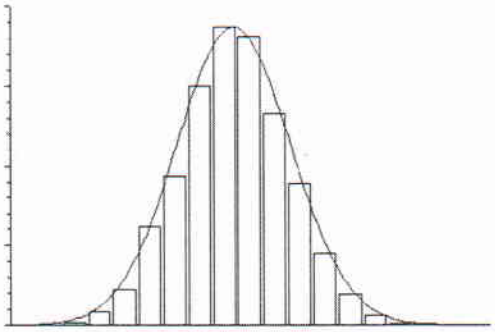
November 20, 2014

BioStatistics Problem Set 2

1. **Describe what a "normal distribution" of data means. What does the graph of data that is normally distributed typically look like?**

   Data is considered to be normally distributed when the bulk of the graphed values lie in the middle of the data spread. As you move further away from the middle, there are fewer and fewer measurements with more extreme values. Normally distributed data generates a bell-shaped curve as shown below:



2. **You are interested in measuring the amount of viral production from the new 2014 influenza strain. DNA constructs from this influenza strain are introduced into an appropriate host cell line and after three days, you measure the amount of viral proteins released in the cell culture media. To get as accurate a measurement as possible for virus production, you repeat the experiment 5 times, recording viral protein levels to generate the data table below. Calculate the sample mean ($\bar{x}$) and standard deviation (s) for the data you collected:**

Experiment #	Viral protein production (µg protein/mL media)
1	12.4
2	18.0
3	11.3
4	15.2
5	23.5

$$\bar{x} = \frac{12.4 + 18.0 + 11.3 + 15.2 + 23.5}{5} = \frac{80.4}{5} = 16.08$$

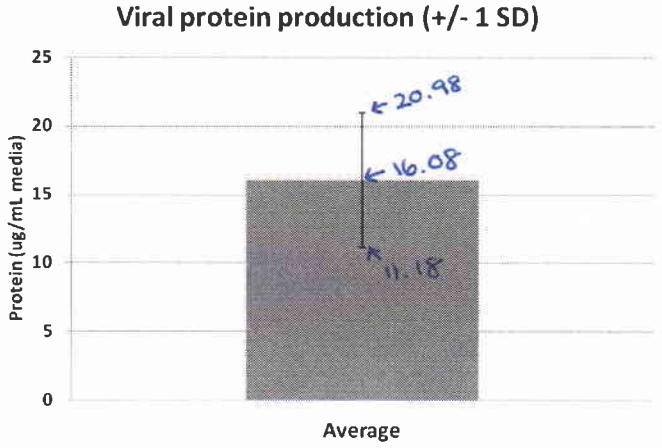
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}} \Rightarrow \sqrt{\frac{(12.4-16.08)^2 + (18-16.08)^2 + (11.3-16.08)^2 + (15.2-16.08)^2 + (23.5-16.08)^2}{(5-1)}}$$

$$= \sqrt{\frac{(-3.68)^2 + (1.92)^2 + (-4.78)^2 + (-0.88)^2 + (7.42)^2}{4}}$$

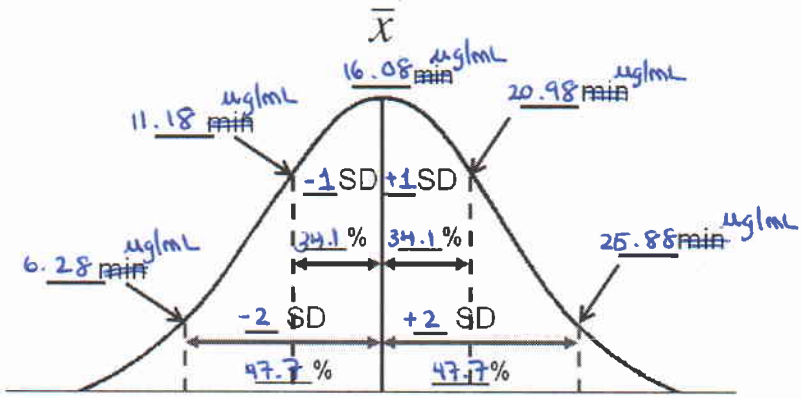
$$= \sqrt{\frac{13.54 + 3.69 + 22.85 + 0.77 + 55.06}{4}} = \sqrt{\frac{95.91}{4}} = \sqrt{23.98} = 4.$$

Standard deviati

Create a bar graph plotting ($\bar{x}$) and use the standard deviation for + and - error bars. Be sure to label all axis' and units correctly.



3. **Fill the values in the distribution curve below using your analysis of the above data.**



4. **In the data from problem 2, what is the sample variance (s^2)? What are the degrees of freedom (df)?**

The sample variance (s^2) is just the standard deviation squared.
In this case it's $(4.9)^2$ = 24.01 ug/mL media

Degrees of freedom for a single sample population = n-1.
Here we have 5 experiments, so DF = 5-1 = 4

5. **Explain the difference between the standard deviation (*s*) and the standard error of the mean (SEM). What happens to each value as your sample size (*n*) increases?**

The sample standard deviation is a measure of how variable measurements within a sample population are from the sample mean. It indicates how "spread out" the data are. The larger the standard deviation, the more widespread the values are within a sample population. Standard error of the mean (SEM) indicates how close the sample mean is to the population mean.

The sample size (*n*) has no significant effect on the standard deviation. The standard deviation only depends on how variable the measurements in the sample population are. If there's a lot of variance in a sample population (whether *n* is large or small), the standard deviation will be large. Unlike standard deviation, the SEM will always decrease as *n* increases because the more measurements you take from the population, the closer your sample mean is to the population mean.

6. **A local farm would like to compare the growth rate of tomatoes grown outside in natural sunlight versus 24-hour artificial light indoors. They are hoping that by increasing the rate of growth, they'll be able to fulfill an unusually high demand for tomatoes. They plant seedlings under both conditions and measure the height of each plant after 8 weeks:**

	Sunlight	Artificial light
Height (cm)	67	91
	55	87
	62	101
	71	98
	58	106
	69	88

a. **Calculate $\bar{x}$, SEM and 95% CI for each of the two groups (round to nearest whole number).**

Sunlight $\bar{x} = \frac{67+55+62+71+58+69}{6} = \frac{382}{6} = 63.67\text{ cm}$

$SEM = \frac{s}{\sqrt{n}}$ ← stand. deviation $\longrightarrow s = \sqrt{\frac{(67-63.67)^2 + (55-63.67)^2 + (62-63.67)^2 + (71-63.67)^2 + (58-63.67)^2 + (69-63.67)^2}{(6-1)}}$

$= \sqrt{\frac{(3.33)^2 + (-8.67)^2 + (-1.67)^2 + (7.33)^2 + (-5.67)^2 + (5.33)^2}{5}} = \sqrt{\frac{11.09 + 75.17 + 2.79 + 53.72 + 32.15 + 28.41}{5}}$

$= \sqrt{\frac{203.33}{5}} = \sqrt{40.66} = 6.38, \rightarrow SEM = \frac{6.38}{\sqrt{6}} = 2.6$

Sunlight, cont. 95% CI = $1.96\left(\frac{S}{\sqrt{n}}\right) \approx 2 \times SEM$, so 95% CI = 2(2.6) = 5.2

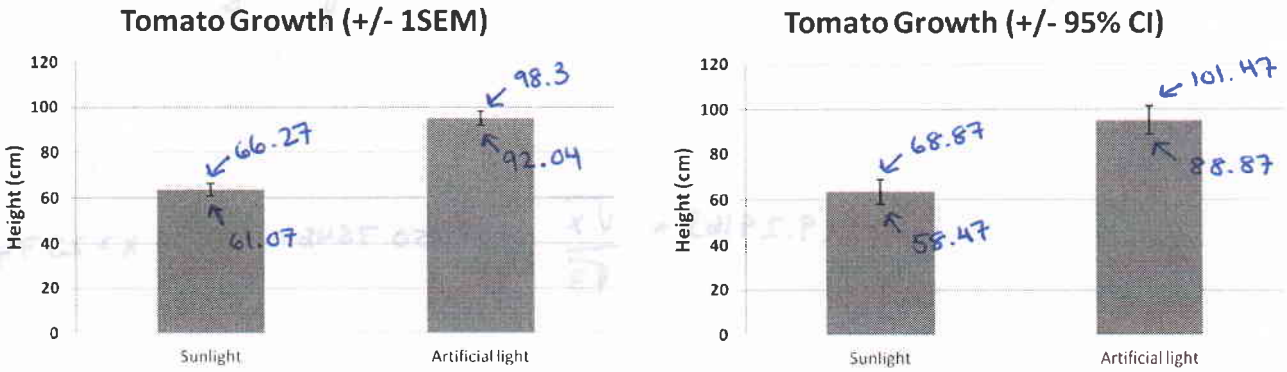
November 20, 2014

Artificial Light $\bar{X} = \frac{91+87+101+98+106+88}{6} = 95.17$

$S = \sqrt{\frac{(91-95.17)^2+(87-95.17)^2+(101-95.17)^2+(98-95.17)^2+(106-95.17)^2+(88-95.17)^2}{(6-1)}} = 7.68$

$SEM = \frac{S}{\sqrt{n}} = \frac{7.68}{\sqrt{6}} = 3.13$ 95% CI = 2(3.13) = 6.30

b. Construct two bar graphs comparing the means of each group. In one graph plot error bars using the SEM. In the other, plot error bars using the 95% CI.



c. Looking at the error bars, would you conclude there's a statistically significant difference in the growth rate of tomatoes grown in sunlight versus artificial light? WHY?

I would conclude there is a statistically significant difference in the growth of tomatoes in sunlight versus artificial light because the error bars between the two groups in each graph do not overlap. It looks like the tomatoes grown in artificial light have grown more than those grown in sunlight. The SEM error bars do not overlap, and because the 95%CI bars also don't overlap, we can say with 95% confidence that tomatoes grown under artificial light have faster rates of growth that those grown in sunlight.

d. When comparing the two groups, what would the null hypothesis (H_0) be?

The null hypothesis would state that there is no difference in the average height between tomatoes grown in sunlight versus artificial light. Any difference observed between the two groups would be purely by chance and not statistically significant.

e. Perform a Student's *t*-test : What is t_{obs} and what is t_{crit} (α=0.05)?

November 20, 2014

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} = \frac{|63.67 - 95.17|}{\sqrt{\left(\frac{6.38^2}{6}\right) + \left(\frac{7.68^2}{6}\right)}} = \frac{|-31.5|}{\sqrt{\left(\frac{40.7}{6}\right) + \left(\frac{58.98}{6}\right)}}$$

$$= \frac{31.5}{\sqrt{6.78 + 9.83}} = \frac{31.5}{\sqrt{16.61}} = \frac{31.5}{4.07} = 7.74 = t_{obs}$$

From the student t-test chart of critical values: Degrees of freedom for a 2-population study = $(n_1 + n_2) - 2$. In this case, DF = $(6+6)-2 = 10$

From the table, the $t_{critical}$ for an $\alpha = 0.05$ with DF = 10 is $2.23 = t_{critical}$

f. **Based on the Student's *t*-test, would you accept or reject the null hypothesis? Is the difference between the two groups statistically significant?**

Because the $t_{obs} > t_{critical}$ $(7.74 > 2.23)$, I would reject the null hypothesis. Therefore, there <u>is</u> a statistically significant difference between the growth rate of tomatoes grown in sunlight vs. artificial light.

7. **Molecular biologists are attempting to clone a DNA polymerase, R1, from a highly replicative strain of bacteria. They are hoping that by isolating this strain, scientists will be able to synthesize DNA constructs in half the time it currently takes using the standard *Taq* polymerase. In separate tubes, they incubate equal amounts of Taq polymerase and R1 polymerase with a DNA template and radioactive nucleotides under conditions optimal for DNA polymerization using both enzymes. After allowing each reaction to incubate for 30 minutes, the free radioactive nucleotides are washed away and the radioactive signal from DNA polymers are measured (in units of radioactive counts per minute – or CPM). The experiments were repeated 6 times, however, in two instances the purity of R1 was questionable and therefore only 4 values were collected for R1:**

	Taq	R1
CPM (counts per minute)	398	500
	421	467
	490	482
	438	431
	390	N/A
	560	N/A

November 20, 2014

a. Calculate $\bar{x}$, SEM and 95% CI for each of the two groups (round to nearest whole number).

Taq : $\bar{x} = \frac{398 + 421 + 490 + 438 + 390 + 560}{6} = 449.5$ or 450

$$s = \sqrt{\frac{(398-449.5)^2 + (421-449.5)^2 + (490-449.5)^2 + (438-449.5)^2 + (390-449.5)^2 + (560-449.5)^2}{6-1}}$$

$s = \sqrt{\frac{20988.72}{5}} = 64.79$; SEM $= \frac{s}{\sqrt{n}} = \frac{64.79}{\sqrt{6}} = 26.45$ or 26

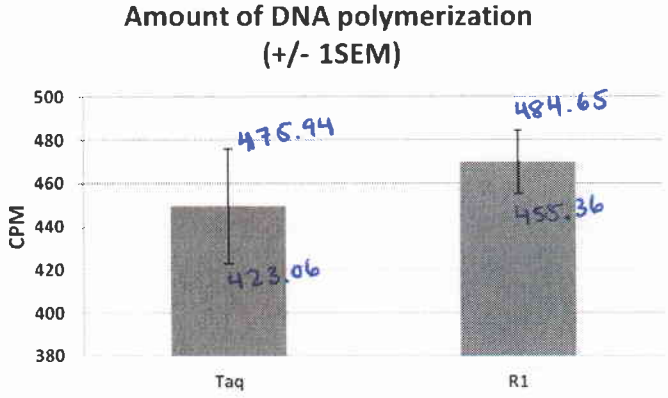
95% CI $\approx 2 \times$ SEM $= 2(26.45) = 52.90$ or 53

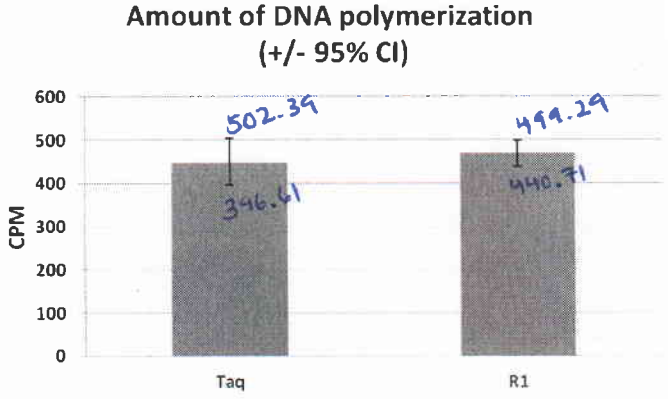
R1 : $\bar{x} = \frac{500 + 467 + 482 + 431}{4} = 470$

$$s = \sqrt{\frac{(500-470)^2 + (467-470)^2 + (482-470)^2 + (431-470)^2}{4-1}} = \sqrt{\frac{2574}{3}} = \sqrt{858} = 30$$

SEM $= \frac{s}{\sqrt{n}} = \frac{30}{\sqrt{4}} = \frac{30}{2} = 15$ 95% CI $\approx 2 \times$ SEM $= 2 \times 15 = 30$

b. Construct two bar graphs comparing the means of each group. In one graph plot error bars using the SEM. In the other, plot error bars using the 95% CI.





c. **Looking at the error bars, would you conclude there's a statistically significant difference in the polymerization rates between the Taq and R1 groups? WHY?**
In this case, because the error bars of the R1 and Taq groups overlap (in both the SEM and 95% CI graphs), I would conclude that there's no statistically significant difference in the polymerization rates between the two enzymes.

d. **When comparing the two groups, what would the null hypothesis (H_0) be?**
The null hypothesis would state that there's no statistically significant difference in the means between the R1 and Taq groups. Any difference observed would be due to chance.

e. **Perform a Student's *t*-test : What is t_{obs} and what is t_{crit} (α=0.05)?**

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} = \frac{|450 - 470|}{\sqrt{\left(\frac{26^2}{6}\right) + \left(\frac{30^2}{4}\right)}} = \frac{|-20|}{\sqrt{112.67 + 225}} = \frac{20}{\sqrt{337.67}} = 1.09 = t_{obs}$$

For $t_{critical}$, first determine degrees of freedom (DF). DF for a 2-population study = $(n_1 + n_2) - 2$
In this case, DF = $(6 + 4) - 2 = 8$.

f. **Based on the Student's *t*-test, would you accept or reject the null hypothesis? Is there a statistically significant increase in the rate of DNA polymerization with the R1 polymerase compared to the regular Taq polymerase?**

Because in this example, $t_{obs} < t_{critical}$ $(1.09 < 8)$, we can't reject the null hypothesis. From this data, ~~At this point,~~ there's no statistically significant difference between the rate of DNA polymerization of Taq versus R1. We can't conclude that the rate of polymerization by R1 is more than Taq.

Table 7. Critical *t*-Values for a Significance Level $\alpha = 0.05$

Degrees of Freedom (df)	t_{crit} ($\alpha = 0.05$)
1	12.71
2	4.30
3	3.18
4	2.78
5	2.57
6	2.45
7	2.36
8	2.31
9	2.26
10	2.23
11	2.20
12	2.18
13	2.16
14	2.14
15	2.13
16	2.12
17	2.11
18	2.10
19	2.09
20	2.09