Teaching the Logic and Scope of Statistical Inference

Allan Rossman, Beth Chance

Department of Statistics

Cal Poly – San Luis Obispo

arossman@calpoly.edu

bchance@calpoly.edu

Logic of statistical inference

- **Significance**: Is there a difference or effect?
  - Strength of evidence, *p*-value
- **Estimation**: How much of a difference or effect?
  - Parameter, confidence interval

Scope of statistical inference

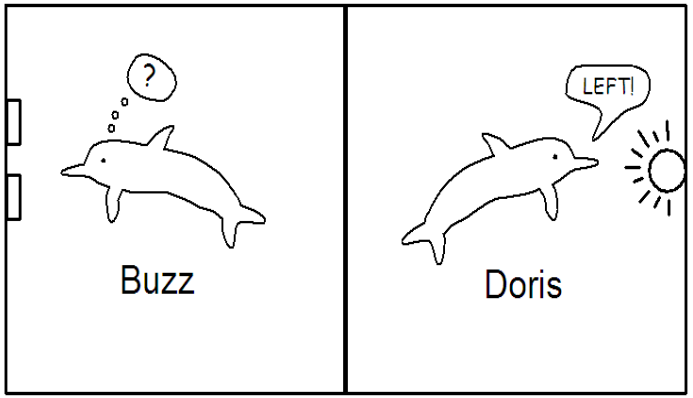
- **Generalizability**: To what population can the results be generalized?
  - Bias, randomly sampling
- **Causation**: Can a cause-and-effect relationship be concluded?
  - Observational study, confounding
  - Experiment, random assignment

Recommendations

- Emphasize these four **pillars** of statistical inference
  - Use real data from genuine studies
  - Focus on concepts, not mechanics
  - Feature simulation-based inference rather than traditional theory-based inference

(Our presentation will try to model these.)

Example 1: Buzz ar



- Can Dolphins Commu)60s)
  - Buzz and Doris trained to push left or right button depending on whether they are shown a steady light or a flashing light
  - Then a curtain is put between them and the light is only shown to Doris. She has to communicate to Buzz which button to push.
  - If Buzz pushes the correct button 15 out of 16 times, are you convinced that dolphins can communicate such abstract ideas?

Example 1: Buzz and Doris

- What are two possible explanations for Buzz getting so many correct?
  - Dolphins can communicate
  - Buzz just got lucky
- Students model the "by chance alone" explanation
  - Flip a coin 16 times, record the number of heads
  - Pool class results together
  - How surprising is 15 heads?
  - Move to technology to expand the simulation model
- p-value = how often do we find a result at least as extreme as 14 out of 16 from a purely 50/50 random process?

Probability of heads: 0.5

Number of tosses: 16

Number of repetitions: 10000

☑ Animate

Draw Samples

Total = 10000

◉ Number of heads

○ Proportion of heads

As extreme as ≥ 14 Count

Proportion of repetitions:
21 / 10000 = 0.0021

☐ Two-sided

☐ Exact Binomial

☐ Normal Approximation

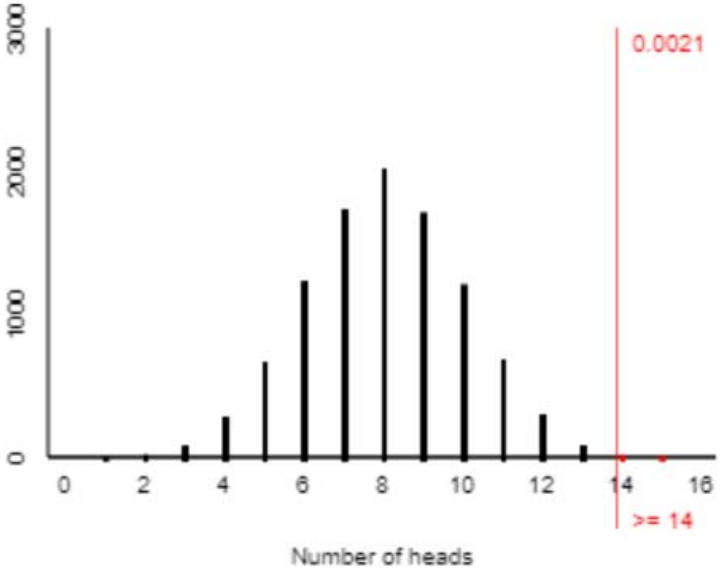
Reset

All Attempts(Last Repetition)

Heads(Last Repetition) = 6

Tails(Last Repetition) = 10

☐ Summary Stats



What if questions

- How many would he have had to get correct to convince you?
- Does this mean dolphins can communicate abstract ideas?
- What about the follow-study with a wooden barricade where Buzz only got 16/28 correct?
  - What changes in the simulation?
  - Does this prove Buzz just got lucky?

Statistical Investigation Process

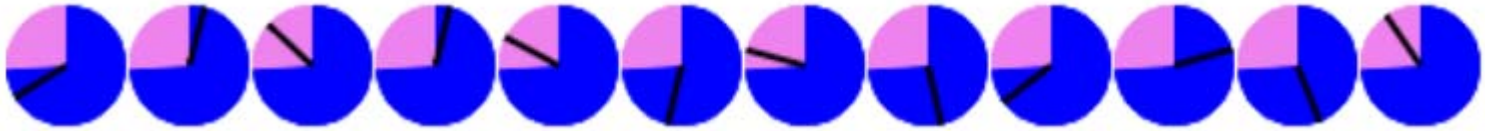
- Statement of research conjecture
  - What is “testable”?
- Data exploration
- Statistical significance – could this result have happened by chance alone?
  - Strength of evidence against chance model
- Summary conclusions/ Suggest follow-up investigations

Example 2: Kissing

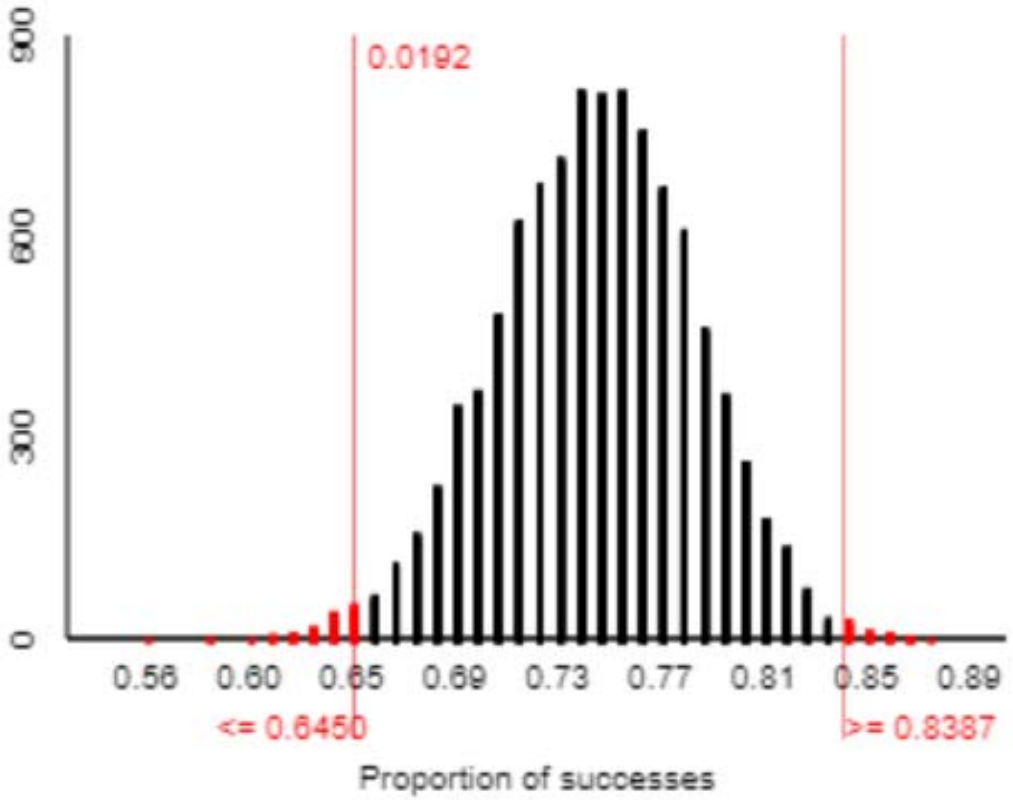
- Do kissing couples tend to lean heads to right?
- Researchers observed kissing couples in public places, recorded direction of head lean
- Sample: 80 of 124 ($\hat{p} \approx .645$) leaned to right
- Is this significantly different from 74%?
- Simulate!

New model

- Every couple is an observation from an identical process
- Probability of "success" is 0.74



- Sample size is 124
- Evidence against null hypothesis?
  - Two-sided p-value
  - More than 2SD away?

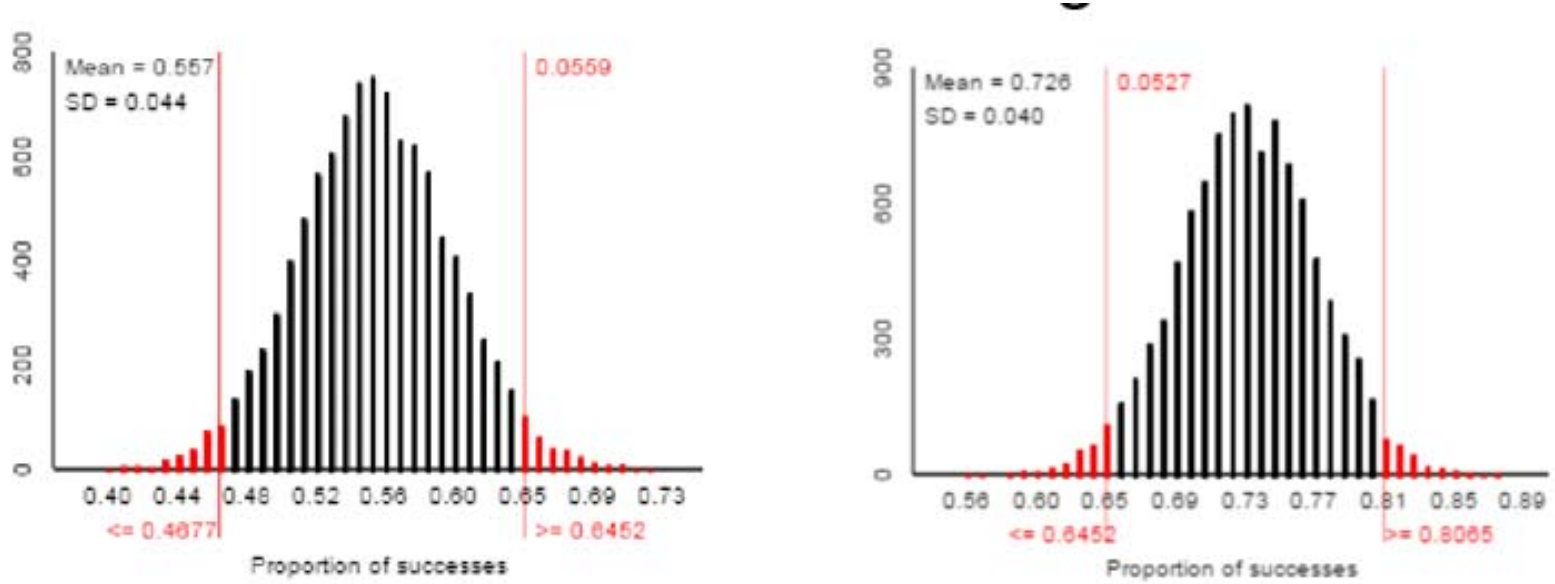


- Sample result would be pretty surprising if probability was 0.74
- Strong evidence the probability ≠ 0.74
- So what is it?

Interval estimation

- Define parameter: π = population proportion of all kissing couples who lean to the right
- Estimate π with interval of values with high degree of confidence
- One option: Use simulation to determine values of π for which sample result is *not* surprising (e.g., two-sided p-value > .05)

Interval estimation



- Boundary values ≈ (.557, .726)
- We can be 95% confident that between 55.7% and 72.6% of all couples lean to right

Interval estimation

- More conventional confidence interval for π:
  - $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - $.645 \pm 1.96 \sqrt{\frac{.645(1-.645)}{124}}$
  - .645 ± .084
  - (.561, .729), very similar to previous interval

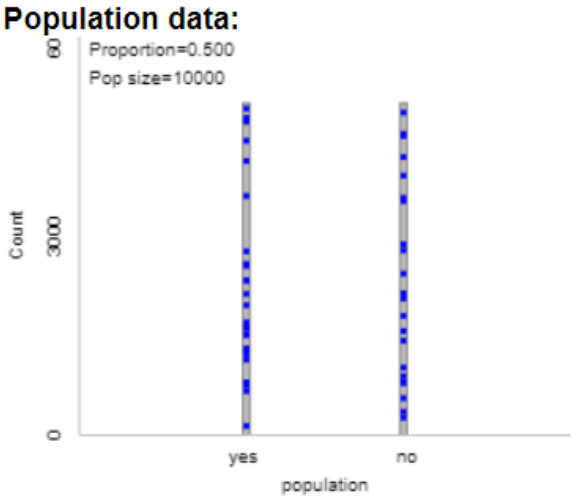
Example 3

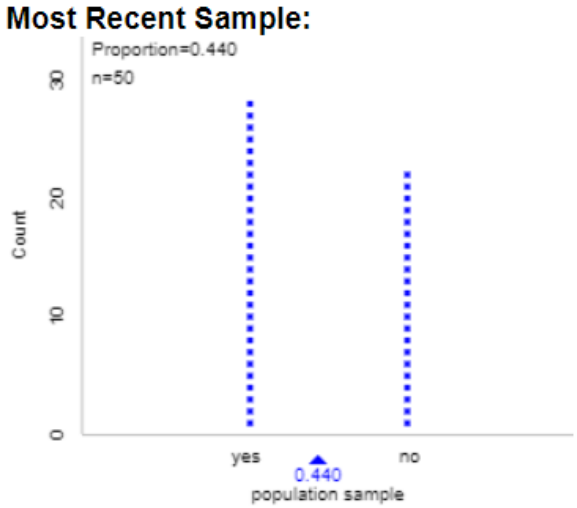
- **Tara Golshan (Vox Media)** Right now, these numbers are showing about 53 percent of Republicans either want [the corporate tax rate] to be raised or stay the same. When selling this idea, do you see that as becoming a problem?
- **Lee Zeldin (R-NY)** What I have come in contact with would reflect different numbers. So it would be interesting to see an accurate poll of 100 million Americans. But sometimes the polls get done of 1,000 [people].

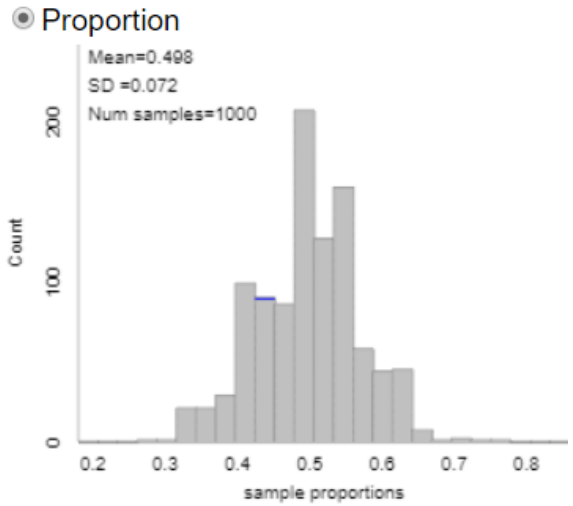
New model

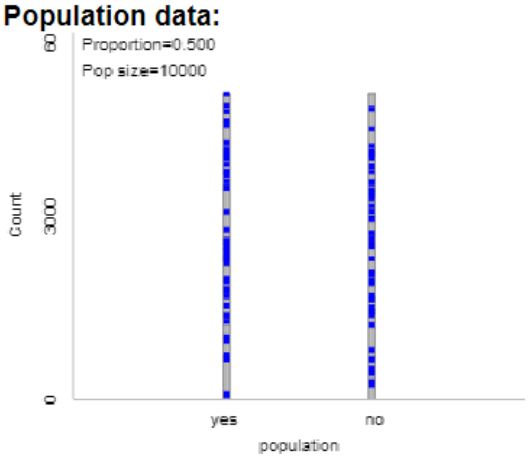
- Is 0.50 a plausible value for the proportion of all US adults who are in favor?
- How do sample proportions behave when taking random samples from a finite population?
  - Depends on sample size?
  - Depends on population size?

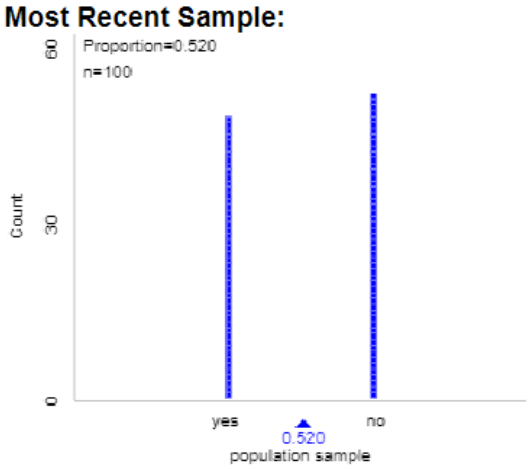
Population of 10,000 yeses/nos

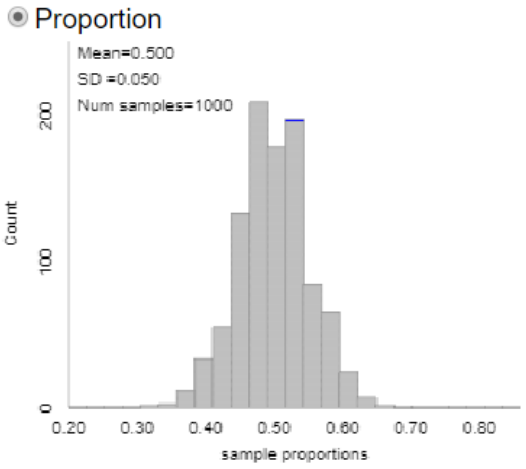


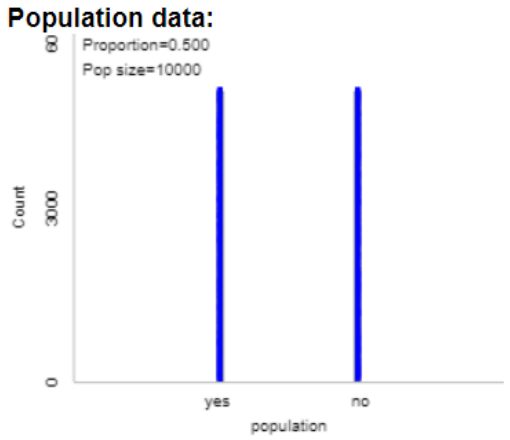


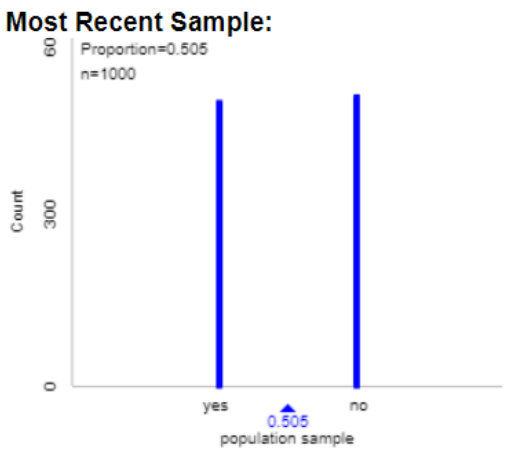


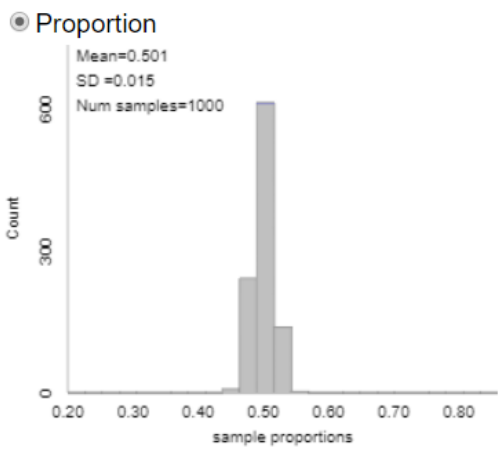




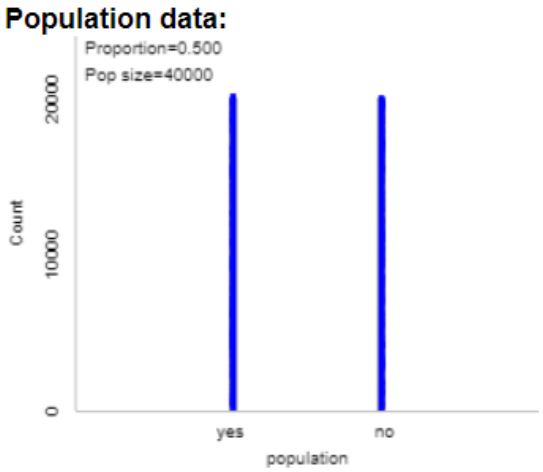


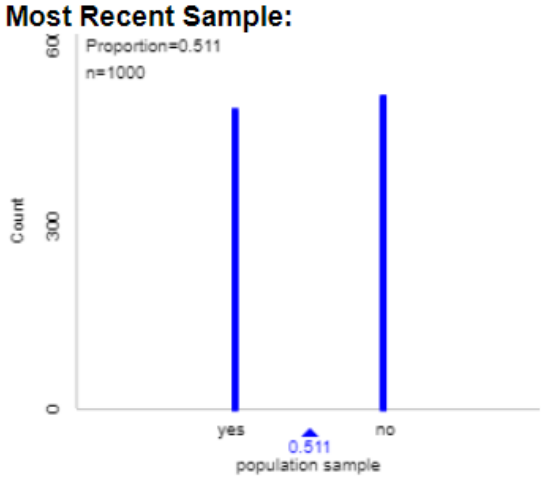


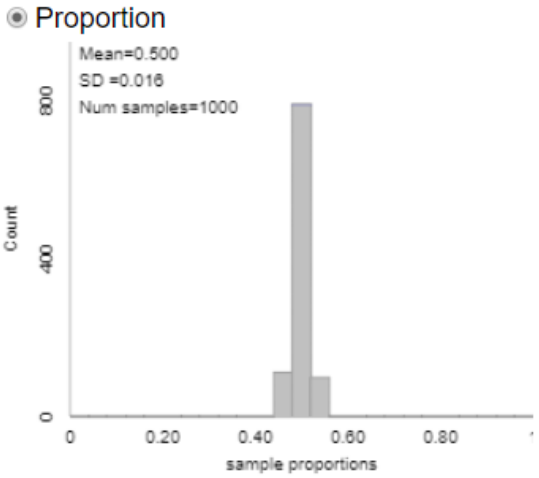




- Population of 40,000?

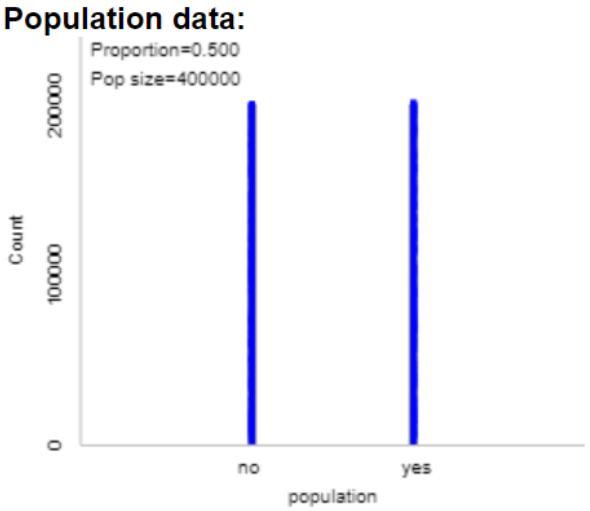


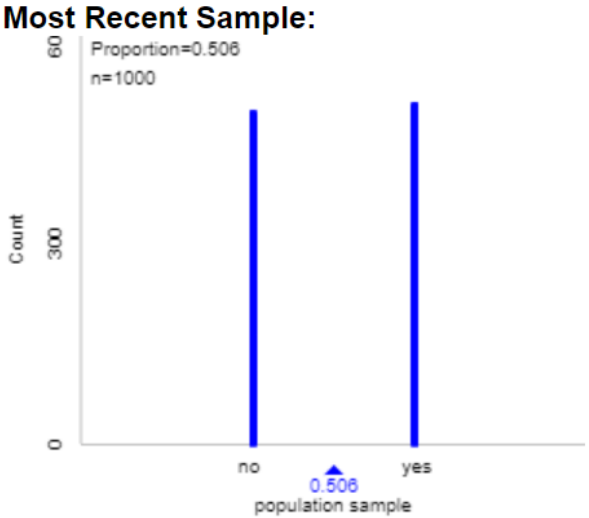


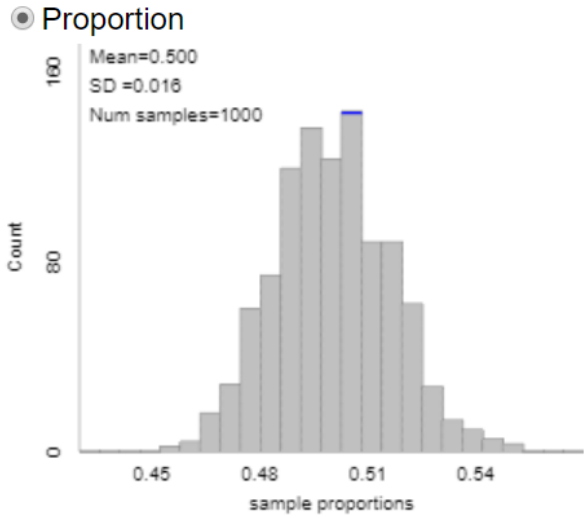


Random sampling

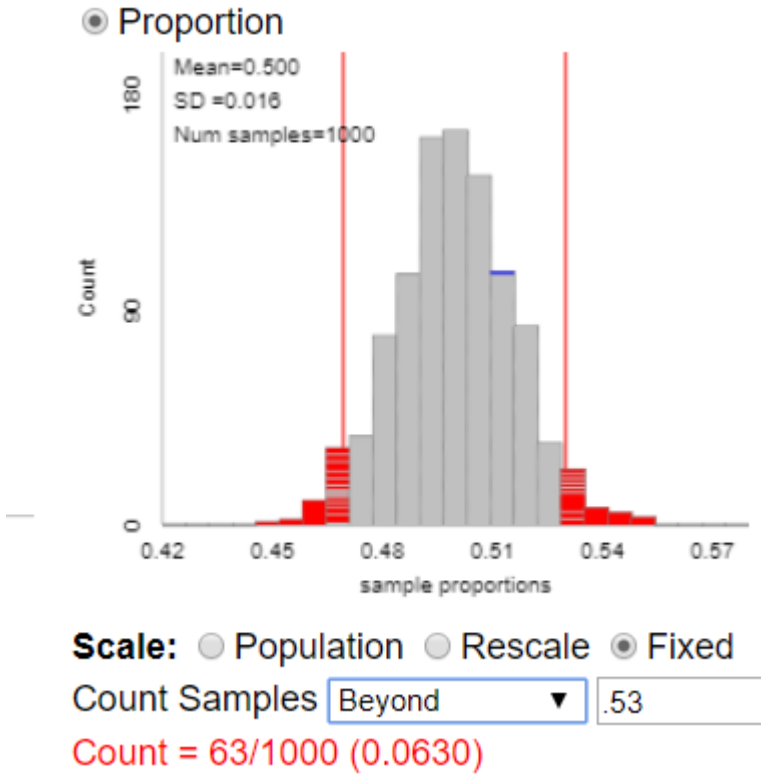
- Population of 400,000







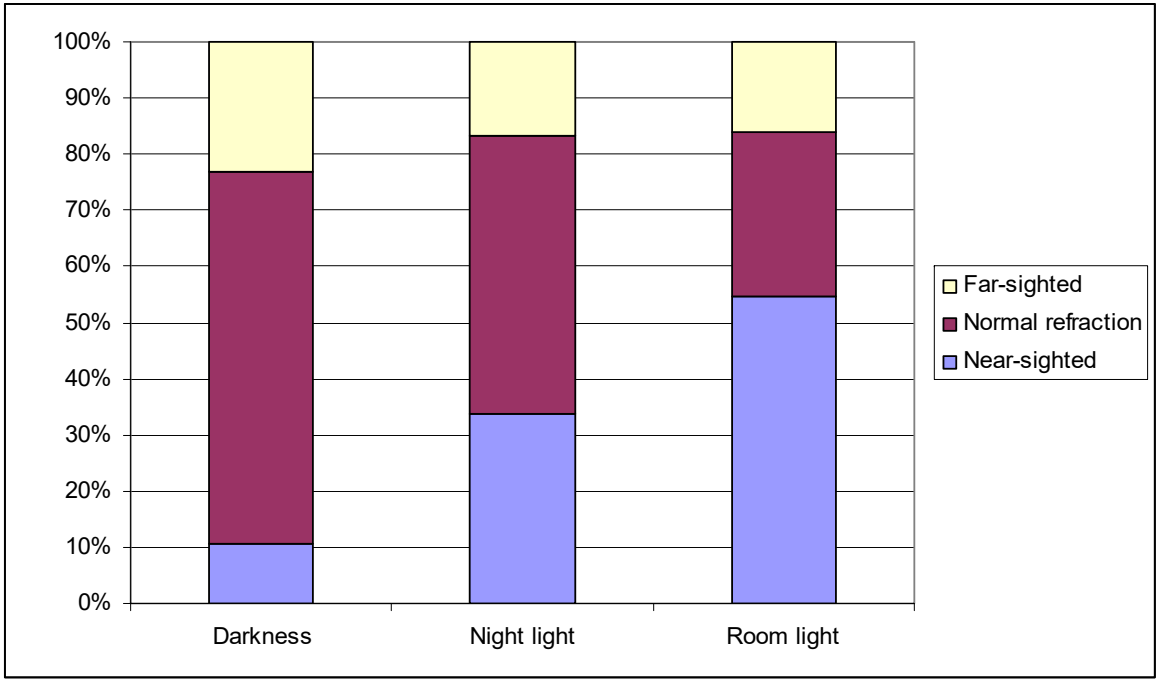
Could this have happened by chance alone?



Example 4: Near-sighted children

- Researchers interviewed 479 parents of children at pediatric ophthalmology clinic
  - Did child sleep in room with full light, night light, or no light before age two?
  - Has child been diagnosed as near-sighted, normal vision, or far-sighted?

Example 4 (cont)



- Chi-square test statistic ≈ 56.5
- P-value ≈ 0

Example 4 (cont)

- Can we draw a *cause-effect* conclusion between type of lighting used in child's room and child's eyesight condition?
- Suggest a *confounding* variable
- Can we draw *any* conclusion here?
- How could we design a study to avoid confounding variables and (potentially) support a cause-effect conclusion?
- What about *generalizability*?

Observation vs. experiment

- **Observational studies** have the potential for confounding variables
  - Cannot establish cause-effect conclusions
  - Can still provide evidence of association
- **Randomized experiments** produce similar groups except for explanatory variable being imposed
  - Support cause-effect conclusions
    - If analysis reveals significant relationship

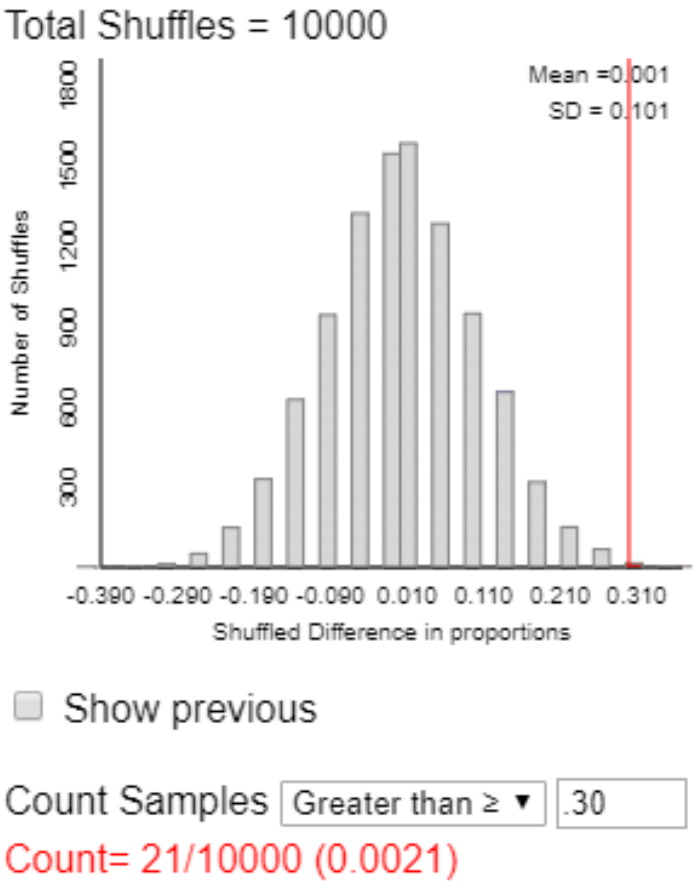
Example 5: Tagging penguins

- Are metal bands harmful to penguins?
- Researchers randomly assigned 50 penguins to receive a metal band and RFID chip, 50 other penguins to receive only RFID chip

	No metal band (control)	Metal band	Total
Survived	31	16	47
Did not survive	19	34	53
Total	50	50	100

- Is this difference in survival proportions (.30) unlikely to have happened by chance alone?

Example 5 (cont)



- Test statistic z ≈ 3.01
- 95% CI for difference in popn proportions: (.113, .487)
- Penguins not randomly sampled, but representative of larger group of penguins?

Example 5 (cont)

- Can we (essentially) rule out random chance/assignment as the explanation for the observed difference between groups?
- Can we rule out confounding variables (e.g., some penguins are just tougher/sturdier than others)?
- Can we legitimately conclude that metal bands cause higher risk of death for penguins?

Example 6: Children and laughter

- A Gallup poll asked adult Americans whether they have children under age 18 in the household and whether they smiled or laughed a lot on previous day

	Adults with children under 18 in household	Adults with no children under 18 in household
% who smiled or laughed a lot on previous day	84.1%	79.6%
Number who were surveyed	36,043	95,116

Example 6 (cont)

- Is this difference significant?
  - Yes: Test statistic $z \approx 18.5$, p-value ≈ 0
- Is this difference large?
  - No: 95% CI for difference in proportions: (.040, .050)
- Can we generalize to all U.S. adults?
  - Yes, random sample
- Can we conclude that kids in household cause more smiling or laughter?
  - No, observational study

Significance vs. importance

- Especially with large sample sizes, a difference can be *statistically significant* without being *practically important*

- Exam question:
  - Suppose someone concludes that having children in the household causes a large increase in the proportion of adults who smile or laugh. How would you respond?

Example 7: Aliens and humans

- Suppose that an alien lands on earth (bear with me please!) and wants to estimate the proportion of human beings who are female
- The alien knows from its statistics course on its home planet (you're still bearing with me, right?) that it can take a sample of human beings and determine a confidence interval
- The alien happens upon the current U.S. Senate as its sample: 21 women, 79 men

Example 7 (cont)

- 95% confidence interval for population proportion of human beings who are female: .21 ± .08
- Is this a reasonable interval estimate?
  - What went wrong?
- Is this a reasonable interval estimate for proportion of current U.S. Senators who are female?
  - Why not?

Example 7 (cont)

- Now suppose that a more statistically savvy alien wants to take a *random sample* of human beings to estimate this proportion
- Among the population of 7.6 billion people on earth, how many would the alien need to sample in order to estimate the population proportion to within ±.03 with 95% confidence?
  - Is necessary sample size closest to 100 or 1000 or 10,000 or 100,000 or 1,000,000 or 10,000,000 or 100,000,000?

Points to keep in mind (ASA Statement on p-values)

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Thanks!

- Please feel free to contact us
  - arossman@calpoly.edu
  - bchance@calpoly.edu
- For example, we'll be very happy to send these slides to you
  - And you'll make us feel good if you ask!