Investigating sampling
Part 2: Confidence intervals

Student worksheet

1. Normal distribution

Let's look again at the data from our 100 different samples of 20 heights (see Appendix A).

Recall:

Mean of the sample means is approximately equal to the population mean. $\mu \cong 164.5$	**The standard deviation of the sample means is approximately equal to standard error.** $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10.2}{\sqrt{20}} = 2.28$

Clearly, the distribution of sample means appears to be normal. (In fact, if all possible samples of 20 heights were taken, the distribution would be perfectly normal.) Based on this result,

a) what percent of the sample means should fall within one standard error of the mean of the sample means?
b) what percent should fall within two standard errors?
c) verify these expectations using the sample means in Appendix A. Are they exactly as you expected? Explain.

What does this say about a single random sample of 20 heights?

This suggests that if we were to select **one** sample of 20 students at random from this population, the mean of the sample has a 68% chance of being within one standard error of the population mean and a 95% chance of being within two standard errors of the population mean.

2. The central limit theorem

We have seen that sample size plays a role in the distribution of the sample means. To ensure a normal distribution of the sample means, we choose 30 samples or more. Thus, we can state the following central limit theorem:

> **When $n \geq 30$, the distribution of the means, $\overline{X}$, of all random samples of size n is approximately normal, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.**

Consider this scenario:

Imagine you are trying to determine the mean height of students at your high school. It is impractical to use the heights of all the students, so you get height data for a random sample of 30 students.

- Do you expect the mean of this sample to be exactly the same as the mean of the population?
- Do you expect the sample mean to be close to the population mean? If so, how close do you expect these means to be?

The mean height of your random sample is called a *point estimate*. It's a single sample statistic used to estimate the population mean. Experience tells us that if a different random sample were chosen, we likely would get a different mean and thus a different point estimate. This is because of variation in the sample means.

To account for this, we determine an interval estimate of the true population mean by taking into consideration the sampling distribution of the mean. Depending on how confident we wish to be that the interval will contain the population mean, we can choose different-sized intervals. The larger the interval is, the more confidence we can claim. Is there a down side to making the interval large? Typically, we use intervals that can claim 95% confidence

A 95% confidence interval means that if we were to choose all possible samples of the same size, in 95% of the samples, the true population mean would be included in the interval around the sample mean in 95% of the samples.

Exercise: Examine different samples

Examine the four example diagrams on the next page. The samples are chosen from the original height data, with $\mu = 164.7$ and $\sigma = 10.21$. Sample 1 is explained below.

Sample 1

Given: $n = 30$ and $\overline{X} = 165.9$

Therefore $\sigma_{\overline{X}} = \dfrac{10.2}{\sqrt{30}} = 1.86$

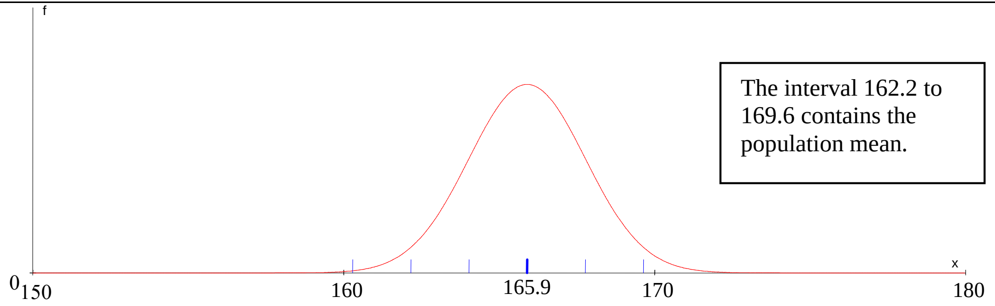
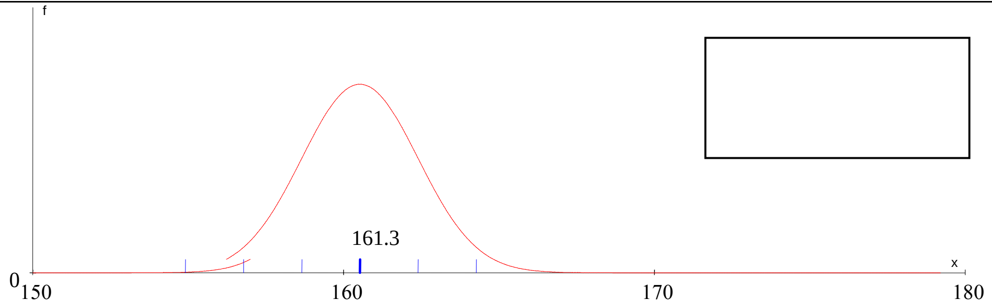
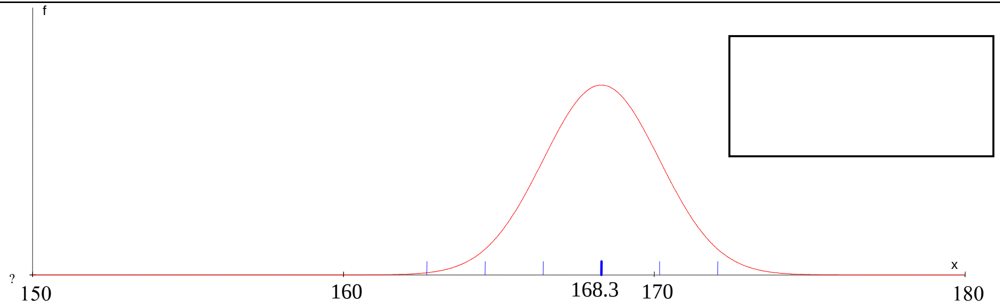
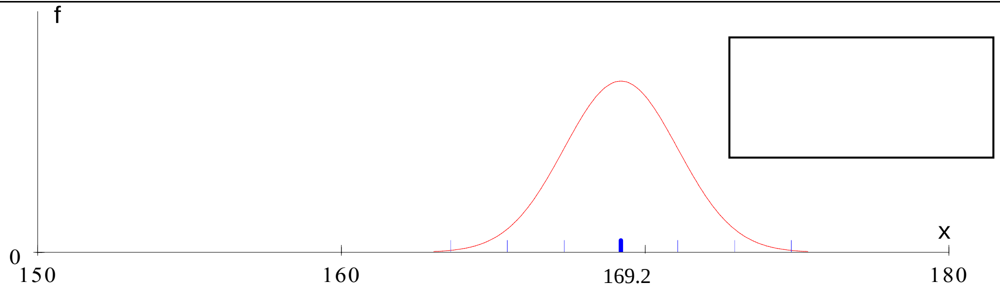
Determine the interval within $2\sigma_{\overline{X}}$ of $\overline{X}$.

$$\overline{X} - 2\sigma_{\overline{X}} = 165.9 - 2(1.86) = 162.2$$

$$\overline{X} + 2\sigma_{\overline{X}} = 165.9 + 2(1.86) = 169.6$$

Does it contain the population mean?

Perform similar calculations for Samples 2, 3 and 4. For each sample, determine whether the interval contains the population mean.

Sample 1 $n = 30$ $\overline{X} = 165.9$ $\sigma_{\overline{X}} = \frac{10.2}{\sqrt{30}} = 1.86$	 The interval 162.2 to 169.6 contains the population mean.
Sample 2 $n = 30$ $\overline{X} = 161.3$ $\sigma_{\overline{X}} = \frac{10.2}{\sqrt{30}} = 1.86$	
Sample 3 $n = 30$ $\overline{X} = 168.3$ $\sigma_{\overline{X}} = \frac{10.2}{\sqrt{30}} = 1.86$	
Sample 4 $n = 30$ $\overline{X} = 169.2$ $\sigma_{\overline{X}} = \frac{10.2}{\sqrt{30}} = 1.86$	

Since the distribution is normal, we can assume that any sample mean has a 95% chance of being within two $\sigma_{\overline{X}}$ of the population mean.

Clearly, not all samples will have a mean within two $\sigma_{\overline{X}}$ of μ. We expect, however, that 95% of the sample means will be within two $\sigma_{\overline{X}}$ of μ.

If we know the standard deviation of the population, we can use the standard error to determine the interval around the sample mean which has a 95% probability of enclosing the population mean.

In reality, we usually don't know the mean or the standard deviation of the population. Remember the scenario where we want to determine the mean height of students in our high school. In that case, we only know the heights of a random sample of 30 students.

What if we don't know what σ is?

We need to look again at the information that we do know and how it relates to σ.

Look at the standard deviations of the samples in the exercise. First, be aware that the calculation for the standard deviation of a sample is slightly different from the calculation of the standard deviation of the population. ($s_X = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$). Notice that your TI-83 calculator's 1-Var Stats menu provides both S_X (sample standard deviation) and σ_X (population standard deviation) functions.

The difference is that instead of dividing by the sample size (*n*), we will divide by *n-1*. In order to compute s_X, we first need to know what $\overline{X}$ is. Therefore, only *n-1* of the sample values are free to vary. The *nth* is determined since $\overline{X}$ has also been determined.

For example, suppose a sample of three values has a mean of 10. Since the mean is 10 and *n* = 3, only two distinct values need to be known before the last one is fixed; it is determined by the information since the sum must be 30.

Look at the s_X values in Appendix A. They are relatively close to the value of σ. In fact, a statistician named William S. Gosset developed a distribution that came to be known as *Student's distribution*. Based on his work, it is reasonable to replace σ by s_X in our formula for $\sigma_{\overline{X}}$, giving $\sigma_{\overline{X}} = \frac{s_X}{\sqrt{n}}$.

Thus, the 95% confidence interval for the population mean is given approximately by the formula:

$$\overline{X} \pm 2\frac{s_X}{\sqrt{n}}\text{, which means that } \overline{X} - 2\frac{s_X}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{s_X}{\sqrt{n}}$$

The variables can be determined from the random sample if $n \ge 30$. As well, both an interval estimate can be stated for the population and a level of confidence can be specified.

3. Project time

Go to the *Census at School* website at www.censusatschool.ca or to any other site where you can find reliable data. Decide what quantitative information you would like to explore and what population you wish to sample. Perform the calculations and write a brief newspaper-style report about the population based on your sample results.

Contributed by Anna Spanik, Math teacher, Halifax West High School, Nova Scotia.