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/sum18

1. Simulating null distributions and Standard Errors.

We observe 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.

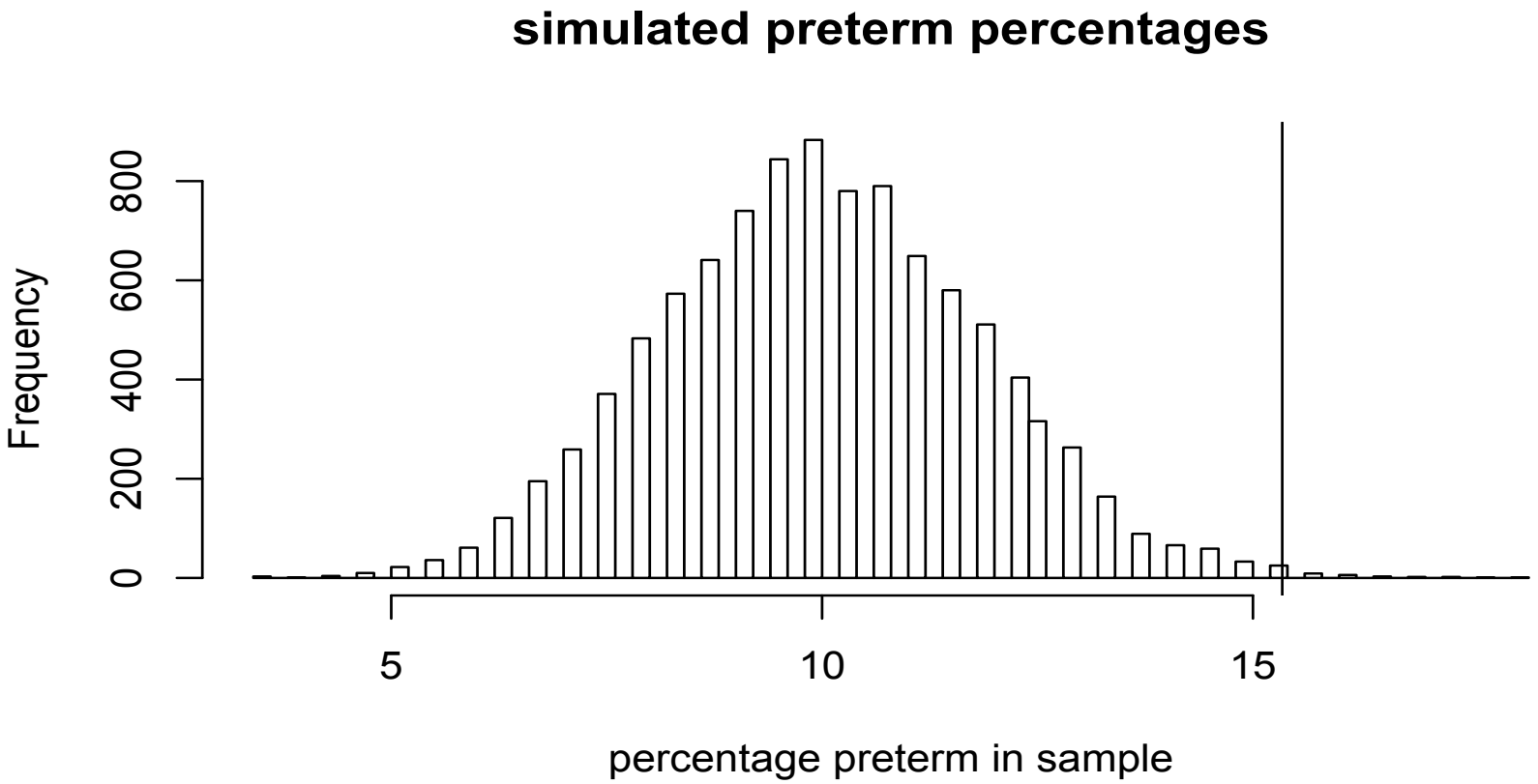
In this example, under Ho, women with HG are just like the rest in terms of probability of delivering preterm. We have a SRS of size 254 from a population with π = 10% having preterm delivery. We can simulate 254 draws on the computer, where each draw is independent of the others and has a 10% chance 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)l
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 or more extreme.

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

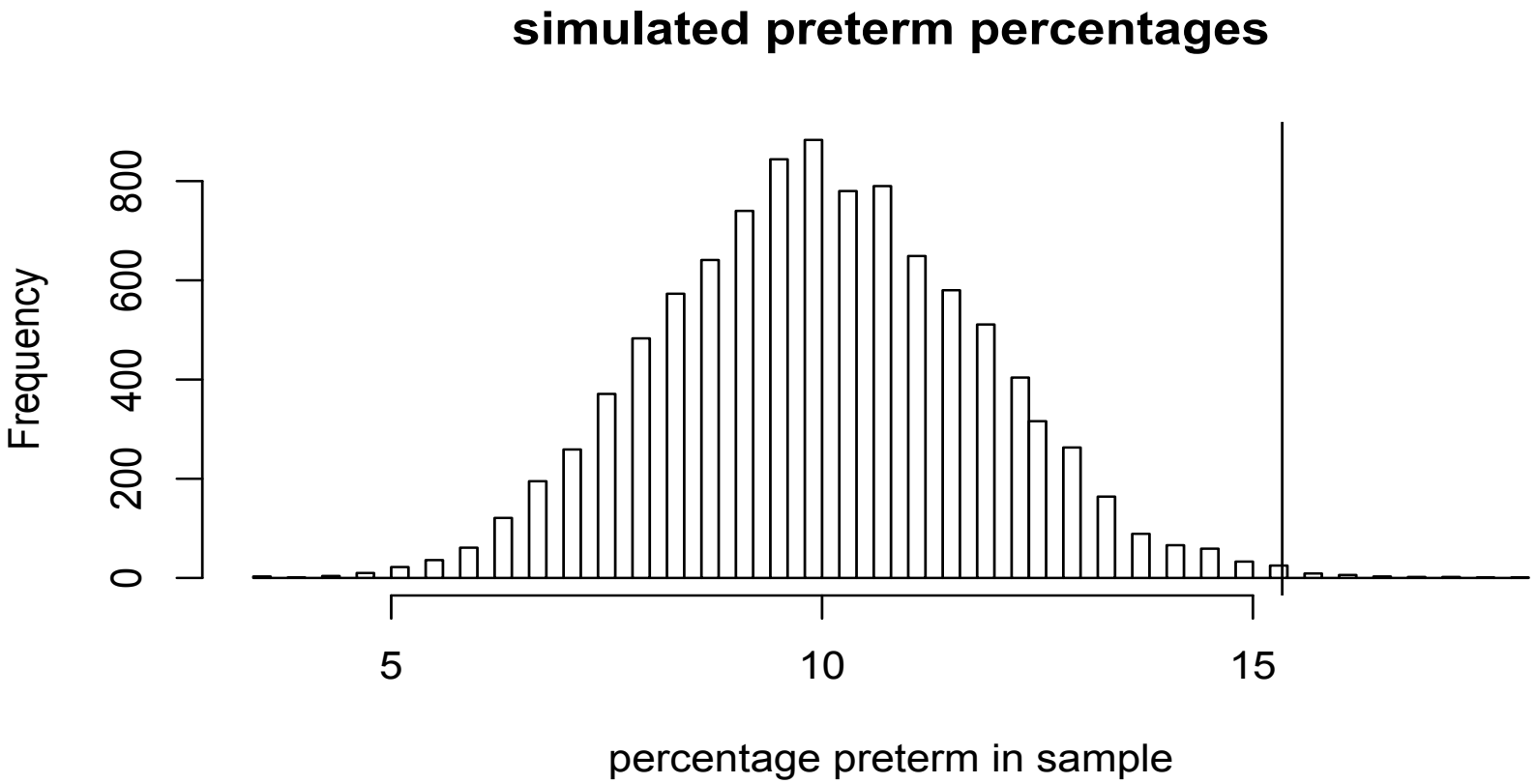
However, what if n = 400, π_o = 50%, and $\hat{p}$ = 201/400? Now the probability of getting 201/400 is 3.97%, but obviously the data are consistent with the null hypothesis that π = 50%.

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 $\leq$ 0.10, moderate evidence against the null hypothesis
- 0.01 < p-value $\leq$ 0.05, strong evidence against the null hypothesis
- p-value $\leq$ 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.

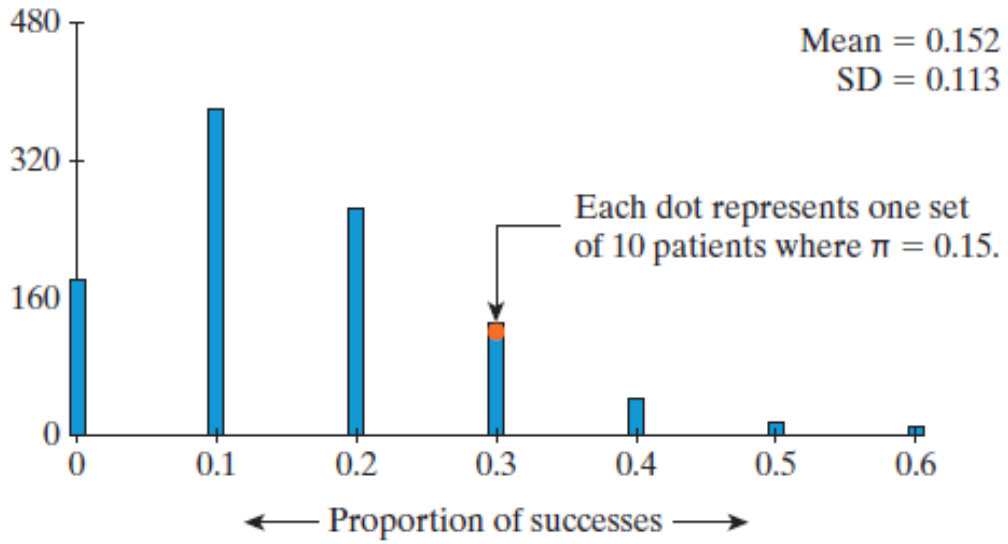
- $\mathbf{H_0}$**:** $\pi = 0.15$ $\mathbf{H_a}$**:** $\pi > 0.15$

- 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?

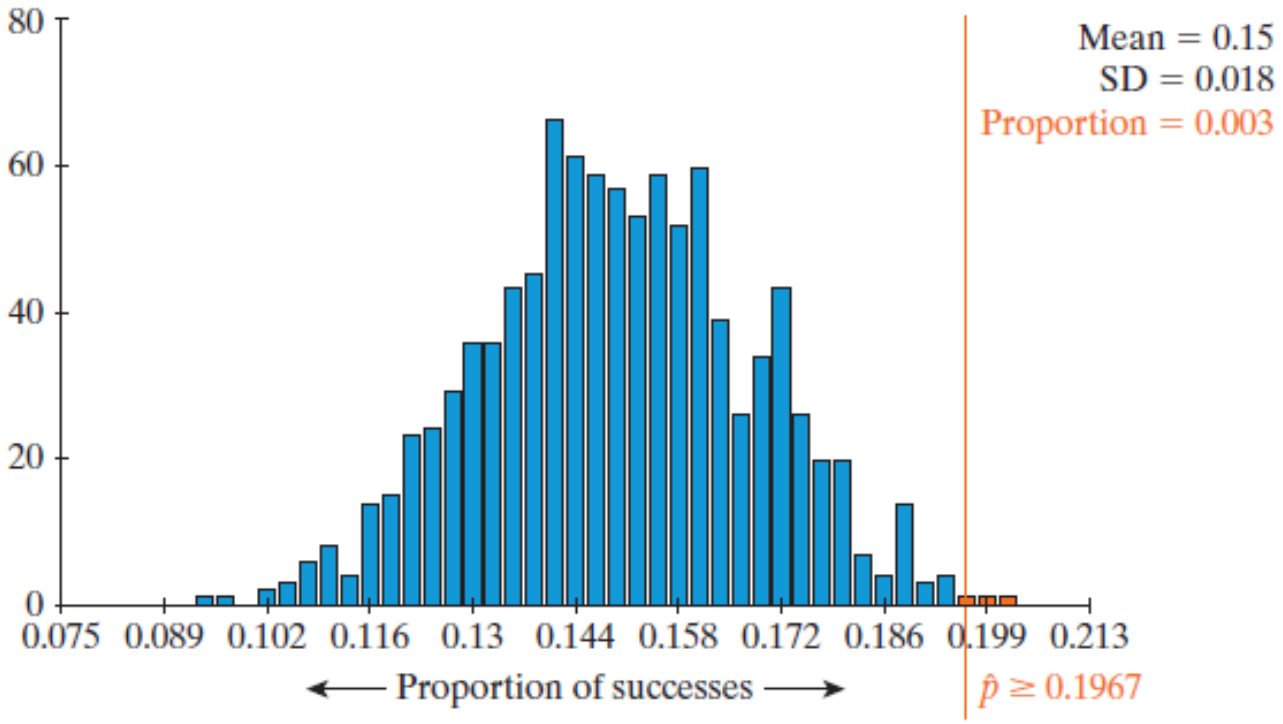
Heart Transplants

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).

- $$z = \frac{\textit{statistic} - \textit{mean of null distribution}}{\textit{standard deviation of null distribution}}$$

- 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

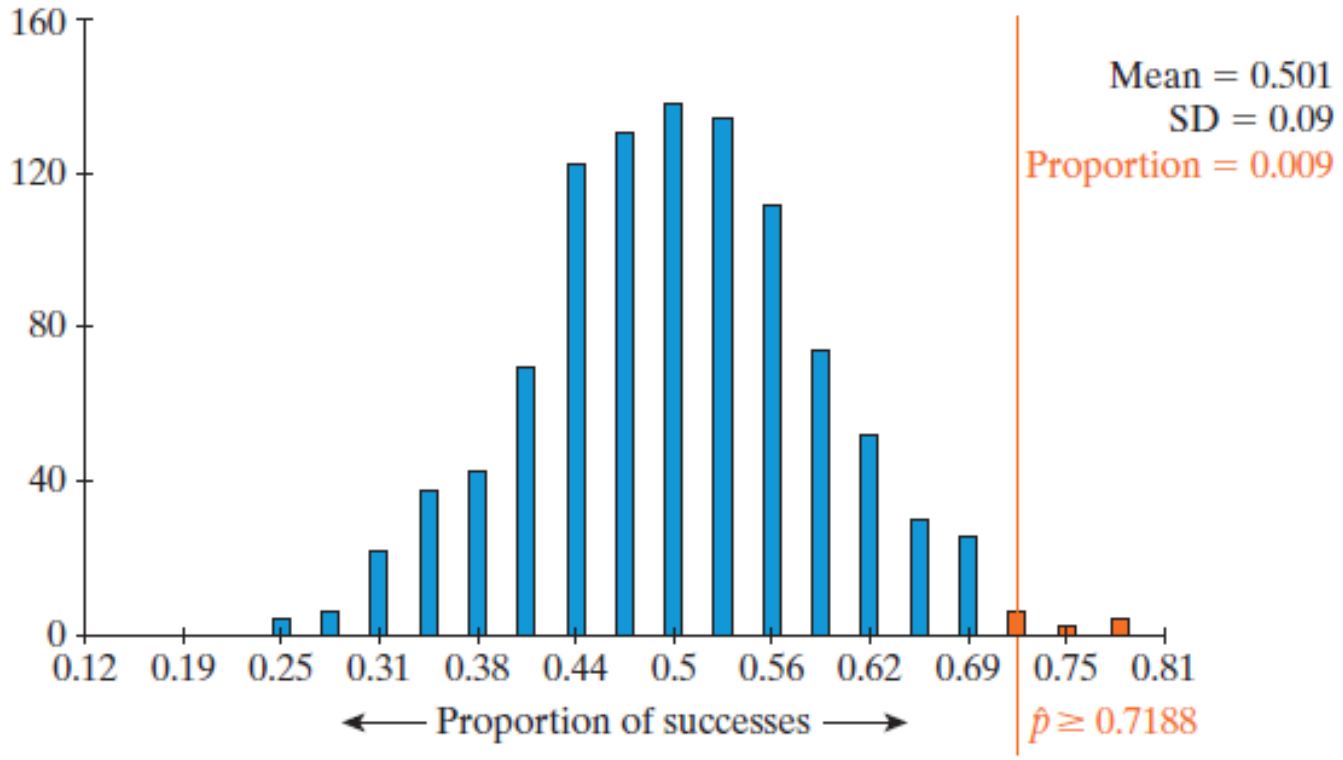
Predicting Elections

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

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

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

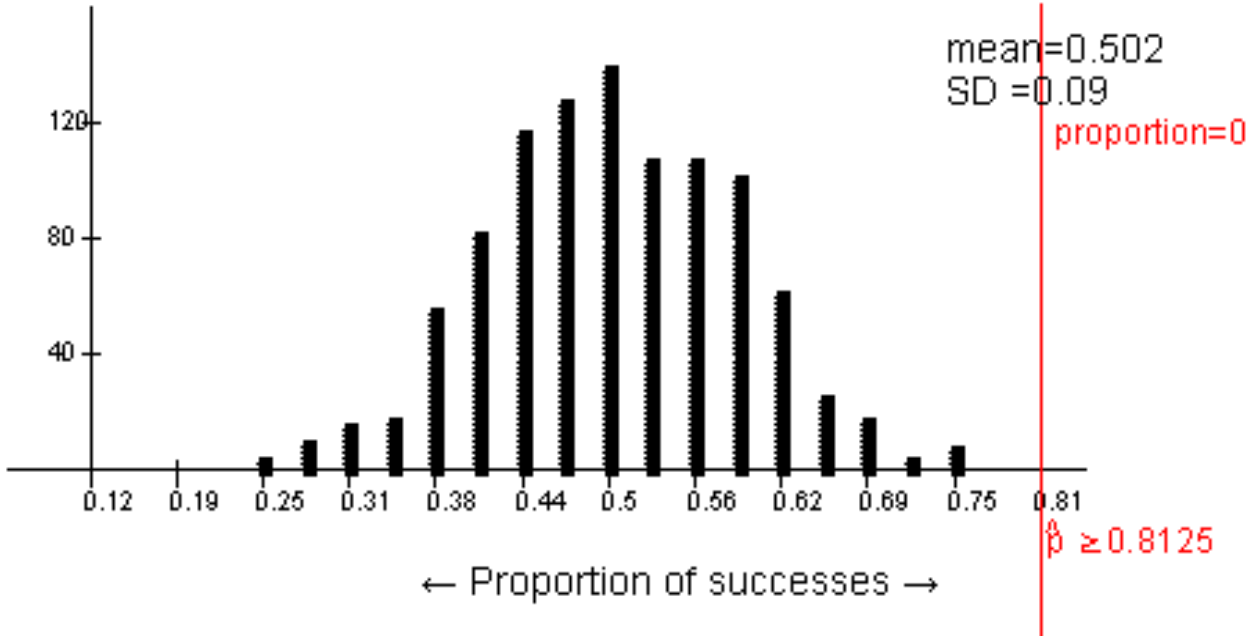
- What do the p-value and standardized statistic mean?

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

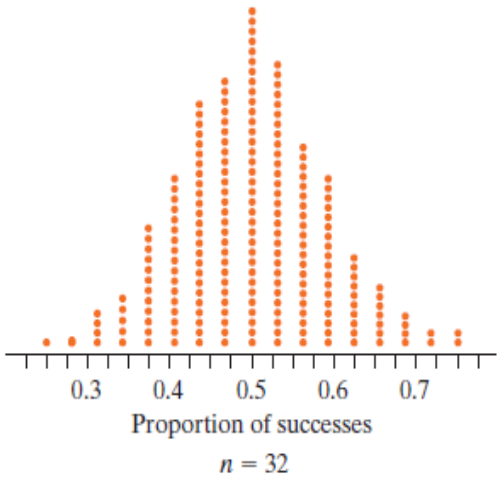
- The farther away the observed statistic is from the average value of the null distribution (or π_0), the more evidence there is against the null hypothesis.

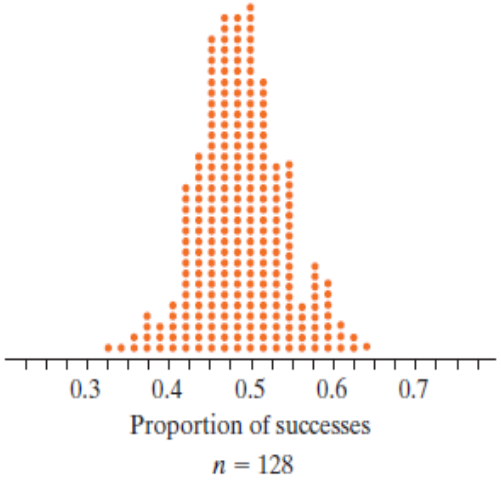
Sample Size

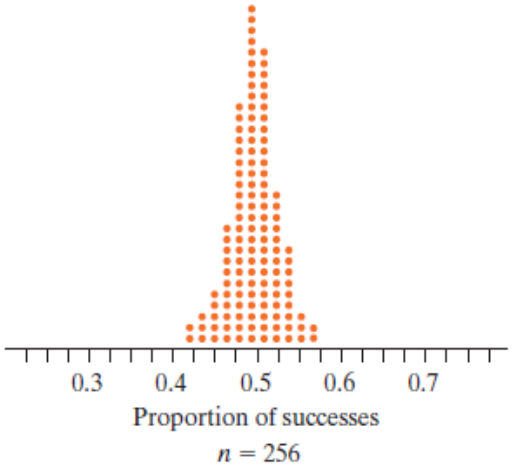
Suppose the sample proportion stays the same, do you think increasing sample size will increase, decrease, or have no impact on the strength of evidence against the null hypothesis?

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.







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?

  $H_0: \pi = 0.5$

  $H_a: \pi > 0.5$

- 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.