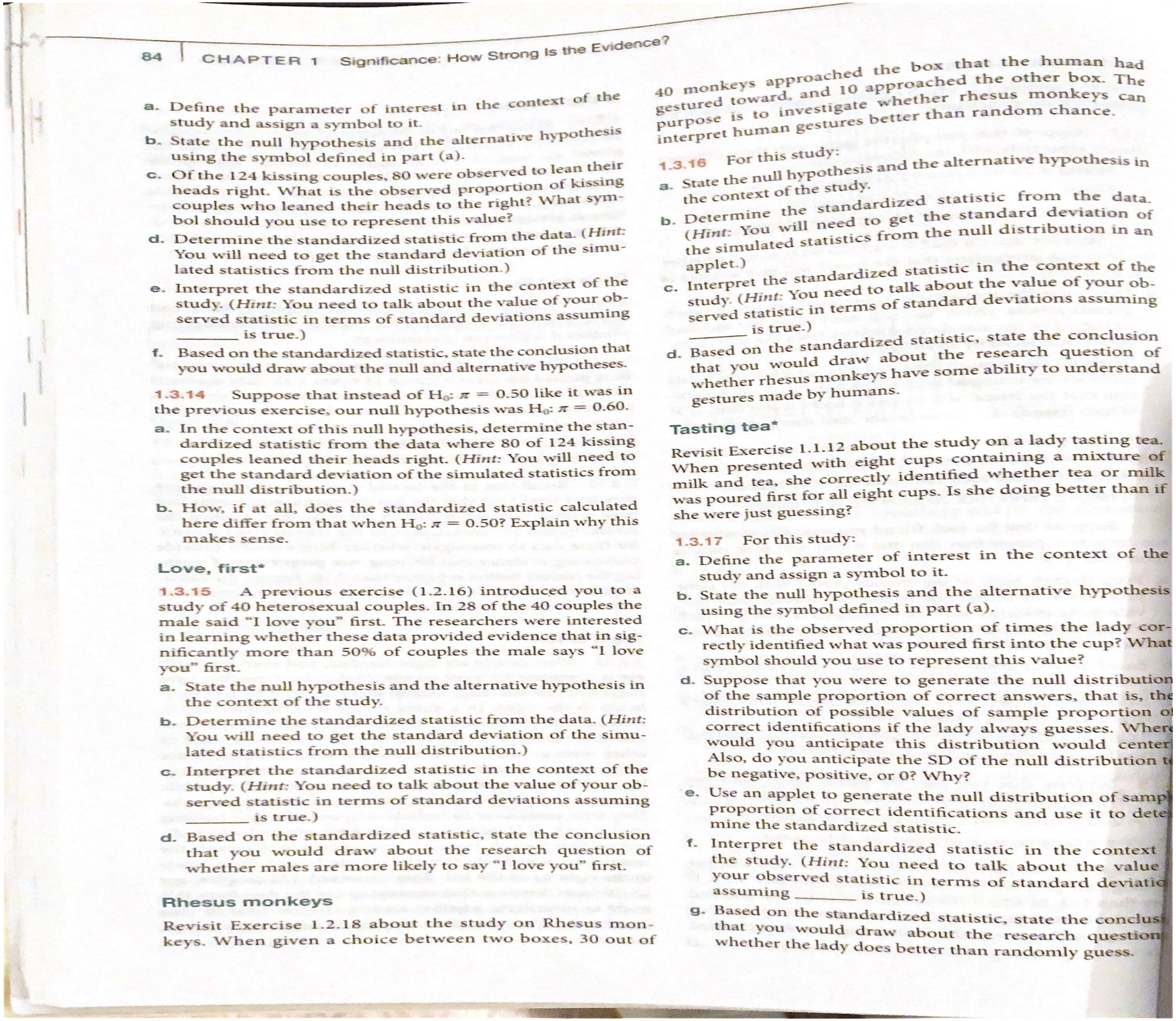
Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Simulating null distributions.
2. p-values.
3. Heart transplant example.
4. Standardized statistic
5. What impacts p-values and strength of evidence

Read through chapter 1.
The course website is http://www.stat.ucla.edu/~frederic/13/sum23



a. Define the parameter of interest in the context of the study and assign a symbol to it.

b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).

c. Of the 124 kissing couples, 80 were observed to lean their heads right. What is the observed proportion of kissing couples who leaned their heads to the right? What symbol should you use to represent this value?

d. Determine the standardized statistic from the data. (*Hint:* You will need to get the standard deviation of the simulated statistics from the null distribution.)

e. Interpret the standardized statistic in the context of the study. (*Hint:* You need to talk about the value of your observed statistic in terms of standard deviations assuming ________ is true.)

f. Based on the standardized statistic, state the conclusion that you would draw about the null and alternative hypotheses.

1.3.14 Suppose that instead of H_0: $\pi = 0.50$ like it was in the previous exercise, our null hypothesis was H_0: $\pi = 0.60$.

a. In the context of this null hypothesis, determine the standardized statistic from the data where 80 of 124 kissing couples leaned their heads right. (*Hint:* You will need to get the standard deviation of the simulated statistics from the null distribution.)

b. How, if at all, does the standardized statistic calculated here differ from that when H_0: $\pi = 0.50$? Explain why this makes sense.

Love, first*

1.3.15 A previous exercise (1.2.16) introduced you to a study of 40 heterosexual couples. In 28 of the 40 couples the male said "I love you" first. The researchers were interested in learning whether these data provided evidence that in significantly more than 50% of couples the male says "I love you" first.

a. State the null hypothesis and the alternative hypothesis in the context of the study.

b. Determine the standardized statistic from the data. (*Hint:* You will need to get the standard deviation of the simulated statistics from the null distribution.)

c. Interpret the standardized statistic in the context of the study. (*Hint:* You need to talk about the value of your observed statistic in terms of standard deviations assuming ________ is true.)

d. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether males are more likely to say "I love you" first.

Rhesus monkeys

Revisit Exercise 1.2.18 about the study on Rhesus monkeys. When given a choice between two boxes, 30 out of 40 monkeys approached the box that the human had gestured toward, and 10 approached the other box. The purpose is to investigate whether rhesus monkeys can interpret human gestures better than random chance.

1.3.16 For this study:

a. State the null hypothesis and the alternative hypothesis in the context of the study.

b. Determine the standardized statistic from the data. (*Hint:* You will need to get the standard deviation of the simulated statistics from the null distribution in an applet.)

c. Interpret the standardized statistic in the context of the study. (*Hint:* You need to talk about the value of your observed statistic in terms of standard deviations assuming ________ is true.)

d. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether rhesus monkeys have some ability to understand gestures made by humans.

Tasting tea*

Revisit Exercise 1.1.12 about the study on a lady tasting tea. When presented with eight cups containing a mixture of milk and tea, she correctly identified whether tea or milk was poured first for all eight cups. Is she doing better than if she were just guessing?

1.3.17 For this study:

a. Define the parameter of interest in the context of the study and assign a symbol to it.

b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).

c. What is the observed proportion of times the lady correctly identified what was poured first into the cup? What symbol should you use to represent this value?

d. Suppose that you were to generate the null distribution of the sample proportion of correct answers, that is, the distribution of possible values of sample proportion o correct identifications if the lady always guesses. Wher would you anticipate this distribution would center Also, do you anticipate the SD of the null distribution t be negative, positive, or 0? Why?

e. Use an applet to generate the null distribution of samp proportion of correct identifications and use it to dete mine the standardized statistic.

f. Interpret the standardized statistic in the context the study. (*Hint:* You need to talk about the value your observed statistic in terms of standard deviatio assuming ________ is true.)

g. Based on the standardized statistic, state the conclus that you would draw about the research question whether the lady does better than randomly guess.

1.4.23 For the "leaning" version of the study from the previous question:

a. *Statistic:* How many times did Krieger choose the correct object? Out of how many attempts? Thus, what proportion of the time did Krieger choose the correct object?

b. *Simulate:* Using an applet, simulate 1,000 repetitions of having the dog choose between the two objects if he is doing so randomly. Report the null and standard deviation.

c. Based on the study's result, what is the p-value for this test?

d. Approximately what proportion of the 10 attempts would Krieger have needed to get correct in order to yield a p-value of approximately 0.05?

1.4.24

a. Based on the study's result, what is the standardized statistic for this test?

b. *Strength of evidence:* What are your conclusions based on the p-value you found in part (d) from the previous exercise? Are the conclusions the same if you base them off the standardized-statistic you found in (a)?

c. Revisit your conjecture in Exercise 1.4.22, part (d). Did the p-value behave the way you had conjectured?

The sign test

o far, the outcome has always been binary—Yes/No, Right/Wrong, Heads/Tails, etc. What if outcomes are uantitative, like heights or percentages? Although there re specialized methods for such data that you will learn in later chapter, you can also use the methods and logic you ave already learned for situations of a very different sort:) outcomes are quantitative, (2) you want to compare wo conditions *A* and *B*, and (3) your data come in pairs, e *A* and one *B* in each pair. To apply the coin toss model, u simply ask for each pair, "Is the *A* value bigger than e *B* value?" The resulting test is called the "sign test" because the difference (*A* − *B*) is either plus or minus. Here's ummary table:

oin toss	Heads	P(Heads)	Null hypothesis	Statistic
zz's guess	Right	$\pi = P$ (Right)	$\pi = 0.50$	$\hat{p}$
ch pair	$A > B$	$\pi = P(A > B)$	$\pi = 0.50$	$\hat{p}$

ine providence

25* Refer to Exercises 1.4.8 to 1.4.12. Dr. Arbuthnot's al analysis was different from the analysis you saw er. Instead of using each individual birth as a coin toss, thnot used a sign test with each of the 82 years as a toss, and a year with more male births counted as a ess."

a. Complete the following table of comparisons:

Analysis method	Sample size n	Null value π_0	Value of $\hat{p}$
A: 1.4.8 – 1.4.12	____	____	____
B: 1.4.25	____	____	____

b. For each method of analysis, rate the strength of evidence against the null hypothesis, as one of: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

Healthy lungs

1.4.26 Researchers wanted to test the hypothesis that living in the country is better for your lungs than living in a city. To eliminate the possible variation due to genetic differences, they located seven pairs of identical twins with one member of each twin living in the country, the other in a city. For each person, they measured the percentage of inhaled tracer particles remaining in the lungs after one hour: the higher the percentage, the less healthy the lungs. They found that for six of the seven twin pairs the one living in the country had healthier lungs.

a. Is the alternative hypothesis one-sided or two-sided?

b. Based on the sample size and distance between the null value and the observed proportion, estimate the strength of evidence: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

c. Here are probabilities for the number of heads in seven tosses of a fair coin:

# Heads	0	1	2	3	4	5	6	7
Probability	0.0078	0.0547	0.1641	0.2734	0.2734	0.1641	0.0547	0.0078

Compute the p-value and state your conclusion.

Bee stings

1.4.27* Scientists gathered data to test the research hypothesis that bees are more likely to sting a target that has already been stung by other bees. On eight separate occasions, they offered a pair of targets to a hive of angry bees; one target in each pair had been previously stung, the other was pristine. On six of the eight occasions, the target that had been previously stung accumulated more new stingers.

a. Is the alternative hypothesis one-sided or two-sided?

b. Based on the sample size and distance between the null value and the observed proportion, estimate the strength of evidence: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

c. Here are probabilities for the number of heads in eight tosses of a fair coin:

# Heads	0	1	2	3	4	5	6	7	8
Probability	0.0039	0.0313	0.1094	0.2188	0.2734	0.2188	0.1094	0.0313	0.0039

Compute the p-value and state your conclusion.

1. Simulating null distributions and Standard Errors.

We observe $\hat{p}$ = 15.34% in our sample, and under Ho, the population percentage π = 10%. So we see a difference of 5.34%. This is our quantity of interest, and it is usually a difference like this. We want to see if that quantity of interest, 5.34%, is bigger than what we'd expect by chance under the null hypothesis.

The Standard Error (SE) is the standard deviation of the quantity of interest under the null hypothesis.

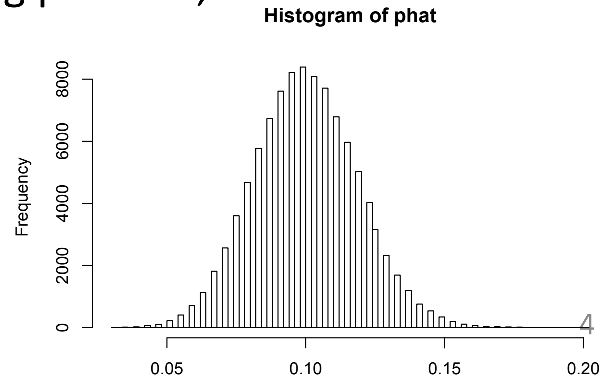
Many stat books just tell you the formulas to get the SE. Your book is different. They want to emphasize that in many cases you can estimate the SE by simulations.

In this example, under Ho, women with HG are just like the rest in terms of probability of delivering preterm. We have a SRS of size 254 from a population with π = 10% having preterm delivery. We can simulate 254 draws on the computer, where each draw is independent of the others and has a 10% chance of being preterm, and then see what results we get. In R, I did

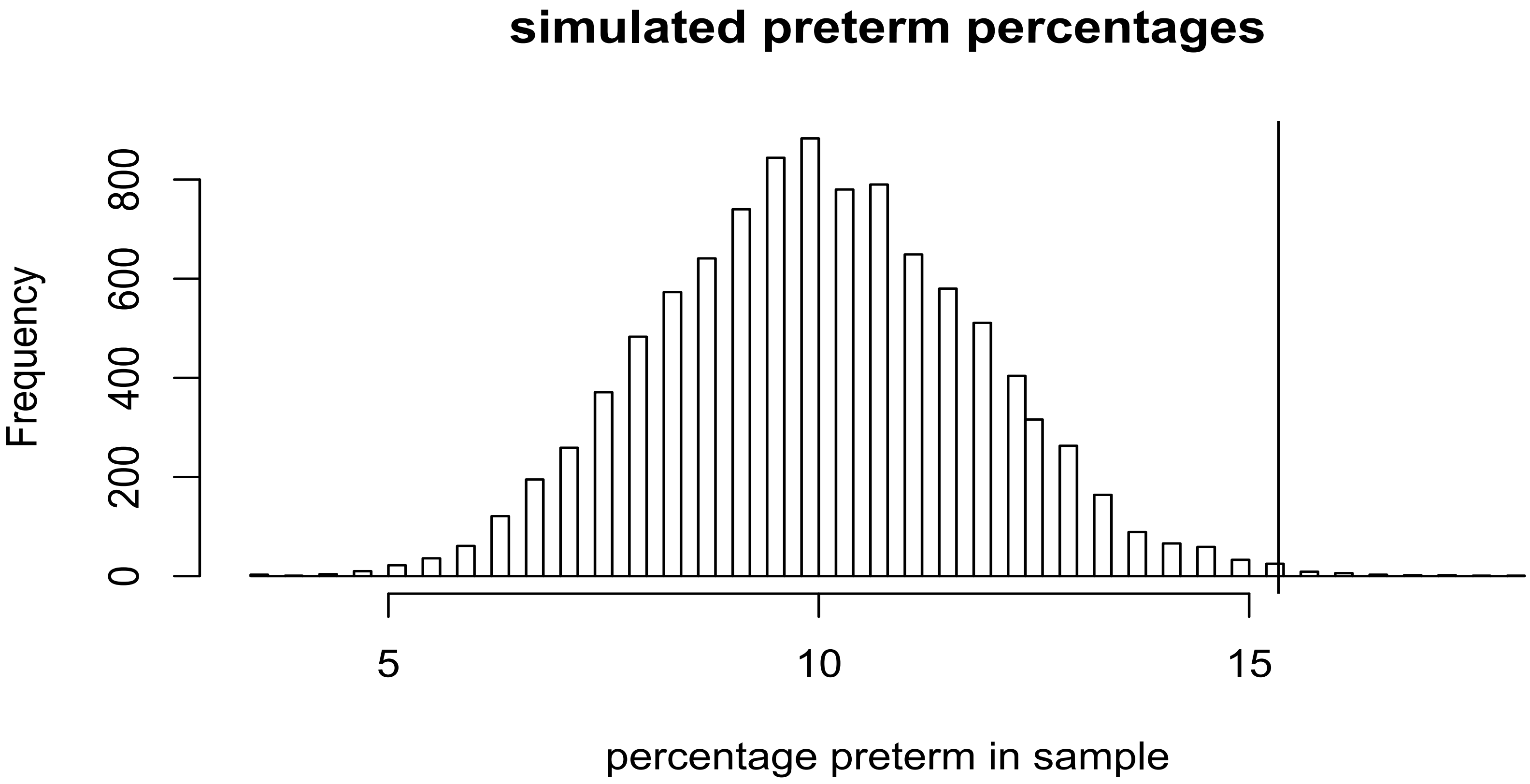
```
x = runif(254)
y = (x<0.1)
phat = mean(y)
```

The first time, I got phat = 0.1259843. 12.60%.

I tried it many times, and here is what I got.



```
a = rep(0,10000)
for(i in 1:10000){ x = runif(254); a[i] = mean(x<.1)}
hist(a*100,main="simulated preterm percentages", nclass=100,
       xlab="percentage preterm in sample")
abline(v=15.34)
sd(a)                        ## 0.01885409
sqrt(.10 * .90 / 254) ## 0.01882367
```



2. p-values.

The p-value is the probability, assuming Ho is true, that the test statistic will be at least as extreme as that observed.

"What are the chances of that?"

The key idea is that the convention is to compute the probability of getting something as extreme as you observed <u>or more extreme</u>.

e.g. n = 5, π_o = 50%, $\widehat{p}$= 4/5. The probability that $\hat{p}$ = 4/5 is 15.625%.

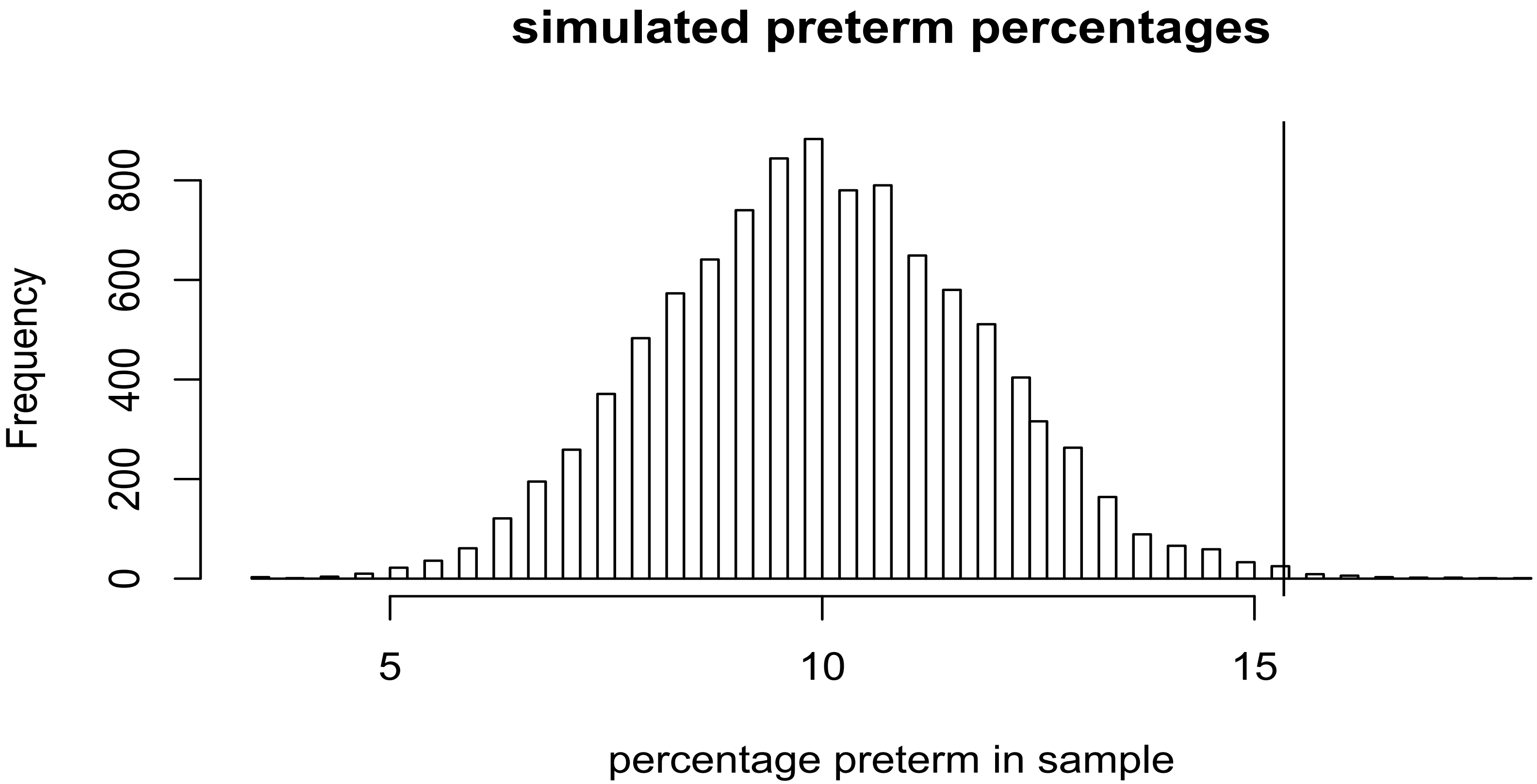
However, what if n = 400, π_o = 50%, and $\hat{p}$ = 201/400? Now the probability of getting 201/400 is 3.97%, but obviously the data are consistent with the null hypothesis that π = 50%.

Typically, one does a two-sided test, which means that by "extreme", we mean extreme in either direction. We want to see how in line our observed value of $\hat{p}$ =15.34% is with our null hypothesis of a population percentage of 10%. Could our sample of 15.34% preterm have come from a population of 10% preterm? A simulation with $\hat{p}$ > 15.34% would be more extreme than what we observed, and also a simulation with $\hat{p}$ < 4.66% would be more extreme than what we observed.

Guidelines for evaluating strength of evidence from p-values

- p-value >0.10, not much evidence against null hypothesis
- 0.05 < p-value ≤ 0.10, moderate evidence against the null hypothesis
- 0.01 < p-value ≤ 0.05, strong evidence against the null hypothesis
- p-value ≤ 0.01, very strong evidence against the null hypothesis

```
phat = rep(0,10000)
for(i in 1:10000){ x = runif(254); phat[i] = mean(x<.1)}
hist(phat*100,main="simulated preterm percentages", nclass=100,
       xlab="percentage preterm in sample")
abline(v=15.34)l
mean(abs(phat-.10)>.0534)        ## 0.0051
```



Continuing the HG example, using simulations of Ho we obtained samples of 254 values, and in 0.51% of these samples, at least 15.34% or more were preterm or less than 4.66% were preterm.

So we'd say the p-value is 0.51% for this two-sided test.

The observed difference is highly significant, and we have strong evidence against the null hypothesis of HG pregnancies having a 10% chance of being preterm like other pregnancies.

3. Heart Transplant Example.

Example 1.3

Heart Transplants

- The *British Medical Journal* (2004) reported that heart transplants at St. George’s Hospital in London had been suspended after a spike in the mortality rate
- Of the last 10 heart transplants, 80% had resulted in deaths within 30 days
- This mortality rate was over five times the national average.
- The researchers used 15% as a reasonable value for comparison.

Heart Transplants

- Does a heart transplant patient at St. George’s have a higher probability of dying than the national rate of 0.15?
- Observational units
  - The last 10 heart transplantations
- Variable
  - If the patient died or not
- Parameter
  - The actual probability of a death after a heart transplant operation at St. George’s

Heart Transplants

- **Null hypothesis:** Death rate at St. George's is the same as the national rate (0.15).
- **Alternative hypothesis:** Death rate at St. George's is higher than the national rate.

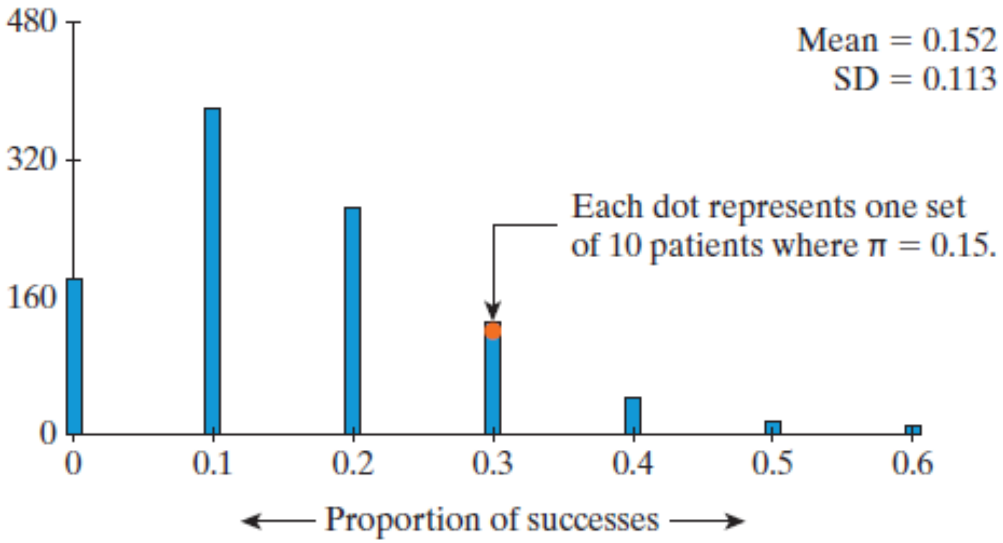
- **H_0:** π = 0.15   **H_a:** π > 0.15

- Our **statistic** is 8 out of 10 ($\hat{p}$ = 0.8)

Heart Transplants

Simulation

- Null distribution of 1000 repetitions of drawing samples of 10 “patients” where the probability of death is equal to 0.15.



What is the p-value?

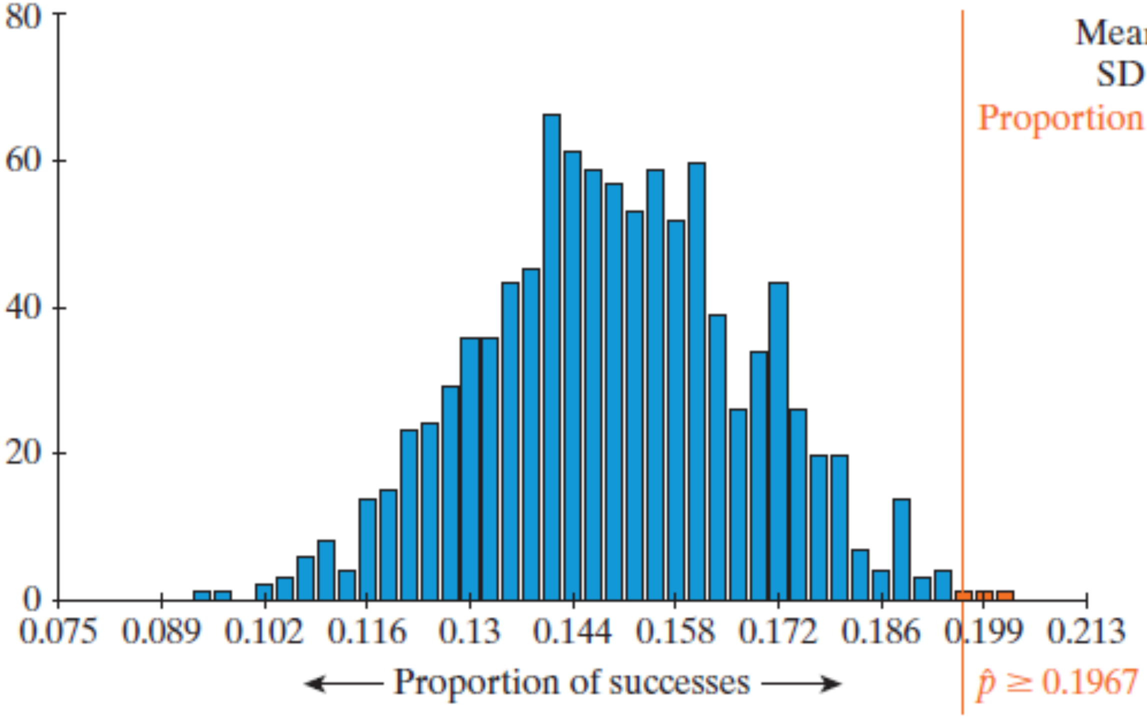
Heart Transplants

Strength of Evidence

- Our p-value is 0, so we have very strong evidence against the null hypothesis.
- Even with this strong evidence, it would be nice to have more data.
- Researchers examined the previous 361 heart transplantations at St. George’s and found that 71 died within 30 days.
- Our new statistic, $\hat{p}$, is 71/361 ≈ 0.1967

Heart Transplants

- Here is a null distribution and p-value based on the new statistic.



Heart Transplants

- The p-value was about 0.003
- We still have very strong evidence against the null hypothesis, but not quite as strong as the first case

- Another way to measure strength of evidence is to ***standardize*** the observed statistic

4. The Standardized Statistic

- The ***standardized statistic*** is the number of standard deviations our sample statistic is above the mean of the null distribution (or below the mean if it is negative).

- $z = \dfrac{\textit{statistic} - \textit{mean of null distribution}}{\textit{standard deviation of null distribution}}$
- The sd of the null distribution is the *standard error.*
- For a single proportion, we will use the symbol *z* for standardized statistic.
- Note: In the formula above, we can either use the mean of the actual null distribution or (better yet) the long-term proportion (probability) given in the null hypothesis.

The Standardized Statistic

- Here are the standardized statistics for our two studies.

$$z = \frac{0.80 - 0.15}{0.113} = 5.75 \qquad z = \frac{0.197 - 0.15}{0.018} = 2.61$$

- In the first, our observed statistic was 5.75 standard deviations above the mean.
- In the second, our observed statistic was 2.61 standard deviations above the mean.
- Both of these are very strong, but we have stronger evidence against the null in the first.

Guidelines for strength of evidence

- If a standardized statistic is below -2 or above 2, we have strong evidence against the null.

Standardized Statistic	**Evidence Against Null**
between -1.5 and 1.5	not much
below -1.5 or above 1.5	moderate
below -2 or above 2	strong
below -3 or above 3	very strong

5. What impacts p-values and strength of evidence?

Section 1.4

Example 1.4

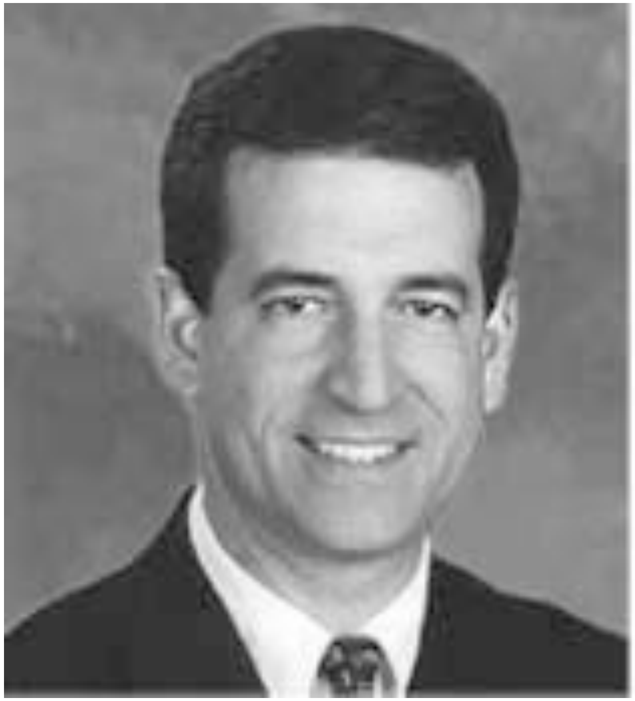
Predicting Elections from Faces

Predicting Elections

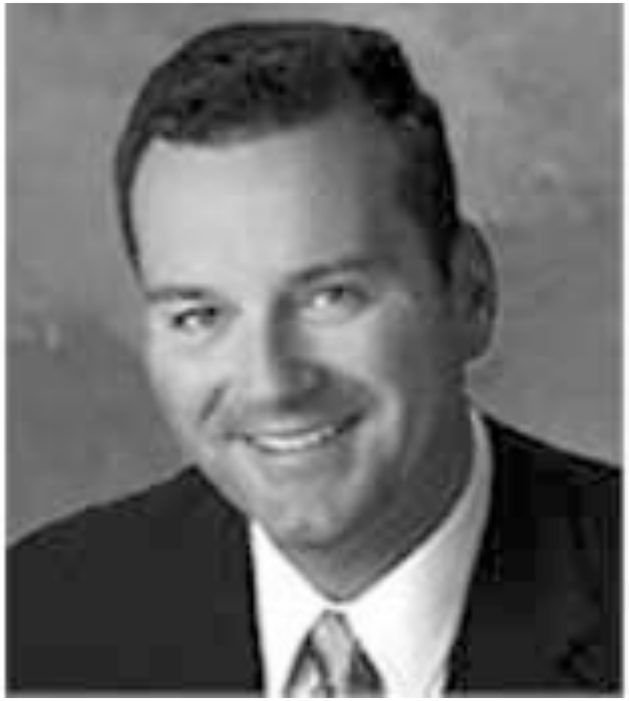
- Do voters make judgments about candidates based on facial appearances?
- More specifically, can you predict an election by choosing the candidate whose face is more competent-looking?
- Participants were shown two candidates and asked who has the more competent-looking face.

Who has the more competent looking face?

- 2004 Senate Candidates from Wisconsin



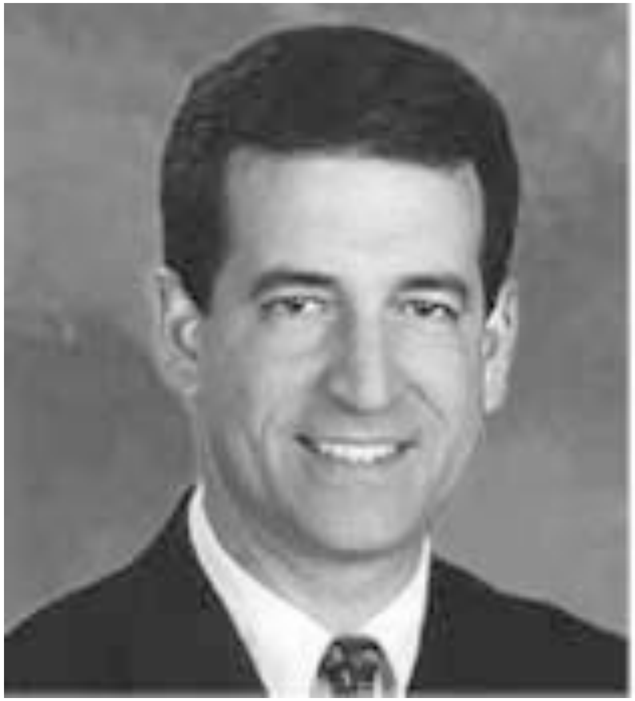
Winner



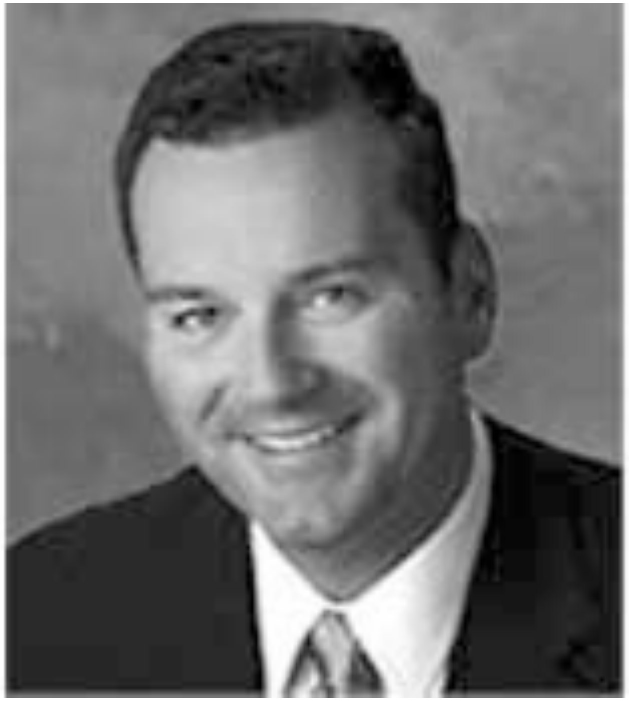
Loser

Bonus: One is named Tim and the other is Russ. Which name is the one on the left?

- 2004 Senate Candidates from Wisconsin



Russ



Tim

Predicting Elections

- They determined which face was the more competent for the 32 Senate races in 2004.
- What are the observational units?
  - The 32 Senate races
- What is the variable measured?
  - If the method predicted the winner correctly

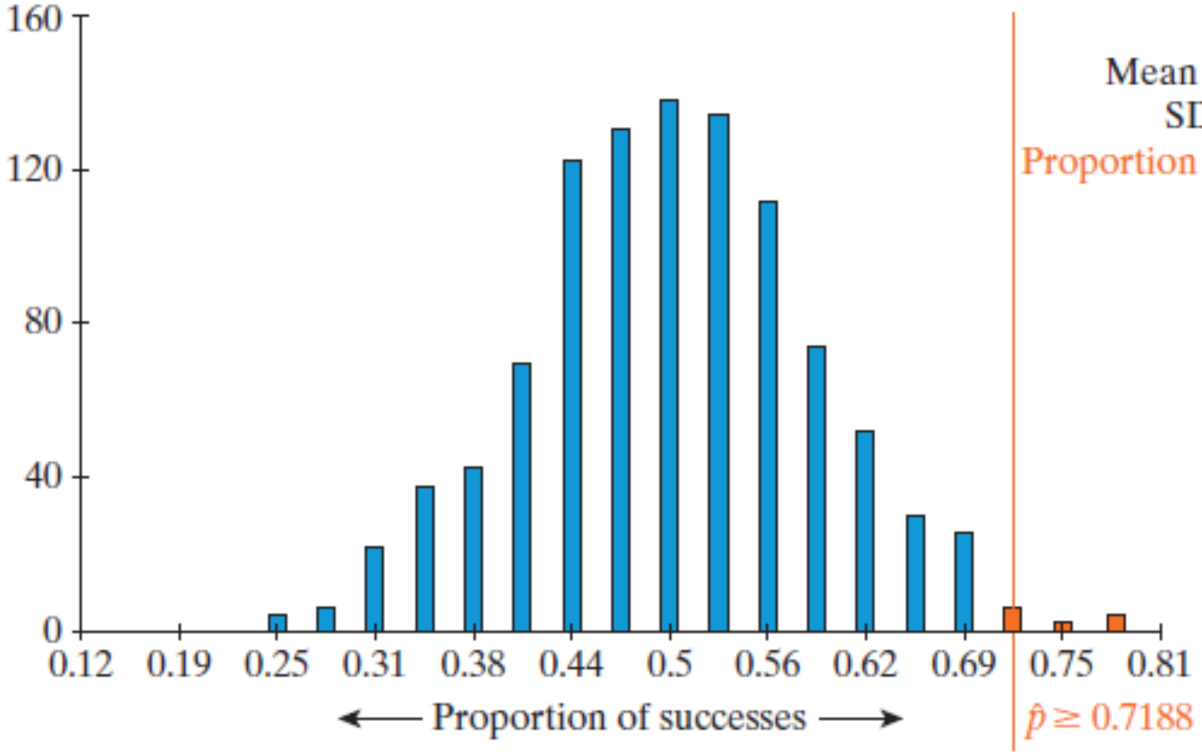
Predicting Elections

- Null hypothesis: The probability this method predicts the winner equals 0.5. (H_0: π = 0.5)
- Alternative hypothesis: The probability this method predicts the winner is greater than 0.5. (H_a: π > 0.5)

- This method predicted 23 of 32 races, hence $\hat{p} =$ 23/32 ≈ 0.719, or 71.9%.

Predicting Elections

1000 simulated sets of 32 races



Predicting Elections

- With a p-value of 0.009 we have strong evidence against the null hypothesis.
- When we calculate the standardized statistic we again show strong evidence against the null.

$$z = \frac{0.7188 - 0.5}{0.09} = 2.43.$$

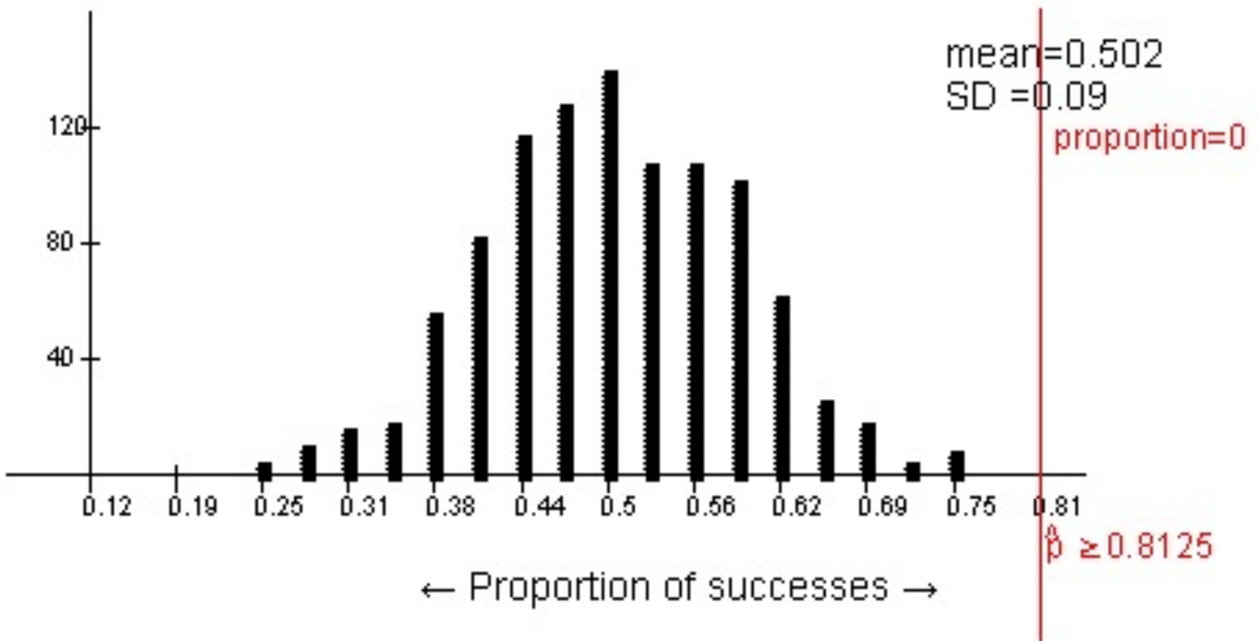
- What do the p-value and standardized statistic mean?

What affects the strength of evidence?

1. The effect size, which is the difference between the observed statistic ($\hat{p}$) and null hypothesis parameter (π_0).
2. Sample size.
3. If we do a one or two-sided test.

Effect size, i.e. the difference between $\hat{p}$ and π_o

- What if researchers predicted 26 elections instead of 23?
  - 26/32 = 0.8125 never occurs just by chance hence the p-value is 0.



Difference between $\hat{p}$ and the null parameter

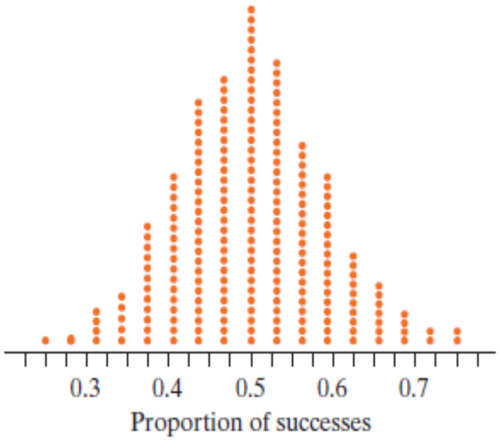
- The farther away the observed statistic is from the average value of the null distribution (or π_0), the more evidence there is against the null hypothesis.

Sample Size

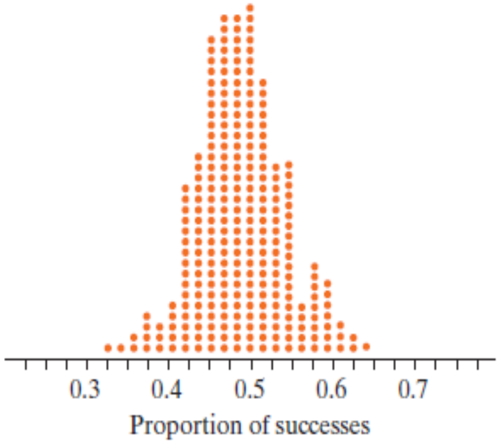
Suppose the sample proportion stays the same, do you think increasing sample size will increase, decrease, or have no impact on the strength of evidence against the null hypothesis?

Sample Size

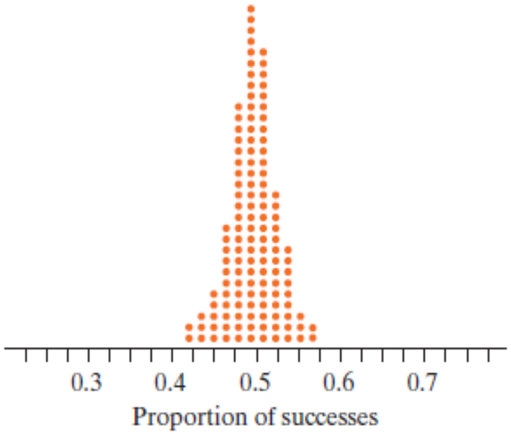
- The null distribution changes as we increase the sample size from 32 senate races to 128 races to 256 races.
- As the sample size increases, the variability (standard error) decreases.



n = 32



n = 128



n = 256

Sample Size

- What does decreasing variability mean for statistical significance (with same sample proportion)?
- 32 elections
  - – p-value = 0.009 and z = 2.43
- 128 elections
  - – p-value = 0 and z =5.07
- 256 elections
  - – Even stronger evidence
  - – p-value = 0 and *z* = 9.52

Sample Size

- As the sample size increases, the variability decreases.
- Therefore, as the sample size increases, the evidence against the null hypothesis increases (as long as the sample proportion stays the same and is in the direction of the alternative hypothesis).

Two-Sided Tests

- What if researchers were wrong; instead of the person with the more competent face being elected more frequently, it was actually less frequently?

H_0: π = 0.5

H_a: π > 0.5

- With this alternative, if we get a sample proportion less than 0.5, we would get a p-value greater than 50%.
- This is a ***one-sided*** test.
- Often one-sided is too narrow
- In fact most research uses two-sided tests.

Two-Sided Tests

- In a two-sided test the null can be rejected when sample proportions are in either tail of the null distribution.

Null hypothesis: The probability this method predicts the winner equals 0.50. (H_0: $\pi = 0.50$)

Alternative hypothesis: The probability this method predicts the winner **is not** 0.50.

(H_a: $\pi \neq 0.50$)

1-sided versus 2-sided tests.

- On my tests, I will tell you explicitly whether to do a 1 or 2 sided test.
- On hw problems, you might have to decide whether to do a 1-sided or 2-sided test.
- With the hw, if in the problem you are given that you are only looking for evidence in one direction, then you do a 1-sided test. If you are looking for *any* difference in proportions, then do a 2-sided test.