Chapter 10

One Sample Hypothesis Tests

In this chapter we will discuss a classical statistical inference that deals with making statements about population parameters. The two main areas of this type of statistical inference are **confidence intervals** (Chapter 9) and **hypothesis testing**.

Hypothesis testing allows us an opportunity to assess evidence for or against a pre-specified hypothesis about population parameters based on the information in the sample. We can measure the amount of evidence for a particular hypothesis using the tools we've learned thus far. Assessing the amount of evidence allows us to make conclusions about the "statistical significance" about a hypothesis.

Hypothesis testing is widely used because it appears to be "objective proof," and they are easy to calculate with software. While we cover the material here, we strongly emphasize the duty to provide a full and detailed picture of any statistical analysis.

Important: The results of a hypothesis test **are never** proof of anything, but they are often helpful for assessing statements about the population.

Recall that the goal of **statistical inference** is to use the information in a sample of individuals to describe a larger population of individuals. In a hypothesis test we will use sample data and the statistics that describe them to assess claims about a population parameter.

Value of Interest	Unknown Parameter	Assumed Value	Statistic
Population mean	μ	μ_0	$\bar{x}$
Population variance	σ^2	σ_0^2	s_x^2
Population proportion	p	p_0	$\widehat{p}$
Population median	M	M_0	$\widehat{m}$

Our understanding about sampling distributions will be paramount to our success in performing a hypothesis test about a population parameter. This chapter heavily relies on Central Limit Theorem and our tools for evaluating probabilities. In a sense, you already know how to complete a hypothesis test – we're just formalizing the technique.

A hypothesis test assesses the probability of an outcome, described by a statistic, if a particular hypothesis about the population, described by a parameter, is assumed to be true. This calculated probability describes whether or not an outcome is "unusual" if that hypothesis was indeed true. If it is "unusual," we take that as evidence of against the assumed hypothesis.

The process of a hypothesis test can be described in five steps. These steps depend on which parameter the hypothesis test is about.

10.1 Parts of a Hypothesis Test

Definition 10.1. A **hypothesis test** is an inferential technique which pits two competing hypotheses versus each other. The goal is to decide which hypothesis is more supported by the observed data. A hypothesis test consists of five basic steps.

1. Develop the hypotheses about a population parameter.
2. Collect data in a manner that satisfies the assumptions of the selected methodology.
   - The sample is representative of the population; e.g., we have a simple random sample.
3. Calculate a test statistic based on the sample.
   - This value compares the observed data with what is expected under an assumed value of the parameter of interest.
4. Calculate a p-value.
   - This is a probability about the observed data given an assumed value of the parameter of interest that specifies the sampling distribution.
5. Interpret the test statistic and p-value to assess evidence about the hypotheses.

10.1.1 Develop a Hypothesis

Definition 10.2. A **null hypothesis**, H_0, states the value of the parameter to be tested. For example, we will consider the following null hypotheses

$$H_0 : p = p_0 \qquad \text{or} \qquad H_0 : \mu = \mu_0 \qquad \text{or} \qquad H_0 : \sigma^2 = \sigma_0^2 \qquad \text{or} \qquad H_0 : M = M_0.$$

In this course, we will take the null hypothesis to be **sharp**; that is, there is only one value of the parameter possible under H_0. We assume that the null hypothesis is true and evaluate if observed data provides evidence against that assumption, in favor of an alternative.

Definition 10.3. An **alternative hypothesis**, H_a describes what values of the parameter we are interested in testing H_0 against. For this type of hypothesis we're choosing from three inequality signs as the alternative to equality, the null hypothesis. The alternative hypothesis is often the hypothesis the research wants to conclude is supported by the data and is often referred to as the "hypothesis of change."

$H_a : \mu < \mu_0$	or	$H_a : \mu > \mu_0$	or	$H_a : \mu \neq \mu_0$
$H_a : \sigma^2 < \sigma_0^2$	or	$H_a : \sigma^2 > \sigma_0^2$	or	$H_a : \sigma^2 \neq \sigma_0^2$
$H_a : p < p_0$	or	$H_a : p > p_0$	or	$H_a : p \neq p_0$
$H_a : M < M_0$	or	$H_a : M > M_0$	or	$H_a : M \neq M_0$

Note: The alternative hypothesis above using "$<$" or "$>$" are called **one-sided alternatives** and those using "$\neq$" ares called a **two-sided alternative**. One-sided alternatives state pointedly which direction we are testing H_0 against. A two-sided alternative does not specify this and is simply interested in difference in either direction.

10.1.2 Assumptions

The assumptions we need to check will vary based on the type of hypothesis testing we are completing, e.g., which parameter is of interest. The goal of a hypothesis test is to characterize a population using a sample and so we must ensure that our procedure takes care to protect us against making bad decisions. Generally, we check

1. The sample is generalizable to the population of interest.
    - The sample is obtained via randomization.
    - The sampling technique or experimental design does not introduce bias.
2. The observations are independent.
    - If an observation affects other observations the observed data may be too similar and lead to a non-representative sample.
3. Distributional assumptions of the test are met.
    - Often, we want to invoke Central Limit Theorem so we can use the Gaussian distribution.

As we introduce hypothesis testing procedures about p, μ and M, we will be more specific about the necessary assumptions for each test. We stress that this is perhaps the most important step of the process in choosing the correct procedure for the observed data and research question.

10.1.3 Test Statistics

Definition 10.4. A **test statistic** is a statistic that is used to test the null versus the alternative hypothesis. We make our decision by comparing the observed value of the test statistic to its sampling distribution under the assumption that the null is true; we will reject a null hypothesis when we have observed data that is unusual if the null hypothesis is assumed to be true.

- If the observed value of the test statistic is consistent with its sampling distribution under the assumption that the null hypothesis is true, then this **is not** evidence for the alternative.
- If the observed value of the test statistic is unusual, with its sampling distribution under the assumption that the null hypothesis is true, then this **is** evidence for the alternative.

10.1.4 P-values

Definition 10.5. When performing a hypothesis test a **probability value, or p-value,** provides the probability of the observed data, or more extreme with respect to the alternative, when the

null hypothesis is true. When we observe data that is "unusual" under the null hypothesis we take that as evidence against H_0.

Within hypothesis testing, p-values have become quite a controversial subject in scientific research. One reason is misinterpretation. The only correct interpretation of a p-value is that to describe them as the probability that we observe what we have (or more extreme in relation to the alternative hypothesis) under the assumption that the null hypothesis is true; e.g.,

$$p\text{-value} = P(\text{observed statistic or more extreme} | H_0 \text{ is true}).$$

Sometimes the p-value is referred to as "the probability that the null hypothesis is true." This interpretation is very incorrect as small p-values are interpreted as evidence that the null-hypothesis is false. This is nonsensical if you consider the probability statement above; to calculate the probability we're assuming the null hypothesis is true!

The probability value (p-value) for a hypothesis test measures how much evidence we have against H_0. It is important to remember the following:

$$\text{the smaller the p-value} \implies \text{the more evidence against } \mathrm{H}_0.$$

10.1.5 Decision Making

Rule: In any hypothesis test, if the test statistic falls in the rejection region then we reject H_0.

Definition 10.6. The **rejection region** specifies for which observed data we will reject the null hypothesis in favor of the alternative. The rejection region is usually located in the tails of a sampling distribution under the assumption that the null hypothesis is true. This is why we take the null hypothesis to be sharp, namely, so that we can construct a single sampling distribution.

Definition 10.7. The **significance level** for a hypothesis test, α, specifies how unusual an observation is required to be for consideration as evidence against the null hypothesis. Thus, if the probability value is less than or equal to α, **we reject** H_0; e.g., the observed data is unusual enough under the null hypothesis that it provides evidence against it. If the probability value is greater than α, **we fail to reject** H_0. By default we will consider observations to be **unusual** when the probability of the event occurring is low (less than 0.05); we use $\alpha = 0.05$. There are, however, times where we may decide to use different levels of alpha.

Remark: Hypothesis tests conducted at with a significance level α are said to have a $(1-\alpha)\times 100\%$ confidence level. For $\alpha = 0.05$, we have a hypothesis test with a confidence level of 95%.

There are two ways to make a decision based on observed data, both of which help us decide if our observed data lies in the rejection region. See Figure 10.1.1 for a depiction of rejection regions under the $t(50)$ model – this provides a visual for hypothesis testing where the sampling distribution is $t(50)$.

Via Critical Values

1. Calculate the test statistic under the assumption of the null hypothesis.
2. Find the percentile of the sampling distribution of interest for the alternative hypothesis of interest.

3. Compare the test statistic with the percentile of the sampling distribution of interest.
4. Reject the null hypothesis if the statistic calculated from the sample is in the rejection region.

Via p-values

1. Calculate the test statistic under the assumption of the null hypothesis.
2. Calculate the p-value for the test statistic under the assumption of the null hypothesis.
3. Compare the p-value to the significance value (α).
4. Reject the null hypothesis if the p-value $< \alpha$.

Question: How small does a p-value have to get before we "reject H_0 in favor of H_1?"

Answer: Unfortunately, there is no right answer to this question. What is commonly done is the following.

- First choose a significance level α that is small. This represents the probability that we will reject a true H_0, that is,
$$\alpha = P(\text{Reject } H_0 | H_0 \text{ true}).$$
- Common values of α chosen beforehand are $\alpha = 0.10$, $\alpha = 0.05$ (the most common), $\alpha = 0.01$.
- The smaller the α is chosen to be, the more evidence one requires to reject H_0. This is a true statement because of the following well-known decision rule:
$$\text{p-value} < \alpha \implies \text{reject } H_0.$$
- Therefore, the value of α chosen by the experimenter (you!) determines how small the p-value must get before H_0 is ultimately rejected

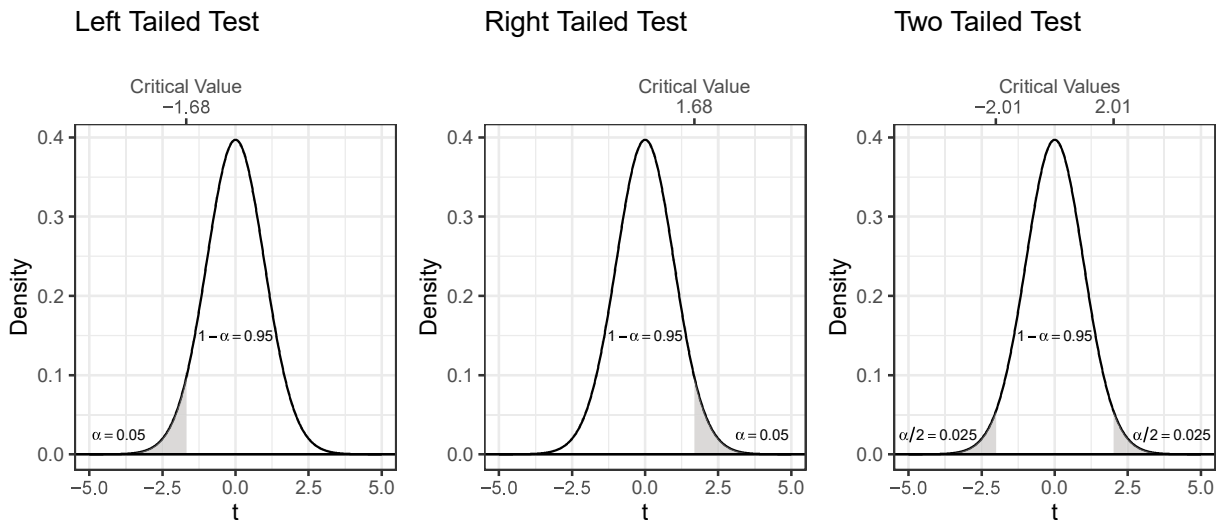

Figure 10.1.1: Rejection regions (in grey) for the three types of alternative hypotheses where the sampling distribution is $t(50)$ and $\alpha = 0.05$.

If we reject the null hypothesis we say "at the α significance level we reject the null hypothesis in favor of the alternative, e.g., there is evidence to suggest that the alternative hypothesis is true." This is known as a "statistically significant" result.

Alternatively, if we do not reject the null hypothesis we say "at the α significance level we fail to reject the null hypothesis in favor of the alternative, e.g., there is not enough evidence to suggest that the alternative hypothesis is true."

We **do not** say that we proved that either hypothesis is true, just that we have evidence that supports or does not support the alternative over the null hypothesis. We do not "prove" anything; even when we reject the null hypothesis our interpretation is that it is unlikely true, not impossible. Hypothesis testing helps scientist formally test which hypothesis is supported more by the observed data.

10.2 Errors in Hypothesis Testing

Table 10.2 summarizes the four possible outcomes from performing a hypothesis test.

	Decision: Reject H_0	Decision: Fail to reject H_0
Truth: H_0	Type I Error	correct decision
Truth: H_a	correct decision	Type II Error

Table 10.2.1: States of testing H_0 versus H_a.

We should recognize a problem with hypothesis testing - we're rejecting if the test statistic calculated on our observed sample is unusual. In fact, we've set up hypothesis testing to reject ($\alpha \times 100$) percent of observations when the null is true. Even though those observations are unusual they are not impossible and so we might reject when that is the incorrect choice.

Consider an experiment where we know our null hypothesis is true. If we collected 1000 random samples of size n and completed a hypothesis test for each sample we would expect to incorrectly reject fifty of them; this is akin to the coverage of confidence intervals.

Definition 10.8. Type I Error occurs when rejecting the null hypothesis when it is actually true. The probability of Type I Error is denoted by α, the significance level. Notationally,

$$\alpha = P(\text{Type I Error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$$

We would like the α to be small and we prespecify this value, usually $\alpha = 0.05$.

Definition 10.9. Type II Error occurs when failing to reject the null hypothesis when the alternative hypothesis is true. The probability of Type II Error is denoted by β. Notationally,

$$\beta = P(\text{Type II Error}) = P(\text{Fail to reject } H_0 | H_a \text{ is true}),$$

where "H_a is true" indicates that the true population parameter takes on some other value. This indicates that β is dependent on the true parameter value.

We would like the β to be small. Unlike the null hypothesis, the alternative hypotheses can be true in many ways so β is different for different values that satisfy the alternative hypothesis.

We will not calculate the probability of making Type II Error but it will be useful to know that the probability changes inversely with α, that is

- if we increase α it will result in smaller β
- if we decrease α it will result in larger β.

Definition 10.10. The **power** of a hypothesis test is defined as the probability that we reject a null hypothesis when, in fact, we should. Noting this is the complement of the probability of Type II Error, we have

$$\begin{aligned}\text{Power} &= P(\text{reject } H_0 | H_a \text{ is true}) \\ &= 1 - \beta.\end{aligned}$$

Example 10.11. Consider an example where a pricing strategist is worried because they think that the mean claim amount of a new class of customers μ_X is going to be greater than the standard class's, \$10,000, and suggests raising prices to cover the added cost. The hypotheses are

$$\begin{aligned}H_0 &: \mu = 10,000 \\ H_a &: \mu > 10,000.\end{aligned}$$

The two errors for this hypothesis are

- **Type I Error:** we deduce that there is evidence to suggest the mean claim amount is higher than \$10,000 when in fact it is not
- **Type II Error:** we deduce that there is not enough evidence to suggest the mean claim amount is higher than \$10,000 when in fact it is.

Type I Error would lead us to raise prices when it is unnecessary perhaps making the company less competitive or causing customers to leave for other companies. A Type II Error would leave the company on the hook for the additional claims without an increased premiums. In this case the severity of each error depends on who you are, but if you're the company a Type II Error is likely less desirable; in that case a larger significance would be desired,e.g., $\alpha = 0.10$ instead of $\alpha = 0.05$ or $\alpha = 0.01$.

Example 10.12. Data is collected to see if there is evidence that the average contaminant concentration level exceeds the acceptable level for fishing, say a measure of more than 1,200 parts per million for total dissolved solids in the water, at a local lake.

With the collection of a random sample of measurements taken at various locations of the lake an official decides to test the hypotheses,

$$\begin{aligned}H_0 &: \text{ Contaminant levels are not unsafe: } \mu_x = 1,200 \\ H_a &: \text{ Contaminant levels are unsafe: } \mu_x > 1,200\end{aligned}$$

The two errors for this hypothesis are

- **Type I Error:** we deduce that there is evidence to suggest the contaminant levels are unsafe when, truly, they are safe.
- **Type II Error:** we deduce that there is not enough evidence to suggest the contaminant levels are unsafe when, truly, they are unsafe.

Type I Error would lead the official to close the lake to fishing when it was actually safe to do so, perhaps inconveniencing those looking to fish the lake for business or pleasure. Type II Error would lead the official to take no action when it was actually unsafe to fish in. This could perhaps lead to people eating contaminated fish and becoming ill. In this case, the severity of each error is obvious. A Type II Error would be a worse outcome and so a larger significance would be desired,e.g., $\alpha = 0.10$ instead of $\alpha = 0.05$ or $\alpha = 0.01$.

10.3 Hypothesis Testing for a Population Mean

As a motivating example consider the data recorded in a study described in Campbell and Mahon (1974). Their study was meant to find distinguishing characteristics for orange and blue rock crabs. The two types of crabs were thought to be of the same species until shown that they are indeed two from separate taxonomic groups.

Unfortunately, when theses rock crabs are preserved, the chemicals used cause them to lose their distinguishing color. This makes it necessary to identify previously collected (but unlabeled) samples without seeing their colors.

The researchers measured $n = 100$ (50 males and 50 females) orange rock crabs in Fremantle, WA and recorded their carapace width (mm), the width of each crabs upper shell. The goal was to develop a criteria for identification without seeing the color of the rock crab.

These data are available in the `MASS` package.

```
> library("MASS")
data(crabs)
orangeCrabs<-crabs[which(crabs$sp=="O"),] #grabs only the data needed
```

Suppose that it is well known that blue rock crabs have a mean carapace widths of 34mm, is there evidence that orange rock crabs from Fremantle, WA have a different mean carapace width based on the sampled data?

Preview: The goal of our hypothesis test for the population mean is to assess the claim that orange rock crabs have a different population mean than 34mm.

10.3.1 Develop a Hypothesis

Again, it is instructive to think about the proposed answer to the research question when trying to choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : \mu_x = \mu_0 & \text{or} & H_0 : \mu_x = \mu_0 \quad \text{or} \quad H_0 : \mu_x = \mu_0 \\ H_a : \mu_x < \mu_0 & & H_a : \mu_x > \mu_0 \qquad\quad H_a : \mu_x \neq \mu_0. \end{array}$$

Research Question: Can rock crabs be identified to the correct taxonomy category without observing the color?

Hypothesis: Yes, the average carapace width of orange rock crabs different than 34mm.

The hypothesis instructs us to consider the alternative hypothesis

$$H_a : \mu_x \neq 34,$$

where we take $\mu_0 = 34$. We then take the sharp null hypothesis of

$$H_0 : \mu_x = 34.$$

For our researchers, the hypothesis of interest is

$$H_0 : \mu_x = 34$$
$$H_a : \mu_x \neq 34$$

where

$\mu =$ the population mean carapace width (mm) of orange rock crabs. This is unknown.
$\bar{x} =$ the sample mean carapace width (mm) of orange rock crabs. This is calculated on the sample.

10.3.2 Assumptions

When completing a hypothesis test for the population mean we have to check our assumptions for using such methodology.

1. The variable of interest is quantitative
2. The sample is generalizable to the population of interest
3. Central Limit Theorem can be invoked to show that the sampling distribution of $\bar{x}$ is approximately normal.

Below, we outline an assessment of the assumptions for our researchers.

1. The carapace width (mm) of a rock crab is a quantitative continuous variable.
2. The 100 orange rock crabs were collected at Fremantle, WA USA and so this analyses may only generalize to rock crabs in Fremantle, WA USA. It's possible that there is geographical variance in rock crabs; e.g, rock crabs from Fremantle, WA USA are larger than rock crabs from Bar Harbor, ME USA.
3. The Central Limit Theorem is satisfied as $n = 100 > 30$. Thus, it follows that

$$\bar{x} \sim \mathcal{AG}\left(\mu_{\bar{x}} = \mu_x, \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$$

Recall, that the goal of hypothesis testing is to assess claims about a population using a sample. Thus, it is highly unlikely that we would know the population standard deviation. When this is the case we will consider the t distribution. Recall that under the same assumptions

$$\frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \sim \mathcal{A}t(df = n - 1).$$

Remark: Using the t-test has been found to be quite robust to the Central Limit Theorem assumptions however, it is only optimal when the assumptions are met.

10.3.3 Test Statistic

The test statistic in the case that we know the population standard deviation the z statistic,

$$z^* = \frac{\bar{x} - \mu_0}{\frac{\sigma_x}{\sqrt{n}}},$$

is the number of standard deviations the observed sample mean is from μ_0 which we assume to be the true population mean.

Since it is infrequently the case that we know the value of σ_x, we most often use a t statistic,

$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}},$$

which is the number of sample standard deviations the observed sample mean is from μ_0 which we assume to be the true population mean.

For our researchers, who don't have information about the population standard deviation of the carapace widths of orange rock crabs, the test statistic for the observed sample of $n = 100$ orange rock crabs is

$$t^* = \frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}} = \frac{38.112 - 34}{\frac{7.540922}{\sqrt{100}}} = 5.452914.$$

Where the sample mean and standard deviation are calculated in R.

```
> (n<-length(orangeCrabs$CW))
[1] 100
> mu0<-34
> (x.bar<-mean(orangeCrabs$CW))
[1] 38.112
> alpha<-0.05
> mu.xbar<-mu0
> (se.xbar<-sd(orangeCrabs$CW)/sqrt(n))
[1] 0.7540922
> (t.obs<-(x.bar-mu.xbar)/se.xbar)  ##test statistic
[1] 5.452914
```

10.3.4 P-values

Recall that a p-value provides the probability of the observed data, or more extreme with respect to the alternative hypothesis, when the null hypothesis is true. The terminology "more extreme" follows from the alternative hypothesis. "More extreme" requires a little more thought under the "not equal" alternative compared to "less" or "greater."

The observation of a mean carapace width of 38.112 mm is 4.112 mm larger than the assumed mean carapace width of 34 mm. When we say "more extreme" under the "not equal" alternative we're interested in observations that are 4.112 mm larger or 4.112 mm smaller or more extreme.

Thus, we can calculate the p-value as follows.

$$\begin{aligned}
p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\
&= P(\bar{x} > 38.112 \cup \bar{x} < 34 - 4.112|\mu_x = \mu_0 = 34) \\
&= P(\bar{x} > 38.112 \cup \bar{x} < 29.888|\mu_x = \mu_0 = 34) \\
&= P(\bar{x} > 38.112|\mu_x = \mu_0 = 34) + P(\bar{x} < 29.888|\mu_x = \mu_0 = 34) && \textbf{[disjoint events]} \\
&= (1 - P(\bar{x} \leq 38.112|\mu_x = \mu_0 = 34)) + P(\bar{x} \leq 29.888|\mu_x = \mu_0 = 34)
\end{aligned}$$

If we knew the population standard deviation, σ_x, we could calculate this probability directly using R. Unfortunately, σ_x is unknown and so we must consider this probability using the t statistic.

$$\begin{aligned}
p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\
&= P(\bar{x} > 38.112 \cup \bar{x} < 29.888|\mu_x = \mu_0 = 34) \\
&= P(T_{100-1} > 5.452914 \cup T_{100-1} < -5.452914|\mu_x = \mu_0 = 34) && [t \textbf{ statistics}] \\
&= 2P(T_{100-1} \leq -5.452914|\mu_x = \mu_0 = 34) && \textbf{[symmetry]}
\end{aligned}$$

This is calculated in R as follows.

```
> (p.value<-2*pt(-abs(t.obs),df=n-1)) ##p value
[1] 3.647644e-07
```

This is visualized in Figure 10.3.2, which is created with the following R code.

```
> ggdat<-data.frame(t=seq(-4.5,4.5,0.01),
+                     f=dt(x=seq(-4.5,4.5,0.01),df=n-1))
> ggdat.highlight<-data.frame(x=c(-t.obs,t.obs),y=c(0,0))
> #Save the observations corresponding to t
> axis.labels<-round(c(-abs(t.obs),qt(alpha/2,df=n-1),0,
+    qt(1-alpha/2,df=n-1),abs(t.obs))*se.xbar+mu.xbar,2)
> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=n-1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,t<=qt(alpha/2,df=n-1)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,t>=abs(t.obs)),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_ribbon(data=subset(ggdat,t<=-abs(t.obs)),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   ggtitle("T Test for the Population Mean",
+           subtitle=bquote(H[0]*":"~mu==34*", versus "*H[a]*":"~mu!=34))+
+   annotate("text", x=3, y=0.07,
+            label= deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
```

```
+   annotate("text", x=-3, y=0.07,
+            label= deparse(bquote(alpha/2==0.025)),parse=TRUE,size=3.5)+
+   annotate("text", x=5.25, y=0.03, label="Observation",size=3.5)+
+   annotate("text", x=-5.25, y=0.06, label="Mirrored \n Observation",size=3.5)+
+   annotate("text", x=5, y=0.25, label="P-value<0.0001",size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+     breaks=c(-abs(t.obs),qt(alpha/2,n-1),0,qt(1-alpha/2,n-1),abs(t.obs)),
+     labels = axis.labels,name="Average Carapace Width (mm)"))
```

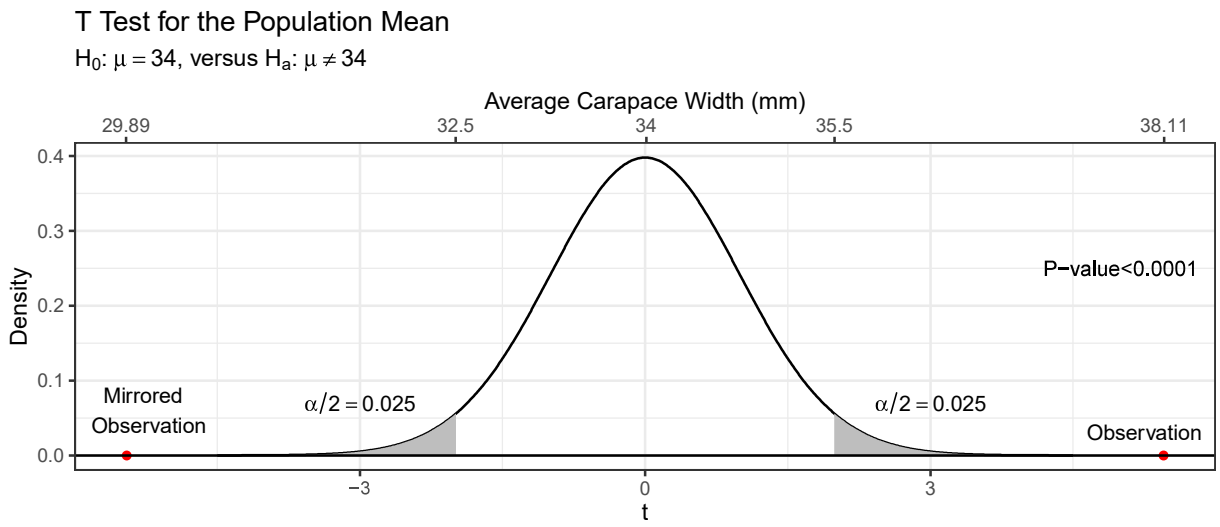

Figure 10.3.2: P-value as calculated in Section 10.3.4 for the t test. The p-value is shaded in red (not visible), the rejection region is shaded in grey and the observation is highlighted with a red point.

10.3.5 Decision Making

We reject the null hypothesis when the observed data is unusual under H_0. A prespecified significance level of $\alpha = 0.05$ dictates that we reject the null hypothesis when the probability of the observed data or more extreme under H_0 is less than α.

Figure 10.3.4 shows that the observed data does fall in the rejection region for this test. While this is clear from the graph, we are curious about how to decide numerically whether or not the observed data is in the rejection region, noting we did this as we asked R to create the graph.

We can ask, and answer, this questions two ways. We discuss the variety of ways to answer the question below for a t test; doing so for the z test would be very similar. While we show both ways below, we note that it is general practice to make this decision using the p-value.

Q1: Is the test statistic t^* in the rejection region?
A1: This question asks if the observed $t^* = 5.452914$ is in the bottom or top $\alpha/2 \times 100$ percentile of T_{100-1} under the null hypothesis. We can ask R for these percentiles of T using the sampling distribution.

```
> qt(p=0.025,df=100-1)
```

```
[1] -1.984217
> qt(p=0.975,df=100-1)
[1] 1.984217
```

Thus, we would reject the null hypothesis for any observation of T less than -1.984217 or greater than 1.984217; these values are termed the **lower critical t value** and **upper critical t value**, respectively.
For our researchers: Since $t^* = 5.452914$, it is in the rejection region and so we reject the null hypothesis.

Remark: We can use the critical values of T to find corresponding critical values for $\bar{x}$ by solving the t statistic formula for $\bar{x}$.

$$\begin{aligned}
t^* &= \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \\
\pm 1.984217 &= \frac{\bar{x} - 34}{\frac{7.540992}{\sqrt{100}}} && \textbf{[plugging in]} \\
\bar{x} - 34 &= \pm 1.984216 \left(\frac{7.540992}{\sqrt{100}}\right) && [\times \tfrac{7.540992}{\sqrt{100}}] \\
\bar{x} &= 34 \pm 1.984216 \left(\frac{7.540992}{\sqrt{100}}\right) && [+34] \\
t^* = -1.984217 &\implies \bar{x} = 32.5037 \\
t^* = 1.984217 &\implies \bar{x} = 35.4963
\end{aligned}$$

Thus, we would reject the null hypothesis for any observation of $\bar{x}$ less than 32.5037 or greater than 35.4963; these values are termed the **lower critical $\bar{x}$ value** and **upper critical $\bar{x}$ value**, respectively.

Q2: Does the p-value indicate that the observation lies in the rejection region?
A2: If the p-value is less than alpha then the observation must lie in the rejection region. Thus, we would reject the null hypothesis for any observation that has a p-value less than 0.05.

For our researchers: The p-value for this test, p-value=0.0000003647641, is less than $\alpha = 0.05$ which indicates that we should reject the null hypothesis in favor of the alternative.

Conclusion: We say "there is sufficient evidence to suggest that the population mean carapace width of orange rock crabs is different from 34 mm ($t = 5.45$, p-value¡0.0001)."

Remark: The answers to these two questions will always match. In practice, quantitative researchers report the test statistic and p-value when drawing their conclusions as above.

10.3.6 Summary

In this section, we succinctly summarize the five steps for the t Hypothesis Test. While it might be simple to check this table and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Note that in the summary below, we are neatly tying chapters worth of material together. Hypothesis testing, in a sense, isn't new material but a new task we can take on given our tools from previous chapters.

Step One	$H_a : \mu_x < \mu_0$	$H_a : \mu_x > \mu_0$	$H_a : \mu_x \neq \mu_0$
Step Two	Check Assumptions		
Step Three	$t^* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$		
Step Four	$P(T_{n-1} < t^*)$	$P(T_{n-1} > t^*)$	$2P(T_{n-1} < -\lvert t^*\rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

For our researchers,

Step One	$H_a : \mu_x \neq 34$
Step Two	Sample may be generalizable – need more information. Observations are independent. Central Limit Theorem assumptions satisfied.
Step Three	$t^* = 5.452914$
Step Four	p-value $= 2P(T_{100-1} < -\lvert t^*\rvert)$=0.0000003647641
Step Five	p-value $< 0.05 \longrightarrow$ reject H_0

10.3.7 Calculation in R

Stock R Command

For the research question considered above, we can ask `R` to calculate the key values.

```
> t.test(x = orangeCrabs$CW,mu = 34,alternative = "two.sided",conf.level = 0.95)

One Sample t-test

data:  orangeCrabs$CW
t = 5.4529, df = 99, p-value = 3.648e-07
alternative hypothesis: true mean is not equal to 34
95 percent confidence interval:
36.61572 39.60828
sample estimates:
mean of x
38.112
```

Here, the inputs are as follows.

$$\begin{aligned} \text{x} &= \text{the data as stored in R} \\ \mu &= \text{testing value} \\ \text{alternative} &= \text{sign in } H_a \text{“greater”, “less”, “two.sided”} \\ \text{conf.level} &= \text{desired level of confidence} = 1-\alpha \end{aligned}$$

The output of this function can be mapped onto our desired output for a t hypothesis test as follows.

$$\begin{aligned} \bar{x} &= \text{mean of x} = 38.112 \\ t^* &= \text{t} = 5.4529 \\ p\text{-value} &= 3.648 \times 10^{-07} \end{aligned}$$

10.4 Hypothesis Testing for a Population Variance

...Empty for now...

10.5 Hypothesis Testing for a Population Proportion

As a motivating example consider the data recorded in a study described in Nelson et al. (2016). This observational study explored data on over 600,000 women from various studies with varied recruitment, and randomization. Among these data were 10,000 women aged 40-49 who were not regularly screened for breast cancer. Thirty-six fatalities were observed among these women.

In 2009, the U.S. Preventive Services Task Force recommended biennial mammography screening for women aged 50 to 74 years and screening those with increased risk or family history for those aged 40 to 49 years. Of interest to the researchers is to review the merit of the suggestion that average-risk women should wait until 50 years of age to start mammogram testing.

If it is the case that it is well known that 0.33% of women between the ages of 40 to 49 years who are regularly screened die of breast cancer, is there evidence that women between the ages of 40 to 49 years who are not regularly screened have a higher population proportion of fatality based on the available data?

First, we can recognize this as a binomial experiment. We can think of each woman in the group of 10,000 as a Bernoulli trial – they either die of breast cancer or they don't. Now, however, we don't know the probability of the event of interest (p); e.g., we don't know the probability that a randomly selected woman between the ages of 40 to 49 years who is not regularly screened dies of breast cancer.

Realizing that our sample is a collection of Bernoulli trials that make up a binomial experiment with success probability p we will conduct a hypothesis test about the population proportion.

Remark: Because this data is not quantitative asking questions about the mean or median does not make sense.

Preview: The goal of our hypothesis test for the population proportion is to assess the claim that women between the ages of 40 to 49 years who are not regularly screened have a higher population proportion of fatality.

10.5.1 Develop a Hypothesis

It is usually instructive to think about the proposed answer to the research question when trying to choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : p = p_0 & \text{or} & H_0 : p = p_0 \\ H_a : p < p_0 & & H_a : p > p_0 \end{array} \qquad \text{or} \qquad H_0 : p = p_0 \quad H_a : p \neq p_0.$$

Research Question: Was the decision in 2009 by the U.S. Preventive Services Task Force to recommended biennial mammography screening of average-risk women should wait until 50 years of age to start mammogram testing acceptable?

Hypothesis: No, women between the ages of 40 to 49 years who are not regularly screened have a higher population proportion of fatality than 0.33%.

The hypothesis instructs us to consider the alternative hypothesis

$$H_a : p > 0.0033,$$

where we take p_0 is the testing value of interest, 0.33%. We then take the sharp null hypothesis of

$$H_0 : p = 0.0033.$$

For our researchers, the hypothesis of interest is

$$\begin{aligned} H_0 &: p = 0.0033 \\ H_a &: p > 0.0033 \end{aligned}$$

where

$p =$ the population proportion of fatality among women between the ages of 40 to 49 years who are not regularly screened. This is unknown.

$\widehat{p} =$ the sample proportion of fatality among women between the ages of 40 to 49 years who are not regularly screened. This is calculated on the sample.

10.5.2 Assumptions

When completing a hypothesis test for the population proportion we have to check our assumptions for using such methodology.

1. The variable of interest is categorical.
2. The sample is generalizable to the population of interest
3. The sample size requirements for Central Limit Theorem are met to show that the sampling distribution of $\widehat{p}$ is approximately normal.

For our researchers,

1. The fatality status of each participant is categorical – the patient either died of breast cancer or did not.

2. The sample was obtained with varied recruitment and randomization.

3. Showing that we can use Central Limit Theorem requires us to lean on the assumption that the null hypothesis is true; e.g., $p = p_0$. Thus,

$$np = np_0 = 10000(0.0033) = 33 > 15$$
$$n(1-p) = n(1-p_0) = 10000(1-0.0033) = 99967 > 15$$

indicates that the Central Limit Theorem can be invoked to say

$$\widehat{p} \sim \mathcal{AG}\left(\mu_{\widehat{p}} = p, \sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}\right).$$

There are two methods for calculating $\sigma_{\widehat{p}}$ since p is unknown. Due to this, we will discuss two methods for testing a hypothesis about a population proportion.

For a **Score Hypothesis Test** we estimate p with the assumed population proportion following what we did to check the assumptions; e.g.,

$$\sigma_{\widehat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Instead, for a **Wald Hypothesis Test**, we listen to the data we've collected and estimate p with $\widehat{p}$ as calculated on the sample; e.g.,

$$\sigma_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

The Wald test statistic approximates the true population proportion by using the sample proportion and the score statistic assumes the testing value in the null hypothesis is the true population proportion. The score statistic is known to have better properties; i.e., it possesses a true significance level which is often closer to α.

Summary: We recommend using the score test for hypothesis testing.

For our researchers, we take

$$\widehat{p} \sim N\left(\mu_{\widehat{p}} = p = p_0 = 0.0033, \sigma_{\widehat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.0033(1-0.0033)}{10000}} \approx 0.000574\right) \quad \textbf{[Score]}$$

$$\widehat{p} \sim N\left(\mu_{\widehat{p}} = p = p_0 = 0.0033, \sigma_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.0036(1-0.0036)}{10000}} \approx 0.000599\right) \quad \textbf{[Wald].}$$

Note: The increased value of the standard error for the Wald test means that the Wald test will require more extreme observations to be considered as evidence against the null hypothesis compared to the Score test.

10.5.3 Test Statistics

The test statistic for a score hypothesis test for a population proportion is

$$z_s = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The test statistic for a Wald hypothesis test for a population proportion is

$$z_w = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}.$$

Both of these test statistics estimate the number of standard deviations the observed sample proportion is from p_0 which we assume to be the true population proportion. Each test uses a different estimate for the standard error and will yield slightly different solutions.

For our researchers, the test statistics for the sample of $n = 10000$ women between the ages of 40 to 49 years who were not regularly screened are

$$z_s = \frac{0.0036 - 0.0033}{0.000574} = 0.5226481,$$

and

$$z_w = \frac{0.0036 - 0.0033}{0.000599} = 0.5008347.$$

10.5.4 P-values

Recall that a p-value provides the probability of the observed data, or more extreme, when the null hypothesis is true. The terminology "more extreme" follows from the alternative hypothesis. Thus, we can calculate the p-value as follows.

$$\begin{aligned} p\text{-value} &= P(\text{observed statistic or more extreme} | H_0 \text{ is true}) \\ &= P(\widehat{p} > 0.0036 | p = p_0 = 0.0033) \end{aligned}$$

```
> x<-36
> n<-10000
> p0<-0.0033
> p.hat<-x/n
> alpha<-0.05
> ##########################
> ## Wald
> ##########################
> mu.phat<-p0
> se.phat<-sqrt(p0*(1-p0)/n)
> pnorm(q=p.hat,mean=mu.phat,sd=se.phat,lower.tail=FALSE)
[1] 0.3004534
> z.obs<-(p.hat-mu.phat)/se.phat  ##test statistic
> pnorm(z.obs,lower.tail=FALSE)   ##p value
[1] 0.3004534
```

```
> ##########################
> ## Score
> ##########################
> mu.phat<-p0
> se.phat<-sqrt(p.hat*(1-p.hat)/n)
> pnorm(q=p.hat,mean=mu.phat,sd=se.phat,lower.tail=FALSE)
[1] 0.3082199
> z.obs<-(p.hat-mu.phat)/se.phat  ##test statistic
> pnorm(z.obs,lower.tail=FALSE)  ##p value
[1] 0.3082199
```

This is visualized in Figure 10.5.3, which is created with the following R code.

```
> ##########################
> ## Wald
> ##########################
> ggdat<-data.frame(z=seq(-4.5,4.5,0.01),
+                   f=dnorm(x=seq(-4.5,4.5,0.01)))
> ggdat.highlight<-data.frame(x=z.obs,y=0)
> #Save the observations corresponding to z
> axis.labels<-c(round(c(0,qnorm(1-alpha))*se.phat+mu.phat,3),
+                round(p.hat,4))
> ggplot(data=ggdat,aes(x=z,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,z>=qnorm(1-alpha)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,z>=z.obs),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept = 0)+
+   theme_bw()+
+   xlab("z")+
+   ylab("Density")+
+   ggtitle("Wald Test for the Population Proportion",
+           subtitle=bquote(H[0]*":"~p==0.0033*", versus "*H[a]*":"~p>0.0033))+
+   annotate("text", x=2.5, y=0.07,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text", x=0, y=0.025, label="Observation",size=3.5)+
+   annotate("text", x=1.8, y=0.25,
+            label=paste("P-value=",round(pnorm(z.obs,lower.tail=FALSE),4),
+                        sep=" "),size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+     breaks=c(0,qnorm(1-alpha),z.obs),
+     labels = axis.labels,name="Proportion of Fatality Due to Breast Cancer"))
> ##########################
> ## Score
> ##########################
> ggdat<-data.frame(z=seq(-4.5,4.5,0.01),
+                   f=dnorm(x=seq(-4.5,4.5,0.01)))
> ggdat.highlight<-data.frame(x=z.obs,y=0)
```

```
> #Save the observations corresponding to z
> axis.labels<-c(round(c(0,qnorm(1-alpha))*se.phat+mu.phat,3),
+                round(p.hat,4))
> ggplot(data=ggdat,aes(x=z,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,z>=qnorm(1-alpha)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,z>=z.obs),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("z")+
+   ylab("Density")+
+   ggtitle("Score Test for the Population Proportion",
+           subtitle=bquote(H[0]*":"~p==0.0033*", versus "*H[a]*":"~p>0.0033))+
+   annotate("text", x=2.5, y=0.07,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text", x=0, y=0.025, label="Observation",size=3.5)+
+   annotate("text", x=1.8, y=0.25,
+            label=paste("P-value=", round(pnorm(z.obs,lower.tail=FALSE),4),
+                 sep=" "),size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+     breaks=c(0,qnorm(1-alpha),z.obs),
+     labels = axis.labels,name="Proportion of Fatality Due to Breast Cancer"))
```

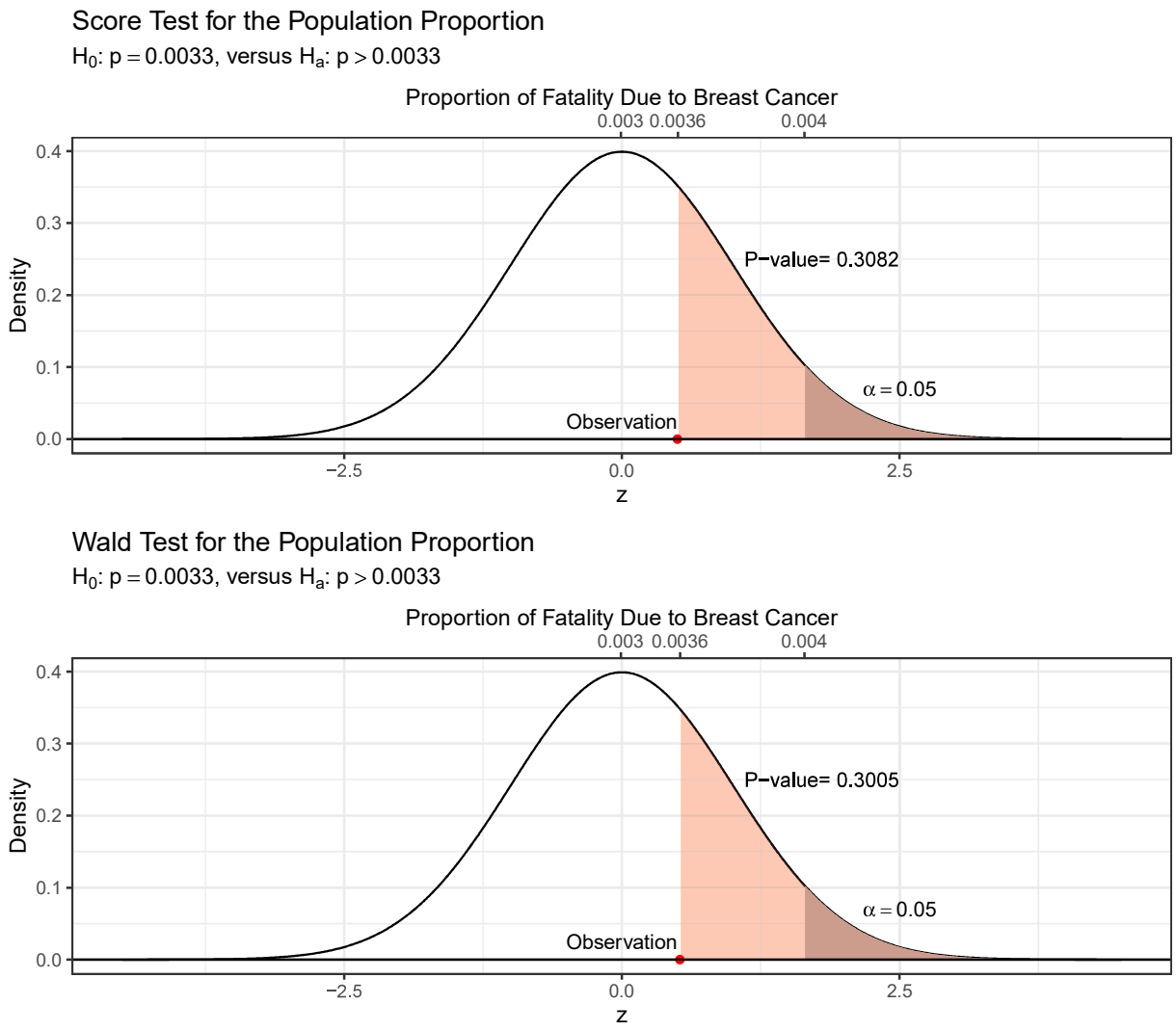

Figure 10.5.3: P-value as calculated in Section 10.5.4 for the Score (top) and the Wald (bottom). The p-values are shaded in red, the rejection region is shaded in grey and the observation is highlighted with a red point.

10.5.5 Decision Making

We reject the null hypothesis when the observed data is unusual under H_0. A prespecified significance level of $\alpha = 0.05$ dictates that we reject the null hypothesis when the probability of the observed data or more extreme under H_0 is less than α.

Figure 10.5.3 shows that the observed data does not fall in the rejection region for this test. While this is clear from the graph, we are curious about how to decide numerically whether or not the observed data is in the rejection region, noting we did this as we asked `R` to create the graph.

We can ask, and answer, this question three ways. We discuss the ways to answer the question below for a score hypothesis test; doing so for the Wald hypothesis test would be the same substituting the key values for that test. While we show all three below, we note that it is general practice to make this decision using the p-value.

Q1: Is $\widehat{p}$ in the rejection region?
A1: This question asks if the observed $\widehat{p} = 0.0036$ is in the top $\alpha \times 100$ percentile of $\widehat{p}$ under the null hypothesis. We can ask R for this percentile of $\widehat{p}$ using the sampling distribution.

```
> qnorm(p=0.05,mean=0.0033,sd=0.000574,lower.tail=FALSE)
[1] 0.004244146
```

Thus, we would reject the null hypothesis for any observation of $\widehat{p}$ greater than 0.004244146; this is termed the **critical $\widehat{p}$ value**.

For our researchers: Since $\widehat{p} = 0.0036$ is not in the rejection region, we fail to reject the null hypothesis.

Q2: Is the test statistic Z in the rejection region?
A2: This question asks if the observed $z_s = 0.5226481$ is in the top $\alpha \times 100$ percentile of Z under the null hypothesis. We can ask R for this percentile of $\widehat{p}$ using the sampling distribution.

```
> qnorm(p=0.05,mean=0,sd=1,lower.tail=FALSE)
[1] 1.644854
```

Thus, we would reject the null hypothesis for any observation of Z greater than 1.644854; this is termed the **critical z value**.

For our researchers: Since $z_s = 0.5226481$, it is not in the rejection region and so we fail to reject the null hypothesis.

Q3: Does the p-value suggest the observation lies in the rejection region?
A3: If the p-value is less than alpha then the observation must lie in the rejection region. Thus, we would reject the null hypothesis for any observation that has a p-value less than 0.05.

For our researchers: The p-value for this test, p-value=0.3006096, is greater than $\alpha = 0.05$ which, again, indicates that we should fail to reject the null hypothesis in favor of the alternative.

Conclusion: We say "there is not sufficient evidence to suggest that the population proportion of fatality among women between the ages of 40 to 49 years who are not regularly screened is greater than 0.0033 ($z_s = 0.5226$, p-value=0.3006)."

Remark: The answers to these three questions will always match. In practice, quantitative researchers report the test statistic and p-value when drawing their conclusions as in the conclusion above.

10.5.6 Summary

In this section, we succinctly summarize the five steps for the hypothesis test about a population proportion. While it might be simple to check these tables and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Note that in the summary below, we are neatly tying chapters worth of material together. Hypothesis testing, in a sense, isn't new material but a new task we can take on given our tools from previous chapters.

Step One	$H_a : p < p_0$	$H_a : p > p_0$	$H_a : p \neq p_0$
Step Two	Check Assumptions		
Step Three	$z_s = \frac{\widehat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ or $z_w = \frac{\widehat{p}-p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}$		
Step Four	$P(Z < z^*)$	$P(Z > z^*)$	$2P(Z < -\lvert z^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if z^* is in the rejection region		

For our researchers,

Step One	$H_0 : p = 0.0033$ $H_a : p > 0.0033$
Step Two	Sample is generalizable. Observations are independent. Central Limit Theorem assumptions satisfied.
Step Three	$z_s = \frac{\widehat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 0.5226481$
Step Four	p-value $= P(Z < z^*) = 0.3006096$
Step Five	p-value$> 0.05 \longrightarrow$ Fail to reject H_0.

10.5.7 Calculation in R

Stock R Command

For the research question considered above, we can ask `R` to calculate the key values of the score test as follows

```
> prop.test(x=36,n=10000,p=0.0033,alternative="greater",
+   conf.level=0.95,correct=FALSE)

1-sample proportions test without continuity correction

data:  36 out of 10000, null probability 0.0033
X-squared = 0.27363, df = 1, p-value = 0.3005
alternative hypothesis: true p is greater than 0.0033
95 percent confidence interval:
0.002740157 1.000000000
sample estimates:
p
0.0036
```

Here, the inputs are as follows.

$$\begin{aligned} \text{x} &= \text{number of observations of interest} \\ \text{n} &= \text{sample size} \\ \text{p} &= \text{testing value} \\ \text{alternative} &= \text{sign in } H_a \text{“greater”, “less”, “two.sided”} \\ \text{conf.level} &= \text{desired level of confidence} = 1-\alpha \\ \text{correct} &= \text{(discussed below)} \end{aligned}$$

The output of this function can be mapped onto our desired output for a score hypothesis test as follows. Note that any differences are due to our rounding in by-hand calculations; the calculations from `R` are more precise.

$$\begin{aligned} \widehat{p} &= p = 0.0036 \\ z_s &= \sqrt{\text{X-squared}} = \sqrt{0.27363} = 0.5230965 \\ p\text{-value} &= 0.3005 \end{aligned}$$

Definition 10.13. A **continuity correction** is often used when using a continuous distribution to approximate a discrete distribution. Here, we are approximating a discrete binomial distribution with a continuous normal distribution as outlined by the Central Limit Theorem. To correct for this we add or subtract a small number to the sample proportion.

For a hypothesis test about a population proportion we take the continuity correction, c to be the following.

$$c = \begin{cases} 0 & |\widehat{p} - p_0| < \frac{1}{2n} \\ \frac{-1}{2n} & \widehat{p} > p_0 \\ \frac{1}{2n} & \widehat{p} < p_0 \end{cases}$$

The test statistic is “corrected” by adding c to $\widehat{p}$.

$$z_s^* = \frac{(\widehat{p} + c) - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

For our researchers,

$$|\widehat{p} - p_0| = |0.0036 - 0.0033| = 0.0003 > \frac{1}{2(10000)} = 0.00005$$

and $\widehat{p} > p_0$. Thus, we take $c = \frac{-1}{2(10000)} = -0.00005$ and correct the test statistic as follows.

$$z_s^* = \frac{(\widehat{p} + c) - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.0036 - 0.00005) - 0.0033}{0.000574} = 0.4355401$$

The p-value for the test is then

$$p\text{-value} = P(Z > z_s^*) = P(Z > 0.4355401) = 0.3315852.$$

Again, we'd like to ask `R` for the corrected version, as the calculations from `R` are more precise.

```
> prop.test(x=36,n=10000,p=0.0033,alternative="greater",
+   conf.level=0.95,correct=TRUE)

1-sample proportions test with
continuity correction

data:  36 out of 10000, null probability 0.0033
X-squared = 0.19002, df = 1,
p-value = 0.3314
alternative hypothesis: true p is greater than 0.0033
95 percent confidence interval:
0.002696945 1.000000000
sample estimates:
p
0.0036
```

The output of the corrected score hypothesis test is as follows.

$$\begin{aligned} \widehat{p} = p &= 0.0036 \\ z_s &= \sqrt{\text{X-squared}} = \sqrt{0.19002} = 0.4359128 \\ p\text{-value} &= 0.3314 \end{aligned}$$

Curiosity: How much "better" does the continuity correction make the estimated probability?
Answer: To answer this question, we can complete an **exact test**. An exact test uses the binomial distribution, not a continuous approximation of it. In the past, when the sample size was large it was difficult to calculate the exact test and so the Gaussian approximation was preferred and proved to be quite accurate with or without correcting.

Now that computation is fast and easy, one might wonder why we don't use the exact test. A drawback of doing this for discrete distributions is that for any given sample size n there are certain values of for significance levels that cannot be obtained exactly – this issue is ameliorated as the sample size grows.

In R we can ask for the exact test as follows.

```
> binom.test(x=36,n=10000,p=0.0033,alternative="greater",conf.level=0.95)

Exact binomial test

data:  36 and 10000
number of successes = 36, number of trials = 10000, p-value = 0.323
alternative hypothesis: true probability of success is greater than 0.0033
95 percent confidence interval:
0.002674223 1.000000000
sample estimates:
probability of success
0.0036
```

The output of the corrected score hypothesis test is as follows.

$$\widehat{p} = p = 0.0036$$
$$p\text{-value} = 0.323$$

Summary: For our motivating example we've conducted a hypothesis test about the population proportion a variety of ways.

Hypothesis Test	Test Statistic	p-value	
Wald	$z_w = 0.5008347$	0.3006096	
Score	$z_s = 0.5230968$	0.3082437	
	$z_s = 0.5230965$	0.3005	(R)
Score (corrected)	$z_s^* = 0.4355401$	0.3315852	
	$z_s^* = 0.4359128$	0.3314	(R)
Exact	–	0.323	(R)

10.5.8 Simulation

Recall that the significance level for a hypothesis test, α, specifies how unusual an observation is required to be for consideration as evidence against the null hypothesis. We choose this value when conducting a hypothesis, generally to be 0.05, and we reject the null when we make an "unusual" observation.

Result: Even when the null hypothesis is true, we expect to reject the null roughly 5% of the time. This expectation is due to specifying $\alpha = 0.05$.

Recall that the **power** of a hypothesis test is defined as the probability that we reject a null hypothesis when, in fact, we should. An example, to evaluate the hypothesis testing procedures discussed is to consider testing the following hypothesis.

$$H_0 : p = 0.75$$
$$H_a : p < 0.75$$

By simulating data we can check how often we reject the null hypothesis when $p = 0.75$ is true (significance), and how often we reject the null hypothesis when $p = 0.72$, $p = 0.65$, or $p = 0.50$ (power).

The simulation results for the case when $p = 0.75$ is true, provide us information about how well the different procedures produce the desired level of significance.

n	Wald	Wald Corrected	Score	Score Corrected	Exact
3	0.014	0.014	0.156	0.014	0.014
5	0.015	0.015	0.101	0.015	0.015
15	0.055	0.017	0.055	0.017	0.017
30	0.053	0.022	0.053	0.053	0.022
50	0.028	0.028	0.056	0.028	0.028
100	0.045	0.028	0.045	0.045	0.045
1000	0.047	0.047	0.054	0.047	0.047
10000	0.043	0.041	0.043	0.043	0.043

Table 10.5.2: Empirical significance levels of hypothesis tests for a population proportion at the 0.05 significance level.

The simulation results for the case when $p \neq 0.75$, provide us information about how well the different procedures are able to pick up the difference in the population proportion based on the sample. Below, we notice that as the true population proportion moves further from the testing value, we can detect the differences with smaller sample sizes.

n	Wald	Wald Corrected	Score	Score Corrected	Exact
3	0.022	0.022	0.188	0.022	0.022
5	0.024	0.024	0.139	0.024	0.024
15	0.095	0.033	0.095	0.033	0.033
30	0.110	0.052	0.110	0.110	0.052
50	0.081	0.081	0.135	0.081	0.081
100	0.160	0.113	0.160	0.160	0.160
1000	0.687	0.687	0.711	0.687	0.687
10000	1.000	1.000	1.000	1.000	1.000

Table 10.5.3: Empirical power of hypothesis tests for a population proportion at the 0.05 significance level taking $p = 0.72$.

n	Wald	Wald Corrected	Score	Score Corrected	Exact
3	0.042	0.042	0.271	0.042	0.042
5	0.051	0.051	0.227	0.051	0.051
15	0.248	0.118	0.248	0.118	0.118
30	0.339	0.218	0.339	0.339	0.218
50	0.370	0.370	0.482	0.370	0.370
100	0.696	0.619	0.696	0.696	0.696
1000	1.000	1.000	1.000	1.000	1.000
10000	1.000	1.000	1.000	1.000	1.000

Table 10.5.4: Empirical power of hypothesis tests for a population proportion at the 0.05 significance level taking $p = 0.65$.

n	Wald	Wald Corrected	Score	Score Corrected	Exact
3	0.123	0.123	0.493	0.123	0.123
5	0.189	0.189	0.505	0.189	0.189
15	0.698	0.494	0.698	0.494	0.494
30	0.900	0.816	0.900	0.900	0.816
50	0.965	0.965	0.982	0.965	0.965
100	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000
10000	1.000	1.000	1.000	1.000	1.000

Table 10.5.5: Empirical power of hypothesis tests for a population proportion at the 0.05 significance level taking $p = 0.50$.

Conclusion: The score hypothesis test almost always returns higher power than any of the other results, but provides a higher than specified significance level for small sample sizes. To account for this, we recommend using the continuity correction when $np_0 < 5$ or $n(1 - p_0) < 5$.

10.6 Hypothesis Testing for a Population Median: Sign Test

As a motivating example consider the data explored by Analytical Methods Committee (1989). The data considered contained $n = 24$ observations of trace amounts of copper in wholemeal flour reported in micro grams of copper per gram of flour ($\mu g\ g^{-1}$).

Suppose that this is a sample of $n = 24$ from a large batch of flour produced. Further suppose that the flour producer is concerned that the this batch of flour yields more than a typical sample of flour; that is, it has more than 4 $\mu g\ g^{-1}$ of copper; they might consider this as part of their quality analysis.

These data are available in the `MASS` package.

```
> library(MASS)
> data(chem)
> copper_amnt<-chem
>  copper_amnt
[1]  2.90  3.10  3.40  3.40  3.70  3.70  2.80  2.50  2.40  2.40  2.70  2.20
[13]  5.28  3.37  3.03  3.03 28.95  3.77  3.40  2.20  3.50  3.60  3.70  3.70
```

A histogram of these data, seen in Figure 10.6.4, is heavily skewed.

```
> ggdat<-data.frame(copper_amnt=copper_amnt)
> ggplot(data=ggdat,aes(x=copper_amnt))+
+   geom_histogram(aes(y=..density..),
+                  bins=20,color="black",fill="lightblue")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote("Amount of Copper ("*mu*g/g*")"))+
+   ylab("Density")+
+   ggtitle("Copper in Wholemeal Flour")
```

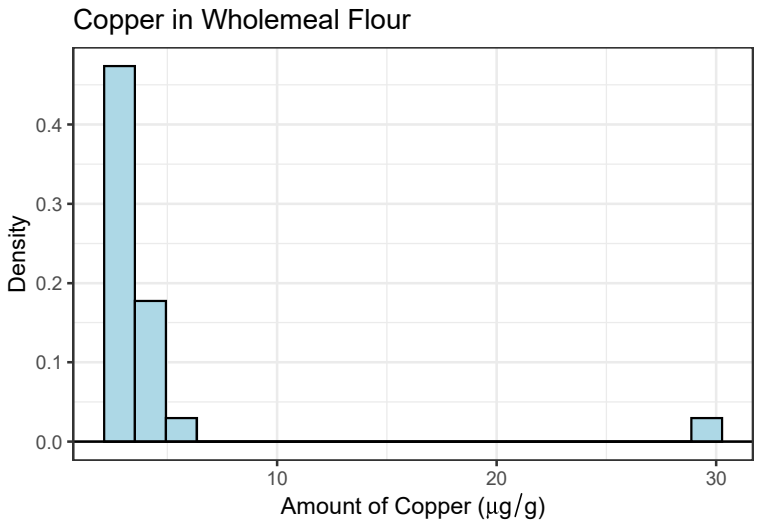

Figure 10.6.4: A histogram of the copper data.

Since our number of observations is quite small we cannot use the Central Limit Theorem with much assurance, particularly due to the heavy skew in the histogram of the data. While the t test is robust to the normality assumption, we can see that we might be more interested in the median as a measure of typical.

Preview: The goal of our hypothesis test for the population median is to assess the claim that the median amount of copper in wholemeal flour is greater than 4 $\mu g\ g^{-1}$.

10.6.1 Develop a Hypothesis

Again, it is instructive to think about the proposed answer to the research question when trying to choose between the three sets of hypotheses,

$$\begin{array}{lll} H_0 : M = M_0 & \text{or} & H_0 : M = M_0 \\ H_a : M < M_0 & & H_a : M > M_0 \end{array} \qquad \text{or} \qquad \begin{array}{l} H_0 : M = M_0 \\ H_a : M \neq M_0. \end{array}$$

Research Question: Does this batch of wholemeal flour pass the quality standards imposed by the producer?

Hypothesis: Yes, the median amount of copper in flour is less than 4 $\mu g\ g^{-1}$.

The hypothesis instructs us to consider the alternative hypothesis

$$H_a : M > 4,$$

where we take $\mu_0 = 4$. We then take the sharp null hypothesis of

$$H_0 : M = 4.$$

For our researchers, the hypothesis of interest is

$$\begin{aligned} H_0 &: M = 4 \\ H_a &: M > 4 \end{aligned}$$

where

M = the population median median amount of copper in wholemeal flour ($\mu g\ g^{-1}$).This is unknown.

$\widehat{m}$ = the population median median amount of copper in wholemeal flour ($\mu g\ g^{-1}$). This is calculated on the sample.

10.6.2 Assumptions

The assumptions for a sign test are

1. the variable of interest is quantitative
2. the sample is generalizable to the population of interest.

For this test it is not required to check any Central Limit Theorem assumptions. This hypothesis test can be used regardless of the sampling distribution of the original data.

10.6.3 Test Statistics

Here, we assume that the population median is M_0 and thus

$$P(X < M_0) = 0.50$$
$$P(X > M_0) = 0.50,$$

recalling that the median is the middle of a distribution.

Consider each observation to be a Bernoulli trial,

$$Y_i = I(X_i > M_0).$$

Each Y_i is 1 if the observation is greater than the assumed population median M_0 and 0 otherwise. It follows that the number of observations greater than M_0, B, is distributed binomial; e.g.,

$$B = \sum_{i=1}^{n} I(X_i > M_0) \sim \text{Binomial}(n,0.50).$$

The test statistic can then be calculated using the data as

$$b^* = \sum_{i=1}^{n} I(x_i > M_0).$$

For our researchers, the test statistic for the sample of $n = 24$ observations of trace amounts of copper in wholemeal flour reported in parts per million is

$$b^* = \sum_{i=1}^{n} I(x_i > 4) = 2.$$

Note, this equation just counts the number of observations larger than 4 $\mu g\ g^{-1}$ in the dataset. If the true population median was 4 $\mu g\ g^{-1}$, we would expect roughly half of the observations (12 of them) to be greater than 4 $\mu g\ g^{-1}$.

For a small dataset like this one, we can simply count the number of observations ourselves, but it is important to know how to ask R to do this when the dataset is much larger. The following code asks R to create a vector of the observations greater than 4 $\mu g\ g^{-1}$, and then to return the length of that vector.

```
> length(which(copper_amnt>4))
[1] 2
```

Remark: This test is useful because it is not dependent on the exact observations but whether or not each observation is larger than a specified value. This is important when analyzing data with outliers. This weakens the effect that extreme observations, like the observation of 28.95 here but at the cost of power; e.g., a decreased probability of rejecting the null hypothesis when we should.

10.6.4 P-values

Recall that a p-value provides the probability of the observed data, or more extreme, when the null hypothesis is true. The terminology "more extreme" follows from the alternative hypothesis. Here, we are interested in the probability of having two or more observations greater than 4 $\mu g\ g^{-1}$.

Thus, we can calculate the p-value as follows.

$$\begin{aligned} p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\ &= P(B \geq 2|M = M_0 = 4) \\ &= 1 - P(B < 2|M = M_0 = 4) \\ &= 1 - P(B \leq 1|M = M_0 = 4) \end{aligned}$$

```
> m0<-4
> (B.obs<-length(which(copper_amnt>m0))) ##test statistic
[1] 2
> (p.value<-1-pbinom(q=(2-1),size=n,prob=0.50)) ##p value
[1] 0.9999985
```

This is visualized in Figure 10.6.4, which is created with the following R code.

```
> ggdat<-data.frame(x=(0:24),
+                   f=dbinom(x=(0:24),size=n,prob=0.50))
> ggdat.highlight.pvalue<-data.frame(x=2:24,f=dbinom(x=2:24,size=n,prob=0.50))
> ggdat.highlight.observation<-data.frame(x=2,y=0)
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f), ymin=0)+
+   geom_linerange(data=ggdat.highlight.pvalue,aes(x=x,ymax=f),color="red", ymin=0)+
+   geom_point(data=ggdat.highlight.observation,aes(x=x,y=y),color="red")+
+   geom_ribbon(data=subset(ggdat,x>=qbinom(1-alpha,size=n,prob=0.50)),aes(ymax=f),ymin=0,
```

```
+                fill="grey",color=NA,alpha=0.5)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Observations Larger than the Median")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Sign Test for the Population Median PMF",
+           subtitle=bquote(H[0]*":"~M==4*", versus "*H[a]*":"~M>4))+
+   annotate("text", x=19, y=0.025,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text", x=B.obs, y=0.025, label="Observation",size=3.5)+
+   annotate("text", x=17, y=0.125, label="P-value=0.9999",size=3.5)
```

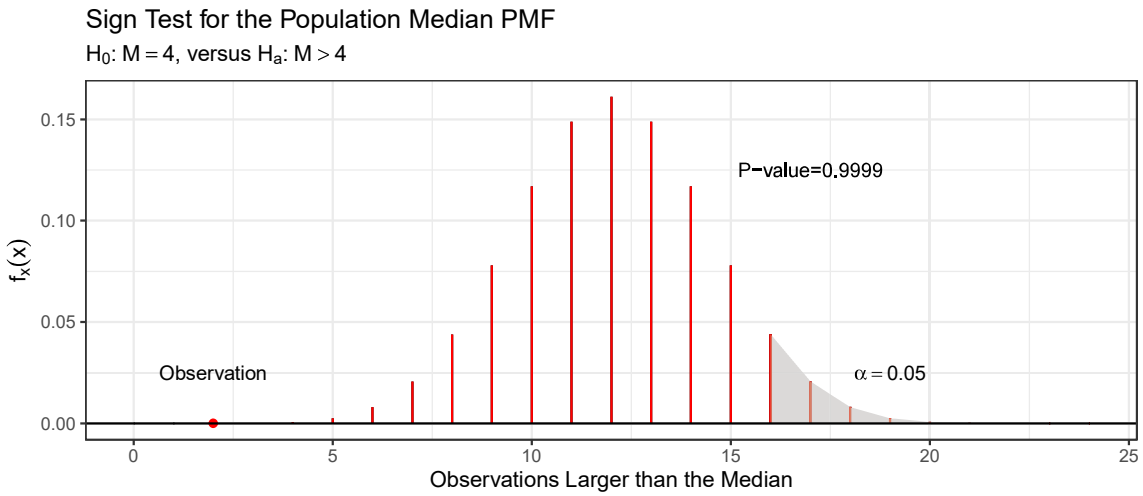

Figure 10.6.5: P-value as calculated in Section 10.6.4 for the sign test.

10.6.5 Decision Making

We reject the null hypothesis when the observed data is unusual under H_0. A prespecified significance level of $\alpha = 0.05$ dictates that we reject the null hypothesis when the probability of the observed data or more extreme under H_0 is less than α.

Figure 10.6.5 shows that the observed data doesn't fall in the rejection region for this test. While this is clear from the graph, we are curious about how to decide this numerically whether or not the observed data is in the rejection region, noting we did this as we asked R to create the graph.

We can ask, and answer, this questions two ways. We discuss the variety of ways to answer the question below for a sign test. While we show both ways below, we note that it is general practice to make this decision using the p-value.

Q1: Is the test statistic b^* in the rejection region?
A2: This question asks if the observed $b^* = 2$ is in the top $\alpha \times 100$ percentile of the binomial$(n = 24, p = 0.50)$ distribution under the null hypothesis. We can ask R for this percentile of B using the sampling distribution.

```
> qbinom(p=0.95,size=24,prob=0.50)
[1] 16
```

Thus, we would reject the null hypothesis for any observation of B greater than 16.
For our researchers: Since $b^* = 2$, it is not in the rejection region and so we fail to reject the null hypothesis.

Q3: Does the p-value indicate that the observation lies in the rejection region?
A3: If the p-value is less than alpha then the observation must lie in the rejection region. Thus, we would reject the null hypothesis for any observation that has a p-value less than 0.05.
For our researchers: The p-value for this test, p-value=0.9999985, is greater than $\alpha = 0.05$ which indicates that we should fail to reject the null hypothesis.

Remark: We don't ask if $\widehat{m}$ is in the rejection region because we don't know the sampling distribution of $\widehat{m}$.

Conclusion: We say "there is not sufficient evidence to suggest that the population median amount of copper in wholemeal flour is greater than 4 $\mu g\ g^{-1}$ ($b = 2$, p-value≈ 1)."

Remark: The answers to these two questions will always match. In practice, quantitative researchers report the test statistic and p-value when drawing their conclusions.

10.6.6 Summary

In this section we succinctly summarize the five steps for the sign Hypothesis Test. While it might be simple to check this table and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Note that in the summary below, we are neatly tying chapters worth of material together. Hypothesis testing, in a sense, isn't new material but a new task we can take on given our tools from Chapters 1-5.

Step One	$H_a : M < M_0$	$H_a : M > M_0$	$H_a : M \neq M_0$
Step Two	Check Assumptions		
Step Three	$b^* = \sum_{i=1}^{n} I(x_i > M_0)$		
Step Four	$P(B \leq b^*)$	$P(B \geq b^*)$	$1 - P(B \in [n - t*, b^*])$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

For our researchers,

Step One	$H_a : M > 4$
Step Two	Sample may be generalizable – need more information. Observations may be independent – need more information.
Step Three	$b^* = 2$
Step Four	p-value $= P(B \geq b^*)$=0.9999985
Step Five	p-value $> 0.05 \longrightarrow$ fail to reject H_0

10.6.7 Calculation in R

Stock R Command

For the research question considered above, we can ask R to calculate the key values using the "BSDA" package.

```
> install.packages("BSDA")
> libary("BSDA")
> SIGN.test(x = copper_amnt,md=4,alternative="greater",conf.level=0.95)
One-sample Sign-Test

data:  copper_amnt
s = 2, p-value = 1
alternative hypothesis: true median is greater than 4
95 percent confidence interval:
2.953506      Inf
sample estimates:
median of x
3.385
```

Here, the inputs are as follows.

$$\begin{aligned} \text{x} &= \text{the data as stored in R} \\ \text{md} &= \text{testing value} \\ \text{alternative} &= \text{sign in } H_a\text{"greater", "less", "two.sided"} \\ \text{conf.level} &= \text{desired level of confidence} = 1-\alpha \end{aligned}$$

The output of this function can be mapped onto our desired output for a sign hypothesis test as follows. Note that any differences are due to our rounding.

$$\begin{aligned} \widehat{m} &= \text{median of x} = 3.385 \\ b^* &= \text{s} = 2 \\ p\text{-value} &= 1 \end{aligned}$$

Part II

Multivariate Statistics with R