

THE UNIVERSITY OF TEXAS AT AUSTIN

SCHOOL OF INFORMATION

MATHEMATICAL NOTES FOR LIS 397.1

INTRODUCTION TO RESEARCH IN

LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys
Last revised: 2003 Jan 15

STATISTICAL HYPOTHESES

What Is a Statistical Hypothesis?

In LIS 397.1 we use the following as our general definition of a hypothesis: "A hypothesis is a statement of a relationship between two or more variables." A statistical hypothesis is simply a particular kind of hypothesis.

What is a statistical hypothesis? A satisfactory definition for our purposes in LIS 397.1 is this: A statistical hypothesis is either (1) a statement about the value of a population parameter (e.g., mean, median, mode, variance, standard deviation, proportion, total), or (2) a statement about the kind of probability distribution that a certain variable obeys.

Here are some examples of statistical hypotheses:

- a. The mean age of all GSLIS students is 23.4 years.
- b. The proportion of GSLIS students who are women is 76 percent.
- c. The variable H_m , representing heights of GSLIS male students, is approximately Gaussianly (i.e., normally) distributed.
- d. The modal grade in LIS 397.1 this semester will be B.
- e. The proportion of books in the Podunk Public Library whose heights exceed 30 cm is less than or equal to 0.13.

A related (and more technical) definition is: A statistical hypothesis that specifies a single value for a population parameter is called a simple hypothesis; every statistical hypothesis that is not simple is called composite. Examples a, b, and d above are simple; c and e are composite.

It is an interesting fact of language that in any situation involving a possible relationship, there are always at least two ways of stating a hypothesis about the relationship: affirmation of the relationship, or denial of it. That is, we can always choose between saying "A stands in relationship R to B" or "A fails to stand in relationship R to B." In terms of typical statistical hypotheses, such a pair of examples might be: "The mean value of variable X equals the mean value of variable Y," and "The means of variables X and Y are different." Note also that further variants are possible, such as "The mean value of variable X is not different from the mean value of variable Y," and "The means of variables X and Y are not equal." It is usually easier to understand hypotheses that are stated positively than those that are stated negatively.

Statistical hypotheses are statements about real relationships; and like all hypotheses, statistical hypotheses may match the reality, or they may fail to do so. Statistical hypotheses have the special characteristic that one ordinarily attempts to *test* them (i.e., to reach a decision about whether or not one believes the statement is correct, in the sense of corresponding to the reality) by observing facts relevant to the hypothesis in a sample. This procedure, of course, introduces the difficulty that the sample may or may not represent well the population from which it was drawn.

Statistical Hypotheses and States of Nature

Depending on whether one's sample is representative of the population or not, one's decision about whether to believe that a certain statistical hypothesis H is true or false may accord with the actual relationship, or it may fail to. We can set up the following schema to represent the possibilities of matching and mismatching between the decision as to what is believed to be the truth or falsity of a hypothesis H and the actual *reality*, the "state of nature." Here the phrase "accept H " represents a decision to believe that the hypothesis H is true, i.e., that it corresponds to, or matches, reality; and the phrase "reject H " means the opposite.

DECISION	STATE OF NATURE	
	H is true	H is false
Accept H :	satisfactory	error
Reject H :	error	satisfactory

Clearly, two of the possible combinations of decision and state of nature are ones in which the decision and the reality match, and two are not. The mismatches between decision and state of nature constitute errors. In general, an error will be made when the sample misrepresents--is atypical of--the population to such an extent that the investigator concludes that the population is rather different from what it is really like.

In real life, it frequently happens that making one of the two possible errors would lead to worse consequences than would making the other error. In tests of statistical hypotheses it is *conventional* to focus attention on the more serious of the possible errors, and to arrange things so that the more serious error is equivalent to "rejecting the hypothesis when it is true."

In a broad sense, one can accomplish this by studying the matters of interest, identifying the more serious of the two errors, and then wording or re-wording the hypothesis in such a way that the more serious error occurs when the decision is to believe that the hypothesis is false even though in reality the hypothesis, as stated, is true. So stated, the hypothesis is called the "null hypothesis." The phrase "null hypothesis" should be taken as an abbreviation for "the hypothesis being tested" (given the arrangement just described as to hypothesis and more serious error), and it should be noted that null hypotheses are not necessarily stated in negative terms.

The Two Types of Errors Possible in Making Decisions about Statistical Hypotheses

The null hypothesis is usually denoted by H_0 (pronounced "H naught"); the more serious error is usually called the "Type I error"; the less serious, the "Type II error." With these conventions, the above schema becomes:

DECISION	STATE OF NATURE	
	H_0 is true	H_0 is false
Accept H_0 :	satisfactory	Type II error
Reject H_0 :	Type I error	satisfactory

It is important to bear in mind that you are merely making a decision concerning what you *believe* about the truth or falsity of the hypothesis; you are not really ascertaining whether the hypothesis is true or false. In other words, if you decide to reject H_0 , that means "I have *decided to believe* that H_0 is false"; it does not necessarily mean that H_0 is actually false. Similarly, if your decision is to accept (or, more precisely, not to reject) H_0 , that means "I have *decided to believe* that H_0 is true"; it does not necessarily mean that H_0 is actually true.

Given the foregoing cautions, it can be helpful to think of the Type I error as "rejecting the null hypothesis when it is true" and the Type II error as "failing to recognize that the null hypothesis is false when it is false." Another way of putting it is that the Type I error amounts to "disbelieving the truth"; the Type II error, to "believing an untruth."

Some examples (cf. Endnote 1) of different types of errors may help to clarify these ideas.

Example 1. Suppose you are giving a party where gambling with dice is occurring, and suppose that you have somehow come to suspect that one of your guests is using loaded dice. What should you do?

We start by noting that you can take either of two actions: do nothing, or get rid of the guest. Assuming that you will want to behave rationally, you will want to make some kind of test of the possibility that the guest is cheating. Such a test could involve observations of his play and consideration of whether his play is what you would expect with fair dice or with loaded dice.

Without going into the details of how to set up such a test, we can note that two errors are possible: (1) doing nothing, i.e., letting the gambling continue, when the dice are, in fact, loaded; and (2) getting rid of the guest when he is actually playing fairly. You will have to decide which of these errors is the more serious, but they are certainly different and will lead to different consequences. (For practice, at this point you should make a decision as to which of the two errors you think is the more important to avoid and then write out the statement of H_0 according to your decision.)

Example 2. A teacher gives an exam consisting of 100 true-false questions. Knowing that students could average 50 correct answers by simply guessing in a random fashion, the teacher realizes that she must set a higher threshold than 50 as the minimum grade for passing the test. Moreover, the teacher, being not only a fair-minded person but also knowledgeable about statistics, recognizes that in reality any test only samples the population of facts, concepts, etc., that each student has learned about the course. This implies, *inter alia*, that a good student may do poorly on a test because the test questions happened, by chance, to over-emphasize the small portion of the course's subject matter that the student had not yet had time to study; clearly, the converse could also occur.

The teacher thus recognizes that no matter what threshold grade she decides on, two errors are possible: (1) that a poor student will happen to guess enough right answers to pass; and (2) that a good student will happen not to know the answers to enough test questions and hence will fail. Which error is the more serious? Could students and teachers differ about which error is the more important to avoid?

Example 3. The process of manufacturing certain drugs is quite complex. Seemingly unimportant departures from the standard procedure may introduce extraneous substances that are highly toxic. Occasionally the toxicity of such impurities is so high that minute quantities, undetectable by ordinary chemical analysis, can be dangerous to persons treated with the drug. As a result, prior to a freshly manufactured batch's being released for sale, it is tested for toxicity by biological methods. Small doses of the drug are injected into a number of experimental animals, such as mice, and the effect of the injections is recorded. If the drug is toxic, then all or most of the animals die; if not, most or all of the animals live.

We can assume that the number K of deaths among the N animals injected with a given dose of the drug is a variable that depends on the toxicity of the drug. (In actual practice, usually several different groups of N animals each are given the drug, each group of animals receiving a different level of dosage of the drug.) The test may lead to either of two possible courses of action: (S) decide that the batch of drug is safe, and put it on the market; (T) decide that the batch is toxic and destroy it. The choice of action S or T will depend on the value observed for the variable K .

But the factors determining K are complex. If the batch of drug is extremely toxic, all of the injected animals may die, i.e., we may have $K = N$; in such a case, we will obviously decide on action T. If the batch is safe, most of the injected animals will live, but some of them might die from other causes, such as accidents while being handled, diseases, or old age. If the batch is moderately toxic, we can expect some injected animals to die and others to survive.

The rule for choosing between actions S and T will take the following form: If K exceeds some threshold level L (where $0 \leq L \leq N$), then action T is to be taken; otherwise, action S is to be taken.

You can see that the problem is to set L low enough so that a dangerous batch of the drug will be recognized as such and destroyed (even if a few exceptionally tough animals survive), yet high enough so that safe batches will be passed and put on sale (even if a few frail, injured, sickly, or aged animals die after being injected). You will recognize further that it could occasionally happen that the batch was toxic but was administered to a group of animals among which a large number were exceptionally tough, and that it could also occasionally happen that the batch was safe but that the group of injected animals would include a large number of unusually frail, injured, sickly, or aged animals. In the first of these possible circumstances, too few animals might die, thus triggering action S inappropriately; in the second, too many might die, thus triggering action T inappropriately.

The two kinds of error that are possible with choosing actions S and T are quite different, and the relative importances of avoiding them are quite unequal. If the manufacturer decides to take action S when action T would have been correct (i.e., would have corresponded to the state of nature: viz., the batch is toxic), then a toxic batch of the drug will reach the market, and human deaths may result. If the manufacturer decides on action T when action S would have been correct (i.e., would have corresponded to the state of nature: the batch is safe), then a safe batch of the drug will be destroyed, and the manufacturer will incur an unnecessary cost. Clearly, it is much more serious to decide on action S when T is appropriate, than to decide on action T when S is appropriate.

The foregoing three examples should suggest to you that there frequently occur situations in which the consequences of the two possible errors in testing a statistical hypothesis are of unequal importance. It is true that there are also

situations in which the relative importances of the two errors are a matter of subjective judgment, and in which one person may consider it more important to avoid the possible errors connected with decision A while some other person may consider it more important to avoid the possible errors connected with an alternative decision B. However, such subjective elements lie outside the field of statistics. The essential point to notice is that, in most cases, the person testing a statistical hypothesis will consider one of the two possible errors to be the more important one to try to avoid.

In Example 3, the more serious error, the Type I error, is to market a batch of the drug that is dangerously toxic. The rejection of a safe batch of the drug is the less serious, or Type II, error. Consequently, the null hypothesis should have the form, " H_0 : This batch of the drug is dangerously toxic." This hypothesis will be more easily recognized as a statistical hypothesis if we re-phrase it as, say, " H_0 : The fatality rate among animals injected with this batch of the drug will be 25% or more."

Because of the effects of random variations (such as there being, by chance, an unusually large number of extremely resistant animals in the group used to test a batch of a drug), there is inevitably a possibility that a toxic batch might be accepted as a safe batch. This is an example of the fact that in any test of a statistical hypothesis by means of a set of observations that are open to random variation, there is inevitably a positive (i.e., greater than zero) probability of making the more serious of the two possible errors, the Type I error.

The inevitability of this positive probability of a Type I error exists because there is always a chance that the sample set of observations will be atypical of the population from which the sample is drawn, and will thus yield a misleading conclusion as to the nature of the population. The probability of making a Type I error is conventionally denoted by α (the lower-case Greek alpha). That is,

$$\Pr[\text{Type I error}] = \alpha$$

This probability is also often called the "level of significance" or the "risk" of the test of the statistical hypothesis.

Similarly, it is conventional to denote the probability of making the less serious error, the Type II error, by β (the lower-case Greek beta). That is,

$$\Pr[\text{Type II error}] = \beta$$

A closely related concept is the probability, $1 - \beta$, of *not* making a Type II error. The value $1 - \beta$ stems from the fact that the two possible events, making a Type II error and not making a Type II error, cover all the possibilities; this means that we must have:

$$\Pr[\text{making a Type II error}] + \Pr[\text{not making a Type II error}] = 1$$

The probability of not making a Type II error is often called the "power" of the test. That is,

$$\Pr[\text{not making a Type II error}] = 1 - \Pr[\text{making a Type II error}] = 1 - \beta = \text{power of the test}$$

Since not making a Type II error is equivalent to correctly recognizing that H_0 is false, the power of a test is the probability of identifying H_0 as false when it is false. In the usual framework for testing a statistical hypothesis, identifying H_0 as false when it is false is equivalent to accepting the hypothesis that is the alternative to the null hypothesis, when this alternative hypothesis is, in fact, true. Hence,

$$\begin{aligned} \text{power of a test} \\ &= \Pr[\text{accepting the alternative hypothesis when the alternative hypothesis is true}] \\ &= \Pr[\text{recognizing that the null hypothesis is false when it is false}] \end{aligned}$$

In testing statistical hypotheses, one's primary concern is to make α as small as feasible, since this the probability of making the more serious error. Unfortunately, one encounters problems such as cost, practicability, and the fact that lowering α often raises β . These problems can all interfere with making α as small as one would ideally like it to be.

The Formal Notion of a Test of a Statistical Hypothesis

We are now ready to formalize the notion of testing statistical hypotheses with the following definition. In full formality, a test of a statistical hypothesis consists of:

1. A statement of the central hypothesis (the null hypothesis);
2. A statement of the alternative hypothesis to be considered (often this is merely left to be understood as the negation of the null hypothesis);
3. A statement of the maximum acceptable probability of a Type I error, i.e., the highest acceptable value of α , the level of significance;
4. A description of an experiment, including specification of:
 - 4.1 The possible outcomes of the experiment (in technical terms, the "sample space"), i.e., the set of possible values that might be observed for the "test statistic" (the number produced by carrying out a specified set of arithmetic procedures on the values observed in the sample);
 - 4.2 The subset of the possible outcomes that will lead to rejection of the null hypothesis (in technical terms, the "critical region").

For our purposes in LIS 397.1, the specification of the critical region will amount to specifying the threshold, obtained from an appropriate table, with which the test statistic is to be compared. If the observed value of the test statistic exceeds the threshold, the decision will be to reject the null hypothesis. In other words, when you use any of the many well developed and well standardized statistical tests (a few of which you will learn about in LIS 397.1), you will find that steps 4.1 and 4.2 have already been built into the procedures of the test. Before you start learning about the standardized statistical-test procedures, however, you need a more detailed understanding of the rationale that underlies them. That is what the rest of this discussion attempts to provide.

In designing an experiment to test a statistical hypothesis, you must take into account a very important probability. This is the probability that the outcome of the experiment will, *by chance*, be in the critical region *even when the null hypothesis is true*. This probability is just the probability, α , of making the Type I error.

What you do is to assume that the null hypothesis *is true*, and to calculate, on the basis of that assumption, the probability of each possible outcome of the experiment. Having ascertained which outcomes are the least likely when the null hypothesis is true, you put these unlikely outcomes successively--starting with the most unlikely ones--into the critical region. You stop adding these unlikely outcomes only when adding the next one would raise the total probability of the critical region (i.e., the probability of all the outcomes in the critical region) above α . You stop there because otherwise you would have a probability exceeding α of the experiment's leading to rejection of the null hypothesis when the null hypothesis is true.

One detail was omitted in the preceding paragraph: In building a critical region, you do not simply pick the outcomes that are the least likely under the null hypothesis (i.e., when the null hypothesis is true). Rather, you pick outcomes that are both unlikely under the null hypothesis and simultaneously likely (or, at any rate, less unlikely) under the alternative hypothesis.

Development of the Ideas of Tests of Statistical Hypothesis

At this point, it will be helpful to return to the situation of the guest suspected of using loaded dice, and to explore various ways of trying to decide whether he is honest or a cheat. We shall show various ways of trying to determine whether the guest appears to be using loaded dice.

Example 4. Checking on the Suspect by Observing Only 10 Tosses of the Dice

Returning to the situation of Example 1, we assume that you have decided that the more serious possible error would be to call the suspect guest a cheat when he is actually playing fairly. That is equivalent to choosing your null hypothesis as H_0 : The suspect guest's dice are fair. An equivalent hypothesis would be that each face of each die has probability $1/6$ of appearing on any given toss of the dice.

A more realistic way of looking at the question of fairness, however, takes account of the fact that two dice are being used and that the outcomes you are interested in are the various sums of upper faces on the dice. In certain games the

outcome, "The total on the two dice is 7" is especially important. The various sums that are possible on the two dice occur in the following ways:

		SECOND DIE						
		1	2	3	4	5	6	
1:	FIRST DIE	2	3	4	5	6	7	SUM
2:		3	4	5	6	7	8	OF
3:		4	5	6	7	8	9	THE
4:		5	6	7	8	9	10	PAIR
5:		6	7	8	9	10	11	OF
6:		7	8	9	10	11	12	DICE

There are 36 different possible pairs of values for the first die together with the second die; of these, exactly 1 pair totals 2, exactly 2 pairs total 3, ..., exactly 2 pairs total 11, and exactly 1 pair totals 12. Thus the probabilities of the various sums are:

$$\begin{array}{llll}
 \text{Pr}[2] = 1/36 & \text{Pr}[5] = 4/36 & \text{Pr}[8] = 5/36 & \text{Pr}[11] = 2/36 \\
 \text{Pr}[3] = 2/36 & \text{Pr}[6] = 5/36 & \text{Pr}[9] = 4/36 & \text{Pr}[12] = 1/36 \\
 \text{Pr}[4] = 3/36 & \text{Pr}[7] = 6/36 & \text{Pr}[10] = 3/36 &
 \end{array}$$

To simplify matters somewhat, we shall assume that you are primarily interested in whether the suspect's dice have the proper, fair-dice probability, $6/36 = 1/6$, of totaling 7, or whether they have some higher, loaded-dice probability of totaling 7. Thus your null hypothesis can be stated as

$$H_0: \text{Pr}[\text{dice total } 7] = 1/6$$

This implies an alternative hypothesis of the form,

$$H_1: \text{Pr}[\text{dice total } 7] \neq 1/6.$$

This is a composite hypothesis, since there are many ways (many different possible degrees of loading) in which $\text{Pr}[\text{dice total } 7]$ can differ from $1/6$.

In order to simplify matters still further in this example while still taking into account your suspicion about the dice, we shall take as the alternative the following simple hypothesis:

$$H_1: \text{Pr}[\text{dice total } 7] = 0.35$$

which says that the suspect's dice have slightly more than twice the proper probability of coming up 7, since $1/6$ is approximately equal to 0.17. (That is, the dice are heavily loaded, so much so that no real gambler would dare use such blatantly loaded dice for fear of being quickly detected. However, badly loaded dice can help us here, by providing a relatively simple example to start with.)

As a first stab at designing a test of this pair of hypotheses, let us consider observing 10 tosses of the dice, so that we can get an idea of how good a test can result from such a limited number of observations of the reality of the dice's behavior. If we use the term "success" for the outcome, "the dice total 7," then the probabilities of observing exactly 0, 1, 2, ..., 10 successes in 10 tosses of the dice are shown in the following table.

We are dealing with with binomial probabilities, and we use a standard notation, $b(k;n,p)$, to represent the probability of observing exactly k successes in n tosses in each of which the probability of success is p .

k	$b(k;10, 1/6)$	$b(k;10, .35)$	k	$b(k;10, 1/6)$	$b(k;10, .35)$
0	.1615 0555	.0134 6274	6	.0021 7064	.0689 0978
1	.3230 1111	.0724 9166	7	.0002 4807	.0212 0301
2	.2907 1002	.1756 5287	8	.0000 1861	.0042 8138
3	.1550 4534	.2522 1951	9	.0000 0083	.0005 1230
4	.0542 6588	.2376 6840	10	.0000 0002	.0000 2759
5	.0130 2381	.1535 7036			

(Note: The separation of each of the eight-digit probabilities into two groups of four digits is just for ease of reading.)

This table shows, for example, that if the dice are fair, the probability of the suspect's getting exactly six occurrences of "dice total 7" in ten tosses of the dice is .00217064 (or roughly .002); whereas if the dice are loaded so heavily that they have a chance equal to .35 of totaling 7 on any one toss, then the probability of the suspect's getting exactly six occurrences of "dice total 7" in ten tosses is .06890978 (or roughly .069), over 30 times higher than the comparable probability for fair dice.

This example is intended to suggest that the critical region should be built up out of those outcomes that are very unlikely to occur if H_0 is true but are more likely (or, at least, less unlikely) to occur if H_1 is true. This should strike you as reasonable, since, for example, you are much more likely to see six occurrences of the event "dice total 7" in ten tosses if H_1 is true than if H_0 is true. Similar arguments apply to other values involving large numbers of occurrences of "dice total 7." For example, getting exactly ten occurrences of "dice total 7" in ten tosses is very unlikely in either case, but it is about $.00002759/.00000002 = 1379.5$ times more likely (or, if you prefer, less unlikely) if H_1 is true than if H_0 is true.

In short, if you were to see exactly ten occurrences of "dice total 7" in ten tosses, you should certainly bet that H_1 is true rather than that H_0 is true. So your critical region will almost surely include the outcome of your seeing exactly ten occurrences of "dice total 7" in ten tosses. Similarly, it will almost surely include the outcome of your seeing exactly nine occurrences of "dice total 7" in ten tosses, and so on. For brevity, in the rest of this example we shall use "k" to represent the "number of occurrences of 'dice total 7'" in ten tosses.

So far we have not discussed how much risk you are willing to run of making a Type I error. Let us assume that you will settle for the widely used level of significance, $\alpha = .05$. What does this choice imply? Well, when you start forming your critical region with the outcome, ten occurrences of "dice total 7" in ten tosses, or for short, with $k = 10$, you will continue through decreasing values of k . That is, you will include the outcome $k = 9$ in the critical region along with $k = 10$; and then you will also include the outcome $k = 8$, and so on, up to some point short of $k = 0$.

How do you decide where to stop? The essential idea is that you look at the probabilities that each of these outcomes has *when H_0 is true*. After all, what you are trying to limit, to the small value α , is the likelihood of incorrectly reaching the decision to reject H_0 *when it is true*. The critical region is just that set of outcomes such that if you see one of them, you will decide to reject H_0 . You keep adding successively smaller values of k , starting down from $k = 10$, till the total probability of all the outcomes included in the critical region would, *when H_0 is true*, exceed $\alpha = .05$ if the next outcome were included.

In this example, therefore, you would include outcomes $k = 10, 9, 8, 7, 6,$ and 5 in the critical region, for their probabilities under H_0 ("under H_0 " means "if H_0 is true") total only

$$.00000002 + .00000083 + .00001861 + .00024807 + .00217064 + .01302381 = .01546198$$

However, you would not include the outcome $k = 4$ in the critical region because to do so would increase the sum to

$$.01546198 + .05426588 = .06972786$$

That is, including the outcome $k = 4$ would raise the probability of the critical region to over .05. To put it another way, the probability of your seeing, *when H_0 is true*, either exactly ten occurrences of "dice total 7" in ten tosses or exactly nine occurrences of "dice total 7" in ten tosses or ... or exactly four occurrences of "dice total 7" in ten tosses) is .06972786. This is greater than .05, but .05 is what you have chosen as your chosen maximum desired probability of making the Type I error. In other words, including the outcome $k = 4$ in the critical region would raise the probability of your seeing one of the outcomes in the critical region (and hence your deciding to reject H_0) *when H_0 is true* to more than .05, your chosen limit.

The test outlined above (viz., rejecting H_0 if $k = 10$ or 9 or ... or 5) will lead only about 1.5% of the time to a decision to reject H_0 when it is actually true. Thus you will not run much of a risk of *erroneously* accusing the suspect of cheating. Unfortunately, you will, on the other hand, not have much chance of deciding that he is cheating even when he is cheating. For, if he *is* cheating with (heavily loaded) dice having $\text{Pr}[\text{dice total } 7] = .35$, the critical region just outlined will have a combined probability of only

$$.00002759 + .00051230 + .00428138 + .02120301 + .06890978 + .15357036 = .24850397$$

That is, the procedure of deciding to reject H_0 if $k = 10$ or 9 or ... or 5 will lead to your correctly calling a cheat a cheat only about 25% of the time. The other 75% of the time you will fail to recognize a cheat as a cheat; in other

words, about 75% of the time you will accept H_0 when it is false. That is, you will make a Type II error about 75% of the time. More precisely, you will have

$$\Pr[\text{Type II error}] = 1 - .24850397 = .75140603 = \beta$$

You can see that keeping your risk of making a Type I error low has resulted in a fairly high probability of your making a Type II error. Note that

$$\text{power of the test} = 1 - \beta = 1 - .75140603 = .24850397$$

is just the probability of your *correctly* recognizing the situation if H_0 is false. The only way of lowering β or, equivalently, raising the power of the test, while keeping α low enough so that $\alpha \leq .05$, is to increase the sample size.

To summarize Example 4, we can say that a test of the null hypothesis that the dice are fair, or specifically, $H_0: \Pr[\text{dice total } 7] = 1/6$, against the alternative hypothesis, $H_1: \Pr[\text{dice total } 7] = .35$, can be constructed as follows:

If five or more occurrences of the outcome "dice total 7" are observed in ten tosses of the dice, you will conclude that the dice are loaded, so loaded, in fact, that $\Pr[\text{dice total } 7] = .35$.

This test will have a risk of only $\alpha = .01546198$, or about 1.5%, of resulting in the Type I error, viz., calling the suspect a cheat when he is actually using fair dice. On the other hand, this test will have a power of only $1 - \beta = .24850397$, or about 24.8%. That is, the test will have a chance of only about 25% (or 1 chance in 4) of resulting in your deciding that the dice are loaded even when they are actually quite heavily loaded.

A possible alternative test, with a larger power, would be this:

If you observe four or more occurrences of "dice total 7", then you will conclude that the dice are loaded.

While this test does have a higher power (viz., $.24850397 + .23766840 = .48617237$), it also has a higher risk (.06972786) of resulting in the Type I error. Thus you will decide not to include $k = 4$ in the critical region even though doing so would increase the power of the test.

Example 5. Checking on the Suspect by Observing 40 Tosses of the Dice

Suppose that, being dissatisfied with the test that you can construct by observing 10 tosses, you decide to try a different test, with a larger sample. Suppose you decide to try observing 40 tosses and counting the number of times you see the outcome "dice total 7." You are interested in working out how good a test this larger number of observations could provide. The pertinent probabilities are $b(k;40,1/6)$ and $b(k;40,.35)$. They are shown in the following table, along with those of $b(k;40,.20)$, which we shall use shortly.

k	$b(k;40,1/6)$	$b(k;40,.20)$	$b(k;40,.35)$
0	.0006 8038	.0001 3292	.0000 0003
1	.0054 4302	.0013 2923	.0000 0071
2	.0212 2777	.0064 7998	.0000 0743
3	.0537 7702	.0205 1995	.0000 5067
4	.0994 8750	.0474 5238	.0002 5238
5	.1432 6201	.0854 1428	.0009 7845
6	.1671 3902	.1245 6249	.0030 7333
7	.1623 6363	.1512 5447	.0080 3795
8	.1339 5001	.1559 8116	.0178 5352
9	.0952 5334	.1386 4993	.0341 8109
10	.0590 5708	.1074 5368	.0570 5612
11	.0322 1295	.0732 6388	.0837 8873
12	.0155 6959	.0442 6360	.1090 3277
13	.0067 0690	.0328 3424	.1264 5221
14	.0025 8695	.0114 9151	.1313 1577

15	.0008	9681	.0049	7965	.1225	6138
16	.0002	8025	.0019	4518	.1031	1654
17	.0000	7913	.0006	8653	.0783	8724
18	.0000	2022	.0002	1931	.0539	3310
19	.0000	0468	.0000	6348	.0336	2631
20	.0000	0098	.0000	1666	.0190	1180
21	.0000	0019	.0000	0397	.0097	4964
22	.0000	0003	.0000	0086	.0045	3392
23	.0000	0001	.0000	0017	.0019	1062
24	.0000	0000	.0000	0003	.0007	2873
25	.0000	0000	.0000	0000	.0002	5113
26	.0000	0000	.0000	0000	.0000	7801
27	.0000	0000	.0000	0000	.0000	2178
28	.0000	0000	.0000	0000	.0000	0545
29	.0000	0000	.0000	0000	.0000	0121
30	.0000	0000	.0000	0000	.0000	0024
31	.0000	0000	.0000	0000	.0000	0004
32	.0000	0000	.0000	0000	.0000	0001
33-40	.0000	0000	.0000	0000	.0000	0000

(Note: In the above table probabilities shown as .00000000 are actually greater than zero, but they are so tiny that when rounded off to a mere eight decimal places, they appear as zero. See Endnote 2.)

As before, we start forming the critical region by including the outcomes that are the least likely to occur if the null hypothesis is true, i.e., with the observation of 40 occurrences of "dice total 7" in 40 tosses, and we work backwards from there, through 39 occurrences in 40 tosses, 38 occurrences in 40 tosses, etc. From the table it is a straightforward task to verify that a critical region consisting of 12 or more occurrences of the outcome "dice total 7" in 40 tosses will have a total probability under H_0 of

$$\begin{aligned}
 &.01556959 + .00670690 + .00258695 + .00089681 + .00028025 + .00007913 + .00002022 + .00000468 \\
 &+ .00000098 + .00000019 + .00000003 + .00000001 + \text{further terms so tiny that they can be ignored} \\
 &= .02614574
 \end{aligned}$$

The power of this test, i.e., the probability of correctly rejecting the null hypothesis when it is false, is just the sum of the probabilities, under H_1 , of the events in the critical region. It is straightforward (though tedious) to verify that the sum of the probabilities $b(k;40,.35)$ for $k = 12, 13, \dots, 40$ is .79471790. This power is quite an improvement over the .24850397 that was the power of the test using observations of only 10 tosses of the dice.

Thus, by using a critical region consisting of 12 or more occurrences of "dice total 7" in 40 tosses, you would have a 79.5% chance of detecting a cheater who was using dice so heavily loaded as to have $\Pr[\text{dice total } 7] = .35$, while running a risk of only 2.6% of calling an honest player a cheat.

This will probably strike you as a reasonable test. However, it is based on the alternative hypothesis,

$$H_1: \Pr[\text{dice total } 7] = 0.35$$

i.e., on the assumption that a cheating player would use dice so heavily loaded as to come up 7s about 35% of the time. Clearly, this is unrealistic, for a gambler cheating so blatantly would quickly be noticed. A more subtle cheater might use dice loaded so as to yield 7s about, say, 20% of the time. Such dice would be unlikely to be noticed by a casual observer but would still give the cheater an unfair advantage over the long run.

Example 6. Detecting a Subtly Cheating Gambler

Let us assume that you think you may be confronting a subtly cheating gambler. Specifically, let us assume that you want to test your original null hypothesis,

$$H_0: \Pr[\text{dice total } 7] = 1/6$$

against a new alternative hypothesis,

$$H_2: \Pr[\text{dice total } 7] = 0.20$$

You will still have the same critical region as in Example 5: viz., if you observe 12 or more occurrences of "dice total 7" in 40 tosses, you will reject H_0 ; and the level of significance of the test will, of course, remain .02614574.

However, the power of the test will be different, because of the new values, under H_2 , of the probabilities of "dice total 7." Using the middle column of the table in Example 5, you can verify that the power of the test of H_0 against H_2 , i.e., the sum of the probabilities under H_2 of getting 12 or more occurrences of "dice total 7" in 40 tosses, is .09650519.

Clearly, this power is not very satisfactory. The test will not call an honest player a cheat very often (only about 2.6% of the time), which is good; but on the other hand, it will not call a cheat a cheat very often either (only about 9.6% of the time), which is bad. The reason is that the two probabilities, $\Pr[\text{dice total } 7]$ under H_0 and $\Pr[\text{dice total } 7]$ under H_2 , are nearly the same, since 0.20 is very close to $1/6$ (which is approximately .17). In such circumstances, the only way to develop a satisfactory test is to observe a much larger number of tosses, i.e., to use a much larger sample.

Example 7. A Second Attempt at Detecting a Subtly Cheating Gambler

Let us consider what could be done with a larger sample. Using the Gaussian approximation to the binomial distribution, we can try tests using various larger numbers of tosses, to see how things work out. After making several such tries, I found that a test using 1200 tosses seems reasonable.

Before going further with the example of the 1200-toss test, let us consider the basic reasoning. With fair dice, we would expect an average of $1200(1/6) = 200$ occurrences of a total of 7 in 1200 tosses. If the dice are loaded to the extent specified by H_2 , then we would expect an average of $1200(.2) = 240$ occurrences of a total of 7 in 1200 tosses. In broad terms, if we observe 1200 tosses, and see that the number of 7s is close to 240 rather than to 200, we should bet that the dice are loaded rather than fair.

The only technical question is deciding just exactly what we mean by "close to 240 rather than to 200." To answer this question, we turn to the Gaussian distribution. We use the fact that the Gaussian distribution is a good approximation to the binomial distribution for large samples, i.e., for large values of n . The approximation involves setting

$$\mu = np \text{ and } \sigma = \sqrt{npq}$$

When we assume that H_0 is true, we have $n = 1200$ and $p = 1/6$, so under H_0 it works out that $\mu = 200$ and $\sigma = 12.9099$. With these numbers we have

$$z = \frac{X - \mu}{\sigma} = \frac{X - 200}{12.9099}$$

as the transformation that links any value of X , the number of occurrences of a total of 7 in 1200 tosses, to the corresponding value of z , the number of standard deviations of the Gaussian distribution.

To see how we can use the Gaussian approximation to the binomial, we first discuss the situation in more detail. If we assume that the dice are fair, i.e., that H_0 is true, then we expect to observe close to 200 occurrences of 7 in 1200 tosses. That is, in any given set of 1200 tosses with fair dice, we are likely to see somewhere between 190 and 210 7s. We are less likely to see 211 tosses of 7 than 210, somewhat less likely to see 212 7s than 211, even less likely to see 213 7s, and so on. (Similar statements could be made about 189 7s, 188 7s, 187 7s, etc., but such statements are of little interest to us here, since we are trying to decide between fair dice and dice loaded so as to make 7s occur more often than they ought to.)

Intuitively, we can feel quite comfortable in deciding to believe that the dice are fair if we see, say, 203 7s. But we would be less comfortable in betting on H_0 if we saw 213 7s, still less comfortable if we saw 223 7s, and so on. We would be extremely uncomfortable in betting on H_0 if we saw 243 7s in 1200 tosses, since 243 is close to—in fact, exceeds—the average number of 7s we would see in 1200 tosses with dice loaded as in H_2 . Nevertheless, it is not absolutely impossible (just very unlikely) for fair dice to yield 243 7s in 1200 tosses.

In this example the Type I error would be to decide that the dice are not fair when, in fact, they are fair. We want to have a less than 5% chance of making the Type I error (i.e., we want to use $\alpha = .05$). Let us use X to represent the number of 7s in 1200 tosses. What we are seeking is a rule along the following lines: If X exceeds some threshold K , then we reject H_0 ; otherwise, we accept H_0 . The threshold K will lie somewhere between 200 and 240, since observing a value of X close to 200 will lead us to bet on H_0 whereas observing a value of X close to 240 will lead us to bet on H_2 . To put it exactly, K will be the (smallest whole) number such that the probability that fair dice will yield more than K 7s in 1200 tosses is less than .05. In symbols, K will be such that $\Pr[X > K] < .05$.

We find the numerical value of K by using the Gaussian approximation. The analog to having $\Pr[X > K]$ be less than .05 is to have $\Pr[z > k]$ be less than .05 for some value k . That is, in Gaussian terms we must find a value k such that the area under the Gaussian curve to the right of k is less than .05. From any table of the Gaussian distribution we can find that $k = 1.645$, i.e., that the area under the Gaussian curve to the right of 1.645 is just under .05. In symbols, we have $\Pr[z > 1.645] < .05$. To find K , we can use the value 1.645 for z in the formula that connects z , X , μ , and σ

$$1.645 = z = \frac{X - \mu}{\sigma} = \frac{X - 200}{12.9099}$$

from which it is easy to find that $X = 221.24$. Thus $K = 221$ is the desired threshold for X .

Our decision rule is this: If we observe 221 or fewer 7s in 1200 tosses of the gambler's dice, we will accept H_0 ; i.e., we will decide to believe that the dice are fair. But if we observe 222 or more 7s in 1200 tosses, we will reject H_0 and accept H_2 ; i.e., we will decide to believe that the dice are loaded so as to have a 20% chance of yielding a total of 7.

This test has $\alpha = .05$, and further calculations would show that its power is equal to .91. That is, with this decision rule we run only a 5% risk of calling the gambler a cheat if his dice are fair, and we have a 91% chance of calling him a cheat if his dice are loaded (i.e., loaded as specified by H_2). This test would be quite satisfactory--provided that we had a chance to observe 1200 tosses of the gambler's dice.

Conclusion

The General Pattern of Carrying Out Tests of Statistical Hypotheses

The foregoing examples should help you understand how one sets up tests of statistical hypotheses. Each of the examples used a simple alternative hypothesis, the least complicated situation. Composite alternative hypotheses look more difficult, but in practice they are sometimes handled by replacing the composite alternative hypothesis by a simple alternative hypothesis that is contained in the composite and is of special interest. More often, however, one works with the composite alternative hypothesis that is the negation of the null hypothesis and, by so doing, forgoes the possibility of dealing, at least in a simple fashion, with the power of the test.

This latter course is the one that we shall find used in the standardized procedures for testing statistical hypotheses that form a large part of the statistical analysis in LIS 397.1. Each of these standardized procedures has a very clear method of dealing with the level of significance of the test (which, after all, is by definition the more important issue). On the other hand, each of these standardized procedures ignores the power of the corresponding test.

The Idea of a Test Statistic

These standardized procedures share a common way of defining the critical region. Each procedure specifies a particular set of arithmetic operations that lead to a number called the "test statistic" of the procedure. When the arithmetic operations are performed on the observations made in a particular investigation, the result is the observed value of the test statistic. This observed value is compared with a threshold value, obtained from a table of such thresholds. If the observed value of the test statistic *exceeds* the tabled threshold value, then the null hypothesis that is built into the procedure is rejected; otherwise, the null hypothesis is accepted.

This is equivalent to saying that the critical region of the particular standardized procedure consists of all those outcomes for which the test statistic exceeds the threshold. The thresholds themselves are calculated so as to provide the desired levels of significance of the test; most of the tables provide thresholds corresponding to levels of significance of $\alpha = .05$ and $\alpha = .01$.

A somewhat more formal summary is this: The probability of making the Type I error (i.e., the more serious error) is denoted α and called the "level of significance" or the "risk" of the test. Thus the level of significance of a test is the probability of (incorrectly) deciding that H_0 is false when it really is true. The probability of making the Type II error (i.e., the less serious error) is denoted β . The probability of making the other possible decision in the circumstance that H_0 is false is denoted $1-\beta$ and is called the "power" of the test. Thus the power of the test is the probability of correctly deciding that H_0 is false when it really is false.

The Traditional Way of Using a Test Statistic

Traditionally the typical method for testing a statistical hypothesis has been this: If the observed value of the pertinent test statistic exceeds the threshold value found from the relevant table (the threshold being selected on the basis of the size of the sample and the pre-chosen level of significance), then you should reject the null hypothesis; otherwise, you should accept it (strictly speaking, not reject it). We can express this rule more concisely as follows:

If observed value of the test statistic \leq tabled threshold, Accept H_0
 $>$ tabled threshold, Reject H_0

An Alternative Way of Using a Test Statistic

In recent years an alternative method of testing a statistical hypothesis has become possible with computer programs that carry out statistical procedures. Typically such programs not only report the observed value of the test statistic, but also calculate the probability that *this observed value will occur, by chance alone, when the null hypothesis is true*.

A *large* value for this probability means that the result that was observed is the kind of result that is *likely* to occur when the null hypothesis is true; hence, the fact that this result did occur suggests strongly that the null hypothesis is true. On the other hand, a *small* value for this probability means that the result that was observed is quite *unlikely* to occur when the null hypothesis is true; hence, the fact that this result did occur suggests strongly that the null hypothesis is false.

In particular, if this probability (viz., the probability of observing the observed value of the test statistic by chance alone when the null hypothesis is true) is small enough to be less than what you chose as the level of significance, α , of the test procedure, then the decision rule is: Reject H_0 . The reasoning is this: What you observed is something that is very unlikely to have occurred when the null hypothesis is true. Nevertheless, it did occur. Therefore, you should decide that its occurrence establishes, beyond the level of reasonable doubt expressed by α , that the null hypothesis is false.

If, however, this probability is equal to or greater than the level of significance, then the decision rule is: Accept H_0 . The reasoning is this: What you observed is something that is likely to have occurred when the null hypothesis is true. Therefore, its occurrence is quite consistent with the idea that the null hypothesis is true. Therefore, you should decide that its occurrence provides a reasonable basis for you to continue to believe that the null hypothesis is true.

Symbolically, we have

If $\Pr[\text{observing the observed value of the test statistic when } H_0 \text{ is true}] \geq \alpha,$ Accept H_0
 $< \alpha,$ Reject H_0

Thus it is convenient to use computer programs that provide a calculation of this probability (the probability of observing the observed value of the test statistic by chance alone when the null hypothesis is true). With such programs, you do not have to compare the observed value of the test statistic with the appropriate threshold value, from the pertinent table, in order to reach a decision about rejecting or accepting the null hypothesis. Instead, you simply observe whether the calculated probability is, or is not, less than α , and make your decision accordingly.

Common Types of Statistical Hypotheses

$$H_0: \mu = \mu_0$$

The population mean is some specified number μ_0 . E.g., "The average daily circulation total is 123."

$$H_0: \mu_1 = \mu_2$$

The mean of Population 1 equals the mean of Population 2. That is, the two populations have the same mean. E.g., "The average [mean] cost per online search using Service A is the same as the average [mean] cost per online search using Service B."

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots \text{ etc.}$$

The means of Populations 1, 2, 3, ..., etc. are all equal. E.g., "Among high-school students, the average [mean] number of library books borrowed per student each semester is the same for sophomores, juniors, and seniors."

$$H_0: \rho_{XY} = 0$$

Variables X and Y are not correlated. E.g., "There is no correlation between the age and the salary of a typical librarian."

$$H_0: \text{Variable A has a distribution of ______ type.}$$

E.g., "The distribution of daily circulation totals is Gaussian."

$$H_0: \text{Variables A and B are not associated.}$$

E.g., "There is no association between the sex of a library patron and the type of book the patron prefers."

Endnotes

1. Examples 1 and 3 of the different types of errors were adapted from: Neyman, Jerzy. *First Course in Probability and Statistics*. New York, NY: Henry Holt; 1950. 350p. LC:50-9260.

2. For example, the probability of getting 7 every time in 40 tosses with fair dice, $b(40;40,1/6)$, is approximately

$$.0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0007\ 8408$$

or 7.841×10^{-32} . If you tossed a pair of dice once every second--day and night without stopping--for the rest of eternity, it would take you about $1/(7.841 \times 10^{-32}) = 1.275 \times 10^{31}$ seconds, on the average, to get a series of 40 tosses in which every toss was a 7. Since there are approximately 31,556,926 seconds in a year, this means that you would toss the dice approximately $(1.275 \times 10^{31})/31,556,926 = 4.040 \times 10^{23}$ years (i.e., about 404 billion trillion years), on the average, between each run of forty 7s in a row with fair dice.