CHAPTER 3

Comparing means

Example 3.1. *A group of 30 cross country runners ran a 5k route at the beginning of their season. After a season of running and training, they ran the same 5k route. Their times were measured and compared. The coach wants to determine if their running speeds significantly increased as a test of his new training program.*

This chapter serves as a bridge between the univariate tests of the past few chapters and the bivariate tests of the next few chapters. This chapter focuses on comparing means, either between samples or between a sample and a proposed population mean. Throughout this chapter, pay attention to the following:

- The measurements on the units of analysis must be categorized uniquely *and* completely; each must fit in one and only one category. This can be group membership or, as in the case above, time measurement.
- The number of categories determines which method you ultimately use. T-tests are for comparing two means; ANOVA is for comparing more than two means.
- If you reject your null hypothesis in the ANOVA procedure, you need to go farther and

determine which categories have statistically different means.

- These methods, as usual, assume that the measurements are Normally distributed within the categories. Violations of this assumption has severe consequences in the applicability of the models. However, there are alternative methods that can be used if the underlying measurements are not distributed Normally. Using these methods, however, will reduce the power of the tests. Remember, you can't get something for nothing.
- Some methods also assume balanced data and homoskedasticity across categories. Not all do, however. Use the appropriate tests from earlier chapters to test these assumptions.

3.1 Units of Analysis

Before we begin discussing t-tests and other methods to compare means, it is time we discuss 'units of analysis'. In the social sciences, the unit of analysis is the entity that serves as the focus of your theory, the item on which (or on whom) you ostensibly perform your measurements. It is very important to be able to articulate the unit of analysis. Without knowing your unit, there is no way of knowing how your variables are supposed to affect it.

There is a difference between a unit of analysis and a level of analysis. The level of analysis refers to the aggregation level of your variable. There are several different ways of categorizing the aggregation levels, however, the four basic levels of analysis are the individual level, the societal (or group) level, the State level, and the system level.

An example should make these differences clearer. In some of my research, I am trying to model the behavior of terrorist groups in their decision to use terrorism. Some of the variables I use include ethnic separation, level of democracy in the State, economic expansion in the State, and the level of globalization in the world. Here, the unit of analysis is the terrorist group. All variables I measure must affect the group in my theory. The variables are taken from three different levels of analysis. The ethnic separation variable is measured at the group level; that is, that variable measures how separate the *group* is from its neighbors. The democracy variable is measured at the state level of analysis; it measures an aspect of the state. In the theory, state-level factors affect the group, therefore it makes sense to include the variable under the guise of 'the democracy the *group* experiences.' The economy is also a state-level variable. It is included because the group

also feels the effects of a poor economy. It affects all people in the world (albeit differently). Finally, globalization is a system-level variable, because its effects are felt on all states in (by all members of) the system. As it affects the states, it also affects the groups within the states.

The missing level is the individual level. In this example, no variable is measured at the individual level. Such measures may include employment status, group membership, family status, etc. With that said, as the unit of analysis in this research is the ethnic group, individual-level variables cannot be used in this research. In fact, there are ontological reasons why lower levels of analysis *cannot* be used to measure higher levels, although the opposite is certainly not the case.

3.2 Comparing one or two means

The next three sections deal with comparing the means of two populations of measurements, or a the mean of a population with a hypothetical value. The differences among the three methods depends on your knowledge about the underlying distribution. The first two methods rely on mathematical relationships between known probability distributions. The final is based relationships within an unknown, yet symmetric, distribution.

3.2.1 The z-test

Let us suppose that you have a single sample of data from a normal population for which you know the variance,[1] and that you wish to test the hypothesis that the population mean, μ, is equal to a specified value, μ_0. That is:

$$H_0 : \mu = \mu_0$$

The thing we must do is create a test statistic, a function of the data, that is large for sample means far away from μ_0. In other words, we want to measure a distance between the sample average and the hypothesized population mean. One possible formula for this is $\left|\bar{x} - \mu_0\right|$, where $|\cdot|$ indicates the absolute value of the argument (the magnitude, the distance from zero).

The good news is that we do know the probability distribution of this difference. Because $X \sim \mathcal{N}(\mu, \sigma^2)$, subtracting μ_0 from each x tells us that $X \sim \mathcal{N}(\mu - \mu_0, \sigma^2)$. Thus, if our null

[1] This requirements is rarely met in reality, but it does offer a good introduction to the next test.

hypothesis is correct, $X \sim \mathcal{N}(0, \sigma^2)$.

We know the formula to calculate the sample mean is $\bar{x} = \frac{1}{n}\sum x_i$. Because this is a linear operator, we know $\overline{X} \sim \mathcal{N}(0, \sigma^2/n)$. Knowing this fact, we have our final test statistic

$$z_{p/2} = \frac{\left|\bar{x} - \mu_0\right|}{\sqrt{\frac{\sigma^2}{n}}} \tag{3.1}$$

This statistic is distributed $Z \sim \mathcal{N}(0, 1)$, the Standard Normal distribution. We use the Standard Normal table (Table **??**) in Appendix **??**. The reason for the "$p/2$" subscript on z is that this is a two-tailed test. For now, just accept this statement. Tailness will be discussed in a little while.

Instead of calculating the p-value, we can calculate confidence intervals for a *given* α level. The formula is just a rewriting of Eqn 3.1. The $1 - \alpha$ percent symmetric confidence interval has endpoints given by Equation 3.2:

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \tag{3.2}$$

As you are exposed to more tests and confidence intervals, you will see a duality between the two. One rejects a null hypothesis if the calculated p-value is too low. One rejects a null hypothesis if the confidence interval does not contain the hypothesized value.

Example 3.2. *Let us assume that an individual's reaction times are Normally distributed with variance* $\sigma^2 = 16$*. A researcher hypothesizes that Bob's mean reaction time is 43 seconds. To test this hypothesis, she measured his reaction time ten times:* $45, 48, 42, 44, 50, 45, 49, 46, 43,$ *and* 48 *seconds. Does the data support the hypothesis? Calculate a 95% confidence interval for the population mean.*

Solution: Our first step is to calculate the sample average, $\bar{x} = \frac{1}{n}\sum x_i = 46$. With this, we calculate our z-statistic:

$$z_{p/2} = \frac{\left|\bar{x} - \mu_0\right|}{\sqrt{\frac{\sigma^2}{n}}} = \frac{|46 - 43|}{\sqrt{\frac{16}{10}}} = 2.37$$

Looking in the z-table in the back,[2] we find that the table probability is 0.0089. As this is $p/2$, we know our p-value is 0.0178. Thus, as the p-value is less than our usual $\alpha = 0.05$ level, we can reject the null hypothesis and conclude that the data suggests Bob's reaction times are different from 43 seconds.

[2]Note that the z-value is found around the edges of the table, while the p-values are in the interior.

Now, we need to calculate the 95% symmetric confidence interval for the population mean. Using Eqn 3.2, we have the endpoints given by:

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 46 \pm 1.96\sqrt{\frac{16}{10}} = (43.52, 48.48)$$

Thus, we are 95% confident that the true population mean is between 43.52 and 48.48. Note that this interval does *not* contain the value 43. Thus, we can also conclude that (at the $\alpha = 0.05$ level) Bob's reaction times are different from 43 seconds. ◇

The R function

Because the assumptions underlying the z-test are not realistic, there is no standard R function to perform the test. However, showing you what the function would look like will give you more insight into programming, into R, and into the test itself.

For this code to actually work, it would require three sections. A section to check that the input is appropriate, a section to prepare the output to be readable, and a section doing the actual calculations. In the interest of brevity, the first two sections will be skipped. Also for the sake of brevity, the listing only shows the case for a two-tailed test.

With those caveats, here is the listing.

```
1  z.test <- function(x, y=NULL,
2                     sigma2,
3                     mu0 = 0,
4                     alternative="two.sided",
5                     conf.level=0.95
6           ) {
7
8  alpha <- 1-conf.level
9  se    <- sqrt( sigma2/length(x) )
10 xbar  <- mean(x)
11 z     <- (xbar-mu0)/se
12
13 if(alternative=="two.sided") {
14          a   <- pnorm(abs(z)) - pnorm(-abs(z))
15          lcp <- xbar - qnorm(1-alpha/2) * se
16          ucp <- xbar + qnorm(1-alpha/2) * se
17 }
18
19 p <- 1-a
```

Lines 1 through 6 initialize a new function, which will be called `z.test`. This function requires two pieces of information, the sample (`x`) and the population variance (`sigma2`). We know this because no default value is given to them. This function also allows you to specify a hypothesized population mean (`mu0`), direction of the test (`alternative`), and the confidence level (`conf.level`). If you do not specify any of these optional parameters, the defaults will be chosen (`0`, `two.sided`, and `0.95`, respectively).

Lines 8 through 11 calculate the alpha level, the standard error, and the z-statistic. The standard error is as usual $se = \sqrt{\frac{\sigma^2}{n}}$, as is the z-statistic, $z = \frac{\bar{x}-\mu_0}{se}$.

The third block, lines 13 through the end, contains the code to calculate the confidence interval and the p-value for a two-sided test. The `pnorm(z)` function returns the probability of a standard normal variable taking on values less than `z`. In other words, `pnorm()` is the cumulative distribution function, $\Phi(z) = \mathbb{P}[Z \leq z]$. The `qnorm(p)` function returns the z-value corresponding to a given probability, `p`. Thus, `qnorm(1-alpha/2)` corresponds to the $z_{\alpha/2}$ in Eqn 3.2.

! **Warning**: *The z-test is* extremely *sensitive to the closeness of the sample variance to the population variance. As a rule of thumb, you should avoid using the z-test. However, I introduce it here to give you an introduction to a typical form of a test statistic.*

⋆ ⋆ ⋆

Thus, we created a perfectly viable test statistic in this section. We started with an idea that we wanted a large difference to result in a large test statistic. We then manipulated that difference so that the test statistic had a *known* probability distribution. This is the reason we had to divide by the standard error (*se*), to change the difference into a known probability distribution.

We will use this same process to create a test statistic for those cases when you do not know the population variance.

3.2.2 The t-test

The drawback to the z-test is that it requires you to know the variance of the population under consideration. Reality suggests that if you do not know the population's mean, then you will not know its variance. A further drawback is that if you do not know the variance, the p-values (and the confidence intervals) calculated from the z-test will most certainly be wrong.

If you do not know the population variance, you may be tempted to substitute the sample variance in its stead. However, this changes the distribution of the test statistic, and this distribution is very different from the Normal distribution for small sample sizes.

Thus, we have to create a new statistic to measure the difference in the mean. Actually, we will use the statistic suggested in the previous paragraph, but with the change suggested above. Instead of using the population variance, we will use the *sample* variance. Recall from Chapter **??** that the formula for the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.3}$$

However, this is a random variable (it is a function of the data). As such, it has a distribution associated with it. In fact, we can prove (should we ever want to) that $\nu S^2 \sim \chi^2_\nu$, where ν (the Greek letter nu) is the number of degrees of freedom ($\nu = n - 1$ here), and χ^2 is the Chi-squared distribution.[3]

With this information, we can create our test statistic *and* know its probability distribution. The statistic will be

$$t = \frac{\left|\bar{x} - \mu_0\right|}{\sqrt{\frac{s^2}{n}}} \tag{3.4}$$

This formula should look very familiar to us; it has the exact same form as Eqn 3.1, but with s^2 substituted for σ^2. As such, the logic of this test statistic is the same as for the z-statistic. However, where the z-statistic was Normally distributed, the t-statistic is distributed t_ν. This is because the t distribution is the ratio of a Normal distribution to the square root of a Chi-squared distribution (divided by its degrees of freedom).[4]

The calculation of the confidence interval parallels that of the z-test, with the exception that

[3]The χ^2_ν distribution is the sum of ν independent squared Normal random variables. That is, if $X_i \sim \mathcal{N}(0,1)$, and if $Y = \sum X_i$, then $Y \sim \chi^2_\nu$.

[4]The t distribution was created by William Sealy Gosset in 1908, while he worked as a statistician for Guinness Brewery in Dublin. The creation of the t is shrouded in legend as befitting a story originating in a brewery. The basics are that Gosset worked with small samples, on which he used the z-test. However, he soon realized that his p-values were not correct. So, he created a distribution that better fit small sample tests. He published under the pseudonym because Guinness did not want its competitors to know they used statisticians for quality control. Gosset's pseudonym was 'Student.' And thus was born the Student's t distribution.

you must be aware of the degrees of freedom.

$$\bar{x} \pm t_{\nu,\alpha/2} \sqrt{\frac{s^2}{n}} \tag{3.5}$$

Example 3.3. *Let us revisit Example 3.3. Instead of unrealistically knowing the variance of the population, let us use the sample to estimate the appropriate variance.*

Solution: The three values of consequence are the sample size ($n = 10$), the sample mean ($\bar{x} = 46$), and the sample variance ($s^2 = 7.111$). With this information, our t-statistic (from Eqn 3.4) is

$$t = \frac{\left|\bar{x} - \mu_0\right|}{\sqrt{\frac{s^2}{n}}} = \frac{|46 - 43|}{\sqrt{\frac{7.1111}{10}}} \approx 3.56$$

Thus, $t = 3.56$, which is distributed as t_9. Using our tables or our computer, we get that the p-value = 0.00614. As this is less than our usual $\alpha = 0.05$, we can reject the null hypothesis that Bob's reaction times really are 43 seconds.

The confidence interval is calculated using Eqn 3.5. As such, our 95% confidence interval for the population mean is

$$\mu \in (44.09, 47.91)$$

Thus, since the proposed mean, $\mu_0 = 43$, is not in the 95% confidence interval, we can reject the null hypothesis at the $\alpha = 0.05$ level and conclude that Bob's reaction time is not 43 seconds. ⋄

Note. *Notice that the two conclusions of Example 3.3 are the same. This will always be the case. The confidence interval and the test statistic are two sides of the same coin.*

! **Warning**: *Also note that while the conclusions of the two examples (3.2 and 3.3) were the same, the p-values were different.*[5] *When the conclusion is obvious, you will usually get the same conclusion with the different p-values. As such, this will usually not be an issue.*

[5] Recall that the p-value is the probability of getting a result as extreme or more extreme assuming the null hypothesis is correct. From a logic standpoint, this means a p-value cannot prove or disprove the null hypothesis; the p-value assumes the null hypothesis is correct. Thus, the p-value only specifies (in a certain sense) how believable it is that the null hypothesis is correct.

However, when the sample mean is close to the proposed population mean, differing p-values may force different conclusions. As such, you will want to avoid using bad hypothesis tests, which give you bad p-values.

Paired t-tests

Sometimes you have just one set of individuals, but two different measurements on each individual. Hypotheses you would usually test in this scheme concern a difference between the two measurements, much like the opening example.

Example 3.4. *The coach mentioned in the opening of this chapter has finally gathered his data. The time* improvement *for a sample of his runners is as follows: 60, 53, 125, 42, -240, -10, 35, 85, -95, and -30. (Negative values indicate that the final 5k time was slower than the initial 5k time.) Did the season and/or his training program affect the times of the runners?*

Solution: First, let us state the implied null hypothesis. If we let D be the time change for the population of runners on the team, then

$$H_0 : D = 0$$

Let us take this null hypothesis one step further. Let us *fully* state the distribution of the null hypothesis. Recall that our test statistic will be distributed t_ν, with $\nu = 9$. Thus, our distributional null hypothesis is

$$H_0 : D \sim t_{\nu=9}$$

When we state our null hypothesis in its distributional form, we realize much more about what we are assuming about the test we are using. Now, as this is our null hypothesis, our alternative hypothesis is that D is *not* distributed in this fashion:

$$H_A : D \nsim t_9$$

Now, to determine the answer, we can either calculate the confidence interval or the test statistic. In general, the test statistic is preferred. As such, let us calculate t and determine if we will reject

Student ID	Pre-test Score	Post-test Score	Student ID	Pre-test Score	Post-test Score
1423	3.4	3.6	6532	1.3	2.1
9683	4.4	4.2	3856	4.0	4.3
4586	3.1	4.2	1685	1.0	1.1
2685	2.6	4.1	2810	2.8	4.1
5945	3.3	2.1	1345	1.3	5.0
3856	3.0	4.1	3099	2.3	4.0

Table 3.1: *Sample of students and their pre- and post-test averages, to accompany Example 3.5.*

the null hypothesis. To calculate the test statistic, we need three pieces of information: the sample size ($n = 10$), the sample mean ($\bar{d} = 2.5$), and the sample variance ($s_d^2 = 11090.06$). Thus, the test statistic is

$$t = \frac{\left|\bar{d} - \mu_0\right|}{\sqrt{\frac{s_d^2}{n}}} = \frac{|2.5 - 0|}{\sqrt{\frac{11090.06}{10}}} \approx 0.0751$$

From this, and the fact that the degrees of freedom are $\nu = 9$, we calculate the p-value = 0.9418. As this is greater than our traditional $\alpha = 0.05$, we cannot reject the null hypothesis. As such, we conclude that the data supports the hypothesis that the training program was ineffective. Said another way, there is no evidence that the training program affected the times of the runners.

The 95% confidence interval is $D \in (-72.83, 77.83)$. I will leave it as an exercise for you to calculate this. Eqn 3.5 should be of some help. ◇

Example 3.5. *A science teacher wanted to increase the attraction of science to her students. She came across an article describing a new unit she could teach to them. She decided to test the efficacy of the unit to increase the interest of the students in science. To that end, she gave her students a pre-test and a post-test that asked the same questions about their feelings concerning science. A sample of the results ($n = 12$) are given in Table 3.1.*

According to the sample, is there sufficient evidence that the unit increased the students' interest in science?

Solution: This is an example of a paired samples t-test because the individuals are specific and repeated measures are taken on them. Thus, Student 1423 had two tests. On the first, she scored 3.4; on the second, she scored 3.6 — repeated measures on an individual.

If we define D as the difference in the test scores, then the null and alternative hypotheses (in distributional form) are

$$H_0 : D \sim t_{11}$$
$$H_A : D \nsim t_{11}$$

So, we perform a t-test on the differences. Doing so gives us our test statistic of $t = 2.4749$. As this is a two-sided test, the p-value will be $p = 0.03085$. At the $\alpha = 0.05$ level, we can reject the null hypothesis and conclude that the data suggest the unit improved the students' interest in science. I leave it as an exercise to determine that the 95% confidence interval is $D \in (0.0959, 1.6374)$. $\diamond$

This is actually the first time we have explicitly compared two samples of data. Previously, we compared a sample to a proposed parameter value. While it is true that this test reduced to a single-sample t-test, such is not always the case. This example relied heavily on the assumption of repeated measured on a single population. If the populations are not the same, then we must find a different test.

Two independent samples; equal variance

The previous section compared the means of two tests on a single sample of individuals. This section deals with comparing means across population samples.

Let us assume that you have *two* categories of individuals. For each individual, you measure a specific characteristic and the group membership. This can be as simple as measuring the height of several people to determine if men or women are taller, or it can be as complex as measuring a latent variable based on a variety of different measures on two different populations of individuals. The key is that you have a single measurement, a group membership (male or female), and you want to compare the means of the two categories.

To solve this question, we need to make a few assumptions: The heights are independent. The heights are distributed Normally in each category. The variances are the same in each category. In

other words, we are assuming the two populations differ *only in their means*.[6]

Under these conditions, we have the following test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.6}$$

Notice that this test statistic also has the basic form of all of the test statistics in this chapter: a difference divided by the standard error. Here, since we are assuming that the populations have a common variance, we are using a weighted average of the two sample variances in the denominator, s_p:

$$s_p = \sqrt{\frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}} \tag{3.7}$$

To test the statistic, you need to know the number of degrees of freedom. In the previous section, it was $n - 1$. Here, since there are two populations, it is $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

Example 3.6. *You decide to test the hypothesis that men and women are the same height. To do this, you measure the heights of 10 men and 15 women. The men had an average height of 70in, with a variance of 4. The women had an average height of 65in, with a variance of 5. Does the data support the hypothesis at the* $\alpha = 0.05$ *level?*

Solution: We are given the necessary information. Let us substitute it into our formulas (Eqns 3.6 and 3.7):

$$\begin{aligned} s_p &:= \sqrt{\frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{(9)4 + (14)5}{23}} \\ &\approx 2.14679 \end{aligned}$$

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{70 - 65}{2.14679 \sqrt{\frac{1}{10} + \frac{1}{15}}} \end{aligned}$$

[6] Again, these assumptions should be tested.

$$= \frac{5}{0.87642}$$
$$\approx 5.70501$$

The number of degrees of freedom are 23. Thus, from the tables, the critical value is 2.0687. As our test statistic $t > 2.0687$, we can reject the null hypothesis that men and women are the same height. ◇

We could actually go a bit farther if we had a computer. We could calculate the p-value that the hypothesis is correct. In `R`, the function is `(1-pt(5.70501, df=23))*2`. This gives a p-value of approximately 1.53×10^{-5}, which is tiny. Thus, even if we had chosen $\alpha = 0.005$, we could still safely reject the hypothesis that men and women are the same height. And, we did this based on 25 data points.

In the formula we used, `(1-pt(5.70501, df=23))*2`, we multiply the p-value by 2 because this is a two-tailed test. Also, the `R` function, `pt()` returns the cumulative probability function of the t distribution; that is, it returns $\mathbb{P}[T \leq t]$. It is easier to remember if you remember that the 'p' represents 'probability'.

Two independent samples; unequal variance

Now, you may be asking, what happens if I am not sure that the variances in the two populations are equal? As the previous formulae were based on the assumption that the variances were equal, relaxing that requirement changes the formula. Actually, the formula is much simpler and interpretable:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.8}$$

Notice that this formula also has the standard form of a t-test. The difference is in the denominator.

Before we heave a sigh of relief and ask why we don't always use Formula 3.8, we have to concern ourselves with the degrees of freedom for the t-statistic. Here is where the complexity

arises. The best general solution is to define the degrees of freedom as

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{3.9}$$

Thankfully, you only have to tell the computer to use this form; you do not have to calculate it yourself. However, you have to *know* to tell the computer.

Many statistical packages (including SPSS and SAS) provide t-statistics values under the assumption that the variances are equal *and* without that assumption. They will also give you p-values on the null hypothesis that the variances are equal. However, it is actually much simpler than that.

If you use the t-test enough times, you will notice that when the variances are equal, the degrees of freedom in Formula 3.9 is equal to the degrees of freedom when one assumes equal variance. And, when the null hypothesis is rejected, you are supposed to use the degrees of freedom in Formula 3.9. In other words, if you always use Formula 3.9, you will get the correct answer.

! **Warning**: *Well, I should place a warning here. Remember, there are* assumptions *underlying the t-test. The most important is that the two populations are Normally distributed. If that is not true (or close to being true), then you cannot use the t-test. This is a rather important assumption, as the next secion demonstrates.*

★ ★ ★

Thus far, we have assumed that we knew the underlying distribution of the data. Not only that, but we assumed that distribution was Normal. Either that, or we assumed the sample size was large enough that the Central Limit Theorem promised the sample mean was Normally distributed. However, in reality, the Central Limit Theorem does not offer quick convergence. In other words, if the underlying distribution is not close to Normal, then the sample size must be on the order of *several hundred* to ensure that the t-tests are applicable.

Figure 3.1 shows the results of a Monte Carlo experiment demonstrating this conclusion. Recall that for an appropriate test, the p-values are uniformly distributed, $p \sim \mathcal{U}(0, 1)$, if the null hypothesis is correct. The graphs above show the distribution of the p-values under different sample sizes

($n = 30, 50, 100, 250, 500,$ and 1000). In each case, $X \sim \mathcal{E}(\lambda = 1)$, which has a mean of $\mu = 1$. If the test is appropriate, all of the bars should be near the horizontal line. The bar that most concerns us is that first one, since that bar is the rate at which we wrongly reject the null hypothesis.

According to this experiment, one will reject the null hypothesis about 40% more often than you should when your sample size is $n = 30$. This proportion slightly improves when the sample size increases to $n = 50$; at that point, you will only wrongly reject the null about 25% more often than you should. It is not until you get to $n = 500$ that the difference is irrelevent.

Most books suggest that a sample size of $n = 30$ is sufficient for the Central Limit Theorem to guarantee the appropriate distribution to make the test work. However, this really depends on how close the underlying distribution is to Normal. When that distribution is not close, you will need a much larger sample size to achieve the correct $\alpha = 0.05$ level.

R code

The code to achieve these results is as follows.

```
1  p  <- numeric()
2  mc <- 100000
3  n  <- 30
4
5  for(i in 1:mc) {
6    x <- rexp(n, r=1)
7    p[i] <- t.test(x, mu=1)$p.value
8  }
9
10 png("ttest-parametric30.png", width=300, height=300)
11 hist(p, breaks=0:20/20, main="n=30", xlab="p-value", ylab="", las=1)
12 abline(h=mc/20, col=2)
13 dev.off()
```

As with all Monte Carlo experiments, there are three main parts: Initialization, Loop, Output. The initialization section (Lines 1 – 3), lets the program know that `p` is going to be a numeric vector, that the number of Monte Carlo trials will be 100,000, and that the sample size for each trial will be $n = 30$.

The loop (Lines 5 – 8) is responsible for actually performing the experiment. It samples n values from an Exponential distribution (Line 6), then performs a t-test on that sample (Line 7), storing the p-value in the variable `p`.

The output section (Lines 10 – 13), plots a histogram (Line 11) and a horizontal line at the expected height of each bar in the histogram (Line 12). Lines 10 and 13 output this histogram to

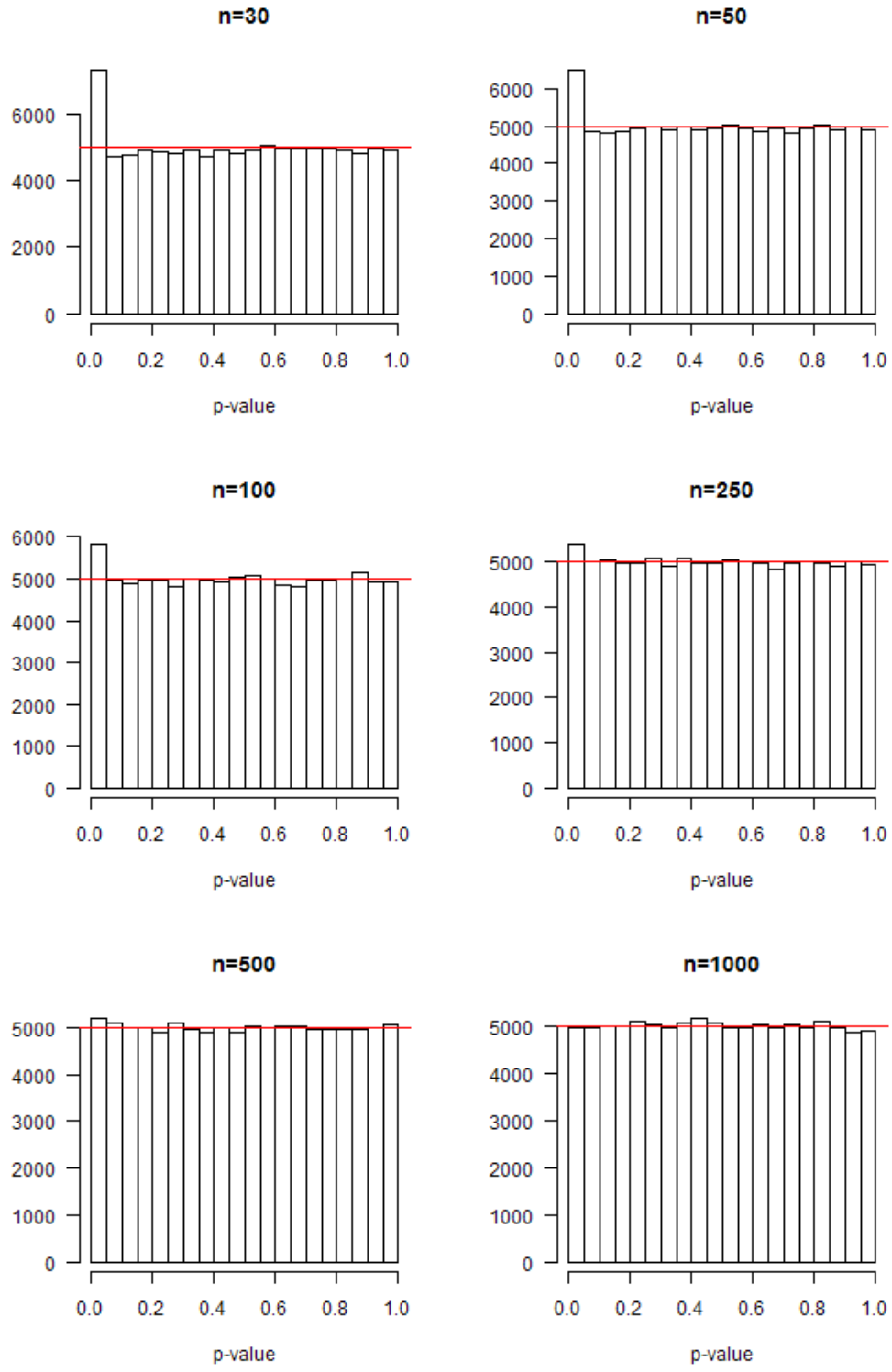

Figure 3.1: *Results from the Monte Carlo experiment comparing the outcomes of a t-test with the expected outcome. The number of replicates is 100,000 in each experiment.*

State	External debt (x_i)	$x_i - \mu_0$	Signed Rank
Australia	920	720	6
Norway	548	348	5
China	347	147	4
Brazil	216	16	1
Ukraine	104	-96	-2
South Korea	334	134	3

Table 3.2: *External debt (x_i), in billions for selected States. Data from the CIA(2009). For this, $\mu_0 = 200$.*

an external file and close that file.

3.2.3 Non-parametric tests

The means tests (so far) have all assumed that the underlying population was distributed Normally. This assumption is rarely true, and the Central Limit Theorem does not save us unless the sample size is quite large. So, what do we do if the sample size is small and the sample is not Normally distributed? In those cases, we need to use non-parametric methods.

Non-parametric tests *do* make assumptions about the underlying distribution, but those assumptions do not require a *specific* distribution. When comparing a single sample to a proposed population mean, the Wilcoxon test assumes the underlying distribution is symmetric. When comparing two samples, the Mann-Whitney test (also known as the Wilcoxon test for two samples) just require that the two samples differ only in their means.

One-sample Wilcoxon test

The Wilcoxon test of one mean requires that the sample be distributed symmetrically. So, let us assume that the population that gave us the measures is symmetrically distributed. Let us select the proposed population mean, μ_0. The first step is to subtract that proposed mean from each data value. Next, rank those differences, carrying the sign of the difference forward. Now, either add up all the negative ranks or all the positive ranks. Finaly, go to the Wilcoxon table and find the p-value (or critical value) corresponding to the test statistic and the sample size.

Example 3.7. *A friend of mine stated that the average external debt for the States in the world was just $200 billion. Using a sample of States, test my friend's assertion.*

Solution: A sample of six States was selected, and the amount of external debt was measured. The data are provided in Table 3.3. As the null hypothesis is that the population mean is \$200 billion, we first subtract 200 from each of the x_i. We then rank those values from smallest (in absolute value) to largest, retaining the sign. Next, we decide to add *either* the positive ranks or the negative ranks.[7] As there are fewer of them, let us sum the negative ranks.

The test statistic is $W_- = 2$ and the sample size is $n = 6$. From these two values, we use the Wilcoxon tables and see that the p-value is approximately $p = 0.10$. Thus, at the traditional level, we cannot reject the null hypothesis that my friends was correct. ◇

The logic behind this test hinges on the same logic as all of the tests we have discussed thus far: when the average is far from the proposed population mean, the null hypothesis should be rejected. Here, we are essentially using the median as our measure of 'average'. The distribution is based on permutations of the sample size; as such, it is very expensive to calculate.[8]

Using the computer makes this, of course, much faster to calculate. In `R`, the function to calculate the two-tailed probability for the above problem, the function is `wilcox.test(x, mu=200, alternative="two.sided")`.

Two-sample Mann-Whitney test

Just as there is a two-sample t-test used to compare means of two samples distributed Normal, there is a two-sample Wilcoxon-style test used to compare the means of two samples. The assumption of the Mann-Whitney test is that the two samples are distributed similarly, regardless of that distribution.

Example 3.8. *The research question is whether or not democratic States have a higher external debt than autocratic States. My friend asserts that autocratic States have a higher external debt than democratic States. Thus, his stated hypothesis is*

$$H_A : \mu_D < \mu_a$$

Notice that this is the alternative hypothesis. The null hypothesis always *include the "no effect" position.*

[7] We do one or the other because we know their sum is completely determined by the sample size. As such, there is no need to use both, and the Wilcoxon table is based on one of them.

[8] However, most statistical programs have a built-in function that calculates the distribution function quickly.

State	External Debt	Rank	Type
Australia	920	11	D
Norway	548	10	D
China	347	9	A
Brazil	216	7	D
Ukraine	104	5	A
South Korea	334	8	D
United Arab Emirates	129	6	A
Kazakhstan	93	4	A
Saudi Arabia	72	3	A
Pakistan	52	2	A
Iraq	50	1	A

Table 3.3: *External debt (x_i), in billions for selected States. Data from the CIA(2009). This table accompanies Example 3.8.*

As my friend stated autocracies have a greater *external debt, he is stating the alternative. Had he said "Autocracies do not have a lower external debt," his statement would be $\mu_d \geq \mu_a$, which includes the null position ($\mu_d = \mu_a$), and would be the null.*[9]

To test his hypothesis (actually, the null hypothesis), he gathered a sample of the external debt of several States in the world (Table 3.3). Assuming he selected a representative sample, does reality support his assertion?

Solution: The steps are quite similar to those of the (one-sample) Wilcoxon test. The first step is to rank the values, from either largest to smallest or smallest to largest. Once they are ranked, you either add up the ranks of either group. For this example, let us add the ranks of the autocratic States (easier to add smaller integers). With that, our test statistic is $W = 30$, and our smaple size is $n = 11$. Looking at the Mann-Whitney table, we find our p-value is $p < 0.01$. As this is a one-sided test, we do not need to double that p-value. As $p < \alpha = 0.05$, we can safely reject my friend's assertion that ◇

Again, in R, determining the p-value is very straight-forward. The applicable function is the same as for the one-sample test. You just pass it two samples instead of one. Thus, for this example,

[9] This comment is actually *very* important. Except when we are making power calculations, we only test th enull hypothesis. This is because the null contains all of the information about the distribution we are using in our test. This is why I have often written the null hypothesis in distributional form. Do not ever forget that these tests are based on the distribution assumed, not necessarily on the stated null hypothesis. The key is to match your statement with the distributional hypothesis.

you would use `wilcox.test(a,d, alternative="less")`. The output gives the test statistic ($W = 2$) and the p-value ($p = 0.01212$).

★ **Notice**: *There is a little disagreement in the literature (and in the statistical software) as to what should be the test statistic. Some assert that the larger of the sum of the ranks should be the test statistic. Others assert that it should be the smaller of the sum. Others do the calculations on the difference in the sums.*

It does not matter, however. The different programs use test distributions adjusted for their specific test statistics. Just be aware. This text will use the smaller of the two rank sums as a test statistic.

★ ★ ★

Non-parametric tests do not assume the distribution of the measures. They do, however, make other assumptions. In order to use the Wilcoxon test, you must assume that the the underlying distribution is symmetric. In order to use the Mann-Whitney test, you must assume the two samples are distributed in the same manner (except for the population mean).

3.3 Analysis of Variance

Thus far, we have only examined tests that help us to compare one sample to a proposed population mean or to compare the means of two samples. Unfortunately, we often have several samples or groups among which we want to compare means. The hypotheses in these cases is that all groups have the same population mean.

To motivate such questions, suppose my friend bets that the South East Conference (SEC) scores more points on average than the other football conferences. Technically, this question only has two populations: SEC and non-SEC teams. The measure is points scored, and the tests are as above. However, what if we ask: Are the major conferences different in terms of the number of points scored? Technically, this question can be read as merely a comparison of the means of two samples. However, let us assume my friend meant that *in general*, the SEC scored more points than the other conferences.

The difference is very subtle. The former question could be answered without question by the current data. The latter question needs to *estimate* the underlying population mean (points scorable by the SEC). The former question just needs a calculation of the means of the points scored. The latter question requires that we estimate the underlying "scoring ability" of the SEC.

To do this, we must estimate the mean of the six populations (one for each of the six major conferences). We cannot perform a t-test with six populations, unless we decide to test the six conferences pairwise — resulting in thirty separate tests. We can, however, use "Analysis of Variance" to find our answer.[10]

Analysis of Variance (ANOVA) is a procedure that generalizes the two-sample, homoskedastic t-test procedure to multiple categories. ANOVA still assumes the underlying populations are Normally distributed. It also still assumes homoskedasticity. As such, you will need to check these assumptions against your data.[11]

If all you want to do is check the point estimates, you do not have to perform ANOVA. If, however, you want to determine if the differences are statistically significant, then you will need to perform ANOVA. Table XX provides some information concerning the 2009 NCAA Football season for the six major conferences. We can use this to refute my friend's assertion that the SEC

[10] We have yet to determine the effects of performing several tests on your data. Let us wait until later to have this discussion.

[11] Furthermore, while the t-test has been altered to reduce the requirement of Normality to the point that many statistical programs implement it, ANOVA's alteration is much more rare.

scored more points, on average, in 2009 than any other major conference. According to Table XX, the SEC teams scored, on average, 29.15 points per game. This is the second highest, after the Big 12's 29.85 points per game. Thus, my friend is not correct.

He now asserts that the difference between the Big 12 and the SEC are not significant, and that the difference between the SEC and the PAC 10 *are* significantly different. Note how the wording has changed. No longer are we merely comparing two means. We are now trying to determine if a 'typical' SEC team scores no higher nor lower than a 'typical' Big 12 team, and if a 'typical' SEC team scores more points than a 'typical' PAC-10 team.

You may also have points scored for each football team in the six major BCS conferences over the course of the season. Your question may be: Does one conference score more than another conference.