F Distribution

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).

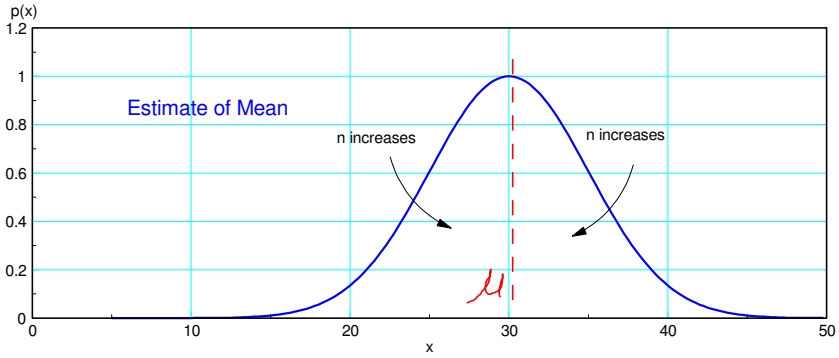
The latter is called an ANOVA (analysis of variance) test and is a fairly common technique.

The F Distribution:

So far, we've covered several typs of distributions in this class. For example, assume X has a normal distribution. The estimated mean has a normal distribution

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

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

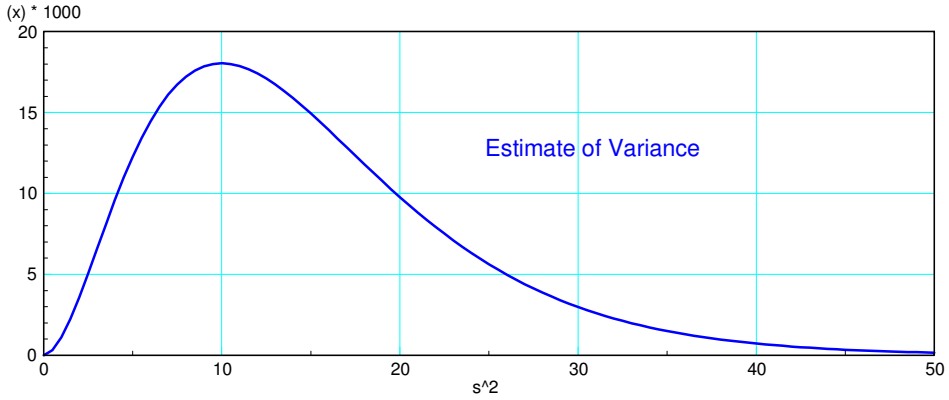


The mean of a sample has a normal distribution

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

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

$$s^2 \sim \Gamma(\sigma^2, n-1)$$



The standard deviation of a sample has a gamma distribution

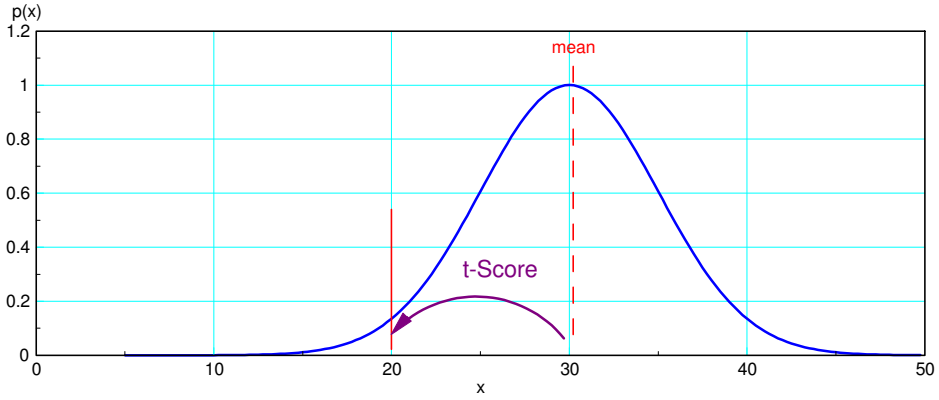
The ratio of

- A Normal distribution and
- A Gamma distribution

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

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

$$\sim t(\bar{x}, s^2, n-1)$$



The ratio of a normal distribution and a gamma distribution is a t-distribution

Finally, the ratio of

- A Gamma distribution and
- A Gamma distribution

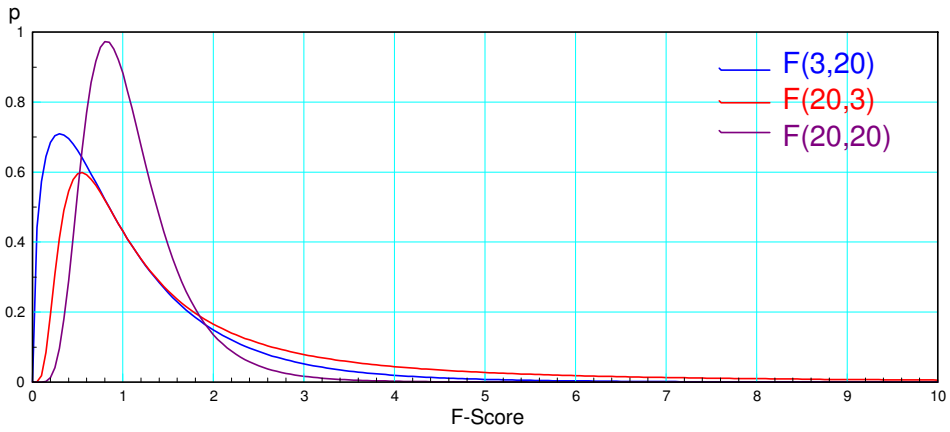
is an F-distribution with

- 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.



The ratio of a gamma distribution and a gamma distgribution is an F-distribution

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
```

```
   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 = 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

F-Distribution Examples:

Do 5% Resistors Have a Uniform Distribution (take 2)

Previously, we looked at modeling 5% tolerance, 1k resistors using a uniform distribution. If the distribution really is a uniform distribution over the range of (950, 1050) Ohms, the mean and variance should be:

Case	Mean	Variance	Sample Size
Uniform	1000	833.330	infinite
Measured	993.57	15.849	56

To see if the measured resitances are consistent with the assumed uniform distribution, you could

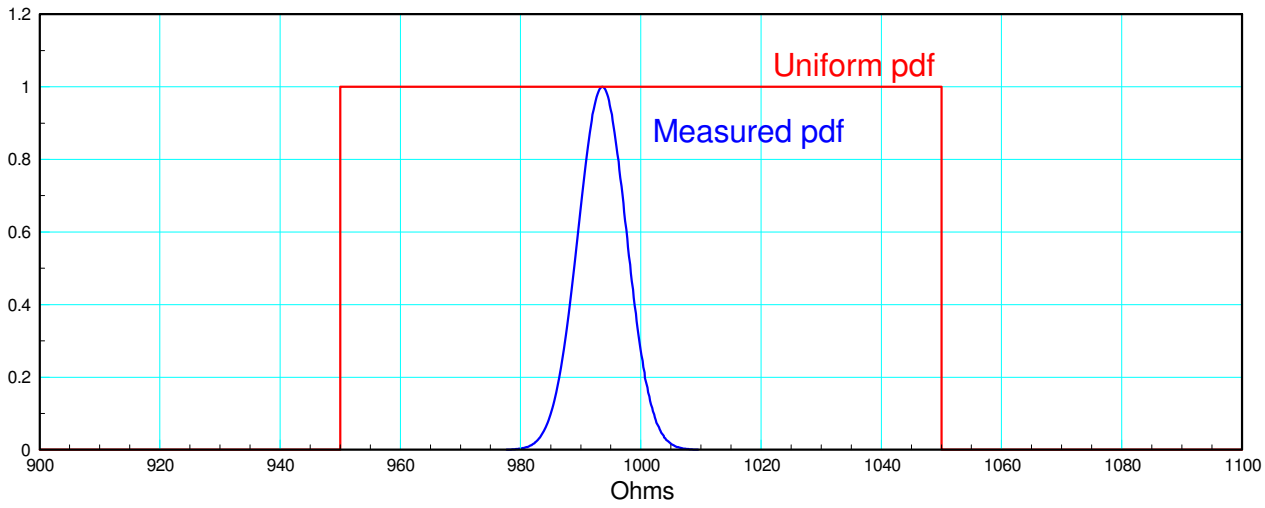
- Use a chi-squared test comparing expected and measured frequency of resistances, or
- Use a t-test to see if the mean is actually 1000, or
- Use an F-test to see if the variance is actually 833.333.

Using the latter, the F-score is

$$F = \left(\frac{833.333}{15.849} \right) = 52.546$$

Using StatTrek with 999 dof in the numerator (should be infinite) and 55 dof in the denominator, this corresponds to a probability of p = 1.000.

Using an F-test, I can be almost certain that 5% tolerance resistors do *not* have a uniform distribution.



3904 Transistors

Four shipments of 3904 transistors are received.

Shipment	Mean	St Dev	Sample Size
1	219.8205	4.1541	39
2	213.6452	3.6382	31
3	165.3333	14.2467	12
4	191.0444	7.6454	45

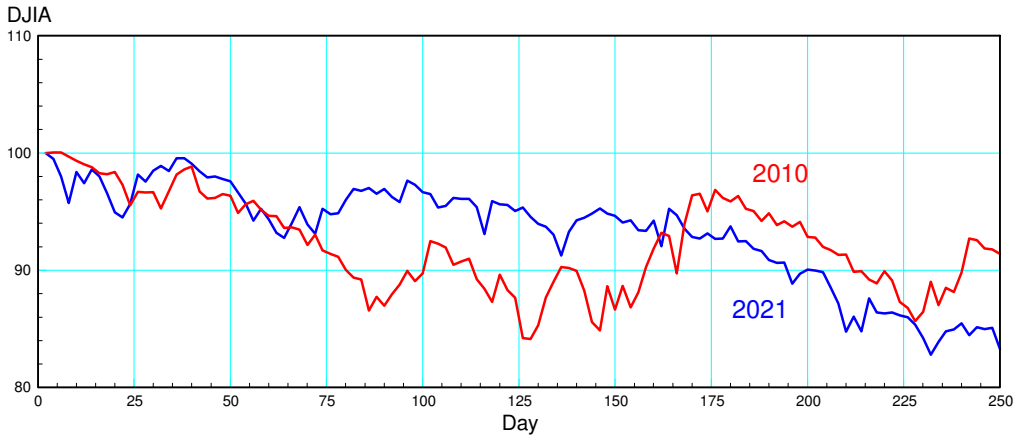
Use an F-test to determine if these transistors have a common supplier and common production run (meaning the variances are the same)

```
>> F12 = var(B1)/var(B2)
F12 =    1.3037
(p = 0.771)

>> F31 = var(B3)/var(B1)
F31 =   11.7620
(p = 1.000)

>> F41 = var(B4)/var(B1)
F41 =    3.3873
(p = 1.000)

>> F32 = var(B3) / var(B2)
F32 =   15.3340
(p = 1.000)

>> F42 = var(B4) / var(B2)
F42 =    4.4160
(p = 1.000)

>> F34 = var(B3) / var(B4)
F34 =    3.4724
(p = 1.000)
```

Shipment #1 and #2 might be from the same source (they were). The rest are almost certainly from different production runs or different manufactureres of 3904 transistors.

Is the Stock Market Going to Crash?

One use of the F-test is to predict when an assembly line is going to crash. Just before the line goes down, the variance in the product typically increases. Using this idea, let's look at trying to predict whether or not the stock market is going to crash by looking at at the variance in the Dow Jones Industrial Average (DJIA) in the year 2010 and 2021. There are several ways to analyze the data. Each gives you a different result.



Normalized Dow Jones Industrial Average in the years 2010 and 2021. Jan 2 is set to 100 in this graph

Using several online sources, the closing value of the DJIA for the first 250 days of each year can be found. Finding the mean and standard deviation of each data set results in the following:

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

The F-score is simplu the ratio of the to variances:

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

From StatTrek, this score can be converted to a probability

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

Conclusion: From this data, it's clear that the stock market was much more variable in 2021 than it was in 2010. It likewise is ripe for a crash.

Actually, 2022, 2023, and 2024 were just fine - suggesting that something was wrong with this analysis. One problem is that the overall average in 2021 was 3.4 times higher than it was in 2010. Likewise, the larger variance may be due to simply being haigher.

DJIA Take 2: A second way to analyze the data is to normalize it so that each year starts out at 100 as in the previous graph. Using this normalized data, the statistics for the DJIA are as follows:

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

Computing the F-score now results in:

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

From StatTrek, this corresponds to a probability of p = 94%. This isn't quite as bad as the previous results, but it again predicts that the stock market is ready for a crash in the coming years - which did not happen.

DJIA Take 3: Another source of problems is that the variance doesn't take into account time. If you take a data set and shuffle the order of the data, you get the same variance. This means that

- Data which has a smooth downward trend has the same variance as
- Data which jumps up and down

What we're trying to measure is the latter case: rapid changing data rather than a smooth overall trend. To fix this problem, adjust the data so that

- Each year starts at 100
- Curve fit as DJIA = at + b
- Subtrack the trend

You could also adjust the start and end point to both be 100 to remove the long-term trend. Doing this results in the following statistics

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

Now, the F-score is

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

From StatTrek, p = 0.9999. This tells you that 2021 is *less* variable than 2010 and the stock market is just fine.

Note that predicting the future is a challenge and you can get different restuls depending upon how you analyze the data. Some methods tell you that the stock market is most certainly on the verge of collapse. Other methods tell you that everything is just fine. So, is it time to buy or sell?

F-Tests and Regression Analysis

Using least squares, you can curve fit data. With an F-test, you can determine if an added term is significant. The procedure here is to first define your function. For a linear curve fit:

$$y_2 = ax + b$$

If you want to see if one of the terms, such as 'ax' is important, set up a second curve fit omitting this term

$$y_1 = b$$

Find the variance of the error in the two curve fits. The ratio is then a F-statistic

$$F = \frac{\sum (y - y_1)^2/(n-1)}{\sum (y - y_2)^2/(n-2)}$$

where the (n-2) and (n-1) terms are the degrees of freedom for y1 and y2 respectively. An equivalent parameter is

$$F = \left(\frac{n-2}{n-1}\right)\left(\frac{var(y - y_1)}{var(y - y_2)}\right)$$

If the F-score is large, it means that the term is significant.

If the F-score is close to 1.000, it means that that term has little impact of estimating y.

For example, consider Arctic sea ice levels.

```
>> x = Data(:,1);
>> y = Data(:,2);
>> n = length(x)

n =    47
```

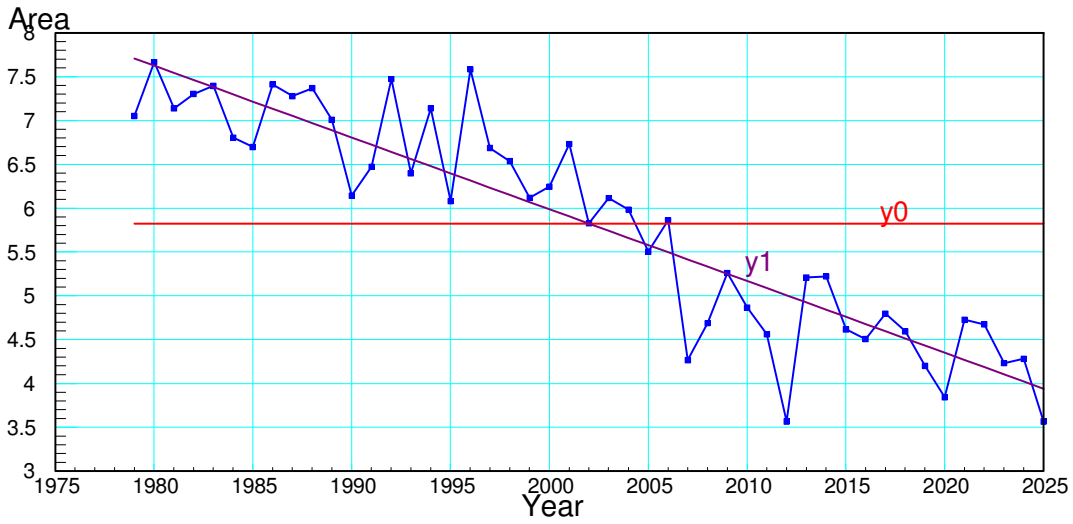
y = a

```
>> B0 = [x.^0];
>> A0 = inv(B0'*B0)*B0'*y

A0 =    5.8230
```

y = a + bx

```
>> B1 = [x.^0, x];
>> A1 = inv(B1'*B1)*B1'*y

a  1.6973e+002
b -8.1872e-002

>> F1 = (n-2)/(n-1) * var(y - B0*A0) / var(y - B1*A1)

F1 =  5.7990
(p = 1.000)
```

It is almost certain that the the minimum Arctic sea ice area contains a 'bx' term

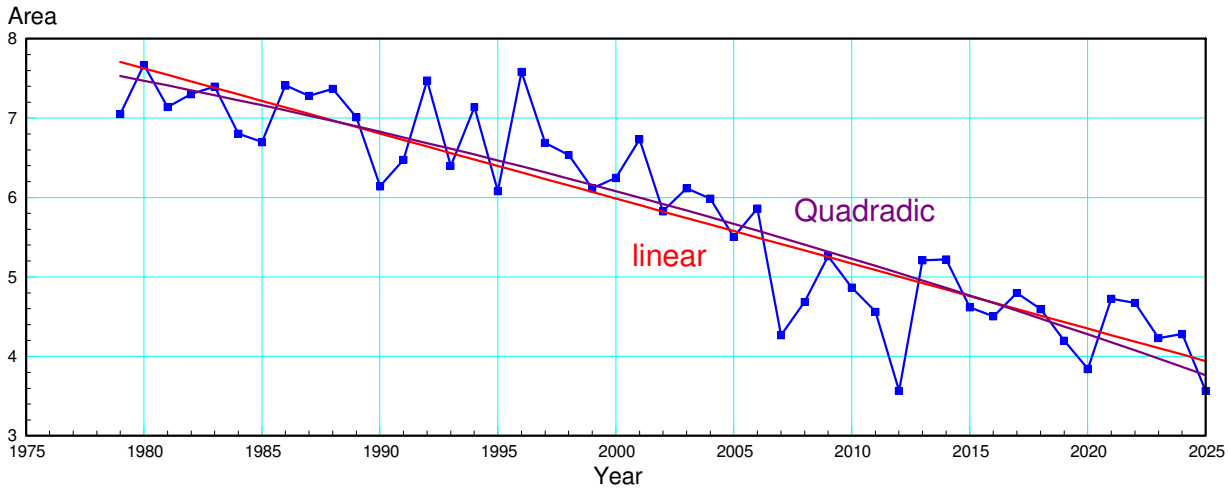


Least squres curve fit witn one term (red) and two terms (purple)

$y = a + bx + cx2$

```
>> B2 = [x.^0, x, x.^2];
>> A2 = inv(B2'*B2)*B2'*y

 -1.8924e+003
  1.9782e+000
 -5.1452e-004

>> F2 = (n-3)/(n-2) * var(y - B1*A1) / var(y - B2*A2)

F2 =  1.0066
(p = 0.508)
```

There is a 50.8% chance that the square term is significant. If the term was significant, it should show up with >90% with an F-test. At this point, there's little justification to using terms higher than 1st-order terms to curve fit the data. We're probably just curve fitting noise.

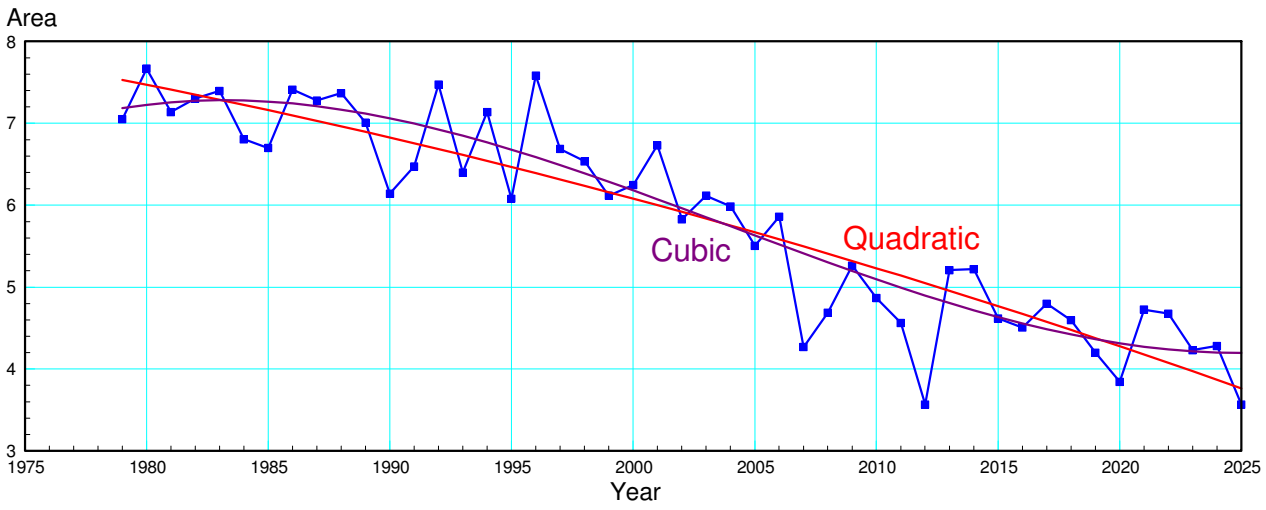


Least squres curve fit witn one term (red) and two terms (purple)

y = a + bx + cx2 + dx3

```
>> B3 = [x.^0, x, x.^2, x.^3];
>> A3 = inv(B3'*B3)*B3'*y

 -6.8554e+005
  1.0265e+003
 -5.1227e-001
  8.5207e-005

>> F3 = (n-4)/(n-3) * var(y - B2*A2) / var(y - B3*A3)

F3 =  1.1173
(p = 0.641)
```

There is a 64.1% chance that the cubic term is significant. Like before, a probability less than 90% usually means we're curve fitting noise rather than the actual trend.



Curve Fit with 3 Terms (red) and 4 terms (purble)

Summary

The F-test is used to compare the variance of two populations. It is useful when trying to determine

- If population A has a larger variance than population B,
- If different terms in a curve fit are significant, or
- If an assembly line is about to crash.