A Basic Introduction to Statistical Inference

James H. Steiger

Introduction

The traditional emphasis in behavioral statistics has been on hypothesis testing logic. This emphasis is changing rapidly, and is being replaced by a new emphasis on effect size estimation and confidence interval estimation. Before we can understand the source of the discontent with traditional hypothesis testing logic, we have to understand what it is, and where it came from. This brief article only gives you a basic introduction. In this chapter, we shall begin by discussing, in the context of a non-statistical example, the general class of circumstances that hypothesis testing was designed for, and then introduce a number of technical terms used in discussions of hypothesis testing. Finally, we shall discuss some of the problems and shortcomings of hypothesis testing.

Decision-Making Under Uncertainty: A Medical Example

Hypothesis testing logic was designed to handle a fundamental situation with these characteristics:

1. A dichotomous decision (yes-no, buy-sell, sick-well) has to be made.
2. The decision is based on some state of the world.
3. There is an available diagnostic indicator of the state of world.
4. This diagnostic indicator is not perfectly reliable.

This kind of situation is not only common in statistics, but in numerous other areas. For example, in medical testing, the laboratory decides whether or not a particular situation (pregnancy, disease) exists on the basis of imperfect information. In signal detection tasks in psychophysics, a participant has to decide whether or not a stimulus is present under conditions where there is substantial uncertainty.

As a simple example of medical testing, suppose there is a new disease called hyperkeluria (HPK). This disease has been discovered recently, and unfortunately is often fatal. Early diagnosis and treatment can cure the disease, however. Testing for the disease is extremely difficult. So far, only one test has been produced. It involves centrifuging part of a blood sample, and placing a drop of the resulting solution into a special reagent. This reagent is initially clear, but changes to a pink or red color when the person tested is infected with HPK.

Unfortunately, the indicator is imperfect. It doesn’t always yield the same color. For people who do not have HPK, there is a range of colors produced. This range extends from perfectly clear to moderately pink for most individuals. Similarly, there is a range of colors produced for people who are infected with HPK. Their test solution colors tend to range from moderately pink to very red. Unfortunately, these indicator distributions
have ranges that overlap. The situation is shown graphically in Figure 1. The indicator distribution for individuals with HPK is plotted in red, the indicator distribution for people who do not have HPK is plotted in blue. The height of the indicator distributions represent how relatively likely a particular color value is to be found in people from that population. The total area under each distribution curve is 1, and the area under the curve between any two points represents the proportion of people who have color values between those two points. Color varies from clear to red as you move from left to right on the $X$-axis. To help dramatize this, I’ve overlaid the actual color of the indicator on the graph.

For example, in the graph in Figure 1, the most common value of color for people who do not have HPK is a light pink, corresponding to a “pink value” of –2 on the somewhat arbitrary “pink scale” plotted on the $X$-axis. Exactly half the area under the blue curve lies to the right of –2, so 50% (a proportion of .50) of the people who do not have HPK produce color values pinker than –2.

In order to perform the test, decision point must be selected. If the solution is redder than this decision point, the individual is judged to have HPK, i.e., the test result for disease is “positive.” If the solution is less red than the decision point, the decision is “no disease, i.e., a “negative” result.

As you can see, there is some overlap in the two curves.

![Figure 1. Frequency Distributions for an imperfect indicator of HPK. The distribution on the left, drawn in blue, represents the relative frequency of occurrence of colors for people who do not have HPK. The distribution on the right, drawn in red, represents the relative frequency of occurrence of test colors for people who do have HPK. The color boundary for deciding whether to declare a positive or negative result is also shown with a vertical line.](image)
Keep in mind that a person either has HPK or doesn’t, and, with this fixed decision rule, the test will decide that the person either is or is not infected. As a result, there are 4 possibilities that we must keep in mind. They are embodied in Table 1.

<table>
<thead>
<tr>
<th>Test Decision</th>
<th>State of the World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient is Infected</td>
<td>Patient is Not Infected</td>
</tr>
<tr>
<td>Positive</td>
<td>Correct Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

Table 1. 2x2 decision table showing the 4 possible outcomes from a standard dichotomous medical testing process. Outcomes representing an error are highlighted.

Two of the four things that can happen are errors. I’ve highlighted them, because we have to be very concerned about errors in medical testing. In this case, however, a False Negative is much more serious than a False Positive, because HPK can be cured if it is caught early.

**Estimating Error Probabilities**

As I mentioned above, in a normal curve, the area under the curve between two points is also the proportion of cases that fall between the two points, or, more formally, the probability of a case being between two points. Since the total area under the normal curve is 1, and the curve is symmetric, we can roughly guess the probability of a False Negative or a False positive by examining areas under the curves in Figure 1.

Let’s try to estimate the probability of a False Negative. A False Negative occurs when a person has HPK, but obtains a test result in the “Negative” region, on the left side of the decision point. For convenience, I’m going to shade in this area in. It turns out, this area is about 2.28% of the total area under the red bell-shaped graph, so the probability of a false negative is .0228. Because the decision point is placed halfway between the centers of the two distributions, the probability of a False Positive is the same as that of a False Negative.
Figure 2. Frequency distributions for an imperfect indicator of HPK, with the False Negative probability (.0228) shaded in. This shaded area represents the False Negative cases, i.e., those cases that occur under the red graph (representing the distribution of test results for people who have HPK), and to the left of the decision point. In this case, you can see from the symmetry of the graph that the probability of a False Positive is the same as that of a False Negative.

NOTE: You have not learned how to calculate probabilities from a normal curve, so you would simply estimate these values, keeping in mind that the total area under each bell-shaped curve is 1. I’m giving you the exact values just for convenience. However, if you draw the picture reasonably accurately, you can do a reasonable job of estimating probabilities.

Suppose that we decided to change our decision criterion to eliminate almost all False Negatives, because the cost to an individual of a False Negative diagnosis is much higher than the cost of a False Positive. This would involve moving our decision point to the left. Below is a picture of the changed decision rule, with the area representing the probability of a False Positive shaded in.
What have we learned so far?

1. In a dichotomous decision process under uncertainty, based on a single imperfect indicator with a fixed decision point, there are 4 possible things that can happen, and two of them represent errors.

2. There is a trade-off between False Positives and False Negatives. Sliding the decision point to the left to eliminate False Negatives increases the probability of a False Positive. Sliding the decision point to the right increases False Negatives while reducing False Positives.

**Statistical Hypothesis Testing**

Statistical hypothesis testing is much like medical testing. In its most common variant, it is designed to produce a dichotomous decision under uncertainty. In this case, the uncertainty again comes from natural variability. But in this case, the variability comes from the “luck of the draw,” i.e., sampling variability.
**The Basic Setup and Terminology**

Suppose we reduce the problem artificially to some very simple terms. We are interested in whether a drug we have invented can increase IQ. Let’s suppose (this is a highly artificial example) that we wanted to test whether (a) the drug did not increase IQ or (b) did increase IQ. We test these alternatives in the following way. We gather two samples of equal size, administer the drug to Group 1, a placebo to Group 2, and then give each group an IQ test. We compute the average IQ for the two groups. We use this as our indicator of how the drug works.

We are interested, of course, in how this drug behaves in the entire population. We cannot afford to experiment on the entire population, unfortunately. We can describe this state of affairs as follows:

1. We are interested in $\mu_1 - \mu_2$, the difference between the population means that would result if all people were administered the drug and the placebo. Call this quantity $\delta$.
2. Unfortunately, we are stuck with an imperfect indicator of $\delta$, namely, $D$, the difference between the two sample means that we hope is representative of the population value $\delta$, but may not be.

Clearly, when $\delta = 0$, $D$ will generally not be 0. It will show some random variation around zero over repeated samples. This distribution of values over repeated performance of the same sampling experiment is called a *sampling distribution*.

So, just as the color of the test solution in HPK testing varies when you test sick people, so does the value of the $D$ indicator vary when you run experiments to try to ascertain the value of $\delta$.

In our situation, let’s pretend for the moment that it is known that (a) the drug cannot decrease IQ, and (b) if it increases IQ, it will do so by exactly 5 points. So our task is to decide whether $\delta = 0$, or $\delta = 5$.

Just in the case of medical testing, we shall used a fixed decision point to decide either in favor of $\delta = 0$, or in favor of $\delta = 5$, depending on the value of $D$. Suppose, for the time being, you decided to use a sample of 72 people in each of your two groups. It is known that, in the general population, IQ scores have a mean of 100 and a standard deviation of 15. Statistical theory that you learn later this year tells you that, with such a sample size, the distribution of $D$ over repeated samples has bell-shaped (normal) distribution with a mean of $\delta$, and a standard deviation of 2.5. So, there are two possible distributions of $D$, one if the drug has no effect, the other if the drug increases IQ by 5 points. The situation is shown in
Figure 4. Distribution of $D$, the difference between the two group means, over repeated samples, when the population difference $\delta = 0$ (the blue distribution on the left) or when $\delta = 5$ (the red distribution on the right). In this case, the population standard deviation is assumed to be 15, and sample sizes are assumed to be 72 in each group.

Suppose that we need to set a decision point, and, if a value of $D$ is greater than this decision point, we decide in favor of $\delta = 5$, otherwise we decide in favor of $\delta = 0$. Once we have set such a decision rule, there are 4 possibilities we need to consider. First, we need some terminology. In this case, the “statistical null hypothesis,” which we will refer to as $H_0$, is that $\delta = 0$. The “statistical alternative hypothesis,” which we will refer to as $H_1$, is that $\delta = 5$. Since either the null hypothesis or the alternative hypothesis must be true, and its opposite must be false, there are only 4 contingencies we need to consider. These contingencies are summarized in Table 2.

<table>
<thead>
<tr>
<th>Test Decision</th>
<th>State of the World</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is True</td>
<td>$H_0$ is False</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>Correct Acceptance</td>
</tr>
<tr>
<td></td>
<td>$(1 - \alpha)$</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type I Error</td>
</tr>
<tr>
<td></td>
<td>$(\alpha)$</td>
</tr>
</tbody>
</table>

Table 2. 2x2 table showing the contingencies in standard hypothesis testing logic. Symbols for the probability of the outcome in each cell are shown in parentheses. Outcomes representing errors are highlighted.
Note that each of the 4 cells has a name, and each of the 4 cells also has a probability in parenthesis. The two probabilities on the left side of the table add up to 1, and the two probabilities in parentheses on the right side of the table also add up to 1. This is because, for any given test, you are either on the right side of the table or the left, because the reality (hidden from you, the person performing the test) is that the null hypothesis ($H_0$) is either true or false. In the parlance of the X-Files, “the truth is out there.” The problem is, we don’t know the truth, so we need to analyze all the contingencies.

Perhaps the most commonly referred to entry in Table 2 is the Type I error rate, $\alpha$, also commonly referred to as the “significance level” of the test. In the original development of hypothesis testing logic, the significance level is fixed in advance to some reasonable value (often .05 or .01) by properly adjusting the decision point for the test. (How this is done is a technical matter to be pursued in your statistics course.) Once the data are observed and $D$ calculated, a decision is made to either accept or reject $H_0$. “Degrees” of acceptance or rejection are not relevant. In other words, one rejection is the same as another, and a miss is as good as a mile.

When statistical software became popular, the notion of “$p$-value” or “$p$ level” became popular as well. Rather than announce “the null hypothesis was rejected,” statistical software programs would print a probability level like “$p = .045$.” Formally, the “$p$-value” is the lowest value of $\alpha$ for which your data would yield a rejection of the null hypothesis. So, for example, if $p = .045$, you would reject the null hypothesis “at the .05 level,” but not “at the .01 level.” The general rule is, if $p$ is less than a value, you would reject the null hypothesis at that level of significance. Perhaps another example will clarify. Suppose the statistical program reports that $p = .0043$. In this case, you would reject the null hypothesis at the .05 significance level (i.e., if $\alpha$ was set at .05), and the .01 level, but not at the .001 level.

The reporting of “$p$-value” was often accompanied by the use of multiple stars in tables of statistical results. For example, it is quite common for statistical tables of correlations to have one, two, or three stars next to correlation coefficients that are significant at the .05, .01, and .001 levels, respectively.

In some journal articles, rather than report an exact $p$-value, the author will make a statement relative to one of the more commonly used significance level values. For example, “$p < .01$” means that the result was significant “beyond the .01 level.”

Because “more significant” results are associated with lower $p$-values within a fixed, specific testing situation, many people came to believe that the $p$-value is somehow useful as a general indicator of the strength of a statistical result across studies, or across circumstances within large studies. It isn’t. Why? With very large sample sizes, a miniscule experimental effect may be “highly statistically significant,” with a $p$-value less than .0001. Moreover, $p$-values depend on several factors beside sample size, and these factors can vary substantially even within a single analysis, when that that analysis has several significance tests (like a 3-way ANOVA). Although a $p$-value is useful as a communication device that allows us to decide at a glance whether a result is significant for a particular significance level, it is not useful as a general indicator of the meaningfulness of a result.
Two Kinds of Hypothesis Testing Logic

Note that two of the possibilities in Table 2 represent errors. And, it turns out, just as in the medical testing example in the previous section, there is a trade-off between the two types of error. All other things being equal, aspects of a test that increase $\alpha$ will decrease $\beta$, and vice-versa. So it is incumbent upon us to understand what each type of error represents in a given situation.

This understanding is made more complicated by the fact that there are two distinctly different kinds of hypothesis testing logic in common use in psychology. One is called “Reject-Support” logic, the other “Accept-Support” logic. “Reject-Support” logic is by far the more common of the two in psychology.

In Reject-Support testing, the statistical null hypothesis, $H_0$, is the opposite of what the experimenter actually believes, so rejecting $H_0$ supports the experimenter’s point of view.

In Accept-Support testing, $H_0$ is what the experimenter believes, so accepting $H_0$ supports the experimenter’s point of view.

Reject-Support Testing and Statistical Power

Let’s begin by considering Reject-Support testing. Notice that a Type I error involves rejecting $H_0$ when it is really true. So in Reject-Support testing, a Type I error is a false confirmation of the experimenter’s actual hypothesis. You might say that a Type I error is a false positive for the experimenter.

Earlier in the course, we discussed how difficult it is to correct false positives once they are published, especially if they are theoretically desirable and/or politically popular. There are many reasons for this, several of which we discussed under the heading of “sociology of science.” Suffice to say, it may take a decade to erase a Type I error from the literature once it becomes popular. Because of this, there is a long tradition in psychology, and in many other sciences, of requiring statistical tests to be performed with Type I error probability (i.e., $\alpha$) controlled at a low level, usually no greater than .05, but also often .01. In your subsequent studies, you will learn that it is relatively straightforward in many instances to control Type I error exactly (or almost exactly) at a desired low level.

A Type II error in Reject-Support testing represents an incorrect disconfirmation of the experimenter’s true belief. This is a result that can be devastating for a researcher, and one that the researcher must take pains to avoid. Imagine being correct on some scientific point, but having the statistical test declare you to be incorrect. Obviously, this kind of result cannot be allowed to happen very often, or science (not to mention the careers of individual experimenters) may suffer. Subsequently, we shall see that the most straightforward way that an experimenter can keep the probability of a Type II error (i.e., $\beta$) low and power high is to increase sample size. However, increasing sample size beyond a certain point can be wasteful, and expensive. Consequently, modern statistical
practice often involves careful calculations to decide how large a sample size is “just large enough.”

Many introductory textbooks in social science research methods act as if Reject-Support testing is the only kind of statistical testing psychologists engage in. As a result, they state that the experimenter’s substantive hypothesis is the reverse of the null hypothesis (this is true in Reject-Support testing, but is not true in Accept-Support testing).

Because of the long tradition of Reject-Support testing in psychology, many of the implicit beliefs that psychologists hold about statistical testing can be incorrect when Accept-Support testing is being performed. However, when Reject-Support testing is performed, the following facts generally hold:

1. Society, in the person of journal editors, reviewers, and readers, will generally insist that $\alpha$ be no greater than .05. $\alpha$ is often referred to as the “significance level” of the test.

2. It is up to the researcher to make certain that sample size and other factors yield a statistical power that is reasonably large. Usually, .80 is the minimum acceptable level for power, and .20 is the largest allowable level of $\beta$.

In Figure 5 below, we show the decision rule for the situation shown in Figure 4, with the areas of the curve representing Type I error and Type II error shaded in.

![Figure 5. Diagram of distributions of $D$ under a true and false null hypothesis. The distribution of $D$ when $\delta = 0$ is shown in blue, the distribution when $\delta = 5$ is shown in red. The decision point is shown, together with shaded areas that represent the probabilities of Type I and Type II errors. Sample size is 72 per group, with population standard deviation = 15. You can see that, with the test criterion set to allow a Type I error probability of .05, the Type II error probability is about .36, and power is therefore only about .64. This](image-url)
power is too low to be considered acceptable. We would like to increase the power, and decrease \( \beta \). In the medical testing example, we simply adjusted our decision criterion in order to reduce False Negatives, but in this case, we have already moved our decision criterion as far to the left as is permissible. What can we do to increase power, while keeping \( \alpha \) at .05?

The key to resolving our predicament is to realize that the width of the bell-shaped curves in Figure 5 is inversely related to the square root of sample size. So, for example, quadrupling sample size will make the curves half as wide. With a large separation between the two curves, power will be large while \( \alpha \) remains low. Figure XXX shows the distributions of \( D \) under \( H_0 \) and under \( H_1 \) (that is, when \( \delta = 0 \) and \( \delta = 5 \), respectively), when the sample sizes have been quadrupled to 288 per group.

![Diagram of distributions of D under a true and false null hypothesis. The distribution of D when \( \delta = 0 \) is shown in blue, the distribution when \( \delta = 5 \) is shown in red. The decision point is shown, together with shaded areas that represent the probabilities of Type I and Type II errors. Sample size is 288 per group, 4 times as large as in Figure 5.](image)

Paradoxes of Accept-Support Testing

In most situations in psychology, Reject-Support testing is performed. However, there are situations where the goal of the study is to demonstrate a small or nonexistent effect. Consider the case of bioequivalence testing, in which two drugs are compared for the
bioavailability of their active ingredient. Often the goal of such testing is to demonstrate that a new, cheaper version of a drug has the same bioavailability levels as a more expensive version of the same drug. In this case, the $D$ might stand for differential bioavailability, and the experimenter may be wishing, ideally, to show that $\delta = 0$.

In this case, it may not be clear at the beginning of the study whether, if the drugs are different, which one has higher bioavailability. Consequently, the null and alternative hypotheses might be stated as follows:

$$
H_0 : \delta = 0 \\
H_1 : \delta \neq 0
$$

(1)

Note that, in this case, values of $\delta$ that are appreciably greater or less than zero might be taken as evidence against the statistical null hypothesis. In this situation, we will need two rejection regions, one on the far right, and one on the far left, and we are engaging in a two-tailed (or two-sided) significance test. The general tradition in psychology when performing a two-sided is to use rejection regions that have equal probability of being entered if the null hypothesis is true. Moreover, note that, in this case, the statistical null hypothesis is what the experimenter hopes (and probably believes) is true. Consequently, accepting the null hypothesis supports the experimenter’s belief, and Accept-Support logic is used.

In this case, the roles of a Type I error and a Type II error are reversed. Now, a Type II error is a false confirmation of the experimenter’s belief, because the experimenter wants not to reject $H_0$. A Type I error is now a false disconfirmation of the experimenter’s belief.

There is a kind of paradox here. If the experimenter simply runs a sloppy, low-power experiment with small sample size, then the desired result (i.e., not to reject the null hypothesis), can be found.

Suppose that we believed that use of a particular drug has no effect on IQ, and we were trying to show that. Suppose sample sizes are 72 for each group of a two group study, and the null and alternative hypotheses are as shown in Equation (1). Then, assuming a population standard deviation of 15, the two rejection regions would be set up as in Figure 7.
Figure 7. A two-sided rejection region for testing the null hypothesis that $\delta = 0$, when both sample sizes are 72, and $\sigma$, the population standard deviation, is 15. The distribution of the statistic $D$ under the null hypothesis is plotted with a solid blue line, while the distribution when $\delta = 5$ is shown in dotted red. Notice how, with a 2-sided rejection region, the right rejection point is farther away from zero than with a one-sided rejection rule. Power is slightly larger than .5. So, despite the fact that the null hypothesis is false, it has about a 50-50 chance of being accepted. Hence, in Accept-Support testing, the experimenter can be “reward” with a low power experiment.

Now, suppose we quadruple sample size to 288 per group. Quadrupling sample size will double precision of estimate, and halve the width of the distribution of $D$. Then the rejection regions would be as in Figure 8. This allows a tighter specification of the rejection region, and increases power.
Figure 8. A two-sided rejection region for testing the null hypothesis that $\delta = 0$, when both sample sizes are 288, and $\sigma$, the population standard deviation, is 15. The distribution of the statistic $D$ under the null hypothesis is plotted with a solid blue line, while the distribution when $\delta = 5$ is shown in dotted red. Power is about .98.

For example, suppose the drug decreases IQ by exactly 1 point. This is a rather trivial effect, but, as shown in Figure 9, the probability of rejecting the null hypothesis when $\delta = 1$ is about .126. Note that there is a modest probability of rejection in the upper rejection region (about .123) and a tiny probability of rejection in the lower region (about .003). In this case, it seems that high precision is actually working against the experimenter in a sense, as there is about a 1 in 8 chance that the null hypothesis should be rejected. In other words, the likelihood of getting a “favorable” result would have been substantially higher for the experimenter if he had used a smaller sample size and run a sloppy experiment.
Figure 9. Rejection regions when the null hypothesis is that $\delta = 0$. The distribution of $D$ is under the null hypothesis shown in blue, while the distribution of $D$ when $\delta = 1$ is shown in dotted red. In this case, there are two samples each of size 288, and the population standard deviation $\sigma$ is equal to 15. You can estimate power from the area under the dotted red curve. The area in the upper rejection region is actually .123, and the area in the lower region .003. So the total power is .126.

Mindful of the ways that Accept-Support testing can reward sloppy experimentation, many writers have argued for the use of confidence intervals, which are numerical intervals that have a high “long run” probability of including the true parameter value. A typical example is the 95% confidence interval for $\delta$, which we can construct using a well-known numerical procedure after observing $D$. A confidence interval statement goes something like this: “I am 95% confident that $D$ is between ____ and ____.”

All other things being equal, the larger the sample size and the greater the precision in an experiment, the narrower the confidence interval will be, reflecting your greater certainty in the information the data have provided about the actual location of $\delta$. So a confidence interval informs you of two things: (1) where the parameter is likely to be located on the number line, and (2) how precisely you have determined this.