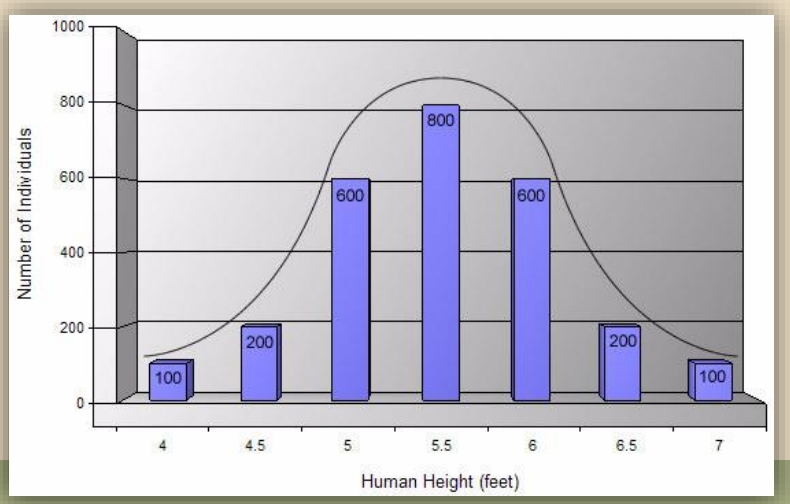
AP Biology Intro to Statistics

Statistics

- Statistical analysis is used to collect a sample size of data which can infer what is occurring in the general population
  - More practical for most biological studies
  - Requires math and graphing data
- Typical data will show a normal distribution (bell shaped curve).
  - Range of data



Statistical Analysis

- Two important considerations
  - How much variation do I expect in my data?
  - What would be the appropriate sample size?

Measures of Central Tendencies

- Mean
  - Average of data set
- Median
  - Middle value of data set
  - Not sensitive to outlying data
- Mode
  - Most common value of data set

Measures of Average

- Mean: average of the data set
  - Steps:
    - Add all the numbers and then divide by how many numbers you added together

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i. \quad = \quad \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots\ldots + x_n}{n}$$

Example: 3, 4, 5, 6, 7

3+4+5+6+7= 25
25 divided by 5 = 5
The mean is 5

Measures of Average

- Median: the middle number in a range of data points
  - Steps:
    - Arrange data points in numerical order. The middle number is the median
    - If there is an even number of data points, average the two middle numbers
- Mode: value that appears most often

Example: 1, 6, 4, 13, 9, 10, 6, 3, 19

1, 3, 4, 6, 6, 9, 10, 13, 19

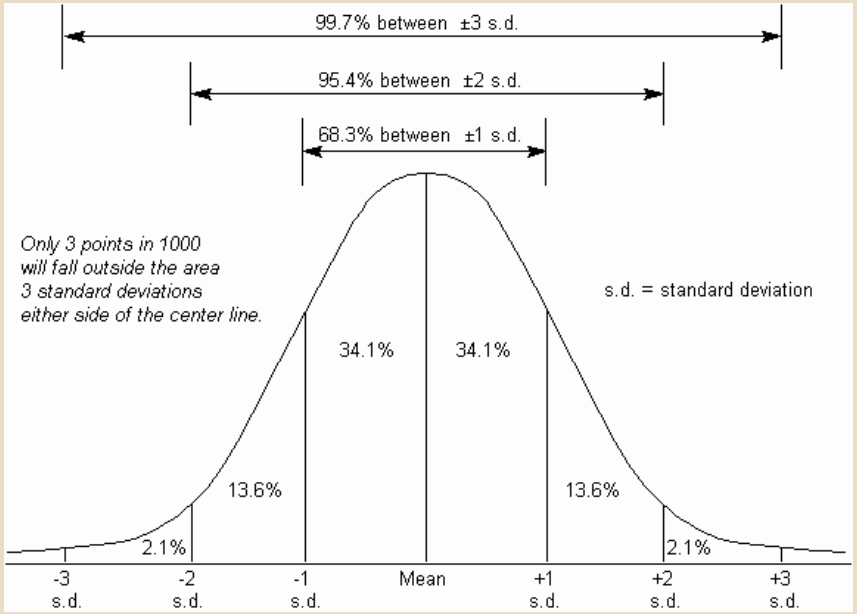
Median = 6

Mode = 6

Measures of Variability

- Standard Deviation
  - In normal distribution, about 68% of values are within one standard deviation of the mean
  - Often report data in terms of +/- standard deviation
- It shows how much variation there is from the "average" (mean).
  - If data points are close together, the standard deviation with be small
  - If data points are spread out, the standard deviation will be larger

Standard Deviation



- 1 standard deviation from the mean in either direction on horizontal axis represents 68% of the data
- 2 standard deviations from the mean and will include ~95% of your data
- 3 standard deviations form the mean and will include ~99% of your data

- Bozeman video: Standard Deviation

Calculating Standard Deviation

Calculation of the standard deviation of a list of numbers can be made easier by using a table:

Difference between measured value and mean
2 - 6.4 = -4.4

Square of number in previous column
-4.4 x -4.4 = 19.6

Measure number	Measured value, x	$(x - \bar{X})$	$(x - \bar{X})^2$
1	2	-4.4	19.36
2	5	-1.4	1.96
3	12	5.6	31.36
...	...	...	...
60	7	0.6	0.36
61	4	-2.4	5.76
Total:	392	Total:	199.66
Mean, $\bar{X}$	6.4	Std. Dev.:	1.82

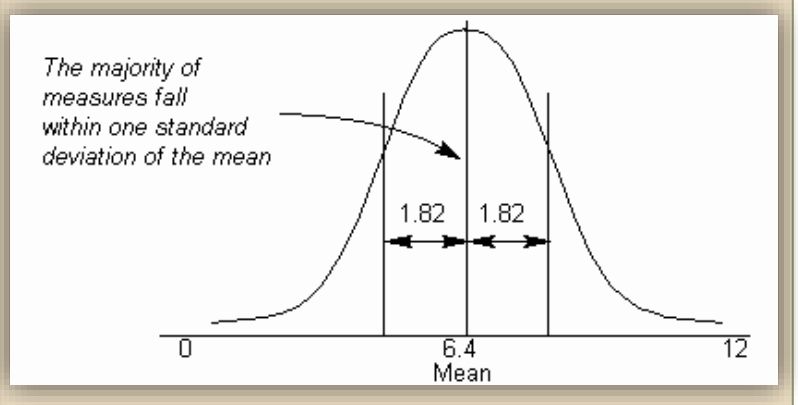
Total number of measurements made, n

Sum of column: $\sum (x - \bar{X})^2$

Average of measured values
$\frac{392}{61} = 6.4$

Standard Deviation = $\sqrt{\frac{\sum (x - \bar{X})^2}{n - 1}} = \sqrt{\frac{199.66}{61 - 1}} = 1.82$

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



Calculating Standard Deviation

Grades from recent quiz in AP Biology:

96, 96, 92, 90, 88, 86, 86, 84, 80, 70

1st Step:

find the mean ($\overline{X}$)

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

Measure Number	Measured Value x	$(x - \overline{X})$	$(x - \overline{X})^2$
1	96	9	81
2	96	9	81
3	92	5	25
4	90	3	9
5	88	1	1
6	86	-1	1
7	86	-1	1
8	84	-3	9
9	80	-7	49
10	70	-17	289
TOTAL	868	TOTAL	546
Mean, X	87	Std Dev	

Calculating Standard Deviation

2nd Step:

determine the deviation from the mean for each grade then square it

$$\sum (X - \overline{X})^2$$

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

Measure Number	Measured Value x	$(x - \overline{X})$	$(x - \overline{X})^2$
1	96	9	81
2	96	9	81
3	92	5	25
4	90	3	9
5	88	1	1
6	86	-1	1
7	86	-1	1
8	84	-3	9
9	80	-7	49
10	70	-17	289
TOTAL	868	TOTAL	546
Mean, $\overline{X}$	87	Std Dev	

Calculating Standard Deviation

Measure Number	Measured Value x	$(x - \overline{X})$	$(x - \overline{X})^2$
1	96	9	81
2	96	9	81
3	92	5	25
4	90	3	9
5	88	1	1
6	86	-1	1
7	86	-1	1
8	84	-3	9
9	80	-7	49
10	70	-17	289
TOTAL	868	TOTAL	546
Mean, $\overline{X}$	87	Std Dev	

Step 3:

Calculate degrees of freedom (n-1)

where n = number of data values

So, 10 – 1 = 9

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

Calculating Standard Deviation

Measure Number	Measured Value x	$(x - \overline{X})$	$(x - \overline{X})^2$
1	96	9	81
2	96	9	81
3	92	5	25
4	90	3	9
5	88	1	1
6	86	-1	1
7	86	-1	1
8	84	-3	9
9	80	-7	49
10	70	-17	289
TOTAL	868	TOTAL	546
Mean, $\overline{X}$	87	Std Dev	8

Step 4:

Put it all together to calculate S

$S = \sqrt{(546/9)}$

$= 7.79$

$= 8$

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

Calculating Standard Error

- So for the class data:
  - Mean = 87
  - Standard deviation (S) = 8

- 1 s.d. would be (87 – 8) thru (87 + 8) or 81-95
  - So, 68.3% of the data should fall between 81 and 95
- 2 s.d. would be (87 – 16) thru (87 + 16) or 71-103
  - So, 95.4% of the data should fall between 71 and 103
- 3 s.d. would be (87 – 24) thru (87 + 24) or 63-111
  - So, 99.7% of the data should fall between 63 and 111

Measures of Variability

- Standard Error of the Mean (SEM)
  - Accounts for both sample size and variability
  - Used to represent uncertainty in an estimate of a mean
  - As SE grows smaller, the likelihood that the sample mean is an accurate estimate of the population mean increases

Calculating Standard Error

Using the same data from our Standard Deviation calculation:

Mean = 87

S = 8

n = 10

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$SE_X = 8/\sqrt{10}$

$= 2.52$

$= 2.5$

Bozeman video: Standard Error

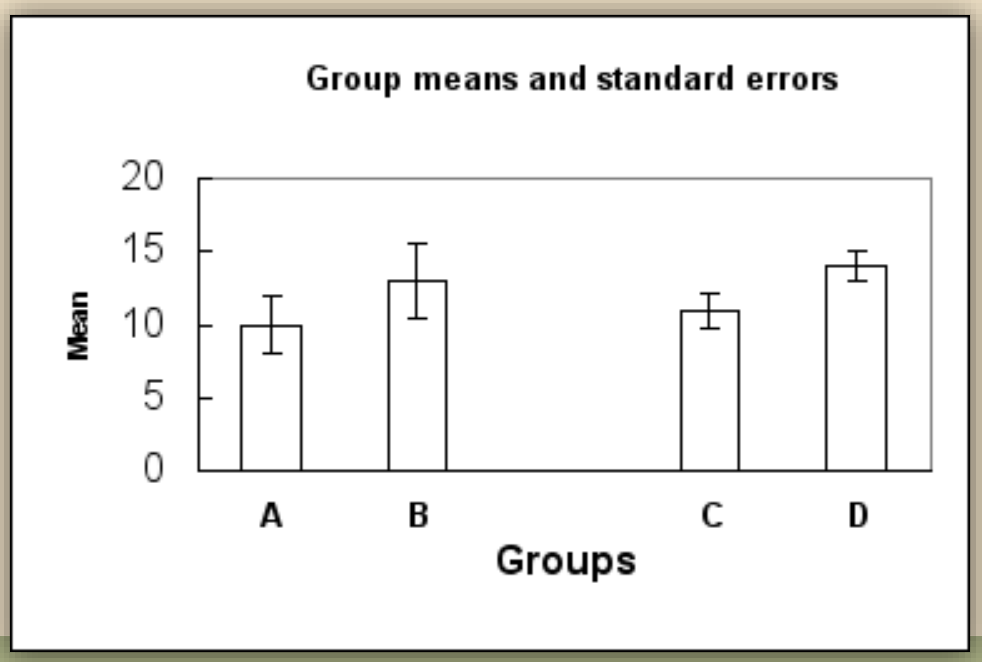
This means the measurements vary by $\pm$ 2.5 from the mean

Graphing Standard Error

- Common practice to add standard error bars to graphs, marking one standard error above & below the sample mean (see figure below). These give an impression of the precision of estimation of the mean, in each sample.

Which sample mean is a better estimate of its population mean, B or C?

Identify the two populations that are most likely to have statistically significant differences?



Practice Problems

- Consider the following three data sets A, B and C.

  A = {9,10,11,7,13}

  B = {10,10,10,10,10} Find

  C = {1,1,10,19,19}

  a) Calculate the mean of each data set.

  b) Calculate the standard deviation of each data set.

  c) Which set has the largest standard deviation?

  d) Is it possible to answer question c) without calculations of the standard deviation?

Practice Problems

- What is the population standard deviation for the numbers: 75, 83, 96, 100, 121 and 125?

- Ten friends scored the following marks in their end-of-year math exam:
  23%, 37%, 45%, 49%, 56%, 63%, 63%, 70%, 72% and 82%

  What was the standard deviation of their marks?