Topic 1: Statistics

1.1.1

* **State that error bars are a graphical representation of the variability of data.**

* **Error bars can be used to show either the range of the data or the standard deviation.**

* **(note: if there are no outliers, the range may be okay, but generally the standard deviation is a better way of showing the data distribution)**
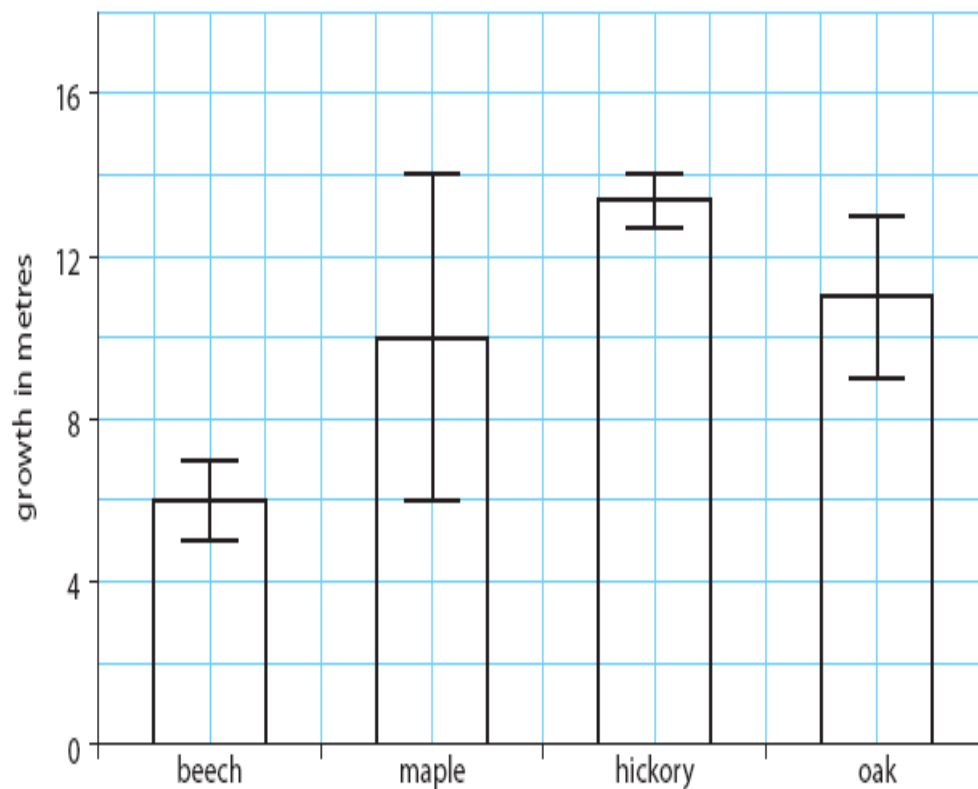
Example showing error bars

Figure 1.1 Rate of tree growth on the Oak–Hickory Dune 2004–05. Values are represented as mean ±1SD from 25 trees per species.

1.1.2

* **Calculate the mean and standard deviation of a set of values.**

* **Students should specify the standard deviation (s), not the population standard deviation.**

* **Students will not be expected to know the formulas for calculating these statistics. They will be expected to use the standard deviation function of a graphic display or scientific calculator.**

1.1.2
Calculate the mean and standard deviation for each group. (Use S_x)

	Group A	Group B
	24	24
	25	29
	26	25
	23	23
	25	29
	25	34
	26	32
	27	31
	23	32
	23	29
Mean		
Standard deviation		

1.1.2- check your answers

	Group A	Group B
	24	24
	25	29
	26	25
	23	23
	25	29
	25	34
	26	32
	27	31
	23	32
	23	29
Mean	**24.7**	**28.8**
Standard deviation	**1.42**	**3.71**

1.1.3

* **State that the term standard deviation is used to summarize the spread of values around the mean, and that 68% of the values fall within one standard deviation of the mean.**

* **This rises to about 95% for ±2 standard deviations.**

* **Note: this assumes a normal distribution (bell curve)**

1.1.3

Figure 21.6 The normal distribution and its SD
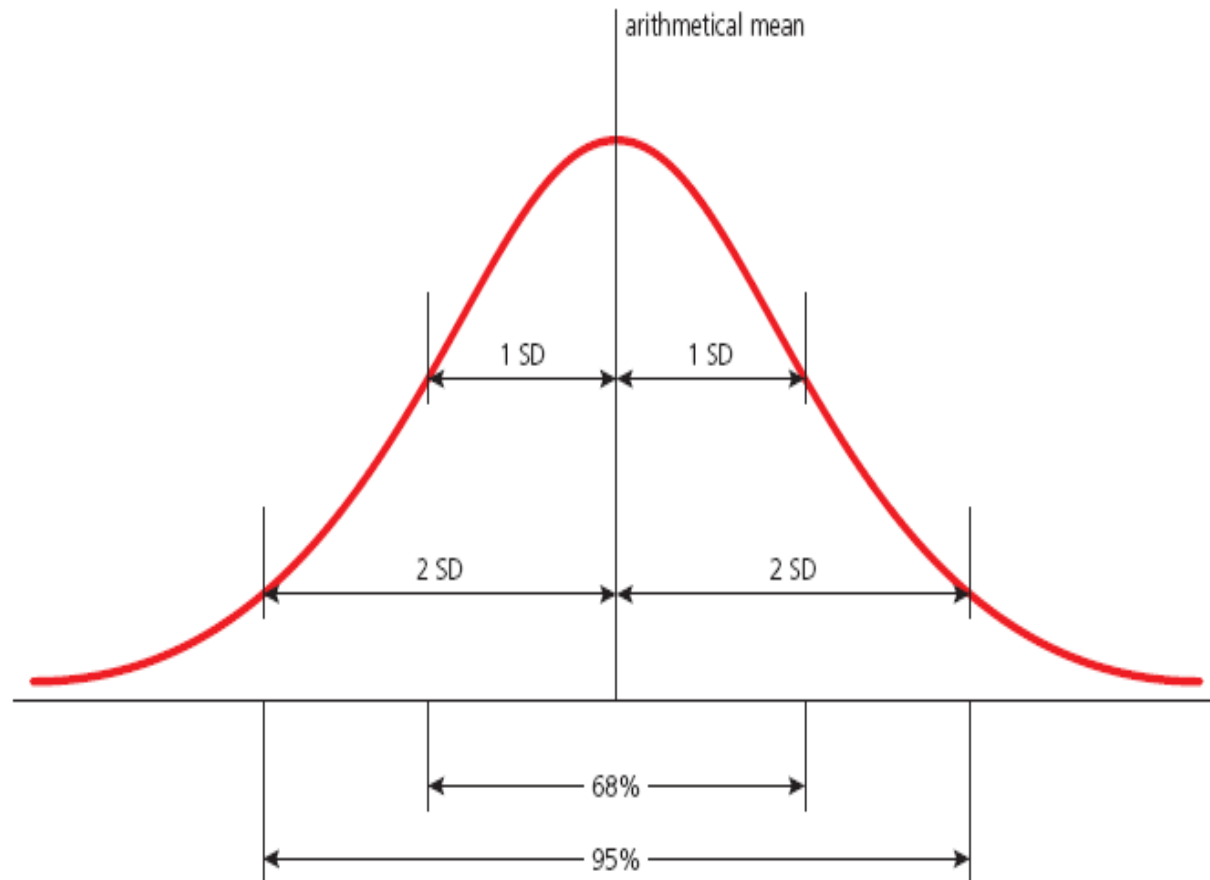
1.1.4

* **Explain how the standard deviation is useful for comparing the means and the spread of data between two or more samples.**

* **A small standard deviation indicates that the data is clustered closely around the mean value. Conversely, a large standard deviation indicates a wider spread around the mean.**

1.1.4

* **Overlapping standard deviations suggest that two data sets <u>are not</u> significantly different.**

Data

# 2 types of Data

Qualitative
Quantitative

Qualitative Data

* Information that relates to **characteristics or description** (observable qualities)

* Information is **often grouped** by descriptive category

Qualitative Data

* Examples
  * Species of plant
  * Type of insect
  * Shades of color
  * Rank of flavor in taste testing
  * Remember: qualitative data can be "scored" and evaluated numerically

Qualitative data, manipulated numerically

* Survey results, teens and need for environmental action
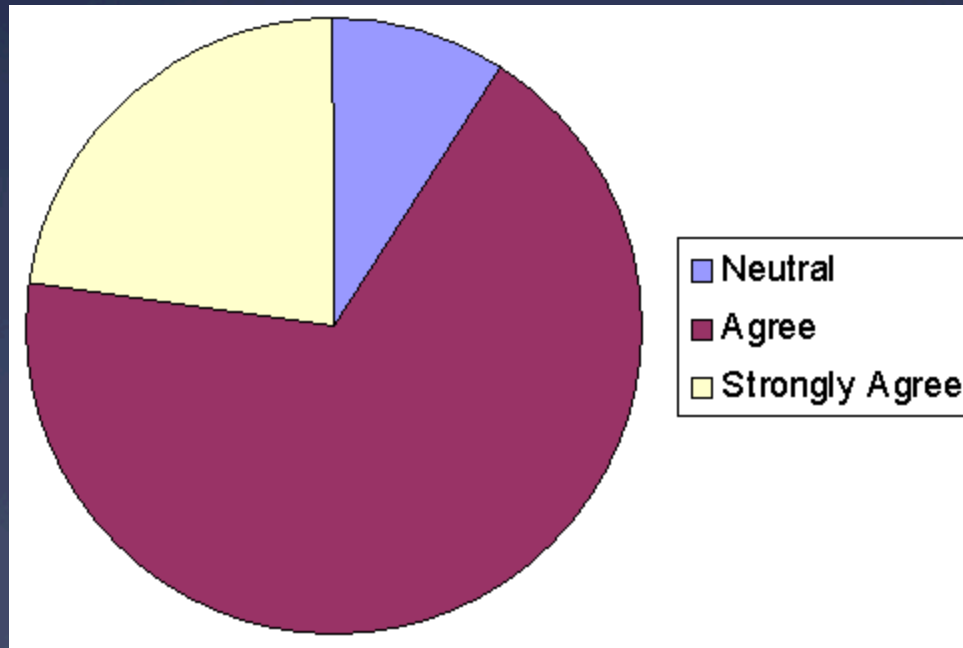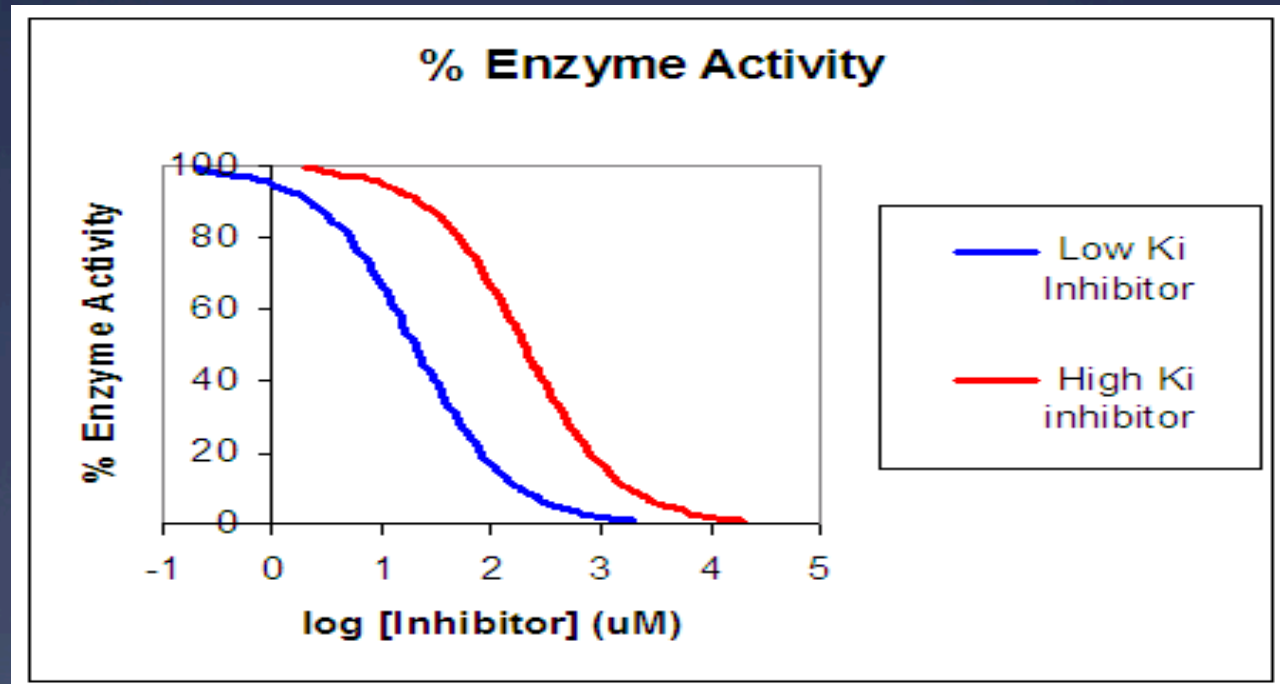
Quantitative data

* Quantitative – measured using a naturally occurring numerical scale

Quantitative data

* Examples
  * Chemical concentration
  * Temperature
  * Length
  * Weight…etc.

Quantitation

* Measurements are often displayed graphically

Quantitation = Measurement

* In data collection for Biology, data must be measured carefully, using laboratory equipment
  * timers
  * metersticks
  * pH meters
  * balances
  * pipettes

Quantitation = Measurement

* The limits of the equipment used add some uncertainty to the data collected.

* All equipment has a certain magnitude of uncertainty.

    * For example, is a ruler that is mass-produced a good measure of 1 cm? 1mm? 0.1mm?

Quantitation = Measurement

* For quantitative testing, ***you must indicate the level of uncertainty of the tool that you are using for measurement!!***

Mean Calculation

Mean

* This is measure of average
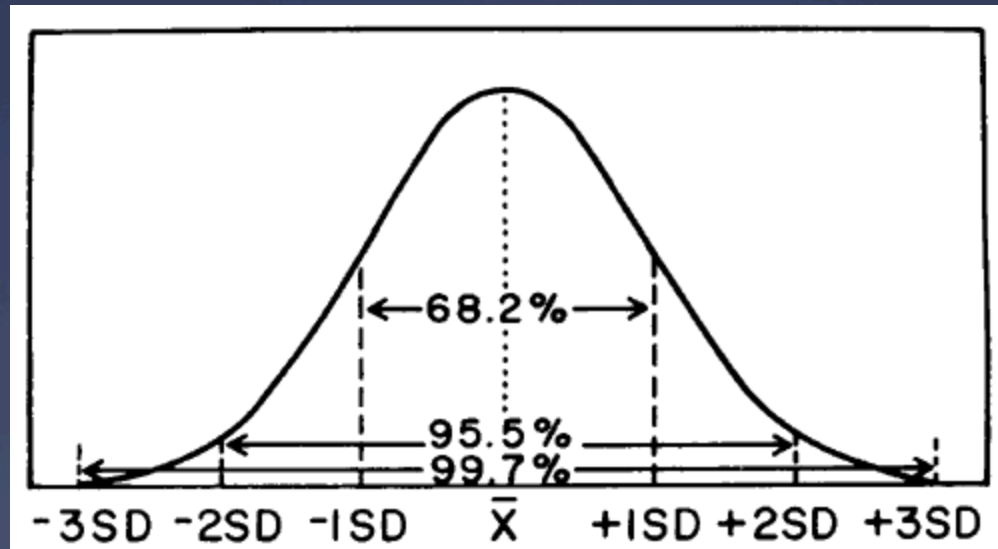

* MEAN = Sum of values/Number of items

Standard Deviation

Standard Deviation

* The degree each data point in the set varies (or deviates) from the mean

* The wider the spread of scores, the larger the standard deviation

Standard Deviation

* For data that has a normal distribution, 68% of the data lies within one standard deviation of the mean.

Standard Deviation

* 1. Calculate the mean (M) of a set of data

$$S = \sqrt{\frac{\sum (X-M)^2}{n-1}}$$

MEAN GPA FOR

MEN: 2.44
WOMEN: 2.78

Standard Deviation

* 2. Subtract the mean from each point of data to determine (X-M)

$$S = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

MEN	WOMEN
0.90-2.44 = -1.54	1.50-2.78=-1.28
2.00-2.44=-0.44	3.00-2.78=0.22
1.40-2.44=-1.04	3.00-2.78=0.22
2.00-2.44=-0.44	2.50-2.78=-0.28
3.00-2.44=-0.56	3.00-2.78=0.22
2.00-2.44=-0.44	3.00-2.78=0.22
3.00-2.44=-.56	4.00-2.78=1.22
4.00-2.44=1.56	3.00-2.78=0.22
3.40-2.44=1.26	2.00-2.78=-0.78

Standard Deviation

* 3. Square each of the resulting numbers to determine $(X-M)^2$

$$S = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

MEN	WOMEN
$(1.54)^2 = 2.37$	$(-1.28)^2 = 1.64$
$(-0.44)^2 = 0.19$	$(0.22)^2 = o.05$
$(-1.04)^2 = 1.08$	$(0.22)^2 = 0.05$
$(-0.44)^2 = 0.19$	$(-0.28)^2 = 0.08$
$(-0.56)^2 = 0.31$	$(0.22)^2 = 0.05$
$(-0.44)^2 = 0.19$	$(0.22)^2 = 0.05$
$(-0.56)^2 = 0.31$	$(1.22)^2 = 1.49$
$(1.56)^2 = 2.43$	$(0.22)^2 = 0.05$
$(1.26)^2 = 1.59$	$(-0.78)^2 = 0.61$

Standard Deviation

* Add the values from the previous step together to get Σ (X-M)^2

$$S = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

MEN	WOMEN
2.37	1.64
0.19	o.05
.08	0.05
0.19	0.08
0.31	0.05
0.19	0.05
0.31	1.49
2.43	0.05
<u>1.59</u>	<u>0.61</u>
8.68	4.06

Standard Deviation

* 5. Calculate (n-1) by subtracting 1 from your sample size. Your sample size is the total number of data points you collected

$$S = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

N-2 for men = 9 - 1 = 8
N-1 for women = 9 -1 = 8

Standard Deviation

* 6. Divide the answer from $\Sigma\ (X-M)^2$ by the answer from (n-1) to find $\dfrac{\Sigma\ (X-M)2}{n-1}$

$$S = \sqrt{\dfrac{\Sigma(X-M)^2}{n-1}}$$

Men 9.68/8 = 1.09
Women 4.06/8 = 0.51

Standard Deviation

* 7.   Calculate the square root of your previous answer to determine the SD

$$S = \sqrt{\dfrac{\sum(X-M)^2}{n-1}}$$
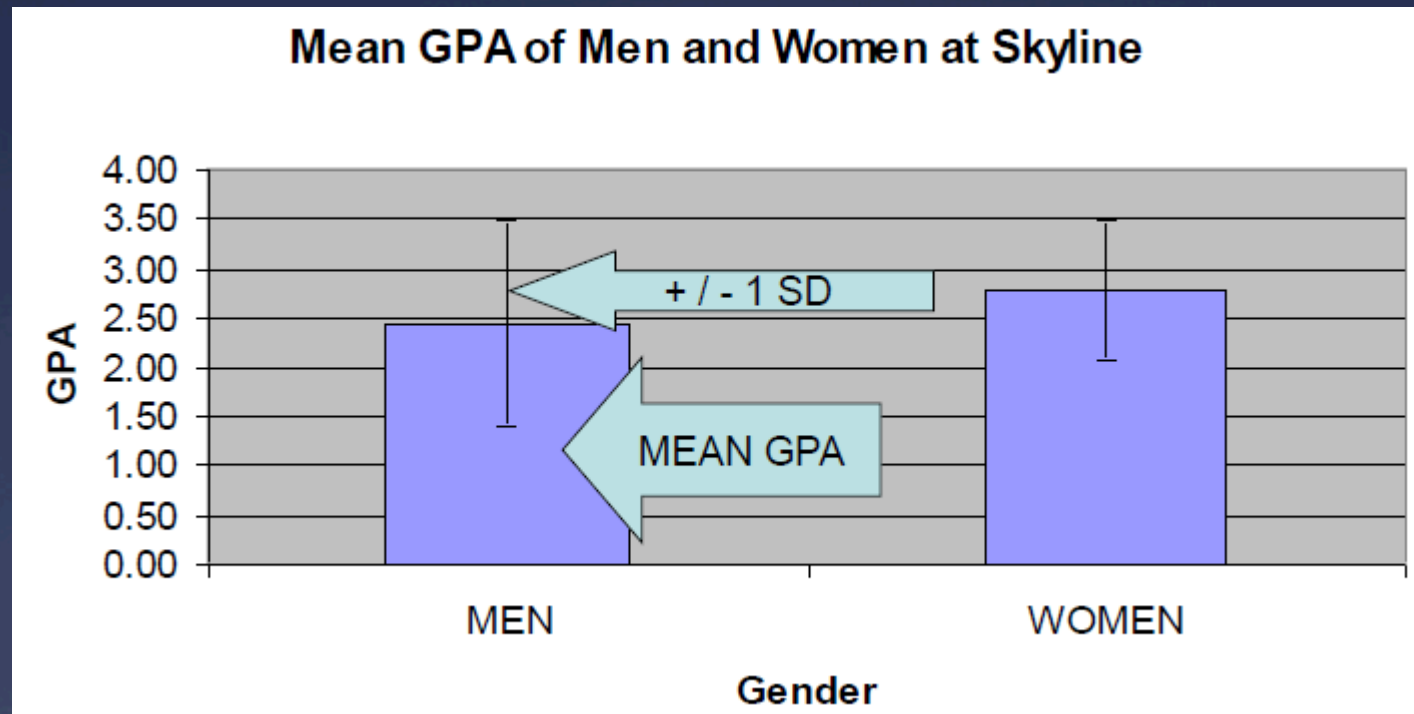
Men $\sqrt{1.09} = 1.04$
Women $= \sqrt{0.51} = 0.71$

* In men, 68% of all students sampled have a GPA that falls within 1.04 grade points of the mean
  * MEAN PLUS 1.04 = 3.49
  * MEAN MINUS 1.04 = 1.40

* In women, 68% of all students sampled have a GPA that falls within 0.71 grade points of the mean
 * MEAN PLUS 0.71 = 3.49
  * MEAN MINUS 0.71 = 2.07

So, are women smarter?

* Graph results to find out:

* Since the SD bars overlap, we conclude that the GPA's are probably NOT significantly different, and that the difference in the means is most likely due to chance in the sampling population

Since the SD bars DO NOT overlap, we conclude the populations are probably significantly different, and that the difference in the means is most likely due to something MORE THAN CHANCE
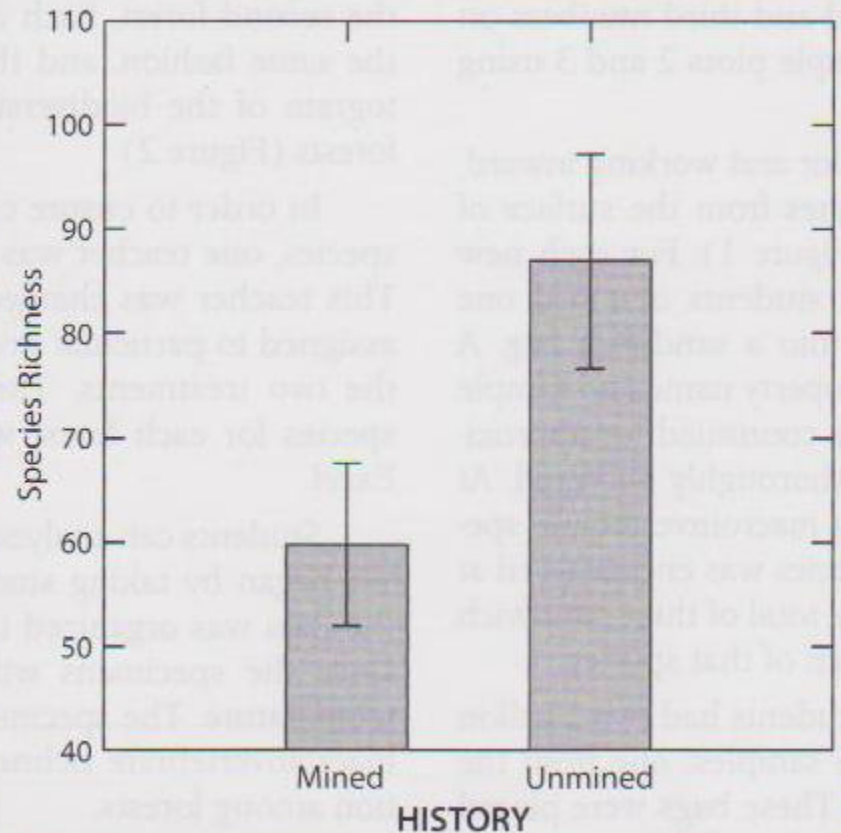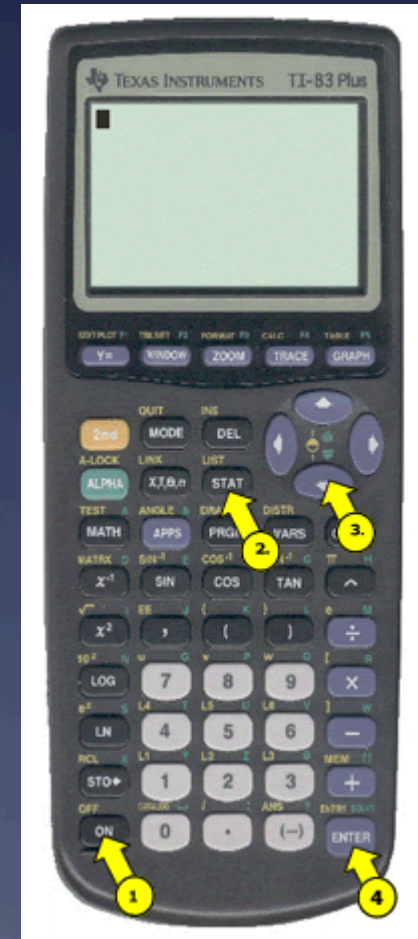
Graphic calculator routines are based on the TI 83plus and the TI 84 plus

Graphic Calculator

Graphic Calculator

* Functions you will be using
  * 1 – On
  * 2 – STAT
  * 3 – Blue Arrow keys
  * 4 - ENTER

Graphic Calculator

* Step 1: Data Entry – One variable
  * Press "Stat"
  * Press "1". You should see 3 columns: L1, L2, L3
  * The cursor should be at L1

Graphic Calculator

* Type in the first number ,then "Enter"

* Type in the second number, the "Enter"

* When fished, press "2nd/Mode". The mode button is to the right of the 2nd button. The text "Quite" is written above the MODE button

Graphic Calculator

* Two Variable Data Entry
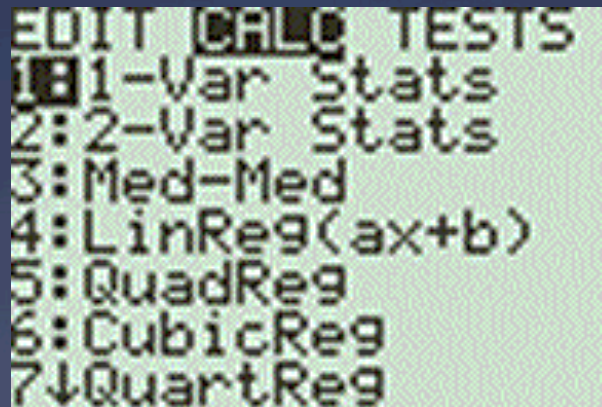 * Begin as before but when you reach the end of the L1 data list press >. The cursor should jump to the top of the second column, L2
  * Enter the y-values (make sure they line up with the corresponding x values in L1
  * When finished, press "2nd /Mode
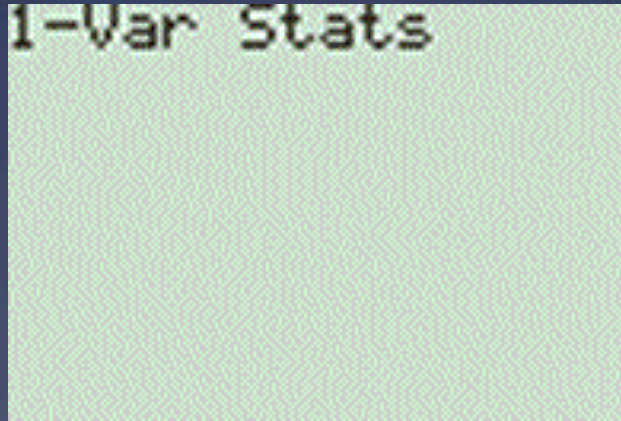
Graphic Calculator

* Descriptive Statistics

  * Press "STAT". Use the blue > to move the highlighted bar over the CALC menu

  * Choose the 1-Var stats option (that is, press "1"

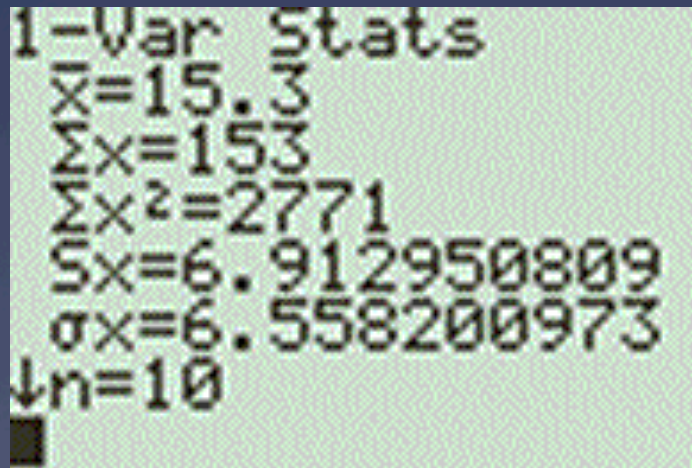# Graphic Calculator

* Descriptive Statistics (cont)
  * You'll see the words 1-Var Stats
  * Press "2nd/1". You should see L1above the key
  * Press "Enter"

Graphic Calculator

* Mean
  * The top value on the screen

* Standard Deviation
  * The SD is the fourth value on the screen
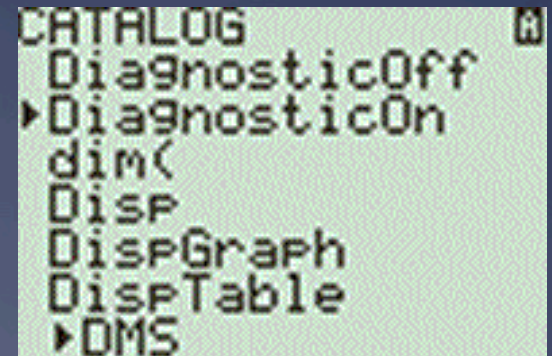
```
1-Var Stats
x̄=15.3
Σx=153
Σx²=2771
Sx=6.912950809
σx=6.558200973
↓n=10
```

Graphic Calculator

* Correlation
    * The TI-83 will only display the correlation in the Diagnostic On Mode
    * Press "2nd/0". You should see the word CATALOG above the key.
    * You will see a screen with an alphabetical list of commands
    * Arrow down to DiagnosticOn. Press "ENTER". The screen will now say "DiagnosticOn.
    * Press "Enter" again
    * You will see the word Done

Graphic Calculator

* Correlation (con't)

* You can continue now

* Press "STAT". Use the blue > to move the highlighted bar over the CALC menu

* Choose the LinReg(a+bx) option (that is press "8") You'll see the words on the screen

LinReg(a+bx) ▌

Graphic Calculator

* Press "2nd/1" (you should see L1 above the key), the ",/2nd/2 (you should see L2 above the key). **You'll** see the words LinReg(a+bx)L1,L2 on the screen

```
LinReg(a+bx) L1,
L2
```

Graphic Calculator

* Press "ENTER".

* The correlation is the fourth number in the list (r=.. \0

* NOTE – You can also find correlation by pressing 4: LinReg(ax+b), instead of 8: LinReg(a+bx).

* In this case, the role of the a and be are switched, but r is the same.

Practice Problems

t-Test

1.1.5

* **The t-test is a statistical test that will also tell if 2 data sets are or are not significantly different.**

1.1.5

* **Deduce the significance of the difference between two sets of data using calculated values for t and the appropriate tables.**

* **For the t-test to be applied, the data must have a normal distribution and a sample size of at least 10. The t-test can be used to compare two sets of data and measure the amount of overlap. (Students will not be expected to calculate values of t. Only a two-tailed, unpaired t-test is expected.)**

T-test steps:

1 The null hypothesis (negative hypothesis) assumes the difference under investigation has arisen by chance. So the null hypothesis is:

'There is no difference in the 2 sets of data.'

The role of the *t*-test is to determine whether to accept or reject the null hypothesis. If it is rejected here, we can have confidence that the difference in data sets is statistically significant.

2 Next, check that the data are normally distributed. This is done by arranging the data in ascending or descending order.

3 *You are not expected to calculate values of t.* This statistic can be found by using a scientific or statistics calculator, or by means of a spreadsheet formula.

4 Once a value of t has been calculated we determine the degrees of freedom (df) for the two samples, using the formula:
df = (total number of values in both samples) − 2
or $(n_a + n_b) − 2$

Now we consult a table of critical values for the t-test.

Table of t values

Degrees of freedom	t values					
1	1.00	3.08	6.31	12.71	63.66	636.62
2	0.82	1.89	2.92	4.30	9.93	31.60
3	0.77	1.64	2.35	3.18	5.84	12.92
4	0.74	1.53	2.13	2.78	4.60	8.61
5	0.73	1.48	2.02	2.57	4.03	6.87
6	0.72	1.44	1.94	2.45	3.71	5.96
7	0.71	1.42	1.90	2.37	3.50	5.41
8	0.71	1.40	1.86	2.31	3.367	5.04
9	0.70	1.38	1.83	2.26	3.25	4.78
10	0.70	1.37	1.81	2.23	3.17	4.590
11	0.70	1.36	1.80	2.20	3.11	4.44
12	0.70	1.36	1.78	2.18	3.06	4.32
13	0.69	1.35	1.77	2.16	3.01	4.22
14	0.69	1.35	1.76	2.15	2.98	4.14
15	0.69	1.34	1.75	2.13	2.95	4.07
16	0.69	1.34	1.75	2.12	2.92	4.02
17	0.69	1.33	1.74	2.11	2.90	3.97
18	0.69	1.33	1.73	2.10	2.88	3.92
19	0.69	1.33	1.73	2.09	2.86	3.88
20	0.69	1.33	1.73	2.09	2.85	3.85
21	0.69	1.32	1.72	2.08	2.83	3.82
22	0.69	1.32	1.72	2.07	2.82	3.79
24	0.69	1.32	1.71	2.06	2.80	3.75
26	0.68	1.32	1.71	2.06	2.78	3.71
28	0.68	1.31	1.70	2.05	2.76	3.67
30	0.68	1.31	1.70	2.04	2.75	3.65
35	0.68	1.31	1.69	2.03	2.72	3.59
40	0.68	1.30	1.68	2.02	2.70	3.55
45	0.68	1.30	1.68	2.01	2.70	3.52
50	0.68	1.30	1.68	2.01	2.68	3.50
60	0.68	1.30	1.67	2.00	2.66	3.46
70	0.68	1.29	1.67	1.99	2.65	3.44
80	0.68	1.29	1.66	1.99	2.64	3.42
90	0.68	1.29	1.66	1.99	2.63	3.40
100	0.68	1.29	1.66	1.99	2.63	3.39
Probability (p) that chance alone could produce the difference	0.50 (50%)	0.20 (20%)	0.10 (10%)	0.05 (5%)	0.01 (1%)	0.001 (0.1%)

5 Look down the column of significance levels (p) at the 0.05 level until you reach the line corresponding to your df. This is your critical value.

6 If the calculated value of t exceeds this critical value at the 0.05 level of significance, <u>it indicates that there is a less than 0.05 probability (5%) that the difference between the two means is due to chance.</u>

Therefore, we reject the null hypothesis, and conclude the difference between the two samples is significant.
(Significant means NOT due to chance.)

T-test for Group A and Group B data

∗ Calculated t = 3.2675

∗ Critical value of t (from t table) = 2.101
df = 18   (there were 20 samples)
p=.05 (less than a 5% probability that the differences observed are due to chance)

Since the calculated t is greater than the critical value we can reject our null hypothesis (that the 2 data sets are the same) and accept the alternate hypothesis (the 2 data sets are statistically different).

T-test summary
(p=.05)

* **If the calculated value of t is GREATER THAN the value from the critical values table, then the 2 sets of data ARE statistically different.**

* **BUT,**

* **If the calculated value of t is LESS THAN the value from the critical values table, then the 2 sets of data ARE NOT statistically different.**

1.1.6

* **Explain that the existence of a correlation does not establish that there is a causal relationship between two variables.**

What is the difference between causation and correlation?

* One of the most common errors we find in the press is the confusion between *correlation* and *causation* in scientific and health-related studies.

* Correlation is just a relationship between two sets of data.

* Causation is a specific reason for a correlation.

* Correlation DOES NOT have to mean causation.

Correlation vs. Causation

* **In theory, these are easy to distinguish — an action or occurrence can *cause* another (such as smoking causes lung cancer), or it can *correlate* with another (such as smoking is correlated with alcoholism).**

* **If one action causes another, then they are most certainly correlated.**

* **But just because two things occur together does not mean that one caused the other, even if it seems to make sense.**

But…

Unfortunately, our intuition can lead us astray when it comes to distinguishing between causality and correlation.

For example, eating breakfast has long been correlated with success in school for elementary school children. It would be easy to conclude that eating breakfast *causes* students to be better learners.

It turns out, however, that those who don't eat breakfast are also more likely to be absent or tardy — and it is absenteeism that is playing a significant role in their poor performance.

When researchers retested the breakfast theory, they found that, independent of other factors, breakfast only helps undernourished children perform better.

Next class you will turn in. . .

*Work through questions 1-6 on pages 10 and 11 of your text. Answer in your science notebook.