The Probability Lifesaver

Steven J. Miller

May 12, 2015

Contents

Chapter 22

Hypothesis Testing

Ernest Rutherford: If your experiment needs statistics, you ought to have done a better experiment.

In 2003–2004 I participated in a data analysis seminar at Ohio State. I remember one speaker mentioning that every day weather satellites beam down more information than is in the entire library of Congress, and forecasters have only a few hours to analyze the data and make their predictions. The wealth of data available is one of the boons of the twenty-first century, as well as one of its greatest challenges. We ignore this data at our own peril. Frequently we have mathematical models for problems of interest, ranging from the weather to the probable travel plans months in advance to choosing a professional sports team to judging the financial impact of regulations and laws to a description of the fundamental particles and forces in physics. Frequently we can gather data related to these issues; the question is whether or not the data supports our beliefs, or contradicts it, and how we make that decision.

This leads us to the very important field of model testing, an important part of **statistics**. As this is a probability book and not a statistics one, our treatment must be brief. I strongly urge you to take a statistics course in the future. When you do, you'll encounter many different tests to determine whether or not the data supports your conjecture. Why can we trust these tests? Probability! The tests in this chapter are consequences of many of our probability results and theorems.

Our point below is to introduce you to some of the major tests, and the reasons why they're true. This can't of course be a complete substitute for a full course on statistics, but hopefully it will give you a good sense of what can be done with probability, and encourage you to continue your studies. The tests in the following section are beautiful and important applications of the earlier material in the book. In a first course, most of the examples are of the following form. We have some population where the quantity of interest is drawn from one of the standard distributions with some unknown parameter. We have an idea what the unknown parameter should equal. Our goal is to see if the data supports our claims for this parameter's value. For example, maybe we believe the wealth of people in America is exponentially distributed with $\lambda = \$60,000$. We then gather our data. It's unlikely that our data

will be a perfect exponential with parameter \$60,000, and thus our goal is to quantify how close it is, and discuss the implications of our result.

22.1 Z-tests

We first describe the null and alternative hypothesis, then talk about significance levels and test statistics, and end with a discussion of one versus two-sided tests. As it's easier to learn material through an example than just pages of theory after theory, we introduce the z-statistic and the z-test, and frame our discussion and examples through this.

Takeaways for the section:

- There is a method for testing the truth of a hypothesis, even when there's randomness involved.
- We assume the hypothesis is true, and then collect data.
- We decide whether or not our assumption is valid by evaluating the likelihood of the data we collect.

22.1.1 Null and Alternative Hypotheses

Suppose McDonald's has come out with a new ad campaign, boasting that the mean time it takes them to fill a typical order is 45 seconds. Clearly every order is different, so there will be some variation around this mean. Being the skeptic you are, you make some observations the next time you visit McDonald's, and in a sample of 20 orders you find the average service time to be 48 seconds with a standard deviation of 8 seconds. Given this data, do you believe McDonald's claim?

There is one major obstacle preventing you from answering this question: the randomness of the sample you drew. Your sample mean suggests that McDonald's is slower than they claim, but it could also be that your sample was slow by chance – perhaps during the course of your observations a Little League baseball team came in to celebrate a win. We need a formal process to determine whether McDonald's is telling the truth, one that takes into account the possibility that your sample is unusual. This process is known as **hypothesis testing**.

There are many circumstances where it's valuable to assess the validity of a claim, from finding out how fast McDonald's can fill an order, to determining whether a new drug works better than existing ones. The first step in hypothesis testing is establishing a **null hypothesis**. The null hypothesis is typically a statement contrary to what the experimenter is attempting to show; the experimenter assumes the null hypothesis to be true, and uses data to try and refute it. In the McDonald's example, since we think the average service time μ is slower than 45 seconds, our null hypothesis (denoted H_0) might look like this:

$$\text{Null Hypothesis: } H_0 : \ \mu \leq 45.$$

That is, we hypothesize that the mean service time is *at most* 45 seconds. After developing our null hypothesis, we also have to articulate the **alternative hypothesis**, which is frequently the statement we want to show true. For the McDonald's

example, our alternative hypothesis (denoted H_a) would be:

$$\text{Alternative Hypothesis: } H_a: \ \mu > 45,$$

which is what we're expecting to find. Note that the alternative and null hypotheses are complements, meaning that together they encompass every possibility for the value of μ. We couldn't have the following set of hypotheses:

$$H_0: \ \mu < 45, \qquad H_a: \ \mu > 46, \tag{22.1}$$

because they don't allow for the possibility that $45 \leq \mu \leq 46$. This is an important point to remember: your null and alternative hypothesis must allow for every possible value of the parameters you're measuring.

One of the most important aspects of hypothesis testing is the way we phrase our arguments. If we want to show that a drug is effective, we would take our null hypothesis to be "The drug is ineffective." If the performance of the drug is inconceivable given the assumption that it's ineffective, we reject the null hypothesis and conclude that the drug works. Notice, though, that we would *not* assume that the drug is effective, and revise our assumption only if the drug gives us good reason to. It is much more convincing to say, "This drug has proven itself," than to say "This drug hasn't screwed up yet."

22.1.2 Significance Levels

Once we have formulated our null and alternative hypothesis, how do we actually test these hypotheses? We assume the null hypothesis to be true, and then look at our data. If the probability of collecting this data under the null hypothesis H_o is sufficiently small, we reject H_o in favor of the alternative hypothesis H_a. This form of argumentation might be new to you, so here's an example. Imagine you're a biologist measuring a newly discovered group of crocodiles. Previous studies have shown that the local adult crocodiles are normally distributed with a mean length of 6 feet and a standard deviation of 6 inches (remember there are 12 inches in a foot). Suppose you measure one of these new crocodiles and find it to be 14 feet long. Do you believe that this is one of the local crocodiles? Probably not, because its length is a whopping 16 standard deviations above the mean! What you are implicitly saying is this: "If the height of adult crocodiles in group A is normally distributed with a mean length of 6 feet and a standard deviation of 6 inches, then the odds of seeing a crocodile in group A as long as I just did is extraordinarily small." Since the evidence you've collected is so unlikely under the assumption that the crocodile is in group A, you decide that the crocodile probably is not in group A (in other words, that your assumption was wrong). Here, H_o would be that the crocodile is in group A, and H_a would be that the crocodile is not in group A. We reject H_o in favor of H_a because it is unlikely that given H_o is true, the crocodile would be 14 feet long.

You might have noticed an issue with the line of reasoning we just gave: how unlikely is unlikely enough? What if the crocodile were 7 feet long? In this case the crocodile is only 2 standard deviations above the mean, which is rare but not unheard of.

To get around this problem, we often establish a **significance level** (significance levels are also known as **α-levels**). A significance level is a limit on how unusual a result we will accept. An α-level of 0.05 means that if our observations from our

collected data would occur less than 5% of the time given that the null hypothesis is true, then we will reject the null hypothesis. The advantage of setting an α-level is that it gives a hard and fast cutoff of how unlikely an event we will accept; this is also a disadvantage, as it gives us no flexibility! We'll discuss the blessing and the curse of significance levels later on.

There is an easy way to visualize α-levels. Suppose we've hypothesized that some population parameter we're interested in is distributed $N(\mu, \sigma^2)$, and we want to test this hypothesis at an α-level of 0.05. Using a z-table, we see that if the null hypothesis is true, we will measure the parameter to be between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ with probability 0.95. In other words, 95% of the time (or 95 out of 100 experiments) yield a value in this range if the null hypothesis is true. So, if we measure something more than 1.96 standard deviations from the mean, we have witnessed an event that would happen less than 5% of the time under the null hypothesis, and we reject the null hypothesis. We can think of the null hypothesis as establishing a **critical region**, where if the measurement we make lands in the critical region then we reject the null hypothesis. When we're using an α-level of 0.05 and the parameter we're testing is normally distributed, the critical region is everywhere more than 1.96 standard deviations from the mean. See Figure 22.1.

In this book, when we say "significance," we use statistical significance rather than practical significance. Just because something is statistically significant does not mean it is practically significant, or that the difference warrants some sort of action. For example, perhaps a newly developed type of long underwear insulates 2% better than the previously best kind of long underwear, and this difference is statistically significant. This does not mean people will choose to buy the new long underwear over the old one solely based on the difference in insulation, for a 2% improvement is not that much better. Thus, we do not have practical significance. Now suppose we have a 30% improvement in insulation. This is much more likely to be practically significant, especially for consumers in Massachusetts during the winter. However, this 30% is not necessarily statistically significant, meaning we do not know whether the large difference is real. We would need to collect more data to show that indeed the difference exists, which we will discuss in a later section when we talk about sample size and power.

Figure 22.1: The probability of finding a normally distributed parameter more than 1.96 standard deviations from its mean is 0.05. This therefore determines the critical region for a two-sided z-test with an α-level of 0.05.

We often have many choices in choosing an α. The most common α-levels are 0.10, 0.05, 0.01, and 0.001. Clearly, the smaller α is, the more difficult it is for the data we have observed to be considered unlikely enough to reject the null hypothesis. Thus, the α we choose should vary based on what we are observing. For example, if we are assessing the effectiveness of a new surgical procedure, we probably want a very low α-level to make sure it is truly significantly different from an older procedure, which most surgeons have likely gotten used to performing, to warrant its adoption. For something such as whether people prefer sprinkles on their ice cream, it may be reasonable to choose 0.10 as the α-level. Because it is difficult to decide whether an α-level is too high or too low, 0.05 is arbitrarily considered the standard

α-level, which of course should not be the only one we use. Most importantly, we must select the α-level before observing the data. The fear is that after observing the data, we will cheat by purposely choosing an α-level that allows us to reject the null hypothesis. For example, if a pharmaceutical company tries to sell a new drug, perhaps it is considered significantly better than the old drug under the α-level of 0.05 but not for an α-level of 0.01. The performance of a drug, especially if used to treat fatal diseases, is very serious and thus should probably be considered under a smaller α-level than 0.05. However, pharmaceutical companies could raise the α-level to 0.05 in order to say that the new drug is significantly better and thus sell more of it.

One more piece of terminology before we move on. For an α-level of 0.05, the 1.96 value we keep mentioning is known as a **critical value**. As you can imagine, the critical value is different for each value of α. We'll give a more precise definition of the critical value in the next section.

22.1.3 Test Statistics

Once we've set our significance level, we're ready to test our hypothesis. All we need to do is formulate a **test statistic**. A test statistic is a measurement we get from the data whose distribution we assume know (remember we start by assuming the null hypothesis is true). Since we know the distribution of our test statistic, we can determine the probability of measuring a test statistic that large or larger; this is the probability we use to decide whether or not to reject the null hypothesis. One common test statistic is the sample mean. After we identify our test statistic, we need to figure out the distribution it follows. This is the most important question in hypothesis testing, since without knowing the distribution of our test statistic, we will have no way to assess whether our result was unusual or not.

As an example, what kind of distribution should the sample mean follow? To figure that out, we recall the Central Limit Theorem: for a group of n identically and independently distributed random variables X_i with mean μ and variance σ^2, the random variable

$$Y_n \;=\; \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$$

is normally distributed in the limit as $n \to \infty$. If each X_i represents the measurement of a member of our population, then Y_n is the sample mean, and the X_i satisfy the conditions above. What are the mean and variance of Y_n? By linearity of expectation we see that

$$\mathbb{E}[Y_n] \;=\; \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\right] \;=\; \frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right) \;=\; \frac{1}{n}\cdot n\mathbb{E}\left[X\right] \;=\; \mu;$$

the expected value of Y_n is the expected value of X. Furthermore, we see that the variance of Y_n is given by

$$\text{Var}\left(Y_n\right) \;=\; \text{Var}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\right) \;=\; \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right). \tag{22.2}$$

Since the X_i's are independent, we know

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \text{Var}(X_i) \;=\; n\text{Var}(X_i),$$

so the variance of Y_n simplifies to

$$\text{Var}\left(Y_n\right) \;=\; \frac{1}{n^2} \cdot n\text{Var}(X_i) \;=\; \frac{\sigma^2}{n}.$$

The wonderful feature about the variance of the sample mean is that it decreases as the sample size increases. To get a feel for why this should be, imagine repeatedly flipping a fair coin (a coin which lands heads or tails with equal probability). Let's call a result "unusual" if we flip fewer than 45% heads or more than 55% heads. With a sample of size 2, there's a 50% chance that we will have an unusual result, since we can only have 2 heads (unusual), 2 tails (unusual), or 1 of each (not unusual). What about a sample of size 100? Now if we flip fewer than 45 heads or more than 55 heads, we will have an unusual result. How many ways can we do this? Recalling our combinatorial knowledge, there are

$$\sum_{i=45}^{i=55} \binom{100}{i} \;=\; 923,796,541,447,310,445,480,620,479,776$$

ways to flip between 45 and 55 heads (i.e., not getting an unusual result), and 2^{100} total ways to flip 100 coins. Therefore the probability of having an unusual result is just one minus the probability of not having an unusual result:

$$\text{Prob(unusual result)} \;=\; 1 - \frac{1}{2^{100}} \sum_{i=45}^{55} \binom{100}{i} \;=\; 0.271.$$

Repeating the same argument for a sample of size 1000, we find a 0.0014 probability of obtaining an unusual result. As the sample size increases, the total number of possible results increases much faster than the number of ways to get an unusual result, so with a larger sample we expect to get a much more faithful estimate of the true mean (see Figure 22.2). This is an example of a very general principle: **more data is always better**. Unfortunately, to double your accuracy (cut the variance in half) you need to collect four times as much data, which in many real-life situations can be expensive or time-consuming.

Figure 22.2: As our sample size increases (that is, as we flip more coins), the odds of obtaining an "unusual result" decrease drastically. This is why we expect the variance to decrease as we take larger and larger samples.

Once we have the distribution of the test-statistic, hypothesis testing is straightforward. Suppose our null hypothesis is that the mean of a random variable X is equal to μ. If the null hypothesis is true, then the sample mean $\bar{x}$ should be distributed $\text{N}\left(\mu, \sigma^2/n\right)$. Then our test statistic is:

Test statistic for a sample mean: z-statistic. Let X be a normal random variable with known variance σ^2 and hypothesized mean μ, and let $x_1, x_2, \ldots, x_n$ be n independent observations drawn from this distribution. Set $\overline{x} = (x_1 + \cdots + x_n)/n$. Then the observed z-test statistic values

$$z \;=\; \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

are normally distributed with mean 0 and variance 1 (so $Z \sim \mathrm{N}(0,1)$. If instead of being normally distributed X is just a nice, well-behaved distribution, then a good rule of thumb is that $\bar{X} = (X_1 + \cdots + X_n)/n$ is nearly normal for if $n \geq 30$. It's very important that the variance is known; if it isn't, more involved tests are needed.

This test is called a **z-test** because our test statistic in normally distributed. This is because a sum of normally distributed random variables is again normally distributed. For any z value we measure, we can use the standard normal table to find the probability of obtaining a test statistic that far or farther from zero. This probability is called the **p-value** (which stands for probability value). If the p-value is less than the α-level, we reject the null hypothesis in favor of the alternative hypothesis. This also gives us a clearer meaning for the **critical value**. For a given hypothesis test, the critical value is the value such that if our test statistic is larger than the critical value (in the absolute value sense), we reject the null hypothesis.

Now we can finally finish off our McDonald's example. From our discussion above, if we accept the null hypothesis and let $\mu = 45$ and assume $\sigma = 8$ (we'll return to this assumption shortly), then in a sample of size 20 we should have

$$\bar{X} \;\sim\; \mathrm{N}\left(45, 8^2/20\right).$$

This means that for our sample of size 20, the wait time should have a mean of 45 seconds and a standard deviation of about 1.79 seconds. We measured $\bar{x} = 48$, so our z value is $(48 - 45)/1.79 = 1.68$. Using a z-table, we see that the odds of measuring a test statistic larger than 1.68 by chance alone is 0.046, or a little over 4%. Since this p-value is less than our significance level of 0.05, we reject the null hypothesis and conclude that McDonald's is indeed slower than they claim.

22.1.4 One-sided versus Two-sided tests

While going through that last example you might have wondered, "Why did you only look at the probability of finding a z-score higher than 1.68? Don't we also need to worry about the possibility of measuring a test statistic smaller than the hypothesized mean, too?" This is an important question which we've been a little lax about up to this but, and it illuminates the core difference between **one-sided** and **two-sided hypothesis tests** . The McDonald's case is an example of a one-sided test, where we're interested in seeing whether the parameter we're measuring is greater than (or less than) a specific value. To do this, we calculate the probability of observing a test statistic as large or larger than the one we have. With a two-sided test, we're

interested in seeing whether the parameter is significantly different from a given value, so we calculate the probability of observing a test statistic as far from the hypothesized mean or further than the one we have. Clearly the p-value we measure depends on which test you're doing. The difference between one and two sided tests is depicted in Figure 22.3.

Figure 22.3: The difference between a two sided z-test and a one sided z-test. Both tests have an α-level of 0.05.

You'll notice that the critical value is smaller for the one-sided test than the two-sided test (your test statistic only needs to be 1.64 standard deviations from the mean instead of 1.96). This is a general property of one and two-sided tests: two-sided tests require more evidence than one-sided tests.

Why do two-sided tests require more evidence, and if so why don't we always use one-sided tests? The issue is we need information to justify performing a one-sided test. In the McDonald's example we could safely assume that they weren't any faster than they advertised, because if they were they most certainly would have advertised it! However, if we were unable to discard that possibility, then we would have needed to use a two-sided test.

One final word about one-sided tests. In our McDonald's example, our null hypothesis was $\mu \leq 45$, however when we actually performed our test, we just let $\mu = 45$. Why don't we need to worry about the possibility that the mean was, say, 43? Well, imagine we did let $\mu = 43$. Then our test statistic would have become $(48 - 43)/1.79 \approx 2.79$ – larger than we had before! For *any* mean less than 45, we would have measured a higher z-value. So $\mu = 45$ is the most difficult case; if we can reject the null hypothesis while letting $\mu = 45$, then we can reject it for any hypothesized mean less than 45. This is a handy property of one-sided tests: you only need to check the most extreme case.

Now that we have laid out the framework for hypothesis testing, let's go through some examples.

Example: Imagine you're measuring the lifetime of lightbulbs, which are known to have a standard deviation of $\sigma = 100$ hours, and you want to test whether their lifetime is significantly different from 2,000 hours. In a random sample of 20 lightbulbs, you find the mean lifetime to be 2050 hours. With an α-level of 0.05, would you reject the hypothesis that the mean lifetime is 2000 hours?

Solution: We've already been given the null hypothesis, namely that $\mu = 2000$. Therefore our alternative hypothesis is

$$H_a: \ \mu \neq 2000.$$

Notice that this is a two-sided test, since we have no reason to dismiss the possibility that lightbulbs last less than 2000 hours. We now need to calculate our test statistic, which is

$$z = \frac{\text{observed mean - hypothesized mean}}{\text{standard deviation}}$$

Be careful! A common mistake people make is to say, "Well the mean I found was 2050 and the hypothesized mean is 2000, and since the standard deviation is 100, my z-score is 1/2. This is too small to conclude that the means are different." However, they're using the wrong standard deviation! The standard deviation of 100 hours is for a sample of *one* lightbulb. For a sample of 20 lightbulbs, you expect to get a more reliable estimate of the mean, and the correct standard deviation to use is $\sigma/\sqrt{n} = 100/\sqrt{20} \approx 22.36$ hours. Thus our z-score is $(2050 - 2000)/22.36 \approx 2.24$. Since we're using a two-sided test, our p-value is the probability of measuring a test statistic greater than 2.24 or less than -2.24, which turns out to be 0.025. This is very compelling evidence that the true mean is not 2000 hours.

Example: Suppose an auto insurance company has recently moved from Boston to Portland, and is trying to get a sense of their new market. In Boston, the percent of their policyholders that filed a claim in a given year was described by a binomial distribution with probability of success (i.e., a driver filing a claim) of $p_b = 0.25$. In their first year in Seattle, the firm finds that 2300 of their 10,000 drivers filed a claim. Using an α-level of 0.05, test whether p_s, the probability of an insured driver in Seattle filing a claim, is less than 0.25.

Solution: What are our null and alternative hypotheses in this case? Since we want to test whether $p_s < 0.25$, our null hypothesis is

$$H_0 : \ p_s \geq 0.25,$$

and our alternative hypothesis is

$$H_a : \ p_s < 0.25.$$

Here we're using a one-sided test, so we need to be careful about justifying why we don't consider the possibility that Seattle drivers have more accidents than Boston drivers. Perhaps previous research suggests that this is the case, or maybe Seattle's smaller population density would make us expect fewer accidents. Assuming our one-sided test is reasonable, once we have our hypotheses, we proceed as we typically do. Let's assume the null hypothesis is true, and take $p_s = 0.25$. If this were the case, the number of claims in Seattle should be described by a binomial distribution of size 10,000 and probability of success 0.25. However, we can make our lives a little bit simpler. For large N, a binomial distribution of size N with probability of success p is well approximated by a normal distribution with mean Np and standard deviation $\sqrt{Np(1-p)}$ (this is just a special case of the Central Limit Theorem; we gave the proof when $p = .5$ in §**??**). In most books we often consider any N-value greater than 30 to be large; as we're at 10,000 here, it's fairly safe to replace the binomial with a normal. Therefore, if $p_s = 0.25$, the number of drivers filing claims out of a group of 10,000 should be approximately normally distributed with mean $10000 \cdot 0.25 = 2500$ and standard deviation $\sqrt{10000 \cdot 0.25 \cdot 0.75} \approx 43.3$.

Now that we know what the distribution of drivers filing for claims should look like, we can proceed to calculate our test statistic. We've shown that under the null hypothesis, the number of drivers filing a claim should be normally distributed with mean 2500 and standard deviation 43.3. We had 2300 drivers file a claim. Therefore our test statistic is

$$\frac{2300 - 2500}{43.3} \approx -4.62.$$

This test statistic should follow a standard normal distribution.

The last step is to compute our p-value, which is the probability of observing a test statistic as or more extreme than the one we did. Since we're testing whether $p_s < 0.25$, we only need to find the probability of measuring a test statistic lower than -4.62, which is just $\Phi(-4.62) \approx 1.92 \times 10^{-6}$ (here Φ is the cumulative distribution function of the standard normal). This is much less than our α-level of 0.05, so we reject the null hypothesis and conclude that the probability of an insured driver in Seattle filing for a claim in a given year is less than 0.25.

One major point to be aware of about hypothesis testing in general: we need to justify the distribution of the test statistics. When we're testing hypotheses about sample means, we can typically appeal to the Central Limit Theorem to conclude that $\bar{X}$ is normally distributed. However, we need to be careful with this method of argumentation. If our underlying distribution is "nice" and our sample size is large, then it's probably fine to take $\bar{X}$ as normally distributed. But if you're ever in a situation where you're dealing with a weird distribution or have only a few data points, you need to be really careful if you're going to appeal to the CLT. To put it another way: if you're going to use a z-test, you need to be able to provide a good reason for why the data you've collected are normally distributed.

One final point to be aware of about z-tests is that they assume perfect knowledge of the variance. In the McDonald's example we actually used the *sample variance* and assumed it to be the true variance. This is technically incorrect, and we will discuss how to fix this situation later in the chapter.

22.2 On p-values

In the last section we developed a method of hypothesis testing. Within the hypothesis testing framework, the most important measurement we make is the p-**value**. The p-value tells us the probability of collecting data as or more extreme then what we have, *assuming the null hypothesis is true*. In this section we provide some more intuition about p-values, and warn against some possible misinterpretations of it.

Takeaways for the section:

1. The p-value is a conditional probability: the odds of collecting the data you have given that the null hypothesis is true.
2. The p-value depends on context.
3. Different tests of the same data can yield different p-values.
4. The p-value is not the probability that the null hypothesis is true.

22.2.1 Extraordinary Claims and p-values

As we outlined in the first section, the p-value is a measure of how likely it is for the event we observed to occur by chance alone, assuming the null hypothesis is true. If we observe an event with a very small p-value, there are two options available to

us: we can either conclude that we have seen a rare event, or conclude that the event we saw was so unlikely given our assumptions that our assumptions were probably wrong (that is, reject the null hypothesis).

One point we'd like to emphasize is that how convincing a p-value is depends on context. How can that be? Imagine a musician walks up to you and tells you he has perfect pitch (he can identify any note just by listening to it). Suppose that to test this, you play 8 different notes for him and he accurately identifies all 8 of them. Would you believe his claim? Probably, since trained musicians tend to have good ears, and if he were guessing the odds of him getting all 8 right are exceedingly small. Now imagine a man walks up to you and says he can tell you what note you're going to play next just by thinking hard about it. To test his claim, you have him write down what note he thinks you're going to play, and you then write down the note you actually play. After 8 notes, you reconcile your lists. Suppose he gets all 8 of the notes right. Would you believe his claim? Even though we have the same evidence as in our first example, you would probably be hesitant to believe that this man can predict your actions (and you would probably want to get more data points!) simply because his claim is so incredible. So even though we have identical situations in terms of evidence (that is, if we defined the null hypothesis that these guys were just guessing, the p-value would be the same in both cases), we're more likely to believe one p-value over the other. As Carl Sagan once said, "Extraordinary claims require extraordinary evidence."

22.2.2 Large p-values

Suppose you want to test whether a coin is fair (that is, lands heads or tails with equal probability), and that in flipping the coin 20 times, you recorded 12 heads. If our null hypothesis is that the coin is fair, what is the p-value? Assuming the coin is fair, the number of heads flipped H should follow a binomial distribution:

$$\Pr(H = x) \;=\; \frac{1}{2^{20}}\binom{20}{x}.$$

The probability of flipping 12 or more heads or 12 or more tails (that is, our p-value) is:

$$p \;=\; \frac{1}{2^{20}}\left(\sum_{j=0}^{8}\binom{20}{j} + \sum_{j=12}^{20}\binom{20}{j}\right) \;\approx\; 0.503.$$

This is a larger p-value than we've seen before, and we definitely could not reject our null hypothesis with this kind of data. So what does this mean? Have we proven that the coin is fair? Certainly not! This data would actually be more consistent with the hypothesis that the probability of flipping a head was 0.6. However, this data is not out of line with what we would expect from a fair coin. Since we cannot say that we have proven the coin fair, we say that we **fail to reject the null hypothesis** that the coin is fair. This is an important part of statistical language – we never "accept" any hypotheses, we merely reject them or fail to reject them. It could be that the coin is not fair, and that with more data points we would see a stronger bias emerge. However, from the sample that we've observed, there's no compelling evidence to make us revise our initial claim.

22.2.3 Misconceptions about p-values

One final point about p-values: the p-value is *not* the probability that the null hypothesis is true. This is a very common mistake, and a reasonable one to make. However, consider the coin example: the coin is either fair or it isn't – there's no randomness to the state of the coin. In this case it doesn't even make sense to talk about the probability of the null hypothesis being true. However, the p-value still makes sense: it's the conditional probability of collecting the evidence we have given that the null hypothesis is true, not the probability that the null hypothesis is true given the data we collected.

Example: You're an office manager and have purchased 20 brand new photocopiers. The manufacturer tells you that the probability of one of them breaking down in a given year is $p_{\text{ph}} = 0.03$. Over the course of the first year, two of the photocopiers break down. Do you now believe the value of p_{ph} the manufacturer told you? At what α level could you reject the manufacturer's claim?

Solution: This is a nice introduction to hypothesis testing with binomial distributions, as our sample is too small to use the normal approximation. We must assume that each photocopier breaking down is independent of whether or not other photocopiers break down in order to use a binomial distribution. As always, we first formulate our null and alternative hypotheses. We suspect that the machines break with frequency greater than 0.03, so let's hypothesize as follows:

$$H_0: \ p_{\text{ph}} \leq 0.03, \qquad H_a: \ p_{\text{ph}} > 0.03. \tag{22.3}$$

Let's assume the null hypothesis to be true, and take $p_{\text{ph}} = 0.03$. If this were the case, then the probability of having x machines break down in a year is

$$\text{Prob}(x \text{ machines break}) = \binom{20}{x}(0.03)^x(1-0.03)^{20-x}.$$

The beautiful thing about hypothesis testing with binomial distributions is that we already know our test statistic – it's 2. Our p-value is just the probability that 2 or more machines break in a given year. We could do this the long way be adding the probability that two fail to the probability that three fail, et cetera, or we could just note that

$$\text{Prob}(2 \text{ or more machines break}) = 1 - \text{Prob}(0 \text{ or } 1 \text{ machines break}),$$

which is just $1 - \left((0.97)^{20} + 20 \cdot 0.03 \cdot (0.97)^{19}\right) \approx 0.12$. This is not a very convincing p-value. Yes having two photocopiers break in a year is rare, but not so rare that you'd begin to wonder whether the manufacturer was lying to you. To be able to reject the null hypothesis, your α-level would have to be at least 0.12, which is too large to be accepted in any formal circumstances.

Example: From 1997-2006, the number of cases of the flu in Hartford – per 10,000 in the population – was modeled by a Poisson distribution with $\lambda = 350$. As a reminder, this means that the probability of seeing k cases of the flu in a particular

year is given by

$$\text{Prob}(k \text{ cases of the flu}) \;=\; \frac{e^{-350} \cdot 350^k}{k!}.$$

Starting in 2007, a law was passed that increased the number of flu shots given out. In the three years since, the number of cases of the flu per 10,000 people in the population have been 330, 320, and 325 respectively. Assuming the new law was the only meaningful change affecting flu rates (this is a big assumption), do you believe that increasing the availability of flu shots helped reduce the number of cases of the flu in Hartford?

Solution: Let's take our null hypothesis to be that the law did nothing, and that the low totals of the past few years are due to nothing but chance. In this case, the number of cases of the flu will be modeled by a Poisson distribution with $\lambda = 350$. What kind of distribution should we have over the course of three years? As we saw back in §**??**, if X follows a Poisson distribution with parameter λ_1 and Y follows a Poisson distribution with parameter λ_2, then $X + Y$ follows a Poisson distribution with parameter $\lambda_1 + \lambda_2$. If we had three Poisson random variables, it would be a Poisson with parameter equal to the sum of the three parameters.

We see that over the course of three years, the number of cases of the flu in Hartford should be modeled by a Poisson distribution with $\lambda = 350 + 350 + 350 = 1050$. We've had 975 cases of the flu over the past three years. The probability of seeing that few or fewer is

$$\sum_{n=0}^{975} \frac{e^{-350} \cdot 350^n}{n!} \;\approx\; 0.010,$$

giving a p-value of 0.01. This is pretty compelling evidence that the law was effective.

Remark: We mentioned that assuming we can attribute all the changes in flu cases to this new law is a big assumption. This is because there are many other factors that may have been responsible for lowering the incidence of the flu. Perhaps a recently launched public health program increased the number of Hartfordians who washed their hands consistently, which helped reduce the transmission of germs. Or maybe a few unseasonably warm winters meant people weren't spending as much time cooped up indoors, and thus were less likely to pass germs to each other. For problems like these, it's always important to recognize the assumptions you're making, and to think about potential factors you've left out. This is important for two reasons: it lets you recognize the shortcomings of your model, and (more importantly) also helps steer you towards important data. For example, if we think weather played a factor in Hartford, maybe we could look at other towns that had warm winters but didn't have a flu vaccine law passed.

22.3 On t-tests

At the end of §22.1 we mentioned one potential issue with z-tests: they assume perfect knowledge of the variance. However, there are very few instances where we

actually know the variance. Most of the time we need to estimate it from the data itself. In this section we discuss how to find an accurate estimator of the variance, and how to use this estimator in our hypothesis testing framework.

Takeaways for the section:

1. The z-test assumes perfect knowledge of the variance, which in many cases we don't have.
2. We can get an estimate of the variance from the data, and use that in place of the actual variance.
3. Using the sample variance changes the distribution of our test statistic.

22.3.1 Estimating the Sample Variance

In the case that we collect data on a population parameter without any prior information about its variance, the most straightforward way to estimate the variance is to calculate the **sample variance** as follows: find the sample mean $\bar{x}$ as always ($\overline{x} = (x_1 + \cdots + x_n)/n$), and then take the sample variance s^2 to be

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

This looks remarkably like our regular formula for variance, except we now have $n-1$ in the denominator. This is due to an important concept in statistics known as **degrees of freedom**. For any given estimate, the degrees of freedom are the number of independent observations that contribute to that estimate. Imagine we know that in Boston the average high temperature in August is $80°$, and we want to get a sense of the variance. If we record the high for the next day to be 85, we could estimate that the variance is $(85-80)^2 = 25$. Suppose, as is usually the case, we didn't know the mean beforehand. Then if we measure the high temperature to be 85, we could use that as our estimate for the mean. However, we can't say anything about the variance, because there's nothing to deviate from the mean. What happens if we try to use the formula above for a sample of size one? You get the indeterminate 0/0 – even the equation knows it shouldn't be used in this instance!

What if we made a second observation, and say the high was 88? Now our sample mean is 86.5, and we can estimate the sample variance. However, since we're using the sample mean to estimate the variance, both observations don't independently contribute to the variance; once you know the mean and the value of one of the highs, you automatically know the value of the second high. Thus only one observation contributes to the sample variance. In general, we use $n-1$ because we want to measure the amount of variation in our sample, per observation that contributes to the variance.

*Remark: Another way to convince yourself that we should use $n-1$ instead of n is to show that s^2 is an **unbiased estimator** of σ^2. This means $\mathbb{E}(s^2) = \sigma^2$, so on average our prediction is good.*

22.3.2 From z-tests to t-tests

With the sample variance in hand, we're ready to fix the problem with the z-tests we mentioned earlier. Let's recall that when performing a z-test, we find a test statistic $\bar{x}$ that is normally distributed with mean μ and variance σ^2/n (under the null hypothesis). Given this, we form the test statistic

$$(\bar{x} - \mu) / \sqrt{\sigma^2/n},$$

which has a standard normal distribution. What if we don't know σ^2? It seems reasonable to use our estimate s^2 in its place. That is, we place $\sqrt{s^2/n}$ in the denominator instead of $\sqrt{\sigma^2/n}$. How should this affect the distribution of our test statistic? Since $\mathbb{E}[\bar{x}] = \mu$ and $\mathbb{E}[s^2] = \sigma^2$, it should be close to the z-statistic. However, the distribution should be a little more spread out than the normal distribution. Why is this? When we knew σ^2, there was only one source of randomness – $\bar{x}$. Now we have two measures that can bounce around – $\bar{x}$ and s^2 – so it makes sense that our distribution should not be as localized as it was before. To put it another way: we shouldn't have better estimates now that we know less!. It turns out that our test statistic does indeed follow a distribution which is symmetric about 0 – a distribution called the t**-distribution**.

The t-distribution. The t-distribution is a family of distributions parametrized by their degrees of freedom. Let $X_1, \ldots, X_n$ be independent standard normal random variables, and let S_n^2 be the sample variance random variable. Then the distribution of $\overline{X}/(S_n/\sqrt{n})$ is called the t-distribution with n degrees of freedom, and is denoted T_n. If instead we draw from a normal with mean μ and variance σ^2 then the test-statistic is

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}.$$

The closed-form expression for the density of a t-distribution with ν degrees of freedom is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The t-distribution looks remarkably like the normal distribution, as shown in Figure 22.4. In fact, in the limit as n goes to infinity, the t-distribution approaches the normal distribution. This makes intuitive sense: we use the t-distribution when we have an underlying normal distribution but need to estimate the variance. When our sample size is large, our estimate of the variance becomes more and more exact, and we don't need to worry as much about the uncertainty in our estimate.

Once we know how our test statistic is distributed, the t-test works exactly the same way as z-tests. We formulate null and alternative hypotheses, calculate our test statistic, and find the corresponding p-value. The intuition is also exactly the same: if

$$\beta = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

Figure 22.4: The t-distribution with (from shortest to tallest): 1, 3, 9, and 20 degrees of freedom. The tallest curve is the standard normal distribution. Notice that the t-distribution approximates the the normal distribution very well once we get to 30 degrees of freedom or so.

follows a t-distribution, then for any value of β we can calculate how likely it is to measure a test-statistic that large or larger. If that probability is sufficiently small, we reject the null hypothesis.

Now that we have the formalism of t-tests at our command, let's return to the McDonald's example we gave at the beginning of the chapter. Instead of taking $\sigma^2 = 48$ as we did before, we can more accurately use the sample variance $s^2 = 48$, which gives

$$t \;=\; \frac{48-45}{\sqrt{\frac{48}{20}}} \;\approx\; 1.68,$$

as we found earlier. The only difference is that now we expect our test statistic to follow a t-distribution with 19 degrees of freedom. It turns out that for this distribution, the cutoff value for a 0.05 α-level is $t = 1.73$, so now we fail to reject the null that $\mu = 45$ at a 5% significance level.

What happened? Why were we able to reject the null when working with a z-test, but were unable to when using a t-test? There are two factors: we weren't as well informed as we thought we were, and significance levels impose arbitrary cutoffs. When we did our first z-test, we claimed we knew that the standard deviation was 8 seconds. However, in using the t-test we had to admit that we were only estimating the variance, and because of that each of our further estimates became a little more uncertain. The other issue is that we ran up against the 5% cutoff level. Even when using the t-test, our p-value was still 0.055 – pretty compelling evidence that McDonald's is slower than they claim, but just over our cutoff. For this reason, many researchers forego rejecting or failing to reject hypotheses, and simply do the analysis and report the p-value.

Example: Suppose you grow tomatoes in your home garden. You've made some observations over the past few years, and have consistently found that the weight of the tomatoes you grow is nearly normally distributed with mean weight 4 ounces. Recently, though, you've seen advertisements for a new fertilizer that claims to increase the size of produce. On an adventurous whim, you decided to test it out. In the next batch of tomatoes you grow there are two 3 ounce tomatoes, four 4 ounce tomatoes, and six 5 ounce tomatoes. From this, can you conclude that the fertilizer increases the yield?

Solution: Since we're interested in seeing whether the fertilizer increases the yield, let's let our null hypothesis be that the fertilizer has no (or perhaps a negative) effect. Denoting the mean tomato size with fertilizer by μ, we have

$$H_0: \;\; \mu \;\leq\; 4, \qquad H_a: \;\; \mu \;>\; 4. \tag{22.4}$$

As usual, let's assume the null hypothesis true and let $\mu = 4$. To calculate our test statistic, we need the sample mean and the sample variance. We see the sample mean is

$$\bar{x} \;=\; \frac{2 \cdot 3 + 4 \cdot 4 + 6 \cdot 5}{12} \;=\; 4.33,$$

and the sample variance is

$$s^2 \;=\; \frac{1}{11}\left(2 \cdot (3 - 4.33)^2 + 4 \cdot (4 - 4.33)^2 + 6 \cdot (5 - 4.33)^2\right) \;\approx\; 0.61.$$

Our test-statistic is therefore

$$\frac{\bar{x} - \mu}{\sqrt{s^2/n}} \;=\; \frac{4.33 - 4}{0.225} \;\approx\; 1.48.$$

How should this test-statistic be distributed? Since we're using the sample variance, our first guess would be a t-distribution. However, to use the t-distribution, we need to have an underlying normal distribution. Do we in this case? Since we said that the distribution of tomato weights was nearly normal, the Central Limit Theorem tells us that a sample of size 12 from this distribution will be be very close to normal. Granted our sample size is a little smaller than we would typically like to apply the CLT, but the assumption of normality seems reasonable.

Since $\bar{x}$ is normally distributed, we know

$$\frac{\bar{x} - \mu}{\sqrt{s^2/n}} \;\sim\; t_{n-1},$$

so our test statistic should follow a t-distribution with 11 degrees of freedom. For this distribution, a t-score of 1.48 corresponds to a p-value of 0.083. While this isn't concrete evidence that the fertilizer increases the size of the tomatoes, I'd probably keep using it until I'd generated some more conclusive data.

22.4 Problems with Hypothesis Testing

Now that we've developed the basic structure and intuition of hypothesis testing, we have a sobering confession to make: hypothesis testing is not perfect. Since we ultimately appeal to probabilistic arguments, we can't be absolutely certain of our conclusions. There are two types of errors we could possibly make: we could falsely reject a true null hypothesis, or we could fail to reject a false null hypothesis. Statisiticians, exhibiting their typical flair for nomenclature, have termed these errors **Type I** and **Type II errors** , respectively.

Takeaways for the section:

1. Probabilistic Arguments are never conclusive.
2. There are two kinds of errors we might make: rejecting a true null hypothesis, or failing to reject a false null hypothesis.
3. There is a tradeoff between Type I and Type II errors.

22.4.1 Type I errors

Let's tackle Type I errors first. As we mentioned above, a **Type I error** is the error of rejecting a true null hypothesis. Why would this happen? Suppose we're testing a hypothesis, and that the null hypothesis we've identified is in fact true. Our α-level (say it's 0.05) means that if we observe a result that would happen less than 5% of the time by chance alone, we'll reject the null hypothesis. How often will we see a result that would happen at most 5% of the time by chance alone? Exactly 5% of the time! More generally, the probability of making a Type I error (conditional on the null hypothesis being true) is exactly the α-level.

22.4.2 Type II errors

There is another type of error we could potentially make. We could identify the wrong null hypothesis, but fail to reject it. This is known as a **Type II error**. Type II errors are trickier than Type I errors because they depend on which false null you've identified. Why is this the case? If your null hypothesis is way off, it should be relatively easy to reject. However, the closer the null hypothesis comes to the truth, the more difficult it is to reject. Let's do an example.

Example: Imagine you're trying to estimate the average height of adult males at your college. Let's assume that the true population mean is 6 feet (1.83 meters) with a standard deviation of 3 inches (7.62 centimeters), and that you draw a sample of size 20. If your null hypothesis is that the mean height is 5 feet (1.52 meters), what is the chance of making a Type II error if your α level is 0.05?

Solution: For simplicity, let's assume you know that the standard deviation is 3 inches, so the sample mean should be roughly normally distributed. This implies that you will reject H_0 if your sample average is more than 1.96 standard deviations away from the mean. With a sample of size 20, your standard deviation is $s = \sigma/\sqrt{n} = 3/\sqrt{20} \approx 0.67$ inches. Thus you will reject H_0 if your sample mean is greater than 5 foot 1.31 inches, or less than 4 foot 10.69 inches. Alternatively, you will make a Type II error if your sample mean falls between 4 foot 10.69 inches and 5 foot 1.31 inches. We know that the true distribution for samples of size 20 is normal with mean 6 feet and standard deviation 0.67 inches. Given this, you can show that the probability of making a Type II error is about 10^{-57} – no real worries here!

What if we had formulated the more reasonable null hypothesis that the average height is 5 foot 11 inches? The same analysis carries through, except now we would make a Type II error if our sample mean was between 5 foot 9.69 inches and 6 foot 0.31 inches. This type of sample would happen over 67% of the time, so it's very likely that we would make a Type II error. The reason is that our null hypothesis is so close to the truth that we will frequently see data that are consistent with the null. The probability of making a Type II error (conditional on the null hypothesis being false) is called β. We will use this idea to describe power, which we will talk about in a later section.

22.4.3 Error Rates and the Justice System

An important concept related to Type I and Type II errors is the error rate. The **Type I error rate** is the probability of committing a Type I error if the null hypothesis is true. Similarly, the **Type II error rate** is the probability of committing a Type II error when the null hypothesis is false. In our last example, the Type II error rate was 0.67.

One way to think about Type I and Type II errors is in the context of a criminal trial. Jurors formulate the null hypothesis that the defendant is innocent, and then listen to testimony. If the evidence presented seems sufficiently unlikely under the premise of innocence, the juror submits a guilty vote, while if the evidence is not substantial enough the juror enters a not guilty vote (note the guilty/not guilty terminology: in much the same manner that we never accept null hypotheses but only reject or fail to reject them, criminal trials never find someone "innocent," only guilty or not guilty). The possible outcomes are summarized in Table 22.1.

	Defendant Innocent	**Defendant Guilty**
Convict	Type I Error	Good
Don't Convict	Good	Type II Error

Table 22.1: The possible outcomes of a criminal trial. The null hypothesis is that the defendant is innocent; had it been that he is guilty, the Type I and Type II errors would be flipped.

We see that in our trial case, a Type I error means convicting an innocent person, while a Type II error would mean letting a guilty person go free (hence Type I errors are also called *false positives*, while Type II errors are called *false negatives*). Neither of these outcomes seems ideal – is there a way we could lower the incidence of both Type I and Type II errors? At first glance, it might seem not. If a juror decided to require more convincing evidence in the hopes of reducing Type I errors, she would automatically make it harder to convict *anyone*, including a guilty defendant. Thus by trying to reduce the incidence of Type I errors, she necessarily increased the likelihood of committing a Type II error. There is, however, one thing we could do to reduce both types of errors – simply listen to more evidence. The clearer we are about what is going on, the less likely we are to make an error in judgement.

Example: A local college is testing its students for swine flu, which is known to raise your white blood cell (WBC) count. Suppose the WBC count of healthy people is normally distributed with a mean of 7000 WBCs per microliter and standard deviation of 1000, and that the WBC count of people with swine flu is normally distributed with a mean of 11,000 WBCs per microliter and a standard deviation of 1500. Since the disease is very communicable, the college wants to quarantine people they suspect of having it. If the college quarantines everyone with a WBC count over 9,000, what are the corresponding Type I and Type II error rates? If the college wants to have a Type II error rate of 0.05, where should they set their threshold?

Solution: Let's first consider the case where the college quarantines anyone with a WBC count over 9,000. What is a Type I error in this case? A Type I error is a false positive, so a Type I error would be deciding that a healthy person had swine flu.

The odds of this are just the odds that a healthy person has a WBC count over 9,000, which happens with probability $1 - \Phi(2) \approx 0.023$ (remember Φ is the cumulative distribution function for the standard normal).

What about Type II errors? These happen when we determine a sick person is actually healthy, which occurs whenever a sick person has a WBC count less than 9,000. This happens with probability $\Phi(-1.33) \approx 0.091$.

If the college wanted a Type II error rate of 0.05, what should their threshold be? Again, a Type II error will occur whenever a sick person has a WBC count that falls below the threshold. Thus we need to find the z such that $\Phi(z) = 0.05$. Using a standard normal table, we find $z = -1.64$. Thus the threshold should be at $11000 - 1.64 * 1500 = 8540$.

Just for fun, what's the Type I error rate in the case where the threshold is a WBC count of 8540? That's the probability that a healthy person has a WBC count over 8540, which is $1 - \Phi(1.54) \approx 0.062$. Again, we see the general phenomenon that lowering the incidence of Type II errors increases the likelihood of Type I errors.

22.4.4 Power

In the case that the null hypothesis is false, we would ideally want our test to correctly reject it. The **power** of the test allows us to measure how likely we are to successfully reject the null hypothesis given that it is indeed false. Thus, since β is the probability of making a Type II error, or failing to reject a false null hypothesis, then the power of the test is the probability that our test successfuly rejects a false null hypothesis, or $1 - \beta$.

Clearly, the lower the power of our test, the more likely we are to make a Type II error. As a result, the first place to look when we fail to reject the null hypothesis is the test's power and ways to increase it. Often, problems emerge from having too small of a sample size. It is very possible that the sample size is too small to be representative of the population we wish to generalize on, resulting in us failing to detect any significant difference when one exists outside of our sample. We will have more data with a larger sample size, meaning we should have more evidence that the null hypothesis is false if it is indeed false. In other words, if the population is on average significantly different, then increasing the sample size, which makes the sample closer to the true population, should increase the likelihood of detecting the difference.

For example, suppose we want to know if 30 passengers on an airplane enjoy a new lunch menu item, fish, more or less than the typically offered meal of steak (assuming no one is vegetarian). The null hypothesis would be that there is no difference in enjoyment. Suppose 25 passengers do enjoy the fish more or less than the steak, but 5 are indifferent. If we have a small sample size of say 4 and happen to select only among the 5 who are indifferent, then we will fail to reject the null hypothesis even though it is clearly false for our population of 30 passengers. If we sampled 10 passengers, we will already detect the difference since only 5 are indifferent. Thus, one way to increase power is to increase sample size. By the end of this chapter, we will know a few ways of increasing power.

22.4.5 Effect Size

As established in the previous section, the power of a test is a quantifier of how far the truth lies from our hypothesized value. This distance between the null value, p_0, and the truth, p, is the effect size. Since we don't know the true value, however, we estimate the effect size as the difference between the null and observed values.

Effect size is crucial to understanding the power of a hypothesis test. A large effect means a smaller Type II error, and thus a larger power. Additionally, a small effect means a larger Type II error and therefore a smaller power. Therefore, know the effect size and sample size helps us determine power. The issue here is that, when designing tests, we won't yet know the observed values and therefore can't calculate the effect size. This means we have to try several different effect sizes and look at their consequences. When researchers design a study, they often know what effect size matters to their conclusions and use that to estimate n, the sample size they'll need.

22.5 Chi-square Distributions, Goodness of fit

Up to this point, we've only done hypothesis tests on the mean. However, there are tests we can run on other interesting parameters as well. In this section we discuss how to perform hypothesis tests on the variance, and how to test a model itself. Before we can do this, however, we need to introduce a very important distribution known as the χ^2 distribution.

22.5.1 Chi-Square distributions and tests of Variance

Suppose X has a standard normal distribution. What distribution does X^2 follow? We can find the density function for X^2 rather easily:

$$\text{Prob}(a \le X^2 \le b) \;=\; \text{Prob}(\sqrt{a} \le X \le \sqrt{b}) + \text{Prob}(-\sqrt{b} \le X \le -\sqrt{a}) \quad (22.5)$$

$$= \; 2 \cdot \text{Prob}(\sqrt{a} \le X \le \sqrt{b}) \;=\; 2\int_{\sqrt{a}}^{\sqrt{b}} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} dx \quad (22.6)$$

Let's make the substitution $u = x^2$. Then our limits of integration go from a to b, and $dx = \frac{du}{2\sqrt{u}}$:

$$\text{Prob}(a \le X^2 \le b) \;=\; 2\int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-u/2} \frac{du}{2\sqrt{u}} \;=\; \int_a^b \frac{1}{\sqrt{2\pi}}\, u^{-1/2}\, e^{-u/2} du \quad (22.7)$$

So we see the density function for X^2 is given by $1/\sqrt{2\pi}\; x^{-1/2} e^{-x/2}$. This is one of the most common densities in statistics, and is called the **chi-square distribution** . Like the t-distribution, the χ^2 distribution is a family of distributions parametrized by their degrees of freedom. More generally, a χ^2 distribution is defined as follows:

Notice that this definition immediately implies that if $Y \sim \chi^2_k$ and $X \sim \chi^2_l$, then $X + Y \sim \chi^2_{k+l}$ so long as X and Y are independent.

The density function for a χ^2 distribution with k degrees of freedom is given by:

χ^2 distribution with k degrees of freedom: Suppose for $1 \leq i \leq k$, X_i are independent, standard normal variables. Then the random variable

$$Y = X_1^2 + X_2^2 + \cdots + X_k^2 \tag{22.8}$$

follows a χ^2 distribution with k degrees of freedom, often denoted χ_k^2.

Figure 22.5: Graphs of the chi-square distribution for different degrees of freedom. MAKE A KEY

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}\, e^{-x/2}, \tag{22.9}$$

where Γ is the gamma function. Graphs of the chi-square distribution are shown in Figure 22.5.

So why on earth are we discussing this funny looking distribution? One reason is that the sample variance of a normally distributed random variable is closely related to the χ^2 distribution. Consider a random variable X which is distributed $\mathrm{N}\left(\mu,\ \sigma^2\right)$, and imagine we've drawn a sample of size n from this random variable. We've seen before that our sample mean $\bar{x}$ should be distributed $\mathrm{N}\left(\mu,\ \sigma^2/n\right)$, so

$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}\left(0, 1\right). \tag{22.10}$$

Now imagine we need to calculate the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \tag{22.11}$$

One natural question to ask is what kind of distribution the sample variance follows. It can't be a normal distribution, because the sample variance is never negative. However, when we calculate the sample variance, we're squaring a whole bunch of $(x_i - \bar{x})$ terms, so maybe we would expect that the χ^2 distribution should come into play. To get at the question of what distribution the sample variance follows, let's first consider something very close to the true variance:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2. \tag{22.12}$$

This is just our regular expression for the variance, except missing a factor of $1/n$ and with an additional $1/\sigma^2$. The reason for these alterations will be made clear in a minute. By adding 0 in a clever manner, we can rearrange this to:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right) + \left(\bar{x} - \mu\right)\right)^2 \tag{22.13}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((x_i-\bar{x})^2+(\bar{x}-\mu)^2+2(x_i-\bar{x})(\bar{x}-\mu)\right). \tag{22.14}$$

Now $\bar{x}-\mu$ is a constant, so $\sum_{i=1}^{n}(\bar{x}-\mu) = n(\bar{x}-\mu)$. We also know that $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$, so $\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) = 0$ as well. Our equation then simplifies to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right). \tag{22.15}$$

One more algebra trick and we're home free:

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\left(\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}\right)^2. \tag{22.16}$$

Let's look at this for a second. We know that $(x_i-\mu)/\sigma$ has a standard normal distribution, so by the definition we gave earlier, the left hand side of this equation is just a χ^2 distribution with n degrees of freedom. Further, we know that $(\bar{x}-\mu)/\sqrt{\sigma^2/n}$ has a standard normal distribution, so the second term on the right hand side of the equation has a χ^2 distribution with one degree of freedom. Therefore the only term left is the first term on the right hand side, which you'll notice looks suspiciously like our formula for the sample variance. It's just $(n-1)s^2/\sigma^2$. So what distribution should it follow? We know that χ^2 distributions have a nice additive property: if $X_1 \sim \chi^2_k$ and $X_2 \sim \chi^2_l$, then $X_1+X_2 \sim \chi^2_{k+l}$, so long as X_1 and X_2 are independent. Therefore, if the first and second terms of the right hand side of equation 22.16 are independent (it turns out they are, although that fact isn't obvious – should we prove this in an Appendix??), we have

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \tag{22.17}$$

So now we see why we've taken so much time to study χ^2 distributions – so long as we're drawing from a normal distribution, the sample variance (more precisely, a multiple of the sample variance) follows a χ^2 distribution! And once we know the distribution of a parameter, we can test hypotheses about it. Let's look at a few examples where we test hypotheses about the variance.

Example: The Coca Cola factory is installing a new machine that dispenses Coke into bottles as they move along the bottling apparatus. To ensure that their bottled products are nearly uniform, they want this machine to dispense 12 ounces of Coke on average, with a standard deviation no greater than 0.05 ounces. Suppose that in a sample of 20 bottles, you find the following amount of liquid:

11.83	12.09	11.93	12.02	11.98
11.97	12.06	12.08	12.06	12.02
12.01	12.10	12.04	11.98	12.04
12.00	12.04	12.09	12.00	11.92.

Is this data consistent with what the company wants?

Solution: We have $\bar{x} = 12.01$, so it seems like the machine is dispensing about the right amount of liquid. If we really wanted to, we could test the mean, but for

now let's concern ourselves with the variance. The sample variance is $s^2 = 0.0045$, meaning the sample standard deviation is 0.067, which is higher than desired. Is this a significant increase, or just due to chance? Let's take our null hypothesis that $\sigma = 0.05$, meaning $\sigma^2 = 0.0025$. From our test statistic in Equation 22.17, we should have (TALK ABOUT THE ASSUMPTIONS YOU'RE MAKING)

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{19}.$$

This is a χ^2 statistic of 34.2. Should we use a one-sided or two-sided test here? Well we're certainly not worried about the variance being too low, so a one-sided test seems reasonable. For a chi-square distribution with 19 degrees of freedom, a test statistic of 34.2 has a one-sided p-value of 0.0174. This is a pretty low p-value, so we'd be concerned that the machine's variance is too high. However, depending on how expensive it is to replace the machine, we might want to collect some more data just to make sure we didn't draw an unusual sample.

Just for fun (and a bit of review), let's test the mean. We want to see whether the machine is dispensing more or less than 12 ounces on average. Taking our null hypothesis to be $\mu = 12$, we know:

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}.$$

This gives us a test-statistic of

$$\frac{12.01 - 12}{\sqrt{\frac{0.0045}{20}}} = 0.67,$$

which should follow a t-distribution with 19 degrees of freedom. This t-value has a corresponding p-value of about 0.5, clearly isn't large enough to reject the null that $\mu = 12$. So while we haven't *proven* that the machine is dispensing more or less than 12 ounces on average, the data we've collected certainly isn't inconsistent with that possibility.

22.5.2 Chi-square distributions and t-distributions

It turns out that χ^2 distributions are related to another distribution we already know and love: the t-distribution. In fact, t-distributions are actually *defined* in terms of χ^2 distributions. The definition goes as follows:

t-distribution with k degrees of freedom: Let Z be the standard normal distribution, and Y_n be the χ^2 distribution with n degrees of freedom. The t-distribution with n degrees of freedom is defined as:

$$\frac{Z}{\sqrt{\frac{Y}{n}}} \sim t_n. \tag{22.18}$$

This might seem like a weird definition, but we'll see in a minute that it's a very natural one. In fact, it comes directly from the situation we found ourselves in a few sections ago: using the sample variance in place of the true variance when performing a z-test. Suppose we have a normally distributed random variable X with mean μ and variance σ^2, and from this variable we draw a random sample of size n. Then we know that

$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1). \tag{22.19}$$

As we showed earlier, the sample variance s^2 obeys:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{22.20}$$

Let's replace σ^2 with s^2 and cleverly multiply by 1:

$$\frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{\bar{x} - \mu}{\sqrt{\frac{(n-1)\sigma^2 \cdot s^2}{(n-1)\sigma^2 \cdot n}}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}}. \tag{22.21}$$

Looking at this equation, we notice that $(\bar{x} - \mu)/\sqrt{\sigma^2/n}$ has a standard normal distribution, and $(n-1)s^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom. Replacing $(\bar{x} - \mu)/\sqrt{\sigma^2/n}$ with Z and $(n-1)s^2/\sigma^2$ with Y_{n-1}, this becomes

$$\frac{Z}{\sqrt{\frac{Y_{n-1}}{n-1}}}, \tag{22.22}$$

which is exactly the definition we gave above for a t-distribution with $n-1$ degrees of freedom.

22.5.3 Goodness of Fit for List Data

One handy thing we can do with hypothesis tests is actually test the models we are working with themselves. Since we're always talking about how important it is to justify the distribution of everything, this is a very useful tool. Suppose we're collecting data and there are k possible outcomes (for example, the birth month of everyone in your probability class, which would have 12 possible outcomes). Then we might generate the list of observed data $\{O_1, O_2, ..., O_k\}$, where O_i is the number of times we observed the i^{th} outcome. We might also have a model for this data in mind, namely that the i^{th} outcome happens with some probability p_i (*e.g.* I think 10% of the people in the probability class were born in January, 7% in February, etc.). Then we can test this hypothesis by using the following test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{22.23}$$

where E_i is the number of times we would have expected to observe event i under the null hypothesis. You'll notice we have suggestively named our test statistic χ^2, and it turns out that this test statistic does follow a χ^2 distribution with $k-1$ degrees of freedom. Notice that the degrees of freedom are determined by the number of categories we have, not the number of data points.

The general proof of why this test statistic follows a χ^2 distribution is rather advanced, but to get a some intuition about the result, let's prove it for the binomial case (that is, when there are only two possible outcomes). Suppose we've gathered n data points and have observed the first outcome O_i times and the second outcome O_2 times. Further suppose we have a model which says that O_1 should happen with probability p_1, and O_2 should happen with probability p_2. Notice that we expect to see the first outcome happen np_1 times, and the second outcome np_2 times. This means our test statistic is:

$$\chi^2 = \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2}. \tag{22.24}$$

We can simplify this a bit, because we know $p_1 + p_2 = 1$, so $p_2 = 1 - p_1$. Also, $O_1 + O_2 = n$, so $O_2 = n - O_1$. Thus we have:

$$\begin{aligned}
\chi^2 &= \frac{(O_1 - np_1)^2}{np_1} + \frac{((n - O_1) - n(1 - p_1))^2}{n(1 - p_1)} \\
&= \frac{(1 - p_1)(O_1 - np_1)^2 + p_1(-O_1 + np_1)^2}{np_1(1 - p_1)} \\
&= \frac{(O_1 - np_1)^2}{np_1(1 - p_1)} = \left(\frac{O_1 - np_1}{\sqrt{np_1(1 - p_1)}}\right)^2
\end{aligned}$$

Since O_1 follows a binomial distribution of size n and probability p_1, the Central Limit Theorem tells us that for large n, $O_1 \approx N(np_1, np_1(1 - p_1))$. Therefore χ^2 is the square of a standard normal distribution, meaning it really does follow a χ^2 distribution with one degree of freedom (when n is large enough to use the Central Limit Theorem). Our proof also helps us see why the number of degrees of freedom is always one less than the number of categories: if there are k possibilities occuring with hypothesized probabilities $\{p_1,\ p_2, ...,\ p_k\}$, then we can use the restrictions $p_1 + p_2 + ... + p_k = 1$ and $O_1 + O_2 + ... + O_k = n$ to eliminate one of the categories from our equation, and then rearrange the remaining terms to get a sum of $k - 1$ squares of standard normal distributions (that is, a χ^2 distribution with $k - 1$ degrees of freedom).

Example: Suppose you're investigating the distribution of birth months among major league baseball players. At first glance, you might expect that birth month should have nothing to do with athletic ability, and would take your null hypothesis to be that birth months are equally distributed among baseball players. Here's the data for American Major Leaguers born after 1950 who debuted before 2005 (THIS DATA TAKEN FROM http://www.slate.com/id/2188866):

Birth Month	Number of Major Leaguers
January	387
February	329
March	366
April	344
May	336
June	313
July	313
August	503
September	421
October	434
November	398
December	371

Given this data, what kind of conclusions can you draw about your hypothesis that birth months should be equally distributed?

Solution: Looking at the data, we notice that there's a considerable bias towards the August through October months. Could this just be chance, or is something going on here? Let's carry out our hypothesis testing as we normally we, and see what happens. Since we're hypothesizing the birth month has no effect, then we would expect that exactly one twelfth of the 4515 players would be born in each month, or 376.25. Our test statistic of interest is the goodness of fit statistic we just developed, which is given by:

$$\chi^2 = \sum_{k=1}^{12} \frac{(\text{Births}_k - 376.25)^2}{376.25} \approx 93.07. \tag{22.25}$$

Since there are 12 months a player could be born in, our test statistic should follow a χ^2 distribution with 11 degrees of freedom. Before we can find our p-value, however, we need to determine whether we want to use a one-sided or two-sided test. Can it really be the case that this test statistic is "too low"? If the data fit our model perfectly, then every term would be zero, and our test statistic would also be zero. Would this make us want to reject our null hypothesis? Not in the slightest! We'd be jumping with joy to have data that good. So it's only cases when our test statistic is large that we worry about, meaning we'll use a one-sided test. For a one-sided test on a chi-square distribution with 11 degrees of freedom, a test statistic of 93.07 has a p-value of 4.1×10^{-15}.

22.6 Two sample tests

One of the most useful applications of hypothesis testing is comparing the means of two (or more!) different samples. These tests are important for any kind of comparative decision making (*e.g.* did more patients regrow hair with drug A than drug B? Do people prefer Coke or Pepsi?). Much as in the one sample case, the exact test we perform depends on what we know about the variances. There are three possibilities (in order of increasing difficulty): we know the variances; we don't know the variances but have reason to believe they are the same; we don't know the variance and they might be different from each other. Let's start with the easiest case.

22.6.1 Two sample z-test: known variances

Suppose the random variables X and Y have variances σ_x^2 and σ_y^2, respectively. Let $\bar{X}$ be the distribution of the sample mean of X in a random sample of size n_x, and $\bar{Y}$ the distribution of the sample mean of Y in a random sample of size n_y. We can compare the means of X and Y by forming the random variable $\bar{X} - \bar{Y}$. If X and Y are independent we have:

$$\text{Var}(\bar{X} - \bar{Y}) \;=\; \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}.$$

Similarly, the expected value of $\bar{X} - \bar{Y}$ is given by:

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_x - \mu_y.$$

If $\bar{X}$ and $\bar{Y}$ are normally distributed, then since sums of normal distributions are still normal, we know

$$\bar{X} - \bar{Y} \;\sim\; \text{N}\left(\mu_x - \mu_y, \sigma_x^2/n_x + \sigma_y^2/n_y\right)$$

We can normalize $\bar{X} - \bar{Y}$ as follows:

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \;\sim\; \text{N}(0, 1) \tag{22.26}$$

Now suppose we wanted to test a hypothesis about the value of $\mu_x - \mu_y$. For simplicity's sake, imagine we wanted to see whether $\mu_x > \mu_y$. We would formulate the null hypothesis that $\mu_x - \mu_y \leq 0$ and take $\mu_x = \mu_y$. We could then sample X and Y, and calculate $\bar{x}$ and $\bar{y}$. Using Equation 22.26 and the fact that under the null hypothesis $\mu_x - \mu_y = 0$, we see that our test statistic is given by:

$$z \;=\; \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}},$$

which should follow a standard normal distribution. As always, we use this z-value to calculate our p-value, and depending on our significance level will decidew whether to reject the null hypothesis or not.

Two sample z-test with known variances: If X and Y are independent random variables with variances σ_x^2 and σ_y^2, respectively, then

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \;\sim\; \text{N}(0, 1)\,.$$

Assumptions: This only holds if $\bar{X}$ and $\bar{Y}$ are normally distributed.

To get some experience using a two-sample test, let's look at an example.

Example: You're working for a sleep researcher who wants to look at the impact of sleep on test scores. To do so, the researcher gathers a group of 28 people, and

randomly assigns half the group to sleep a full 8 hours before coming in the next morning, while assigning the other half to only sleep 4 hours. The next morning, he administers a test to both groups and records their scores. Suppose he finds the following data:

X = Sleep Group	Y = Sleep-Deprived Group
73	76
95	65
93	74
89	59
79	75
90	76
86	71
91	76
98	74
74	84
91	71
90	77
50	96
70	81
$\mu_1 = 83.5$	$\mu_2 = 75.4$
$s_1^2 = 168.6$	$s_2^2 = 73.32$

Given this data, can you conclude at a 5% significance level that sleep helps test performance?

Solution: After collecting this data and calculating the sample mean and variances for the two populations, we formulate the following hypotheses:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 \leq 0 \\ H_a &: \mu_1 - \mu_2 > 0 \end{aligned} \tag{22.27}$$

That is, we assume that sleep has no benefit, and hope to reject the null hypothesis in favor of the alternative hypothesis that sleep is indeed beneficial. Notice that we're performing a one-sided test, which seems justified because so much evidence suggests that sleep certainly isn't detrimental. What should our test statistic be in this case? If we assume the null hypothesis, we can take $\mu_1 - \mu_2 = 0$. Since we're interested in testing whether the means are different from each other, it seems natural to consider the random variable $\bar{X} - \bar{Y}$. If we assume both $\bar{X}$ and $\bar{Y}$ are normally distributed, then we know from our discussion above that

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim \mathrm{N}(0, 1).$$

(You might be wondering why we lost the $\mu_x - \mu_y$ term: remember that we're assuming the null hypothesis to be true, which says that $\mu_x - \mu_y = 0$.) You might've spotted an issue here: the formula requires us to use σ_x^2 and σ_y^2, but all we have is s_x^2 and s_y^2! While we will ultimately fix this problem by appealing to a t-distribution (just as we did for one sample tests), for now let's assume we've gotten the variances right and let $\sigma_x^2 = 168.6$ and $\sigma_y^2 = 73.32$. Then our test statistic is:

$$z = \frac{83.5 - 75.4}{\sqrt{\frac{168.6}{14} + \frac{73.32}{14}}} \approx 1.95$$

This z-score (remember we're doing a one-sided test!) corresponds to a p-value of about 0.0256, which provides pretty solid evidence that sleep is indeed beneficial to test performance.

22.6.2 Two sample t-test: unknown but same variances

As we saw in the last example, using a z-test for a two sample hypothesis tests requires us to have knowledge about the variances. Typically, we do not know the variances beforehand, and will estimate them from the sample. We can then use the sample variances in place of the real variances, at the cost of making all of our estimates a little more uncertain. There are actually two cases to consider: when the variances of the samples is the same, and when the variances of the samples might not be the same. The "same variances" case might seem a little artificial, but we begin with it because it contains all the intuition of the general case, and the math is little more straightforward.

Suppose we know X and Y have the same variance σ^2, but we don't know what it is. In that case, much as with our first introduction to the t-test, we estimate the variance. Before we generated the sample variance s_x^2. However, since we have two variables now, we calculate both s_x^2 and s_y^2. How do we combine them to form one estimate? One possibility would be to disregard s_y^2 altogether and only use s_x^2. This would work, since we know s_x^2 is an unbiased estimator of σ^2. But we can do better than this. Since we've sampled both X and Y, only using s_x^2 would be the equivalent of throwing data away (which is never a good idea!). So how should we combine s_x^2 and s_y^2 to get a better estimate? We could weight the two equally and take our estimate to be $s_p^2 = \frac{1}{2}\left(s_x^2 + s_y^2\right)$ (we call it s_p^2 because we are pooling the variances - this is known as the *pooled variance*) . But this also can't be ideal: imagine we had 1000 samples from X and 10 from Y. Would we really want to give s_y^2 half the weight in our estimate? Probably not, since X should be a much better estimate. It turns out the best estimate we can make is a weighted average:

Pooled Variance for two samples: If X and Y are independent random variables with common variance σ^2 then the best estimate of the variance is the *pooled variance*, s_p^2, given by

$$s_p^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}.$$

The equation for the pooled variance isn't too complicated, but it does look a bit confusing. Thankfully we can clean it up a little. Let $r = \frac{n_x - 1}{n_x + n_y - 2}$. Notice the numerator is just the number of degrees of freedom for s_x^2, and the denominator is the sum of the degrees of freedom for s_x^2 and s_y^2. The equation then becomes

$$s_p^2 = r \cdot s_x^2 + (1 - r) \cdot s_y^2,$$

so this is indeed a weighted averaged of s_x^2 and s_y^2, with the sample variances being weighted by their degrees of freedom.

Now that we have our estimate for the variance, we proceed just as we did in the one sample case. We replace σ^2 by s_p^2, and our test statistic goes from following a normal distribution to following a t-distribution . More specifically:

Two sample t-test - same but unknown variance: If X and Y are independent random variables with common variance σ^2 and estimated pooled variance s_p^2, then

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} \sim t_{n_x+n_y-2}.$$

Assumptions: We need to know that $\bar{X}$ and $\bar{Y}$ are normally distributed.

Notice that our test statistic follows a t-distribution with $n_x + n_y - 2$ degrees of freedom, which is just the sum of the degrees of freedom for s_x^2 and s_y^2.

Important: As always, to use the t-test we need to have an underlying normal distribution. In this case, we need to know that

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n_x}+\frac{\sigma^2}{n_y}}} \sim \mathrm{N}(0,1).$$

Thankfully, for most reasonable sample sizes and distributions the Central Limit Theorem assures us that $\bar{X}$ and $\bar{Y}$ will be nearly normal, even if X and Y are not.

Once we have our test statistic, hypothesis testing is business as usual. Let's go through an example to make sure everything's clear.

Example: Hi
Solution: Bye

22.6.3 Unknown and Different Variances

Now that we've discussed what to do for unknown but equal variances, the final (and most general) case to talk about is when we know neither σ_x^2 nor σ_y^2, and σ_x^2 might not equal σ_y^2. In this case, we need to estimate both variances. As usual, we calculate s_x^2 and s_y^2 and we use them in place of σ_x^2 and σ_y^2. In an ideal world, we would now state that

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}$$

follows a t-distribution with some easy to calculate number of degrees of freedom. Unfortunately, this is not an ideal world, and the above equation is not true. Why not? You'll remember from earlier that the definition of a t-distribution is

$$\frac{Z}{\sqrt{Y/n}} \sim t_n,$$

where Z is the standard normal random variable, and Y has a χ^2 distribution with n degrees of freedom. In order to prove, in the one sample case, that replacing σ^2 with s^2 gives us a t-distribution, we had to show that

$$\frac{(n-1)\,s^2}{\sigma^2} \;\sim\; \chi^2_{n-1}.$$

The problem we run into now is this: before, we were able to say that s^2/n was some multiple of a χ^2 distribution. But in the two sample case, we *cannot* say that $s_x^2/n_x + s_y^2/n_y$ is a multiple of a χ^2 distribution, and so we cannot conclude that the expression above follows a t-distribution. It has *some* distribution, but unfortunately it can't be expressed nicely in terms of the distributions we've met already. So now what? Generally in mathematics, when we cannot find an analytic expression we do the next best thing: approximate. Thankfully in this case there's a well known and easy to use approximation which states that in the case of unknown variances, the test statistic

$$\frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

is approximately distributed t_ν, where ν is the number of degrees of freedom, given by

$$\nu \;=\; \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{s_x^4}{n_x^2(n_x-1)} + \frac{s_y^4}{n_y^2(n_y-1)}}.$$

As a bit of terminology, this approximate test is known as Welch's t-test. So we're almost in an ideal world – our test statistic nearly follows a t-distribution. You'll notice, however, that we do not have a "some easy to calculate number of degrees of freedom." The equation for ν is a bit daunting, so let's take a look at it. It's not too hard to show that ν is bounded above by $n_x + n_y - 2$, and below by $\min(n_x - 1, n_y - 1)$. This makes sense, since we shouldn't have more degrees of freedom than s_x^2 and s_y^2 have combined, and we shouldn't have fewer than the most uncertain measurement we made. Another limit to consider is when n_x is large (the same argument holds for when n_y is large). When n_x is large, ν approaches $n_y - 1$. This is nice, since when n_x is really big we're not worried about the uncertainty in s_x^2, so our limiting factor becomes the degrees of freedom for s_y^2.

Another question you might have is, "Does this equation always give us an integer number of degrees of freedom, and, if not, what does a non-integer number of degrees of freedom mean?" This equation certainly does not always return an integer! However, we're not too worried if the equation tells us $\nu = 19.394$ because this is an approximation. Remember, the typical interpretation for the degrees of freedom is the number of independent observations that contribute to a measurement (and in that case, 19.934 degrees of freedom makes no sense). But with an approximation like this, that interpretation isn't valid. All the equation is telling us is that this test statistic behaves like a t-distribution with 19.394 "degrees of freedom." Since the formula for the t-distribution uses the degrees of freedom as an input, this is a perfectly well-defined mathematical function.

Once we have our test statistic and know the distrubtion it follows (approximately), hypothesis testing is the same as always. Table 22.6.3 summarizes two sample tests in the case of unknown and potentially unequal variances.

Two sample t-test - unknown, potentially unequal variances: If X and Y are independent random variables with variances σ_x^2 and σ_y^2, respectively, then

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} \approx t_\nu$$

with ν degrees of freedom given by

$$\nu = \frac{\left(\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}\right)^2}{\frac{s_x^4}{n_x^2(n_x-1)}+\frac{s_y^4}{n_y^2(n_y-1)}}.$$

Assumptions: As always, we need $\bar{X}$ and $\bar{Y}$ to be normally distributed.

As an example, let's return to the sleep researcher problem we gave in the first section.

Example: The same setup as (REFERENCE EXAMPLE)

Solution: The analysis we gave in Section 1 still holds, right up through the p-value we gave. As a reminder, we had

$$\begin{aligned} \mu_x &= 83.5;\ \mu_y = 75.4 \\ s_x^2 &= 168.6;\ s_y^2 = 73.32, \end{aligned} \tag{22.28}$$

and the null hypothesis that $\mu_x - \mu_y = 0$. Our test statistic is:

$$t = \frac{83.5-75.4}{\sqrt{\frac{168.6}{14}+\frac{73.32}{14}}} \approx 1.95.$$

Now, however, we also need to calculate the number of degrees of freedom. Using the equation from Table 22.6.3, we see

$$\nu = \frac{\left(\frac{168.6}{14}+\frac{73.32}{14}\right)^2}{\frac{168.6^2}{14^2(14-1)}+\frac{73.32^2}{14^2(14-1)}} \approx 22.5.$$

For a t-distribution with 22.5 degrees of freedom, a t-score of 1.95 corresponds to a p-value of 0.0318. So while our result is a little less certain (with a z-test we had a p-value of 0.025), we can still safely reject the null at a 5% significance level.

22.7 Summary

Ernest Rutherford: `If your experiment needs statistics, you ought to have done a better experiment.`

In this chapter, we discussed the probabilities associated with hypothesis testing in statistics. We learned about one- and two-side z- and t-tests, which are used for

testing the truth of a hypothesis, namely when randomness is involved. A z-test assumes knowledge of the population standard deviation, while a t-test uses degrees of freedom to adjust the estimated t-distribution based on sample size and variaous other factors. A one-sided test is used to determine if a result is strictly greater than or less than the null hypothesis value, whereas a two-sided test determines if a result is simply different than the null value (either above or below). All of these tests assume the hypothesis to be true and use collected data to try and disprove it. The probability value (p-value) associated with these tests represents the probability of measuring a value as or more extreme than the observed value, given the null hypothesis. When this value is below a certain theshold, known as the α-level, we can conclude statistically significant evidence against the null and assume the alternative hypothesis.

Additionally, we learned how to compute test-statistics, the values used to determine the rarity of observed events, which are quantified as standardized differences from the expected result. We then elaborated more on p-values and the problems, including Type I and II errors, and misconceptions related to them, such as the p-value being the probability that the null hypothesis is true. We also discuss the power of a test and how it relates to effect size, sample size, and Type II error, or β. Lastly, we covered the Chi-square distribution and its application in the Goodness of Fit hypothesis test.

22.8 Additional Problems

Problem 22.8.1 *Imagine that systolic blood pressure in healthy people is known to be distributed according to $N \sim (110, 100)$. If a patient in a hospital has a systolic blood pressure of 137, find the Z-score for this patient, the associated p-value and use it to test a hypothesis about whether this persons blood pressure is "normal" (in the common usage).*

Problem 22.8.2 *Imagine we have an eight sided dice which we role twice. Our assumption is that the die is fair. Our null hypothesis is that the dice is fair. We roll it twice and the sum of the die is 16. Is this sufficient evidence to say that the die is not fair at a significance level of $\alpha = .05$?*

Problem 22.8.3 *Plot the number of heads we would need to call the number of heads in n tosses of a fair coin 'unusual' if we define unusual as a result extreme enough we expect to see it only 5% of the time. Plot this number of heads as a percentage of n for n from 5 to 100.*

Problem 22.8.4 *Plot the number of heads we would need to call the number of heads 'unusual' if we define unusual as a result extreme enough we expect to see it only 5% of the time in n tosses of a coin. Use a normal approximation for the distribution of heads. Plot this number of heads for as a percentage of n for n from 5 to 100. Above what n does the number of heads needed begin to line up pretty well with the binomial model?*

Problem 22.8.5 *Show that s^2 is an unbiased estimator for σ^2.*

Problem 22.8.6 *Show that as $\nu \to \infty$ the t-distribution approaches the standard normal distribution.*

Problem 22.8.7 *Imagine you are performing a statistical test to determine if employment has actually increased in the past 3 years, or whether the change in employment can be explained by random fluctuations. What is type I error in this case? What is a type II error?*

Problem 22.8.8 *What are three ways of increasing power? Explain.*

Problem 22.8.9 *Suppose a group of people are either lactose intolerant or not. The probability that a lactose intolerant person eats vanilla ice cream is 30% and the probability that someone who is not lactose intolerant eats vanilla ice cream is 65%. Here, the null hypothesis is the person is not lactose intolerant, whereas the alternative hypothesis is the person is lactose intolerant. Suppose you are told that if a person does not eat ice cream, then we reject the null hypothesis. What is the probability of a Type I error? A Type II error?*

Problem 22.8.10 *Imagine that we planted two types of peppers in the garden, which look very similar. One is a hot pepper and the other is not so we do not want to confuse them, but we unfortunately forgot to label them. There are an equal number of both types. Luckily, the hot peppers are smaller than the sweet peppers. The sizes of both peppers are normally distributed with standard deviation 1 inch, but the mean length of the hot peppers is 3 inches and the mean length of the hot peppers is 4.5 inches. We randomly pick a pepper. Our null hypothesis is that it is hot. There is a penalty if we miss classify it, since we will put it in the wrong food. If we misclassify it as hot when it is sweet, our utility function is -5. If we misclasify it as hot when it is sweet, our utility is negative 3. If we correctly classify it, our utility is 10 either way. Set a cutoff that maximizes our utility function for a randomly selected pepper.*

Problem 22.8.11 *Your friend (who is not the most honest person) claims he has made a die that will roll one through 3 10% of the time each 4 20% of the time and 5 and 6 both 25% of the time. In 200 rolls, there were 15 1s, 20 2s, 22 3s, 45 4s, 53 5s, 45 6s. Should you trust your friend?*

Problem 22.8.12 *There are two probability classes, each with 30 students. On the first exam, the average score for the first class is 84% and the standard deviation is 9%. The second class scores an average of a 90% with a standard deviation of 6%. Is this significant evidence to say that the classes have significantly different mean scores? Comment on the conditions for the test.*