F Test and ANOVA

ECE 341: Random Processes

Lecture #27

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

F-Test

F-tests compare the variance of two distributions.

This is useful

- In manufacturing: one indication that a manufacturing process is about to go out of control (i.e. fail) is the variance in the output starts to increase.
- In stock market analysis: A similar theory holds that increased volatility in the stock market is an indicator of an upcoming recession.
- In comparing the means of 3 or more populations. (t-test is used with one or two populations).

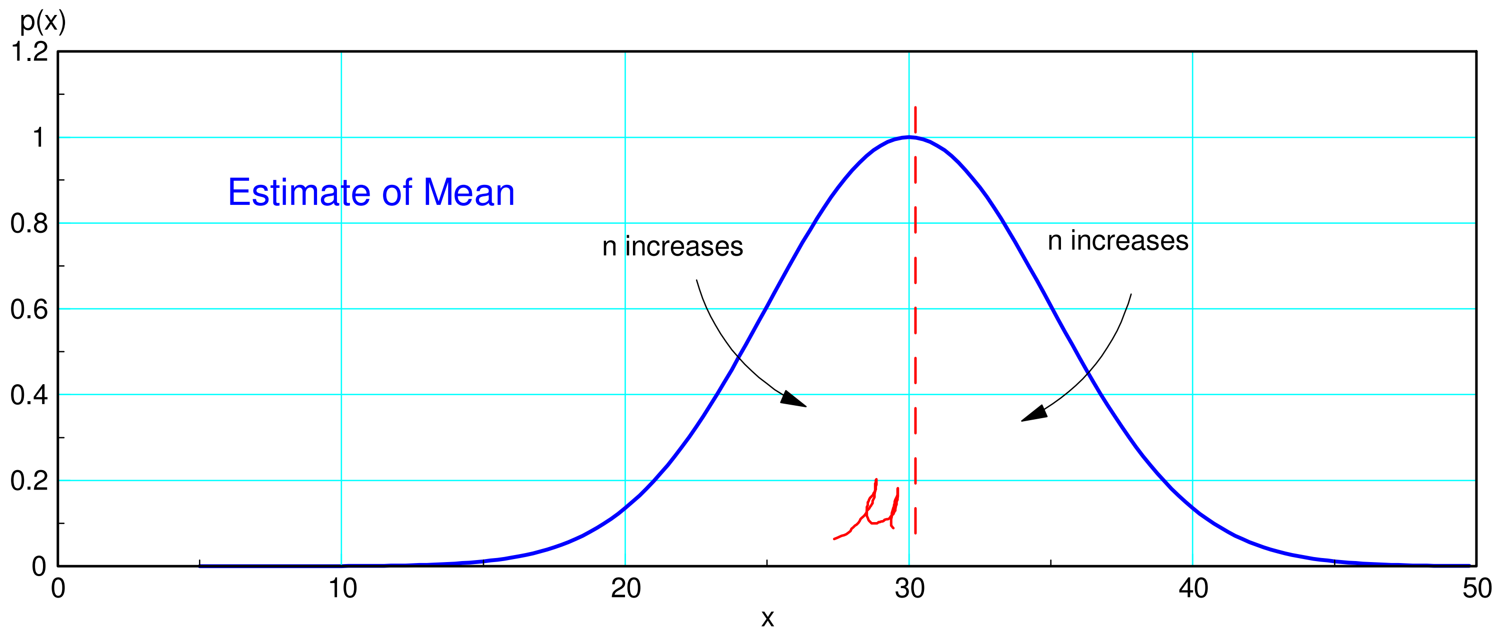
The latter is called an ANOVA (analysis of variance) test and is a fairly common technique.

Distribution of Computed Parameters:

- Assume X has a normal distribution.

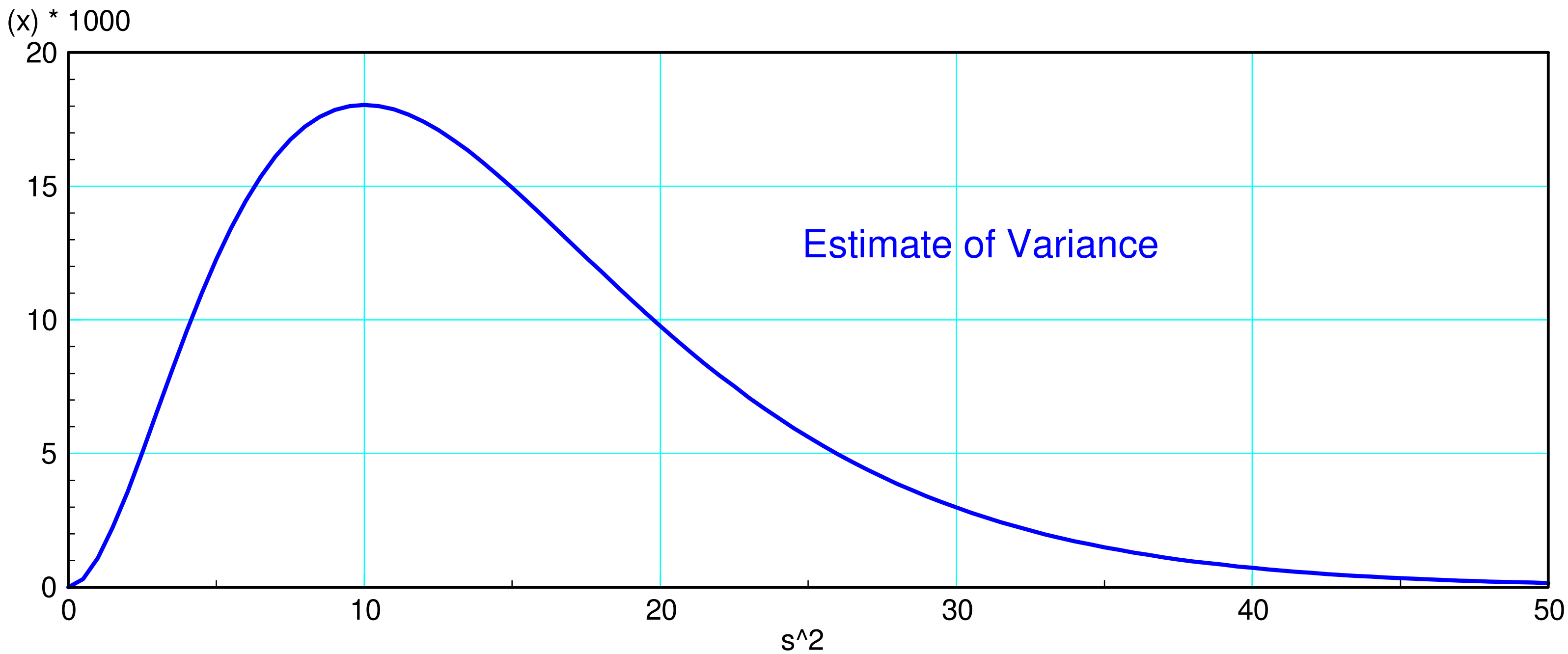
The estimated mean has a normal distribution

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



The estimated variance has a Gamma distribution with n-1 d.o.f.

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \qquad s^2 \sim \Gamma(\sigma^2, n-1)$$

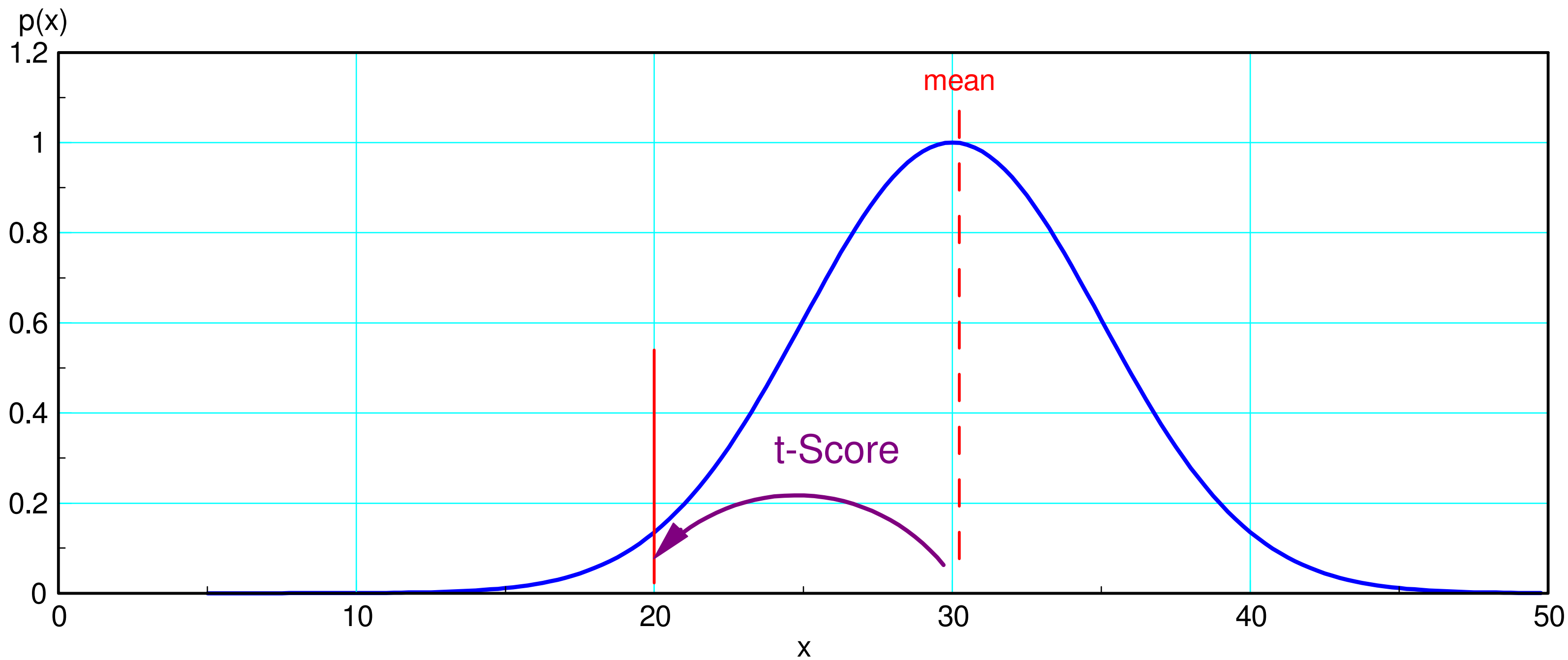


The ratio of

- A Normal distribution and
- A Gamma distribution

is a Student t-distribution with n-1 d.o.f.

$$t = \left(\frac{\beta - \bar{x}}{s}\right) \qquad \sim t(\bar{x}, s^2, n-1)$$



The ratio of

- A Gamma distribution and
- A Gamma distribution

is an F-distribution with

- n-1 (numerator) and
- m-1 (denomionator)

degrees of freedom

$$F = \frac{s_n^2}{s_m^2}$$

Essentially, F distributions are used when you want to compare the variance of two populations.

F-Test

- X is a random variable with unknown mean and variance with m observations
- Y is a random variable with unknown mean and variance with n observations

Test the following hypothesis:

$$H_0: \quad \sigma_x^2 < \sigma_y^2 \qquad \text{or} \qquad H_1: \quad \sigma_x^2 > \sigma_y^2$$

Procedure: Find the sample variance of X and Y:

$$s_x^2 = \left(\frac{1}{m-1}\right)\sum (x_i - \bar{x})^2 \qquad s_y^2 = \left(\frac{1}{n-1}\right)\sum (y_i - \bar{y})^2$$

Define a new variable, F:

$$F = \frac{s_x^2}{s_y^2}$$

Reject the null hypothesis with a confidence level of alpha if V > c

- c is a constant from an F-table.

This is called an F-test.

F-tables tend to be fairly large

- m (numerator dof), n (denominator dof)
- different F-table for each alpha (confidence level).

F-Table for alpha = 0.1
www.statsoft.com/textbook/distribution-tables/

	m = 1	m = 2	m = 3	m = 4	m = 5	m = 10	m = 20	m = 40	m = INF
n = 1	39.86	49.5	53.59	55.83	57.24	60.2	61.74	62.53	63.33
n = 2	8.53	9	9.16	9.24	9.29	9.39	9.44	9.47	9.49
n = 3	5.54	5.46	5.39	5.34	5.31	5.23	5.18	5.16	5.13
n =4	4.55	4.33	4.19	4.11	4.05	3.92	3.84	3.8	3.76
n =5	4.06	3.78	3.62	3.52	3.45	3.3	3.21	3.16	3.11
n =10	3.29	2.92	2.73	2.61	2.52	2.32	2.2	2.13	2.06
n =20	2.98	2.59	2.38	2.25	2.16	1.94	1.79	1.71	1.61
n =40	2.84	2.44	2.23	2.09	2	1.76	1.61	1.51	1.38
n =inf	2.71	2.3	2.08	1.95	1.85	1.6	1.42	1.3	1

Example 1:

Let X and Y be normally distributed:

$$X \sim N(50, 20^2)$$

$$Y \sim N(100, 30^2)$$

Take

- 5 samples from X
- 11 samples from Y

Determine if the variance is different:

$$H_0 : \quad \sigma_x^2 < \sigma_y^2$$

F-Test: Procedure:

- Generate 5 random numbers for X
- Generage 11 random numbers for Y:

```
X = 20*randn(5,1) + 50

   60.7533
   86.6777
    4.8231
   67.2435
   56.3753
```

```
Y = 30*randn(11,1) + 100

   60.7694
   86.9922
  110.2787
  207.3519
  183.0831
   59.5034
  191.0477
  121.7621
   98.1084
  121.4423
   93.8510
```

Find the variance of X and Y.

- If the ratio is less than one, inverse F
- F is always larger than 1.000

```
F = var(X) / var(Y)

F =    0.3542

F = 1 / F

F =    2.8235
```

To convert this F-score to a probability, refer to an F-table.

- The numerator (Y) has 10 degrees of freedom (m = 10)
- The denominator (X) has 4 degrees of freedom (n = 4)

F < 3.92

- No conclusion at a 90% confidence level

F-Table for alpha = 0.1
www.statsoft.com/textbook/distribution-tables/

	m = 1	m = 2	m = 3	m = 4	m = 5	m = 10	m = 20	m = 40	m = INF
n = 1	39.86	49.5	53.59	55.83	57.24	60.2	61.74	62.53	63.33
n = 2	8.53	9	9.16	9.24	9.29	9.39	9.44	9.47	9.49
n = 3	5.54	5.46	5.39	5.34	5.31	5.23	5.18	5.16	5.13
n =4	4.55	4.33	4.19	4.11	4.05	3.92	3.84	3.8	3.76
n =5	4.06	3.78	3.62	3.52	3.45	3.3	3.21	3.16	3.11
n =10	3.29	2.92	2.73	2.61	2.52	2.32	2.2	2.13	2.06
n =20	2.98	2.59	2.38	2.25	2.16	1.94	1.79	1.71	1.61
n =40	2.84	2.44	2.23	2.09	2	1.76	1.61	1.51	1.38
n =inf	2.71	2.3	2.08	1.95	1.85	1.6	1.42	1.3	1

An F-score of 3.920 or more is required to reject the null hypothesis (variances are the same) with 90% certainty

You can also use StatTrek:

- An F-score of 2.8325 means
- p = 0.84

I am 84% certain that the two populations have different variances.

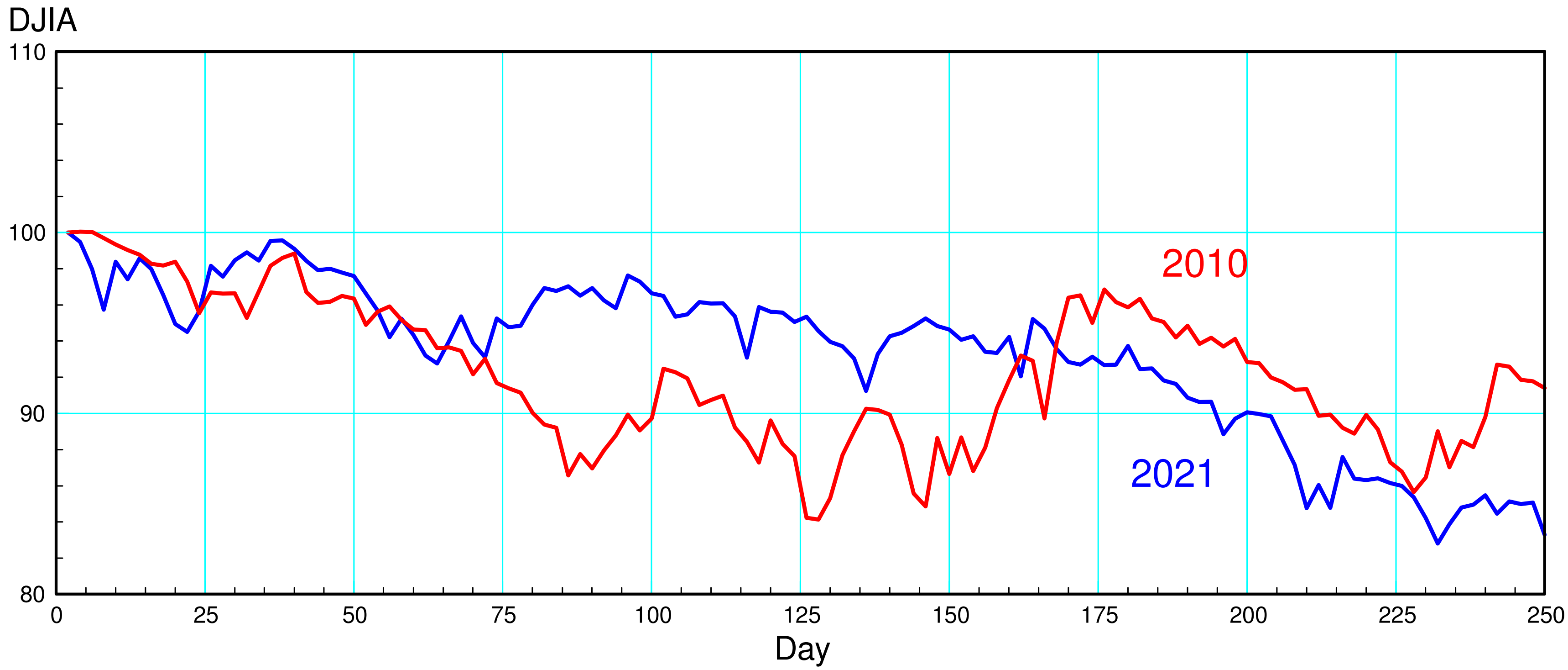
- Enter values for degrees of freedom.
- Enter a value for one, and only one, of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Degrees of freedom (v_1)	10
Degrees of freedom (v_2)	4
Cumulative prob: $P(F \leq 2.8235)$	0.84
f value	2.8235

Example 2: Stock Market

Is the stock market getting more variable?

- Increase in the variance indicated an upcoming crash
- Compare closing price of the DJIA in 2010 and 2021



Data:

Year	Mean	St Dev	# data points
2010	10,664	456.93	251
2021	34,036	1,610.39	250

F-Test

$$F = \left(\frac{1610.39}{456.93}\right)^2 = 12.4212$$

Compute the F-score

- m = 249 dof
- n = 250 dof
- p = 1.0000 (from StatTrek)

Conclusion:

- Yes, the stock market is much more variable than it was 11 years ago
- It's ready for a crash

Data:

- Scale the data so each year starts at 100
- A variation of 100 points relative to 10,000 points is the same as a variation of 300 points relative to a mean of 30,000

Year	Mean	St Dev	# data points
2010	0.9218	0.0395	251
2021	0.9328	0.0441	250

Compute the F-score

$$F = \left(\frac{0.0441}{0.0395}\right)^2 = 1.2465$$

p = 94%

- It still looks like the stock market is much more variable than it was in 2010
- It's ready for a crash

Data (take 3):

- Remove the long-term trend
  - An upward or downward trend is different than more variability
  - Scale the data so each year starts at 100
  - Curve fit as DJIA = at + b
  - Subtrack the trend

Year	Mean	St Dev	# data points
2010	0	0.0368	251
2021	0	0.0223	250

Compute the F-score

$$F = \left(\frac{0.0368}{0.0223}\right)^2 = 2.7232$$

From StatTrek, p = 0.9999

- 2021 is *less* variable than 2010
- The stock market is just fine...

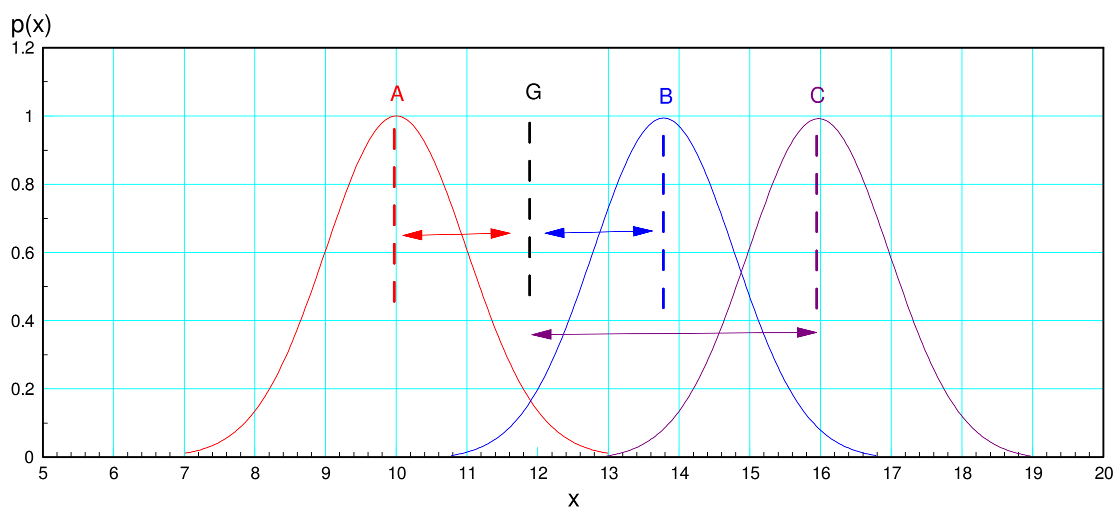
ANOVA

Analysis of Variance

A second use of F distributions it to compare the means of 3+ populations. This is called an Analysis of Variance (ANOVA) test.

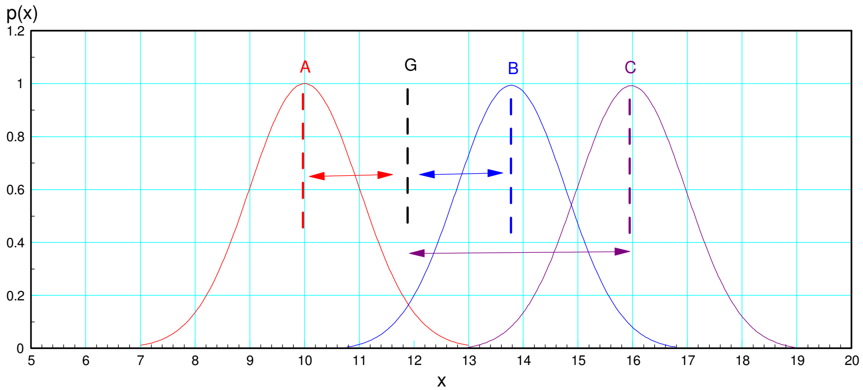
The basic idea is this:

- Assume you have samples from three populations with unknown means and variances
  - Each population will have a mean and a variance
  - The whole sample size will have a mean and a variance

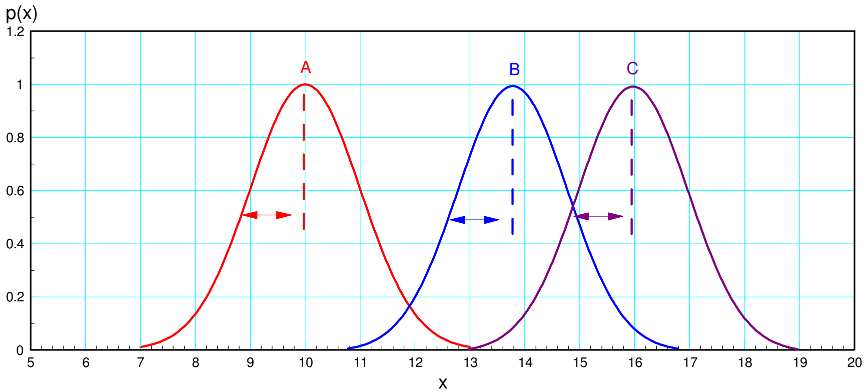


If the variances are different, the means are different (F-test)

$$F = \frac{\mathrm{MSS}_b}{\mathrm{MSS}_w} = \frac{\text{mean sum of squares between data sets}}{\text{mean sum of squares within data sets}}$$



MSSb: The weighted distance (squared) from each populations mean to the global mean (G)



MSSw: The distance (squred) from each data point to it's respective mean

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:

MSSB: Mean Sum Squared Distance Between Columns

MSSb measures the sum squared distance between columns. To take into account sample size, the number of data points in each population is used.

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

The degrees of freedom is k-1: there are k data sets (means) being used in this calculation

d.f.: k - 1

MSSw: Mean Sum Squred Distance Within Columns

MSSw measures the total variance of each population. Two (equivalent) equations are:

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

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

The degrees of freedom are N - k (na-1 + nb-1 + nc-1)

d.f. = N - k

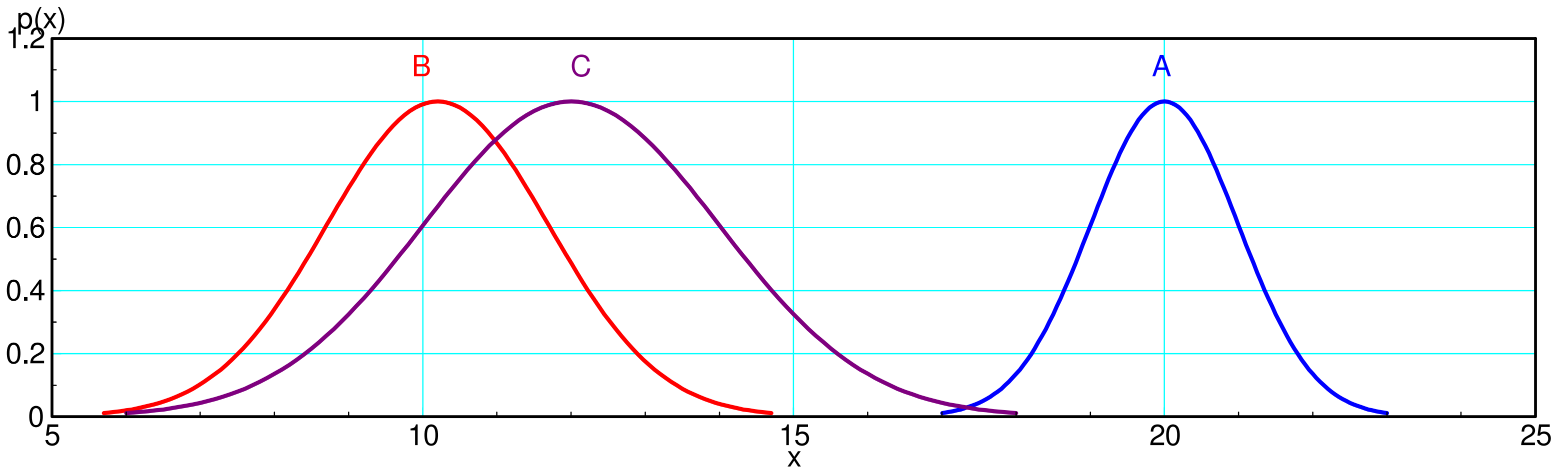
F-value: The F-value is then the ratio

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

ANOVA Example:

Do the following populations have the same mean?

Population A	**Population B**	**Population C**
`A = 1*randn(8,1)+20;`	`B = 1.5*rand(8,1)+10.2`	`C = 2*rand(8,1)+12`
18.2501	20.7599	21.6631
20.9105	20.2525	21.5629
20.8671	24.2810	23.0827
19.9201	18.3500	22.7785
20.8985	17.3186	23.5025
20.1837	18.3890	25.5565
20.2908	18.4600	24.4461
20.1129	19.4496	19.4335



Procedure:

Compue the F-score

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

where

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

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

In Matlab:

Input the data

```
A = [.....];
B = [.....];
C = [.....];
```

Calculate the global average

```
Na = length(A);
Nb = length(B);
Nc = length(C);
N = Na + Nb + Nc;
k = 3;

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

G =   20.8633
```

Calculate MSSb: Mean sum squared difference between populations:

```
MSSb = ( Na*(mean(A)-G)^2 + Nb*(mean(B)-G)^2 + Nc*(mean(C)-G)^2 ) / (k-1)

MSSb =   21.9743
```

Calculate MSSw:

- mean sum squared difference within populations.
- Either equation works: they are equivalent

```
MSSw=( sum( (A-mean(A)).^2) + sum( (B-mean(B)).^2) + sum(
(C-mean(C)).^2))/(N-k)

MSSw =    3.0268

MSSw =  (Na-1)*var(A) + (Nb-1)*var(B) + (Nc-1)*var(C) ) / (N - k)

MSSw =    3.0268
```

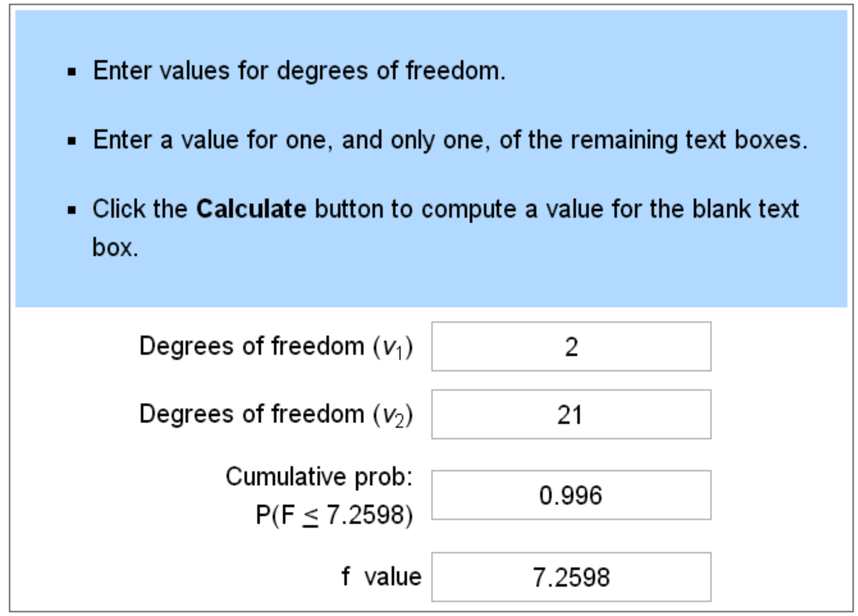
Calculate the F-value:

```
F = MSSb / MSSw

F =    7.2598
```

Convert to a probability:

From StatTrek,

- numerator has 2 d.f. (k - 1)
- denominator has 21 d.f. (N - k)
- F-value = 7.2598
- p = 0.996



There is a 99.6% chance that the means for the data are different

- *You will have to compare means using a t-test to determine which one(s) are the out-liers.*

Matlab Code

```
A = 1*randn(8,1)+20;
B = 1.5*rand(8,1)+10.2;
C = 2*rand(8,1)+12;

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

G    = 20.7588
N    = 24
MSSb = 21.97
MSSw = 3.0268
F    = 7.2585
```

ANOVA Table

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

A	B	C	$\left(a_i - \bar{A}\right)^2$	$\left(b_i - \bar{B}\right)^2$	$\left(c_i - \bar{C}\right)^2$
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

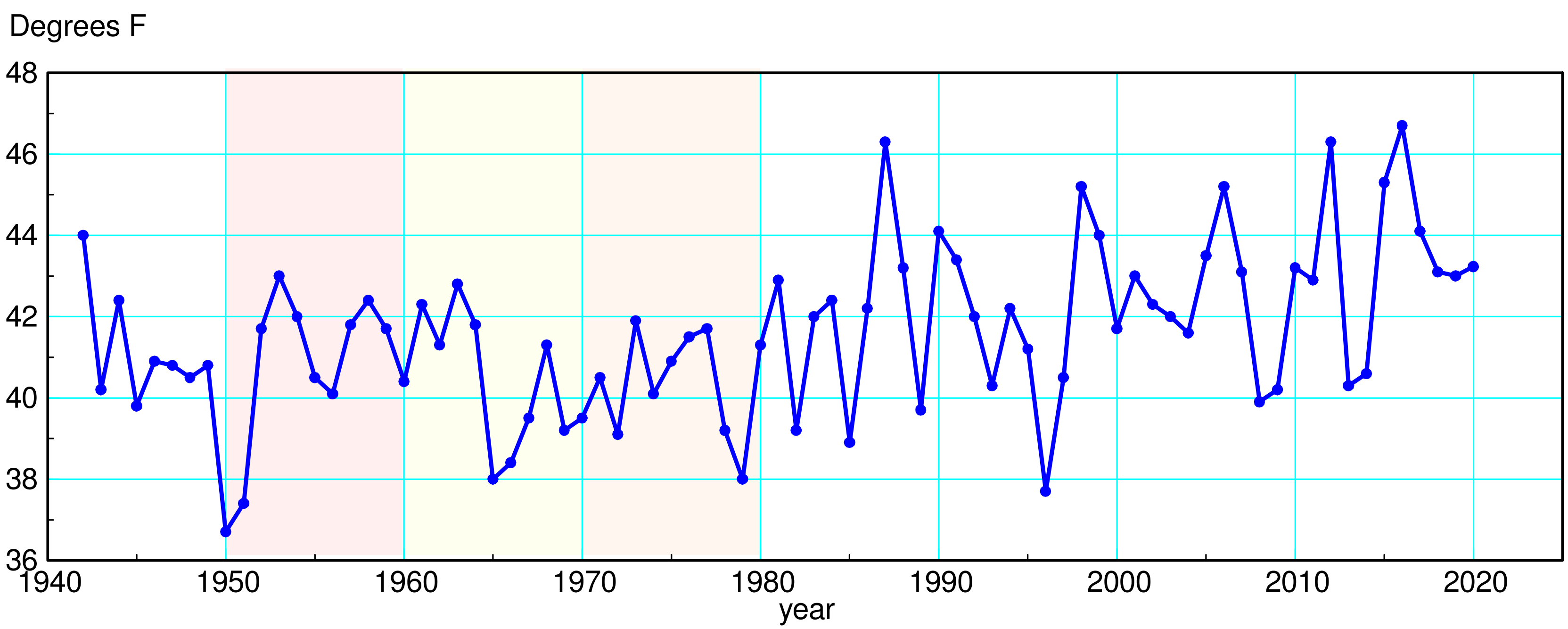
$$F = \left(\frac{MSSb}{MSSw}\right) = 7.2585$$

ANOVA Example:

Compare the average yearly temepratures in Fargo for

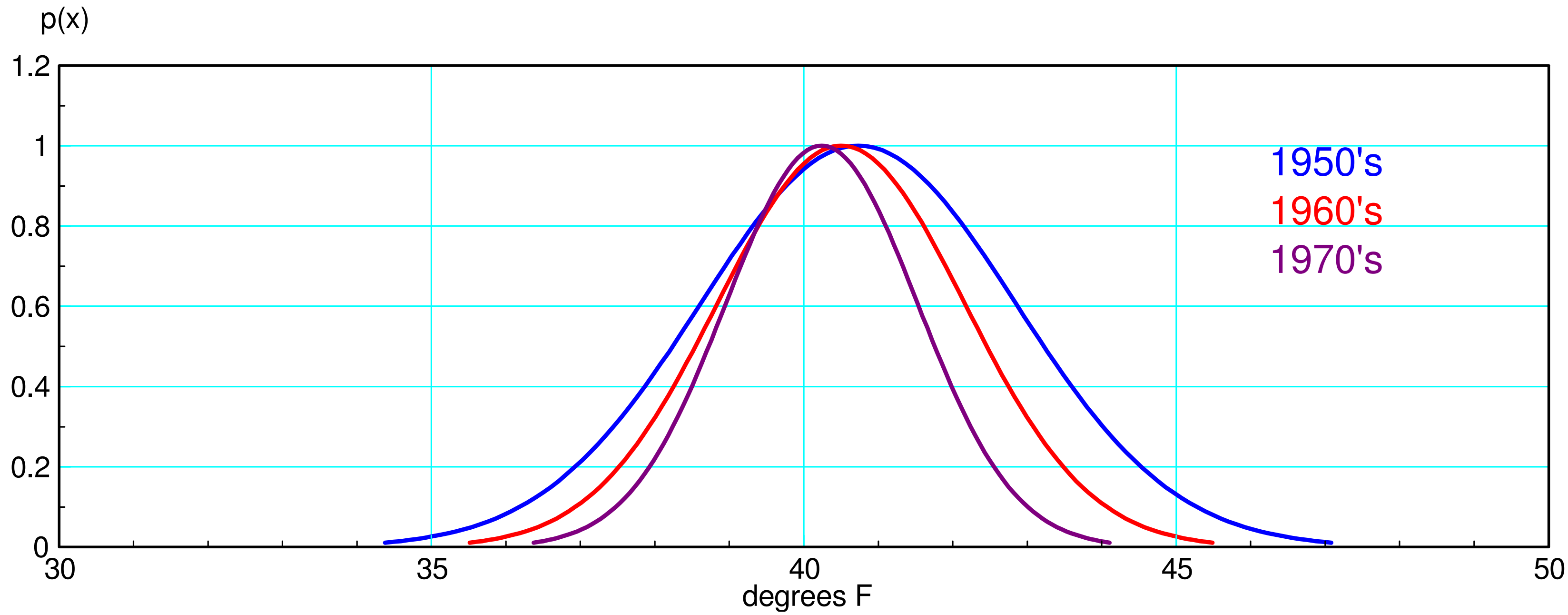
- 1950-1959
- 1960-1969
- 1970-1979

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

```
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

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

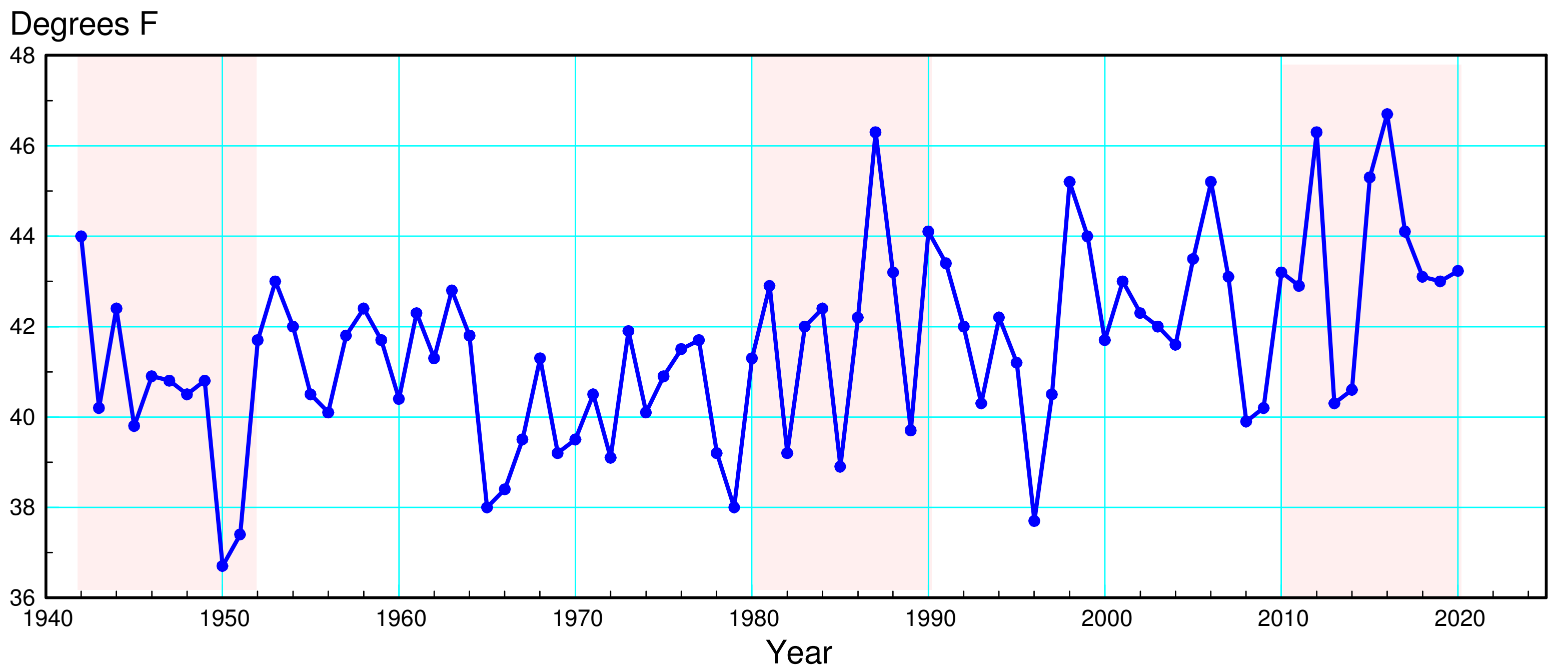
F<1 means no difference in the means

ANOVA Example:

Compare the average yearly temepratures in Fargo for

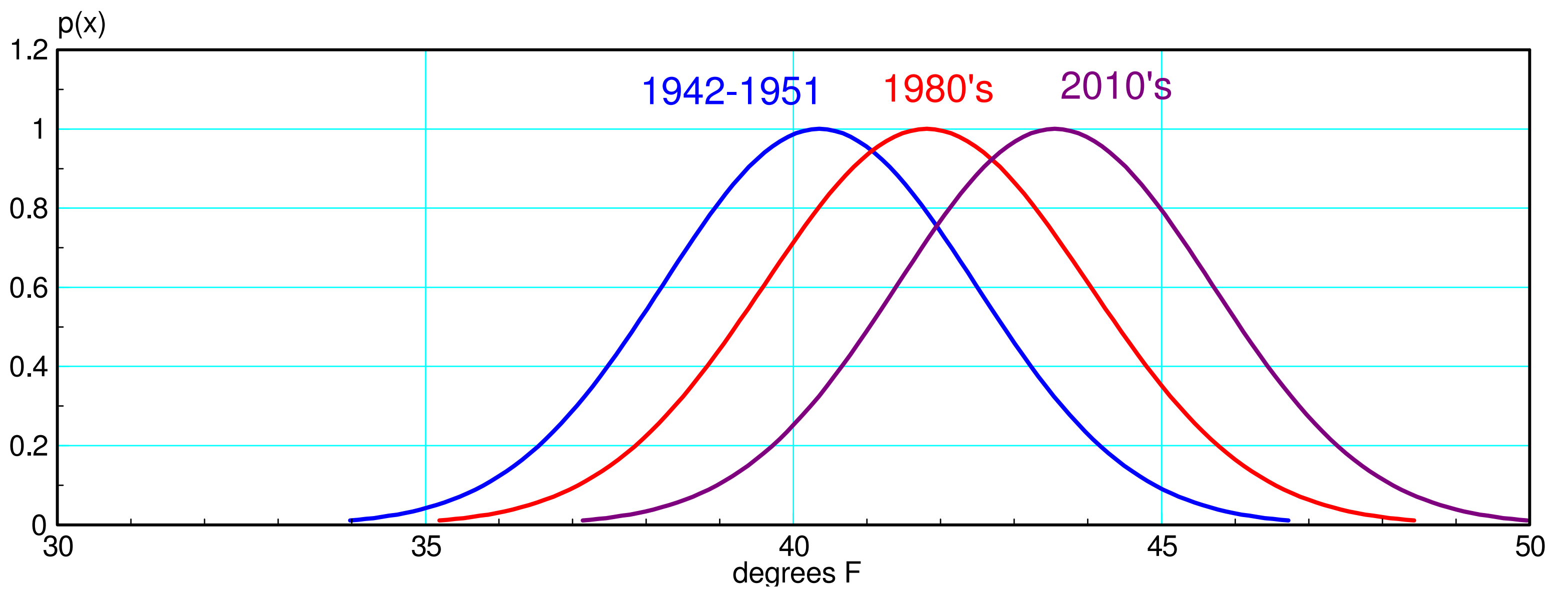
- 1942-1951
- 1980-1989
- 2010-2019

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	$\left(a_i - \bar{A}\right)^2$	$\left(b_i - \bar{B}\right)^2$	$\left(c_i - \bar{C}\right)^2$
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

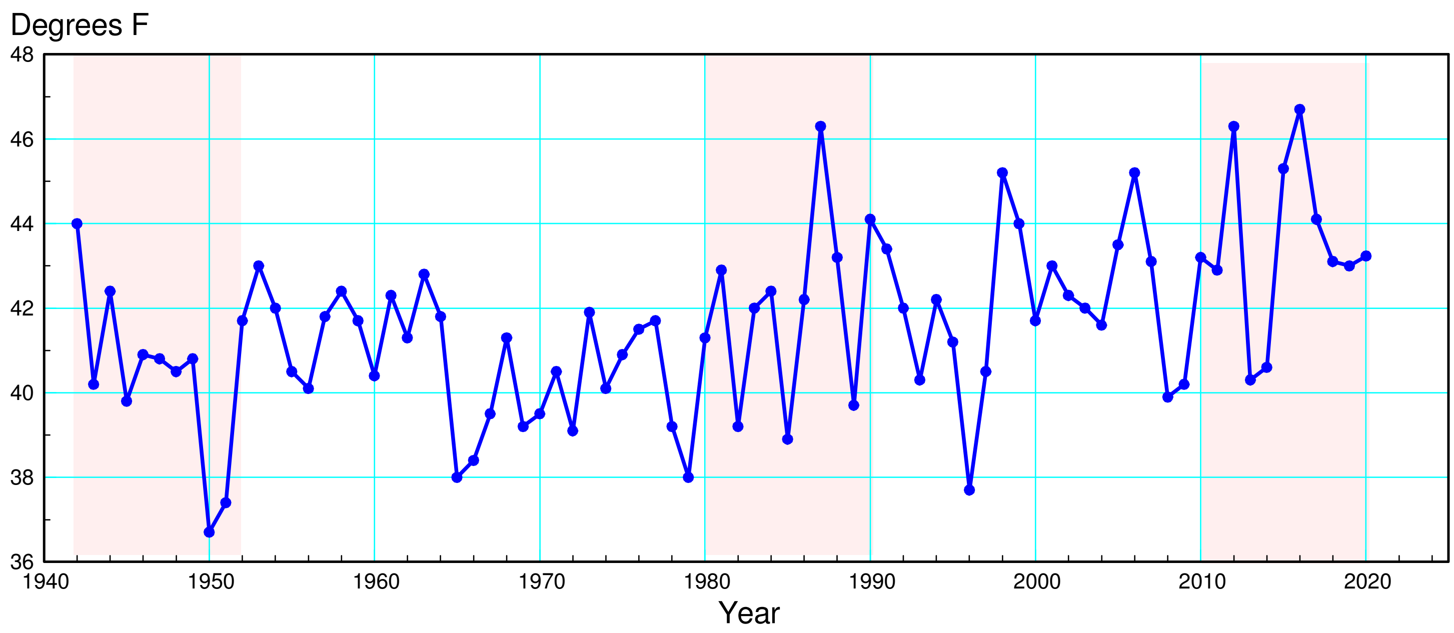
```
A = T(1:10);
B = T(39:48);
C = T(69:78);

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.