Chapter 7

One-Sample Hypothesis Tests

Chapter Objectives

- Explain the meanings of the terms hypothesis, test statistic, level of significance, p-value, statistical significance.
- Carry out a one-sample t test for a population mean using the rejection region and p-value approaches.
- Recognize data snooping and state why it can lead to incorrect conclusions in hypothesis testing.
- Use confidence intervals to test hypotheses.
- Distinguish between Type I and II errors.
- State the relationship between the level of significance and the probability of a Type I error.
- Differentiate between statistical significance and practical importance.
- Carry out a one-sample sign test for a population median.
- Decide which test (the t test or the sign test) is more appropriate for a given set of data.
- Carry out a one-sample z test for a population proportion.

Key Takeaways

- Hypothesis tests involve using data to decide between two hypotheses about one or more population parameters.
- An alternative hypothesis can be one-sided or two-sided. If a one-sided alternative is tested, its direction should be stated prior to examining the data.
- The level of significance determines how strong the evidence against the null hypothesis needs to be in order to reject that hypothesis. A result is called statistically significant, at a given level of significance, if the null is rejected.
- The observed value of a test statistic is compared to the sampling distribution that the statistic would follow if the null hypothesis was true. If, based on that comparison, the test statistic is unlikely to have occurred by chance, the null is rejected. A decision rule dictates when the test statistic is deemed unlikely to have occurred by chance.
- Two approaches to hypothesis testing are the rejection region approach and the p-value approach. The two approaches always lead to the same conclusion.
- The p-value is the probability that just by chance you'd get a result that's as incompatible with the null hypothesis as the result you got.
- The one-sample t test is a parametric test for a population mean that requires either that the sample is from a normal population or the sample size is large. We can assess normality by graphing the data. A log transformation can make right skewed data more normal prior to conducting a t test.
- Data snooping means choosing the direction for the alternative hypothesis, after looking at the data, to fit what you already see in the data. It can lead to incorrect conclusions in hypothesis testing via artificially small p-values.

- Confidence intervals can be used to carry out hypothesis tests. If the null-hypothesized value isn't in the interval, the null hypothesis is rejected.
- The two types of errors in hypothesis testing are Type I errors (false positives) and Type II errors (false negatives).
- Statistical significance doesn't necessarily imply practical importance.
- The one-sample sign test is a nonparametric test for a population median that does not require a normality assumption or a large sample size.
- The one-sample z test for a population proportion is used when the sample size is large.

7.1 Introduction to Hypothesis Testing

7.1.1 Null and Alternative Hypotheses

A statistical ***hypothesis test*** is a procedure for *deciding* between two competing claims, or ***hypotheses***, about the values of the parameters of one or more populations, such as their means. A hypothesis test could be used, for example, to decide if the mean contaminant level μ at a remediation site is below a regulatory standard, or to decide if there's any difference between the means at the control and impact sites in an impact assessment study. In every hypothesis test, one of the two claims, called the ***alternative hypothesis***, will be of primary interest, and the decision between this claim and the opposing claim, called the ***null hypothesis***, will be based on whether or not data provide sufficient support for the alternative hypothesis.

Example 7.1: Null and Alternative Hypotheses

The diameters of trees are often used as a proxy for their ages when designating a forest as *old growth*. For example, the U.S. Forest Service will classify a Douglas-fir stand in Oregon's Willamette National Forest as old growth only if its mean tree stem diameter μ is 32 inches or larger.

Suppose that a logging company is required to provide convincing evidence that a forest isn't old growth before the government will allow the company to log the forest. The company successfully argues that rather than having to measure *every* tree in the forest to show that μ is less than 32 inches, it be allowed make its case using the mean diameter $\bar{X}$ in a random *sample* of trees.

In this example, the *alternative hypothesis* is the claim the logging company seeks to validate – that the true mean diameter μ is less than 32 inches. We'll see how the hypothesis test is carried out in later examples.

A hypothesis test must take into account chance variation (sampling variation) in the value of a statistic, such as $\bar{X}$. In this chapter, we'll look at two hypothesis test procedures based on a *single* sample from a population:

1. The one-sample t test
2. The one-sample sign test

The first one is a test for a population mean μ, and requires either that the population is normal or the sample size is large. The second one is a test for a population median $\tilde{\mu}$ and doesn't have the normality or large sample size requirement.

We denote the null hypothesis by $\boldsymbol{H_0}$ and the alternative by $\boldsymbol{H_a}$. The null hypothesis often reflects the status quo and the alternative a departure therefrom. For example, after a remediation project, the null hypothesis might be that a regulatory standard still isn't met and the alternative that it is. In an impact

assessment study, the null hypothesis might be that the impact site still doesn't differ from the control site and the alternative that it does.

When deciding which hypothesis should be designated H_0 and which H_a, the following may help.

Alternative Hypothesis H_a	**Null Hypothesis H_0**
The hypothesis that's of primary interest.	The hypothesis that's of little or no interest.
The hypothesis we seek evidence *for* in the data.	The hypothesis we seek evidence *against* in the data.
The hypothesis that "there's a difference" when comparing two or more population means.	The hypothesis that "there's no difference" when comparing two or more population means.
The hypothesis that "there's an effect" when conducting an experiment or impact assessment study.	The hypothesis that "there's no effect" when conducting an experiment or impact assessment study.

Example 7.2: Null and Alternative Hypotheses

In Example 7.1, the logging company is seeking evidence for their claim that the true (unknown) population mean tree diameter μ is less than 32 inches. This is the alternative hypothesis. The null hypothesis, therefore, is that μ is 32 inches or greater. These are stated as

$$H_0 : \mu \geq 32 \tag{7.1}$$
$$H_a : \mu < 32. \tag{7.2}$$

The null and alternative hypotheses must be contrary to each other, in that it should be impossible for them to both be true. A "borderline" value separating the two hypotheses, such as $\mu = 32$ in the last example, belongs in the null hypothesis.

An alternative hypothesis is called ***one-sided*** if it specifies a particular direction (either "$>$" or "$<$"), as the one in Example 7.2 does. The direction of a one-sided alternative hypothesis is intended to reflect what we suspect, *going into the study*, to be true. If we don't have a particular direction in mind going into the study, we use a ***two-sided*** alternative hypothesis, which doesn't specify a direction (only "$\neq$").

7.1.2 Test Statistic and Decision Rule

At the conclusion of any hypothesis test, we'll either *reject* or *fail to reject* the null hypothesis. The decision will be based on the observed value of a ***test statistic*** and whether it provides compelling evidence against the null. If it does, we reject the null. Otherwise, we fail to reject it. The decision as to whether the evidence is compelling or not is based on a ***decision rule*** that dictates whether the observed value of the test statistic is among the values that would be unlikely to occur just by chance if the null hypothesis was true.

Note: "Failing to reject" the null hypothesis only means there was insufficient evidence to reject it, *not* that there was sufficient evidence to accept it. So "failing to reject" the null isn't the same as "accepting"

it. Hypothesis tests are able to establish (with reasonable certainty) that an alternative hypothesis is true (by rejecting the null), but they're not able to establish that a null hypothesis is true.

Example 7.3: Test Statistic and Decision Rule

A test of the hypotheses (7.1) and (7.2) of Example 7.2 would involve taking a random sample of, say, $n = 100$ trees and calculating the sample mean diameter $\bar{X}$. Then, because $\bar{X}$ is an estimate of the true mean diameter μ, evidence that μ is less than 32 inches would come in the form of an $\bar{X}$ value below 32.

But the value of $\bar{X}$ is subject to sampling variation, so it could fall below 32 just by chance even if μ was equal to (or greater than) 32. Thus, even if $\bar{X}$ ended up below 32, the logging company would still have to convince the government that it didn't get that result by chance variation (sampling variation).

Chance becomes a less plausible explanation the farther $\bar{X}$ falls below 32, and if it falls far enough below, chance can be ruled out altogether. A *decision rule* is used to state how far below 32 $\bar{X}$ needs to be before chance is no longer be a viable explanation. If $\bar{X}$ fell more than about 1.65 standard errors below 32, chance would be a far-fetched explanation because such an outcome would occur by chance no more than 5% of the time (according to the Central Limit Theorem, which says the sampling distribution of $\bar{X}$ is normal). So, letting $\sigma_{\bar{X}}$ denote the standard error of $\bar{X}$, if the *test statistic*

$$Z = \frac{\bar{X} - 32}{\sigma_{\bar{X}}} \tag{7.3}$$

ended up being less than -1.65, we could reasonably conclude that it wasn't just the result of chance variation but rather the result of μ being less than 32.

A sensible *decision rule*, therefore, would be

Reject H_0 if $Z < -1.65$.
Fail to reject H_0 if $Z \geq -1.65$.

The hypothesis test procedure just described is called a ***one-sample z test for μ***. Other implementations of the *z test* would use the same test statistic, but with a different value in place of 32, depending on the hypotheses being tested, and a decision rule that might use a different value in place of -1.65. The *z test* requires knowing the value of the standard error $\sigma_{\bar{X}}$, though, which in turn requires knowing the population standard deviation σ (since $\sigma_{\bar{X}} = \sigma/\sqrt{n}$). In practice, knowing the population standard deviation would be rare. Instead, in practice, we use the test statistic

$$t = \frac{\bar{X} - 32}{S_{\bar{X}}}.$$

where

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

is the estimated standard error. The test is then called the *one-sample t test*, the details of which are covered in Section 7.2.

Note: Often a set of hypotheses such as

$$H_0 : \mu \geq 32 \tag{7.4}$$
$$H_a : \mu < 32 \tag{7.5}$$

in which the alternative is one-sided, is stated as

$$H_0 : \mu = 32 \tag{7.6}$$
$$H_a : \mu < 32 \tag{7.7}$$

with the implication being that (7.6) really means (7.4). The decision rule for the two sets of hypotheses will be the same because if the evidence provided by a given test statistic value is strong enough to reject (7.6) in favor of (7.7), then it's also strong enough to reject (7.4) in favor of (7.5). In practice, either set of hypotheses would be legitimate.

7.1.3 Two Approaches to Hypothesis Testing

There are two ways we can form the decision rule for a hypothesis test, the ***rejection region approach*** and the ***p-value approach***. Both are based on an assessment of whether the observed test statistic value is one that would rarely occur by chance if the null hypothesis was true and, if so, rejecting the null. As we'll see, the rarity of the observed test statistic value under the null hypothesis is assessed by comparing that value to the sampling distribution that the statistic would follow *if* the null was true.

Rationale of a Hypothesis Test

1. A test statistic is selected for it's capacity to discern between the null and alternative hypotheses.
2. The sampling distribution that the statistic would follow if the null hypothesis was true is used to identify values of the statistic that would be unlikely to occur by chance if the null was true.
3. If the observed test statistic value is among those that would be unlikely to occur just by chance, the null hypothesis is rejected.

The rejection region approach: This approach involves choosing a ***rejection region***, which is the set of all test statistic values that are considered so contradictory to the null hypothesis (and consistent with the alternative) that they would rarely occur by chance if the null was true. This is the approach that was used in Example 7.3. Here's the decision rule.

Decision Rule for Rejection Region Approach:

Reject H_0 if the observed test statistic value falls in the rejection region.
Fail to reject H_0 if it doesn't fall in the rejection region.

The p-value approach: This approach involves determining the ***p-value***, which is the probability that the test statistic value would end up being as contradictory to the null hypothesis (and consistent with the alternative) as the observed value *if* the null hypothesis was true. A small p-value means the observed test statistic value would rarely occur by chance if the null hypothesis was true, so small p-values provide compelling evidence against the null and the smaller the p-value, the more compelling the evidence. Here's the decision rule.

Decision Rule for P-Value Approach:

Reject H_0 if the p-value is less than a pre-specified level (such as 0.05).
Fail to reject H_0 if it's not less than that level.

Example 7.4: Rejection Region and P-Value Approaches

The test statistic Z in Example 7.3 will lie below zero whenever the sample mean tree diameter $\bar{X}$ lies below 32, so negative Z values count as evidence against the null hypothesis and in favor of the alternative.

Using the *rejection region approach*, the decision rule given in that example was

Reject H_0 if $Z < -1.65$
Fail to reject H_0 if $Z \geq -1.65$

which is another way of saying we should reject the null if $\bar{X}$ falls more than 1.65 standard errors below 32. As an example, if a random sample of trees resulted in, say, $Z = -2.1$, we'd reject the null hypothesis and conclude that μ is smaller than 32 inches.

Using the *p-value approach*, with pre-specified level 0.05, the decision rule would be

Reject H_0 if p-value < 0.05
Fail to reject H_0 if p-value ≥ 0.05

If the test statistic ended up being $Z = -2.1$, the p-value would be 0.018, which is the probability of getting a Z value as far below zero as -2.1 just by chance (from a standard normal table). Here again, we'd reject the null and conclude that μ is less than 32.

7.1.4 The Level of Significance α

Regardless of whether the rejection region or p-value approach is used, a *level of significance* must be chosen prior to deciding between the two hypotheses. The ***level of significance***, denoted $\boldsymbol{\alpha}$, is a value that determines how strong the evidence against the null hypothesis needs to be before we're willing to reject that hypothesis. Smaller α values require stronger evidence. The most common choices for α are 0.01, 0.05, and 0.10.

Role of α in the rejection region approach: In the rejection region approach, the boundary value for the rejection region used in the decision rule will depend on the value chosen for α. Thus α determines how contradictory to the null hypothesis the test statistic needs to be in order to lie in the rejection region. We'll see that a smaller value of α requires a more null-contradictory test statistic value.

Role of α in the p-value approach: In the p-value approach, α is the threshold level to which the p-value is compared in the decision rule. Thus α determines how strong the evidence needs to be in order to reject the null hypothesis. A smaller α requires a smaller p-value (stronger evidence).

For a given choice of α, if the null hypothesis is rejected, we say the result is ***statistically significant*** at the ***α level***. For example, using $\alpha = 0.05$, we'd say the result is statistically significant at the 5% level. A statistically significant result is one for which chance alone cannot adequately explain the observed evidence supporting the alternative hypothesis.

7.1.5 Steps in Carrying Out a Hypothesis Test

There are a multitude of specific hypothesis test procedures, each designed to address a particular type of research question under a specific set of conditions (assumptions). Although the details of the procedures

differ, *every* hypothesis test involves the following steps.

Steps in Performing a Hypothesis Test

1. Identify the population parameter(s) of interest.
2. State the null and alternative hypotheses.
3. Determine an appropriate test procedure and verify any assumptions.
4. Choose a level of significance and use it to form the decision rule.
5. Compute the test statistic value.
6. Find the p-value or determine if the test statistic lies in the rejection region.
7. State the conclusion and interpret the result.

7.2 The One-Sample *t* Test for a Population Mean

7.2.1 The One-Sample *t* Test Procedure

The ***one-sample t test*** is a hypothesis test for the (unknown) value of a population mean μ. Suppose we have a random sample from the population, and we want to use the data to test the null hypothesis

$$H_0 : \mu = \mu_0$$

that μ is equal to some ***hypothesized value*** $\boldsymbol{\mu_0}$ versus one of the three alternative hypotheses

1. $H_a : \mu > \mu_0$ (upper-tailed test)
2. $H_a : \mu < \mu_0$ (lower-tailed test)
3. $H_a : \mu \neq \mu_0$ (two-tailed test)

The null-hypothesized value μ_0 should be chosen for its relevance to the study's research question, and the alternative hypothesis should reflect what the study is attempting to substantiate. If the direction specified by the alternative is ">", the test is called an ***upper-tailed*** test, and if the direction is "<", it's ***lower-tailed***. If no direction is specified ("≠"), the test is called ***two-tailed***.

The ***one-sample t test statistic***, denoted $\boldsymbol{t}$, is defined as follows.

One-Sample *t* Test Statistic:

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}, \tag{7.8}$$

where

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}.$$

Note that because $\bar{X}$ is an estimator of the true (unknown) population mean μ, if the null hypothesis was true, and μ equal to the null-hypothesized value μ_0, we'd expect $\bar{X}$ to equal μ_0 approximately, in which case t would be near zero. Any discrepancy between t and zero would be due purely to chance (sampling variation). On the other hand, if the alternative hypothesis was true, and μ different from μ_0 in the direction specified by that hypothesis, we'd expect $\bar{X}$ to differ from μ_0 in that direction too, in which case t would differ from zero in that same direction. Moreover, the denominator of t is an estimate of the standard error of $\bar{X}$. Therefore, we have the following.

Interpretation of the Test Statistic: The value of t measures (approximately) how many standard errors $\bar{X}$ is away from μ_0.

1. *Large positive* values of t provide evidence in favor of $H_a : \mu > \mu_0$.
2. *Large negative* values of t provide evidence in favor of $H_a : \mu < \mu_0$.
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \mu \neq \mu_0$.

To decide whether an observed value of t provides statistically significant evidence to support the alternative hypothesis, we'll determine if it's among the values that would be unlikely to occur just by chance under the null hypothesis. For this, we'll need to know the sampling distribution that t would follow if the null hypothesis was true. We know (Section 6.4 of Chapter 6) that if a sample is from a population whose mean is μ, and either the population is *normal* or n is *large*, then

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} \sim t(n-1) \tag{7.9}$$

(at least approximately), the t distribution with $n-1$ degrees of freedom. Thus, because the test statistic t is obtained by replacing the (unknown) population mean μ in (7.9) by its null-hypothesized value μ_0, we have the following.

Sampling Distribution of t Under H_0: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population whose mean is μ, and either the population is normal or n is large. Then when

$$H_0 : \mu = \mu_0$$

is true,

$$t \sim t(n-1).$$

Thus, if the null hypothesis was true, chances are the test statistic t would fall close to zero, the center of its null distribution. Values of t in the tail of the distribution, in the direction (or directions) specified by the alternative hypothesis, would be unlikely and would therefore cast doubt on the null hypothesis, instead supporting the alternative. How far in the tail t would need to be before we'd reject the null altogether will depend on our choice for the level of significance α. We'll see in a bit that a smaller value of α will require t to be farther out in the tail. For now, though, here's a table summarizing the test procedure for both the rejection region and p-value approaches.

One-Sample t Test for μ

Assumptions: The data $X_1, X_2, \ldots, X_n$ are a random sample from a population and either the population is *normal* or n is large.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic value: $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t distribution with $n-1$ d.f.:	Rejection region = t values such that:*
$H_a : \mu > \mu_0$	to the right of t	$t > t_{\alpha,n-1}$
$H_a : \mu < \mu_0$	to the left of t	$t < -t_{\alpha,n-1}$
$H_a : \mu \neq \mu_0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-1}$ or $t < -t_{\alpha/2,n-1}$

* $t_{\alpha,n-1}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-1$ d.f.

The rejection region and p-value given in the table above are depicted graphically below for each of the three alternative hypotheses. The t *critical value* (or values) used to demarcate the rejection region were discussed in Chapter 6 (Section 6.5) and are obtained from a t distribution table or statistical software. In either case, the level of significance α needs to be chosen first. The p-value is also obtained from the table or using software, and requires observing the test statistic value t first.

1. $H_a : \mu > \mu_0$ (Upper-Tailed Test)

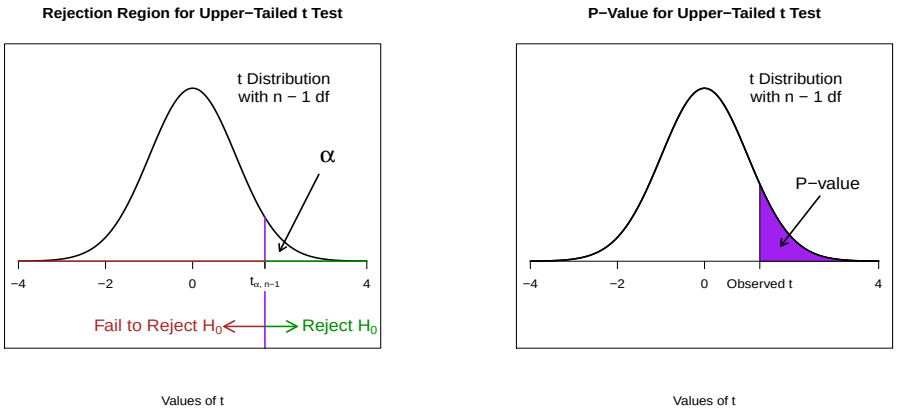


Figure 7.1

2. $H_a : \mu < \mu_0$ (Lower-Tailed Test)

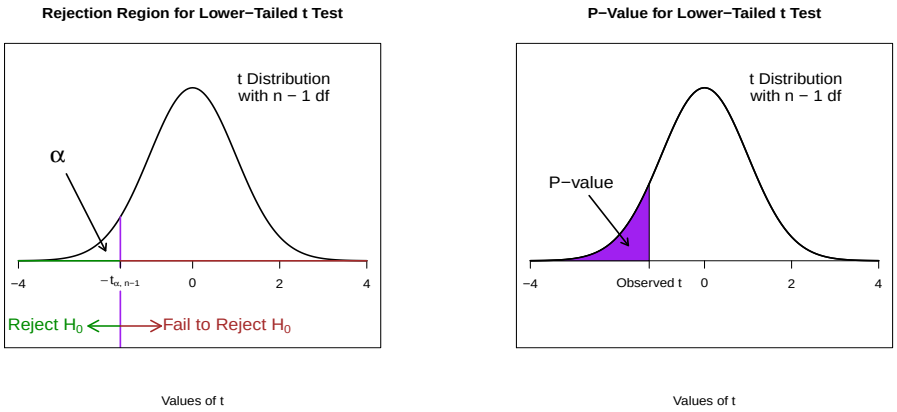


Figure 7.2

3. $H_a : \mu \neq \mu_0$ (Two-Tailed Test)

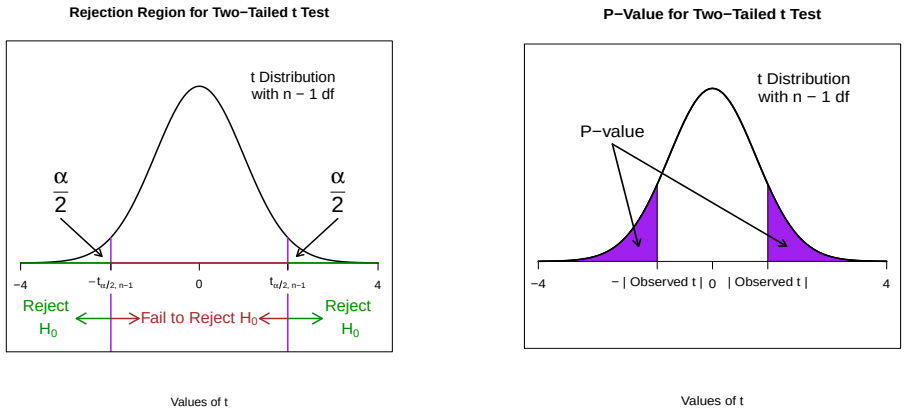


Figure 7.3

Comment: It's worthwhile at this point to look more closely at the role that the level of significance α plays in hypothesis testing.

- For the rejection region approach, α determines the critical value $t_{\alpha,n-1}$ (or $t_{\alpha/2,n-1}$), which demarcates the rejection region. The smaller α is, the farther away from zero the critical value will be, and so the farther away from zero the test statistic t will need to be before we're willing to reject the null hypothesis.

- For the p-value approach, the smaller α is, the smaller the p-value will need to be before we're willing to reject the null hypothesis. And because smaller p-values result from t being farther away from zero, a smaller α again means t will have to be farther away from zero before we're willing to reject the null.

Thus, in either case, *using a smaller α means we require stronger evidence in favor of the alternative hypothesis before we're willing to reject the null.*

7.2.2 Carrying Out the One-Sample *t* Test

We'll now turn to several examples. Examples 7.5 and 7.6 illustrate the rejection region and p-value approaches, respectively, for a one-tailed test. Examples 7.7 and 7.8 illustrate the two approaches for a two-tailed test. In all examples, the seven steps listed in Subsection 7.1.5 are followed.

Example 7.5: One-Sample *t* Test Using the Rejection Region Approach

In previous examples, a logging company was interested in showing that the true mean tree diameter μ is less than 32 inches. The hypotheses are

$$H_0 : \mu \geq 32 \tag{7.10}$$
$$H_a : \mu < 32. \tag{7.11}$$

Equivalently, the hypotheses can be stated as in (7.6) and (7.7).

Suppose that in a random sample of $n = 100$ trees, the sample mean and standard deviation of the diameters are

$$\bar{X} = 30.3$$
$$S = 8.16.$$

Suppose also that a histogram and a normal probability plot indicate that the data can be assumed to be from a normally distributed population, so the one-sample t test is appropriate. In practice, the t test would be appropriate even if the normality assumption wasn't met because n is large.

The decision whether or not to reject the null hypothesis will be based on the observed value of the test statistic

$$t = \frac{\bar{X} - 32}{S_{\bar{X}}}.$$

Using a level of significance $\alpha = 0.05$, the decision rule for the rejection region approach is

Reject H_0 if $t < -t_{0.05,99}$
Fail to reject H_0 if $t \geq -t_{0.05,99}$

From a table of t distribution critical values, using $n - 1 = 99$ degrees of freedom, $t_{0.05,99} = 1.66$, so the decision rule is

Reject H_0 if $t < -1.66$
Fail to reject H_0 if $t \geq -1.66$

The figure below shows the rejection region for the test.

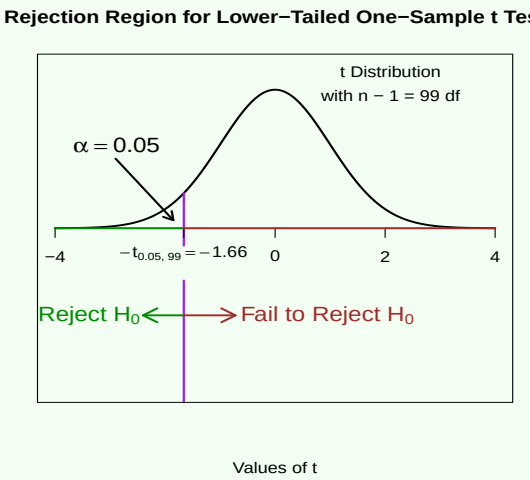


Figure 7.4: The rejection region is demarcated by the t critical value $-t_{0.05,99} = -1.66$.

The (estimated) standard error of $\bar{X}$ is

$$S_{\bar{X}} = \frac{8.16}{\sqrt{100}} = 0.82,$$

so the observed t value is

$$t = \frac{30.3 - 32}{0.82} = -2.08.$$

Therefore we reject the null hypothesis. This says that the observed difference between the sample mean ($\bar{X} = 30.3$) and the null-hypothesized population mean ($\mu_0 = 32$) is *statistically significant*, that is, not likely the result of chance variation (sampling variation). It appears that the logging company's claim that μ is less than 32 inches is correct.

Now we'll repeat the hypothesis test of the last example using the p-value approach.

Example 7.6: One-Sample *t* Test Using the P-Value Approach

For the p-value approach, the hypotheses and test statistic t are the same as before. Again we'll use $\alpha = 0.05$. Because we're performing a lower-tailed test and the sample size is $n = 99$, the p-value is the area to the left of the observed test statistic value, $t = -2.08$, under the $t(99)$ curve. The decision rule is

Reject H_0 if p-value < 0.05
Fail to reject H_0 if p-value ≥ 0.05

The p-value is the probability that we'd get a test statistic value as far below zero as the observed one, $t = -2.08$, if the null hypothesis was true. It's shown in the figure below.

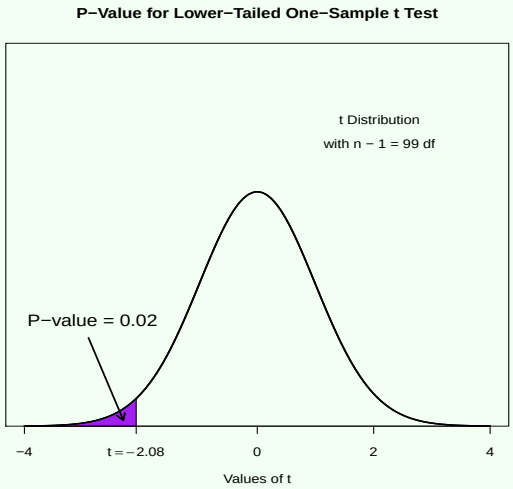


Figure 7.5: The p-value is the area to the left of the observed test statistic value, $t = -2.08$, under the $t(99)$ distribution.

From a table of t distribution tail areas, using $n - 1 = 99$ degrees of freedom, the p-value is found to be 0.0201. This says that we'd end up with a test statistic value as far below zero as -2.08 by chance only 2.01% of the time if the null hypothesis was true, that is, if the population mean tree diameter was 32 inches.

Because the p-value is less than 0.05, we reject the null hypothesis using the p-value approach just as we did using the rejection region approach. Once again, we find that the evidence supporting the hypothesis that the true (unknown) mean diameter μ is less than 32 inches is *statistically significant.*

The next example illustrates a *two-tailed* test (using the rejection region approach).

Example 7.7: One-Sample *t* Test Using the Rejection Region Approach

A laboratory quality assurance study was carried out to look for signs of systematic bias in a lab's method for measuring total organic carbon (TOC), a measure of water quality [**?**]. Certified standard solutions having 50 mg/L TOC were randomly inserted into the lab's normal work stream. The lab analysts were unaware of the presence of these standard solutions.

If there was no systematic bias, a 50 mg/L standard solution would, on average, give a 50 mg/L measurement reading. But if the measurement method was biased, the reading would differ from 50, on average, in the direction of the bias. The relevant hypotheses are

$$H_0 : \mu = 50 \tag{7.12}$$
$$H_a : \mu \neq 50 \tag{7.13}$$

where μ is the true (unknown) mean measurement reading for a standard solution containing a true concentration of 50 mg/L. The null hypothesis says there's no bias. The alternative says there's bias, but doesn't specify a direction for that bias.

Here are the measurement readings for $n = 16$ of the standard solutions inserted into the lab's work stream:

50.3	51.2	50.5	50.2	49.9	50.2	50.3	50.5
49.3	50.0	50.4	50.1	51.0	49.8	50.7	50.6

The sample mean and sample standard deviation are

$$\begin{aligned} \bar{X} &= 50.31 \\ S &= 0.46. \end{aligned}$$

A histogram and normal probability plot of the data are shown below.

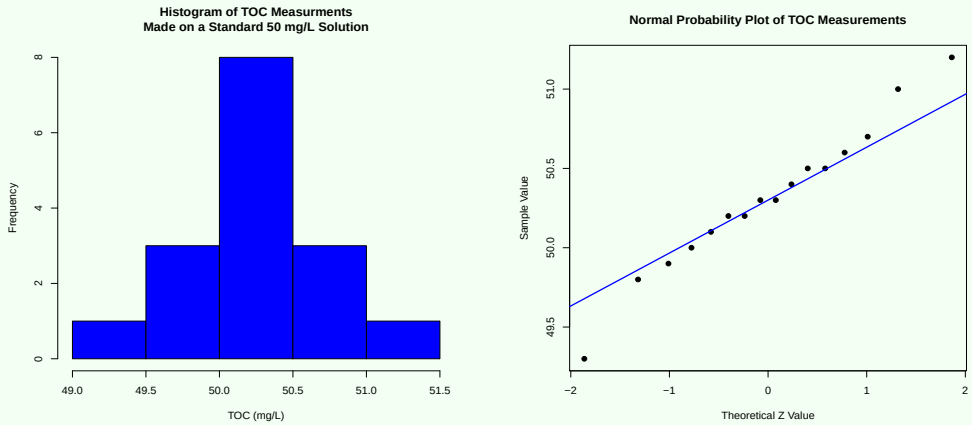


Figure 7.6: Histogram (left) and normal probability plot (right) of $n = 16$ measurements of TOC made on 50 mg/L standard solutions.

The plots indicate that it's reasonable to assume the data are a random sample from a normal distribution, so a one-sample t test is appropriate.

For the rejection region approach, using a level of significance $\alpha = 0.01$, the decision rule is

Reject H_0 if $t < -t_{0.005,15}$ or $t > t_{0.005,15}$
Fail to reject H_0 if $-t_{0.005,15} \leq t \leq -t_{0.005,15}$

From a table of t distribution critical values, using $n - 1 = 15$ degrees of freedom), $t_{0.005,15} = 2.95$, so the decision rule is

Reject H_0 if $t < -2.95$ or $t > 2.95$
Fail to reject H_0 if $-2.95 \leq t \leq 2.95$

The figure below shows this rejection region along with the t distribution with $n - 1 = 15$ degrees of freedom.

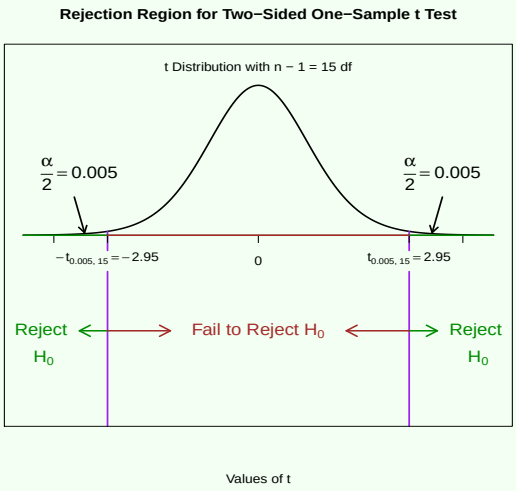


Figure 7.7: The rejection region is demarcated by the t critical values $-t_{0.005,15} = -2.95$ and $t_{0.005,15} = 2.95$.

The (estimated) standard error of $\bar{X}$ is

$$S_{\bar{X}} = \frac{0.46}{\sqrt{16}} = 0.115,$$

The observed value of the test statistic is

$$\begin{aligned} t &= \frac{50.31 - 50}{0.115} \\ &= 2.70. \end{aligned}$$

Because $t = 2.70$ isn't in the rejection region, we fail to reject the null hypothesis. The observed difference between the sample mean $\bar{X} = 50.31$ and the certified concentration 50 mg/L can be explained by chance variation (sampling variation). In other words, there's no convincing evidence that the lab's measurement method is biased.

Now we'll repeat the hypothesis test of the last example using the p-value approach.

Example 7.8: One-Sample *t* Test Using the P-Value Approach

For the p-value approach, the hypotheses and test statistic t are the same as before. Again we'll use $\alpha = 0.01$. Because we're performing a two-tailed test and the sample size is $n = 15$, the p-value is the area to the right of the observed test statistic value, $t = 2.70$, *and* to the left of its negative, $-t = -2.70$, under the $t(14)$ curve. The decision rule is

Reject H_0 if p-value < 0.01
Fail to reject H_0 if p-value ≥ 0.01

Because we're conducting a two-tailed test, the p-value is the sum of the two tail areas to the right of 2.70 and left of -2.70 under the $t(15)$ curve, as shown below.

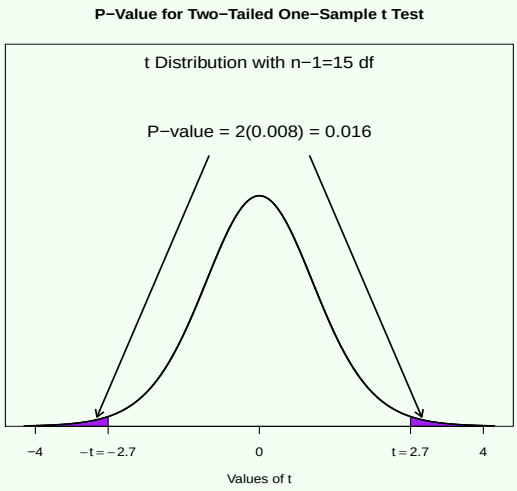


Figure 7.8: The p-value is the sum of the two tail areas to the right of $t = 2.70$ and left of its negative, -2.70, under the $t(15)$ distribution.

From a table of t distribution tail areas, using $n - 1 = 15$ degrees of freedom, the p-value is found to be $2(0.0082) = 0.0164$. This says that if the null hypothesis was true and there was no bias in the lab results, we'd get a test statistic value as far away from zero as 2.70 (in either direction) just by chance about 1.64% of the time.

Because the p-value is greater than 0.01, we fail to reject H_0, and conclude, just as we did in Example 7.7, that there's no convincing evidence for bias. Any bias suggested by the sample isn't statistically significant, and can be explained by chance variation (sampling variation).

7.3 Data Snooping and its Consequences

The term ***data snooping*** refers to looking at the data *before* having decided on a direction for the alternative hypothesis, and then, *after* looking at the data, choosing the direction for the alternative hypothesis that best fits what you already see in the data. Data snooping is considered "cheating" because, as we'll see, it results in an artificially small p-value, which can lead to mistakenly declaring a spurious result to be real.

We avoid data snooping by only using a one-sided alternative hypothesis if we have a specific direction in mind *prior* to looking at the data. This way, the direction of the alternative is like a prediction of what we think the data will show, and if they end up showing it, we can be more sure that we're making the correct decision if we reject the null. If we don't have a specific direction in mind for the alternative hypothesis *prior* to looking at the data, we should perform a two-tailed test.

The next example shows how data snooping can lead to an artificially small p-value. In particular, the p-value ends up being only half as large as it would've ended up being if we hadn't data snooped.

Example 7.9: Data Snooping

In the laboratory quality assurance study of Examples and 7.7 and 7.8, the investigator had no reason to suspect, in advance, that the bias would be one specific direction or the other, so the appropriate alternative hypothesis was the two-sided one given by 7.13.

Suppose, though, that the investigator had data snooped, and decided, *after* noticing that $\bar{X} = 50.31$ is greater than 50, to do a one-sided, upper-tailed test of

$$\begin{aligned} H_0 : \mu &= 50 \\ H_a : \mu &> 50 \end{aligned}$$

using $\alpha = 0.01$ again.

The test statistic would still be $t = 2.70$, but now the p-value would be just the upper tail area of the $t(15)$ distribution, to the right of 2.70, which was found to be 0.0082 in Example 7.8. In this case the p-value would be half of what it was for the two-tailed test, and using $\alpha = 0.01$, the null hypothesis of no bias would be rejected. Here, the investigator would (mistakenly) conclude that the lab's results are biased in the positive direction, and might unnecessarily recommend corrective actions.

The warning to avoid data snooping *doesn't* say that it's inappropriate to use data to form hypotheses about the world. Indeed, much of scientific progress stems from forming hypotheses by looking at data. What it *does* say is that we need to be cautious of using the *same* data set to both form *and* test a hypothesis. The scientific method entails using data to form theoretical hypotheses, but then testing those hypotheses using independent sets of data.

7.4 Equivalence of the Rejection Region and P-value Approaches

In the tree diameters and laboratory quality assurance examples, the rejection region and p-value approaches led to the same conclusions. In general, the two approaches always lead to the same conclusion, regardless of whether a t test or some other type of test is being carried out. This is a consequence of the following fact.

Fact 7.1 For any hypothesis testing procedure, using level of significance α, the test statistic will fall in the rejection region if and only if the p-value is less than α.

To see why, in the context of an upper-tailed one-sample t test using $\alpha = 0.05$, note that the area under the $t(n-1)$ curve to the right of the critical value $t_{0.05,n-1}$ is 0.05. See Fig. 7.1 in Subsection 7.2.1. Thus if the observed test statistic value is greater than $t_{0.05,n-1}$, that is, if it's in the rejection region, the area to its right, which is the p-value, will be less than 0.05. Similarly, if the test statistic is less than $t_{0.05,n-1}$, the area to its right, or p-value, will be greater than 0.05.

In practice the p-value approach is used much more often than the rejection region approach. A reported p-value indicates how strong the evidence against the null hypothesis is (smaller p-values indicate stronger evidence), and the end user can compare that p-value to whatever significance level they choose. We'll see in Subsection 7.6.3 how to choose a significance level.

7.5 Using Confidence Intervals to Test Hypotheses

In Chapter 6, confidence intervals were presented as a method for *estimating* an unknown population parameter such as μ. Hypothesis tests, on the other hand, are a method for *deciding* ("yes" or "no") whether μ (or some other parameter) is different from a particular value μ_0. It turns out that we can use confidence intervals to make this decision too, by checking whether μ_0 is in the interval.

Using Confidence Intervals to Test Hypotheses: We can use a $100(1-\alpha)\%$ one-sample t (or z) confidence interval for an unknown population mean μ to test the hypotheses

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_a : \mu &\neq \mu_0 \end{aligned}$$

by invoking the decision rule

Reject H_0 if the confidence interval doesn't contain μ_0
Fail to reject H_0 if it does contain μ_0

The significance level of the test performed in this manner will be α.

The intuition is that if μ_0 isn't contained in the interval, it's not a plausible value for μ, so the null hypothesis can be rejected. Notice that the alternative hypothesis is two-sided. A one-sided alternative hypothesis can be tested in a similar manner using a one-sided confidence interval.

The next fact says that testing hypotheses using a confidence interval, as just described, leads to the same the conclusion as testing them using a one-sample t (or z) test.

Fact 7.2 In a one-sample t (or z) test of the two-sided hypotheses

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_a : \mu &\neq \mu_0 \end{aligned}$$

using level of significance α, we would reject H_0 if and only if μ_0 lies outside the $100(1-\alpha)\%$ one-sample t (or z) confidence interval.

Example 7.10: Using Confidence Intervals to Test Hypotheses

Using the TOC data from the laboratory quality assurance study of Example 7.7, a 99% one-sample t confidence interval for μ, the true (unknown) mean result when a 50 mg/L standard solution is measured at the lab, is

$$(49.97,\ 50.65)\,.$$

Values inside this interval are plausible for μ. Values outside it are implausible. To use the confidence interval to *decide* between the hypotheses

$$\begin{aligned} H_0 : \mu &= 50 \\ H_a : \mu &\neq 50 \end{aligned}$$

the decision rule is

Reject H_0 if the confidence interval doesn't contain 50
Fail to reject H_0 if it does contain 50

Using this method, we fail to reject the null hypothesis because 50 is within the confidence interval. We arrived at the same conclusion using the t test in Examples 7.7 and 7.8.

The intuition behind Fact 7.2 is that the distance away from μ_0 that $\bar{X}$ would need to be in order for the null hypothesis to be rejected is exactly the same as the margin of error in the confidence interval. To

demonstrate this formally, consider the decision rule in a t test using rejection region approach,

Reject H_0 if $t < -t_{\alpha/2,n-1}$ or $t > t_{\alpha/2,n-1}$
Fail to reject H_0 otherwise

where

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}.$$

Thus we'd reject the null hypothesis if

$$\frac{\bar{X} - \mu_0}{S_{\bar{X}}} < -t_{\alpha/2,n-1} \qquad \text{or} \qquad \frac{\bar{X} - \mu_0}{S_{\bar{X}}} > t_{\alpha/2,n-1}.$$

A little algebra shows that this is equivalent to rejecting the null hypothesis if

$$\mu_0 > \bar{X} + t_{\alpha/2,n-1}S_{\bar{X}} \qquad \text{or} \qquad \mu_0 < \bar{X} - t_{\alpha/2,n-1}S_{\bar{X}}.$$

But this says we reject the null if μ_0 is above the upper endpoint of the $100(1-\alpha)\%$ confidence interval or below the lower endpoint.

The use of confidence intervals to test hypotheses can be generalized beyond just the one-sample t procedures.

Using Confidence Intervals to Test Hypotheses: We can use a $100(1-\alpha)\%$ confidence interval for *any* unknown population parameter to test the hypotheses

$$\begin{aligned} H_0 &: \text{Population parameter} = \text{Claimed value} \\ H_a &: \text{Population parameter} \neq \text{Claimed value} \end{aligned}$$

by invoking the decision rule

Reject H_0 if the confidence interval doesn't contain the claimed value
Fail to reject H_0 if it does contain the claimed value

The significance level of the test performed in this manner will be α.

7.6 Type I and II Errors and Their Probabilities

7.6.1 Introduction

Any time we carry out a hypothesis test, there's a possibility that we might draw the wrong conclusion. There are two types of errors in hypothesis testing, analogous to "false positive" and "false negative" in medical testing. A ***Type I error*** ("false positive") occurs when we mistakenly *reject* the null hypothesis even though in fact the null is true. A ***Type II error*** ("false negative") occurs when we mistakenly *fail to reject* the null hypothesis even though the alternative is in fact true. The table below summarizes the two error types.

Type I and II Errors

		True State of Nature	
		$\mathbf{H_0}$	$\boldsymbol{H_a}$
Your Decision	**Reject $\mathbf{H_0}$**	Type I Error	Correct Decision
	Fail to Reject $\mathbf{H_0}$	Correct Decision	Type II Error

Example 7.11: Type I and II Errors

To decide whether a lake's radioactivity level is safe, a random sample of $n = 50$ radioactivity measurements (pCi/L) is made in the lake. The value 5 pCi/L is considered the dividing line between safe and unsafe water.

To decide whether the lake is safe or not, we could test

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5 \end{aligned}$$

where μ is the lake's true mean radioactivity level. The null hypothesis says the water is hazardous. The alternative says it's safe.

A Type I error would occur if in reality the lake's water is unsafe, but we conclude that it's safe. A Type II error would occur if in reality the water is safe, but we conclude that it's hazardous.

Notice in this case that a Type I error would have *more serious consequences* than a Type II error. We'll return to this point in Subsections 7.6.3 and 7.6.4.

7.6.2 The Level of Significance as the Probability of a Type I Error

When evidence strong enough to reject the null hypothesis occurs just by chance, a Type I error results. It turns out that chance of this happening is determined by our choice of the level of significance α.

Fact 7.3 Consider a one-sample t (or z) test, using level of significance α, with null hypothesis

$$H_0 : \mu = \mu_0$$

When H_0 is true,

$$P(\text{Type I error}) = \alpha.$$

Thus the chance of making a Type I error is the same as the significance level α. To see why, consider a one-sample t test of

$$H_0 : \mu = \mu_0$$

using the rejection region approach and significance level, say, $\alpha = 0.05$. The rejection region would be comprised of t values in the most extreme 5% of the $t(n-1)$ distribution in the direction(s) specified by the alternative hypothesis. See the left-side graphs of Figures 7.1 - 7.3. But the $t(n-1)$ distribution is

the distribution that the test statistic t would follow if the null hypothesis was true. In other words, the rejection region consists of test statistic values that would occur just by chance 5% of the time if the null hypothesis was true, and each time it occurred, a Type I error would result.

The level of significance α represents the Type I error probability for other tests too, not just the one-sample t and z tests.

Fact 7.4 Consider a test for an unknown population parameter, using level of significance α, with null hypothesis of the form

$$H_0 : \text{Population parameter} \quad = \quad \text{Claimed value}$$

When H_0 is true,

$$P(\text{Type I error}) \;=\; \alpha$$

if the test statistic follows a *continuous* distribution, and

$$P(\text{Type I error}) \;\leq\; \alpha$$

if it follows a *discrete* distribution.

7.6.3 Choosing a Level of Significance

We saw in Subsection 7.2.1 that the level of significance determines how strong the evidence needs to be before we're willing to reject the null hypothesis – a smaller α requires stronger evidence because it requires that the observed t value be farther away from zero.

We now have a more intuitive way of looking at this. If we use $\alpha = 0.05$, we require that the evidence against the null hypothesis be strong enough that evidence as strong would occur just by chance only 5% of the time. Using $\alpha = 0.01$ requires even stronger evidence. It requires evidence so strong that it would occur by chance only 1% of the time.

Because using a smaller α requires stronger evidence against the null hypothesis, if we end up rejecting the null we can be more sure we're making the correct decision, that is, more sure we're not committing a Type I error. It's customary to choose α to be either 0.01, 0.05, or 0.10. The choice will depend on the consequences of making a Type I error. If the consequences are very serious, a small value for α should be used (such as $\alpha = 0.01$ or even smaller).

Example 7.12: Choosing a Level of Significance

Refer to the study to decide if a lake's radioactivity level is safe, as described in Example 7.11. We want to test

$$\begin{aligned} H_0 : \mu &\geq 5 \\ H_a : \mu &< 5, \end{aligned}$$

where μ is the true mean radioactivity level and the value 5 pCi/L is considered the dividing line between safe and unsafe water.

A Type I error consists of declaring the water safe even though it's not (see Example 7.11). This could have very serious consequences for people's health, so we'd want to use a small α, say $\alpha = 0.01$. With this choice of α, the chance of mistakenly deeming the lake safe if it wasn't would only be 1%.

Using a small significance level to lessen the chance of mistakenly rejecting a true null hypothesis comes

at a price, though. We'll see (Subsection 7.6.5) that using a smaller α raises the chance of making a Type II error when the alternative hypothesis is true.

7.6.4 Proof of Safety Versus Proof of a Hazard

As mentioned in Subsection 7.1.2, hypothesis tests are able to establish (beyond a reasonable doubt) that an alternative hypothesis is true, but they can't establish that a null hypothesis is true. When we "fail to reject" the null hypothesis, we're not necessarily "accepting" it.

In studies to decide if the contaminant level at a site is safe or hazardous, therefore, we're faced with the question of whether to establish that the site is safe or to establish that it's hazardous. For example, in the study of a lake's radioactivity level (Examples 7.11 and 7.12), we'd be faced with a choice between testing

$$\begin{array}{lcl} H_0: \mu & \geq & 5 \\ H_a: \mu & < & 5 \end{array} \qquad \text{or} \qquad \begin{array}{lcl} H_0: \mu & \leq & 5 \\ H_a: \mu & > & 5 \end{array} \tag{7.14}$$

where μ is the mean radioactivity level. In the first set of hypotheses, we could establish that the lake is safe, but not that it's hazardous. In the second set, we could establish that the lake is hazardous, but not that it's safe.

It's customary in such situations to set up the hypotheses so that we can establish that the site is *safe*. In other words, it's customary for the alternative hypothesis to say that the site is safe and the null to say that it's hazardous. There are two reasons for doing it this way:

1. If we tested the opposite, with the alternative saying the site is *hazardous* (and the null saying it's safe), then even if we failed to reject the null hypothesis, we couldn't necessarily conclude that the site is safe – only that there was insufficient evidence to conclude that it's hazardous.
2. When the alternative hypothesis says the site is *safe* (and the null says it's hazardous), a Type I error (concluding the site is safe when in fact it's hazardous) has more serious consequences than a Type II error (concluding the site is hazardous when it's actually safe). We can then reduce the chance of making the more serious (Type I) error by using a small α (Sections 7.6.2 and 7.6.3).

In the study of a lake's radioactivity level, we'd test the first set of hypotheses in (7.14) and use a small α value. That way, we'd reduce the chance of making a Type I error, which would be deeming the water to be safe when in fact it's hazardous.

7.6.5 Power and the Probability of a Type II Error

Recall that a Type II error occurs when we (mistakenly) fail to reject the null hypothesis even though in reality the alternative hypothesis is true. A Type II error results when, by happenstance, we end up with data that provide insufficient evidence to reject the null even though the alternative is true.

We use $\boldsymbol{\beta}$ to denote the chance of making a Type II error, that is,

$$\beta = P(\text{Type II error}).$$

The ***power*** of a test is the probability that you (correctly) reject the null hypothesis when in fact the alternative is true. It's the chance that you *don't* make a Type II error when the alternative is true, and therefore equals one minus the chance that you *do* make one.

Power of a Test: For *any* hypothesis test,

$$\begin{aligned} \text{Power} &= 1 - P(\text{Type II error}) \\ &= 1 - \beta. \end{aligned}$$

The *power* of a test measures the capacity of the test to detect departures from the null hypothesis. A test with high power is likely to detect a departure from the null. A test with low power is unlikely to detect it. In an experiment, with the null hypothesis saying that there's no treatment effect and the alternative saying there is one, the power measures ability of the hypothesis test to detect an effect if there is one. If our goal is to disprove the null hypothesis, *a more powerful test is better*. Some hypothesis test procedures are more powerful than others. When it's applicable, the one-sample t test procedure is more powerful than any other test procedure that might also be applicable, including the one-sample sign test discussed in Section 7.8. Furthermore, regardless of which test procedure is used, *the power will be higher the larger the sample size n is*.

Maximizing the Power: To ensure that when the alternative hypothesis is true, the power for rejecting the null is as high as possible, first choose from among the applicable test procedures the one that's most powerful. Then use the largest feasible sample size n.

For the one-sample t procedure (and other commonly used test procedures), the power of the test can be determined, for any given sample size n, from tables found in many books on hypothesis testing. In practice, these tables can also be useful for sample size determination by following the steps below.

1. Decide on a level of significance (for example $\alpha = 0.05$).
2. Decide on a desired power (for example $1 - \beta = 0.80$).
3. Look up in the table how big n would need to be to attain the desired power.

7.6.6 Statistical Significance Versus Practical Importance

There's an important distinction between a study result that's *statistically significant* and one that's of *practical importance*. The former only means that a difference or effect was *detected*, *not* that the difference or effect is large enough to have any consequential impacts. For example, one study found that logging a forest can raise temperatures of nearby streams (via increased solar radiation), but that further research was needed to determine the biological implications of those temperature changes, suggesting that the effects of logging are statistically significant, but may not be of practical importance for stream biology [**?**].

While it's true that a large difference or effect will lead to a low p-value, a small, unimportant difference or effect can also produce a low p-value if the sample size n is very large. Large sample sizes give very precise estimates of even small differences and effects, and this precision, reflected in the standard errors of the estimates, can translate into highly statistically significant results.

When a difference or effect is found to be statistically significant, it's a good idea, therefore, to report not only the p-value, which measures how strong the evidence for a difference or effect is, but also a confidence interval, which measures the size of the difference or effect.

7.6.7 The Trade-Off Between Type I and II Error Probabilities

Using a smaller level of significance requires stronger evidence against null hypothesis before we're willing to reject that hypothesis (Subsection 7.6.2), and therefore lessens the chance of making a Type I error. But unfortunately, it also means we reject the null less often *even when the alternative hypothesis is true*. In other words, using a smaller value for α *simultaneously* reduces the risk of making a Type I error *and* increases the risk of making a Type II error.

In practice, we want α to be small enough that the chance of making a Type I error is of little concern, but not so prohibitively small that detecting a difference or effect (when there is one) becomes nearly impossible. This line of reasoning has led practitioners to settle on the values 0.01, 0.05, or 0.10 for α,

with the choice depending on the severity of the consequences of making a Type I error – the more severe, the smaller the value of α should be (Subsection 7.6.3).

7.7 Dealing With Non-Normal Data: Transformations and Nonparametric Procedures

Many hypothesis testing procedures, including the one-sample t procedure, rest on an assumption that the sample was drawn from a normal population (or the sample size n is large). If the normality assumption isn't met (and n isn't large), there are two possible remedies.

1. **Transform the data to normality**: Sometimes it's possible to transform non-normal data so that the transformed values are approximately normally distributed, for example by taking their logs or using another transformation from the Ladder of Powers (Section 6.10.4), and then carry out the test on the transformed data.
2. **Carry out a nonparametric test**: Another options is to use a so-called ***nonparametric*** test procedure, meaning a test procedure that *doesn't* rely on an assumption that the population is normal (nor does it require that the population follows any other particular distribution). The *sign test* described in the next section is a nonparametric alternative to the one-sample t test.

7.8 The One-Sample Sign Test for a Population Median

7.8.1 Introduction

The ***one-sample sign test*** is a *nonparametric* test for the (unknown) value of a population *median* $\tilde{\mu}$. Unlike the one-sample t test, it doesn't require a normality assumption about the population from which the sample was drawn, so it can be viewed as a nonparametric alternative to the t test (despite the fact that it tests for the median, not the mean). Example 7.13 presents a study for which the t test wouldn't be appropriate, but the sign test would.

Example 7.13: One-Sample Sign Test

Emissions from automobiles contain platinum (Pt) as a byproduct of their catalytic converters. A study was carried out to investigate Pt contamination in roadside soils in the Manoa basin in southeast Oahu, Hawaii [**?**].

Platinum concentrations (ng/g) were measured in soil specimens extracted from a depth of 7.5 - 10 cm at $n = 11$ roadside locations.

For each location, a measure of the roadside Pt concentration relative to background concentrations was computed. The observed values of the measure, called the enrichment ratio (ER), are given in the table below.

Roadside Pt Enrichment Ratios at 7.5 - 10 cm Depth

Soil Specimen	Enrichment Ratio
1	5.3
2	4.1
3	0.9
4	9.7
5	0.6
6	0.9
7	0.6
8	2.9
9	1.1
10	2.6
11	0.6

Enrichment ratio values greater than one indicate that the roadside Pt concentration is higher than background concentrations, and support the hypothesis that automobile emissions contribute to soil Pt concentrations. Formally, the ERs used in the study were defined to be

$$\text{ER} = \frac{X}{\tilde{X}_{bg} + 2\,(\text{MAD}_{bg})},$$

where X is a roadside Pt concentration, and $\tilde{X}_{bg}$ and MAD_{bg} are the median and median absolute deviation of Pt concentrations measured at background locations. They're interpreted as ratios of roadside Pt concentrations to a typical (large) background concentration.

We want to decide if these data provide convincing evidence that roadside Pt ERs tend to be greater than one. A normal probability plot and dotplot (below) indicate that it wouldn't be reasonable to assume that the sample is from a normal distribution, and because the sample size isn't large, a one-sample t test wouldn't be appropriate. Instead, in Example 7.14, we'll carry out a one-sample sign test.

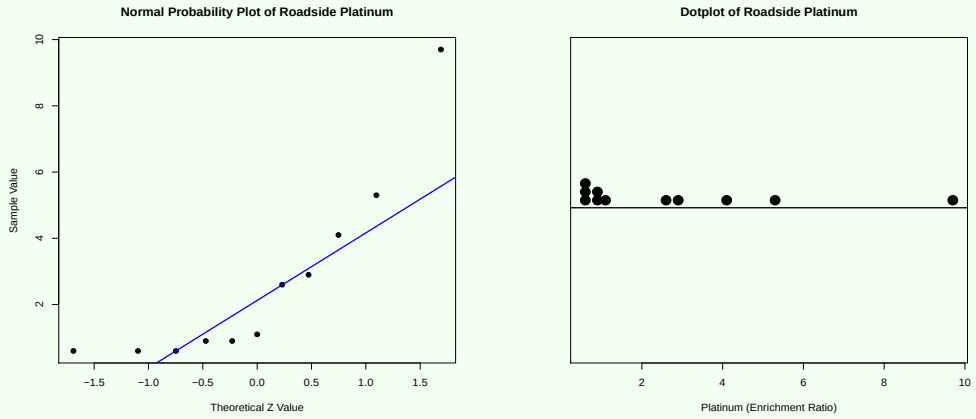


Figure 7.9: Normal probability plot (left) and dot plot (right) of the roadside Pt enrichment ratios for a soil depth of 7.5 - 10 cm.

7.8.2 The One-Sample Sign Test Procedure

Suppose we have a random sample from *any* continuous population (not necessarily normal) whose median is $\tilde{\mu}$. In the population, half of the individuals lie above $\tilde{\mu}$ and the other half below it, so each individual selected for the sample has a 50/50 chance of falling above or below $\tilde{\mu}$, and we'd expect about half of the sample to fall above $\tilde{\mu}$ and the other half below it.

We'll want to test the null hypothesis

$$H_0 : \tilde{\mu} \;=\; \tilde{\mu}_0, \tag{7.15}$$

that the true (unknown) population median is equal to some claimed value $\boldsymbol{\tilde{\mu}_0}$, versus one of the alternative hypotheses

1. $H_a : \tilde{\mu} \;>\; \tilde{\mu}_0$ (upper-tailed test)
2. $H_a : \tilde{\mu} \;<\; \tilde{\mu}_0$ (lower-tailed test)
3. $H_a : \tilde{\mu} \;\neq\; \tilde{\mu}_0$ (two-tailed test)

The null-hypothesized value $\tilde{\mu}_0$ should be chosen for its relevance to the study's research question, and the alternative hypothesis should reflect what the study is attempting to substantiate.

If the null hypothesis was true, the population median $\tilde{\mu}$ would equal the claimed value $\tilde{\mu}_0$, and each individual selected for the sample would have a 50/50 chance of falling above or below $\tilde{\mu}_0$. In this case, we'd expect about half of the individuals in the sample, or $n/2$ of them, to fall above $\tilde{\mu}_0$ (and the other half below it). The ***one-sample sign test statistic***, denoted $\boldsymbol{S^+}$, is just the number of observations in the sample that fall above $\tilde{\mu}_0$.

Sign Test Statistic: For a sample $X_1, X_2, \ldots, X_n$,

$$S^+ \;=\; \text{Number of } X_i\text{'s that are greater than } \tilde{\mu}_0.$$

If any of the X_i's equal $\tilde{\mu}_0$, they're discarded prior to computing S^+, and the sample size n is reduced by the number of discarded X_i's.

It the null hypothesis was true, we'd expect S^+ to approximately equal $n/2$, meaning about half of the observations in the sample fall above $\tilde{\mu}_0$. If substantially more than half or substantially less than half of fall above $\tilde{\mu}_0$, that is, if S^+ differs substantially from $n/2$, it's evidence that the true (unknown) population median $\tilde{\mu}$ differs from $\tilde{\mu}_0$. More specifically, we have the following.

1. *Large* values of S^+ (larger than $n/2$) provide evidence in favor of $H_a : \tilde{\mu} > \tilde{\mu}_0$.
2. *Small* values of S^+ (smaller than $n/2$) provide evidence in favor of $H_a : \tilde{\mu} < \tilde{\mu}_0$.
3. *Both large and small* values of S^+ (larger or smaller than $n/2$) provide evidence in favor of $H_a : \tilde{\mu} \neq \tilde{\mu}_0$.

To decide whether an observed value of S^+ provides statistically significant evidence in support of the alternative hypothesis, we'll need to know its sampling distribution under the null hypothesis.

Sampling Distribution of $\boldsymbol{S^+}$ Under $\boldsymbol{H_0}$: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from any continuous population whose median is $\tilde{\mu}$. Then when

$$H_0 : \tilde{\mu} \;=\; \tilde{\mu}_0$$

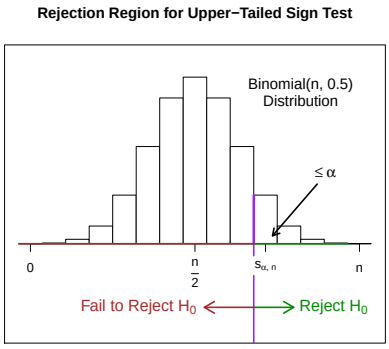
is true,

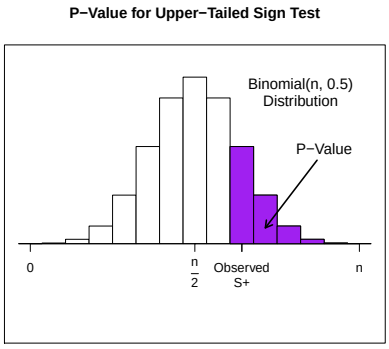
$$S^+ \;\sim\; \text{binomial}(n,\, 0.5).$$

This is easy to see if we think of the n observations in the sample as independent trials, each of which results in a *success* if the observation is greater than $\tilde{\mu}_0$, which happens with probability 0.5 when the null hypothesis is true, and a *failure* otherwise. Then the test statistic S^+ just counts how many of the trials were *successes*, that is, it's a binomial random variable (Section 4.3).

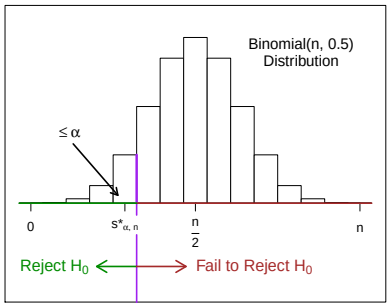
The mean of the binomial$(n, 0.5)$ distribution is $n/2$, and values of S^+ in the extreme tail of the distribution (in the direction specified by the alternative hypothesis) provide evidence against the null hypothesis. Thus the rejection region is composed of S^+ values in the the extreme $100\alpha\%$ of the distribution, and the p-value is the tail probability outward from the observed S^+ value, as shown in the figures below for the three choices of the alternative hypothesis.

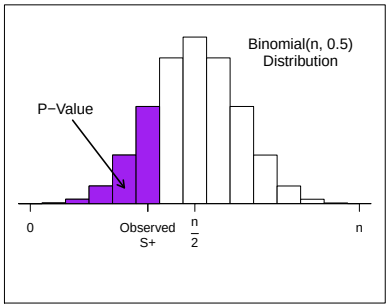
1. $H_a : \tilde{\mu} > \tilde{\mu}_0$ (Upper-Tailed Test)



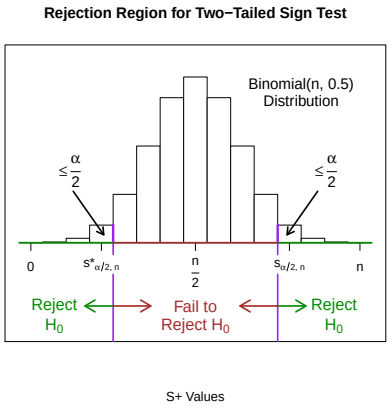


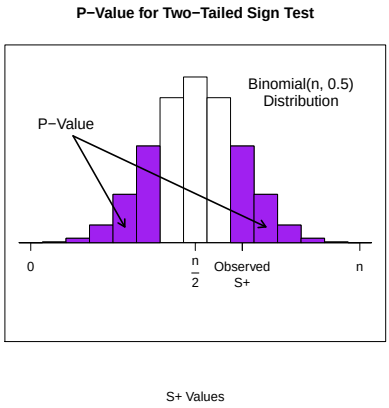
1. $H_a : \tilde{\mu} < \tilde{\mu}_0$ (Lower-Tailed Test)





1. $H_a : \tilde{\mu} \neq \tilde{\mu}_0$ (Two-Tailed Test)





The steps for carrying out a ***one-sample sign test*** for $\tilde{\mu}$ are summarized below and in the table that follows.

Steps for carrying out the one-sample sign test:

1. If any of the sample observations $X_1, X_2, \ldots, X_n$ equal $\tilde{\mu}_0$, remove those X_i's from the data set and carry on without them, reducing the sample size by the number of deleted X_i's.
2. Count the number of remaining observations that are greater than $\tilde{\mu}_0$. This gives the observed test statistic value S^+.
3. Obtain the p-value by referring the observed value of S^+ to a table of tail probabilities for the binomial(n, 0.5) distribution.

One-Sample Sign Test for $\tilde{\mu}$

Assumptions: $x_1, x_2, \ldots, x_n$ is a random sample from *any* continuous population.

Null hypothesis: $H_0 : \tilde{\mu} = \tilde{\mu}_0$.

Test statistic value: s^+ = number of x_i's greater than $\tilde{\mu}_0$.

Decision rule: Reject H_0 if p-value $< \alpha$ or s^+ is in rejection region.

Alternative hypothesis	P-value = tail probability of the binomial$(n, 0.5)$ distribution: *	Rejection region = s^+ values such that: **
$H_a : \tilde{\mu} > \tilde{\mu}_0$	to the right of (and including) s^+	$s^+ \geq s_{\alpha,n}$
$H_a : \tilde{\mu} < \tilde{\mu}_0$	to the left of (and including) s^+	$s^+ \leq s^*_{\alpha,n}$
$H_a : \tilde{\mu} \neq \tilde{\mu}_0$	2·(the smaller of the tail probabilities to the right of (and including) s^+ and to the left of (and including) s^+)	$s^+ \leq s^*_{\alpha/2,n}$ or $s^+ \geq s_{\alpha/2,n}$

* For a given sample size (after deleting the $\tilde{\mu}_0$-valued x_i's) n, the p-value for a one-tailed test is obtained from a binomial$(n, 0.5)$ distribution table by locating the upper or lower tail probability (depending on the direction of H_a) associated with the observed S^+ value. For a two-tailed test, locate both the upper and lower tail probabilities and multiply the smaller of these by two.

** For a given sample size (after deleting $\tilde{\mu}_0$-valued x_i's) n and level of significance α, $s_{\alpha,n}$ is obtained from a binomial$(n, 0.5)$ distribution table by locating the smallest s for which the upper tail probability is less than α. $s^*_{\alpha,n}$ is obtained by locating the largest s for which the lower tail probability is less than α. For the two-tailed test, $s_{\alpha/2,n}$ and $s^*_{\alpha/2,n}$ are defined analogously but with $\alpha/2$ used in place of α. In practice, due to the discreteness of the distribution, it's not always possible obtain a rejection region having exact probability α.

Here are two examples illustrating the sign test, the first with a one-sided alternative hypothesis and the second a two-sided alternative.

Example 7.14: One-Sample Sign Test

Returning to the study of platinum (Pt) in roadside soils (Example 7.13), we'll us a one-sample sign test to test the hypotheses

$$\begin{aligned} H_0 : \tilde{\mu} &= 1 \\ H_a : \tilde{\mu} &> 1 \end{aligned}$$

where $\tilde{\mu}$ is the true (unknown) population median Pt enrichment ratio (ER) in roadside soils in the Manoa basin. The null says median ER is one, meaning Pt concentrations are no greater in roadside soils than in background soils. The alternative says the median is greater than one, meaning Pt concentrations are higher in roadside soils.

From the data in Example 7.13, we see that six of the $n = 11$ ERs (slightly more than half) are greater than one, so the observed test statistic value is

$$S^+ = 6.$$

This provides *some* evidence in favor of the alternative hypothesis, but it's very weak.

The p-value is the probability that we'd get as many as six ER values greater than one just by chance if the null hypothesis was true. If the null was true, S^+ would follow a binomial(11, 0.5) distribution, and for the upper-tailed test, the p-value is the probability that a binomial(11, 0.5) random variable would be as large as six (or larger), as shown in the probability histogram below.

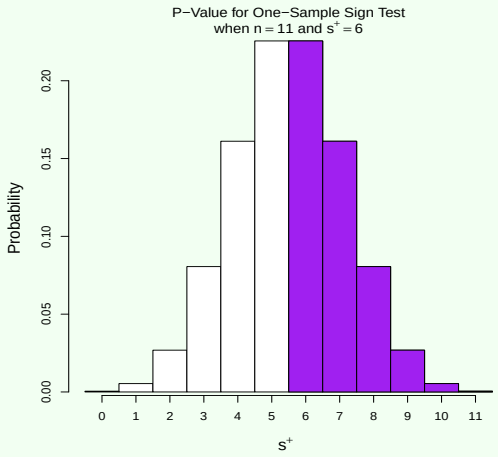


Figure 7.10: Sampling distribution of S^+ under the null hypothesis, a binomial(11, 0.5) distribution, and the p-value for an upper-tailed sign test when $n = 11$ and the observed test statistic value is $S^+ = 6$.

Although we could obtain the p-value from a table of binomial distribution tail areas, in this example, for expository purposes, we'll use probabilities obtained from the binomial(11, 0.5) probability function of Chapter 4. The table below shows these probabilities.

s^+	0	1	2	3	4	5	6	7	8	9	10	11
$p(s^+)$	0.0005	0.0054	0.0269	0.0806	0.1611	0.2256	0.2256	0.1611	0.0806	0.0269	0.0054	0.0005

From the binomial probabilities given above, the p-value is

$$\text{P-value} = 0.2256 + 0.1611 + 0.0806 + 0.0269 + 0.0054 + 0.0005 = 0.5000.$$

Using a level of significance $\alpha = 0.05$, since the p-value is greater than α, we fail to reject the null hypothesis. There's no statistically significant evidence that the true (unknown) median ER for roadside soil is greater than one.

The next example illustrates the sign test when the alternative hypothesis is two-sided.

Example 7.15: One-Sample Sign Test

A study was carried out to determine the accuracy of a new a method developed by the U.S. Environmental Protection Agency for measuring sulfur dioxide (SO_2) emissions from coal-burning power plants [**?**].

Nine lab technicians were trained in the new method and then asked to use it to measure the SO_2 concentration in an EPA audit cylinder containing a known concentration of 447 ppm. Their measurement results are shown below.

SO_2 Measurements on EPA Audit Cylinder Containing 447 ppm

Technician	SO_2 Measurement
1	688
2	478
3	524
4	447
5	2135
6	434
7	712
8	464
9	478

According to the cited paper, the very large SO_2 result may have been due to an equipment malfunction, but in this example we'll treat it as a legitimate observation.

We want to decide if there's any statistically significant evidence that the new method produces biased measurements, either ones that are systematically too large or ones that are systematically too small.

A normal probability plot and dotplot of the data are below.

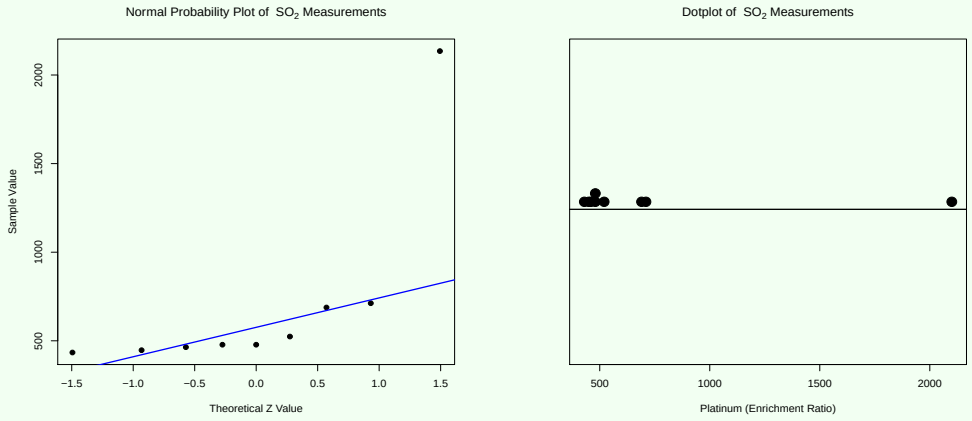


Figure 7.11: Normal probability plot (left) and dot plot (right) of SO_2 measurements on an audit cylinder containing 447 ppm SO_2.

Due to the extreme outlier, it wouldn't be reasonable to treat the data as a sample from a normal distribution, and therefore, because n is small, a one-sample t test wouldn't be appropriate. Instead, we'll carry out a sign test.

If there was no systematic bias, we'd expect measurements to be greater than 447 about half of the time and less than 447 the other half of the time. In other words, the population median would be $\tilde{\mu} = 447$. To decide whether this is the case, we'll test

$$\begin{aligned} H_0 : \tilde{\mu} &= 447 \\ H_a : \tilde{\mu} &\neq 447 \end{aligned}$$

One of the observations in the sample is equal to 447, so before proceeding we discard this observation and use a reduced sample size of $n = 8$.

Because seven of the observations in the (reduced) sample are greater than 447, the test statistic value is

$$S^+ = 7.$$

If the method was unbiased, we'd expect about four of the eight (remaining) measurements to be greater than 447.

The p-value for the two-tailed test is the probability that the number of measurements greater than 447 would differ from four by as much as the observed number, seven, if the the method was unbiased. If the method was unbiased, S^+ would follow a binomial(8, 0.5) distribution. The p-value is the sum of the two tail probabilities from this distribution, as shown in the figure below.

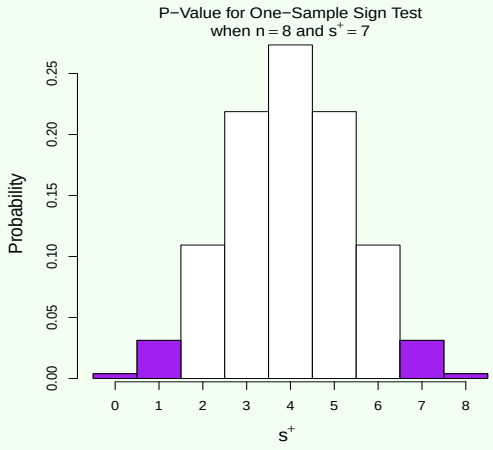


Figure 7.12: The p-value for the two-tailed, one-sample sign test when $n = 8$ and the observed test statistic value is $S^+ = 7$.

From a table of binomial distribution tail areas using $n = 8$, the upper tail probability associated with $S^+ = 7$ is 0.0352, and the lower tail probability is 0.9961. The p-value is two times the smaller of these, which is to say,

$$\text{P-value} = 2(0.0352) = 0.0704.$$

Thus using $\alpha = 0.05$, we fail to reject the null hypothesis. The data do not provide statistically significant evidence at the 5% level that the new method is biased.

7.8.3 Large Sample Version of the Sign Test

When the sample size n is large, it turns out that the sign test statistic S^+ follows (approximately) a *normal* distribution. In this case, p-values can be obtained from the normal distribution that S^+ would follow under the null hypothesis rather than from the binomial distribution.

Recall (Chapter 4) that the mean and standard deviation of a binomial(n, p) distribution are

$$\mu_{\text{bin}} = np \qquad \text{and} \qquad \sigma_{\text{bin}} = \sqrt{np(1-p)}.$$

Since $S^+ \sim \text{binomial}(n, 0.5)$ when the null hypothesis is true, the mean and standard error of its sampling distribution under the null are given by the following.

Mean and Standard Error of the Sampling Distribution of S^+ Under H_0: The mean of the sampling distribution of S^+ (when H_0 is true) is

$$\mu_{s^+} = \frac{n}{2} \tag{7.16}$$

and the standard error is

$$\sigma_{s^+} = \sqrt{\frac{n}{4}}. \tag{7.17}$$

These are valid regardless of whether or not n is large. Recall also (Chapter 4) that when n is large, the binomial distribution is approximately bell-shaped. The graphs below show the sampling distribution of S^+ under the null hypothesis for various sample sizes n.

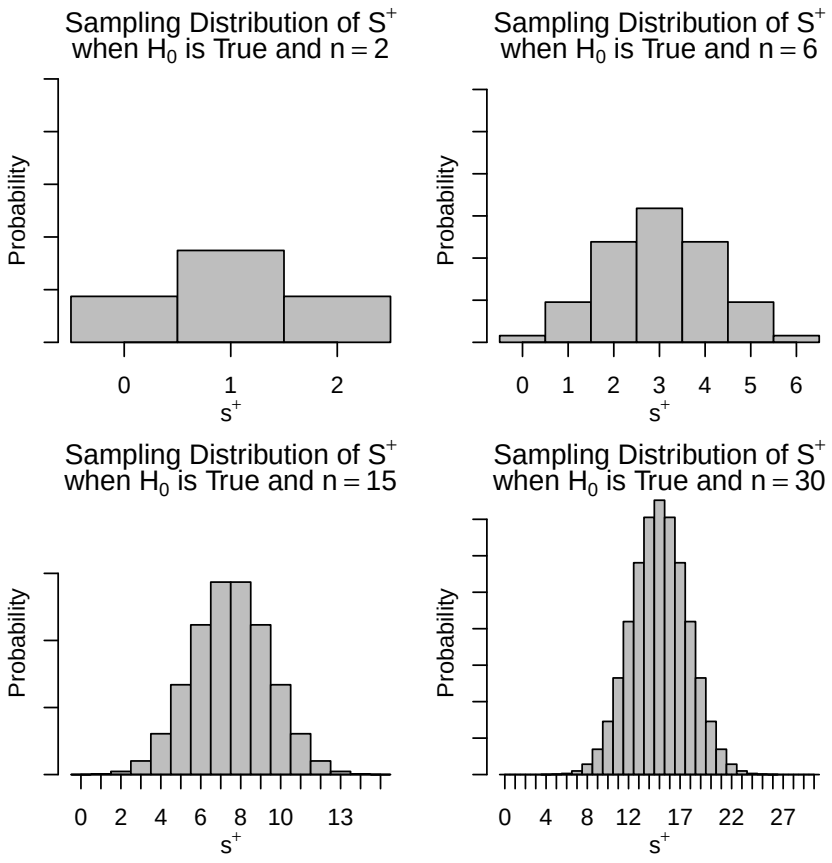


Figure 7.13: Sampling distributions of the one-sample sign test statistic S^+ under the null hypothesis for various values of n.

The larger n is, the more closely the binomial distribution is to a perfect normal distribution. This is stated more formally by the following fact.

Fact 7.5 Suppose $X_1, X_2, \ldots, X_n$ is a random sample from *any* continuous population whose median is $\tilde{\mu}$. Then if n is large ($n \geq 30$ is sufficient) and

$$H_0 : \tilde{\mu} \;=\; \tilde{\mu}_0$$

is true,

$$S^+ \;\sim\; \mathrm{N}(\mu_{s^+},\, \sigma_{s^+})$$

(approximately), where μ_{s^+} and σ_{s^+} are given by (7.16) and (7.17).

As a consequence, if we standardize S^+, the resulting random variable Z follows (approximately) a standard normal distribution, that is,

$$Z \;=\; \frac{S^+ - \mu_{s^+}}{\sigma_{s^+}} \;\sim\; \mathrm{N}(0,\, 1).$$

When n is large, the appropriate test statistic for the sign test is the ***large-sample sign test statistic***, denoted $\boldsymbol{Z^+}$, given by the following.

Large Sample Sign Test Statistic:

$$Z^+ = \frac{S^+ - \mu_{s^+}}{\sigma_{s^+}}$$

where μ_{s^+} and σ_{s^+} are given by (7.16) and (7.17).

P-values (and critical values for the rejection region approach) are obtained from the tails of the standard normal distribution in the direction(s) specified by the alternative hypothesis.

Comment: Most statistical software programs actually use a slightly more accurate ***continuity corrected*** version of Z^+ when computing p-values for the large-sample version of the sign test. The continuity correction adjusts for the fact that a continuous distribution (the standard normal) is used to approximate a discrete one (the true distribution of Z^+). Details about the continuity correction can be found in many statistics textbooks, including [**?**].

7.9 Which Test Should Be Used, the *t* Test or the Sign Test?

If the sample is from a non-normal population, and we don't want to transform the data, then unless the sample size n is large, the one-sample t test shouldn't be used. In this case we have little choice but to use a nonparametric test such as the sign test.

But if the normality assumption *is* met, we have a choice between the t test and the sign test. It can be shown that when the normality assumption is met, the t test is *more powerful* than the sign test (or other nonparametric test). In other words, when the alternative hypothesis is true, the t test is less likely to lead to a Type II error and thus more likely to find a statistically significant difference or effect. Therefore, whenever we have a choice, *the t test is preferred.*

The intuition behind why the t test is more powerful is that it's based on the actual numerical values of the data (via $\bar{X}$) as opposed to just whether or not each value is greater than $\tilde{\mu}_0$ (yes/no). Therefore the t test makes use of the complete information that's contained in the data, whereas the sign test doesn't. As a result, there's some loss of power in the sign test.

The lack of power in the sign test can be particularly annoying when the sample size n is small. Problem 7.9 gives an example in which n is small enough that the sign test is *unable* to reject the null hypothesis *at all*, regardless of which values make up the sample!

7.10 One-Sample *Z* Test for a Population Proportion

Consider now data on a dichotomous categorical variable that takes values *success* and *failure*, and suppose the data are a sample from a population whose proportion of successes is p. We saw how to compute a confidence interval for p in Chapter 6. We'll now see how to test the null hypothesis

$$H_0 : p = p_0$$

that the true (unknown) population proportion p is equal to some hypothesized value p_0 versus one of the alternative hypotheses

1. $H_a : p > p_0$ (upper-tailed test)
2. $H_a : p < p_0$ (lower-tailed test)
3. $H_a : p \neq p_0$ (two-tailed test)

The null-hypothesized value p_0 should be chosen for its relevance to the research question, and the alternative hypothesis should reflect what the study is attempting to substantiate.

The ***one-sample z test statistic***, denoted $\boldsymbol{Z}$, is defined as follows.

One-Sample Z Test Statistic (for a Proportion):

$$Z = \frac{\hat{P} - p_0}{\sigma_{\hat{P}}}, \tag{7.18}$$

where

$$\sigma_{\hat{P}} = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Because the sample proportion $\hat{P}$ is an estimate of the (unknown) population proportion, if H_0 was true, and p equal to p_0, we'd expect the sample proportion to be approximately equal to p_0 too, in which case Z would be close to zero. Any discrepancy between Z and zero would be due purely to chance (sampling variation). On the other hand, if the alternative hypothesis was true, and p different from p_0 in the direction specified by that hypothesis, we'd expect $\hat{P}$ to differ from p_0 in that direction too, in which case Z would differ from zero in that same direction. Moreover, the denominator of Z is the standard error of $\hat{P}$ (under the null hypothesis). Therefore, we have the following.

Interpretation of the Test Statistic: The value of Z measures how many standard errors $\hat{P}$ is away from p_0.

1. *Large positive* values of Z provide evidence in favor of $H_a : p > p_0$.
2. *Large negative* values of Z provide evidence in favor of $H_a : p < p_0$.
3. *Both large positive and large negative* values of Z provide evidence in favor of $H_a : p \neq p_0$.

To decide whether an observed value of Z provides statistically significant evidence to support the alternative hypothesis, we'll determine if it's among the values that would be unlikely to occur just by chance under the null hypothesis. For this, we'll need to know the sampling distribution that Z would follow if the null hypothesis was true.

We know (Section 5.4 of Chapter 5) that if n is *large*, then

$$\frac{\hat{P} - p}{\sigma_{\hat{P}}} \sim N(0, 1) \tag{7.19}$$

(approximately). Thus, because the test statistic Z is obtained by replacing the (unknown) population proportion p in (7.19) by its null-hypothesized value p_0, we have the following.

Sampling Distribution of Z Under H_0: Suppose we have a random sample of size n from a dichotomous population. Let p denote the proportion of *successes* in the population. Then if n is large and

$$H_0 : p = p_0$$

is true,

$$Z \sim \mathrm{N}(0, 1)$$

(approximately).

The sample size n can be considered large enough for Z to follow (approximately) a standard normal distribution if both

$$np_0 \geq 10 \qquad \text{and} \qquad n(1-p_0) \geq 10\,.$$

Because values of Z that differ from zero in the direction(s) specified by the alternative hypothesis count as evidence in favor of that hypothesis, p-values (and critical values for the rejection region approach) are obtained from the corresponding tail (or tails) of the standard normal distribution, as summarized below.

One-Sample $\boldsymbol{Z}$ Test for $\boldsymbol{p}$

Assumptions: The data are a random sample of size n from a dichotomous population, and n is large (using the criteria $np_0 \geq 10$ and $n(1-p_0) \geq 10$).

Null hypothesis: $H_0 : p = p_0$.

Test statistic value: $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Decision Rule: Reject H_0 if p-value $< \alpha$ or z is in rejection region.

Alternative hypothesis	P-value = area under standard normal distribution:	Rejection region = z values such that:*
$H_a : p > p_0$	to the right of z	$z > z_\alpha$
$H_a : p < p_0$	to the left of z	$z < -z_\alpha$
$H_a : p \neq p_0$	to the left of $-\lvert z\rvert$ and right of $\lvert z\rvert$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

* z_α is the $100(1-\alpha)$th percentile of the standard normal distribution.

Here's an example illustrating an upper-tailed one-sample z test for p.

Example 7.16: One-Sample $\boldsymbol{Z}$ Test for $\boldsymbol{p}$

Example 5.11 described a program for monitoring water quality in the Annapolis River, Nova Scotia, Canada. Of the 106 water specimens sampled in 2008, 29 had unsafe *E. Coli* levels and the other 77 had safe levels. The sample proportion that were unsafe was

$$\hat{P} \;=\; \frac{29}{106} \;=\; 0.27.$$

One criterion that's sometimes used to decide if a river's *E. Coli* levels are too high is if there's convincing evidence that more than 10 percent are unsafe. We'll decide whether the evidence is convincing by testing

$$\begin{aligned} H_0 : p &= 0.10 \\ H_a : p &> 0.10 \end{aligned}$$

where p is the true (unknown) population proportion of the river's *E. Coli* levels that are unsafe.

The standard error of $\hat{P}$ (under the null hypothesis) is

$$\sigma_{\hat{p}} \;=\; \sqrt{\frac{0.10(1-0.10)}{106}} \;=\; 0.029\,,$$

and thus the test statistic is

$$Z \;=\; \frac{0.27 - 0.10}{0.029} \;=\; 5.86\,,$$

indicating that the observed sample proportion is 5.86 standard errors above 0.10.

The p-value is the area to the right of $Z = 5.86$ under the standard normal curve.

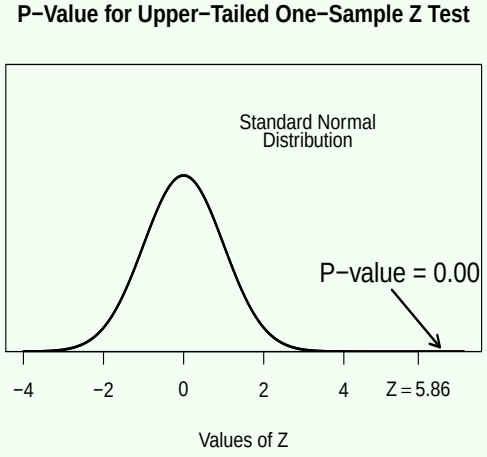


Figure 7.14: The p-value is the area to the right of the observed test statistic value, $Z = 5.86$, under the $N(0, 1)$ distribution.

From the standard normal table, the p-value is 0.0000. Using a level of significance $\alpha = 0.05$, we reject the null hypothesis and conclude that there's convincing (statistically significant) evidence that more than 10 percent of the river's *E. Coli* levels are unsafe.

To justify the hypothesis test result, we need to check that the sample size n is large enough. We need $np_0 \geq 10$ and $n(1 - p_0) \geq 10$, where $p_0 = 0.10$. We have $n = 106$, so $106(0.10) = 10.6$ and $106(1 - 0.10) = 95.4$. Therefore n is large enough.

7.11 Problems

7.1 For each of the following, state whether it's a legitimate set of statistical hypotheses. For any that aren't legitimate, say why.

a) $H_0 : \mu \leq 50$ versus $H_a : \mu \geq 50$

b) $H_0 : \bar{X} \leq 50$ versus $H_a : \bar{X} > 50$

c) $H_0 : \mu \leq 50$ versus $H_a : \mu > 50$

d) $H_0 : \mu < 50$ versus $H_a : \mu \geq 50$ (Note the equality in H_a.)

7.2 Each of the following statements violates one of the ideas presented in this chapter. Explain what's wrong with each statement.

a) "We tested the null hypothesis that there's a difference between the mean contaminant levels at two sites versus the alternative that there's no difference. We rejected the null hypothesis and concluded that there's no difference."

b) "We tested the null hypothesis that there's no difference between the mean contaminant levels at two sites versus the alternative that there's a difference. We rejected the alternative hypothesis and concluded that there's no difference."

7.3 A one-sample t test of the null hypothesis $H_0 : \mu = \mu_0$ using a sample size $n = 25$ gives a test statistic $t = 1.2$.

a) What's the p-value if the alternative hypothesis is $H_a : \mu > \mu_0$?

b) What's the p-value if the alternative hypothesis is $H_a : \mu < \mu_0$?

c) What's the p-value if the alternative hypothesis is $H_a : \mu \neq \mu_0$?

7.4 Quality assurance tests were carried out to validate a new method of measuring metal concentrations in canned fish [**?**]. They measured metals in reference fish tissues having known, certified metal concentrations using their new method, called inductively coupled plasma-optical emission spectrometry (ICP-OES).

The table below shows the means and standard deviations (mg/kg dry weight) of the ICP-OES measurements for two metals, chromium and copper, as well as the certified values. For both metals, the ICP-OES sample size was $n = 10$.

Metal Measurements on Certified Fish Tissues

	Chromium	Copper
Certified Value	34.7	2.34
ICP-OES Sample Mean	32.5	2.35
ICP-OES Sample Standard Deviation	5.5	0.01

a) Carry out a one-sample t test using the rejection region approach to decide if the true (unknown) population mean chromium measurement using ICP-OES differs from the certified value 34.7 mg/kg. Use a level of significance $\alpha = 0.05$.

b) Now carry out the test of part a using the p-value approach.

c) Carry out a one-sample t test using the rejection region approach to decide if the true (unknown) population mean copper measurement using ICP-OES differs from the certified value 2.34 mg/kg. Use a level of significance $\alpha = 0.05$.

d) Now carry out the test of part c using the p-value approach.

7.5 Problem 3.8 in Chapter 3 described a study of heavy metal contamination in soil due to industrialization around the Manali area in Chennai, Southern India [**?**]. Heavy metals were measured in soil specimens from a depth of 10 cm at 32 sites in the region. The table below shows the concentrations (mg/kg) for three of the metals, cobalt (Co), chromium (Cr), and copper (Cu).

Metals in Soil

Site	Co	Cr	Cu
M-1	32.9	207.4	44.9
M-2	30.8	150.2	55.9
M-3	32.1	156.1	44.1
M-4	15.9	150.7	24.4
M-5	17.1	158.3	22.4
M-6	23.1	191.0	36.1
M-7	9.2	230.0	25.1
M-8	11.1	150.0	26.4
M-9	13.4	240.0	37.1
M-10	11.5	197.0	24.6
M-11	58.5	149.8	181.0
M-12	10.5	218.0	193.0
M-13	7.96	151.0	178.0
M-14	3.4	215.0	160.0
M-15	6.7	306.0	175.0
M-16	12.3	172.0	186.0
M-17	15.4	157.0	190.0
M-18	12.9	385.0	237.0
M-19	14.5	395.0	251.0
M-20	12.4	255.0	211.0
M-21	15.3	201.0	326.0
M-22	12.6	183.0	335.0
M-23	15.2	204.0	277.0
M-24	9.8	158.0	372.0
M-25	14.8	247.0	300.0
M-26	20.4	246.0	225.0
M-27	11.3	309.0	198.0
M-28	16.7	418.0	150.0
M-29	19.2	328.0	171.0
M-30	11.1	251.0	87.5
M-31	22.3	154.0	94.6
M-32	20.8	161.0	102.0

The cited paper also lists concentrations of these metals in Earth's crust, which serve as reference values to which the Manali area concentrations can be compared. The reference values are shown in the table below along with sample means and standard deviations computed from the Manali data above.

	Co	Cr	Cu
Reference Value	10.0	35.0	25.0
Manali Sample Mean	16.9	221.7	154.4
Manali Sample Standard Deviation	10.3	76.6	103.9

a) Is there statistically significant evidence that the true (unknown) population mean cobalt (Co) concentration in the Manali area is higher than the reference value 10.0 mg/kg? Carry out a one-sample t test using $\alpha = 0.05$.

b) Is there statistically significant evidence that the true (unknown) population mean chromium (Cr) concentration in the Manali area is higher than the reference value 35.0 mg/kg? Carry out a one-sample t test using $\alpha = 0.05$.

c) Is there statistically significant evidence that the true (unknown) population mean copper (Cu) concentration in the Manali area is higher than the reference value 25.0 mg/kg? Carry out a one-sample t test using $\alpha = 0.05$.

7.6 Problem 6.4 in Chapter 6 described a toxicity study to investigate the reduction of thyroidal iodide uptake in humans due to perchlorate. Volunteer human subjects were given different doses of perchlorate through drinking water, and the percent change in thyroidal uptake after 24 hours was measured. Results for two dose groups, Low (given 0.007 mg/kg/day) and High (given 0.020 mg/kg/day), are below.

	Low dose	High dose
Number of subjects in group	7	10
Sample mean percent change in thyroidal uptake	-1.8	-16.4
Sample standard deviation	22.0	12.8

We want to decide if either of these two perchlorate dose levels reduces thyroidal iodide uptake.

a) Carry out a one-sample t test to decide if the population mean percent change is less than zero for the Low dose. Use a level of significance $\alpha = 0.05$.

b) Carry out a one-sample t test to decide if the population mean percent change is less than zero for the High dose. Use a level of significance $\alpha = 0.05$.

7.7 A 95% confidence interval for a population mean is $(47,\ 65)$.

a) How does the value 67 compare to the upper endpoint of the confidence interval? Could you reject the null hypothesis in a test of $H_0 : \mu = 67$ versus $H_a : \mu \neq 67$ at the 5% significance level? Explain.

b) How does the value 62 compare to the two endpoints of the confidence interval? Could you reject the null hypothesis in a test of $H_0 : \mu = 62$ versus $H_a : \mu \neq 62$ at the 5% significance level? Explain.

7.8 For a test of $H_0 : \mu = 20$ versus $H_a : \mu \neq 20$, the p-value is 0.04. Thus the null hypothesis would be rejected at both the 0.05 and 0.10 significance levels.

a) Would a 95% confidence interval for μ contain the value 20? Explain.

b) Would a 90% confidence interval for μ contain the value 20? Explain.

7.9 The Colorado Human Health Standard for nitrate and nitrite concentrations in groundwater is 10 mg/L [**?**]. Concentrations below this value are considered safe and concentrations above it are considered hazardous.

Nitrate and nitrite concentrations are to be measured in a sample of wells and a one-sample t test carried out to verify either that the water is safe or that it's hazardous.

a) If the hypotheses are

$$\begin{aligned} H_0 : \mu &\leq 10 \\ H_a : \mu &> 10 \end{aligned}$$

describe the Type I and II errors in terms of the safety of the groundwater.

b) In part *a*, which type of error has more serious consequences?

c) Now suppose instead that the hypotheses are

$$\begin{aligned} H_0 : \mu &\geq 10 \\ H_a : \mu &< 10 \end{aligned}$$

Describe the Type I and II errors in terms of the safety of the groundwater.

d) In part *c*, which type of error has more serious consequences?

e) The set of hypotheses in part a would be used if we wished to establish that the water is hazardous. The set in part c would be used if we wanted to establish that it's safe. If the hypothesis test results are to be used to make a recommendation as to whether people should use the water or not, which set of hypotheses should be tested? Why?

7.10 Suppose we have a random sample of size $n = 25$ from a population whose mean is μ, and we want to carry out a one-sample t test of

$$\begin{aligned} H_0 : \mu &= 40 \\ H_a : \mu &> 40 \end{aligned}$$

a) If the decision rule (using the rejection region approach) is

Reject H_0 if $t > 2.49$
Fail to reject H_0 if $t \leq 2.49$

what's the level of significance for the test? **Hint**: The value 2.49 is the t critical value $t_{\alpha,n-1}$. What must be the value of α?

b) What's the probability that we'd make a Type I error if the null hypothesis was true? **Hint**: What's the probability that t would fall above 2.49 just by chance?

7.11 The Metrogro Farm is a 70 mi^2 property 75 miles east of Denver, Colorado owned by Metro Wastewater Reclamation District (MWRD). In 1993 the MWRD began applying biosolids from Denver's municipal sewage treatment as fertilizer to farmland on the Metrogro Farm. Biosolids can improve soil nutrient quality, but can also contaminate groundwater by contributing to its nitrification.

Because of concerns about contamination, the U.S. Geological Survey implemented a program for monitoring groundwater on the Metrogro Farm and surrounding areas [**?**]. Dissolved nitrate and nitrite (as nitrogen in mg/L) was measured in several wells in the region. The $n = 4$ observed values for one of the wells are

11.7 13.1 14.4 14.6

The Colorado Human Health Standard for nitrogen is 10 mg/L. Concentrations below this value are considered safe and concentrations above it are considered hazardous.

The purpose of this problem is to illustrate the lack of power of the sign test when the sample size is small.

a) Carry out a one-sample sign test to decide if the true (unknown) population median nitrogen concentration for this well is greater than the health standard, 10 mg/L. Does the test find statistically significant evidence that the water is hazardous? Use level of significance $\alpha = 0.05$.

b) If a one-sample t test is used instead, is there statistically significant evidence that the true (unknown) mean nitrogen concentration is greater than 10 mg/L? Does this test find statistically significant evidence that the water is hazardous? Use level of significance $\alpha = 0.05$.

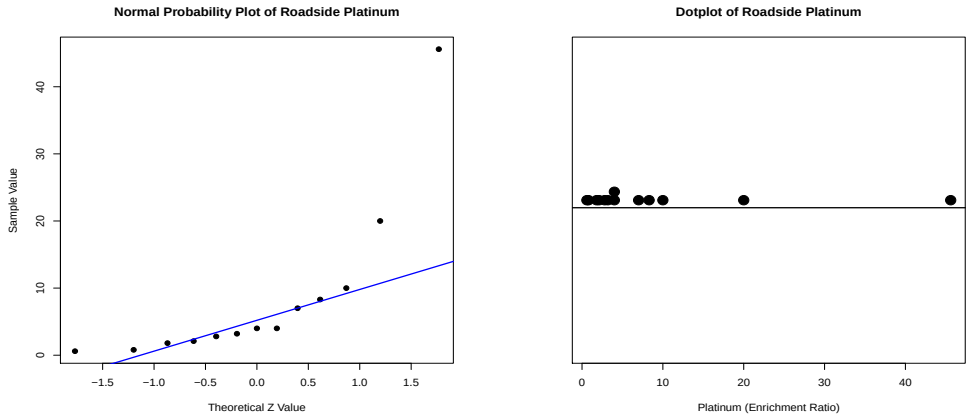
7.12 In the study of platinum (Pt) in roadside soils in Oahu, Hawaii (Examples 7.13 - 7.14), Pt was also measured in soil from a depth of 0 - 2.5 cm at $n = 13$ roadside locations. The enrichment ratios (ERs), as described in Example 7.13, are below.

Roadside Pt Enrichment Ratios at 0 - 2.5 cm Depth

Soil Specimen	Enrichment Ratio
1	7.0
2	0.6
3	20.0
4	4.0
5	10.0
6	0.8
7	8.3
8	45.6
9	4.0
10	3.2
11	2.8
12	2.1
13	1.8

Recall that enrichment ratios greater than one indicate that the roadside Pt concentration is higher than background concentrations.

A normal probability plot and dotplot (below) indicate that the sample is from a very right skewed population, so because the sample size is small, a one-sample t test isn't appropriate.



Carry out a one-sample sign test to decide if there's statistically significant evidence that the true (unknown) population median ER is greater than one.

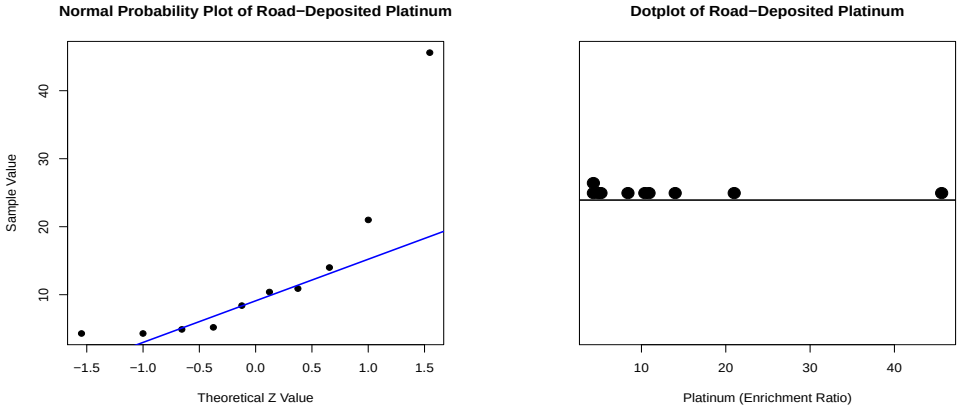
7.13 In the study of platinum (Pt) in roadside soils in Oahu, Hawaii (Examples 7.13 - 7.14), Pt was also measured in *road-deposited sediment* at $n = 10$ sites. The enrichment ratios (ERs), as described in Example 7.13, are below.

Pt Enrichment Ratios in Road-Deposited Sediment

Sediment Specimen	Enrichment Ratio
1	21.0
2	14.0
3	5.2
4	45.6
5	4.3
6	10.9
7	4.3
8	4.9
9	8.4
10	10.4

Recall that enrichment ratios greater than one indicate that the roadside Pt concentration is higher than background concentrations.

A normal probability plot and dotplot (below) indicate that the sample is from a very right skewed population, so because the sample size is small, a one-sample t test isn't appropriate.



Carry out a one-sample sign test to decide if there's statistically significant evidence that the true (unknown) population median ER is greater than one.

7.14 In the study of the accuracy of a new method for measuring sulfur dioxide in coal plant emissions (Example 7.15), the method's accuracy for measuring carbon dioxide (CO_2) was also investigated.

Nine lab technicians were trained in the new method, then they used it to measure the CO_2 concentration in an EPA audit cylinder containing a known concentration 6.04%. Their measurement results (%) are shown below.

CO_2 Measurements on EPA Audit Cylinder Containing 6.04% CO_2

Technician	CO_2 Measurement
1	6.96
2	5.97
3	6.07
4	5.94
5	5.79
6	7.08
7	10.40
8	5.69
9	5.13

A normal probability plot and dotplot indicate that the sample is from a right skewed population, so because the sample size is small, a one-sample t test isn't appropriate.

Carry out a one-sample sign test to decide if the true (unknown) population median measurement result using the new method differs from the certified value 6.04%. Use a level of significance $\alpha = 0.05$.

7.15 Problem 6.12 (Chapter 6) discussed a sample survey of $n = 79$ people who did not participate in decision making processes involving an environmental assessment of a proposed hog slaughtering facility and associated wastewater treatment plant in Brandon, Manitoba, Canada.

Of the 79 people surveyed, 51 said that "The ultimate decisions were foregone" was an important reason for their decision not to participate.

Carry out a one-sample z test (for a proportion) to decide if the true (unknown) population proportion for whom "The ultimate decisions were foregone" was an important reason for their decision not to participate is greater than 0.5. Use a level of significance $\alpha = 0.05$.

7.16 Problem 6.13 (Chapter 6) discussed a study to assess the risk of farmers' exposure to salmonella through the application of biosolids to farmlands in Ohio.

In a sample of $n = 92$ biosolids specimens, 22 tested positive for salmonella.

Carry out a one-sample z test (for a proportion) to decide if the true (unknown) population proportion of biosolids specimens that would test positive for salmonella is less than 0.25. Use a level of significance $\alpha = 0.05$.

Bibliography

[1] The national study of chemical residues in lake fish tissue. Technical Report EPA-823-R-09-006, United States Environmental Protection Agency, Sept 2009.

[2] Battelle. Heavy-duty truck activity data. Technical report, Office of Highway Information Management, Office of Technology Applications, Federal Highway Administration, 1999.

[3] Kimberlee B. Beckmen et al. Factors affecting organochlorine contaminant concentrations in milk and blood of northern fur seal (*Callorhinus ursinus*) dams and pups from St. George Island, Alaska. *The Science of the Total Environment*, 231:183–200, 1999.

[4] Alan Diduck and John A. Sinclaire. Public involvement in environmental assessment: The case of the nonparticipant. *Environmental Management*, 29(4):578 – 588, 2002.

[5] Electa Draper. Feds raided Rocky Flats 25 years ago, signaling the end of an era. *The Denver Post*, Jun. 1 2014.

[6] Electa Draper. Former Rocky Flats site stirs concerns for some neighbors. *The Denver Post*, Feb. 9 2014.

[7] Richard O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley and Sons, 1987.

[8] M. A. Greer, G. Goodman, R. C. Pleus, and S. E. Greer. Health effects assessment for environmental perchlorate contamination: the dose-response for inhibition of thyroidal radioiodine uptake in humans. *Environmental Health Perspectives*, 110:927–937, 2002.

[9] Lee Hsiang Liow, Navjot S. Sodhi, and Thomas Elmqvist. Bee diversity along a disturbance gradient in tropical lowland forests of south-east Asia. *Journal of Applied Ecology*, 38(1):180 – 192, February 2001.

[10] C. H. Nelson and P. J. Lamothe. Heavy metal anomalies in the Tinto and Odiel river and estuary system, Spain. *Estuaries*, 16(3A):496 – 511, 1993.

[11] Abramo C. Ottolenghi and Vincent V. Hamparian. Multiyear study of sludge application to farmland: Prevalence of bacterial enteric pathogens and antibody status of farm families. *Applied and Environmental Microbiology*, 53(5):1118 – 1124, May 1987.

[12] Christopher B. Pepper et al. Oganochlorine pesticides in chorioallantoic membranes of Morelet's crodocile eggs from Belize. *Journal of Wildlife Diseases*, 40(3):493–500, 2004.

[13] G. Perin et al. A five-year study on the heavy-metal pollution of Guanabara Bay sediments (Rio de Janeiro, Brazil) and evaluation of the metal bioavailability by means of geochemical speciation. *Water Resources*, 31(12):3017 – 3028, 1997.

[14] Yue Rong. Statistical methods and pitfalls in environmental data analysis. *Environmental Forensics*, 1:213 – 220, 2000.

[15] John E. Till, George G. Killough, Arthur S. Rood, Jill Weber Aanenson, Kathleen R. Meyer, Helen A. Grogan, and Warren K. Sinclair. Final report, task 5: Independent calculation. Technical Report RAC Report No. 16-RSALOP-RSAL-1999-FINAL, Risk Assessment Corporation, Feb 2000. Report submitted to the Radionuclide Soil Action Level Oversight Panel.

[16] David Ting, Robert A. Howd, Anna M. Fan, and George V. Alexeeff. Development of a health-protective drinking water level for perchlorate. *Environmental Health Perspectives*, 114(6):881–886, June 2006.

[17] Randall S. Wells et al. Integrating life-history and reproductive success data to examine potential relationships with organochlorine compounds for bottlenose dolphins (*Tursiops truncatus*) in Sarasota Bay, Florida. *Science of the Total Environment*, 349:106–119, 2005.