Ch. 7: Estimates and Sample Sizes

Overview of Chapter 8 Concepts

The best point estimate of the population mean is the same mean, because:

1) It is more consistent, meaning that it has less variation, than any other estimator for the population
2) It is an unbiased estimator of the population mean, meaning that it symmetrically estimates over and under the population mean.

However, a better estimate than a **point estimate** exists. It is a confidence interval, which is likely to contain the true population mean, based upon Normal theory.

Here is how and what we will be doing in this chapter:

- We will use the standard normal distribution to define the confidence interval with, alpha (symbol: α) This is the probability that the true value of the mean is outside the interval created.

- We will say that the level of confidence/degree of confidence that the interval actually contains the mean is $1 - \alpha$. What this means is that if we were to create 100 confidence intervals based upon 100 different samples, then we should expect $1 - \alpha$ of those intervals to contain the true population mens.

§7.2 Estimating p in the Binomial Distribution

Notation

p = population proportion

$\hat{p} = \frac{x}{n}$ the sample proportion (read: p-hat)

x = # of successes (recall binomial)

n = # of trials

$\hat{q} = 1 - \hat{p}$ the complement of the sample proportion

α = probability that r.v. is in the tails of the dist.

$1 - \alpha$ = Level of Confidence (be careful of how you interpret this!)

$z_{\alpha/2}$ the critical value for our confidence interval

Given this notation, then we can create $(1-\alpha)\%$ confidence interval for the true population mean in the following way:

If the following assumptions are met:

Binomial Assumptions
1) Fixed number of trials
2) Trials are independent
3) Two categories for the outcomes
4) Probabilities remain constant for each trial

Normality Assumption ➔ 5) $np \geq 5$ & $nq \geq 5$

(1– α)% Confidence Interval for Population Proportion

With margin of error: $\mathbf{E = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}}$

The (1 – α)% CI for p is: $\hat{p} \pm E$ which creates an interval for which one has a (1–α) level of confidence, that it contains the true population proportion.

This interval can be written in the form: $\hat{p} - E < p < \hat{p} + E$ **or** $(\hat{p} - E, \hat{p} + E)$

Now we will do a quick example so you can get the hang of the computation and then we will do an example with the data that you created in our last lab on random number generation and finding sample proportions.

Example: A sample survey at ta supermarket showed that 204 of 300 shoppers use Cent's-Off coupons regularly. Find a 99% confidence interval for the true population proportion of shoppers that use Cent's-Off coupons.

Step 1: Determine the values of: n = ________ & x = _______

Step 2: Calculate $\hat{p}$ & $\hat{q}$ $\hat{p}$ = ________ & $\hat{q}$ = ________

Step 3: Find α [(1 – Level of confidence)] α = _________

Step 4: Find $z_{\alpha/2}$ by looking up in

1) large t, in t-table using two-tailed test for α
2) use critical value for 99% listed on z-table
3) look up (Level of Confidence) + $^{\alpha}/_{2}$ in the body of positive z-value table (right-tail)
4) look up $^{\alpha}/_{2}$ in the body of the negative z-table (left-tail)

Step 5: Calculate the margin of error E = ___________

Step 6: Give the confidence interval in interval notation

Here are a couple more examples that we can work together or you can practice on your own.

Example: In a random sample of 250 T.V. viewers in a certain area, 190 had seen a certain controversial program. Find a 90% CI for the true population proportion.

Example: A study is being made to estimate the proportion of voters in a sizeable community who favor the construction of a nuclear power plant. It is found that only 140 of the 400 voters selected at random favor the project, find a 95% CI for the proportion of all voters in this community who favor the project.

Our knowledge of confidence intervals can also be used in a different way. We can use it to calculate a sample size to poll in order to achieve a set level of confidence and margin of error. This formula is found by solving the margin of error for n.

$$\mathbf{n} = \frac{[z_{\alpha/2}]^2 \hat{p}\,\hat{q}}{E^2}$$

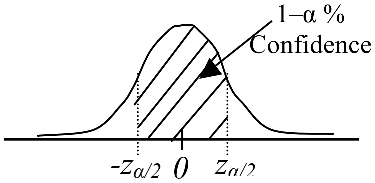
You may wonder why you would want to do this! Here is a real life example of how I used it in my consulting work. A large city wanted to conduct a survey to find out customer satisfaction with their garbage service. They want me to come up with a plan for conducting the survey, so I asked them a few questions pertaining to margin of error and set a confidence level at a fairly hefty size of 90% (survey's seldom achieve that great of level of confidence due to sampling errors). With my calculation I indicated to them that they should poll around a thousand people (mind you this was in an area of around 100,000).

Example: If your client wants a 90% confidence level and a margin of error equal to 0.2, calculate the sample size needed to ascertain that 90% confidence is achieved if a previous poll showed that p-hat was 0.47.

***Note:** When a p-hat is not known, we use p-hat=q-hat=0.5.

§7.3 Estimating Population Mean: Sigma Known

We know that sample means are normally distributed and thus there is a small probability that they fall in the tails of the normal distribution. We will define the probability that a value falls within the tail of the normal distribution as α. Due to the symmetry of the normal distribution there is only $^1/_2$ α probability that a value will fall in either tail. We'll write that as $^{\alpha}/_2$.



The values that are associated with the 1–α % Confidence Region are called **critical values** and are denoted by $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. We are going to practice finding critical values based the z-table. We will also use the table in your book to find the critical values, and I will also show you how to use a t-table to find critical values for the normal distribution.

Example: Find the critical value corresponding to 90% level of confidence.

Step 1: Find α (remember that 1–α% is the level of confidence)

Step 2: Split α into two tails and draw a picture

Step 3: Find the z-score $P(Z < z_{\alpha/2}) = {}^{\alpha}/_2$

***Note**: This is the same as finding normal values given a pre-defined probability. On your calculator this is done with the invnormal($^{\alpha}/_2$*

There is an alternate ways of finding the z-score associated with a 1–α% level of confidence.

Alternate: The t-table A-3 on page 606

Step 1: In the left column find the ∞ (the degrees of freedom where t is approximately a standard normal)

Step 2: Along the top in the One-Tail find $^{\alpha}/_2$ or in the Two-Tail find α

Step 3: Pinpoint the value that corresponds to step 1 & step 2 in the "body" of the t-table.

Note: *The draw back is that there are only values for 50%, 75%, 80-95% by increments of 5, 98%, 99% & 99.9%. Another draw back, depending on the table, is that the values may differ from those in a z-table (although they are typically more accurate than a z-table).*

Now, you might be wondering why we find critical values in the way that we do, and how that relates to the normal distribution, so I will include it in my notes, although I may not take the time in the class to go over the derivation as it is mostly Algebra.

We are discussing the distribution of the sample means and therefore when we are discussing the standard normal we will be talking about a standard normal where

$$\mu = \text{population mean} \quad \text{and} \quad \sigma = {}^{\sigma} / \sqrt{n}$$

$$\therefore\ z = \frac{\text{x-bar} - \mu}{{}^{\sigma} / \sqrt{n}}$$

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ and knowing that $z = \dfrac{\text{x-bar} - \mu}{{}^{\sigma} / \sqrt{n}}$

$P(-z_{\alpha/2} < \dfrac{\text{x-bar} - \mu}{{}^{\sigma} / \sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$ and using a little Algebra to solve a compound inequality – for μ

$P(\text{x-bar} - z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n}) < \mu < \text{x-bar} + z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n})) = 1 - \alpha$

Thus, you see that the mean is between the values of x-bar minus $z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n})$ and x-bar plus $z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n})$ with probability $1–\alpha$. We call the $z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n})$ the **margin of error** and denote it with a "E".

Finding Confidence Intervals when σ is known

1) You must know that the value comes from an approximately normal distribution unless $n \geq 30$
2) Standard deviation of the population must be known
3) Find the critical values
4) Compute the margin of error. $E = z_{\alpha/2} \bullet ({}^{\sigma} / \sqrt{n})$
5) Compute the interval x-bar ± E

Example: If a random sample of size n = 20 from a normal population with variance 225 has mean x-bar of 64.3. Construct a 95% confidence interval for the population mean.

Example: The axial loads of 175 cans have a sample mean of 267.1 lbs and population standard deviation of 22.1 lbs. Find a 99% confidence interval for the population mean.

Sometimes we might need to conduct a study in order to get sample data by which to make inferences about the population. If this is the case, we need to make sure that we have a large enough sample in order to make the desired inferences. In order to decide how large of a sample to compute, we need only to rewrite the margin of error in order to find the desired sample size to make inferences about the population mean.

$E = z_{\alpha/2} \bullet (\sigma / \sqrt{n})$ so this means that $E\sqrt{n} = z_{\alpha/2} \bullet \sigma$ which means that $\sqrt{n} = \frac{z_{\alpha/2} \bullet \sigma}{E}$

and this means that $n = \left[\frac{z_{\alpha/2} \bullet \sigma}{E}\right]^2$

What do we need to decide on sample size

1) Decide on the confidence level
2) Decide on the margin of error
3) Know σ
4) Find n based upon $n = \left[\frac{z_{\alpha/2} \bullet \sigma}{E}\right]^2$

Example: Suppose that the weights of all fox terriers dogs are normally distributed with population standard deviation 2.5 kg. How large a sample should be taken in order to be 95% confident that the sample mean doesn't differ from the population mean by more than 0.5kg?

Example: You want to find an estimate for the population height of men in the US. You want to be 90% confident that the true population mean is within 2 inches of the true population mean. How large of sample must you draw to have this level of confidence and this margin of error?

§7.4 Estimating a Population Means: Sigma Not Known

When the standard deviation of the population is unknown the sampling distribution of the sample means does not follow the normal distribution but another related distribution called Student's t-distribution. This distribution was discovered by a brewer for Guiness Brewing Company, but he was not allowed to publish his findings under his true name and although it was really Gosset's distribution, it has come to be known in the field of science as Student's t-distribution and thus it shall always be known as that!

Facts about Student's t-distribution

1) When s is used to approximate σ, the distribution of x-bar's follows this distribution
2) Student's t-distribution is symmetric about its mean (zero) and approximately bell-shaped
3) Student's t-distribution is wider, flatter with thicker tails than the normal distribution
4) As the number of samples increases the distribution becomes more and more like the normal distribution (approximately 1000 is fairly normal)
5) The degrees of freedom, d.f. is $n - 1$ (in advanced texts this is referred to as ν, the Greek letter nu)
6) Still must be known that the data comes from an approximately normal distribution and if this is unknown then must be shown that the variables are approximately normally distributed (see the last chapter) or $n \geq 30$

First, let's practice using the table and then learn how to find the values on your calculator. I'll need to upload a nice little program to everyone's calculator before we can use our calculators to do that, however.

Finding critical values for a confidence interval when the population standard deviation is unknown is the equivalent of finding critical values for the normal distribution, we just need an extra piece of information – the degrees of freedom $(n - 1)$.

Example: Using the t-table A-3 on p. 606 find the critical values for a sample size of 23 for which we want a 90% confidence interval for the population mean.

Step 1: Find α

Step 2: Compute the degrees of freedom – $n - 1$

Step 3: Look in the top row for area in two-tails to find α

Step 4: Look along the left column to find the degrees of freedom

Step 5: Find the critical value in the body of the table

Example: Using the t-table on p. 606 find the critical values for a sample of size 48 for which we want a 85% confidence interval for the population mean.

Note: *There is no df = 47 in the table. Convention says that when there is no corresponding degrees of freedom that we will use the* ***next lower degrees of freedom available****.*

Now, let's put it all together and find confidence intervals for the population mean based upon Student's t-distribution.

Finding Confidence Intervals when σ is unknown

1) You must know that the value comes from an approximately normal distribution unless $n \geq 30$
2) Standard deviation of the sample must be computed/computable
3) Find the critical values, $t_{n-1,\alpha}$
4) Compute the margin of error. $E = t_{n-1,\alpha} \bullet ({}^{s} / {}_{\sqrt{n}})$
5) Compute the interval x-bar ± E

Example: We have a sample of size 37 with a sample mean of 20 and sample standard deviation of 2. Find a 90% confidence interval for the true population mean.

Example: The data below represents the ages of people when they were 1st diagnosed with cancer. Show that a CI is relevant and give a 95% CI for the true population mean.

Stem (x10)	Leaf(x1)
0	2
1	5 7
2	7 7 9
3	5 5 6 9
4	0 1
5	0

Step 1: Calculate the sample mean & std. dev. & median

Step 2: Investigate normality

a) View stem-and-leaf

b) Normal probability plot

c) Mean vs Median

d) Pearson's Skewness

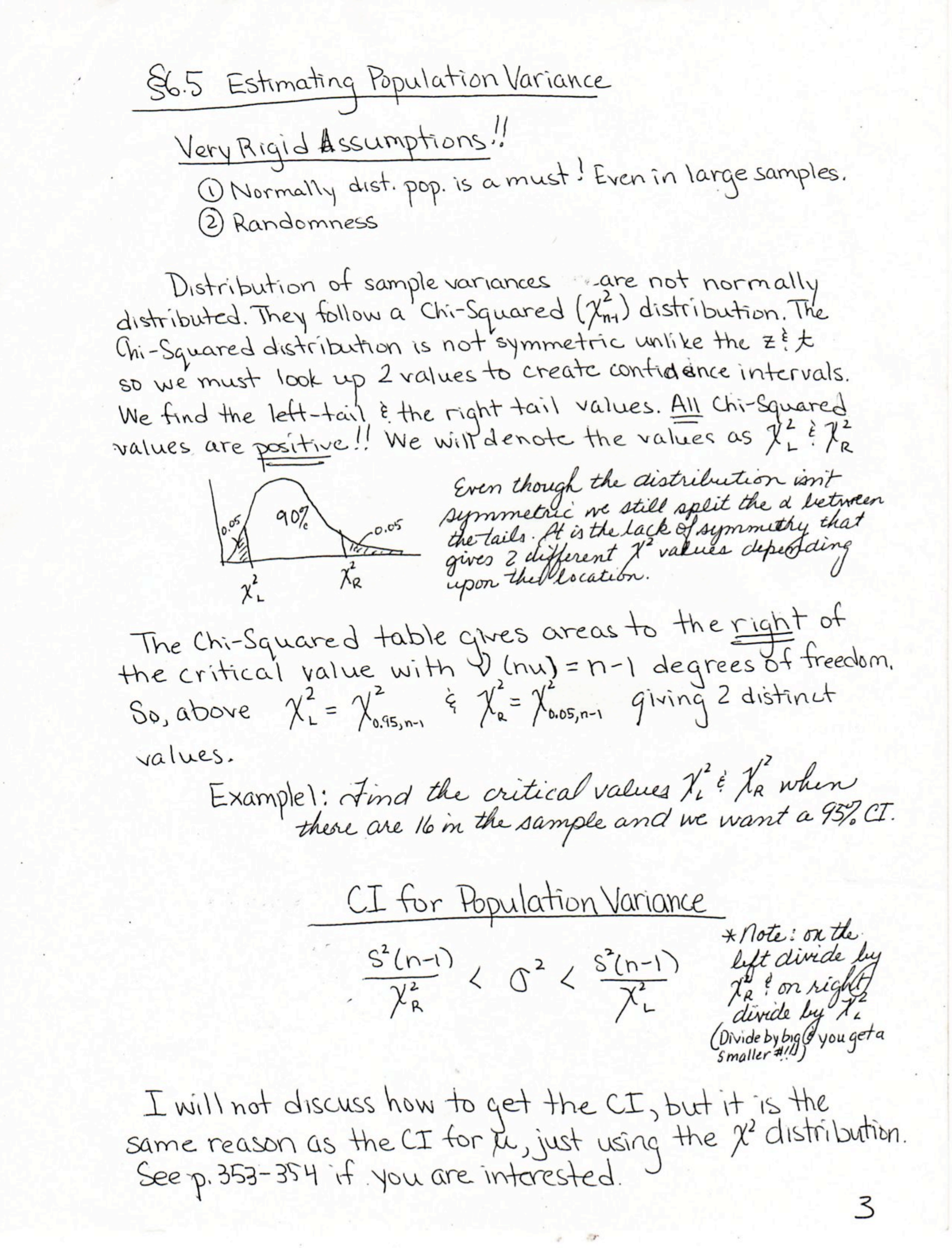
e) Outliers?

Step 3: Find the critical value: $t_{n-1,\alpha}$

Step 4: Calculate the margin of error, E

Step 5: Compute the interval

§7.5 Estimating a Population Variance



§6.5 Estimating Population Variance

Very Rigid Assumptions!!

① Normally dist. pop. is a must! Even in large samples.

② Randomness

Distribution of sample variances are not normally distributed. They follow a Chi-Squared (χ^2_{n-1}) distribution. The Chi-Squared distribution is not symmetric unlike the z & t so we must look up 2 values to create confidence intervals. We find the left-tail & the right tail values. All Chi-Squared values are positive!! We will denote the values as χ^2_L & χ^2_R

Even though the distribution isn't symmetric we still split the α between the tails. It is the lack of symmetry that gives 2 different χ^2 values depending upon the location.

The Chi-Squared table gives areas to the right of the critical value with ν (nu) = n-1 degrees of freedom. So, above $\chi^2_L = \chi^2_{0.95,n-1}$ & $\chi^2_R = \chi^2_{0.05,n-1}$ giving 2 distinct values.

Example1: Find the critical values χ^2_L & χ^2_R when there are 16 in the sample and we want a 95% CI.

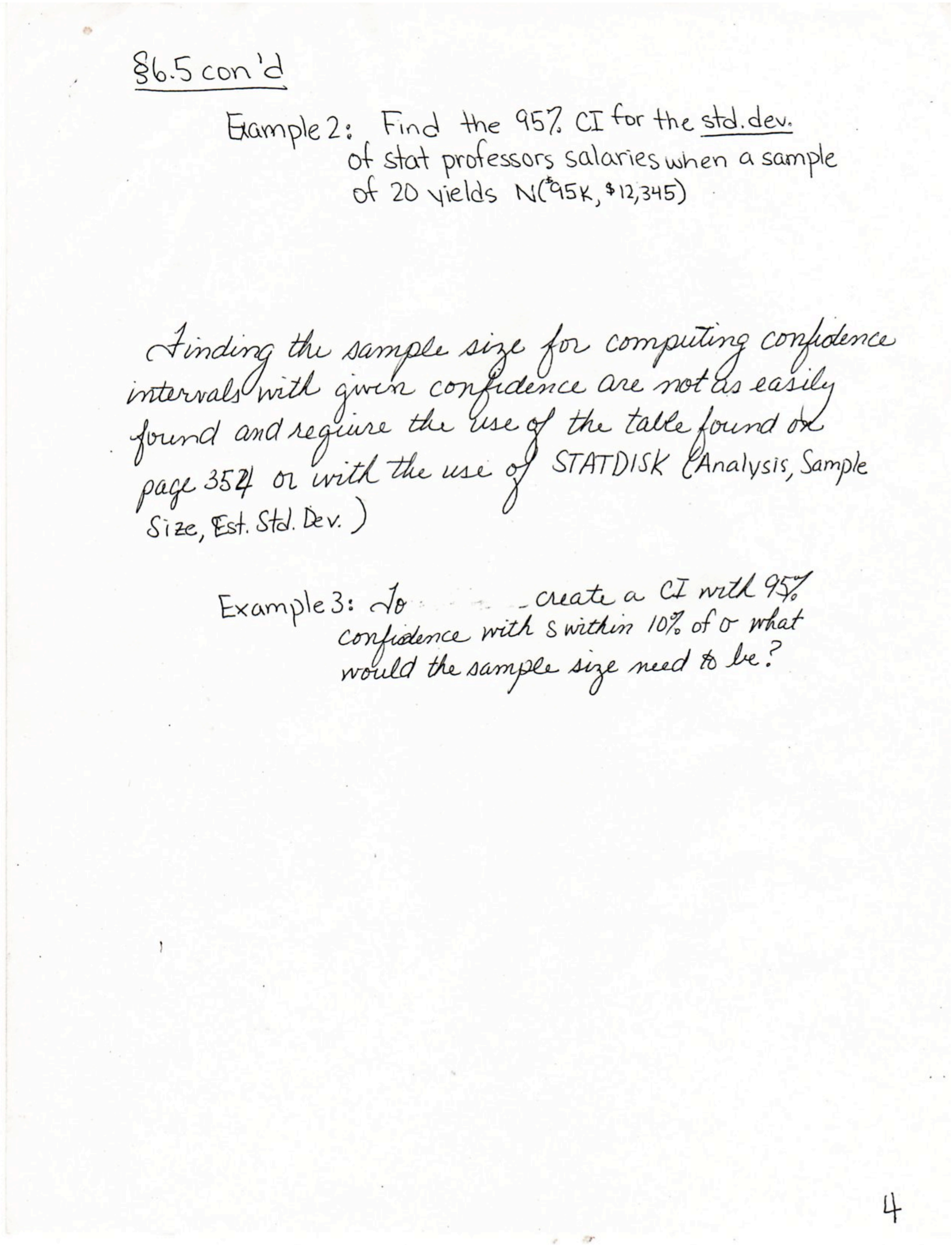
CI for Population Variance

$$\frac{s^2(n-1)}{\chi^2_R} < \sigma^2 < \frac{s^2(n-1)}{\chi^2_L}$$

*Note: on the left divide by χ^2_R & on right divide by χ^2_L (Divide by big & you get a smaller #!!)

I will not discuss how to get the CI, but it is the same reason as the CI for μ, just using the χ^2 distribution. See p. 353-354 if you are interested.

3



§6.5 con'd

Example 2: Find the 95% CI for the std. dev. of stat professors salaries when a sample of 20 yields N($95K, $12,345)

Finding the sample size for computing confidence intervals with given confidence are not as easily found and require the use of the table found on page 354 or with the use of STATDISK (Analysis, Sample Size, Est. Std. Dev.)

Example 3: To create a CI with 95% confidence with s within 10% of σ what would the sample size need to be?

4