CHAPTER 13: F PROBABILITY DISTRIBUTION

$F_{5,20}$

$F_{20,5}$
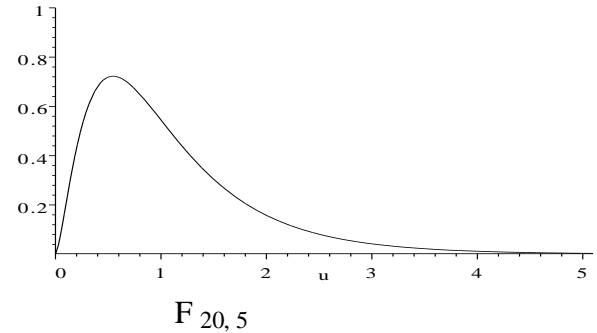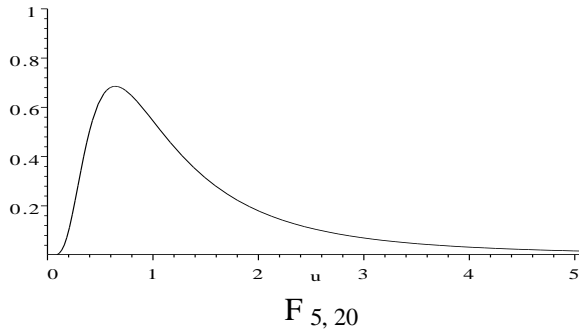
- **continuous probability distribution**
- **skewed to the right**
- **variable values on horizontal axis are ≥ 0**
- **area under the curve represents probability**
- **horizontal asymptote – extends to infinity along positive horizontal axis**
    curve gets closer to horizontal axis but does not 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.

    F distribution with 5 and 20 degrees of freedom is written $F_{5,20}$
    The "first" number for degrees of freedom is associated with the numerator;
    The "second" number for degrees of freedom is associated with the denominator.
    The mathematical formula for the F probability distribution is a ratio that involves two
    chi square distributions, so the two values for degrees of freedom are associated with
    the numerator and the denominator of the ratio.

    $F_{5,20}$ has: 5 degrees of freedom for the _____

    20 degrees of freedom for the _____.

    $F_{5,20}$ 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 $X \sim F_{5,20}$ above, shade the area and find $P(X > 2)$

        Fcdf(_____ , _____, _____, _____) = _____

On the graph of $X \sim F_{20,5}$ 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
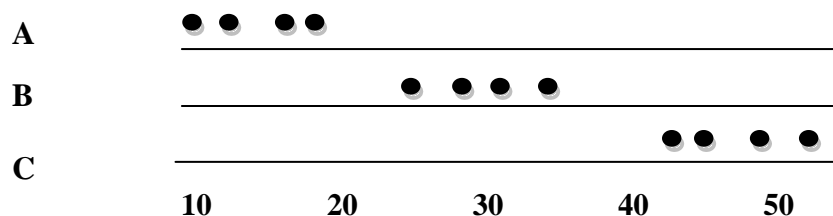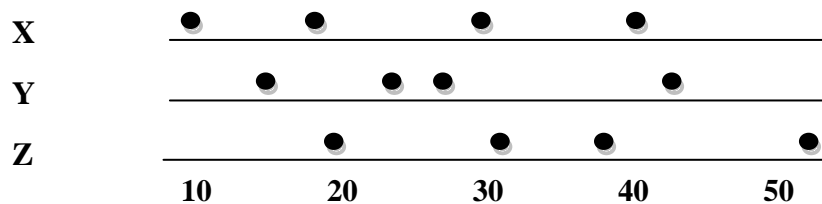

EXAMPLE 2: Means may all be the same
Variation between groups is not large compared to the variation within groups.

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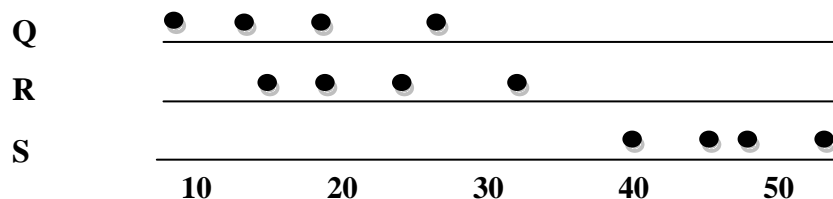
NULL HYPOTHESIS: **Ho: All the means are equal to each other**

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

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

Seeds for the same type of plant are grown in the same soil using three different types of fertilizer (Fertilizers Q, R, S). Are average plant heights the same for the three different types of fertilizer? Below are height data for samples of 7 plants grown for four weeks with each type of fertilizer.

Assume the plant heights for each fertilizer are approximately normally distributed with equal standard deviations (equal variances).

Q (Quick-Grow)	R (Rapid-Grow)	S (Super-Grow)	
43	41	41	N = 21 plants
49	40	46	k = 3 groups
44	42	46	
43	41	47	Average of all
39	40	50	sample values:
44	37	47	
46	45	45	

$\bar{x}_Q = 44$ $\quad\bar{x}_R = 40.9$ $\quad\bar{x}_S = 46$ $\qquad \bar{\bar{x}} = 43.6$

$n_Q = 7$ $\qquad n_R = 7$ $\qquad n_S = 7$

$s_Q = 3.06$ $\qquad S_R = 2.41$ $\qquad s_S = 2.71$ $\qquad$ Std Dev

$s_Q^2 = 3.06^2 = 9.36$ $\quad s_R^2 = 2.41^2 = 5.81$ $\quad s_S^2 = 2.71^2 = 7.34$ $\quad$ Variance

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.

Variation between Groups: $\quad$ SSbetween = $7(44 - 43.6)^2 + 7(40.9 - 43.6)^2 + 7(46 - 43.6)^2 = 94.1$
  (also called Factor, Treatment) $\quad$ MSbetween = SSbetween/(k − 1) = 94.1/(3 −1)= 94.1/2 = 47.05

Variation within Groups: $\quad$ SS within = $(7 − 1)(3.06)^2 + (7 − 1)(3.41)^2 + (7 − 1)(2.71)^2 = 134.86$
  (also called Error) $\quad$ MSwithin = SSwithin /(N − k) = 134.86/(21−3)= 134.86/18 = 7.49

Test Statistic F = MSbetween ÷ MSwithin = _____ / _____ =

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 = _____(_____, _____, ____, ____) = _____

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

Decision: _____ $\quad$ Reason for decision_____

Conclusion:_____

_____

CHAPTER 13: One Way ANALYSIS OF VARIANCE (ANOVA)

If we reject the null hypothesis and ecide that some of the means differ from each other we should use statistical software to decide which means are different. Statistical software provides comparisons of the 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 two sample t-tests on pairs of groups are not the correct tests to use. The tests need to take into account a "joint" significance or confidence level for all the groups at once, not just two groups at a time. Statistical software does both the hypothesis test and the analysis of which means are as part of the ANOVA analysis.

3) Our TI-84 does not do the "Tukey multiple comparisons", so we can't tell which means are different. Using the TI-84 instead of statistical software, we could guess about which means are different from each other by using the two sample t-tests, but it is not mathematically correct to do that, and it 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: Q., R., S.
```
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       3  Q., R., S.
```

Analysis of Variance
```
Source  DF  Adj SS  Adj MS  F-Value  P-Value
Factor   2   94.10  47.048     6.28    0.009
Error   18  134.86   7.492
Total   20  228.95

Factor  N    Mean  StDev       95% CI
Q.      7   44.00   3.06  ( 41.83,  46.17)
R.      7   40.86   2.41  ( 38.68 , 43.03 )
S.      7   46.00   2.71  ( 43.83,  48.17)
Pooled StDev = 2.73716
```

Tukey Pairwise Comparisons
Grouping Information Using the Tukey Method and 95% Confidence
```
Factor  N     Mean  Grouping
Q.      7    44.00    A  B
R.      7    40.86       B
S.      7    46.00    A
Means that do not share a letter are significantly different.
```

Fertilizers Q and R do not have different means. (Q and R share a letter   B in the Minitab output)
Fertilizers Q and S do not have different means. (Q and S share a letter   A in the Minitab output)
Fertilizers R and S have different means. (R and S do not share a letter in the Minitab output table for the Tukey multiple comparisons)

Conclusion: We think the mean heights for plants grown with fertilizers R and S are not equal.

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 the 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:

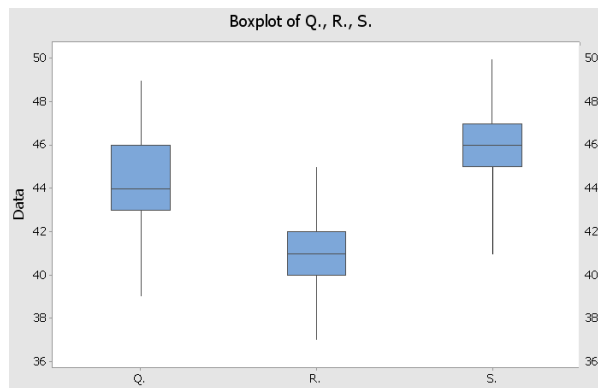 **The boxplots should be <u>approximately symmetric and should not be very skew.</u>**

 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)

 The boxplots should not have a very long box with very short whiskers; however, if the sample size is very small, this may be acceptable and may not be an indicator of non-normality.

ONE WAY ANALYSIS OF VARIANCE ANOVA EXAMPLE 3:

Investigated whether average plant heights are the same for the three different types of fertilizer.

Do the assumptions appear to be satisfied so that it is valid to use ANOVA?

Boxplot of Q., R., S.

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 speed, in miles per hour, was 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).

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 the assumptions appear to be satisfied so that it is valid to use ANOVA?

Boxplot of SCB, DAB, ST, MC

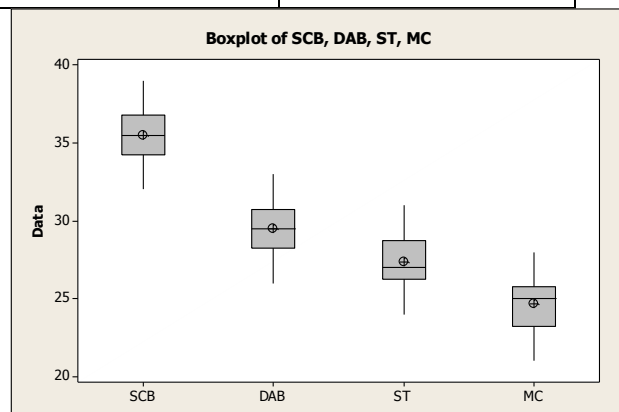## EXAMPLE 5: ONE WAY ANALYSIS OF VARIANCE ANOVA

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 speed, in miles per hour, was 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).

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:

Distribution to use for this test:

Calculator Output:

Decision:

Conclusion:

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    0.000
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
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
```
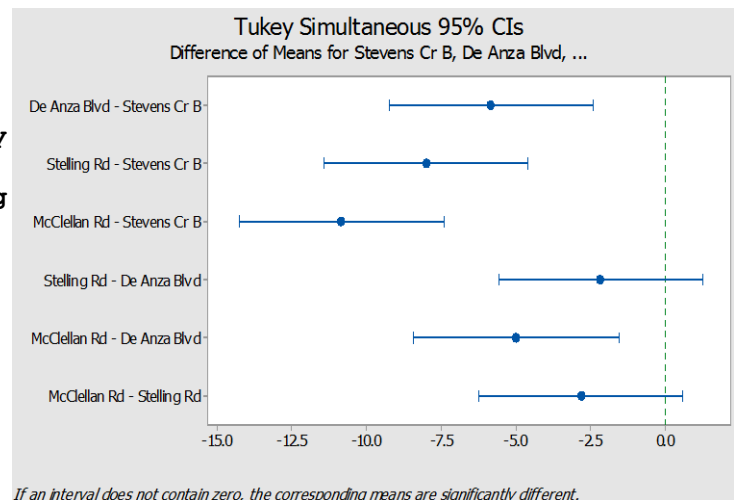
Tukey Pairwise Comparisons

```
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.
```


Tukey Simultaneous 95% CIs
Difference of Means for Stevens Cr B, De Anza Blvd, ...

If an interval does not contain zero, the corresponding means 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

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.

Example of ANOVA TABLE - created from statistical software

```
Source  DF      SS      MS        F        P
Factor   2    94.10   47.048    6.28     0.009
Error   18   134.86    7.492
Total   20   228.95
```

EXAMPLE 6:

The sample data show the numbers of customers visiting the pharmacy per hour at three different locations L1, L2, L3 of a chain pharmacy store.

At each store during one week, the number of customers was recorded for 6 randomly selected hours.

Are the average number of customers per hour the same at all locations?

L1	L2	L3	One Way ANOVA
14	21	16	F = 4
18	25	20	P=.0405
20	27	22	Factor
21	28	23	df = 2
23	30	25	SS = 156
27	34	29	MS = 78
ANOVA(L1, L2, L3)			Error
			df = 15
			SS = 292.5
			MS = 19.5
			Sxp= 4.41588

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
        $SST = \Sigma \,(\text{data value} - \text{overall mean})^2$
    Factor (Between Groups): use sum over all groups
        $SSF = \Sigma \,[(\text{sample size for group})\,(\text{group mean} - \text{overall mean})^2]$
        SSF may also be referred to as SSB or SSG for between groups or SST for treatments
    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

Practice Problems for Analysis of Variance:

ANOVA PRACTICE PROBLEM #7:

The management of a car rental agency wants to know if the average fuel efficiency for cars rented from its various rental agency locations are different or the same.

The data in the table represent the fuel efficiency in miles per gallons for samples of 8 rental cars at a car rental agency at several locations.

Do the data show evidence that there is a difference in the population average fuel efficiency for car rentals at these agency locations?

Rental Agency Location		
A	B	C
17	28	22
14	27	20
22	34	17
23	20	27
21	21	25
31	14	36
26	20	30
18	27	23

ANOVA PRACTICE PROBLEM #8:

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?

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

NOTES: each column is sorted in ascending order so data are not shown in order of the week of occurrence
        Assume that the data come from approximately normally distributed distributions with approximately equal variances and standard deviations.

ANOVA PRACTICE PROBLEM #9:

   For practice problem #8, write the ANOVA table in the standard horizontal form.

ANOVA PRACTICE PROBLEM #10:

The data below represent samples of weights of three types of small tomatoes; 8 tomatoes of each type A, B, C were selected and weighed.

Are the population average weights the same for these three types of tomatoes?

A	B	C
34.4	26.2	40.8
30.8	24	39.7
40.4	20.7	47.4
41.6	31.7	32
39.2	29.5	33.1
51.2	41.6	25.4
45.2	35	32
35.6	27.3	39.7

   Assume that the data come from approximately normally distributed distributions with approximately equal variances and standard deviations.