6 Estimation in the One-Sample Situation

SW Chapter 6

Standard Errors and the $t-$Distribution

We need to add one more small complication to the sampling distribution of $\overline{Y}$. What we saw last time, and in SW Chapter 5, is that if Y_1, Y_2, ..., Y_n is a random sample from a normal population and that population has mean μ and standard deviation σ, then $\overline{Y}$ looks like it is a single number randomly selected from a normal distribution also with mean $\mu_{\overline{Y}} = \mu$ but with standard deviation $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$. We get from this that $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}} = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} = Z$ is a standard normal random variable, and we can use the table on the inside front cover of SW to compute probabilities involving $\overline{Y}$.

Unfortunately, in the context in which we need to use this result, we would need to know σ in order to apply the result. We are sampling from a population in order to find out something about it, so almost certainly we do not know what σ is. What works well is to estimate the population standard deviation σ with sample standard deviation S calculated from the random sample. Our best guesses of the population mean and standard deviation μ and σ are the corresponding sample values $\overline{Y}$ and S. While μ and σ are constants (we do not know the actual values, but they are constants), $\overline{Y}$ and S depend upon the actual sample randomly selected from the population. If we repeated the experiment and drew a second random sample of n observations, we would get different values for $\overline{Y}$ and S, which is to say $\overline{Y}$ and S are random variables.

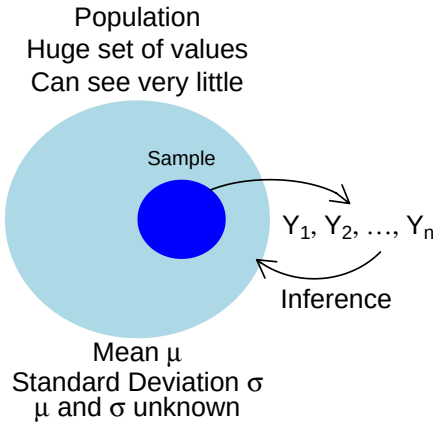
If we are going to estimate σ with S, then of course we would estimate $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$ with $\frac{S}{\sqrt{n}}$. That is exactly what we do, and we give this quantity the name **Standard Error of** $\overline{Y}$, $SE_{\overline{Y}}$. Instead of standardizing $\overline{Y}$ with the expression $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}}$ we use the new expression $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$. Using $SE_{\overline{Y}}$ in the denominator introduces extra variability, though, so this no longer looks like a random number that came from a Z distribution. Provided our assumptions are correct (random sampling from a normally distributed population), then $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ looks like a single random number randomly selected from a Student's $t-$distribution with $n-1$ degrees of freedom (df). The amount of extra variability introduced depends upon the sample size n; if n is very small it is a lot, but by the time n is 30 or so, there is very little difference from a Z, and in fact df $= \infty$ makes the $t-$ and Z distributions the same. SW on p. 187 show how the distribution compares to the normal – it doesn't look like a big difference, but the probability statements we can make are different enough to matter.

Table 4 p. 677 in SW is a standard table of the $t-$distribution. It is organized differently from the Normal Table, since it gives areas under the curve across the top and lets you look up the "critical" values that generate those areas, while the Z table gives you critical values across the side and top and lets you look up areas. We will go through some examples of reading this table during the lecture.

Inference for a Population Mean

Suppose that you have identified a population of interest where individuals are measured on a single quantitative characteristic, say, weight, height or IQ. You select a random or representative sample from the population with the goal of estimating the (unknown) **population mean** value, identified by μ.

This is a standard problem in statistical inference, and the first inferential problem that we will tackle. For notational convenience, identify the measurements on the sample as Y_1, Y_2, ..., Y_n, where n is the sample size. Given the data, our best guess, or estimate, of μ is the sample mean:



$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1+Y_2+\cdots+Y_n}{n}$.

There are two main methods that are used for inferences on μ: **confidence intervals** (CI) and **hypothesis tests**. The standard CI and test procedures are based on the sample mean and the sample standard deviation, denoted by s. We will consider CIs in this lecture, and hypothesis tests in the next lecture.

Let's apply the results of the preceding section, and then lay out the mechanics of the procedure. The main idea behind a CI is this: $\overline{Y}$ should be a pretty good guess as to what μ is, but while μ is a constant (we don't know the value, though), $\overline{Y}$ is a random variable (every possible sample gives a different value), so most assuredly $\overline{Y} \neq \mu$. Still, $\overline{Y}$ should not be too far from μ, but how far away from μ do we think $\overline{Y}$ could be? As a specific example, suppose we randomly sample $n = 9$ values from a normal population and get $\overline{Y} = 22$ and $S = 6$. What could μ be?

To answer such a question, apply the $t-$distribution. $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ looks like a single random number sampled from a $t-$distribution with 8 df, so it *should have* come out somewhere in the middle of that distribution. The middle 95% of that distribution is between -2.306 and 2.306 (from the table). So, we *had* a 95% chance that $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ would fall in that range. Substituting the actual values of $\overline{Y}$ and S we obtained, we are 95% confident that $\frac{22-\mu}{6/\sqrt{9}} = \frac{22-\mu}{2}$ is between -2.306 and 2.306, or equivalently we are 95% confident that $22-\mu$ is between -4.612 and 4.612. This says that μ should be within 4.612 of 22, or in the range $22-4.612$ to $22+4.612$, i.e. between 17.388 and 26.612. We still do not know what μ is, but to have gotten data like this μ must be somewhere between 17.388 and 26.612.

The interval $17.388 \leq \mu \leq 26.612$ is referred to as a 95% confidence interval for μ. It is improper to say there is a 95% chance that μ is in that range: If it is in that range, say 25, there is a 100% chance it is in that range, while if it is not in that range, say 30, there is a 0% chance it is in that range. The 95% refers to how often using this technique works (like a lifetime batting average) - this interval either worked in capturing μ or it did not work, and we cannot know which is true.

Mechanics of a CI for μ

A CI for μ is a range of plausible values for the unknown population mean μ, based on the observed data. To compute a CI for μ:

1. Specify the **confidence coefficient**, which is a number between 0 and 100%, in the form $100(1-\alpha)\%$. Solve for α.

2. Compute the $t-$critical value: $t_{crit} = t_{.5\alpha}$ such that the area under the $t-$ curve ($df = n-1$) to the right of t_{crit} is $.5\alpha$.

3. The desired CI has lower and upper endpoints given by $L = \bar{Y} - t_{crit}SE_{\overline{Y}}$ and $U = \bar{Y} + t_{crit}SE_{\overline{Y}}$, respectively, where $SE_{\overline{Y}} = s/\sqrt{n}$ is the standard error of the sample mean. The CI is often written in the form $\bar{Y} \pm t_{crit}SE_{\overline{Y}}$.

In practice, the confidence coefficient is large, say 95% or 99%, which correspond to $\alpha = .05$ and .01, respectively. The value of α expressed as a percent is known as the **error rate** of the CI.

The CI is determined once the confidence coefficient is specified and the data are collected. Prior to collecting the data, the interval is unknown and is viewed as random because it will depend on the actual sample selected. Different samples give different CIs. The "confidence" in, say, the 95% CI (which has a 5% error rate) can be interpreted as follows. If you repeatedly sample the population and construct 95% CIs for μ, then 95% of the intervals will contain μ, whereas 5% will not. The interval you construct from your data will either cover μ, or it will not.

The length of the CI

$$U - L = 2t_{crit}SE_{\overline{Y}}$$

depends on the accuracy of our estimate $\overline{Y}$ of μ, as measured by $SE_{\overline{Y}} = s/\sqrt{n}$ the standard error of $\overline{Y}$. Less precise estimates of μ lead to wider intervals for a given level of confidence.

Assumptions for Procedures

I described the classical CI. The procedure is based on the assumptions that the data are a random sample from the population of interest, and that the population frequency curve is normal. The population frequency curve can be viewed as a "smoothed histogram" created from the population data.

The normality assumption can be checked using a stem-and-leaf display, a boxplot, or a normal scores plot of the sample data (probably the more the better).

Example

Let us go through a hand-calculation of a CI, using Minitab to generate summary data. I will then show you how the CI is generated automatically in Minitab. The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows: 54 42 51 54 49 56 33 58 54 64 49.

```
Data Display

 AgeTran
    54   42   51   54   49   56   33   58   54   64   49

 Stem-and-Leaf Display: AgeTran
```

```
Stem-and-leaf of AgeTran  N  = 11
Leaf Unit = 1.0

 1   3  3
 1   3
 2   4  2
 4   4  99
(4)  5  1444
 3   5  68
 1   6  4
```

```
Descriptive Statistics: AgeTran

Variable   N  N*   Mean  SE Mean  StDev  Minimum    Q1  Median      Q3 Maximum
AgeTran   11   0  51.27     2.49   8.26    33.00 49.00   54.00  56.00   64.00
```

The summaries for the data are: $n = 11$, $\overline{Y} = 51.27$, and $s = 8.26$ so that $SE_{\overline{Y}} = 8.26/\sqrt{11} = 2.4904$. The degrees of freedom are $df = 11 - 1 = 10$.

A necessary first step in every problem is to define the population parameter in question. Here, let

$$\mu = \text{mean age at time of first transplant for population of patients.}$$

Let us calculate a 95% CI for μ. The degrees of freedom are $df = 11 - 1 = 10$. For a 95% CI $\alpha = .05$, so we need to find $t_{crit} = t_{.025} = 2.228$.

Now $t_{crit} SE_{\overline{Y}} = 2.228 * 2.4904 = 5.55$. The lower limit on the CI is $L = 51.27 - 5.55 = 45.72$. The upper endpoint is $U = 51.27 + 5.55 = 56.82$.

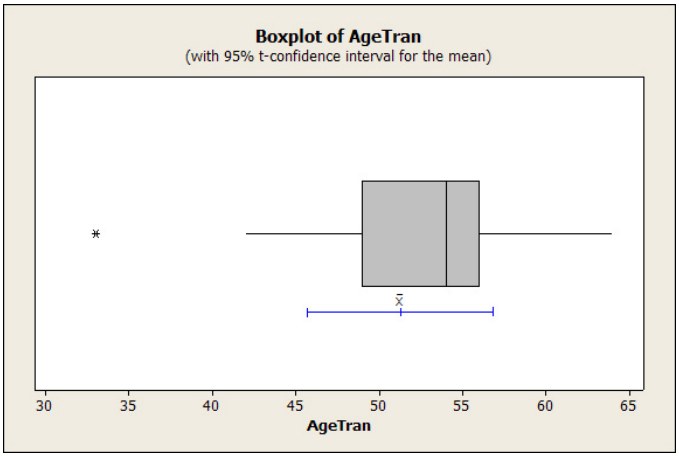
I insist that the results of every CI be summarized in words. For example, I am 95% confident that the population mean age at first transplant is between 45.7 and 56.8 years (rounding off to 1 decimal place).

Minitab does all this very easily. Follow the menu path `Stat > Basic Statistics > 1-Sample t` (be careful that you don't select the `1-Sample Z` — it will treat S as if it is actually σ and give you incorrect bounds). Under `Options...` select Confidence Level of 95 (the default) and Alternative: not equal (we will understand that next week). Under `Graphs` check Boxplot. Do not check `Summarized data` or `Perform hypothesis test`. You get the following results:

```
One-Sample T: AgeTran

Variable   N     Mean   StDev  SE Mean         95% CI
AgeTran   11  51.2727  8.2594   2.4903  (45.7240, 56.8215)
```

We might be a little concerned about the outlier and the possible skewness indicated in the boxplot below, since that could be evidence we did not sample from a normal distribution. It will be worth trying one of the nonparametric procedures we will learn about later, since the assumption of normality is not made there.



The Effect of α on a Two-Sided CI

A two-sided $100(1-\alpha)\%$ CI for μ is given by $\overline{Y} \pm t_{crit} s/\sqrt{n}$. The CI is centered at $\overline{Y}$ and has length $2t_{crit}s/\sqrt{n}$. The confidence coefficient $100(1-\alpha)\%$ is *increased* by *decreasing* α, which increases t_{crit}. That is, increasing the confidence coefficient makes the CI wider. This is sensible: to increase your confidence that the interval captures μ you must pinpoint μ with less precision by making the CI wider. For example, a 95% CI is wider than a 90% CI.

SW Example 6.9 page 192: Let us compute a 90% and a 95% CI by hand.

Note: For large n the Central Limit Theorem gives us the ability to treat $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}}$ as a Z random variable even without sampling from a normal distribution. Some texts would suggest using the `1-Sample Z` procedure in this case (although that still begs the issue of not knowing σ). In practice what we do about large n is to worry a little less about lack of normality in the population we sampled from (outliers and extreme skewness are still problems, just slightly different ones), but continue to use the t-procedures. Remember for large n we get large df, and for large df there is little difference between Z and t.

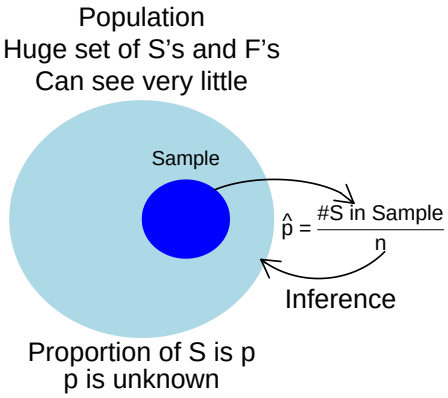
Inference for a Population Proportion

Assume that you are interested in estimating the proportion p of individuals in a population with a certain characteristic or attribute based on a random or representative sample of size n from the population. The **sample proportion** $\hat{p}$ =(# with attribute in the sample)/n is the best guess for p based on the data.

This is the simplest **categorical data** problem. Each response falls into one of two exclusive and exhaustive categories, called success and failure. Individuals with the attribute of interest are in the success category. The rest fall into the failure category. Knowledge of the **population proportion** p of successes characterizes the distribution across both categories because the population proportion of failures is $1-p$.

As an aside, note that the probability that a randomly selected individual has the attribute of interest is the population proportion p with the attribute, so the terms population proportion and probability can be used interchangeably with random sampling.

The diagram of this is very similar to the earlier one. Note that a random sample of size n now becomes just a set of S's and F's.



A CI for p

The derivation of the CI follows the same basic ideas as before, except we do not have the idea of df since we are considering n as large ($np \geq 5$ and $n(1-p) \geq 5$). $\hat{p}$ is a random variable (it almost surely is not p), and it looks like a single number randomly selected from a normal distribution with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, so $\frac{\hat{p}-p}{\sigma_{\hat{p}}}$ looks like a Z. We have the same problem as before – to use this as we wish, we need to compute the denominator, but we need to know p to compute it. We estimate it instead, and call the estimated standard deviation of $\hat{p}$ the standard error of $\hat{p}$, $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Everything proceeds as before.

A two-sided CI for p is a range of plausible values for the unknown population proportion p, based on the observed data. To compute a two-sided CI for p:

1. Specify the confidence level as the percent $100(1-\alpha)\%$ and solve for the error rate α of the CI.

2. Compute $z_{crit} = z_{.5\alpha}$ (i.e. area under the standard normal curve to the right of z_{crit} is $.5\alpha$.)

3. The $100(1-\alpha)\%$ CI for p has endpoints $L = \hat{p} - z_{crit}SE$ and $U = \hat{p} + z_{crit}SE$, respectively, where the "CI standard error" is
$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$
The CI is often written as $\hat{p} \pm z_{crit}SE$.

The CI is determined once the confidence level is specified and the data are collected. Prior to collecting data, the CI is unknown and can be viewed as random because it will depend on the

actual sample selected. Different samples give different CIs. The "confidence" in, say, the 95% CI (which has a .05 or 5% error rate) can be interpreted as follows. If you repeatedly sample the population and construct 95% CIs for p, then 95% of the intervals will contain p, whereas 5% (the error rate) will not. The CI you get from your data either covers p, or it does not.

The length of the CI

$$U - L = 2z_{crit}SE$$

depends on the accuracy of the estimate $\hat{p}$, as measured by the standard error SE. For a given $\hat{p}$, this standard error decreases as the sample size n increases, yielding a narrower CI. For a fixed sample size, this standard error is maximized at $\hat{p} = .5$, and decreases as $\hat{p}$ moves towards either 0 or 1. In essence, sample proportions near 0 or 1 give narrower CIs for p. However, the normal approximation used in the CI construction is less reliable for extreme values of $\hat{p}$.

Example:

The 1983 Tylenol poisoning episode highlighted the desirability of using tamper-resistant packaging. The article "Tamper Resistant Packaging: Is it Really?" (Packaging Engineering, June 1983) reported the results of a survey on consumer attitudes towards tamper-resistant packaging. A sample of 270 consumers was asked the question: "Would you be willing to pay extra for tamper resistant packaging?" The number of yes respondents was 189. Construct a 95% CI for the proportion p of all consumers who were willing in 1983 to pay extra for such packaging.

Here $n = 270$ and $\hat{p} = 189/270 = .700$. The critical value for a 95% CI for p is $z_{.025} = 1.96$. The CI standard error is given by

$$SE = \sqrt{\frac{.7 * .3}{270}} = .028,$$

so $z_{crit}SE = 1.96 * .028 = .055$. The 95% CI for p is $.700 \pm .055$. You are 95% confident that the proportion of consumers willing to pay extra for better packaging is between .645 and .755. (How much extra?).

Appropriateness of the CI

The standard CI is based on a large sample standard normal approximation to

$$z = \frac{\hat{p} - p}{SE}.$$

A simple rule of thumb requires $np \geq 5$ and $n(1-p) \geq 5$ for the method to be suitable. The population proportion p is unknown so you should use $\hat{p}$ in these formulae to check the suitability of the CI. Given that $n\hat{p}$ and $n(1-\hat{p})$ are the observed numbers of successes and failures, you should have at least 5 of each to apply the large sample CI.

In the packaging example, $n\hat{p} = 270*(.700) = 189$ (the number who support the new packaging) and $n(1-\hat{p}) = 270*(.300) = 81$ (the number who oppose) both exceed 5. The normal approximation is appropriate here.

More Accurate Confidence Intervals

Large sample CIs for p should be interpreted with caution in small sized samples because the true error rate usually exceeds the assumed (nominal) value. For example, an assumed 95% CI, with a nominal error rate of 5%, may be only an 80% CI, with a 20% error rate. The large sample CIs are usually overly optimistic (i.e. too narrow) when the sample size is too small to use the normal approximation.

SW use the following method, originally suggested by Alan Agresti, for a 95% CI. The standard method computes the sample proportion as $\hat{p} = y/n$ where y is the number of individuals in the sample with the characteristic of interest, and n is the sample size. Agresti suggested estimating the proportion with $\tilde{p} = (y+2)/(n+4)$, with a standard error of

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

and using the "usual interval" with these new summaries: $\tilde{p} \pm 1.96SE$. This appears odd, but just amounts to adding two successes and two failures to the observed data, and then computing the standard CI.

This adjustment has little effect when n is large and $\hat{p}$ is not close to either 0 or 1, as in the Tylenol example. Let us do examples using SW's proposed CI.

SW Examples 6.16 and 6.17, page 208-9

Minitab Implementation

A CI for p can be obtained in Minitab from summary data from the menu path `Stat > Basic Statistics > 1 Proportion`, check `Summarized data`, enter `Number of trials` (n) and `Number of events` (# Successes), click `Options`, enter `Confidence level` in percent (95.0 usually), ignore `Test proportion` for now, select `Alternative: not equal`, and check `Use test and interval based on normal distribution`.

The above choices produce a CI based upon $\hat{p}$. In order to use SW's CI based on $\tilde{p}$, add 4 to n and 2 to # Successes. Finally, to get the best interval (arguably *the* correct one), **do not** check `Use test and interval based on normal distribution`. This third choice produces what is known as an exact interval – it is a lot harder to explain how we get it (I'll indicate where it comes from next week), but the confidence level and error rate are correct and not subject to approximation like the other two intervals. Minitab is a little unique in providing this. Let us examine Minitab results from two examples:

The Tylenol Example:

Using $\hat{p}$:

```
Sample    X    N  Sample p         95% CI         Z-Value  P-Value
1       189  270  0.700000  (0.645339, 0.754661)     6.57    0.000
```

Using $\tilde{p}$:

```
Sample    X    N  Sample p         95% CI         Z-Value  P-Value
1       191  274  0.697080  (0.642670, 0.751490)     6.52    0.000
```

Using exact interval:

```
                                                    Exact
Sample    X    N  Sample p         95% CI         P-Value
1       189  270  0.700000  (0.641500, 0.754047)    0.000
```

Ignore the Z-Value and P-Value entries for now. You can see that the intervals all agree for any practical interpretation.

Example 6.17 p. 209 of SW

Using $\hat{p}$:

```
 Sample  X   N  Sample p    CI    Z-Value  P-Value
 1       0  11  0.000000  (*, *)    -3.32    0.001

 * NOTE * The normal approximation may be inaccurate for small samples.
```

Using $\tilde{p}$:

```
Sample  X   N  Sample p         95% CI         Z-Value  P-Value
1       2  15  0.133333  (0.000000, 0.305361)    -2.84    0.005

* NOTE * The normal approximation may be inaccurate for small samples.
```

Using exact interval:

```
                                                  Exact
Sample  X   N  Sample p         95% CI         P-Value
1       0  11  0.000000  (0.000000, 0.238404)    0.001
```

The only one of these I would trust is the exact one. The one based on $\tilde{p}$ is surprisingly informative, though. Minitab's warning on the other two should not be ignored.