16.3: The Process of Null Hypothesis Testing

We can break the process of null hypothesis testing down into a number of steps:

1. Formulate a hypothesis that embodies our prediction (*before seeing the data*)
2. Collect some data relevant to the hypothesis
3. Specify null and alternative hypotheses
4. Fit a model to the data that represents the alternative hypothesis and compute a test statistic
5. Compute the probability of the observed value of that statistic assuming that the null hypothesis is true
6. Assess the "statistical significance" of the result

For a hands-on example, let's use the NHANES data to ask the following question: Is physical activity related to body mass index? In the NHANES dataset, participants were asked whether they engage regularly in moderate or vigorous-intensity sports, fitness or recreational activities (stored in the variable $PhysActive$). The researchers also measured height and weight and used them to compute the *Body Mass Index* (BMI):

$$BMI = \frac{weight(kg)}{height(m)^2}$$

16.3.1 Step 1: Formulate a hypothesis of interest

For step 1, we hypothesize that BMI is greater for people who do not engage in physical activity, compared to those who do.

16.3.2 Step 2: Collect some data

For step 2, we collect some data. In this case, we will sample 250 individuals from the NHANES dataset. Figure 16.1 shows an example of such a sample, with BMI shown separately for active and inactive individuals.

Table 16.1: Summary of BMI data for active versus inactive individuals

PhysActive	N	mean	sd
No	131	30	9.0
Yes	119	27	5.2

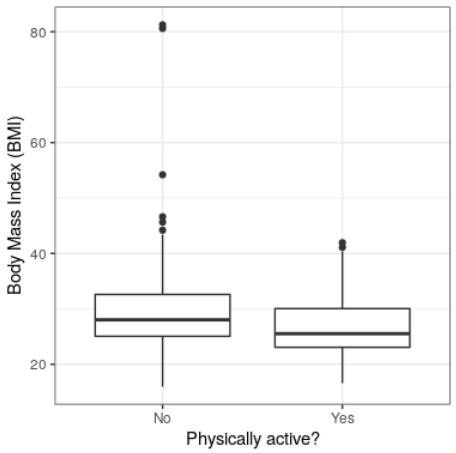

Figure 16.1: Box plot of BMI data from a sample of adults from the NHANES dataset, split by whether they reported engaging in regular physical activity.

16.3.3 Step 3: Specify the null and alternative hypotheses

For step 3, we need to specify our null hypothesis (which we call H_0) and our alternative hypothesis (which we call H_A). H_0 is the baseline against which we test our hypothesis of interest: that is, what would we expect the data to look like if there was no effect? The null hypothesis always involves some kind of equality (=, $\leq$, or $\geq$). H_A describes what we expect if there actually is an effect.

The alternative hypothesis always involves some kind of inequality ($\neq$, >, or <). Importantly, null hypothesis testing operates under the assumption that the null hypothesis is true unless the evidence shows otherwise.

We also have to decide whether to use *directional* or *non-directional* hypotheses. A non-directional hypothesis simply predicts that there will be a difference, without predicting which direction it will go. For the BMI/activity example, a non-directional null hypothesis would be:

$H0 : BMI_{active} =$

and the corresponding non-directional alternative hypothesis would be:

$HA : BMI_{active} \neq BMI_{inactive}$

A directional hypothesis, on the other hand, predicts which direction the difference would go. For example, we have strong prior knowledge to predict that people who engage in physical activity should weigh less than those who do not, so we would propose the following directional null hypothesis:

$H0 : BMI_{active} \geq BMI_{inactive}$

and directional alternative:

$HA : BMI_{active} <$

As we will see later, testing a non-directional hypothesis is more conservative, so this is generally to be preferred unless there is a strong *a priori* reason to hypothesize an effect in a particular direction. Any direction hypotheses should be specified prior to looking at the data!

16.3.4 Step 4: Fit a model to the data and compute a test statistic

For step 4, we want to use the data to compute a statistic that will ultimately let us decide whether the null hypothesis is rejected or not. To do this, the model needs to quantify the amount of evidence in favor of the alternative hypothesis, relative to the variability in the data. Thus we can think of the test statistic as providing a measure of the size of the effect compared to the variability in the data. In general, this test statistic will have a probability distribution associated with it, because that allows us to determine how likely our observed value of the statistic is under the null hypothesis.

For the BMI example, we need a test statistic that allows us to test for a difference between two means, since the hypotheses are stated in terms of mean BMI for each group. One statistic that is often used to compare two means is the *t-statistic*, first developed by the statistician William Sealy Gossett, who worked for the Guiness Brewery in Dublin and wrote under the pen name “Student” - hence, it is often called “Student’s t-statistic”. The t-statistic is appropriate for comparing the means of two groups when the sample sizes are relatively small and the population standard deviation is unknown. The t-statistic for comparison of two independent groups is computed as:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the means of the two groups, S_1^2 and S_2^2 are the estimated variances of the groups, and n_1 and n_2 are the sizes of the two groups. Note that the denominator is basically an average of the standard error of the mean for the two samples. Thus, one can view the the t-statistic as a way of quantifying how large the difference between groups is in relation to the sampling variability of the means that are being compared.

The t-statistic is distributed according to a probability distribution known as a *t* distribution. The *t* distribution looks quite similar to a normal distribution, but it differs depending on the number of degrees of freedom, which for this example is the number of observations minus 2, since we have computed two means and thus given up two degrees of freedom. When the degrees of freedom are large (say 1000), then the *t* distribution looks essentialy like the normal distribution, but when they are small then the *t* distribution has longer tails than the normal (see Figure 16.2).

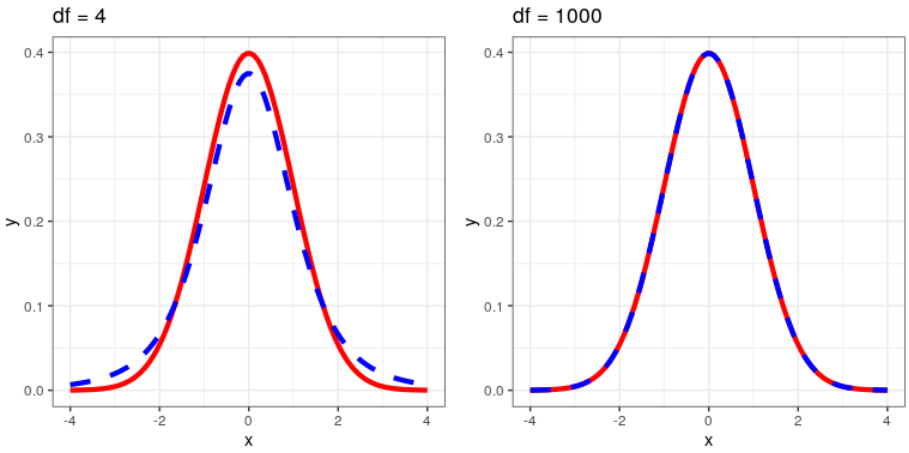

Figure 16.2: Each panel shows the t distribution (in blue dashed line) overlaid on the normal distribution (in solid red line). The left panel shows a t distribution with 4 degrees of freedom, in which case the distribution is similar but has slightly wider tails. The right panel shows a t distribution with 1000 degrees of freedom, in which case it is virtually identical to the normal.

16.3.5 Step 5: Determine the probability of the data under the null hypothesis

This is the step where NHST starts to violate our intuition – rather than determining the likelihood that the null hypothesis is true given the data, we instead determine the likelihood of the data under the null hypothesis - because we started out by assuming that the null hypothesis is true! To do this, we need to know the probability distribution for the statistic under the null hypothesis, so that we can ask how likely the data are under that distribution. Before we move to our BMI data, let's start with some simpler examples.

16.3.5.1 Randomization: A very simple example

Let's say that we wish to determine whether a coin is fair. To collect data, we flip the coin 100 times, and we count 70 heads. In this example, $H_0 : P(heads) = 0.5$ and $H_A : P(heads) \neq 0.5$, and our test statistic is simply the number of heads that we counted. The question that we then want to ask is: How likely is it that we would observe 70 heads if the true probability of heads is 0.5. We can imagine that this might happen very occasionally just by chance, but doesn't seem very likely. To quantify this probability, we can use the *binomial distribution*:

$$P(X < k) = \sum_{i=0}^{k} \binom{N}{k} p^i (1-p)^{(n-i)}$$

This equation will tell us the likelihood of a certain number of heads or fewer, given a particular probability of heads. However, what we really want to know is the probability of a certain number or more, which we can obtain by subtracting from one, based on the rules of probability:

$$P(X \geq k) = 1 - P(X < k)$$

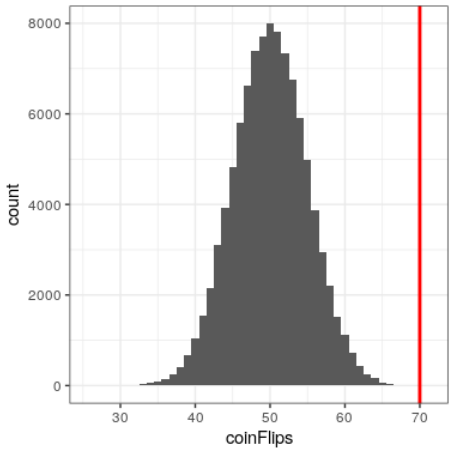

Figure 16.3: Distribution of numbers of heads (out of 100 flips) across 100,000 simulated runsl with the observed value of 70 flips represented by the vertical line.

We can compute the probability for our example using the `pbinom()` function. The probability of 69 or fewer heads given P(heads)=0.5 is 0.999961, so the probability of 70 or more heads is simply one minus that value (0.000039) This computation shows us that the likelihood of getting 70 heads if the coin is indeed fair is very small.

Now, what if we didn't have the `pbinom()` function to tell us the probability of that number of heads? We could instead determine it by simulation – we repeatedly flip a coin 100 times using a true probability of 0.5, and then compute the distribution of the number of heads across those simulation runs. Figure 16.3 shows the result from this simulation. Here we can see that the probability computed via simulation (0.000030) is very close to the theoretical probability (.00004).

Let's do the analogous computation for our BMI example. First we compute the t statistic using the values from our sample that we calculated above, where we find that (t = 3.86). The question that we then want to ask is: What is the likelihood that we would find a t statistic of this size, if the true difference between groups is zero or less (i.e. the directional null hypothesis)?

We can use the t distribution to determine this probability. Our sample size is 250, so the appropriate t distribution has 248 degrees of freedom because lose one for each of the two means that we computed. We can use the `pt()` function in R to determine the probability of finding a value of the t-statistic greater than or equal to our observed value. Note that we want to know the probability of a value greater than our observed value, but by default `pt()` gives us the probability of a value less than the one that we provide it, so we have to tell it explicitly to provide us with the "upper tail" probability (by setting `lower.tail = FALSE`). We find that (p(t > 3.86, df = 248) = 0.000), which tells us that our observed t-statistic value of 3.86 is relatively unlikely if the null hypothesis really is true.

In this case, we used a directional hypothesis, so we only had to look at one end of the null distribution. If we wanted to test a non-directional hypothesis, then we would need to be able to identify how unexpected the size of the effect is, regardless of its direction. In the context of the t-test, this means that we need to know how likely it is that the statistic would be as extreme in either the positive or negative direction. To do this, we multiply the observed *t* value by -1, since the *t* distribution is centered around zero, and then add together the two tail probabilities to get a *two-tailed* p-value: (p(t > 3.86 or t< -3.86, df = 248) = 0.000). Here we see that the p value for the two-tailed test is twice as large as that for the one-tailed test, which reflects the fact that an extreme value is less surprising since it could have occurred in either direction.

How do you choose whether to use a one-tailed versus a two-tailed test? The two-tailed test is always going to be more conservative, so it's always a good bet to use that one, unless you had a very strong prior reason for using a one-tailed test. In that case, you should have written down the hypothesis before you ever looked at the data. In Chapter 32 we will discuss the idea of pre-registration of hypotheses, which formalizes the idea of writing down your hypotheses before you ever see the actual data. You should *never* make a decision about how to perform a hypothesis test once you have looked at the data, as this can introduce serious bias into the results.

16.3.5.2 Computing p-values using randomization

So far we have seen how we can use the t-distribution to compute the probability of the data under the null hypothesis, but we can also do this using simulation. The basic idea is that we generate simulated data like those that we would expect under the null hypothesis, and then ask how extreme the observed data are in comparison to those simulated data. The key question is: How can

we generate data for which the null hypothesis is true? The general answer is that we can randomly rearrange the data in a particular way that makes the data look like they would if the null was really true. This is similar to the idea of bootstrapping, in the sense that it uses our own data to come up with an answer, but it does it in a different way.

16.3.5.3 Randomization: a simple example

Let's start with a simple example. Let's say that we want to compare the mean squatting ability of football players with cross-country runners, with $H_0 : \mu_{FB} \leq \mu_{XC}$ and $H_A : \mu_{FB} >$. We measure the maximum squatting ability of 5 football players and 5 cross-country runners (which we will generate randomly, assuming that $\mu_{FB} = 300$, $\mu_{XC} = 140$, and $\sigma = 30$).

Table 16.2: Squatting data for the two groups

group	squat
FB	335
FB	350
FB	230
FB	290
FB	325
XC	115
XC	115
XC	170
XC	175
XC	215

Table 16.2: Squatting data after randomly scrambling group labels

squat	scrambledGroup
335	FB
350	FB
230	XC
290	FB
325	FB
115	XC
115	XC
170	FB
175	XC
215	XC

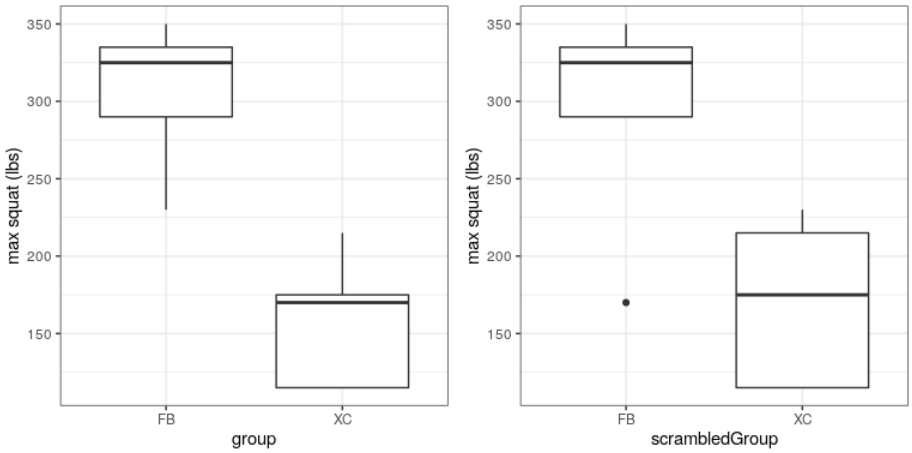

Figure 16.4: Left: Box plots of simulated squatting ability for football players and cross-country runners.Right: Box plots for subjects assigned to each group after scrambling group labels.

From the plot in Figure 16.4 it's clear that there is a large difference between the two groups. We can do a standard t-test to test our hypothesis, using the `t.test()` command in R, which gives the following result:

```
##
##  Two Sample t-test
##
## data:  squat by group
## t = 5, df = 8, p-value = 4e-04
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##   95 Inf
## sample estimates:
## mean in group FB mean in group XC
##              306              158
```

If we look at the p-value reported here, we see that the likelihood of such a difference under the null hypothesis is very small, using the *t* distribution to define the null.

Now let's see how we could answer the same question using randomization. The basic idea is that if the null hypothesis of no difference between groups is true, then it shouldn't matter which group one comes from (football players versus cross-country runners) – thus, to create data that are like our actual data but also conform to the null hypothesis, we can randomly reorder the group labels for the individuals in the dataset, and then recompute the difference between the groups. The results of such a shuffle are shown in Figure **??**.

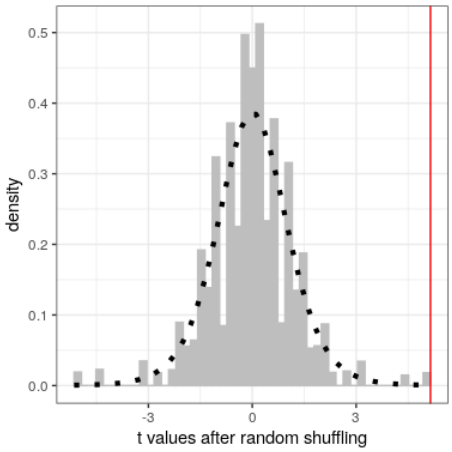

Figure 16.5: Histogram of t-values for the difference in means between the football and cross-country groups after randomly shuffling group membership. The vertical line denotes the actual difference observed between the two groups, and the dotted line shows the theoretical t distribution for this analysis.

After scrambling the labels, we see that the two groups are now much more similar, and in fact the cross-country group now has a slightly higher mean. Now let's do that 10000 times and store the t statistic for each iteration; this may take a moment to complete. Figure 16.5 shows the histogram of the t-values across all of the random shuffles. As expected under the null hypothesis, this distribution is centered at zero (the mean of the distribution is -0.016. From the figure we can also see that the distribution of t values after shuffling roughly follows the theoretical t distribution under the null hypothesis (with mean=0), showing that randomization worked to generate null data. We can compute the p-value from the randomized data by measuring how many of the shuffled values are at least as extreme as the observed value: $p(t > 5.14, df = 8)$ using randomization = 0.00380. This p-value is very similar to the p-value that we obtained using the t distribution, and both are quite extreme, suggesting that the observed data are very unlikely to have arisen if the null hypothesis is true - and in this case we *know* that it's not true, because we generated the data.

16.3.5.3.1 Randomization: BMI/activity example

Now let's use randomization to compute the p-value for the BMI/activity example. In this case, we will randomly shuffle the `PhysActive` variable and compute the difference between groups after each shuffle, and then compare our observed t statistic to the distribution of t statistics from the shuffled datasets. Figure 16.6 shows the distribution of t values from the shuffled samples, and we can also compute the probability of finding a value as large or larger than the observed value. The p-value obtained from randomization (0.0000) is very similar to the one obtained using the t distribution (0.0001). The advantage of the randomization test is that it doesn't require that we assume that the data from each of the groups are normally distributed, though the t-test is generally quite robust to violations of that assumption. In addition, the randomization test can allow us to compute p-values for statistics when we don't have a theoretical distribution like we do for the t-test.

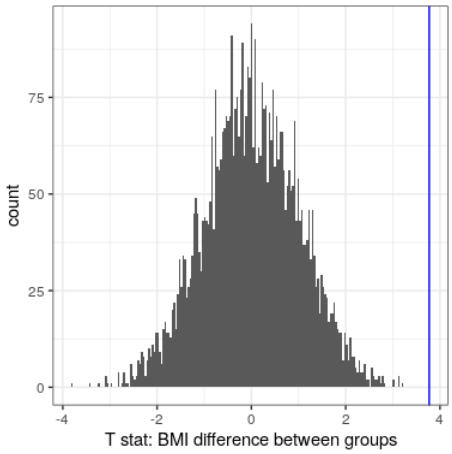

Figure 16.6: Histogram of t statistics after shuffling of group labels, with the observed value of the t statistic shown in the vertical line, and values at least as extreme as the observed value shown in lighter gray

We do have to make one main assumption when we use the randomization test, which we refer to as *exchangeability*. This means that all of the observations are distributed in the same way, such that we can interchange them without changing the overall distribution. The main place where this can break down is when there are related observations in the data; for example, if we had data from individuals in 4 different families, then we couldn't assume that individuals were exchangeable, because siblings would be closer to each other than they are to individuals from other families. In general, if the data were obtained by random sampling, then the assumption of exchangeability should hold.

16.3.6 Step 6: Assess the "statistical significance" of the result

The next step is to determine whether the p-value that results from the previous step is small enough that we are willing to reject the null hypothesis and conclude instead that the alternative is true. How much evidence do we require? This is one of the most controversial questions in statistics, in part because it requires a subjective judgment – there is no "correct" answer.

Historically, the most common answer to this question has been that we should reject the null hypothesis if the p-value is less than 0.05. This comes from the writings of Ronald Fisher, who has been referred to as "the single most important figure in 20th century statistics"(Efron 1998):

> *"If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 … it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials" (Fisher 1925)*

However, Fisher never intended $p < 0.05$ to be a fixed rule:

> *"no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" [fish:1956]*

Instead, it is likely that it became a ritual due to the reliance upon tables of p-values that were used before computing made it easy to compute p values for arbitrary values of a statistic. All of the tables had an entry for 0.05, making it easy to determine whether one's statistic exceeded the value needed to reach that level of significance.

The choice of statistical thresholds remains deeply controversial, and recently (Benjamin et al., 2018) it has been proposed that the standard threshold be changed from .05 to .005, making it substantially more stringent and thus more difficult to reject the null hypothesis. In large part this move is due to growing concerns that the evidence obtained from a significant result at

Unexpected text node: '32.'

16.3.6.1 Hypothesis testing as decision-making: The Neyman-Pearson approach

Whereas Fisher thought that the p-value could provide evidence regarding a specific hypothesis, the statisticians Jerzy Neyman and Egon Pearson disagreed vehemently. Instead, they proposed that we think of hypothesis testing in terms of its error rate in the long run:

> *"no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong" (Neyman and Pearson 1933)*

That is: We can't know which specific decisions are right or wrong, but if we follow the rules, we can at least know how often our decisions will be wrong on average.

To understand the decision making framework that Neyman and Pearson developed, we first need to discuss statistical decision making in terms of the kinds of outcomes that can occur. There are two possible states of reality (H_0 is true, or H_0 is false), and two possible decisions (reject H_0, or fail to reject H_0). There are two ways in which we can make a correct decision:

- We can decide to reject H_0 when it is false (in the language of decision theory, we call this a *hit*)
- We can fail to reject H_0 when it is true (we call this a *correct rejection*)

There are also two kinds of errors we can make:

- We can decide to reject H_0 when it is actually true (we call this a *false alarm*, or *Type I error*)
- We can fail to reject H_0 when it is actually false (we call this a *miss*, or *Type II error*)

Neyman and Pearson coined two terms to describe the probability of these two types of errors in the long run:

- P(Type I error) = α
- P(Type II error) = β

That is, if we set Unexpected text node: '18.3, which is the complement of Type II error.'

16.3.7 What does a significant result mean?

There is a great deal of confusion about what p-values actually mean (Gigerenzer, 2004). Let's say that we do an experiment comparing the means between conditions, and we find a difference with a p-value of .01. There are a number of possible interpretations.

16.3.7.1 Does it mean that the probability of the null hypothesis being true is .01?

No. Remember that in null hypothesis testing, the p-value is the probability of the data given the null hypothesis ($P(data|H_0)$). It does not warrant conclusions about the probability of the null hypothesis given the data ($P(H_0|data)$). We will return to this question when we discuss Bayesian inference in a later chapter, as Bayes theorem lets us invert the conditional probability in a way that allows us to determine the latter probability.

16.3.7.2 Does it mean that the probability that you are making the wrong decision is .01?

No. This would be $P(H_0|data)$, but remember as above that p-values are probabilities of data under H_0, not probabilities of hypotheses.

16.3.7.3 Does it mean that if you ran the study again, you would obtain the same result 99% of the time?

No. The p-value is a statement about the likelihood of a particular dataset under the null; it does not allow us to make inferences about the likelihood of future events such as replication.

16.3.7.4 Does it mean that you have found a meaningful effect?

No. There is an important distinction between *statistical significance* and *practical significance*. As an example, let's say that we performed a randomized controlled trial to examine the effect of a particular diet on body weight, and we find a statistically significant effect at p<.05. What this doesn't tell us is how much weight was actually lost, which we refer to as the *effect size* (to be discussed in more detail in Chapter 18). If we think about a study of weight loss, then we probably don't think that the loss of ten ounces (i.e. the weight of a bag of potato chips) is practically significant. Let's look at our ability to detect a significant difference of 1 ounce as the sample size increases.

Figure 16.7 shows how the proportion of significant results increases as the sample size increases, such that with a very large sample size (about 262,000 total subjects), we will find a significant result in more than 90% of studies when there is a 1 ounce weight loss. While these are statistically significant, most physicians would not consider a weight loss of one ounce to be practically or clinically significant. We will explore this relationship in more detail when we return to the concept of *statistical power* in Section 18.3, but it should already be clear from this example that statistical significance is not necessarily indicative of practical significance.

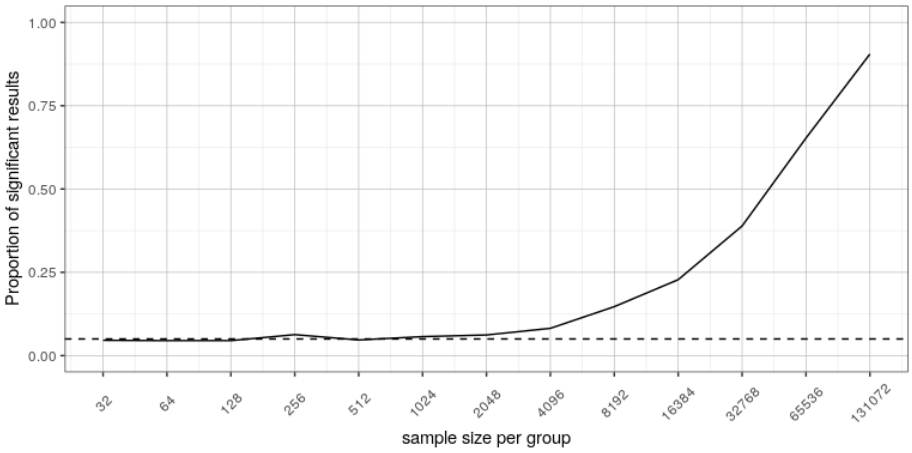

Figure 16.7: The proportion of signifcant results for a very small change (1 ounce, which is about .001 standard deviations) as a function of sample size.
