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

0. Hand in HW1.
1. Strength of evidence, and 1 and 2 sided tests.
2. Normal distribution, CLT, and Halloween candy example.
3. Validity conditions for testing proportions.
4. Failing to reject the null vs. accept the null, wealth and echinacea examples.
5. Sampling, bias, and students example.

Read chapter 2.
http://www.stat.ucla.edu/~frederic/13/F17 .

0. Hand in HW1.

1. What affects the strength of evidence?

A. The difference between the observed statistic ($\hat{p}$) and null hypothesis parameter (π_0).
B. Sample size.
C. If we do a one or two-sided test.

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

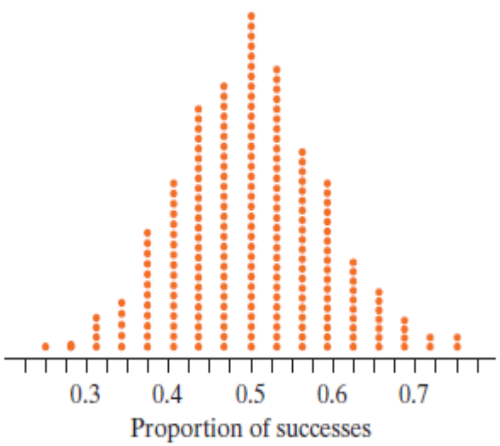
- The farther away the observed statistic is from the average value of the null distribution (or π_0), the more evidence there is against the null hypothesis.

Sample Size

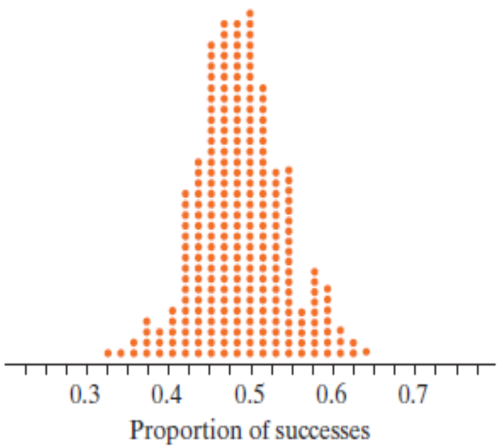
Suppose the sample proportion stays the same, do you think increasing sample size will increase, decrease, or have no impact on the strength of evidence against the null hypothesis?

Sample Size

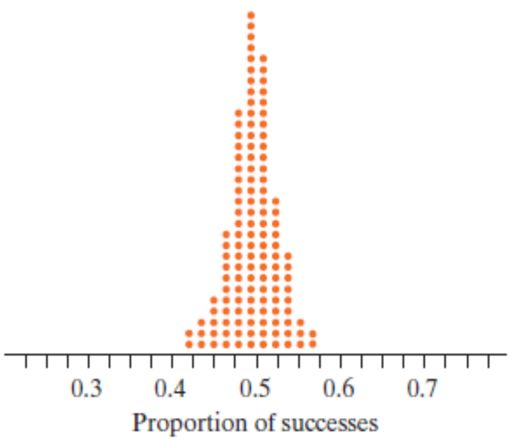
- The null distribution changes as we increase the sample size from 32 senate races to 128 races to 256 races.
- As the sample size increases, the variability (standard error) decreases.



$n = 32$



$n = 128$



$n = 256$

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).

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 large p-value, > 50%.
- This is a ***one-sided*** test.
- In practice, most research uses two-sided tests.

2-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 trustworthy face 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.

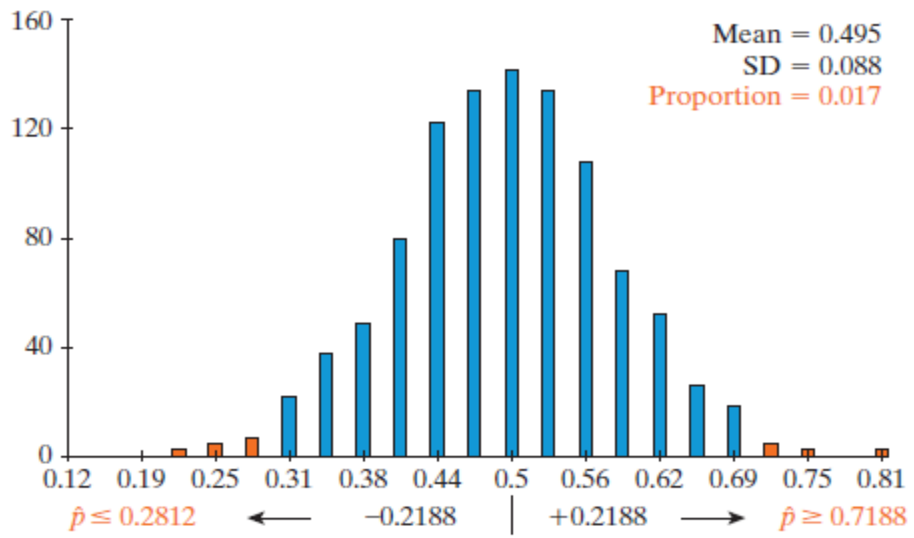
(H_a: $\pi \neq 0.50$)

2-Sided Tests

- The change to the alternative hypothesis also affects how we compute the p-value.
- Remember that the p-value is the probability (assuming the null hypothesis is true) of obtaining a proportion that is equal to or **more extreme** than the observed statistic
- In a *two-sided test*, **more extreme** goes in both directions.

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 $189/279 \approx 0.677$, 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

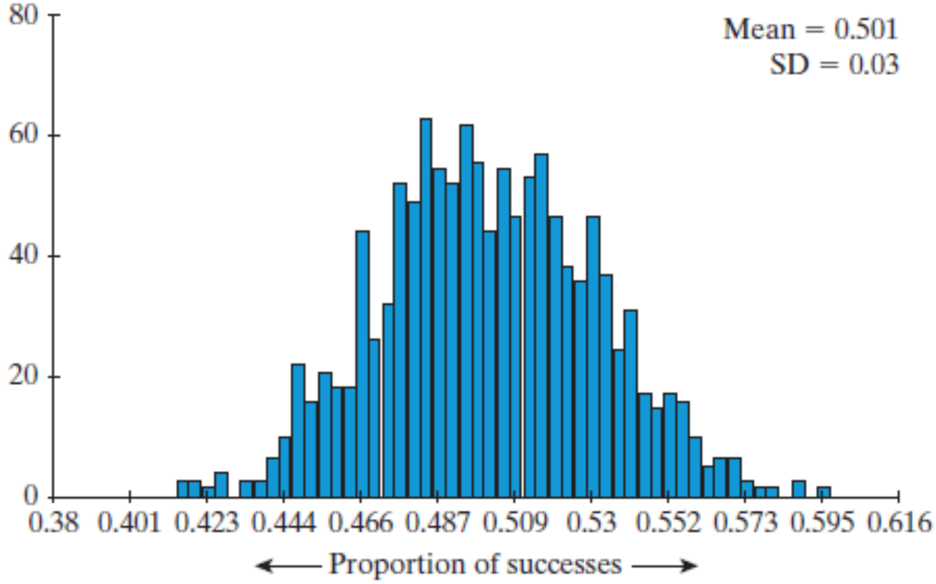
- The statistic in the House is 0.677 compared to 0.719 in the Senate
- Slight decrease in the effect size.

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

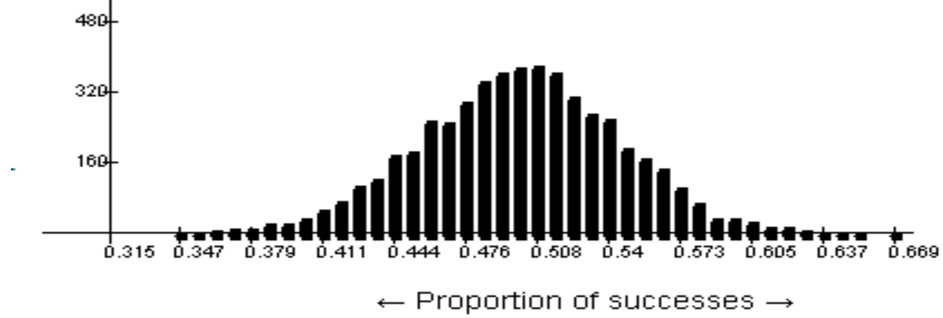
- What about the standardized statistics?
  - For the Senate it was 2.43
  - For the House is 5.90.
- The larger sample size for the House outweighed the smaller effect size in this particular case. We have stronger evidence against the null using the data from the House.

1-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.

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

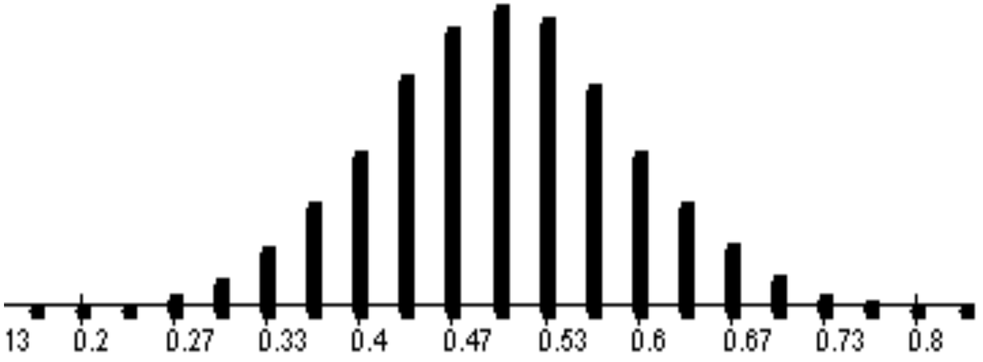
Section 1.5

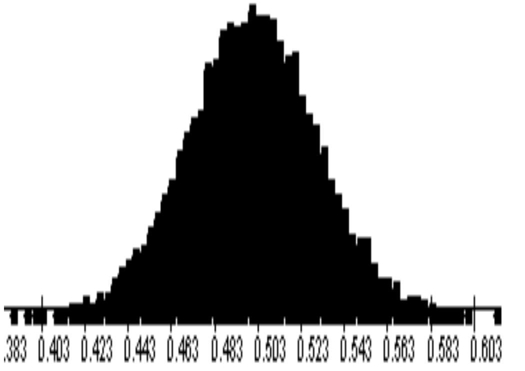


- The shape of most of our simulated null distributions always seems to be bell shaped. This shape is called the normal distribution.
- The Central Limit Theorem (CLT) dictates that, as *n* gets large, the sample mean or proportion becomes approximately normally distributed.
- When we do a test of significance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

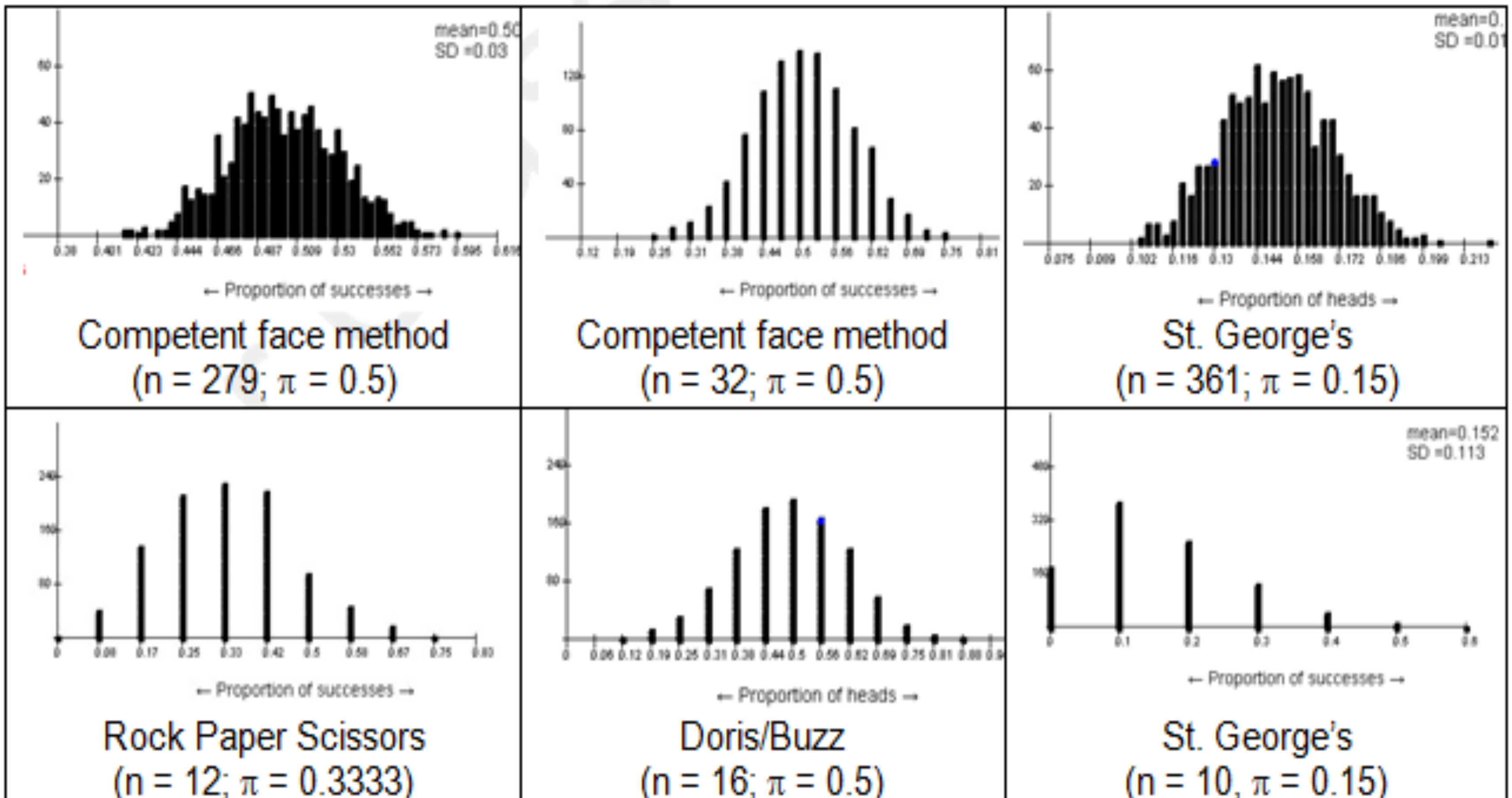
The Normal Distribution

- 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 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 *p*

- Under the null distribution, the standard deviation of 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 any evidence against the null.

Halloween Treats

- To compute the p-value in *R*,
  2*(1-pnorm(.774)) ~ 0.439.

pnorm(x) means the prob. of a standard normal < x, so

1-pnorm(.774) is the prob. of a std. normal being ≥ x,

and 2 times this is the prob. of it being ≥ x or ≤ -x.

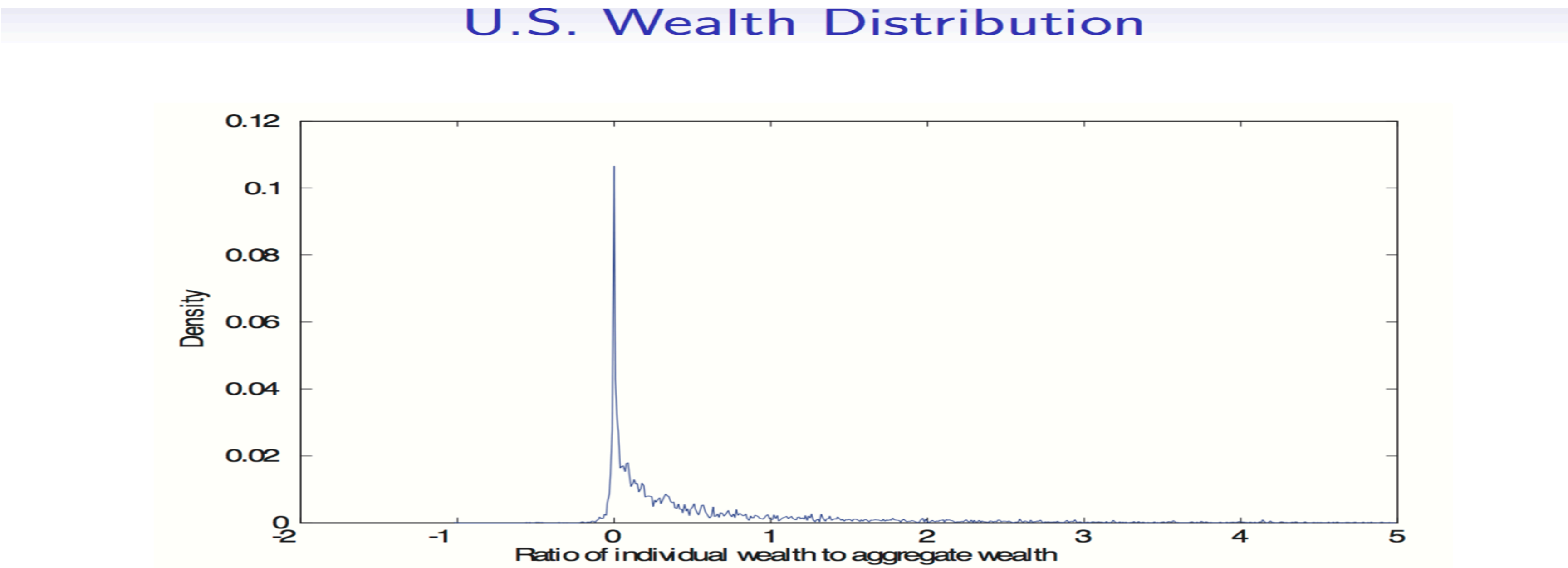
- The theory-based p-value is 0.439 so if half of the population of trick-or-treaters preferred candy, then there's a **43.9%** chance that a random sample of 283 trick-or-treaters would have 148 or more, or 135 or fewer, candy choosers.
- Since 43.9% is not a small p-value, we don't have strong (or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

3. 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 when we talk about sampling in chapter 2.

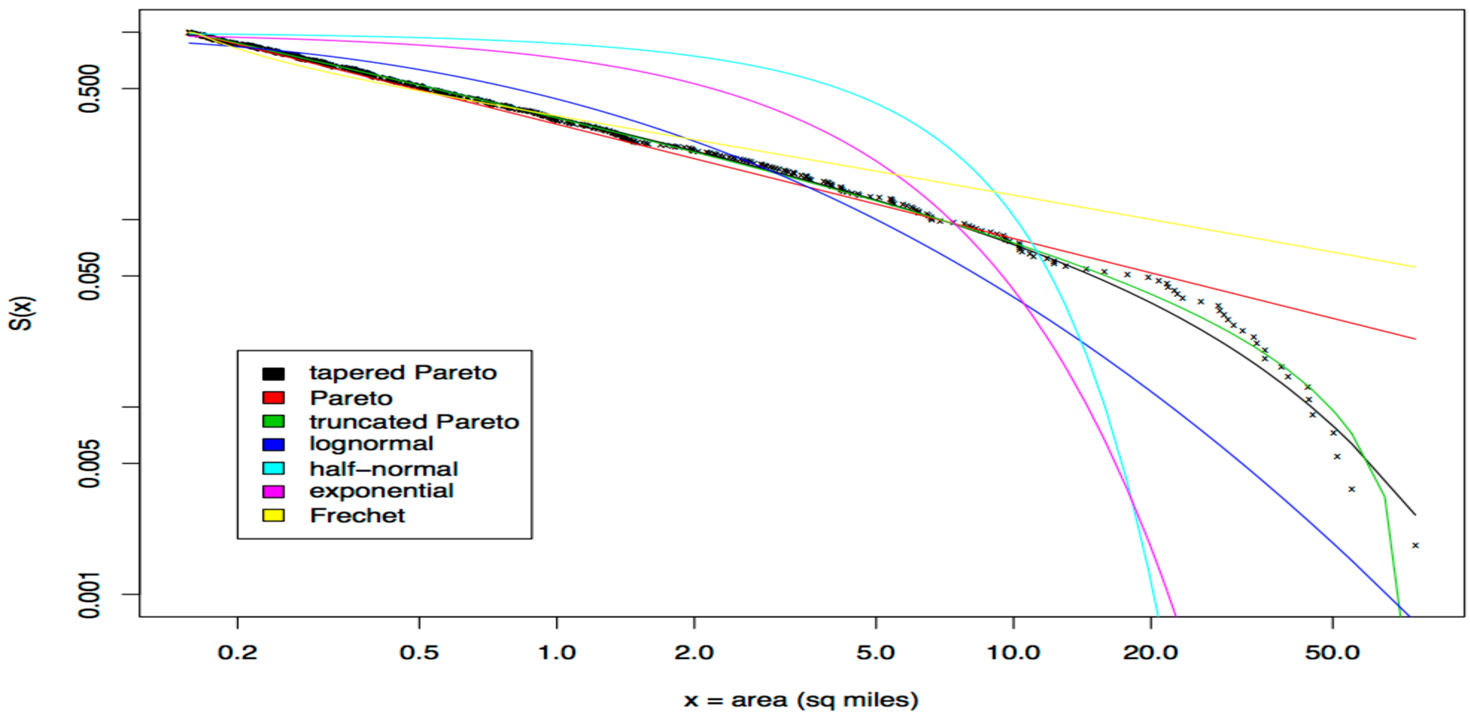
4. Failing to reject the null vs. accepting the null.

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



4. 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".



4. 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.

4. 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 ($z = -0.42$; $P = .67$)."

4. Failing to reject vs. accepting

- 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".

4. Failing to reject Ho vs. accepting Ho.

Cold and flu on NBCNEWS.com

Got a cold? Sorry, echinacea won't help much

Study shows the popular herbal remedy may bring milder symptoms — but that could be due to chance

Recommend 7

CNN Health » Diet + Fitness | Living Well | Parenting + Family

Echinacea fails to curb the common cold

4. Failing to reject vs. accepting.

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."

Sampling Students

Example 2.1A

Sampling Students

- We will look at data collected from the registrar's office from the College of the Midwest for ALL students for Spring 2011

Student ID	Cumulative GPA	On campus?
1	3.92	Yes
2	2.80	Yes
3	3.08	Yes
4	2.71	No
5	3.31	Yes
6	3.83	Yes
7	3.80	No
8	3.58	Yes
…	…	…

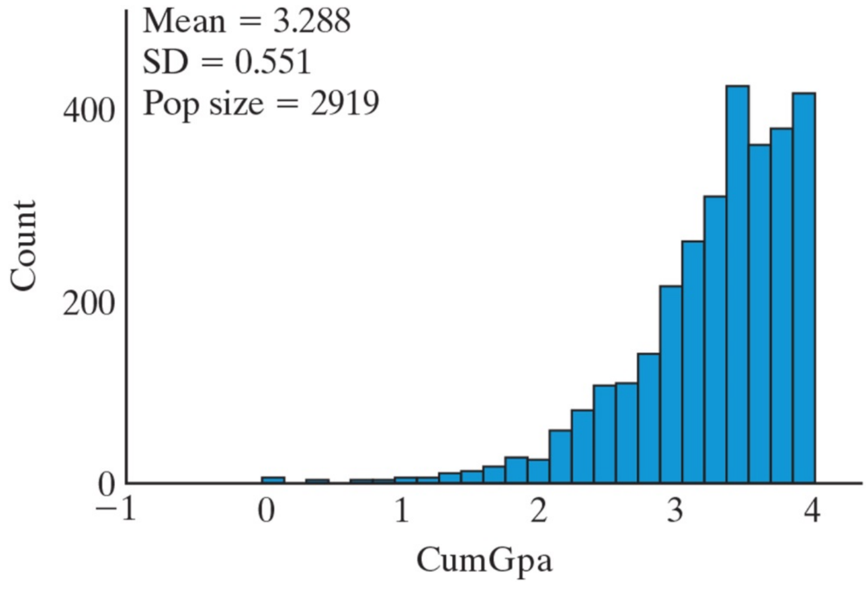
Sampling Students

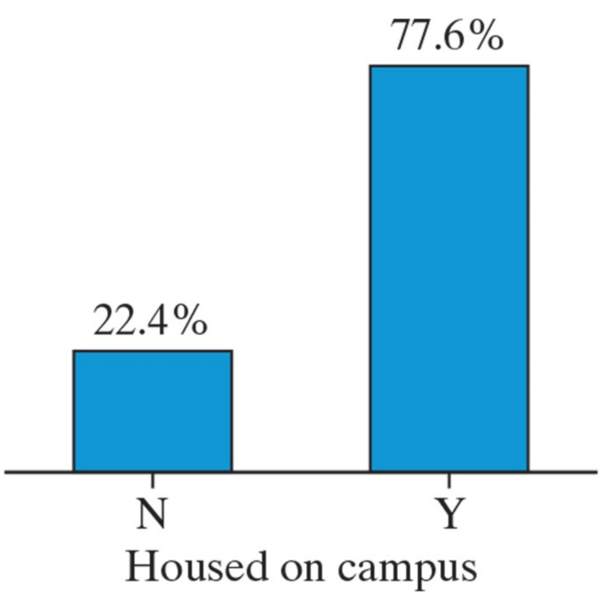
- What type of variable is “On campus”?
- What type is Cumulative GPA?

Student ID	Cumulative GPA	On campus?
1	3.92	Yes
2	2.80	Yes
3	3.08	Yes
4	2.71	No
5	3.31	Yes
6	3.83	Yes
7	3.80	No
8	3.58	Yes
…	…	…

Sampling Students

- Here are graphs (a histogram and a bar graph) representing all of the 2919 students at the College of the Midwest for our two variables of interest.





Sampling Students

- We usually don’t have information on an entire population like we do here.
- We usually need to make inferences about a population based on a sample.
- Suppose a researcher asks the first 30 students she finds on campus one morning whether they live on campus. This would be a quick and convenient way to get a sample.

Sampling Students

For this scenario:

- What is the population?
- What is the sample?
- What is the parameter
- What is the statistic?
- Do you think this quick and convenient sampling method will result in a similar sample proportion to the population proportion?

Sampling Students

- The researcher's sampling method might overestimate the proportion of students that live on campus because if it is taken early in the morning most of those that live off campus might not have arrived yet.
- We call this sampling method ***biased***.
- A sampling method is ***biased*** if statistics from samples over or under-estimate the population parameter on average.

Sampling Students

- Bias is a property of a sampling *method*, not the sample
  - A method must *consistently* (on average) produce non-representative results to be considered biased
- Sampling bias also depends on what is measured
  - Would the morning sampling method be biased in estimating the average GPA of students at the college?
  - What about estimating the proportion of students wearing orange shirts?

Sampling Students

- What's a better way of selecting a representative sample?
- Use a ***random*** mechanism to select the observational units
- Don't rely on ***convenience samples***
- A ***Simple Random Sample (SRS)*** is where every collection of size n is equally likely to be the sample selected from the population.

Sampling Students

- How could we take a Simple Random Sample of 30 students from the College of the Midwest?
- Represent each student by ID numbers 1 to 2919
- Have the computer randomly select 30 numbers between 1 and 2919

Sampling Students

IDs of the 30 people selected, along with their cumulative GPA and residential status

ID	Cum GPA	On campus?	ID	Cum GPA	On campus?	ID	Cum GPA	On campus?
827	3.44	Y	844	3.59	N	825	3.94	Y
1355	2.15	Y	90	3.30	Y	2339	3.07	N
1455	3.08	Y	1611	3.08	Y	2064	3.48	Y
2391	2.91	Y	2550	3.41	Y	2604	3.10	Y
575	3.94	Y	2632	2.61	Y	2147	2.84	Y
2049	3.64	N	2325	3.36	Y	2590	3.39	Y
895	2.29	N	2563	3.02	Y	1718	3.01	Y
1732	3.17	Y	1819	3.55	N	168	3.04	Y
2790	2.88	Y	968	3.86	Y	1777	3.83	Y
2237	3.25	Y	566	3.60	N	2077	3.46	Y

Sampling Students

- What is the average cum GPA for these 30 students?
  - $\bar{x}$ is the sample average
  - $\bar{x}$ = 3.24
- What proportion live on campus?
  - $\hat{p}$ is the sample proportion
  - $\hat{p}$ = 0.80
- μ is the population mean.
- π is the population proportion.

Sampling Students

- How do we know if $\bar{x}$ and $\hat{p}$ are close to the population values, μ and π?
- A different sample of 30 students would probably have had different values.
- How are these statistics useful in estimating the population parameter values?
- Let's take more simple random samples of 30 students to examine the null distribution of the statistics from other samples.