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

1. Guidelines for strength of evidence.
2. What affects strength of evidence and p value? Elections and faces example.
3. Normal distribution, CLT, and halloween candy example.
Read chapter 2.
Hw1 is due Tue 10/9. 1.3.16, 1.4.26, and the names and emails of 2 students.
http://www.stat.ucla.edu/~frederic/13/F18 .

1. 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

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

Section 1.4

Example 1.4

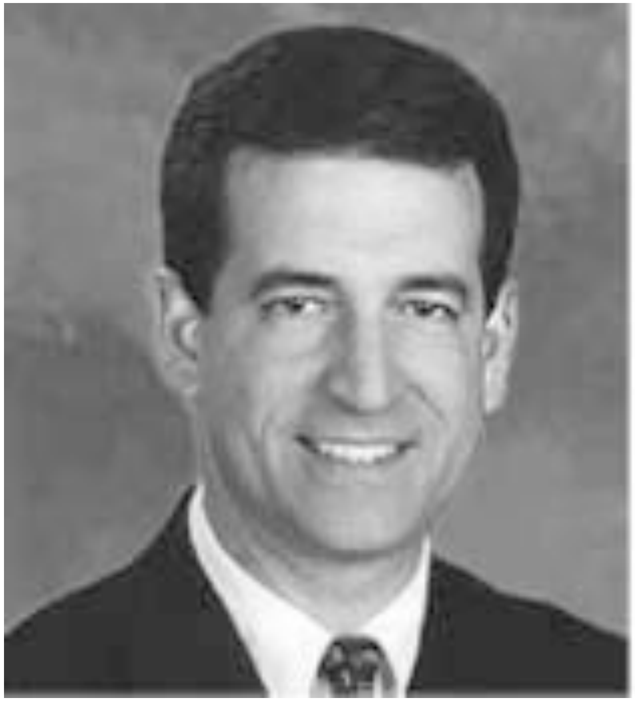
Predicting Elections from Faces

Predicting Elections

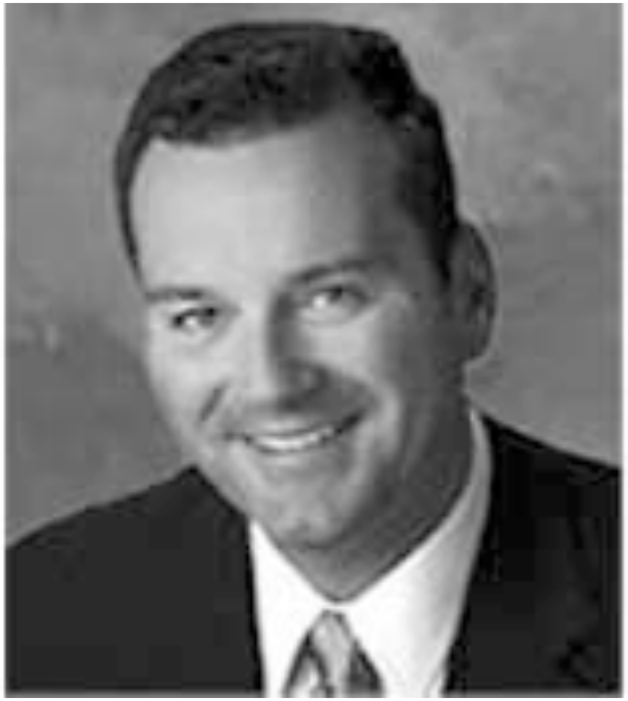
- Do voters make judgments about candidates based on facial appearances?
- More specifically, can you predict an election by choosing the candidate whose face is more competent-looking?
- Participants were shown two candidates and asked who has the more competent-looking face.

Who has the more competent looking face?

- 2004 Senate Candidates from Wisconsin



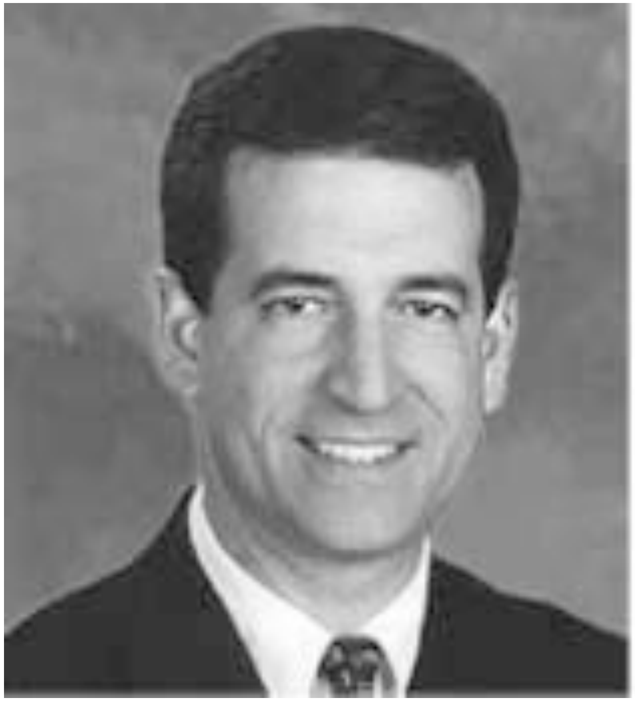
Winner

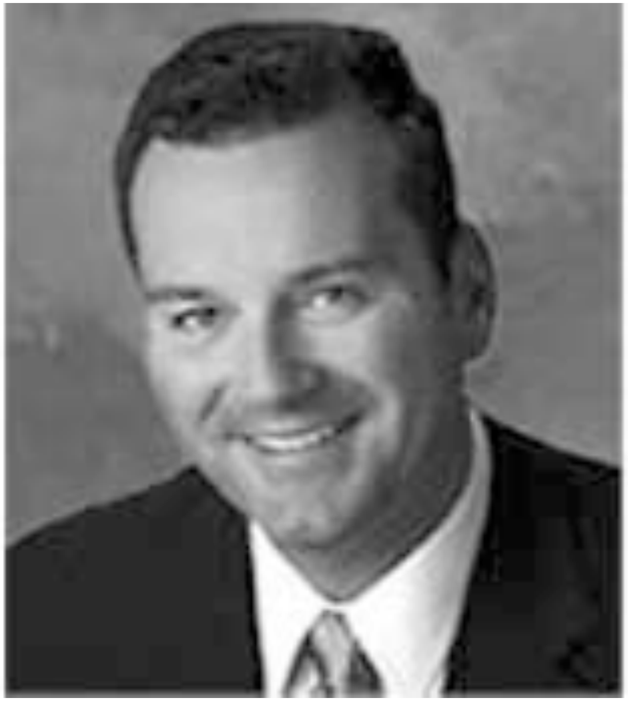


Loser

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

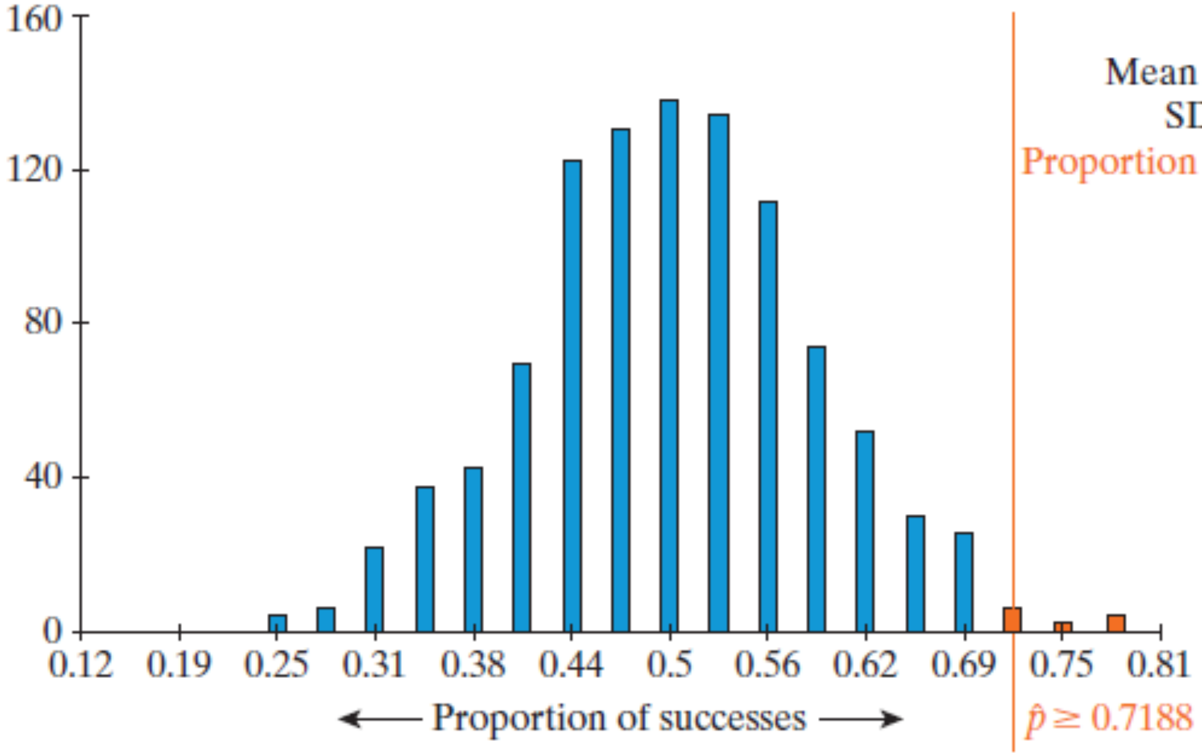
Predicting Elections

- 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.

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

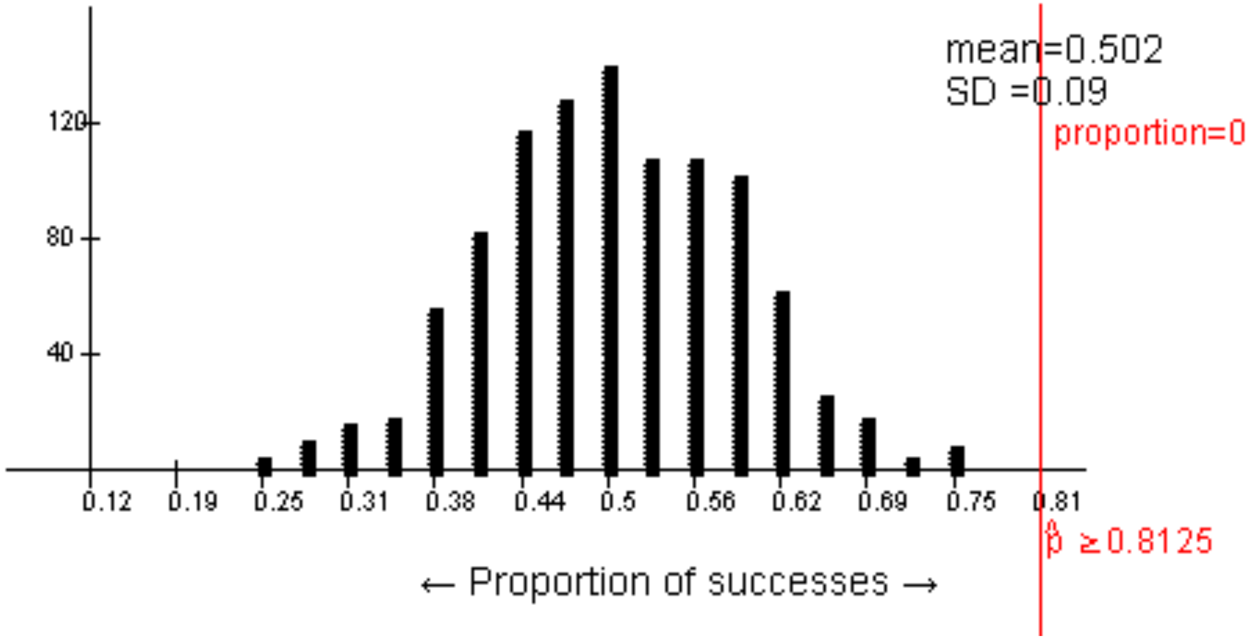
- What do the p-value and standardized statistic mean?

What affects the strength of evidence?

1. The 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.

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

- 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

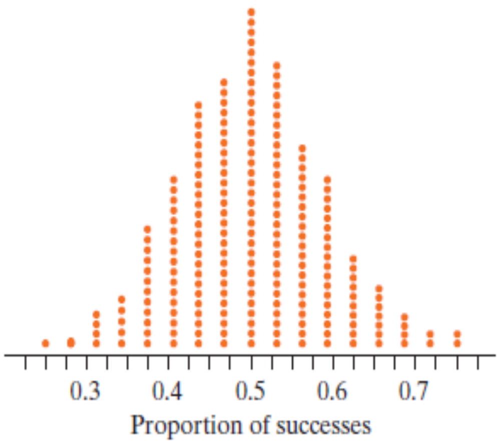
- The effect size is the distance between the observed statistic from what you would expect under the null hypothesis (π_0). If this is larger then there is more evidence against the null hypothesis.

Sample Size

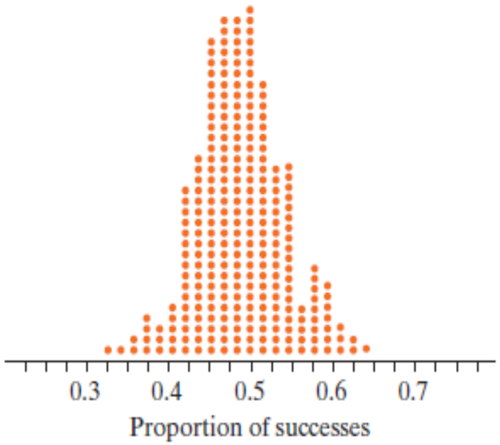
Suppose the sample proportion stays the same, do you think increasing sample size will increase, decrease, or have no impact on the strength of evidence against the null hypothesis?

Sample Size

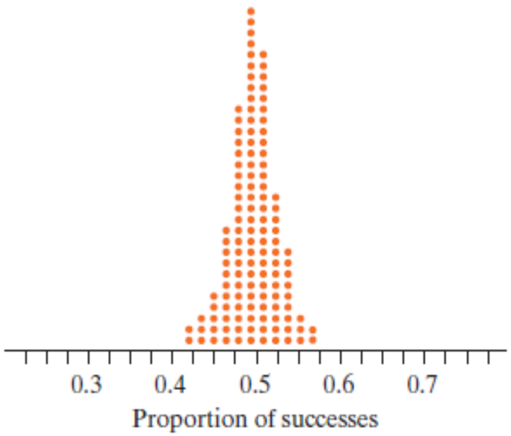
- The null distribution changes as we increase the sample size from 32 senate races to 128 races to 256 races.
- As the sample size increases, the variability (standard error) decreases.



$n = 32$



$n = 128$



$n = 256$

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?

$H_0: \pi = 0.5$

$H_a: \pi > 0.5$

- With this alternative, if we get a sample proportion less than 0.5, we would get a very large p-value.
- 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.

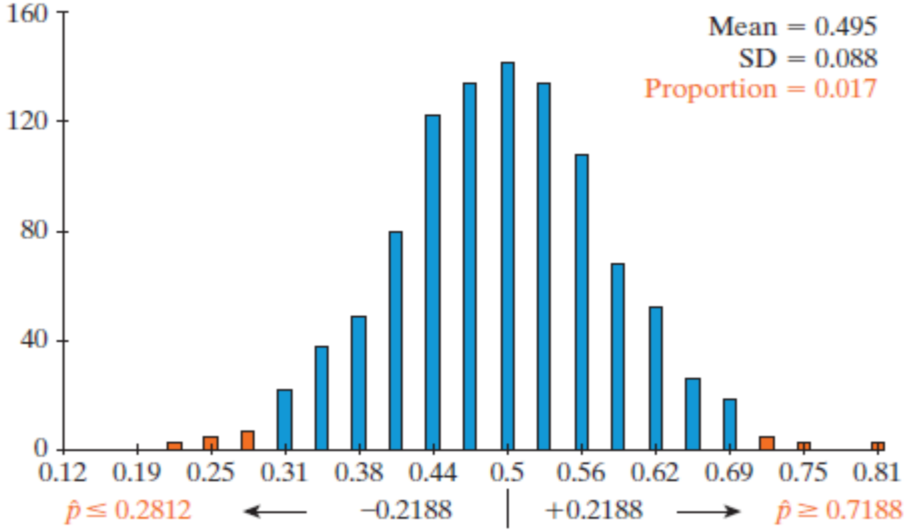
(H_a: $\pi \neq 0.50$)

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.

Two-Sided Tests

- Continuing with the example of predicting elections based on faces, since our sample proportion was 0.7188 and 0.7188 is 0.2188 *above* 0.5, we also need to look at 0.2188 *below* 0.5.
- The p-value will include all simulated proportions 0.7188 and above as well as those 0.2812 and below.



Two-Sided Tests

- 0.7188 or greater was obtained 9 times
- 0.2812 or less was obtained 8 times
- The p-value is (8 + 9 = 17)/1000 = 0.017.
- Two-sided tests increase the p-value (it about doubles) and hence decrease the strength of evidence.
- Two-sided tests are said to be more conservative. More evidence is needed to reject the null hypothesis.

Predicting House Elections

- Researchers also predicted the 279 races for the House of Representatives in 2004.
- The correctly predicted the winner in 189/279 ≈ 0.677, or 67.7% of the races.
- The House's sample percentage (67.7%) is a bit smaller than the Senate (71.9%), but the sample size is larger (279) than for the senate races (32).
- Do you expect the strength of evidence to be stronger, weaker, or essentially the same for the House compared to the Senate?

Predicting House Elections

The effect size, the distance from the observed statistic to the null hypothesis value

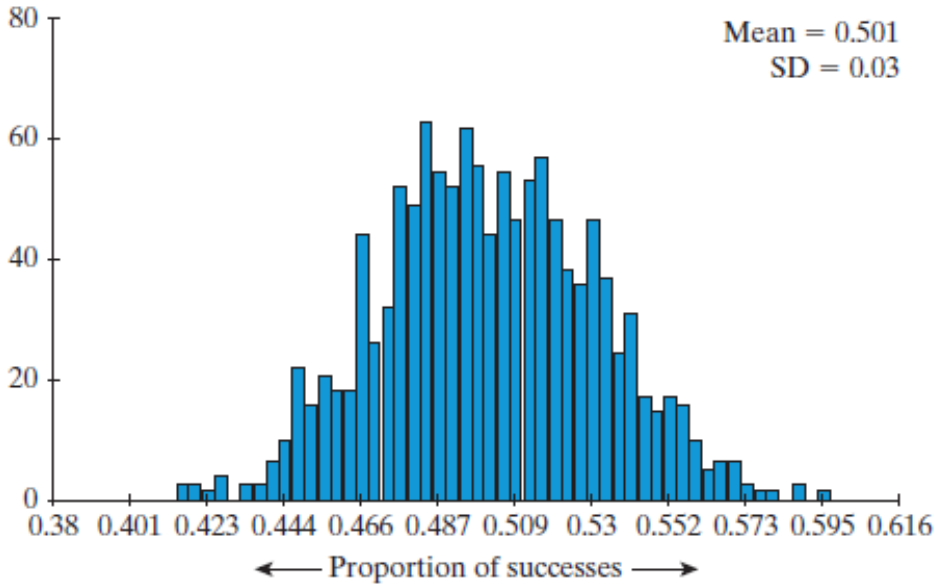
- The statistic in the House is 0.677 compared to 0.719 in the Senate
- Slight decrease in the effect size.

Sample size

- The sample size is almost 10 times as large (279 vs. 32)
- This will increase the strength of evidence.

Predicting House Elections

Null distribution of 279 sample House races



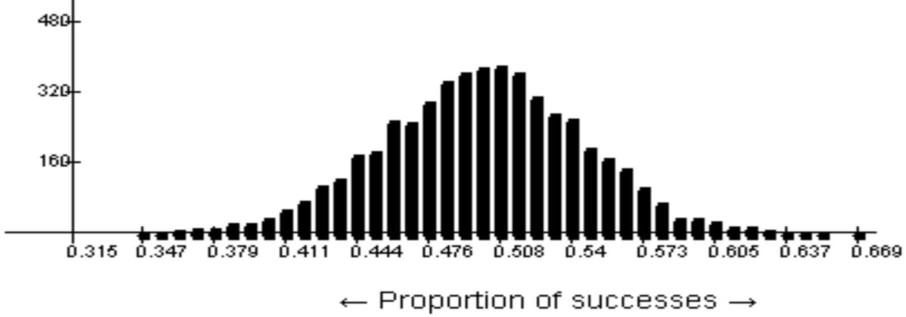
Simulated statistics ≥0.677 didn’t occur at all so the estimated p-value is 0

Predicting House Elections

- What about the standardized statistics?
  - For the Senate it was 2.43
  - For the House is 5.90.
- The larger sample size for the House outweighed the smaller effect size in this particular case. We have stronger evidence against the null using the data from the House.

Normal distribution, CLT, and halloween candy example.

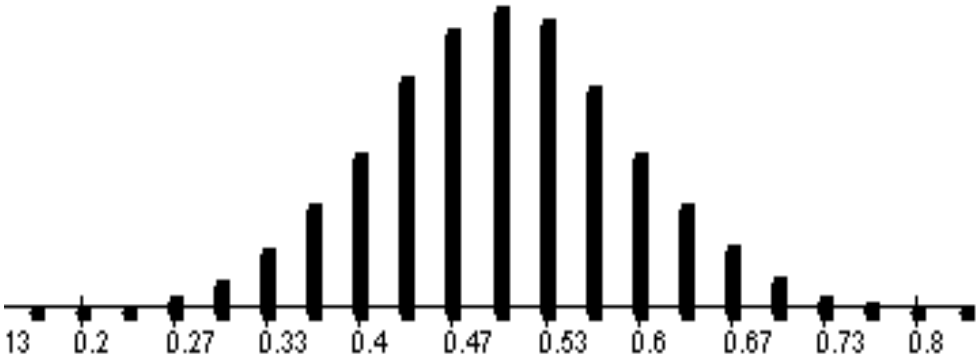
Section 1.5

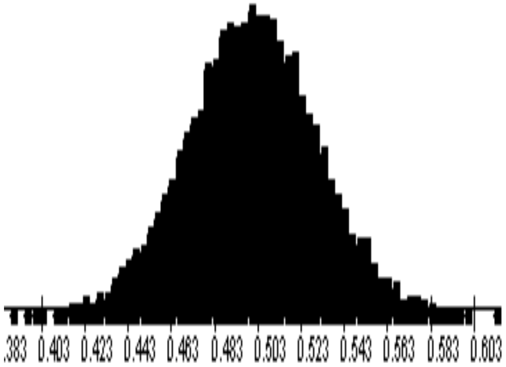


- The shape of most of our simulated null distributions always seems to be bell shaped. This shape is called the normal distribution.
- The Central Limit Theorem (CLT) dictates that, as *n* gets large, the sample mean or proportion becomes approximately normally distributed.
- When we do a test of significance using theory-based methods, only how our p-values are found will change. Everything else will stay the same.

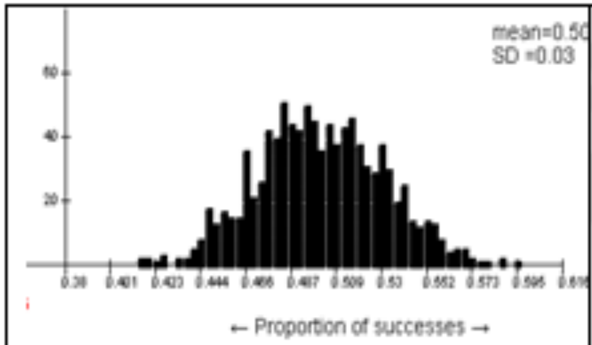
The Normal Distribution

- Both of these are centered at 0.5.
  - The one on the left represents samples of size 30.
  - The one on the right represents samples of size 300.
  - Both could be described as normal distributions.

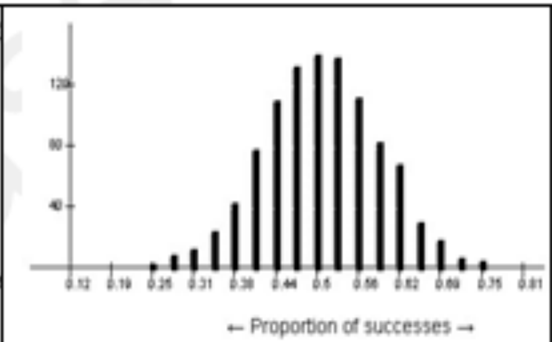




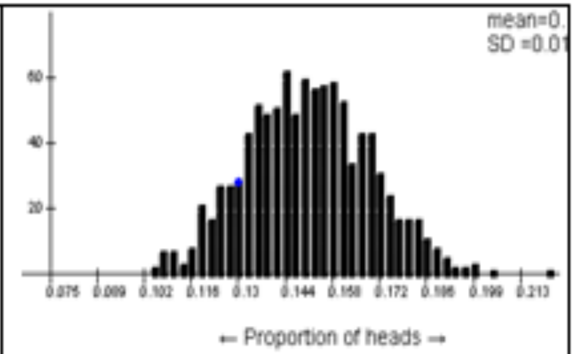
- Which ones will normal distributions fit?



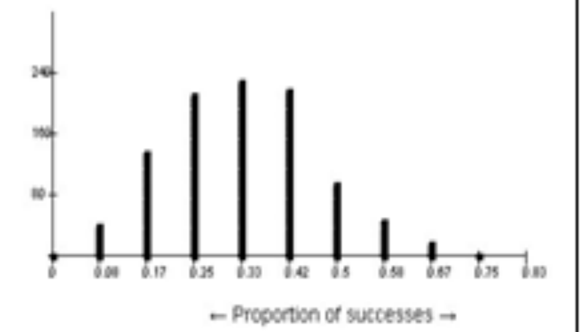
Competent face method
(n = 279; π = 0.5)



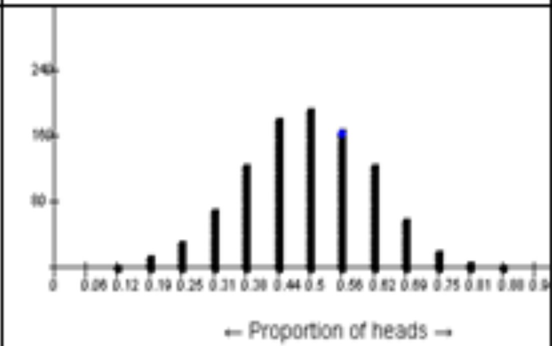
Competent face method
(n = 32; π = 0.5)



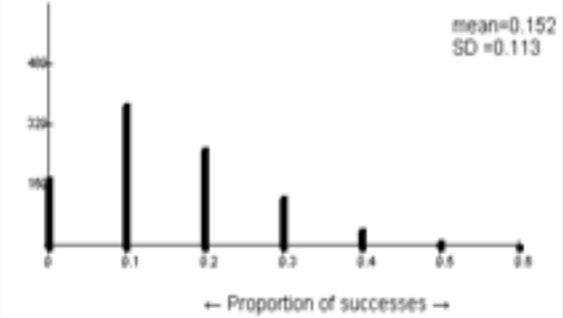
St. George's
(n = 361; π = 0.15)



Rock Paper Scissors
(n = 12; π = 0.3333)



Doris/Buzz
(n = 16; π = 0.5)



St. George's
(n = 10, π = 0.15)

When can I use a theory-based test that uses the normal distribution?

- The shape of the randomized null distribution is affected by the sample size and the proportion under the null hypothesis.
- The larger the sample size the better.
- The closer the null proportion is to 0.5 the better.
- For testing proportions, you should have at least 10 successes and 10 failures in your sample to be confident that a normal distribution will fit the simulated null distribution nicely.

Advantages and Disadvantages of Theory-Based Tests

- **Advantages of theory-based tests**
  - No need to set up some randomization method
  - Fast and Easy
  - Can be done with a wide variety of software
  - We all get the same p-value.
  - Determining confidence intervals (we will do this in chapter 3) is easier.
- **Disadvantages of theory-based tests**
  - They all come with some validity conditions (like the number of success and failures we have for a single proportion test).

Example 1.5: Halloween Treats

- Researchers investigated whether children show a preference to toys or candy
- Test households in five Connecticut neighborhoods offered children two plates:
  - One with candy
  - One with small, inexpensive toys
- The researchers observed the selections of 283 trick-or-treaters between ages 3 and 14.

Halloween Treats

- Null: The proportion of trick-or-treaters who choose candy is 0.5.
- Alternative: The proportion of trick-or-treaters who choose candy is not 0.5.

- H_0: π= 0.5
- H_a: $\pi \neq 0.5$

- 283 children were observed
  – 148 (52.3%) chose candy
  – 135 (47.7%) chose toys

for percentage problems, standard error = sd of $\hat{p}$

- Under the null distribution, the SE, which is the standard deviation of $\hat{p}$, is $\sqrt{\pi(1-\pi)/n}$ where π is the proportion under the null and n is the sample size.
- $\sqrt{\frac{0.5(1-0.5)}{283}} = 0.0297$.

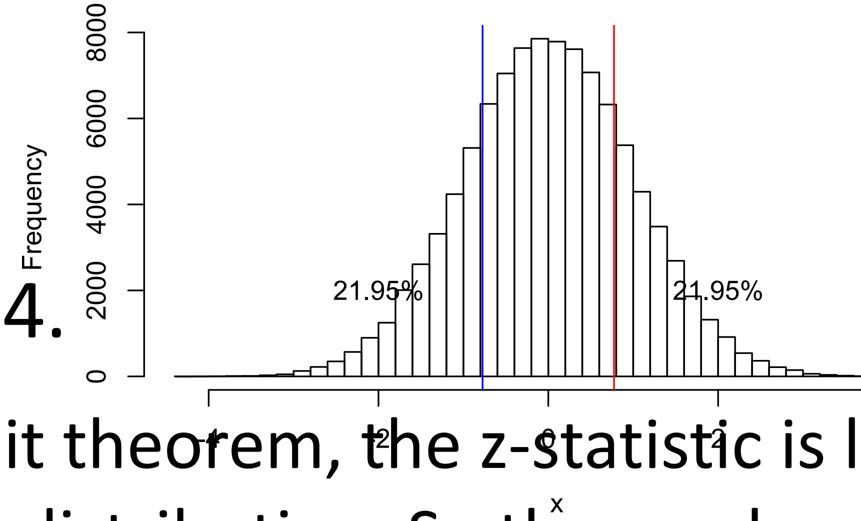
Theory-Based Inference

- The theory-based standard error works if we have a large enough sample size.
- We have 148 successes and 135 failures. Is the sample size large enough to use the theory-based method?

Standardized Statistic

- $\frac{0.523-0.5}{.0297}$ = 0.774.
- This is our Z-statistic, meaning the sample proportion is 0.774 SEs above the mean.
- Remember that a standardized statistic of more than 2 indicates that the sample result is far enough from the hypothesized value to be unlikely if the null were true.
- We had a standardized statistic that was not more than 2 (or even 1) so we do not have strong evidence against the null.

Test and p-value.



The z statistic is $\frac{0.523-0.5}{.0297}$ = 0.774.

Since n is large, by the central limit theorem, the z-statistic is l *a* draw from the standard normal distribution. So the p-value the probability of a standard normal being larger than .774 in absolute value, for a 2-sided test.

In *R*, pnorm(.774) is the prob. of a std. normal being < .774.

pnorm(.774,lower=F) = .2195 is the prob. std. normal ≥ .774.

So 2 * pnorm(.774,lower=F) is the prob. it is ≥ .774 or ≤ -.774.

2 * pnorm(.774,lower=F) = 43.9%. Not stat. sig.