t Tests with a Single Population

ECE 341: Random Processes

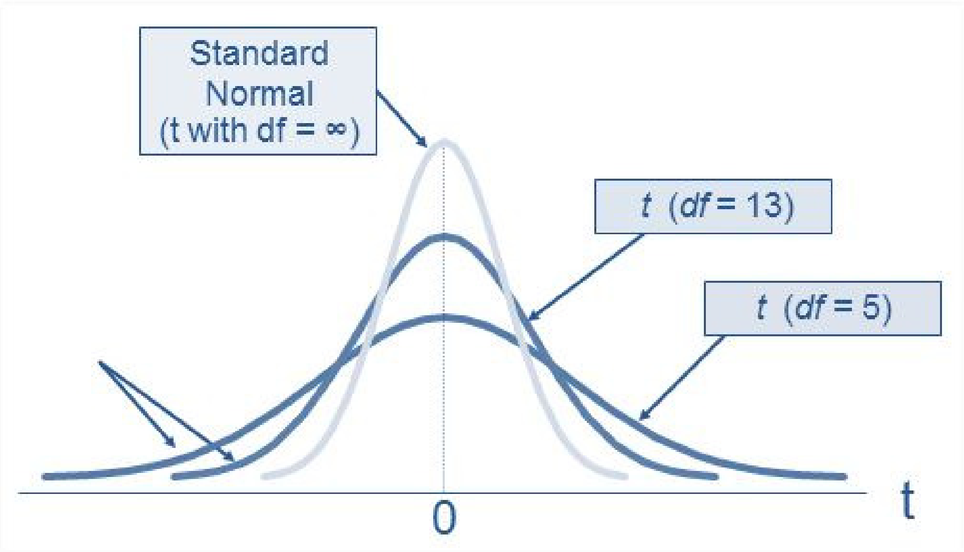
Lecture #26

note: All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

Student t Distribution

The previous lecture introducted the Student t-Test

- Assume a normal distibution
- Collect n samples
- Estimate the mean and variance



Using a t-test, you can find...

- $p(y > c)$
  - *One-Sided test*
- $p(a < y < b)$
  - *Two-sided test*
- The 90% confidence interval for the population's mean
  - *Test of a population*

This Lecture will Present Several Examples

Weather in Fargo

- Is June going to break 100F?
- What's the 90% confidence interval?

10d6:

- What's the probability that 10d6 > 44?
- What's the 90% confience interval?
- What's the 90% confidence interval for the mean?

5-Card Stud Poker

- What's the confidence interval for the probability of a full-house

Do 1k 5% Resistors Have a Uniform Distribution?

- Is the mean what you would expect?
- Is the variance what you would expect?

Weather in Fargo

Previous lectures looked at the monthly high in Fargo.

- Data from Hector Airport
- Data collected since 1942

We assumed the data was a normal distribution

- known mean
- known standard deviation

This is *actually* a t-distribution

- Finite sample size

Month	mean	st dev	npt
Jan	38.5892	6.4993	83
Feb	41.1602	7.3822	83
Mar	56.0964	10.8387	83
Apr	77.8675	7.8706	83
May	87.8554	4.5455	83
June	92.2422	4.6143	83
July	94.6627	3.9549	83
Aug	94.5795	4.5300	83
Sep	89.6386	5.5800	83
Oct	79.7108	6.9097	83
Nov	59.5060	7.4431	83
Dec	41.9036	6.8103	83

Is June going to break 100F?

- Single sided t-Test
- Very similar to previous analysis

Step 1: Find the mean, std, sample size

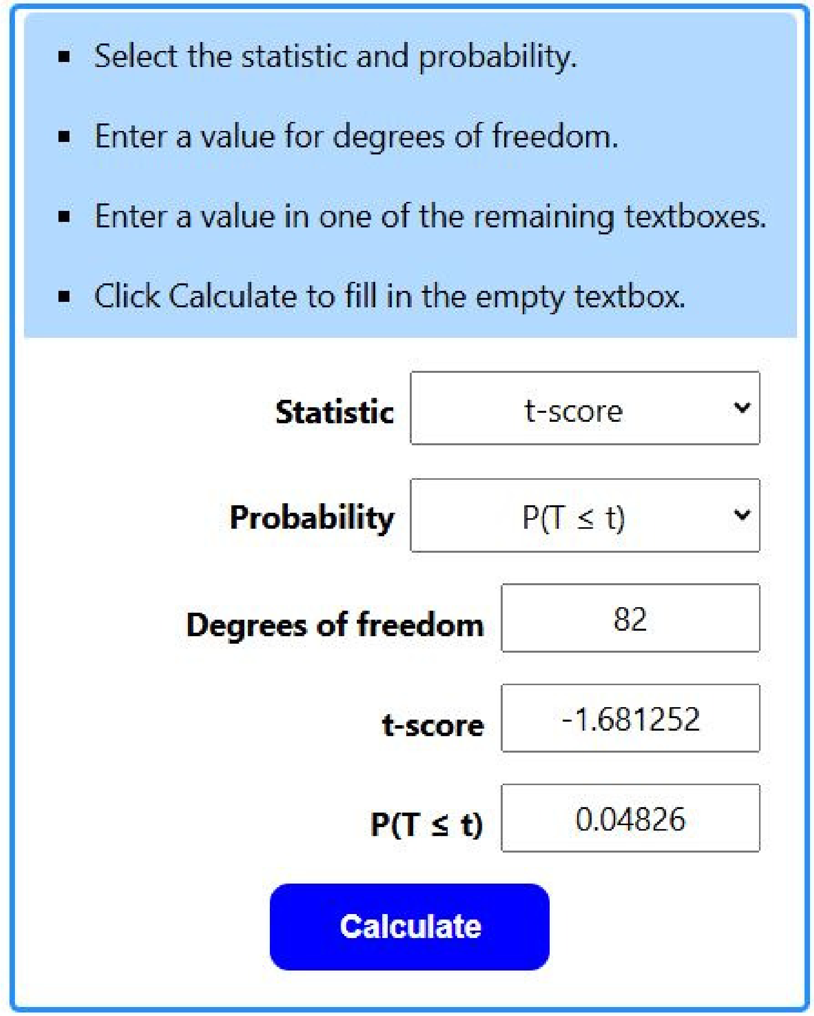
- $\bar{x} = 92.2422F$
- $s = 4.6143F$
- $n = 83$

Step 2: Find the t-score

$$t = \frac{100 - 92.2422}{4.6143} = 1.6812$$

Step 3: Convert to a probability

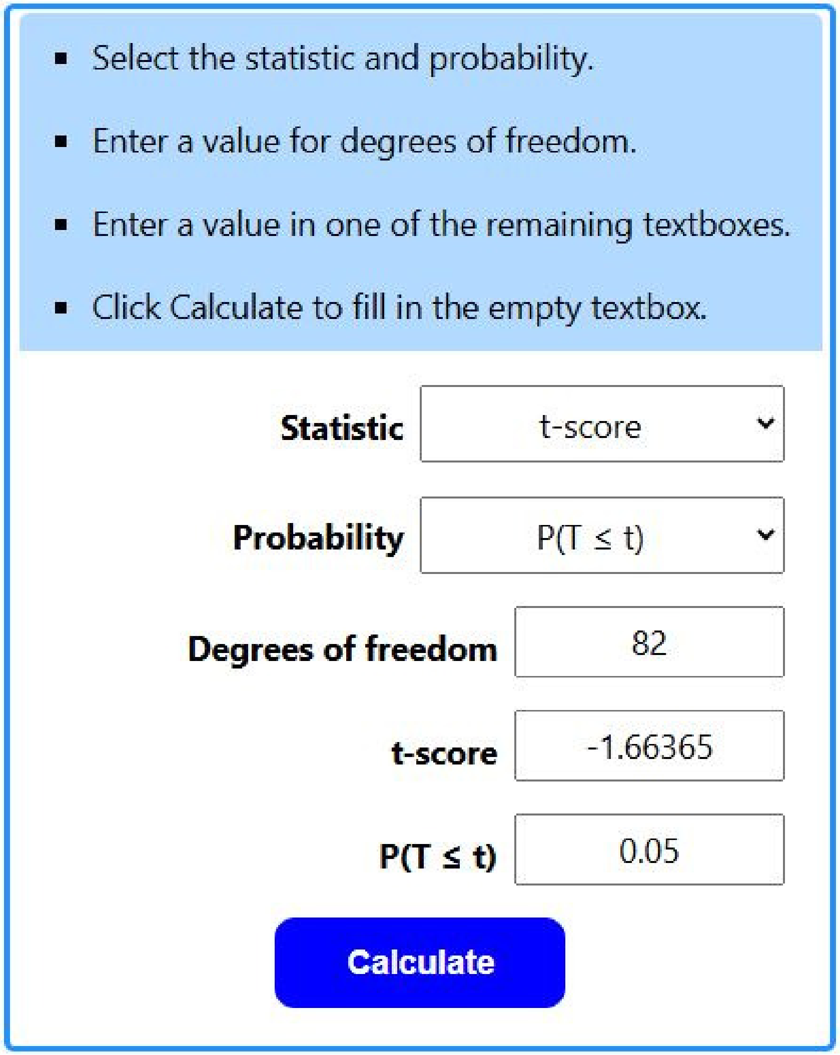
- StatTrek
- p = 0.04826
- Slightly different than before
  - *This is a t pdf not a Normal pdf*



What is the 90% Confidence Interval?

Step 1: Find the t-score for 5% tails

- StatTrek
- t = 1.66365
- Slightly different than before
  - *This is a t distribution*
  - *Not a normal distribution*



Step 2: Find the 90% confidence interval

$$\bar{x} - 1.66365s < high < \bar{x} + 1.66365s$$

$$84.56F < high < 99.92F$$

What if you only use the last 5 years of data?

- Climate is changing

Using just the last 5 years

- $\bar{x} = 99.2000F$
- $s = 2.2804F$
- $n = 5$

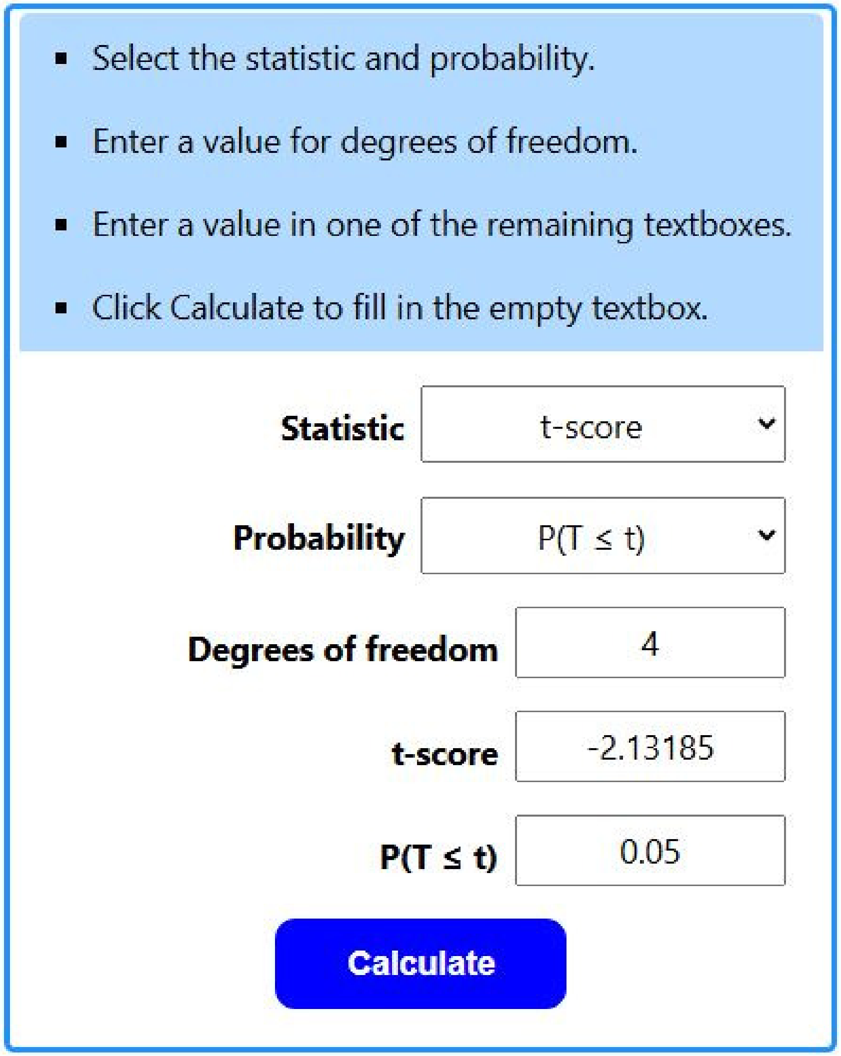
Find the t-score for 5% tails

- $t = 2.13185$
- Changes with sample size

Find the 90% confidence interval

$$\bar{x} - 2.13185s < high < \bar{x} + 2.13185s$$

$$94.33F < high < 104.06$$



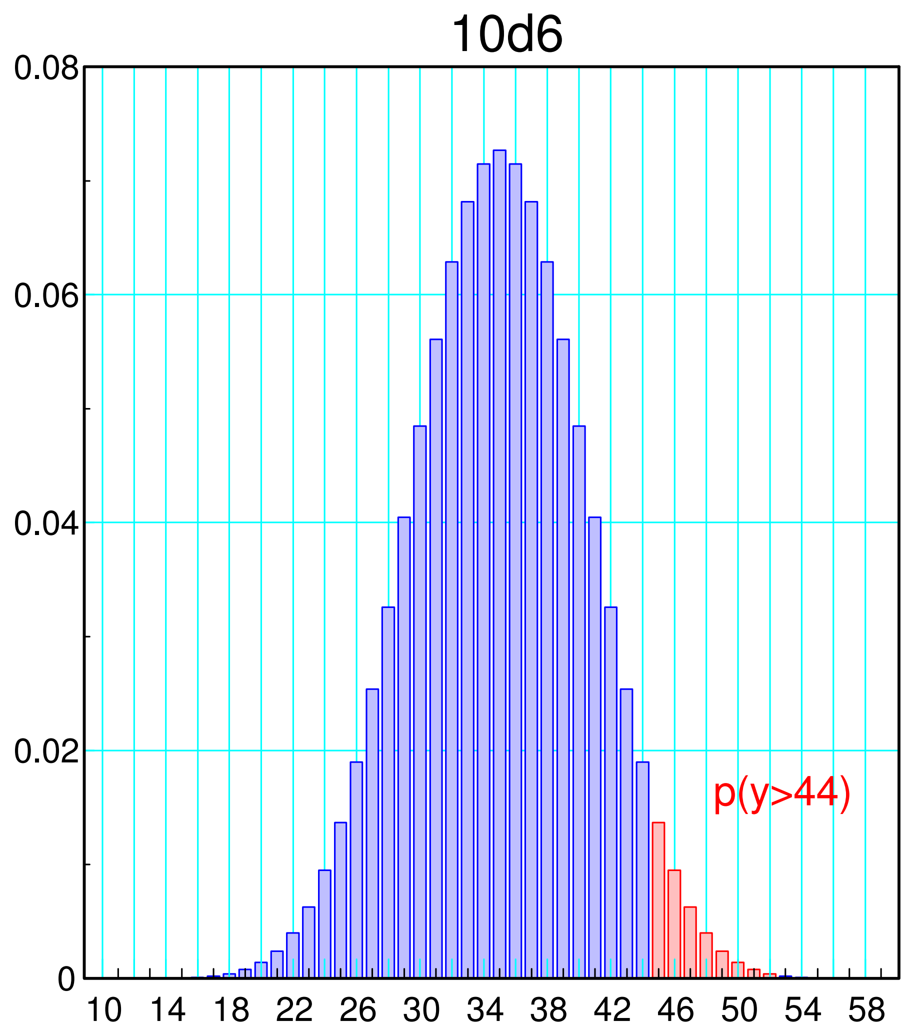
Y = 10d6:

Start with the ten 6-sided dice

The probability that Y > 44 can be computed using colvolution

```
>> d6 = [0,1,1,1,1,1,1]' / 6;
>> r2 = conv(d6,d6);
>> r4 = conv(r2,r2);
>> r8 = conv(r4,r4);
>> r10 = conv(r2,r8);
>> sum(r10(46:61))

ans =     0.0390
```

Can you get similar results using a small sample size and a t Test?



Step 1: Collect Data

Let n=10

- sample size = 10

From the sample, compute

- the mean and
- the variance

The results vary

```
n = 10;
Y = zeros(n,1);

for i=1:n
   d6 = ceil( 6*rand(1,10) );
   Y(i) = sum(d6);
end

x = mean(Y)
s = std(Y)
dof = n-1
```

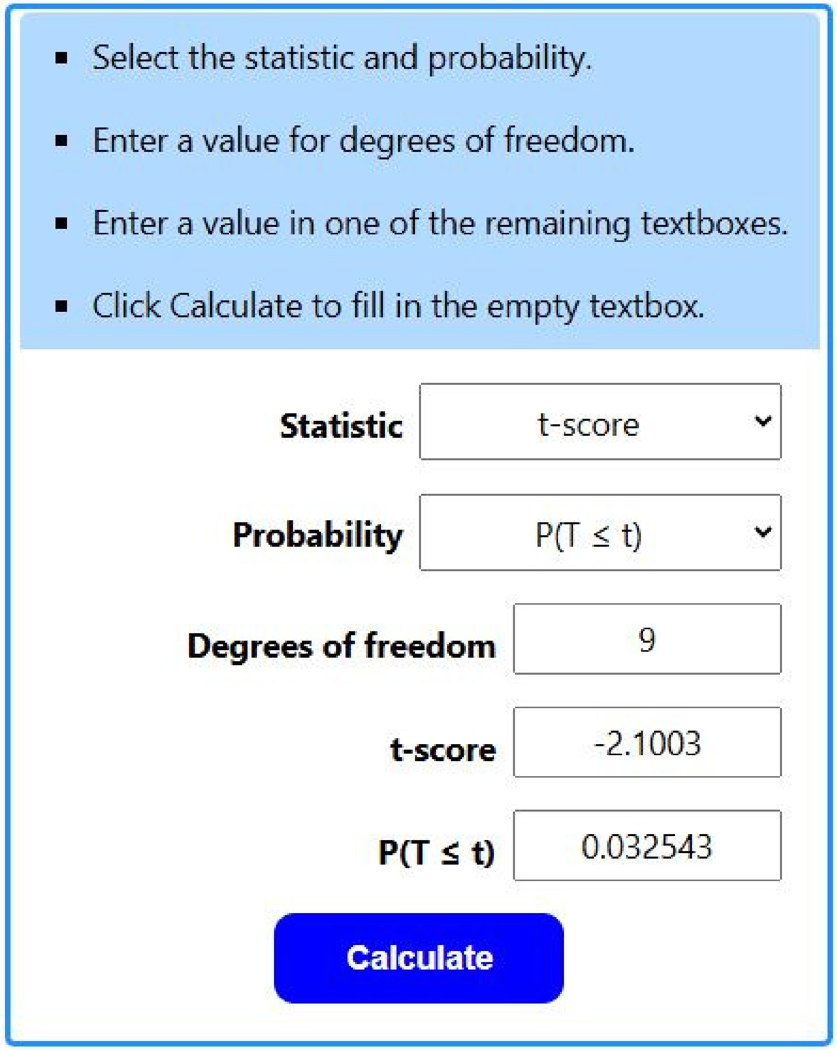
```
x =   33.6000
s =    5.1897
dof =  9
```

Step 2: Compute the t-score

```
>> t = (44.5 - x) / s

t =     2.1003
```

Once computed, convert the t-score to a probability using a Student t Table

- or StatTrek



$p(10d6 > 44.5) = 3.2543\%$

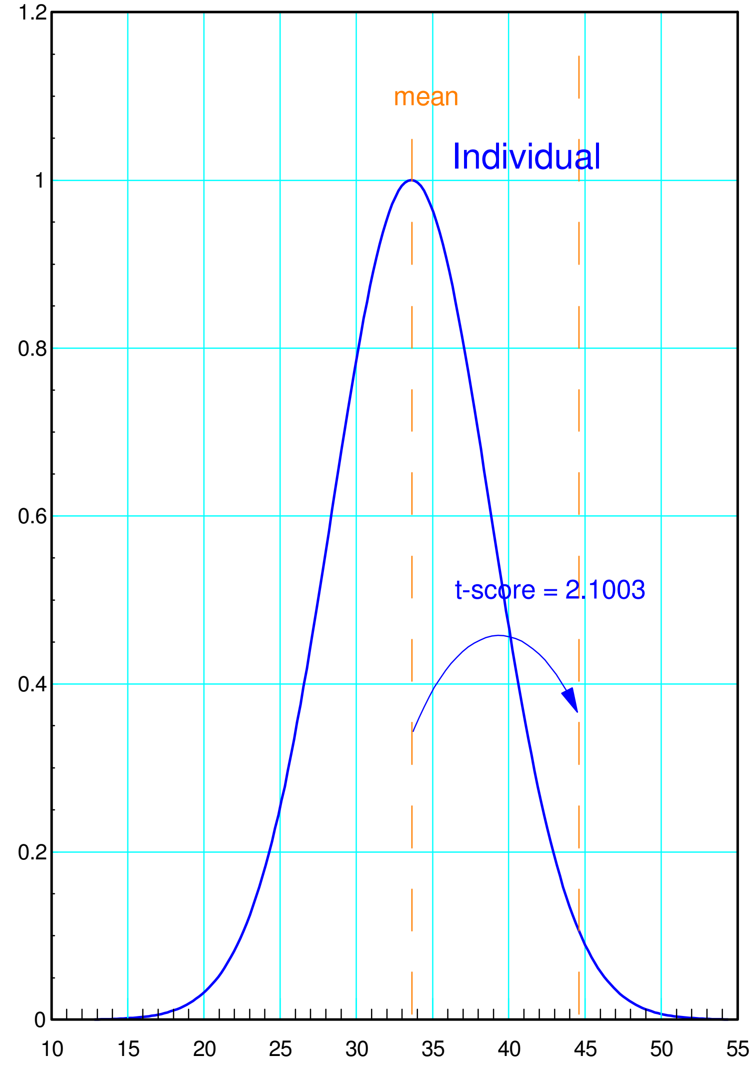
What's Going On?

The t-score is very similar a z-score

- Distance from the mean
- In terms of standard deviations

A table is similar to a Normal table

- It converts t-scores to probabities
- The area to the left of the t-score
- Also takes sample size into account



What's the probability that the *mean* of 10d6 > 44?

Population question

- Divide the variance by the sample size

Results in the t-score be larger

- by the square root of the sample size
- t = 6.6418

Results in smaller tails

- We know more about populations than individuals

```
n = 10;
Y = zeros(n,1);

for i=1:n
   d6 = ceil( 6*rand(1,10) );
   Y(i) = sum(d6);
end

x = mean(Y)
s = std(Y) / sqrt(n)
dof = n-1

x =   33.6000
s =    1.6411
dof =  9

t = (44.5 - x) / s

t =    6.6418
```

Convert t Score to a Probability

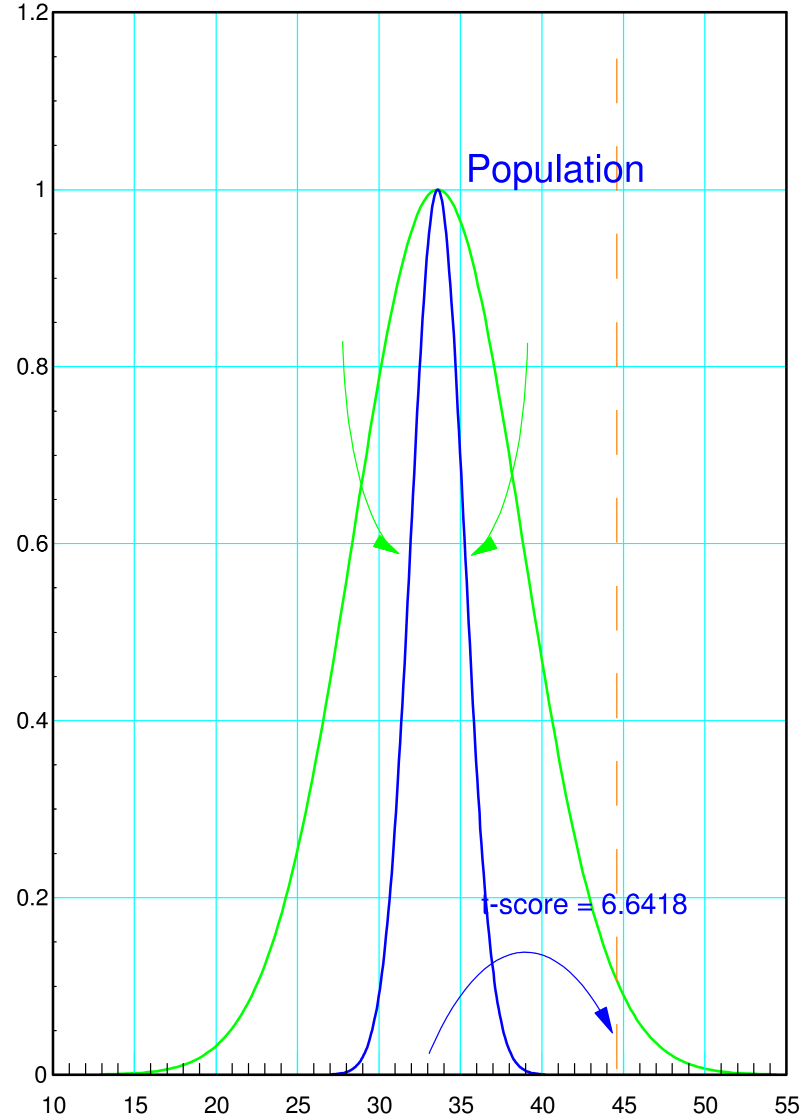
Using StatTrek

- t score is 6.6418
- p = 0.000047

Meaning

- The chance that the *population's* mean is more than 44.5 is 0.000 047

You know more about populations than individuals.



What's the 90% confidence interval for y = 10d6?

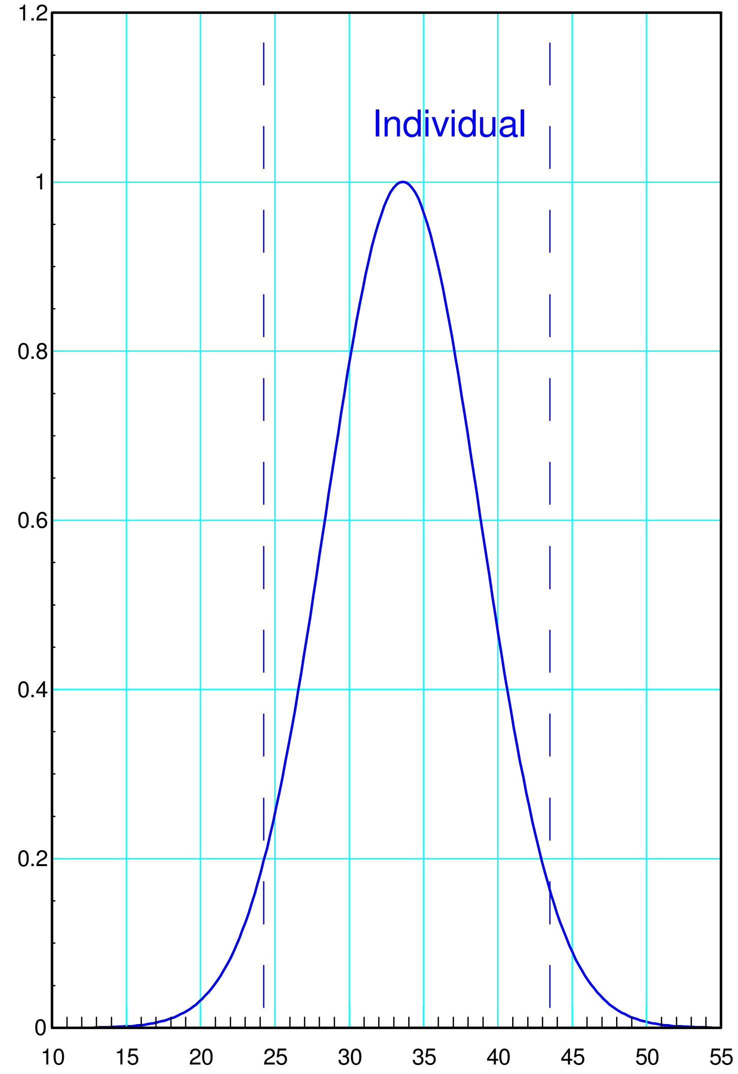
This is a 2-sided test.

- Find the t-score for 5% tails
- 9 degrees of freedom
- t = 1.83311
  - *StatTrek*

The confidence interval is

$$\bar{x} - t \cdot s < y < \bar{x} + t \cdot s$$

$$24.08 < y < 43.11$$

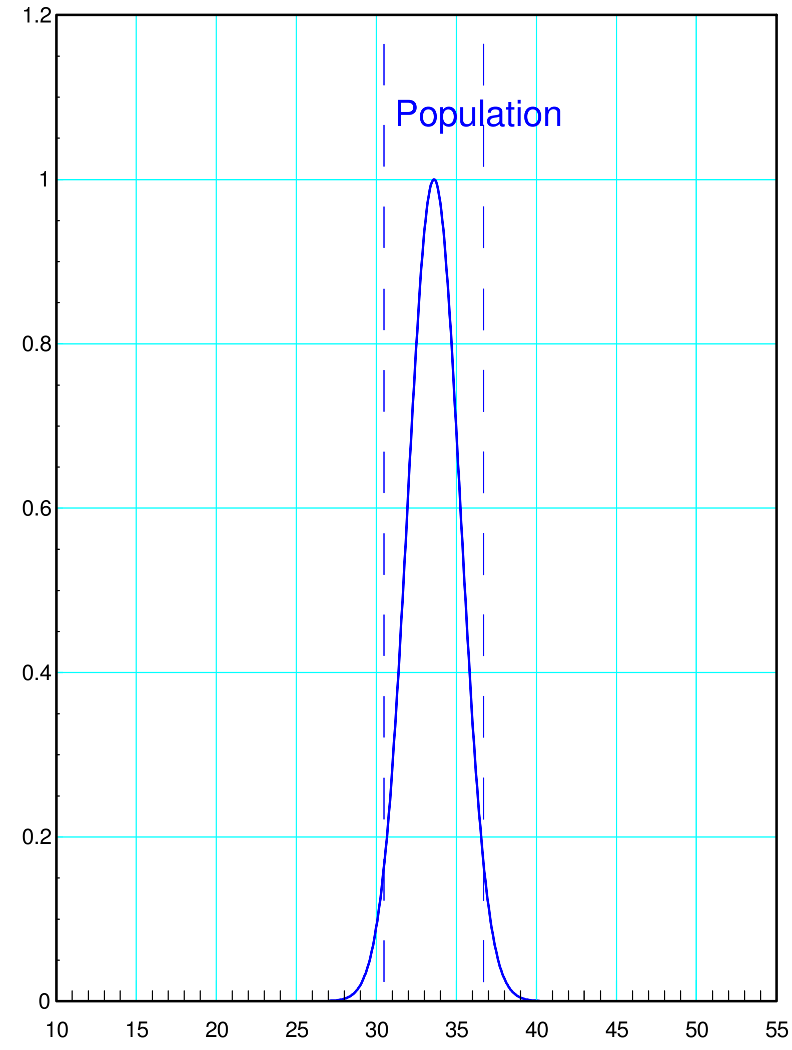


What's the 90% confidence interval for the *mean* of 10d6?

- Population question
- Divide the variance by sample size

$$\bar{x} - t \cdot \frac{s}{\sqrt{10}} < y < \bar{x} + t \cdot \frac{s}{\sqrt{10}}$$

$$30.59 < mean < 36.61$$



Note

- As the sample size goes to infinity
- The spread goes to zero

5-Card Stud Poker

Earlier in the semester we wrote Matlab programs to deal random poker hands

- Monte Carlo experments with 100,000 hands per trial

What is the 90% confidence interval for

- The number of full-houses each time you run the experiment?
- The actual odds for being dealt a full-house?

	St-Fl	4ok	Full-Hou	Flush	Str	3ok	2-Pair	Pair	High-C
Calc	1.53	24.01	145.21	196.54	392.46	1,997.41	4,753.9	42,256.9	50,117.73
run1	1	27	124	218	402	2,175	4,689	42,187	50,177
run2	1	20	144	203	423	2,145	4,800	42,219	50,145
run 3	1	36	153	203	411	2,090	4,767	42,362	49,977

Number of Hands Containing a Full House

Individual question

From the data

- Find the mean and variance
- Find the t-score for 5% tails
  - *StatTrek*
  - *t = 2.91999*
- Find the 90% confidence interval

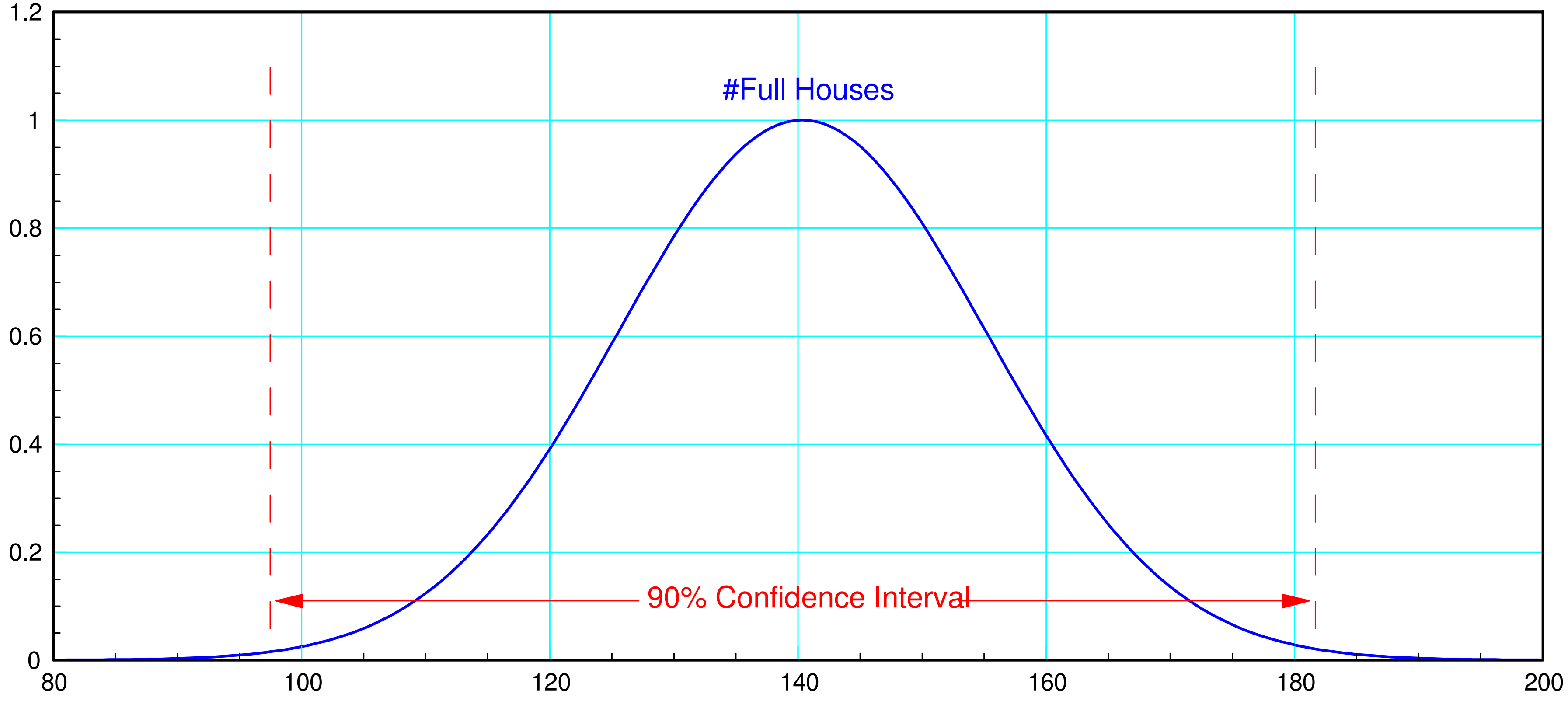
Result

- 96.99 < # hands < 183.68

```
>> Data = [124, 144, 153];
>> x = mean(Data);
>> s = std(Data);

>> high = x + 2.91999*s
high =  183.6766

>> low = x - 2.91999*s
low = 96.9901
```

pdf for Number of Full Houses in 100,000 Poker Hands

- along with 90% confidence interval



Probability of Being Dealt a Full House

Population question

- actual = 145.21 in 100,000 hands

From the data

- Find the mean and variance
  - *Divide variance by sample size*
- Find the t-score for 5% tails
  - *StatTrek*
  - *t = 2.91999*
- Find the 90% confidence interval

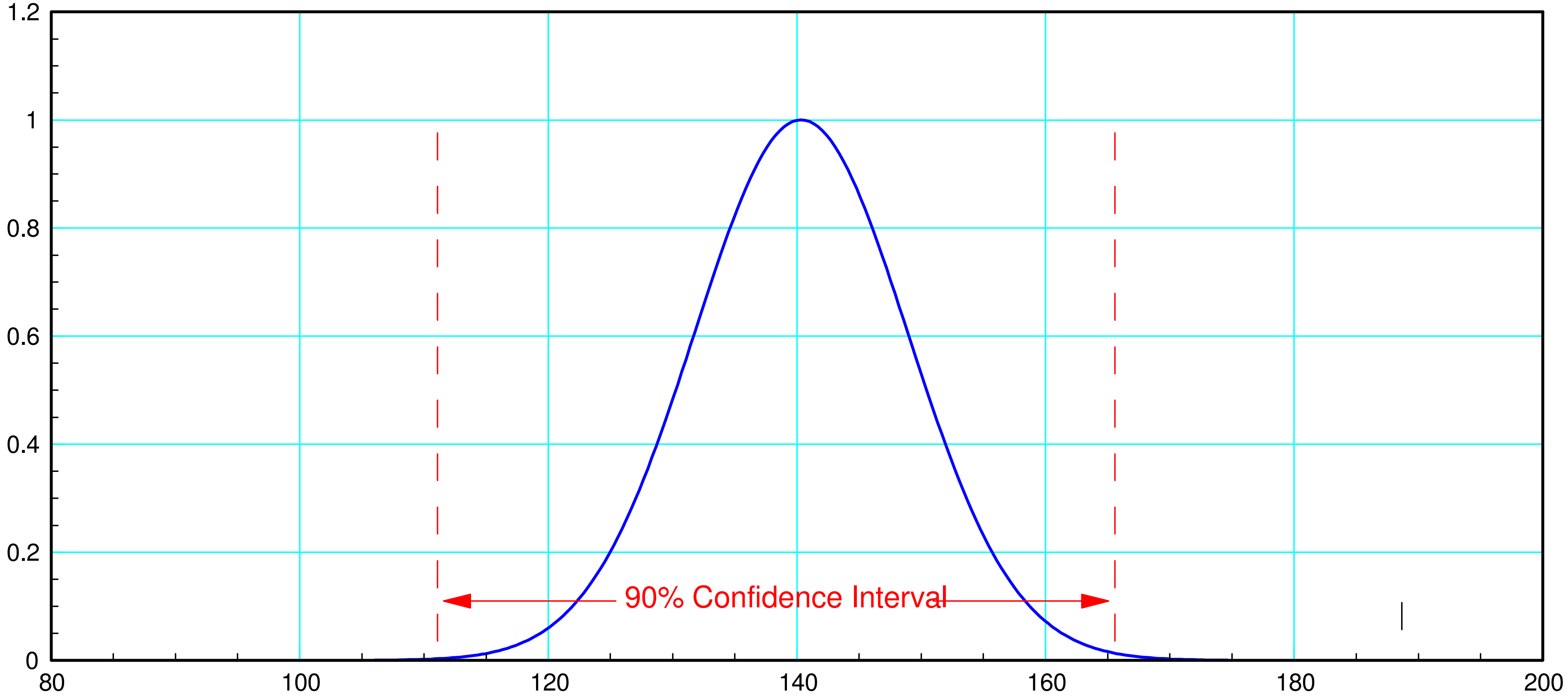
Result

- 115.31 < p(full house) < 165.36

```
>> Data = [124, 144, 153];
>> n = length(Data);
>> x = mean(Data);
>> s = std(Data) / sqrt(n);

>> high = x + 2.91999*s
high =  165.3576

>> low = x - 2.91999*s
low = 115.3091
```

pdf for p(Full Houses) in 100,000 Poker Hands

- Variance is divided by 3 (sample size)
- More Monte Carlo runs would produce a tighter estimate

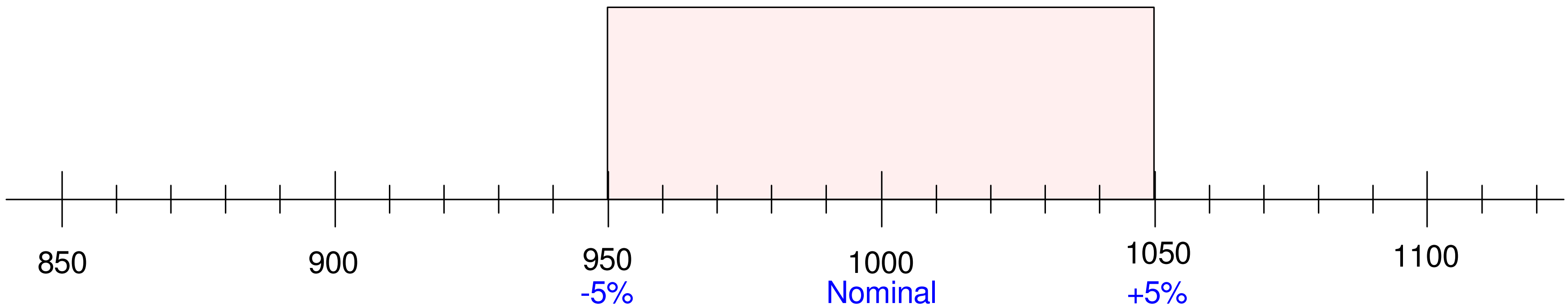


Do Resistors Have a Uniform Distribution?

Previous lecture

- 1k, 5% tolerance resistor
- Model as a uniform distribution

Is this model correct?

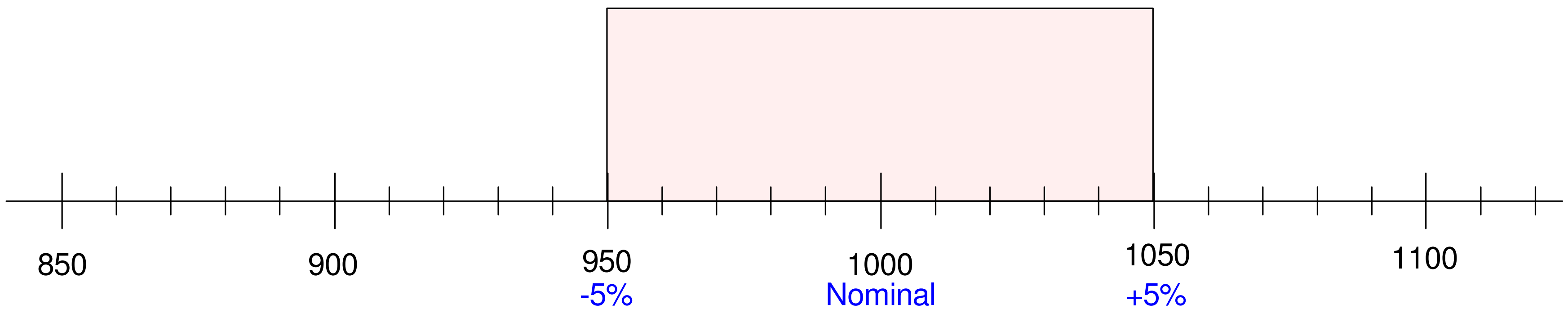


How to Test?

There are several ways to answer this question using a t-Disribution.

- If the assumed distribution is correct, the mean should be 1000 Ohms.
- If the assumed distribution is correct, the standard deviation should be 28.86 Ohms

Given some data, I can check each of these.



Step 1: Collect data.

Measure 56 resistors

```
989, 996, 993, 991, 993,  991, 997, 996, 995, 995, 991, 997,
1008, 995,996, 995, 996, 995, 998,  996, 995, 990, 981, 988,
994, 999,  990, 992, 997, 992, 995, 994, 990,  990, 994, 992,
996, 992, 992, 994,  995, 988, 984, 993, 992, 994, 999, 1000,
994, 995, 990, 997, 991, 993,  992, 993
```

Step 2: Analysis.

- Is the population's mean 1000 Ohms?

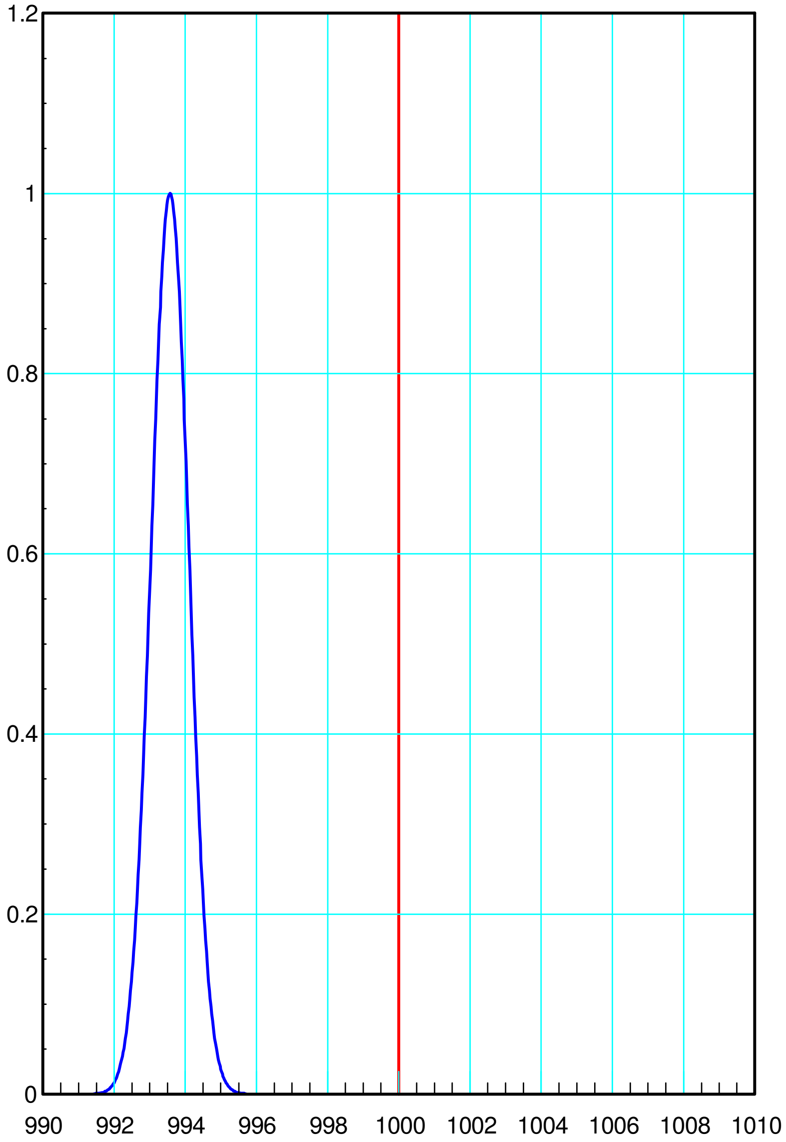
In Matlab:

```
>> x = mean(R)
x =  993.5714

>> s = std(R)
s =     3.9811

>> n = length(R)
n =     56
```

90% confidence interval

```
>> high = x + 1.6733*s/sqrt(n)
high =    994.4616

>> low = x - 1.6733*s/sqrt(n)
low =  992.6812
```

Nope



Is the standard deviation 28.86 Ohms?

- Uniform distribution over (-5%. +5%)
- Standard deviation should be 28.86

```
>> x = mean(R)

x =  993.5714

>> s = std(R)

s =    3.9811
```

How to you check this?

- Standard deviation is actually a gamma distribution
  - *can't be negative*
- Use t-test anyway

Using the entire population doesn't work

- Gives a single number
- Can't do statistics with a single number

Is the standard deviation 28.86 Ohms?

- Take 2
- Split data into four populations

Analyze the data

- Find the standard deviations
- 4 data points

For the 4 resulting data points

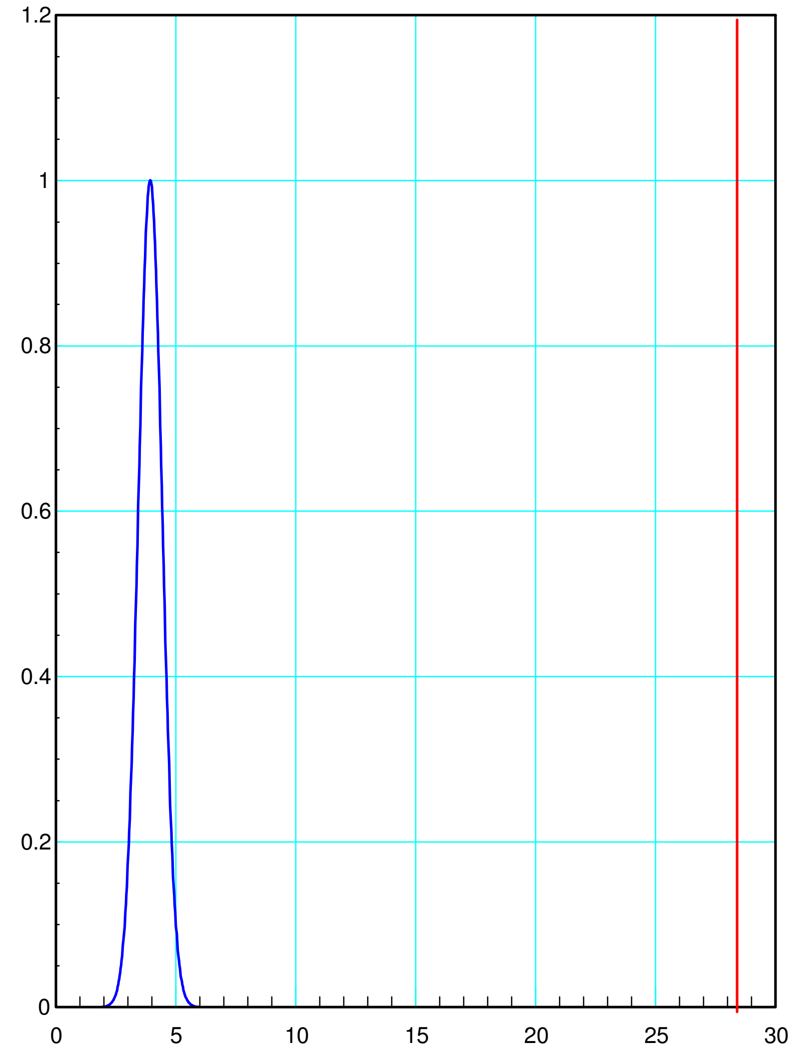
- Find the mean
- Find the standard deviation

```
>> s1 = std(R1);
>> s2 = std(R2);
>> s3 = std(R3);
>> s4 = std(R4);

>> Data = [s1,s2,s3,s4]

Data =    4.5603
          4.7097
          2.5257
          3.9342

>> x = mean(Data)
x =    4.0664

>> s = std(Data)
s =    1.7102
```

Is the standard deviation 28.86 Ohms?

Find the 90% confidence interval

```
>> x = mean(Data)
x =     3.9325

>> s = std(Data)
s =     0.9962

>> x + 2.35336*s
ans =      6.2769

>> x - 2.35336*s
ans =      1.5880
```



Check

- Nope: 28.86 isn't anywhere close

R does *not* have asumed distribution

More Data

A common theme with t-tests is that more data is good.

More data gives you

- A better estimate of μ and σ^2
- A smaller estimate of the population's μ
- A smaller t-score

But...

- Data can be difficult or expensive to obtain

How can you increase the sample size?

Example: Full-House in 6-Card Stud Poker

A program to deal out 100,000 hands of 6 cards was written.

The number of hands with a full-house were:

811	805	809	804	837
830	841	770	889	821
850	754	786	763	754
855	785	724	815	823

What's the 90% confidence interval for being dealt a full-house?

Solution: t-Test

```
>> x = mean(Data)
x =  806.3000

>> s = std(Data)
s =   40.2219

>> n = length(Data)
n =    20

>> x + 1.58488*s/sqrt(n)
ans =  820.5542

>> x - 1.58488*s/sqrt(n)
ans =  792.0458

>>
```

More data would be nice

- I only have 20 points though

Can I create more data?

Group the data

```
A:  811   805   809   804   837
B:  830   841   770   889   821
C:  850   754   786   763   754
D:  855   785   724   815   823
```

Create sets of 10 data points:

- {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}

It now looks like there are

- 6 samples
- 10 measurements each

Matlab Results

- Is this valid?
  - *It is done*
  - *I would call it making up data*

```
>> A = [811   805   809   804   837];
>> B = [830   841   770   889   821];
>> C = [850   754   786   763   754];
>> D = [855   785   724   815   823];
>> a = [A,B];
>> b = [A,C];
>> c = [A,D];
>> d = [B,C];
>> e = [B,D];
>> f = [C,D];
>> x = [mean(a), mean(b), mean(c), mean(d), mean(e), mean(f)]

x =  821.7000  797.3000  806.8000  805.8000  815.3000  790.9000

>> s = [std(a), std(b), std(c), std(d), std(e), std(f)]

s =   31.1735   33.0590   34.8482   46.9084   46.2867   43.7784
```

Summary

A t-test is a test of a mean.

With it, you can take a small sample from a population and determine

- The probability that a random sample will be more than a threshold (single-sided test), or
- The range over which 90% of the data will lie (two-sided test).

You can test to see if the population's mean is

- More than a threshold (single-sided test), or
- Within a given range (two-sided test).

The main difference is

- For individual tests, you find the variance as usual (Matlab function *var()*)
- For population tests, you divide the variance by the sample size.