Hypothesis Testing

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Hypothesis Testing

Introduction

- In the preceding lecture, we learned that, over repeated samples, the sample mean M based on n independent observations from a population with mean μ and variance σ^2 has a distribution that, under fairly general conditions, can be assumed to have a normal distribution with a mean of μ and a standard deviation (called the "standard error of the mean") equal to $\sigma/\sqrt{n}$.
- We saw that we can view the standard error of the mean as a kind of "noise factor" in estimation process. Increasing n may cost us more money and/or effort in the short run, but it also increases the long run accuracy of the estimation process (and probably the short run accuracy as well, although that can never be guaranteed).

Introduction

- In the preceding lecture, we learned that, over repeated samples, the sample mean M based on n independent observations from a population with mean μ and variance σ^2 has a distribution that, under fairly general conditions, can be assumed to have a normal distribution with a mean of μ and a standard deviation (called the "standard error of the mean") equal to $\sigma/\sqrt{n}$.
- We saw that we can view the standard error of the mean as a kind of "noise factor" in estimation process. Increasing n may cost us more money and/or effort in the short run, but it also increases the long run accuracy of the estimation process (and probably the short run accuracy as well, although that can never be guaranteed).

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Introduction

- Now we want to take a big step forward and see how this result leads to the procedures that we use to test our theoretical hypotheses with data.
- The first thing we need to realize is that hypothesis testing is a special case of decision-making under uncertainty.
- We are going to gather data that we hope will give us an accurate picture of the world.
- We are going to use those data to make a decision.
- We know in advance that the decision might be wrong.
- What are the things that can happen?

Dichotomous Decision-Making Under Uncertainty

- The classic approach to hypothesis testing requires us to decide which of two possible decisions is the correct one.
- There are plenty of "real world" decisions where one must make a choice between two courses of action, or two states of the world, on the basis of information in the presence of uncertainty.
- Can you give me some examples?

Dichotomous Decision-Making Under Uncertainty

- The classic approach to hypothesis testing requires us to decide which of two possible decisions is the correct one.
- There are plenty of "real world" decisions where one must make a choice between two courses of action, or two states of the world, on the basis of information in the presence of uncertainty.
- Can you give me some examples?

Dichotomous Decision-Making Under Uncertainty

- The classic approach to hypothesis testing requires us to decide which of two possible decisions is the correct one.
- There are plenty of "real world" decisions where one must make a choice between two courses of action, or two states of the world, on the basis of information in the presence of uncertainty.
- Can you give me some examples?

Dichotomous Decision-Making Under Uncertainty

- Invest in stocks vs. Invest in bonds
- You are pregnant vs. You are not pregnant
- Build the reservoir at location A vs. Build the reservoir at location B
- Operate on the tumor vs. Use chemotherapy, watch, and wait
- Buy Chrysler stock vs. Sell Chrysler stock

Dichotomous Decision-Making Under Uncertainty

- Invest in stocks vs. Invest in bonds
- You are pregnant vs. You are not pregnant
- Build the reservoir at location A vs. Build the reservoir at location B
- Operate on the tumor vs. Use chemotherapy, watch, and wait
- Buy Chrysler stock vs. Sell Chrysler stock

Dichotomous Decision-Making Under Uncertainty

- Invest in stocks vs. Invest in bonds
- You are pregnant vs. You are not pregnant
- Build the reservoir at location A vs. Build the reservoir at location B
- Operate on the tumor vs. Use chemotherapy, watch, and wait
- Buy Chrysler stock vs. Sell Chrysler stock

Dichotomous Decision-Making Under Uncertainty

- Invest in stocks vs. Invest in bonds
- You are pregnant vs. You are not pregnant
- Build the reservoir at location A vs. Build the reservoir at location B
- Operate on the tumor vs. Use chemotherapy, watch, and wait
- Buy Chrysler stock vs. Sell Chrysler stock

Dichotomous Decision-Making Under Uncertainty

- Invest in stocks vs. Invest in bonds
- You are pregnant vs. You are not pregnant
- Build the reservoir at location A vs. Build the reservoir at location B
- Operate on the tumor vs. Use chemotherapy, watch, and wait
- Buy Chrysler stock vs. Sell Chrysler stock

Dichotomous Decision-Making Under Uncertainty

- These situations share a common framework:
  1. There are two possible states of the world
  2. You are not sure which is the true state
  3. You make a decision in favor of one state or the other
- In such a situation, one of two things can happen.

Dichotomous Decision-Making Under Uncertainty

- These situations share a common framework:
  1. There are two possible states of the world
  2. You are not sure which is the true state
  3. You make a decision in favor of one state or the other
- In such a situation, one of two things can happen.

Dichotomous Decision-Making Under Uncertainty

- These situations share a common framework:
  1. There are two possible states of the world
  2. You are not sure which is the true state
  3. You make a decision in favor of one state or the other
- In such a situation, one of two things can happen.

Dichotomous Decision-Making Under Uncertainty

- These situations share a common framework:
  1. There are two possible states of the world
  2. You are not sure which is the true state
  3. You make a decision in favor of one state or the other
- In such a situation, one of two things can happen.

Dichotomous Decision-Making Under Uncertainty

- These situations share a common framework:
  1. There are two possible states of the world
  2. You are not sure which is the true state
  3. You make a decision in favor of one state or the other
- In such a situation, one of two things can happen.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- 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.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- 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.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- 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.

Dichotomous Decision-Making Under Uncertainty

An Example – 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.
- Here is a picture of the situation. The blue distribution, for the non-infected people, has a mean of -2 and the red distribution for infected people has a mean of 2. Both distributions have a standard deviation of 1.

Dichotomous Decision-Making Under Uncertainty

An Example – 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.
- Here is a picture of the situation. The blue distribution, for the non-infected people, has a mean of -2 and the red distribution for infected people has a mean of 2. Both distributions have a standard deviation of 1.

Dichotomous Decision-Making Under Uncertainty

An Example – 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.
- Here is a picture of the situation. The blue distribution, for the non-infected people, has a mean of -2 and the red distribution for infected people has a mean of 2. Both distributions have a standard deviation of 1.

Dichotomous Decision-Making Under Uncertainty

An Example – 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.
- Here is a picture of the situation. The blue distribution, for the non-infected people, has a mean of -2 and the red distribution for infected people has a mean of 2. Both distributions have a standard deviation of 1.

Dichotomous Decision-Making Under Uncertainty

An Example – 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.
- Here is a picture of the situation. The blue distribution, for the non-infected people, has a mean of -2 and the red distribution for infected people has a mean of 2. Both distributions have a standard deviation of 1.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

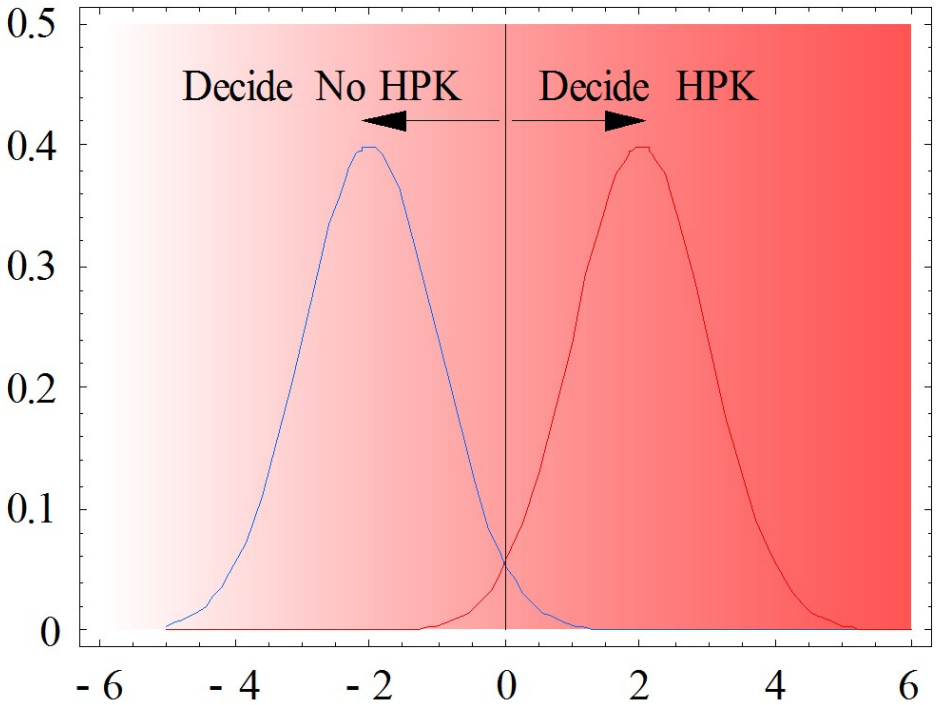


Figure 1. Frequency Distributions for an imperfect indicator of HPK. The distribution on the left, drawn in blue, represents the relative frequency of occurrence of colors for people who do not have HPK. The distribution on the right, drawn in red, represents the relative frequency of occurrence of test 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.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- Going into the diagnostic process, you know that you have to "draw the line" somewhere.
- You will see a color. You will not know for certain what that color means.
- Where you "draw the line" has implications.
- In our diagram, the line has been drawn at the color value of 0, corresponding to a moderate level of pink.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- Going into the diagnostic process, you know that you have to "draw the line" somewhere.
- You will see a color. You will not know for certain what that color means.
- Where you "draw the line" has implications.
- In our diagram, the line has been drawn at the color value of 0, corresponding to a moderate level of pink.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- Going into the diagnostic process, you know that you have to "draw the line" somewhere.
- You will see a color. You will not know for certain what that color means.
- Where you "draw the line" has implications.
- In our diagram, the line has been drawn at the color value of 0, corresponding to a moderate level of pink.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- Going into the diagnostic process, you know that you have to "draw the line" somewhere.
- You will see a color. You will not know for certain what that color means.
- Where you "draw the line" has implications.
- In our diagram, the line has been drawn at the color value of 0, corresponding to a moderate level of pink.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

There are 4 possibilities, summarized in the table below:

	State of the World	
Test Decision	*Patient is Infected*	*Patient is Not Infected*
Positive	Correct Positive	False Positive
Negative	False Negative	Correct Negative

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

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- 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.
- But you know how to calculate this probability, don't you! ?

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

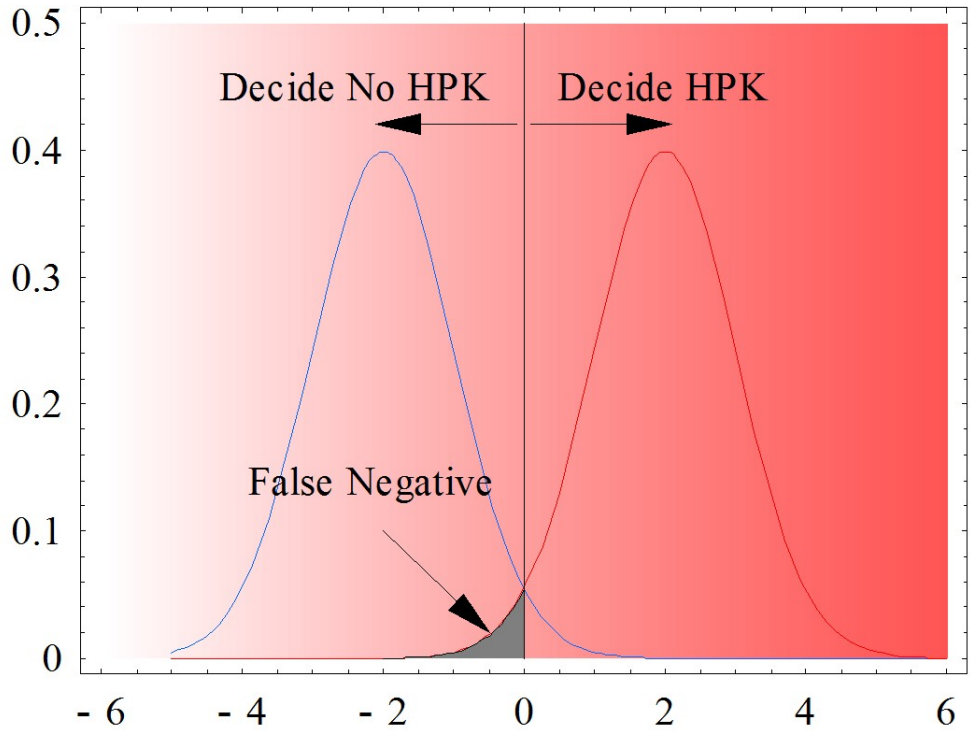


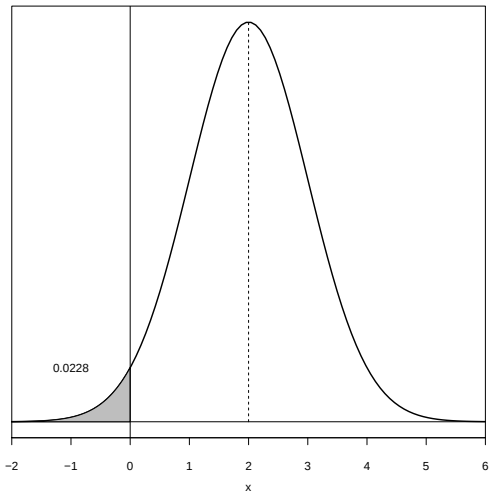
Figure 2. Frequency distributions for an imperfect indicator of HPK, with the False Negative probability (.0228) shaded in. This shaded area represents the False Negative cases, i.e., those cases that occur under the red graph (representing the distribution of test results for people who have HPK), and 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.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

```
> pnorm(0,2,1)

[1] 0.02275013
```

```
> nc(2,1)
> cn(-4,0,2,1,x=-1)
> abline(v=0)
```



Dichotomous Decision-Making Under Uncertainty

An Example – HPK

- 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, from 0 to -1.
- Below is a picture of the changed decision rule, with the area representing the probability of a False Positive shaded in.
- Note that, although there are two distributions drawn on the graph, only one can be "true."
- We need to consider the two distributions *alternately, not at the same time*, even though it is common to draw both of them on the same graph.
- It is essential that you realize this point!

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

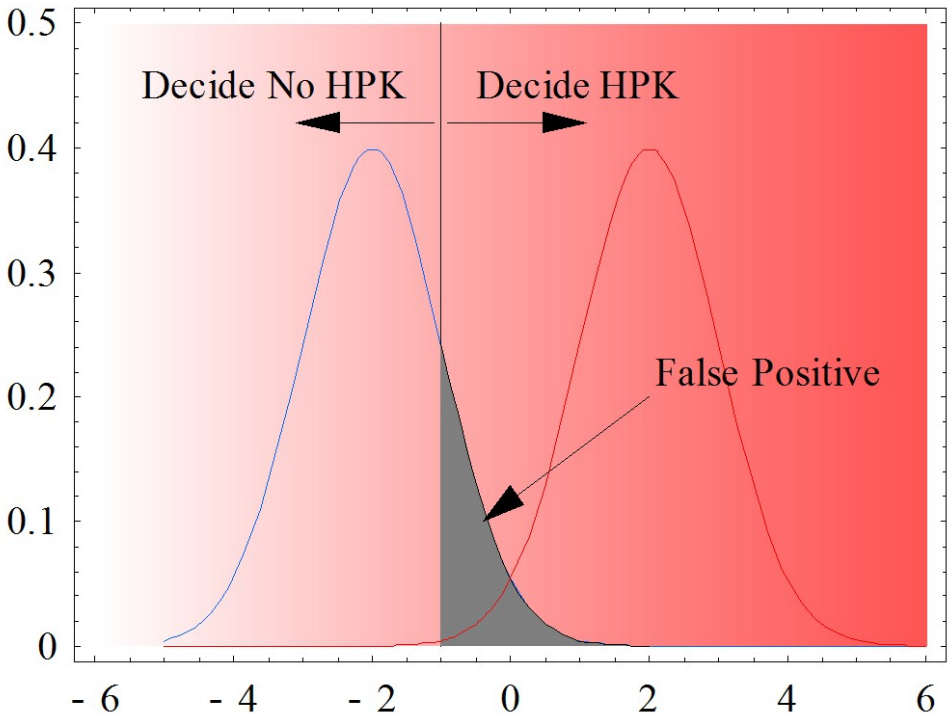
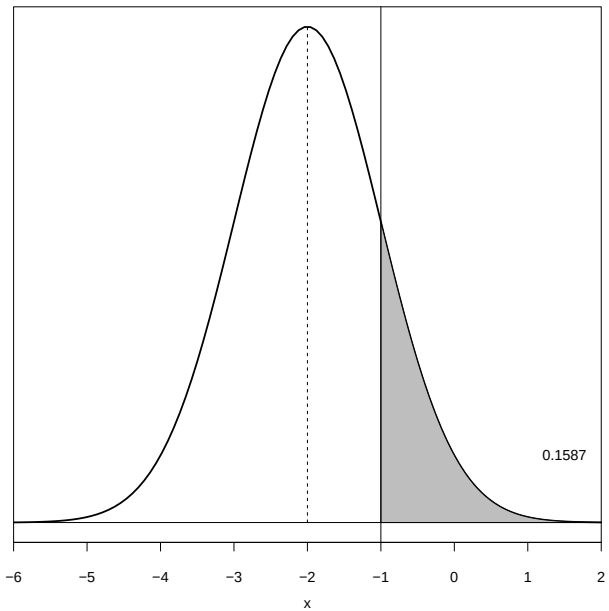


Figure 3. Frequency distributions for an imperfect indicator of HPK, with the decision point altered to virtually eliminate False Negatives. This was accomplished by moving the decision point to the left. The result is that the False Positive rate has increased from .0228 to .1587, while the False negative rate is almost zero.

Dichotomous Decision-Making Under Uncertainty

An Example – HPK

```
> nc(-2,1)
> cn(-1,4,-2,1)
> abline(v=-1)
```



Dichotomous Decision-Making Under Uncertainty

An Example – HPK

What have we learned so far?

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

Statistical Hypothesis Testing

The Basic Setup

- Let's suppose (this is a highly artificial example) that we wanted to test whether the average IQ score of a population of students known to have received a certain kind of post-natal vitamin regimen differs from the known average of $\mu = 100$ in the general population.
- It is possible that the group of students has an average IQ lower than the average, and it is also possible that the group of students has an IQ that is above the average.
- We will assume furthermore that if the special vitamin regimen has any effect (positive or negative), this effect is *additive*, in the sense that it simply displaces IQ scores by a constant, rather than multiplying them by a constant.
- The result of this assumption is that, regardless of the effect of the vitamin regimen, the population standard deviation of IQ scores in the treated group is assumed to be 15, the same as in the general population.

Statistical Hypothesis Testing

The Basic Setup

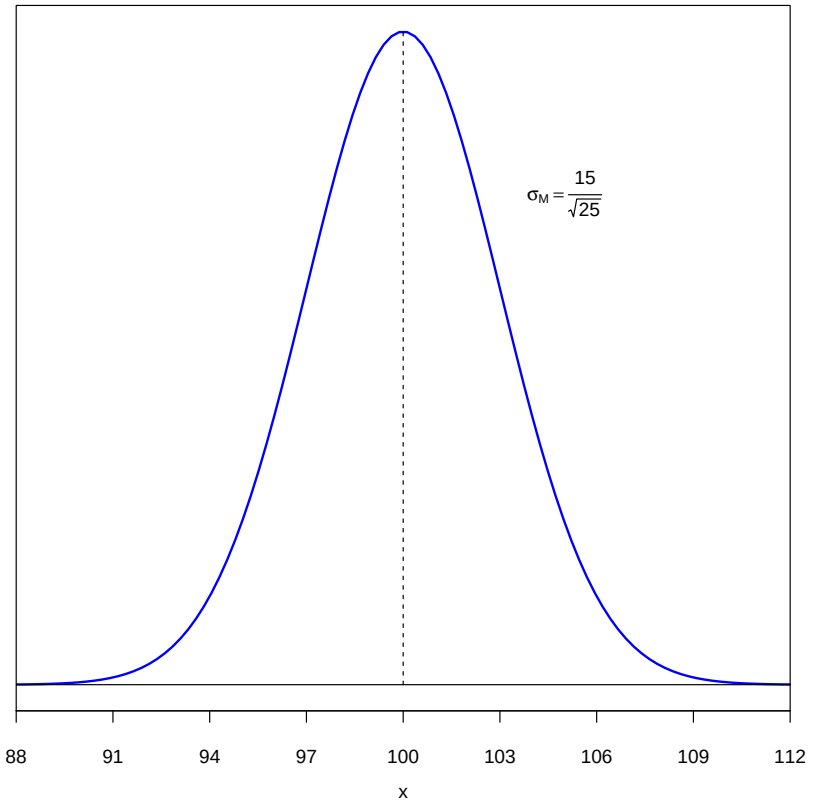
- We take a random sample of $n = 25$ students from this special group, and compute the sample mean M.
- What do we know about the behavior of the sample mean over repeated samples?
- Suppose the post-natal treatment had no effect. Then the mean of the sample means would remain at $\mu = 100$.
- The standard error of the mean is

$$\sigma_M = \frac{\sigma}{n} = \frac{15}{\sqrt{25}} = 3$$

- So the sample mean would have a distribution like the picture in the next slide.

Statistical Hypothesis Testing

The Basic Setup



Statistical Hypothesis Testing

The Basic Setup

- We see that, just like with the medical testing situation, the information we get from our indicator M is imperfect.
- *Regardless of what the value of* μ *is*, the sample mean M will, in the long run, include "noise" mixed in with the signal. As we can see from the plot, most of the time the sample mean M will be within about 6 points of μ, but a fair percentage of the time it will be off by at least 3 points.

Statistical Hypothesis Testing

The Null and Alternative Hypotheses

- Suppose we had to decide between two possible states of the world.
  1. H_0, the *statistical null hypothesis*, states that $\mu = 100$. That is, there is no difference between our special group and the general population.
  2. H_1, the *alternative hypothesis*, states that $\mu \neq 100$.
- These two hypotheses are *mutually exclusive and exhaustive*. One has to be true, and both cannot be true.

Statistical Hypothesis Testing

Reject-Support Testing

- In most research in education and psychology, the statistical null hypothesis is the opposite of what the experimenter actually wishes to show.
- Rejecting the null hypothesis actually supports the experimenter's belief.
- Consequently, the approach is sometimes called "Reject-Support" testing.
- In Reject-Support testing, a Type-I error is a false positive for the experimenter's belief, while a Type-II error is a false negative with respect to the experimenter's belief.

Statistical Hypothesis Testing

Accept-Support Testing

- Occasionally, researcher's want to "accept the null" in order to support their belief system.
- Some researchers have attempted to employ the standard statistical setup in order to "prove the null."
- Such Accept-Support testing turns the standard conventions around, and in general works very poorly. If you want to demonstrate that something has no effect, you need special methods to do it, and these methods are not discussed in your text.
- For now, you are well advised to avoid Accept-Support testing. In particular, if someone asserts that two groups are equal because they "did not find a significant difference," you should be very skeptical.

Statistical Hypothesis Testing

The Null and Alternative Hypotheses

- Suppose that, just as in the medical testing example, we "draw the line," that is, produce a completely objective rule for deciding, on the basis of the value of M that we get from our sample of $n = 25$, whether to favor H_0 or H_1.
- Let's assume we have such a rule.
- Before we ever discuss what that rule is, or how we might derive it, we can say that several possibilities exist.
- They are summarized in the table on the next slide.

Statistical Hypothesis Testing

The 2×2 Table

	State of the World	
Test Decision	H_0 is True	H_0 is False
Accept H_0	Correct Acceptance $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Rejection $(1-\beta = \text{Power})$

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.

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- We begin with an informal definition of our decision rule.
- The null hypothesis specifies that $\mu = 100$
- We will observe a sample mean M.
- If M is sufficiently different from 100, we will reject the null hypothesis.

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- But how far is far enough?
- Suppose we chose a point 5.88 units away from 100 in either direction (i.e., less than 94.12 or greater than 105.88. (Don't worry yet about how I came up with the 5.88).
- If M is farther away from 100 than either of these two points, then we reject H_0.
- What would be the probability of a Type I Error?

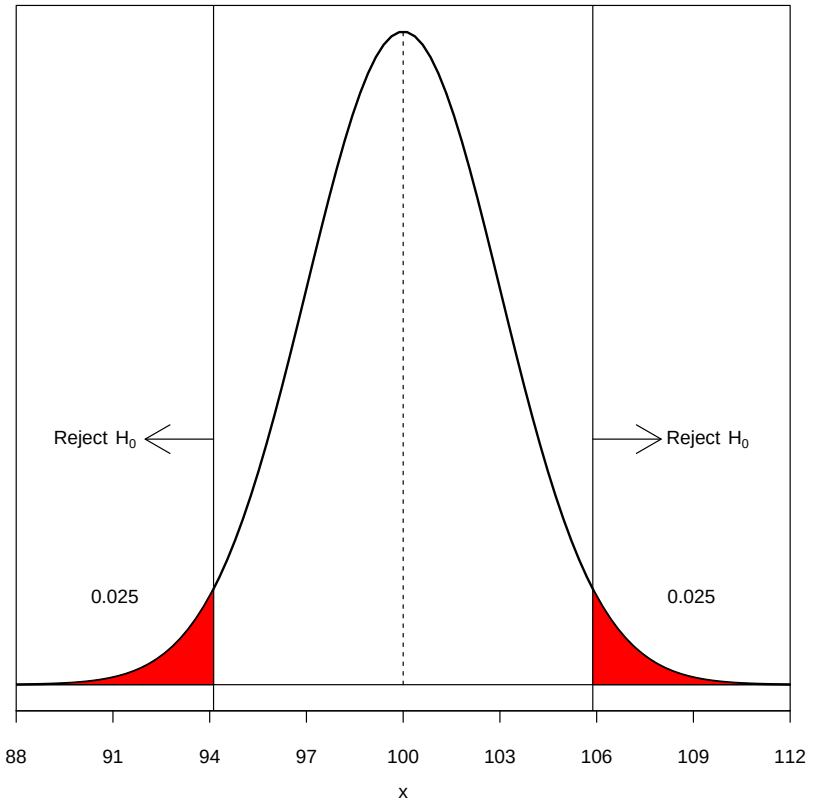
Statistical Hypothesis Testing

Drawing the line — Decision Regions

- Remember, a Type I Error can only occur if the null hypothesis is true.
- If the null hypothesis is true, $\sigma = 15$, and $n = 25$, the sample mean M has a normal distribution with a mean of 100 and a standard deviation of 3.
- So our question becomes a normal curve problem.
- On the next slide, I diagram the normal curve, the rejection regions, with the Type I error probability shown in red. The rejection regions are often referred to as "critical regions" and the rejection points as "critical values."
- When there are two rejection regions at either end of the distribution, the test is called "two-sided" or "two-tailed."

Statistical Hypothesis Testing

Drawing the line — Decision Regions



Statistical Hypothesis Testing

Drawing the line — Decision Regions

- We can see that the total probability of a Type I Error is, in this case, 0.05.
- Here is the R calculation.

```
> area.below <- pnorm(94.12,100,3)
> area.above <- 1 - pnorm(105.88,100,3)
> Type.I.Error.Rate <- area.below + area.above
> round(Type.I.Error.Rate,5)

[1] 0.05
```

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- That probability, 0.05, is one of the standard values used in psychological research.
- But where did I come up with critical values of 94.12 and 105.88?

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- I obviously didn't snatch the critical values out of thin air.
- They were chosen so that I would have an α of 0.05, the the probability of a rejection split symmetrically into 0.025 in each of the two tails.

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- I *worked backwards* from the fact that for the two-tailed test to have a total α of 0.05 and be "balanced," it needs to have an area of 0.025 in each tail.
- Look at the upper tail first.
- To have an area above the critical value of 0.025, the area under the curve below the critical value must be 0.025, so the critical value must have a percentile value of 97.5.

```
> qnorm(0.975,100,3)

[1] 105.8799
```

Statistical Hypothesis Testing

Drawing the line — Decision Regions

Similarly, the lower critical value must be at

```
> qnorm(0.025,100,3)

[1] 94.12011
```

Statistical Hypothesis Testing

Drawing the line — Decision Regions

- Suppose you wanted the α to be 0.01, rather than 0.05.
- With everything else remaining the same, what would you have to push your critical values out to in order to reduce the total α to be 0.01.
- *Hint*: What probability would have to be in each tail of the sampling distribution?

The Z-Statistic

Why We Need It

- With modern software like R, "drawing the line" in terms of a critical value of M, the sample mean, is easier than it used to be.
- But it *still usually wastes time*, because each time you entertain a different sample size, or address a new situation, you need to compute a new critical value.
- For example, in the situation in which the null hypothesis is thT $\mu = 100$, $\sigma = 15$, but now $n = 100$, what would be the new critical values for an *alpha* of 0.05?

The Z-Statistic

Why We Need It

- First, we realize that the standard error of the mean, σ_M, has changed. It was 3, but now it is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$$

- We quadrupled n so we halved σ_M and doubled our precision.
- So we can make our rejection points (critical values) half as far away from 100 as they were.

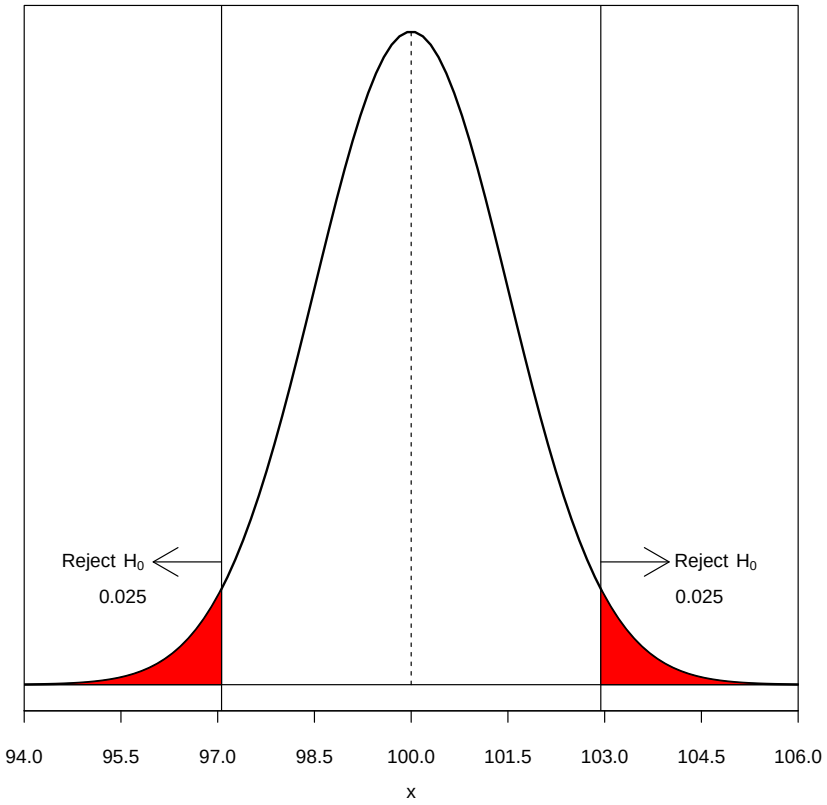
The Z-Statistic

Why We Need It

```
> qnorm(.975,100,1.5)

[1] 102.9399

> qnorm(.025,100,1.5)

[1] 97.06005
```

The Z-Statistic

Why We Need It



The Z-Statistic

Why We Need It

- Wouldn't it be easier if you could have the same rejection point every time you ran this test with $\alpha = 0.05$, two-tailed?
- People realized a long time ago that you could make the process simpler.
- Why? Because although the raw score rejection point for M changes, the $Z - Score$ value stays the same.
- For example, we worked two problems in which *alpha* was 0.05, two-tailed.
- In one case, the rejection point was 105.88 in a normal distribution with a mean of 100 and a standard deviation of 3. That's a Z score of ??
- In the other case, the rejection point was 102.94 in a normal distribution with a mean of 100 and a standard deviation of 1.5. That's a Z score of ??.

The Z-Statistic

Why We Need It

- In both cases, the Z-score value was 1.96. In the early days of statistics, people realize that if they used the *test statistic*

$$Z = \frac{M - \mu_0}{\sigma_M} = \frac{M - \mu_0}{\sigma/\sqrt{n}} \tag{1}$$

  then they would only have to memorize a few "magic numbers" from the normal curve. (*Note*: I use the notation μ_0 to stand for the null-hypothesized value of μ.)
- For a two-sided test, the magic numbers are 1.96 for $\alpha = 0.05$, and 2.576 for $\alpha = 0.01$.
- So rather than compute critical values (rejection points) for M, and see if M is in the critical region, one computes a Z statistic and sees if the Z statistic exceeds one of the magic numbers in either the positive or negative direction.

The Z-Statistic

How it Works

- Let's go back to our first example. When $\sigma = 15$, $n = 25$, and the null hypothesis was that $\mu = 100$, we examine M and reject the null hypothesis if it is less than 94.12 or greater than 105.88.
- With the Z-statistic approach, we compute the Z statistic

$$Z = \frac{M - \mu_0}{\sigma_M} = \frac{M - \mu_0}{\sigma/\sqrt{n}} \tag{2}$$

  and reject the null hypothesis if Z is less than -1.96 or greater than 1.96.
- The two rejection rules are equivalent — it is easy to verify that the Z statistic just reaches 1.96 when M barely reaches 105.88, and the Z statistic just reaches -1.96 when M equals 94.12.
- The Z statistic rejection point, unlike the rejection point for M, remains the same if you change the sample size.

One-Tailed Hypothesis Tests

- So far, we've been talking about situations in which an M either substantially higher than μ_0 or an M substantially lower than μ_0 would be reason to reject H_0.
- In some situations, a researcher has enough information about the situation to make the hypothesis *directional*.
- For example, suppose you are trying to show that a vitamin treatment works, by elevating μ above 100. Your null hypothesis in the Reject-Support framework is that $\mu \leq 100$.
- In this case, only an M greater than 100 could provide a reason to reject the null hypothesis.
- So, to control α, we put our rejection point at the upper $1 - \alpha$ quantile of the sampling distribution.

One-Tailed Hypothesis Tests

Example (One-Tailed Test)

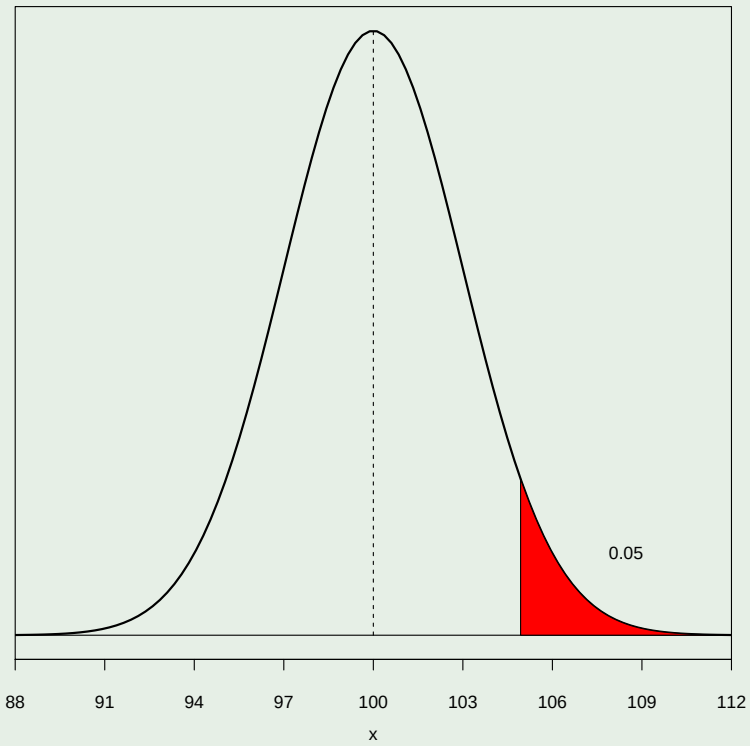
We wish to test the null hypothesis that $\mu \leq 100$, with $\sigma = 15$ and $n = 25$, with $\alpha = 0.05$. There is only one rejection point, in the upper tail of the sampling distribution. The rejection point for M is at the 0.95 quantile of the sampling distribution, which is

```
> qnorm(0.95,100,3)
[1] 104.9346
```

One-Tailed Hypothesis Tests

Example (One-Tailed Test (continued))

Here is a plot of the critical value and the rejection region:



A Flow Chart

In this course, we will be dealing with situations in which the null hypothesis and alternative hypothesis form mutually exclusive and exhaustive opposites. In such a case, we can construct a simple decision tree for deciding (1) whether our 1-sample Z test is 1-tailed or 2-tailed, and (2) how to set up the critical value(s).

- Is the null hypothesis $\mu = \mu_0$?
- If *Yes*, then
  1. The test is two-tailed. Evidence that μ is either larger or smaller than μ_0 is reason to reject the null hypothesis.
  2. The upper rejection point is at the $1 - \alpha/2$ quantile. The lower rejection point is at the $\alpha/2$ quantile.
- If *No*, then
  1. The test is 1-tailed.
  2. If the null hypothesis is that $\mu \leq \mu_0$, then only evidence that μ is above μ_0 on the number line can be cause for rejection. The critical value is in the upper tail, and is at the $1 - \alpha$ quantile.
  3. If the null hypothesis is that $\mu \geq \mu_0$, then only evidence that μ is below μ_0 on the number line can be cause for rejection. The critical value is in the lower tail, and is at the α quantile.

Statistical Power

- Power is the probability of getting a result in the rejection region when the null hypothesis is, in fact, false.
- In general, the null hypothesis can be wrong in infinitely many ways, and to different degrees.
- For example, It can be "barely false," or "overwhelmingly false."
- In general, all other things being equal, the more false the null hypothesis is, the larger the power is.
- Computing power requires you to specify a degree of falsity of the null hypothesis.

Statistical Power

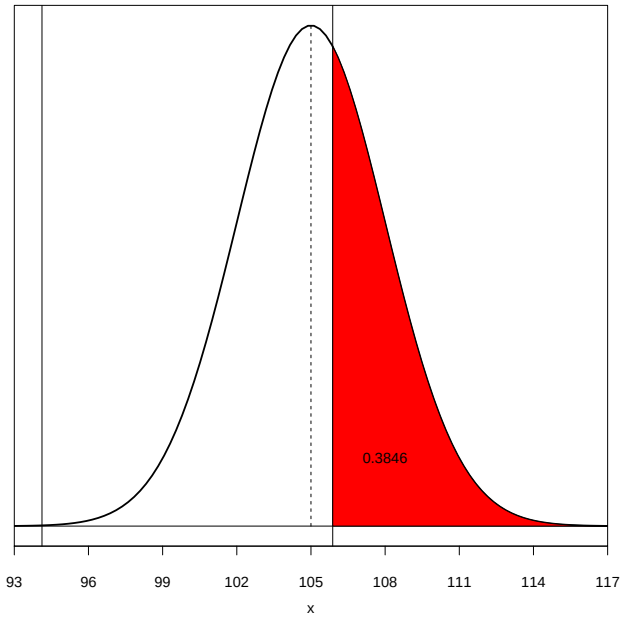
- For example, recall our test that $\mu = 100$, with $\sigma = 15$ and $n = 25$.
- We set up rejection points at 94.12 and 105.88 in order to control α at 0.05.
- What would the power be if the null hypothesis is false the vitamin regimen does have an effect, and the true μ is $\mu = 105$?

Statistical Power

- We draw the true distribution of M on a map of the rejection regions, and compute the probability.

Statistical Power

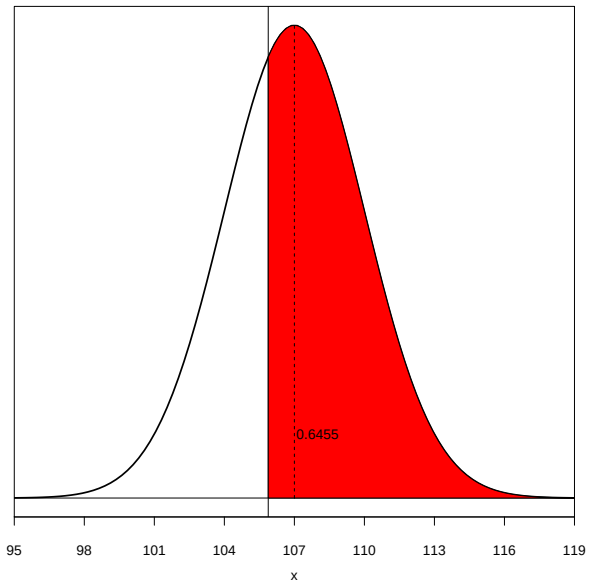
```
> nc(105,3)
> abline(v=105.88)
> abline(v=94.12)
> cn(105.88,117,105,3,,x=108,color="red")
```



Statistical Power

What would power be if $\mu = 107$?

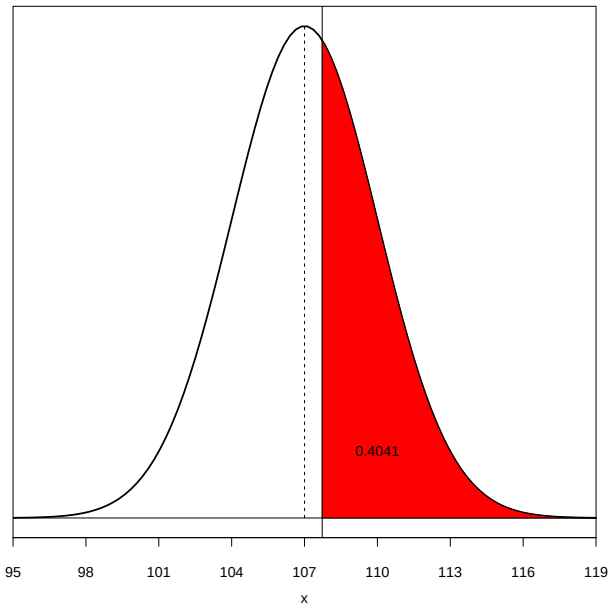
```
> nc(107,3)
> abline(v=105.88)
> abline(v=94.12)
> cn(105.88,119,107,3,,x=108,color="red")
```



Statistical Power

What would power be if $\mu = 107$, but α were set to 0.01?

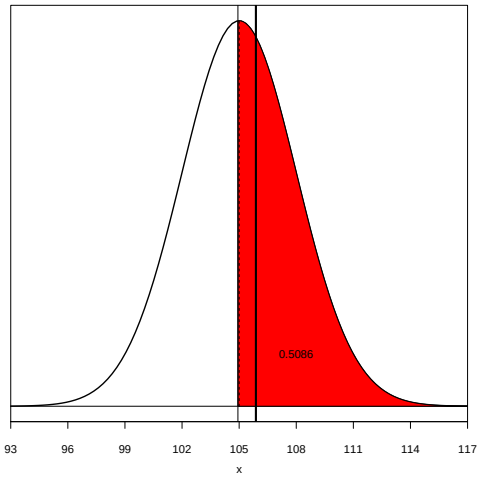
```
> nc(107,3)
> abline(v=100+2.576*3)
> cn(100+2.576*3,119,107,3,,x=110,color="red")
```



Statistical Power

What would power be if $\mu = 105$, α were set to 0.05, but we had a one-tailed test? Notice in the graph below that the critical value has moved to the left, because with a one-tailed test, it is at the 0.95 quantile, rather than the 0.975 quantlile. The old two-tailed critical value is also marked in bold on the plot below, so you can see how much power was gained by shifting to the one-tailed test.

```
> nc(105,3)
> cn(100+1.645*3,119,105,3,,x=108,color="red")
> abline(v=100+1.645*3)
> abline(v=105.88,lwd=3)
```



Statistical Power and Standardized Effect Size

- There is a simpler, more direct way to compute power for the 1-Sample Z test than the approach we took in the preceding section. It requires just a single equation, plus the introduction of a new concept
- We will not derive the equation, rather I'll simply present the equation and demonstrate its use on the same problems we worked by the standard method.

Statistical Power and Standardized Effect Size

Standardized Effect Size

- We saw in the preceding power calculations that an important factor influencing power is *Effect Size*, defined as the *the amount by which the null hypothesis is wrong.*

$$\text{Effect Size} = \mu - \mu_0 \tag{3}$$

- As μ moves past μ_0, and effect size increases, the distribution of M moves into the rejection region and power increases.

Statistical Power and Standardized Effect Size

Standardized Effect Size

- The *standardized effect size* E_s converts the Effect Size into standardized units by dividing by σ, i.e.,

$$E_s = \frac{\mu - \mu_0}{\sigma} \tag{4}$$

- Standardized effect size is also known as *Cohen's d*. It is *the amount by which the null hypothesis is wrong in standard deviation units.*

Statistical Power and Standardized Effect Size

Standardized Effect Size

- The *standardized effect size* E_s converts the Effect Size into standardized units by dividing by σ, i.e.,

$$E_s = \frac{\mu - \mu_0}{\sigma} \tag{5}$$

- Standardized effect size is also known as *Cohen's d*. It is *the amount by which the null hypothesis is wrong in standard deviation units.*

Statistical Power and Standardized Effect Size

Standardized Effect Size

- The standardized effect size has a very significant advantage over the unstandardized effect size — it is, in an important sense, *metric-free*.
- If you linearly rescale (change inches into centimeters, for example), the standardized effect size remains the same.
- Because the effect is standardized, it is possible to suggest general standards for evaluating it.

Statistical Power and Standardized Effect Size

Standardized Effect Size

- Cohen proposed the following standards for effect size:
  1. Small Effect — 0.20
  2. Medium Effect — 0.50
  3. Large Effect — 0.80

Statistical Power and Standardized Effect Size

Calculating Power

- Let T be the number of tails (either 1 or 2).
- Assume for simplicity that if the test is 1-tailed, the critical value of the Z statistic is positive, and that the effect is in the direction for rejection. (There is not much point computing power for a 1-tailed test if the effect is in the wrong direction.)
- $E_s = d = (\mu - \mu_0)/\sigma$ is the standardized effect size, n the sample size
- $\Phi()$ is the normal distribution cumulative probability function (`pnorm` in R), and $\Phi^{-1}()$ the normal curve quantile function (`qnorm` in R).
- Then power is calculated as

$$\text{Power} = \Phi\left(\sqrt{n}E_s - Z_{crit}\right) \tag{6}$$

- Z_{crit} is the critical value for the Z-test, and is calculated as

$$Z_{crit} = \Phi^{-1}\left(1 - \alpha/T\right) \tag{7}$$

Statistical Power and Standardized Effect Size

Calculating Power

- To employ the equations, calculate the critical value first. For a given α and number of tails (T), it will not change.
- Suppose you are doing a two-tailed test that $\mu = 100$ with $\alpha = 0.05$. The critical value is

```
> alpha <- 0.05
> T <- 2
> Z.crit <- qnorm(1 - alpha/T)
> Z.crit
[1] 1.959964
```

Statistical Power and Standardized Effect Size

Calculating Power

- Next, calculate the standardized effect size.
- In some cases, you will be given μ_0, and σ, and will use them to calculate E_s, for some hypothetical value of μ.
- For example, if $\mu_0 = 100$, $\sigma = 15$, what is power if the true μ is 107 and $n = 25$?
- The standardized effect size is $E_s = (107 - 100)/15 = 7/15$. The full calculation is shown below.

```
> Es = (107-100)/15
> Es
[1] 0.4666667
> n<-25
> Power <- pnorm(sqrt(n)*Es - Z.crit)
> Power
[1] 0.6455632
```

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Power of 0.646 is generally not considered to be adequate. In most applications, power of 0.80 is considered minimal and 0.90 a reasonable target.
- From the calculation on the previous slide, it is clear that increases in sample size will increase power, since the square root of n is multiplied by Es in the formula.
- But how large an n do we need?
- With a bit of manipulation, you can produce a formula to calculate the sample size that will produce a given level of power for a given standardized effect size, α, and testing situation.

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Required n can be calculated as

$$n = \left(\frac{\Phi^{-1}(1 - \alpha/T) + \Phi^{-1}(\text{Power})}{E_s} \right)^2 \quad (8)$$

$$= \left(\frac{Z_{crit} + \Phi^{-1}(\text{Power})}{E_s} \right)^2 \quad (9)$$

$$= \left(\frac{Z_{crit} + Z_{power}}{E_s} \right)^2 \quad (10)$$

- This formula is much simpler than it appears at first glance. The numerator is the square of the sum of two normal curve values.
- One is the critical value for the Z-test (1.96 in this case), the other the value corresponding to power. In this case, the 0.90 quantile of the normal curve corresponding to power is 1.282.

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- The numerator is $1.96 + 1.28 = 3.24$, so the required n will be the square of the ratio of $3.24/E_s$. For example, if $E_s = 7/15$, required n is

```
> Power <- 0.90
> Z.power <- qnorm(Power)
> n <- ((Z.crit + Z.power) /(7/15))^2
> n
[1] 48.24837
```

- Should we use 48.24 as our sample size? Or should we use 49? (It won't make too much difference, but there is a convention in use, and we discuss it in the following slides.)

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- What will the required n be if the standardized effect size is "small" according to Cohen's standards?
- Since the small effect is $0.20 = 1/5$, dividing 3.24 by the effect size of $1/5$ is the same as multiply it by 5. so the required n should be the square of 16.2 or about 263.
- More precisely

```
> Z.power <- qnorm(0.90)
> ((Z.crit + Z.power) / (0.2))^2
[1] 262.6856
```

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Since sample size 0f 262.7 is required to get power of 0.90, an n of 262 will yield power of slightly less than 0.90. Since n must be an integer, we "bump" the value up to $n = 263$ to guarantee power greater than 0.90. This act of moving a number n with a decimal fraction to the smallest integer that is greater than or equal to n is called the *ceiling* function, and is part of R.

```
> E.s <- 0.5
> ceiling( ((Z.crit + Z.power) /E.s )^2  )
[1] 43
```

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- To summarize:
  1. Compute two normal curve values, one, Z_{crit}, corresponding to (the absolute value of) the critical value for the Z-test, the other, Z_{power} corresponding to the desired level of power.
  2. Add the two values together.
  3. Divide by the standardized effect size.
  4. Square the result.
- Let's do an example, and then let's try another, graphical approach to get the same answer.

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Suppose the standardized effect size you are anticipating in your experiment is a "medium effect" of $E_s = 0.50$, and you decide to run a 2-sided hypothesis test with $\alpha = 0.01$.
- How big a sample size n will you need to guarantee power of at least 0.95?

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- The normal curve value corresponding to an alpha of 0.01 is at the $1 - \alpha/2 = 0.995$ quantile. This is one of our "magic numbers" (2.576) from the normal curve.

```
> Z.crit <- qnorm(.995)
```

- The normal curve corresponding to desired power is

```
> Z.power <- qnorm(0.95)
```

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Adding these two and squaring, we get

```
> Z.crit + Z.power
```

```
[1] 4.220683
```

- To calculate the required n, we divide this value by the standardized effect size, then square the result.

```
> E.s <- 0.5
> ceiling( ((Z.crit + Z.power)/Es)^2 )
```

```
[1] 82
```

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

- Most hypothesis tests use one of just a few "magic numbers" from the normal curve.
- Most hypothetical sample size calculations use one of a few common "target values" of power, specifically, 0.80, 0.90, 0.95, and a few "target values" of E_s, like 0.20, 0.50, 0.80.
- The Z_{power} values corresponding to power of 0.80, 0.90, and 0.95 are, respectively, 0.842, 1.282, and 1.645.
- With a little bit of practice, you can get very fast at computing the required n for the 1-Sample Z-statistic, even without using R.
- With R, it is a snap. In my experience, students (and profs) tend to make the most errors by
  1. Mixing up the 1-tailed Z_{crit} values with the 2-tailed values.
  2. Forgetting to square the quotient as a final step, thereby ending up with an n of 8 instead of, say, 64.
  3. Miscalculating E_s in problems where you are given μ, μ_0, and σ and are required to calculate E_s as a first step.
- So, be careful!

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

Example (Test Yourself – Sample Size Estimation)

Suppose you wish to test a 1-Sided Hypothesis with $\alpha = 0.05$. What sample size would you need to detect a standardized effect size of $E_s = 0.4$ with power of 0.80?

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

Example (Test Yourself – Sample Size Estimation)

Suppose you wish to test a 1-Sided Hypothesis with $\alpha = 0.05$. What sample size would you need to detect a standardized effect size of $E_s = 0.40$ with power of 0.80?

Answer. We need two values from the normal curve, $Z_{crit} = \Phi^{-1}(1 - \alpha)$ (the one-tailed critical value for the Z-statistic) and $Z_{power} = \Phi^{-1}(\text{Power})$, the normal curve quantile corresponding to the desired power. These values are

```
> Z.crit <- qnorm(1-.05)
> Z.crit
[1] 1.644854
> Z.power <- qnorm(0.80)
> Z.power
[1] 0.8416212
```

(continued on next slide . . .)

Statistical Power and Standardized Effect Size

Calculating Required Sample Size

Example (Test Yourself – Sample Size Estimation (ctd))

Answer(continued). Using the values from the preceding slide, required n is then calculated as

$$\begin{aligned} n &= \text{ceiling}\left(\left(\frac{(Z_{crit} + Z_{power})}{E_s}\right)^2\right) \\ &= \text{ceiling}\left(\left(\frac{(1.6449 + 0.8416)}{0.4}\right)^2\right) \\ &= \text{ceiling}\left(\left(\frac{2.4865}{0.40}\right)^2\right) \\ &= \text{ceiling}(38.64) \\ &= 39 \end{aligned}$$