F Distribution and ANOVA Test

F-tests detect whether the variance of two distributions are significantly different. 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.

The F-distribution

Assume X has a normal distribution. The estimated mean

$$\bar{x} = \frac{1}{n} \Sigma x_i$$

will have a normal distribution.

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The estimated standard variance

$$s^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$$

results in the standard deviation (s) having a Gamma distribution with n-1 degrees of freedom.

If you take the ratio of a normal distribution divided by a Gamma distribution (i.e. a t-score), the result is a student-t distribution with n-1 degrees of freedom

$$t = \left(\frac{\beta - \bar{x}}{s}\right)$$

If you take a Gamma distribution and divide it by a Gamma distribution (i.e. take the ratio of two variances), the result is an F-distribution with (n-1) and (m-1) 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

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

Test the following hypothesis:

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

$$H_1 : \quad \sigma_x^2 > \sigma_y^2$$

Procedure: Find the sample variance of X and Y:

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

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

Define a new variable, V:

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

You reject the null hypothesis with a confidence level of alpha if V > c where c is a constant from an F-table. This is called an F-test.

F-tables tend to be fairly large since you have m and n degrees of freedom. There is different F-table for each level of confidence, alpha. A shortened version follows.

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	39.863	49.500	53.593	55.833	57.240	60.195	61.740	62.529	63.328
n = 2	8.526	9.000	9.162	9.243	9.293	9.392	9.441	9.466	9.491
n = 3	5.538	5.462	5.391	5.343	5.309	5.230	5.184	5.160	5.134
n =4	4.545	4.325	4.191	4.107	4.051	3.920	3.844	3.804	3.761
n =5	4.060	3.780	3.619	3.520	3.453	3.297	3.207	3.157	3.105
n =10	3.285	2.924	2.728	2.605	2.522	2.323	2.201	2.132	2.055
n =20	2.975	2.589	2.380	2.249	2.158	1.937	1.794	1.708	1.607
n =40	2.835	2.440	2.226	2.091	1.997	1.763	1.605	1.506	1.377
n =inf	2.706	2.303	2.084	1.945	1.847	1.599	1.421	1.295	1.000

MATLAB Example: Let X and Y be normally distributed:

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

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

With 5 samples from X and 11 samples from Y, determine if you can detect whether X has a smaller variance than Y:

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

Procedure: Generate 5 random numbers for X and 11 random numbers for Y:

```
>>

   X = 20*randn(5,1) + 50                         Y = 30*randn(11,1) + 100

      60.7533                                       60.7694
      86.6777                                       86.9922
       4.8231                                      110.2787
      67.2435                                      207.3519
      56.3753                                      183.0831
                                                    59.5034
                                                   191.0477
                                                   121.7621
                                                    98.1084
                                                   121.4423
                                                    93.8510
```

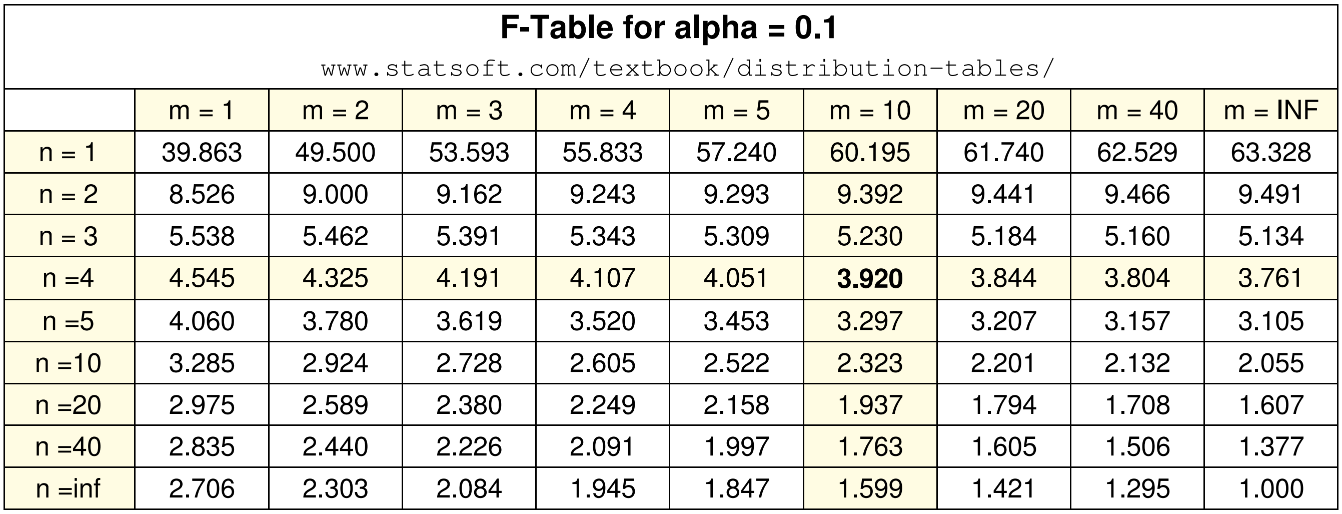
Find the variance of X and Y. If the ratio is less than one, inverse F (F is always larger than 1.000)

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

  F =    0.3542

  F = var(Y) / var(X)

  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)

From the F-table with α =0.1, you need an F-score of at least 3.920 to be 90% certain that the two populations have a different variance. Since the F-score is less than this, there is no conclusion.



F-Table for alpha = 0.1

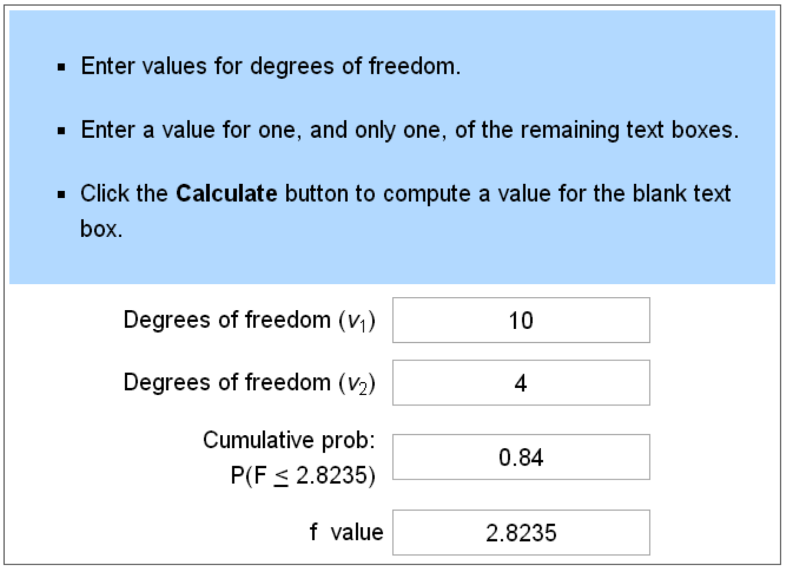
www.statsoft.com/textbook/distribution-tables/

	m = 1	m = 2	m = 3	m = 4	m = 5	m = 10	m = 20	m = 40	m = INF
n = 1	39.863	49.500	53.593	55.833	57.240	60.195	61.740	62.529	63.328
n = 2	8.526	9.000	9.162	9.243	9.293	9.392	9.441	9.466	9.491
n = 3	5.538	5.462	5.391	5.343	5.309	5.230	5.184	5.160	5.134
n =4	4.545	4.325	4.191	4.107	4.051	**3.920**	3.844	3.804	3.761
n =5	4.060	3.780	3.619	3.520	3.453	3.297	3.207	3.157	3.105
n =10	3.285	2.924	2.728	2.605	2.522	2.323	2.201	2.132	2.055
n =20	2.975	2.589	2.380	2.249	2.158	1.937	1.794	1.708	1.607
n =40	2.835	2.440	2.226	2.091	1.997	1.763	1.605	1.506	1.377
n =inf	2.706	2.303	2.084	1.945	1.847	1.599	1.421	1.295	1.000

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 allows you to reject the null hypothesis with a probability of 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 ≤ 2.8235) 0.84

f value 2.8235

ANOVA

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

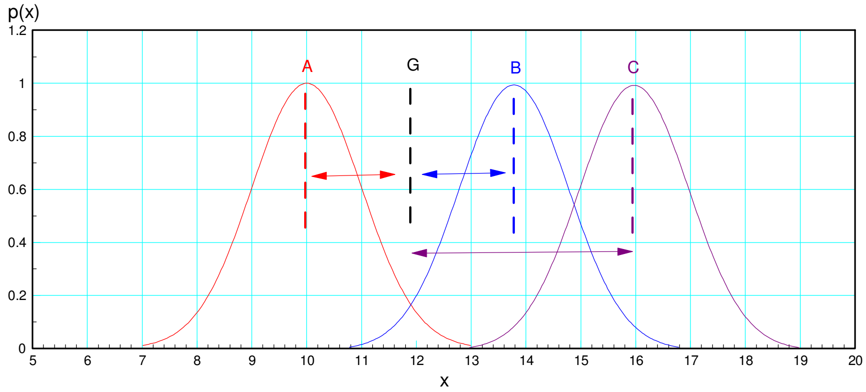
- Each population will have a mean and a variance
- The whole sample size will have a mean and a variance

Now take two measurements:

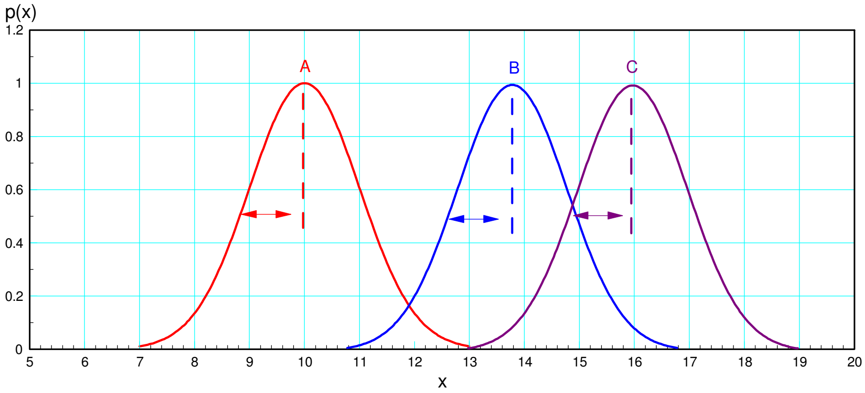
- One measures the mean sum-squared distance from each population to the global mean (mean sum squared between populations, or MSSB)
- And the other measures the mean sum squared distance from each population to that populations' mean (mean sum square within populations, or MSSW)

From these two measurements, form an F-statistic comparing the variances:

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



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

If all populations have the same mean, the two numbers should be the same (and the ratio should be one)

$$F = \frac{MSS_b}{MSS_w} \approx 1$$

If one (or more) populations has a mean which is significantly different, then the ration should be much larger than one:

$$F = \frac{MSS_b}{MSS_w} > 1$$

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 (variance of all data points treated as one population)

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 Table: Usually you set up an ANOVA table to compute the resulting F-value. If you're using Matlab, you only need two compuations (MSSb and MSSw), which aren't too hard.

If you really want to use an ANOVA table, you can split the following compuations into about a dozen computations. You'll get the same result...

ANOVA Example:

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

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

where

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

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

In Matlab:

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

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

G =   20.8633
```

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
```

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
```

F-value:

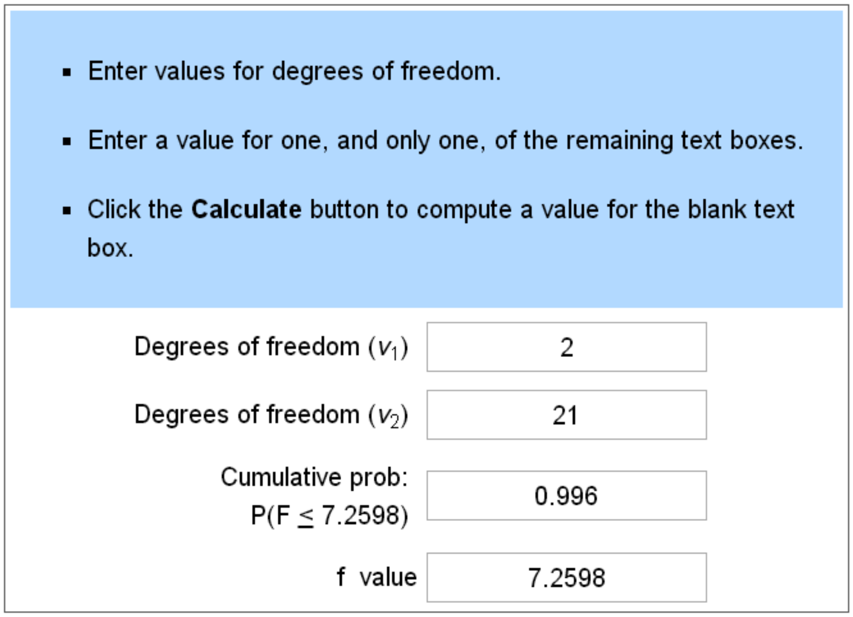
```
F = MSSb / MSSw

F =    7.2598
```

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.



- 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)	2
Degrees of freedom (v_2)	21
Cumulative prob: P(F ≤ 7.2598)	0.996
f value	7.2598

ANOVA Table

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

A	B	C	$\left(a_i - \overline{A}\right)^2$	$\left(b_i - \overline{B}\right)^2$	$\left(c_i - \overline{C}\right)^2$
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	19.6576	22.7532	5.33	33.21	25.02
mean(A)	mean(B)	mean(C)			
20.7588			63.5638		
global mean (G)			SSw		
8	8	8	3.0268		
na	nb	nc	MSSw		
24					
N					
43.95					
SSb					
21.97					
MSSb					

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)`

Summary:

ANOVA analysis produces an F-value:

$$F = \frac{MSSb}{MSSw}$$

Standard Way to ANOVA

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

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

Numerator has k-1 degrees of freedom

Denominator has N-k degrees of freedom