Math-3

Lesson 7-3

Sampling for Statistical Studies

(how we obtain the “numbers”)

Quiz

1. What portion of the data is between 1 and 3 standard deviations above the mean of the data?

2. Danny's math score: 78, class mean math score: 84, SDEV = 3

   Danny's science score: 65, class mean science score: 80, SDEV = 15 On which test did Danny do "better" on (relative to his peers)?

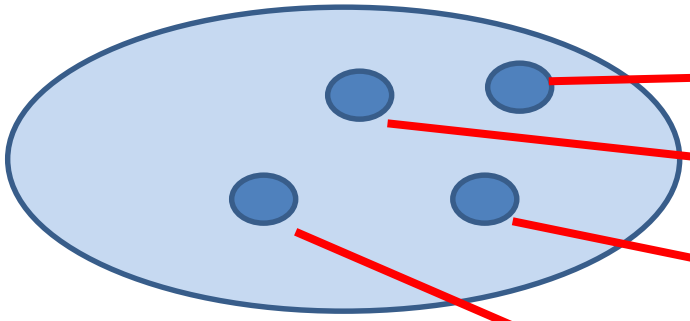
3. Explain your answer to problem 3.

Quiz Problem 4.

What are the 3 types of Statistical studies and what is the purpose of each?

Study Type	Purpose?
Sample	Find a statistic for the population
Experiment	Is a treatment effective?
Observational	Is there a correlation between parameters in the population?

1. <u>"Sample Study"</u> - The purpose of a sample study is to *<u>estimate</u>* a certain parameter of a population.

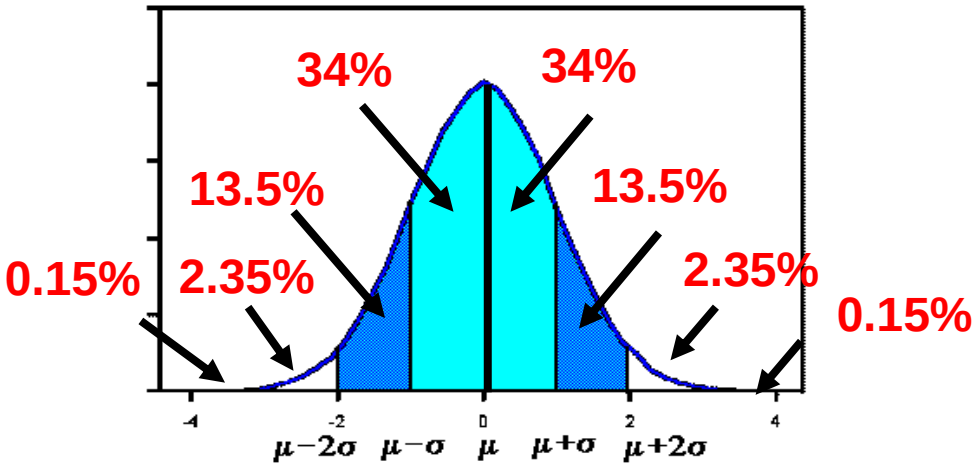


Large Population

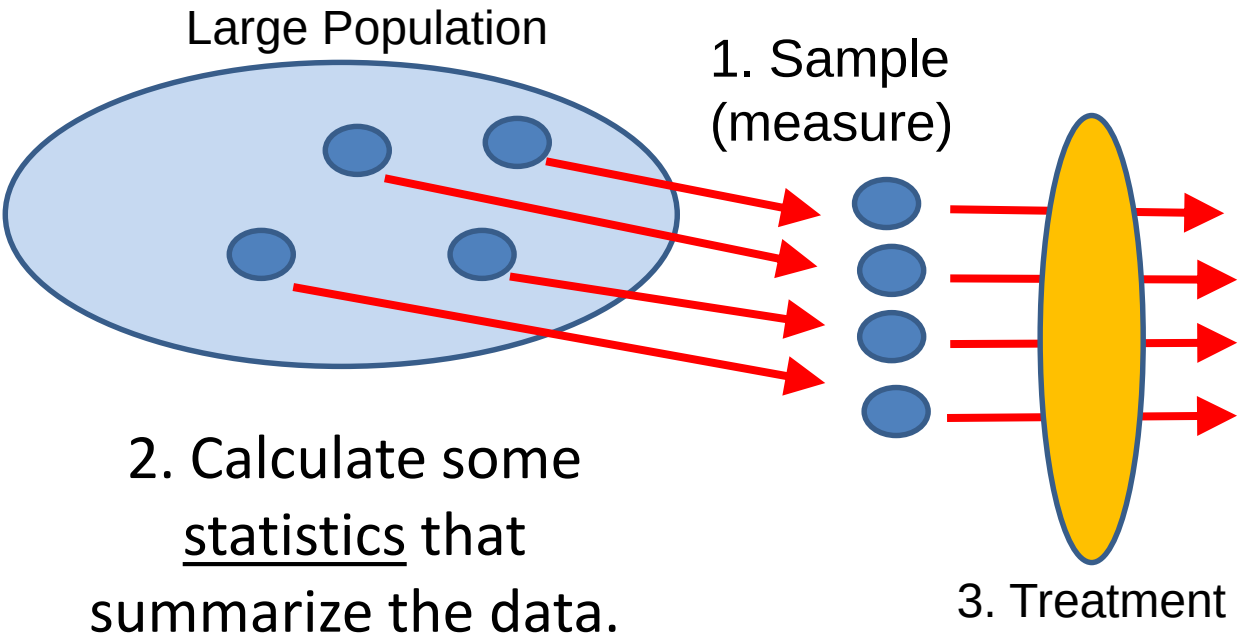
Sample
(measurement)

Data set
105
100
100
100
95
90
85
80
80
75
75
75
75
75
70
65
55
50
50
50
45
35

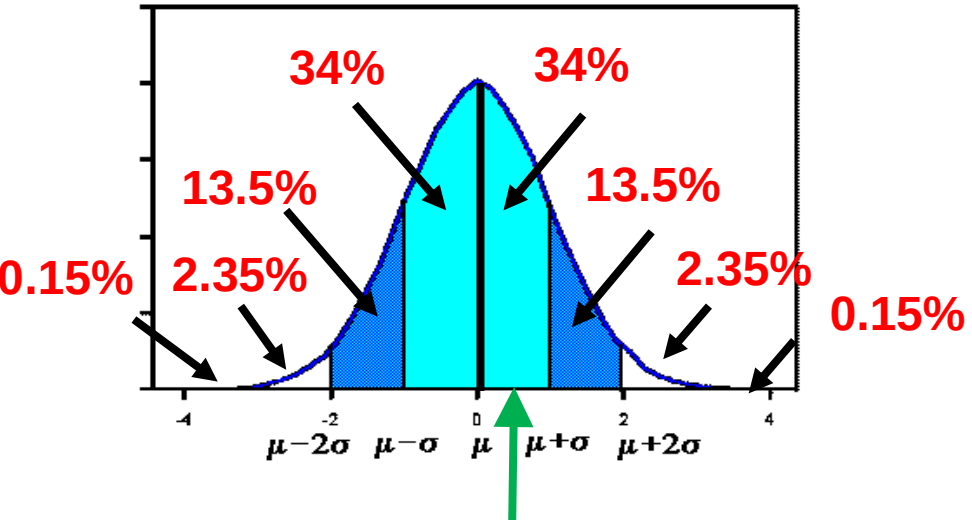
Calculate some <u>statistics</u> that summarize the data.



2. Experimental study: Purpose: to determine of a treatment has an effect on the population.



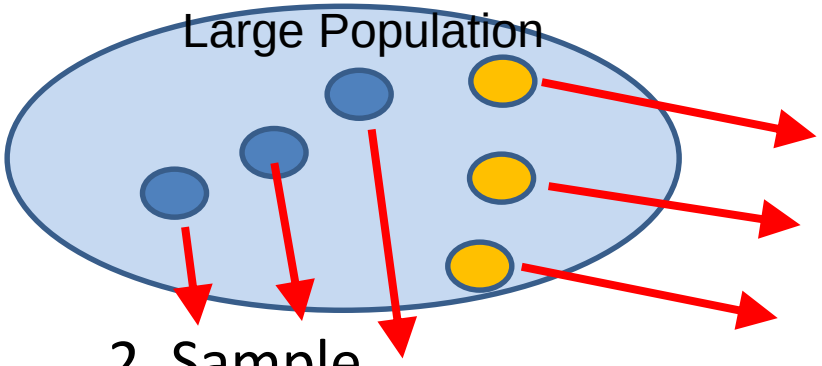
2. Calculate some statistics that summarize the data.



5. Calculate some statistics that summarize the data.

6. Compare statistics from treatment to the "expected" values (pre-treatment)

3. Observational study: Purpose: to see if there is a correlation between parameters of the population.



1. Sample (measure 'A')

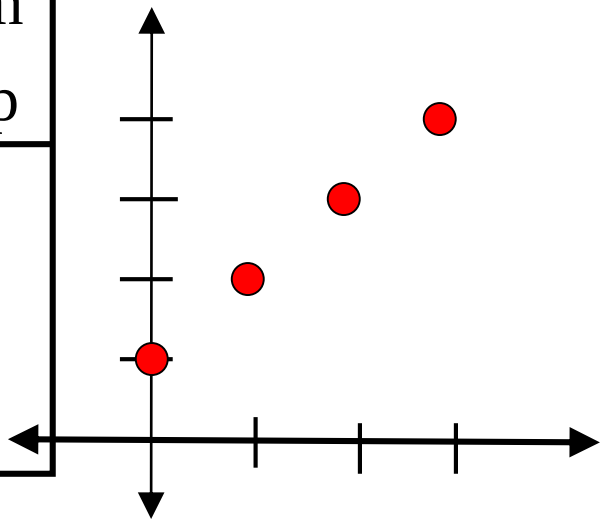
2. Sample (measure 'B')

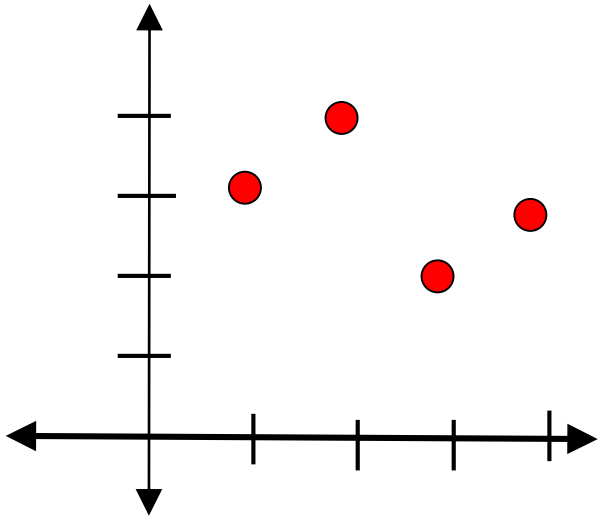
3. Compare 'A' to 'B' by graphing

4. Is there correlation?

# problems / HW assignment	Ave Test score
20	80
30	85
40	74
50	77

logs / hr	room temp
0	65
1	70
2	74
3	78





If we want to study the variation in height of individuals in a certain population what statistics would we want from that population?

Mean

Standard deviation

If the population is very large, it is unpractical to measure every member of the population. What do we do?

Take a sample from the population and obtain statistics from the sample. We than assume that the sample statistics reflect the statistics of the underlying population.

What can we do to maximize the probability that the statistics of the sample are representative of the statistic of the population?

The sample must be completely randomized so that every individual has the same chance of being selected.

Types of Samples:

1. Self-selected sample. completely randomized ?

   Members of the sample group volunteer to participate in the sample.

2. Convenience sample. completely randomized ?

   Easy-to-reach members of the population are used.

3. Systematic sample. completely randomized ?

   A rule is used to select individuals.

4. Random sample. completely randomized ?

   Each member of the population has an equal chance of being selected.

In statistics, **sampling bias** is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

Sampling bias merely represents a mathematical property, no matter if it is deliberate or either unconscious or due to imperfections in the instruments used for observation.

Are the following examples of sample bias? If so, would this indicate that the underlying population is healthier or less healthy than it actually is?

1. Sampling workers at a factory to measure the health of the general population.

sample is likely healthier than the general population.

2. Sampling current companies to measure the health of the economy?.

sample is likely healthier than the general population.

3. Using questions by using words like “sometimes” and “often” in your survey.

Requires individual interpretation.

Are the following examples of sample bias? If so, would the sample be healthier or less healthy than the underlying population?

4. Taking temperatures of people in a hospital waiting room.

sample is likely less healthy than the general population.

5. Measuring cholesterol in participants of P90X exercise class?

sample is likely healthier than the general population.

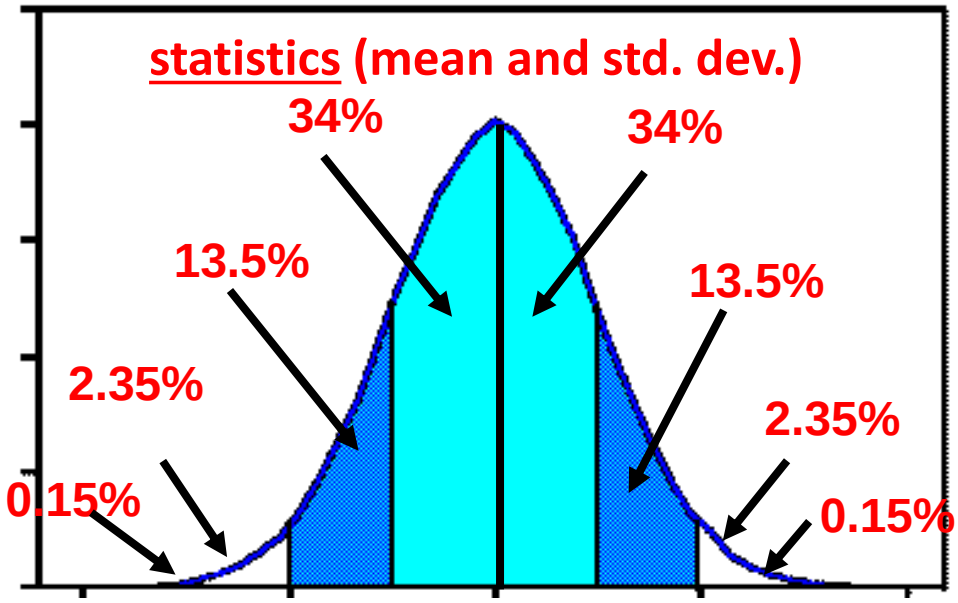
6. Measuring IQ scores on campus at Harvard University.

Sample is likelier higher than the general population.

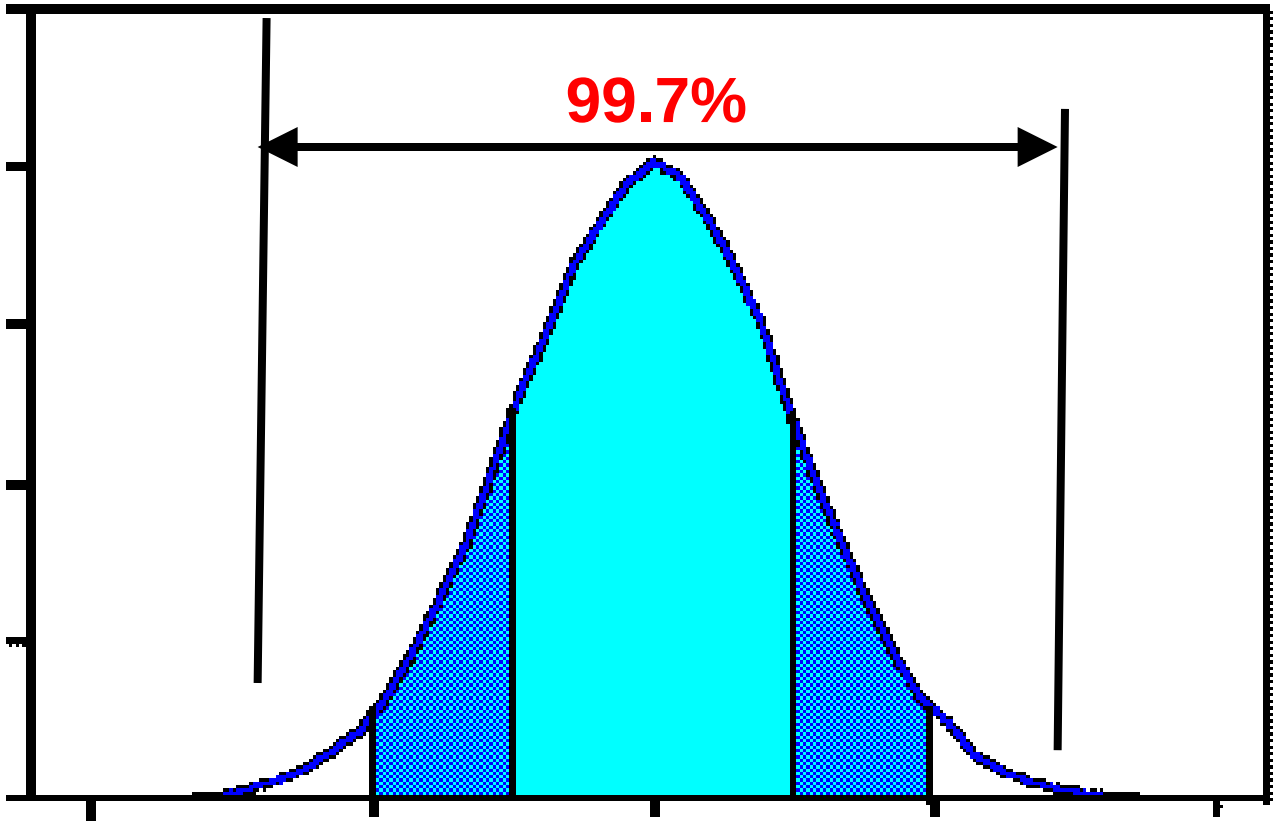
Types of Samples and the potential for bias:

Sample Type	Potential for bias?
Self-selected	often
convenience	often
Systematic	sometimes
random	Unbiased!

If our sample is randomly selected, then the variability in the sample is due to randomness of the underlying population rather than some bias in the sampling method.

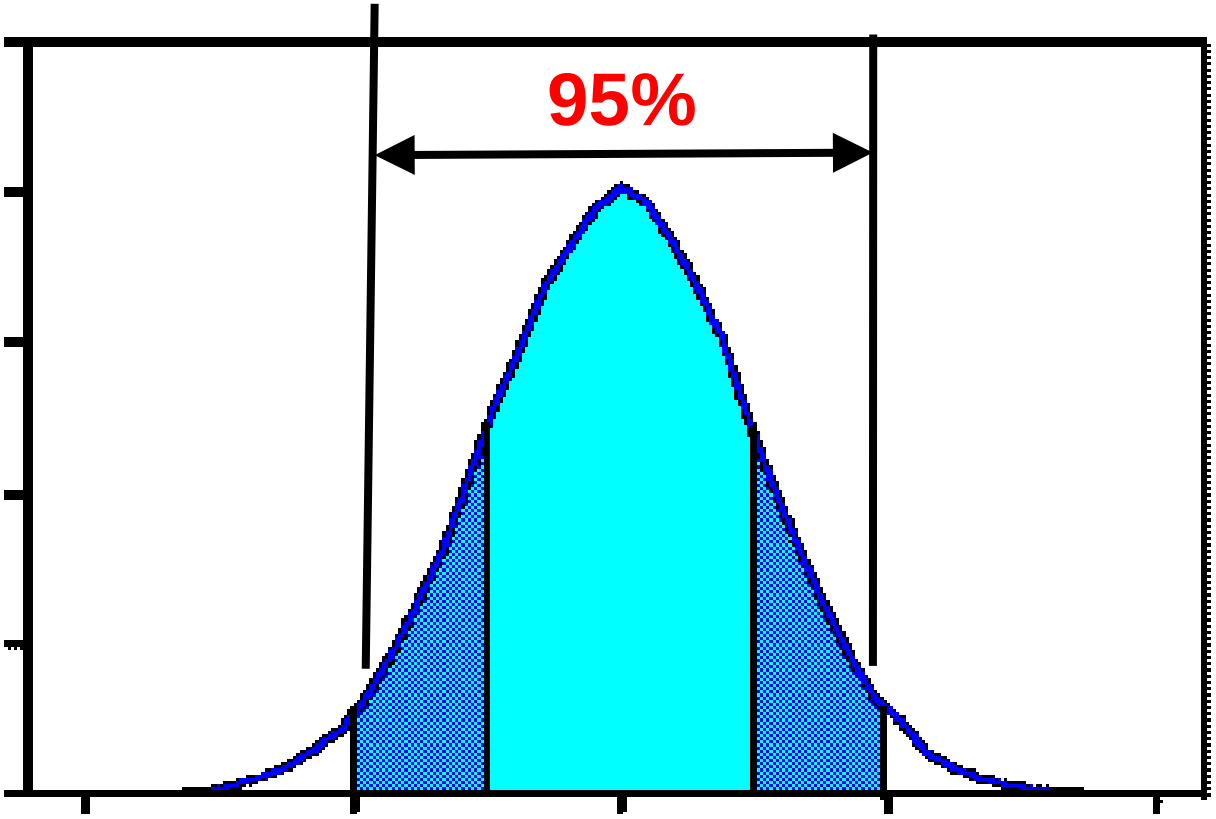


99.7% Confidence Interval: if we repeated the same sampling method to select different data sets, and re-computed the +/- 3 SDEV for each of these different data sets, we would expect the mean of the underlying population to fall within +/- 3 SDEV 99.7% of the time.



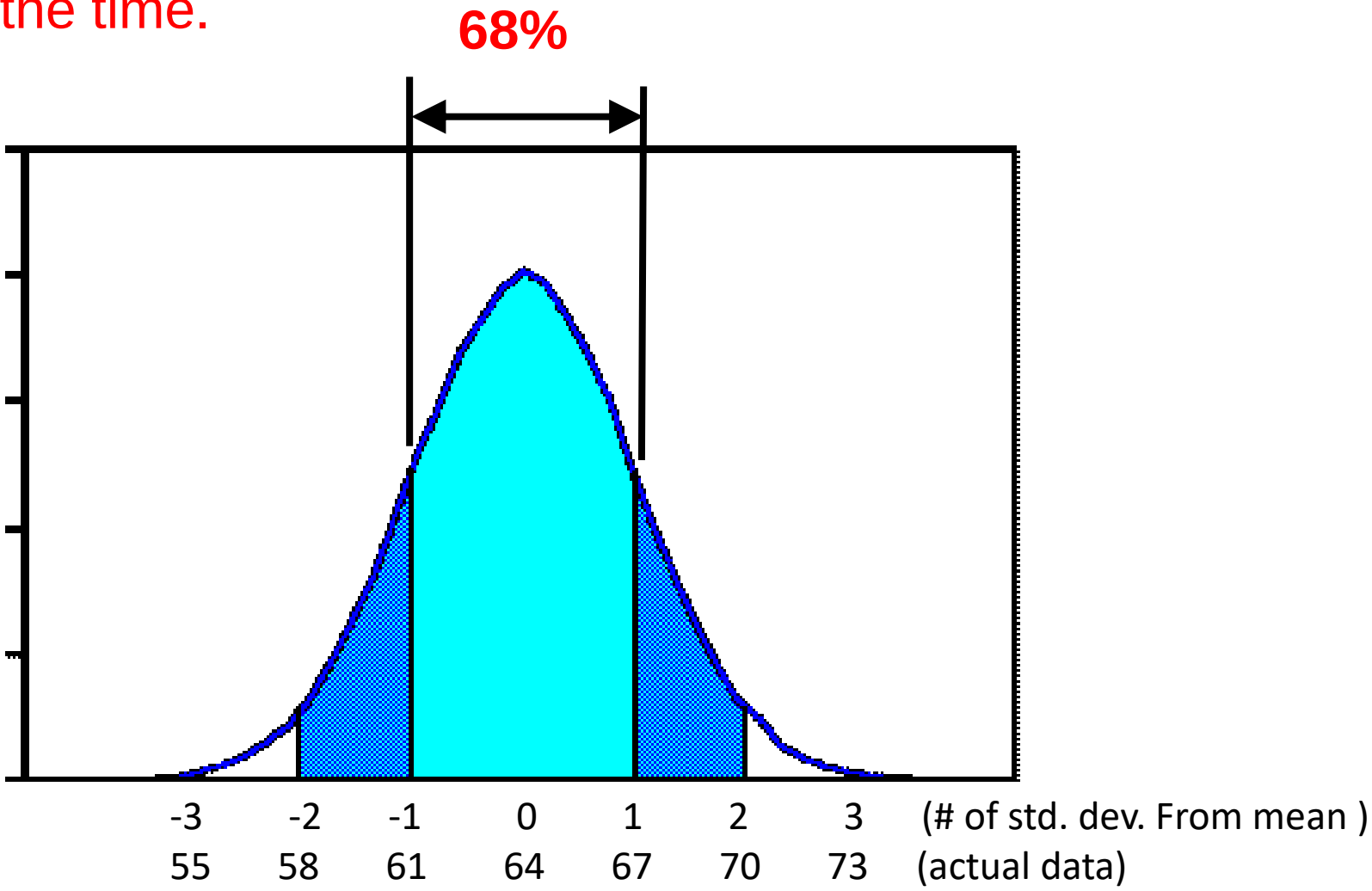
-4	-3	-2	-1	0	1	2	3	(# of std. dev. From mean)
	55	58	61	64	67	70	73	(actual data)

95% Confidence Interval: if we repeated the same sampling method to select different data sets, and re-computed the +/- 2 SDEV for each of these different data sets, we would expect the mean of the underlying population to fall within +/- 2 SDEV 95% of the time.

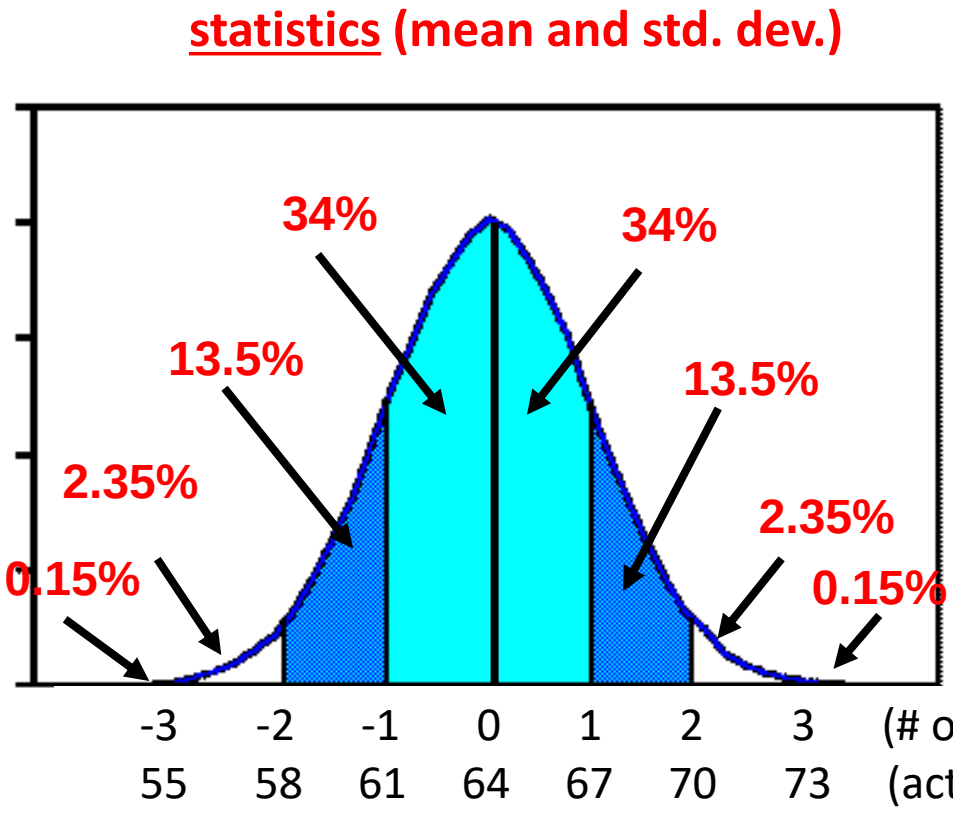


-3	-2	-1	0	1	2	3	(# of std. dev. From mean)
55	58	61	64	67	70	73	(actual data)

68% Confidence Interval: if we repeated the same sampling method to select different data sets, and re-computed the +/- 1 SDEV for each of these different data sets, we would expect the mean of the underlying population to fall within +/- 1 SDEV 68% of the time.



95% Confidence Interval: the colored region.



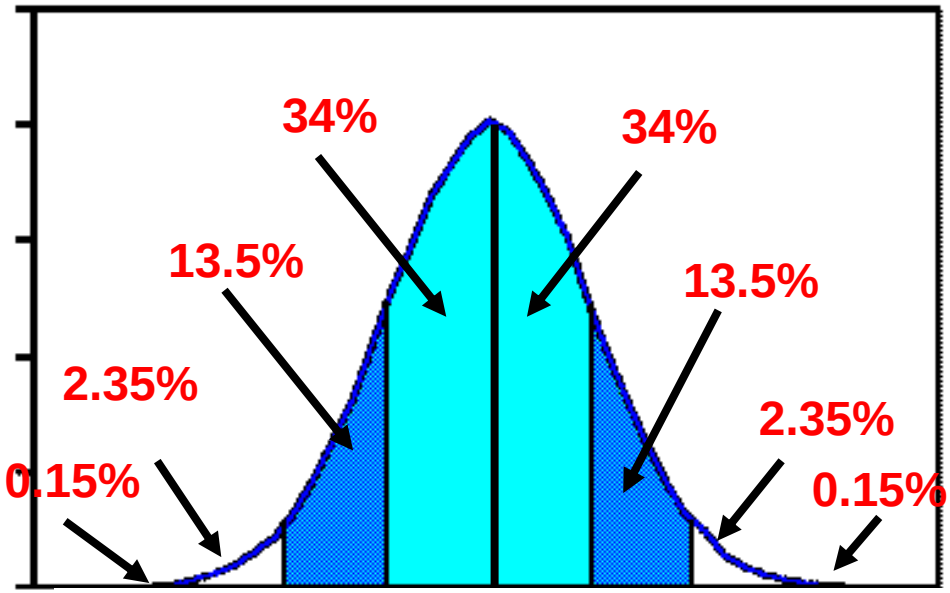
If our sample is randomly selected, then the variability in the sample is due to random processes rather than some bias in the sampling method.

The "confidence interval" combines an estimate of the interval in which the population mean falls, and a probability of the repeatability of where the mean would fall given the same sample procedure.

Study: Does taking a caffeine pill raise heart rate?

statistics (mean and std. dev.) prior to taking the pill

Treatment: take caffeine pill



-3 -2 -1 0 1 2 3 (std. dev. Above/below mean)

55 58 61 64 67 70 73 (heart beats/minute)

Mean heart rate after pill

1 2 3

At what point do you say the pill has an effect?

Experimental Studies: Comparing Treatment to No Treatment

"Null hypothesis": assume that the treatment has no effect on the outcome. Mathematically this means your assumption is: mean of the "treated population = untreated population mean"

$$H_0 : \mu_{\text{treatment population}} = \mu_{\text{no-treatment population}}$$

1. Conduct the experiment (treatment and no treatment)

2. Compare treatment group and control group means

3. Very simplistically, if treatment mean is outside the 95% confidence interval of the non-treatment distribution, we must reject the null hypotheses.

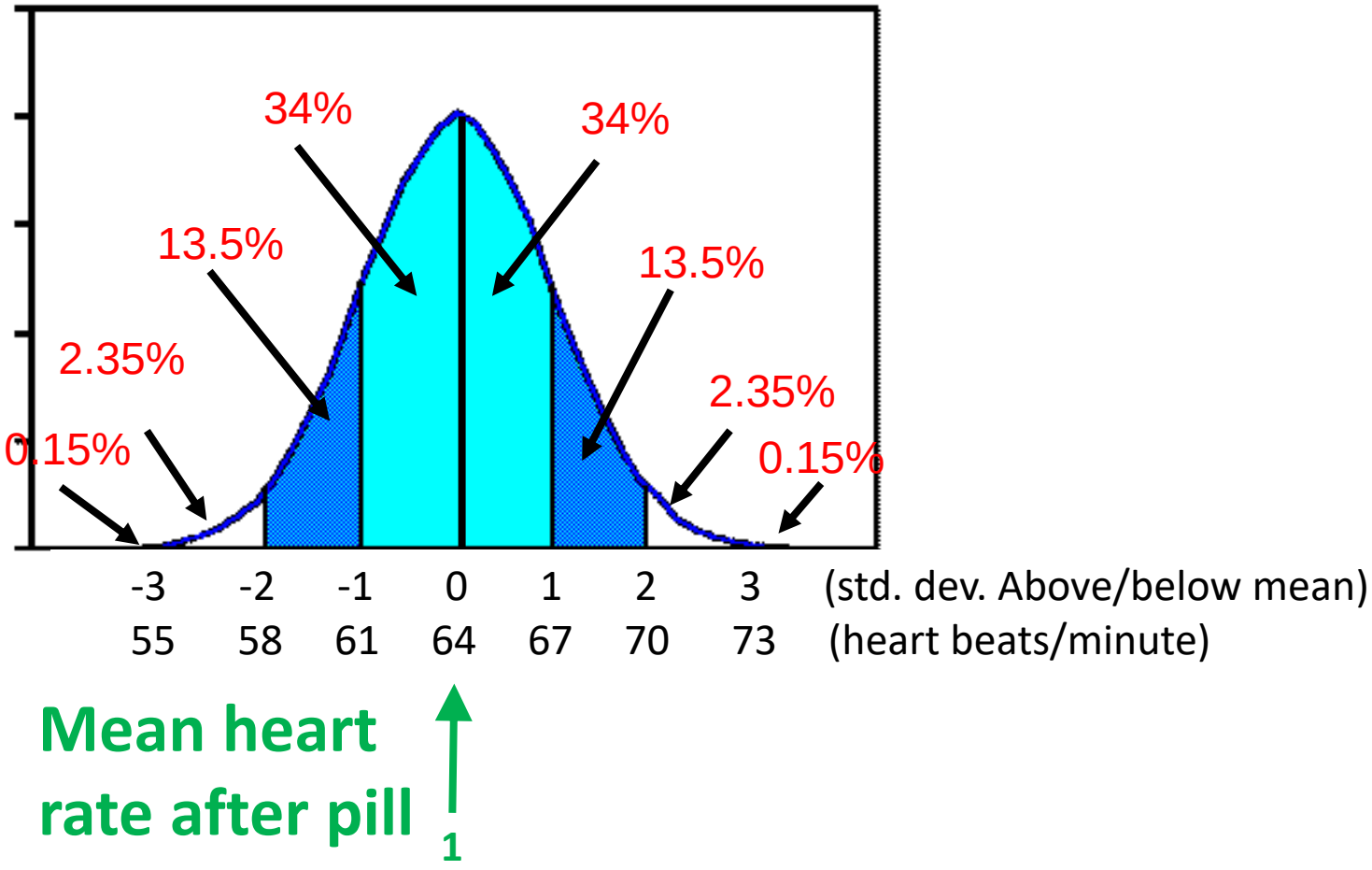
"Reject the Null hypothesis": What does this mean?

The shift if the mean due to the treatment is most likely due to some factor other than random variation of the sample mean.

95% Confidence Interval: the colored region.

At what point do you say the pill has an effect?

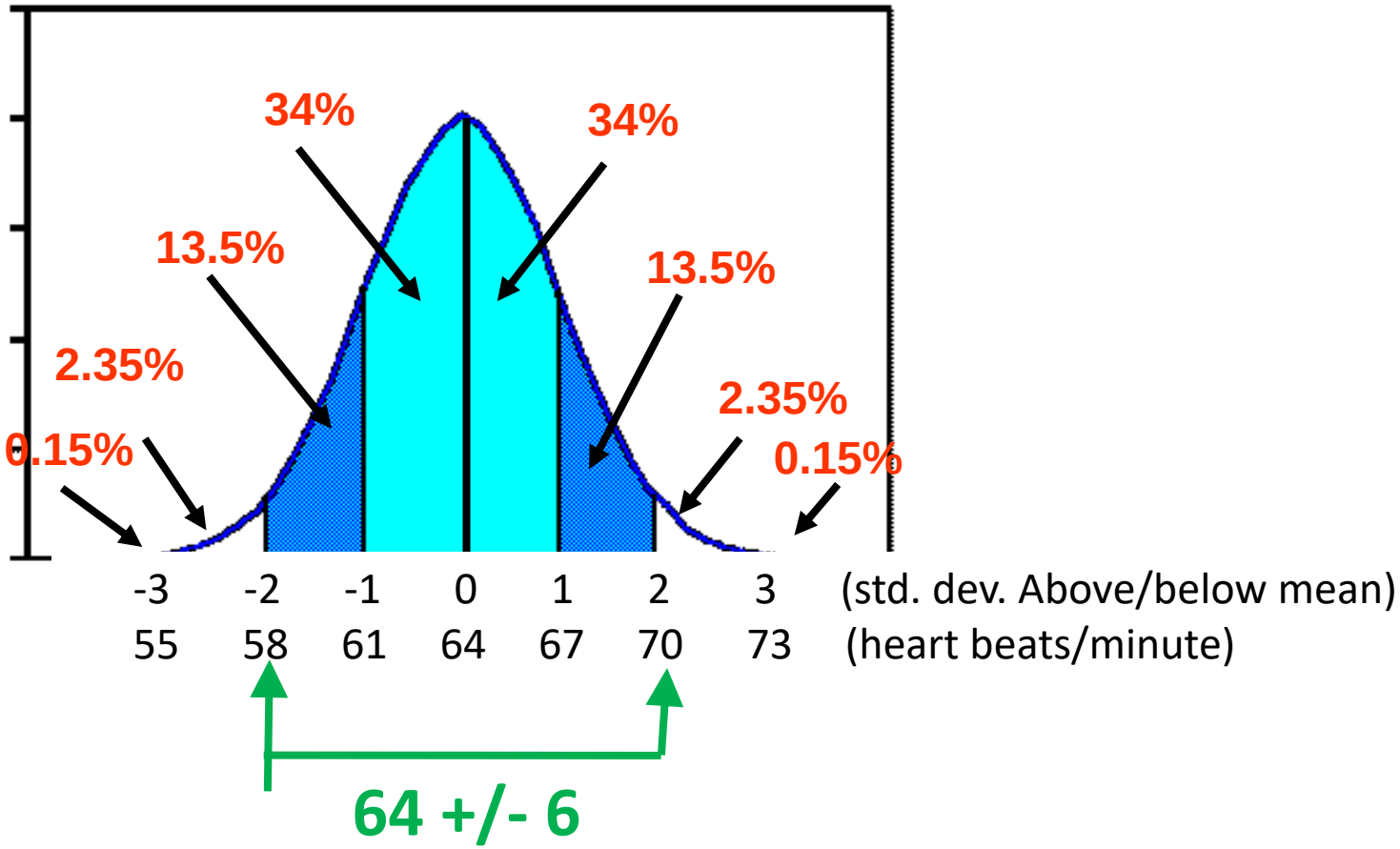
statistics (mean and std. dev.) **prior to** taking the pill



95% Confidence Interval:

What is the 95% confidence interval for the data below?

statistics (mean and std. dev.) **prior to** taking the pill



95% confidence interval: 58 to 70 BPM

You want to conduct an experiment to determine if adding fertilizer results in more flowers blooming on petunias.

1. What is your null hypothesis?

<u>Null Hypothesis</u>: fertilizer does not change the number of flowers on petunias.

$$H_0 : \mu_{\text{treatment population}} = \mu_{\text{no-treatment population}}$$

You want to conduct an experiment to determine if adding fertilizer results in more flowers blooming on petunias.

2. How would you set up the experiment so that you could test this factor only?

a. Randomly select ‘x’ number of Petunias seeds.

b. Number each pot (from 1 to ‘x’).

c. Put numbers 1 to ‘x’ into a hat and select ‘x’/2 numbers. This is the control group. The other plants are the treatment group.

d. Place all plants randomly (x spots, randomly selected from a ”hat”) in a green house and provide the same soil, water, sunlight, etc.

e. Control group gets same amount of fertilizer.

You want to conduct an experiment to determine if adding fertilizer results in more flowers blooming on petunias.

3. Why can we say the statistics from this experiment represent that of the underlying population?

Random selection of seeds for control and treatment, along with random positioning inside the green house ensures the sampling method does not introduce bias into the experiment.

The random variation within the population is not being over-represented in either the control or treatment samples. Therefore the sample results should represent the underlying population.

If we repeated the same sampling method to select different seeds and re-computed the +/- 2 SDEV for the parameter of interest (mean number of blooms), we would expect the mean of the underlying population to fall within this interval 95% of the time.

Control group: mean: 15 flowers/plant
standard deviation: 2.

Treatment group: mean: 20 flowers/plant.

4. What does accepting the null hypothesis mean?

Applying fertilizer <u>does not</u> significantly change the number of flowers on petunias.

5. What does rejecting the null hypothesis mean?

Applying fertilizer <u>DOES</u> significantly change the number of flowers on petunias.

Control group: mean: 15 flowers/plant, standard deviation: 2.

Treatment group: mean: 20 flowers/plant.

6. How do you decide whether to accept or reject the null hypothesis?

This involves a more detailed analysis and calculations. That I will summarize on the next few slides but I don't expect you to learn it for homework or for tests.

1. Calculate “standard error” (SE):

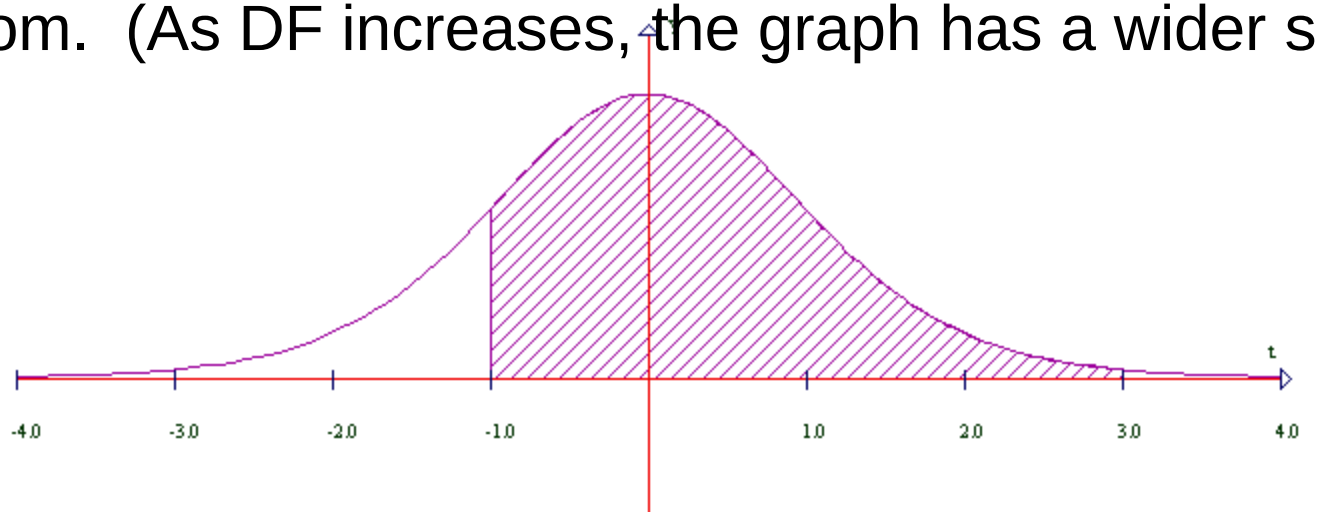
$$SE = \sqrt{\frac{(S_T)^2}{n_T} + \frac{(S_C)^2}{n_c}}$$

2. Calculate “degrees of freedom” (round to the nearest whole number)

$$DF = \frac{\frac{(S_T)^2}{n_T} + \frac{(S_C)^2}{n_c}}{\frac{\left((S_T)^2 / n_T\right)^2}{(n_T - 1)} + \frac{\left((S_C)^2 / n_C\right)^2}{(n_C - 1)}}$$

3. Calculate the "test statistic" $t = \dfrac{\bar{x}_T - \bar{x}_c}{SE}$

4. Use the "t-distribution" to calculate the probability that the "t-statistic" is outside of the range [-t, +t] (that we calculated). This graph has the same shape as the Normal distribution but has different portions of the total data that falls in the "bins" that the normal distribution has (0.15%, 2.35%, 13.5% and 34%). These portions depend upon the number of degrees of freedom. (As DF increases, the graph has a wider spread).



5. If the probability of the t-statistic is outside of the 95% confidence interval, we would reject the "null hypothesis".