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

1. Predicting faces example.
2. 2-sided tests.
3. Predicting house elections.
4. Normal distribution, CLT, and Halloween candy example.
5. Validity conditions for testing proportions.
6. Failing to reject the null vs. accepting the null, wealth and echinacea examples.

http://www.stat.ucla.edu/~frederic/13/W24 .

Read chapters 2 and 3.

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 of correct identifications if the lady always guesses. Where would you anticipate this distribution would center? Also, do you anticipate the SD of the null distribution to be negative, positive, or 0? Why?
e. Use an applet to generate the null distribution of sample proportion of correct identifications and use it to determine the standardized statistic.
f. 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.)
g. Based on the standardized statistic, state the conclusion that you would draw about the research question of 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, tight/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 o 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" beuse the difference ($A - B$) is either plus or minus. Here's ummary table:

Coin toss	Heads	P(Heads)	Null hypothesis	Statistic
zz's guess	Right	$\pi = P(\text{Right})$	$\pi = 0.50$	$\hat{p}$
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?

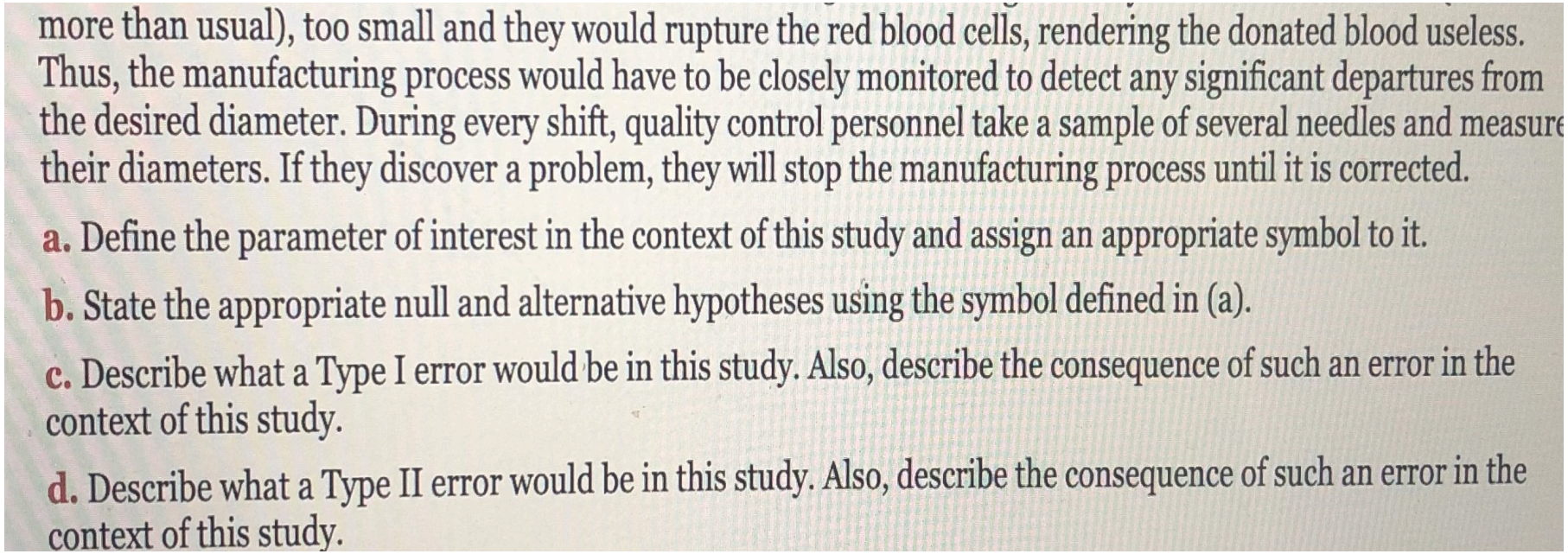
b. Based on the sample size and distance between the null value and the observed proportion, estimate the strength of evidence: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

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.

2.3.15 starts "Consider a manufacturing process that is producing hypodermic needles that will be used for blood donations. These needles need to have a diameter of 1.65mm – too big and they would hurt the donor (even



more than usual), too small and they would rupture the red blood cells, rendering the donated blood useless. Thus, the manufacturing process would have to be closely monitored to detect any significant departures from the desired diameter. During every shift, quality control personnel take a sample of several needles and measure their diameters. If they discover a problem, they will stop the manufacturing process until it is corrected.

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

b. State the appropriate null and alternative hypotheses using the symbol defined in (a).

c. Describe what a Type I error would be in this study. Also, describe the consequence of such an error in the context of this study.

d. Describe what a Type II error would be in this study. Also, describe the consequence of such an error in the context of this study.

3.3.18 starts "Reconsider the investigation of the manufacturing process that is producing hypodermic needles. Using the data from the most recent sample of needles, a 90% confidence interval for the average diameter of needles is...."
4.1.23 starts "In November 2010, an article titled 'Frequency of Cold Dramatically Cut with Regular Exercise' appeared in *Medical News Today*."

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

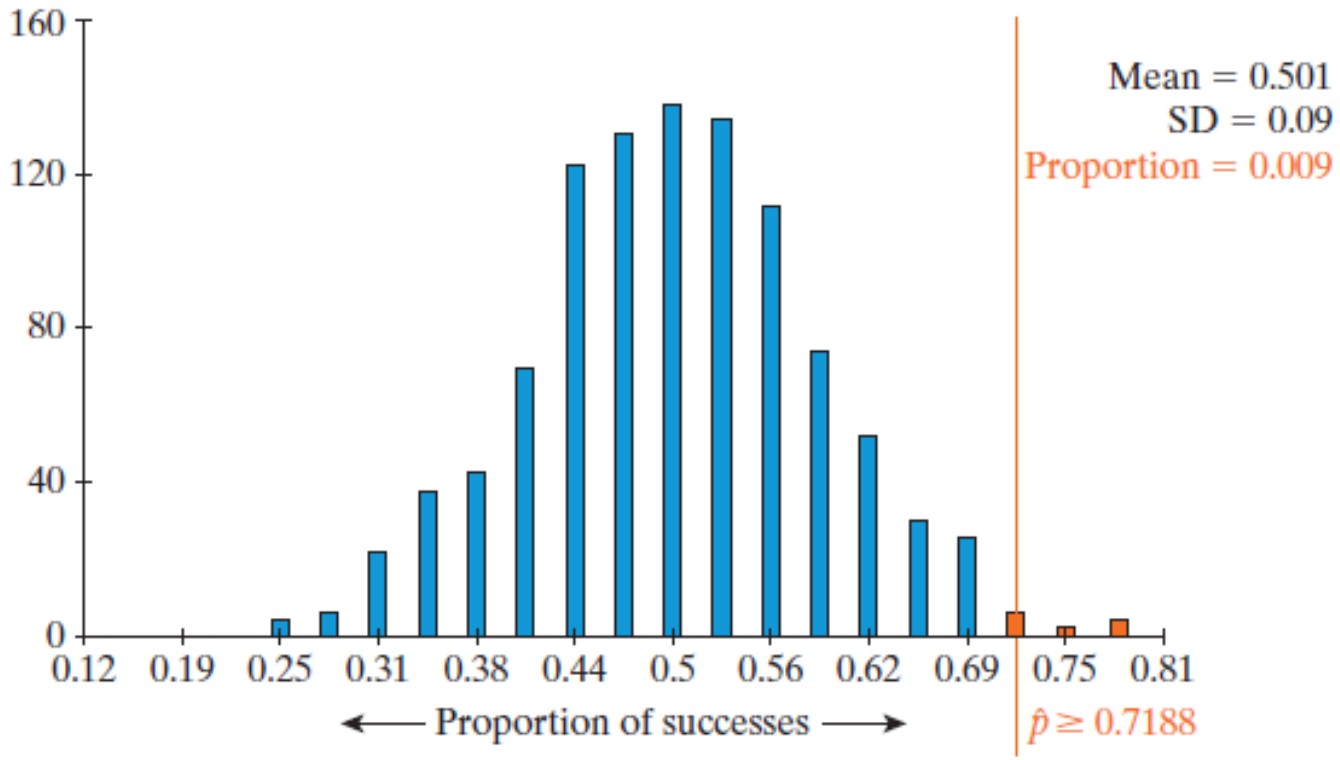
Predicting Elections

- 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 $\hat{p} = 23/32 \approx 0.719$, 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.

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

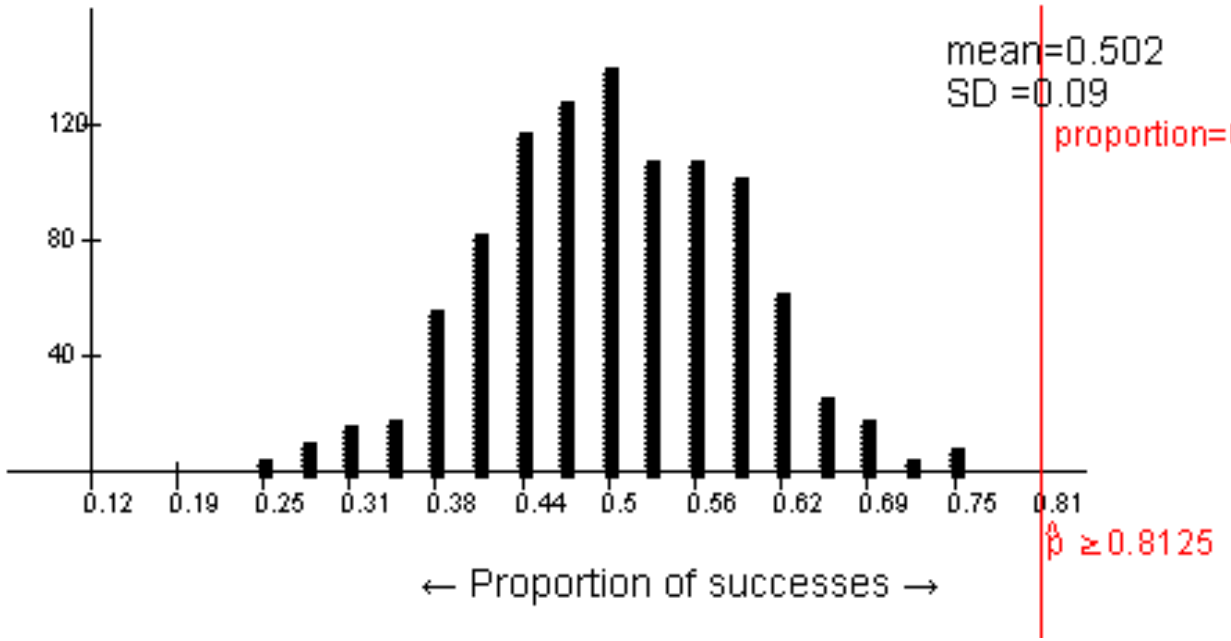
- What do the p-value and standardized statistic mean?

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

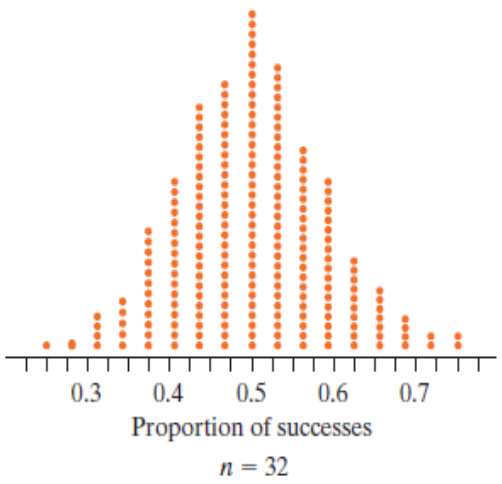
- The farther away the observed statistic is from the average value of the null distribution (or π_0), the more evidence there is against the null hypothesis.

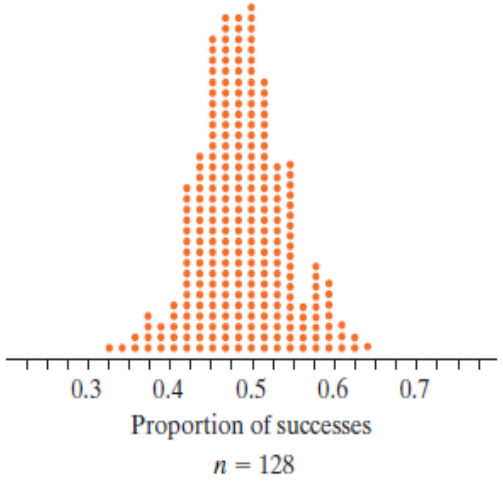
Sample Size

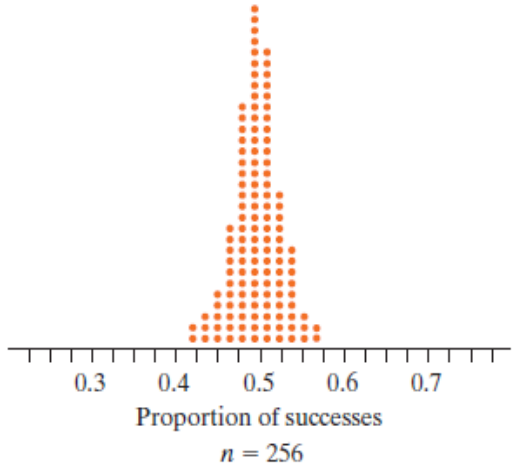
Suppose the sample proportion stays the same, do you think increasing sample size will increase, decrease, or have no impact on the strength of evidence against the null hypothesis?

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.

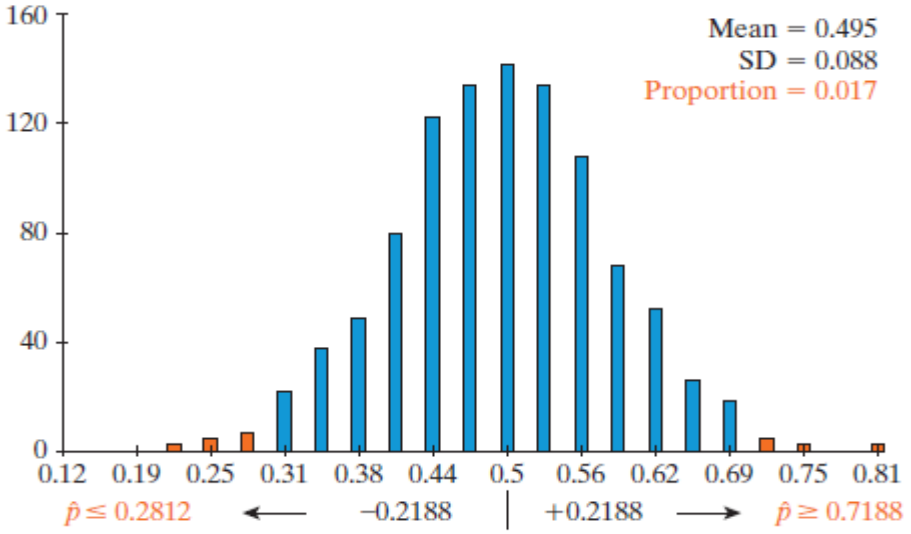
Null hypothesis: The probability this method predicts the winner equals 0.50. (H_0: $\pi = 0.50$)

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

(H_a: $\pi \neq 0.50$)

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.
- They 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

Distance of the observed statistic to the null hypothesis value

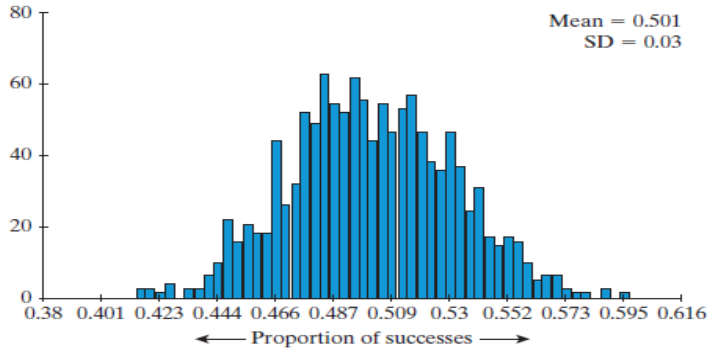
- The statistic in the House is 0.677 compared to 0.719 in the Senate
- Slight decrease in the effect size.

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 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.

1. What impacts p-values and strength of evidence? Faces example,

Section 1.4.

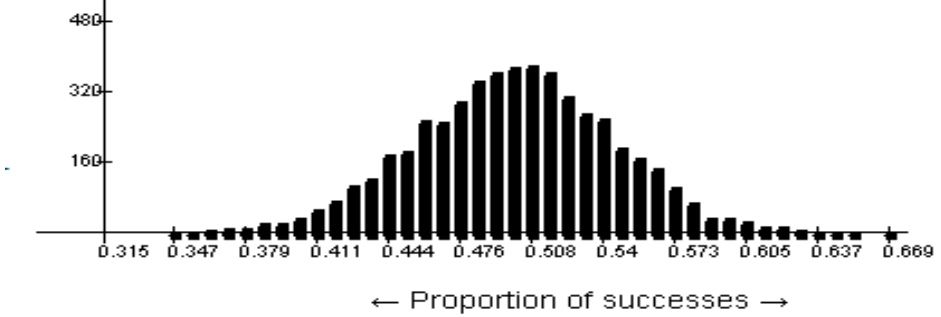
Predicting Elections from Faces

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.

4. Normal distribution, CLT, and halloween candy example.

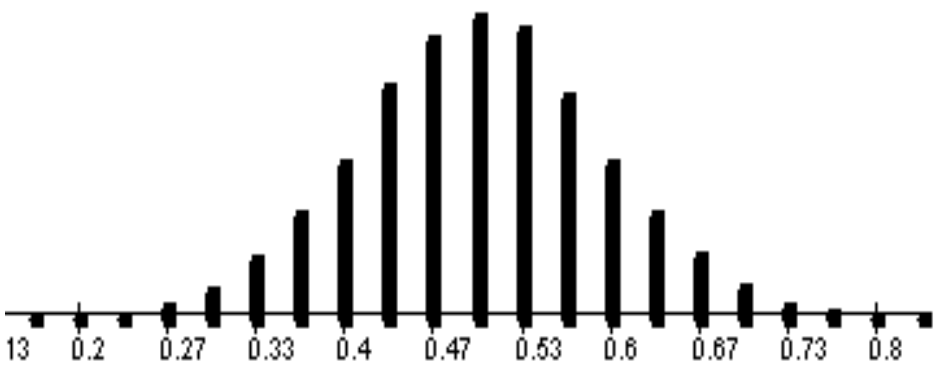
Section 1.5

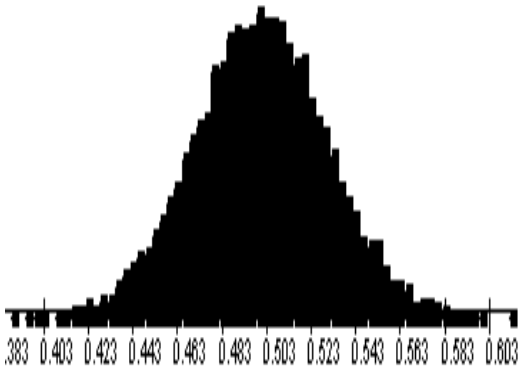


- The shape of most of our simulated null distributions always seems to be bell shaped. This shape is called the normal distribution.
- The Central Limit Theorem (CLT) dictates that, as *n* gets large, the sample mean or proportion becomes approximately normally distributed.
- When we do a test of significance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

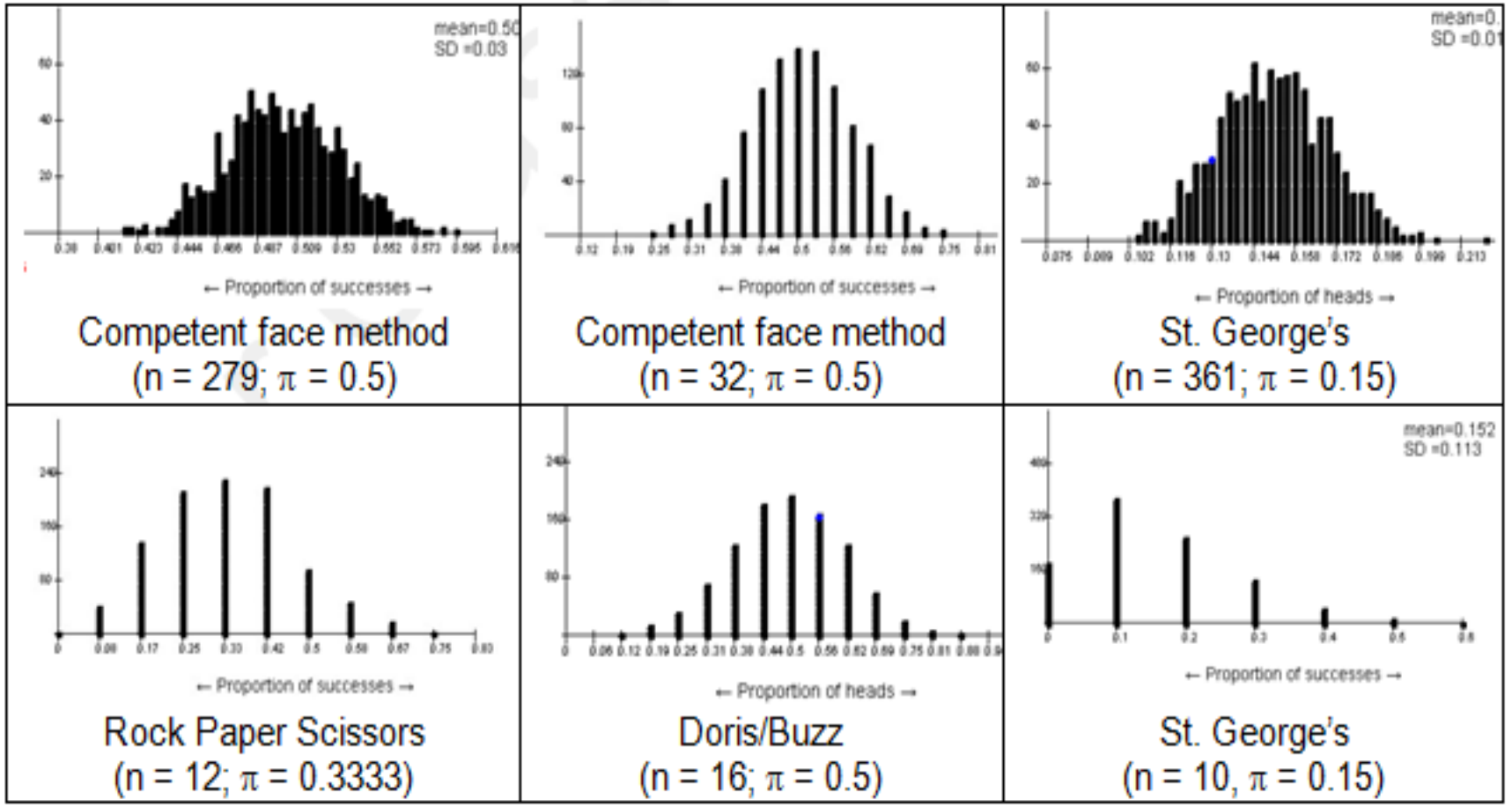
The Normal Distribution

- 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?



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 much 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: $\pi = 0.5$
- H_a: $\pi \neq 0.5$

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

Standard Deviation of $\hat{p}$

- Under the null distribution, 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 don’t really have strong evidence against the null.

Halloween Treats

- To compute the p-value in *R*,
  2*(1-pnorm(.774)) ~ 0.439.
- The theory-based p-value is 0.439 so if half of the population of trick-or-treaters preferred candy, then there's a **43.9%** chance that a random sample of 283 trick-or-treaters would have 148 or more, or 135 or fewer, candy choosers.
- Since 43.9% is not a small p-value, we don't have strong (or even moderate) evidence that trick-or-treaters prefer one type of treat over the other. We cannot reject the null hypothesis.

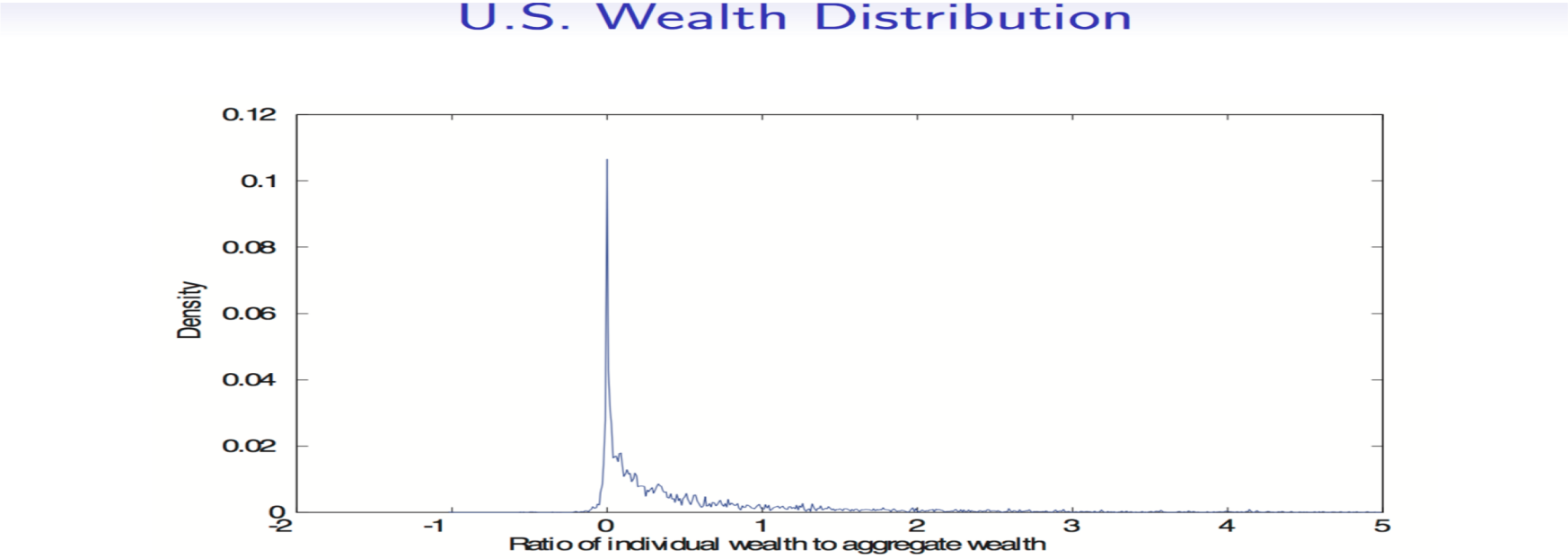
5. Validity conditions for testing proportions.

- You should have at least 10 successes and 10 failures in your sample to be confident a normal distribution will fit the simulated null distribution nicely.
- Your observations should be (at least approximately) independent. We will discuss what this means later in this lecture when we talk about sampling.

6. Failing to reject the null hypothesis vs. accepting the null hypothesis.

- Benoit Mandelbrot.

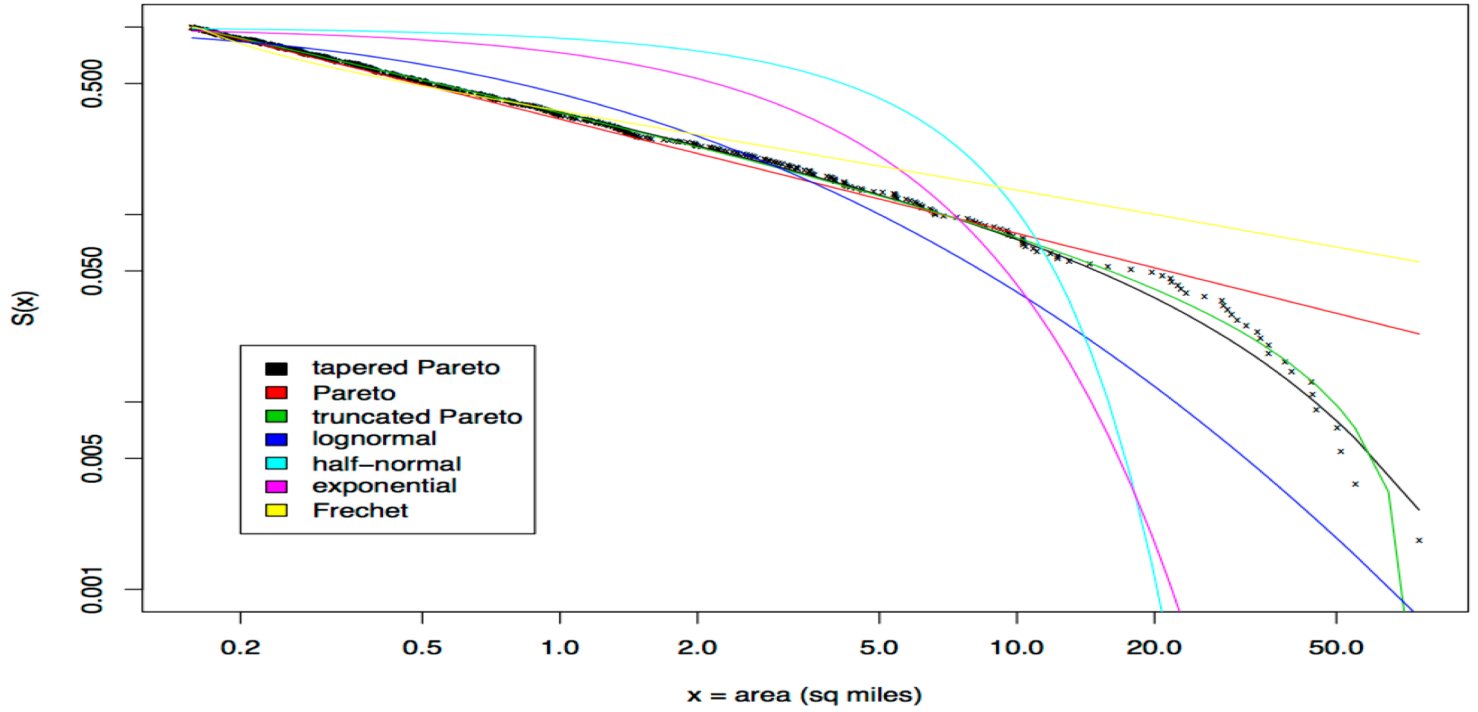
  We've tested it on many datasets and found the Pareto distribution "fits perfectly".
- from B. Moll (2012).



Failing to reject the null vs. accepting the null.

- Benoit Mandelbrot.

  We've tested it on many datasets and found the Pareto distribution "fits perfectly".



Failing to reject the null vs. accepting the null.

- Benoit Mandelbrot.
  We've tested it on many datasets and found the Pareto distribution "fits perfectly".
- Think about it. What is the null hypothesis of the test. Is it possible to show that the model fits perfectly?
- You might not reject the null with a certain n, and then as n grows, you reject it.
- Nowadays people are using the tapered Pareto distribution instead of the Pareto.
- Echinacea vs. placebo. n = 58. Oneil et al. 2008.

Failing to reject the null vs. accepting the null.

- 28 in echinacea group and 30 in placebo group.
- "[V]olunteers recruited from hospital personnel were randomly assigned to receive 3 capsules twice daily of either placebo (parsley) or E. purpurea [echinacea] for 8 weeks during the winter months. Upper respiratory tract symptoms were reported weekly during this period.
- "Individuals in the echinacea group reported 9 sick days per person during the 8-week period, whereas the placebo group reported 14 sick days ($z = -0.42$; $P = .67$)."

Failing to reject the null vs. accepting the null.

- conclusion in Oneil et al. (2008), "commercially available E. purpurea capsules did not significantly alter the frequency of upper respiratory tract symptoms compared with placebo use."
- From sciencebasedmedicine.org, "[The study] added to the evidence that *Echinacea* is not useful for prevention of colds or flus. They found no difference in incidence of cold symptoms."
- ABC News headline "Study: Echinacea no help for colds".

Failing to reject the null vs. accepting the null.

Failing to reject the null vs. accepting the null.

Today, most of the evidence seems to indicate that echinacea does boost the immune system a little bit and help to fight colds. From WebMD: "Extracts of echinacea do seem to have an effect on the immune system, your body's defense against germs. Research shows it increases the number of white blood cells, which fight infections. A review of more than a dozen studies, published in 2014, found the herbal remedy had a very slight benefit in preventing colds."