I. Introduction

Chapter 23 Extends the material covered in Chapters 19, 20, and 21 to means (averages)

In other words, the Central Limit Theorem is still important

We need random samples
Of reasonable size
We need a normally distributed sampling distribution

II. The sampling distribution for means

The mean of the sampling distribution (the mean of all possible sample means $\bar{y}$) will be μ_y (mu) the mean of the population.

The standard deviation of the sampling distribution will be $SD(\bar{y}) = \sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$ that is the population standard deviation σ_y divided by the square root of the size of the sample.

Previously in Chapters 19-21 we looked at sample proportions. Your text points out that when you don't have the value of the population parameter p (for proportions) you can't find the standard deviation of the sampling distribution $SD(p) = \sigma_p = \sqrt{\frac{pq}{n}}$ BUT if you have the sample proportion $\hat{p}$ then you can figure out $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ because you can find $\hat{q}$ from knowing $\hat{p}$

But for means, we have a slightly different problem – if we know $\bar{y}$ it doesn't really tell us anything about $SD(\bar{y})$ so in practice what we do is SUBSTITUTE s for σ_y and then we get $SE(\bar{y}) = \frac{s}{\sqrt{N}}$ the standard error of the mean.

And it turns out that AS LONG AS YOUR SAMPLES ARE LARGE (let's say that large is at least size 50 but not greater than 10% of your population) this works fine. But people noticed problems when you had small samples and the sampling distribution did not appear to be normal.

III. Gosset's solution – the t distribution (optional)

Basically, what Gosset discovered is that a t-distribution has fatter "tails" than the normal model it's not as tall either but it is still bell-shaped. Consequently things like our 68-95-99.7 rule don't apply here. The t-distribution's shape changes with the sample size – the larger the sample the more normal it becomes.

Basically, the t-distribution and it's results (t-test, t-confidence interval) is what is used when you do not know the value of σ_y the true population standard deviation. When you have information on σ_y, you should use that information and use the normal (Z) table. When you do not have information on σ_y, you should use a t-table.

IV. Confidence Intervals and Hypothesis Tests when σ_y is known

Basic formula for a confidence interval for a mean

$\bar{y} \pm Z^* \frac{\sigma_Y}{\sqrt{n}}$ where the value of Z* depends on the level of confidence

Basic Z test for hypothesis tests involving means

$$Z = \frac{\bar{y} - \mu}{\sigma_Y / \sqrt{n}}$$

will result in a "p-value" and comparison against a pre-determined alpha level will determine if the result is statistically significant.

Using either of these formulas assumes that a random sample was drawn (and the process could be replicated if necessary) and that the samples are of reasonable size (at least size 50 but not larger than 10% of the population) and that the population standard deviation σ_y is known.

If σ_y was not known, you would need the t distribution, not the normal, to construct confidence intervals and perform hypothesis tests.

Hypothesis test example

In 1960, census results indicated that the age at which American men first married had a mean of 23.3 years with a standard deviation of 3.1 years. It is widely suspected that American men are waiting longer to get married. We want to find out if the mean age at first marriage has increased during the past 40 years. A random sample of 40 men who married for the first time was drawn and we calculated that their mean age was 24.2 years with a standard deviation of 5.3 years. Please assume a sample of size 40 was large enough.

$$H_0 : \mu_Y = 23.3 years$$
$$H_1 : \mu_Y > 23.3 years$$
$$Z = \frac{\bar{y} - \mu_y}{\sigma_y / \sqrt{n}} = \frac{24.2 - 23.3}{3.1/\sqrt{40}} = \frac{.9}{.4902} = +1.84$$

This test statistic says, in a way, that if the true parameter was 23.3 years and the distribution of all possible samples (sampling distribution) has a variation of .4902 years then the chance that you could have picked a sample of size 40 with a mean of 24.2 is about .0329 or about 3.29%. If your alpha=.05 then we would reject the null hypothesis and conclude that men are marrying at later ages.

Confidence Example

Suppose a psychologist wants to know the average IQ of the 28,000 students at USC. Suppose he takes a simple random sample of 100 and the sample average turns out to be 90. Suppose it is known that the population standard deviation of USC students is 50 but the population average is unknown. The psychologist will need your help in order to construct a 99% confidence interval to present the findings to the university president. Please show the psychologist how to do this.

$$90 \pm 2.58 * (\frac{50}{\sqrt{100}}) = 90 \pm 12.9$$

A 99% confidence interval for the mean IQ of all USC students is 77.1 to 102.9