Data Analysis & Student t Test

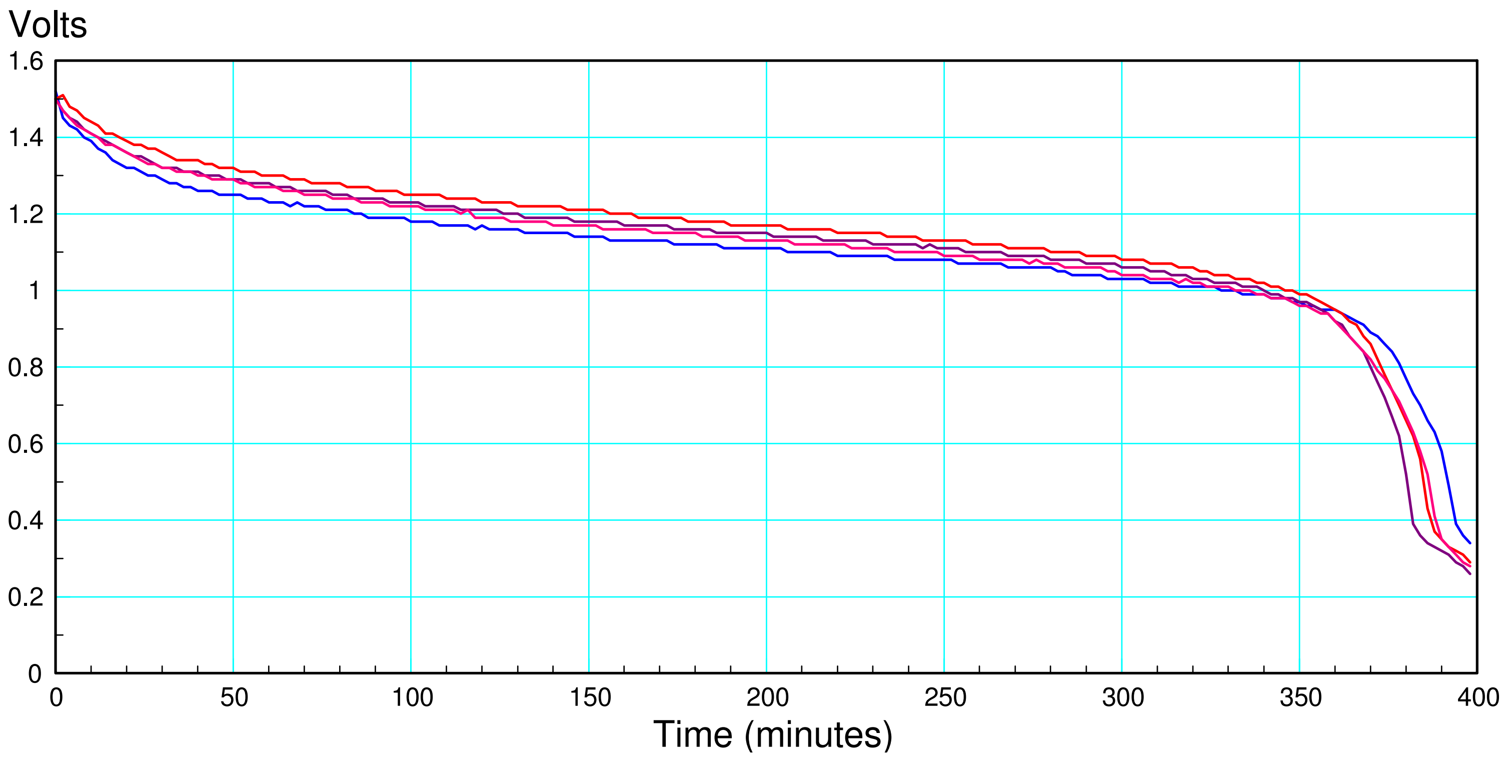
ECE 376 Embedded Systems

Jake Glower - Lecture #16

Please visit Bison Academy for corresponding
lecture notes, homework sets, and solutions

Problem:

- Every time you run an experiment, you get different results
  - *Example: Voltage across a AA battery as it discharges*
- How do you analyze such data?



Student t Test

Most common test for data analysis

- 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

- Which battery has more energy: A or B?
- Does adding a spoon to a hot cup of tea affect its cooling rate?
- Does adding a lid affect the cooling rate?

Central Limit Theorem

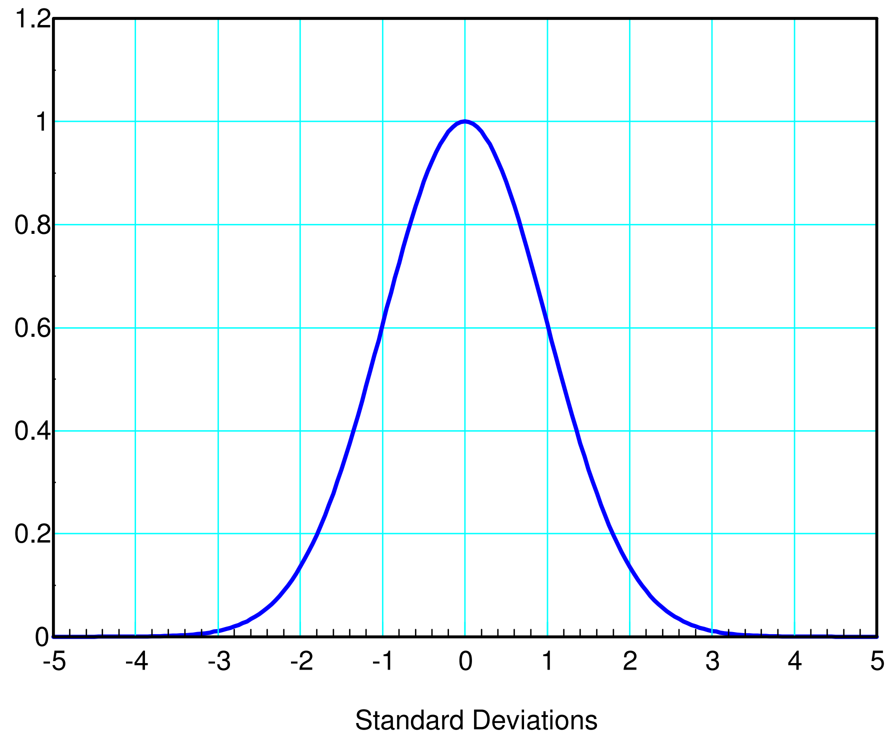
t-Tests assume your data has a normal distribution

Not a bad assumption

Central Limit Theorem:

- 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.



Student t Distribution

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

d.f. = n-1

The pdf looks like a Normal distribution, only with slightly wider tails

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
```

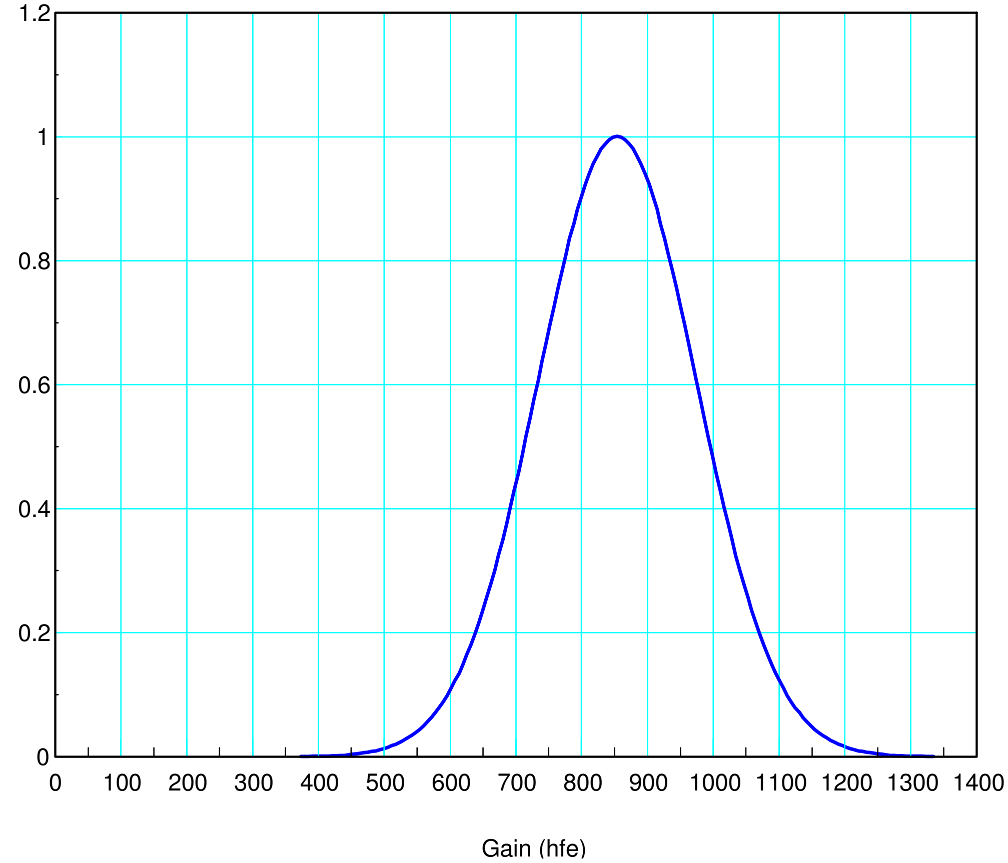
Determine

- Probability density function for the gain
- p(gain > 500)
- 90% confidence interval

Step 1: Determine

- Mean
- Standard Deviation
- Degrees of Freedom
  - *sample size minus 1*

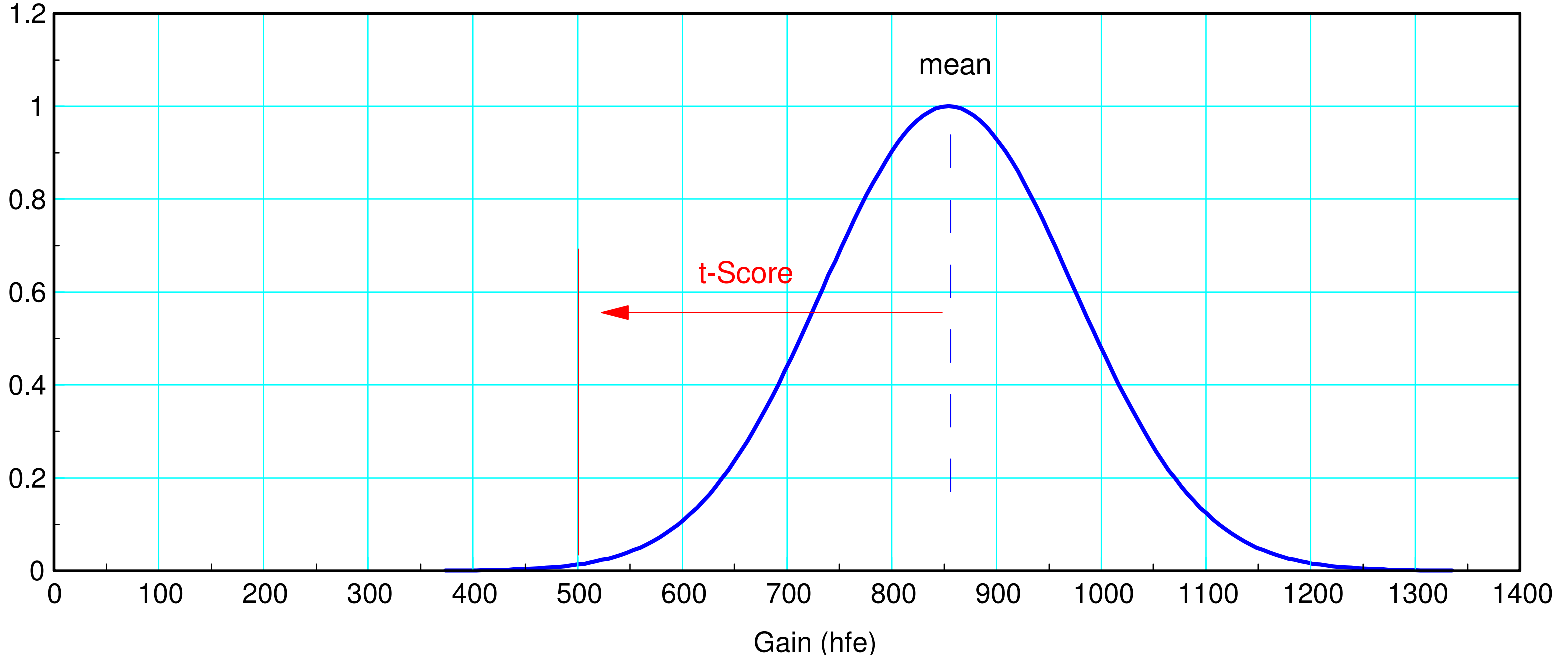
```
hfe = [  <paste data here> ]

x = mean(hfe)
x =  854.1290

s = std(hfe)
s =  120.2034

df = length(hfe) - 1
df = 61


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



What is the probability that the gain is more than 500?

Compute the t-score

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



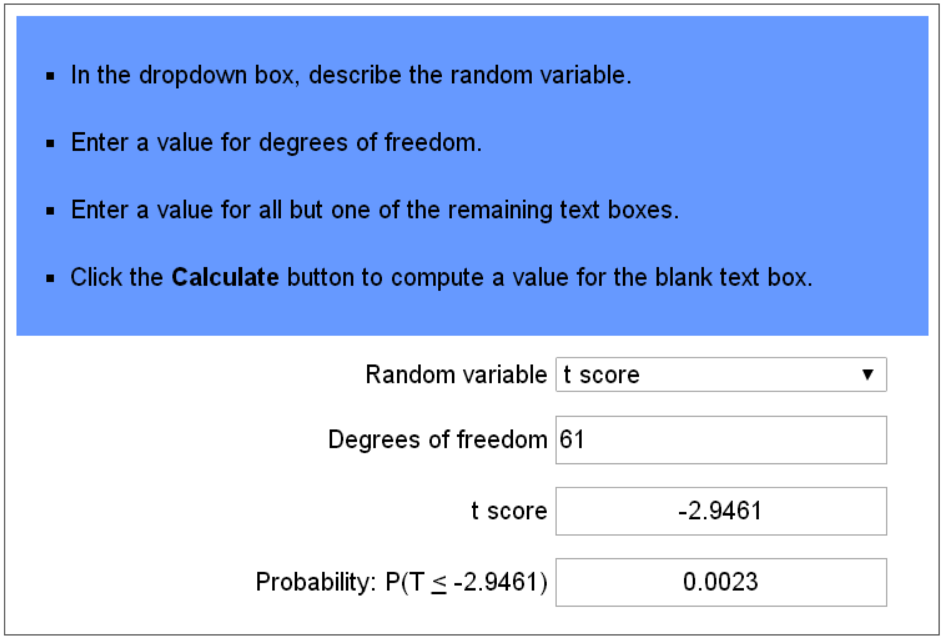
Convert t = -2.9461 to a probability (t-table)

The probability that the gain is less than 500 is 0.0023

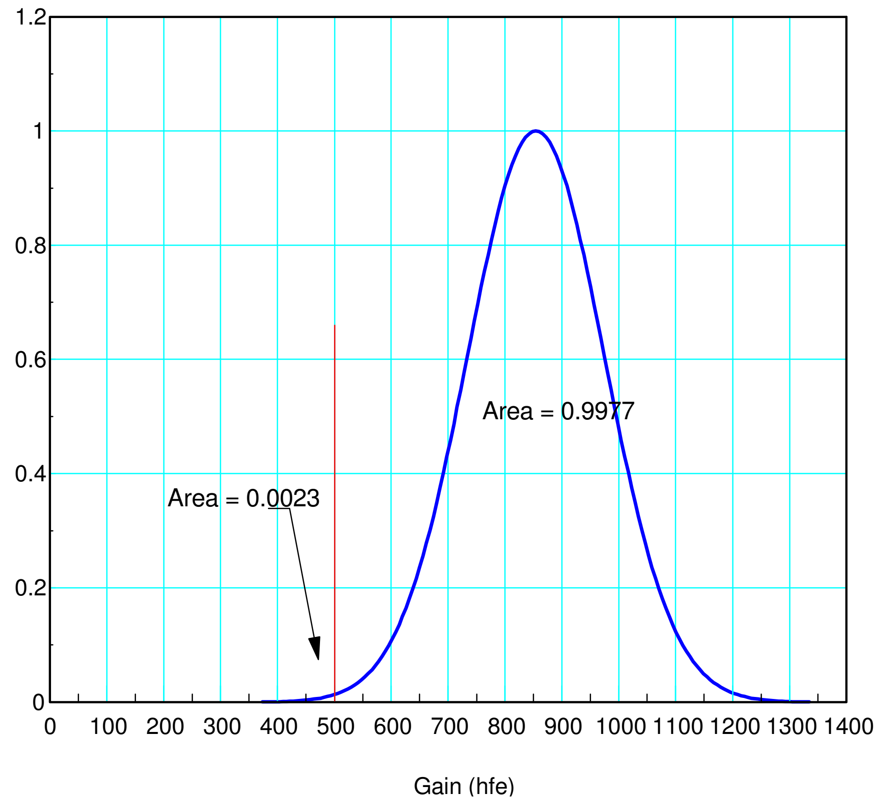
The probability that the gain is more than 500 is 0.9977

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

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



www.StatTrek.com

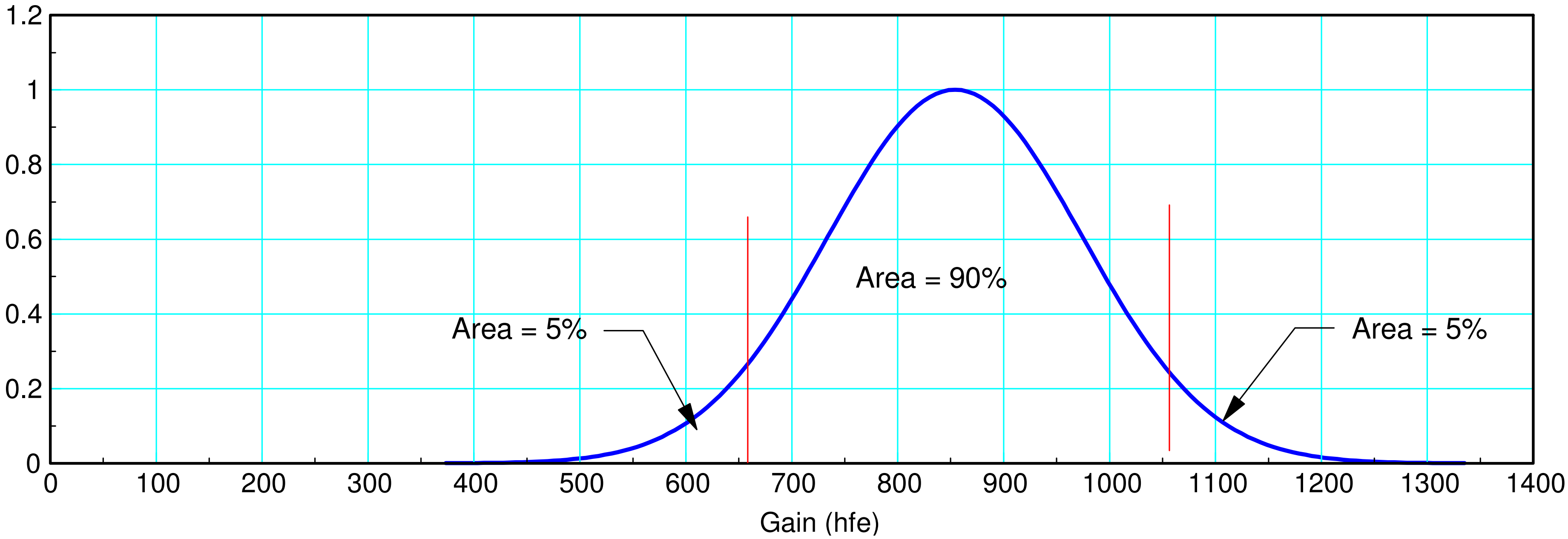


iii) What is the 90% confidence interval for the gain?

- Each tail is 5% (leaving 90% in the middle)
- 5% tails means t = 1.67

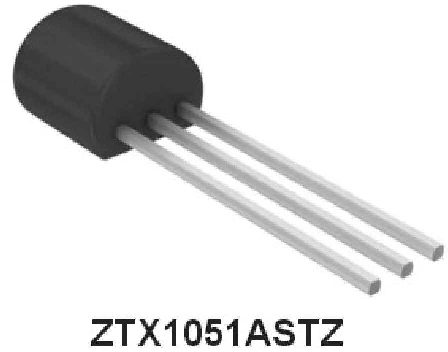
$$\bar{x} - 1.67s < gain < \bar{x} + 1.67s$$

$$653 < gain < 1055$$



Note: Individuals vs. Populations

- You know more about populations than individuals

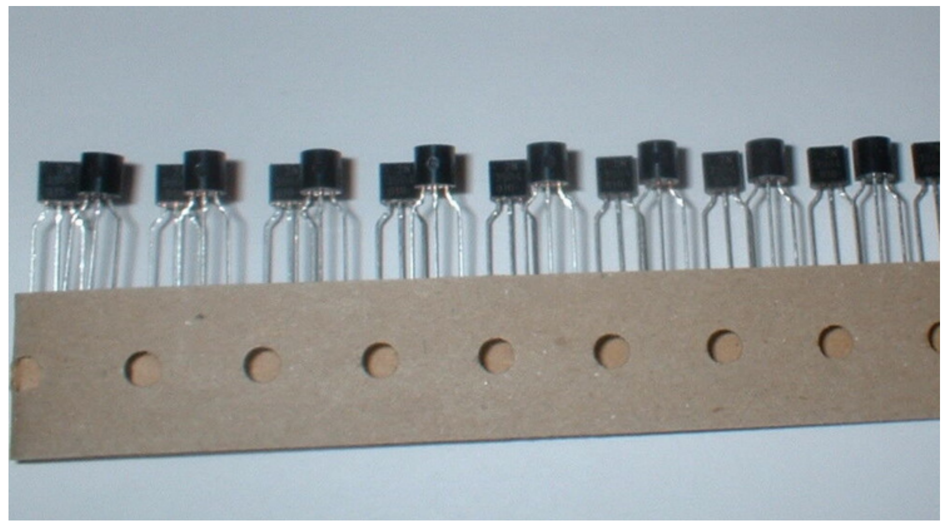


Individuals:

- $x_i \sim N(\mu, \sigma^2)$
  - *Normal distribution if you know them*
  - *t-distribution if you estimate them from data*
- Shape doesnt' change with sample size

Population:

- $\bar{x} \sim N\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right)$
- The more data you have, the more certain you are of the *true* mean
- As the sample size goes to infinity, you eventyally know the population mean *exactly*



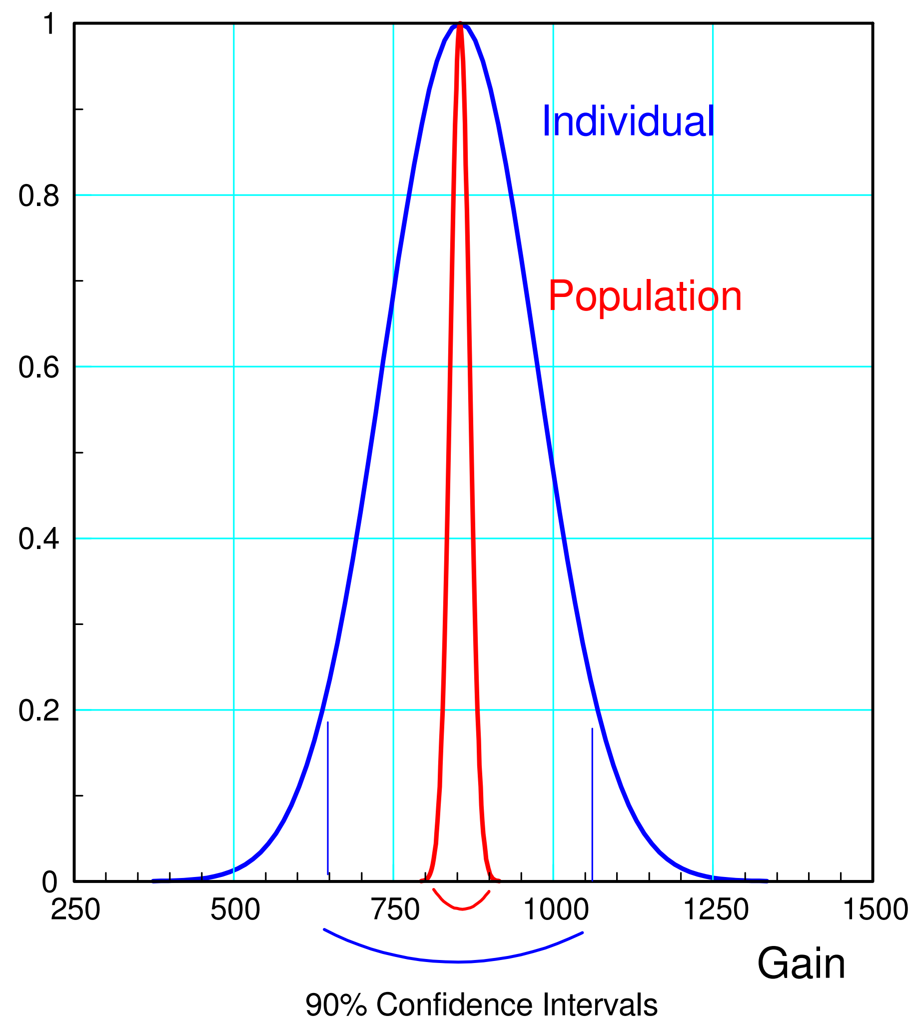
Example:

Individual: What is the 90% confidence interval for the gain of any given transistor?

- x = 854.129
- s = 120.2034
- 653 < gain < 1055

Population: What is the 90% confidence interval for the gain of *all* 1051a transtors?

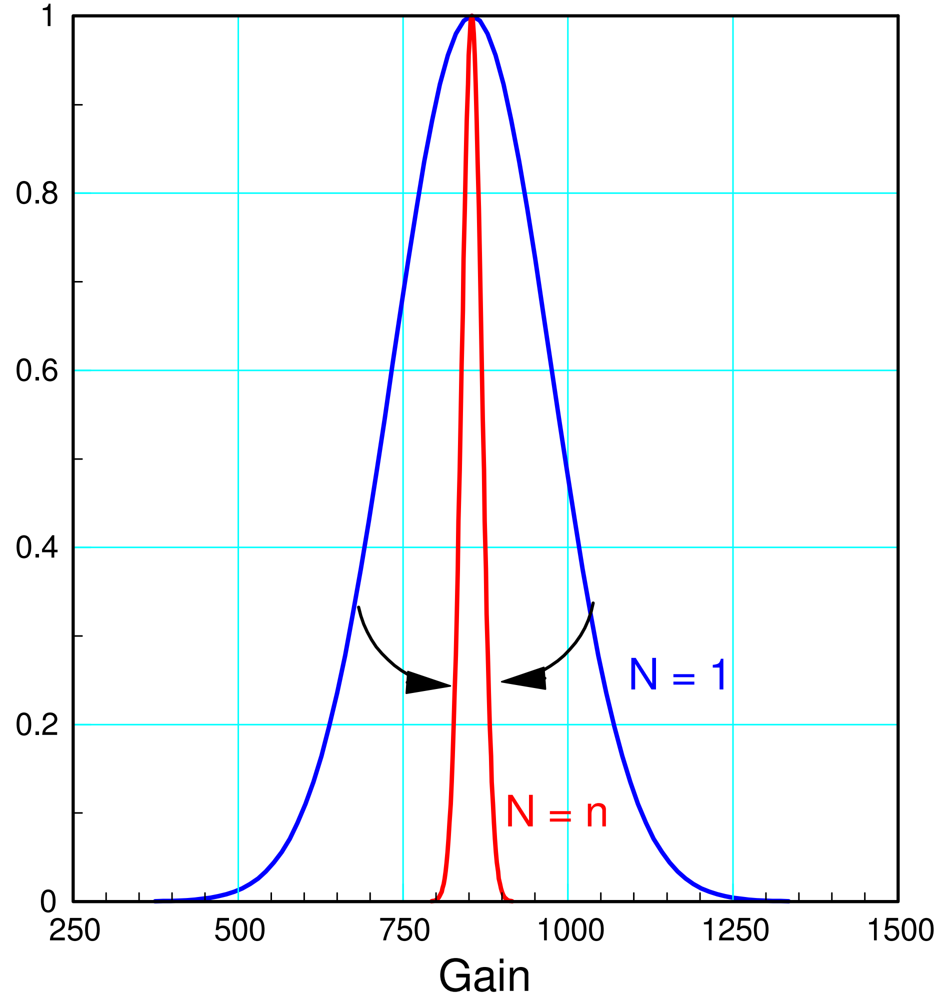
- x = 854.129
- $s = \frac{120.2034}{\sqrt{62}} = 15.27$
- 828.63 < gain < 879.63



How many transistors do I have to sample to know the gain within 10 with 90% certainty?

- t-score ≈ 1.67
- $1.67 \cdot \frac{120.2034}{\sqrt{n}} = 10$
- $n = 402.96$

If I sample 403 transistors, I will know the true mean within 10 with 90% certainty



Design of Experiment

"If you don't know where you are going, any road can take you there"

Lewis Carroll, Alice in Wonderland

Before starting an experiment, think about...

- 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?

Design of Experiment

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)

Energy in a AA battery

How much energy does a AA battery contain?

What data do we need?

Energy is hard to measure, Voltage is easy.

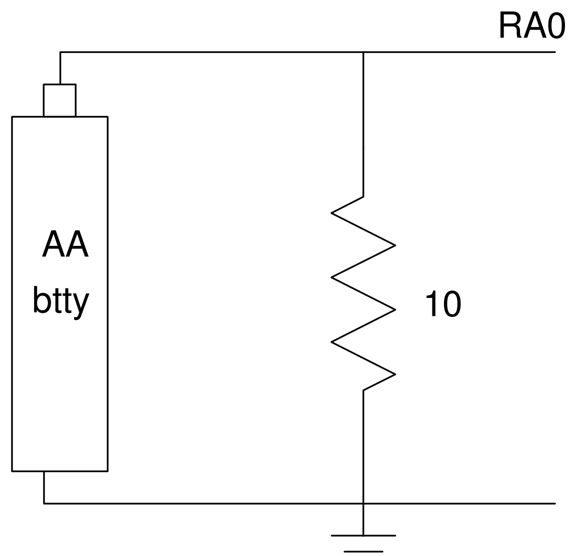
- Short the battery across a 10 Ohm resistor, and
- Measure the voltage every 6 seconds,

Measure voltage and computer power

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

Run experiment until battery is dead

$$E = \int P\, dt \qquad \text{Joules}$$



How Much Data do you Need?

- One data point (discharging one battery) is meaningless
- Two data points work but give a large t-value
- 10+ data points give diminishing returns
- *Make 4 measurements (3 degrees of freedom)*

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

Significance:

- 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?

Sloppy procedures give sloppy results

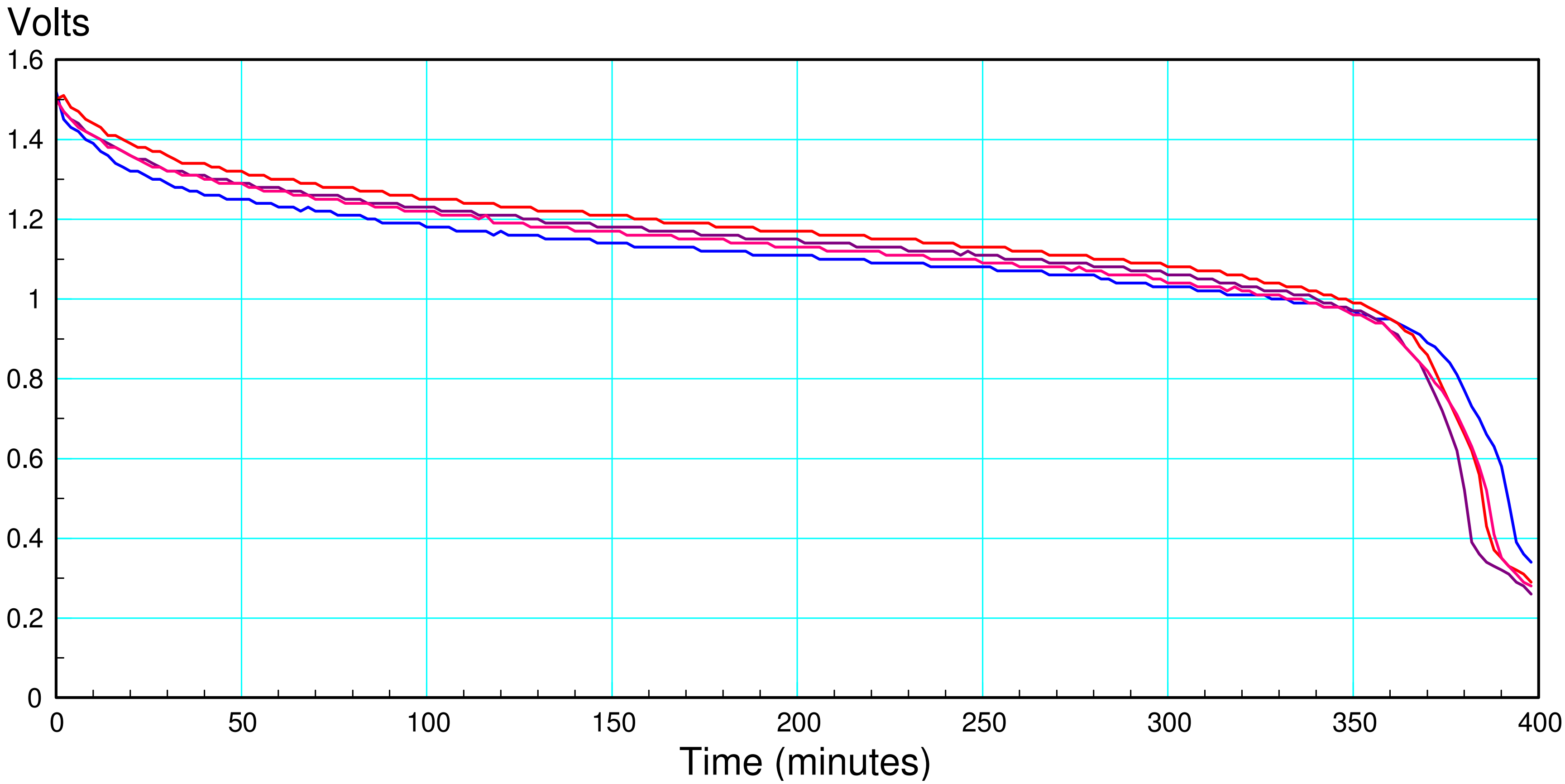
- Large standard deviation
- *Energy is in the range of -2000 Joules to +20,000 Joules*

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

- Measure the voltage of 4 batteries for 6+ hours



Step 3: Data Analysis

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.

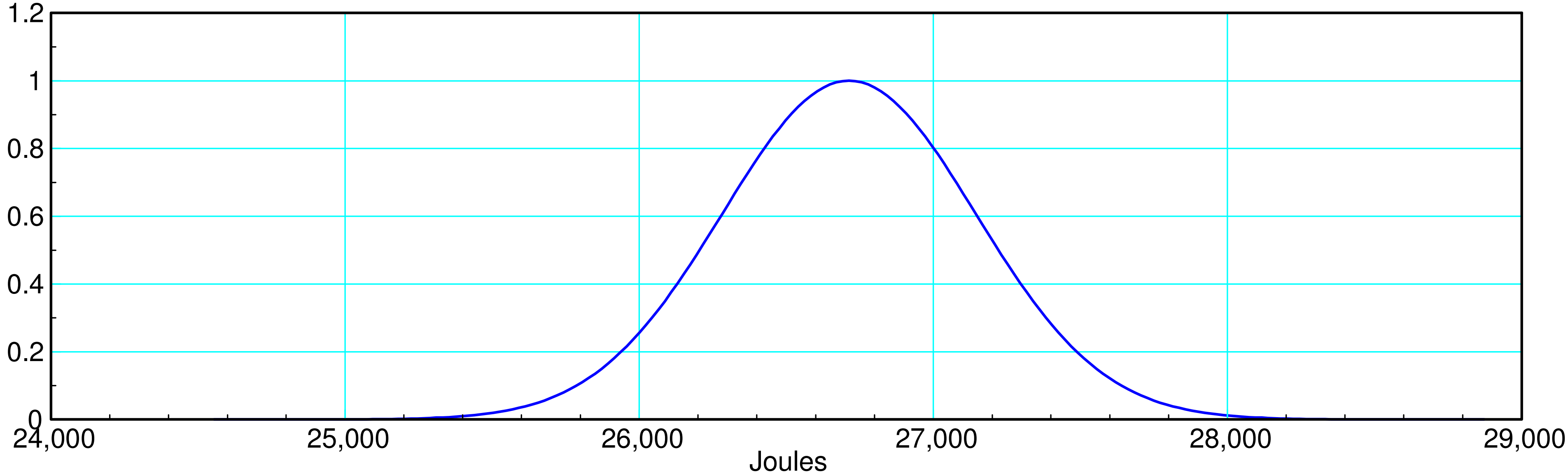
$$P = \frac{V^2}{R} = 0.1V^2 \text{ Watts}$$

$$E = 0.6 \sum (V^2) \text{ Joules}$$

E = {26,332 26,648 27,330 26,543} Joules

The mean & Standard Deviation are:

```
x = mean(Joules) = 26,713
s = std(Joules) = 431.6950
```



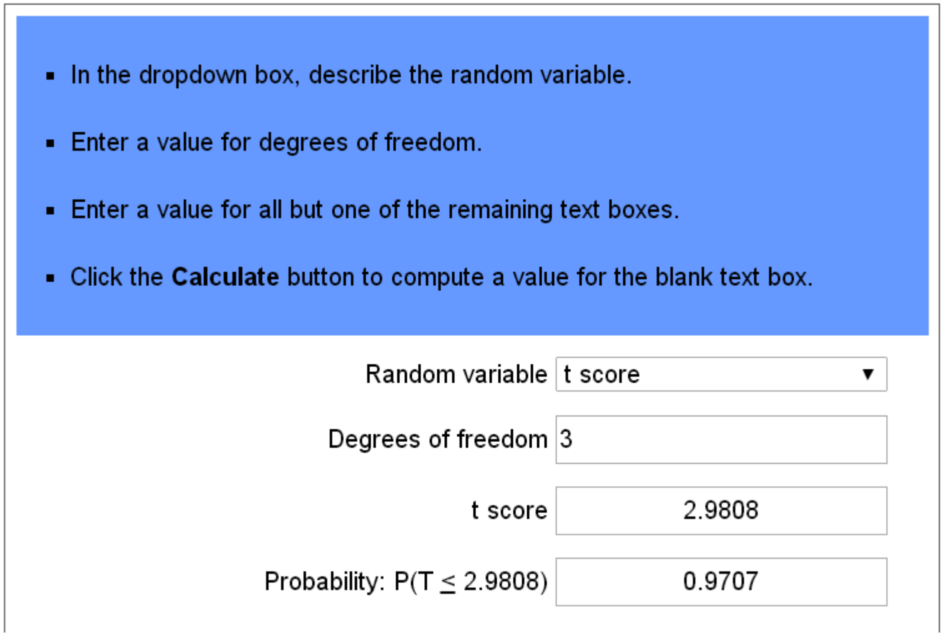
Energy in a AA battery: Mean = 26,713, standard deviation = 431 Joules

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.

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

$$t = 2.9808$$



Convert to a probability with a t-table

- p(energy < 28,000 Joules) = 0.9707
- p(energy > 28,000 Joules) = 0.0293

all probabilities add to one

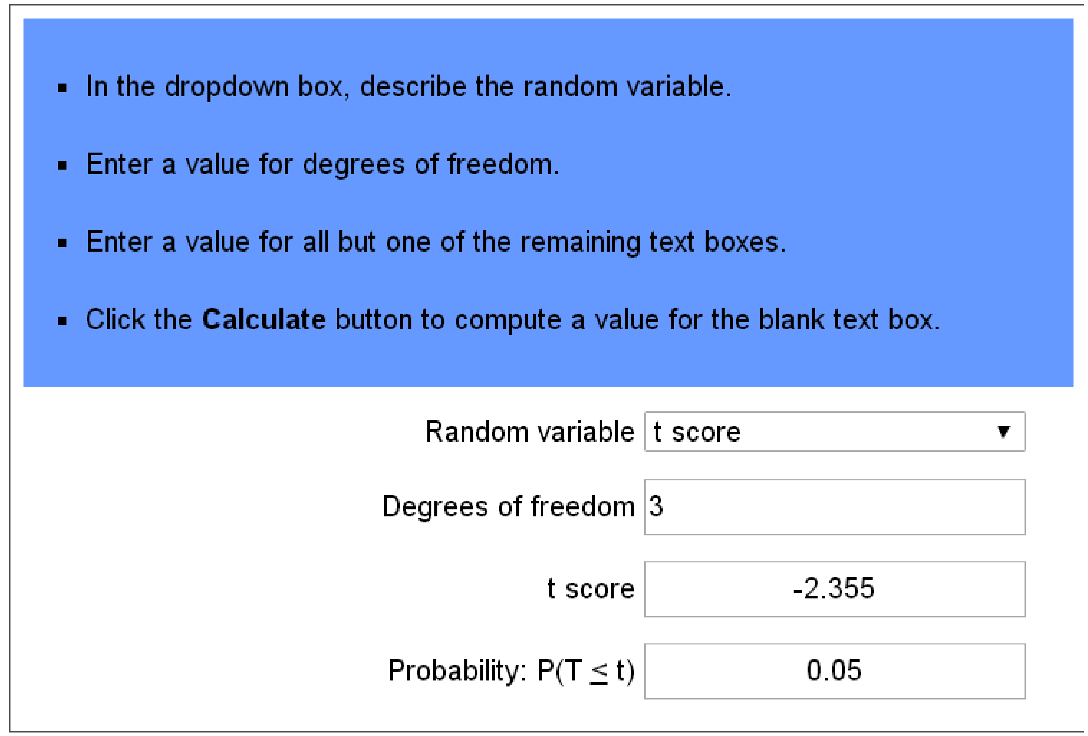
What is the 90% confidence interval for any given AA battery?

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

The 90% confidence interval will be

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

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

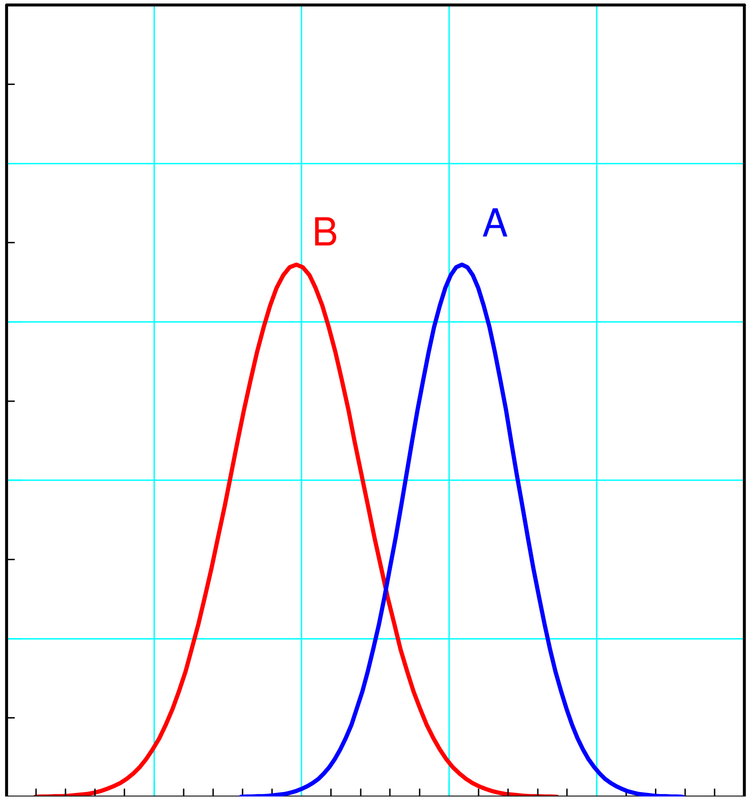


Comparison of Means

- Two **populations**: A and B
- Which has the higher mean?

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



Create a new variable, W

- W = A - B
- $\bar{x}_w = \bar{x}_a - \bar{x}_b \qquad s_w^2 = \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}$

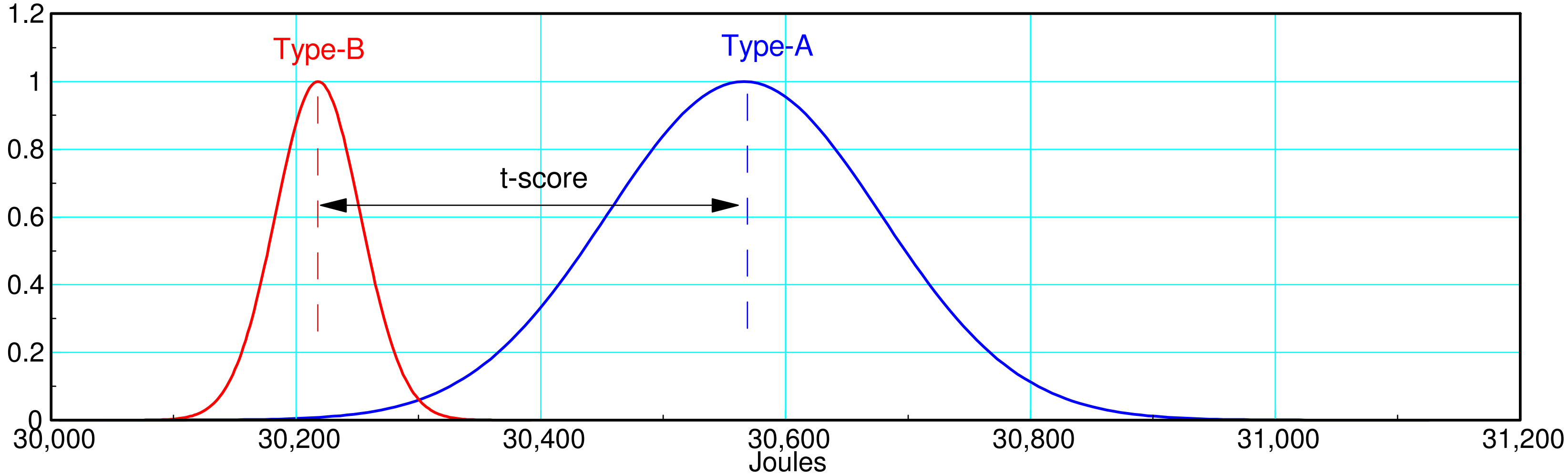
p(A > B) = p(W > 0)

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

Which battery has more energy:

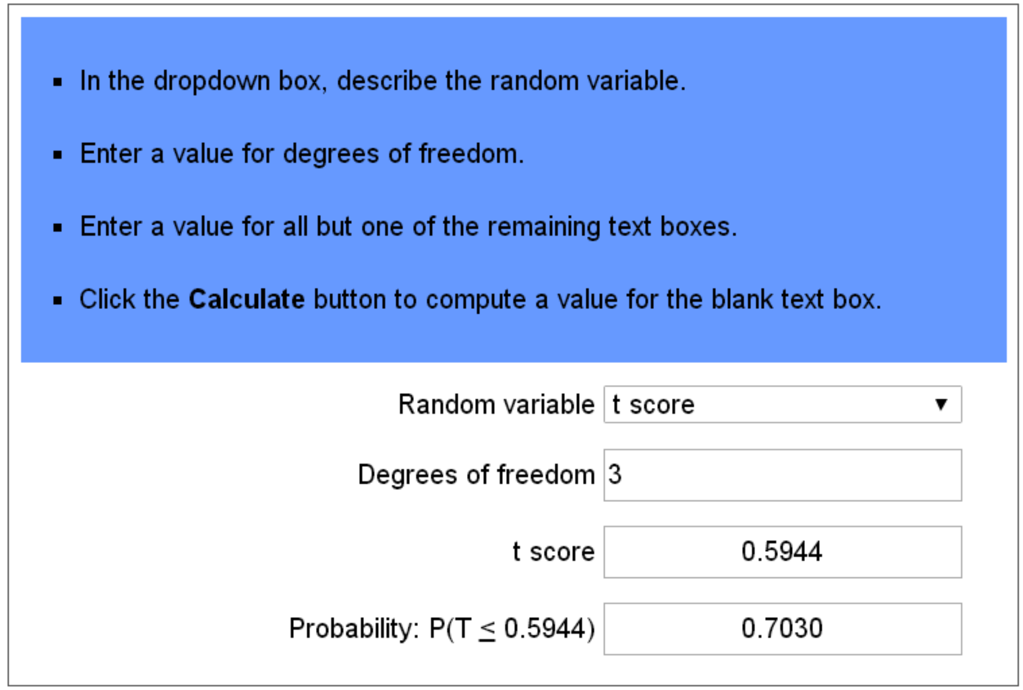
- A = {29,376 30,639 32,048 30,200} Joules
- B = {30,186 30,197 30,668 29,820} Joules
- W = A - B

	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



t-Score: p(A > B) = p(W > 0)

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



Convert to a probability

- p(A > B) = 70.30%