Section 1. Short answers (1 point each).

1. To determine how vulnerable basements are to raccoon break-ins in the Stony Brook area, a BIO211 student sets up camera traps at basement entrances throughout dorms at SBU. This sample is inadequate because:
   1. All basements in the area have equal probability of being sampled,

**Incorrect! If they did the sample would be adequate**

* 1. All raccoons in the area have equal probability of being sampled,

**Incorrect! We are interested in the basements and their vulnerability, raccoons may or may not be randomly selected, what matters is the basements not the raccoons themselves.**

* 1. A few raccoons may responsible for most break-ins generating pseudoreplication,

**Incorrect! We are interested in the basements and their vulnerability, raccoons may be biased, but what matters is the basements not the raccoons themselves.**

* 1. The SBU dorms are a biased sample of basements in the area,

**Correct! The SBU dorms are a particular subset of the basements in the area.**

* 1. The SBU dorms are a random sample of basements in the area.

**Incorrect! Dorms are definitely not random compared to all the basements in the area.**

1. Your classmate is tasked with determining the mean size of oysters harvested off the North Fork. However, she is unsure as to how sampling different numbers of oysters will change her estimates. Thus, your suggestion to her is:
   1. As you measure more oysters, the standard error of the mean will be smaller,

**Correct! The standard error of the mean (or the standard deviation of the sampling distribution of the mean) decreases as sample size increases,**

* 1. As you measure more oysters, the mean size will be smaller,

**Incorrect! The estimates of the mean may change as you sample more, but as you sample more, the error associated with the estimate of the mean decreases.**

* 1. As you measure more oysters, the variance of sampling distribution of the mean will be larger,

**Incorrect! The variance of the sampling distribution of the mean is given by the square of the standard error of the mean. Therefore as sampling size increases, this variance decreases.**

* 1. As you measure more oysters, the standard deviation of the sampling distribution will be larger,

**Incorrect! The standard deviation of the sampling distribution of the mean is given by the standard error of the mean. Therefore as sampling size increases, this standard deviation decreases.**

* 1. As you measure more oysters, the standard deviation of the population distribution will be larger.

**Incorrect! The standard deviation of the population as a whole is not dependent on whether the sample is small or large, it is a fixed parameter not associated with my sampling.**

1. In a population of 7000 evening primrose (Oenothera spp.), there is an allele for self-incompatibility that has historically been found at a frequency of 60%. For your honors thesis you decide to examine allele frequency for this population since it was last measured >35 years ago by randomly sampling 10 individuals for a total of 20 alleles. What proportion of the self-incompatibility allele in your sample would you consider to be different from the historical frequency?
   1. More than 60% of the sample,

**Incorrect! In such a relatively small sample you may find >60% frequency for the allele of interest just by chance.**

* 1. Less than 50% of the sample,

**Incorrect! In such a relatively small sample you may find <40% frequency for the allele of interest just by chance.**

* 1. Anything 10 points away from 50% because there is a 50/50 chance of being the allele of interest or not

**Incorrect! The probability or chance of being the allele of interest is not 50/50, it is instead given by the proportion of the underlying population.**

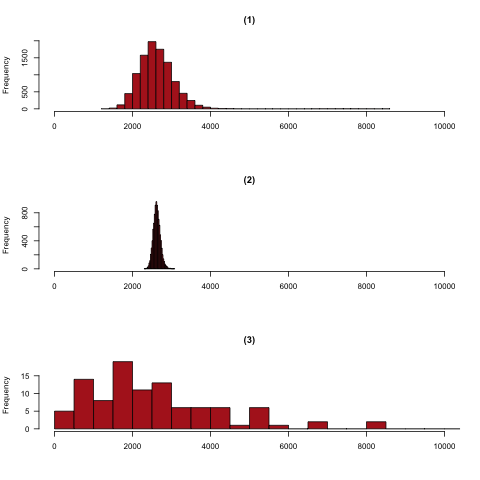
* 1. Less than 50% or more than 70% because there is a 50/50 chance of being the allele of interest or not, but the sample won’t be exact

**Incorrect! The probability or chance of being the allele of interest is not 50/50, it is instead given by the proportion of the underlying population.**

* 1. Only extreme values close to 0 or 100% because the sample is small and could estimate the underlying proportion poorly.

**Correct! The sample is relatively small and only very extreme values would not be compatible with the 60% underlying frequency of the population.**

1. The following three histograms (1 ,2, and 3) show information about the mean length of genes in the human genomes. One of them is shows the frequency distribution of individual values in a random sample of 100 genes (A). Another shows the distribution of sample means for samples of size 20 taken from all the genes in the genome (B). Another shows the distribution of the sample means for samples of size 500 (C). Which one is which?



* 1. 1 is A, 2 is B, 3 is C.

**Incorrect! The means will have lower spread than the raw values from a sample, and their spread will depend on the size of the sample taken from all the data.**

* 1. 3 is A, 2 is B, 1 is C.

**Correct! The means have lower spread than the raw values from a sample, and their spread depends on the size of the sample taken from all the data.**

* 1. 2 is A, 3 is B, 1 is C.

**Incorrect! The means will have lower spread than the raw values from a sample, and their spread will depend on the size of the sample taken from all the data.**

* 1. 3 is A, 1 is B, 2 is C.

**Incorrect! The spread of the mean with fewer samples should be less than with the mean from larger samples taken from all the data.**

* 1. 2 A, 1 is B, 3 is C.

**Incorrect! The means will have lower spread than the raw values from a sample, and their spread will depend on the size of the sample taken from all the data.**

1. The fecal egg count (FEC) is a measure of parasite load in populations of black rhinoceros. Stringer et al. (2014) analyzed FEC from many samples and examined the confidence intervals of estimates of the mean as they took larger samples (with more observations). In general, the confidence intervals of the mean FEC became:
   1. Independent from the size of the samples,

**Incorrect! Confidence intervals of the estimated mean change with sampling size.**

* 1. Wider as more observations were collected,

**Incorrect! The CI of the mean becomes wider with fewer observations.**

* 1. Narrower as more observations were collected,

**Correct! The CI of the mean becomes narrower with more observations.**

* 1. Narrower as smaller samples were collected,

**Incorrect! The CI of the mean becomes wider with fewer observations and smaller samples have fewer observations.**

* 1. Narrower as fewer samples were collected.

**Incorrect! The CI of the mean becomes wider with fewer observations and fewer samples will widen them.**

1. The height of students in BIO211 had a frequency distribution that was both bimodal and skewed to the right. Which of the following is the best explanation for this frequency distribution?
   1. Differences in height between men and women explain the bimodality, and a few short students explain the skew,

**While bimodality is linked to drawing a sample from populations that differ (e.g., in height), the skew is to the right so short students cannot explain it.**

* 1. Differences in height between men and women explain the bimodality, and a few tall students explain the skew,

**Correct! The skew is linked to extreme values, while bimodality to drawing a sample from populations that differ (in this case men and women differ in height).**

* 1. Differences in height between men and women explain the skew, and a few short students explain the bimodality,

**Incorrect! The skew is linked to extreme values, while bimodality to drawing a sample from populations that differ (in this case in height).**

* 1. Differences in height between men and women explain the skew, and a few tall students explain the bimodality,

**Incorrect! The skew is linked to extreme values, while bimodality to drawing a sample from populations that differ (in this case in height).**

* 1. Differences in height between men and women explain the bimodality, and the normal distribution explains the skew.

**Incorrect! The normal distribution is symmetrical and cannot explain the skew.**

1. Genetic bottlenecks have shaped the genetic diversity of many human populations. There is also a lively debate on whether certain alleles that elevate risks of disease are found in human populations persist because they also have a positive effect, or simply because bottlenecks have been common as a result of migration and as a result these “bad” alleles are more common than expected. Based on the in-class exercise, and considering an allele that elevates the risk of disease could be any of the colors you studied, which of the following statements do you agree with?
   1. On average, alleles were lost but there were few increases in frequency of any alleles,

**Incorrect! If alleles are lost that means other alleles go up in frequency, so this cannot be right.**

* 1. On average, alleles went up in frequency because of selection for certain colors,

**Incorrect! If the sample going through the bottleneck was random, no selection for or against any color took place.**

* 1. On average, some alleles went up in frequency because of the loss of certain other alleles during the drastic reduction in population size from the start to the first generation,

**Correct! The loss of certain alleles through sampling brings about the increase in frequencies of other alleles.**

* 1. On average, the bottleneck was not long enough to change allele frequencies,

**Incorrect! Although an individual group might by chance have samples with the original proportions, on average different groups had highly variable frequencies/proportions.**

* 1. On average, when the population went to 10 individuals, allele frequencies were conserved.

**Incorrect! Although an individual group might by chance have a sample with the original proportions, on average the groups changed frequencies/proportions, to the point of losing certain alleles completely.**

1. A random sample of measurements of the weight of 100 yellow-rump warblers is taken. Suppose a 95% confidence interval (CI) for the mean weight is 10 to 15 g. If a 99% CI is also calculated, then:
   1. The 99% CI is the better estimate,

**Incorrect! The 99% refers to the probability the CI includes the population mean, hence confidence, not whether on is better than the other.**

* 1. The 99% confidence interval will be wider than the 95% CI,

**Correct! To increase the probability that the CI include the parameter, the spread of the 99% CI has to be larger than that of the 95%.**

* 1. The 99% confidence interval will be narrower than the 95% CI,

**Incorrect! To increase the probability that the CI include the parameter, the spread of the 99% CI has to be larger than that of the 95%.**

* 1. 95% and 99% confidence intervals will be the same,

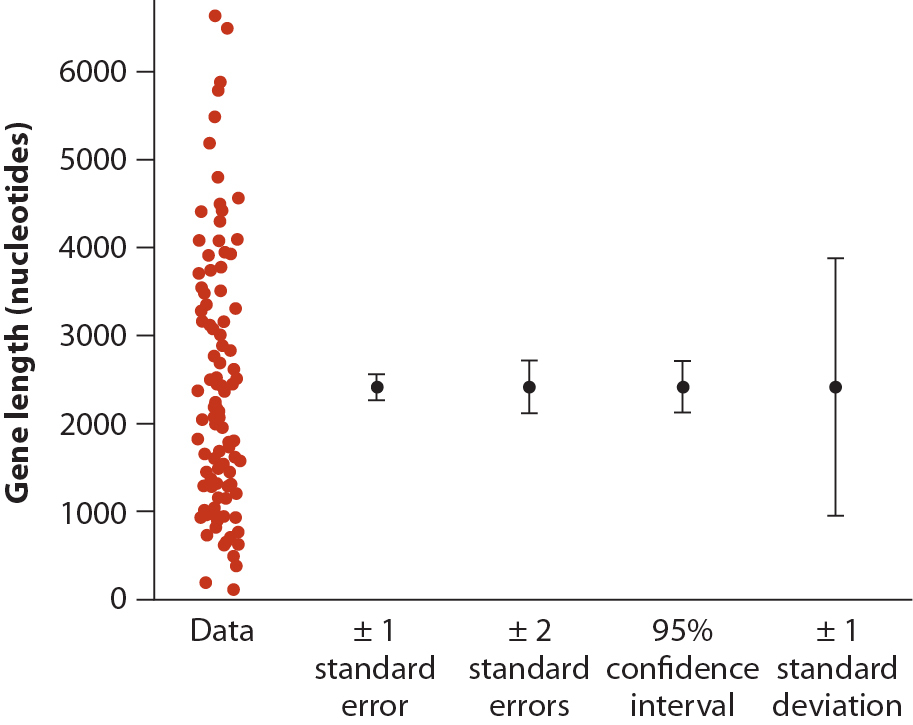
**Incorrect! To increase the probability that the CI include the parameter, the spread of the 99% CI has to be larger than that of the 95%.**

* 1. One cannot make a general statement about whether the 95% confidence interval would be narrower, wider or the same as the 99%.

**Incorrect! To increase the probability that the CI include the parameter, the spread of the 99% CI has to be larger than that of the 95%.**

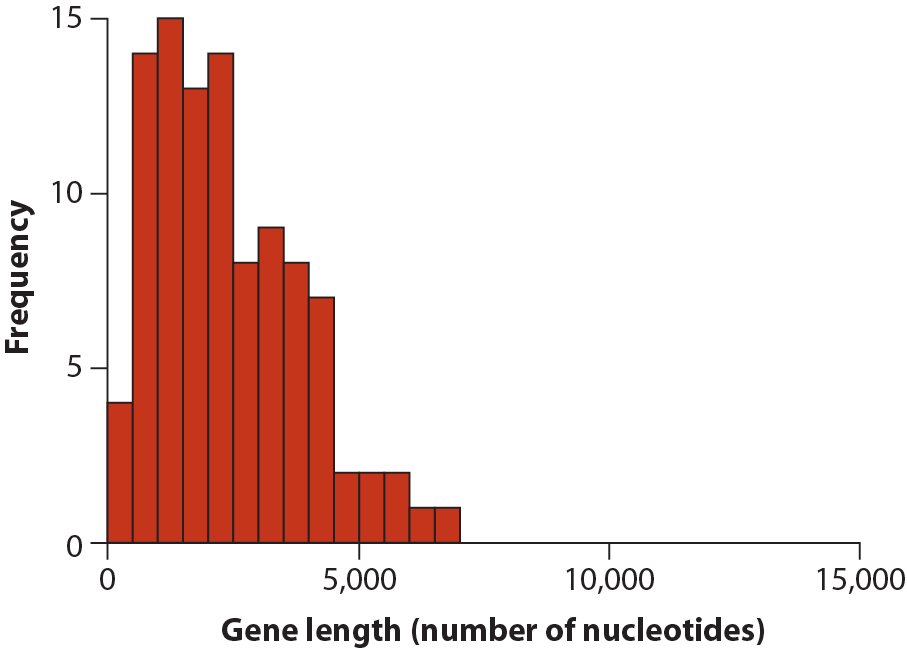
Part 2. Problems.

1. The following shows the length of a sample of 100 genes in the human genome, standard error and confidence intervals of the mean, and the standard deviation.



* 1. Based on these data, do you expect the frequency distribution of the genes to be symmetrical, left-skewed, or right-skewed? **Right-skewed**
  2. If instead of sampling 100 genes, you sampled 1000 which of the error bars shown do you expect not to change? **The standard deviation.**

1. The histogram below shows the frequency distribution of a unique random samples of *n* = 100 genes from the human genome ( = 2411.8, *s* = 1463.5). Go to <http://www.onlinestatbook.com/stat_sim/sampling_dist/index.html> and using the ‘custom’ setting, generate a frequency distribution similar to that of the data shown here (don’t worry about matching the mean etc., the scale is off) . Then, draw samples of N=2, to generate a sampling distribution of the mean. Repeat the procedure, this time with N=5. Compare and contrast the two sampling distributions of the mean. How do they differ? How are they similar? What is your conclusion regarding estimates of the mean from samples of N=2?



**The sampling distribution of the mean from N=2 is more similar to the frequency distribution of of the data. The sampling distribution of the mean from N=5 is symmetrical and normal or normal-like. Estimates of the mean from N=2 are not normal/don’t behave the same as larger samples/are skewed similar to the original data.**