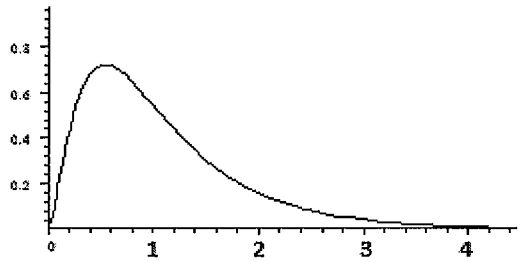
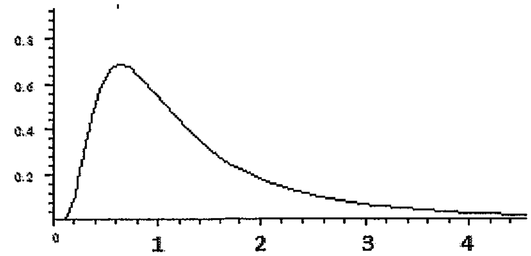
CHAPTER 13: F PROBABILITY DISTRIBUTION



$F_{5,20}$



$F_{20,5}$

- **continuous probability distribution**
- **skewed to the right**
- **variable values on horizontal axis are ≥ 0**
- **area under the curve represents probability**
- **horizontal asymptote – extends to infinity along positive horizontal axis**
  curve gets closer to horizontal axis but does not touch it as X gets large
- **The shape of the F distribution is determined by two values for "degrees of freedom".**
  The degrees of freedom are both written as subscripts.

  The theoretical mathematical formula for the F probability distribution is a ratio, so the two values for degrees of freedom are associated with the numerator and the denominator of the ratio.

  The "first" number for degrees of freedom is associated with the numerator;
  The "second" number for degrees of freedom is associated with the denominator.

  Notation F $_{\text{df for "numerator", df for "denominator"}}$

  F distribution with 5 and 20 degrees of freedom is written $\mathbf{F_{5,20}}$

  5 degrees of freedom for the numerator
  20 degrees of freedom for the denominator

  $F_{5,20}$ and $F_{20,5}$ are not the same because the values of degrees of freedom are not in the same order. Graphs at the top of the page show that they have somewhat different shapes – they are not identical.
- The mean is $\mu = d/(d-2)$ where d is number of degrees of freedom for the denominator.

  $F_{5,20}$ has mean $\mu=20/18 = 1.111$ and $F_{20,5}$ has mean $\mu=5/3 = 1.667$

TI-83+,84+: Finding a right tailed probability with the F distribution

2nd DISTR Fcdf(left boundary, right boundary, df numerator, df denominator)

Use 10^99 for the right boundary if finding a right tailed probability (area to the right)

On the graph of X ~ $F_{5,20}$ above, shade the area and find P(X > 2)

Fcdf(________ , ________ , ________ , ________) = ________

On the graph of X ~ $F_{20,5}$ above, shade the area and find P(X > 2)

Fcdf(________ , ________ , ________ , ________) = ________

We will use the F probability distribution to perform a hypothesis test called ANALYSIS OF VARIANCE which is often abbreviated as ANOVA

CHAPTER 13: One Way ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA) is a hypothesis test of whether the means for several populations are all equal to each other, or if there are differences between some of the means.

- Purpose is similar to a test of two population means (Chapter 10)
- Allows us to compare more than two population means at once, using several samples of data
- Analysis of Variance compares the variance between groups to the variance within groups.
- Comparison of variance uses a ratio (not as a difference).
- F distribution is used to compare the variance between groups to the variance within groups.

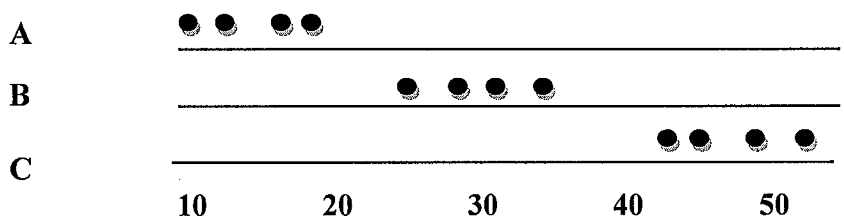
We will study "ONE WAY" ANALYSIS OF VARIANCE in Math 10.

EXAMPLE 1: Means are different.

Variation between groups is large compared to the variation within groups

Amounts of money spent by individual customers at restaurants A, B, C

It appears that the average amounts of money spent by customers at restaurants A, B, C are different

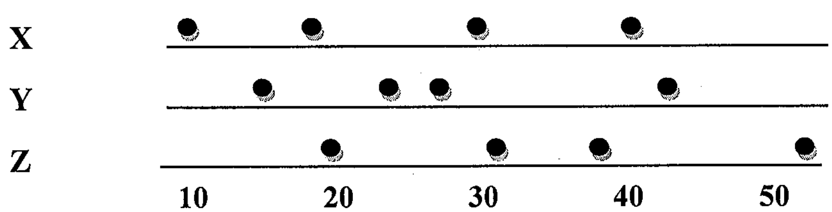


EXAMPLE 2: Means may all be the same

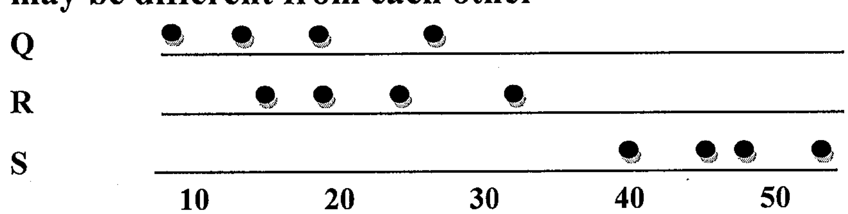
Variation between groups is not large compared to the variation within groups.

Amounts of money spent by individual customers at restaurants X, Y, Z

The sample data do not appear to give us reason to believe that the average amounts of money spent by customers at restaurants X, Y, Z are different. The averages may all be the same.



EXAMPLE 3: Some means may be the same as each other and some means may be different from each other



NULL HYPOTHESIS: **Ho: All the means are equal to each other**

$\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

means are being compared in k populations using k samples of data

ALTERNATE HYPOTHESIS: **H_A: Some of the means are different from each other**

must be written as a sentence – can NOT be written symbolically

EXAMPLE 4: ONE WAY ANALYSIS OF VARIANCE ANOVA

Does the average length of a song differ for songs of different genres or are the average song lengths the same for each genre? The sample data show the lengths of songs, in minutes, for random samples of Pop, Jazz, and Rock songs. Assume the song lengths for each genre are approximately normally distributed with equal standard deviations (equal variances). Perform a hypothesis test to determine if the average song length is the same for all three genres; use a 5% level of significance.

	Pop	Jazz	Rock
	3.6	4.6	3.8
	4.2	4.5	4.3
	3.7	4.8	4.3
	3.5	4.6	4.5
	3.1	4.5	4.8
	3.7	4.1	4.4
	3.9	5.2	4.2
Sample Mean	$\bar{x}_{Pop} = 3.67$	$\bar{x}_{Jazz} = 4.61$	$\bar{x}_{Rock} = 4.33$
Sample Size	$n_{Pop} = 7$	$n_{Jazz} = 7$	$n_{Rock} = 7$
Sample Std Deviation	$S_{Pop} = 0.340$	$S_{Jazz} = 0.334$	$s_{Rock} = 0.304$
Variance	$S_{Pop}^2 = 0.340^2 = .116$	$S_{Jazz}^2 = 0.334^2 = .112$	$s_{Rock}^2 = 0.304^2 = .092$

N = 21 songs
k = 3 groups

Average of all sample values:
$\bar{\bar{x}} = 4.21$

Ho: ____________________

Ha: ____________________

Analysis of Variance compares the variation between the sample means to the variation between the data points within each group. ANOVA measures variation by looking at **variance**.

Remember from chapter 2: Variance = (Standard Deviation)^2

Variance and Standard Deviation are calculated using the Sum of Squares.
SS stands for Sum of Squares. MS stands for Mean Square.
Mean Square = Sum of Squares/Degrees of Freedom

Variation between Groups: (also called Factor, Treatment)

Sum of Squares between groups: SSF = $7(3.67-4.21)^2 + 7(4.61-4.21)^2 + 7(4.33-4.21)^2 = 3.262$

Mean Square between groups: MSF = SSF/(k − 1) = 3.262/(3 −1)= 3.262/2 = **1.631**

Variation within Groups: (also called Error)

Sum of Squares within groups: SSE = $(7-1)(0.340)^2 + (7-1)(0.334)^2 + (7-1)(0.304)^2 = 1.92$

Mean Square within groups: MSE = SSE /(N − k) = 1.92/(21−3)= 1.92/18 = **0.107**

Note: Hand Calculations may vary slightly from results using calculator/computer due to rounding.

We compare whether the variation between groups is large compared to variation within groups by using a ratio instead of a difference:

Test Statistic F = MSF ÷ MSE = __________ / __________ =

Distribution to use for this test: ______________

Degrees of freedom for numerator = (number of groups) − 1 = k − 1
Degrees of freedom for denominator = (total number data values) − (number of groups) = N − k

pvalue = ________(________, ________, ____, ____) = ______

Decision: ______________ Reason for decision______________________

EXAMPLE 4: Performing ANOVA using the TI – 83 or 84 calculator

Does the average length of a song differ for songs of different genres or are the average song lengths the same for each genre? The sample data show the lengths of songs, in minutes, for random samples of Pop, Jazz, and Rock songs. Assume the song lengths for each genre are approximately normally distributed with equal standard deviations (equal variances). Perform a hypothesis test to determine if the average song length is the same for all three genres; use a 5% level of significance.

	Pop	Jazz	Rock
	3.6	4.6	3.8
	4.2	4.5	4.3
	3.7	4.8	4.3
	3.5	4.6	4.5
	3.1	4.5	4.8
	3.7	4.1	4.4
	3.9	5.2	4.2
Sample Mean	$\bar{x}_{Pop} = 3.67$	$\bar{x}_{Jazz} = 4.61$	$\bar{x}_{Rock} = 4.33$
Sample Size	$n_{Pop} = 7$	$n_{Jazz} = 7$	$n_{Rock} = 7$
Sample Std Deviation	$S_{Pop} = 0.340$	$S_{Jazz} = 0.334$	$s_{Rock} = 0.304$
Variance	$S_{Pop}^2 = 0.340^2 = .116$	$S_{Jazz}^2 = 0.334^2 = .112$	$s_{Rock}^2 = 0.304^2 = .092$

N = 21 songs
k = 3 groups

Average of all sample values: $\bar{\bar{x}} = 4.21$

Hypotheses: Ho: ______________________________

Ha: ______________________________

Calculations using TI 83+, 84+ ANOVA TEST: Put data into lists **L1, L2, L3**
STAT TESTS ANOVA (L1, L2, L3)

Test Statistic: ____ = ________ pvalue = ________ Distribution:____________

Draw shade and label a graph:

Decision: ______________ Reason for decision______________________

Conclusion:______________________________

____

____

Rewrite "ANOVA Table" from calculator (scrolls vertically) to paper (organized horizontally)

Analysis of Variance

Source	DF	SS	MS	F	p
Factor/Treatment (between groups)					
Error (within groups)					
Total					

One Step Further – Which means differ from each other?

If we reject the null hypothesis and decide that some of the means differ from each other, we want to know which means are different.
We should use statistical software to decide which means are different.
Statistical software compares means using a method called "Tukey multiple comparisons"

Why use ANOVA instead of doing a lot of two sample t tests?

1) It saves work – if there is no difference and all the means are the same, you are done with ONE test ANOVA, and don't have to investigate which pairs of means are different from each other.

2) Using several two-sample t-tests on pairs of groups is not correct. The tests need to use a "joint" significance or confidence level for all groups at once, not just two groups at a time. Doing a bunch of paired tests results in a higher significance level (less confidence) than doing all the tests at once.

3) Our TI-84 does not do the "Tukey multiple comparisons", so we can't tell which means are different. Using the TI-84, we could guess about which means are different from each other by using the two-sample t-tests, but it is not mathematically correct, and might sometimes give wrong results; using the two-sample t-tests would not have the correct significance level because it does not consider all the differences jointly.

EXAMPLE 4: One Step Further – Which means differ from each other?

MINITAB OUTPUT One-way ANOVA: Pop, Jazz, Rock

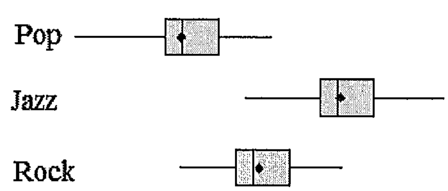
```
Null hypothesis          All means are equal
Alternative hypothesis   At least one mean is different
Significance level       α = 0.05
Equal variances were assumed for the analysis.

Factor   Levels   Values
Factor        3   Pop, Jazz, Rock
```

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	2	3.272	1.636	15.36	0.00013
Error	18	1.917	0.107		
Total	20	5.190			

C4	N	Mean	StDev	95% CI w/pooled StDev
Pop	7	3.671	0.340	(3.412, 3.931)
Jazz	7	4.614	0.334	(4.355, 4.873)
Rock	7	4.329	0.304	(4.069, 4.588)

Pooled StDev = 0.3264



Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping	
Pop	7	3.6714		B
Jazz	7	4.6143	A	
Rock	7	4.3286	A	

Means that do not share a letter are significantly different.

Conclusion:
The sample data do not show evidence that the average lengths of Rock and Jazz songs are different. Therefore we assume that the average lengths of Rock and Jazz songs are the same.
The averages lengths of Jazz and Pop songs differ from each other.
The averages lengths of Pop and Rock songs differ from each other.

CHAPTER 13: One Way ANALYSIS OF VARIANCE (ANOVA)

Assumptions needed to use ANOVA

Populations must be approximately normally distributed.
Distributions of populations must have equal population standard deviations (equal variances).

Checking assumptions:

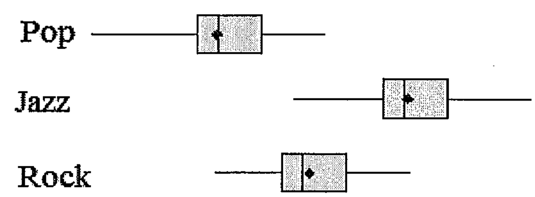
There are many ways to check whether sample data seem to come from populations that satisfy the above assumptions. One somewhat inexact but easy visual way to check if assumptions appear to be satisfied is to compare graphs, such as boxplots, of the samples:

- **The boxplots should have <u>approximately equal variance</u>** (we can visually examine spread by looking at both the range, which is max – min, and at the IQR represented by the box).
- Another rule of thumb is that the ratio of largest sample variance to smallest sample variance should be less than 4 (ratio of largest to smallest sample standard deviation should be less than 2).
- **The boxplots should be <u>approximately symmetric and should not be very skew.</u>**
- The data should be more concentrated toward center of distribution. The boxplots should not have a very long box with very short whiskers. (However, if the sample size is very small, short whiskers compared to the box may be acceptable and may not be an indicator of non-normality.)

EXAMPLE 4 Revisited:

Boxplots for data for song lengths:

Do these data appear to satisfy the assumptions needed for ANOVA?

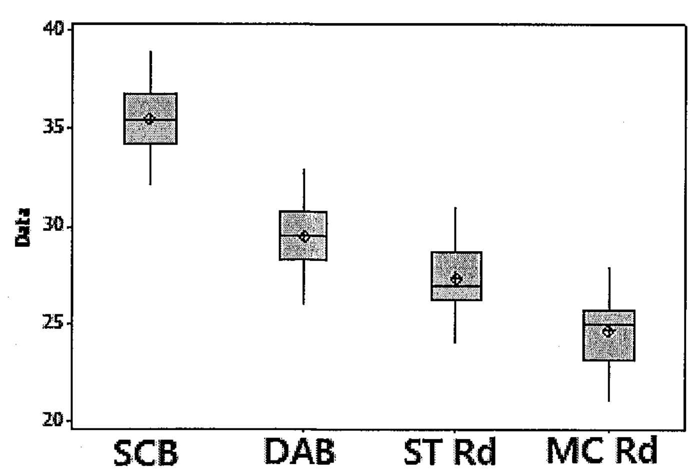


EXAMPLE 5: ONE WAY ANALYSIS OF VARIANCE ANOVA

We want to determine whether the true population average speeds of vehicles on four roads is the same, or whether the average speeds differ on some of the roads.
Vehicle speeds, in miles per hour, were recorded for a random sample of 6 vehicles on each road.

Stevens Creek Blvd	De Anza Blvd	Stelling Rd	McClellan Rd
36	30	27	25
35	26	28	25
38	29	31	27
32	33	24	21
35	29	27	24
36	30	27	25

Do these data appear to satisfy the assumptions needed for ANOVA?



EXAMPLE 5: ONE WAY ANALYSIS OF VARIANCE ANOVA

We want to determine whether the true population average speeds of vehicles on four roads is the same, or whether the average speeds differ on some of the roads.
Vehicle speeds, in miles per hour, were recorded for a random sample of 6 vehicles on each road. Assume the speeds of individual vehicles on each road are approximately normally distributed with equal standard deviations (equal variances). Use a 5% significance level.

Stevens Creek Blvd	De Anza Blvd	Stelling Rd	McClellan Rd
36	30	27	25
35	26	28	25
38	29	31	27
32	33	24	21
35	29	27	24
36	30	27	25

Hypotheses: Ho: ______________________________

Ha: ______________________________

Calculations using TI 83+, 84+ ANOVA TEST: Put data into lists **L1, L2, L3, L4**
STAT TESTS ANOVA (L1, L2, L3, L4)

Test Statistic: ____ = ________ pvalue = ________ Distribution:____________

Draw shade and label a graph:

Decision: ________________ Reason for decision______________________________

Conclusion:______________________________

____

____

____

Rewrite "ANOVA Table" from calculator (scrolls vertically) to paper (organized horizontally)

Analysis of Variance

Source	DF	SS	MS	F	p
Factor/Treatment (between groups)					
Error (within groups)					
Total					

EXAMPLE 5: ONE WAY ANALYSIS OF VARIANCE ANOVA

One-way ANOVA: Stevens Cr Blvd, De Anza Blvd, Stelling Rd, McClellan Rd

Method

Null hypothesis	All means are equal
Alternative hypothesis	At least one mean is different
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	4	Stevens Cr Blvd, De Anza Blvd, Stelling Rd, McClellan Rd

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	379.67	126.556	28.23	0.0000002
Error	20	89.67	4.483		
Total	23	469.33			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.11739	80.89%	78.03%	72.49%

Means

Factor	N	Mean	StDev	95% CI with pooled StDev
Stevens Cr Blvd	6	35.333	1.966	(33.530, 37.136)
De Anza Blvd	6	29.500	2.258	(27.697, 31.303)
Stelling Rd	6	27.333	2.251	(25.530, 29.136)
McClellan Rd	6	24.500	1.975	(22.697, 26.303)

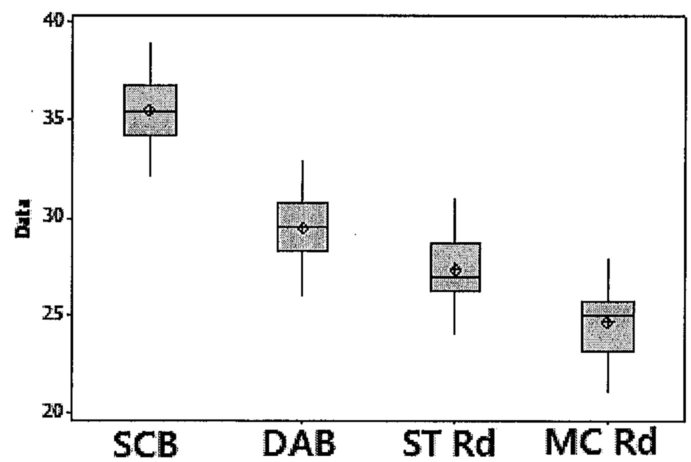
Pooled StDev = 2.11739

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
Stevens Cr Blvd	6	35.333	A
De Anza Blvd	6	29.500	B
Stelling Rd	6	27.333	B C
McClellan Rd	6	24.500	C

Means that do not share a letter are significantly different.



Write a conclusion:

Which PAIRS OF ROADS have means that are the same?
Which PAIRS OF ROADS have means that are different from each other?

ANOVA TABLE

As we saw in the previous examples, the output from ANOVA on your calculator scrolls down the screen vertically because of the limitations of the shape of the small calculator screen.

ANOVA tables are usually written horizontally. Since you may be reading journal articles or using statistical software in your future educational endeavors, you need to be familiar with the horizontal presentation of an ANOVA table.

Examples of ANOVA Tables created from statistical software appear earlier in these notes for Examples 4 and 5.

ANOVA PRACTICE PROBLEM #6:

The "Statistics Club at Hilltop College wonders whether the Chemistry, Math and Physics Departments all have the same mean class size.

The sample data show the numbers of students per class for samples of classes in these three departments at Hilltop College.

Are the average number class sizes the same for all these departments?

Chem	Math	Physics
L1	**L2**	**L3**
19	26	20
24	31	25
25	32	27
26	32	28
28	34	26
32	39	32
ANOVA(L1, L2, L3)		

One Way ANOVA
F = 4.67
P=.0266
Factor
df = 2
SS = 161.78
MS = 80.89
Error
df = 15
SS = 260
MS = 17.33
Sxp= 4.1633

Rewrite the TI-calculator output above into a standard horizontal ANOVA Table

Source	DF	SS	MS	F	P
Factor					
Error					
Total					

SUMMARY OF FORMULAS FOR ANOVA TABLES and TI-83 & 84 ANOVA OUTPUT:

Degrees of Freedom (df)

Factor (Between Groups): df = (number of groups) – 1
Error (Within Groups): df = (total number of data values) – (number of groups)
Total: df = (total number of data values) – 1

Sums of Squares (SS)

Total: use sum over all data values
SST = Σ (data value – overall mean)^2
Factor (Between Groups): use sum over all groups
SSF = Σ [(sample size for group) (group mean – overall mean)^2]
SSF may also be referred to as SSB or SSG for between groups
Error (Within Groups): SSE = SST – SSF
SSE may also be referred to as SSW for within Groups

Mean Square MS = SS/df
F = Test Statistic = MS Factor / MS Error = MS Between Groups / MS Within Groups
p = pvalue = Fcdf(FTestStatistic, 10^99, df Factor, df Error)
Sxp = square root of MS Error

To see these formulas in symbolic form, check the textbook and references for this class

Practice Problems for Analysis of Variance:

- For each practice problem, assume that the data come from approximately normally distributed distributions with approximately equal variances and standard deviations.
- For problems #7, #8 perform ANOVA and examine and use your calculator to compare the boxplots for the three samples in the problem.
- For #9, we can only compare the boxplots if using statistical software or drawing boxplots by hand – the calculator can only show up to 3 boxplots at one time on the screen.

ANOVA PRACTICE PROBLEM #7:

The Transit Commissioner wants to know if the average time between subway trains on 3 subway routes are different or the same.

The data in the table represent the time between trains, in minutes, for samples of size 8 on each route.

Do the data show evidence that there is a difference in the population average time between trains on these routes? Use a 5% significance level.

Train Route		
A	B	C
11	17	16
7	17	13
13	20	10
14	13	17
12	14	15
20	8	22
16	12	18
10	16	15

ANOVA PRACTICE PROBLEM #8:

Party Pizza specializes in meals for students. Hsieh Li, President, recently developed a new tofu pizza. Before making it a part of the regular menu she decides to test it in several of her restaurants. She would like to know if there is a difference in the mean number of tofu pizzas sold per day at the Cupertino, San Jose, and Santa Clara pizzerias.

At the 5% significance level, can Hsieh Li conclude that there is a difference in the mean number of tofu pizzas sold per day at the three pizzerias?

Cupertino	San Jose	Santa Clara
13	10	18
12	12	16
14	13	17
12	11	17
		17

- *Note the sample sizes at the three locations are not all equal. While all other examples and problems in these notes have equal sample sizes (balanced samples), in ANOVA it is OK to have samples with different sizes. The procedure as summarized on bottom of page 9 allows for unequal sample sizes.*
- *Tukey analysis further investigating differences between means is available at the source below.*

Source: Problem #8 is from **Inferential Statistics and Probability: A Holistic Approach,** by Maurice Geraghty, De Anza College, 1/1/2018, Rev 2/4/2019
http://professormo.com/holistic/HolisticStatisticsRev190204.pdf on 3/4/2019

ANOVA PRACTICE PROBLEM #9:

A statistics instructor wonders whether her average commute time varies by day of the week. She records her commute times, in minutes, for a random sample of 8 weeks.

The data are shown in the table.

Do the data show evidence that for the population of all commutes, the average commute times are differ by day of the week? Use a 5% significance level.

Mon	Tues	Wed	Thurs	Fri
27	32	29	31	26
30	35	32	33	28
31	36	33	35	29
32	37	35	36	30
33	38	36	37	31
34	39	37	38	32
35	40	38	39	33
35	40	38	39	33

NOTE: Each column for day of week is sorted in ascending order so data are not shown in the actual order that the data were collected by the week of occurrence.