Contents

[Z Test – For known Variance 1](#_Toc521682112)

[T Test – For Unknown Variance 2](#_Toc521682113)

[One sample T test 4](#_Toc521682114)

[2 Sample T test /Independent T test 5](#_Toc521682115)

[Matched Pair T Test /Paired sample T test/Dependent sample T Test 6](#_Toc521682116)

[Nonparametric Methods 7](#_Toc521682117)

[Contingency Table 7](#_Toc521682118)

[Chi – square test 8](#_Toc521682119)

[ANOVA (analysis of variance) 8](#_Toc521682120)

[F – test 8](#_Toc521682121)

[Mean Square 9](#_Toc521682122)

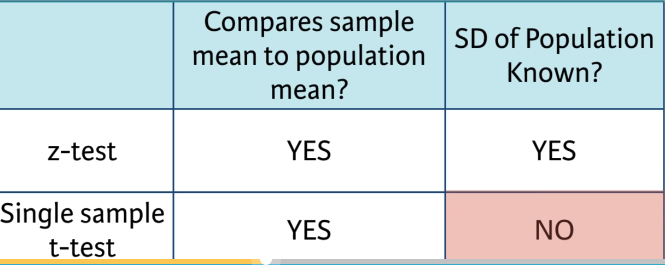
[Multivariate Testing 9](#_Toc521682123)

[Differences 9](#_Toc521682124)

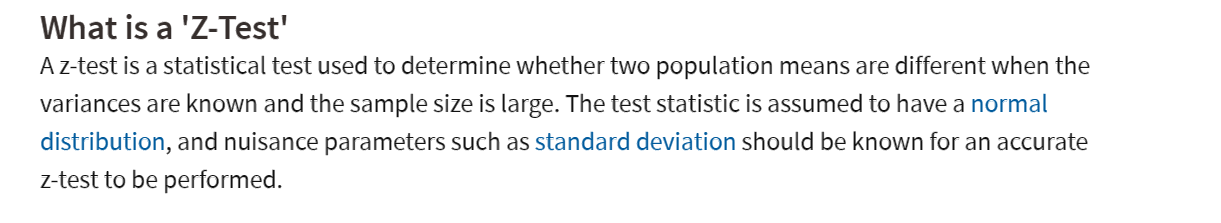
[Difference between chi square and F test 9](#_Toc521682125)

[Difference between F test and t test 10](#_Toc521682126)

Z test and T test are called significance test. Both are designed to compare means



# Z Test – For known Variance



* **1 sample z test**
* **2 sample z test**

Example

**Problem 4**: (15 pts) *NormalData.csv* contains random numbers generated from the normal distribution with variance 1. Test the null hypothesis *H*0: *μ* = 28.41.

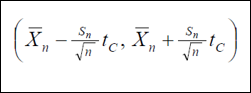
**Solution**: Since the data is normally distributed with known population variance we use the (*one-sample*) ***z-test***. It is a ***two-sided*** test by the problem statement, since no indication is given whether the population mean should be larger or smaller than 28.41. Apparently *R* has no standard *z*-test function, so we can write a custom one (done below, adopted

from *Session08+.R*) or import a function from some package.

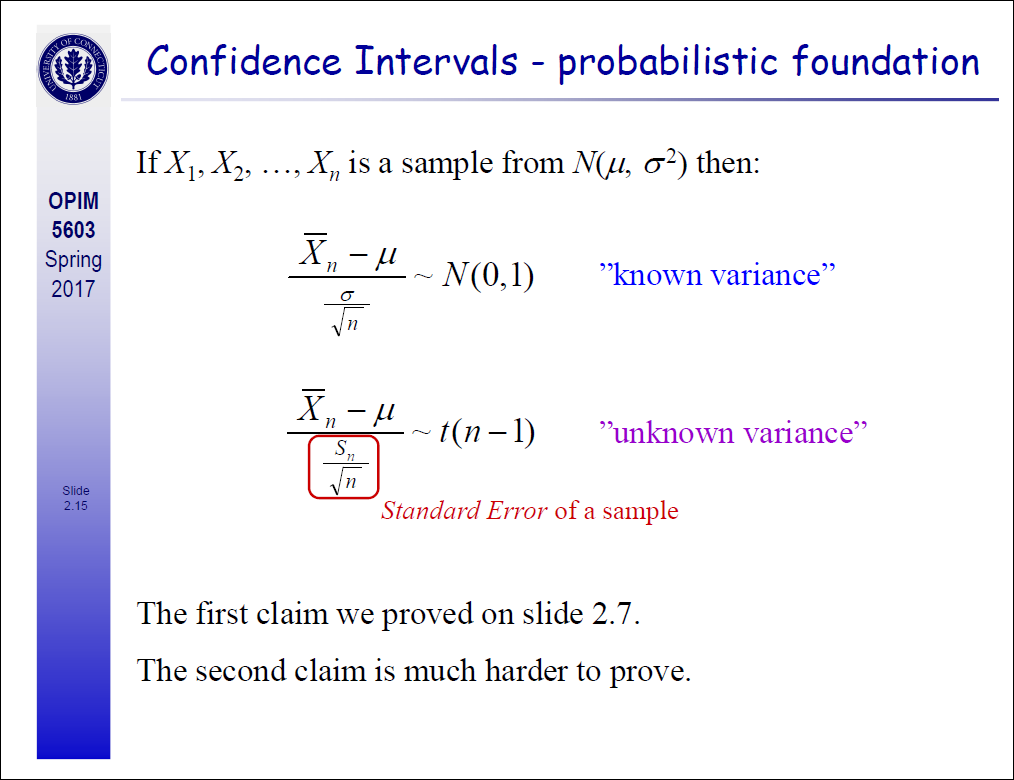
# T Test – For Unknown Variance

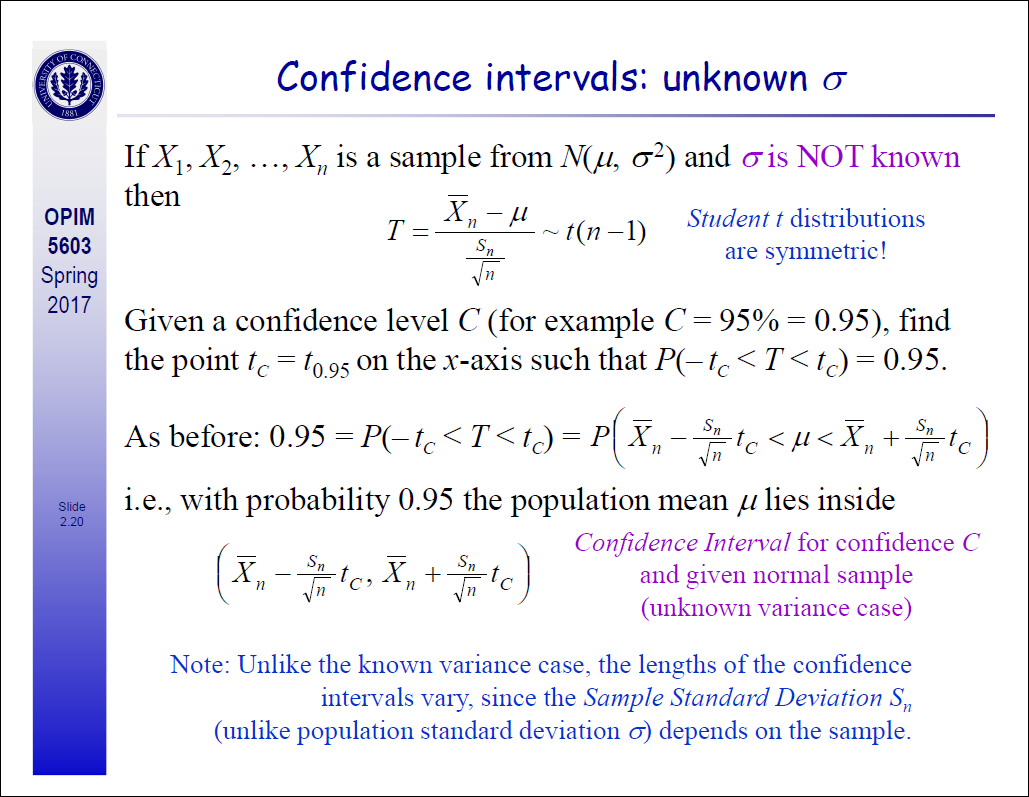
Confidence interval

(Xbar n – standard error \* t value , Xbar n + standard error \* t value )



Std error = std deviation / sqrt sample size





## One sample T test

**Problem 1**: (20 pts) A botanist is trying to determine whether the seeds treated with an experimental growth substance resulted in a higher average height of plants than the standard height of 305.3 mm. The botanist treated a random sample of 38 seeds with the extract and subsequently obtained the height data provided in the file *PlantHeight.csv*. What is his null hypothesis and should it be accepted at 5% significance level?

It is an example of t test – one sample- Upper tailed -

**Null Hypothesis:**

**H0 =>** Mu0 <= Mu(305.3)

**H1** => Mu0> Mu(305.3mm) (The mean height is larger than 305.3mm

**One sample upper tailed t test**

**R code:**

|  |
| --- |
| PlantHeight =read.csv("C:/Uconn msba/studies/stat/assignment/assignment 5/PlantHeight.csv")  View(PlantHeight)  mu0=305.33  result = ttest(PlantHeight$Height,mu0,0.05)  result    #upper tailed T test  ttest = function(v,mu0,alpha=0.05)  {  n = length(v)  tstat = (mean(v)-mu0)/(sd(v)/sqrt(n))  aux = pt(tstat,df=n-1)    return(1-aux,aux<alpha)  }    #OR we can do  t.test(PlantHeight$Height,alternative ="greater",mu=305.3,conf.level = 0.95) |

Test Conclusion:

Since p value is less than alpha , so we reject the null hypothesis that means mean height is less than or equal to 305.3mm

**Problem 3**: (20 pts) A manufacturer claims that the thickness of the steel plates it produces is 61 mils (thousandth of an inch). Car manufacturer’s quality control officer regularly checks the claim. From a recent shipment he took a random sample of 50 plates and measured their thickness. The data obtained is in the file *PlateThickness.csv*. In order to verify that the manufacturer’s claim is accurate, what should his null hypothesis be and should it be accepted at 10% significance level?

*t* test one sample two tailed

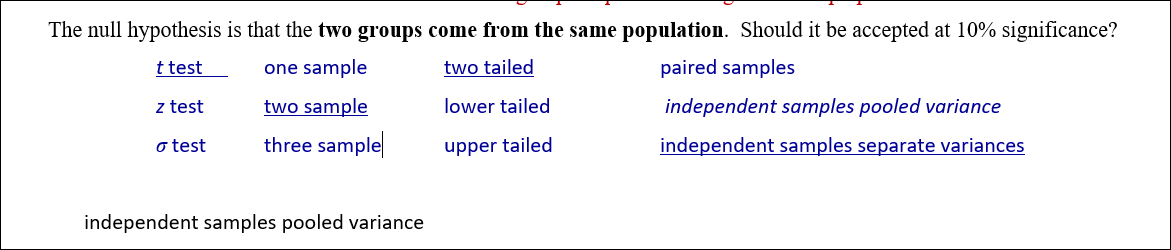
## 2 Sample T test /Independent T test

* 2 separate group of people
* E.g. Man vs Woman
* Drug vs Placebo

**example**

**Problem 2**: (20 pts) The addition of bran to the diet has been reported to benefit patients with diverticulosis. Several different bran preparations are available, and a clinician wants to test the efficacy of two of them on patients, since favorable claims have been made for each. Among the consequences of administering bran that requires testing is the transit time through the alimentary canal.

By random allocation the clinician selects two groups of patients aged 40-65 with diverticulosis of comparable severity. *SampleA.csv* contains 45 patients who are given treatment A, and *SampleB.csv* contains 62 patients who are given treatment B. The transit times of food through the gut are measured by a standard technique with marked pellets and the results are recorded. Does it differ in the two groups of patients taking these two preparations?



Solution:

Null hypothesis: Populations A and B have the same population means: μA = μB.

Alternative hypothesis: Populations A and B have different population means: μA ≠ μB.

We have two samples of different sizes so the independent samples (unpaired) two-sample (two-tailed) t-test should be used. The difference in sample sizes is large (62 vs. 45) so separate variances test (and not the pooled variance) should be used. We shall compute the sample variances and their ratio, but even without this calculation the use of separate variances test is more appropriate based on the large difference in sample sizes.

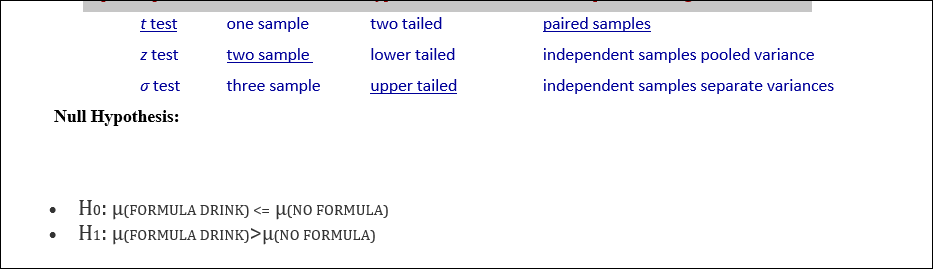
Null hypothesis: Populations A and B have the same population means: μA = μB.

Alternative hypothesis: Populations A and B have different population means: μA ≠ μB.

## Matched Pair T Test /Paired sample T test/Dependent sample T Test

* Difference between 2 related sample
* E.g. same population measured twice at different point of time
* Post Test score – Pre Test Score
* R code - t.test(immer$Y1, immer$Y2, paired=TRUE)

**Problem 5**: (20 pts) A company researcher wants to test a new formula for a sports drink that has been designed to improve running performance. To carry out the experiment, the researcher recruited 41 middle distance runners. All of these participants performed two trials (on different days) in which they had to run as far as possible for 2 hours on a treadmill. In one of the trials, all participants drank from a bottle containing the new formula. In the other trial, the same participants drank from a bottle containing no formula. At the end of the two trials, the distance each participant ran (in km) was recorded and stored in *Distance.csv*. The researcher wants to know whether the formula truly improves performance. What is the null hypothesis and should it be accepted at 5% significance level?



# [Nonparametric Methods](https://onlinecourses.science.psu.edu/stat414/node/291)

* + [Contingency Tables](https://onlinecourses.science.psu.edu/stat414/node/229)
  + [Chi-Square Goodness-of-Fit Tests](https://onlinecourses.science.psu.edu/stat414/node/228)
  + [Order Statistics](https://onlinecourses.science.psu.edu/stat414/node/230)
  + [Distribution-Free CIs for Percentiles](https://onlinecourses.science.psu.edu/stat414/node/231)
  + [The Wilcoxon Tests](https://onlinecourses.science.psu.edu/stat414/node/232)
  + [Run Test and Test for Randomness](https://onlinecourses.science.psu.edu/stat414/node/233)
  + [Kolmogorov-Smirnov Goodness-of-Fit Test](https://onlinecourses.science.psu.edu/stat414/node/234)

## Contingency Table

* Helps organize two or more categorical variables
* Used to study patterns that may exist between the responses of two or more categorical variables
* Cross tabulates or tallies jointly the responses of the categorical variables
* For two variables the tallies for one variable are located in the rows and the tallies for the second variable are located in the columns

## F – test

The F-statistic is simply a ratio of two variances. Variances are a measure of dispersion, or how far the data are scattered from the mean. Larger values represent greater dispersion.

**F = variation between sample means / variation within the samples**

Variance is the square of the standard deviation. For us humans, standard deviations are easier to understand than variances because they’re in the same units as the data rather than squared units. However, many analyses actually use variances in the calculations.

F-statistics are based on the ratio of mean squares. The term “[mean squares](http://support.minitab.com/minitab/17/topic-library/modeling-statistics/anova/anova-statistics/understanding-mean-squares/)” may sound confusing but it is simply an estimate of population variance that accounts for the [degrees of freedom (DF)](http://support.minitab.com/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/basic-concepts/df/)used to calculate that estimate.

For example We’ll analyze four samples of plastic to determine whether they have different mean strengths.

## Mean Square

Mean squares represent an estimate of population variance. It is calculated by dividing the corresponding sum of squares by the degrees of freedom.

# Differences

## Difference between chi square and F test

**Chi-Square test**

A chi-squared test is any statistical hypothesis test wherein the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. In simple way, we can say that any statistical test that uses the chi square distribution can be called chi square test.

**What is Chi-squared distribution?**

Chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables (Z).

χ2k=∑ki=1Z2i(1)(1)χk2=∑i=1kZi2

We can use Chi-square test for different purpose.

* **Chi-square test of single variance**: It is used to test a hypothesis on a specific value of the population variance.
  + **Example**: H0:σ2=15,H0:σ2=15, H1:H1:σ2σ2≠15≠15
* **Chi-square goodness of fit:**Chi square test for testing goodness of fit is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.
  + **Example**: We want to test the hypothesis that there is an equal probability of six sides of dice.
    - Expected(Theoretical) Value: (16,16,16,16,16,16)(16,16,16,16,16,16)
    - Observed (Experimental) Value: Throw the dice n times and find the frequency probability for each side of a dice.
* **Chi-square test of independence:** Chi square test for independence of two attributes. Suppose N observations are considered and classified according two characteristics say A and B. We may be interested to test whether the two characteristics are independent. In such a case, we can use Chi square test for independence of two attributes.
  + **Example:** H0:H0: In the population, the two categorical variables are independent., H1:H1:In the population, two categorical variables are dependent.

**F-test**

An F-test is any statistical test in which the sampling distribution of test statistic has an F-distribution when the null hypothesis is true. Similarly, any statistical test that uses the F distribution can be called F test.

We can use F-test for different purpose.

* **F-test of the equality of two variances:** It is used to compare the variances of two quantitative variables.
  + **Example**: H0:σ21=σ22,H1:σ2≠σ22H0:σ12=σ22,H1:σ2≠σ22
* **ANOVA**: We use ANOVA to compare more than two means. F-test in ANOVA is used to assess whether the expected values of a quantitative variable within several pre-defined groups differ from each other.
  + **Example**: Medical trial compares four treatments. The ANOVA F-test can be used to assess whether any of the treatments is on average superior, or inferior, to the others versus the null hypothesis that all four treatments yield the same mean response.
* **Regression problems**: We do t-test for individual coefficient significance in regression. We can use F-test for overall significance of the model . It helps to compare the fit of different linear model for same data.
  + Example: H0:β1=β2=β3………=βk=0,H1:H0:β1=β2=β3………=βk=0,H1:At least one coefficient is significant.

## Difference between F test and t test

T-test and F-test are completely two different things.

1. T-test is used to estimate population parameter, i.e. population mean, and is also used for hypothesis testing for population mean. Though, it can only be used when we are not aware of population standard deviation. If we know the population standard deviation, we will use Z-test.

For eg. Suppose a data suggests that the average height of boys between 10-16 years in city X is 6 Feet. So, we want to test this hypothesis,whether the height of boys between 10-16 years in city X is less than, more than, or equal to 6 Feet. For doing so, we will take some samples, say 2000, and find out the height of boys between age 10 to 16 years. We will calculate the standard deviation of the 2000 boys, and calculate the**t-statistic=**

X bar = sample mean

u= pop mean

S= sample standard deviation

n= sample size

df= n-1

Once we calculate t-statistic, we will compare it with the critical value. Suppose, we take **α**=.05, as it is two tailed test,**α/2= .025,**we then look at the table value for t, with **degrees of freedom** =n-1, 2000-1, and **α/2=0.25** which is=±1.96.

Once we get the t-value, we will compare whether our t-statistic is greater than +1.96 or less than -1.96. . If it is greater than +1.96 or less than -1.96, we reject the null hypothesis, which means, that the average height of boys in city X is not equal to 6 feet. If our t-statistic is between ±1.96, we fail to reject the null hypothesis, which means, that the average height of boys in city X is equal to 6 feet. This is when we conduct hypothesis testing.

2. We can also use t-statistic to estimate population mean:

Eg. Suppose,a large conglomerate like TCS(Indian IT company) which has employees more than 300,000. So, TCS wants to estimate average over time an employee works for the company, in a week. So, it might not be possible to get required data(hypothetical situation, though these days it might be possible) from all employees. Therefore, the company takes a sample, say 3000, and finds the number of extra hours of work employees have done in week. With the help of the sample mean and sample standard deviation; for the entire population- one can estimate the range of average number of extra hours of work, employees have done in week.

Confidence interval to find out the range= x̅±tα/2,n-1 \*S/√n **≤Ų≤** x̅±tα/2,n-1\* S/√n.

3. t-statistic is also used for finding out the difference in two population mean with the help of sample means.

For eg, suppose, if we want to understand buying behaviour of customers from two cities for a particular product. We want to understand whether there is any difference in buying behaviour in these cities,or is it similar.

**F-statistic**

Z statistic or t-statistic is used to estimate population parameters- population mean & proportion. It is also used for testing hypothesis for population mean or population proportion.

Unlike Z-statistic or t-statistic, where we deal with mean & proportion, Chi-square or F-test is used for finding out whether there is any variance within the samples. F-test is the ratio of variance of two samples.

Eg. Suppose, in a manufacturing plant there are 2 machines producing same products, and the management wants to understand, whether there is any variability among the products produced by these two machines. Researcher will take samples from both the machines and find out the variability, and test it against the null hypothesis, i.e. the prescribed limit.

F-statistic also forms the basis for ANNOVA.