Student t Distribution

Probability and hypothesis testing

t-Test: Test of a mean.

- -What is the probability that April will break 90F this year?
- -What is the 90% confidence interval for the hottest it will get this coming April
- -Is April warmer than October?

Normal Distribution (a.k.a. Gaussian Distribution)

A t-test is a test of a mean. The heart of a t-test is the Central Limit Theorem. This states that all distributions converge to a Normal distribution (the bell-shaped curve you're probably familiar with). Furthermore, if you add a Normal distribution to a Normal distribution, the result is a Normal distribution.

A Normal distribution is defined by two parameters:

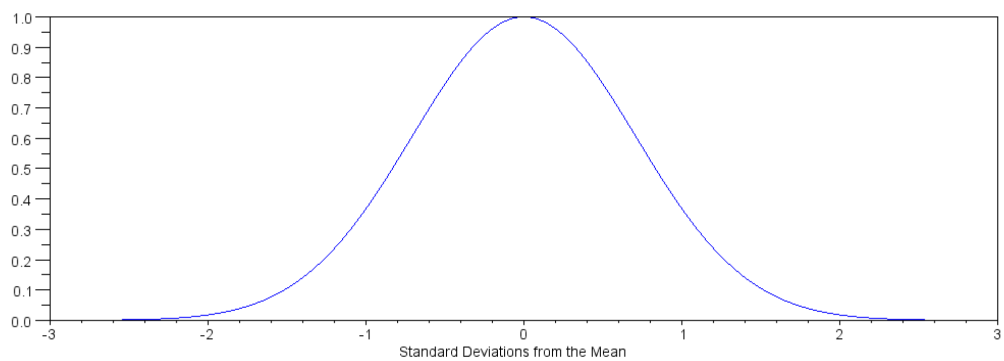
- $N(\mu, \sigma^2)$
- μ: The mean (average) value
- σ^2: The variance (a measure of the spread)

These are computed as

$$\mu = \frac{1}{n}\sum x_i$$

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2$$

A *standard* Normal distribution is a special case where the mean is zero and the standard deviation is one. It has a probability density function proportional to the following



Standard Normal Distribution: N(0, 1)

With this, you can compute different probabilities.

Example: Suppose you took the sum or rolling 10 six-sided dice (10d6). What is the probability you'll roll 50 or higher?

Solution: Here's where we use the Central Limit Theorem. The mean and standard deviation of rolling a single 6-sided die is

$$\mu = 3.5$$

$$\sigma^2 = 2.917$$

If you add normal distributions,

- The mean adds, and
- The variance adds (s^2)

For 10d6

$$\mu = 35$$

$$\sigma^2 = 29.17$$

$$\sigma = 5.401$$

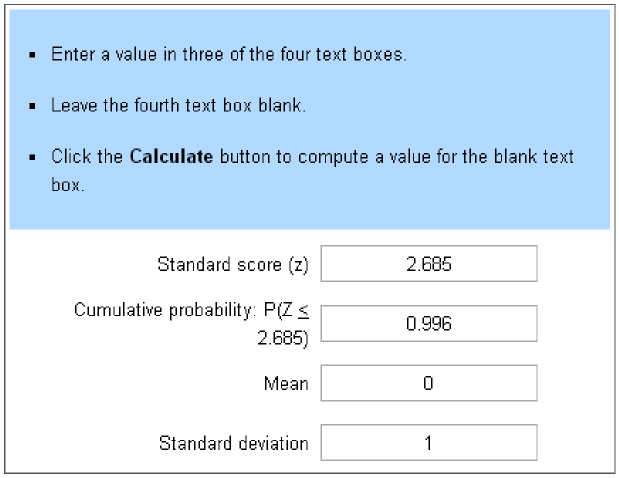
To find the probability of the total being 50 or more, find the area under the curve to the right of 49.5. To find this, determine how many standard deviations 49.5 is to the right of the mean

$$z = \left(\frac{49.5-\mu}{\sigma}\right) = \left(\frac{49.5-35}{5.401}\right) = 2.685$$

Use a Normal table to determine the area of the tail that's 2.685 standard deviations away from the mean

Normal Distribution (area of tail)									
0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.29

This table tells you that the tail has an area between 0.005 and 0.001. You can also go to StatTrek to get the same result:



The probability of rolling less than 50 is 0.996. The probability of rolling more than 50 is 0.004. (www.StatTrek.com)

This tells you that the probability of rolling 50 or more with 10d6 is 0.004.

Student t-Distribution (t-test)

If instead of knowing the mean and standard deviation you estimate it using data, you get a Student t-Distribution. This is similar to a Normal distribution except that the tails move out to compensate for having only limited data. In the limit where the sample size goes to infinity, the Student t-Distribution converges to a Normal distribution. A Student-t distribution has three parameters:

- The mean ($\bar{x}$)
- The standard deviation (s), and
- The degrees of freedom (p) - equal to the sample size minus one.

These are computed slightly differently from a Normal distribution

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

$$\sigma^2 \rightarrow s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$ *slight difference from the Normal distribution*

$$p = n - 1$$ *degrees of freedom equals sample size minus one*

A Student-t Table looks like the following:

Student t-Table (area of tail)										
p	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	0.82	1.06	1.39	1.89	2.92	4.3	6.97	9.93	22.33	31.6
3	0.77	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.22	12.92
4	0.74	0.94	1.19	1.53	2.13	2.78	3.75	4.6	7.17	8.61
5	0.73	0.92	1.16	1.48	2.02	2.57	3.37	4.03	5.89	6.87
10	0.7	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14	4.59
20	0.69	0.86	1.06	1.33	1.73	2.09	2.53	2.85	3.55	3.85
infinity	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.29

Student t-Table. Note that as the sample size goes to infinity (p = infinity), you converge to a Normal distribution.

The procedure to use a Student t-Table is almost identical to that of using a Normal distribution -only you look in the row corresponding to your sample size minus one (p = degrees of freedom).

Example: The monthly high for the month or April in Fargo was

- 2018: 81F
- 2017: 76F

What is the probability it will break 90F in this coming April?

Solution: Using this data, compute the mean and standard deviation

$$\bar{x} = \frac{1}{n}\Sigma(x_i) = 78.5F$$

$$s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2 = 12.5$$

$$s = 3.536$$

Compute the t-score (how many standard deviations 90F is away from the mean)

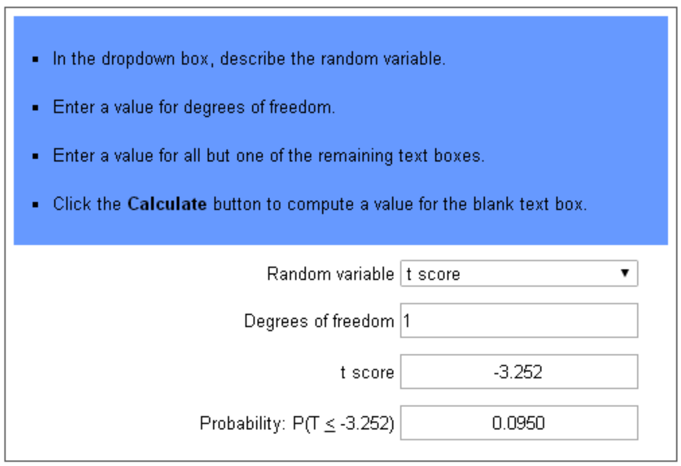
$$t = \left(\frac{90-78.5}{3.536}\right) = 3.252$$

Convert this t-score to a probability using a Student t-Table. (note: you have one degree of freedom since the sample size is only two)

Student t-Table (area of tail)										
p	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	0.82	1.06	1.39	1.89	2.92	4.3	6.97	9.93	22.33	31.6
3	0.77	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.22	12.92

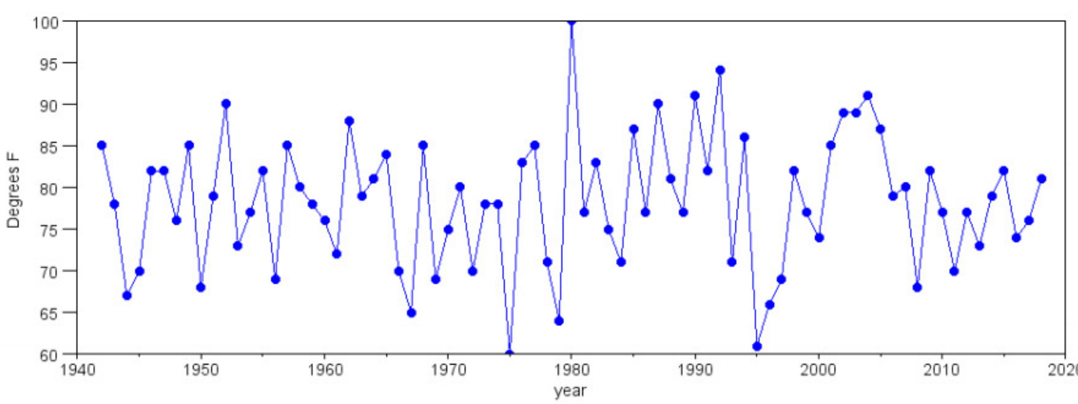
Student t-Table. With t = 3.252, the probability of being warmer than 90F is in-between 0.05 and 0.025

You can also use StatTrek



The probability of breaking 90F this coming April is 0.0950 (www.StatTrek.com)

With more data, you can get a more accurate answer. This shows up in the t-Table with more degrees of freedom. For example, determine the probability of breaking 90F in April of this year using 77 years of data:

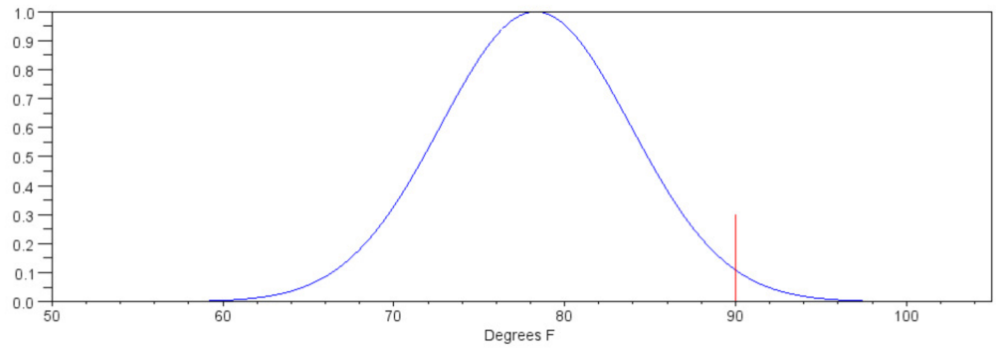


High for the Month of April since 1942 (National Weather Service)

To determine this probability, compute the mean and standard deviation:

- $\bar{x} = 78.299F$
- s = 7.869F
- p = 76 *77 data points = 76 degrees of freedom*

This tells you that the probability distribution for the high for the month of April looks like the following. The probability of it breaking 90F this year is the area to the right of 90F.

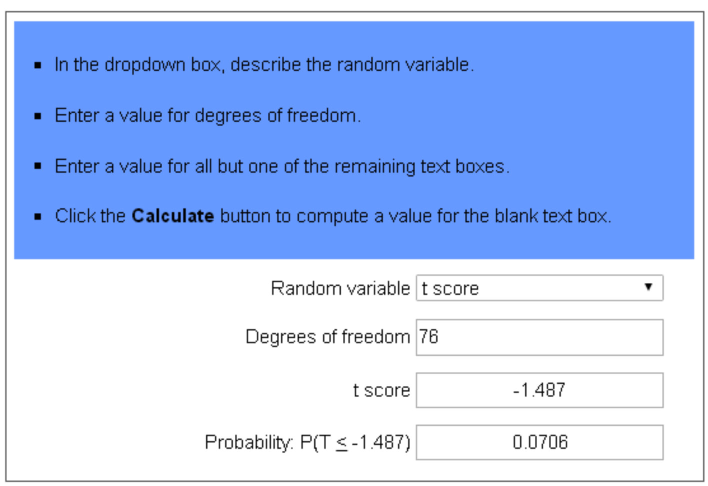


Probability Distribution for the High of the Month of April

Compute the t-score, calculate the distance of 90F to the mean in terms of standard deviations:

$$t = \left(\frac{90-78.299}{7.896}\right) = 1.487$$

Use a t-Table with 76 degrees of freedom to convert this t-score to a probability. Using StatTrek:



Using 77 years worth of data (76 degrees of freedom), the probability of it breaking 90F this coming April is 0.0706

Confidence Intervals:

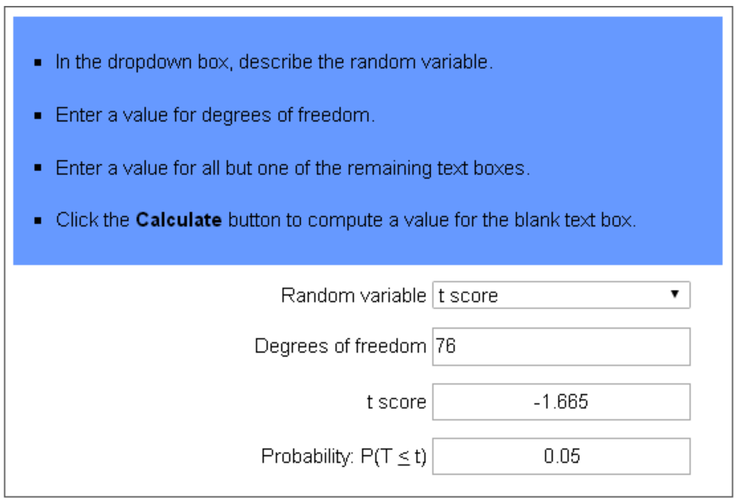
A second use of a Student t-Table is to determine the confidence interval. This is the range where you expect the next data point to fall.

Example: Determine the 90% confidence interval for high for the month of April.

Solution: Collect the data and determine the mean and standard deviation

- mean = 78.299F
- st dev = 7.869F
- sample size = 77

Use a Student t-Table to determine how far away from the mean you have to go for each tail to have an area of 5% (leaving 90% in the middle)

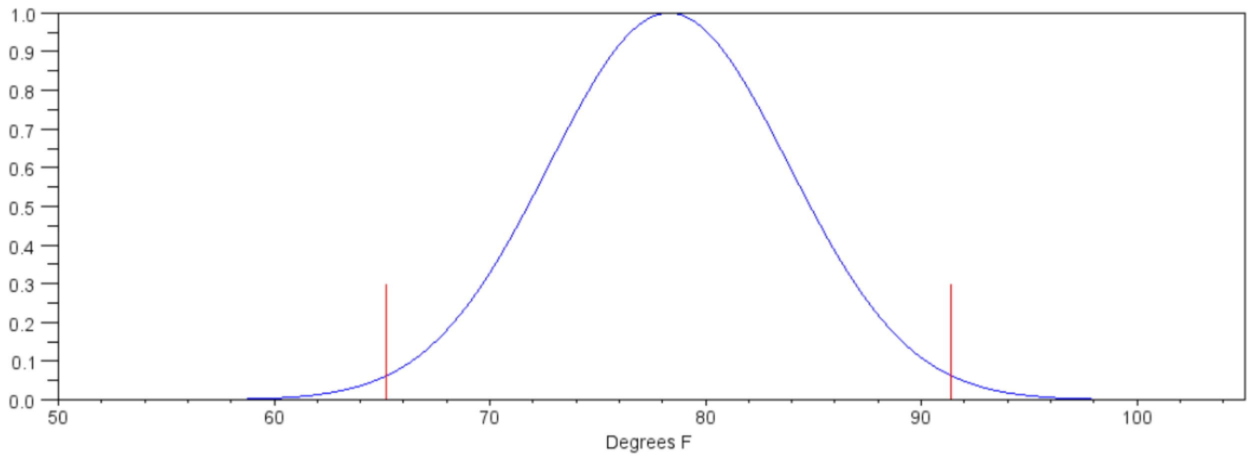


The 90% confidence interval is

$$\bar{x} - 1.665s < high < \bar{x} + 1.665s$$

$$65.19F < high < 91.40F$$

It is 90% likely that the high for the month or April will be in the range of (65.19F, 91.40F)



90% Confidence Interval for the High for the Month of April

Sidelights:

i) You have to have an interval. The area of a point is zero.

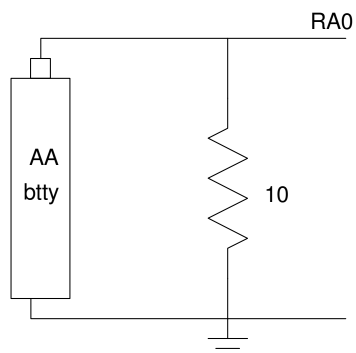
ii) You have to specify a probability less than 1.000. If you want to be 100% certain, the t-score is infinity.

Statistic Examples:

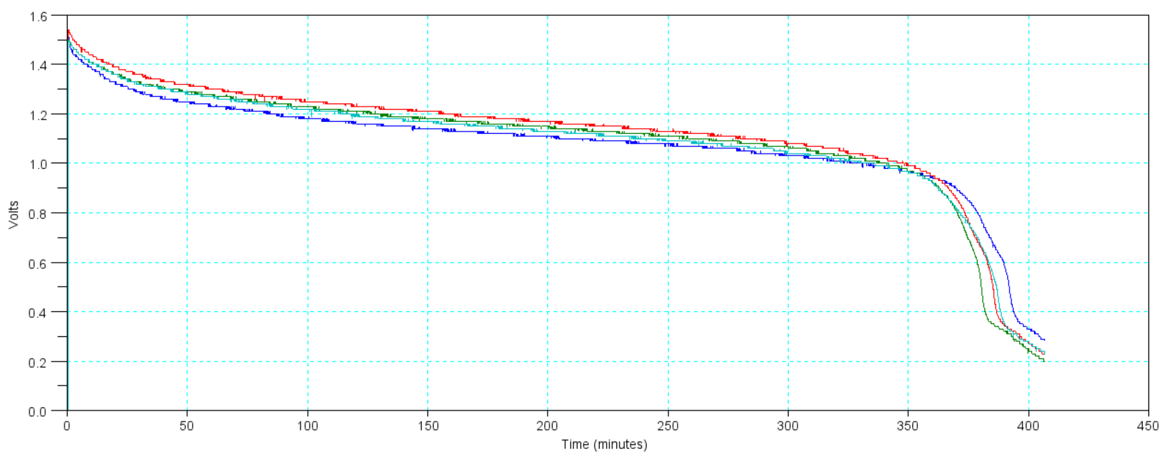
1. How much energy is in a AA battery?

Design of Experiment:

- Discharge a AA battery across a 10 Ohm resistor.
- Measure the voltage every 6 seconds for 7 hours.
- Compute the power at any time as $P = \frac{V^2}{R}$
- Integrate to get the energy dissipated across the 10 Ohm resistor



Data:



Statistical Analysis:

To analyze this, you have to convert each plot to a number. Once you have a number, you can compute the mean, standard deviation, and run a t-test.

Question 1) What is the total energy in this battery in Joules?

Convert the data for each battery into a single number: the energy in Joules. To do this, compute the energy at any given time as

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

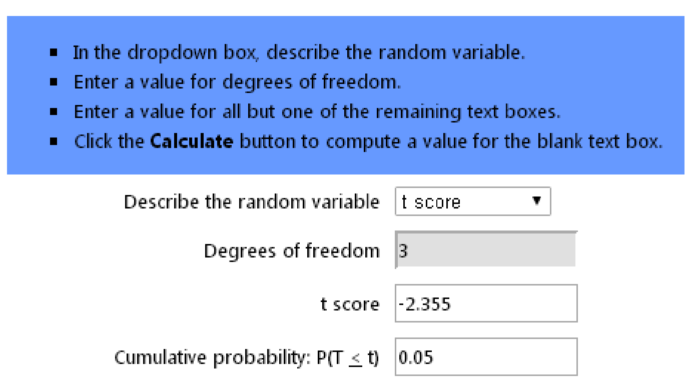
Integrate to find the total energy in Joules. Since the sampling rate was 6 seconds, the MATLAB code would be

```
Watts = ( V .^ 2 ) / 10;
Joules = sum(Watts) * 6

 Joules  =

    2937.5546    3063.8801    3204.829    3019.9991
```

Now, find the mean and the standard deviation:

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

    3056.5657

-->s = stdev(Joules)

    111.85742
```

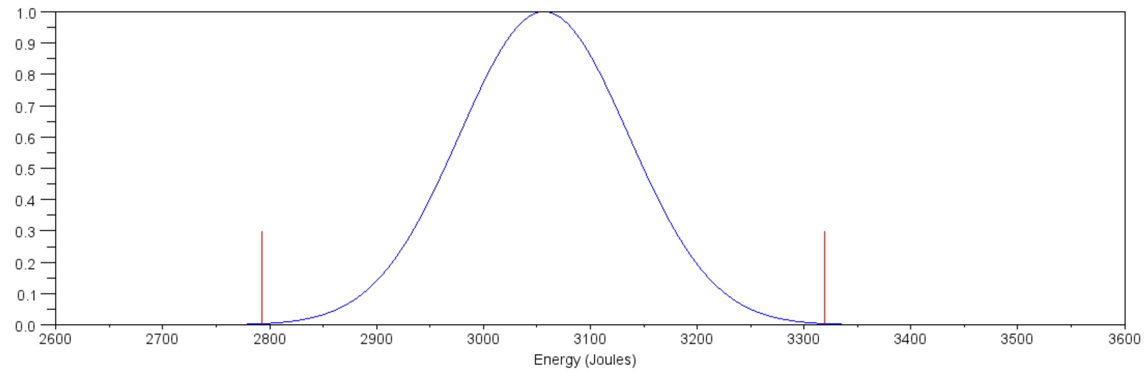
From StatTrek.com, with 3 degrees of freedom and 5% tails, you need to go out +/- 2.355 deviations



StatTrek.com

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

$$2793 < \text{Joules} < 3319$$



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

Question 2: This AA Batteries is rated at 500 mAh. What percentage of batteries have at least this much energy?

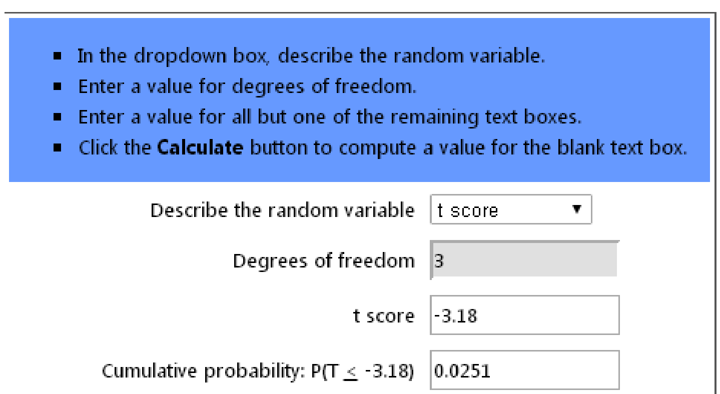
Convert 500mAh to Joules. Assuming 1.5V (the rated voltage):

$$1.5V \cdot 500mAh \cdot \left(\frac{60\min}{hr}\right)\left(\frac{60s}{\min}\right) = 2700 \text{ Joules}$$

2700 Joules is to the left of the mean by 3.18 standard deviations:

```
-->(x - 2700) / s

      3.1876804
```

From a t-table with 3 degrees of freedom, the tail has an area of 2.51%

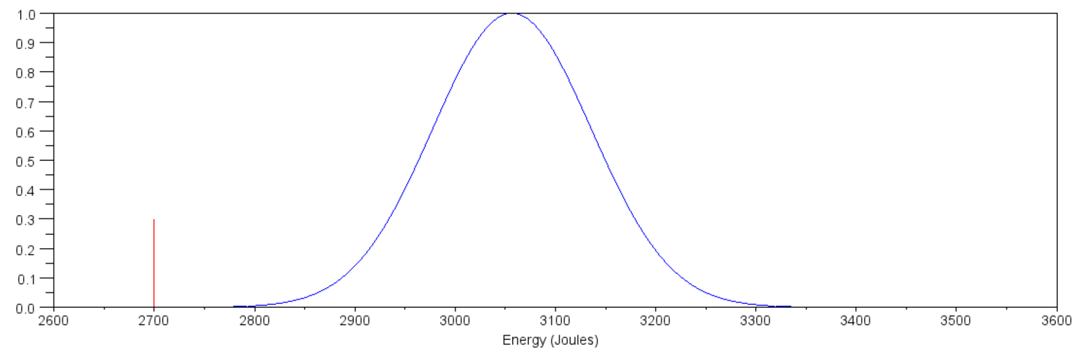


StatTrek.com

Result:

97.49% of the motors should meet the specifications.

Based upon this data, 97.49% of all batteries should have a total energy in excess of 500mAh



97.49% of thebatteries should have a total energy more than 500mAh (2700 Joules)

Problem #2: How cold is my freezer?

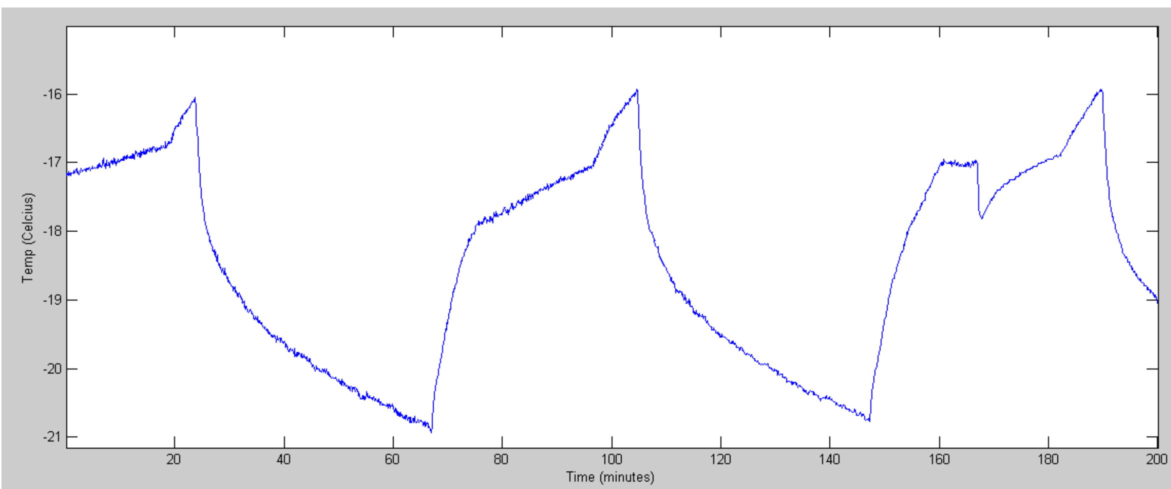
Design of Experiment: Place a thermistor in the freezer where

$$R \approx 1000 \cdot e^{-0.04(T-25)}\ \Omega$$

Turn resistance in to voltage using a 1k resistor and a voltage divider.

Measure the voltage with a PIC processor every 6 seconds for 3 hours.

Data:



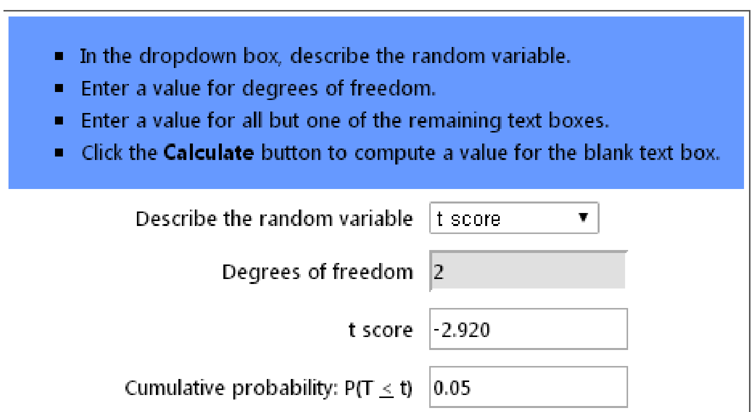
Analysis: To use statistics, you have to convert this plot to a number.

Take the hottest point at each cycle:

Cycle	Max Temp
1	-16.04
2	-15.94
3	-15.94
mean	-15.97
st dev	0.06

```
>> DATA = [-16.0438, -15.9363, -15.9363]

DATA =

  -16.0438  -15.9363  -15.9363

>> x = mean(DATA)

  -15.9721

>> s = std(DATA)

    0.0621
```

From StatTrek, you need to go 2.92 standard deviations on both sides for a probability of 90% (5% for each tail):



Based upon this data, I am 90% certain that the hottest my freezer gets during any cycle is

-15.7909C < T < -16.1534C

Student t-Table (area of tail) (http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf)										
p	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	0.82	1.06	1.39	1.89	2.92	4.3	6.97	9.93	22.33	31.6
3	0.77	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.22	12.92
4	0.74	0.94	1.19	1.53	2.13	2.78	3.75	4.6	7.17	8.61
5	0.73	0.92	1.16	1.48	2.02	2.57	3.37	4.03	5.89	6.87
6	0.72	0.91	1.13	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.71	0.9	1.12	1.42	1.9	2.37	3	3.5	4.79	5.41
8	0.71	0.89	1.11	1.4	1.86	2.31	2.9	3.36	4.5	5.04
9	0.7	0.88	1.1	1.38	1.83	2.26	2.82	3.25	4.3	4.78
10	0.7	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14	4.59
11	0.7	0.88	1.09	1.36	1.8	2.2	2.72	3.11	4.03	4.44
12	0.7	0.87	1.08	1.36	1.78	2.18	2.68	3.06	3.93	4.32
13	0.69	0.87	1.08	1.35	1.77	2.16	2.65	3.01	3.85	4.22
14	0.69	0.87	1.08	1.35	1.76	2.15	2.62	2.98	3.79	4.14
15	0.69	0.87	1.07	1.34	1.75	2.13	2.6	2.95	3.73	4.07
16	0.69	0.87	1.07	1.34	1.75	2.12	2.58	2.92	3.69	4.02
17	0.69	0.86	1.07	1.33	1.74	2.11	2.57	2.9	3.65	3.97
18	0.69	0.86	1.07	1.33	1.73	2.1	2.55	2.88	3.61	3.92
19	0.69	0.86	1.07	1.33	1.73	2.09	2.54	2.86	3.58	3.88
20	0.69	0.86	1.06	1.33	1.73	2.09	2.53	2.85	3.55	3.85
25	0.68	0.86	1.06	1.32	1.71	2.06	2.49	2.79	3.45	3.73
30	0.68	0.85	1.06	1.31	1.7	2.042	2.46	2.750	3.39	3.646
40	0.68	0.85	1.05	1.3	1.68	2.02	2.42	2.7	3.31	3.55
60	0.68	0.848	1.05	1.3	1.67	2	2.390	2.660	3.232	3.46
infinity	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.29