Analysis of Variance
ANOVA

ECE 341: Random Processes

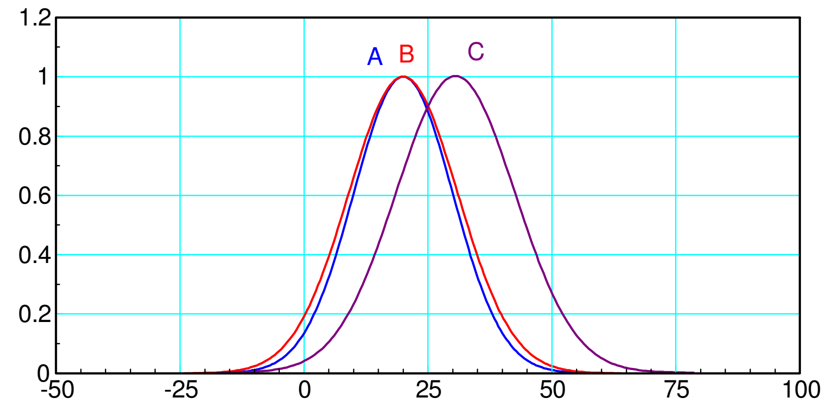
Lecture #32

note: All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

ANOVA

A second use of F distributions it to compare the means of 3+ populations.

- Termed Analysis of Variance (ANOVA)



ANOVA tests the hypothesis:

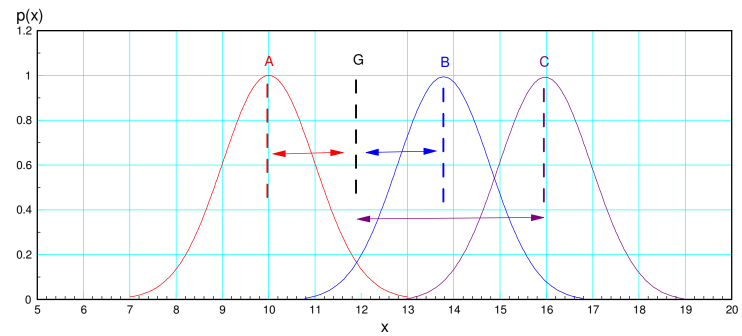
- H0:All populations have the same mean
- H1: At least one population's mean is different

This results in an F-test

ANOVA Idea

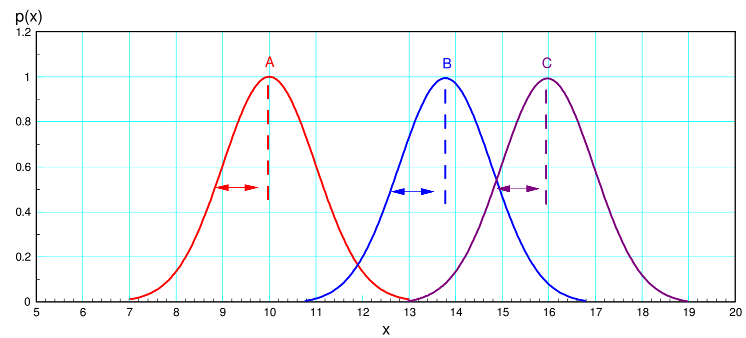
The basic idea is this:

- Assume you have samples from three populations with unknown means and variances
  - Each sample will have a mean and a variance
  - The entire data set will have a mean and a variance



Compute two variances:

- MSSb: Mean sum squared distance of the data to the global mean
- MSSw: Mean sum squared distance to each data set's mean



The ration is the F-value

$$F = \frac{MSS_b}{MSS_w}$$

A large F-value indicates the means are different

ANOVA Equations:

Define

k	the number of data sets (assume k = 3 here)
a_i, b_i, c_i	samples from data sets A, B, and C
$\overline{A}, \overline{B}, \overline{C}$,	the means of each data set
n_a, n_b, n_c	the number of data points in each data set
s_a^2, s_b^2, s_c^2	the variance of each data set
$N = n_a + n_b + n_c$	the total number of data points
$\overline{G}$	the global average (average of all data points)
s_g^2	the global variance

ANOVA Calculations: (non-standard)

MSSb: Mean Sum Squared Distance Between Columns

G is the average of all of the data

$$\overline{G} = \frac{1}{N}\left(\sum a_i + \sum b_i + \sum c_i\right)$$

MSSb is the variance of the entire data set

$$MSS_b = \left(\frac{1}{N-1}\right)\left(\sum \left(a_i - \overline{G}\right)^2 + \sum \left(b_i - \overline{G}\right)^2 + \sum \left(c_i - \overline{G}\right)^2\right)$$

or equivalently

$$MSS_b = \left(\frac{1}{N-1}\right)\left((n_a - 1)s_a^2 + (n_b - 1)s_b^2 + (n_c - 1)s_c^2\right)$$

MSSb has N-1 degrees of freedom

- N data points,
- Minus one computed mean (G)

ANOVA Calculations (cont'd)

MSSw: Mean Sum Squared Distance Within Columns

MSSw is the variance of the entire data set relative to their respective means

$$MSS_w = \left(\frac{1}{N-k}\right)\left(\sum\left(a_i - \overline{A}\right)^2 + \sum\left(b_i - \overline{B}\right)^2 + \sum\left(c_i - \overline{C}\right)^2\right)$$

MSSw has N-k degrees of freedom

- N data points
- minus k computed means (A, B, C)

F is the ratio of the variances

$$F = \frac{MSS_b}{MSS_w}$$

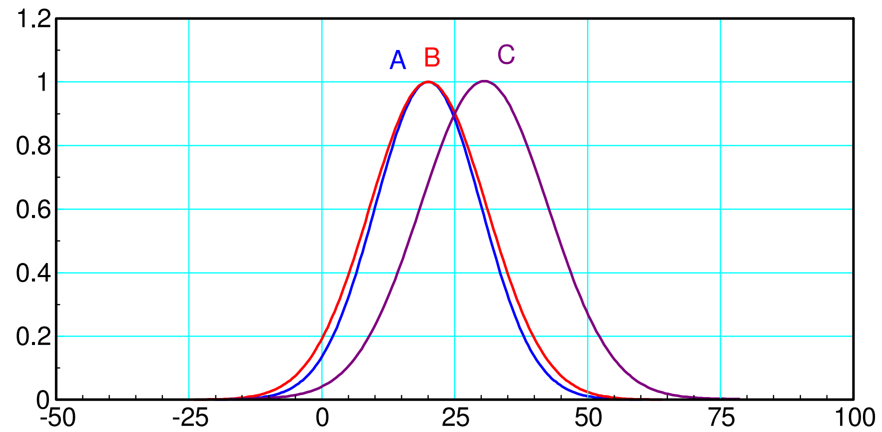
ANOVA Example #1

To illustrate this, consider three populations

$$A \sim N(20, 10^2)$$

$$B \sim N(20, 11^2)$$

$$C \sim N\left(\overline{C}, 12^2\right)$$



Let C's mean vary from 20 to 50

Can you detect that C's mean is different

- With sample sizes of 20 for A, B, and C?
- With sample sizes of 100?

Example 1: Matlab Code

Use a Monte-Carlo simulation

- Generate random numbers for A, B, C
- Compute MSSb and MSSw
- Compute the resulting F-value
- Repat 100,000 times

```
for i=1:1e5
    A = 10*randn(20,1) + 20;
    B = 11*randn(20,1) + 20;
    C = 12*randn(20
,1) + 50;
    Na = length(A);
    Nb = length(B);
    Nc = length(C);
    N = Na + Nb + Nc;
    k = 3;
    G = mean([A; B; C]);
    MSSb = var([A; B; C]);
    MSSw = 1/(N-k) * ((Na-1)*var(A) +
(Nb-1)*var(B) + (Nc-1)*var(C));
    F = MSSb / MSSw;
    n = round(F/dx);
    n = max(1, n);
    n = min(length(y),n);
    y(n) = y(n) + 1;
end
```

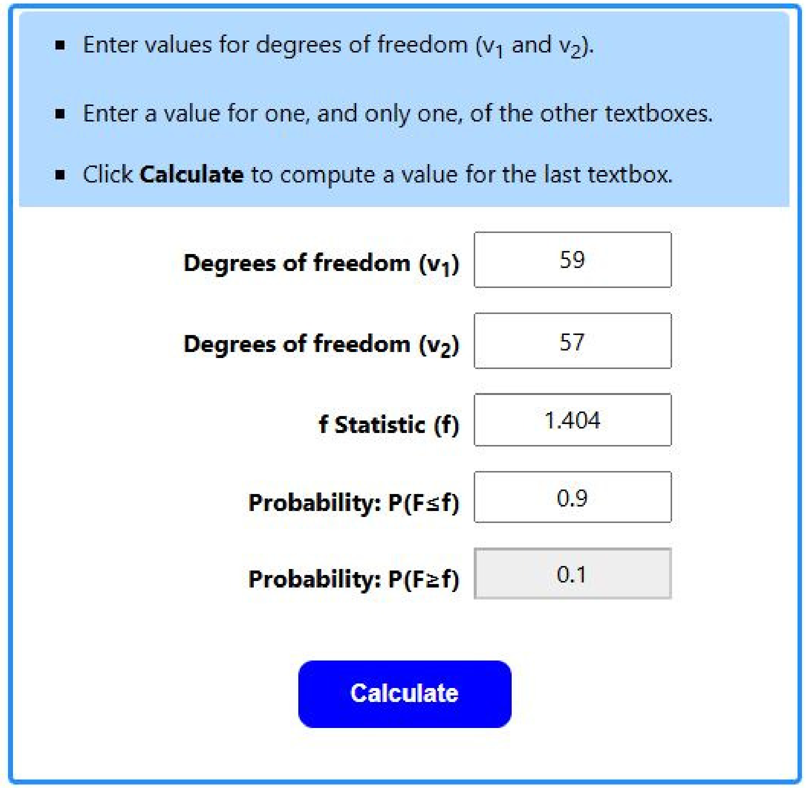
F Critical Values

If you collect 20 samples from each population, the F-value you're looking for is either

- F > 1.404 for p = 90%
- F > 2.818 for p = 99%

This can be found using StatTrek with

- 59 degrees of freedom in the numerator (N-1 = 59)
- 57 degrees of freedom in the denominator (N-k = 57)



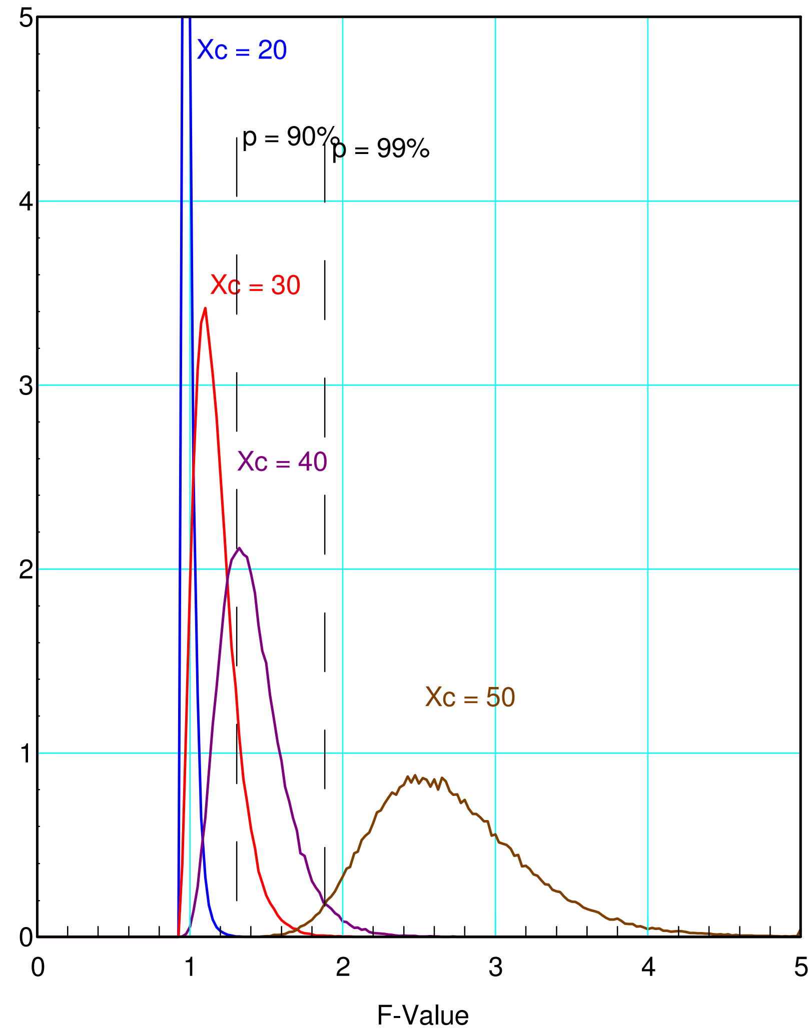
Resulting pdf for F-values

- Sample Size = 20
- Graph to the right

Interpriting the results:

- At 90% certainty, you can usually detect a difference in the means when population C's mean is 2x the means of populations A and B
- At 99% certainty, you can almost always detect a difference in the means when C's means is 2.5x larger

These results change if the variance of {A, B, C} change



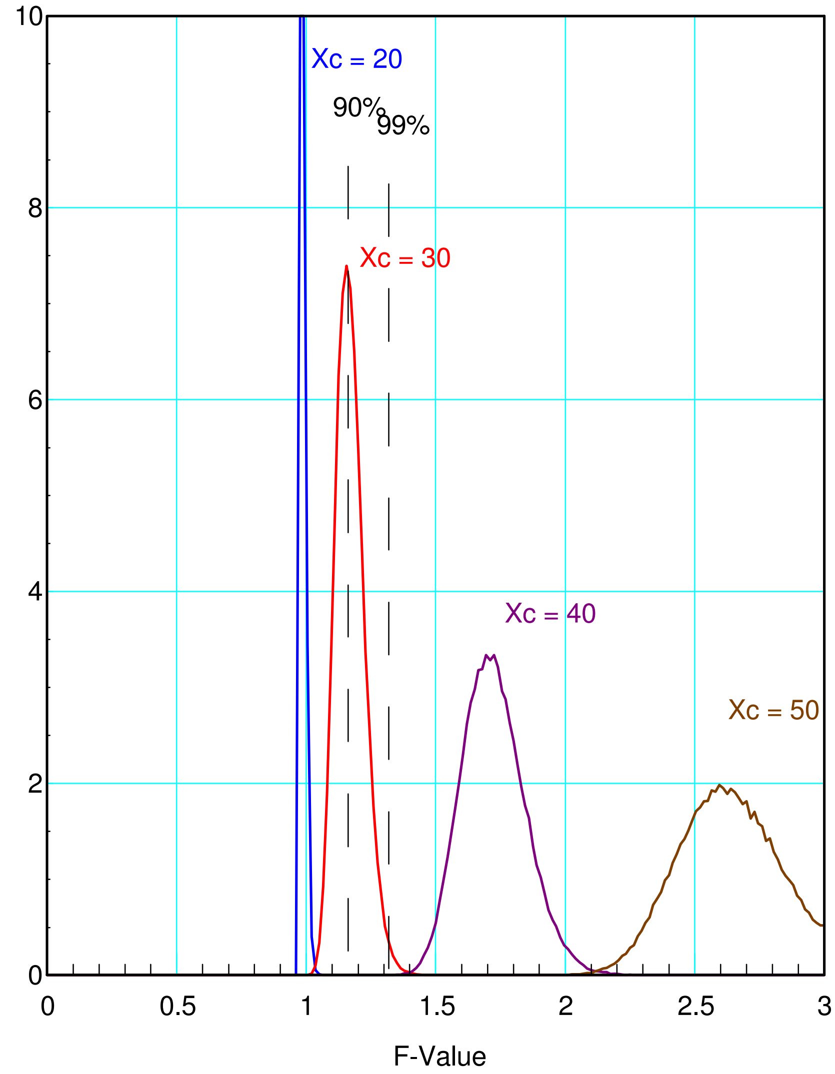
Resulting pdf for F-values

- Sample Size = 100
- Graph to the right

Interpriting the results:

- You can usually detect a difference of 10 in the means with 90% certainty,
- You can almost always detect a difference in means of 20 (Xc = 40) with 99% certainty

You can detect smaller differences in the means with more data



ANOVA Equations: Variation #2 (Standard Method)

While the previous way of computing MSSb and MSSw is the *correct* way (in my opinion), it's not the standard way of computing them.

The standard way only uses each population's

- Mean,
- Variance, and
- Sample size.

This is sometimes good

- Sometimes you don't have access to the raw data
- You can still proceed in this case

MSSb ('correct' way)

The previous way to compute MSSb was

$$MSS_b = \left(\frac{1}{N-1}\right)\left(\sum\left(a_i - \overline{G}\right)^2 + \sum\left(b_i - \overline{G}\right)^2 + \sum\left(c_i - \overline{G}\right)^2\right)$$

$$dof = N - 1$$

The correct way is

$$MSS_b \approx \left(\frac{1}{k-1}\right)\left(n_a\left(\overline{A} - \overline{G}\right)^2 + n_b\left(\overline{B} - \overline{G}\right)^2 + n_c\left(\overline{C} - \overline{G}\right)^2\right)$$

$$dof = k - 1$$

This assumes the variance of {A, B, C} is small so that

$$\sum\left(a_i - \overline{G}\right)^2 \approx n_a\left(\overline{A} - \overline{G}\right)^2$$

MSSw remains unchanged

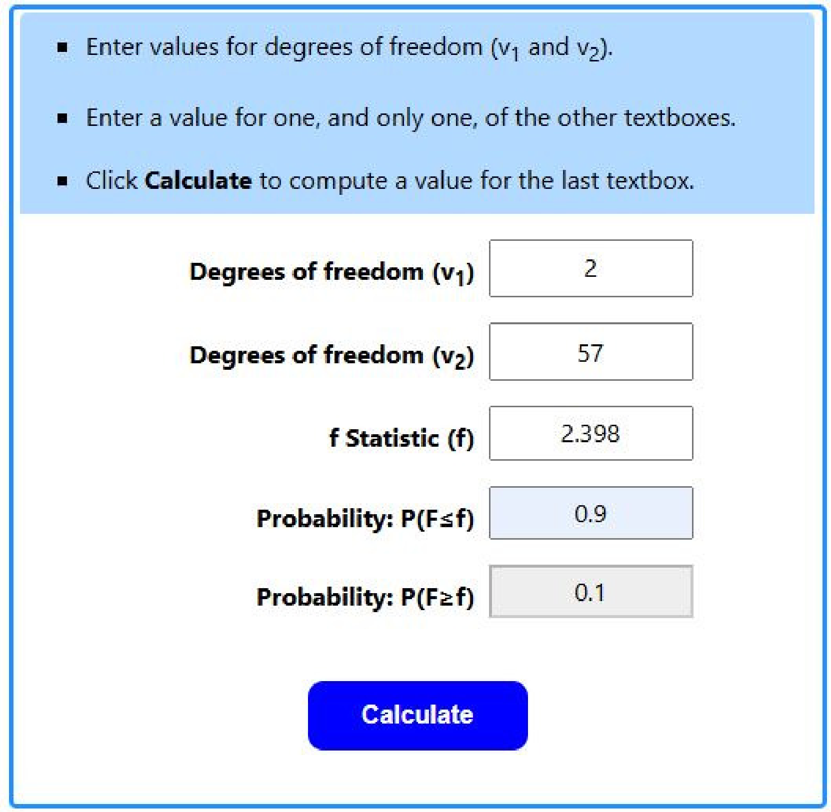
F Critical Values

If you collect 20 samples from each population, the F-value you're looking for is either

- F > 2.398 for p = 90%
- F > 4.983 for p = 99%

This can be found using StatTrek with

- 2 degrees of freedom in the numerator (k-1 = 2)
- 57 degrees of freedom in the denominator (N-k = 57)



ANOVA Example #2

Use a Monte-Carlo simulation

- Generate random numbers for A, B, C
- Compute population's mean, variance, and sample size
  - (only data used from this point onwards)
- Compute MSSb and MSSw
- Compute the resulting F-value
- Repeat 100,000 times

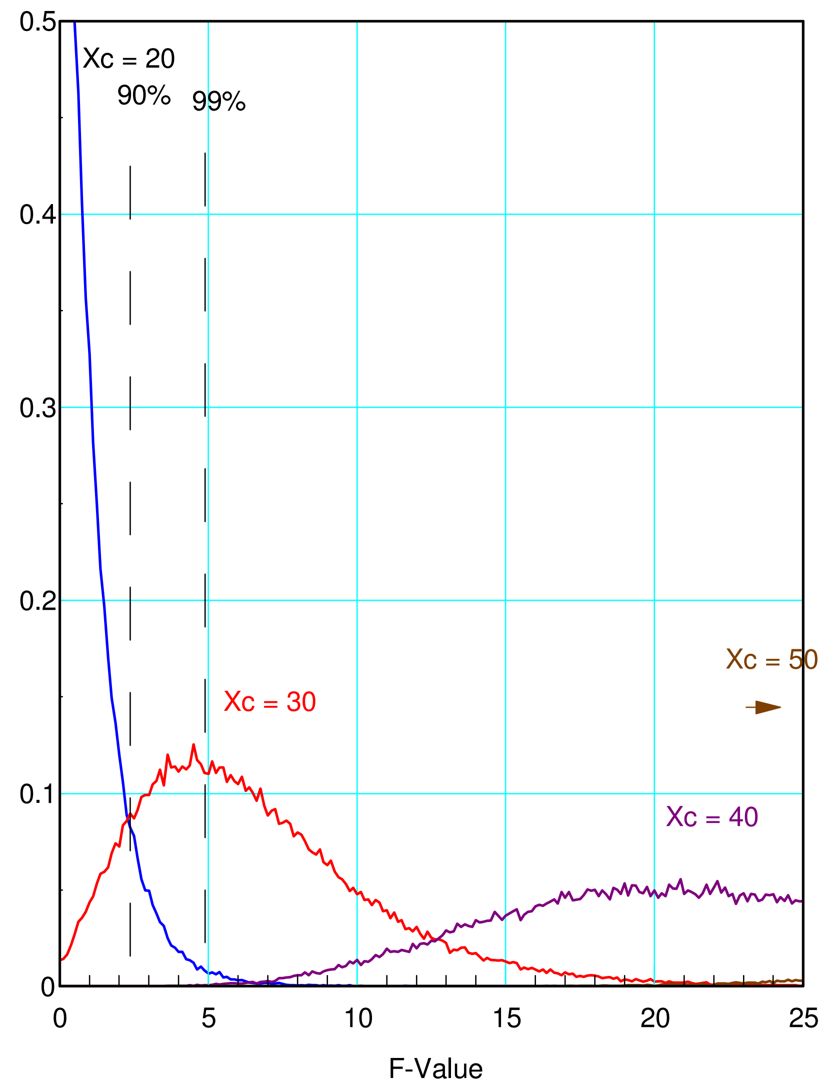
```
for i=1:1e5
    A = 10*randn(N0,1) + 20;
    B = 11*randn(N0,1) + 20;
    C = 12*randn(N0,1) + 35;

    Xa = mean(A);
    Xb = mean(B);
    Xc = mean(C);
    Na = length(A);
    Nb = length(B);
    Nc = length(C);
    Va = var(A);
    Vb = var(B);
    Vc = var(C);

    N = Na + Nb + Nc;
    k = 3;
    G = 1/N * (Na*Xa + Nb*Xb + Nc*Xc);
    MSSb = (1/(k-1)) * (Na*(Xa-G)^2 +
Nb*(Xb-G)^2 + Nc*(Xc-G)^2);
    MSSw = 1/(N-k) * ((Na-1)*Va +
(Nb-1)*Vb + (Nc-1)*Vc);

    F = MSSb / MSSw;
end
```

Resulting pdf

- Sample size = 20
- Numerator = 2 dof
- Denominator = 57 dof

Results:

- You can usually detect a 50% difference in the mean with 90% certainty,
- You can almost always detect a 100% difference in mean (Xc = 40) with 99% certainty, and
- There is a lot more noise in the resulting F value.

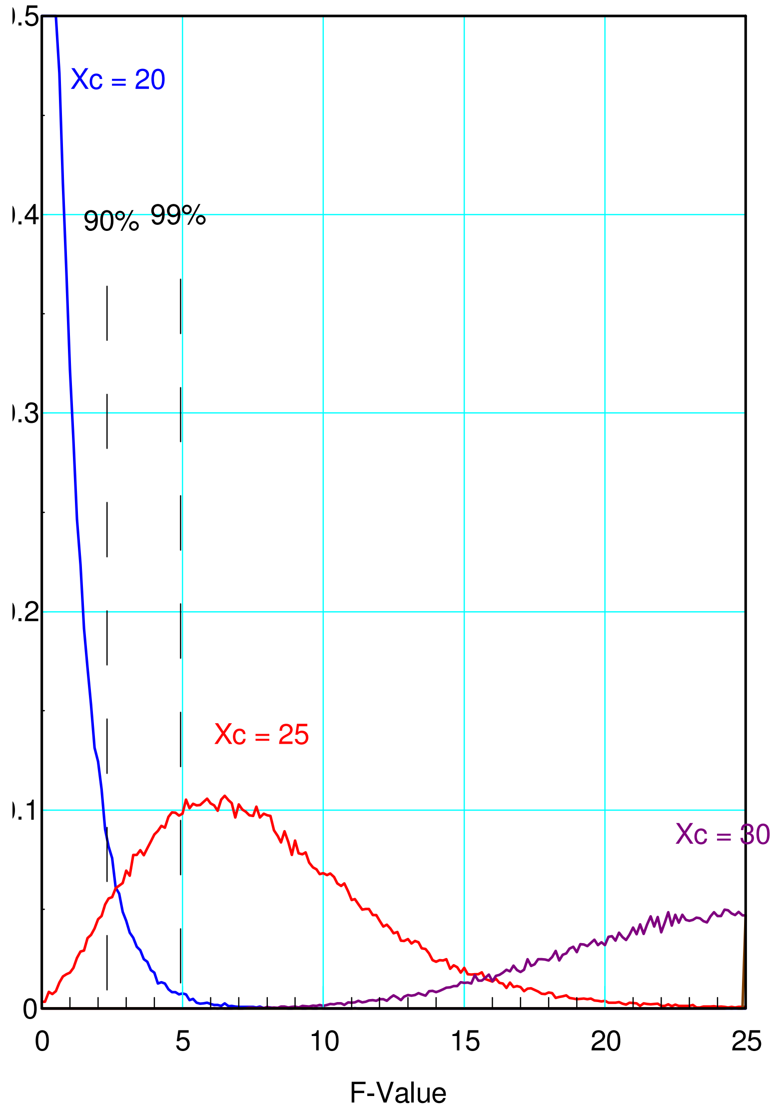


Resulting pdf

- Sample size = 100
- Numerator = 2 dof
- Denominator = 297 dof

Results:

- You can usually detect a 25% difference in the mean (Xc = 25) with 90% certainty,
- You can almost always detect a 50% difference in mean (Xc = 40) with 99% certainty, and
- There is a lot more noise in the resulting F value.



ANOVA Table

The typical (and equivalent) way to compute F is with an ANOVA table.

A	B	C	$(a_i - \bar{A})^2$	$(b_i - \bar{B})^2$	$(c_i - \bar{C})^2$
18.2501	20.7599	21.6631	3.7215	1.2151	1.1884
20.9105	20.2525	21.5629	0.5348	0.3539	1.4169
20.8671	24.2810	23.0827	0.4732	21.3761	0.1086
19.9201	18.3500	22.7785	0.0671	1.7098	0.0006
20.8985	17.3186	23.5025	0.5174	5.4708	0.5614
20.1837	18.3890	25.5565	0.0000	1.6093	7.8584
20.2908	18.4600	24.4461	0.0125	1.4342	2.8658
20.1129	19.4496	19.4335	0.0044	0.0433	11.0206
19.9649 mean(A)	19.6576 mean(B)	22.7532 mean(C)	5.33	33.21	25.02
20.7588 global mean (G)			63.5638 SSw		
8 na	8 nb	8 nc	3.0268 MSSw		
24 N					
43.95 SSb					
21.97 MSSb					

F = MSSb / MSSw

F = 7.2585

Step 1: Start with the data (shown in yellow)

Step 2: Calculate MSSb (shown in blue)

- Find the mean of A, B, C

```
mean(A)
```

- Find the global mean, G

```
G = mean( [A;B;C] )
```

- Find the number of data points in A, B, C

```
Na = length(A)
```

- Find the total number of data points

```
N = Na + Nb + Nc
```

- Compute the sum-squared total between columns

```
SSb = Na*(mean(A)-G)^2 + Nb*(mean(B)-G)^2 + Nc*(mean(C)-G)^2
```

- Compute the mean sum-squared to tal between columns

```
MSSb = SSb / (k-1)
```

Step 3: Calculate MSSw (shown in pink)

- Compute $\left(a_i - \overline{A}\right)^2$

```
(A - mean(A)).^2
```

- Find the total

```
sum( (A-mean(A)).^2 )
```

- Add them up

```
SSw = sum((A-mean(A)).^2) + sum((B-mean(B)).^2) + sum((C-mean(C)).^2)
```

- Find MSSw

```
MSSw = SSw / (N-k)
```

Compute F

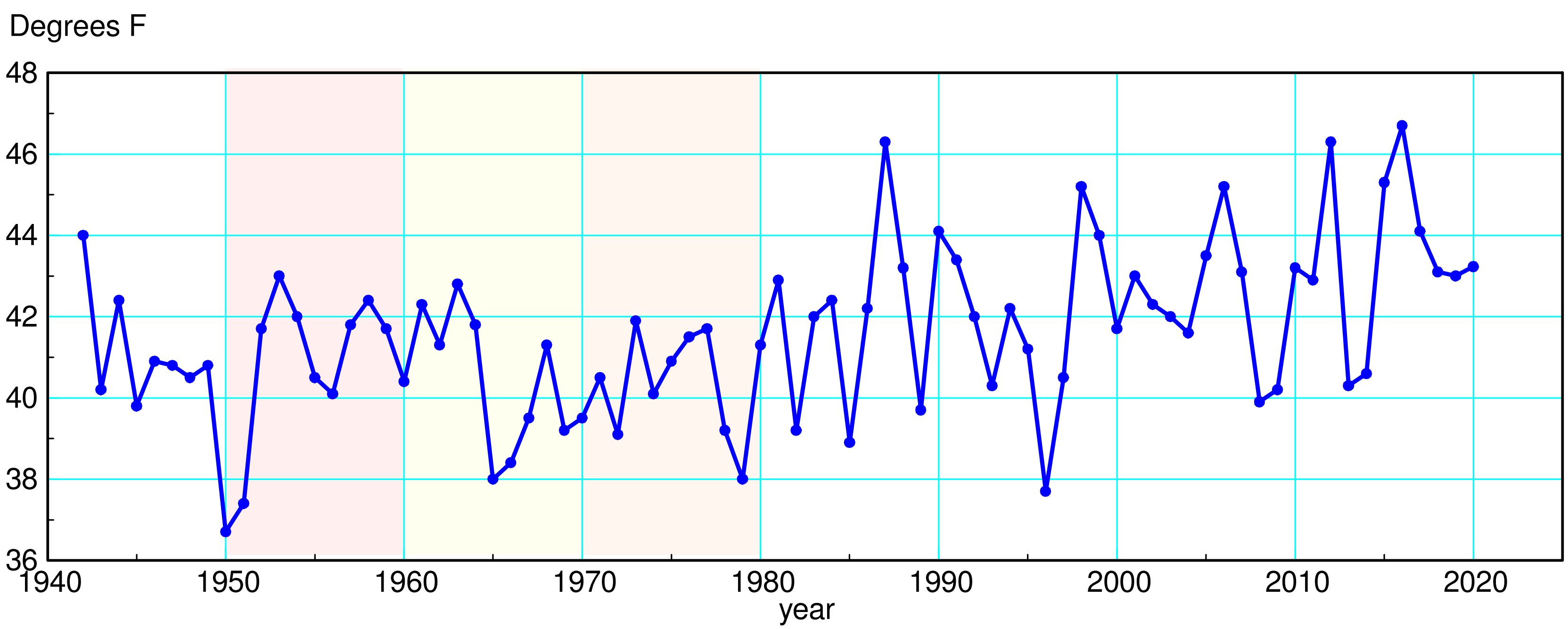
$$F = \left(\frac{MSSb}{MSSw}\right) = 7.2585$$

ANOVA Example:

Compare the average yearly temepratures in Fargo for

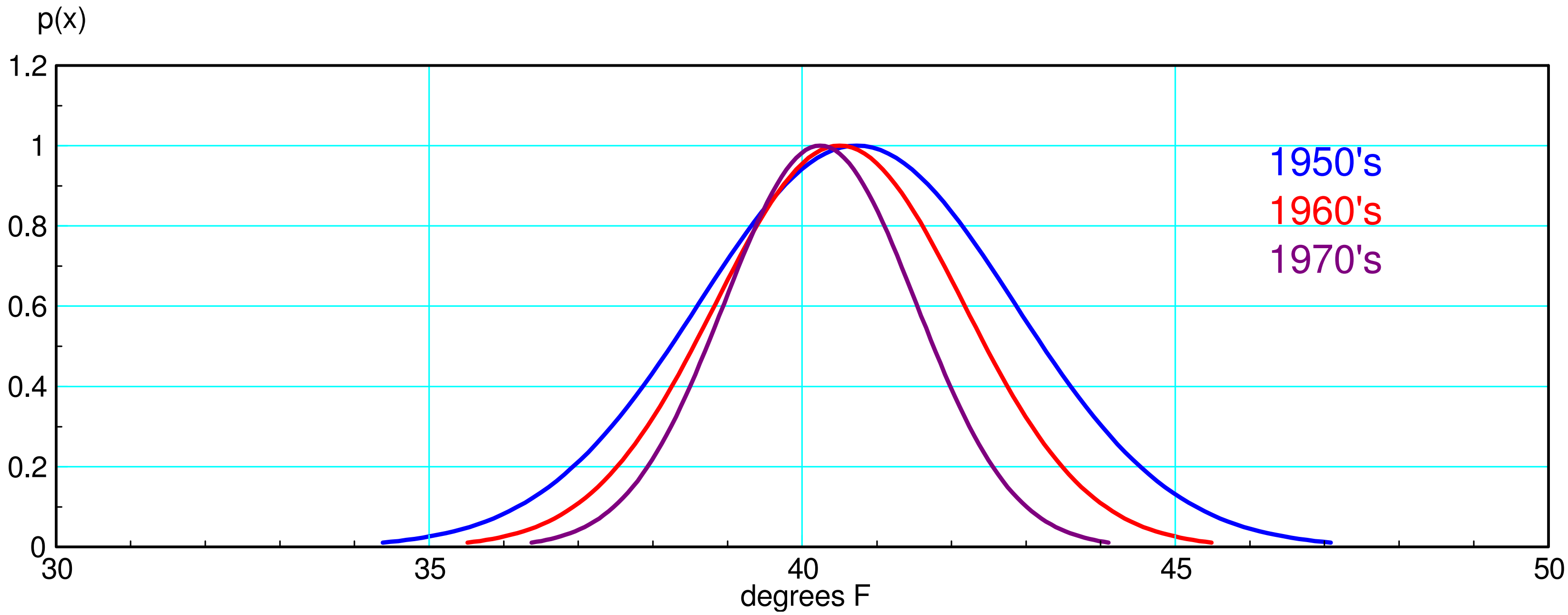
- 1950-1959
- 1960-1969
- 1970-1979

Is the mean temperature for each decade the same?



Data:

	Decade	Mean	St. Dev	N
A	1950-1959	40.73	2.12	10
B	1960-1069	40.5	1.66	10
C	1970-1979	40.24	1.29	10



Matlab Code (standard method)

Placing that algorithm into Matlab

Result:

```
N =       30
G =       40.4900
MSSb =     0.6010
MSSw =     2.9691
F =        0.2024
```

F<1 means no difference in the means

Matlab Code

```
A = T(9:18);
B = T(19:28);
C = T(29:38);

Xa = mean(A);
Va = var(A);
Xb = mean(B);
Vb = var(B);
Xc = mean(C);
Vc = var(C);
Na = length(A);
Nb = length(B);
Nc = length(C);
k = 3;
N = Na + Nb + Nc
G = (Na*Xa + Nb*Xb + Nc*Xc) / N
MSSb = ( Na*(Xa-G)^2 +
         Nb*(Xb-G)^2 +
         Nc*(Xc-G)^2 ) / (k-1)
MSSw = ( (Na-1)*Va +
         (Nb-1)*Vb +
         (Nc-1)*Vc) / (N-k)
F = MSSb / MSSw
```

Matlab Code (non-standard method)

The standard way to do ANOVA is *slightly* wrong

- The reason the F-score is less than 1

The correct way is as follows:

- F > 1 as it should be

```
N   =    30
G   =   -0.2684
MSSb =  0.0147
MSSw =  0.0133
F =     1.1056
```

But, this isn't how ANOVA is computed

Matlab Code

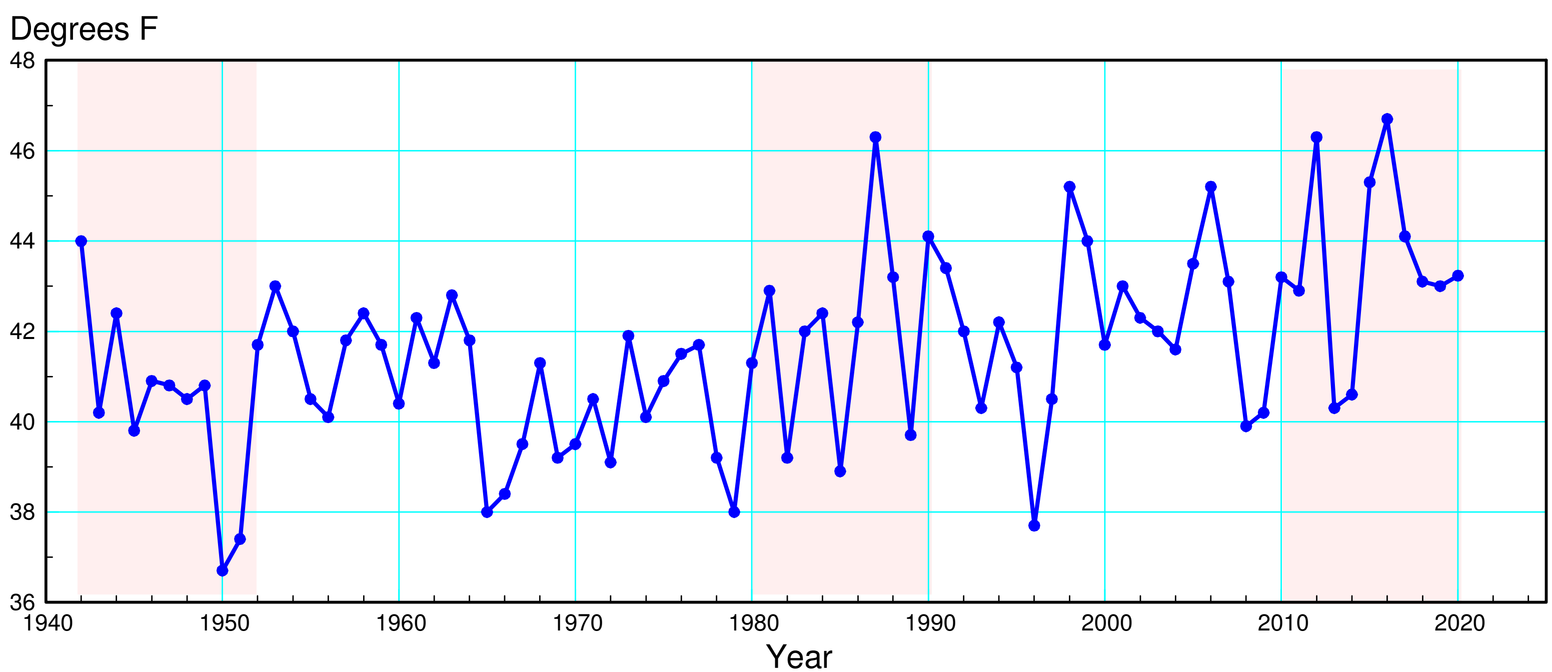
```
A = T(9:18);
B = T(19:28);
C = T(29:38);

Xa = mean(A);
Va = var(A);
Xb = mean(B);
Vb = var(B);
Xc = mean(C);
Vc = var(C);
Na = length(A);
Nb = length(B);
Nc = length(C);
k = 3;
N = Na + Nb + Nc
G = (Na*Xa + Nb*Xb + Nc*Xc) / N
MSSb = ( sum((A-G).^2) +
         sum((B-G).^2) +
         sum((C-G).^2) ) / (N-1)
MSSw = ( sum((A-Xa).^2) +
         sum((B-Xb).^2) +
         sum((C-Xc).^2) ) / (N-k)
F = MSSb / MSSw
```

ANOVA Example:

Compare the average yearly temepratures in Fargo for

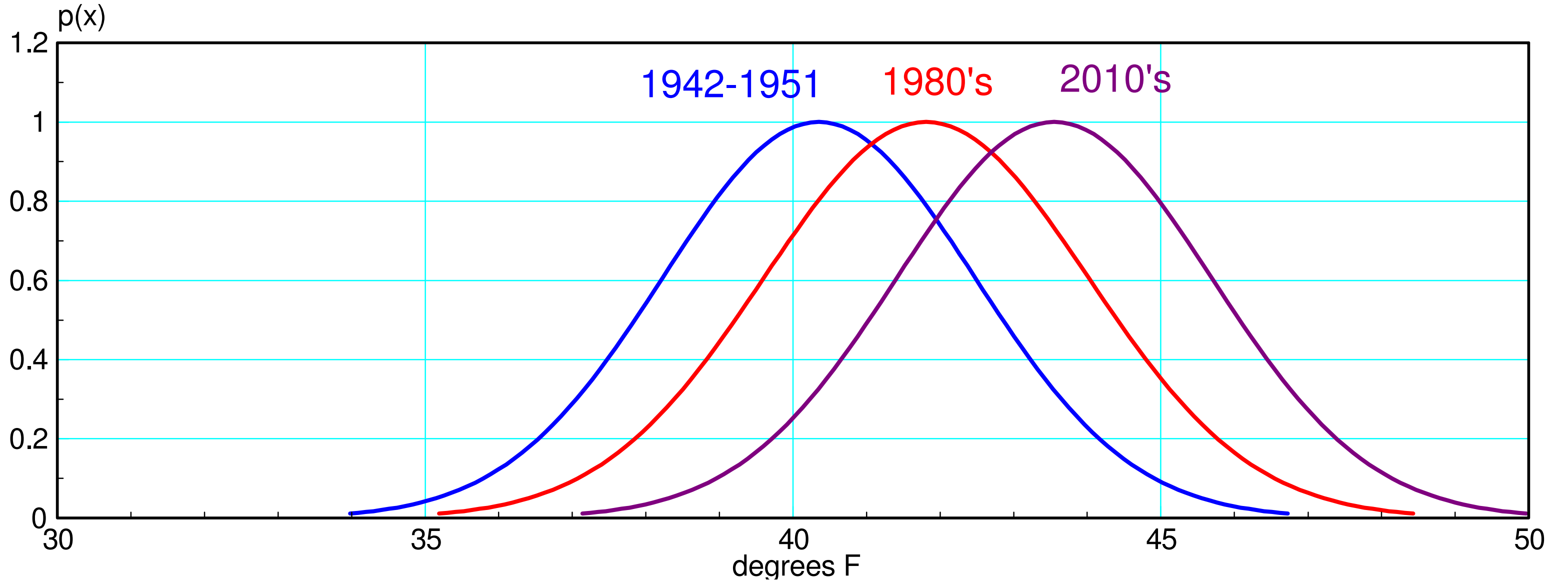
- 1942-1951
- 1980-1989
- 2010-2019

Is the mean temperature for each decade the same?



Data

	Decade	Mean	St. Dev	N
A	1942 - 1951	40.35	2.12	10
B	1980 - 1089	41.81	2.21	10
C	2010 - 2019	43.55	2.14	10



ANOVA Table

A	B	C	$(a_i - \bar{A})^2$	$(b_i - \bar{B})^2$	$(c_i - \bar{C})^2$
44.0	41.3	43.2	13.32	0.26	0.12
40.2	42.9	42.9	0.02	1.19	0.42
42.4	39.2	46.3	4.20	6.81	7.56
39.8	42.0	49.3	0.30	0.03	10.56
40.9	42.4	49.6	0.30	0.34	8.70
40.8	38.9	45.3	0.20	8.46	3.06
40.5	42.4	46.7	0.02	0.15	9.92
40.8	46.3	44.1	0.20	20.16	0.30
36.7	43.2	43.1	13.32	1.93	0.20
37.4	39.7	43.0	8.70	4.45	0.30
40.35 mean(A)	41.81 mean(B)	43.55 mean(C)	40.60 var(A)*9	43.81 var(B)*9	41.16 var(C)*9
41.9 G			125.58 SSw		
10 na	10 nb	10 nc	4.6511 MSSw = SSw / 27		
30 N					
51.3106 SSb					
25.6653 MSSb					

F = MSSb / MSSw

F = 5.5182

Matlab Code

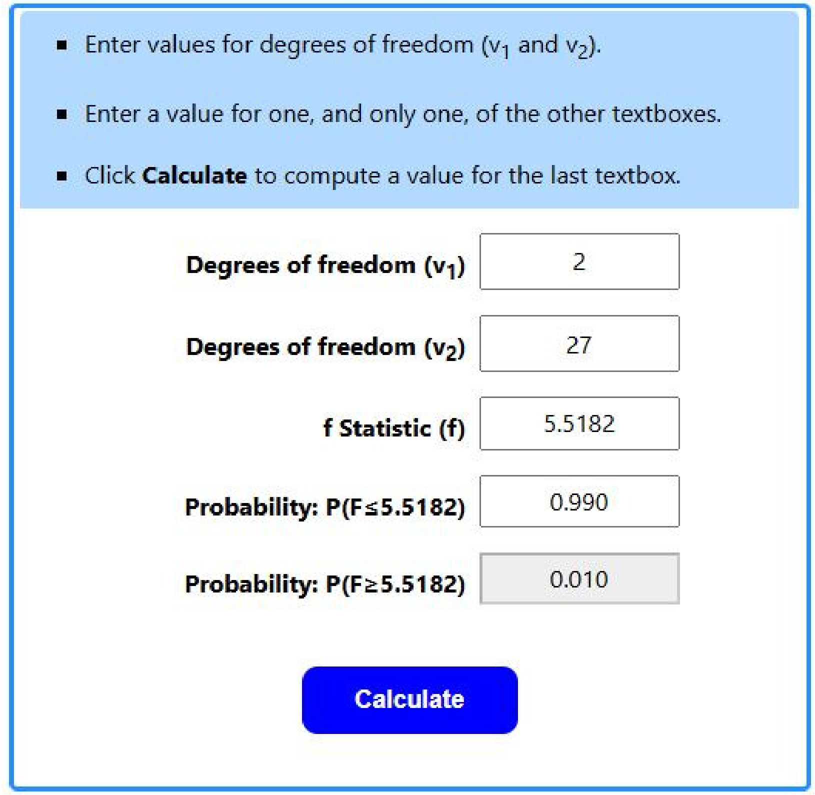
```
A = T(1:10);
B = T(39:48);
C = T(69:78);

Xa = mean(A);
Va = var(A);
Xb = mean(B);
Vb = var(B);
Xc = mean(C);
Vc = var(C);
Na = length(A);
Nb = length(B);
Nc = length(C);
k = 3;
N = Na + Nb + Nc
G = (Na*Xa + Nb*Xb + Nc*Xc) / N
MSSb = (Na*(Xa-G)^2 + Nb*(Xb-G)^2 + Nc*(Xc-G)^2) / (k-1)
MSSw = ((Na-1)*Va + (Nb-1)*Vb + (Nc-1)*Vc) / (N-k)
F = MSSb / MSSw

N =       30
G =       41.9033
MSSb =    25.6653
MSSw =     4.6511
F =        5.5182
```

From StatTrek

- m = 2 dof (numerator)
- d = 27 dof (denominator)
- F = 5.5182
- p = 0.990

It is 99% likely that the three decades have different means

- Something is changing



Summary:

F-Tests allow you to compare the variance

- A large F-score indicates the variance is changing
- A change in variance indicates a manufacturing process is about to fail

ANOVA allows you to compare the mean of 3+ populations

- Result is an F-test
- A large F-score indicated that the means are different
  - The data comes from different populations
  - Something is changing with the system
- A t-test is then needed to see *which* population is the outlier.

- Enter values for degrees of freedom (v_1 and v_2).
- Enter a value for one, and only one, of the other textboxes.
- Click **Calculate** to compute a value for the last textbox.

Degrees of freedom (v_1)	2
Degrees of freedom (v_2)	27
f Statistic (f)	5.5182
Probability: P(F≤5.5182)	0.990
Probability: P(F≥5.5182)	0.010

Calculate