Statistical Analysis

Probability and hypothesis testing

With the analog input, a PIC microcontroller is able to measure voltage, light, temperature, etc. You can also generate random numbers, such as a 6-sided die. For each of these, if you take several measurements or roll a die several times, you'll get a different number each time. This means you're dealing with uncertainty. To analyze a system with uncertainty, you should use statistics.

In this lecture, we'll look at two different tests:

- Chi-Squred Test: Test of a distribution.
  - *When you program a PIC to generate random numbers from 1..6, are the numbers really random? Do they have equal probability? Is the die loaded?*
- 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?*

Chi-Squred Test

The Chi-Squred test is used to determine if your data is consistent with an assumed distribution. It is used to test

- Whether a die is fair (each number has equal probability)
- Whether a distribution is Normal (vs. Poisson or geometric)

For example, suppose you program your PIC to generate a 6-sided die using the following code:

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

When you press RB0, the PIC processor counts really fast mod 6. When you release the button, whatever number was stored in DIE is the random number.

Note that this may or may not be a fair die. The corresponding assembler code simply counts, except when you get to 6. In that case, you do something slightly different. This difference might show up as spending a few extra clocks on one number, resulting in that number showing up too often.

To run a Chi-Squred test to see if this really *is* a fair die, do the following:

- Divide the results into M bins (6 bins in this case: numbers 0 .. 5)
- Collect n data points.
- Count how many times the data fell into each of the M bins
- Compute the Chi-Squred total for each bin as

`

$$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$$

  - *np is the expected number of times data should fall into each bin*

- *N is the actual number of times data fell into each bin*

- Use a Chi-Squared table to convert the resulting Chi-Squred score to a probability. Note that the degrees of freedom is equal to the number of bins minus one.

Example: n = 300 die rolls

Die Roll (bin)	p theortical probability	n*p expected results	N actual results	$\chi^2 = \left(\frac{(N-np)^2}{np}\right)$
1	1/6	50	48	0.08
2	1/6	50	52	0.08
3	1/6	50	53	0.18
4	1/6	50	48	0.08
5	1/6	50	51	0.02
6	1/6	50	48	0.08
			Total:	0.52

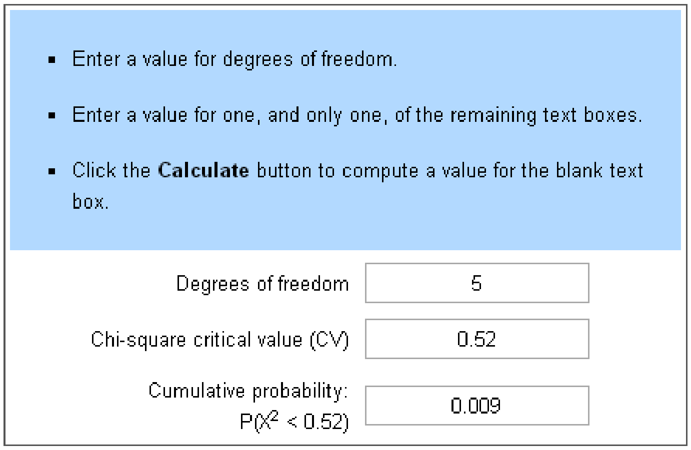
From a Chi-Squred table with 5 degrees of freedom, a Chi-Squred total of 0.52 corresponds to a probability of about 1%. This tells you that, based upon this data, there is only a 1% chance that the die is loaded. The code we've been using appears to generate a fair die.

Chi-Squared Table

Probability of rejecting the null hypothesis

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	**0.55**	0.41

You can also go to StatTrek.com to get the same result



Converting a Chi-Squred value to a probability using www.StarTrek.com

Example 2: Loaded Die

Suppose instead you had a loaded die:

- 90% of the time, the die is fair (all results have equal probability)
- 10% of the time, the result is always a 6.

Can you detect that the die is fair?

Code:

```
while(!RB0);
while(RB0) {
   DIE = (DIE + 1) % 6;
   Load = (Load + 1) % 100;
   }
if (Load < 10) DIE = 6;
```

To check if this die is loaded, again collect data and compute the Chi-Squred score. Running this code 300 times results in the following (note: each time you do this you'll get different results. It's random)

Die Roll	p	n*p	N (actual)	$\chi^2 = \left(\frac{(N-np)^2}{np}\right)$
1	1/6	50	44	0.72
2	1/6	50	47	0.18
3	1/6	50	42	1.28
4	1/6	50	50	0
5	1/6	50	56	0.72
6	1/6	50	61	2.42
			Total:	5.32

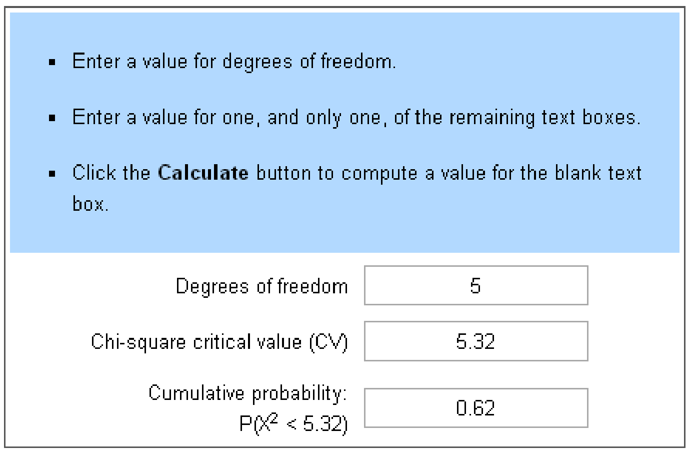
A Chi-Squred table allows you to convert the Chi-Squred score to a probability

Chi-Squared Table

Probability of rejecting the null hypothesis

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	**9.24**	**1.61**	1.15	0.83	**0.55**	0.41

Form the Chi-Squared table, the probability is between 10% and 90%. From StatTrek, it's actually 0.62



StarTrek.com. A Chi-Squared score of 5.32 corresponds to a probability of 0.62

Based upon this data, there is a 62% chance that this die is loaded. That's still not enough to accuse someone of cheating. Loaded dice are difficulty to spot unless you have a LOT of data. By that time, you're probably broke. But then, you'll know why.

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:

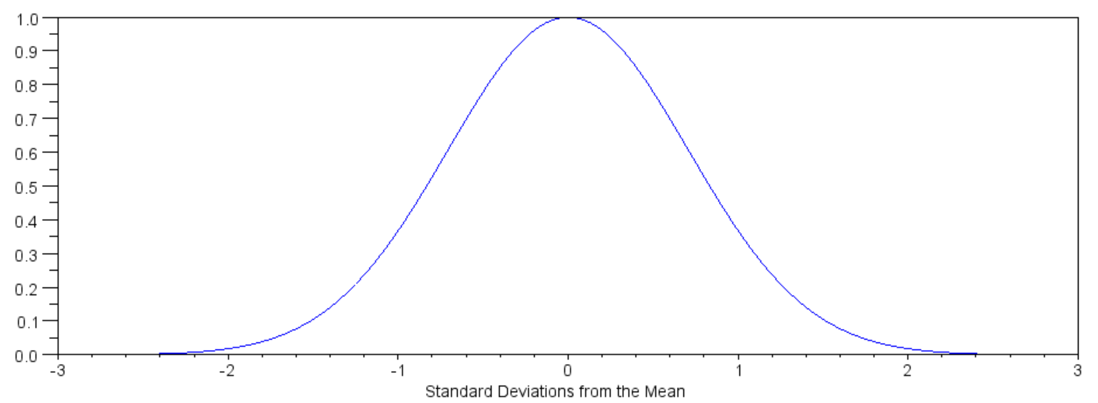
- $N(\bar{x}, s)$
- $\bar{x}$: The mean (average) value
- s: The standard deviation (a measure of the spread)

These are computed as

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

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

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

$$\bar{x} = 3.5$$

$$s^2 = 2.917$$

If you add normal distributions,

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

For 10d6

$$\bar{x} = 35$$

$$s^2 = 29.17$$

$$s = 5.401$$

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

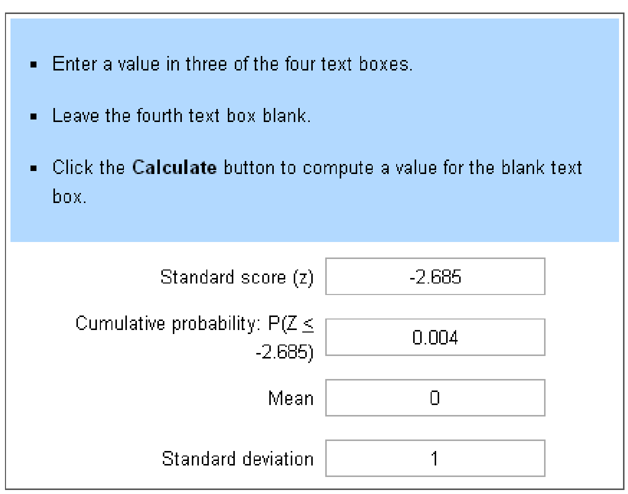
$$z = \left(\frac{49.5 - \bar{x}}{s}\right)$$

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

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

Normal Distribution									
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 thatthe tail has an area between 0.005 and 0.001. You can also go to StatTrek to get the same result:



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

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

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

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

A Student-t Table looks like the following:

Student t-Table (http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf)										
p	0.25	0.2	0.15	0.1	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 April 2019?

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

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

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

$$s = 3.536$$

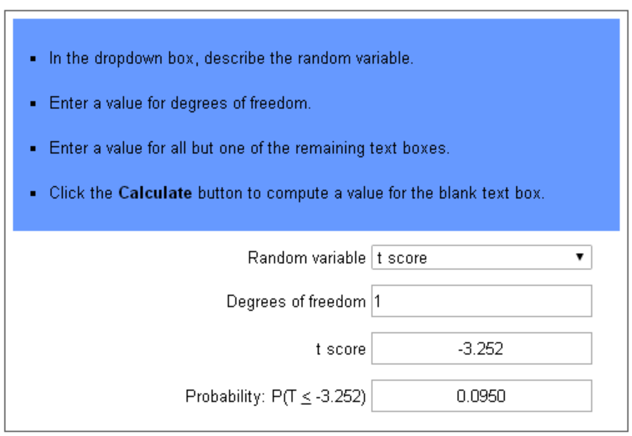
Compute the t-score (how many standard deviations 90F is away from the mean)

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

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)

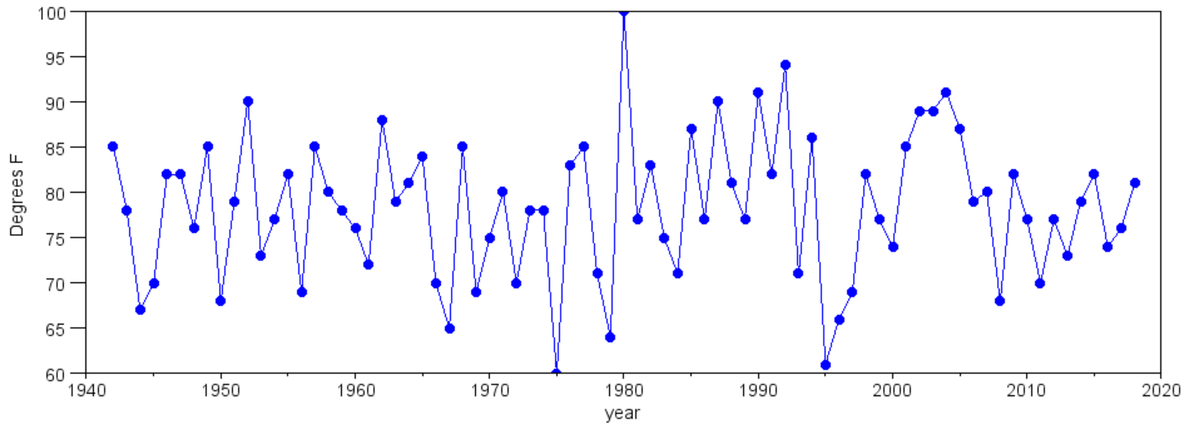
Student t-Table (http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf)										
p	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	1	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	0.82	1.06	1.39	1.89	2.92	4.3	6.97	9.93	22.33	31.6
3	0.77	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.22	12.92

You can also use StatTrek



The probability of breaking 90F in April 2019 is 0.095 (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:

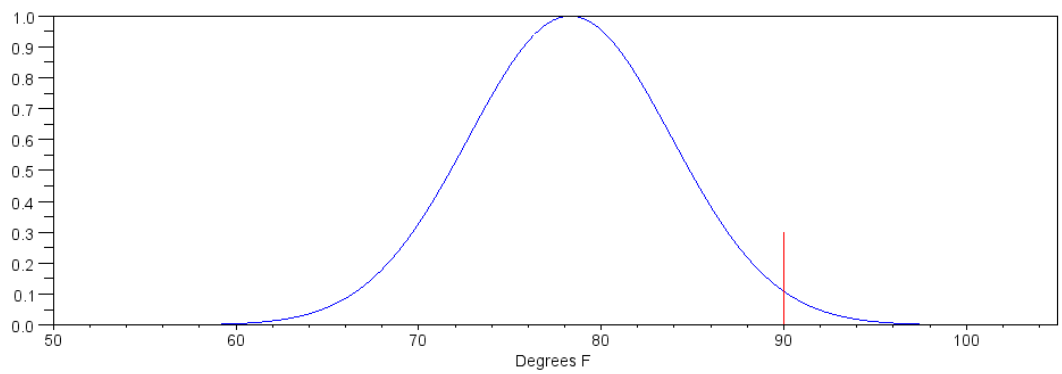


High for the Month of April since 1942 (National Weather Service)

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.

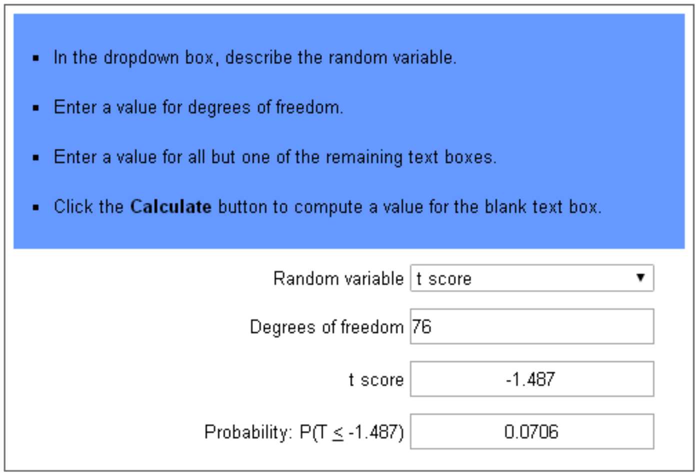


Probability Distribution for the High of the Month of April

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

$$t = \left(\frac{90-78.299}{7.869}\right) = 1.487$$

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:

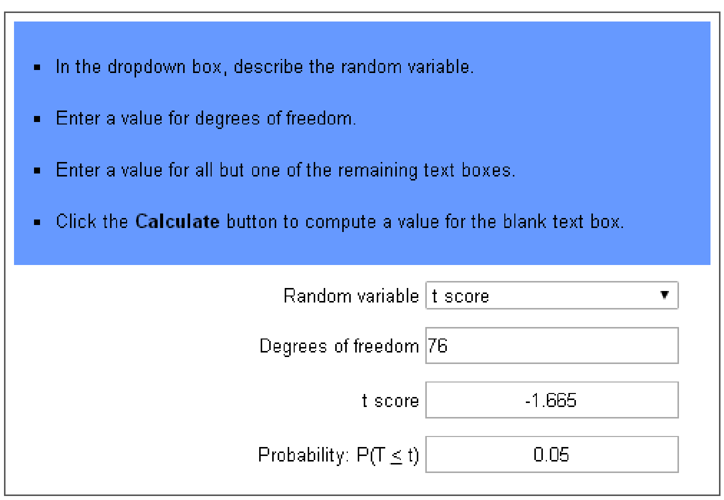
A second use of a Student t-Table is to determine the confidence interval. This is the range where you expect the next data point to fall.

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)

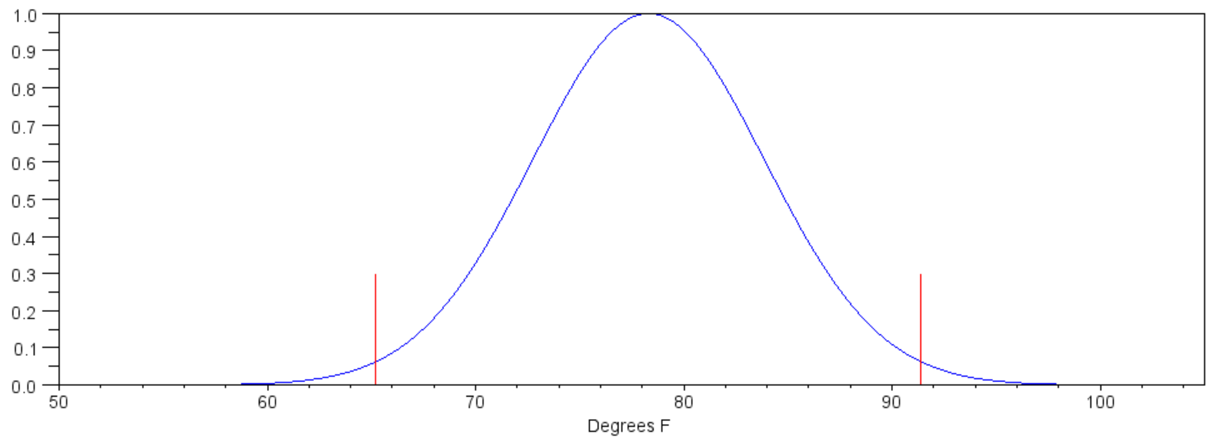


The 90% confidence interval is

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

$$65.19F < high < 91.40$$

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.

Comparing Two Means

With a Student t-Distribution, you can also compare two distributions. For example, which month is hotter on average, April or October?

To compare two means, take two distributions (call it A and B). Create a variable, W, equal to the difference

W = A - B

The mean and standard deviation of W is then

$$\bar{x}_w = \bar{x}_a - \bar{x}_b$$

$$s_w^2 = \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}$$

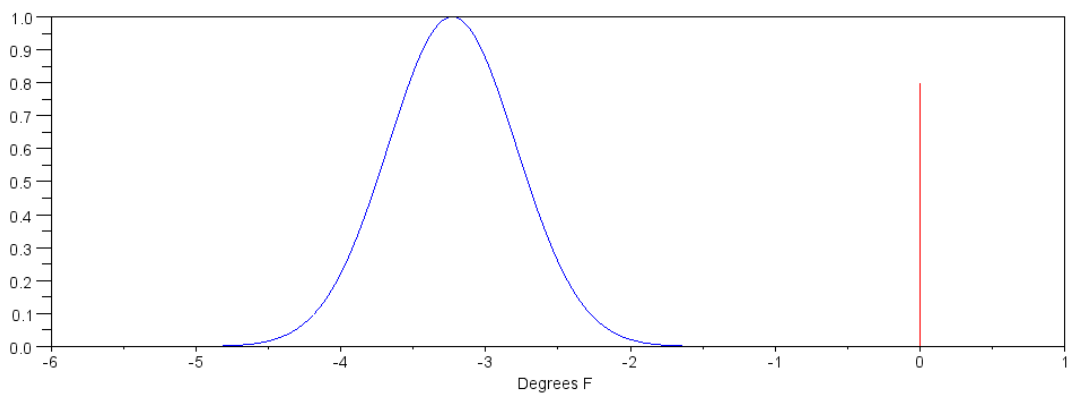
The t-score is the area under the curve to the left of zero

$$t = \frac{\bar{x}_w}{s_w}$$

Example: Which month is hotter on average: April or October? Start with the data

	April Avg Temp	October Avg Temp	W April - October
mean	42.97	46.2	-3.23
s	4.26	3.49	0.63
sample size	77	77	min(77, 77) = 77

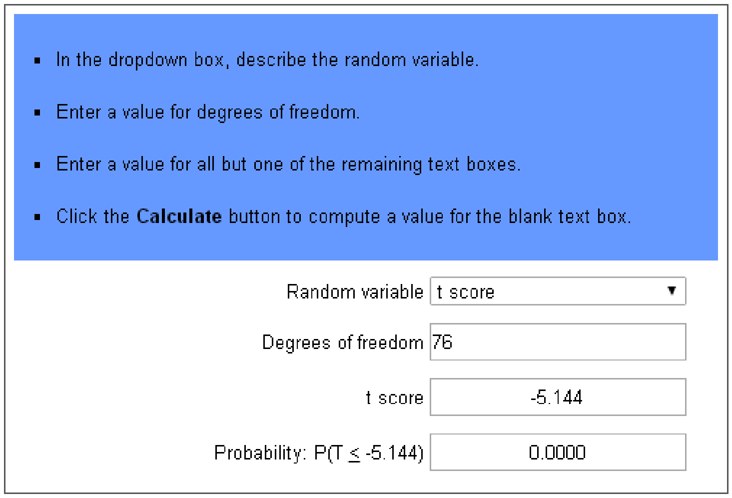
This tells you that the distribution for W (April - October) looks like this:



The probability that April is warmer than October is the area to the right of the red line. To compute this, find the t-score

$$t = \left(\frac{0 - 3.23}{0.63}\right) = 5.144$$

Use a Student t-Table to convert this t-score to a probability



The probability that April is warmer than October is less than 0.00005

Chi-Squared Table

Probability of rejecting the null hypothesis
http://people.richland.edu/james/lecture/m170/tbl-chi.html

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41
6	18.55	16.81	14.45	12.59	10.65	2.2	1.64	1.24	0.87	0.68
7	20.28	18.48	16.01	14.07	12.02	2.83	2.17	1.69	1.24	0.99
8	21.96	20.09	17.54	15.51	13.36	3.49	2.73	2.18	1.65	1.34
9	23.59	21.67	19.02	16.92	14.68	4.17	3.33	2.7	2.09	1.74
10	25.19	23.21	20.48	18.31	15.99	4.87	3.94	3.25	2.56	2.16
11	26.76	24.73	21.92	19.68	17.28	5.58	4.58	3.82	3.05	2.6
12	28.3	26.22	23.34	21.03	18.55	6.3	5.23	4.4	3.57	3.07
13	29.82	27.69	24.74	22.36	19.81	7.04	5.89	5.01	4.11	3.57
14	31.32	29.14	26.12	23.69	21.06	7.79	6.57	5.63	4.66	4.08
15	32.8	30.58	27.49	25	22.31	8.55	7.26	6.26	5.23	4.6
16	34.27	32	28.85	26.3	23.54	9.31	7.96	6.91	5.81	5.14
17	35.72	33.41	30.19	27.59	24.77	10.09	8.67	7.56	6.41	5.7
18	37.16	34.81	31.53	28.87	25.99	10.87	9.39	8.23	7.02	6.27
19	38.58	36.19	32.85	30.14	27.2	11.65	10.12	8.91	7.63	6.84
20	40	37.57	34.17	31.41	28.41	12.44	10.85	9.59	8.26	7.43
30	53.67	50.89	46.98	43.77	40.26	20.6	18.49	16.79	14.95	13.79
40	66.77	63.69	59.34	55.76	51.81	29.05	26.51	24.43	22.16	20.71
50	79.49	76.15	71.42	67.51	63.17	37.69	34.76	32.36	29.71	27.99
60	91.95	88.38	83.3	79.08	74.4	46.46	43.19	40.48	37.49	35.53
70	104.22	100.43	95.02	90.53	85.53	55.33	51.74	48.76	45.44	43.28
80	116.32	112.33	106.63	101.88	96.58	64.28	60.39	57.15	53.54	51.17
90	128.3	124.12	118.14	113.15	107.57	73.29	69.13	65.65	61.75	59.2
100	140.17	135.81	129.56	124.34	118.5	82.36	77.93	74.22	70.07	67.33

Student t-Table (http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf)										
p	0.25	0.2	0.15	0.1	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