



Chapter 9

Hypothesis Testing

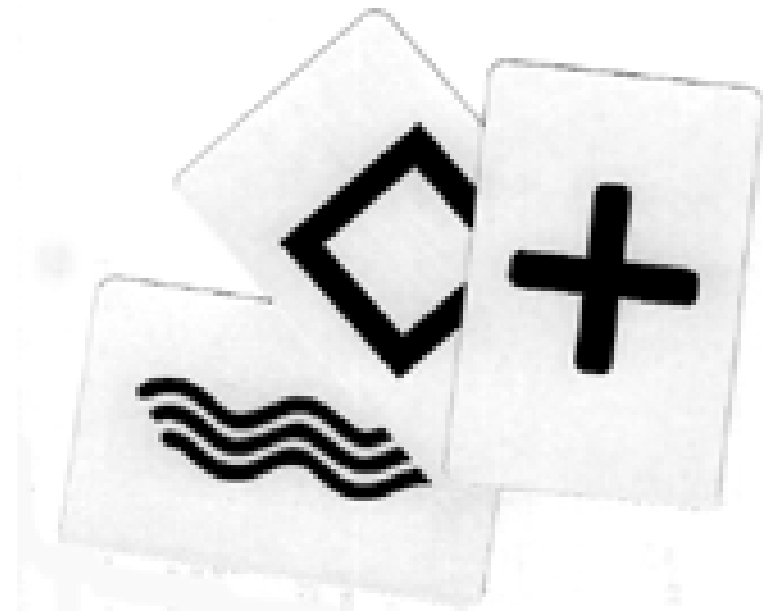


Testing vs. Estimation

- ◆ Confidence Intervals are used to estimate an unknown population parameter using sample data.
- ◆ A Hypothesis test is used when theory suggests a particular value for an unknown population parameter. The question then becomes whether the sample data is consistent with the theory or not.

A Theory about the value of a population parameter?

- ◆ One popular test of ESP uses what is called a Rhine deck, which consists of cards with five different images on them.
- ◆ The cards are put in random order, and the subject must guess the card that has been drawn without seeing it.





The Rhine Deck and ESP

- ◆ The population parameter you are investigating is the proportion of correct answers in an infinitely long sequence of guesses.
- ◆ The sample statistic is the sample proportion of correct answers in a fixed number of guesses



The Theory



- ◆ There is one obvious theory here: Nothing is going on and the person is just guessing. By guessing the person should get 20% correct. This is the value of the population parameter we wish to test.
- ◆ Given sample data, such as 25 correct in 100 trials, the question becomes: Is this sample evidence enough to disprove the theory that the person is just guessing?

The Null Hypothesis

- ◆ The theoretical value of the unknown population parameter we wish to test is called *the null hypothesis*.
- ◆ H_0 is the universal notation for a null hypothesis.
- ◆ Often the null corresponds to the idea “nothing interesting or unexpected is going on.”

$$H_0 : p = p_0$$

In this application,

$$H_0 : p = .20$$

The Alternative Hypothesis

- ◆ The alternative hypothesis comes in three flavors

- Less than
- Not equal to (two-sided)
- Greater than

$$H_A : p < p_0$$

- ◆ H-sub-A signifies the alternative hypothesis

$$H_A : p \neq p_0$$

- ◆ Which flavor is a matter of judgment: What do you expect to be true if the null is false?

$$H_A : p > p_0$$

ESP and the Alternative Hypothesis

- ◆ If you believe that ESP (if it exists) will manifest itself in the person getting more than 20% correct, you'd do a greater than test.

$$H_0 : p = .20$$

$$H_A : p > .20$$

- ◆ This leaves the possibility $p < .20$ hanging, so generally it is tossed into the null hypothesis.

$$H_0 : p \leq .20$$

$$H_A : p > .20$$

Be careful about your notation!

Correct

$$H_0 : \mu = 12$$

$$H_0 : p = .20$$

$$H_0 : \sigma^2 = 7$$

Incorrect

$$H_0 : \bar{X} = 12$$

$$H_0 : \bar{p} = .20$$

$$H_0 : s^2 = 7$$

- ◆ Null and alternate hypotheses are *always* about unknown population parameters; they are *never* about sample statistics.

Isn't this nitpicking?

- ◆ Confusing sample statistics and population parameters leads to complete nonsense.

- ◆ The statement is a perfect example.

- Before the sample is drawn, a sample statistic is a random variable, and therefore can't be equal to a constant.

- After the sample is drawn, the sample mean is just a number. No fancy statistical theory is required to compare two known numbers.

$$H_0 : \bar{X} = 10$$

Another Hypothesis Test

Mendel's Carnations

- ◆ One of Mendel's experiments involved crossing pink carnations.
- ◆ These pink carnations produce red, white, and pink offspring.
- ◆ Mendel believed each pink carnation had one red gene and one white gene, and that the pink offspring got one of each.

	Gene taken from the other parent		
Gene taken from one parent		R	W
	R	RR	RW
	W	WR	WW

The Hypothesis

- ◆ Mendel further believed that each of the four outcomes was equally likely, so that in an infinitely long trial, a quarter of the offspring should be red, a quarter white, and half pink.
- ◆ So, for example, the population proportion of pink offspring is predicted to be $p = .50$
- ◆ Suppose that among 100 offspring, 56 are pink. Is this consistent with Mendel's theory, or does it refute the theory?

The Alternative Hypothesis

- ◆ Suppose we test a hypothesis about the proportion of pink carnations. Our null is that the proportion is one half.
- ◆ If the null is false, do we expect p bigger than .5, smaller than .5, or could either be plausible?
- ◆ Both are plausible, so a Not Equals, or two-sided test, is appropriate.

$$H_0 : p = .50$$

$$H_A : p \neq .50$$

A Less-than test

- ◆ Advertising claims are sometimes taken as the basis for a null hypothesis.
- ◆ Suppose Duracell says their batteries will power a flashlight for at least 4 hours.
- ◆ If we wish to test this claim by testing a random sample of their batteries, these would be the null and alternative hypotheses. (Mu is pop mean lifetime)

$$H_0 : \mu \geq 4$$

$$H_A : \mu < 4$$



How do we know we have the correct alternative?

- ◆ In homework and exams, the kind of test is usually signaled with key words: “test to see if psychics do *better than* pure guessers” would signal a G.T. test.
- ◆ The words “different from” signal a two-sided or N.E. test – For example, “Test to see if the population proportion is *different from* one half.”



Selecting the alternative hypothesis

- ◆ A question such as “Does the statistical evidence suggest Duracell’s advertised claim is incorrect, and their batteries are in fact *less* durable?” is signaling a less-than test.
- ◆ In real-world applications, deciding on the alternative is sometimes difficult; ambiguous cases ought to be treated as two-sided.

Picking the alternative: a problem even in fiction

- ◆ In the movie “Man on a Swing” Cliff Robertson plays a small town sheriff.
- ◆ A woman is brutally murdered, and he has no suspects.
- ◆ Joel Grey appears, says he is psychic, and offers to help; he then describes details of the crime scene only a true psychic or the true killer would know.



As the movie poster says . . .

- ◆ Psychic
- ◆ Occultist
- ◆ Murderer
- ◆ Which?





We discover Joel Grey is creepy.

- ◆ Grey knows he is the prime suspect, and enjoys it. Grey's smug satisfaction suggests he is getting away with murder. Robertson desperately wants to nail Grey for the killing, but lacks evidence.
- ◆ Voila! In comes a statistics professor, who arrives with a Rhine deck to test Grey. Cliff Robertson, who arranged the test, paces nervously outside the door, fingering his handcuffs.



The professor emerges . . .

- ◆ “I have good news and bad news,” he says.
- ◆ “Give me the good news,” Robertson replies.
- ◆ “He didn’t do as well as the law of odds,” says the professor. [Robertson pulls out his handcuffs.] “The bad news is . . .”
- ◆ “he did much worse than can be explained by chance . . .”



Huh?

- ◆ Robertson wheels to face the professor.
“What does it mean?” he snarls.
- ◆ “Beats me,” the professor replies, “that’s your problem.”



The strange thing is . . .



- ◆ Parapsychologists have done tests of this kind for many decades now, and believe they have documented cases where the outcome is too far from random to be explained by chance.
- ◆ But “psychics” sometimes perform worse than random guessers, so that professional parapsychologists now do two sided tests.

Accepting and Rejecting

- ◆ In a hypothesis test, after examining the sample data you make a choice:
 - Either you accept the null hypothesis, or
 - You reject the null hypothesis in favor of the alternative.
- ◆ Accepting the null hypothesis means you found no evidence contradicting it; *it does not prove that the null is true.*

A useful analogy

- ◆ A hypothesis test is like a jury trial.
 - In a jury trial, the null hypothesis is innocence. “Innocent until proven guilty.”
 - Only if the evidence proves guilt beyond a reasonable doubt do you reject the null and find the defendant guilty.
 - Acquittal (accepting the null) is no guarantee of innocence: it means insufficient evidence to convict.
- ◆ Example: The O. J. Simpson case. The jury acquitted based on doubts about the evidence, not because they were convinced O. J. was innocent.



This is why . . .



- ◆ Some authors dislike the phrase “accept the null hypothesis” -- it implies the null has been shown to be true.
- ◆ They prefer to say “fail to reject the null” instead.
- ◆ However, since almost everyone uses the phrase “accept the null,” so will I.

Two Kinds of Errors

- ◆ When you accept or reject a Null, there are two distinct kinds of errors that can be made.
- ◆ Rejecting a true null, or Type I error.
- ◆ Accepting a false null, or Type II error.

		Truth	
		Null True	Null False
You say	Accept null	Ok	Type II
	Reject null	Type I	Ok

How likely are these errors?

- ◆ The probability of a Type I error is called α .
- ◆ The probability of a Type II error is called β .
- ◆ These are really conditional probabilities.
 - Alpha is the probability of rejecting the null *given that it is true*.
 - Beta is the probability of accepting the null *given that it is false* (plus a particular false value).

The ESP example

- ◆ I am going to treat this as a greater than test, despite the objections of parapsychology.
- ◆ Suppose we propose to test with a trial of 100 cards.
- ◆ And we propose to reject the null and proclaim the subject psychic if they get 30 or more correct.

$$H_0 : p \leq .20$$

$$H_A : p > .20$$

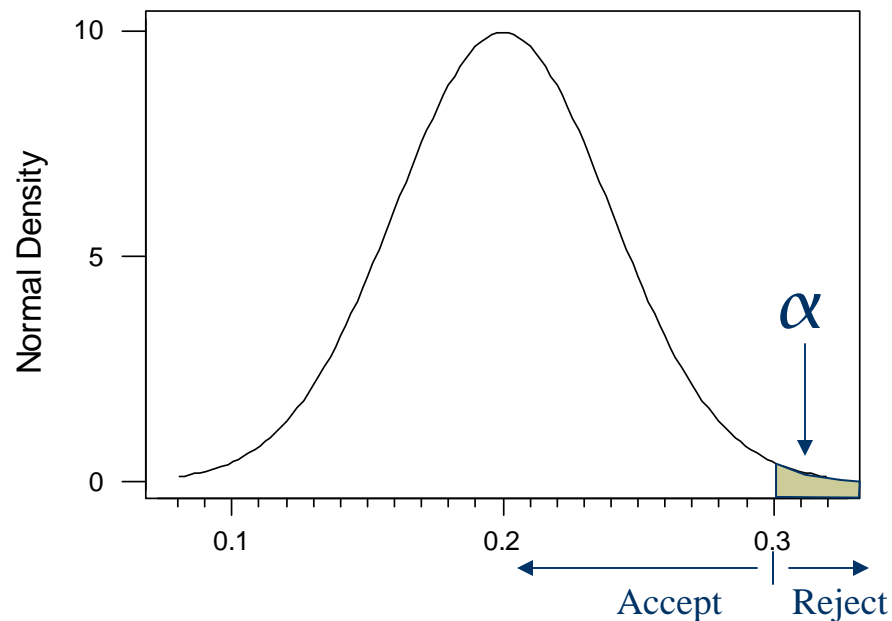
What are the properties of the test?

- ◆ We can easily evaluate the chance of committing a type I error using this rule.
- ◆ It is the chance a person just guessing succeeds in getting 30 or more correct.
- ◆ Binomial, $n=100$, $p = .20$

$$\begin{aligned}\alpha &= \text{prob}(x \geq 30) \\ &= \sum_{x=30}^{100} \binom{100}{x} (.20)^x (.80)^{100-x}\end{aligned}$$

Here is the picture for the normal approximation.

Distribution of sample proportion if the null is true.



And the corresponding computation of alpha

- ◆ α is a probability *given that the null is true*.
- ◆ The null says $p=.20$.
- ◆ Therefore, we use $p=.20$ in all the probability calculations.
- ◆ Ans: $\alpha = .0062$

$$E(\bar{p}) = p = .20$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.2 \times .8}{100}} = .04$$

$$z = \frac{.300 - .200}{.04} = 2.5$$

$$\text{Prob}(z > 2.5) = .0062$$

How about Type II error?

- ◆ Type II error occurs when the person is genuinely psychic, and we fail to detect it.
- ◆ How likely this is depends on how talented the psychic is; a psychic who never makes a mistake would be easily detected.
- ◆ *Power* is the chance a false null is correctly rejected. It is equal to $1-\beta$.
- ◆ Consider some specific examples: psychics who get 25% correct, and psychics who get 35% correct.

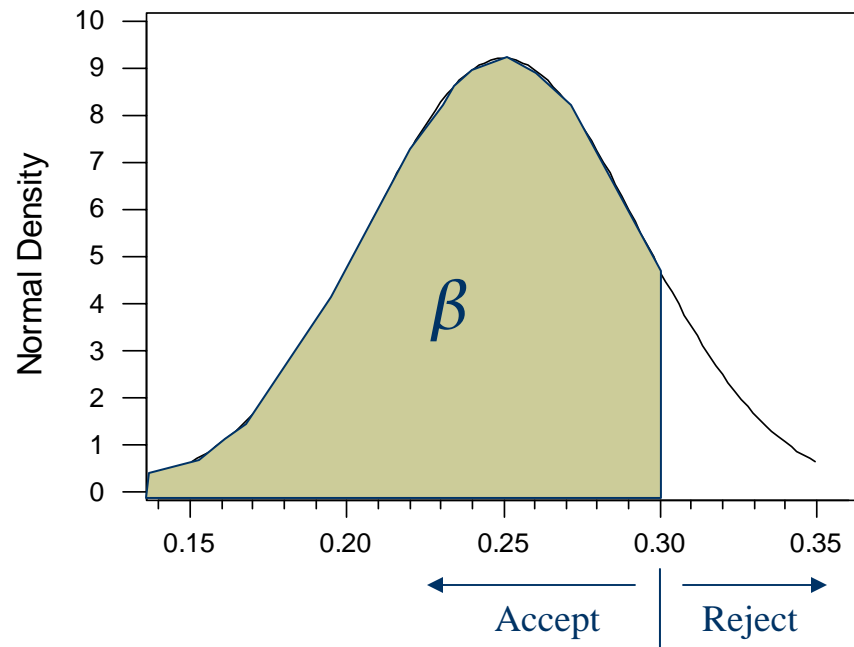
What is β when $p = .25$?

- ◆ We can easily evaluate the chance of committing a type II error using this rule.
- ◆ It is the chance a person capable of getting 25% correct gets fewer than 30 correct in 100 tries.
- ◆ Binomial, $n=100$, $p = .25$

$$\begin{aligned}\beta &= \text{prob}(x < 30) \\ &= \sum_{x=0}^{29} \binom{100}{x} (.25)^x (.75)^{100-x}\end{aligned}$$

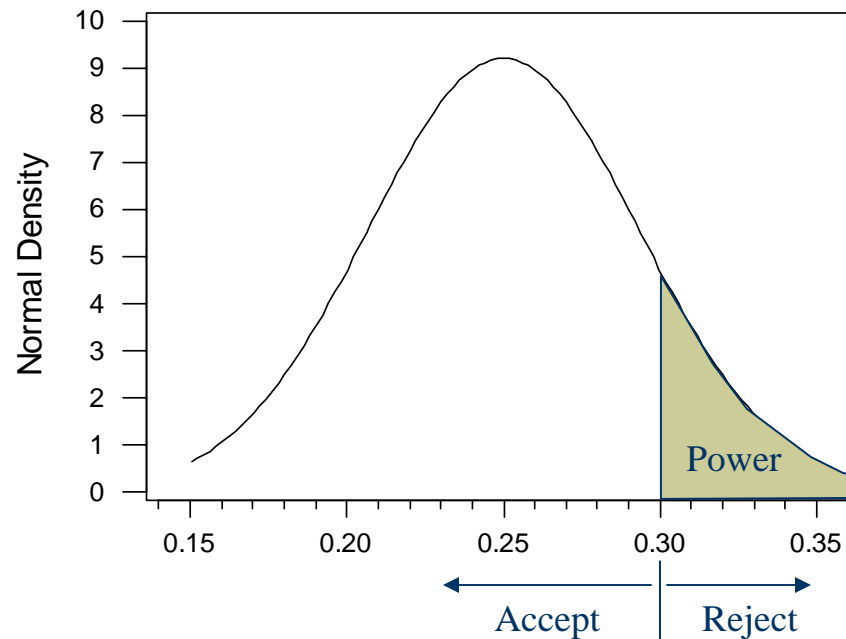
Here is the picture of β for $p = .25$

Distribution of Sample Proportion
Psychic who gets 25% Correct



Here is the *Power* of the test for $p = .25$

Distribution of Sample Proportion
Psychic who gets 25% Correct



And the corresponding computation of β

- ◆ β is a probability *given that the null is false*.
- ◆ Here the null is false, and p is actually .25.
- ◆ Therefore, we use $p=.25$ in all our calculations.
- ◆ Ans: $\beta = .8749$; power is $1-\beta = .1251$

$$E(\bar{p}) = p = .25$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.25 \times .75}{100}} = .0433$$

$$z = \frac{.300 - .25}{.0433} = 1.15$$

$$\beta = \text{Prob}(z < 1.15) = .8749$$

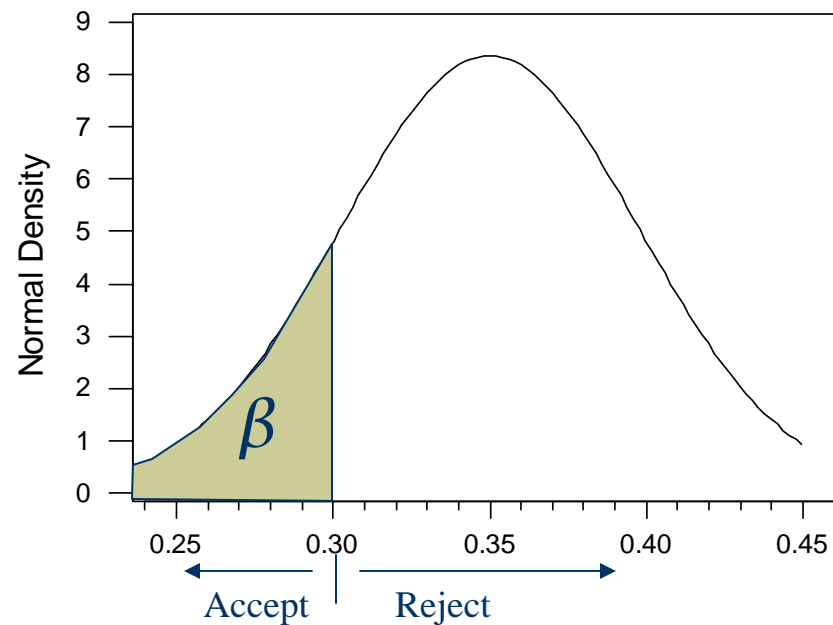
What is β when $p=.35$?

- ◆ We can easily evaluate the chance of committing a type II error using this rule.
- ◆ It is the chance a person capable of getting 35% correct gets fewer than 30 correct in 100 tries.
- ◆ Binomial, $n=100$, $p = .35$

$$\begin{aligned}\beta &= \text{prob}(x < 30) \\ &= \sum_{x=0}^{29} \binom{100}{x} (.35)^x (.65)^{100-x}\end{aligned}$$

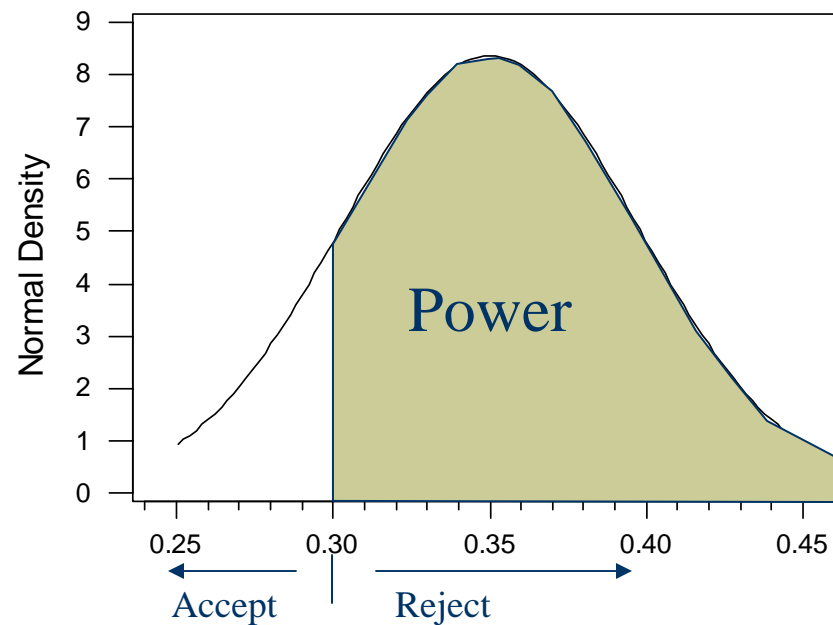
Here is the picture of β for $p = .35$

Distribution of Sample Proportion
Psychic who gets 35% Correct



Here is the *Power* of the test when $p = .35$

Distribution of Sample Proportion
Psychic who gets 35% Correct



And the corresponding computation of β

- ◆ β is a probability *given that the null is false*.
- ◆ Here the null is false, and p is actually .35.
- ◆ Therefore, we use $p=.35$ in all our calculations.
- ◆ Ans: $\beta = .1469$; power is $1-\beta = .8531$

$$E(\bar{p}) = p = .35$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.35 \times .65}{100}} = .0477$$

Neglecting ccf,

$$z = \frac{.30 - .35}{.0477} = -1.05$$

$$\beta = \text{Prob}(z < -1.05) = .1469$$

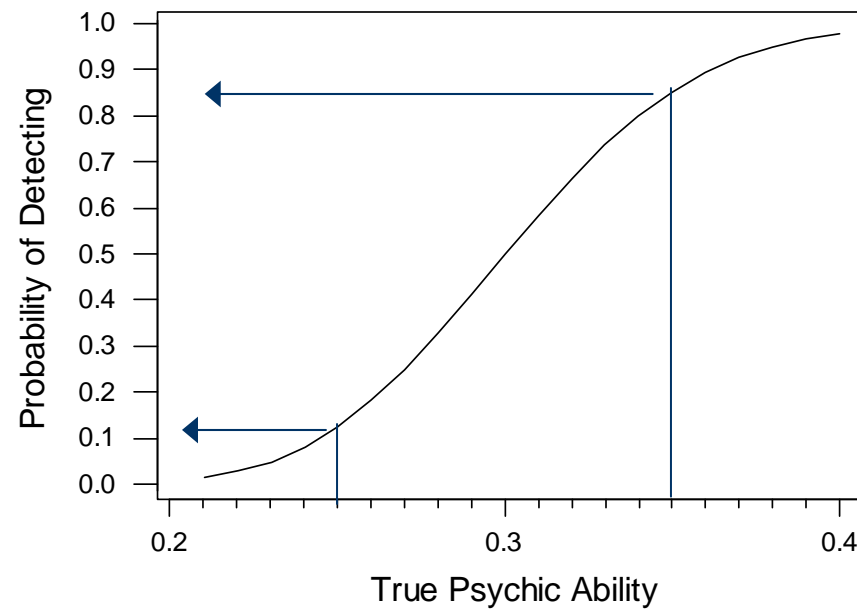


Repeat for many values of H_A

- ◆ Then plot the power of the test as a function of the true value of psychic ability.
- ◆ This plot of power versus alternative values is called a *power curve*.
- ◆ The power curve and α together tell you how well the test performs.

The Power Curve

The Power Curve for this Test





In an ideal world . . .

- ◆ Investigators would consider many factors in designing a test.
 - Any prior knowledge regarding the plausibility of the null hypothesis.
 - The relative cost of Type I and Type II errors.
 - The costs and benefits of getting a bigger sample.
- ◆ They would then pick a sample size and rejection region that produced an acceptable α and power curve in light of these considerations.



But what many really do is . . .

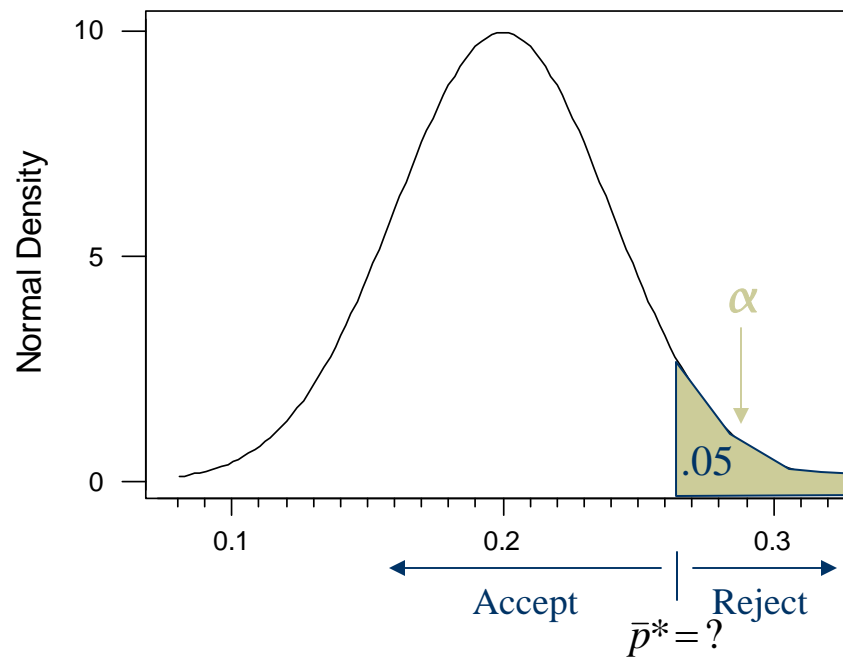
- ◆ Set α at a conventional level, typically 5%.
- ◆ Select the rejection region to achieve the desired level for α .
- ◆ Completely ignoring
 - Type II Error and Power
 - Costs/Benefits
 - Prior Knowledge

Why?

- ◆ There is no good reason; it is a sociological phenomenon. Doing tests with $\alpha=.05$ has been standard practice for many decades; many practitioners now regard it as a hallmark of proper scientific procedure.
- ◆ When R.A. Fisher (a famous statistician) popularized hypothesis tests, he was asked what would be an appropriate α . He opined that 5% seemed to him to be about right; this was the origin of the practice.

If we want $\alpha = .05$ in the ESP example, here is the picture.

Distribution of sample proportion if the null is true.



How do we find the boundary of the rejection region?

- ◆ The boundary value, \bar{p} -star can be found from the following equation.
- ◆ Therefore, our test is: If the psychic gets more than 26.58% right in 100 trials, reject H_0 .

$$z_{.05} = 1.645 = \frac{\bar{p}^* - .20}{\sqrt{\frac{.2 \times .8}{100}}}$$

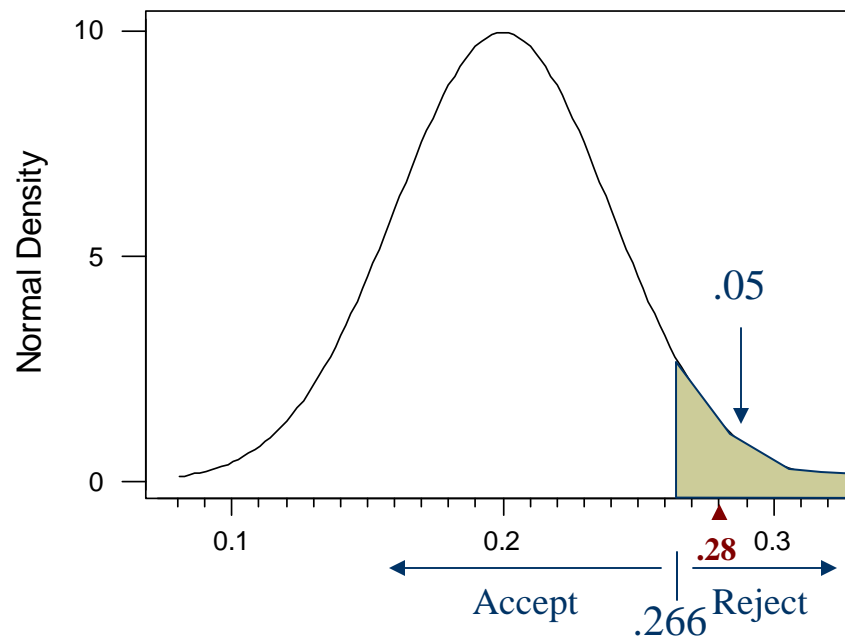
$$\bar{p}^* = .20 + 1.645 \times .04 = .2658$$

If our sample outcome were .28

- ◆ Our decision rule is
 - If the sample proportion is less than or equal to .266, accept H_0
 - If the sample proportion is greater than .266, reject H_0 .
- ◆ Since $.28 > .266$, it is in the rejection region, and we reject H_0 .

Here is the picture

Distribution of sample proportion if the null is true.




More Terminology

- ◆ If you test a null hypothesis using a rejection rule designed to give $\alpha = .05$, and you reject the null, the result is said to be *statistically significant at the 5% level*.
- ◆ If you test a null hypothesis using a rejection rule designed to give $\alpha = .01$, and you reject the null, the result is said to be *statistically significant at the 1% level*. And so on.



What *statistically significant* means, and doesn't mean.

- ◆ If a result is *statistically significant*, what it means is that you believe you have enough sample data to discern a difference between the null and the truth.
- ◆ The difference you discern may or may not be of any practical importance: *statistical significance* shouldn't be confused with *practical significance*.

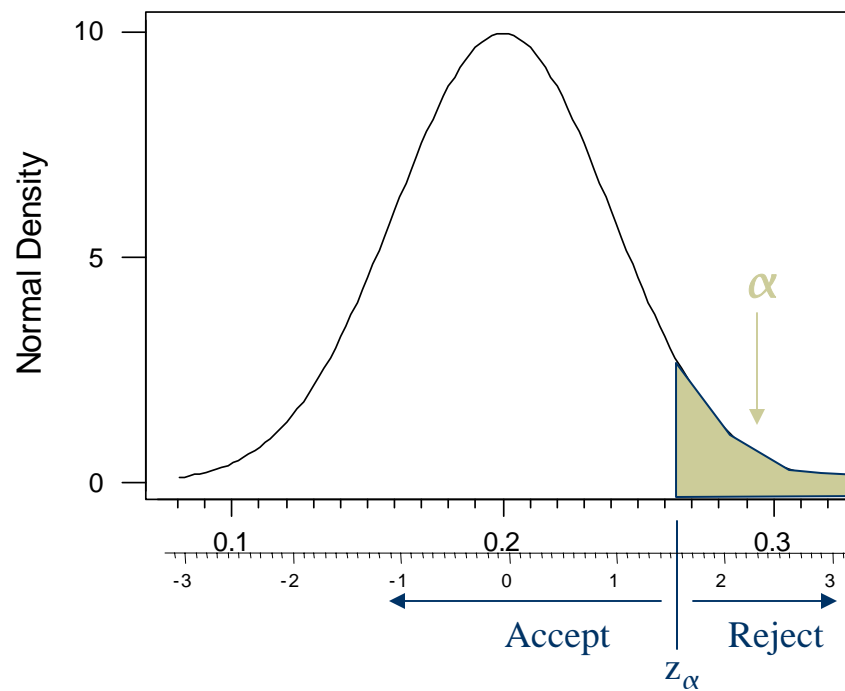


There are other ways to do a hypothesis test.

- ◆ The method most often used in Anderson, Sweeny, and Williams involves comparing a critical value of z to a computed value.
- ◆ This amounts to nothing more than relabeling the x axis in our diagram.

Look at the picture in terms of z

Distribution of sample proportion if the null is true.



Compute the z that corresponds to our observed value

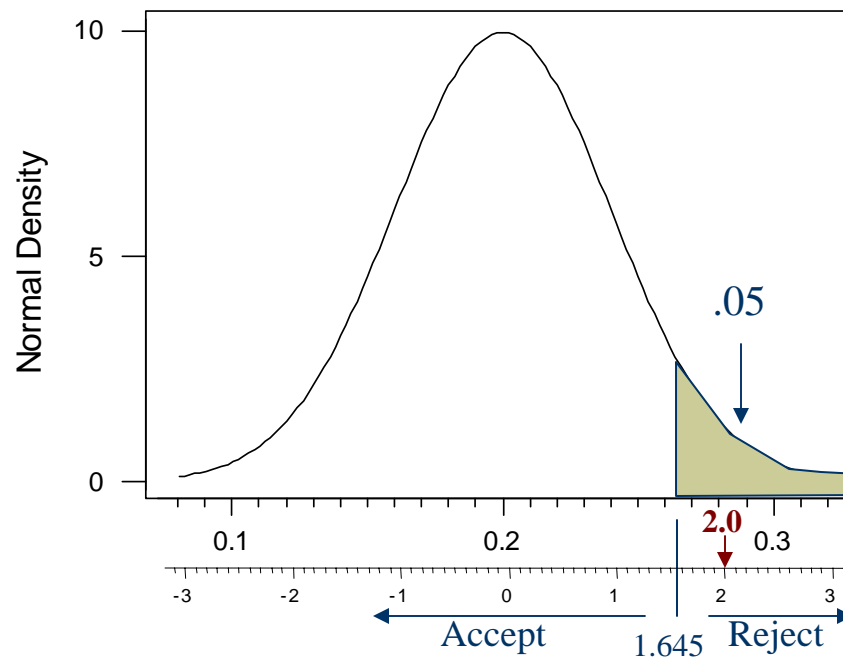
- ◆ Again, suppose our subject got 28 correct in 100 trials.
- ◆ The z-score is 2.00.
- ◆ Which is greater than the critical z-score of 1.645.
- ◆ So we reject the null hypothesis.

$$z_{obs} = \frac{.28 - .20}{\sqrt{\frac{.2 \times .8}{100}}}$$

$$z_{obs} = 2.00 > 1.645 = z_{.05}$$

Here is the picture

Distribution of sample proportion if the null is true.

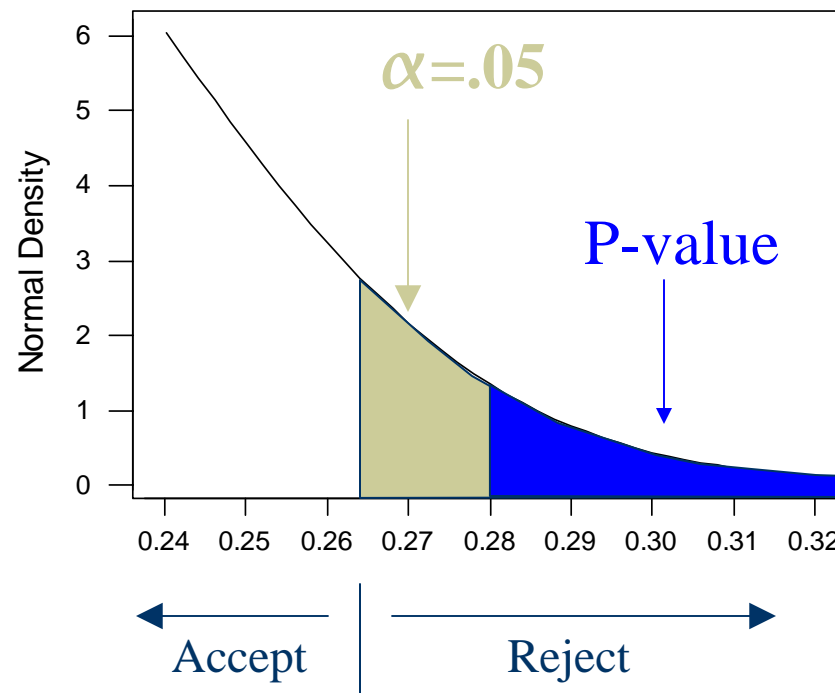


P-values

- ◆ Another way to test a hypothesis is by using what are known as p-values.
- ◆ If $\alpha > p\text{-value}$, reject H_0 ; if $\alpha < p\text{-value}$, accept H_0
- ◆ For a greater-than test, *the p-value is the probability of getting an outcome as big or bigger than what you got in your sample, when the null hypothesis is true.*

A Comparison of tail areas

Distribution of sample proportion correct
Right tail area



Computing the p-value

- ◆ The chance of getting 28 or more correct answers, while just guessing, is .0228.
- ◆ This is the p-value.
- ◆ Since $\alpha = .05 > .0228$, reject H_0 .

$$\text{p-value} = \text{Prob}(\bar{p} \geq .28)$$

$$\text{p-value} = \text{Prob}\left(z \geq \frac{.28 - .20}{.04}\right)$$

$$= \text{Prob}(z \geq 2.00)$$

$$= .0228$$

Why are p-values important?

- ◆ It is a superior technique
 - Allows the reader to use whatever α the reader prefers.
 - Communicates the strength of the result better than simply saying “statistically significant at the 5% level.”
 - Our p-value of .0228, for example, is statistically significant at the 5% level, but not the 1% level, and would probably not convince a skeptic.
- ◆ Practical considerations
 - Many computer programs (including Minitab) report p-values.

A 2-sided test: Mendel's carnations

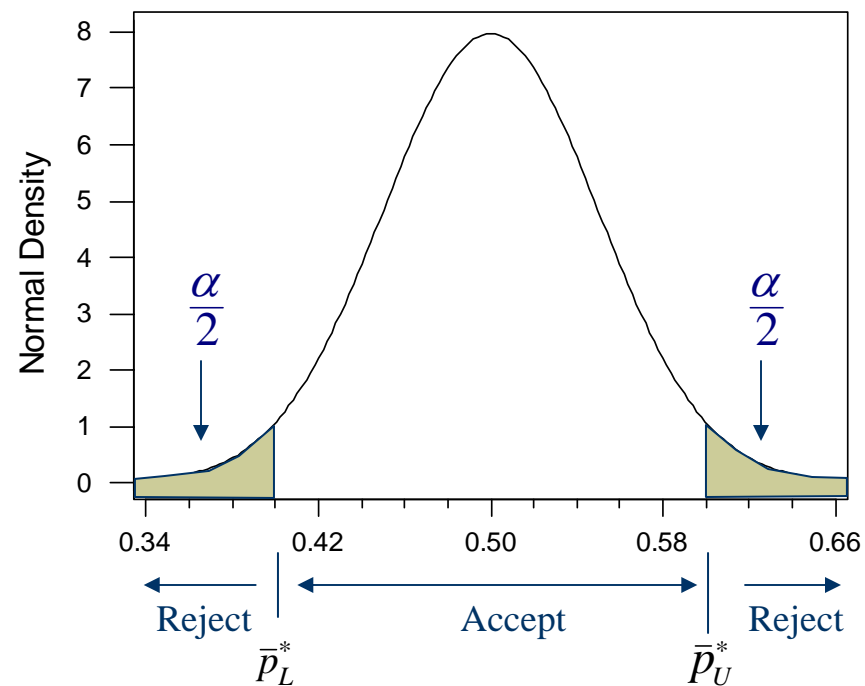
- ◆ Two-sided (or “Not Equal to”) tests are a bit different. Since we have no prior belief that the true value must be above or below the null, we reject for both very large and very small sample outcomes.
- ◆ Testing whether the proportion of pink offspring is different from $p = .50$ illustrates the technique.

$$H_0 : p = .50$$

$$H_A : p \neq .50$$

A Two-Sided Test

Distribution of Sample Proportion of Pink Carnations
when Null Hypothesis is true.



Computing Boundaries of the Rejection Region

- ◆ The test is one where we accept H_0 for all values of the sample proportion between an upper and lower limit.
- ◆ For $\alpha=.05$, these limits are computed to be roughly .40 and .60.

$$z_{.025} = 1.96 = \frac{\bar{p}_U^* - .50}{\sqrt{\frac{.5 \times .5}{100}}}$$

$$\bar{p}_U^* = .50 + 1.96 \times .05 \approx .60$$

$$-z_{.025} = -1.96 = \frac{\bar{p}_L^* - .50}{\sqrt{\frac{.5 \times .5}{100}}}$$

$$\bar{p}_L^* = .50 - 1.96 \times .05 \approx .40$$

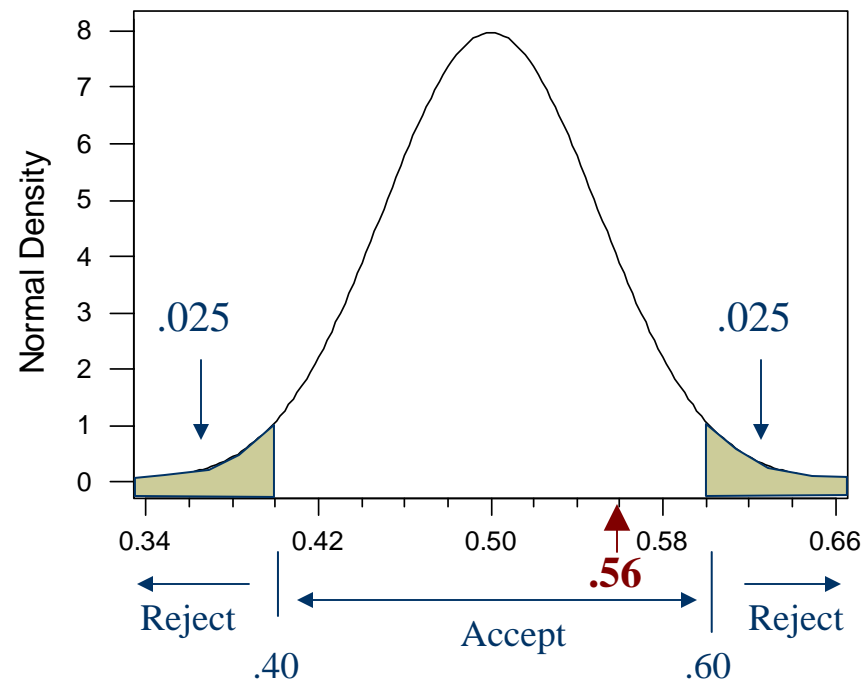


So our conclusion is . . .

- ◆ If the sample proportion of pink carnations is between .40 and .60, the evidence is consistent with Mendel's theory, and we accept the null.
- ◆ If the sample proportion is greater than .60 or less than .40, we reject the null.
- ◆ So if we got 56 pink carnations, to take a specific example, we'd accept the null.

This is the Picture

Distribution of Sample Proportion of Pink Carnations
when Null Hypothesis is true.

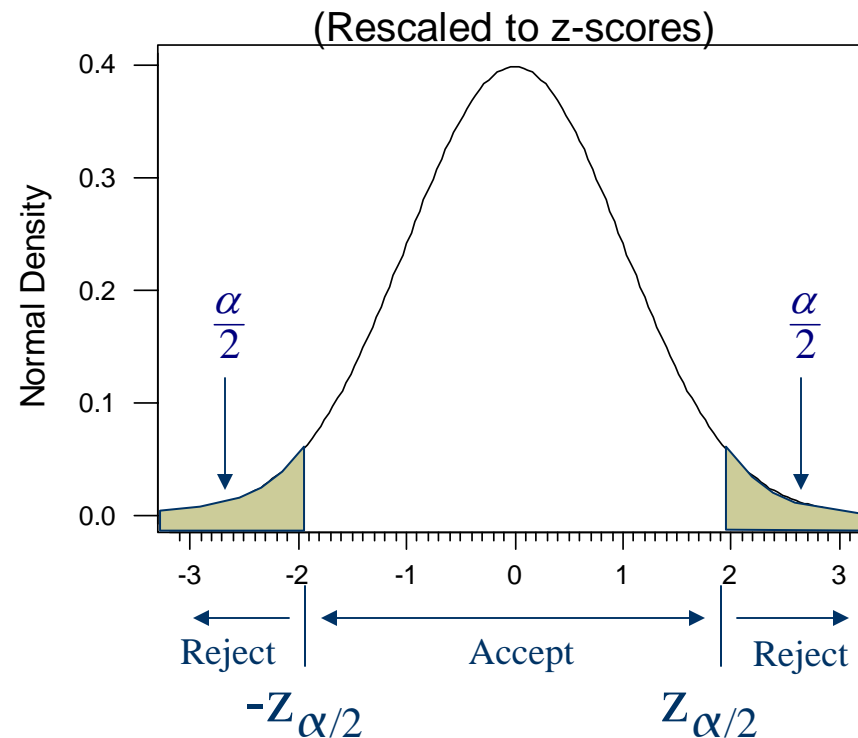


Other ways to do the test

- ◆ As in the case of Greater-than tests, there are two other ways to perform the test.
 - Compare observed z-values to critical z-value(s).
 - Compare a p-value to α .
- ◆ Let's repeat the test, using the z-value technique.

Acceptance and Rejection Regions in z-values

Distribution of Sample Proportion of Pink Carnations
when Null Hypothesis is true.



Computing an observed z

- ◆ Our sample outcome is that 56 of 100 offspring are pink, so the observed z is 1.20.
- ◆ Note we use .50, not .56, in computing the standard deviation. This is because α is a probability *given the null is true*, and the null says $p = .50$.
- ◆ For $\alpha = .05$, we accept H_0 .

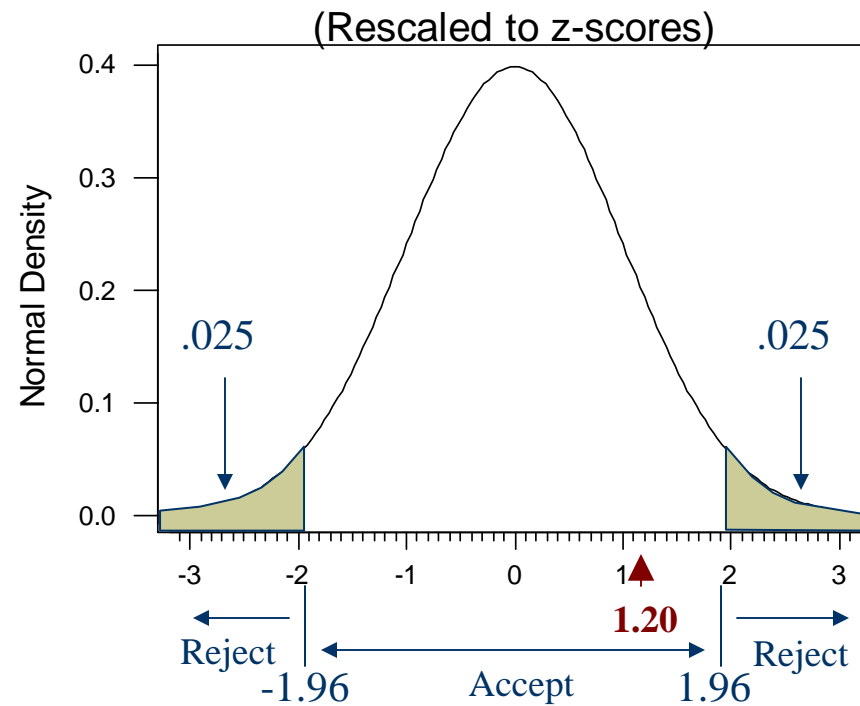
$$z_{obs} = \frac{.56 - .50}{\sqrt{\frac{.5 \times .5}{100}}}$$

$$z_{obs} = 1.20$$

$$-z_{.025} = -1.96 < 1.20 < 1.96 = z_{.025}$$

This is the picture.

Distribution of Sample Proportion of Pink Carnations
when Null Hypothesis is true.

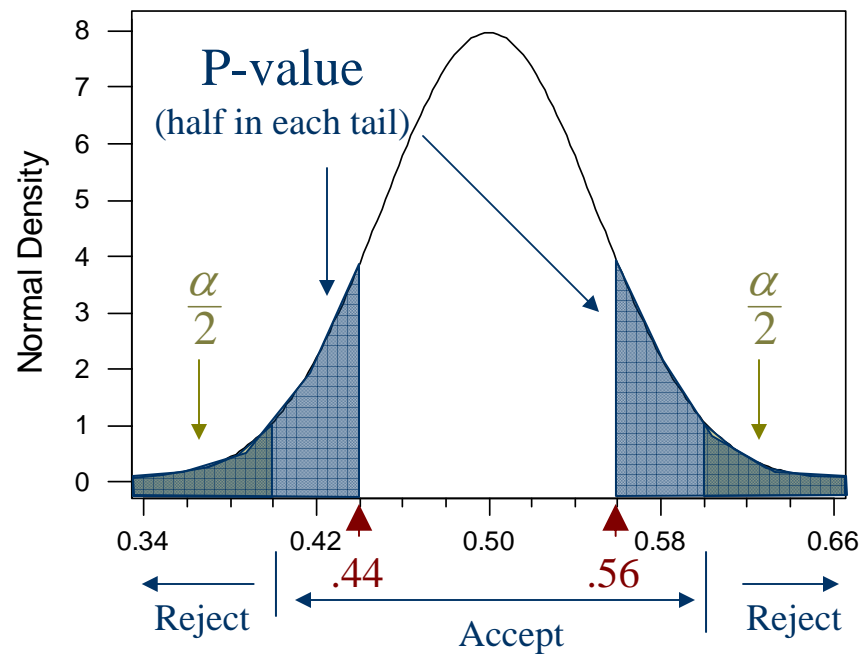


P-values again

- ◆ The other way to test a hypothesis is by using a p-value.
- ◆ Once again, the rule is:
If $\alpha > \text{p-value}$, reject H_0 ; if $\alpha < \text{p-value}$, accept H_0
- ◆ For a *two-sided test*, when the sampling distribution is symmetric, the p-value is the probability of getting an outcome *as far or further* from H_0 as what you got in your sample, when the null hypothesis is true.

A Two-Sided p-value

Distribution of Sample Proportion of Pink Carnations
when Null Hypothesis is true.



Computing the p-value

- ◆ We compute the p-value, and compare it to α .
- ◆ Since $.2302 > .05$, we accept H_0 in this case.
- ◆ There is a 23% chance of getting a proportion as far or further from $.50$ as what we got; it is not an unlikely outcome when H_0 is true.

$$\text{P-value} = P(\bar{p} \leq .44 \cup \bar{p} \geq .56)$$

$$z_1 = \frac{.56 - .50}{.05} = 1.20$$

$$z_2 = \frac{.44 - .50}{.05} = -1.20$$

From the z table we find

$$P(z \leq -1.20 \cup z \geq 1.20) = .2302$$

A less-than test: Testing the durability of batteries

- ◆ Our examples so far have tested propositions about population proportions. Tests of population means are also common.

$$H_0 : \mu \geq 4$$

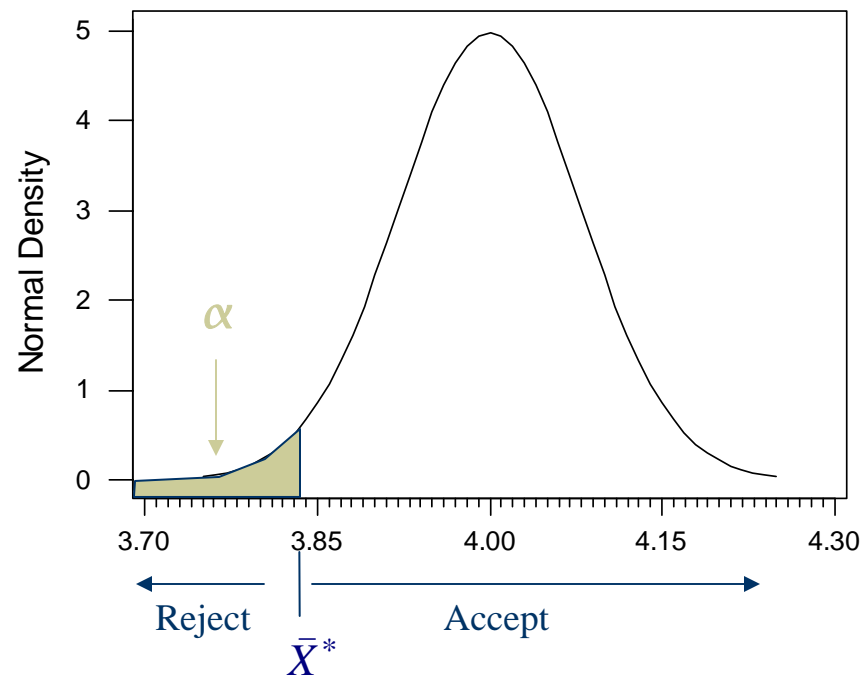
- ◆ Testing a manufacturer's claim that its batteries last 4 or more hours in a particular application is a concrete example.

$$H_A : \mu < 4$$

- ◆ Suppose $\sigma = 0.8$ and $n=100$.

The Less-than Test Illustrated

Distribution of the Sample Mean
when $\mu = 4$



Computing Boundaries of the Rejection Region

- ◆ For a change of pace, let's use $\alpha = .02$ this time.
- ◆ For this α we accept H_0 whenever the observed sample mean is greater than 3.836; we reject H_0 when the observed sample mean is below 3.836.

$$-z_{.02} = -2.05 = \frac{\bar{X}^* - 4.0}{\frac{.8}{\sqrt{100}}}$$

$$\bar{X}^* = 4.0 - 2.05 \times .08 = 3.836$$



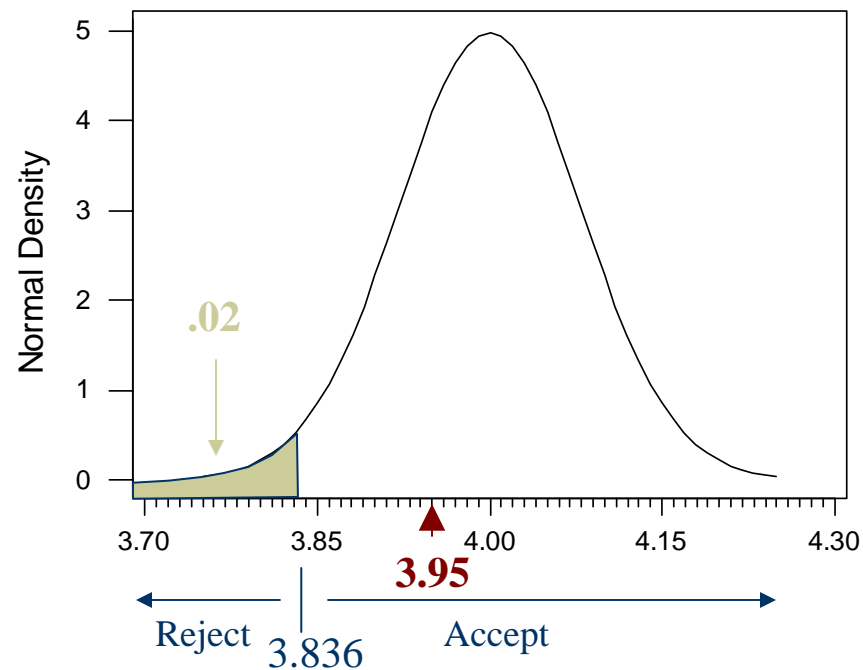
So our conclusion is . . .



- ◆ Suppose our sample yielded a sample mean of 3.95 hours.
- ◆ Since $3.95 > 3.836$, we would accept the null hypothesis.

This is the picture.

Distribution of the Sample Mean
when $\mu = 4$



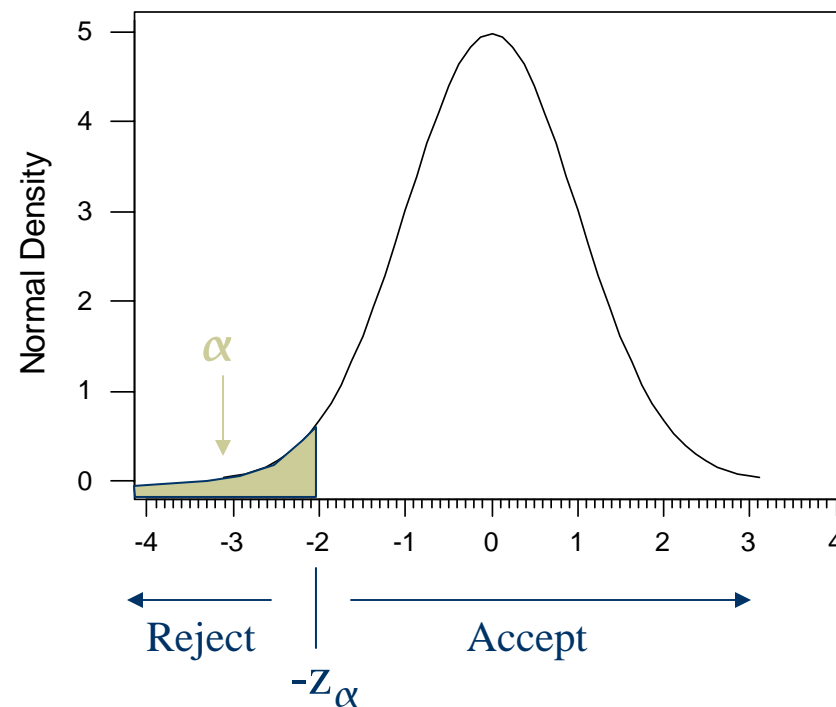


Other ways to do the test

- ◆ Once again, there are two other ways to perform the test.
 - Compare observed z-values to critical z-value(s).
 - Compare a p-value to α .
- ◆ Let's repeat the test, using the z-value technique.

The test, scaled in z-scores

Distribution of the Sample Mean
when $\mu = 4$, measured by z-score



Computing the observed z

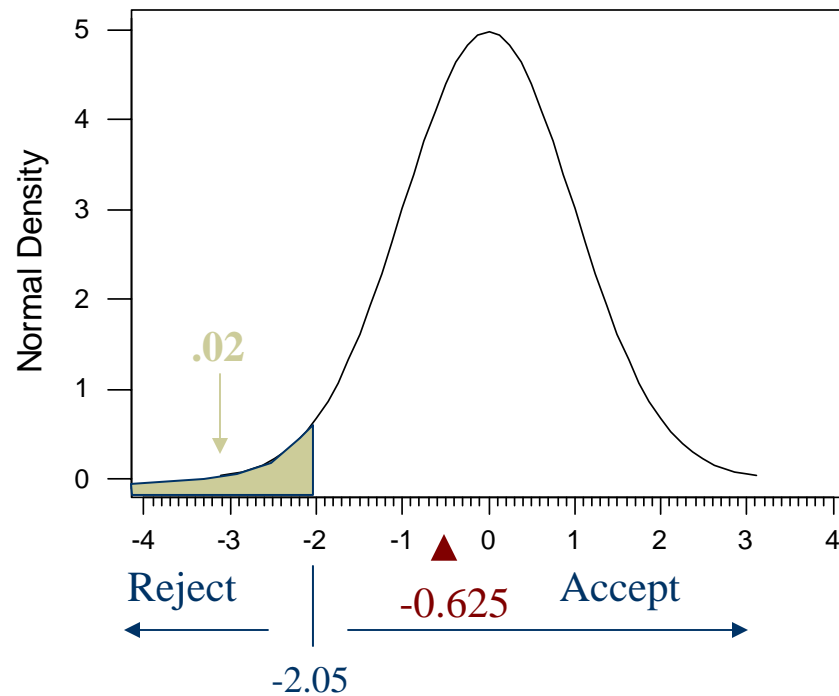
- ◆ The z-value corresponding to our observed sample mean of 3.95 turns out to be $-.625$.
- ◆ Since $-2.05 < -0.625$, we accept the null hypothesis.

$$z_{obs} = \frac{3.95 - 4.00}{\frac{.8}{\sqrt{100}}}$$

$$z_{obs} = \frac{-.05}{.08} = -.625$$

This is the picture.

Distribution of the Sample Mean
when $\mu = 4$, measured by z-score

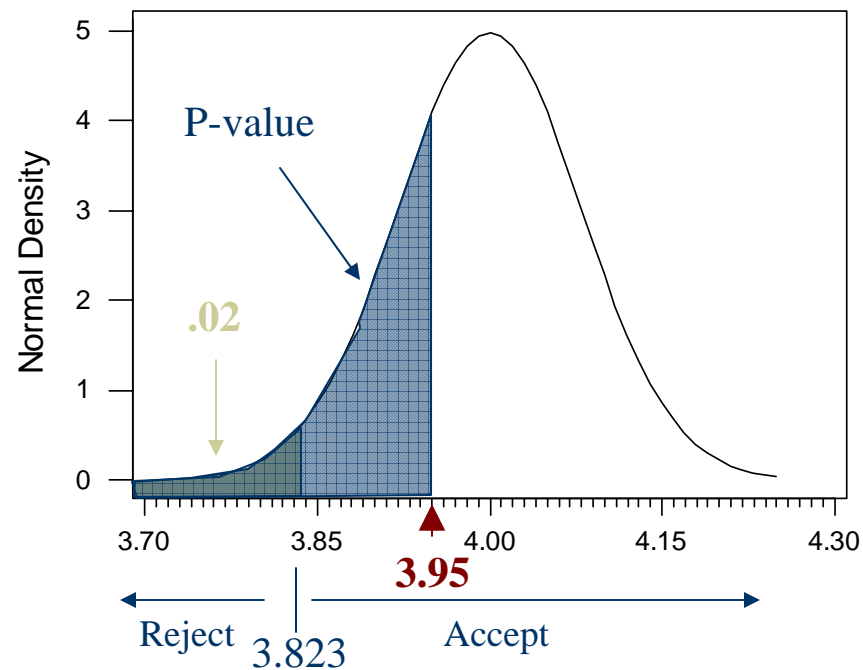


P-values again

- ◆ The other way to test a hypothesis is by using a p-value.
- ◆ For all tests (greater-than, less-than, and not-equal), the rule is always the same:
If $\alpha > \text{p-value}$, reject H_0 ; if $\alpha < \text{p-value}$, accept H_0
- ◆ For a *less-than test*, the p-value is the probability of getting an outcome *as small or smaller* as what you got in your sample, when the null hypothesis is true.

The p-value illustrated.

Distribution of the Sample Mean
when $\mu = 4$



Computation of the p-value

- ◆ The p-value is .266, which means that even if the advertising claim is true, and battery life averages 4 hours, there is almost a 27% chance of getting a sample mean battery life as low as 3.95 hours.
- ◆ $.266 > .02$, so we accept H_0

$$\begin{aligned} \text{p-value} &= P(\bar{X} < 3.95) \\ z &= \frac{3.95 - 4.00}{\frac{8}{\sqrt{100}}} = -.625 \end{aligned}$$

From table or Minitab cdf command

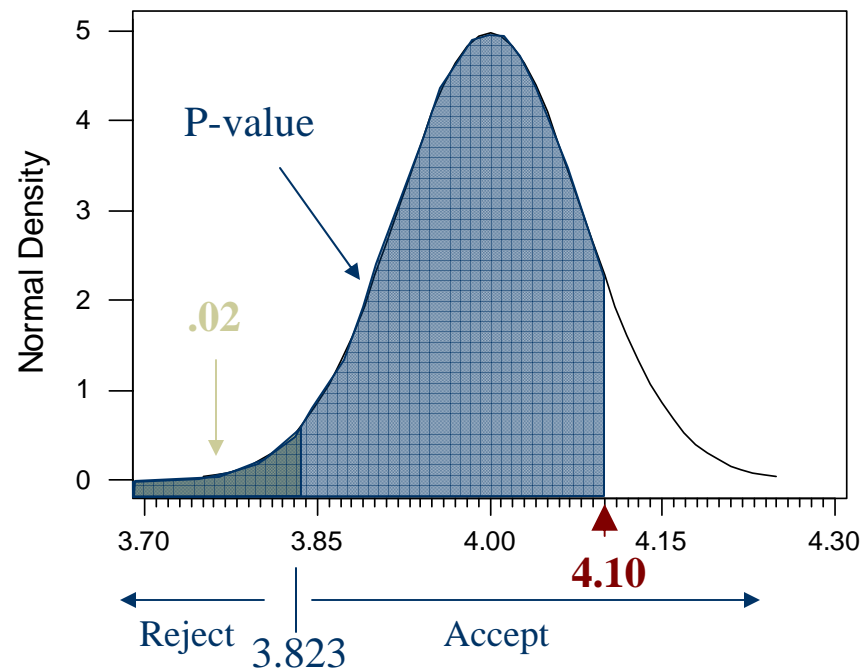
$$P(z < -0.625) = .266$$

A warning about p-values

- ◆ When doing a greater-than or less-than test, a p-value is *not necessarily* a tail area.
- ◆ My examples, and the book's problems, may mislead you into thinking so.
- ◆ For a less-than test, the p-value is the chance of getting a result *as small or smaller* than what you got in your sample, when H_0 is true.
- ◆ Example: suppose our sample mean had been 4.10 instead of 3.95.

This is what the p-value would have been!

Distribution of the Sample Mean
when $\mu = 4$



When do these situations arise?

- ◆ In a one-sided test, when the sample value of the statistic is on the side predicted by the null hypothesis, rather than the side predicted by the alternate hypothesis.
- ◆ The large p-value simply tells you the outcome is very likely if the null is true.
- ◆ Example: the p-value in the last picture is .8944, which says a sample mean of 4.10 hours is entirely consistent with the null.



Other Distributions

- ◆ The logic of hypothesis testing can be applied anytime we know how the sample statistic is distributed when the null hypothesis is true.
- ◆ Our examples thus far have all been based on the normal distribution, but other distributions, such as the t , frequently arise.

A t-distribution problem

- ◆ On the average, a housewife with a husband and two children is estimated to work 55 hours or less per week on household-related activities. The hours worked during a week for a sample of eight housewives are: 58, 52, 64, 63, 59, 62, 62, and 55. Test the null that μ is less than or equal to 55 against the alternative it is greater than 55, using $\alpha=.05$.

Using the t distribution

- ◆ The parameter μ is mean hours of housework per week being done by housewives with two children.
- ◆ We are asked to test this hypothesis using $\alpha=.05$ and a sample of eight observations.

$$H_0 : \mu \leq 55$$

$$H_A : \mu > 55$$

Why the t distribution?

- ◆ The t arises in hypothesis testing just as it does in confidence intervals: when the sample size is small and the population standard deviation is unknown. Here, eight is a small sample, and σ is unknown.
- ◆ Strictly speaking the observations must come from a normal distribution, but the t distribution is known to be a good approximation if the population is *anywhere near* being normally distributed.

A preliminary step

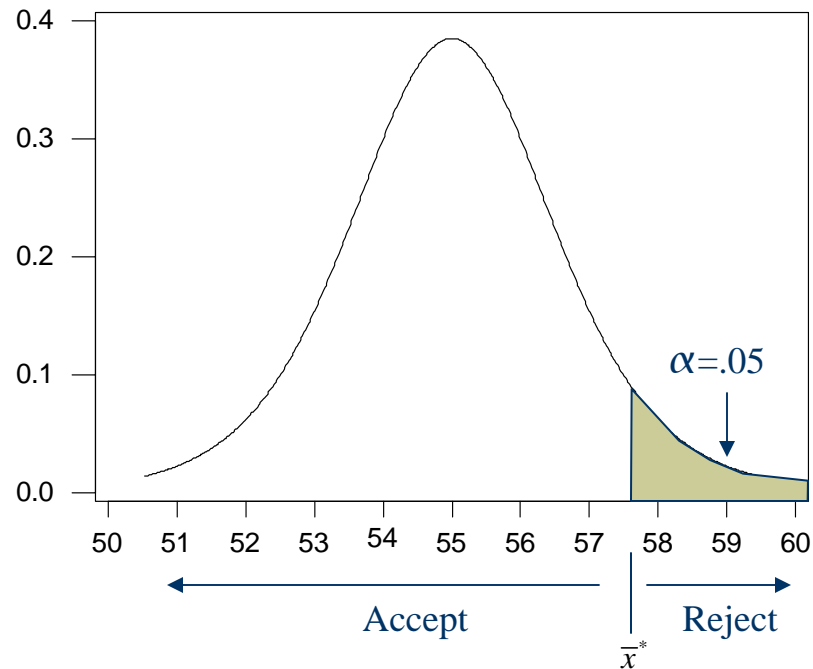
- ◆ Before doing the hypothesis test, we must first use the eight observations given to compute the sample mean and standard deviation.

$$\bar{x} = \frac{\sum x_i}{n} = 59.37$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}} = 4.21$$

The test, pictured

Distribution of the Sample Mean
when $\mu = 55$



Finding the accept/reject boundary

- ◆ Here is the calculation to find the value of the sample mean that marks the boundary of the accept/reject region.
- ◆ Our sample x-bar is 59.37. Since $59.37 > 57.82$, we reject the null.

$$t_{.05} = \frac{\bar{x}^* - \mu_0}{s / \sqrt{n}}$$

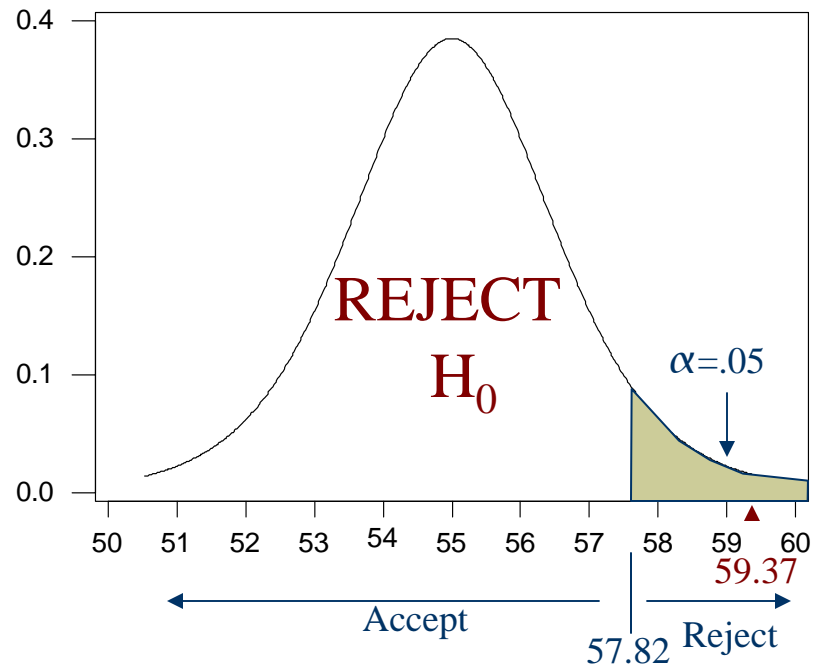
$$1.895 = \frac{\bar{x}^* - 55}{4.21 / \sqrt{8}}$$

$$\bar{x}^* = 55 + 1.895 \times \frac{4.21}{\sqrt{8}}$$

$$\bar{x}^* = 57.82$$

This is the picture.

Distribution of the Sample Mean
when $\mu = 55$



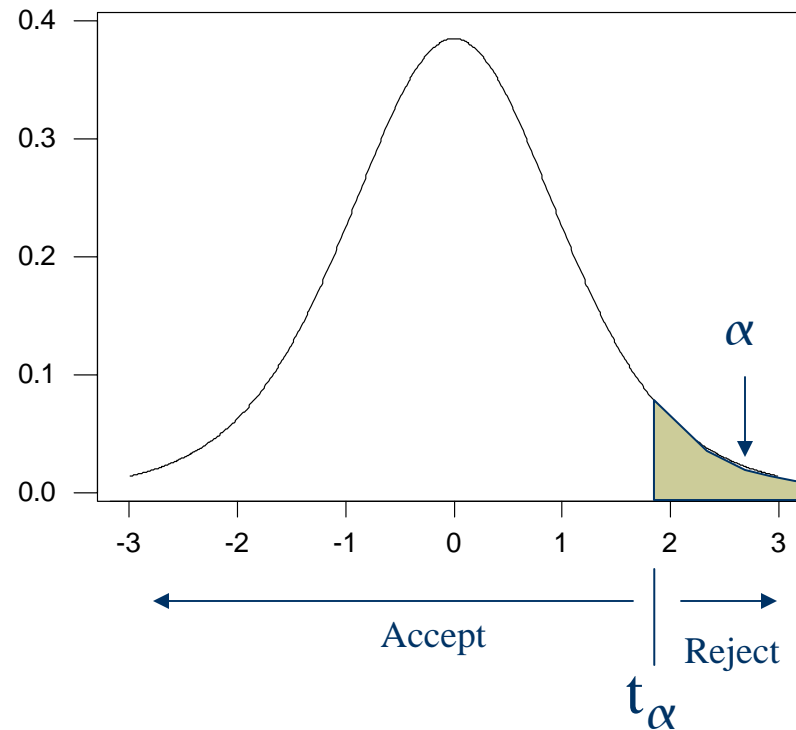


Two other methods

- ◆ As in our earlier examples, we can perform this test in two other ways.
 - By comparing a critical value of the test statistic with an observed value, or
 - By comparing α to a p-value.
- ◆ Let's begin by looking at the distribution of the t statistic when the null hypothesis is true.

Accept/Reject in t values

The Distribution of the t-statistic
when the Null is true



Compute the observed t value

- ◆ To compare the critical and observed t values, we must first convert our observed sample mean to an observed t value.
- ◆ For $\alpha=.05$, our critical t value is 1.895.

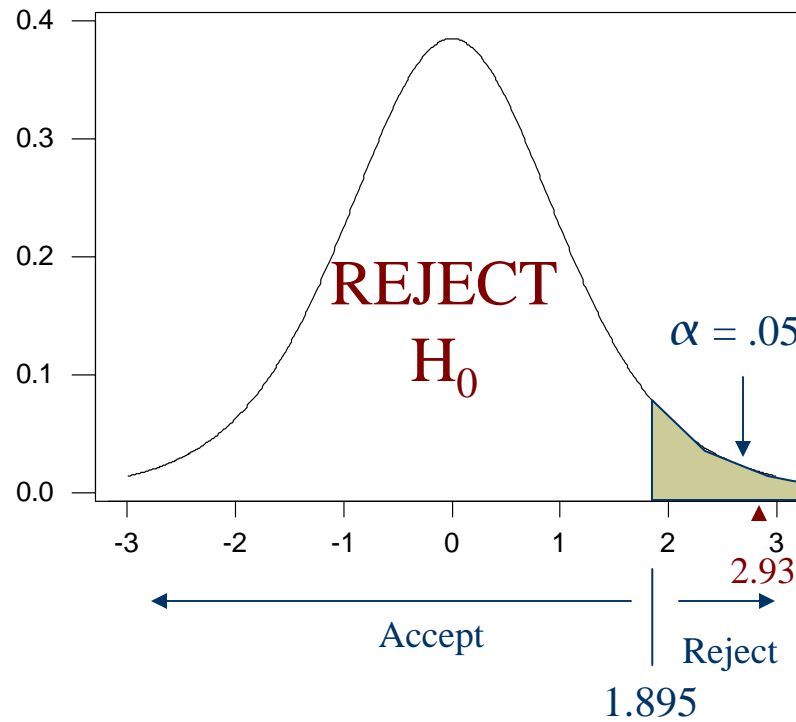
$$t_{obs} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$t_{obs} = \frac{59.37 - 55}{4.21 / \sqrt{8}}$$

$$t_{obs} = 2.93$$

This is the picture.

The Distribution of the t-statistic
when the Null is true



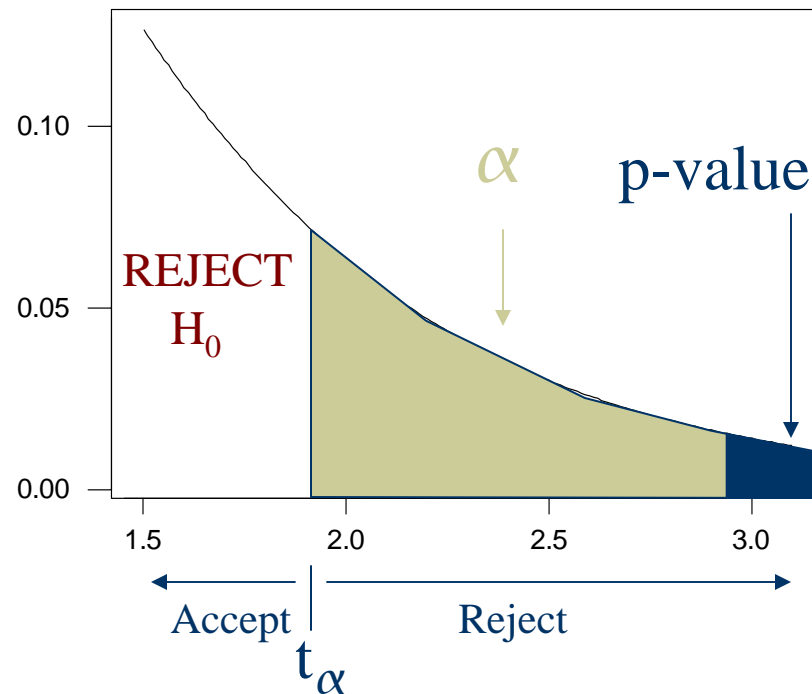


Using the p-value

- ◆ Since this is a greater-than test, the p-value is the probability of getting a sample outcome *as big or bigger* than what you observed, *when the null hypothesis is true*.
- ◆ If the p-value is smaller than α we reject the null.

The p-value illustrated

Right Hand tail of the Distribution
when the Null is true



Bracketing the p-value

- ◆ When you look for $t = 2.93$ in the table, you won't find it, but . . .
- ◆ Since $2.365 < 2.93 < 2.998$, we can conclude that $.025 > \text{p-value} > .01$. Therefore $\alpha = .05 > \text{p-value}$.

Area in Upper Tail			
Degrees of Freedom	.05	.025	.01
7	1.895	2.365	2.998

Type II error redux

- ◆ Suppose you want to find the chance of Type II error when $\mu = 18$ for the following hypothesis test.
- ◆ Givens:
 - $\sigma = 10$
 - $n = 200$
 - $\alpha = .05$
- ◆ A, S, & W, # 47(a), p. 374.

$$H_0 : \mu = 20$$

$$H_A : \mu \neq 20$$

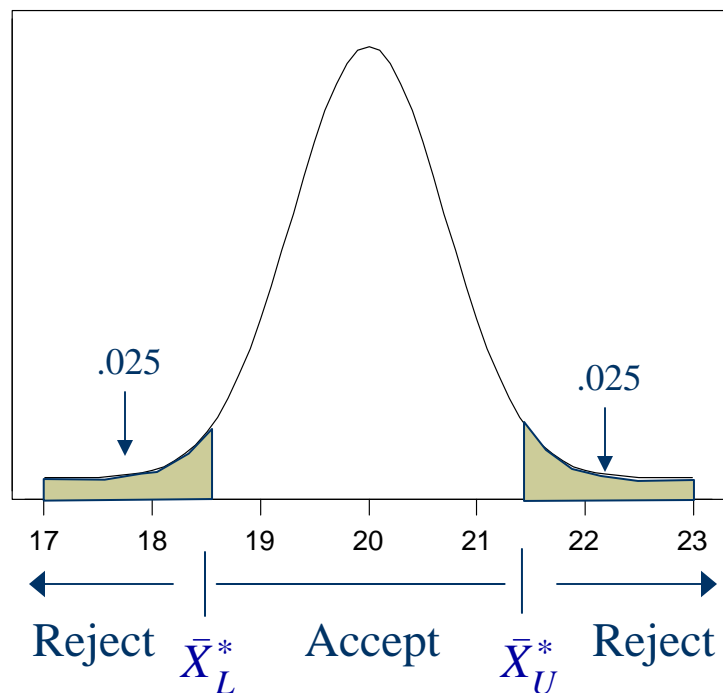


It's complicated

- ◆ Before we can begin to compute the chance of a type II error, we must first solve to find the boundaries of the acceptance and rejection region.
- ◆ The boundaries must be found in terms of values of the test statistic (in this case, the sample mean).
- ◆ That means we *must* use the first technique.

Here is the picture

Distribution of the Sample Mean
when the null hypothesis is true



Solving for the Accept/Reject region

- ◆ One accepts the null hypothesis for all values of the sample mean between 18.61 and 21.39.
- ◆ Now we can compute the chance of type II error.

$$z_{.025} = 1.96 = \frac{\bar{X}_U^* - 20}{10/\sqrt{200}}$$

$$\bar{X}_U^* = 20 + 1.96 \left(\frac{10}{\sqrt{200}} \right)$$

$$\bar{X}_U^* = 21.39$$

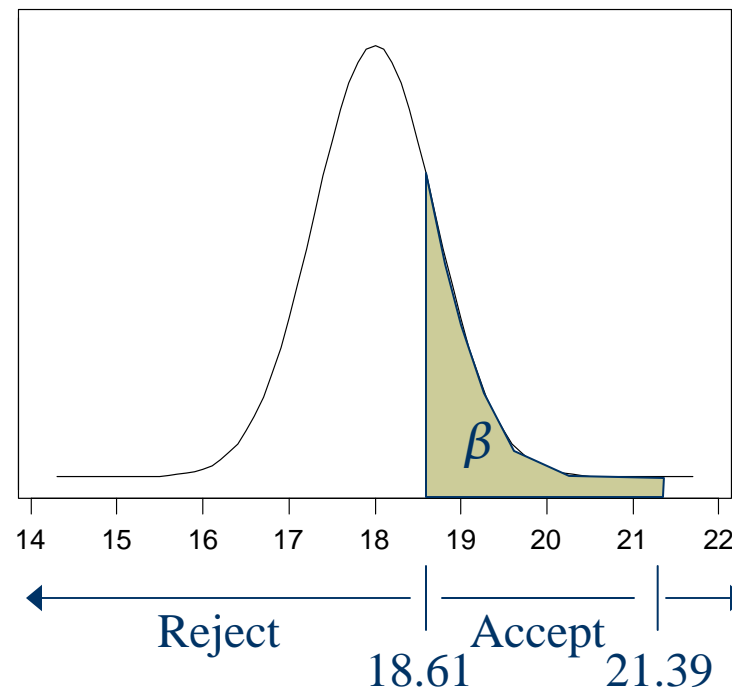
$$-z_{.025} = -1.96 = \frac{\bar{X}_L^* - 20}{10/\sqrt{200}}$$

$$\bar{X}_L^* = 20 - 1.96 \left(\frac{10}{\sqrt{200}} \right)$$

$$\bar{X}_L^* = 18.61$$

Probability of type II error illustrated

Distribution of the Sample Mean
when μ is actually 18



Calculating β

- ◆ The z values corresponding to the boundaries of the accept/reject region are computed at the right.
- ◆ The answer is $\beta = .1949$

$$z_0 = \frac{18.61 - 18}{10 / \sqrt{200}} = .86$$

$$z_1 = \frac{21.39 - 18}{10 / \sqrt{200}} = 4.79$$

$$\beta = \text{Prob}(.86 < z < 4.79)$$

$$\beta = .1949$$



Acceptance Sampling

- ◆ Acceptance sampling is a quality control tool used by organizations buying items in bulk from suppliers. The organization doing the purchasing tests a random sample of items to see if they meet specifications.
- ◆ Acceptance sampling is a widely used tool in business, and its use illustrates both the strengths and weaknesses of hypothesis tests.



World War II and Acceptance Sampling

- ◆ Commonly during America's wars, suppliers have tried to pass off shoddily made goods on government procurement officers.
- ◆ During the Civil War, there were numerous instances of shoe companies delivering boots that would immediately fall apart, etc.
- ◆ When America entered WWII, the military took steps to protect itself.

The military adopted acceptance sampling.

- ◆ Suppose a supplier of combat boots delivered a train load of boots to Hampton Roads.
- ◆ The supply contract usually specified an acceptable fraction of defective boots, such as 2%.
- ◆ Procurement officers would select a sample and perform a hypothesis test, such as the one at the right; if the shipment failed, it would be not be accepted.

$$H_0 : p \leq .02$$

$$H_A : p > .02$$



The Result

- ◆ The military's use of acceptance sampling during the war was credited with preventing the worst abuses of previous wars.
- ◆ After the war, demobilized procurement officers spread the technique in American industry and in classrooms where statistical techniques for business were taught.



A True Story



- ◆ The weaknesses of hypothesis testing, as described earlier, are as follows:
 - It ignores prior information.
 - It fails to consider costs and benefits.
 - It frequently ignores Type II errors.
- ◆ These weaknesses are nicely illustrated in an acceptance sampling incident that involved my father.



The problem

- ◆ Back in the 1960s, a company in Washington, D.C. specialized in microfilming newspapers for libraries.
- ◆ Since they microfilmed all the newspapers published in the United States, they used a great deal of Kodak microfilm.
- ◆ My father was an engineer for Kodak.



My father dispatched. . .

- ◆ A call came into Kodak one day reporting that this company had a malfunction in the machine that develops microfilm, and the local Kodak representative couldn't fix it.
- ◆ Dad went to investigate, despite it being outside his usual duties.
- ◆ When he arrived, they took him to the machine, where the developed microfilm was coming out still wet.



Why was the film wet??

- ◆ Dad looked at the machine and said, “There is a fan in there that blows air across the film to dry it. It must not be working.”
- ◆ “Oh no,” the managers countered, “there is a gauge here and it says the fan is working.”
- ◆ My father stuck his hand into the machine. “Your gauge is broken too, because there is no air moving in here.”



Which was all there was to the problem.

- ◆ One sidelight you'll appreciate: the manager's response was "Oh, we hired this college student, Jack, to help out over the summer. He put his hand inside and said the same thing, but since he was a college student we figured he didn't know anything."
- ◆ Dad now had the whole day to kill, so they gave him a tour of the factory.



The CEO was a Harvard MBA

- ◆ He was especially proud of his company's acceptance sampling scheme, and proceeded to tell my father about it.
- ◆ Since they used so much microfilm, they routinely tested shipments from Kodak to see if it was of acceptable quality – not more than a small percentage defective.



My father was thunderstruck.

- ◆ It was inconceivable to him that Kodak would *ever* produce film incapable of photographing black ink on white paper.
- ◆ He asked to visit the quality control lab, and the CEO led him there. The lab was full of cameras, color wheels, and open boxes of Kodak microfilm. A couple people worked there full time.
- ◆ “How long have you been doing this?” Dad asked.



More questions



- ◆ “Over five years,” the CEO replied.
- ◆ “How many rolls to you test?” he asked the chief of quality control.
- ◆ “Dozens every day,” was the reply.
- ◆ “How many rolls of defective film have you found in that time?” Dad continued.



Idiocy revealed!

- ◆ “None.”
- ◆ The CEO’s jaw dropped. He’d never thought to ask that obvious question.
- ◆ “When you get done with the testing, what do you do?” Dad asked.
- ◆ “We fill out this form and send it to Tom on the factory floor.”



Dad took the CEO to see Tom.

- ◆ “Do you recognize this form?” he asked Tom, when they found him on the factory floor.
- ◆ “Sure,” Tom replied, “I get one every day. I have no idea what it is for; I just throw them away.”



What did the CEO do wrong?

- ◆ **Prior Knowledge:** My father knew Kodak film would photograph black ink on white paper. The CEO should have caught on much more quickly.
- ◆ **Costs and Benefits:** What happens if Kodak delivers a roll of bad film? A newspaper must be redone: hardly a catastrophic event. Testing all that good film, however, was quite costly.
- ◆ **Mechanical and unthinking application of hypothesis tests, and a total lack of common sense.**



Fraud and Deceit

- ◆ An acquaintance offers to settle some dispute between you with a coin toss.
- ◆ You ask if the coin is fair, and he replies “Sure, I tested it myself. I tossed it 200 times and got 100 heads.”
- ◆ Is the person lying?

You ought to be suspicious

- ◆ On average, a fair coin comes up heads 100 times in 200 tosses, but the standard deviation of the number of heads is about 7.
- ◆ The chance of getting an answer this perfect is a binomial calculation with $n = 200$, $p = .50$, and $x = 100$; the probability works out to .056. Not terribly likely, but not wildly implausible.



This is basically a p-value

- ◆ But a peculiar one, since the rejection area is in the middle of the distribution. We reject the null of honesty when the results are “too good to be true.”
- ◆ Testing at $\alpha = .05$, we’d accept the null of honesty here, but only barely.

Change the numbers . . .

- ◆ Suppose the acquaintance had claimed to have tossed the coin 10,000 times and to have gotten exactly 5,000 heads.
- ◆ The odds of getting such a perfect outcome are small; binomial, $n=10,000$, $p=.50$, $x=5000$.
- ◆ The actual p-value is just 0.008, which should make you *very* suspicious.



Sampling error is oft forgotten

- ◆ Particularly by people who are falsifying their data.
- ◆ Therefore, *sample statistics that appear to perfectly support the author's hypothesis often signal that the author has falsified or tampered with the data.*

Speaking of too good to be true

- ◆ In Mendel's famous genetics experiments that you studied in high school, Mendel often reported sample statistics that corresponded *precisely* with the population parameter he was looking for.
- ◆ R.A. Fisher, a famous statistician, reviewed Mendel's work in the 1930s and concluded Mendel falsified his data.
- ◆ Reference: <http://www.nih.gov/about/director/ebiomed/mendel.htm>



Even if the population parameter was wrong!

- ◆ At one point, Mendel misunderstood the implications of his own theory, and incorrectly asserted that the ratio of one kind of offspring to another should be 2:1.
- ◆ Actually, his theory predicts 1.7:1.
- ◆ But his reported sample outcome was 2:1, an outcome with p-value of .0005 under the correct null hypothesis.



Many Geneticists revere Mendel

- ◆ Their furious reaction is that Mendel is the victim of a creationist conspiracy attempting to discredit Mendel's contribution to the theory of evolution.
- ◆ Repeated attacks on Fisher's conclusions have resulted in statisticians making some concessions.
- ◆ However, the probability of results as favorable as those Mendel reports are still thought to be .00003 and the consensus among statisticians is still that Mendel fabricated his data.



Example: Sir Cyril Burt

- ◆ The most prestigious, powerful, and influential psychologist of his generation.
- ◆ Chair of psychology at London's University College.
- ◆ Knighted by King George VI.
- ◆ Received the Thorndike award from the American Psychological Association.
- ◆ Reference: <http://www.discovery.org/lewis/bettleheim.html>



What did Burt believe?

- ◆ Burt's career was based upon his statistical studies of the intelligence of identical twins, showing that poverty was due to inferior intelligence of the working class.
- ◆ In the 1940s Burt was involved in setting up the British school system which segregated students on the basis of an IQ test they took at age eleven.

The evidence of fabrication

- ◆ After Burt died in 1971, a Princeton psychologist noted that in three different studies of different numbers of identical twins, Burt reported the same sample correlation of IQ scores *to the third decimal point*. This is extraordinarily improbable. The same problem appeared in Burt's reported correlation of height and weight.
- ◆ In 1976, the *London Sunday Times* tried to locate Burt's field investigators and co-authors, and concluded they never existed.



Burt still has his defenders

- ◆ A *Fortune* magazine article in 1990 hinted that Burt had been the victim of a vast left wing conspiracy of academics who wanted to discredit his belief that nature, not nurture, was the primary determinant of ability.
- ◆ Which tells you more about *Fortune* magazine than about Burt's critics.

“Women & Love”

- ◆ A 1987 book by self-styled feminist Shere Hite.
- ◆ According to Ms. Hite, the book was based on 4500 responses to questionnaires she distributed to women.
- ◆ Her basic conclusion: Men of every ethnicity and social group are the worst kind of *pond scum*.





Academics questioned the methodology . . .

- ◆ Hite said she mailed 100,000 questionnaires to a variety of women's groups in 43 states, ranging from feminist organizations to church groups to garden clubs.
- ◆ Hite admitted it was not a truly scientific survey: "It's 4500 people. That's enough for me."



It isn't a random sample, critics said.

- ◆ Recipients of the questionnaire were not selected randomly.
- ◆ Since only 4.5% were returned, response bias is an issue – only the most motivated responded.
- ◆ Some answers appeared peculiar, for instance 98% of all women wanted to make basic changes in their relationships. As one pollster noted “any question you asked that got 98% is either a wrong question or wrongly phrased.”

The critics are kind or gullible.

- ◆ Here is a representative table from her book.
- ◆ Note the amazingly constant fraction of discontented women; nearly 84% for each sub-group.

Basic dissatisfactions with the current emotional contract

84% of women are not satisfied emotionally with their relationships.

AGE		EDUCATION	
—%	under 18	86%	up to high school graduate
86%	18–34	85%	some college
85%	35–50	84%	college graduate
84%	51–70		
81%	71 and over		
INCOME: annual		OCCUPATION/EMPLOYMENT	
89%	under \$5,000	89%	homemaker and/or mother (full-time)
89%	\$6,000–\$14,000	75%	full-time employment
82%	\$15,000–\$39,000	84%	part-time employment
84%	\$40,000–\$74,000	88%	unemployed/student
76%	over \$75,000		
RACE/ETHNICITY		MARRIED/SINGLE	
84%	White	92%	single, never married
85%	Black	86%	divorced, separated, or widowed
84%	Hispanic	74%	married
82%	Middle Eastern	70%	1–5 years
83%	Asian-American	79%	6–15 years
84%	other	75%	16–25 years
		72%	over 25 years
		74%	



Sample sizes for some subgroups were very small

- ◆ In her tables, she reported only 0.3% of her respondents were Middle-eastern, 1.8% Asian, and 1.8% Hispanic.
- ◆ The implied sample size of Middle-eastern women is 14; of Asian women and Hispanic women is 81.

Sampling errors should be substantial with n this small.

- ◆ So even if exactly 84% of all Middle-eastern women were dissatisfied, you'd expect the sample proportion to be about 10% away.
- ◆ Or in the case of Hispanics or Asians, 4% away.

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\sqrt{\frac{.84(.16)}{14}} = .098$$

$$\sqrt{\frac{.84(.16)}{81}} = .041$$



So? She got lucky.

- ◆ There are over 100 pages of tables, and they *all* look pretty much the same.
- ◆ Non-random samples and response bias do not account for this kind of discrepancy.
- ◆ *Either* the true proportions are very similar over ethnic groups, and she was *unbelievably lucky* in her sampling, or . . .



Hmm . . .

- ◆ On page 806, Hite reports that 99% of all Middle eastern women say they would like their husband or lover to talk more about his feelings.
- ◆ With a sample size of 14, you can get 100% or 93%, but you simply *cannot* get 99%.

And the fun continues . . .

- ◆ On page 884 Hite's tables report that 96% of Middle Eastern women in *gay* relationships say they feel loved in a satisfying way.
- ◆ For 96% to be a possible number, you'd need a sample size of about 25, but there are not that many Middle eastern women in the sample!
- ◆ And if they are *all gay*, who are the Middle Eastern women dissatisfied with their men?
- ◆ You don't suppose she fooled *Time* magazine and made the data up, do you?

Would she really do that?

- ◆ Hmm . . . Let's see. Hite quit the the Ph. D. program in History at Columbia after being accused of plagiarism.
- ◆ She later posed nude for *Oui* and *Playboy* because, she told *Time*, “the money was good.”
- ◆ Her first two books earned her \$2.5 million dollars.
- ◆ But Hite sues people for complaining that her samples are not random and there is response bias.
- ◆ So you decide. . .