Introduction

Prerequisites
Binomial Distribution

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that Martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from the binomial distribution and the binomial distribution calculator shows it to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Binomial Calculator

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of normal weight. The chart a particular physician viewed was determined randomly. Thirty three physicians viewed charts of average-weight patients and

38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for average-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. *Random assignment* of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. Therefore it is possible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the differences are due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in another section, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.016 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing.) It is not the probability that a state of world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim,

the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial calculator, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower, perhaps lower that 0.0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird can tell whether a number is divisible by 7). In statistics, it is conventional to refer to possible states of the world as *hypotheses* since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called *Bayesian statistics* provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and therefore are difficult to apply in some contexts.

The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the *null hypothesis*. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average}$$

or as

$$\mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high-school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a *parameter* is 0, there are occasions on which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the *alternative hypothesis*) is accepted. The alternative hypothesis is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{obese} = \mu_{average}$$

is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average}$$

$$\mu_{obese} > \mu_{average}.$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null

hypothesis that two populations means are equal does not justify a conclusion about which population mean is larger.