6.3: Basic Hypothesis Testing

Let's start with a motivating example, described somewhat more casually than the rest of the work we usually do, but whose logic is exactly that of the scientific standard for hypothesis testing.

EXAMPLE 6.3.1. Suppose someone has a coin which they claim is a fair coin (including, in the informal notion of a fair coin, that successive flips are independent of each other). You care about this fairness perhaps because you will use the coin in a betting game.

How can you know if the coin really is fair?

Obviously, your best approach is to start flipping the coin and see what comes up. If the first flip shows *heads* [*H*], you wouldn't draw any particular conclusion. If the second was also an *H*, again, so what? If the third was still *H*, you're starting to think there's a run going. If you got all the way to ten *H*s in a row, you would be very suspicious, and if the run went to 100 *H*s, you would demand that some other coin (or person doing the flipping) be used.

Somewhere between two and 100 *H*s in a row, you would go from bland acceptance of fairness to nearly complete conviction that this coin is not fair – why? After all, the person flipping the coin and asserting its fairness could say, correctly, that it is possible for a fair coin to come up *H* any number of times in a row. Sure, you would reply, but it is very unlikely: that is, given that the coin is fair, the conditional probability that those long runs without *T*s would occur is very small.

Which in turn also explains how you would draw the line, between two and 100 *H*s in a row, for when you thought the the improbability of that particular run of straight *H*s was past the level you would be willing to accept. Other observers might draw the line elsewhere, in fact, so there would not be an absolutely sure conclusion to the question of whether the coin was fair or not.

It might seem that in the above example we only get a probabilistic answer to a yes/no question (is the coin fair or not?) simply because the thing we are asking about is, by nature, a random process: we cannot predict how any particular flip of the coin will come out, but the long-term behavior is what we are asking about; no surprise, then, that the answer will involve likelihood. But perhaps other scientific hypotheses will have more decisive answers, which do not invoke probability.

Unfortunately, this will not be the case, because we have seen above that it is wise to introduce probability into an experimental situation, even if it was not there originally, in order to avoid bias. Modern theories of science (such as quantum mechanics, and also, although in a different way, epidemiology, thermodynamics, genetics, and many other sciences) also have some amount of randomness built into their very foundations, so we should expect probability to arise in just about every kind of data.

Let's get a little more formal and careful about what we need to do with hypothesis testing.

The Formal Steps of Hypothesis Testing

1. State what is the population under study.
2. State what is the variable of interest for this population. *For us in this section, that will always be a quantitative variable* X.
3. State which is the resulting population parameter of interest. *For us in this section, that will always be the population mean* μ_X *of* X.
4. State two hypotheses about the value of this parameter. One, called the **null hypothesis** and written H_0, will be a statement that the parameter of interest has a particular value, so

$$H_0 : \mu_X = \mu_0 \tag{6.3.1}$$

   where μ_0 is some specific number. The other is the interesting alternative we are considering for the value of that parameter, and is thus called the **alternative hypothesis**, written H_a. The alternative hypothesis can have one of three forms:

$$\begin{aligned} H_a &: \mu_X < \mu_0 \text{ ,} \\ H_a &: \mu_X > \mu_0 \text{ , or} \\ H_a &: \mu_X \neq \mu_0 \text{ ,} \end{aligned}$$

   where μ_0 is the same specific number as in H_0.
5. Gather data from an SRS and compute the sample statistic which is best related to the parameter of interest. *For us in this section, that will always be the sample mean* $\overline{X}$
6. Compute the following conditional probability

$$p = P\left(\begin{array}{c}\text{getting values of the statistic which are as extreme,}\\ \text{or more extreme, as the ones you did get}\end{array}\;\middle|\; H_0\right)\ . \tag{6.3.2}$$

This is called the ***p*-value of the test**.

7. If the p-value is sufficiently small – typically, $p < .05$ or even $p < .01$ – announce

   "We reject H_0, with $p = \langle \text{number here} \rangle$."

   Otherwise, announce

   "We fail to reject H_0, with $p = \langle \text{number here} \rangle$."

8. Translate the result just announced into the language of the original question. As you do this, you can say *"There is strong statistical evidence that ..."* if the p-value is very small, while you should merely say something like *"There is evidence that..."* if the p-value is small but not particularly so.

Note that the hypotheses H_0 and H_a are *statements*, not numbers. So ***don't** write something like* $H_0 = \mu_X = 17$; you might use

$$H_o: \text{ " } \mu_X = 17 \text{ "} \tag{6.3.3}$$

or

$$H_o: \ \mu_X = 17 \tag{6.3.4}$$

(we always use the latter in this book).

How Small is Small Enough, for *p*-values?

Remember how the p-value is defined:

$$p = P\left(\begin{array}{c}\text{getting values of the statistic which are as extreme,}\\ \text{or more extreme, as the ones you did get}\end{array}\;\middle|\; H_0\right)\ . \tag{6.3.5}$$

In other words, if the null hypothesis is true, maybe the behavior we saw with the sample data would sometimes happen, but if the probability is very small, it starts to seem that, under the assumption H_0 is true, the sample behavior was a crazy fluke. If the fluke is crazy enough, we might want simply to say that since the sample behavior actually happened, it makes us doubt that H_0 is true at all.

For example, if $p = .5$, that means that under the assumption H_0 is true, we would see behavior like that of the sample about every other time we take an SRS and compute the sample statistic. Not much of a surprise.

If the $p = .25$, that would still be behavior we would expect to see in about one out of every four SRSs, when the H_0 is true.

When p gets down to .1, that is still behavior we expect to see about one time in ten, when H_0 is true. That's rare, but we wouldn't want to bet anything important on it.

Across science, in legal matters, and definitely for medical studies, we start to reject H_0 when $p < .05$. After all, if $p < .05$ and H_0 is true, then we would expect to see results as extreme as the ones we saw in fewer than one SRS out of 20.

There is some terminology for these various cut-offs.

DEFINITION 6.3.2. When we are doing a hypothesis test and get a p-value which satisfies $p < \alpha$, for some real number α, we say the data are **statistically significant at level α**. Here the value α is called the **significance level** of the test, as in the phrase *"We reject H_0 at significance level α,"* which we would say if $p < \alpha$.

EXAMPLE 6.3.3. If we did a hypothesis test and got a p-value of $p = .06$, we would say about it that the result was statistically significant at the $\alpha = .1$ level, but not statistically significant at the $\alpha = .05$ level. In other words, we would say *"We reject the null hypothesis at the $\alpha = .1$ level,"* but also *"We fail to reject the null hypothesis at the $\alpha = .05$ level,"*.

FACT 6.3.4. The courts in the United States, as well as the majority of standard scientific and medical tests which do a formal hypothesis test, use the significance level of $\alpha = .05$.

In this chapter, when not otherwise specified, we will use that value of $\alpha = .05$ as a default significance level.

EXAMPLE 6.3.5. We have said repeatedly in this book that the heights of American males are distributed like $N(69, 2.8)$. Last semester, a statistics student named Mohammad Wong said he thought that had to be wrong, and decide to do a study of the

question. MW is a bit shorter than 69 inches, so his conjecture was that the mean height must be less, also. He measured the heights of all of the men in his statistics class, and was surprised to find that the average of those 16 men's heights was 68 inches (he's only 67 inches tall, and he thought he was typical, at least for his class[12]). Does this support his conjecture or not?

Let's do the formal hypothesis test.

The population that makes sense for this study would be all adult American men today – MW isn't sure if the claim of American men's heights having a population mean of 69 inches was *always* wrong, he is just convinced that it is wrong *today*.

The quantitative variable of interest on that population is their height, which we'll call X.

The parameter of interest is the population mean μ_X.

The two hypotheses then are

$$H_0 : \mu_X = 69 \quad \text{and}$$
$$H_a : \mu_X < 69 \, ,$$

where the basic idea in the null hypothesis is that the claim in this book of men's heights having mean 69 is true, while the new idea which MW hopes to find evidence for, encoded in alternative hypothesis, is that the true mean of today's men's heights is less than 69 inches (like him).

MW now has to make two bad assumptions: the first is that the 16 students in his class are an SRS drawn from the population of interest; the second, that the population standard deviation of the heights of individuals in his population of interest is the same as the population standard deviation of the group of all adult American males asserted elsewhere in this book, 2.8. These are definitely **bad assumptions** – particularly that MW's male classmates are an SRS of the population of today's adult American males – but he has to make them nevertheless in order to get somewhere.

The sample mean height $\overline{X}$ for MW's SRS of size $n = 16$ is $\overline{X} = 68$.

MW can now calculate the p-value of this test, using the Central Limit Theorem. According to the CLT, the distribution of X is N (69, 2.8/ $\sqrt{16}$). Therefore the p-value is

$$p = P\left(\begin{array}{c} \text{MW would get values of } \bar{X} \text{ which are as} \\ \text{extreme, or more extreme, as the ones he did get} \end{array} \middle| H_0 \right) = P(\bar{X} < 69)$$

Which, by what we just observed the CLT tells us, is computable by

normalcdf(−9999, 68, 69, 2.8/ $\sqrt{16}$

on a calculator, or

NORM.DIST(68, 69, 2.8/SQRT(16), 1)

in a spreadsheet, either of which gives a value around .07656 .

This means that if MW uses the 5% significance level, as we often do, the result is not statistically significant. Only at the much cruder 10% significance level would MW say that he rejects the null hypothesis.

In other words, he might conclude his project by saying

> *-2.5mm"My research collected data about my conjecture which was statistically insignificant at the 5% significance level but the data, significant at the weaker 10% level, did indicate that the average height of American men is less than the 69 inches we were told it is ($p = .07656$)."*

People who talk to MW about his study should have additional concerns about his assumptions of having an SRS and of the value of the population standard deviation

Calculations for Hypothesis Testing of Population Means

We put together the ideas in §3.1 above and the conclusions of the Central Limit Theorem to summarize what computations are necessary to perform:

FACT 6.3.6. Suppose we are doing a formal hypothesis test with variable X and parameter of interest the population mean μ_X. Suppose that somehow we know the population standard deviation σ_X of X. Suppose the null hypothesis is

$$H_0 : \mu_X = \mu_0 \tag{6.3.6}$$

where μ_0 is a specific number. Suppose also that we have an SRS of size n which yielded the sample mean $\overline{X}$. Then exactly one of the following three situations will apply:

1. If the alternative hypothesis is $H_a : \mu_X < \mu_0$ then the p-value of the test can be calculated in any of the following ways
   1. the area to the left of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
   2. **normalcdf**$(-9999, \overline{X}, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
   3. NORM.DIST($\overline{X}$, μ_0, σ_X/SQRT(n), 1) on a spreadsheet.
2. If the alternative hypothesis is $H_a : \mu_X > \mu_0$ then the p-value of the test can be calculated in any of the following ways
   1. the area to the right of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
   2. **normalcdf**$(\overline{X}, 9999, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
   3. 1-NORM.DIST($\overline{X}$, μ_0, σ_X/SQRT(n), 1) on a spreadsheet.
3. If the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ then the p-value of the test can be found by using the approach in exactly one of the following three situations:
   1. If $\overline{X} < \mu_0$ then p is calculated by any of the following three ways:
      1. two times the area to the left of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
      2. 2 **normalcdf**$(-9999, \overline{X}, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
      3. 2 NORM.DIST($\overline{X}$, μ_0, σ_X/SQRT(n), 1) on a spreadsheet.
   2. If $\overline{X} > \mu_0$ then p is calculated by any of the following three ways:
      1. two times the area to the right of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
      2. 2 **normalcdf**$(\overline{X}, 9999, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
      3. 2 (1-NORM.DIST($\overline{X}$, μ_0, σ_X/SQRT(n), 1)) on a spreadsheet.
   3. If $\overline{X} = \mu_0$ then $p = 1$.

Note the reason that there is that multiplication by two if the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ is that there are two directions – the distribution has two tails – in which the values can be more extreme than $\overline{X}$. For this reason we have the following terminology:

DEFINITION 6.3.7. If we are doing a hypothesis test and the alternative hypothesis is $H_a : \mu_X > \mu_0$ or $H_a : \mu_X < \mu_0$ then this is called a **one-tailed test**. If, instead, the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ then this is called a **two-tailed test**.

EXAMPLE 6.3.8. Let's do one very straightforward example of a hypothesis test:

A cosmetics company fills its best-selling 8-ounce jars of facial cream by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ounces per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either underfill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value .22 ounce. A quality control engineer randomly selects 30 jars from the assembly line each day to check the amounts filled. One day, the sample mean is $\overline{X} = 8.2$ ounces. Let us see if there is sufficient evidence in this sample to indicate, at the 1% level of significance, that the machine should be recalibrated.

The population under study is all of the jars of facial cream on the day of the 8.2 ounce sample.

The variable of interest is the weight X of the jar in ounces.

The population parameter of interest is the population mean μ_X of X.

The two hypotheses then are

$$H_0 : \mu_X = 8.1 \quad \text{and}$$
$$H_a : \mu_X \neq 8.1 \, .$$

The sample mean is $\overline{X} = 8.2$, and the sample – which we must assume to be an SRS – is of size $n = 30$.

Using the case in Fact 6.3.6 where the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ and the sub-case where $\bar{X} > \mu_0$, we compute the p-value by

2 * (1-NORM.DIST(8.2, 8.1, .22/SQRT(30), 1))

on a spreadsheet, which yields p = .01278 .

Since p is not less than .01, we fail to reject H_0 at the $\alpha = .01$ level of significance.

The quality control engineer should therefore say to company management

> *-2.5mm"Today's sample, though off weight, was not statistically significant at the stringent level of significance of $\alpha = .01$ that we have chosen to use in these tests, that the jar-filling machine is in need of recalibration today ($p = .01278$)."*

Cautions

As we have seen before, the requirement that the sample we are using in our hypothesis test is a valid SRS is quite important. But it is also quite hard to get such a good sample, so this is often something that can be a real problem in practice, and something which we must assume is true with often very little real reason.

It should be apparent from the above Facts and Examples that most of the work in doing a hypothesis test, after careful initial set-up, comes in computing the p-value.

Be careful of the phrase *statistically significant*. It does not mean that the effect is large! There can be a very small effect, the $\overline{X}$ might be very close to μ_0 and yet we might reject the null hypothesis if the population standard deviation σ_X were sufficiently small, or even if the sample size n were large enough that $\sigma_X/\sqrt{n}$ became very small. Thus, oddly enough, a statistically significant result, one where the conclusion of the hypothesis test was statistically quite certain, might not be *significant* in the sense of mattering very much. With enough precision, we can be very sure of small effects.

Note that the meaning of the p-value is explained above in its definition as a conditional probability. So p ***does not*** compute the probability that the null hypothesis H_0 is true, or any such simple thing. In contrast, the Bayesian approach to probability, which we chose not to use in the book, in favor of the frequentist approach, does have a kind of hypothesis test which includes something like the direct probability that H_0 is true. But we did not follow the Bayesian approach here because in many other ways it is more confusing.

In particular, one consequence of the real meaning of the p-value as we use it in this book is that sometimes we will reject a true null hypothesis H_0 just out of bad luck. In fact, if p is just slightly less than .05, we would reject H_0 at the $\alpha = .05$ significance level even though, in slightly less than one case in 20 (meaning 1 SRS out of 20 chosen independently), we would do this rejection even though H_0 was true.

We have a name for this situation.

DEFINITION 6.3.9. When we reject a true null hypothesis H_0 this is called a **type I error**. Such an error is usually (but not always: it depends upon how the population, variable, parameter, and hypotheses were set up) a **false positive**, meaning that something exciting and new (or scary and dangerous) was found even though it is not really present in the population.

EXAMPLE 6.3.10. Let us look back at the cosmetic company with a jar-filling machine from Example 6.3.8. We don't know what the median of the SRS data was, but it wouldn't be surprising if the data were symmetric and therefore the median would be the same as the sample mean $\overline{X} = 8.2$. That means that there were at least 15 jars with 8.2 ounces of cream in them, even though the jars are all labelled "8oz." The company is giving way at least $.2 \times 15 = 3$ ounces of the very valuable cream – in fact, probably much more, since that was simply the overfilling in that one sample.

So our intrepid quality assurance engineer might well propose to management to increase the significance level α of the testing regime in the factory. It is true that with a larger α, it will be easier for simple randomness to result in type I errors, but unless the recalibration process takes a very long time (and so results in fewer jars being filled that day), the cost-benefit analysis probably leans towards fixing the machine slightly too often, rather than waiting until the evidence is extremely strong it must be done.