

Chapter 7

Hypothesis Testing I: An Introduction

Hypothesis testing is the heart of science. In its broadest sense, it is the procedure used to evaluate our competing ideas about how the aspects of our world are structured. Hypothesis testing in statistics is merely one formulation of the more general procedure operative in science. In archaeology, the ideas subject to evaluation are our ideas about how the archaeological record is structured.

Statistical procedures have been developed that allow us to address many different kinds of hypotheses, but our discussion in this chapter will be necessarily limited to the most general, and most useful, of these applications. It is common for archaeologists, as well as scientists in general, to be interested in statistical procedures for hypothesis testing that involve comparisons of one kind or another. For example, we are frequently interested in comparing a variate to a population mean, a sample mean to a population mean, or two sample means to each other. Often the relationships we wish to evaluate are conceptually difficult to express in common language, yet it is crucial to state the hypothesis of interest unambiguously so that others can understand our research and to prevent ourselves from being confused and making preventable mistakes. A uniform method of statistical notation has consequently developed to help simplify the statement of hypotheses. Let us take examples of common comparisons to illustrate the appropriate symbolism of hypothesis testing. For organizational purposes, this will be accomplished over the course of the next three chapters: Chapters 7, 8, and 9.

Hypothesis of Interest

Statistical hypothesis testing is a mathematical means of calculating the *probability* that some relationship we posit is correct. Often, we evaluate specific relationships we think are true. These relationships can be referred to as our hypotheses of interest.

Our discussion of the normal distribution in the previous chapter has already laid the foundation for understanding hypothesis testing where we are concerned with establishing the comparability of a variate to a population mean. The z-score provides a means of determining the probability of a variate within a normal distribution. By determining the probabilities of variates we wish to consider, we can decide if it is likely or unlikely that each variate is part of a distribution.

For example, the specific function of ceramic vessels has been heavily researched throughout the world. We might hypothetically have two measurements of the maximum wall thickness from two ceramic vessels we believe were used strictly for water storage. We might argue that the properties including maximum wall thickness required for water storage are different from those properties required for cooking. As a result, we might wish to compare each of our two variates to a population of known cooking vessels to determine if either or both of them have wall thicknesses characteristic of cooking vessels.

Our hypothesis of interest in this case could be stated as, “Are the maximum wall thickness of our two vessels (Y_1 and Y_2) different from the average maximum wall thickness for the cooking vessels (μ_1)?” Using the z-score, we could determine that our

hypothetical variates Y_1 and Y_2 are 0.73σ and 3.5σ from μ_1 , respectively. We know that an individual variate does not differ meaningfully from a population mean if it is less than 1σ away from μ . In fact, variates such as Y_1 that are within $\pm 1\sigma$ are very likely. But any Y_i more than $\pm 3\sigma$ away from μ is unlikely (only 0.26% of the variates in a normal distribution are more than $\pm 3\sigma$ from the mean), and we therefore recognize that it is extremely unlikely that Y_2 was drawn from population μ_1 . We may intuitively conclude that while Y_1 is possibly from the population with μ_1 , Y_2 is probably not.

Formal Hypothesis Testing and the Null Hypothesis

While we might be satisfied with this type of intuitive interpretation when the probabilities associated with the variates are as extreme as the previous example, we will likely find it dissatisfying in most circumstances. This intuitive approach leaves us with many questions such as: “What is the value demarcating variates that probably did not come from a population from those that might have?” and “What formal statistical relationship are we really addressing?” To answer these sorts of questions, statisticians have developed a system of formal hypothesis testing that is designed to ensure the hypotheses under consideration are clearly defined, the terms of rejection of hypotheses are understood, and the outcome of our hypothesis testing is unambiguously presented.

The hypothesis tested above can be more formally stated as $H_0 : Y_i = \mu$. H_0 is by convention the symbol for the *null hypothesis*; the hypothesis of no difference. As you know, Y_i represents our variates and μ represents the population mean. In plain English,

the null hypothesis can be summarized as the statistical proposition that there is no difference between a variate Y_i and the population parameter μ .

With respect to hypothesis testing, the presentation of the null hypothesis is the basic statement formalizing our intent to compare. When comparing, two possible outcomes come to mind. The first is, of course H_0 : no difference between the two things that are being compared. The second possible outcome is that there is indeed a meaningful difference. This outcome is symbolized by the alternative hypothesis, H_a . After all, if the relationship specified in the null hypothesis is false, some other relationship must be true. Therefore, once a null hypothesis is defined, one or more alternate hypothesis must also be defined. In this example, the alternate hypothesis can simply be the inverse of the null hypothesis. It can be written as $H_a : Y_i \neq \mu$ (e.g., the maximum wall thickness of our two ceramic vessels is significantly different than the average maximum wall thickness of the cooking vessels). So, the general terminology is as follows:

$H_0 : Y_i = \mu$ constitutes the null hypothesis specifying no meaningful difference, and $H_a : Y_i \neq \mu$ constitutes the alternative hypothesis where there is indeed a meaningful difference.

An astute reader will notice that the null hypothesis (there is no difference between the maximum wall thicknesses) states exactly the opposite of our hypothesis of interest (that there is a difference) and that the alternate hypothesis agrees with the hypothesis of interest. That is true. Null hypotheses are statistical propositions that specifically state the statistical relationship we will evaluate, not necessarily the specific relationship that interests us.

In our example, we evaluate the hypotheses with a z-score. As you learned in the previous chapter, the z-score is designed to determine the probability of the occurrence of a set of observations (variates) as defined by an area under the normal distribution. To test our hypothesis of interest, we must determine if the variate is meaningfully different than the mean. *By convention, the null hypothesis is that there is no difference.* If we find that the data support our belief that the maximum wall thickness of one of our ceramic vessels is meaningfully different than the average cooking vessel maximum wall thickness, we will *reject* the null hypothesis and conclude that the alternate hypothesis is more likely true. If we find that the data suggest the null hypothesis cannot be rejected, and no difference is present in the maximum wall thickness, we must conclude that the alternate hypothesis and our hypothesis of interest are probably false.

We used the z-score in the example, but the relationship between the null and alternate hypotheses characterizes all hypothesis testing. The actual statistical test used to evaluate competing hypotheses is dictated by the data and the nature of the hypotheses themselves, but the philosophy of hypothesis testing remains the same.

Let us now work through a practical example. In Chapter 5 we considered the likelihood of specimen number HC-215 with $Y_i=6.4$ being drawn from the population of *Thomomys bottae* alveolar lengths with a $\mu = 5.7$ and $\sigma = .48$.

Formally, we were testing the null hypothesis:

$$H_0 : Y_i = \mu$$

or

$$H_0 : 6.4 = 5.7$$

and implicitly testing the alternative hypothesis:

$$H_a : Y_i \neq \mu$$

or

$$H_a : 6.4 \neq 5.7 .$$

However, we have not yet drawn a formal conclusion as to whether the variate is or is not from a distribution of *Thomomys bottae* alveolar lengths. We merely concluded that it is relatively unlikely to observe a *Thomomys bottae* alveolar length of 6.4mm. But which hypothesis is more likely correct, the null or the alternate hypothesis? The rules of hypothesis testing state that it is important to specify the probability at which you, the researcher, will reject the null hypothesis of no difference. Put another way, you must choose at what point you find it so unlikely that the null hypothesis is correct that you are willing to accept the alternate hypothesis instead. The boundary is completely arbitrary and is based on your understanding of the data, and the consequences of making a wrong decision. We will discuss this issue in detail below.

For the sake of convenience, let us specify that we will consider H_0 rejected if the probability of $z < .05$. That is, we will reject H_0 if the probability for obtaining a value of z or greater is smaller than .05. You will recall that the value of .05 corresponds roughly to the probability of a variate being greater than $2\sigma \pm$ the mean (the exact value is 1.96). What is being specified here is how close is close enough for our null hypothesis to be satisfied. Reconsider our null hypothesis: $H_0 : 6.4 = 5.7$. We know of course that $6.4 \neq 5.7$. What we are really assessing is whether or not 6.4 is *close enough* to 5.7 to likely be from that population the mean of 5.7 describes. The specification of the value at which the null hypothesis will be rejected defines what is *close enough* for our purposes. After specifying this value, the hypothesis testing continues by calculating z :

$$z = \frac{Y_i - \mu}{\sigma}$$

$$z = \frac{6.40 - 5.70}{.48}$$

$$z = 1.46 .$$

Once z was calculated in the previous chapter, we went to Appendix A to find the probability of a variate falling between μ and z . That probability is .4279, which indicates a common event. Yet, to test the hypothesis formally, we are really interested in the probability of an event as large as z or greater. That probability is equal to $1 - (.5 + .4279) = .0721$. This is a relatively rare event (about 7 times out of 100), but is not smaller than our level of rejection (probability of $z < .05$). Therefore, we cannot reject H_0 , and must conclude that $Y_i = 6.4$ is not significantly different than $\mu = 5.7$.

While this procedure is useful, it is not always necessary to calculate the specific probabilities for a hypothesis test. Oftentimes, knowing the z -score is sufficient. The hypothesis evaluation process described above can be streamlined by using the z -score 1.96 in a way we have not previously conceptualized; we can use it as a *critical value*. Critical values are the values that mark the cut-off line that describes the point where we either reject or fail to reject the null hypothesis $H_0 : Y_i = \mu$. From our previous discussions, we know that z -scores greater than 1.96 represent variates that are outside of the 95% probability space and z -scores less than 1.96 represent variates that are common in our distribution. In this example $z = 1.46$ which is less than 1.96. We therefore can conclude that the variate 6.4 is a common variate in this distribution and is not

significantly different from μ . By simply looking at the z-score values, we can conclude that we cannot reject $H_0 : Y_i = \mu$.

You will notice, of course, that we used the clumsy phrase *cannot reject the null hypothesis* instead of the more elegant phrase *accept the null hypothesis* in the sentence above. This is a result of the nature of hypothesis testing. If a value is close (say within 1σ) to the mean, it is a value with a high probability within that distribution μ_1 . However, can we conclude that it is by all certainty a member of the population represented by μ_1 ? No. It could be a member of population μ_2 , μ_3 , or any other population, but could still be close to μ_1 .

In other words, simply because a variate is close to a particular mean does not necessarily indicate that the variate is drawn from the population represented by the mean. However, if the variate is far from the mean (say 5σ from the mean), it is extremely unlikely that the variate is from the population μ_1 . We are more confident, then, in *rejecting* null hypotheses than *accepting* them.

This point can be illustrated using the Alyawara settlement size data presented in Chapter 4. Based on our sample of settlement sizes, we determined that the average Alyawara settlement has a mean of 23.375 individuals with a standard deviation of 11.795. If we found a village with 23 individuals, can we conclude the village is an Alyawara settlement? Of course not. The village could be anywhere in the world with residents of any number of ethnic groups. While the number of residents in our hypothetical village is not significantly different from the average number of residents in an Alyawara camp, we cannot conclude that it is an Alyawara settlement. That is, we cannot conclude that it wasn't drawn from another population. If, however, we found a

village with 230 residents, is it likely that the settlement is *not* an Alyawara camp?

Assuming our sample of Alyawara settlements is representative, yes. Based on our data, it is extremely unlikely that an Alyawara settlement has 230 people, suggesting our hypothetical settlement is not Alyawaran. But a settlement with 23 people may or may not be.

Using hypothesis testing, then, we can often conclude a null hypothesis is false, but we cannot conclude it is true. When we do fail to reject the null hypothesis, we can merely conclude that the null hypothesis *may* be true, then provide logical arguments justifying its validity.

The discussion above should have driven one point home; hypothesis testing in a statistical framework is not a magical gateway to the truth. Statistics merely provide a framework to assess the probability that a relationship we are interested in is or is not present. You might even say statistics merely tell which of our educated guesses are plausible and which are not. Even then, we may be wrong. This leads us to our next topic of discussion: errors in hypothesis testing.

Errors in Hypothesis Testing

Because hypothesis testing is based on the assessment of probabilities, we are never certain that any specific conclusion derived from our statistical analyses is necessarily correct. There are a number of errors and correct decisions that can be made during hypothesis testing. Figure 7.1 illustrates these possibilities.

Figure 7.1. Possible outcomes of hypothesis testing.

Null Hypothesis is:	Results of Statistical Analysis cause Null Hypothesis to be:	
	Accepted	Rejected
True	Correct Decision	Type I Error (alpha)
False	Type II Error (beta)	Correct Decision

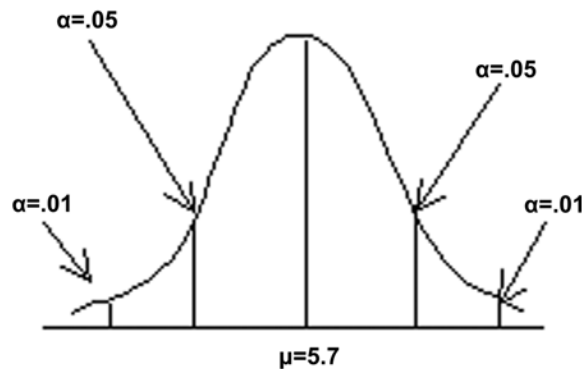
Two correct decisions are possible. We may fail to reject the null hypothesis when it is in fact true, and reject the null hypothesis when it is in fact false. Likewise, two kinds of errors may be made. We may reject the null hypothesis when it is indeed true (thereby making a Type I or alpha [α] error), or we may accept the null hypothesis when it is indeed false (a Type II or beta [β] error).

Knowledge of the properties of statistical distributions allow us to specify α , or the probability of rejecting the null hypothesis when it is in fact true, before we begin our analysis. This is our probability of being wrong in a very specific way; rejecting the null hypothesis when it is in reality true. A Type I error may be expressed as a probability, or as a percentage. For example, setting α at .05 allows 5% of our samples to lead us to a type I or α error. Setting α at 0.10 necessitates that we will commit a type I error 10% of the time. When expressed as a percentage, the alpha value is known as a *significance level*.

The importance of the critical value is illustrated by Figure 7.2, which presents a hypothetical distribution of the *Thomomys bottae* alveolar lengths previously discussed. It is composed of a distribution of individual alveolar lengths normally distributed around

μ . In this figure, the areas of rejection for $\alpha = 0.05$ and $\alpha = 0.01$ are illustrated. When evaluating the null hypothesis $H_0 : Y_i = \mu$ using an α of 0.05, we will expect to incorrectly reject the null hypothesis (i.e. conclude the variate is not representative of a *Thomomys bottae* when in fact it is) 5% of the time. Likewise, when evaluating the null hypothesis with an $\alpha = 0.01$ we would expect to incorrectly reject the null hypothesis 1% of the time. As illustrated by this figure, the larger α is, the more often we will expect to commit a Type I error.

Figure 7.2. Illustration of the areas on a normal distribution associated with alpha levels of .05 and .01.



Many researchers use, by convention, α of .05. Unfortunately, there is neither any particular reason for choosing this value as opposed to any other, nor a rule of thumb dictating which α is the most useful value for any given situation. The investigator simply needs to decide how frequently he or she is comfortable with rejecting the null hypothesis when it is indeed true, fully considering the consequence of this decision.

The natural inclination from the previous discussion of Type I errors is to make α as small as possible. After all, if an α of 0.05 suggests we will incorrectly reject a null hypothesis when it is true 5% of the time (i.e., 5 times out of a hundred), and an α of 0.001 suggests we will reject a null hypothesis when it is true only 0.1% of the time (i.e., 1 time out of a thousand), why not minimize our error and make the α level 0.00001 or even 0.0000000001?

The answer is that to do so would make it more likely that we will incorrectly fail to reject a null hypothesis when it is in fact false and another hypothesis is true. In other words, by reducing α we are making it more likely that we will not identify variates that actually were drawn from another population. After all the only way to be sure of never committing an α error is to never reject the null hypothesis. Yet if we do this, we are guaranteed to incorrectly conclude that the variates from all other possible distributions are possibly from μ_1 , that is, we are guaranteed to commit a β error if in fact the variate is from a different distribution. Therefore, when setting our α level, we must consider how willing we are to commit a β error.

Oddly enough, archaeologists seldom determine β . It seems to us that this is probably the result of three factors. First, archaeologists seldom compare hypotheses that are mutually exclusive. As a result, the potential for β errors is extreme and potentially damaging to any number of otherwise useful arguments. Archaeologists therefore find it convenient to ignore the possibility of β errors instead of trying to justify their conclusions further. Second, the probability of committing a β error must actually be calculated instead of being arbitrarily set as is the case with α . Calculating β may be a labor-intensive process for individuals with little statistical or mathematical background.

Finally, there has been a common misperception that β errors cannot be calculated at all. Regardless of the reasons, the failure of archaeologists to consider β is unfortunate, as it is only with the determination of β that we are able to determine the *power* of a hypothesis test. We will return to this issue and demonstrate how the probability of committing a β error can be calculated later in Chapter 9. First, though, we need to understand the role of *confidence limits* and their *critical regions* in evaluating hypotheses. That is the subject of Chapter 8.