Chapter 5:

A Single Population

This chapter deals with drawing conclusions about the center of a *single* population based on a sample of data. These measures of center may be the mean or the median. Throughout this chapter, pay attention to the following:

- The number of categories is important; it helps to determine which method you ultimately use. For this chapter, we will use only single categories.
- Some of these methods assume that the measurements are Normally distributed within the categories. Violations of this assumption have consequences in the applicability of the models (as discovered in Chapter 1). 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 cannot get something for nothing.
- Other methods require the measurements come from a symmetric distribution. If this is violated, then one cannot easily draw conclusions about the population mean, except through simulation (Monte Carlo).

❧❧❧

The mayor of İstanbul decided that the average response times for the 30 fire stations in his city were too great. To reduce response times, he required all 30 stations to use a new GPS mapping system, with the expectation that the increased cost would be balanced out by a significantly reduced average response time.

5.1: Units of and Levels of Analysis

Before we begin discussing appropriate methods to draw conclusions on a single sample, we should discuss 'units of analysis'. In the 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.[1] It is very important to be able to articulate the unit of analysis clearly and precisely. Without knowing your unit, there is no way of knowing how your variables are supposed to affect it.

unit of analysis

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*, not of your unit. There are several different ways of categorizing the aggregation levels; however, the four basic levels of analysis in the social sciences are the individual level, the societal (or group) level, the state level, and the system level.

level of analysis

An example should make these differences more clear: In some research, we try to model the behavior of groups in their decision to use terrorism. Some of the variables used include ethnic separation, level of democracy in the state, economic expansion in the state, and the level of globalization in the world (Forsberg 2007). Here, the unit of analysis is the group. Therefore, all measured variables must affect *the group*. 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 state (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. Table 5.1 diagrams the positions of these four variables.

variable

unit of analysis

level of analysis

The missing level in this example is the individual level. In this example, no variable is measured at the individual level. Such measures may include employment status, group membership, and family status *of the individual*. With that said, as the unit of analysis in this research is the ethnic

[1] The experimental sciences will frequently term these 'experimental units.'

	The Four Levels	Variables
	System	`globalization`
	State	`democracy`, `economic expansion`
The unit of analysis ⇒	Group	`ethnic separation`
	Individual	

Table 5.1: *Schematic of the levels of analysis, including the variables discussed in the text and the unit of analysis.*

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.

5.2: The z-test

The next sections deal with drawing conclusions about the population mean based on the sample of data from a *single* group. The differences among the methods depends on your knowledge about the underlying distribution of your measurements. The first two methods rely on mathematical relationships between known probability distributions. The final is based relationships within an unknown, yet symmetric, distribution.

power

> ***Note***: In general, where allowed, presented tests are given in order of declining power. That is, the most powerful tests are mentioned first in the discussion, followed by tests of less power. Also note that earlier tests (the more powerful ones) also tend to have the greatest number of assumptions behind them. That is the usual rule; you cannot get something (more power) for nothing (more assumptions).

Let us suppose that you have a sample of data (of size n) from a population with a measurement that is Normally distributed. Let us also assume that you know the variance of the population, σ^2. Finally, let us assume that you wish to test the hypothesis that the population mean, μ, is equal to a specified value, μ_0. That is, your null hypothesis is

null hypothesis

$$H_0 : \mu = \mu_0$$

for some specified numeric value μ_0. We can also state this null hypothesis in distributional form, which makes manifest many of the above assumptions:

$$H_0 : X_i \sim \mathcal{N}(\mu_0, \sigma^2)$$

Of course, we would fail to reject the null hypothesis if our sample mean exactly equaled our hypothesized mean. But, rarely does this happen in reality. Thus, let us suppose our sample mean does not equal the proposed mean. What is our rejection rule; when do we conclude that our null hypothesis is incorrect and conclude that our alternative hypothesis is preferable?

Should we reject the null hypothesis if the hypothesized mean and the sample mean differ by 1 unit? If not, then what about a difference of 5 units? In other words, where is the boundary between 'able to reject the null' and '*not* able to reject the null'? **sample mean**

The answer: *It depends*.

First, it depends on how willing you are to make a mistake in rejecting a correct null hypothesis. **Type I Error**

Second, it depends on the data and its spread.

5.2.1 Type I Error Rate

You have previously discussed the α level of a test. The level is also known as the Type I Error rate — the long-run proportion of times we **reject a true null hypothesis** (see Table 5.2). As a rate, the α-level ranges between 0 and 1. Smaller values are better, as it is an error rate. We can even set the Type I Error rate equal to zero: This means we *never* reject a true null hypothesis. Unfortunately, this also means we will fail to reject all of the *false* null hypotheses. **level**

Failing to reject a false null hypothesis is called a Type II Error (see Table 5.2). The symbol for the Type II Error rate is β. As with α, smaller is better since it is an error rate. Unfortunately, decreasing either results in increasing the other (although not linearly). In fact, setting $\alpha = 0$ results in $\beta = 1$ (and vice-versa). **power**

Statisticians made a decision long ago to focus on the Type I Error rate: We would rather continue what we are doing than wrongly switch; switching costs resources. However, we are not fanatical about it. Thus, we do not set $\alpha = 0$. Taking a cue from the legal system, we decided upon a default value for our Type I Error rate: $\alpha = 0.05$. There is no fundamental reason for selecting this as our α value, it is just tradition.

	Reality	
	H_0 true	H_0 false
Reject H_0	Type I Error (α)	
Fail to Reject H_0		Type II Error (β)

Table 5.2: *Table showing the two types of error for a test.*

p-value

Once computer time became inexpensive, we began to focus more on the calculated p-value and not the selected α-level. The p-value is the largest α for which we would reject the null hypothesis, given this data.

random variable

Note the subtlety of this point: The α is selected *before* we collect the data; the p-value is a *function* of the data. As the p-value is a function of the data, and as the data is a realization of a random variable, we know that the p-value is a random variable. Furthermore, as we wish this p-value to correspond to an *a priori* α-level, it must be distributed $P \sim \mathcal{U}(0,1)$. This was the basis of the Monte Carlo testing example in Section 1.4.

$\mathcal{U}(0,1)$

5.2.2 The Test Statistic

Now that we have chosen an α-level, we need to devise a way of determining when the observed data is "too extreme" for the null hypothesis. In order to determine whether or not to reject the null hypothesis, we need to create a test statistic. As it is a statistic, the test statistic is a function of the data. Ideally the test statistic should be easy to calculate, should be easy to use, should have a known distribution, and should be obviously related to the (population) parameter we wish to estimate.

test statistic

Not all test statistics meet these ideal criteria. Some tests, like the Kolmogorov-Smirnov statistic D, only meets one of the four suggestions — the fourth. Others, like the z-test, have all four.[2]

hypothesized distribution

For the opening example of this section, we have a ready-made statistic. If, as the null hypothesis states, $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$, then we know $\overline{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Thus, to calculate the p-value, we merely calculate the $\overline{x}$ and compare it to its hypothesized distribution.

[2] This is where it becomes imperative that you understand the assumptions behind the tests. This is also why stating the null hypothesis in distributional form is helpful — it makes manifest the distributional assumptions of the hypothesis you are testing.

How do we know the distribution of $\overline{X}$? By the following theorem:

Theorem 5.1. *Let $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ for a random sample of size n. Then $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.*

Proof. Let us define

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

First, we know the expected value of the sum of those random variables is the sum of the expected values. Thus,

$$\begin{aligned}
\mathbb{E}\left[\overline{X}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_1] \\
&= \frac{1}{n} n \mathbb{E}[X_1] \\
&= \mu
\end{aligned}$$

Second, we know that the variance of a sum of *independent* random variables is sum of the individual variances. Also we recall from its definition that $\mathbb{V}[aX] = a^2\mathbb{V}[X]$. Thus, we have

$$\begin{aligned}
\mathbb{V}\left[\overline{X}\right] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_1] \\
&= \frac{1}{n^2} n \mathbb{V}[X_1] \\
&= \sigma^2/n
\end{aligned}$$

Finally, we know the sum of Normally distributed random variables is a Normally distributed random variable (*v.i.* Appendix B).

Thus, we can conclude

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

□

p-value

Since the sample mean has a known distribution, we can calculate the p-value for any given sample mean and hypothesized population mean, μ_0. Of course, it is usually easier if we standardize things a bit. So, let us use the two-tailed transformation discussed in Appendix C:

$$z_{p/2} = \frac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}} \tag{5.1}$$

standard Normal distribution

non-directional

This statistic is distributed $Z \sim \mathcal{N}(0,1)$, the standard Normal distribution. We use the standard Normal table (Table C.1) in Appendix C. The reason for the "$p/2$" subscript on z is to remind us that this is a *two*-tailed test; we want to test if the observed mean is significantly *different* from the theorized mean.

Example 5.1: Let us assume that we know any forest's plant mass density (in kg/m^2) is Normally distributed with variance $\sigma_x^2 = 16$. A researcher hypothesizes that the plant mass density in the Niepołomice Forest in Poland is $43 kg/m^2$. To test this hypothesis, she measured plant mass density in ten randomly-selected places around the 42 sq mi forest: 45, 48, 42, 44, 50, 45, 49, 46, 43, and 48 kg/m^2. Do the data support the hypothesis?

Solution: The first step is usually to translate reality into probability statements. If we let X_i be a measurement of plant density in the forest, then our null hypothesis in distributional form is

$$H_0 : X_i \sim \mathcal{N}(\mu_0 = 43, \sigma_x^2 = 16)$$

Because of Theorem 5.1, this is equivalent to

$$H_0 : \overline{X}_i \sim \mathcal{N}(\mu_0 = 43, \sigma_{\overline{x}}^2 = 1.6)$$

z-test

Since we know the variance of the distribution, and since that distribution is the Normal distribution, we can use the z-test. To use this test, we

need to calculate the sample mean:

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = 46$$

With this, we calculate our test statistic:

$$z_{p/2} := \frac{\overline{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{46 - 43}{\sqrt{\frac{16}{10}}} = 2.37$$

Looking in the z-table in Appendix C,[3] we find that the table probability is 0.0089. As this is a two-tailed test, we know our p-value is twice that: 0.0178. Thus, as the p-value is less than our usual $\alpha = 0.05$ level, we *reject* the null hypothesis and conclude that the plant density in the Niepołomice Forest is significantly different from $43 kg/m^2$. **conclude**

We can use the computer to do the calculations for us:

```
2*pnorm(46, m=43, s=sqrt(1.6), lower.tail=FALSE),
```

which gives the same answer. ◇

5.2.3 Confidence intervals

Thus, from Example 5.1, we have more information about what the actual average plant density of the Niepołomice Forest: we know it is highly unlikely to be $43 kg/m^2$. That answer is not entirely satisfying; we are hardly closer to knowing the *actual* plant density than we were before.

So, what *is* the actual average plant density of the Niepołomice Forest?

Our best estimate is that the average plant density is $46 kg/m^2$ (this is referred to as our point estimate), our sample mean. However, how sure are we of that value? Is it likely that the average plant density is 44 instead of $46 kg/m^2$? What about 47 or 47.1432? **point estimate**

To answer this question, we need to find a range of likely values, an interval of values that satisfies our confidence requirement (based on our chosen α). This interval is called a confidence interval: It is based on the **confidence interval**

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

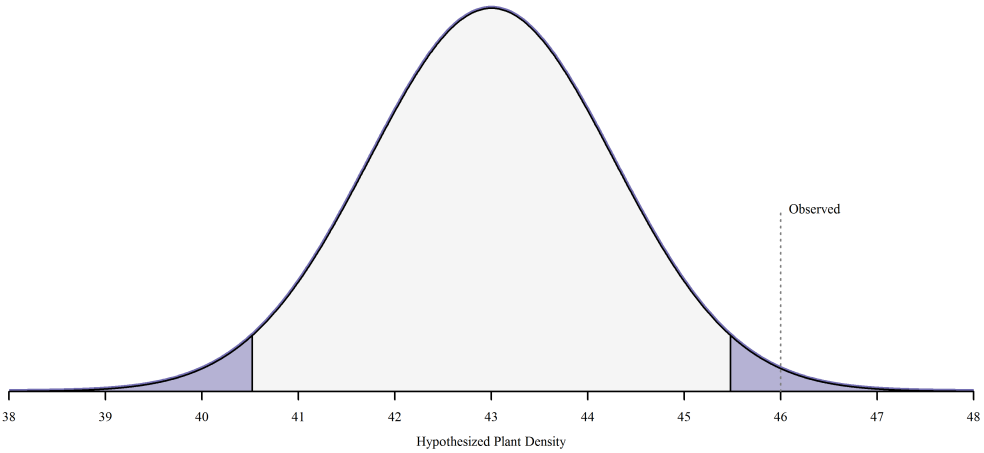

Figure 5.1: *Plot of the hypothesized distribution for the Niepołomice Forest problem. Note that the value of the (observed) sample mean, $\mu_0 = 46$, is located in the rejection region (dark blue). Thus, we reject the null hypothesis at the $\alpha = 0.05$ level and conclude that the plant density for the Niepołomice Forest is not equal to the assumed value of $43kg/m^2$.*

definitions of the p-value and of the test statistic used.

Recall that our test statistic is

$$z_{p/2} := \frac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}}$$

symmetry

Changing the p to α and the μ_0 to μ, using the symmetry of the Normal distribution, and solving for μ gives us:

$$\mu \in \overline{x} \pm Z_{\alpha/2}\sqrt{\sigma^2/n} \tag{5.2}$$

Type I Error rate

This formula gives both endpoints of the $100\times(1-\alpha)\%$ confidence interval for μ, the population mean. A likely value[4] for the population mean is between these two endpoints.

[4]Note that we are defining 'likely' in terms of our previously selected Type I Error rate, α. Different values of α will produce different confidence intervals for the population mean.

Applying this to the Niepołomice Forest (Example 5.1), we find our 95% confidence interval in a direct use of Equation 5.2:

$$\begin{aligned}\text{Interval endpoints} &= \overline{x} \pm Z_{\alpha/2}\sqrt{\sigma^2/n} \\ &= 46 \pm 1.96\sqrt{16/10} \\ &= (43.52, 48.48)\end{aligned}$$

Thus, we are 95% sure that the real average plant density in the Niepołomice Forest is between 43.52 and 48.48kg/m^2.

Note: The originally hypothesized average plant density was 43kg/m^2. This value is not in the 95% confidence interval. Thus, we would reject the null hypothesis that the average plant density was 43kg/m^2 (at the $\alpha = 0.05$ level). This is not a coincidence; there is a duality between hypothesis tests and confidence intervals. If one rejects the null hypothesis at the α level, then the $100 \times (1-\alpha)$% confidence interval will not contain the hypothesized value.

duality

5.2.4 The R function

Because the assumptions underlying the z-test are unrealistic, there is no standard `R` function to perform the test (there is a z-test function in the `RFS` package and on the book's website to `source`). 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.[5]

With those caveats, here is the partial listing:

```
1  z.test <- function(x, sigma2,
2                     mu0=0,
3                     alternative="two.sided",
4                     conf.level=0.95
5            ) {
6
7  alpha <- 1-conf.level
8  se    <- sqrt( sigma2/length(x) )
```

[5] To see the entire function, turn your web browser to `http://rfs.kvasaheim.com/Rfctn/z.test.R`.

```
 9  xbar  <- mean(x)
10  z     <- (xbar-mu0)/se
11
12  if(alternative=="two.sided") {
13          a   <- pnorm(abs(z)) - pnorm(-abs(z))
14          lcp <- xbar - qnorm(1-alpha/2) * se
15          ucp <- xbar + qnorm(1-alpha/2) * se
16  }
17
18  p <- 1-a
```

default value

Lines 1 through 6 initialize a new function, called `z.test`. This function requires two pieces of information (parameters), the sample (`x`) and the *known* population variance (`sigma2`). We know this because no default value is given for 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 used (`0`, `two.sided`, and `0.95`, respectively).

test statistic

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

confidence interval

non-directional

CDF

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, `pnorm(z)` $= \Phi(z) = \mathbb{P}[Z \leq z]$. The `qnorm()` function returns the z-value corresponding to a given probability, `p`; that is, if $\Phi(z) = p$, then `qnorm(p)` $= \Phi^{-1}(p) = z$. Thus, `qnorm(1-alpha/2)` corresponds to the $z_{\alpha/2}$ in Eqn 5.2.

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

❧ ❧ ❧

test statistic

Thus, we created a perfectly viable test statistic in this section. We started with an idea that we wanted a large difference between μ and μ_0 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*).

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

5.2.5 Why is this a Bad Test?*

One of the assumptions of the z-test is that we know the population variance. If we do not, we should *not* use the test. However, one is often tempted to substitute the sample variance for the population variance and still use the test. **That does not work**. If you use the sample variance, the distribution of the test statistic is no longer standard Normal, it is Student's t distribution (*v.i.*, Section 5.3). But, how bad is it if we violate the assumption?

population variance

Recall that one of the requirements for a test to be appropriate is for the p-values to be distributed standard Uniform: $P \sim \mathcal{U}(0,1)$. As such, we can check the appropriateness of the z-test using Monte Carlo methods. Remember the parts to running a Monte Carlo experiment? Refresh your memory (*v.s.*, Section 1.4) before reading through the code that follows.

$\mathcal{U}(0,1)$

Monte Carlo

Here is the code. Make sure you understand everything in it; you may need to determine whether you can use a given test under non-appropriate circumstances in your future.

```
#####
# Monte Carlo test of the z-test

# Preamble
set.seed(577)

# Initialize variables
p <- numeric()     # to be a vector of p-values
t <- 1000000       # number of trials to run
n <- 35            # the size of each sample

# The loop
for(i in 1:t) {
  x    <- rnorm(n)
  s    <- sd(x)
  p[i] <- z.test(x, mu=0, sigmax=s)$p.value
}

# Graphical test
hist(p, yaxt="n", xlab="p-value", main="", ylab="")
abline(h=t/20, col=4, lty=2)


# Kolmogorov-Smirnov test
ks.test(p, punif)
```

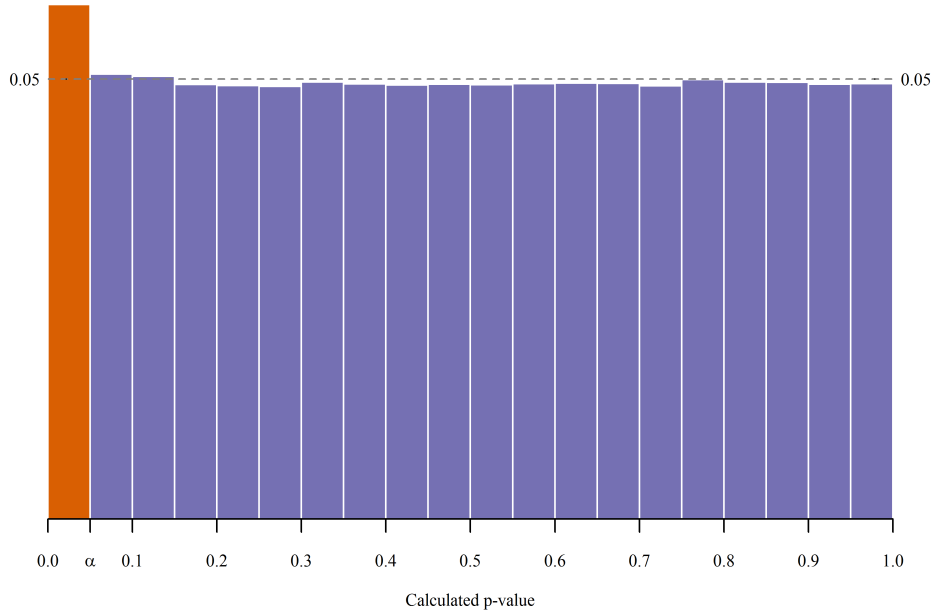

Figure 5.2: *The results of 1,000,000 Monte Carlo trials wherein the z-test is used, but the standard deviation is estimated from the data ($n = 35$). Note that, as Gosset discovered, one rejects more often than one should.*

histogram

When this is run, we get the histogram similar to Figure 5.2. Note that the first bar is significantly taller than it should be. Performing the Kolmogorov-Smirnov test (line 25) indicates that we can reject the null hypothesis that $P \sim \mathcal{U}(0,1)$. Thus, we conclude that using the sample variance in lieu of the population variance invalidates this test (for a sample size of $n = 35$). I leave it as an exercise to see how large of a sample size is necessary before one can use the z-test under these circumstances.

5.3: The t-test

population mean

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 confidence intervals and conclusions) 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 different from the Normal distribution (*v.s.*, Section 5.2.5).

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 that the formula for the sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{5.3}$$

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

Roman majuscule

$$\nu S^2 \sim \chi^2_\nu$$

where ν (the Greek letter nu) is the number of degrees of freedom (for the one-sample case, $\nu = n-1$), and χ^2 is the Chi-squared distribution. [6]

Chi-squared

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

$$t := \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}} \tag{5.4}$$

This formula should look very familiar to us; it has the exact same form as Eqn 5.1, 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 had a standard Normal distribution, the t-statistic is distributed t_ν, since the ratio of a Normal distribution to the square root of a Chi-squared distribution (divided by its degrees of freedom) is the t distribution.[7]

distribution

[6] The χ^2_ν distribution is the sum of ν independent squared Standard Normal random variables. That is, if $Z_i \sim \mathcal{N}(0,1)$, and if $Y = \sum_{i=1}^{\nu} Z_i$, then $Y \sim \chi^2_\nu$. For more about the Chi-squared distribution (see Appendix B.4).

[7] 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, but substituting s for σ. However, he soon realized that his p-values were not correct — he rejected far too often. So, he created a distribution that better fit small sample tests. He published under a 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.

degrees of freedom

The calculation of the confidence interval parallels that of the z-test, with the exception that you must be aware of the degrees of freedom. The endpoints of the $100 \times (1-\alpha)\%$ symmetric confidence interval are

$$\overline{x} \pm t_{\nu,\alpha/2}\sqrt{s^2/n} \tag{5.5}$$

Example 5.2: Let us revisit Example 5.1. Instead of unrealistically knowing the variance of the population, let us use the sample to estimate the appropriate variance and test the null hypothesis that the plant density in the Niepołomice Forest is $43 kg/m^2$.

Solution: The three values of consequence are the sample size ($n = 10$), the sample mean ($\overline{x} = 46$), and the sample variance ($s^2 = 7.111$). With this information, our t-statistic (from Eqn 5.4) is

test statistic

$$\begin{aligned} t &:= \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}} \\ &= \frac{46 - 43}{\sqrt{7.1111/10}} \\ &\approx 3.56 \end{aligned}$$

Thus, $t = 3.56$, which is distributed as $t_{\nu=9}$. Using our tables or our computer, we get that the p-value is 0.00614. As this is less than our usual $\alpha = 0.05$, we can reject the null hypothesis and conclude that the plant density in the Niepołomice Forest is significantly different from $43 kg/m^2$.

The confidence interval is calculated using Eqn 5.5. As such, our 95% confidence interval for the population mean is

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

confidence interval

Again, 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 once again conclude that the plant density is not $43 kg/m^2$.

More importantly, we are 95% confident that the real average plant density in the Niepołomice Forest is between 44.09 and $47.91 kg/m^2$. ◇

Note: Notice that the two substantive conclusions of Example 5.2 are the same. This will always be the case when the test statistic has a continuous

distribution. 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 (5.1 and 5.2) were the same, the confidence intervals and p-values were different.*[8] *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. 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.*

Example 5.3: The mayor mentioned in the opening of this chapter has finally gathered his data. The time *differences* for a random sample of those 30 fire stations is as follows: 0, -5, -12, 4, -24, -1, 3, -8, -19, and -15 minutes. (Negative values indicate that the average response time in 2011 was lower (faster) than that in 2010.) Does the data support the contention that the new GPS system helped change the response time from the 30 fire stations?

Solution: First, let us state the implied null hypothesis. If we define D to be the change in average response times for the population of fire stations, then

define

$$H_0 : D = 0$$

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

$$H_0 : D \sim \mathcal{N}(0, \sigma^2)$$

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:

$$H_A : D \nsim \mathcal{N}(0, \sigma^2)$$

[8] Recall that the p-value is the probability of getting data as extreme or more extreme than you did, 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.

Now, to determine the answer, we can either calculate the confidence interval or the test statistic. In general, the test statistic and the p-value are preferred, but calculating all will give a conclusion that is much more informative.

As such, let us calculate t and determine if we will reject the null hypothesis. To calculate the test statistic, we need three pieces of information: the sample size ($n = 10$), the sample mean ($\overline{d} = -7.7$), and the sample variance ($s_d^2 = 92.01$). Thus, the test statistic is

$$\begin{aligned} t &:= \frac{\overline{d} - \mu_0}{\sqrt{s_d^2/n}} \\ &= \frac{-7.7 - 0}{\sqrt{92.01/10}} \\ &\approx -2.538 \end{aligned}$$

From this, and the fact that the degrees of freedom are $\nu = 9$, we calculate the p-value to be 0.0318. As this is less than our usual $\alpha = 0.05$, we reject the null hypothesis and conclude that the data supports the hypothesis that the new GPS system was successful in reducing average response time in the İstanbul fire stations.

The 95% confidence interval is $D \in (-14.56, -0.84)$. Thus, we are 95% sure that the real improvement in response time for the İstanbul fire stations is between 0.84 and 14.56 minutes. ◇

remote data

Now, let us use R to perform the t-test. Thus far, we have use data on the Internet (using `read.csv`). Let us now see how to input the data directly into R. For this example, it is just two lines:

```
improvement = c(0, -5, -12, 4, -24, -1, 3, -8, -19, -15)
t.test(improvement)
```

The first line stores the data into the variable `improvement`. This line uses the `c()` construct, which combines the comma-separated list of values into a single vector of data. The second line performs the t-test using default values.

When you run these two lines, R outputs the following:

```
        One Sample t-test

data:  improvement
t = -2.5385, df = 9, p-value = 0.03179
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -14.5618767  -0.8381233
sample estimates:
mean of x
     -7.7
```

The output gives us the test statistic (`t = -2.5385`), the number of degrees of freedom (`df = 9`), the p-value (`p-value = 0.03179`), the mean of the time changes (`-7.7`), and the 95% confidence interval (`-14.562` to `-0.838`).

As the p-value is less than our usual $\alpha = 0.05$, we reject the null hypothesis and conclude that the response times are lower this year than last. The confidence interval tells us that we are 95% sure that this improvement is between 0.84 and 14.56 minutes.

reject

The improvement is statistically significant. It is up to the mayor of İstanbul to determine if the time improvement (somewhere from 0.84 to 14.56 minutes) is worth the cost of the GPS system.

Note: There is a difference between statistical significance and *practical* significance. We used the t-test to show statistical significance. It is only a function of the data. Practical significance depends on the confidence interval *and* the cost of the change. If the new GPS system cost \$1.00 total, then the switch would be worth it. If the GPS system cost \$1,000,000,000,000, then the switch would *not* be worth it. Where is the cutoff? That is a policy question and beyond the scope of this book.

practical significance

policy

This is actually the first time we have explicitly compared two samples of data (response time of the previous year and the response time of this year). 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

paired-sample t-test

	Scores			Scores	
Student ID	Pre-test	Post-test	Student ID	Pre-test	Post-test
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 5.3: *Sample of students and their pre- and post-test averages, to accompany the Science Unit example, 5.4.*

on a *single population*. If the populations are not the same, then we must find a different test (see Chapter 6).

Example 5.4: A science teacher wants 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 in increasing 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 5.3.

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

Solution: This is another example of what is termed 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. The other important aspect (as with the İstanbul example) is that we only care about the differences between the two measurements (tests).

repeated measures

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

hypotheses

$$H_0 : D \sim \mathcal{N}(0, \sigma^2)$$
$$H_A : D \nsim \mathcal{N}(0, \sigma^2)$$

differences

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. non-directional

Furthermore, we are 95% confident that the real increase in warmth toward science due to this teaching unit is between 0.0959 and 1.6374 points. ◇

The code to run this in R is

```
pre  = c(3.4,4.4,3.1,2.6,3.3,3.0, 1.3,4.0,1.0,2.8,1.3,2.3)
post = c(3.6,4.2,4.2,4.1,2.1,4.1, 2.1,4.3,1.1,4.1,5.0,4.0)
change = post - pre
t.test(change)
```

Note that I opted to let R calculate the differences (`change`) based on the pre-test (`pre`) and post-test (`post`) scores imported manually.

The R output is

```
        One Sample t-test

data:  change
t = 2.4749, df = 11, p-value = 0.03085
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.09592496 1.63740837
sample estimates:
mean of x
0.8666667
```

Solution: Thus, we know that the average increase in warmth toward science is 0.867 points, with a 95% confidence interval from 0.096 to 1.637 points. We also reject the null hypothesis and conclude that there was a statistically significant change in feelings of warmth toward science ($t = 2.47; \nu = 11; p = 0.031$). ◇

5.4: Testing the Assumption

As mentioned above, in formulating the t-test we assumed that the measurements came from a Normally-distributed population. If the measurements

do not, then the t-test is *not* appropriate. How, then, do we test this assumption?

Q-Q plot

Testing for Normality is straight forward. Graphically, one can use a Normal quantile-quantile plot or a histogram. If the plotted points fall near the diagonal line in the quantile-quantile plot, then there is sufficient evidence that the measurements are Normally distributed. Likewise if the histogram is bell-shaped, we make the same conclusion.

In addition to the graphical methods, we can use numeric methods to test the assumption (null hypothesis) of Normality. For a table of many Normality tests, see Table 13.1 of Chapter 13.

Shapiro-Wilk test

In lieu of using all of those tests, let us simply rely on the venerable Shapiro-Wilk test (1965). While other Normality tests are better under different circumstances, the Shapiro-Wilk test seems to be sufficient.

The R function for this test is `shapiro.test`. It takes only the sample to be tested for Normality.

To see this function in action, let us return to the Niepołomice example (Example 5.2). Running

```
massDensity = c(45,48,42,44,50,45,49,46,43,48)
shapiro.test(massDensity)
```

tests if the sample violates the Normality assumption. The code produces this output

```
        Shapiro-Wilk normality test

data:  massDensity
W = 0.9578, p-value = 0.7606
```

$p > \alpha \Rightarrow$ pass

As the p-value ($p = 0.7606$) is greater than our usual $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the data pass the assumption of Normality. Thus, we can use the t-test on this data.

Example 5.5: Recall the İstanbul example (Example 5.3). Determine if the sample violates the Normality assumption.

Solution: Running

```
improvement = c(0,-5,-12,4,-24,-1,3,-8,-19,-15)
shapiro.test(improvement)
```

tests if the sample violates the Normality assumption. The code produces this output

```
        Shapiro-Wilk normality test

data:  improvement
W = 0.9459, p-value = 0.6204
```

As the p-value ($p = 0.6204$) is greater than our usual $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the data pass the assumption of Normality. Thus, we can use the t-test on this data. ◇

$p > \alpha \Rightarrow$ **pass**

Example 5.6: Recall the science example (Example 5.4). Determine if the sample violates the Normality assumption.

Solution: We already defined the variable `change` as the improvement from the pre-test to the post-test. With that, we only need to run the command `shapiro.test(change)` to test the Normality assumption. Doing so gives this output

```
        Shapiro-Wilk normality test

data:  change
W = 0.9399, p-value = 0.4966
```

As the p-value ($p = 0.4966$) is greater than our usual $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the data pass the assumption of Normality. Thus, we can use the t-test on this data as well. ◇

$p > \alpha \Rightarrow$ **pass**

Note that all three passed the Shapiro-Wilk test of Normality. As such, the t-test is appropriate in all three cases. If one or more of the tests had failed the assumption test, we would not be able to use the t-test. What can we use?

5.4.1 The Effect of Non-Normality*

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 approximately Normally distributed (Appendix C). However, in reality, the Central Limit Theorem does not always offer quick convergence: if the

Normal

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 acceptable.

Figure 5.3 shows the results of a Monte Carlo experiment demonstrating this very 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 show the distribution of the p-values under different sample sizes ($n = 30, 50, 100, 250, 500$, and 1000). In each case, $X \sim \mathcal{E}xp(\lambda = 1)$, which has a mean of $\mu = 1/\lambda = 1$. 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 (recall $\alpha = 0.05$).

$\mathcal{U}(0,1)$

Type I Error rate

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 $n = 250$ that the difference appears irrelevant.

sample size

Warning: *Many 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 to achieve an α-level that is close to stated.*

As an aside, it also depends on how important your findings and how expensive it is to be wrong. The more expensive, the larger the needed sample size.

R CODE: The code to achieve similar results is as follows.

```
p <- numeric()        ## Vector to hold the p-values
B <- 1e6              ## Number of Monte Carlo trials
n <- 30               ## Tested sample size

for(i in 1:B) {       ## The loop
  x    <- rexp(n, rate=1)
  p[i] <- t.test(x, mu=1)$p.value
}

                      ## The graphic and the test
```

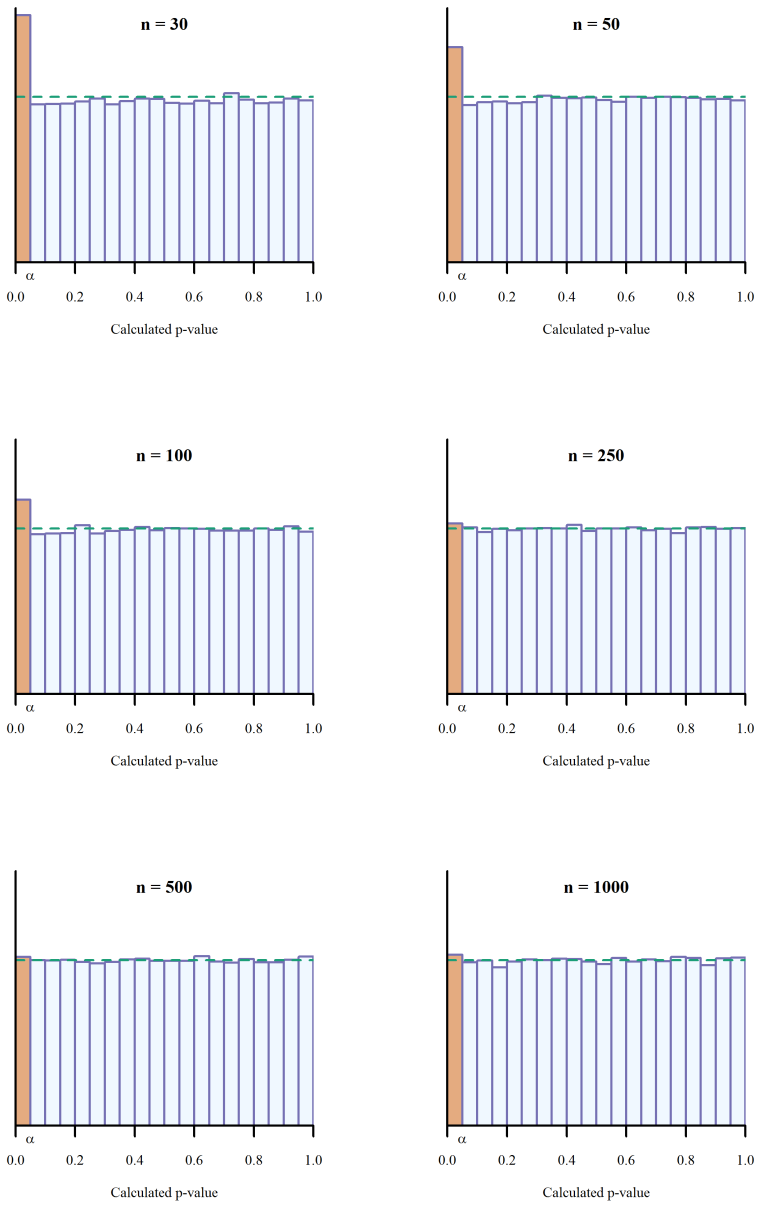

Figure 5.3: *Results from the Monte Carlo experiment comparing the outcomes of a t-test with the expected outcome. The number of replicates is 100,000 in each experiment.*

```
hist(p, breaks=0:20/20, main="n=30")
abline(h=B/20, col=4, lty=2)
ks.test(p,"punif")
```

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 1,000,000, and that the sample size for each trial will be $n = 30$.

p-value

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

Exponential

The analysis section (Lines 10–12), plots a histogram (Line 10) and a horizontal line at the expected height of each bar in the histogram (Line 11). This utilitarian graphic is followed by the Kolmogorov-Smirnov test to determine if the observed p-values have a $\mathcal{U}(0,1)$ distribution.

$\mathcal{U}(0,1)$

5.5: Non-Parametric Means Tests I

The tests of means (thus 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 or the distribution of the measurements is close to Normal. So, what do we do if the sample size is small and the sample is not sufficiently Normal? In those cases, we can use non-parametric methods.

CLT

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 continuous and *symmetric*. The binomial test only assumes the distribution is continuous.

symmetry

5.5.1 Wilcoxon test

The Wilcoxon test for the population mean requires that the population be distributed symmetrically (and continuously). So, let us assume that the population that gave us the continuous 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, add up

rank

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

Table 5.4: *External debt (x_i), in billions for selected States. Data from the CIA(2009). For this, $\mu_0 = 200$.*

the values of the negative ranks (or the positive ranks). Finally, go to the Wilcoxon table and find the p-value (or critical value) corresponding to the test statistic and the sample size.

Example 5.7: An associate of mine stated that the average external debt for the States in the world was just \$200 billion. Using a sample of six States, test the associate's assertion.

Solution: A sample of six States was randomly selected from all 190+ States, and the amount of external debt was measured. The data are provided in Table 5.4. 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.[9] As there are fewer, 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. In other words, the data support my friend's hypothesis. ◇

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 using the

[9] 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.

median

median as our measure of 'average' (as we assume the distribution is symmetric). The distribution is based on permutations of the sample size; as such, it is very expensive to calculate.[10]

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 is

```
wilcox.test(x, mu=200, alternative="two.sided")
```

Note that R uses the sum of the positive ranks as its test statistic.

Example 5.8: This same associate later stated that the average external debt for the States in the world was $500 billion. Using the same sample of six States, test the associate's new assertion.

Solution: Because the p-value is greater than our usual $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that there is not enough evidence to conclude my friend is wrong again ($V = 7; p = 0.5625$). ◇

The R code is just `wilcox.test(x, mu=500, alternative="two.sided")`.

5.5.2 A Binomial test

median test

To better understand the Wilcoxon test, let us remove some assumptions and formulate a basic test for the *median* value of a continuous population. We know that if $\tilde{\mu}$ is the median of the population and X is a value from that population, $\mathbb{P}[X \leq \tilde{\mu}] = 0.500$. In a sample of size n, we will have n such tests. The number of values less than the median, T, is a Binomial random variable, with parameters n and $\pi = 0.500$; that is, $T \sim Bin(n, 0.500)$.

test statistic

Thus, a natural test statistic is this variable T; it fits the four requirements of an ideal test statistic (*v.s.* Section 5.2.2). Note that this test is for the median of the population, not the mean. We can only apply it to the mean if the distribution is symmetric (the assumption of the Wilcoxon test).

symmetric

Example 5.9: Let us return to my associate's first assertion: the average external debt for the States in the world was just $200 billion. To use this Binomial test, we need to interpret "average" as median.

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

Solution: We can either count the number of States with external debt below (or equal to) the hypothesized median ($200 billion), or we can count the number above. Since $\pi = 0.500$, the distribution of T is symmetric.

Using the data in Table 5.4, our test statistic is $T = 1$. We know $T \sim Bin(6, 0.500)$. With this, we calculate the p-value as

$$\begin{aligned} p &:= 2 \times \mathbb{P}[T \leq 1] \\ &= 2(\mathbb{P}[T = 0] + \mathbb{P}[T = 1]) \\ &= 2(0.015625 + 0.09375) \end{aligned}$$

Thus, our p-value is 0.21875. As this is greater than our usual $\alpha = 0.05$, we fail to reject our null hypothesis and conclude that the data support my colleague's assertion. ◇

Warning: *As usual, we* cannot *conclude that the population median* is *$200 billion. We do not know the actual value of the population median. We only know that it is not too far away from the hypothesized median.*

In fact, with this sample size of $n = 6$, we can only conclude that we are 95% confident the population median is between $104 billion and $920 billion.

Note: The conclusion of this test is different from that of the Wilcoxon test (Example 5.7). This starkly illustrates why we cannot *accept* null hypotheses. The ability to reject a false null hypothesis is the power of the test. Here, we showed that our binomial test is less powerful than the Wilcoxon test.

accept

power

This is not too surprising. This binomial test only required that the measurements were continuous. The Wilcoxon test required symmetry. Also, this test only counted the number of values less than the proposed median. The Wilcoxon test incorporated the ranks of the values.

Tests that make more assumptions and use more information tend to be more powerful. However, one needs to test the viability of the assumptions. Is the distribution of external debt really symmetric? If so, the Wilcoxon is the right test. If not, it is not.

Non-parametric tests do not assume the *specific* 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.

non-parametric tests

I will leave it as an exercise for you to investigate the effect of different distributions on the applicability of the Wilcoxon tests.

5.6: Non-Parametric Means Tests II*

At some point, it may be necessary to estimate the mean of a population that violates both of the two assumptions. It is not Normal. It is not symmetric. What can you do? You can simulate means from the population using the sample and a process called non-parametric bootstrapping.

The difference between the parametric bootstrap of Section 1.4 (page 16) and the non-parametric bootstrap is that the parametric bootstrap draws its random sample from a distribution, whereas the non-parametric bootstrap draws its random sample from the data itself.

The non-parametric bootstrap has the advantage of being usable even when you do not know the distribution of the population. It has the disadvantage of requiring the sample to be representative of the population *and* the disadvantage of tending to reject at a different rate than the selected α level. When testing means, it tends to reject at a higher rate than α, and its confidence interval tends to be too narrow.

Review the three steps of a Monte Carlo experiment given in Section 1.4 (page 16). For the non-parametric bootstrap, the random samples are taken from the given data and the "test statistic" is the sample mean.

Example 5.10: Let us return to the data and hypothesis of Example 5.7. The original hypothesis was that the *average* external debt was $200 billion. However, the Wilcoxon test requires that the data come from a symmetric distribution in order to draw conclusions about the population mean. A histogram of the data suggest that it is skewed right. The Hildebrand rule concurs. Thus, there is evidence that the underlying distribution of external debts is skewed left (positive). As such, the Wilcoxon test may not be appropriate.

Since we wish to draw conclusions about the population *mean*, we cannot use the median test. Thus, we can use bootstrapping. Since we do not know the population distribution of external debt, we cannot use the parametric bootstrap. We use the non-parametric bootstrap.

The code for the non-parametric bootstrap in this case is

```
theData = c(920,216,347,548,334,104)
n = length(theData)

B = 1e4
m = numeric()
set.seed(370)

for( b in 1:B ) {
 thisSample = sample(theData,n,replace=TRUE)
 m[b] = mean(thisSample)
}

quantile(m,c(0.025,0.975))
```

Notice how it is similar to the Monte Carlo code from Section 1.4.

The above code produces the following output:

```
    2.5%    97.5%
220.1667 638.5000
```

From this, we can conclude that we are 95% confident that the true mean external debt in the world is between \$220.2 billion and \$638.5 billion. Since the hypothesized mean of \$200 billion is outside this interval, we would reject the hypothesis and conclude that the average external debt in the world is not \$200 billion.

Warning: *This method produces confidence intervals that are of the wrong size, but are close. Drawing black-and-white conclusions from this is dangerous when the hypothesized mean is close to either endpoint, such as here. This effect is especially strong when the sample size is small, also such as here.*

In other words, the estimates are poor, but better than nothing.

5.7: Further Examples

To further illustrate some of these processes, this section provides several additional examples.

Example 5.11: The `HeartOfTheValleyTriathalon` dataset contains a sample of the intermediate and the finishing times for participants in the

May 26, 2014, Heart of the Valley Triathlon held in Corvallis, Oregon. One of the racers hypothesized that her time of 1:20:50.15 was less than the average time for the entire group. Test her hypothesis.

Solution: Her (research) hypothesis is $\mu > 1:20:50.15$. In seconds, this is $\mu > 4850.15$. As this does not contain the "equals" position, it is also the alternative hypothesis.

As we are testing the average racing time for a single population, we would like to use the one-sample t-test. However, it requires that the measurements come from a Normally-distributed population. To test this, I will use the Shapiro-Wilk test. According to this test, the data are not from a Normally distributed distribution ($p = 0.03092$). Thus, we cannot use the one-sample t-test.

We can use the Wilcoxon test, but that only deals with the mean if the data are from a symmetric distribution. According to the Hildebrand rule, there is no evidence of the distribution being skewed. Thus, we can use the Wilcoxon test.

According to the Wilcoxon test, we have strong evidence that the mean completion time for the race is greater than 1:20:50.12 ($p = 0.007425$). A 95% confidence interval for the mean completion time is from 1:21:26:53 to 1:26:45.55.

The following is the code used for this analysis:

```
tri = read.csv("http://rfs.kvasaheim.com/data/
   HeartOfTheValleyTriathalon.csv")
attach(tri)

time = TOTALH*3600 + TOTALM*60 + TOTALS

shapiro.test(time)

hildebrand.rule(time)

hypAvg=60*60+60*20+50.15
wilcox.test(time, mu=hypAvg, alternative="greater")
wilcox.test(time, conf.int=TRUE)
```

◇

Example 5.12: The `football1` dataset contains a sample of the results from NCAA football games in the SEC and the Big 12 from 2009. An associate hypothesized that the mean number of points scored in NCAA football games in 2009 is just three touchdowns (21 points). Test this hypothesis.

Solution: The research hypothesis is $\mu = 21$. Since this contains the equals position, the alternative hypothesis is $\mu \neq 21$.

As we are testing the average number of points scored for a single population (all NCAA teams), we would like to use the one-sample t-test. However, it requires that the measurements come from a Normally-distributed population. To test this, I will use the Shapiro-Wilk test. According to this test, the data are not from such a population ($p = 0.001$). Thus, we cannot use the one-sample t-test.

We can use the Wilcoxon test, but that only deals with the mean if the data are from a symmetric distribution. According to the Hildebrand rule, there is no evidence of the distribution being skewed. Thus, we can use the Wilcoxon test.

According to the Wilcoxon test, we have strong evidence that the mean number of points scored in NCAA football games in 2009 is not 21 ($p \ll 0.0001$). A 95% confidence interval for the mean number of points scored is from 27.5 to 31.0.

The following is the code used for this analysis:

```
fb = read.csv("http://rfs.kvasaheim.com/data/football1.csv"
    )
attach(fb)

shapiro.test(score)

hildebrand.rule(score)

wilcox.test(score, mu=21)
wilcox.test(score, conf.int=TRUE)
```

◇

Warning: *An understood assumption is that the sample is representative of the target population. Without that assumption being true, all conclusions are suspect. This is the reason simple random sampling (SRS) is so important. When using SRS, the sample is (on average) representative of the population.*

In this case, the sample was not drawn using SRS. Without SRS, one must now provide sufficient evidence that the sample is representative of the population. Without that evidence, the results mean nothing.

Example 5.13: The `clf` contains a sample of production ratios for several states of the world. The data collector hypothesized that the mean production ratio in the world is greater than 50. Test this hypothesis.

Solution: The research hypothesis is $\mu > 50$. Since this does not contain the equals position, it is the alternative hypothesis.

As we are testing the average production ratio for a single population (all states in the world), we would like to use the one-sample t-test. However, it requires that the measurements come from a Normally-distributed population. To test this, I will use the Shapiro-Wilk test. According to this test, the data are not from such a population ($p \ll 0.0001$). Thus, we cannot use the one-sample t-test.

We can use the Wilcoxon test, but that only deals with the mean if the data are from a symmetric distribution. According to the Hildebrand rule, this data comes from a positively-skewed distribution. Thus, we should not use the Wilcoxon test. We can, however, use non-parametric bootstrapping.

According to the non-parametric bootstrap test, we have strong evidence that the mean production ratio in the world is greater than 50 ($p = 0.0177$). A central 95% confidence interval for the mean production ratio is from 51.4 to 100.8.

The following is the code used for this analysis:

```
pr = read.csv("http://rfs.kvasaheim.com/data/clf.csv")
attach(pr)

shapiro.test(productionRatio)
hildebrand.rule(productionRatio)

## Non-Parametric Bootstrapping
 theData = productionRatio
 n = length(theData)
 B = 1e4
 m = numeric()
```

```
for(b in 1:B) {
 thisSample = sample(theData,n,replace=TRUE)
 m[b] = mean(thisSample)
}

mean(m<50)                    ## p-value
quantile(m,c(0.025,0.975))    ## confidence interval
```

◇

Example 5.14: The `studentHeight` contains a sample of heights for several ($n = 20$) students, both male and female. I hypothesized that the mean student height is greater than 170cm. Test this hypothesis.

Solution: The research hypothesis is $\mu > 170$. Since this does not contain the equals position, it is the alternative hypothesis.

As we are testing the average height for a single population (students), we would like to use the one-sample t-test. However, it requires that the measurements come from a Normally-distributed population. To test this, I will use the Shapiro-Wilk test. According to this test, the data are from such a population ($p = 0.8094$). Thus, we should use the one-sample t-test.

According to the one-sample t-test, we have evidence that the mean height of students is not greater than 170cm ($p = 0.7543$). A central 95% confidence interval for the mean student height is from 160.4 to 174.8cm.

```
sh = read.csv("http://rfs.kvasaheim.com/data/studentHeight.
   csv")
attach(sh)

shapiro.test(height)

t.test(height, mu=170, alternative="greater")
t.test(height)
```

◇

5.8: Conclusion

In this chapter, you have learned how to perform some tests of center (means and medians) regarding a single population. You have also examined two classes of tests: parametric (assumes distribution of your data) and non-parametric (does not assume a specific distribution for your data).

Non-parametric tests are useful if your data has an obviously non-Normal distribution or if the sample size is small. However, the weakness of all non-parametric tests is that they tend to have lower power than the parametric tests. As such, when the parametric assumptions are not met, one should run the non-parametric test. If the non-parametric test fails to reject the null hypothesis, data transformation should be attempted (*v.i.* Section 14).

Frequently, we wish to compare the centers of two populations, either independent populations or repeated measures on a single population. For the latter, the tests can be handled using the methods of this chapter. For the former, we use the two-sample t-test, the Mann-Whitney test, or a non-parametric bootstrapping test. The one we select depends on characteristics of the distributions that gave us the data.

5.9: End of Chapter Materials

5.9.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Statistics:

ks.test(x,y) This performs a Kolmogorov-Smirnov test, which determines if the two provided samples (`x`,`y`) come from the same distribution. This test is often used to determine if a sample is Normally distributed, in which case `x` will be the data and `y` will be `pnorm`.

length(x) Returns the number of values in the vector x, if x is a vector. Returns the length of the character string x, if x is a character string.

t.test(·) This function preforms a t-test of the provided data. The four types of t-tests can be specified as

`t.test(x, mu=)`	1-sample t-test
`t.test(x,y)`	2-sample t-test, unequal variances
`t.test(x,y, var.equal=TRUE)`	2-sample t-test, equal variances
`t.test(x,y, paired=TRUE)`	2-sample, paired t-test

wilcox.test(x) Performs a one- or two-sample Wilcoxon test (known as the Mann-Whitney test when comparing two samples).

Probability:

dnorm(x) Returns the likelihood (or *d*ensity) for an x-value according to the specified Normal distribution: `dnorm(1,m=3,s=6)` returns the value of the pdf at 1 corresponding to the $\mathcal{N}(\mu = 3, \sigma = 6)$ distribution, 0.0628972.

pnorm(x) Returns the cumulative probability for an x-value according to the specified Normal distribution: `pnorm(1.96,m=0,s=1)` returns the value of the CDF at 1.96 corresponding to the $\mathcal{N}(\mu = 0, \sigma = 1)$ distribution, 0.975.

qnorm(p) Returns the value of x corresponding to the p-value provided according to the specified Normal distribution: `qnorm(0.95,m=5,s=1)` returns the x-value such that $\mathbb{P}[X < x] = 0.95$, where X is distributed as $\mathcal{N}(\mu = 5, \sigma = 1)$.

rexp(n) Returns n *r*andom numbers from the specified Exponential distribution: `rexp(100,r=3)` gives 100 random numbers drawn from an $\mathcal{E}xp(\lambda = 3)$ distribution.

rnorm(n) Returns n *r*andom numbers from the specified Normal distribution: `rnorm(100,m=3,s=6)` gives 100 random numbers drawn from a $\mathcal{N}(\mu = 3, \sigma = 6)$ distribution.

Graphing:

abline() Draws a line on a currently open plot: `abline(h=3)` draws a horizontal line at $y = 3$; `abline(v=6)` draws a horizontal line at $x = 6$; `abline(a=3,b=1)` draws a line with intercept $a = 3$ and slope $b = 1$.

hist(x) Calculates (and draws) a histogram corresponding to the vector X.

Mathematics:

abs(x) Returns the magnitude of the argument: `abs(-3)` = 3.

sqrt(x) Returns the positive square root of the argument: `sqrt(9)` = 3.

Programming:

attach(d) Connects the dataset d to the current working environment so that one does not need to use '$' notation to access its variables and values. This is rather handy if you are only using one dataset in your analysis. If, however, you are using several, then it becomes rather easy to forget that the value you are requesting may not be the one you actually want. As such, use this with care.

for(){} Creates a loop in your script, allowing statements contained within the braces to be performed more than once. This statement is invaluable when performing Monte Carlo analysis.

function(){} Creates a user-defined function, whose parameters (required or options) are contained in the parentheses immediately following `function`, and whose statements are contained in the braces following `function`.

names(d) Returns the variables contained in the *d* variable, which can be a dataframe, a list, or a matrix/array.

read.csv(f) Imports a dataset from *f*, the specified file location. If the first row (header) of the dataset contains variable names, you may specify the optional parameter `header=TRUE` in the function call; otherwise, you must specify `header=FALSE`.

5.9.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 5 folder. For each of the problems using `R`, please save the associated `R` script in the chapter folder as `ext0x.R`, where `x` is the problem number.

Summary:

1. Why should a research *not* use a level of analysis that is lower than the unit of analysis?
2. Select an article from Section 5.9.3. Provide the unit of analysis, the variables used, and the level of analysis for each variable in that article.
3. What is the major drawback to the z-test? Why does one not just use the sample variance in place of the population variance?
4. What is the name of the distribution of the t-test test statistic? How many parameters does it take? What do they (does it) represent?
5. What is the major difference between the t-test and the Wilcoxon test? When would you use one over the other? In general, why is the t-test preferred?
6. Explain why the test statistic for the Binomial test has the Binomial distribution of $T \sim Bin(n, 0.500)$.
7. Using the appropriate formula, calculate the confidence intervals for Examples 5.3 and 5.4 by hand.

Data:

8. According to the `patrickHenry` datafile, what is the average SAT Mathematics score (`math`)? Make sure to include the 95% confidence interval. A research hypothesizes that the average SAT Mathematics score at Patrick Henry College is 600. Do the data support this contention? Produce a box-and-whiskers plot to illustrate your findings.
9. According to the `patrickHenry` datafile, what is the average SAT Verbal score (`reading`)? Make sure to include the 95% confidence interval. A research hypothesizes that the average SAT Verbal score at Patrick Henry College is 600. Do the data support this contention? Produce a box-and-whiskers plot to illustrate your findings.

10. According to the `patrickHenry` datafile, what is the average difference between SAT Mathematics and SAT Verbal scores? A researcher hypothesizes that the average SAT Verbal score is lower than the average SAT Mathematics score. Do the data support this contention? Produce a box-and-whiskers plot to illustrate your findings.

11. According to the `patrickHenry` datafile, what is the average GPA? A researcher hypothesizes that the average GPA is less than 2.0. Do the data support this contention? Produce a box-and-whiskers plot to illustrate your findings.

12. According to the `positioningtubes` datafile, does the data support the contention that the average diameter of the positioning tubes is 11.99? Produce a box-and-whiskers plot to illustrate your findings.

Monte Carlo:

13. Create a random dataset (of size 500) from an Exponential distribution, with mean 4 (rate, $\lambda = 0.25$). Use a seed value of 3. Test the null hypothesis that this population has a mean of 4. Save this script in your chapter folder as `ext01.R`.

    a) If you wanted to use the parametric test, is the sample size large enough?

    b) Which test should you use?

    c) Does that test reject the null hypothesis?

    d) What is the appropriate conclusion based on the test results?

    e) Knowing what you know about the *actual* variable, is the population mean 4?

14. Create a random dataset (of size 10) from the Normal distribution, with mean 4 and standard deviation 1. Use a seed value of 3. Test the null hypothesis that this distribution has a mean of 8. Save this script in your chapter folder as `ext02.R`.

    a) If you wanted to use the parametric test, is the sample size large enough?

    b) Which test should you use?

    c) Does that test reject the null hypothesis?

    d) What is the appropriate conclusion based on the test results?

    e) Knowing what you know about the actual variable, is the population equal to 8?

5.9.3 Applied Research

This section offers some applied research works that are connected with the topics in this chapter.

- Charles Bérubé and Pierre Mohnen. (2009) "Are Firms That Receive R&D Subsidies More Innovative?" *The Canadian Journal of Economics / Revue canadienne d'Economique*. **42**(1):206–25.
- Matthew S. Bothner, Edward Bishop Smith, and Harrison C. White. (2010) "A Model of Robust Positions in Social Networks." *American Journal of Sociology*. **116**(3): 943–92.
- Kevin Denny and Orla Doyle. (2009) "Does Voting History Matter? Analysing Persistence in Turnout." *American Journal of Political Science*. **53**(1): 17–35.
- Ole J. Forsberg. (2007) *Terrorism and Nationalism: Theories, causes, and causers*. Saarbr ucken, Germany: VDM Verlag.
- Esther Godson and John D. Stednick. (2010) "Modeling Post-Fire Soil Erosion." *Fire Management Today* **70**(3): 32–36.
- Matthijs Kalmijn. (2010) "Consequences of Racial Intermarriage for Children's Social Integration." *Sociological Perspectives*. **53**(2): 271–86.
- John R. Lott, Jr. (2009) "Non-Voted Ballots, the Cost of Voting, and Race." *Public Choice*. **138**(1/2):171–97.
- Tonya L. Putnam. (2009) "Courts without Borders: Domestic Sources of U.S. Extraterritoriality in the Regulatory Sphere" *International Organization*. **63**(3): 459–90.

5.9.4 References and Additional Readings

This section provides a list of statistical works. Those works cited in the chapter are here. Also here are works that complement the chapter's topics.

- Lee Bain and Max Englehardt. (1992) *Introduction to Probability and Mathematical Statistics*, 2nd edn. Brooks/Cole: Belmont, CA.
- William Navidi. (2006) *Statistics for Engineering and Scientists*, 2nd edn. McGraw-Hill: New York.
- Samuel S. Shapiro and Martin B. Wilk. (1965) "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* **52**(3–4): 591–611.