STA 570 Spring 2011

Lecture 15 *Thursday, March 10*

- **Hypothesis Tests**

Hypothesis Testing

- Fact: It is easier to *prove* that a parameter **isn't** equal to a particular value than it is to prove it **is** equal to a particular value
- Hypothesis testing: *Proof by contradiction*:
  - we set up the belief we wish to disprove as the **null hypothesis** (**H_0**) and the belief we wish to prove as our **alternative hypothesis** (**H_1**) (or: research hypothesis)

What about those errors?

Two possible errors:

- Type I error: Rejecting the null when we shouldn't have [P(Type I error) = α]

- Type II error: Not rejecting the null when we should have [P(Type II error) = β]

More Terminology

- Note: We set the probability that the correct decision is made when H_0 is true. This probability is typically termed $1-\alpha$.
- Thus, α is the probability of making a mistake when H_0 is true (rejecting when you should NOT reject).
- Once α is set, the cutoff is determined. The most common value for α is 0.05.
- Aside from making α small (it is the chance of a mistake) there is no absolute justification for this choice compared to others.

More Terminology

- The choice of α determines the cutoff, and thus the probability of making the correct decision when H_1 is true (when H_1 is true the correct decision is to reject the null hypothesis).
- This probability is called the ***power*** of the test.
- Thus, we want 1-α and the power to be high (close to 1).

Four Possibilities

- There are two possible states of the world – either the null hypothesis is true or the alternative hypothesis is true.
- You can make two decisions – either reject the null hypothesis or do not reject the null hypothesis.
- Thus, there are four possibilities. Two correspond to correct decisions and two are errors.

The four possibilities, with terminology

	Choose H_0 ("***do not reject H_0***")	Choose H_1 ("***reject H_0***")
H_0 is true	Correct answer	WRONG! (**Type I error**)
H_1 is true	WRONG! (**Type II Error**)	Correct answer

Type I and Type II errors

- Type I error corresponds to rejecting H_0 when H_0 is in fact true. The probability of making this mistake when H_0 is true is α
- Type II error corresponds to *not* rejecting H_0 when H_1 is in fact true. The probability of making this mistake when H_1 is true is 1 minus the power of the test.

Another Example

- A person claims to have ESP.
- You test him/her by having one of four images displayed on a computer screen to a second individual.
- The individual claiming to have ESP has to guess what is on the computer screen.
- Just by random guessing, you would expect the "ESP person" to get 25% of the images correct.
- Thus, a claim of ESP would be strengthened if the person guessed more than 25% of the images correctly.

ESP Example in Statistical Terms

- Suppose you show the person n=100 images. Each guess is correct or incorrect (binary, dichotomous).
- The proportion of correct guesses is p-hat.
- The true proportion of correct guesses (unknown to us) is p.
- Under random guessing, p=0.25
- If the person has ESP, p>0.25

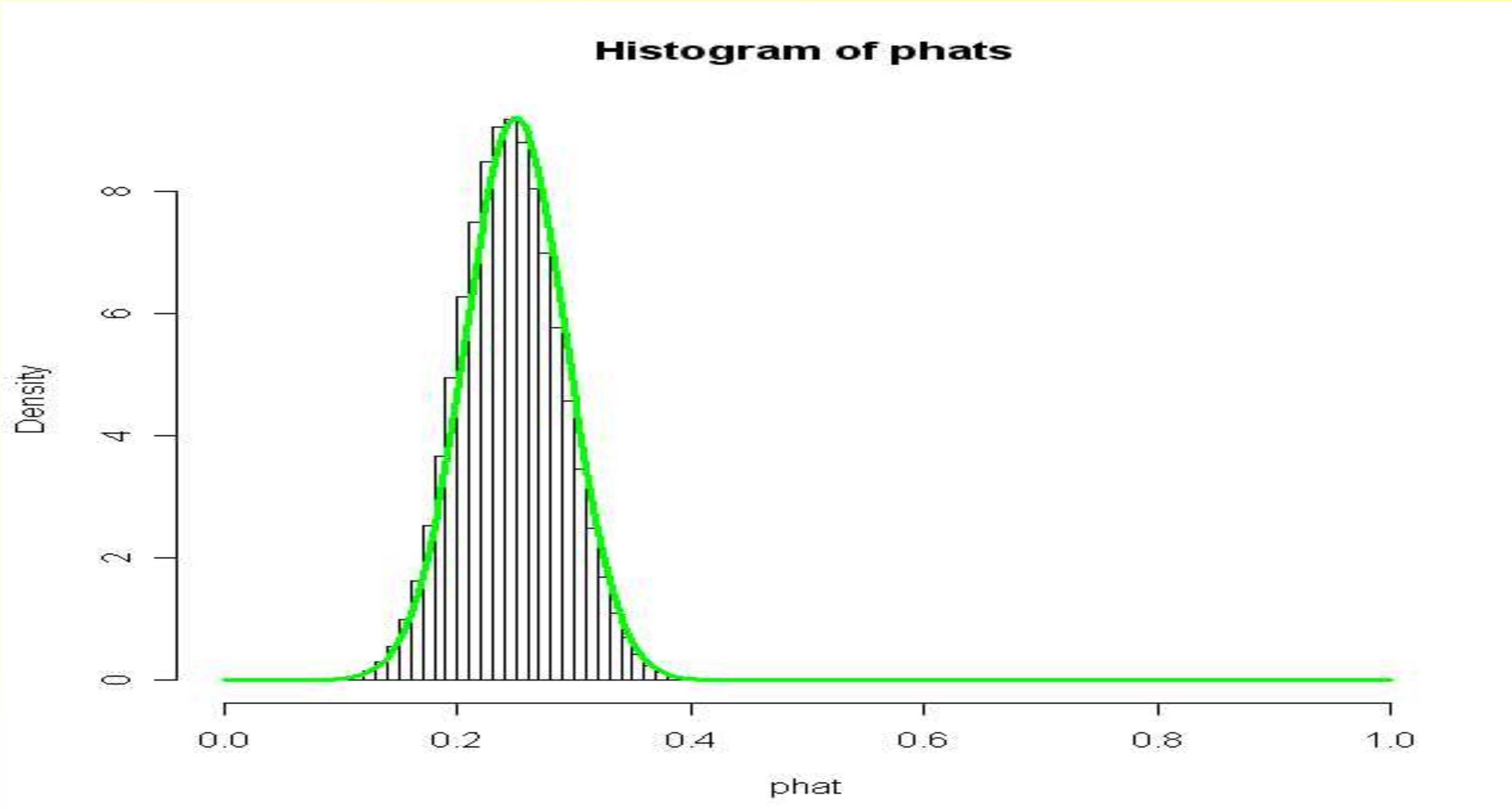
Null and Alternative Hypotheses

- The cutoff is based on the null hypothesis.
- H_0 : p=0.25 allows us to find a sampling distribution.
- The hypothesis “p>0.25” would not determine a “p” and a sampling distribution.
- It is a ***composite, one-sided alternative.***
- The alternative hypothesis is the “interesting conclusion” which would get the paper published.

Null Distribution

- Under the null hypothesis, p=0.25.
- n=100.
- Thus, the null distribution is normal with mean 0.25 and standard deviation sqrt (0.25*0.75/100)=0.0433).
- The null distribution here is the distribution of correct guesses under random guessing.

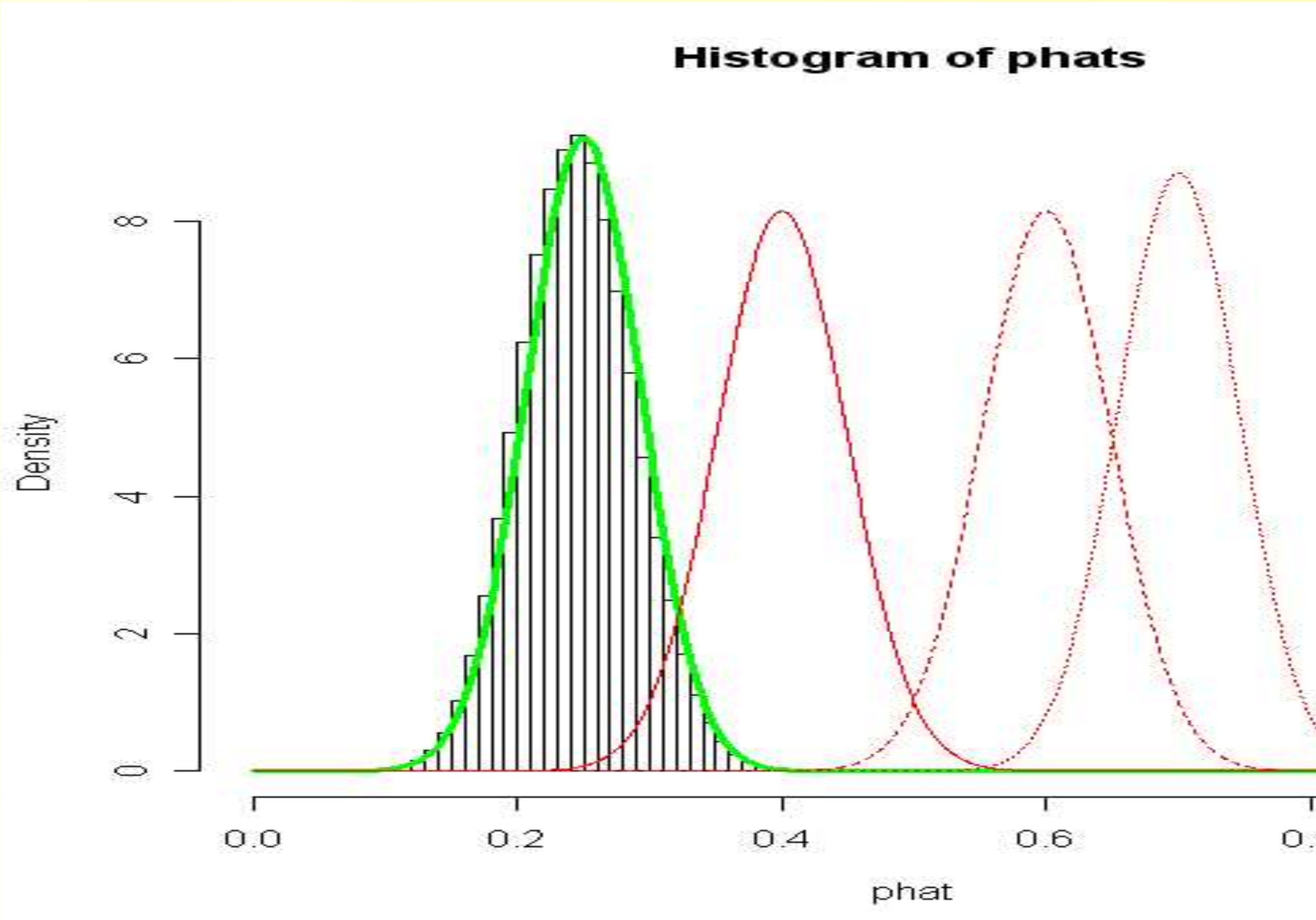
Sampling Distribution Under $H_0 : p=0.25$



Difficulty of Composite Alternative Hypotheses

- Finding the sampling distribution under H_1 is impossible.
- H_1 : $p>0.25$ only specifies a range for p.
- The true proportion p might be 0.8, it might be 0.3, it might be 0.99, we don’t know.
- This is different from the “Bill vs. Maria” where the two possible models where exactly specified.

Any of the Red Curves Could Be the Alternative Distribution



The alternative hypotheses contains a range of values for p. The sampling distributions for three of these values are shown with the null distribution.

Where to Place the Cutoff?

- The alternative hypothesis is designed to be the "interesting conclusion", thus the statistical test is designed to avoid mistakenly choosing H_1.
- Thus, we give H_0 the "benefit of the doubt" by choosing a small probability of type I error.

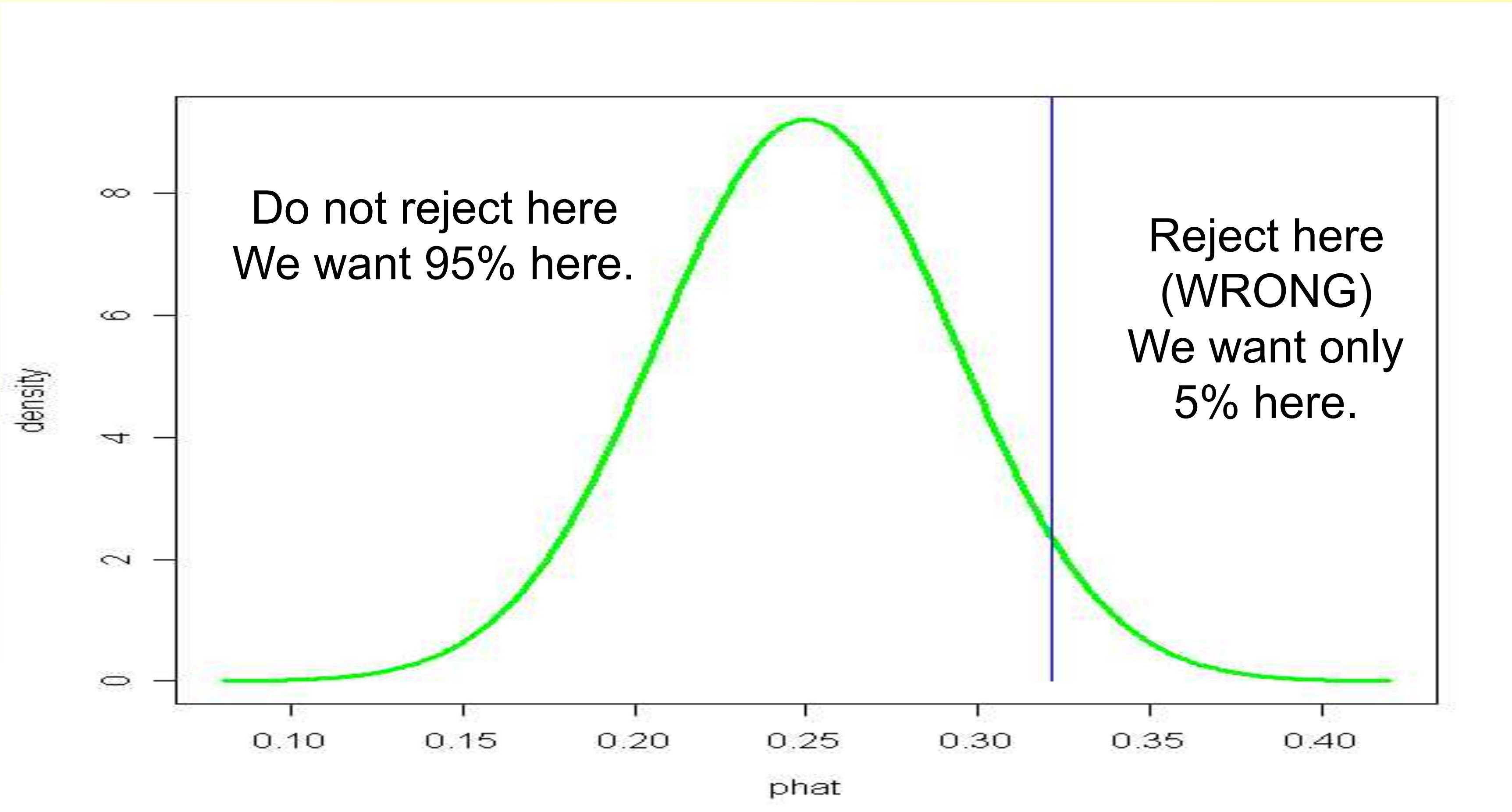
Why Give H_0 the Benefit of the Doubt?

- In the ESP example, we probably would not be convinced if someone did just a little better than 25%.
- While 25% is the expected number of correct guesses, you could do a little better based on chance alone.
- So how much better does the person have to do before you'd start to believe them?

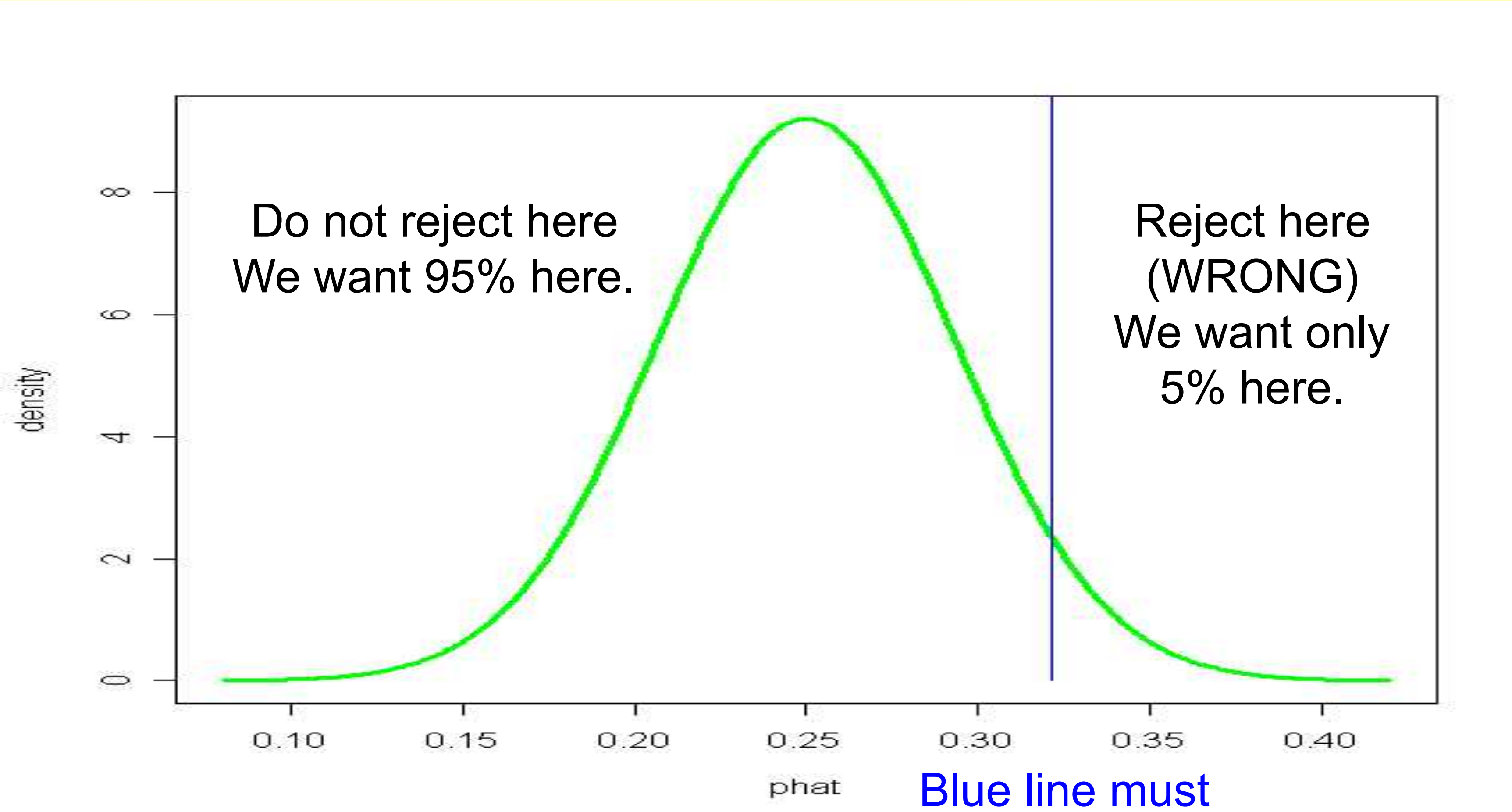
Finding the Cutoff

- Again let's choose $\alpha=0.05$
- The null hypothesis contains $p=0.25$ while the alternative hypothesis contains $p>0.25$. Thus, we reject H_0 for large values of phat.
- We need 95% of the null distribution below the cutoff.

Null Distribution



Null Distribution

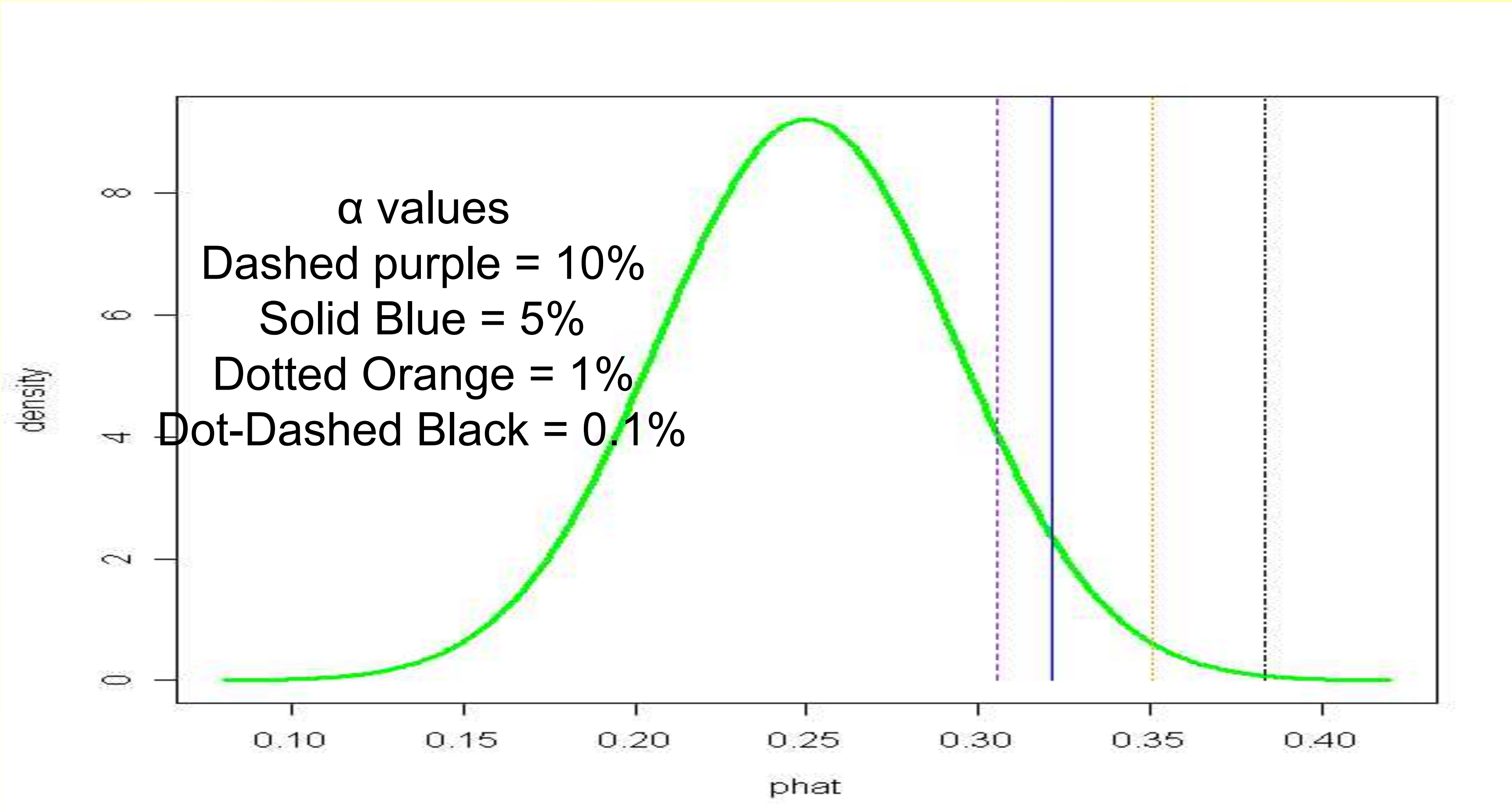


Blue line must be the 95^{th} percentile

Finding the 95^{th} Percentile

- The null distribution is normal with mean 0.25 and standard deviation 0.0433
- The corresponding Z-score for the 95^{th} percentile is 1.64 and thus the cutoff is (1.64)(0.0433)+0.25=0.3210.
- This cutoff says:
  - Getting a little more than 25% correct out of 100 isn't enough to convince us.
  - Our benefit of the doubt says we don't believe H_1 until you get at least 32.10% correct.

Different α Values Result in Different Cutoffs



Type I and Type II Errors

	Choose H_0 ("***do not reject H_0***")	Choose H_1 ("***reject H_0***")
H_0 is true	Correct answer	WRONG! (**Type I error**)
H_1 is true	WRONG! (**Type II Error**)	Correct answer

Type I and Type II Errors

- We reject p=0.25 if p-hat is greater than 0.3210.
- The chance of making a type I error (rejecting H_0 when H_0 is in fact true) is 0.05.
- Calculating the type II error probability is more complicated.
- For each different value in the alternative, the type II error probability is different.
- For values close to 0.25 (e.g., 0.26), the type II error probability is high because the sampling distribution overlaps much with the null distribution
- For values far from 0.25 (e.g., 0.9), the probability of type II error is low.

More on Type II Error

- One issue that we will study is how many observations should we collect to achieve particular error probabilities.
- *Power and sample size calculations.*

Review

- We have *n* observations. The true proportion, *p*, is unknown.
- We have a null value of $H_0 : p=p_0$, and depending on the setting may want to test $H_1 : p>p_0$ or $H_1 : p<p_0$, or $H_1 : p \neq p_0$.
- The alternative depends on the "scientifically interesting" conclusion.
- It may only matter if p is above p_0, or p below p_0, or perhaps simply being different.

Review

- We use the data (specifically p-hat, the sample proportion) to make one of two conclusions.
- One conclusion is to reject H_0, indicating that the data are not consistent with H_0.
- The other possible conclusion is to "not reject H_0", which indicates the data is consistent with H_0.
- *This is not the same as saying H_0 is true, since it is impossible to distinguish, with finite data, "$p=p_0$" from "p very close to p_0."*

Significance Tests: Summary

- A significance test checks whether data agrees with a hypothesis
- A hypothesis is a statement about a characteristic of a variable or a collection of variables
- If the data is very unreasonable under the hypothesis, then we will reject the hypothesis
- Usually, we try to find evidence ***against*** the hypothesis

Logical Procedure

1. State a hypothesis that you would like to find evidence against
2. Get data and calculate a statistic (for example: sample mean)
3. The hypothesis (for example: population mean equals 5) determines the sampling distribution of our statistic
4. If the calculated value in 2. is very unreasonable given 3., then we conclude that the hypothesis was wrong

Significance Test

- A ***significance test*** is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis
- Data that fall far from the predicted values provide ***evidence against the hypothesis***

Elements of a Significance Test

- Assumptions
- Hypotheses
- Test Statistic
- P-value
- Conclusion

Assumptions

- What type of data do we have?
  - Qualitative or quantitative?
  - Different types of data require different test procedures
- What is the population distribution?
  - Is it normal? Symmetric?
  - Some tests require normal population distributions
- Which sampling method has been used?
  - We always assume simple random sampling
  - Other sampling methods are discussed in STA 675
- What is the sample size?
  - Some methods require a minimum sample size (like $n=30$)

Hypotheses

- The ***null hypothesis (H_0)*** is the hypothesis that we test (and try to find evidence against)
- The name null hypothesis refers to the fact that it often (not always) is a hypothesis of “no effect” (no effect of a medical treatment, no difference in characteristics of countries, etc.)
- The ***alternative hypothesis (H_a)*** is a hypothesis that contradicts the null hypothesis
- When we reject the null hypothesis, the alternative hypothesis is judged acceptable
- Often, the alternative hypothesis is the actual research hypothesis that we would like to “prove” by finding evidence against the null hypothesis (proof by contradiction)

Hypotheses

The hypothesis is always a statement about one or more population parameters.

Test Statistic

- The ***test statistic*** is a statistic that is calculated from the sample data
- Often, the test statistic involves a point estimator of the parameter about which the hypothesis is stated
- For example, the test statistic may involve the sample mean or sample proportion if the hypothesis is about the population mean or population proportion

P-Value

- How unusual is the observed test statistic when the null hypothesis is assumed true?
- The ***P-value*** is the probability, assuming that H_0 is true, that the test statistic takes values at least as contradictory to H_0 as the value actually observed
- The smaller the P-value, the more strongly the data contradict H_0

Conclusion

- In addition to reporting the P-value, a formal decision is made about rejecting or not rejecting the null hypothesis
- Most studies choose a cutoff of 5%.
- This corresponds to rejecting the null hypothesis for P-values smaller than 0.05.
- Smaller P-values provide more significant evidence against the null hypothesis
- “The results are significant at the 5% level”

Elements of a Significance Test

- Assumptions
  - Type of data, population distribution, sample size
- Hypotheses
  - Null and alternative hypothesis
- Alpha-level (Type I error probability)
  - Specify alpha-level before looking at data
  - Alpha-level determines rejection region
- Test Statistic
  - Compares point estimate to parameter value under the null hypothesis
- P-value
  - Uses sampling distribution to quantify evidence against null hypothesis
  - Small P is more contradictory
- Conclusion
  - Report P-value
  - Rejection if test statistic in rejection region or P-value < alpha-level

P-Value

- How unusual is the observed test statistic when the null hypothesis is assumed true?
- The ***p-value*** is the probability, assuming that H_0 is true, that the test statistic takes values at least as contradictory to H_0 as the value actually observed
- *The **p-value** is **not** the probability that the hypothesis is true*
- The smaller the *p*-value, the more strongly the data contradict H_0

Alpha-Level

- Alpha-level (significance level) is a number such that one rejects the null hypothesis if the *p*-value is less than or equal to it.
- Often, alpha=0.05
- Choice of the alpha-level reflects how cautious the researcher wants to be
- Significance level alpha needs to be chosen ***before*** analyzing the data

Rejection Region

- The rejection region is a range of values such that if the test statistic falls into that range, we decide to reject the null hypothesis in favor of the alternative hypothesis

Type I and Type II Errors

- Type I Error: The null hypothesis is rejected, even though it is true.
- Type II Error: The null hypothesis is not rejected, even though it is false.

Type I and Type II Errors

- Terminology:
  - ***Alpha*** = Probability of a Type I error
  - ***Beta*** = Probability of a Type II error
  - ***Power*** = 1 – Probability of a Type II error
- The smaller the probability of Type I error, the larger the probability of Type II error and the smaller the power
- If you ask for very strong evidence to reject the null hypothesis, it is more likely that you fail to detect a real difference

Type I and Type II Errors

- In practice, alpha is specified, and the probability of Type II error could be calculated, but the calculations are usually difficult
- **How to choose alpha?**
- If the consequences of a Type I error are very serious, then alpha should be small.
- For example, you want to find evidence that someone is guilty of a crime
- In exploratory research, often a larger probability of Type I error is acceptable
- If the sample size increases, both error probabilities decrease

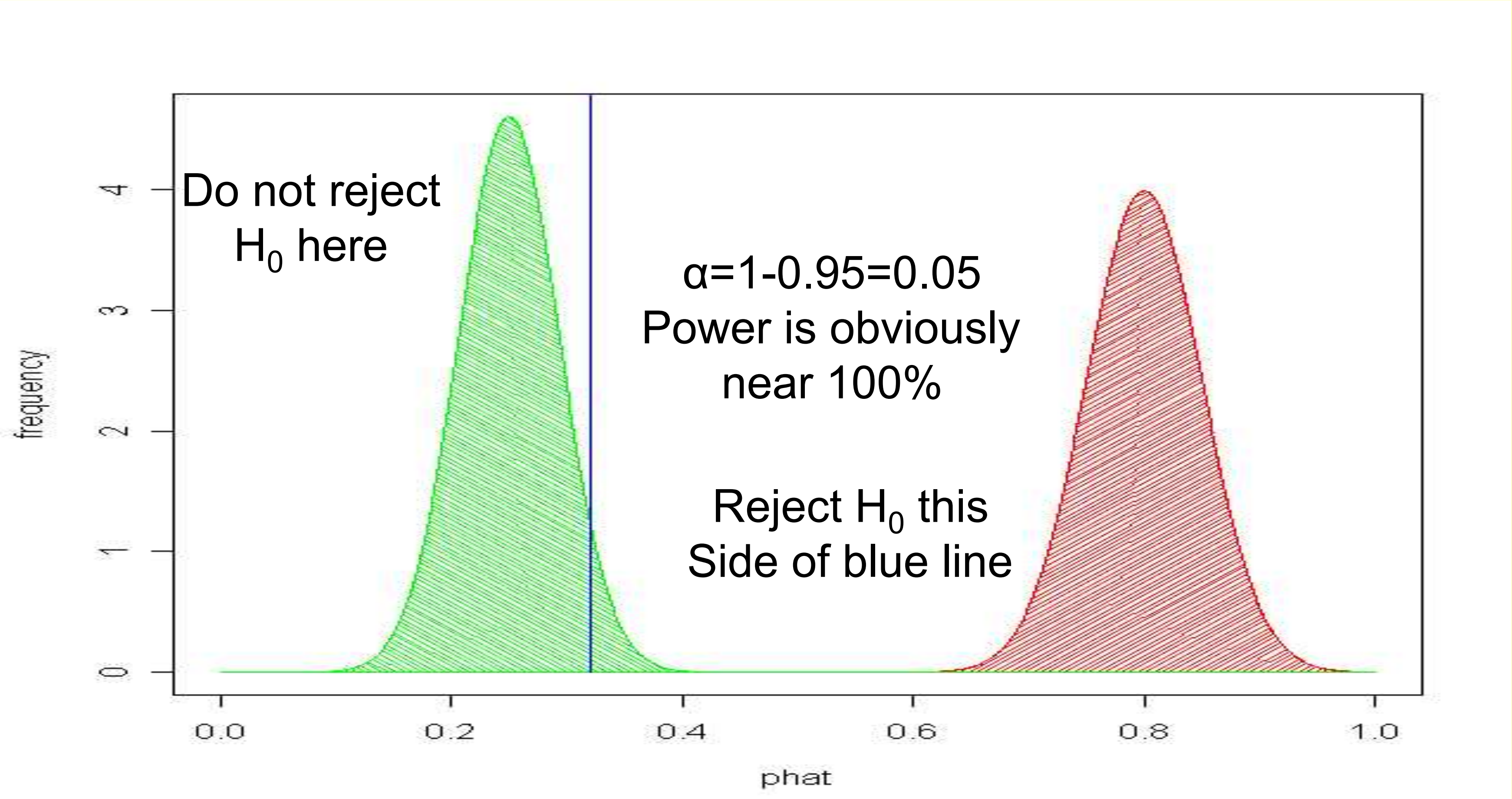
Power Calculations
Recall our ESP Example

- Hypothesis H_0 : p=0.25 against
  H_1 : p>0.25.
- The null distribution is normal with mean 0.25 and standard deviation sqrt (0.25*0.75/n)
- The cutoff is the 1-α percentile of this null distribution. We reject H_0 if p-hat is above this cutoff.

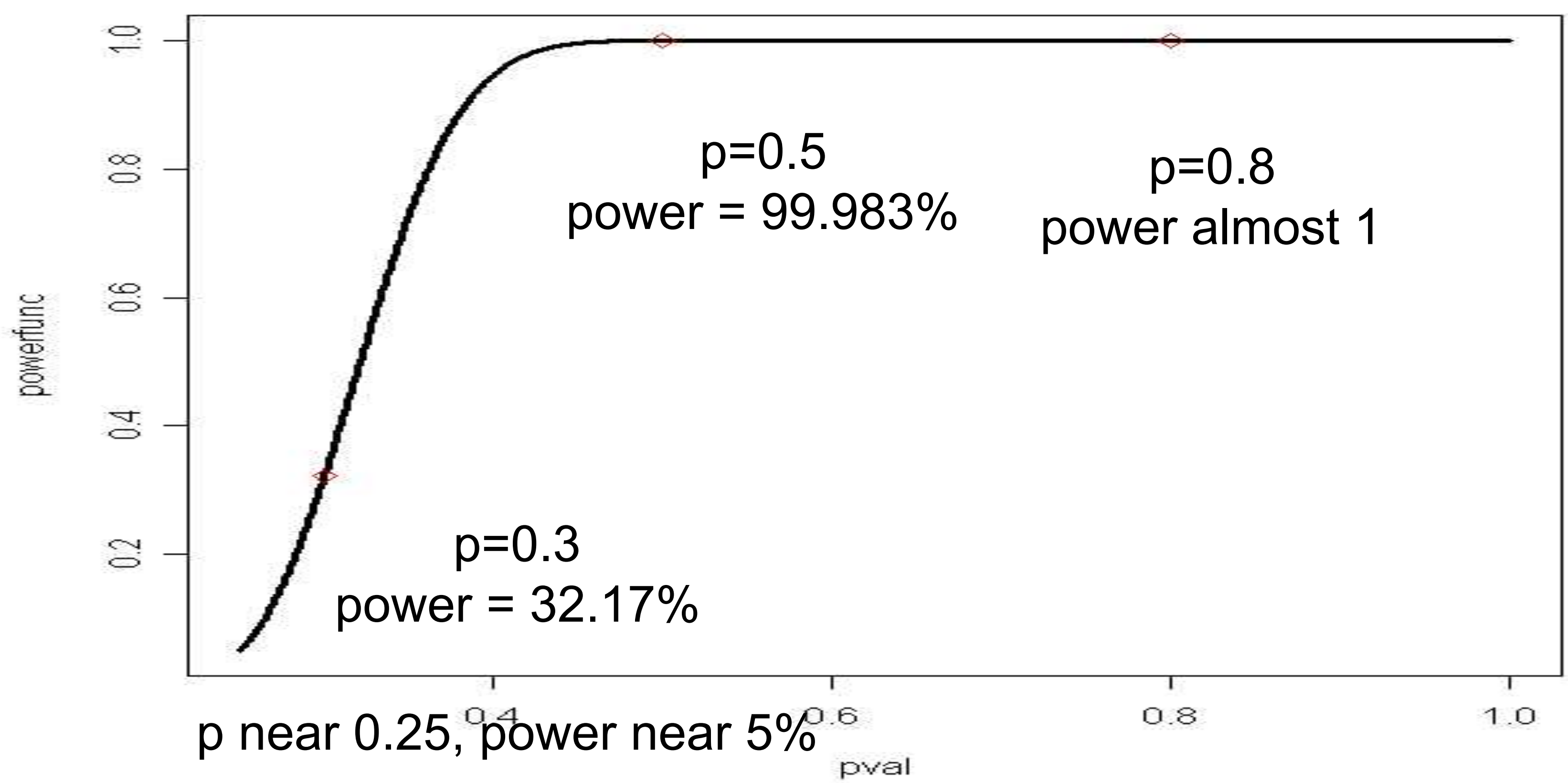
Power Function

- What about the power of the test?
- Unfortunately under H_1 we know nothing more about p than p>0.25
- BUT we can compute the power for ***each*** p>0.25

What if the Alternative Is H_1:p=0.8?



Power Function



Questions…

- What happens to the power function when we use $\alpha=0.001$ instead of $\alpha=0.05$?
- What happens to the power function when the sample size n is increased?
- How can we be sure we get a particular amount of power in our experiment?

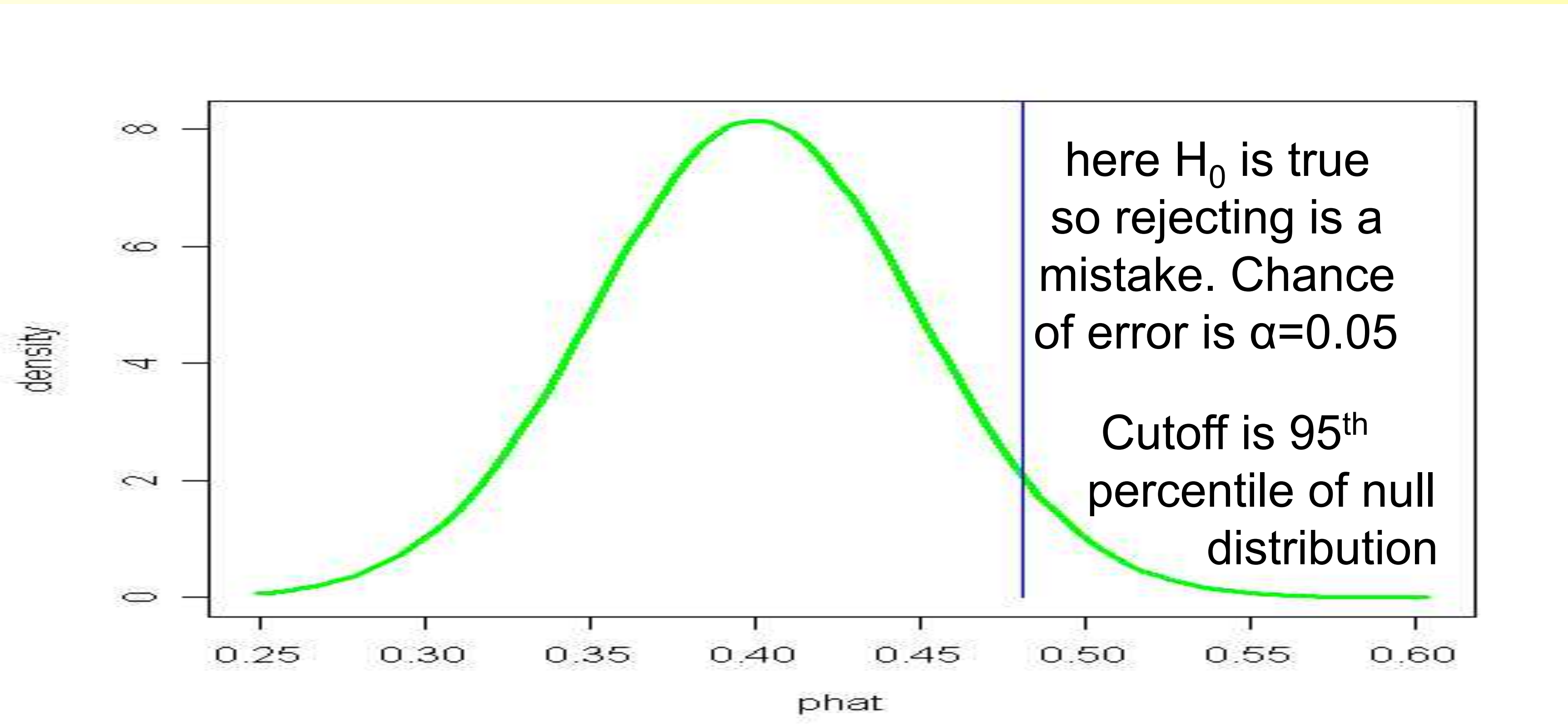
Power – Why Important?

- Example
  - A pharmaceutical company knows that the old treatment for a disease cures 40% of the people.
  - They hope their new treatment is better.
  - They hope they can get the cure rate to 45%.

Power Example Continued

- Our hypothesis test will test the null hypothesis H_0 : p=0.4 (the old treatment proportion) against H_1 : p>0.4 (we want our treatment to do better, hence this alternative).
- We intend to give the treatment to 100 people, and using α=0.05
- Cutoff?
- 95th percentile of the null distribution.
- Z=1.64, thus Y = (0.049)(1.64) + 0.4 = 0.4804

Null Distribution Centered at p_0=0.4



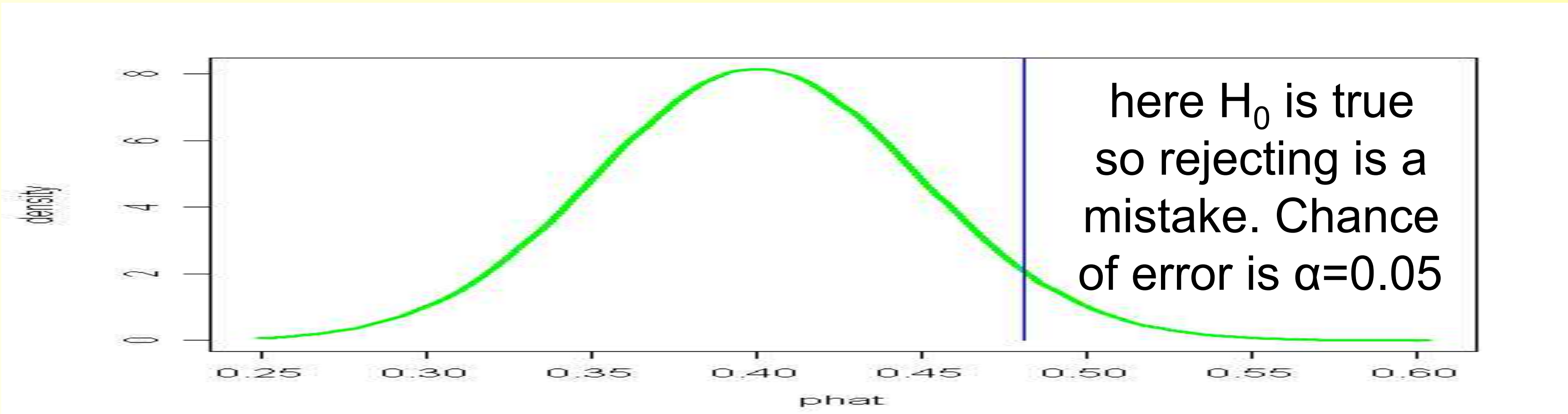
do not reject this side of blue line

reject this side of blue line

Can this Experiment Find Anything?

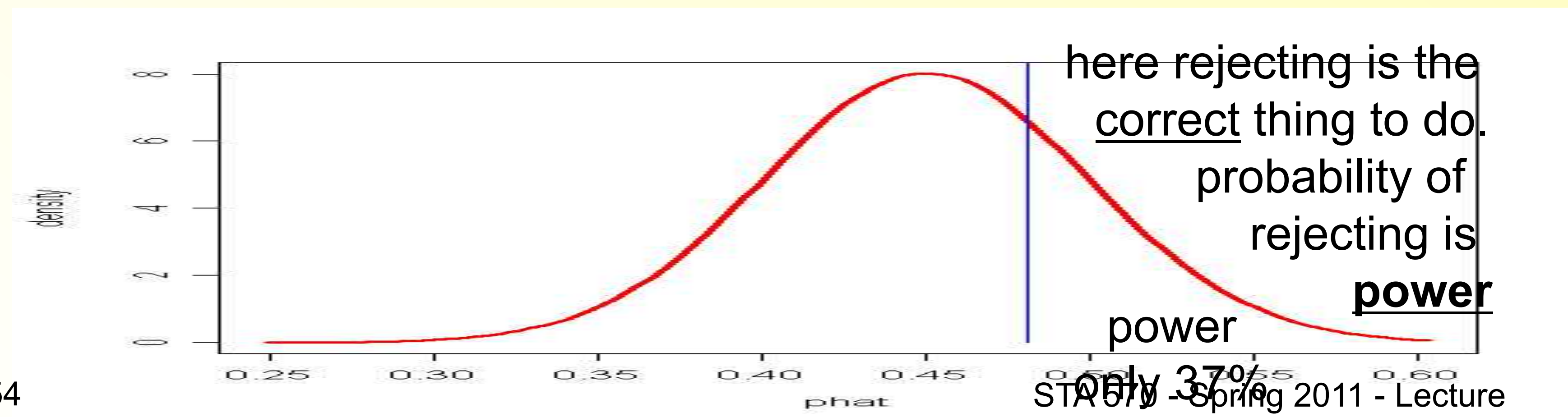
- We only are ***guessing*** our treatment can get the cure rate to 45%.
- What is the power for 45%?
- Remember, the power is the chance that, ***when p=0.45***, we reject H_0 (the right decision in that case).
- We reject when p-hat>0.4804.
- Thus, we need the probability that p-hat is greater than 0.4804, ***given*** p=0.45.

Green curve is distribution for p=0.4
Red curve is distribution for p=0.45



do not reject this side of blue line

reject this side of blue line



What This Means…

- Suppose our treatment works (this is the assumption under which the power is calculated).
- Then we only have a 37.09% chance of getting a "reject H_0" conclusion.
- That is not great. We could have a beneficial treatment and miss it.
- Solution – choose a higher sample size.

Power Equation

- We need to solve for the n that satisfies

$$\frac{0.4899(1.645)}{\sqrt{n}} + 0.4 = \frac{0.4975(-1.28)}{\sqrt{n}} + 0.45$$

$$\frac{1.4427}{\sqrt{n}} = 0.05$$

- n=832.5, so n must be at least 833