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 Module 1. 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:

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

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

Module 2.2

Simple Random Sampling

- Another useful measure is the coefficient of variation CV(Xbar) 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.

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

Module 2.2

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 30s or in the 60s.

The Central Limit Theorem and The Sampling Distribution of the Mean

- In fact, the Central Limit Theorem gives us a theoretical way to determine the sampling distribution of the sample 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 less symmetric.

The Central Limit Theorem and The Sampling Distribution of the Mean

- If Y is Normal (μ,σ) and N is relatively small, then $\bar{Y}$ is Normal with mean μ and variance 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 Central Limit Theorem and The Sampling Distribution of the Mean

- Note that the dispersion of a statistic is called its standard error.
- It's the standard deviation of the distribution that you would get if you took repeated samples and calculated the statistic for each sample and then formed a distribution of all of the sample statistics.
- These are estimated, of course, by replacing σ with s.

The Central Limit Theorem and 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

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

The Central Limit Theorem and The Sampling Distribution of the Mean

- So, the bottom line is that if Y is Normal (μ,σ)
- Then the sampling distribution of $\bar{Y}$ is Normal(μ,σ/sqrt(n))

The Central Limit Theorem and The Sampling Distribution of the Mean

- Additionally, if Y is not Normal but has mean μ and standard deviation σ
- Then the sampling distribution of $\overline{Y}$ is not Normal but it does approach N(μ,σ/sqrt(n)) as n becomes large

The Central Limit Theorem and The Sampling Distribution of the Mean

- This is an important result because it allows us to make inferences, do hypothesis testing, and create confidence intervals on the mean from the sample mean and sample standard deviation
- But, keep in mind they are only completely correct if the data distribution is normal
- If not, then the sample size n should be large

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, 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:

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

Confidence Intervals on the Mean

- The Student's t Distribution is like the Normal but wider and with fatter tails to account for the additional uncertainty that comes with not knowing σ and having to estimate it with the sample standard deviation, s.

Example Confidence Interval on the Mean

- Ok, let's do an example to show you how to construct a 95% confidence interval on the mean.
- You have a set of data
- You do a histogram, it's slightly skewed right
- n=30
- That's large enough that you are probably ok to use the standard Normal to get the confidence interval but I'm still going to use the t distribution, you just can't go wrong using the t distribution rather than Z.

Example Confidence Interval on the Mean

$\overline{Y}$ = 10.5

s = 2.1

n = 30

The 95% confidence interval on μ is

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

Example Confidence Interval on the Mean

So, you know everything to plug into the formula except the value from the t distribution

Go to Manly's table of the t statistic

Since n = 30, the degrees of freedom df = 30 - 1 = 29

Since it's a 95% CI, alpha is 5%, alpha/2 = 2.5%

Example Confidence Interval on the Mean

The reason you divide alpha by 2 is that half of the probability that is outside the confidence interval is in the upper tail and half is in the lower tail

Table A2.2 on page 274 of Manly shows the Critical Values for the t-Distribution in Upper Tail Probabilities. So, that's alpha/2.

Example Confidence Interval on the Mean

Go to the column for Upper Tail Probability = 0.025 (that's the same as 2.5%)

So, go to the row for df = 29

And what do you get?

$T_{0.025,29} = 2.045$

Example Confidence Interval on the Mean

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

So, $\bar{Y} = 10.5$

$T_{0.025,29} = 2.045$

$s = 2.1$

$n = 30$

Example Confidence Interval on the Mean

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

10.5 +/- 2.045(2.1/sqrt(30))
= 10.5 +/- 2.045(0.383406)
= 10.5 +/- 0.784065

95% CI on the true mean μ is (9.7, 11.3)

Example Confidence Interval on the Mean

95% CI on the true mean μ is (9.7, 11.3)

This means that, if you took repeated samples of size 30 from this same population that has mean μ, 95% of them would contain μ.

Some people say that this CI has a 95% probability of containing m, which, for technical reasons, isn't quite the way to say it.

Example Confidence Interval on the Mean

- **Tips:** Pay attention to the order of operations, it matters in what order you do the calculation. Start with the innermost parenthesis and work out.
- Always carry along as many digits as you can while doing the calculations.
- An easy way to do this is to use an Excel worksheet to do the calculations rather than a calculator.
- Always roundoff when you are ready to report the results, never report the 6 or so digits Excel gives you. Always round off to the number of decimal places in the data or, at most, one more.

Conclusion

- Similarly to means, other statistics such as 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.