4. Hypothesis testing

Hypothesis testing is the other widely used form of inferential statistics. It is different from estimation because you start a hypothesis test with some idea of what the population is like and then test to see if the sample supports your idea. Though the mathematics of hypothesis testing is very much like the mathematics used in interval estimation, the inference being made is quite different. In estimation, you are answering the question "what is the population like?" While in hypothesis testing you are answering the question "is the population like this or not?"

A hypothesis is essentially an idea about the population that you think might be true, but which you cannot prove to be true. While you usually have good reasons to think it is true, and you often hope that it is true, you need to show that the sample data supports your idea. Hypothesis testing allows you to find out, in a formal manner, if the sample supports your idea about the population. Because the samples drawn from any population vary, you can never be positive of your finding, but by following generally accepted hypothesis testing procedures, you can limit the uncertainty of your results.

As you will learn in this chapter, you need to choose between two statements about the population. These two statements are the hypotheses. The first, known as the "null hypothesis", is basically "the population is like this". It states, in formal terms, that the population is no different than usual. The second, known as the "alternative hypothesis", is "the population is like something else". It states that the population is different than the usual, that something has happened to this population, and as a result it has a different mean, or different shape than the usual case. Between the two hypotheses, all possibilities must be covered. Remember that you are making an inference about a population from a sample. Keeping this inference in mind, you can informally translate the two hypotheses into "I am almost positive that the sample came from a population like this" and "I really doubt that the sample came from a population like this, so it probably came from a population that is like something else". Notice that you are never entirely sure, even after you have chosen the hypothesis which is best. Though the formal hypotheses are written as though you will choose with certainty between the one that is true and the one that is false, the informal translations of the hypotheses, with "almost positive" or "probably came", is a better reflection of what you actually find.

Hypothesis testing has many applications in business, though few managers are aware that that is what they are doing. As you will see, hypothesis testing, though disguised, is used in quality control, marketing, and other business applications. Many decisions are made by thinking as though a hypothesis is being tested, even though the manager is not aware of it. Learning the formal details of hypothesis testing will help you make better decisions and better understand the decisions made by others.

The next section will give an overview of the hypothesis testing method by following along with a young decision-maker as he uses hypothesis testing. The rest of the chapter will present some specific applications of hypothesis tests as examples of the general method.

The strategy of hypothesis testing

Usually, when you use hypothesis testing, you have an idea that the world is a little bit surprising, that it is not exactly as conventional wisdom says it is. Occasionally, when you use hypothesis testing, you are hoping to confirm that the world is not surprising, that it is like conventional wisdom predicts. Keep in mind that in either case you are asking "is the world different from the usual, is it surprising?" Because the world is usually not surprising and because in statistics you are never 100 per cent sure about what a sample tells you about a population, you cannot say that your sample implies that the world is surprising unless you are almost positive that it does. The dull, unsurprising, usual case not only wins if there is a tie, it gets a big lead at the start. You cannot say that the world is surprising, that the population is unusual, unless the evidence is very strong. This means that when you arrange your tests, you have to do it in a manner that makes it difficult for the unusual, surprising world to win support.

The first step in the basic method of hypothesis testing is to decide what value some measure of the population would take if the world was unsurprising. Second, decide what the sampling distribution of some sample statistic would look like if the population measure had that unsurprising value. Third, compute that statistic from your sample and see if it could easily have come from the sampling distribution of that statistic if the population was unsurprising. Fourth, decide if the population your sample came from is surprising because your sample statistic could not easily have come from the sampling distribution generated from the unsurprising population.

That all sounds complicated, but it is really pretty simple. You have a sample and the mean, or some other statistic, from that sample. With conventional wisdom, the null hypothesis that the world is dull and not surprising, tells you that your sample comes from a certain population. Combining the null hypothesis with what statisticians know tells you what sampling distribution your sample statistic comes from if the null hypothesis is true. If you are "almost positive" that the sample statistic came from that sampling distribution, the sample supports the null. If the sample statistic "probably came" from a sampling distribution generated by some other population, the sample supports the alternative hypothesis that the population is "like something else".

Imagine that Thad Stoykov works in the marketing department of Pedal Pushers, a company that makes clothes for bicycle riders. Pedal Pushers has just completed a big advertising campaign in various bicycle and outdoor magazines, and Thad wants to know if the campaign has raised the recognition of the Pedal Pushers brand so that more than 30 per cent of the potential customers recognize it. One way to do this would be to take a sample of prospective customers and see if at least 30 per cent of those in the sample recognize the Pedal Pushers brand. However, what if the sample is small and just barely 30 per cent of the sample recognizes Pedal Pushers? Because there is variance among samples, such a sample could easily have come from a population in which less than 30 percent recognize the brand—if the population actually had slightly less than 30 per cent recognition, the sampling distribution would include quite a few samples with sample proportions a little above 30 per cent, especially if the samples are small. In order to be comfortable that more than 30 per cent of the **population** recognizes Pedal Pushers, Thad will want to find that a bit more than 30 per cent of the **sample** does. How much more depends on the size of the sample, the variance within the sample, and how much chance he wants to take that he'll conclude that the campaign did not work when it actually did.

Let us follow the formal hypothesis testing strategy along with Thad. First, he must explicitly describe the population his sample could come from in two different cases. The first case is the unsurprising case, the case where there is no difference between the population his sample came from and most other populations. This is the case where the ad campaign did not really make a difference, and it generates the null hypothesis. The second case is the

surprising case when his sample comes from a population that is different from most others. This is where the ad campaign worked, and it generates the alternative hypothesis. The descriptions of these cases are written in a formal manner. The null hypothesis is usually called " H_o " The alternative hypothesis is called either " H_1 :" or " H_a :". For Thad and the Pedal Pushers marketing department, the null will be :

H_o : proportion of the population recognizing Pedal Pushers brand ≤ .30 and the alternative will be:

H_a : proportion of the population recognizing Pedal Pushers brand >.30.

Notice that Thad has stacked the deck against the campaign having worked by putting the value of the population proportion that means that the campaign was successful in the alternative hypothesis. Also notice that between H_o : and H_a : all possible values of the population proportion—>,=, and < .30 — have been covered.

Second, Thad must create a rule for deciding between the two hypotheses. He must decide what statistic to compute from his sample and what sampling distribution that statistic would come from if the null hypothesis, H_o :, is true. He also needs to divide the possible values of that statistic into "usual" and "unusual" ranges if the null is true. Thad's decision rule will be that if his sample statistic has a "usual" value, one that could easily occur if H_o : is true, then his sample could easily have come from a population like that described in H_o :. If his sample's statistic has a value that would be "unusual" if H_o : is true, then the sample probably comes from a population like that described in H_a :. Notice that the hypotheses and the inference are about the original population while the decision rule is about a sample statistic. The link between the population and the sample is the sampling distribution. Knowing the relative frequency of a sample statistic when the original population has a proportion with a known value is what allows Thad to decide what are "usual" and "unusual" values for the sample statistic.

The basic idea behind the decision rule is to decide, with the help of what statisticians know about sampling distributions, how far from the null hypothesis' value for the population the sample value can be before you are uncomfortable deciding that the sample comes from a population like that hypothesized in the null. Though the hypotheses are written in terms of descriptive statistics about the population—means, proportions, or even a distribution of values—the decision rule is usually written in terms of one of the standardized sampling distributions—the t, the normal z, or another of the statistics whose distributions are in the tables at the back of statistics books. It is the sampling distributions in these tables that are the link between the sample statistic and the population in the null hypothesis. If you learn to look at how the sample statistic is computed you will see that all of the different hypothesis tests are simply variations on a theme. If you insist on simply trying to memorize how each of the many different statistics is computed, you will not see that all of the hypothesis tests are conducted in a similar manner, and you will have to learn many different things rather than learn the variations of one thing.

Thad has taken enough statistics to know that the sampling distribution of sample proportions is normally distributed with a mean equal to the population proportion and a standard deviation that depends on the population proportion and the sample size. Because the distribution of sample proportions is normally distributed, he can look at the bottom line of a t-table and find out that only .05 of all samples will have a proportion more than 1.645 standard deviations above .30 if the null hypothesis is true. Thad decides that he is willing to take a 5 per cent chance that he will conclude that the campaign did not work when it actually did, and therefore decides that he will

conclude that the sample comes from a population with a proportion that has heard of Pedal Pushers that is greater than .30 if the sample's proportion is more than 1.645 standard deviations above .30. After doing a little arithmetic (which you'll learn how to do later in the chapter), Thad finds that his decision rule is to decide that the campaign was effective if the sample has a proportion which has heard of Pedal Pushers that is greater than .375. Otherwise the sample could too easily have come from a population with a proportion equal to or less than .30.

alpha	0.1	0.05	0.03	0.01
df infinity	1.28	1.65	1.96	2.33

Exhibit 5: The bottom line of a t-table, showing the normal distribution

The final step is to compute the sample statistic and apply the decision rule. If the sample statistic falls in the usual range, the data supports H_o :, and the world is probably unsurprising and the campaign did not make any difference. If the sample statistic is outside the usual range, the data supports H_a :, and the world is a little surprising, the campaign affected how many people have heard of Pedal Pushers. When Thad finally looks at the sample data, he finds that .39 of the sample had heard of Pedal Pushers. The ad campaign was successful!

A straight-forward example: testing for "goodness-of-fit"

There are many different types of hypothesis tests, including many that are used more often than the "goodness-of-fit" test. This test will be used to help introduce hypothesis testing because it gives a clear illustration of how the strategy of hypothesis testing is put to use, not because it is used frequently. Follow this example carefully, concentrating on matching the steps described in previous sections with the steps described in this section; the arithmetic is not that important right now.

We will go back to Ann Howard's problem with marketing "Easy Bounce" socks to volleyball teams. Remember that Ann works for Foothills Hosiery, and she is trying to market these sports socks to volleyball teams. She wants to send out some samples to convince volleyball players that wearing "Easy Bounce" socks will be more comfortable than wearing other socks. Her idea is to send out a package of socks to volleyball coaches in the area, so the players can try them out. She needs to include an assortment of sizes in those packages and is trying to find out what sizes to include. The Production Department knows what mix of sizes they currently produce, and Ann has collected a sample of 97 volleyball players' sock sizes from nearby teams. She needs to test to see if her sample supports the hypothesis that volleyball players have the same distribution of sock sizes as Foothills is currently producing—is the distribution of volleyball players' sock sizes a "good fit" to the distribution of sizes now being produced?

Ann's sample, a sample of the sock sizes worn by volleyball players, as a frequency distribution of sizes:

size	**frequency**
6	3
7	24
8	33
9	20
10	17

From the Production Department, Ann finds that the current relative frequency distribution of production of "Easy Bounce" socks is like this:

size	re. frequency
6	0.06
7	0.13
8	0.22
9	0.3
10	0.26
11	0.03

If the world in "unsurprising", volleyball players will wear the socks sized in the same proportions as other athletes, so Ann writes her hypotheses:

H_o : Volleyball players' sock sizes are distributed just like current production.

H_a : Volleyball players' sock sizes are distributed differently.

Ann's sample has n=97. By applying the relative frequencies in the current production mix, she can find out how many players would be "expected" to wear each size if her sample was perfectly representative of the distribution of sizes in current production. This would give her a description of what a sample from the population in the null hypothesis would be like. It would show what a sample that had a "very good fit" with the distribution of sizes in the population currently being produced would look like.

Statisticians know the sampling distribution of a statistic which compares the "expected" frequency of a sample with the actual, or "observed" frequency. For a sample with c different classes (the sizes here), this statistic is distributed like χ^2 with c-1 df. The χ^2 is computed by the formula:

$$sample\ \chi^2 = \sum \frac{(O-E)^2}{E}$$

where:

O = observed frequency in the sample in this class

E = expected frequency in the sample in this class.

The expected frequency, E, is found by multiplying the relative frequency of this class in the H_o:hypothesized population by the sample size. This gives you the number in that class in the sample if the relative frequency distribution across the classes in the sample exactly matches the distribution in the population.

Notice that χ^2 is always ≥ 0 and equals 0 only if the observed is equal to the expected in each class. Look at the equation and make sure that you see that a larger value of goes with samples with large differences between the observed and expected frequencies.

Ann now needs to come up with a rule to decide if the data supports H_o : or H_a :. She looks at the tableand sees that for 5 df (there are 6 classes—there is an expected frequency for size 11 socks), only .05 of samples drawn from a given population will have a χ^2 > 11.07 and only .10 will have a χ^2 > 9.24. She decides that it

would not be all that surprising if volleyball players had a different distribution of sock sizes than the athletes who are currently buying "Easy Bounce", since all of the volleyball players are women and many of the current customers are men. As a result, she uses the smaller .10 value of 9.24 for her decision rule. Now she must compute her sample χ^2 . Ann starts by finding the expected frequency of size 6 socks by multiplying the relative frequency of size 6 in the population being produced by 97, the sample size. She gets E = .06*97=5.82. She then finds O-E = 3-5.82 = -2.82, squares that and divides by 5.82, eventually getting 1.37. She then realizes that she will have to do the same computation for the other five sizes, and quickly decides that a spreadsheet will make this much easier. Her spreadsheet looks like this:

sock size	**frequency in sample**	**population relative frequency**	**expected frequency = 97*C**	**(O-E)^2/E**
6	3	0.06	5.82	1.3663918
7	24	0.13	12.61	10.288033
8	33	0.22	21.34	6.3709278
9	20	0.3	29.1	2.8457045
10	17	0.26	25.22	2.6791594
11	0	0.03	2.91	2.91
	97			X^2 = 26.460217

Exhibit 6: Ann's Excel sheet

Ann performs her third step, computing her sample statistic, using the spreadsheet. As you can see, her sample χ^2 = 26.46, which is well into the "unusual" range which starts at 9.24 according to her decision rule. Ann has found that her sample data supports the hypothesis that the distribution of sock sizes of volleyball players is different from the distribution of sock sizes that are currently being manufactured. If Ann's employer, Foothill Hosiery, is going to market "Easy Bounce" socks to volleyball players, they are going to have to send out packages of samples that contain a different mix of sizes than they are currently making. If "Easy Bounce" are successfully marketed to volleyball players, the mix of sizes manufactured will have to be altered.

Now, review what Ann has done to test to see if the data in her sample supports the hypothesis that the world is "unsurprising" and that volleyball players have the same distribution of sock sizes as Foothill Hosiery is currently producing for other athletes. The essence of Ann's test was to see if her sample χ^2 could easily have come from the sampling distribution of χ^2 's generated by taking samples from the population of socks currently being produced. Since her sample χ^2 would be way out in the tail of that sampling distribution, she judged that her sample data supported the other hypothesis, that there is a difference between volleyball players and the athletes who are currently buying "Easy Bounce" socks.

Formally, Ann first wrote null and alternative hypotheses, describing the population her sample comes from in two different cases. The first case is the null hypothesis; this occurs if volleyball players wear socks of the same sizes in the same proportions as Foothill is currently producing. The second case is the alternative hypothesis; this occurs if volleyball players wear different sizes. After she wrote her hypotheses, she found that there was a sampling

distribution that statisticians knew about that would help her choose between them. This is the χ^2 distribution. Looking at the formula for computing χ^2 and consulting the tables, Ann decided that a sample χ^2 value greater than 9.24 would be unusual if her null hypothesis was true. Finally, she computed her sample statistic, and found that her χ^2 , at 26.46, was well above her cut-off value. Ann had found that the data in her sample supported the alternative, H_a :, that the distribution of volleyball players' sock sizes is different from the distribution that Foothill is currently manufacturing. Acting on this finding, Ann will send a different mix of sizes in the sample packages she sends volleyball coaches.

Testing population proportions

As you learned in the chapter “Making estimates”, sample proportions can be used to compute a statistic that has a known sampling distribution. Reviewing, the z-statistic is:

$$z = \frac{p - \pi}{\sqrt{\frac{(\pi)(1-\pi)}{n}}}$$

where: p = the proportion of the sample with a certain characteristic

π = the proportion of the population with that characteristic.

These sample z-statistics are distributed normally, so that by using the bottom line of the t table, you can find what portion of all samples from a population with a given population proportion, π, have z-statistics within different ranges. If you look at the table, you can see that .95 of all samples from any population have a z-statistics between ±1.96, for instance.

If you have a sample that you think is from a population containing a certain proportion, π, of members with some characteristic, you can test to see if the data in your sample supports what you think. The basic strategy is the same as that explained earlier in this chapter and followed in the "goodness-of-fit" example: (a) write two hypotheses, (b) find a sample statistic and sampling distribution that will let you develop a decision rule for choosing between the two hypotheses, and (c) compute your sample statistic and choose the hypothesis supported by the data.

Foothill Hosiery recently received an order for children's socks decorated with embroidered patches of cartoon characters. Foothill did not have the right machinery to sew on the embroidered patches and contracted out the sewing. While the order was filled and Foothill made a profit on it, the sewing contractor's price seemed high, and Foothill had to keep pressure on the contractor to deliver the socks by the date agreed upon. Foothill's CEO, John McGrath has explored buying the machinery necessary to allow Foothill to sew patches on socks themselves. He has discovered that if more than a quarter of the children's socks they make are ordered with patches, the machinery will be a sound investment. Mr McGrath asks Kevin Schmidt to find out if more than 25 per cent of children's socks are being sold with patches.

Kevin calls the major trade organizations for the hosiery, embroidery, and children's clothes industries, and no one can answer his question. Kevin decides it must be time to take a sample and to test to see if more than 25 per cent of children's socks are decorated with patches. He calls the sales manager at Foothill and she agrees to ask her salespeople to look at store displays of children's socks, counting how many pairs are displayed and how many of

those are decorated with patches. Two weeks later, Kevin gets a memo from the sales manager telling him that of the 2,483 pairs of children's socks on display at stores where the salespeople counted, 716 pairs had embroidered patches.

Kevin writes his hypotheses, remembering that Foothill will be making a decision about spending a fair amount of money based on what he finds. To be more certain that he is right if he recommends that the money be spent, Kevin writes his hypotheses so that the "unusual" world would be the one where more than 25 per cent of children's socks are decorated:

H_o: $\pi_{\text{decorated socks}} \leq .25$

H_a:$\pi_{\text{decorated socks}} > .25$

When writing his hypotheses, Kevin knows that if his sample has a proportion of decorated socks well below .25, he will want to recommend against buying the machinery. He only wants to say the data supports the alternative if the sample proportion is well above .25. To include the low values in the null hypothesis and only the high values in the alternative, he uses a "one-tail" test, judging that the data supports the alternative only if his z-score is in the upper tail. He will conclude that the machinery should be bought only if his z-statistic is too large to have easily have come from the sampling distribution drawn from a population with a proportion of .25. Kevin will accept H_a : only if his z is large and positive.

Checking the bottom line of the t-table, Kevin sees that .95 of all z-scores are less than 1.645. His rule is therefore to conclude that his sample data supports the null hypothesis that 25 per cent or less of children's socks are decorated if his sample z is less than 1.645. If his sample z is greater than 1.645, he will conclude that more than 25 per cent of children's socks are decorated and that Foothill Hosiery should invest in the machinery needed to sew embroidered patches on socks.

Using the data the salespeople collected, Kevin finds the proportion of the sample that is decorated:

$$p = \frac{716}{2483} = .288$$

Using this value, he computes his sample z-statistic:

$$z = \frac{p - \pi}{\sqrt{\frac{(\pi)(1-\pi)}{n}}}$$

$$= \frac{.288-.25}{\sqrt{\frac{(.25)-(1-.25)}{2483}}} =$$

$$= \frac{.0380}{.0087} = 4.368.$$

Because his sample z-score is larger than 1.645, it is unlikely that his sample z came from the sampling distribution of z's drawn from a population where $\pi \leq .25$, so it is unlikely that his sample comes from a population with $\pi \leq .25$. Kevin can tell John McGrath that the sample the sales people collected supports the conclusion that

more than 25 per cent of children's socks are decorated with embroidered patches. John can feel comfortable making the decision to buy the embroidery and sewing machinery.

Summary

This chapter has been an introduction to hypothesis testing. You should be able to see the relationship between the mathematics and strategies of hypothesis testing and the mathematics and strategies of interval estimation. When making an interval estimate, you construct an interval around your sample statistic based on a known sampling distribution. When testing a hypothesis, you construct an interval around a hypothesized population parameter, using a known sampling distribution to determine the width of that interval. You then see if your sample statistic falls within that interval to decide if your sample probably came from a population with that hypothesized population parameter.

Hypothesis testing is a very widely used statistical technique. It forces you to think ahead about what you might find. By forcing you to think ahead, it often helps with decision-making by forcing you to think about what goes into your decision. All of statistics requires clear thinking, and clear thinking generally makes better decisions. Hypothesis testing requires very clear thinking and often leads to better decision-making.