Hypothesis Testing and Interval Estimation

James H. Steiger

November 17, 2003

1. Topics for this Module

1. An Idealistic Special Case — When σ is Known.

2. Confidence Interval Estimation

   (a) Taking a Stroll with Mr. Mu

3. Hypothesis Testing

   (a) Parameter Spaces and Sample Spaces

   (b) Partitioning the Parameter Space

   (c) Partitioning the sample space

   (d) Raw Score Rejection Rules

   (e) Error Rates

      i. Type I Error

      ii. Type II Error

   (f) Statistical Power

      i. Power vs. Precision

4. Test Statistics

5. Standardized Rejection Rules

Home Page

Print

Title Page

◀◀ ▶▶

◀   ▶

Page 2 of 37

Go Back

Full Screen

Close

Quit

6. 1-tailed vs. 2-Tailed Tests

7. Influences on Power

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

Home Page

Print

Title Page

◀◀   ▶▶

◀   ▶

Page 3 of 37

Go Back

Full Screen

Close

Quit

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

2. An Idealistic Special Case — Statistical Procedures When σ is Known.

In this section, we will employ a distribution that is a standard "teaching device" in a number of behavioral statistics texts — statistical procedures related to the distribution of the sample mean $\overline{X}_\bullet$ when the population standard deviation σ is known. This situation is unrealistic, in the sense that we are no more likely to know σ than we are to know μ. However, it turns out that, to a surprising extent, it actually doesn't matter much that we don't know σ.

When the population distribution is normal, the distribution of the sample mean is exactly normal, with mean μ, and standard deviation $\sigma_{\overline{X}_\bullet} = \sigma/\sqrt{N}$. The normality follows from the fact that the sample mean is a linear combination of independent normal observations, and that any linear combination of multivariate normal normal random variables is itself normally distributed. The mean and standard deviation of the distribution follow from linear combination theory.

When the population distribution is not normal, the distribution of the sample mean will often still be very close to normal in shape, because of the Central Limit Theorem we discussed previously.

We shall proceed, for a while, as if the distribution of the sample mean can be assumed to be normal to a high degree of accuracy. We will now examine two key topics: interval estimation and hypothesis testing.

3.   Confidence Interval Estimation

Assume, for the time being, that we know that the distribution of $\overline{X}_{\bullet}$ over repeated samples is as pictured below:
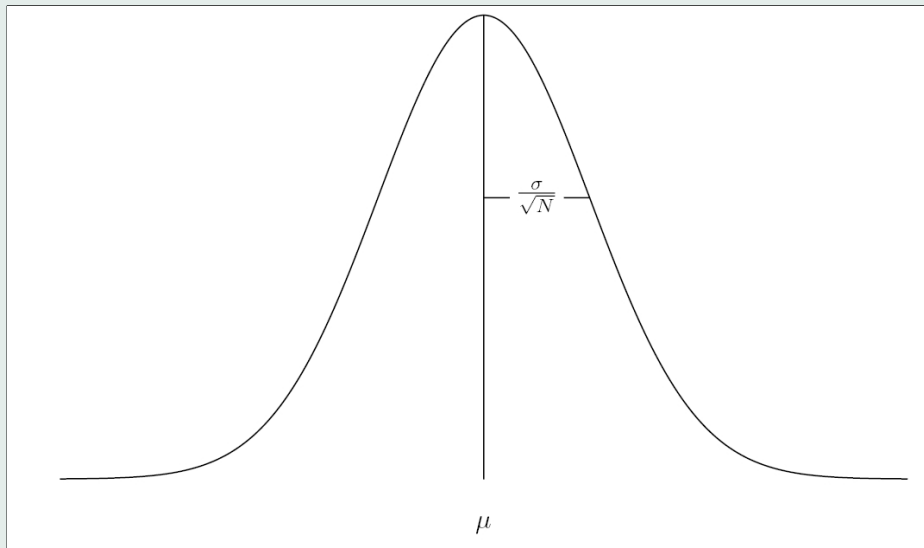

Figure 1: The Sampling Distribution of $\overline{X}_{\bullet}$

This graph demonstrates the distribution of $\overline{X}_{\bullet}$ over repeated samples.

In the above graph, as in any normal curve, 95% of the time a value will be between Z-score equivalents of -1.96 and $+1.96$. These points are at a raw score that is 1.96 standard deviations below the mean and 1.96 standard deviations above the mean. Consequently, if we mark points on the graph at $\mu - 1.96\sigma/\sqrt{N}$ and $\mu + 1.96\sigma/\sqrt{N}$, we will have two points between which $\overline{X}_\bullet$ will occur 95% of the time.

We can then say that

$$\Pr\left(\mu - 1.96\frac{\sigma}{\sqrt{N}} \leq \overline{X}_\bullet \leq \mu + 1.96\frac{\sigma}{\sqrt{N}}\right) = .95 \tag{1}$$

and, after applying some standard manipulations of inequalities, we can manipulate the μ to the inside of the equality and the $\overline{X}_\bullet$ to the outside, obtaining

$$\Pr\left(\overline{X}_\bullet - 1.96\frac{\sigma}{\sqrt{N}} \leq \mu \leq \overline{X}_\bullet + 1.96\frac{\sigma}{\sqrt{N}}\right) = .95 \tag{2}$$

Equation 2 implies that, if we take $\overline{X}_\bullet$ and add and subtract the "critical distance" $1.96\sigma/\sqrt{N}$, we obtain an interval that contains the true μ, in the long run, 95% of the time.

3.1. Taking a Stroll with Mr. Mu

Even if you are not familiar with the manipulation of inequalities, there is a way of seeing how Equation 2 follows from Equation 1. The first

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

Home Page

Print

Title Page

◀◀ ▶▶

◀ ▶

Go Back

Full Screen

Close

Quit

inequality states that there is a critical distance, $1.96\sigma/\sqrt{N}$. and $\overline{X}_\bullet$ is within that distance of μ 95% of the time, over repeated samples. Now imagine that you had a friend named Mr. Mu, and you went for a stroll with him. After a certain length of time, he turned to you and said, "You know, about 95% of the time, you've been walking within 2 feet of me." You could, of course, reply that he has also been within 2 feet of you 95% of the time. The point is, if $\overline{X}_\bullet$ is within a certain distance of μ 95% of the time, it must also be the case (because distances are symmetric) that μ is within the same distance of $\overline{X}_\bullet$ 95% of the time.

3.2. Constructing a Confidence Interval

The confidence interval for μ, when σ is known, is, for a $100(1-\alpha)\%$ confidence level, of the form

$$\overline{X}_\bullet \pm z^*_{1-\alpha/2} \frac{\sigma}{\sqrt{N}} \tag{3}$$

where z^* is a critical value from the standard normal curve. For example, $z^*_{.975}$ is equal to 1.96. At first this notation is somewhat difficult to master. When you are talking about a 95% confidence interval, α is equal to .05, and $1-\alpha/2$ is .975.

Example 3.1 (A Simple Confidence Interval) *Suppose you are interested in the average height of Vanderbilt male undergraduates, but you*

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

only have the resources to sample about 64 *men at random from the general population. You obtain a random sample of size* 64*, and find that the sample mean is* 70.6 *inches. Suppose that the population standard deviation is somehow known to be* 2.5 *inches. What is the* 95% *confidence interval for* μ*?*

Solution 3.1 *Simply process the result of Equation* 3*. We have*

$$70.6 \pm 1.96 \frac{2.5}{\sqrt{64}}$$

or

$$70.6 \pm .6125$$

We are 95% confident that the average height for the population of interest is between 69.99 and 71.21 inches.

Remark 3.1 *A confidence interval provides an indication of precision of estimation (narrower intervals indicate greater precision), while also indicating the location of the parameter. Note that the width of the confidence interval is related inversely to the square root of* N*, i.e., one must quadruple* N *to double the precision.*

4.   Hypothesis Testing

Hypothesis testing logic was imported into psychology from the "hard sciences," and has dominated the landscape in behavioral statistics ever since. Hypothesis testing is most appropriate in situations where a dichotomous decision (pass-fail, infected-clean, buy-sell) needs to be made on the basis of data in the face of uncertainty. Unfortunately, many situations in the social sciences do not fit this mold. To see why, we need to investigate the logic of hypothesis testing carefully.

4.1.   Parameter Spaces and Sample Spaces

Suppose we are interested in a particular parameter, say a population mean μ. The *parameter space* Ω is the set of all possible values of the parameter. Often this is modeled to be the entire real number line from $-\infty$ to $+\infty$.

4.2.   Partitioning the Parameter Space

Generally, we perform statistical tests to ascertain whether the parameter satisfies some restriction. The most common restriction is that it is a particular value, or falls within some specified range of values. This restriction is commonly stated as a *statistical hypothesis*, which confines the parameter to a particular region of the parameter space.

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

Definition 4.1 (*Statistical Hypothesis*) *A statistical hypothesis is a statement delineating a region of the parameter space in which the parameter may be found. A statistical hypothesis partitions the parameter space. Usually in our applications there will be two regions.*

Hypotheses about μ restrict it to a particular region of the parameter space. Here are some examples:

$$\mu = 100$$

$$\mu < 100$$

$$60 \leq \mu \leq 70$$

A very common form of statistical inference pits two hypotheses about a statistical parameter against each other. The goal of the statistical decision process is to decide between the two hypotheses. One hypothesis, labeled H_0, is called the *null hypothesis*, and the other, labeled H_1, is called the *alternative hypothesis*. Usually, these two hypotheses are mutually exclusive and exhaustive, i.e., they are opposites that, taken together, exhaust all possibilities, so that one or the other must be true.

Often, in practice, the statistical null hypothesis is exactly the opposite of what you believe, so your goal is to gather enough evidence to reject it in favor of the alternative hypothesis. Such a situation is called *Reject-Support Testing*, and is by far the most common form of statistical testing in the behavioral sciences.

Home Page

Print

Title Page

◀◀   ▶▶

◀   ▶

Page 10 of 37

Go Back

Full Screen

Close

Quit

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

On occasion, the situation is reversed — the null hypothesis is what the experimenter believes, so accepting the null hypothesis supports the experimenter's theory. In such a case, the test is called *Accept-Support Testing*. Some of the common conventions of behavioral statistics are based on Reject-Support logic, and may be inappropriate or illogical when Accept-Support testing is being performed.

We begin with an example of Reject-Support testing. Suppose you wished to prove that a particular group of men has an above-average height. The average height in the general population of men is known to be 70 inches. You state the statistical null hypothesis as

$$H_0 : \mu \leq 70$$

and the alternative hypothesis as

$$H_1 : \mu > 70$$

In this case, you actually believe the alternative hypothesis. Since either H_0 or H_1 (but not both) must be true, the falsity of H_0 implies the truth of H_1 and vice-versa. The two hypotheses partition the number line from $-\infty$ to $+\infty$ into two mutually exclusive and exhaustive regions, as shown in Figure 2. (The arrow pointing to H_0 indicates that the null hypothesis region includes the value 70.)
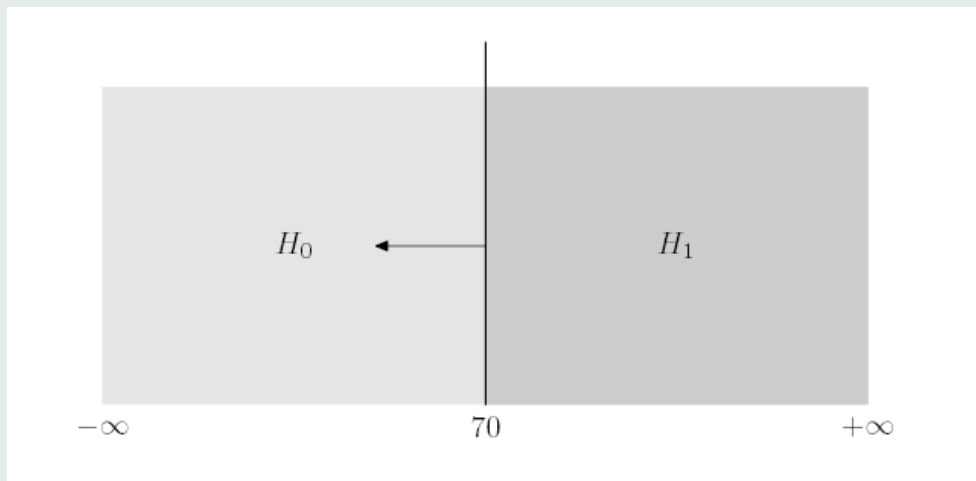
Figure 2: Partitioning the Parameter Space with Null and Alternative Hypotheses

4.3. Partitioning the sample space

In the preceding example, we arrived at a pair of statistical hypotheses that might be used to assess whether a particular group has an above-average height. To evaluate these hypotheses about μ, we seldom have the luxury of examining all the members of the group of interest. Consequently,

Home Page

Print

Title Page

◀◀ ▶▶

◀ ▶

Page 12 of 37

Go Back

Full Screen

Close

Quit

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

we must base our decision on a sample. Typically, we assume simple random sampling from the population of interest, i.e., a sample of size N is selected so that all members of the population have an equal probability of being sampled. (In practice, of course, true random sampling is seldom achieved.)Once the sample of size N is obtained, we calculate a *statistic*, examine its value, and try to decide which of the two hypotheses, H_0 or H_1, is more reasonable. The major problem is that the statistic, being based on a finite sample, provides an estimate of the parameter that almost certainly is inaccurate to a degree. The amount by which the statistic is incorrect is called *sampling error*, which varies randomly from sample to sample. In any particular sample, you may experience relatively large or small sampling error. An experimenter cannot control the luck of the draw, but decision rules can be employed that control the long run probabilities of making the wrong decision. To devise an effective decision rule for deciding between two hypotheses on the basis of a statistic, it is extremely useful to have good information about the sampling distribution of the statistic. Suppose a sample of size $N = 25$ is taken from a population with a normal distribution, a mean of 70, and a standard deviation of 10. Based on our preceding discussion, we can state that, over repeated samples, the sample mean, $\overline{X}_\bullet$, will have a distribution that is normal, with a mean of 70, and a standard deviation of 2. This fact can be exploited to create a *statistical decision rule* for testing the null and alternative hypotheses discussed above. The statistical decision rule declares, in advance, which

values of the statistic $\overline{X}_\bullet$ will result in acceptance or rejection of the null hypothesis.

4.4. A Raw Score Rejection Rule

The *Sample Space* of the statistic is the set of all possible values of the statistic. Formally, we say that the decision rule *partitions the sample space* of the statistic. For example, suppose we are trying to decide between the hypotheses in the previous example on the basis of a sample mean $\overline{X}_\bullet$ obtained from a sample of 25 observations, in a case where the population is normal and the standard deviation is known to be 10. Our construction of a decision rule is complicated by the fact that the statistic $\overline{X}_\bullet$ has sampling error. If $\overline{X}_\bullet$ were a perfect indicator of μ, our decision rule would be the same as the diagram in Figure 2. That is, we would decide in favor of H_1 if we observed an $\overline{X}_\bullet$ greater than 70, otherwise we would retain H_0. However, because there is sampling error, any statistical decision rule we construct will have leave open the possibility of decision errors. Typically, this involves constructing a decision rule that is similar in appearance to the diagram of the statistical hypotheses, but is "a little fuzzy" to allow for sampling error. Suppose, for example, we decide on a one-sided or "one-tailed" decision rule shown in Figure 2. With this rule, if a value of $\overline{X}_\bullet$ is greater than or equal to 73.29, we decide in favor of H_1, otherwise we retain H_0.

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

Note that, so far, I haven't described how I arrived at the decision rule of Figure 3. The rule is based on a consideration of error rates.
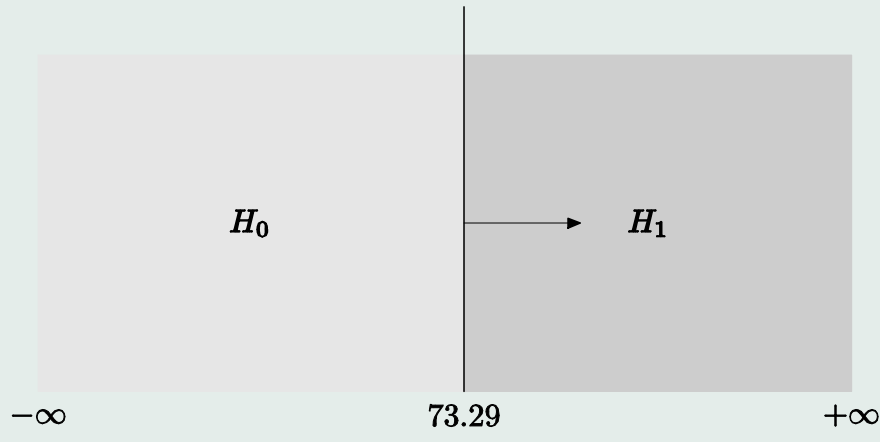

Figure 3: A Statistical Decision Rule for the Hypotheses in Figure 2

4.5. Error Rates

Since there are two possible states of the world (H_0 is either true or false) and there are only two possible decisions, there are only 4 possible out-

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

comes, shown in Table 1.

	State of the World	
Decision	H_0 True	H_0 False
Accept H_0	Correct Acceptance $(1 - \alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Rejection $(1 - \beta)$

Table 1: 2×2 Statistical Decision Table

4.5.1. Type I Error

A Type I error is an incorrect rejection of H_0, and occurs with probability α. The formal definition of a Type I error remains the same whether one is performing Accept-Support or Reject-Support testing, but the implications of a Type I error are different in the two situations. In Reject-Support testing, a Type I error represents a false positive for the experimenter's belief. Mindful of this, "society" in the person of journal editors and other authority figures requires a statistical test to be designed so that α, the probability of a Type I error, is small. In Accept-Support testing, however, a Type I error is an incorrect rejection of the experimenter's belief.

In the previous example, the decision rule was fixed at a value of 73.29. Values of $\overline{X}_\bullet$ above this cut-off resulted in rejection of H_0. For a particular decision rule, we can immediately compute bounds for α, the probability

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

of a Type I error, under the supposition that H_0 is true. For example, consider the decision rule diagrammed in Figure 3. If the null hypothesis is true, the sample mean $\overline{X}_\bullet$ will have a normal distribution with a mean no greater than 70, and a standard deviation of $\sigma/\sqrt{N} = 10/\sqrt{25} = 2$. The mean of the sampling distribution may be less than 70 and the H_0 will still be true, but in that case it is easy to see that the probability of a Type I error will be lower than when $\mu = 70$. So, to compute an upper bound for α with the decision rule shown in Figure 3, we plot the distribution of $\overline{X}_\bullet$, and compute the probability of obtaining a result in the rejection region. The situation is shown in Figure 4.

To compute α, we simply calculate the probability of obtaining a value higher than 73.29 in a Normal(70,2) distribution. This can be calculated easily by converting 73.29 to its Z-score equivalent, i.e.,

$$Z = \frac{73.29 - 70}{2} = 1.645$$

The probability above 1.645 in a standard normal distribution is .05. Of course, the value 73.29, which yields the familiar .05 value for α, did not arrive out of thin air. Rather, it was calculated, deliberately, to yield the value .05. Specifically, we know that, for the one-sided test to yield an α value of .05, the area under the normal curve to the left of the decision point must be .95, and the area to its right must be .05. Scanning down the normal curve table, we find that the Z-score value for the decision criterion must be 1.645. Next, we must convert this value to a point in the sampling

Home Page

Print

Title Page

◀◀   ▶▶

◀   ▶

Page 17 of 37

Go Back

Full Screen

Close

Quit

distribution of the sample mean, which, under the null hypothesis, has a mean of 70 and a standard deviation of 2. Since we must have

$$\frac{\overline{X}_\bullet - 70}{2} = 1.645$$

it trivially follows that, in this case, the critical value of $\overline{X}_\bullet$ separating the H_0 and H_1 decision regions must be

$$\overline{X}_\bullet = (2)(1.645) + 70 = 73.29$$

Note that this calculation is tedious, and that the raw score rejection point will generally change with each new situation. Fortunately, there is a way around this tedium.

4.5.2. Type II Error

A Type II error is an incorrect acceptance of H_0, and occurs with probability β. In Reject-Support testing, a Type II error represents an incorrect failure to support the experimenter's belief, i.e., a false negative for the experimenter's belief. A Type II error represents (in Reject-Support testing) a potential debacle for the experimenter — the experimenter's belief is correct, but the statistical test fails to detect this! It is therefore important for the Reject-Support tester to take steps to assure that β is small. Once a decision rule is set, β may be calculated by presupposing an effect size, i.e., an amount by which the null hypothesis is false.

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

4.6.   Statistical Power

Statistical Power is defined as $1 - \beta$. The amount by which the null hypothesis is false is called an *experimental effect*. One can think of an experimental effect as a signal, and statistical power as the ability to detect the signal. For example, earlier we decided on a rejection rule that controls α at or below .05. Suppose that the null hypothesis is false. Rather than having a value of 70, μ is actually 75. What will the statistical power be in this situation?

To compute power, one must draw the actual distribution of $\overline{X}_{\bullet}$ and see what percentage of the area falls in the rejection rejion. From the general sampling distribution of the sample mean we can calcualte that with $\mu = 75$, $\sigma = 10$, and $N = 25$, the sample mean will have a Normal(75,2) distribution. So the power of the test is the probability of obtaining a rejection with this distribution. This is the probability of obtaining a value greater than or equal to 73.29 in this distribution, as diagrammed in Figure 5.

To compute the area, we convert 73.29 to a Z-score, obtaining $Z = (73.29 - 75)/2 = -.855$. Consulting a standard normal curve table, we find that the area above $-.855$ is approximately .804.

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

4.6.1.   Power vs. Precision

Precision of estimation is reflected in the narrowness of a sampling distribution. In general, the greater the precision of estimation, the greater the statistical power, because of the "sharpening of vision" that occurs with greater precision. However, if the experimental effect is very large, even a low precision experiment may have high power — so power and precision are not the same thing.

Home Page

Print

Title Page

◀◀   ▶▶

◀   ▶

Page 20 of 37

Go Back

Full Screen

Close

Quit
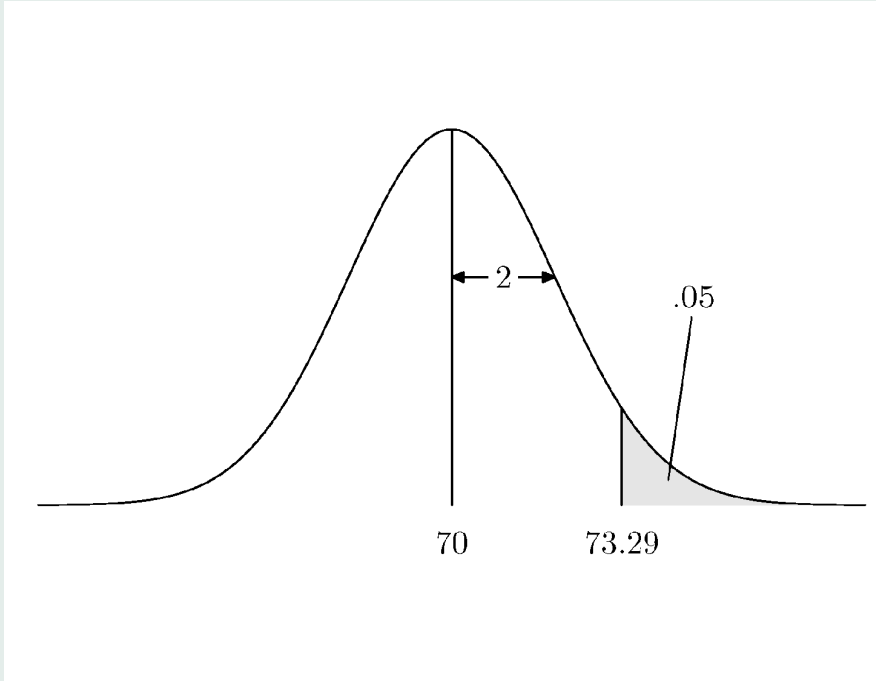
Figure 4: Calculating α for a simple decision rule

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power
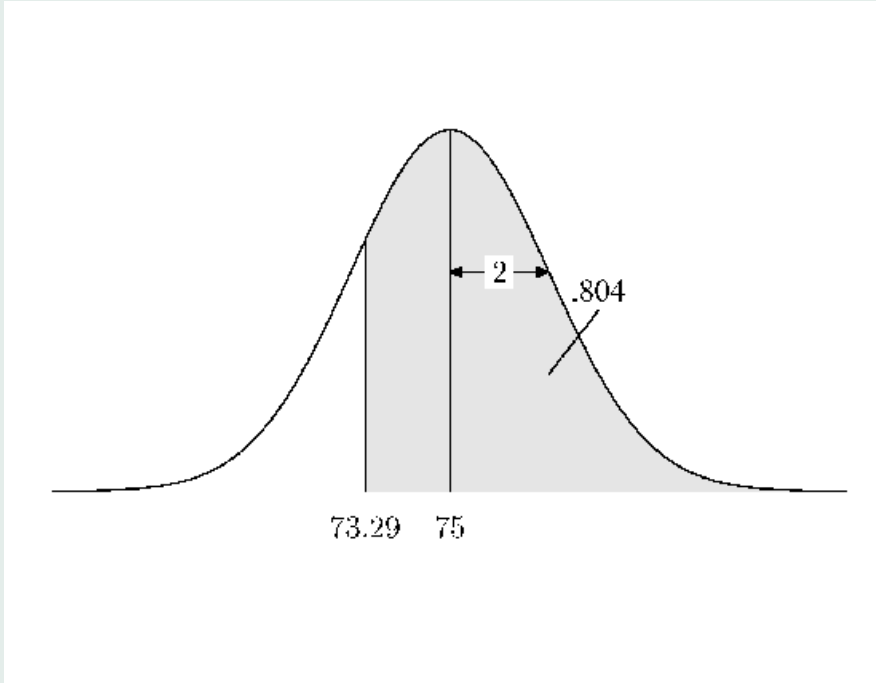
Figure 5: Calculating power when $\mu = 75$

5. Test Statistics and Standardized Rejection Rules

In the preceding section, we determined that, with a sample size of 25, the power for our one-sided test to detect a false null hypothesis when $\mu = 75$ is about .80. We calculated the power by

- Calculating a *statistical decision rule* (and associated rejection region for H_0) for the statistic $\overline{X}_\bullet$;

- Drawing the sampling distribution of $\overline{X}_\bullet$ under the true state of the world, i.e., when $\mu = 75$;

- Calculating the area of this sampling distribution that falls in the rejection region.

This method of power calculation for the test on a single mean is the one taught to most undergraduates in the social sciences. Ironically, it is neither efficient nor realistic. In practice, one seldom calculates a single power value. Rather, one calculates power for a range of sample sizes, and for a range of possible values of μ, for several reasons. First, if one actually knew μ, there would be little point performing the experiment, so a range of possible values of μ may need to be considered. Second, an initial power calculation performed by hand often turns out to be disappointing.

Home Page

Print

Title Page

◀◀ ▶▶

◀ ▶

Go Back

Full Screen

Close

Quit

Suppose, for example, you felt that a power of .95 was necessary in order to undertake a particular experiment, and you were trying to determine, in advance, the sample size N required to obtain that level of power when the null hypothesis is false by 5 points (i.e., when μ is at least 75). The calculation we performed in the preceding section would be disappointing to you, and your next question, after determining that power is "only" .80, might well be "How large an N do I need to have a power of .95 or greater?"

The method of calculation described in the preceding section is rather inefficient for answering this question. If you keep your (algebraic) wits about you, you might stumble on the realization that you can solve for the minimum N required to yield a desired power by solving a system of two simultaneous equations, an approach taught in a number of textbooks. The reason the answer is not simpler is that, for each N, the rejection region for $\overline{X}_\bullet$ changes, *and* the width of the distribution of $\overline{X}_\bullet$ also changes. The fact that two important determinants of power are changing simultaneously as a function of N adds to the complexity of the problem.

However, there is a much simpler approach to power calculation in this situation that, surprisingly, is seldom taught in textbooks. Recall something that is almost inevitably taught, namely that, rather than calculating a new rejection value for $\overline{X}_\bullet$ for each new situation, there is an easier way, based on the use of a "standardized" version of $\overline{X}_\bullet$. Suppose, for example, you wish to perform the single-sided test with $\alpha = .05$. Simply employ the

following decision rule for testing the hypothesis that $\mu \leq \mu_0$ against the alternative that $\mu > \mu_0$. First, compute the "test statistic"

$$Z = \frac{\mu - \mu_0}{\sigma/\sqrt{N}}$$

Then adopt the decision rule to reject H_0 in favor of H_1 whenever the value of Z reaches the 95th percentile in the standard normal curve, i.e., whenever $Z \geq 1.645$. . Referring back to the numerical values used in the specific example in the previous section, you can quickly determine that Z will reach 1.645 if an only if $\overline{X}_\bullet$ reaches 73.29, so the two decision rules are equivalent. The advantage of the standardized approach is that, for a single-sided hypothesis $H_0 : \mu \leq \mu_0$ with an α of .05, the Z-statistic will always have the same rejection point (1.645). So if μ_0 changes, or σ changes, or N changes, the rejection rule remains the same.

5.1. Standardized Effect Size, Power, and Sample Size

It is somewhat ironic that although most textbooks are quick to recognize the value of the "standardized test statistic" approach to streamlining hypothesis testing, they fail to recognize the even more significant advantages of this approach in power calculation and sample size estimation. We will now investigate the extension of the standardized test statistic approach

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

to power calculation, and, in the process, discover some important general principles that hold for many different types of power calculations. First, we recall a key result regarding the Z-statistic.

Proposition 5.1 *Suppose a sample of N observations is taken from a normal distribution with mean μ and standard deviation σ, and the test statistic*

$$Z = \frac{\overline{X}_\bullet - \mu_0}{\sigma/\sqrt{N}}$$

is calculated. Then Z will have a distribution that is Normal($\sqrt{N}\,E_s$, 1), where

$$E_s = \frac{\mu - \mu_0}{\sigma}$$

The parameter E_s, often referred to as a measure of "standardized effect size," may be thought of as the amount by which the null hypothesis is wrong, expressed in standard deviation units. Return again to our previous power calculation. We contemplating a test of the null hypothesis that μ is less than or equal to 70, in a case where the true μ is 75, sample size is $N = 25$, and the population standard deviation is $\sigma = 10$. With this single-sided significance rejection region, the decision rule for the Z statistic is to reject H_0 for any value of the Z-statistic greater than or equal to the 95th percentile of its distribution when $\mu = \mu_0$. Note that, if $\mu = \mu_0$, then $E_s = 0$, and the Z-statistic has a distribution that is Normal(0,1). So the

critical value for our decision rule is at 1.645. This rule is diagrammed in Figure 6 \figureref{StandardizedDecisionRule}.
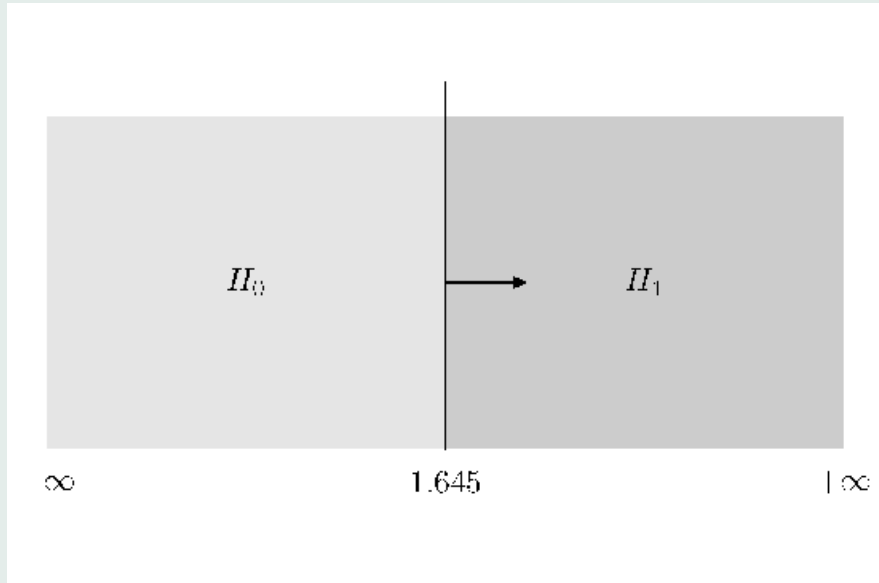

Figure 6: A standardized decision rule

To calculate power for the case where $\mu = 75$, we calculate the distribution of the Z-statistic, superimpose it on the decision rule, and calculate

the probability of a rejection. In this case, $E_s = (75 - 70)/10 = .5$, and, from Proposition 5.1, we find that the Z-statistic has a distribution that is Normal($\sqrt{25}.5$, 1), or Normal(2.5, 1). As shown in Figure 7, the power of the test is the probability of obtaining a value greater than 1.645 in a normal distribution with a mean of 2.5 and a standard deviation of 1. To calculate this probability, we compute the Z-score equivalent of 1.645.
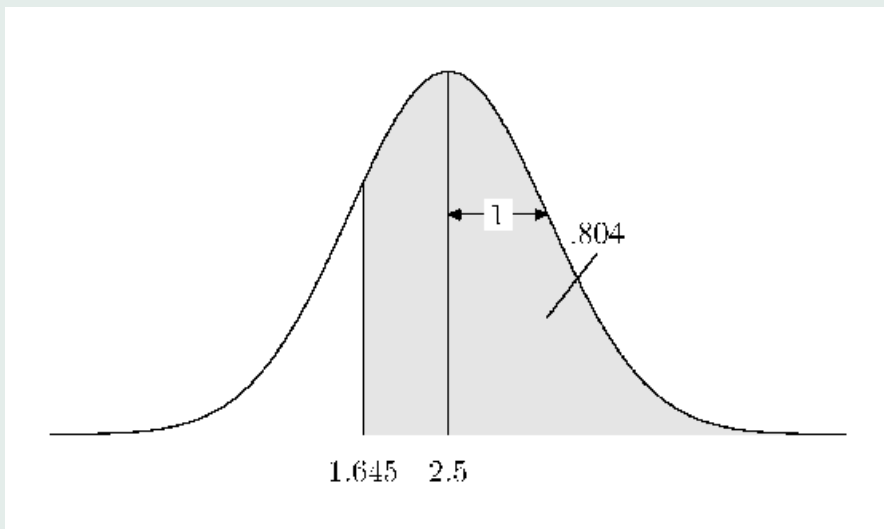

Figure 7: Power calculation using a standardized rejection rule

This is $(1.645 - 2.5)/1 = -.855$. Note how the calculation is made easier by the fact that the standard deviation of the test statistic is 1, so there is no necessity to divide by it. The area to the right of $-.855$ in the standard normal distribution is .804.

We have reviewed two techniques for performing the same power calculation. The second method turns out to be substantially simpler, although the magnitude of the advantage may not yet be apparent. We now investigate how the standardized approach allows one to

- Calculate power directly from the normal curve table;

- Grasp much more easily the importance of various influences on power, and

- Calculate sample size (N) required to yield a given level of power without solving a simultaneous equation system.

To begin with, recall the final steps we took in solving for power. Having computed our rejection point (1.645), and the mean of the sampling distribution of the Z-statistic (2.5), we computed the area to the right of 1.645 in a normal distribution with a mean of 2.5 and a standard deviation of 1. Reducing the operation to its essentials, we subtracted 2.5 from 1.645 (obtaining $-.855$) and computed the area to the right of $-.855$ in the standard normal distribution.

Home Page

Print

Title Page

◀◀   ▶▶

◀   ▶

Page 29 of 37

Go Back

Full Screen

Close

Quit

Topics for this Module

An Idealistic Special . . .

Confidence Interval . . .

Hypothesis Testing

Test Statistics and . . .

1-tailed and 2-tailed . . .

Influences on Power

At this point, the typical reader might still be a few steps short of "seeing the forest for the trees." It is quite common in statistics for some exceedingly simple realities to be concealed in a blizzard of formulas, but it is also the case that casting an operation in the proper mathematical notation can sometimes reveal some important details. First, suppose we call the rejection point R, and the mean of the sampling distribution M. Computing power involves simply computing the area to the right of $R-M$ in the standard normal distribution. Since normal curve tables are more likely to give the cumulative probability, it might be useful to alter the procedure slightly. Recalling that the area to the right of a value x in a standard normal distribution is equal to the area to the left of $-x$ in the same distribution, we can also calculate power as the area to the left of $M - R$. In other words, all we need to do is compute $M - R$, and look up its cumulative probability on the standard normal curve table. So the standard normal curve table is also a power table for this test. Notice that this one table can be used to solve for power for *all values* of μ, σ, and N.

To make the discussion more succinct, let's proceed to develop some notation for talking about the standard normal curve table. Z-score values are listed on the left side of the table, cumulative probabilities on the right. We will refer to a Z-score value as z, the cumulative probability as $\Phi(z)$. If we graph $\Phi(z)$, we see an S-shaped curve as shown in Figure 8.

Note that $\Phi()$ is a monotonic, strictly increasing function that is invertible, i.e., each value of $\Phi(z)\$ corresponds to a unique value of z. We

denote the inverse function $\Phi^{-1}(p)$. Some examples will help make the notions of the standard normal CDF and its inverse more concrete. The standard normal curve CDF is tabled in Glass and Hopkins. To compute $\Phi(z)$, one scans down the left column labeled z, finds the value, then scans over to the value of $\Phi(z)$ ("area below z) in the same row. So, for example, if one were evaluating $\Phi(1.0)$, one would scan down until the value 1.0 is encountered in the z column, then scan over in the same row to the value of $\Phi(1.0)$, which is .8413. To compute $\Phi^{-1}(p)$, one scans down the "area below z" column until the value closest to p is found, then moves to the left column in the same row to find the value of z. For example, $\Phi^{-1}(.9772) = 2.00$.

Armed with this notation, let us now return to our standardized approach to power calculation. We decided that the power of the test could be described as the area to the left of the quantity $(M - R)$ in the standard normal curve, where M is the mean of the sampling distribution of $\overline{X}_\bullet$, and R is the rejection point used for the Z-statistic. Using our new notation, we can write that same quantity as $Power = \Phi(M - R) = \Phi\left(\sqrt{N}E_s - R\right)$.

Since the $\Phi()$ function has an inverse, $\Phi^{-1}(\Phi(x)) = x$, and we can actually solve this equation to determine the minimum N required to produce a given power. Specifically, let P be the required power. Then

$$P = \Phi\left(\sqrt{N}E_s - R\right)$$

so
$$\Phi^{-1}(P) = \sqrt{N} E_s - R$$
and
$$N = \left(\frac{\Phi^{-1}(P) + R}{E_s}\right)^2$$

Usually N in the above will not be an integer, and to exceed the required power, you will need to use the smallest integer that is not less than N. This value is called ceiling(N). Moreover, in a single-sided (1-tailed) test, we can write
$$R = \Phi^{-1}(1 - \alpha)$$

So we can make the resulting expression look really complicated! For a 1-sided test such as the one currently under consideration,
$$N = \text{ceiling}\left[\left(\frac{\Phi^{-1}(P) + \Phi^{-1}(1 - \alpha)}{E_s}\right)^2\right]$$

Example 5.1 *Suppose that the null hypothesis is that $\mu \leq 70$, but the true state of the world is that $\mu = 75$, while $\sigma = 10$. Find the minimum sample size needed to achieve a statistical power of .90 when Type I error is set at $\alpha = .05$.*

Solution 5.1 *Scanning down the normal curve table we find that $\Phi^{-1}(.90) = 1.283$, and $\Phi^{-1}(1 - .05) = \Phi^{-1}(.95) = 1.645$. The standardized effect size*

is $(\mu - \mu_0)/\sigma = .5$. So the minimum N is

$$
\begin{aligned}
N &= \text{ceiling}\left[\left(\frac{1.282 + 1.645}{.5}\right)^2\right] \\
&= \text{ceiling}\left[5.854\right]^2 \\
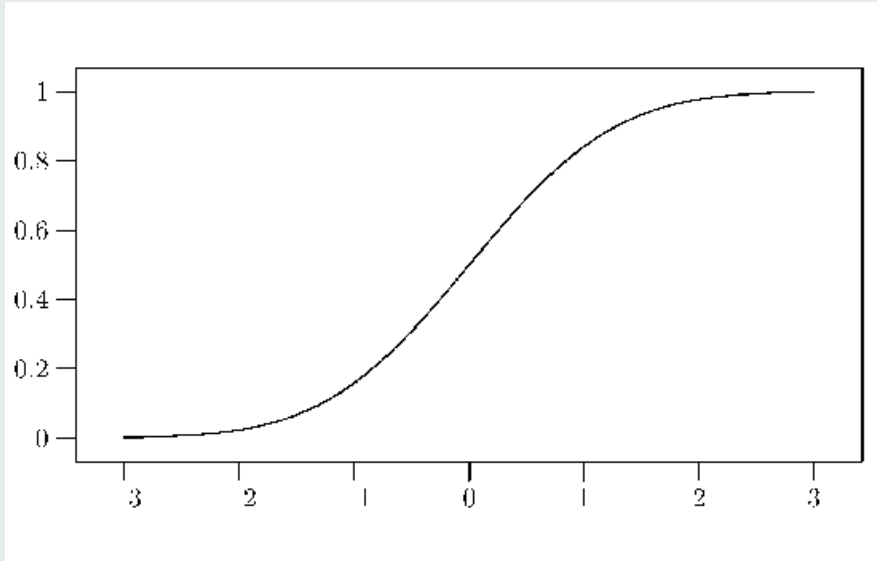&= \text{ceiling}[34.27] = 35
\end{aligned}
$$

Figure 8: Standard Normal Curve Cumulative Probability Function

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

6. 1-tailed and 2-tailed Tests

So far we have examined a situation where the null hypothesis would be rejected only on the basis of evidence that the parameter is on one side of the acceptance region. In some situations, however, the rejection region is two sided. The classic case is the null hypothesis of the form $\mu = \mu_0$. A value of $\overline{X}_\bullet$ far above μ_0 or far below μ_0 should result in rejection of the null hypothesis. Consequently, there will be two rejection regions. Such a hypothesis test is commonly called "two-sided" or "two-tailed."

Commonly, two-tailed tests have *symmetric rejection regions* in the sense that half the α is assigned to each tail. One consequence of having a two-tailed test as opposed to a one-tailed test is that the rejection points will be different. Consider a test with $\alpha = .05$. In the 1-tailed version, the rejection point R for the Z-statistic is either 1.645 *or* -1.645, but not both. In the 2-tailed version, the rejection points are both -1.96 and $+1.96$. The bottom line is that if the null hypothesis is false, and if you run a 1-tailed test with the rejection region on the correct side, you will have greater power, and a smaller required N, because R will be less.

We can revise our formula for required N to take into account T, the number of tails, as follows

$$N = \text{ceiling}\left[\left(\frac{\Phi^{-1}(P) + \Phi^{-1}(1 - \alpha/T)}{E_s}\right)^2\right] \tag{4}$$

Home Page

Print

Title Page

◀◀ ▶▶

◀ ▶

Page 35 of 37

Go Back

Full Screen

Close

Quit

The equation for power becomes

$$
\begin{aligned}
P &= \Phi\left[\sqrt{N}E_s - \Phi^{-1}\left(1 - \alpha/T\right)\right] \quad (5) \\
&= \Phi\left[\sqrt{N}\frac{\mu - \mu_0}{\sigma} - \Phi^{-1}\left(1 - \alpha/T\right)\right] \quad (6)
\end{aligned}
$$

Topics for this Module

An Idealistic Special...

Confidence Interval...

Hypothesis Testing

Test Statistics and...

1-tailed and 2-tailed...

Influences on Power

7. Influences on Power

By examining Equations 4 and 6 we can see more clearly the factors that influence power, and, in some cases, allow us to manipulate it. Since both $\Phi()$ and $\Phi^{-1}()$ are strictly increasing in their arguments, it follows that anything that increases the argument within brackets in Equation 4 will increase the required sample size, and anything that increases the argument within brackets in Equation 6 will increase power. We will discuss these in class

1. Sample Size (N)

2. Effect Size $(\mu - \mu_0)$

3. Variation (σ)

4. Type I Error Rate (α)

5. Number of Tails in the Rejection Region (T)