Module 2: Environmental Sampling

2.2 Simple Random Sampling
The Sampling Distribution of the Mean
Confidence Intervals on the Mean

Simple Random Sampling

- A simple random sample (SRS) is one that gives each sample unit an equal chance of being selected to be in the sample.
- Using SRS, the sample statistics are the same as shown in Modules 1.2 and 1.4. Other sampling designs, such as stratified and systematic sampling, result in different equations for calculating the values of the sample statistics.

Simple Random Sampling

- Sample statistics under SRS:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

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

Simple Random Sampling

- Another useful measure is the coefficient of variation CV(y) which is the standard deviation divided by the mean and gives the dispersion of the data as a proportion of the average. When multiplied by 100, it gives the dispersion as a percentage of the mean.

$$CV = \frac{s}{\overline{X}}$$

The Sampling Distribution of the Mean

- Just like data having underlying distributions that describe their probability characteristics, sample statistics have distributions as well. Distributions of statistics are called sampling distributions.
- The sampling distribution of the sample mean is a distribution of particular importance.

The Sampling Distribution of the Mean

- Imagine taking repeated samples of size n from a population and calculating the sample means. Those means could then be complied together into a histogram to display their probability distribution.
- With some thought, you can see that this distribution should be narrower than the distribution of the data since really high or low data values would be, to some extent, moderated by other data points when calculating the means.

The Sampling Distribution of the Mean

- For example, if the data follows a normal distribution with mean 50 and standard deviation of 10, then most of the data are in the range between 20 and 80. Means of samples of size 10 would tend to be close to 50 and would rarely be as low as in the 20s or in the 70s.

The Sampling Distribution of the Mean

- In fact, the central limit theorem gives us a theoretical way to determine the sampling distribution of the mean when data are drawn from a normal distribution. It is approximately true for data from any distribution, although it is less and less true as the underlying data distribution becomes more asymmetric.

The Sampling Distribution of the Mean

- If X is Normal (μ,σ) and N is relatively small, then X is Normal with mean μ and standard error:

$$\sigma^2(\bar{y}) = Var(\bar{y}) = \left(\frac{\sigma^2}{n}\right)\left(1 - \frac{n}{N}\right)$$

$$\sigma(\bar{y}) = SE(\bar{y}) = \sqrt{\left(\frac{\sigma^2}{n}\right)\left(1 - \frac{n}{N}\right)}$$

The Sampling Distribution of the Mean

- Note that the dispersion of a statistic is called its standard error .
- These are estimated, of course, by replacing σ with s.

The Sampling Distribution of the Mean

- When N is infinite, or even vary large, then the n/N term is zero or close to it so it disappears from the equations. Then

$$\sigma(\bar{y}) = SE(\bar{y}) = \sqrt{\left(\frac{\sigma^2}{n}\right)} = \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals on the Mean

- The sample mean estimates the population mean but would rarely be expected to be equal to it.

- A confidence interval is an interval, calculated from the sample mean and its standard error, that has a specified probability of containing the true parameter.

Confidence Intervals on the Mean

- For data from a normal distribution (at least approximately) and large samples (n>25), a (1-α)% confidence interval on μ is

$$\bar{y} \pm Z_{a/2} SE(\bar{y}) = \bar{y} \pm Z_{a/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \bar{y} \pm Z_{a/2}\left(\frac{s}{\sqrt{n}}\right)$$

Confidence Intervals on the Mean

- For small sample sizes, the sample standard deviation is not a very good estimate of the population standard deviation. Use the Student's t distribution instead of the normal:

$$\bar{y} \pm t_{a/2,n-1} SE(\bar{y}) = \bar{y} \pm t_{a/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Conclusion

- Similarly, population totals and proportions can be calculated. They have sampling distributions and standard errors, and confidence intervals can be calculated.
- These topics are covered more thoroughly in the text.