prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, correct = TRUE)

prop.test(153, 177, p = 0.9, alternative="less", conf.level = 0.95, correct = FALSE)
```

R practices in this section - 3

- Example 3-1 (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?

$H_0: \sigma^2 = 250$

$H_1: \sigma^2 < 250$

Rejection region: $\chi^2 < \chi^2_{1-\alpha,n-1} = \chi^2_{.90,9} = 4.17$

$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(10-1)(210.22)}{250} = 7.57$, p-value = .4218. There is not enough evidence to infer that the population variance has decreased.

```
mydata<- data.frame(Xr12_72)
View(mydata)
str(mydata)
install.packages("EnvStats")
require(EnvStats)
#varTest(x, alternative = "two.sided", conf.level = 0.95, sigma.squared = 1, data.name = NULL)
varTest(mydata$Marks, alternative = "less", conf.level = 0.90, sigma.squared = 250, data.name = NULL)
```

ARE YOU READY?

R practices - Exercise 1

- 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 in parentheses. 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?

$H_0: p = .50$

$H_1: p > .50$

$z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \dfrac{.57 - .50}{\sqrt{.50(1-.50)/100}} = 1.40$, p-value = $P(Z > 1.40) = 1 - .9192 = .0808$. There is not enough evidence to conclude that more than 50% of all business students would rate the book as excellent.

```
mydata<- data.frame(Xr12_112)
View(mydata)
str(mydata)
mydata$fTextbook<- as.factor(mydata$Textbook)
table(mydata$fTextbook)
#prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, correct = TRUE)
prop.test(57, 100, p = 0.5, alternative="greater", conf.level = 0.99, correct = FALSE)
```

R practices - Exercise 2

- 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 20 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?

$H_0 : \mu = 20$

$H_0 : \mu > 20$

Rejection region: $t > t_{\alpha,n-1} = t_{.05,19} = 1.729$

$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{20.85 - 20}{6.76 / \sqrt{20}} = .56$, p-value = .2902. There is not enough evidence to support the doctor's claim.

```
mydata<- data.frame(Xr12_25)
View(mydata)
#> t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired =
FALSE, var.equal = FALSE, conf.level = 0.95)
str(mydata)
t.test(mydata$Overweight, alternative="greater", mu=20)
```

R practices - Exercise 3

- Exercise 3 (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.

$H_0 : \sigma^2 = 18$

$H_1 : \sigma^2 > 18$

Rejection region: $\chi^2 > \chi^2_{\alpha,n-1} = \chi^2_{.10,244} = 272.704$ (from Excel)

$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(245-1)(22.56)}{18} = 305.81$; p-value = .0044. There is enough evidence to infer that the population variance is greater than 18.

```
mydata<- data.frame(Xr12_76)
View(mydata)
str(mydata)
install.packages("EnvStats")
require(EnvStats)
#varTest(x, alternative = "two.sided", conf.level = 0.95, sigma.squared = 1, data.name = NULL)
varTest(mydata$Speeds, alternative = "greater", conf.level = 0.90, sigma.squared = 18, data.name = NULL)
```

Where are we and where are we going ?

Populations and Samples

Continuous probability

- Business decision
  - Estimation
  - Hypothesis testing

MORE!!!