

Data Analysis & Student t Test

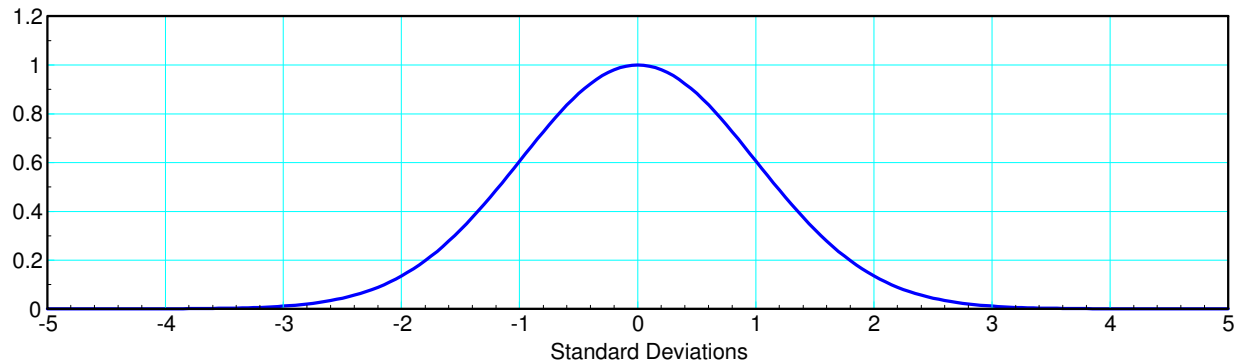


Figure 1: Standard normal distribution: the heart of a t-test

Probably the most common test used in data analysis is the Student t test: this test is a test of the mean. With it, you can determine the 90% confidence interval for

- The gain of a transistor,
- The energy in a AA battery,
- The value of a capacitor, or
- The thermal time constant of a coffee cup.

You can also compare two populations and determine

- Does type A battery have more energy than type B?
- Does adding a spoon to a cup of hot water make it cool off faster?
- Does adding a lid help keep it warm?

The heart of the Student t-Test is the standard normal distribution. This is the bell-shaped curve you've encountered many times (grade distribution, sum of rolling 10 dice, height of people, etc.) When doing a t-test, you're implicitly assuming that the data you're analyzing has a normal distribution.

This actually isn't that bad of an assumption. The Central Limit Theorem states that, under some very general assumptions,

- The sum or average of random variables converges to a normal distribution, and
- The sum of normal distributions is a normal distribution.

Translating: everything converges to a normal distribution. Once you get there, you're stuck with a normal distribution.

This lecture covers how to analyze data like we collected before.

Student t-Test

The Student-t distribution is described by three parameters: the mean, standard deviation, and degrees of freedom

Mean: The average of your data

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

Standard Deviation: A measure of the spread

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

Degrees of Freedom: Sample size minus one

$$\text{d.f.} = n-1$$

The probability density function is very similar to a normal distribution, only the tails are slightly extended in accordance to the sample size. As the sample size goes to infinity, the t-distribution converges to the normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(x-\bar{x})^2}{2s^2}\right)$$

The probability of getting a value is related to how far you are from the mean in terms of standard deviations, termed the t-score

$$t = \frac{x-\bar{x}}{s}$$

It's probably easiest to explain this through an example.

Example 1: Gain of a Zetex Transistor

The gain of a Zetex 1051a transistor was measured resulting in the following data:

915, 602, 963, 839, 815, 774, 881, 912, 720, 707, 800, 1050, 663, 1066, 1073, 802, 863, 845, 789, 964, 988, 781, 776, 869, 899, 1093, 1015, 751, 795, 776, 860, 990, 762, 975, 918, 1080, 774, 932, 717, 1168, 912, 833, 697, 797, 818, 891, 725, 662, 718, 728, 835, 882, 783, 784, 737, 822, 918, 906, 1010, 819, 955, 762

Determine

- Probability density function for any given Zetex 1051a transistor
- The probability that the gain for any given Zetex 1051a transistor being more than 500
- The 90% confidence interval for any given Zetex 1051a transistor, and

Solution: First, determine the mean, standard deviation, and sample size. In Matlab

```

hfe = [ <paste data here> ]
x = mean(hfe)
x = 854.1290
s = std(hfe)
s = 120.2034
df = length(hfe) - 1
df = 61

```

You can now plot the probability density function by starting with a standard normal distribution (mean = 0, standard deviation = 1) and scaling it

```

x1 = [-4:0.05:4]';
p = exp(-x1.^2 / 2);
plot(x1*s+x, p);

```

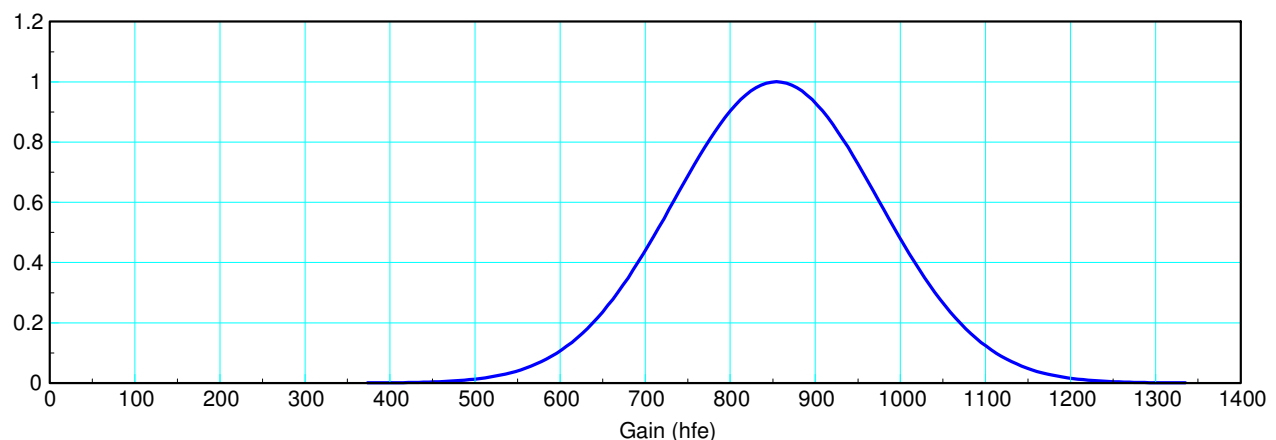


Figure 2: Normalize probability density function for a Zetex 1051a transistor

With this, you can now answer several questions.

i) What is the probability density function for any given Zetex 1051a transistor?

answer: The graph above.

ii) What is the probability that the gain of any given Zetex transistor is more than 500?

answer: This is the area to the right of 500. First, compute the t-score (the distance from the mean in terms of standard deviations)

$$t = \left(\frac{500 - \bar{x}}{s} \right) = \left(\frac{500 - 854.129}{120.2} \right) = -2.9461$$

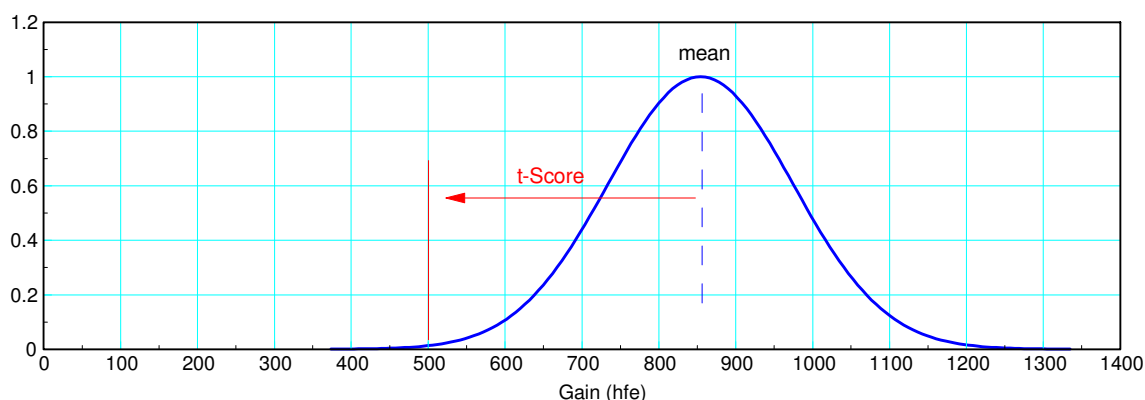


Figure 3: The t-score is the distance from the mean in terms of standard deviations

To convert this t-score to a probability, use a t-table.

To use a t-table,

- Go to the row with 61 degrees of freedom (which isn't on the table so use 60)
- Look for the number 2.9461 (the sign doesn't matter)
- The top of the table tells you the area of the tail

The area of the tail is about 0.002 (?)

The probability that the gain is less than 500 is 0.0023

The probability that the gain is more than 500 is 0.9977

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
10	0.7	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14	4.59
15	0.69	0.87	1.07	1.34	1.75	2.13	2.6	2.95	3.73	4.07
20	0.69	0.86	1.06	1.33	1.73	2.09	2.53	2.85	3.55	3.85
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

Table 1: Part of a student t-table

Another (easier) way to do this is to go to StatTrek. The area of the tail is 0.0023

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable	t score ▼
Degrees of freedom	61
t score	-2.9461
Probability: $P(T \leq -2.9461)$	0.0023

Figure 4: A t-score of -2.9461 corresponds to a probability of 0.0025 (www.StatTrek.com)

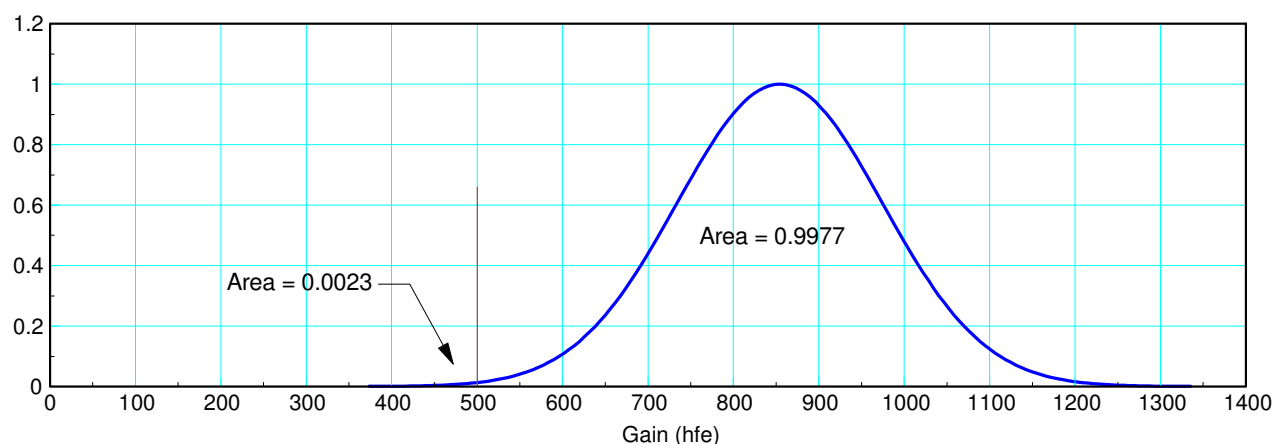


Figure 5: The probability that any given Zetex 1051a transistor has a gain more than 500 is 0.9977

iii) What is the 90% confidence interval for the gain of a Zetex 1051a transistor?

answer: For 90% of the area to be in the middle, each tail needs to be 5%. From the t-table, this corresponds to

$$t = 1.67$$

The 90% confidence interval is thus

$$\bar{x} - 1.67s < \text{gain} < \bar{x} + 1.67s$$

$$653 < \text{gain} < 1055$$

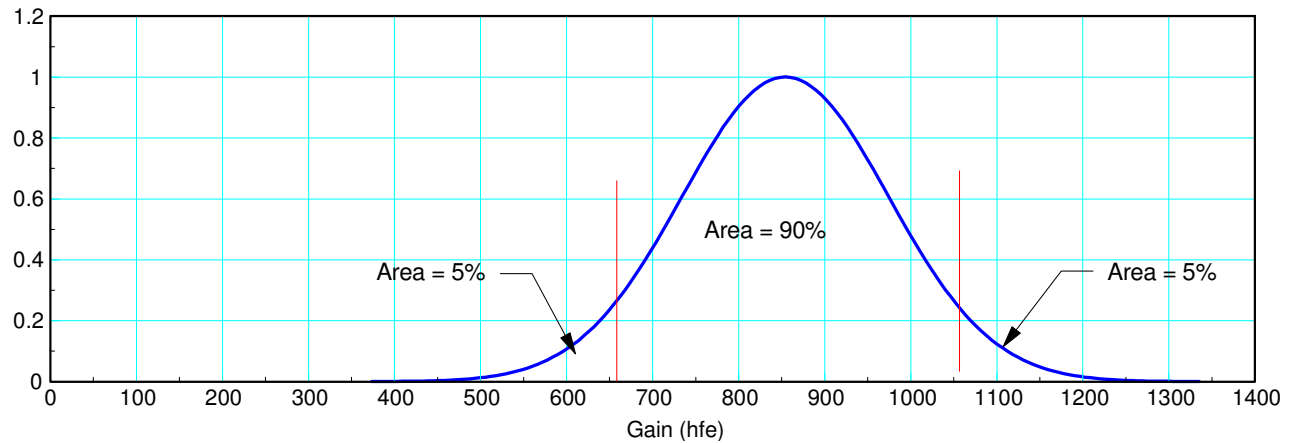


Figure 6: 90% confidence interval for the gain of a Zetex 1051a transistor

“If you don't know where you are going any road can take you there”

Lewis Carroll, Alice in Wonderland

Design of Experiment

Suppose you want to know

How much energy does a AA battery contain?

Before starting, think about how you will answer this question: ask the following:

- What question you want to answer?
- What data you need to answer that question?
- How much data you need?
- How you will go about collecting that data?
- How you will analyze that data?

The point behind this is to

- Collect the right data (don't waste time collecting data you can't use)
- Collect the right amount of data (don't waste time collecting too much or too little data)
- Make the experiment as repeatable as possible (minimize the variation in the data)

Example 2: Energy in a AA battery

Suppose you want to know

How much energy does a AA battery contain?

What data do we need?

Energy is hard to measure, Voltage is easy. If you

- Short the battery across a 10 Ohm resistor, and
- Measure the voltage every 6 seconds,

You can measure the power being dissipated in Watts

$$P = \frac{V^2}{R} = 0.1 V^2 \quad \text{Watts}$$

If you let the experiment run until the battery is discharged, you'll have the energy in Joules

$$E = \int P dt \quad \text{Joules}$$

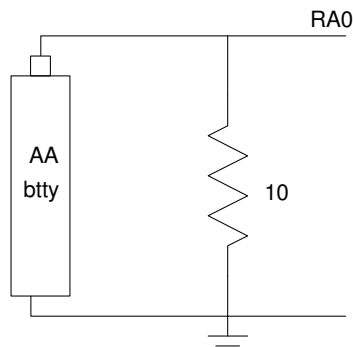


Figure 7: To determine the energy in a AA battery, the voltage is measured as it discharges across a 10 Ohm resistor

How Much Data do you Need?

This is where t-tables are kind of insightful.

- One data point (discharging one battery) tells you nothing. One battery has zero degrees of freedom
- Two data points actually let you analyze the data and answer your questions
- If you can afford to test three batteries, the t-score drops drastically (see the column for 0.01: the t-score drops from 31.82 to 6.97 by going from 1 to 2 degrees of freedom)
- You start to get diminishing returns once you go past 10

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
10	0.7	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14	4.59
15	0.69	0.87	1.07	1.34	1.75	2.13	2.6	2.95	3.73	4.07
20	0.69	0.86	1.06	1.33	1.73	2.09	2.53	2.85	3.55	3.85
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

Table 2: As the sample size increases, the t-score drops with diminishing returns

This is important. In industry, you typically have to scrap every item you test: it's no longer new.

- If you test all of your products, you have good data for statistical analysis. You're also broke since you no longer have any product to sell.
- If you test none of your products, you have no idea what you're selling.
- All you really need is a sample size of two. You can do statistical analysis with a sample size of two.
- Given a choice, a sample size of 4 or 5 would be nice. That gives you a lot more information and you only lose 4 or 5 from your inventory. These you can probably sell on ebay as "like new."

Long story short, let's test four batteries (for 3 degrees of freedom)

How will you collect that data?

This is where you get really picky about the experimental procedure. The goal is to follow a set procedure precisely. The hope is that if you do everything the same each time you run the experiment, you'll get the same data.

If you don't follow a procedure and are sloppy in collecting data, you tend to get wildly varying results (which shows up as a large standard deviation). If you have a large standard deviation, you'll end up saying something like

I am 90% certain that the energy in a AA battery is in the range of (-2000 Joules to +20,000 Joules)

A not terribly helpful conclusion.

For this experiment, the procedure was

- Purchase a pack of 4 batteries from the grocery store
- Connect a 10 Ohm resistor across each battery
- Measure the voltage across each battery using a PIC processor, sampled every 6 seconds

- Run the experiment for each battery for 10 hours.

Step 2: Data Collection

Once you have the procedure, collect data. The result is as follows:

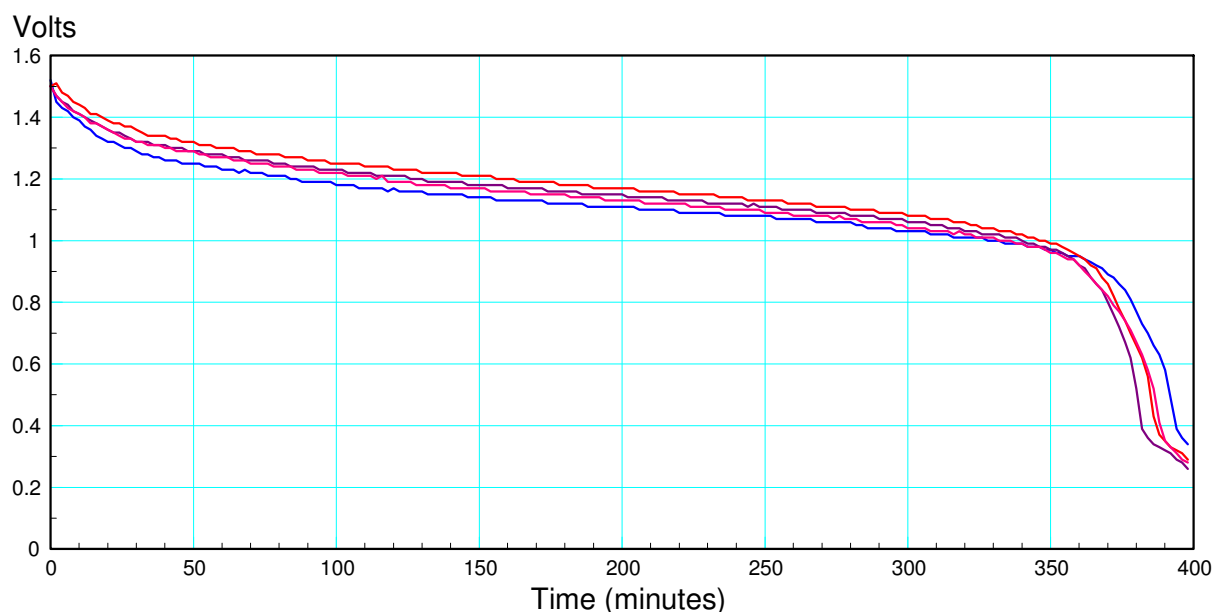


Figure 8: Voltage across four AA batteries as they discharge across a 10 Ohm resistor

Step 3: Data Analysis

In order to analyze the data, you need to convert each data set to a number.

- The average of the data is a number. It doesn't tell me much though.
- The time it takes to discharge down to 1.00V is a number. It sort of tells me the life of a battery.
- The energy contained in the battery in Joules is a number. That's actually useful information.

So, convert each graph to the energy contained in Joules.

The power dissipated is

$$P = \frac{V^2}{R} = 0.1 V^2 \text{ Watts}$$

The energy is the integral of the power. Since the sampling rate is 6 seconds

$$E = 0.6 \sum (V^2)$$

giving four numbers (one for each data set)

$$E = \{26,332 \quad 26,648 \quad 27,330 \quad 26,543\} \text{ Joules}$$

Now that you have four numbers, you can do some statistical analysis.

The mean is

```
x = mean(Joules)
x = 26,713
```

```
s = std(Joules)
s = 431.6950
```

The normalized probability of the energy in a given AA battery is thus

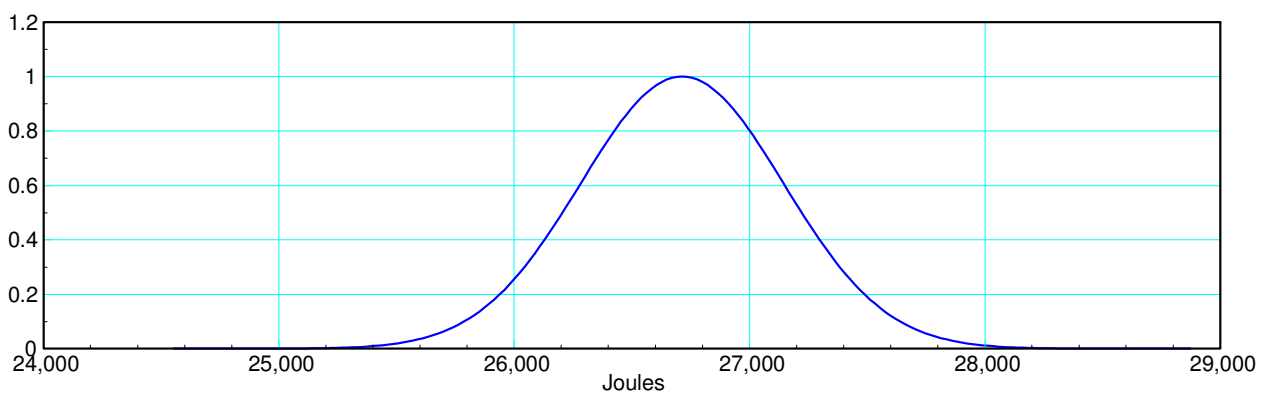


Figure 9: Energy in a AA battery: Mean = 26,713, standard deviation = 431 Joules

With this, you can answer some questions.

Question 1) What is the probability that a given batter will have more than 28,000 Joules?

To answer this, determine the distance from mean to 28,000 in terms of standard deviations.

$$t = \left(\frac{28,000 - \bar{x}}{s} \right) = \left(\frac{28,000 - 26,713}{431.69} \right) = 2.9808$$

Use a t-table to convert this to a probability

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable	t score ▼
Degrees of freedom	3
t score	2.9808
Probability: $P(T \leq 2.9808)$	0.9707

Figure 10: With a sample size of four, there are three degrees of freedom

- The probability that the energy is less than 28,000 is 97.07%
- The probability that the energy is more than 28,000 is 2.93%

Question 2: What is the 90% confidence interval for any given AA battery?

Answer: Use a t-table to convert 5% tails to a t-score

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable	t score ▼
Degrees of freedom	3
t score	-2.355
Probability: $P(T \leq t)$	0.05

Figure 11: For 5% tails and 3 degrees of freedom, the t-score is 2.355

The 90% confidence interval will be

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

$$25,6897 < \text{Joules} < 27,730$$

Comparison of Means

Suppose you collect data from two populations, A and B. What is the probability that the mean of A is more than the mean of B?

This is a common problem

- For batteries, you might want to know which one has more energy
- For coffee cups, you might want to know which one has better insulation

To solve this problem, create a new variable, W

$$W = A - B$$

The mean and variance of W will be

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

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

The degrees of freedom is a little more tricky. It's equation is

$$d.f. = \frac{\left(\left(\frac{s_1^2}{n_1} \right) + \left(\frac{s_2^2}{n_2} \right) \right)}{\left(\frac{(s_1^2/n_1)}{n_1-1} \right) + \left(\frac{(s_2^2/n_2)}{n_2-1} \right)}$$

or, to be slightly conservative, it's the smaller of the degrees of freedom between A and B.

The probability that the mean of A is more than the mean of B is the probability that $W > 0$. This has a t-score of

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

Example 3: Which battery has more energy:

- A = {29,376 30,639 32,048 30,200} Joules
- B = {30,186 30,197 30,668 29,820} Joules

Solution: Create a new variable, $W = A - B$. The mean and standard deviation are then:

	A	B	$W = A - B$
mean	30,566	30,218	34.811
st dev	111.86	34.76	58.56
d.f.	3	3	3

Table 3: Normalized probability distribution of type A and type B batteries

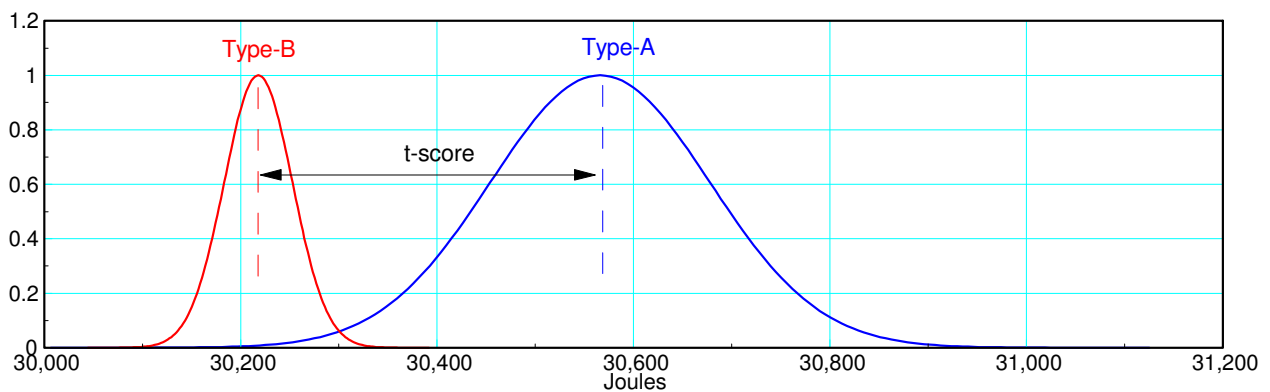


Figure 12: Probability distributions for Type-A batteries and Type-B batteries

The t-score is then the distance between the means relative to their standard deviations:

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{\bar{x}_w}{s_w} = \frac{34.811}{56.56} = 0.5944$$

From StatTrek, this corresponds to a probability of 0.7030

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable t score ▼

Degrees of freedom 3

t score 0.5944

Probability: $P(T \leq 0.5944)$ 0.7030

Figure 13: StatTrek: There is a 70.30% chance that battery A has more energy than battery B.

Matlab Code

```

A = [ 29376    30639    32048    30200];
B = [ 30186    30197    30668    29820];

Xa = mean(A)

Xa =    30566

Xb = mean(B)

Xb =    30218

Xw = Xa - Xb

Xw =    34.8110

Sa = std(A)

Sa =    111.8574

Sb = std(B)

Sb =    34.7628

Sw = sqrt( Sa^2 / 4 + Sb^2 / 4 )

Sw =    58.5673

df = (Sa^2/4 + Sb^2/4) / ( Sa^2/4/3 + Sb^2/4/3 )

df =    3

```

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