# 3 Estimation in One Sample Problems

### SW Chapter 6

I will talk in class about standard errors and the t- distribution prior to these notes.

### 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  $\mu$ . You cannot see much of the population, but you would like to know what is typical in the population ( $\mu$ ). The only information you can see is that in the sample.

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  $\mu$  is the sample mean:



 $\bar{Y} = \underbrace{\sum_{i} Y_i}_{n} = \underbrace{Y_1 + Y_2 + \dots + Y_n}_{n}.$ 

There are two main methods that are used for inferences on  $\mu$ : 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.

# CI for $\mu$

A CI for  $\mu$  is a range of plausible values for the unknown population mean  $\mu$ , based on the observed data, of the form Best Guess  $\pm$  Reasonable Error of the Guess. To compute a CI for  $\mu$ :

- 1. Specify the **confidence coefficient**, which is a number between 0 and 100%, in the form  $100(1-\alpha)\%$ . Solve for  $\alpha$ .
- 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 $\alpha$ . See SW page 677 for a t-table.

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  $\alpha$  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  $\mu$ , then 95% of the intervals will contain  $\mu$ , whereas 5% will not. The interval you construct from your data will either cover  $\mu$ , or it will not.

The length of the CI

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

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

After going through a few calculations, I will try to justify where this procedure comes from, and why it "works".

### 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 never be completely verified without having the entire population data. You can assess the reasonableness of this assumption using a stem-and-leaf display or a boxplot of the sample data. The stem-and-leaf display from the data should resemble a normal curve.

### Example: Age at First Heart Transplant

Let us go through a hand-calculation of a CI, using Minitab to generate summary data. I show you later how to generate the CI 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. Summaries for the data are: n = 11,  $\overline{Y} = 51.27$ , and s = 8.26 so that  $SE_{\overline{Y}} = 8.26/\sqrt{11} = 2.4904$ .

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

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

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

### The effect of $\alpha$ on a two-sided CI

A two-sided  $100(1-\alpha)\%$  CI for  $\mu$  is given by  $\overline{Y} \pm t_{crit}SE_{\overline{Y}}$ . The CI is centered at  $\overline{Y}$  and has length  $2t_{crit}SE_{\overline{Y}}$ . The confidence coefficient  $100(1-\alpha)\%$  is **increased** by **decreasing**  $\alpha$ , 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  $\mu$  you must pinpoint  $\mu$  with less precision by making the CI wider. For example, a 95% CI is wider than a 90% CI.

SW Example 6.9 page 196: Let us compute by hand a 90% and a 95% CI

# Hypothesis Testing for $\mu$

Suppose you are interested in checking whether the population mean  $\mu$  is equal to some prespecified value, say  $\mu_0$ . This question can be formulated as a two-sided hypothesis test, where you are trying to decide which of two contradictory claims or hypotheses about  $\mu$  is more reasonable given the observed data. The **null hypothesis**, or the hypothesis under test, is  $H_0: \mu = \mu_0$ , whereas the **alternative hypothesis** is  $H_A: \mu \neq \mu_0$ .

I will explore the ideas behind hypothesis testing later. At this point, I focus on the mechanics behind the test. The steps in carrying out the test are:

- 1. Set up the null and alternative hypotheses:  $H_0: \mu = \mu_0$  and  $H_A: \mu \neq \mu_0$ , where  $\mu_0$  is specified by the context of the problem.
- 2. Choose the size or significance level of the test, denoted by  $\alpha$ . In practice,  $\alpha$  is set to a small value, say, .01 or .05, but theoretically can be any value between 0 and 1.
- 3. Compute the critical value  $t_{crit} = t_{.5\alpha}$  from the *t*-distribution table with degrees of freedom df = n 1.
- 4. Compute the **test statistic**

$$t_s = \frac{\bar{Y} - \mu_0}{SE_{\overline{Y}}},$$

where  $SE_{\overline{Y}} = s/\sqrt{n}$  is the standard error.

5. Reject  $H_0$  in favor of  $H_A$  (i.e. decide that  $H_0$  is false, based on the data) if  $|t_s| > t_{crit}$ . Otherwise, do not reject  $H_0$ . An equivalent rule is to Reject  $H_0$  if  $t_s < -t_{crit}$  or if  $t_s > t_{crit}$ . I sometimes call the test statistic  $t_{obs}$  to emphasize that the computed value depends on the observed data.

The process is represented graphically below. The area under the t-probability curve outside  $\pm t_{crit}$  is the size of the test,  $\alpha$ . One-half  $\alpha$  is the area in each tail. You reject  $H_0$  in favor of  $H_A$  only if the test statistic is outside  $\pm t_{crit}$ .



### Assumptions for procedures

I described the classical t-test, which assumes that the data are a random sample from the population and that the population frequency curve is normal. These are the same assumptions as for the CI.

#### Example: Age at First Transplant (Revisited)

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. Summaries for these data are:  $n = 11, \overline{Y} = 51.27, s = 8.26$  and  $SE_{\overline{Y}} = 2.4904$ . Test the hypothesis that the mean age at first transplant is 50. Use  $\alpha = .05$ .

As in the earlier analysis, define

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

We are interested in testing  $H_0: \mu = 50$  against  $H_A: \mu \neq 50$ , so  $\mu_0 = 50$ .

The degrees of freedom are df = 11 - 1 = 10. The critical value for a 5% test is  $t_{crit} = t_{.025} = 2.228$ . (Note  $.5\alpha = .5 * .05 = .025$ ). The same critical value was used with the 95% CI.

For the test,

$$t_s = \frac{\bar{Y} - \mu_0}{SE_{\overline{Y}}} = \frac{51.27 - 50}{2.4904} = 0.51$$

Since  $t_{crit} = 2.228$ , we do not reject  $H_0$  using a 5% test. Notice the placement of  $t_s$  relative to  $t_{crit}$  in the picture below. The results of the hypothesis test should not be surprising, since the CI tells you that 50 is a plausible value for the population mean age at transplant. Note: All you can say is that the data *could have* come from a distribution with a mean of 50 – this is not convincing evidence that  $\mu$  actually is 50.



# **P-values**

The **p-value**, or **observed significance level** for the test, provides a measure of plausibility for  $H_0$ . Smaller values of the p-value imply that  $H_0$  is less plausible. To compute the p-value for a two-sided test, you

- 1. Compute the test statistic  $t_s$  as above.
- 2. Evaluate the area under the t-probability curve (with df = n 1) outside  $\pm |t_s|$ .



The p-value is the total shaded area, or twice the area in either tail. You can only get bounds on the p-value using SW's t-table. A useful interpretation of the p-value is that it is the chance of obtaining data that favor  $H_A$  by this much or more *if*  $H_0$  actually is true. If the p-value is small then the sample we obtained is pretty unusual to have obtained if  $H_0$  is true – but we actually got the sample, so probably it is not very unusual, so we would conclude  $H_0$  is false (it would not be unusual if  $H_A$  is true).

Most, if not all, statistical packages, including Minitab, summarize hypothesis tests with a pvalue, rather than a decision (i.e reject or not reject at a given  $\alpha$  level). You can make a decision to reject or not reject  $H_0$  for a size  $\alpha$  test based on the p-value as follows - reject  $H_0$  if the p-value is less than or equal to  $\alpha$ . This decision is identical to that obtained following the formal rejection procedure given earlier. The reason for this is that the p-value can be interpreted as the smallest value you can set the size of the test and still reject  $H_0$  given the observed data.

There are a lot of terms to keep straight here.  $\alpha$  and  $t_{crit}$  are constants we choose (actually, one determines the other so we really only choose one, usually  $\alpha$ ) to set how rigorous evidence against  $H_0$  needs to be.  $t_s$  and the p-value (again, one determines the other) are random variables because they are calculated from the random sample. They are the evidence against  $H_0$ .

### Example: Age at First Transplant

The picture below is used to calculate the p-value. Using SW's table, all we know is that the p-value is greater than .40. (Why?) The exact p-value for the test (generated with Minitab) is 0.62. For a 5% test, the p-value indicates that you would not reject  $H_0$  (because .62 > .05).



Minitab output for the heart transplant problem is given below. Let us look at the output and find all of the summaries we computed. Also, look at the graphical summaries to assess whether the t-test and CI are reasonable here.

# **COMMENTS:**

1. The data were entered into the worksheet as a single column (C1) that was labelled **agetran**.

- 2. To display the data follow the sequence Data > Display Data, and fill in the dialog box.
- 3. To get the stem and leaf display, follow the sequence Graph > Stem and Leaf ..., then fill in the dialog box.
- 4. To get a one-sample t-test and CI follow the sequence: STAT > BASIC STATISTICS > 1-sample t... . In the dialog box, select the column to analyze (C1). For the test, you need to check the box for Perform Hypothesis Test and specify the null mean (i.e.  $\mu_0$ ) and the type of test (by clicking on OPTIONS): not equal gives a two-sided test (default), less than gives a lower one-sided test, and greater than gives an upper one-sided test. The results of the test are reported as a p-value. We have only discussed two-sided tests up to now. Click on the Graphs button and select Boxplot of data.
- 5. I would also follow Stat > Basic Statistics > Display Descriptive Statistics to get a few more summary statistics. The default from the test is a bit limited.
- 6. If you ask for a test, you will get a corresponding CI. The CI level is set by clicking on Option in the dialog box. If you want a CI but not a test, do not check Perform Hypothesis Test in the main dialog box. A 95% CI is the default.
- 7. The boxplot will include a CI for the mean.
- 8. The plots generated with Stat > Basic Statistics > Graphical Summary include a CI for the population mean.

```
Data Display
```

agetran 42 49 49 51 54 54 54 56 58 64 33 Stem-and-Leaf Display: agetran Stem-and-leaf of agetran Ν = 11 Leaf Unit = 1.0 334455 3 1 2 4 (4) 3 1 2 99 1444 68 4 One-Sample T: agetran Test of mu = 50 vs not = 5095% CI T P (45.7240, 56.8215) 0.51 0.620 StDev SE Mean Variable Ν Mean agetran 11 51.2727 8.2594 2.4903 Descriptive Statistics: agetran Ν N\* SE Mean StDev Maximum Variable Mean Minimum Q1 Median Q3 33.00 49.00 54.00 56.00 51.27 2.49 64.00 agetran 0 8.26 11



# **Example:** Meteorites

One theory of the formation of the solar system states that all solar system meteorites have the same evolutionary history and thus have the same cooling rates. By a delicate analysis based on measurements of phosphide crystal widths and phosphide-nickel content, the cooling rates, in degrees Celsius per million years, were determined for samples taken from meteorites named in the accompanying table after the places they were found.

Suppose that a hypothesis of solar evolution predicted a mean cooling rate of  $\mu = .54$  degrees per million year for the Tocopilla meteorite. Do the observed cooling rates support this hypothesis? Test at the 5% level. The boxplot and stem and leaf display (given below) show good symmetry. The assumption of a normal distribution of observations basic to the *t*-test appears to be realistic.

Meteorite	Cooling rates
Walker County	$0.69\ 0.23\ 0.10\ 0.03\ 0.56\ 0.10\ 0.01\ 0.02\ 0.04\ 0.22$
Uwet	$0.21 \ 0.25 \ 0.16 \ 0.23 \ 0.47 \ 1.20 \ 0.29 \ 1.10 \ 0.16$
Tocopilla	$5.60\ 2.70\ 6.20\ 2.90\ 1.50\ 4.00\ 4.30\ 3.00\ 3.60\ 2.40\ 6.70\ 3.80$

Let

 $\mu$  = mean cooling rate over all pieces of the Tocopilla meteorite.

To answer the question of interest, we consider the test of  $H_0: \mu = .54$  against  $H_A: \mu \neq .54$ . I will explain later why these are the natural hypotheses here. Let us go carry out the test, compute the p-value, and calculate a 95% CI for  $\mu$ . The sample summaries are n = 12,  $\overline{Y} = 3.892$ , s = 1.583. The standard error is  $SE_{\overline{Y}} = s/\sqrt{n} = 0.457$ .

Minitab output for this problem is given below. For a 5% test (i.e.  $\alpha = .05$ ), you would reject  $H_0$  in favor of  $H_A$  because the  $p - value \leq .05$ . The data strongly suggest that  $\mu \neq .54$ . The 95% CI says that you are 95% confident that the population mean cooling rate for the Tocopilla meteorite is between 2.89 and 4.90 degrees per million years. Note that the CI gives us a means to assess how different  $\mu$  is from the hypothesized value of .54.

### **COMMENTS:**

- 1. The data were entered as a single column in the worksheet, and labelled **Toco**.
- 2. Remember that you need to specify the null value for the mean (i.e. .54) in the 1-sample t dialog box!
- 3. I generated a boxplot within the 1-sample t dialog box. A 95% CI for the mean cooling rate is superimposed on the plots.

```
Data Display
Toco
5.6
            2.7
                    6.2
                            2.9
                                    1.5
                                            4.0
                                                    4.3
                                                            3.0
                                                                    3.6
                                                                            2.4
                                                                                    6.7
                                                                                            3.8
Stem-and-Leaf Display: Toco
Stem-and-leaf of Toco
Leaf Unit = 0.10
                              Ν
                                  = 12
          5
4
79
      12233445566
1245(2)
(2533321
          0
          68
03
          6
2
7
One-Sample T: Toco
Test of mu = 0.54 vs not = 0.54
                                          SE Mean
0.45684
                              StDev
1.58255
                                                      95% CI
(2.88616, 4.89717)
Variable
              Ν
                      Mean
                                                                                T P
7.34 0.000
             12
                  3.89167
Toco
Descriptive Statistics: Toco
             N
12
                                                         nimum Q1
1.500 2.750
                                                                                     Q3
5.275
Variable
                                 SE Mean
                                             StDev
                  N*
                         Mean
                                                      Minimum
                                                                           Median
                                                                                              Maximum
                    0
                                    0.457
                                                                            3.700
Тосо
                        3.892
                                             1.583
                                                                                                 6.700
```



# The mechanics of setting up hypothesis tests

# SW Section 7.10

When setting up a test you should imagine you are the researcher conducting the experiment. In many studies, the researcher wishes to establish that there has been a change from the **status quo**, or that they have developed a method that produces a **change** (possibly in a specified direction) in the typical response. The researcher sets  $H_0$  to be the **status quo** and  $H_A$  to be the **researcher hypothesis** - the claim the researcher wishes to make. In some studies you define the hypotheses so that  $H_A$  is the **take action** hypothesis - rejecting  $H_0$  in favor of  $H_A$  leads one to take a radical action.

Some perspective on testing is gained by understanding the mechanics behind the tests. A hypothesis test is a decision process in the face of uncertainty. You are given data and asked which of two contradictory claims about a population parameter, say  $\mu$ , is more reasonable. Two decisions are possible, but whether you make the correct decision depends on the true state of nature which is unknown to you.

Decision	If $H_0$ true	If $H_A$ true
Reject $H_0$ in favor of $H_A$	Type I error	correct decision
Do not Reject [accept] $H_0$	correct decision	Type II error

For a given problem, only one of these errors is possible. For example, if  $H_0$  is true you can make a Type I error but not a Type II error. Any reasonable decision rule based on the data that tells us when to reject  $H_0$  and when to not reject  $H_0$  will have a certain probability of making a Type I error if  $H_0$  is true, and a corresponding probability of making a Type II error if  $H_0$  is false and  $H_A$  is true. For a given decision rule, define  $\alpha = \operatorname{Prob}(\operatorname{Reject} H_0 \text{ given } H_0 \text{ is true}) = \operatorname{Prob}(\operatorname{Type I error})$ 

and

 $\beta = \operatorname{Prob}(\text{ Do not reject } H_0 \text{ when } H_A \text{ true }) = \operatorname{Prob}(\text{ Type II error }).$ 

The mathematics behind hypothesis tests allows you to prespecify or control  $\alpha$ . For a given  $\alpha$ , the tests we use (typically) have the smallest possible value of  $\beta$ . Given the researcher can control  $\alpha$ , you set up the hypotheses so that committing a Type I error is more serious than committing a Type II error. The magnitude of  $\alpha$ , also called the **size** or **level** of the test, should depend on the seriousness of a Type I error in the given problem. The more serious the consequences of a Type I error, the smaller  $\alpha$  should be. In practice  $\alpha$  is often set to .10, .05, or .01, with  $\alpha = .05$  being the scientific standard. By setting  $\alpha$  to be a small value, you reject  $H_0$  in favor of  $H_A$  only if the data **convincingly indicate** that  $H_0$  is false.

Let us piece together these ideas for the meteorite problem. Evolutionary history predicts  $\mu = .54$ . A scientist examining the validity of the theory is trying to decide whether  $\mu = .54$  or  $\mu \neq .54$ . Good scientific practice dictates that rejecting another's claim when it is true is more serious than not being able to reject it when it is false. This is consistent with defining  $H_0: \mu = .54$  (the status quo) and  $H_A: \mu \neq .54$ . To convince yourself, note that the implications of a Type I error would be to claim the evolutionary theory is false when it is true, whereas a Type II error would correspond to not being able to refute the evolutionary theory when it is false. With this setup, the scientist will refute the theory only if the data overwhelmingly suggest that it is false.

### The effect of $\alpha$ on the rejection region of a two-sided test

For a size  $\alpha$  test, you reject  $H_0: \mu = \mu_0$  if

$$t_s = \frac{\overline{Y} - \mu_0}{SE_{\overline{Y}}}$$

satisfies  $|t_s| > t_{crit}$ .



Rejection Regions for .05 and .01 level tests

The critical value is computed so that the area under the t-probability curve (with df = n - 1) outside  $\pm t_{crit}$  is  $\alpha$ , with .5 $\alpha$  in each tail. Reducing  $\alpha$  makes  $t_{crit}$  larger. That is, reducing the size of the test makes rejecting  $H_0$  harder because the rejection region is smaller. A pictorial representation is given above for the Tocopilla data, where  $\mu_0 = 0.54$ , n = 12 and df = 11. Note that  $t_{crit} = 2.201$  and 3.106 for  $\alpha = 0.05$  and 0.01, respectively.

The mathematics behind the test presumes that  $H_0$  is true. Given the data, you use

$$t_s = \frac{Y - \mu_0}{SE_{\overline{Y}}}$$

to measure how far  $\overline{Y}$  is from  $\mu_0$ , relative to the spread in the data given by  $SE_{\overline{Y}}$ . For  $t_s$  to be in the rejection region,  $\overline{Y}$  must be significantly above or below  $\mu_0$ , relative to the spread in the data. To see this, note that rejection rule can be expressed as: **Reject**  $H_0$  if

$$\overline{Y} < \mu_0 - t_{crit} SE_{\overline{V}}$$
 or  $\overline{Y} > \mu_0 + t_{crit} SE_{\overline{V}}$ .

The rejection rule is sensible because  $\overline{Y}$  is our best guess for  $\mu$ . You would reject  $H_0 : \mu = \mu_0$ only if  $\overline{Y}$  is so far from  $\mu_0$  that you would question the reasonableness of assuming  $\mu = \mu_0$ . How far  $\overline{Y}$  must be from  $\mu_0$  before you reject  $H_0$  depends on  $\alpha$  (i.e. how willing you are to reject  $H_0$ if it is true), and on the value of  $SE_{\overline{Y}}$ . For a given sample, reducing  $\alpha$  forces  $\overline{Y}$  to be further from  $\mu_0$  before you reject  $H_0$ . For a given value of  $\alpha$  and s, increasing n allows smaller differences between  $\overline{Y}$  and  $\mu_0$  to be **statistically significant** (i.e. lead to rejecting  $H_0$ ). In problems where small differences between  $\overline{Y}$  and  $\mu_0$  lead to rejecting  $H_0$ , you need to consider whether the observed differences are important.

In essence, the t- distribution provides an objective way to calibrate whether the observed  $\overline{Y}$  is typical of what sample means look like when sampling from a normal population where  $H_0$  is true. If all other assumptions are satisfied, and  $\overline{Y}$  is inordinately far from  $\mu_0$ , then our only recourse is to conclude that  $H_0$  must be incorrect.

### Two-sided tests, CI and p-values

An important relationship among two-sided tests of  $H_0: \mu = \mu_0$ , CI, and p-values is that

size  $\alpha$  test rejects  $H_0 \Leftrightarrow 100(1-\alpha)\%$  CI does not contain  $\mu_0 \Leftrightarrow$  p-value  $\leq \alpha$ .

size  $\alpha$  test does not reject  $H_0 \Leftrightarrow 100(1-\alpha)\%$  CI contains  $\mu_0 \Leftrightarrow$  p-value >  $\alpha$ .

For example, an  $\alpha = .05$  test rejects  $H_0 \Leftrightarrow 95\%$  CI does not contain  $\mu_0 \Leftrightarrow$  p-value  $\leq .05$ . The picture below illustrates the connection between p-values and rejection regions.



Either a CI or a test can be used to decide the plausibility of the claim that  $\mu = \mu_0$ . Typically, you use the test to answer the question is there a difference? If so, you use the CI to assess how much of a difference exists. I believe that scientists place too much emphasis on hypothesis testing.

# Statistical versus practical significance

Suppose in the Tocopilla meteorite example, you rejected  $H_0: \mu = .54$  at the 5% level and found a 95% two-sided CI for  $\mu$  to be .55 to .58. Although you have sufficient evidence to conclude that the population mean cooling rate  $\mu$  differs from that suggested by evolutionary theory, the range of plausible values for  $\mu$  is small and contains only values close to .54. Although you have shown statistical significance here, you need to ask yourself whether the actual difference between  $\mu$  and .54 is large enough to be important. The answer to such questions is always problem specific.

# Design issues and power

An experiment may not be sensitive enough to pick up true differences. For example, in the Tocopilla meteorite example, suppose the true mean cooling rate is  $\mu = 1.00$ . To have a 50% chance of rejecting  $H_0$ :  $\mu = .54$ , you would need about n = 30 observations. If the true mean is  $\mu = .75$ , you would need about 140 observations to have a 50% chance of rejecting  $H_0$ . In general, the smaller the difference between the true and hypothesized mean (relative to the spread in the population), the more data that is needed to reject  $H_0$ . If you have prior information on the expected difference between the true and hypothesized mean, you can design an experiment appropriately by choosing the sample size required to likely reject  $H_0$ .

The **power** of a test is the probability of rejecting  $H_0$  when it is false. Equivalently,

power = 1 - Prob( not rejecting  $H_0$  when it is false ) = 1- Prob( Type II error ).

For a given sample size, the tests I have discussed have maximum power (or smallest probability of a Type II error) among all tests with fixed size  $\alpha$ . However, the actual power may be small, so sample size calculations, as briefly highlighted above, are important prior to collecting data. See your local statistician.

### **One-sided tests on** $\mu$

There are many studies where a one-sided test is appropriate. The two common scenarios are the **lower one-sided test**  $H_0: \mu = \mu_0$  (or  $\mu \ge \mu_0$ ) versus  $H_A: \mu < \mu_0$  and the **upper one-sided test**  $H_0: \mu = \mu_0$  (or  $\mu \le \mu_0$ ) versus  $H_A: \mu > \mu_0$ . Regardless of the alternative hypothesis, the tests are based on the *t*-statistic:

$$t_s = \frac{Y - \mu_0}{SE_{\overline{Y}}}.$$

### For the upper one-sided test

- 1. Compute the critical value  $t_{crit}$  such that the area under the *t*-curve to the **right** of  $t_{crit}$  is the desired size  $\alpha$ , that is  $t_{crit} = t_{\alpha}$ .
- 2. Reject  $H_0$  if and only if  $t_s \ge t_{crit}$ .
- 3. The p-value for the test is the area under the t-curve to the **right** of the test statistic  $t_s$ .

The **upper one-sided test** uses the **upper tail** of the t- distribution for a rejection region. The p-value calculation reflects the form of the rejection region. You will reject  $H_0$  only for large positive values of  $t_s$  which require  $\overline{Y}$  to be significantly greater than  $\mu_0$ . Does this make sense?

For the lower one-sided test

- 1. Compute the critical value  $t_{crit}$  such that the area under the *t*-curve to the **right** of  $t_{crit}$  is the desired size  $\alpha$ , that is  $t_{crit} = t_{\alpha}$ .
- 2. Reject  $H_0$  if and only if  $t_s \leq -t_{crit}$ .
- 3. The p-value for the test is the area under the t-curve to the left of the test statistic  $t_s$ .

The lower one-sided test uses the lower tail of the t- distribution for a rejection region. The calculation of the rejection region in terms of  $-t_{crit}$  is awkward but is necessary for hand calculations because SW only give upper tail percentiles. Note that here you will reject  $H_0$  only for large negative values of  $t_s$  which require  $\overline{Y}$  to be significantly less than  $\mu_0$ .

As with two-sided tests, the p-value can be used to decide between rejecting or not rejecting  $H_0$  for a test with a given size  $\alpha$ . A picture of the rejection region and the p-value evaluation for one-sided tests is given on the next page.



#### **Example: Weights of canned tomatoes**

A consumer group suspects that the average weight of canned tomatoes being produced by a large cannery is less than the advertised weight of 20 ounces. To check their conjecture, the group purchases 14 cans of the canner's tomatoes from various grocery stores. The weights of the contents of the cans to the nearest half ounce were as follows: 20.5, 18.5, 20.0, 19.5, 19.5, 21.0, 17.5, 22.5, 20.0, 19.5, 18.5, 20.0, 18.0, 20.5. Do the data confirm the group's suspicions? Test at the 5% level.

Let  $\mu$  = the population mean weight for advertised 20 ounce cans of tomatoes produced by the cannery. The company claims that  $\mu = 20$ , but the consumer group believes that  $\mu < 20$ . Hence the consumer group wishes to test  $H_0: \mu = 20$  (or  $\mu \ge 20$ ) against  $H_A: \mu < 20$ . The consumer group will reject  $H_0$  only if the data overwhelmingly suggest that  $H_0$  is false.

You should assess the normality assumption prior to performing the t-test. The stem and leaf display and the boxplot suggest that the distribution might be slightly skewed to the left. However, the skewness is not severe and no outliers are present, so the normality assumption is not unreasonable.

Minitab output for the problem is given below. Let us do a hand calculation using the summarized data. The sample size, mean, and standard deviation are 14, 19.679, and 1.295, respectively. The standard error is  $SE_{\overline{Y}} = s/\sqrt{n} = .346$ . We see that the sample mean is less than 20. But is it sufficiently less than 20 for us to be willing to publicly refute the canner's claim? Let us carry out the test, first using the rejection region approach, and then by evaluating a p-value.

The test statistic is

$$t_s = \frac{\overline{Y} - \mu_0}{SE_{\overline{Y}}} = \frac{19.679 - 20}{.346} = -.93$$

The critical value for a 5% one-sided test is  $t_{.05} = 1.771$ , so we reject  $H_0$  if  $t_s < -1.771$  (you can get that value from Minitab or from the table). The test statistic is not in the rejection region. Using the t-table, the p-value is between .15 and .20. I will draw a picture to illustrate the critical region and p-value calculation. The exact p-value from Minitab is .185, which exceeds .05.

Both approaches lead to the conclusion that we do not have sufficient evidence to reject  $H_0$ . That is, we do not have sufficient evidence to question the accuracy of the canner's claim. If you did reject  $H_0$ , is there something about how the data were recorded that might make you uncomfortable about your conclusions?

# COMMENTS:

- 1. The data are entered into the first column of the worksheet, which was labelled **cans**.
- 2. You need to remember to specify the lower one-sided test as an option in the 1 sample t-test dialog box.

Descriptive Statistics: Cans

SE Mean Minimum Median 19.750 Variable Ν N\*Mean StDev Q3 20.500 18.500 Cans 14 0 19.679 0.346 1.295 17.500 Variable Maximum 22.500 Cans Stem-and-Leaf Display: Cans Stem-and-leaf of Cans Leaf Unit = 0.10 Ν = 14 5 0 12447742111 18 18 19 20 21 22 22 22 55 555 000 5

```
One-Sample T: Cans
Test of mu = 20 \text{ vs} < 20
                                                     95%
                                                   Upper
Variable
             Ν
                 Mean
19.6786
                             StDev
                                      SE Mean
                                                   Bound
                                                                    P
0.185
                                                            -0.93
            14
Cans
                                                 20.2915
                            1.2951
                                       0.3461
```



How should you couple a one-sided test with a CI procedure? For a lower one-sided test, you are interested only in an **upper bound** on  $\mu$ . Similarly, with an **upper one-sided test** you are interested in a **lower bound** on  $\mu$ . Computing these type of bounds maintains the consistency between tests and CI procedures. The general formulas for lower and upper  $100(1-\alpha)\%$  confidence bounds on  $\mu$  are given by

$$\overline{Y} - t_{crit}SE_{\overline{Y}}$$
 and  $\overline{Y} + t_{crit}SE_{\overline{Y}}$ 

respectively, where  $t_{crit} = t_{\alpha}$ .

In the cannery problem, to get an upper 95% bound on  $\mu$ , the critical value is the same as we used for the one-sided 5% test:  $t_{.05} = 1.771$ . The upper bound on  $\mu$  is

$$\overline{Y} + t_{.05}SE_{\overline{V}} = 19.679 + 1.771 * .346 = 19.679 + .613 = 20.292.$$

Thus, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29. As expected, this interval covers 20.

If you are doing a one-sided test in Minitab, it will generate the correct one-sided bound. That is, a lower one-sided test will generate an upper bound, whereas an upper one-sided test generates a lower bound. If you only wish to compute a one-sided bound without doing a test, you need to specify the direction of the alternative which gives the type of bound you need. An upper bound was generated by Minitab as part of the test we performed earlier. The result agrees with the hand calculation.

Quite a few packages, including only slightly older versions of Minitab, do not directly compute one-sided bounds so you have to fudge a bit. In the cannery problem, to get an upper 95% bound on  $\mu$ , you take the upper limit from a 90% two-sided confidence limit on  $\mu$ . The rational for this is that with the 90% two-sided CI,  $\mu$  will fall above the upper limit 5% of the time and fall below the lower limit 5% of the time. Thus, you are 95% confident that  $\mu$  falls below the upper limit of this interval, which gives us our one-sided bound. Here, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29, which agrees with our hand calculation.

One-Sample	Τ:	Cans				
Variable	N	Mean	StDev	SE Mean	90%	CI
Cans	14	19.6786	1.2951	0.3461	(19.0656,	20.2915)

The same logic applies if you want to generalize the one-sided confidence bounds to arbitrary confidence levels and to lower one-sided bounds - always double the error rate of the desired one-sided bound to get the error rate of the required two-sided interval! For example, if you want a lower 99% bound on  $\mu$  (with a 1% error rate), use the lower limit on the 98% two-sided CI for  $\mu$  (which has a 2% error rate).