# STA 570 Spring 2011

Lecture 14 *Tuesday, March 8*

- **Hypothesis Tests**

# Midterm Exam

- Recall:
  - **Midterm Exam 20%**
  - Final Exam 30%
  - Homework 30%
  - Quizzes 20%

| | Example 1 | Example 2 |
|---|---|---|
| **Midterm** | **78** | **97** |
| Final | 90 | 75 |
| Homework | 96 | 75 |
| Quizzes | 94 | 73 |
| ***Overall*** | ***90*** | ***79*** |

# 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)

# Analogy: Court trial

- In American court trials, the jury is instructed to think of the defendant as innocent:

  $H_0$: Defendant is innocent

- District attorney, police involved, plaintiff, etc., bring every evidence, hoping to prove

  $H_1$: Defendant is guilty

- Which hypothesis is correct?
- Does the jury make the right decision?

# Back to statistics …

Critical Concepts

- Two hypotheses:  the null and the alternative
- Process begins with the assumption that the null is *true*
- We calculate a test statistic to determine if there is enough evidence to infer that the alternative is true
- Two possible decisions:
  - Conclude there is enough evidence to reject the null (and therefore accept the alternative)
  - Conclude that there is not enough evidence to reject the null
- Two possible errors?

# What about those errors?

Two possible errors:

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

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

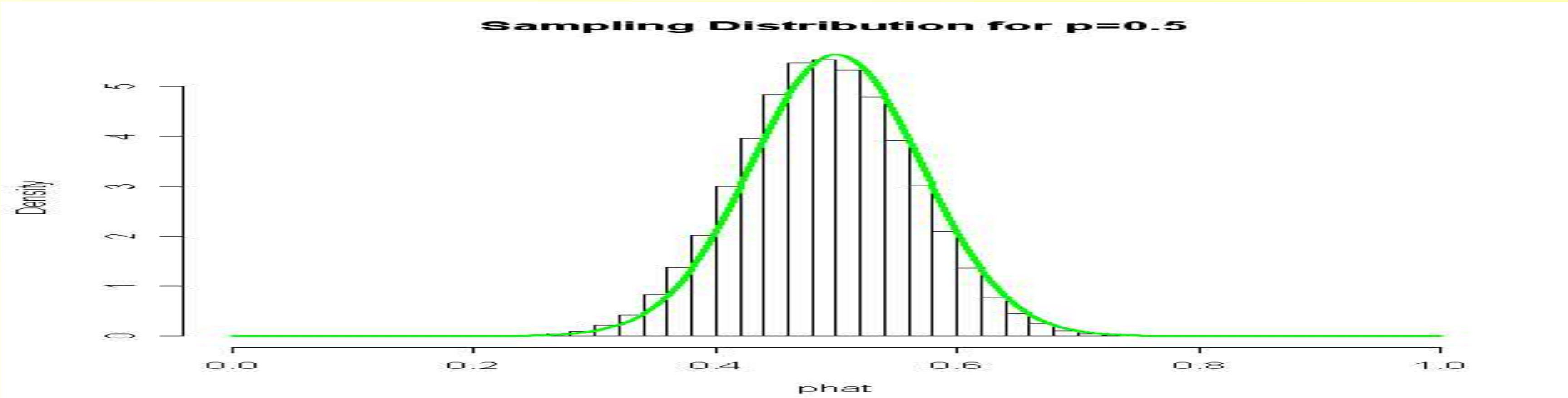
# Hypothesis Testing Example

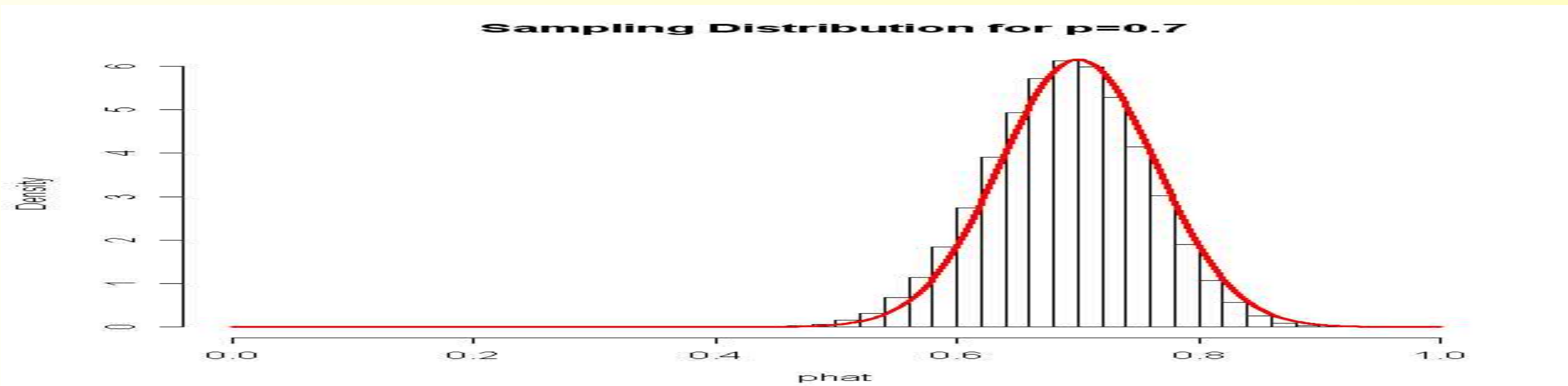
- You are given a coin. You know the coin might be fair (50% heads, 50% tails), but the coin might also be weighted (70% heads, 30% tails).
- You flip the coin 50 times and get 29 heads. Is the coin fair or weighted?

# Sampling Distributions for Each Kind of Coin

- Suppose the coin is weighted, so p=0.7.
- If you flip the coin n=50 times, the sampling distribution of the proportion of heads p-hat is normal with mean p=0.7 and standard deviation sqrt(p(1-p)/n) = sqrt(0.7*0.3/50) = 0.0648.
- If the coin is fair, with p=0.5, the sampling distribution of p-hat is normal with mean p=0.5 and standard deviation sqrt (0.5*0.5/50) = 0.0707.

# Need Cutoff to Separate These Groups.



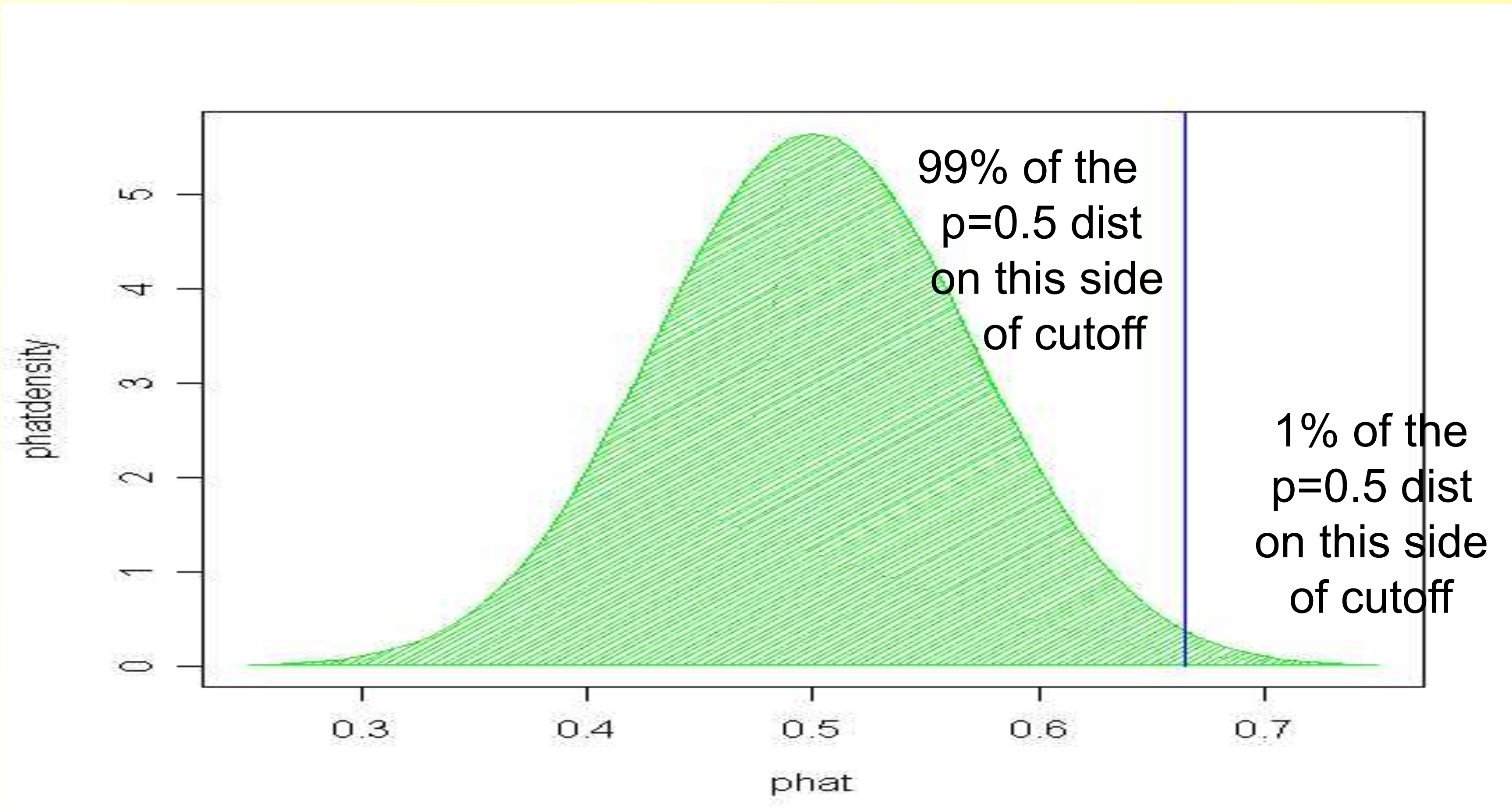
# Determining a Cutoff

- We choose the cutoff to either
  - set the proportion of fair coins that are classified correctly, or
  - set the proportion of weighted coins that are classified correctly.
- We assume that the coin is fair (null hypothesis), and we try to find evidence against this.
- That is, $H_0 : p=0.5$ vs. $H_1 : p=0.7$

# Example, contd.

- Return to n=50, which resulted in
  p-hat ~ N(0.5,0.0707) for $H_0$ : p=0.5
  p-hat ~ N(0.7,0.0648) for $H_1$ : p=0.7
- Let us determine a cutoff where the probability that a p=0.5 coin is classified correctly is 99%.
- We classify a coin as p=0.5 if p-hat is below the cutoff, so we need a cutoff that is the 99$^{th}$ percentile of a N(0.5,0.0707) distribution.

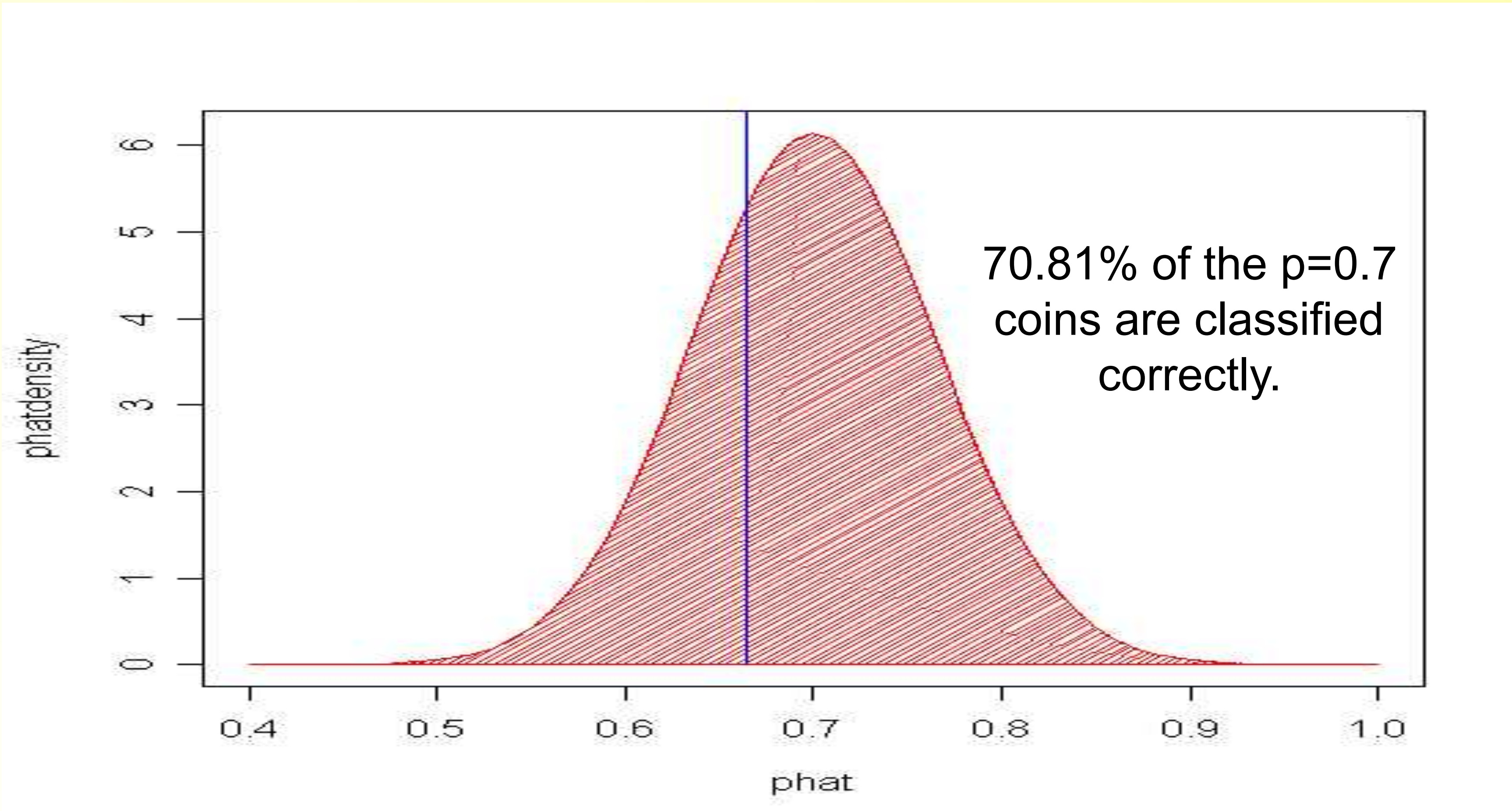
# A cutoff of 0.6645 results in 99% correct classification of p=0.5 coins.



# What about the weighted (p=0.7) coins?

- For the chosen cutoff of 0.6645, we can also find the probability a p=0.7 coin is correctly classified.
- We need to find the probability a N (0.7,0.0648) is greater than the cutoff of 0.6645, which is 70.81%.

# 70.81% of the p=0.7 coins are classified correctly with cutoff=0.6645



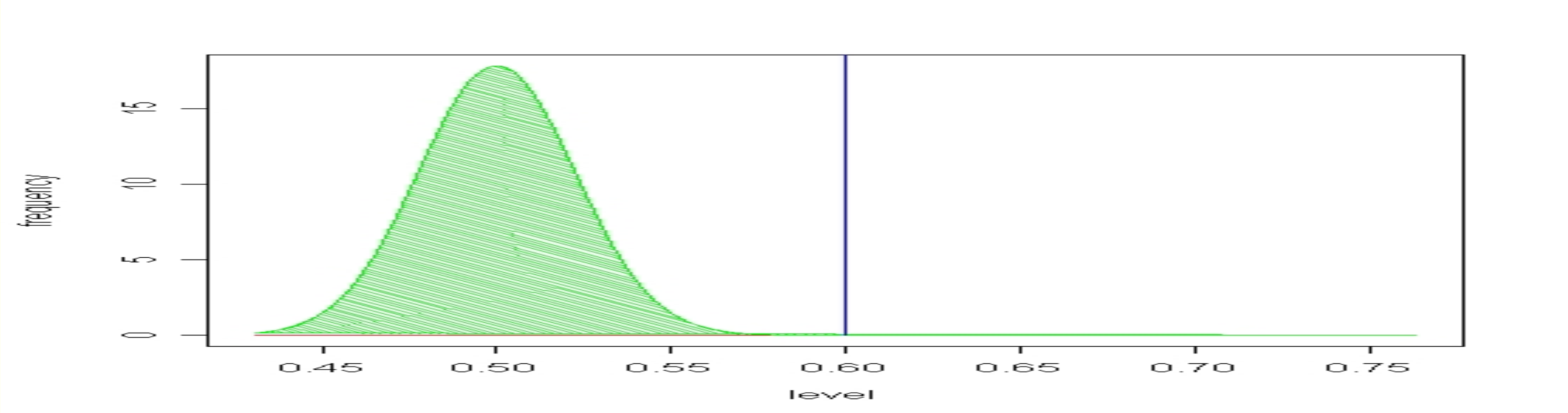
# Larger Sample Sizes Increase Accuracy

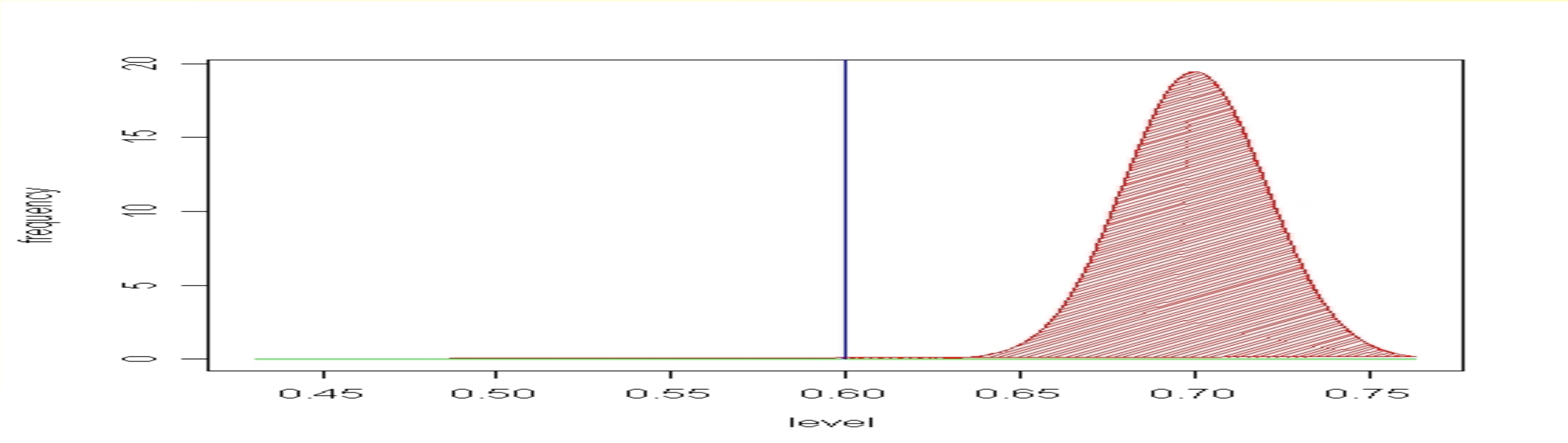
- Simple way to make a stronger test:
- Flip the coin more than 50 times.
- This changes both sampling distributions, reducing the standard deviation of both.

# Suppose we flip the coin n=500 times

- When p=0.5, the sampling distribution is normal with mean 0.5 and standard deviation sqrt(0.5*0.5/500) = 0.0224
- When p=0.7, the sampling distribution is normal with mean 0.7 and standard deviation sqrt(0.7*0.3/500) = 0.0205
- The two groups are now well separated.

# Now the groups are well separated



# Terminology

- Choosing the group corresponding to $H_1$ is called "rejecting the null hypothesis"
- Choosing the group corresponding to $H_0$ is called "not rejecting the null hypothesis"

# 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, $\alpha$ is the probability of making a mistake when $H_0$ is true (rejecting when you should NOT reject).
- Once $\alpha$ is set, the cutoff is determined. The most common value for $\alpha$ is 0.05.
- Aside from making $\alpha$ 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

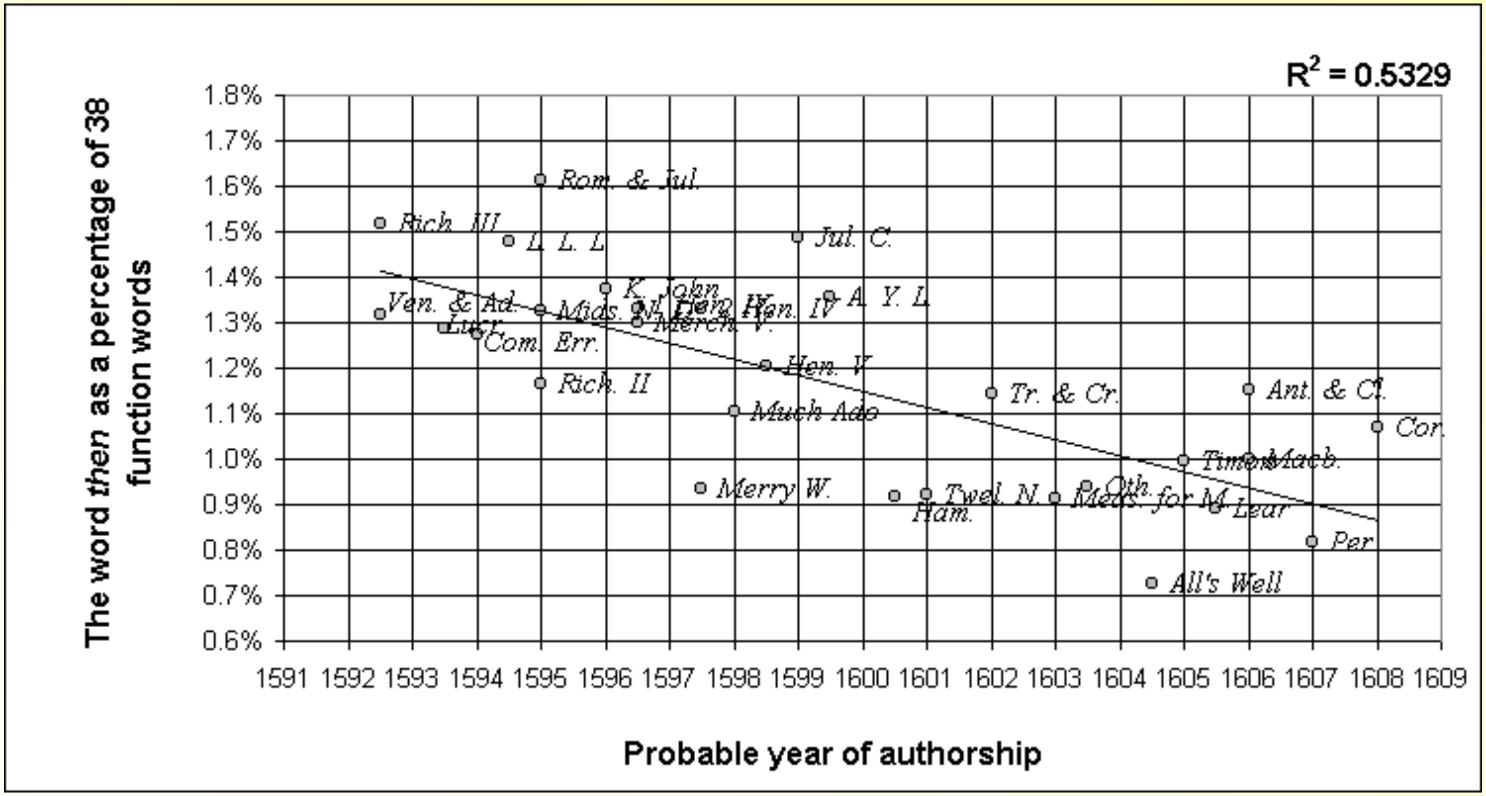
| | Choose $H_0$ ("***do not reject $H_0$***") | Choose $H_1$ ("***reject $H_0$***") |
|---|---|---|
| $H_0$ is true | Correct answer | WRONG! (**Type I error**) |
| $H_1$ is true | WRONG! (**Type II Error**) | Correct answer |

# 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 $\alpha$
- 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.

# By the way..regression is also useful for this

- Which of the Shakespeare Sonnets were really within the Shakespeare canon?



# Example

- Suppose you have a document that may or may not have been written by author Bill
- You observe that Bill uses a particular form X of the word 80% of the time
- The document in question has 84 instances of the word choice, and word X is used 58 times.

# Example continued

- Null hypothesis $H_0$: p=0.8 vs. $H_1$ : p≠0.8
- The ***null distribution*** (when $H_0$ is true) is normal with mean 0.8 and standard deviation sqrt(0.8*0.2/84)=0.0436.
- The null distribution is the sampling distribution of the sample statistic when the null hypothesis is assumed true.

# Example continued

- We have the null distribution N(0.8,0.0436)
- Let's choose $\alpha=0.05$
- We will reject $H_0$ for p-hats far away from 0.8.
- The cutoff to be the $2.5^{th}$ and the $97.5^{th}$ percentile of the null distribution. This corresponds to Z=(-1.96) and Z=1.96, and a cutoff of
- (-1.96)(0.0436)+0.8=0.7145 and
- (1.96)(0.0436)+0.8=0.8855

# Example continued

- (-1.96)(0.0436)+0.8=0.7145 and
- (1.96)(0.0436)+0.8=0.8855
- Our ***rejection region*** consists of those values of p-hat that are not between 0.7145 and 0.8855
- Our actual observed phat was 58/84=0.6904.
- Thus our conclusion is to reject $H_0$. We would conclude that $H_1$ is true and that Bill did not write the document.

# Example continued

- How likely is 58/84=0.6904 under the null hypothesis?
- The z-score is (0.6904-0.8)/0.0436= –2.51.
- Beyond (Here="below") a z-score of –2.51 is probability 0.00597.
- All values ***at least as extreme as*** a z-score of -2.51 have together probability 2 x 0.00597 = 0.0119.
- This is the ***P-value: The probability, assuming that the null hypothesis is true, of observing anything at least as extreme as what we actually observed.***
- ***"As extreme" = "providing as much, or more evidence against the null hypothesis"***

# Example continued

- By construction, the probability of type I error is $\alpha=0.05=5\%$. What is the probability of type II error?
- Type II error is not rejecting $H_0$ when in fact $H_1$ is true.
- Assume that another author, Maria, could have written the manuscript, and she chooses form X 60% of the time.
- The sampling distribution for her would be the ***alternative distribution*** (when $H_1$ is true), normal with mean 0.6 and standard deviation sqrt(0.6*0.4/84) =0.0535.
- We need to find the probability that the alternative distribution places above the cutoff 0.7145.
- The Z-score for a N(0.6,0.0535) is $Z=(0.7145-0.6)/0.0535=2.14$ and the probability of type II error is 0.01609=1.6%

# 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

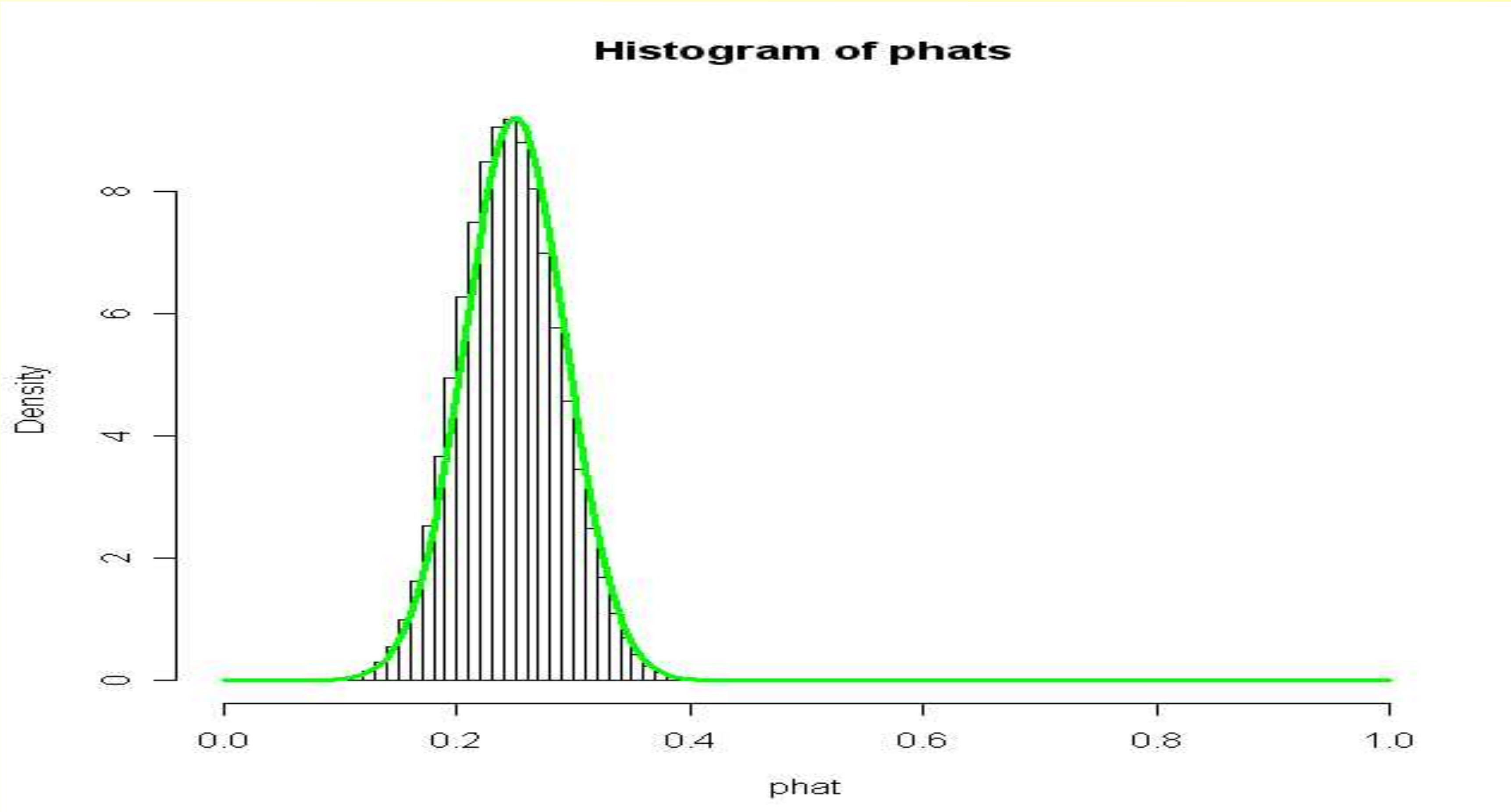
# Null and Alternative Hypotheses

- 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.

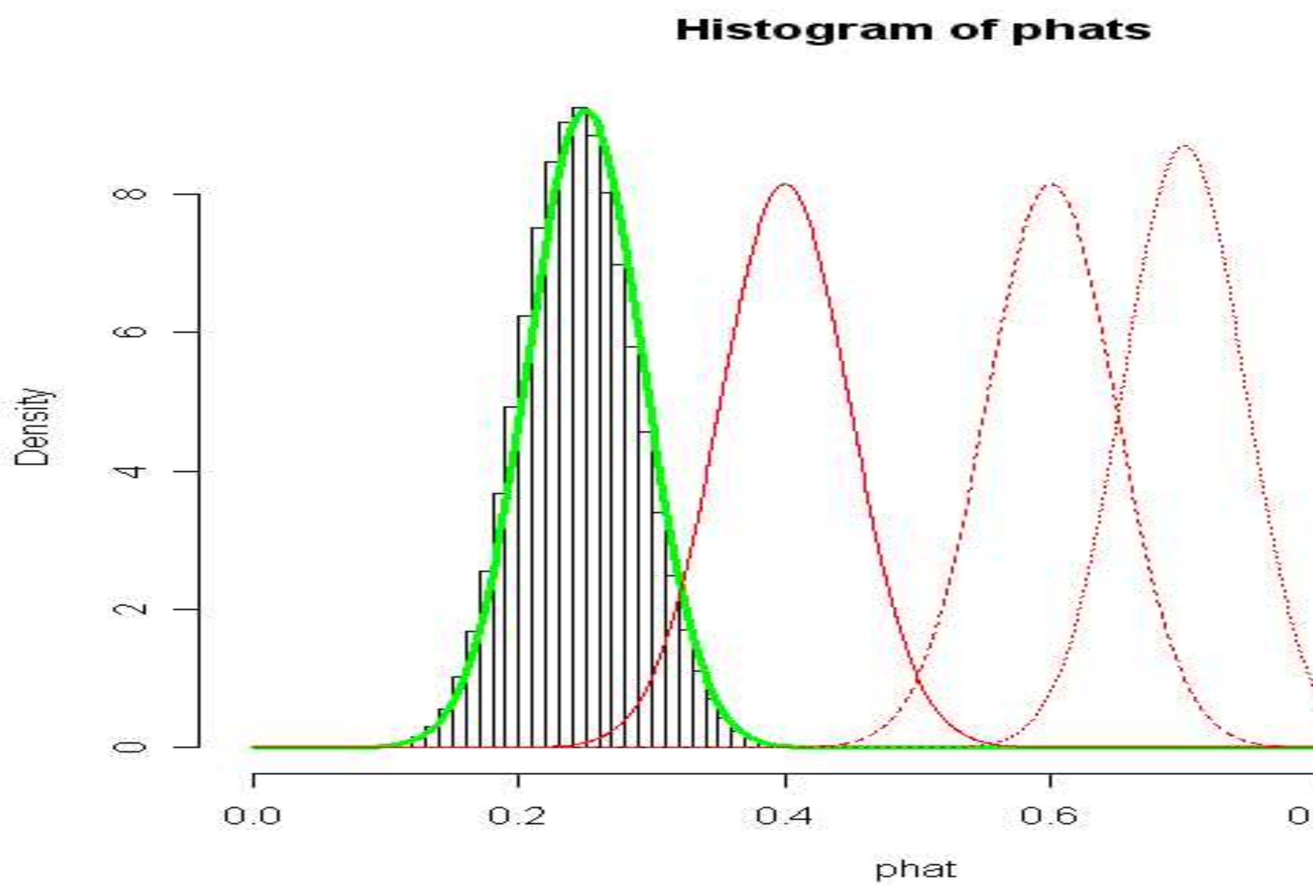
# Sampling Distribution Under $H_0 : p=0.25$



## 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.

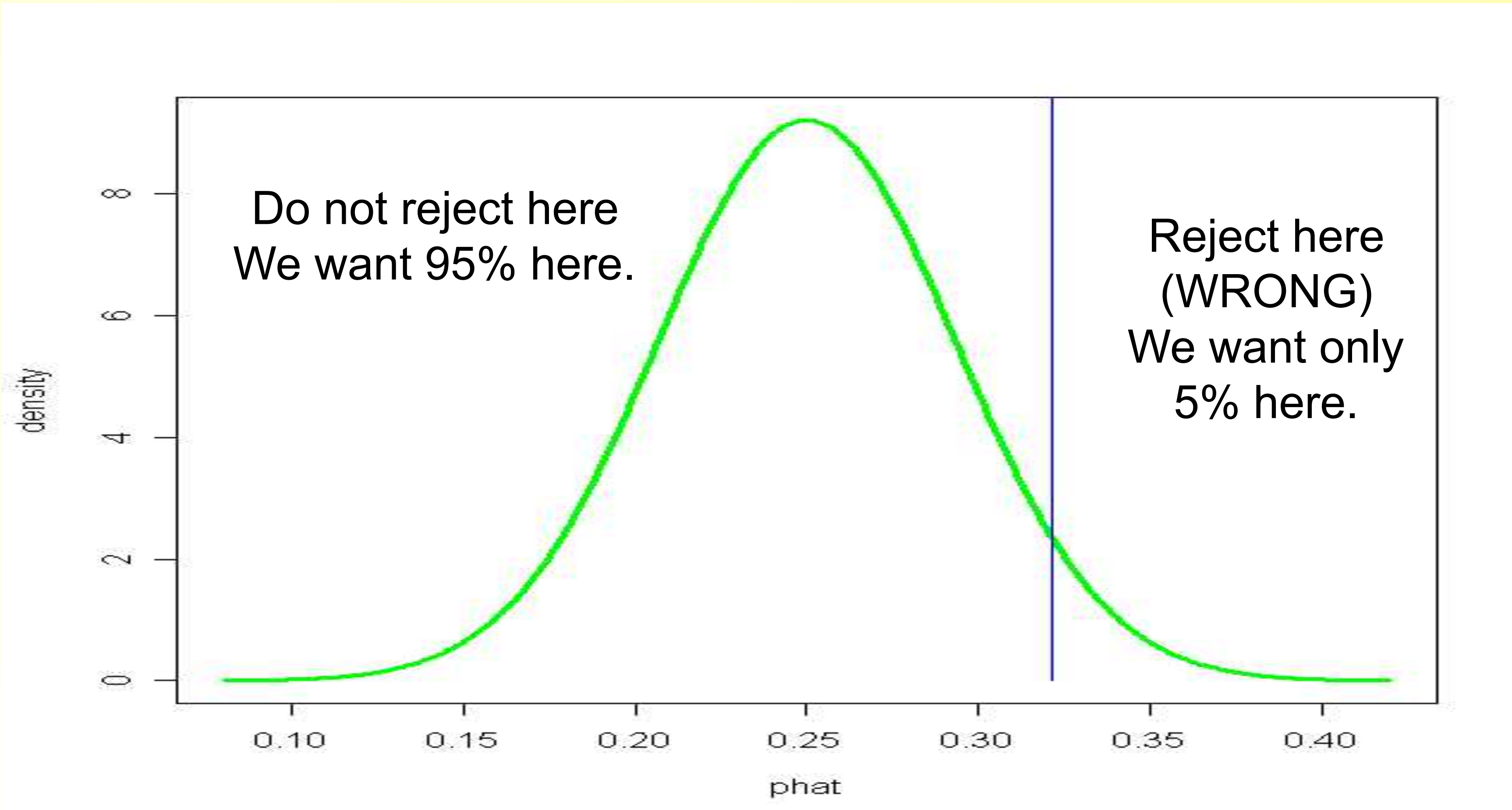
# Why Give $H_0$ the Benefit of the Doubt?

- 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?

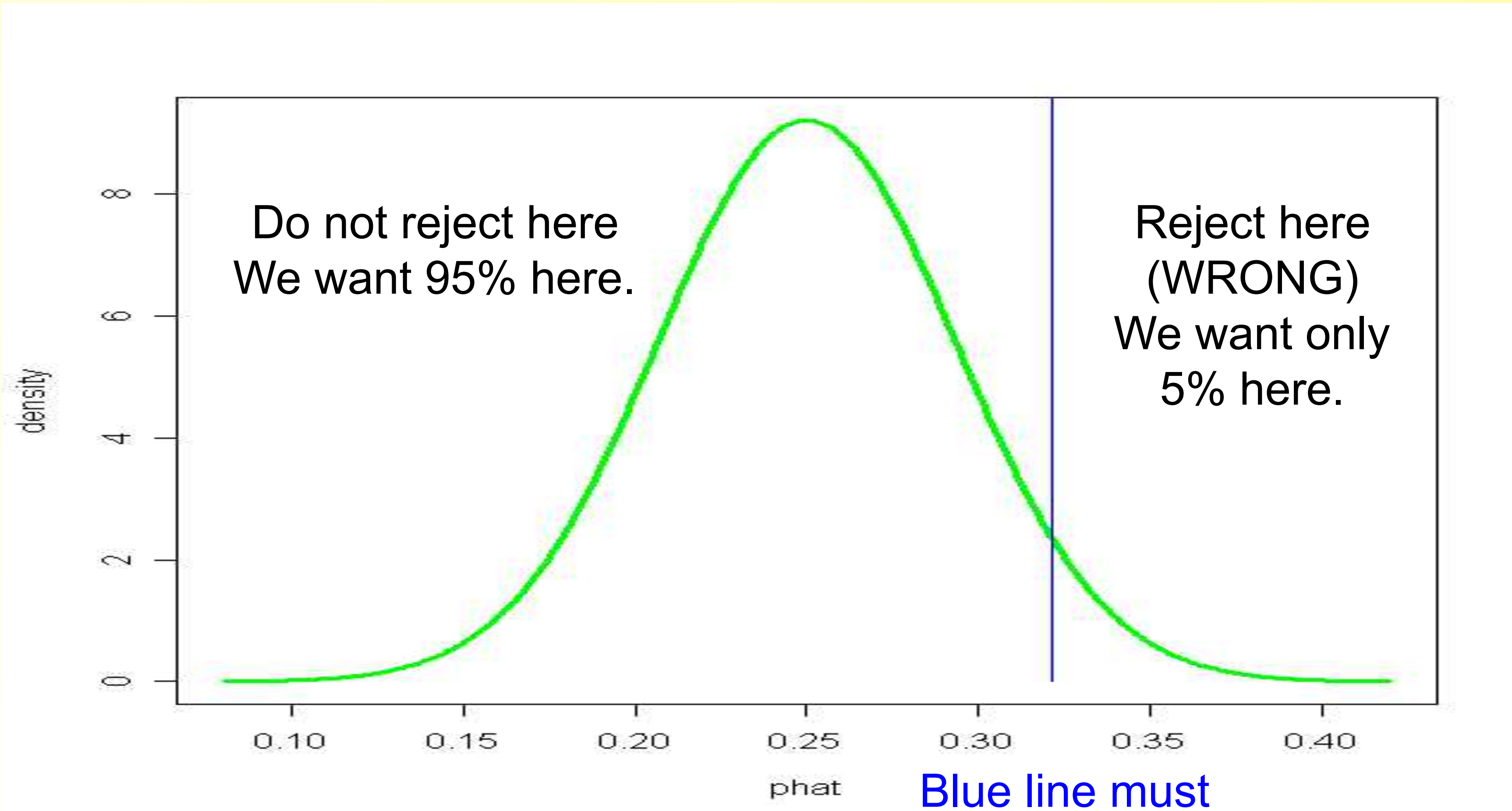
# Finding the Cutoff

- 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

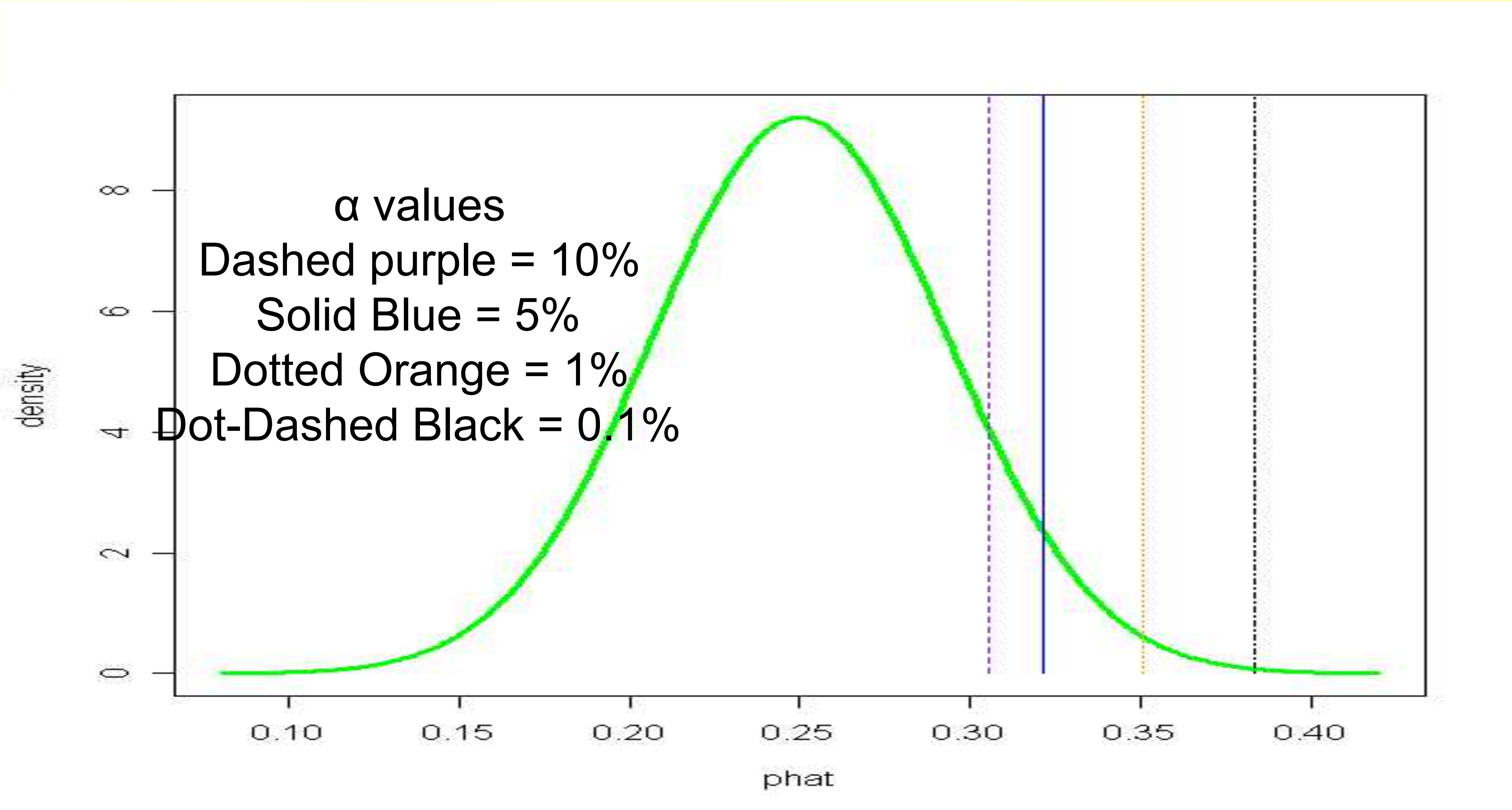



Blue line must be the 95th percentile

# Finding the 95th Percentile

- The null distribution is normal with mean 0.25 and standard deviation 0.0433
- The corresponding Z-score for the 95th 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”

# Quiz!