F Test

ECE 341: Random Processes

Lecture #31

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).

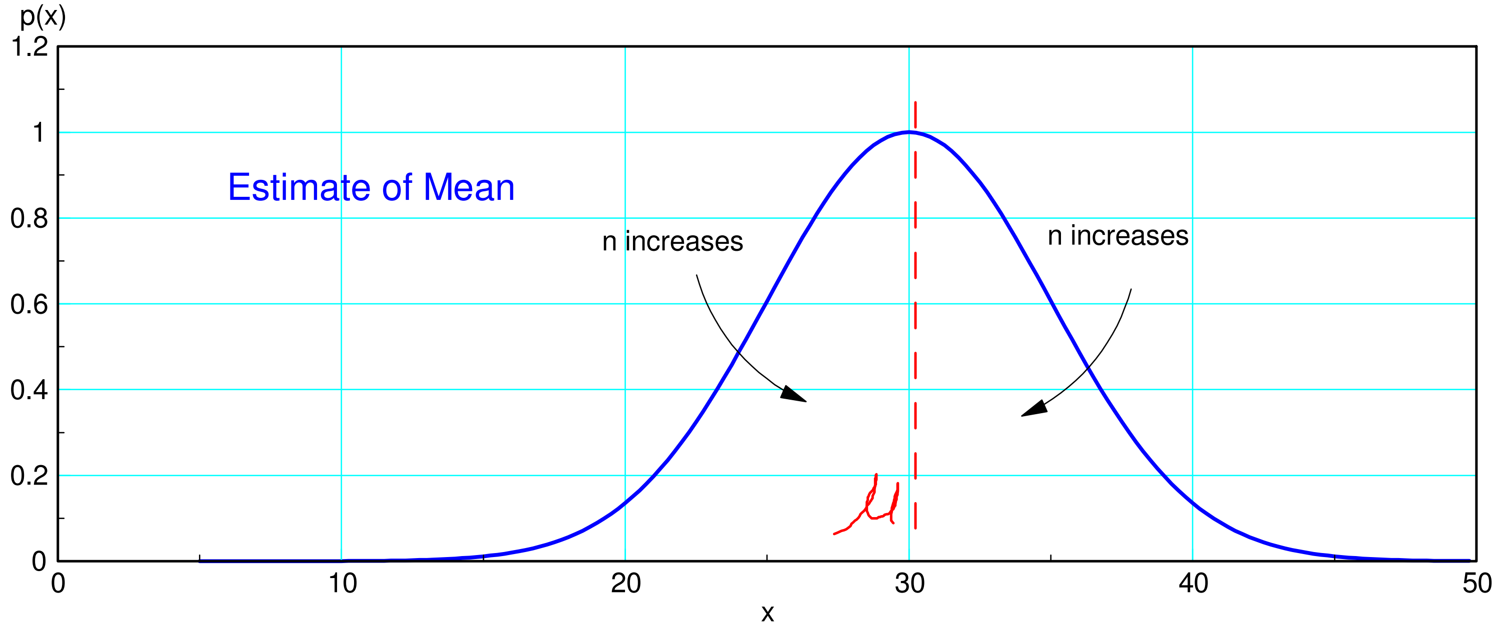
The latter is called an ANOVA (analysis of variance) test and is a fairly common technique.

Distribution of Computed Parameters:

- Assume X has a normal distribution.

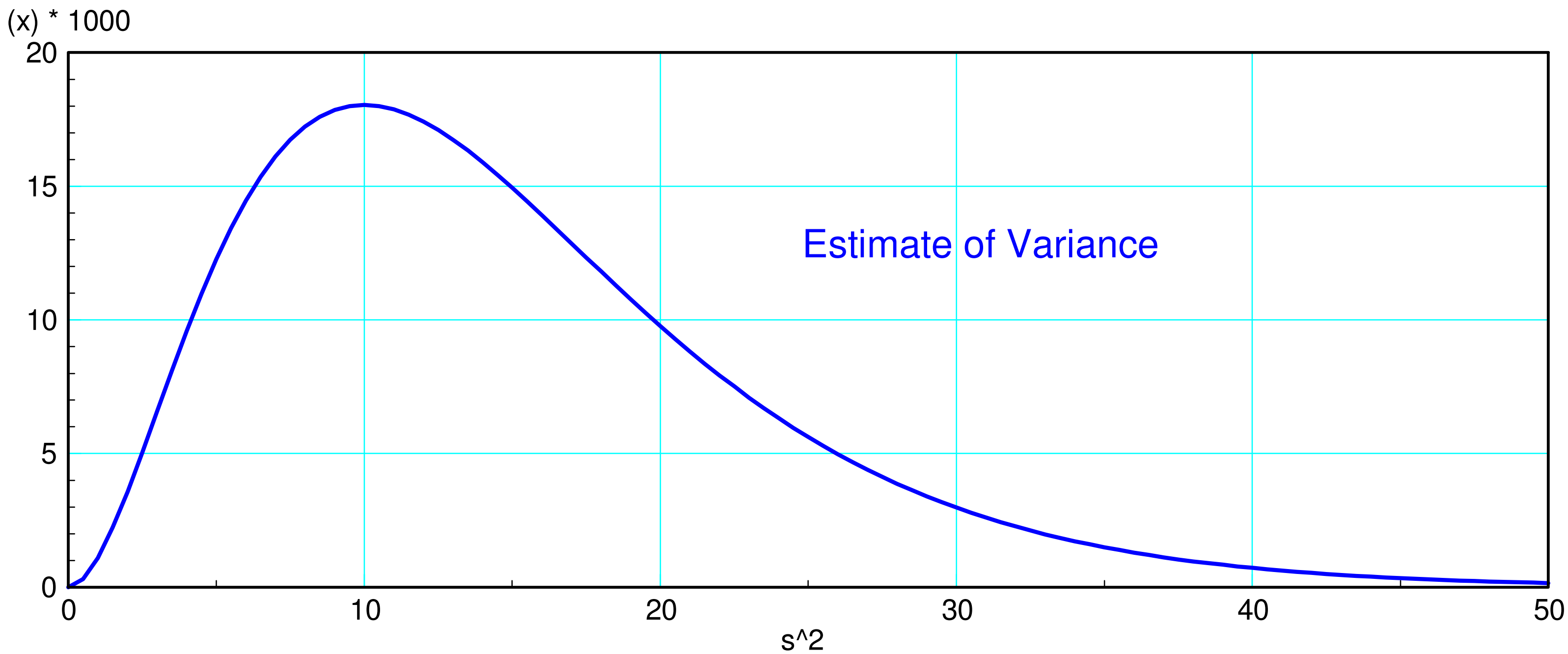
The estimated mean has a normal distribution

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



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

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

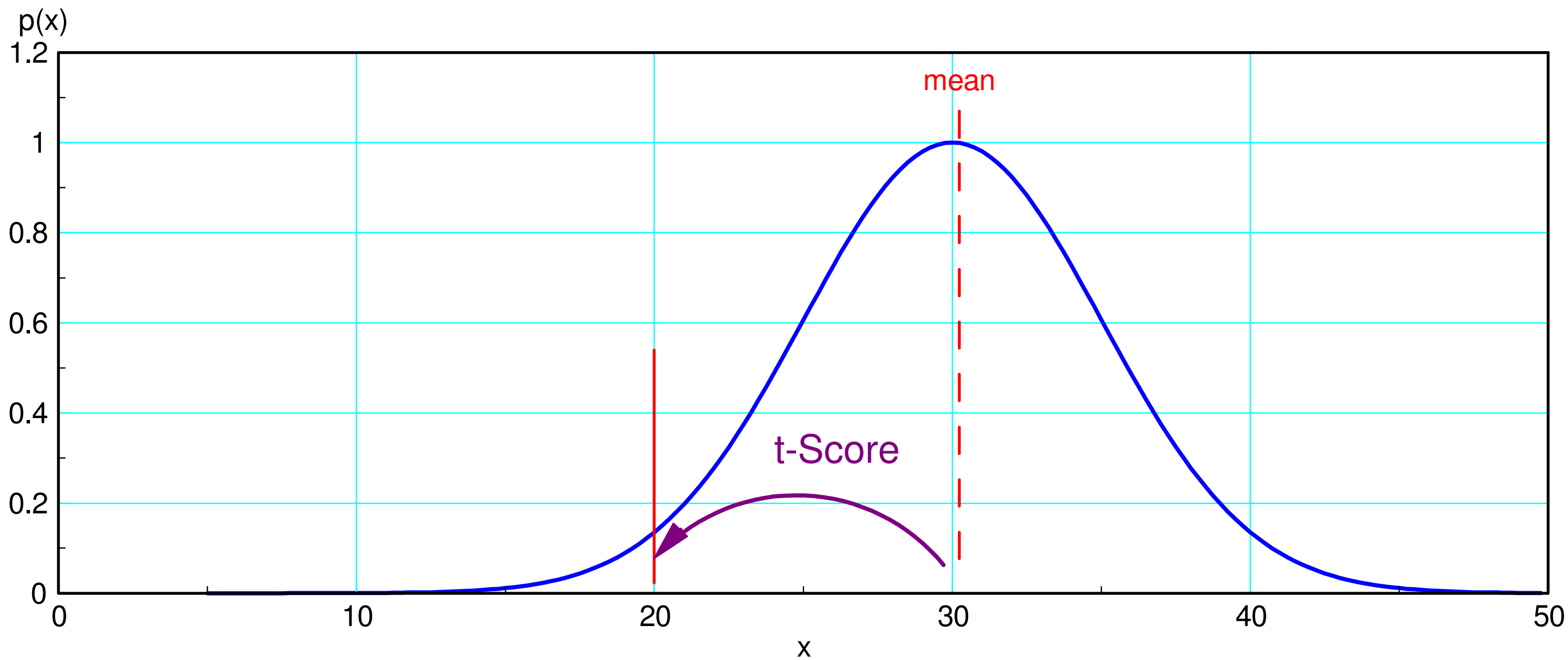


The ratio of

- 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

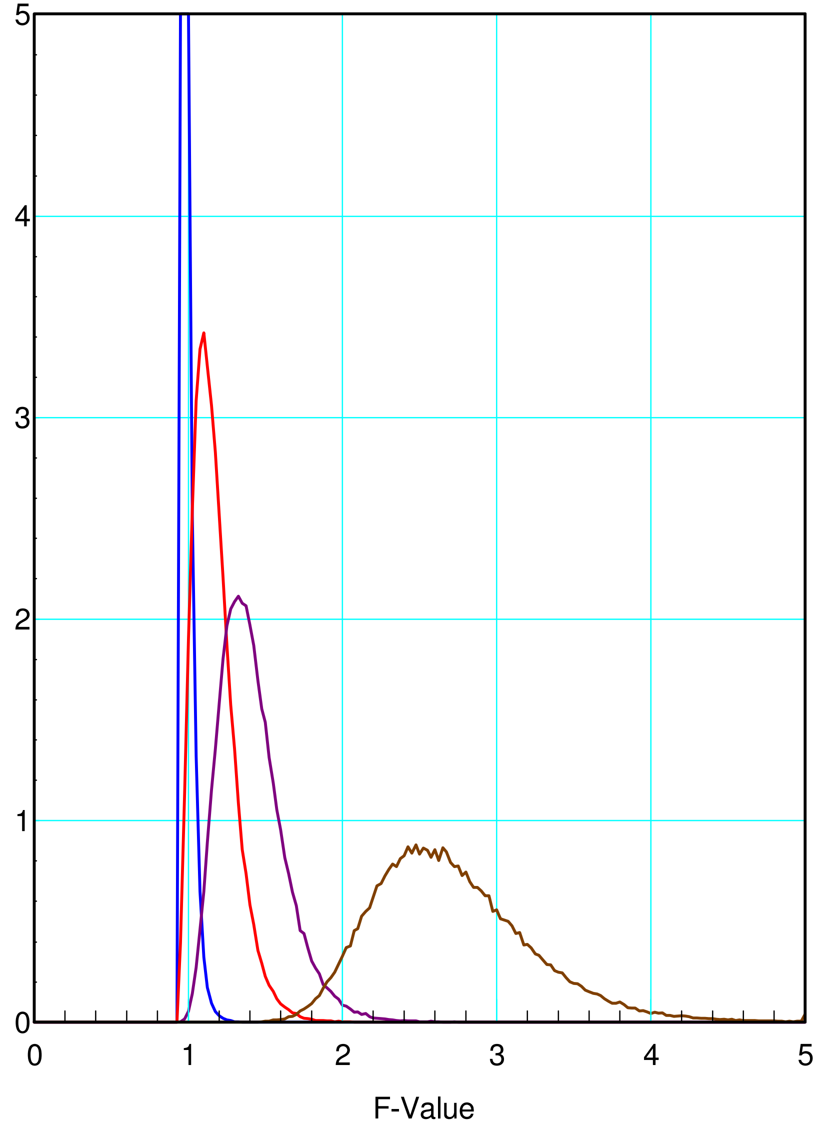
is an F-distribution with

- 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:

$H_0 : \quad \sigma_x^2 < \sigma_y^2$ or $H_1 : \quad \sigma_x^2 > \sigma_y^2$

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/

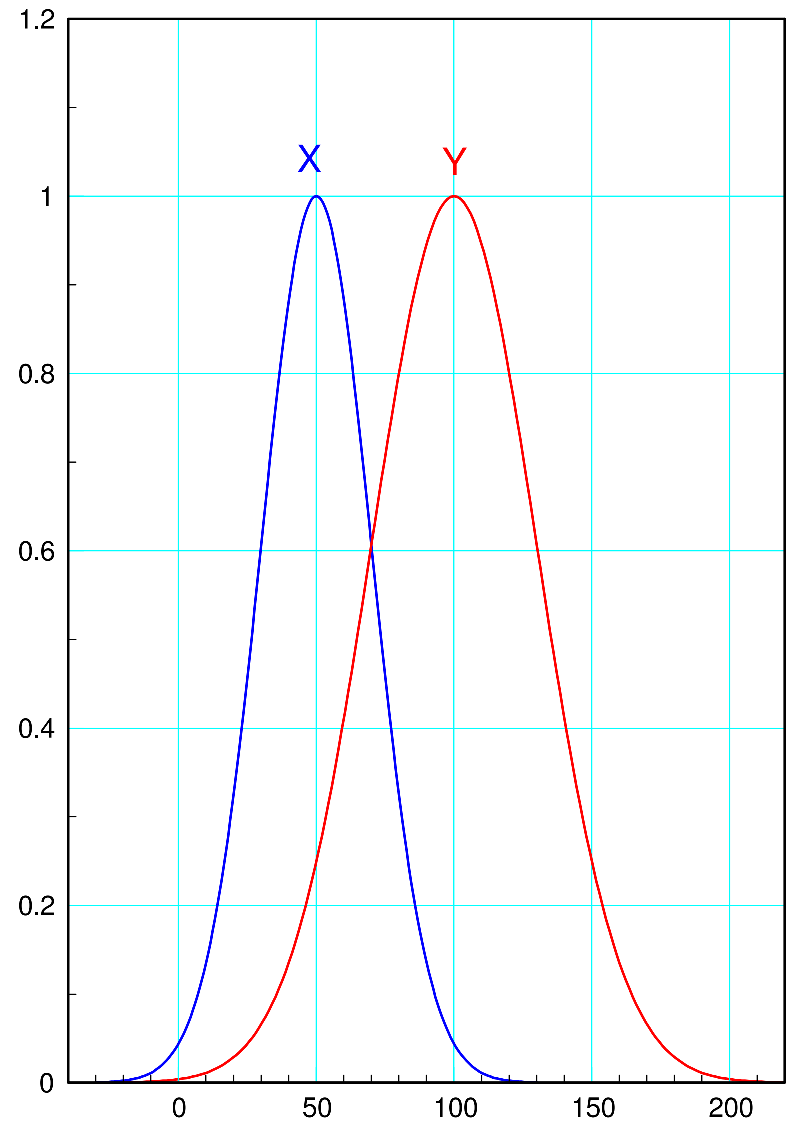
	m = 1	m = 2	m = 3	m = 4	m = 5	m = 10	m = 20	m = 40	m = INF
n = 1	39.86	49.5	53.59	55.83	57.24	60.2	61.74	62.53	63.33
n = 2	8.53	9	9.16	9.24	9.29	9.39	9.44	9.47	9.49
n = 3	5.54	5.46	5.39	5.34	5.31	5.23	5.18	5.16	5.13
n =4	4.55	4.33	4.19	4.11	4.05	3.92	3.84	3.8	3.76
n =5	4.06	3.78	3.62	3.52	3.45	3.3	3.21	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:

Step 1: Collect Data

- 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
```

Step 2: Compute the variance and the F-value

- 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

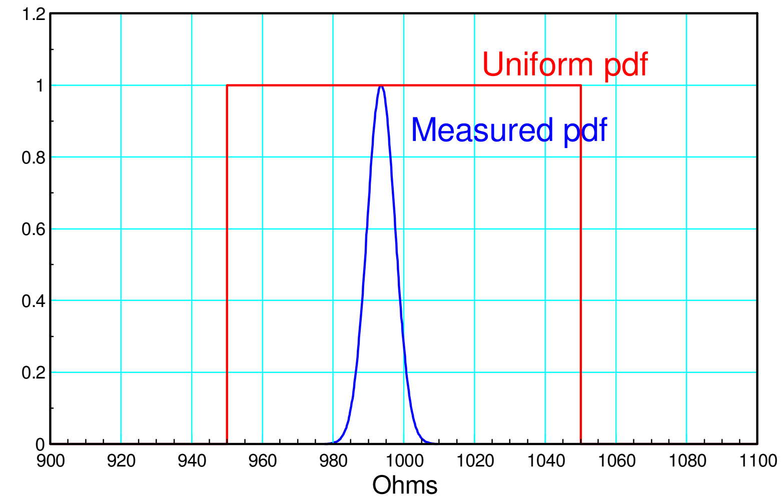
Example 2: 5% Resistors

- Do 5% Resistors have a Uniform Distribution?
- H0: The resistance has a uniform distribution over the range of (95%, 105%)

Step 1: Collect Data

- 56 resistors are measured

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

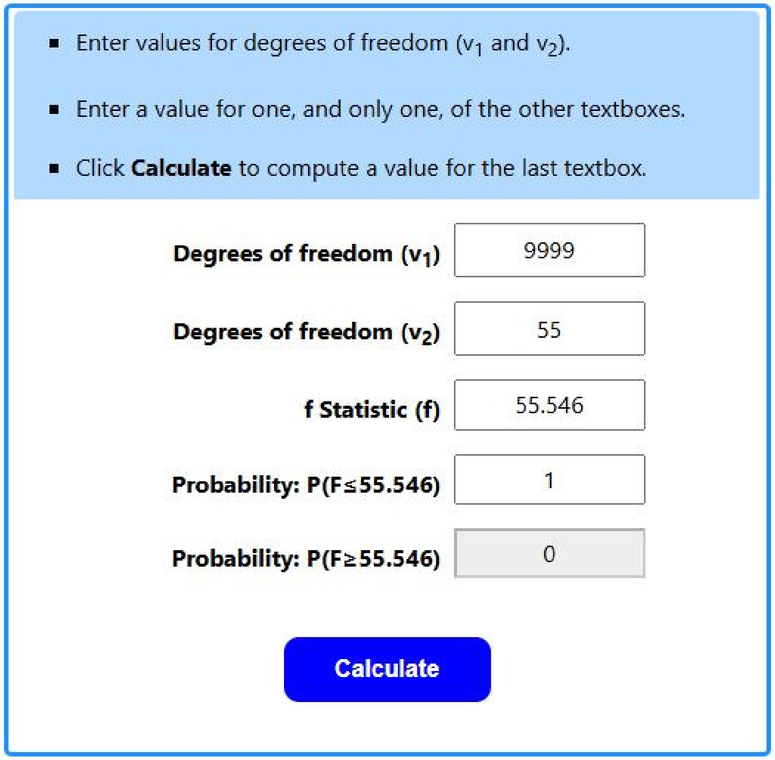


Step 2: Compute the F-value

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

Step 3: Convert to a probability

- Use StatTrek
- p = 1.0000 (rounded)



I'm almost 100% certain that these resistors do *not* have a uniform distributoin

Example 3: 3904 Transistors

Four shipments of 3904 NPN transtrors were 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 the same variance

- Common supplier, or
- Common production run
- Similar prodution process

3904 Transistors (cont'd)

- Compute the F-values
- Convert to a probability using an F-table

Result:

- All four shipments have different variances
- They probably have different manufacturers or different production runs

Note:

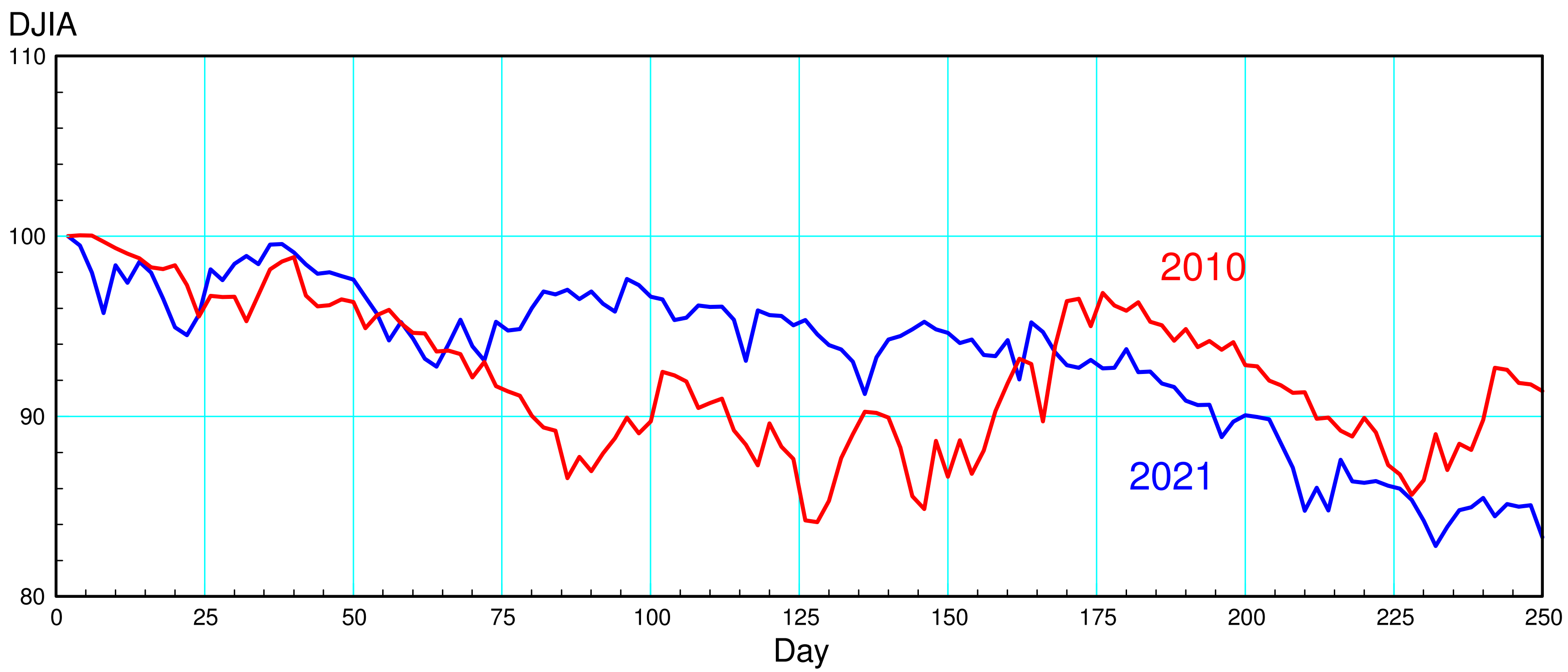
- All are within specs
- 100 < hfe < 300

```
>> 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)
```

Example 4: Is the Stock Market Going to Crash?

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



Stock Market: 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

- numerator = 249 dof
- denominator = 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

Stock Market (take 2):

- 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

Stock Market (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
  - Do a linear curve fit and find the residual (deviation from a line)

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...

So, is the stock market heading towards a crash?

F-Test and Regression Analysis

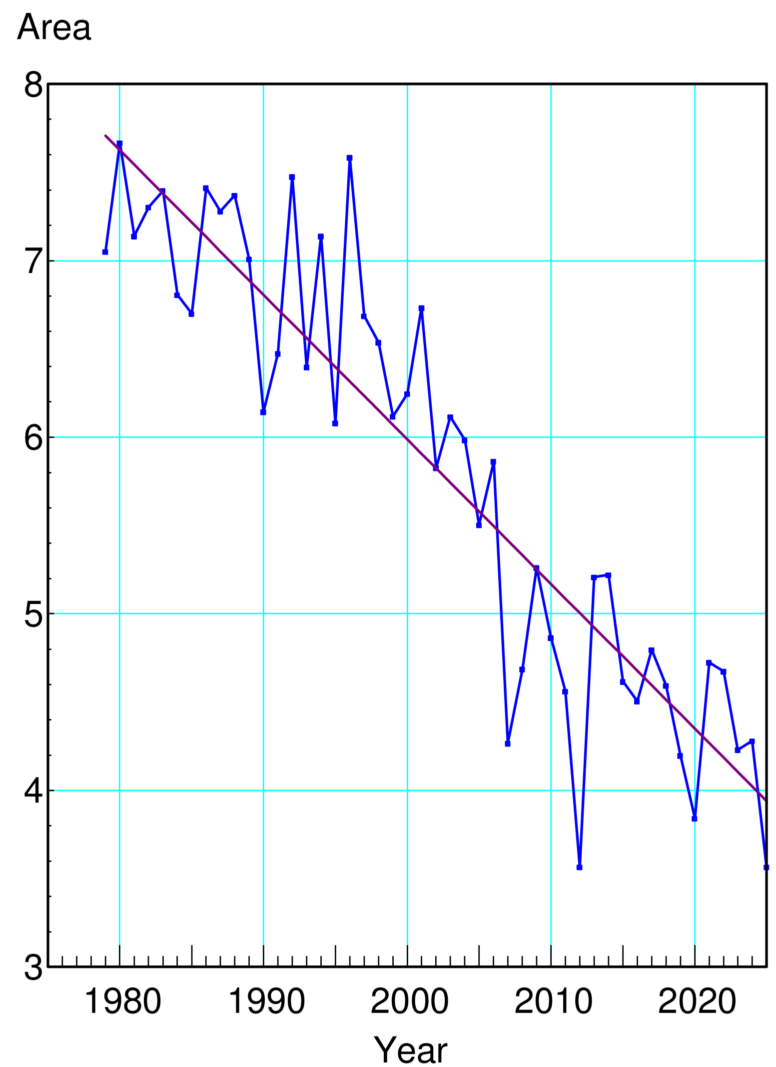
F-Tests can also be used to determine how significant each term is in a least-squares curve fit

Example: Curve fit Arctic Sea Ice to a line

$$y = ax + b$$

Is the term 'ax' significant?

Is the term 'b' significant?



Sea Ice: Linear Term

To check if the linear term is significant, compare two curve fits

$$y_1 = b$$

$$y_2 = ax + b$$

Find the variance of the residual for each curve fit

$$s_1^2 = \frac{1}{n-1} \sum (y_i - y_1)^2$$

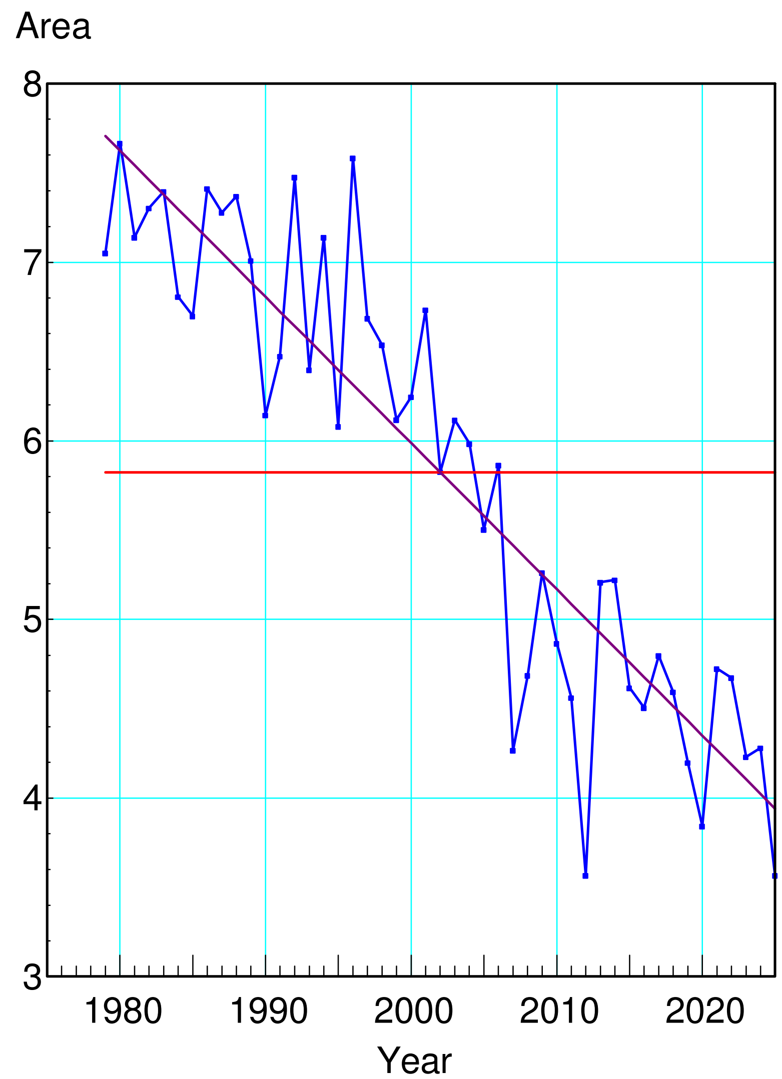
n-1 degrees of freedom

$$s_2^2 = \frac{1}{n-2} \sum (y_i - y_2)^2$$

n-2 degrees of freedom

The F-value is then

$$F = \frac{s_1^2}{s_2^2}$$



Linear Term in Matlab

The resulting F-score is

- F = 5.7990

This corresponds to p =1.000

- I'm almost 100% certain that the linear term is significant
- There is a trend with the data

Matlab Code

```
>> x = Data(:,1);
>> y = Data(:,2);
>> n = length(x)
n =    47

# y = b curve fit
>> B0 = [x.^0];
>> A0 = inv(B0'*B0)*B0'*y
b =   5.8230

# y = ax + b curve fit
>> 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)
```

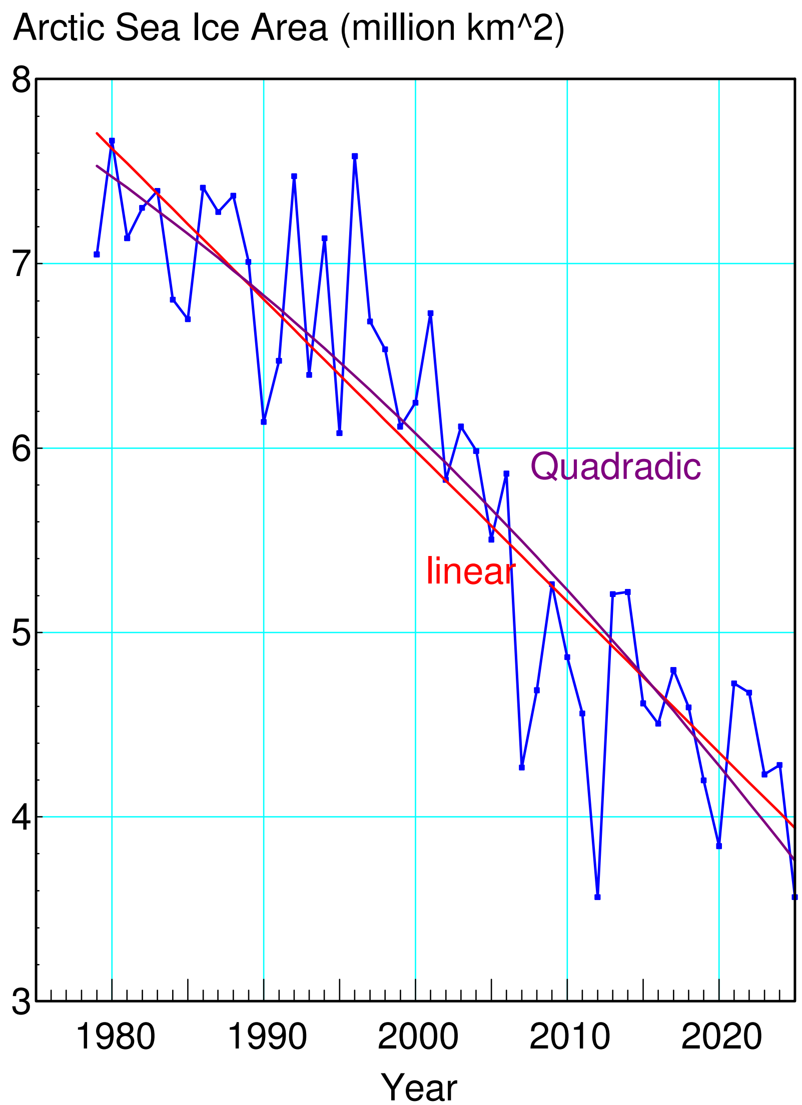
Sea Ice: Quadratic Term

- Is the rate of melt increasing year by year?
- Is a quadratic term significant?

Test with a F-text

- H0: $y = a + bx$
- H1: $y = a + bx + cx^2$

Does c reduce the variance in the residual?



Quadratic Term in Code

Degrees of freedom change

- n-2 for the numerator
- n-3 for the denominator

The resulting F-score is

- F = 1.0066

This corresponds p = 0.508

- It's 50/50 whether the quadratic term is fitting data or noise

The data does not support the notion that Arctic sea ice is melting at a faster rate each year

- The data only supports a linear model

Matlab Code

```
>> # H0: Linear curve fit
>> B0 = [x.^0, x];
>> A0 = inv(B0'*B0)*B0'*y
a -8.1872e-002
b  1.6973e+002

>> # H1: Quadratic curve fit
>> B1 = [x.^0, x, x.^2];
>> A1 = inv(B1'*B1)*B1'*y
a -1.8924e+003
b  1.9782e+000
c -5.1452e-004

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

F2 =  1.0066
(p = 0.508)
```

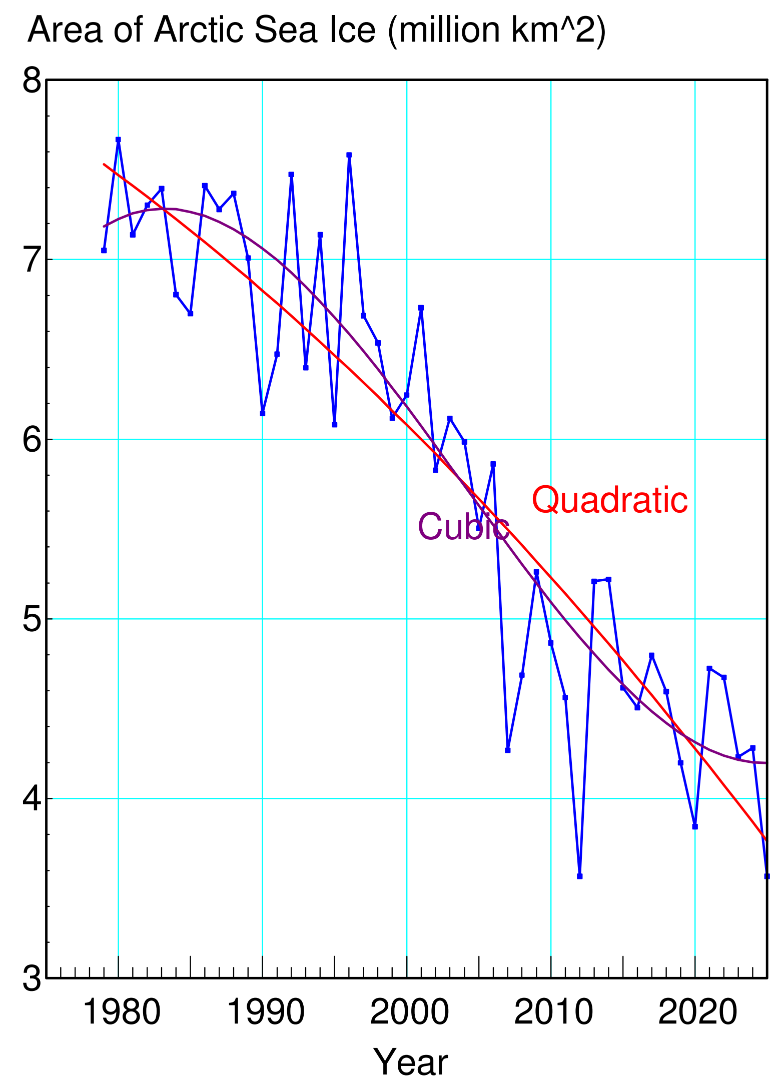
Arctic Sea Ice: Cubic Term

- Is the sea ice melting and an increasing increasing rate?
- Does a cubit term fit data or noise?

Test with a F-text

- H0: $y = a + bx + cx^2$
- H1: $y = a + bx + cx^2 + dx^3$

Does d reduce the variance in the residual?



Cubic Term: Matlab Code

Degrees of freedom change

- n-3 for the numerator
- n-4 for the denominator

The resulting F-score is

- F = 1.1173

This corresponds p = 0.641

- 64% chance the cubic term is fitting data rather than noise noise

The data does not support the notion that change in Arctic sea ice is changing at a faster rate each year

Matlab Code

```
>> # H0: Quadratic curve fit
>> B0 = [x.^0, x, x.^2];
>> A0 = inv(B0'*B0)*B0'*y
a -1.8924e+003
b  1.9782e+000
c -5.1452e-004

>> # H1: Cubic curve fit
>> B3 = [x.^0, x, x.^2, x.^3];
>> A3 = inv(B3'*B3)*B3'*y

a -6.8554e+005
b  1.0265e+003
c -5.1227e-001
d  8.5207e-005

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

F3 =  1.1173
(p = 0.641)
```

Summary

The F-test is used to compare the variance of two populations.

F-tests are useful when trying to determine

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

- 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: $P(F \leq 5.5182)$	0.990
Probability: $P(F \geq 5.5182)$	0.010

Calculate