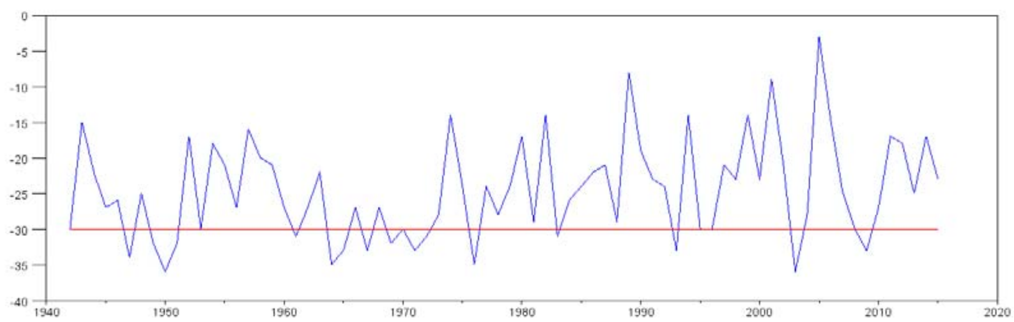
ECE 376 - Fun with Statistics and Curve Fitting

t-Test: Temperature in Fargo

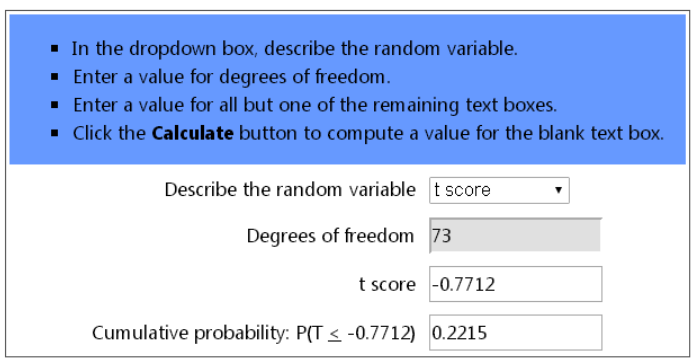
The low for the month for the month of January in Fargo, ND since 1943 is as follows:



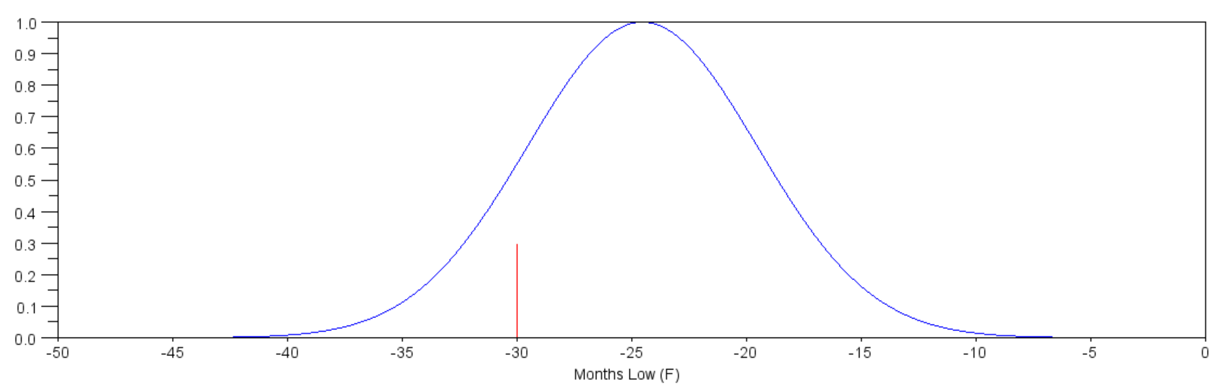
http://weather-warehouse.com/WeatherHistory/PastWeatherData_FargoHectorIntlArpt_Fargo_ND_January.html

1) What is the probability that it will get colder than -30F in January 2017

mean (x)	standard dev (s)	npt	(x - (-30F)) / s
-24.505F	7.0792F	74	0.7712



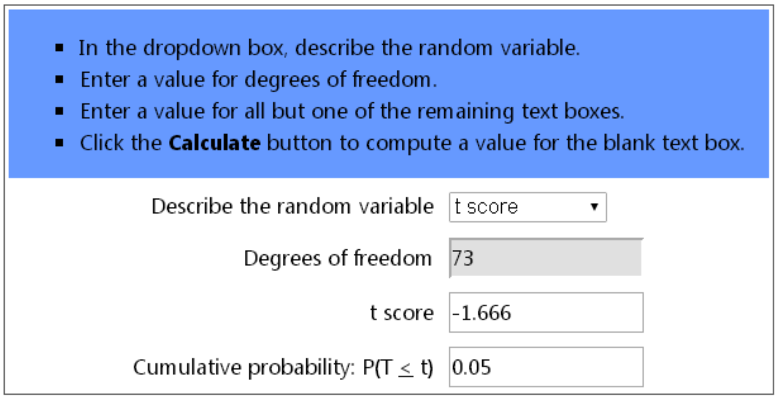
StatTrek.com



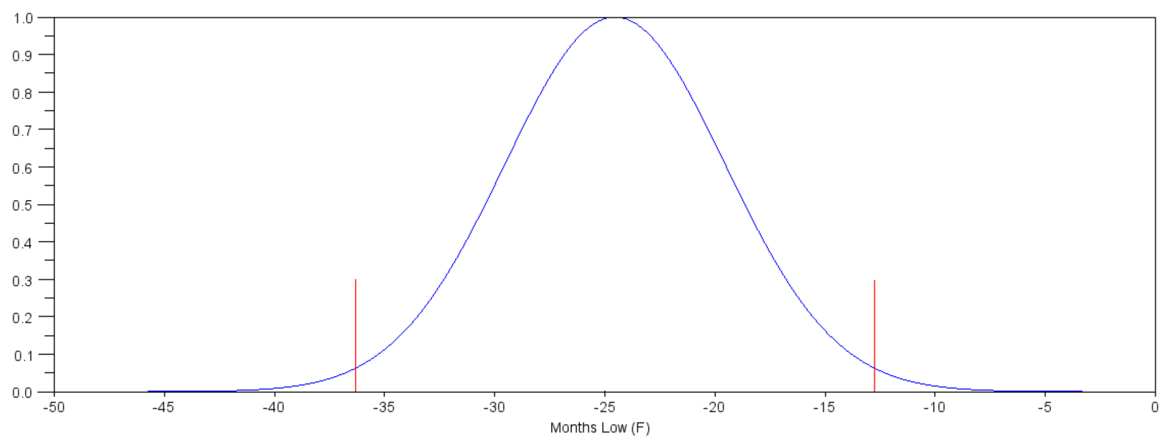
There is a 22.15% chance it will get colder than -30F this coming January.

What is the 90% confidence interval for the month's low?

For 5% tails, you need to go out 1.666 standard deviations




```
>> x + 1.666*s
  -12.7466

>> x - 1.666*s
  -36.3345
```



It is 90% likely that the month's low will be in the range of (-36.33F < low < -12.75F)

Comparison of Means:

If you want to compare two populations, create a new variable, W

W = sample A - sample B

W will have statistics of

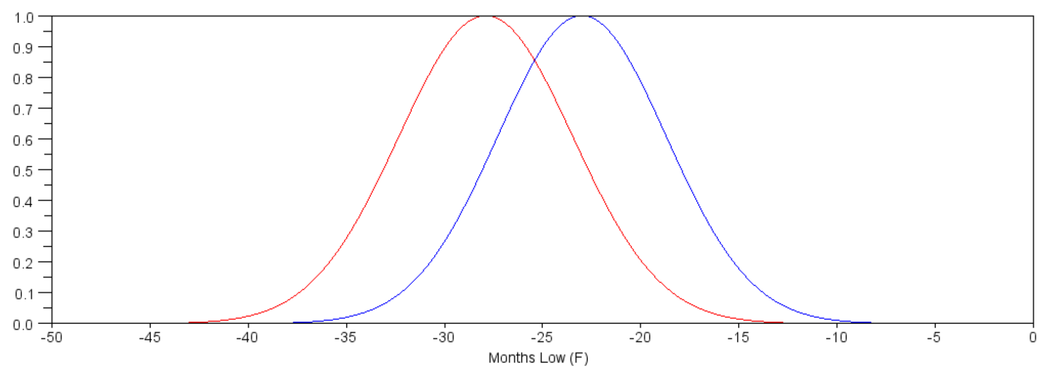
mean(W) = mean(A) - mean(B)

$$\text{std dev(W)} = \sqrt{\left(\frac{s_a^2}{n_a}\right) + \left(\frac{s_b^2}{n_b}\right)}$$

degrees of freedom = min(na, nb) - 1

Comparison of Means: Weather in Fargo

What is the probability that the last 10 years (2006-2015) were warmer the first 10 years (1943 - 1952)



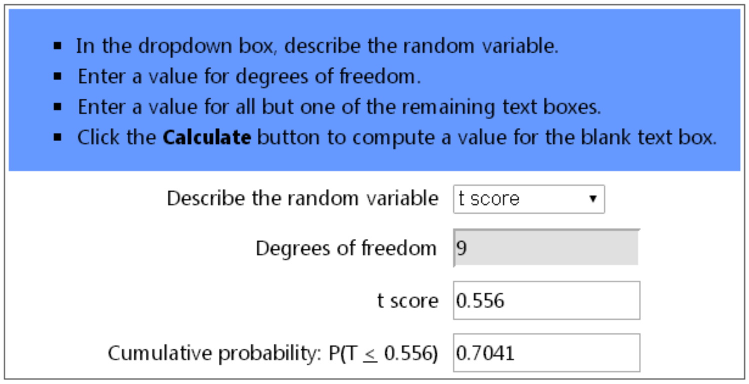
Low for January over 1943 - 1952 (red) and 2006-2015 (blue)

	1943-1952	2006-2015	Diff
mean	-27.9F	-23.00F	+1.54F
st dev	6.27F	6.0919F	2.77F
npt	10	10	10

```
->sa = std(DATA(1:10))
    6.091889

-->sb = std(DATA(65:74))
    6.2795966

-->sqrt((sa^2)/10 + (sb^2)/10)
    2.7666667
```

$$t = \left(\frac{1.54F}{2.77F}\right) = 0.556$$

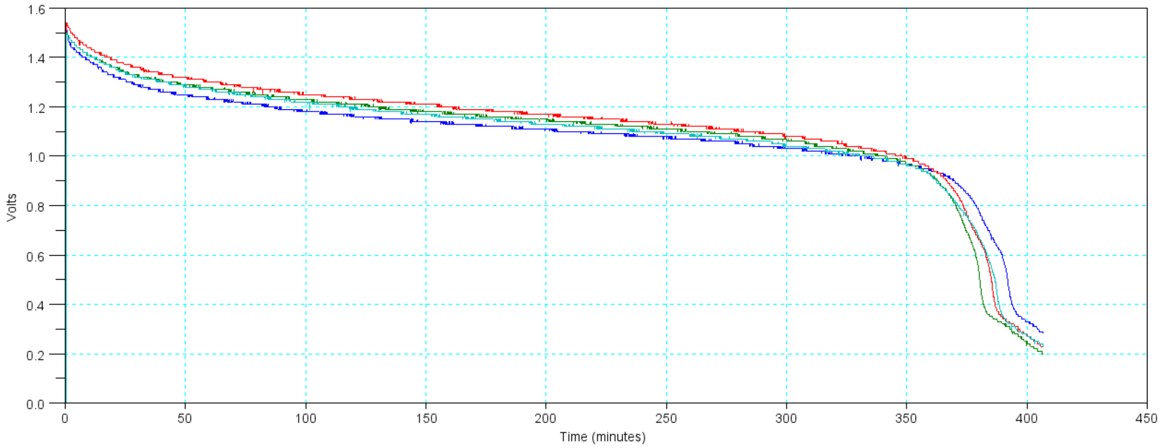
The last 10 years were 0.556 standard deviations warmer than 65 years ago. This works out to a probability level of 0.7041:



It is 70.41% likely that January has been warmer over the last 10 years vs. 65 years ago.

t-Test: Energy in a AA Battery

The voltage across a AA battery driving a 10-Ohm resistor was measured



Determine the 90% confidence interval for the energy in any given battery.

Solution: Turn the data in to a number. The instantaneous power in Watts is

$$P = \frac{V^2}{R}$$

The energy in Joules is the integral of the power. This results in the energy in the four batteries being

```
Joules =     2937.5546    3063.8801    3204.829    3019.9991
```

Take the mean and standard deviation:

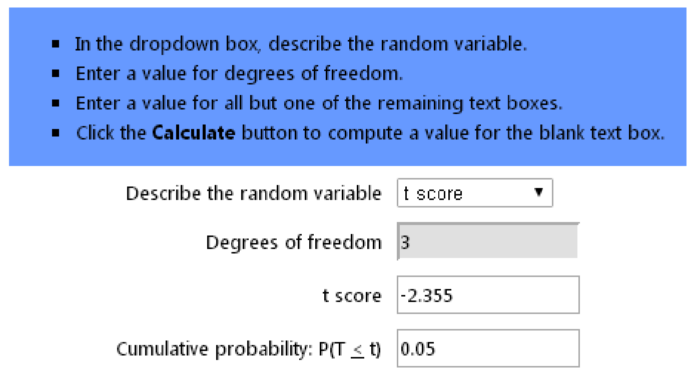
```
-->x = mean(Joules)

    3056.5657

-->s = stdev(Joules)

    111.85742
```

Use a t-table (StatTrek.com) to find our how many deviations you have to go to capture 90% of the area,

From StatTrek.com,

- with 3 degrees of freedom (sample size minus one) and
- 5% tails, (two tails of 5% each leaves 90% in the middle)

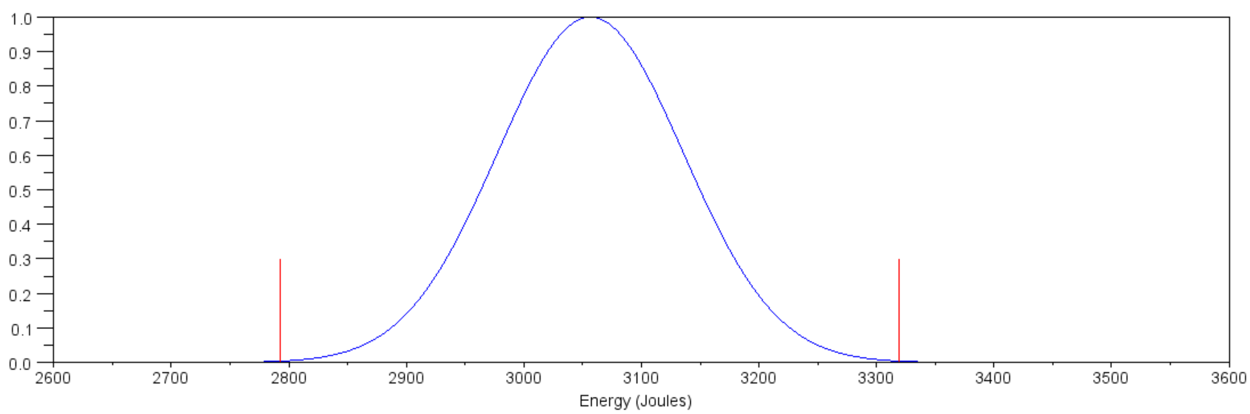
you need to go out +/- 2.355 deviations



Therefore, I can be 90% certain that the total energy in any given type D battery will be

$$\bar{x} - 2.355s < Joules < \bar{x} + 2.355s$$

2793 < Joules < 3319 (p = 0.9)



90% Confidence Interval for the Total Energy in a type-D AA battery

Random Number Generator with a PIC:

To generate a random number from 1 to 6, the following C code is used:

```
while(RB0)  DIE = (DIE + 1) % 6;
```

Is this a uniform distribution?

Design of Experiment: Toss the die 3014 times. The resulting frequency vs. number is

Roll	1	2	3	4	5	6
Frequency	498	482	495	538	465	536

t-Test:

If this is a fair die, the mean should be 3.5

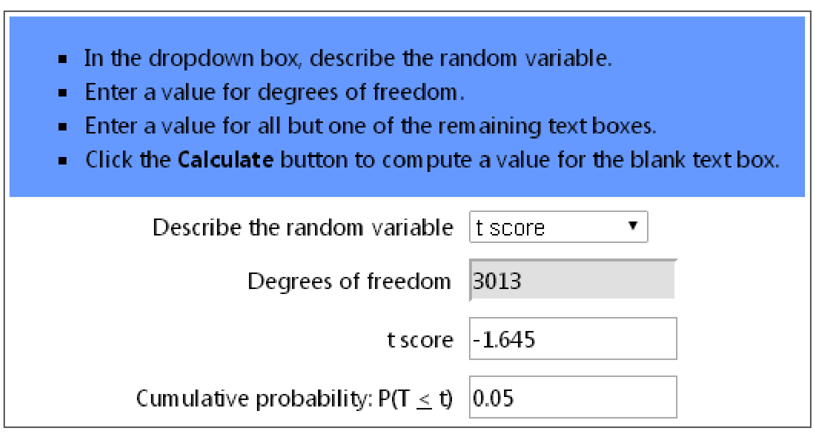
The actual data for this die is

```
-->x = mean(Die)

    3.5301924

-->s = std(Die)

    1.7137263

>npt = length(Die)

    3014.
```

With a sample size of 3014, the population's mean drops as the square root of the sample size:

```
-->sp = s / sqrt(npt)

    0.0312155
```

The 90% confidence interval for a sample size of 3014 (3013 degrees of freedom) has 5% tails. To get 5% tails, you need to go out 1.645 standard deviations:



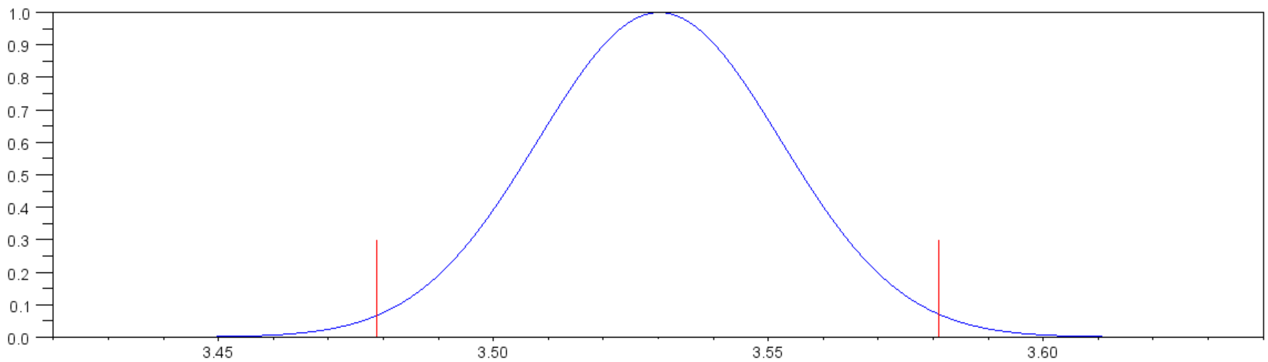
This makes the 90% confidence interval for the mean of the die rolls

$$\bar{x} - 1.645 \cdot \frac{s}{\sqrt{3014}} < mean < \bar{x} + 1.645 \cdot \frac{s}{\sqrt{3014}}$$

3.478843 < mean < 3.5815419

3.50 is in this area, so I cannot call this a loaded die with a probability of 90% (no result).

Note: If you make the sample size large enough, you eventually will be able to see tiny differences in the means.



90% Confidence Interval for the Mean of the Random Number Generator

Note: This also tells you how large your sample size has to be to detect that this is a loaded die with a probability of 0.9.

3.50 differs from the sample mean by

- x - 3.5 = 0.0301924

For the tail to be 5%, you need to go out 1.645 standard deviations

- $0.0301924 = 1.645 \cdot \left(\frac{s}{\sqrt{n}} \right)$

The sample standard deviation is 1.7137263. Solving for n

$$0.0301924 = 1.645 \cdot \left(\frac{1.7137}{\sqrt{n}} \right)$$

$$n = 8717$$

I should be able to detect that this is a loaded die with a probability of 0.9 if I roll the dice 8717 times (or more)

F-Test

F-Tests test the standard deviation. If this is a fair die, the standard deviation of the sample should match the standard deviation of a fair die.

If this were a fair die, the statistics should be

- mean = 3.50
- standard deviation = 1.7078

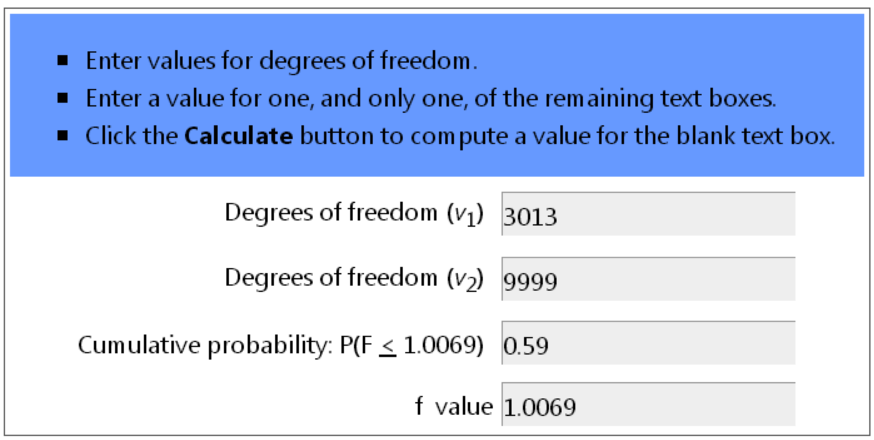
For the 3014 die rolls, the statistics are

- mean = 3.5301924
- standard deviation = 1.7137263

The F-test takes the ratio of the two standard deviations:

$$F = \left(\frac{1.7137}{1.7078}\right)^2 = 1.0069523$$

From an F-table



There's a 59% chance this distribution is not uniform

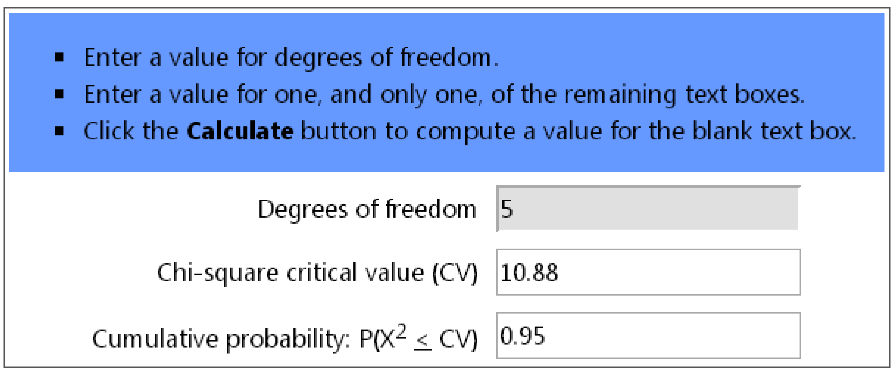
Chi-Squred Test

Chi-Squared Tests test the distribution. If this is a fair die, each number should occure 1/6th of the time

To see if this is the case, set up a table:

Roll	Actual (N)	Expected (np)	$\chi^2 : \left(\frac{(N-np)^2}{np}\right)$
1	498	517.3	0.72
2	482	517.3	2.41
3	495	517.3	0.96
4	538	517.3	0.83
5	465	517.3	5.29
6	536	517.3	0.68
total			10.88

Use a Chi-Squared Table (from StatTrek.com) with 5 degrees of freedom (there are 6 bins)

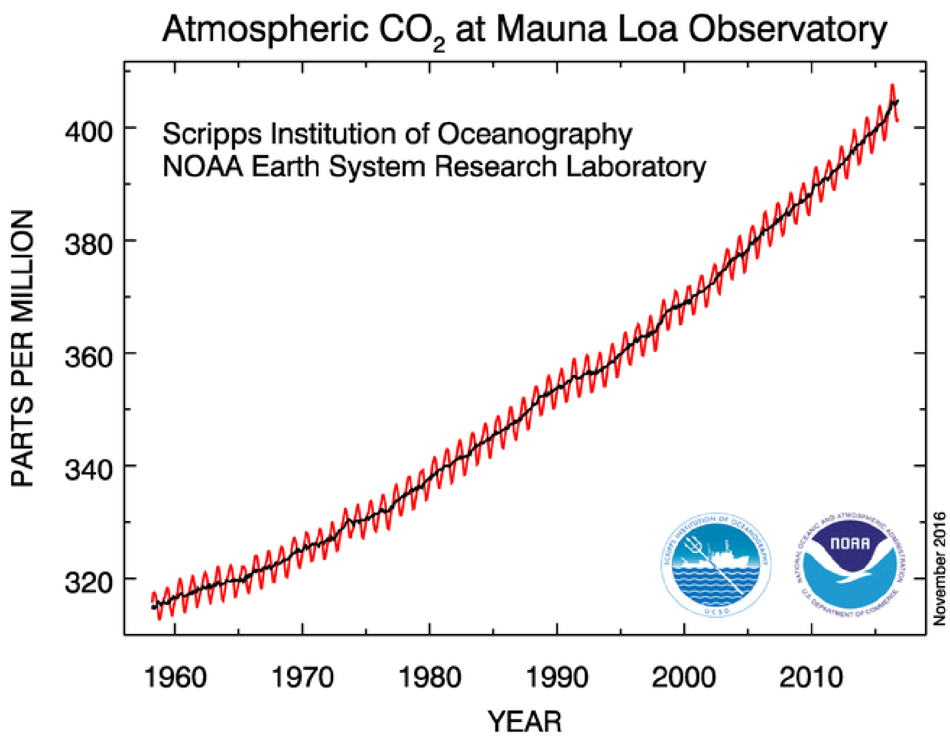


Based upon this data, I'm 95% certain this is not a uniform distribution.

There's a 5% chance that I just got unlucky and some numbers came up too often. There's a 95% chance that this program doesn't create truely random numbers. I wouldn't use this random number generator in Vegas.

Comparison of Means: CO2 Levels

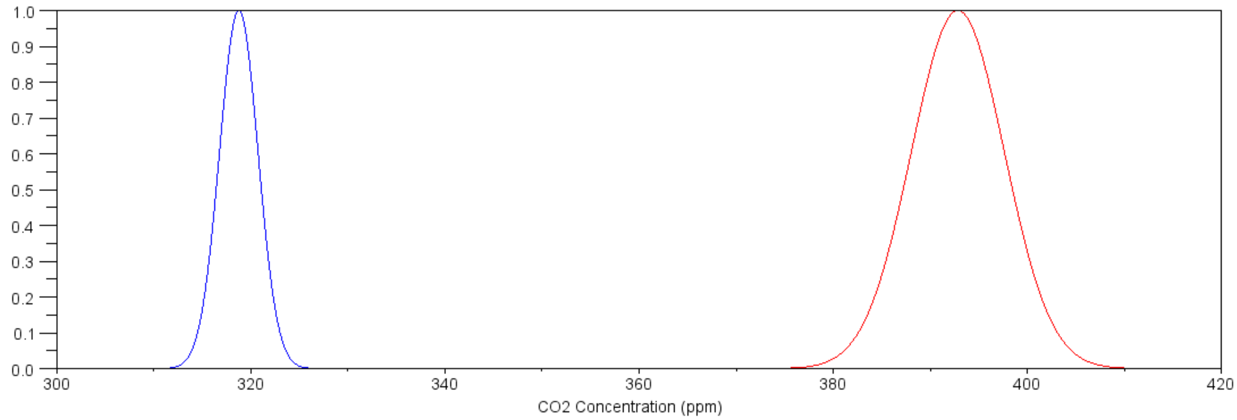
The CO2 level in the atmosphere is recorded at Mauna Loa Observatory:



http://www.esrl.noaa.gov/gmd/ccgg/trends/full.html -
data available at ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_mm_mlo.txt

What is the probability that CO2 levels are going up? (The probability that the CO2 levels from 2007-2016 are higher than the CO2 levels for 1961 - 1970?)

	1960 - 1969	2007-2016	difference
mean	318.76	392.83	74.07
st dev	2.8	6.76	2.39
npt	120	120	

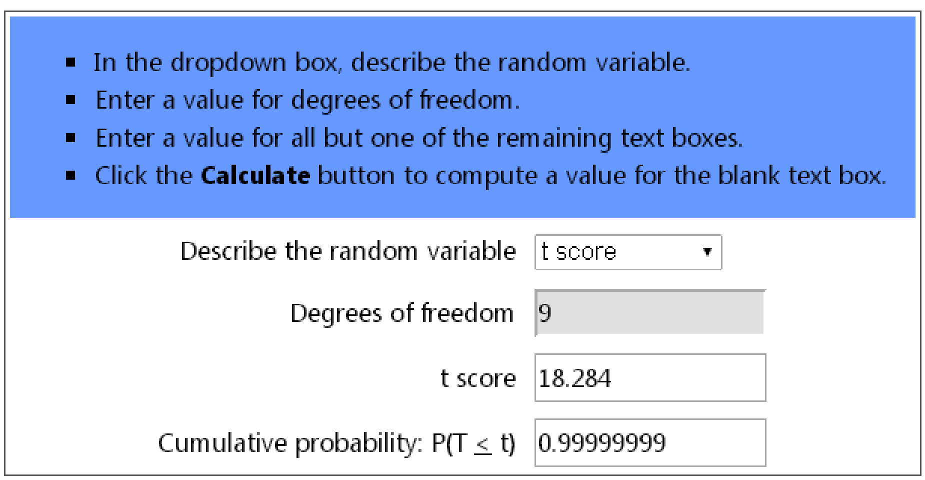


probability distribution of CO2 levels from 1960-1969 (blue) and 2007-2016 (red)

The difference in means is

$$t = \left(\frac{74.07}{2.31}\right) = 32.01 \text{ standard deviations.}$$

A probability of 0.99999999 corresponds to a t-value of 18.284 standard deviations



It is almost certain that the CO2 levels are going up.

There is less than a 0.00000001% chance that the variation in CO2 levels is due to chance.

Chi-Squared Test:

You can check if data is consistent with a given distribution with a Chi-Squared test.

i) Place the data into N bins (meaning N-1 degrees of freedom)

ii) Compare the actual number of events in each bin vs. the expected number based upon the probability

iii) Sum up the chi-squared difference

$$\chi^2 = \sum \left(\frac{(N_i - np_i)^2}{np_i} \right)$$

iv) Check this number on a Chi-Squared table with n-1 degrees of freedom.

Chi-Squared Test: World Weather

16 of the last 17 months have been record highs. What is the probability that this is due to chance?

With 135 years of data, you would expect the probability that any given month being:

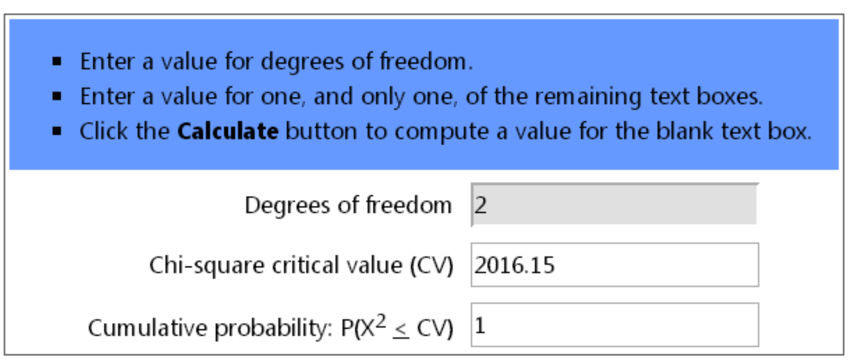
- Record high: $p = 1/135$
- Record low: $p = 1/135$
- Neither record high or low: $p = 133/135$

16 of the last 17 months have been record highs. What is the probability that this is due to chance? (hint: do a Chi-squared test).

	p	np	actual	$\chi^2 = \left(\frac{N-np}{np}\right)$
Record High	1/135	0.1259	16	2,001.48
no record	133/135	16.7481	1	14.54
Record Low	1/135	0.1259	0	0.13
Total				2,016.15

From a Chi-Squared table with 2 degrees of freedom, p = 1 (or 0.9999999999999)

Based upon this data we are almost certain that this is not due to random chance



Sidelight: To be 99.999% certain, you would need a Chi-Square valueof 24. We got 2016. Something is happening.

Curve Fitting:

Given a set of data (x, y), you can come up with a calibration curve fit of the form

$$y = ax^2 + bx + c$$

i) Place in matrix form

$$y = \begin{bmatrix} x^2 & x & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$Y = XA$$

ii) Solve

$$A = (X^T X)^{-1} X^T Y$$

Curve Fitting and Global Sea Ice: http://nsidc.org/

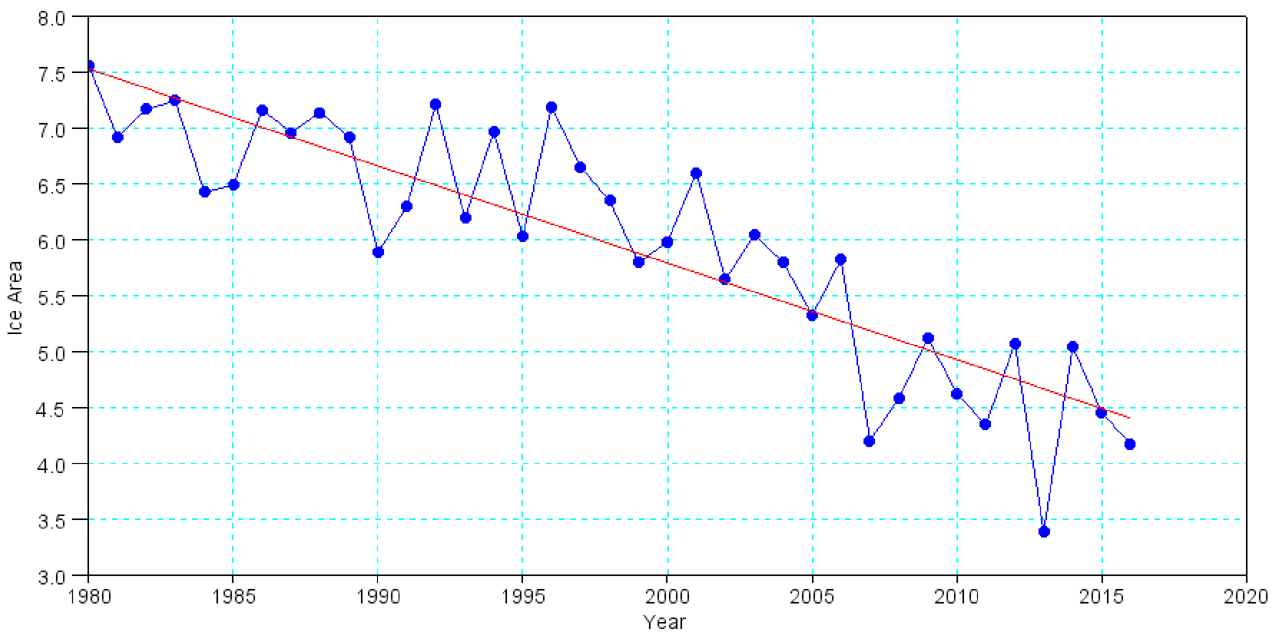
The area covered by sea ice is recorded by the National Snow and Ice Data Center:

Plot the sea ice level on September 15 from 1986 - 2016 (approximately the minimum sea ice level)

Approximate this data with a line

$$Area \approx ay + b$$

```
-->year = DATA(:,1);
-->ICE = DATA(:,2);
-->plot(year,ICE)

-->X = [year, year.^0];
-->A = inv(X'*X)*X'*ICE

a  - 0.0868113
b    179.4173

-->plot(year,ICE,'b.-',year,X*A,'r')
-->xlabel('Year');
-->ylabel('Ice Area');
```



http://nsidc.org/arcticseaicenews/charctic-interactive-sea-ice-graph/

From this curve fit, when do you expect the Arctic to be ice free?

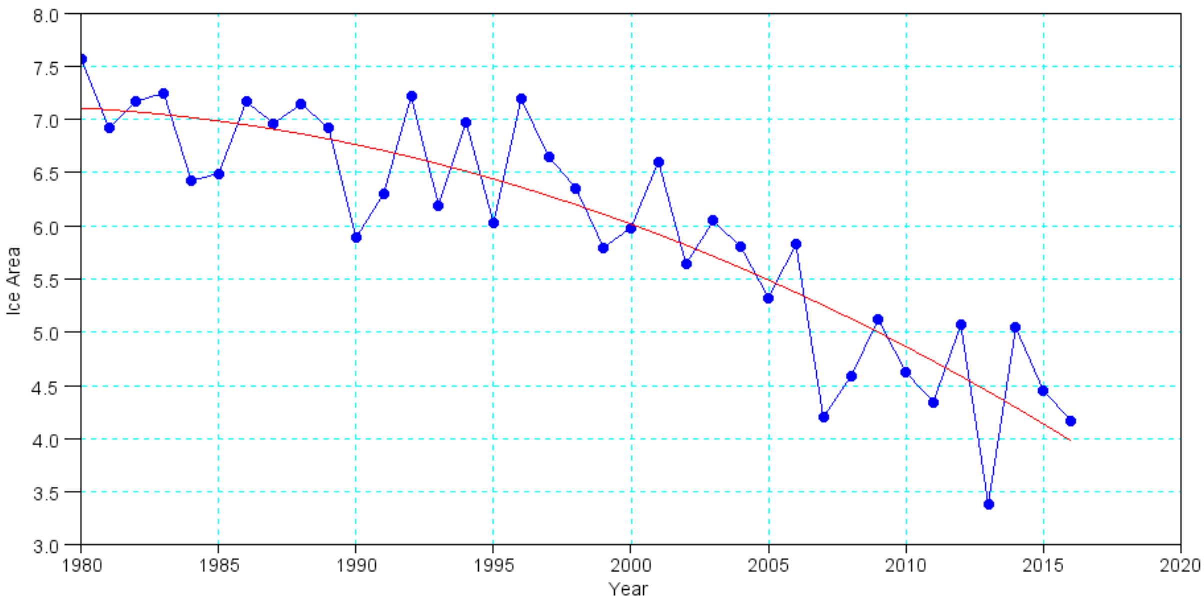
```
-->roots([A(1),A(2)])

    2066.7509
```

At this rate, the Arctic will be ice free in 2066 (49 years from now)

Sidelight: This assume a linear model. If there is positive feedback (less ice means more exposed water which causes more melting), a parabolic curve fit might be a better model.

$$Area \approx ay^2 + by + c$$

```
-->X = [year.^2  year  year.^0];

-->A = inv(X'*X)*X'*ICE

a  - 0.0020153
b    7.9664276
c  - 7865.5386

-->plot(year,ICE,'b.-',year,X*A,'r')
-->xgrid(4)
-->xlabel('Year');
-->ylabel('Ice Area');
```



http://nsidc.org/arcticseaicenews/charctic-interactive-sea-ice-graph/

With a parabolic model the zero crossing (a.k.a. the roots to the polynomial) are:

```
-->roots([A(1),A(2),A(3)])

    2035.9547
    1916.9697
```

If there's positive feedback, the Arctic might be ice free in 2035 (18 years from now)

Curve Fitting CO2 Levels

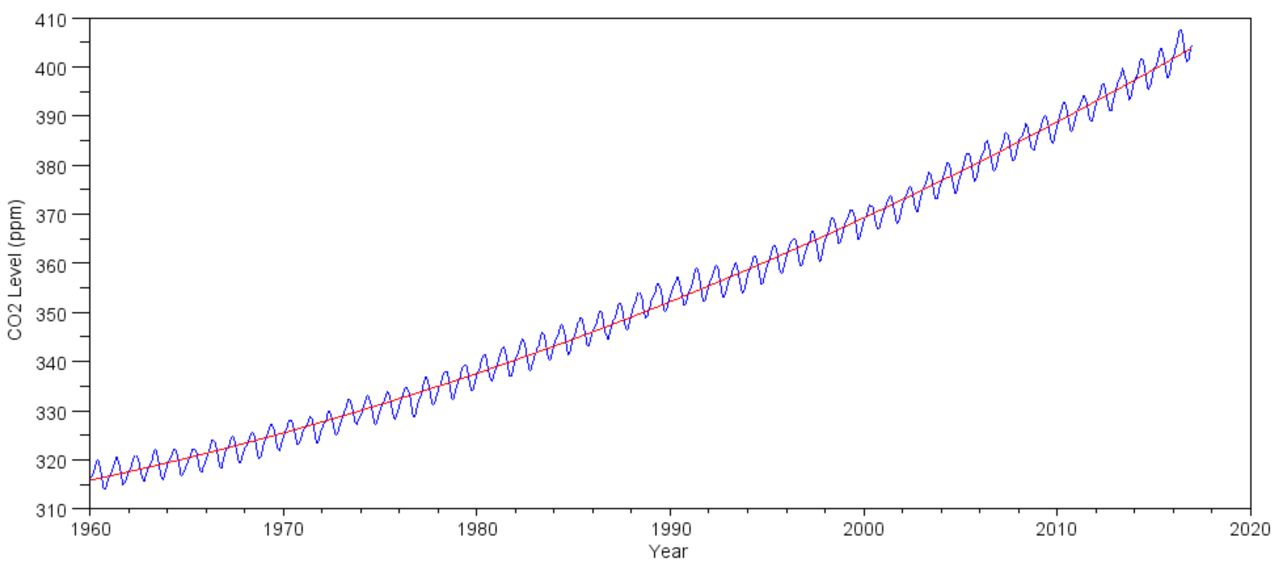
Determine a parabolic curve fit for this data in the form of

$$CO_2 \approx ay^2 + by + c$$

where 'y' is the year.

In Matlab: data is pasted into Matlab array DATA:

```
-->year = DATA(:,3);
-->CO2 = DATA(:,5);
-->X = [year.^2  year  year.^0];
-->A = inv(X'*X)*X'*CO2

a     0.0123653
b   - 47.629524
c     46167.283

-->plot(year,CO2,year,X*A);
-->xlabel('Year');
-->ylabel('CO2 Level (ppm)');
```



http://www.esrl.noaa.gov/gmd/ccgg/trends/full.html -

data available at ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_mm_mlo.txt

From this data, when do you predict that we will hit 2000 ppm of CO2? (the same as what was observed during the Permian extinction)

You can solve this as roots to a polynomial:

The year we'll hit 400ppm (2015 on the graph)

$$400 = ay^2 + by + c$$

$$0 = ay^2 + by + c - 400$$

```
>roots([A(1),A(2),A(3)-400])

    2015.1988              check:  2015
    1836.6808
```

The year we'll hit 700ppm (the level some models predict will trigger the Gulf Stream stopping 300 years later)

$$700 = ay^2 + by + c$$

$$0 = ay^2 + by + c - 700$$

```
>roots([A(1),A(2),A(3)-700])

    2105.4633              2105:  700 ppm
    1746.4164
```

Assuming this model is correct, we have 88 years to solve our energy problem. If not, people 300 years from now will have problems.

The year we'll hit 2000 ppm (the level that triggered the Permian extinction)

```
>roots([A(1),A(2),A(3)-2000])

    2296.5634              2296:  279 years from now
    1555.3162
```

According to this data and curve fit, we'll find out if 2000ppm triggers another mass extinction in 279 years.

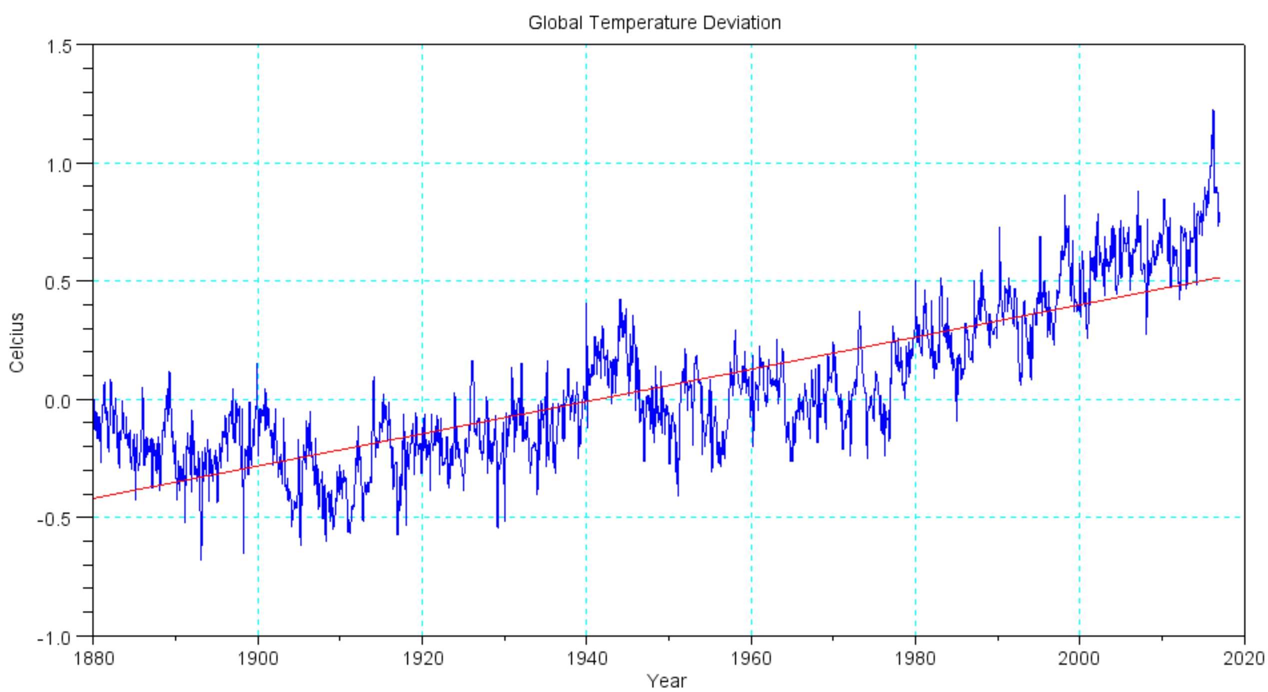
Curve Fitting Example: World Temperatures

NASA Goddard has been keep records since 1880 (135 years of data). Determine a least-squares curve fit for this data in the form of

$$\delta T \approx ay + b$$

In Matlab:

```
-->dT = DATA(:,2);
-->size(DATA)

    1644.    2.

-->year = 1880 + [1:1644]' / 12;
-->plot(year,dT)
-->X = [year, year.^0];
-->A = inv(X'*X)*X'*dT

a    0.0068204
b  -13.240968

-->plot(year,dT,year,X*A,'r')
-->xlabel('Year');
-->ylabel('Celcius');
-->title('Global Temperature Deviation');
```



https://www.ncdc.noaa.gov/cag/time-series/global/globe/land_ocean/p12/12/1880-2016.csv

Based upon this data, predict when we will see a 10 degree temperature increase if nothing changes.

$$\delta T = 10 = ay + b$$

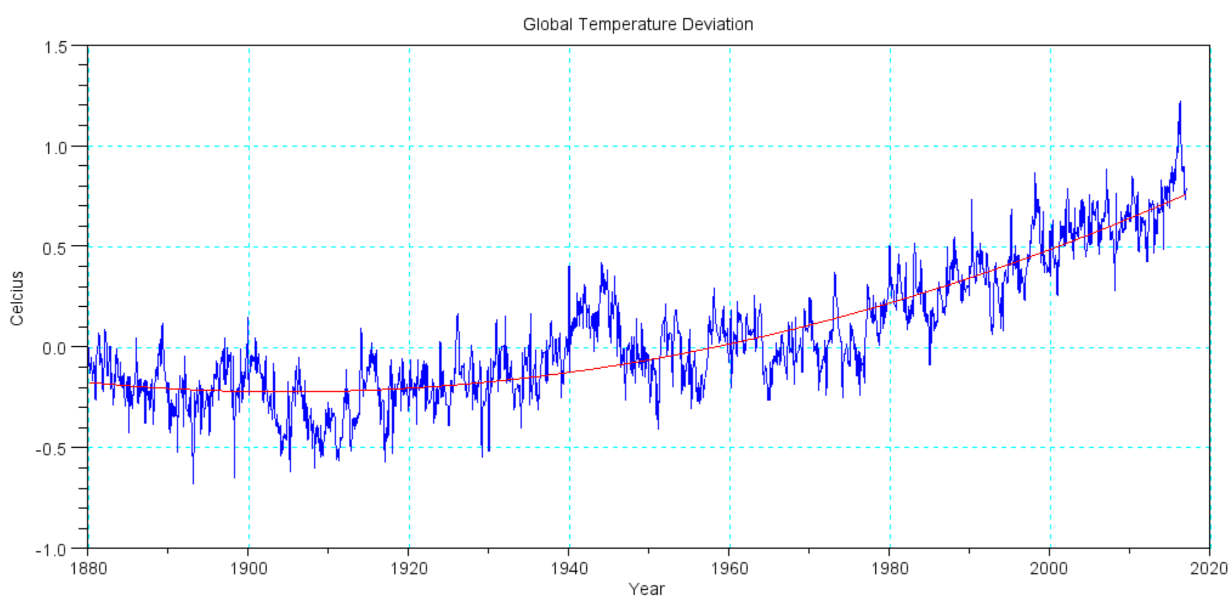
$$0 = ay + b - 10$$

```
-->roots([A(1),A(2)-1])

    2088.0067              1 degree rise (actually it was 2016)

-->roots([A(1),A(2)-2])

    2234.6264              2 degree rise

-->roots([A(1),A(2)-4])

    2527.8659              4 degree rise

-->roots([A(1),A(2)-10])

    3407.5842              10 degree rise
```

Repeat using a parabolic curve fit:

$$\delta T = ay^2 + by + c$$

```
-->X = [year.^2  year  year.^0];
-->A = inv(X'*X)*X'*dT

a     0.0000778
b    -0.2965230
c   282.17592

-->plot(year,dT,year,X*A,'r')
-->xlabel('Year');
-->ylabel('Celcius');
-->title('Global Temperature Deviation');
```



What does a temperature rise of 10 degrees mean for the planet?

From http://globalwarming.berrens.nl/globalwarming.htm

One Degree: 2030   Summers like 2003 where a heat wave in France caused 10,000 deaths become the norm. Flows of the Po and Rhine river decrease. Crop production drops.

```
-->roots([A(1),A(2),A(3)-1])

    2030.045              1 degree rise by 2030
    1779.4164
```

Two Degrees: 2073.   Oceans absorb less CO2 (too hot) and soils start to release CO2. Vacations to the Mediterranean in the summer are just too hot. Crop failures in Africa and Central America cause mass migration. Coastal cities flood. 1/3rd of species face extinction.

```
-->roots([A(1),A(2),A(3)-2])

    2073.7005             2 degree rise by 2073
    1735.761
```

Three Degrees: 2108.   Crop failures in China cause the migration of more than 1 billion people. Collapse of equatorial governments.

```
-->roots([A(1),A(2),A(3)-3])
 ans  =

    2108.1954
    1701.266
```

Four Degrees: 2137.   Spain becomes a desert. Mass migration to Northern latitudes.   Rain forests burn up.

```
-->roots([A(1),A(2),A(3)-4])

    2137.6362             4 degree rise by 2137
    1671.8252
```

Six Degrees: 2187.   Ice caps are gone. Methane hydrates become unstable raising temperatures in a positive-feedback loop. Ocean circulation stops. Hydrogen sulfide producing bacteria flourish poisoning the air. The Ozone layer dissipates leaving the land sterilized with UV radiation.   End-Permian-like conditions make life nearly impossible.

```
-->roots([A(1),A(2),A(3)-6])

    2187.4659             6 degree rise by 2187
    1621.9955
```

Ten Degrees: 2267: See six degrees.   Only more-so.

```
-->roots([A(1),A(2),A(3)-10])

    2267.1222
    1542.3392
```

What does a temperature rise of 10 degrees mean for the planet?

From http://globalwarming.berrens.nl/globalwarming.htm

One Degree: 2030 Summers like 2003 where a heat wave in France caused 10,000 deaths become the norm. Flows of the Po and Rhine river decrease. Crop production drops.

```
-->roots([A(1),A(2),A(3)-1])

    2030.045              1 degree rise by 2030
    1779.4164
```

Two Degrees: 2073. Oceans absorb less CO_2 (too hot) and soils start to release CO_2. Vacations to the Mediterranean in the summer are just too hot. Crop failures in Africa and Central America cause mass migration. Coastal cities flood. 1/3rd of species face extinction.

```
-->roots([A(1),A(2),A(3)-2])

    2073.7005             2 degree rise by 2073
    1735.761
```

Three Degrees: 2108. Crop failures in China cause the migration of more than 1 billion people. Collapse of equatorial governments.

```
-->roots([A(1),A(2),A(3)-3])
 ans  =

    2108.1954
    1701.266
```

Four Degrees: 2137. Spain becomes a desert. Mass migration to Northern latitudes. Rain forests burn up.

```
-->roots([A(1),A(2),A(3)-4])

    2137.6362             4 degree rise by 2137
    1671.8252
```

Six Degrees: 2187. Ice caps are gone. Methane hydrates become unstable raising temperatures in a positive-feedback loop. Ocean circulation stops. Hydrogen sulfide producing bacteria flourish poisoning the air. The Ozone layer dissipates leaving the land sterilized with UV radiation. End-Permian-like conditions make life nearly impossible.

```
-->roots([A(1),A(2),A(3)-6])

    2187.4659             6 degree rise by 2187
    1621.9955
```

Ten Degrees: 2267: See six degrees. Only more-so.

```
-->roots([A(1),A(2),A(3)-10])

    2267.1222
    1542.3392
```