Quantitative Sociological Analysis

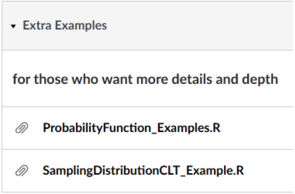
Reinforcing Understanding Surrounding Sample Data & Uncertainty A Foundation in Probability

Exercise 5

March 4 and maybe 6, 2025

Follow up after CELT feedback: thank you

- R programming: too little vs too much, and issues with Mac
  - Too little and/or Mac issues: only way to learn is to practice using RStudio
    - ask me, Google, YouTube, maybe even ChatGPT
  - Too much: switching gears soon to focus on statistical concepts
    - expectations for R programming: download and save script and data, load data, and run script to get results
- Huge range of statistics and technology backgrounds and competencies
  - Overwhelmed: talk to me. Underwhelmed: talk to me, also see



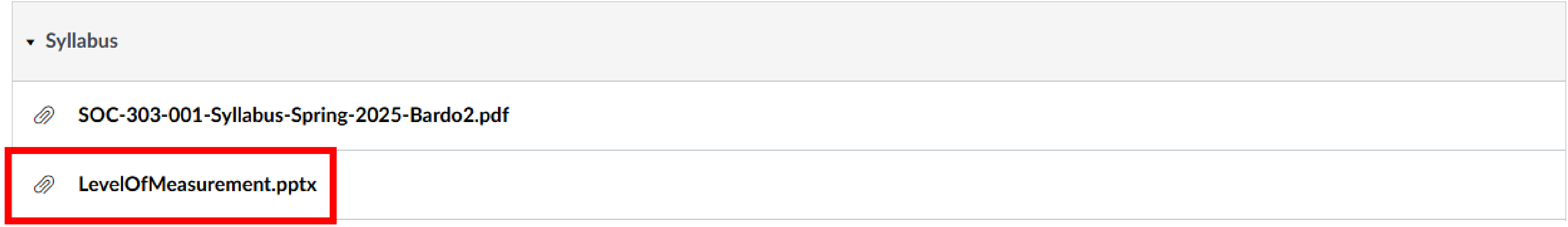
▾ Extra Examples

for those who want more details and depth

ProbabilityFunction_Examples.R

SamplingDistributionCLT_Example.R

- Level of measurement review



▾ Syllabus

SOC-303-001-Syllabus-Spring-2025-Bardo2.pdf

LevelOfMeasurement.pptx

- Finally, any questions/concerns about descriptive statistics exercise?

Exercise 5: preface

- The vibe I picked up on last Thursday was that the material (PPT_7) was
    - a lot and too abstract to fully digest

- Purpose of today's exercise is to make that material more concrete
    - while keeping math to a minimum

- First, let's review the learning objectives from PPT_7
    - Note: please review PPT_7 before our next class while this exercise is fresh in your head

Part 5

Learning objective: begin to understand that sample data have uncertainty due to chance, which must be addressed to make generalizable statements that can be applied to the broader population

recognize how:

probability theory underlies sampling

the Central Limit Theorem (CLT) connects probability and sampling

differences between a population and sample due to chance can be addressed

Takeaway: descriptive statistics help summarize sample data, but they cannot produce generalizable conclusions because they do not account for sampling variability

Pair-Share: Can anyone describe any part of this in their own words? What seems to make most or least sense?

The following

- is an abbreviated review of key terms and concepts we covered in PPT_7
  - Let’s set the stage for our in-class exercise…

Sample

- subset of a population selected for data collection
  - Sampling: process of selecting a subset of individuals or entities from a population
    - goal is to draw a sample from which generalizable conclusions can be made

- Sampling variability: difference in sample vs population due to chance

- Thus, uncertainty is inherent in sample data, which
  - needs to be accounted for to make statements that can be applied to the broader population

Inferential statistics

- are used to quantify the likelihood that a sample statistic (e.g., $\hat{p}$) approximates its true population parameter (e.g., p), because they
  - enable us to address uncertainty due to chance, sampling variability
    - Can anyone describe in own words difference between a statistic and a parameter?

- example of an inferential statistic…
- Standard error (SE) measures the variability of a sample statistic across repeated samples, which can be used to help determine
  - likelihood that sample results (e.g., p) reflect the true population parameter (e.g., $\hat{p}$)

Probability

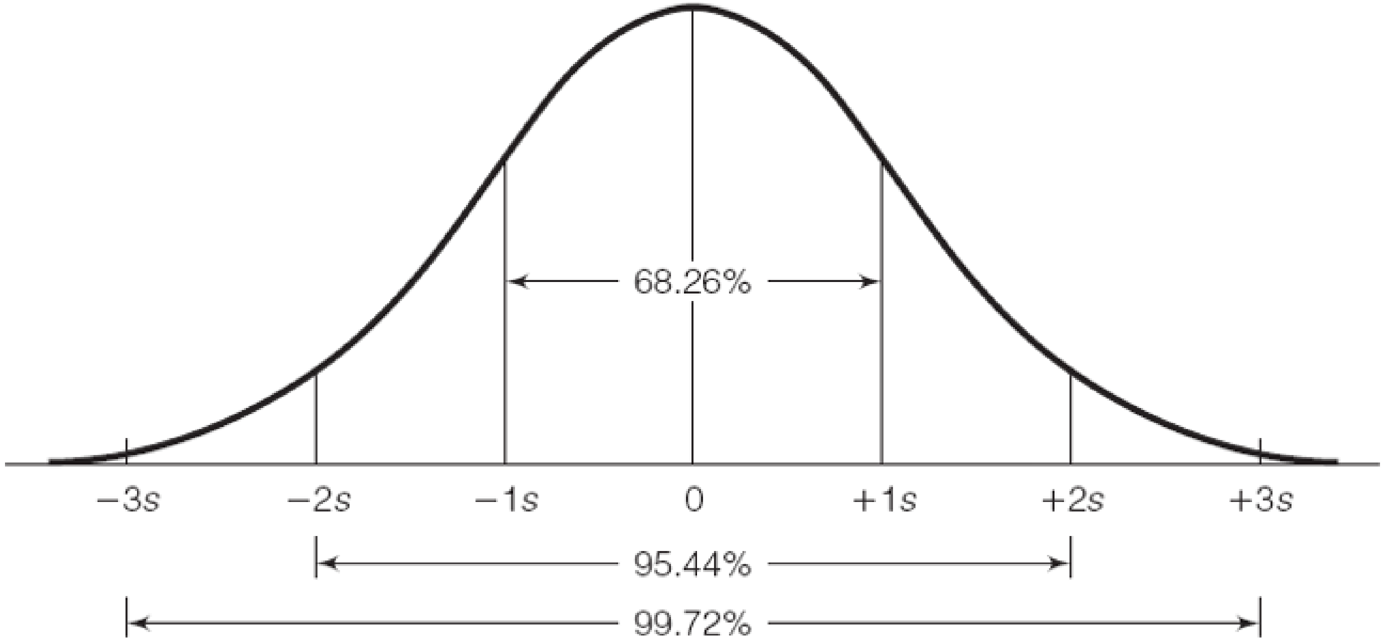
- mathematical measurement of the chance that an event will occur
  - foundational for inferential statistics

- Essential for addressing <u>sampling error</u>
  - difference between a sample statistic (e.g., $\hat{p}$) and the true population parameter (e.g., p)
    - due to chance
      - Can anyone in own words describe how this is related to but distinct from sampling variability?

- Some understanding of probability becomes especially useful for understanding the importance of the…

Central Limit Theorem, which holds that the

- sampling distribution of the sample mean ($\bar{X}$) or proportion ($\hat{p}$) approaches a normal distribution as the sample size increases, when
  - samples are randomly selected and independent
    - Can anyone explain difference between sample mean and proportion, and what this has to do with level of measurement?



Can anyone describe any part of this in their own words?

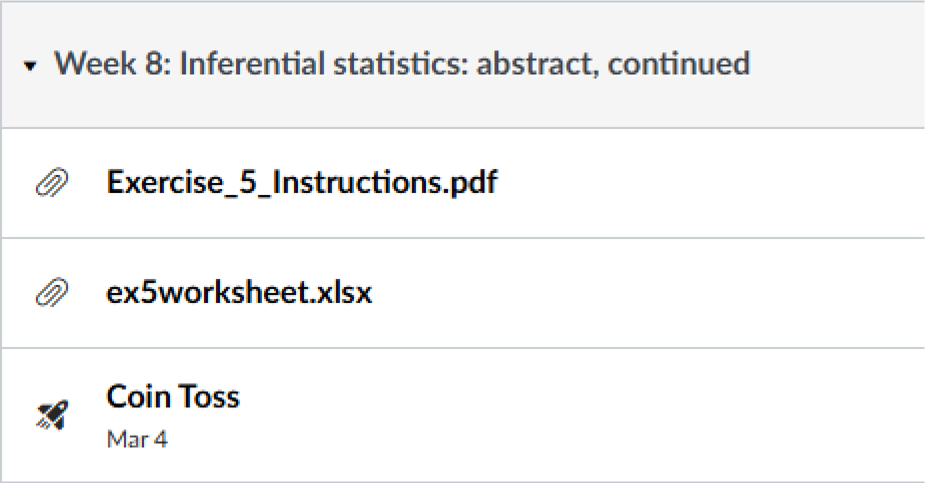
Let's consider the following exercise…

Exercise 5: coin toss trials

- The purpose of this exercise is to reinforce learning objectives surrounding sample data and uncertainty, to strengthen a foundation in probability and build up to inferential statistics.

- Let’s work in our groups that we’ve been using for the Netflix exercises. Before beginning, each member must download these two files…

- After which,
- complete exercise and then
- take practice quiz
- Finally, we’ll review the exercise results and practice quiz
  - collectively as a class and make connections with key learning objectives



▾ Week 8: Inferential statistics: abstract, continued

Exercise_5_Instructions.pdf

ex5worksheet.xlsx

Coin Toss
Mar 4

Note

- I’ll update this PPT with results from the exercise, and
  - upload an updated version in Week 8 module
    - after class today

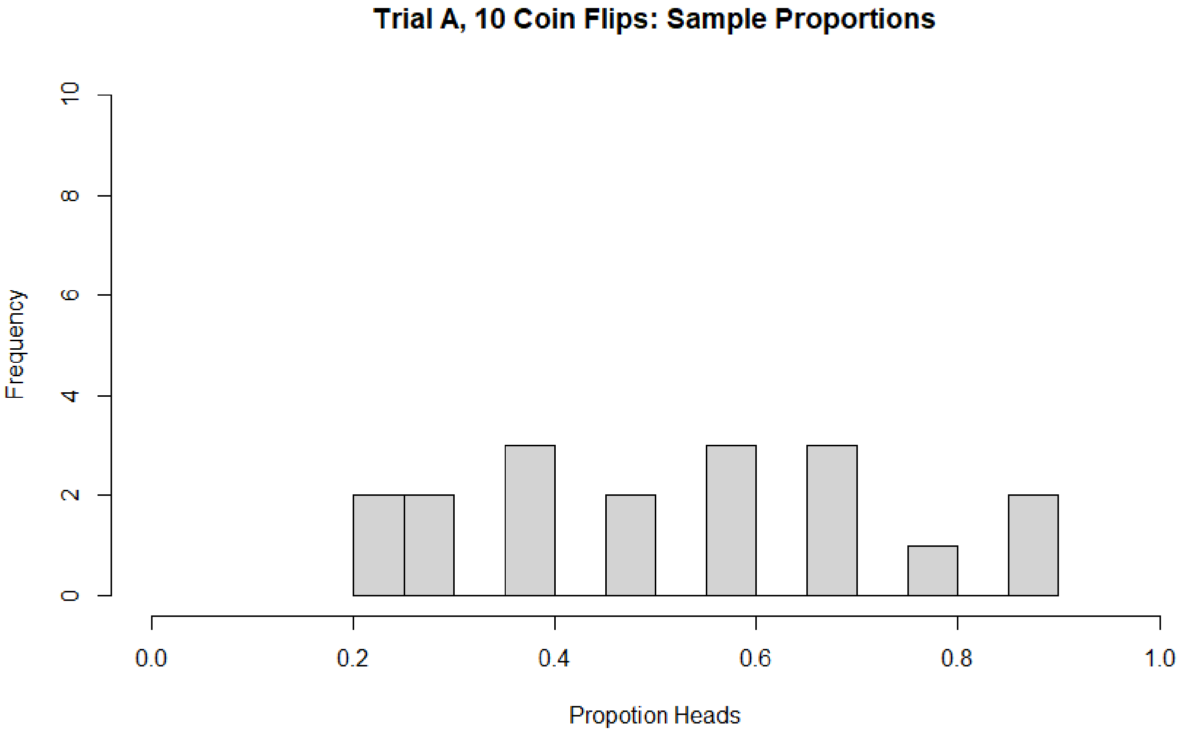
- For those who are seeking more depth and details,
  - I’ll upload the RScript I developed to summarize results from the exercise
    - this will be in the Extra Examples module

Review: exercise, preface

- If you missed, or feel uncertain about any questions on the practice quiz, keep this in mind during exercise review to see if it helps with your understanding
  - If it does not, don’t worry because we will review the quiz afterwards
- As always, please stop me and ask questions or express concerns at any point
  - This is critical so I can design fair exams and assignments that meet you where you’re at

Sample proportions

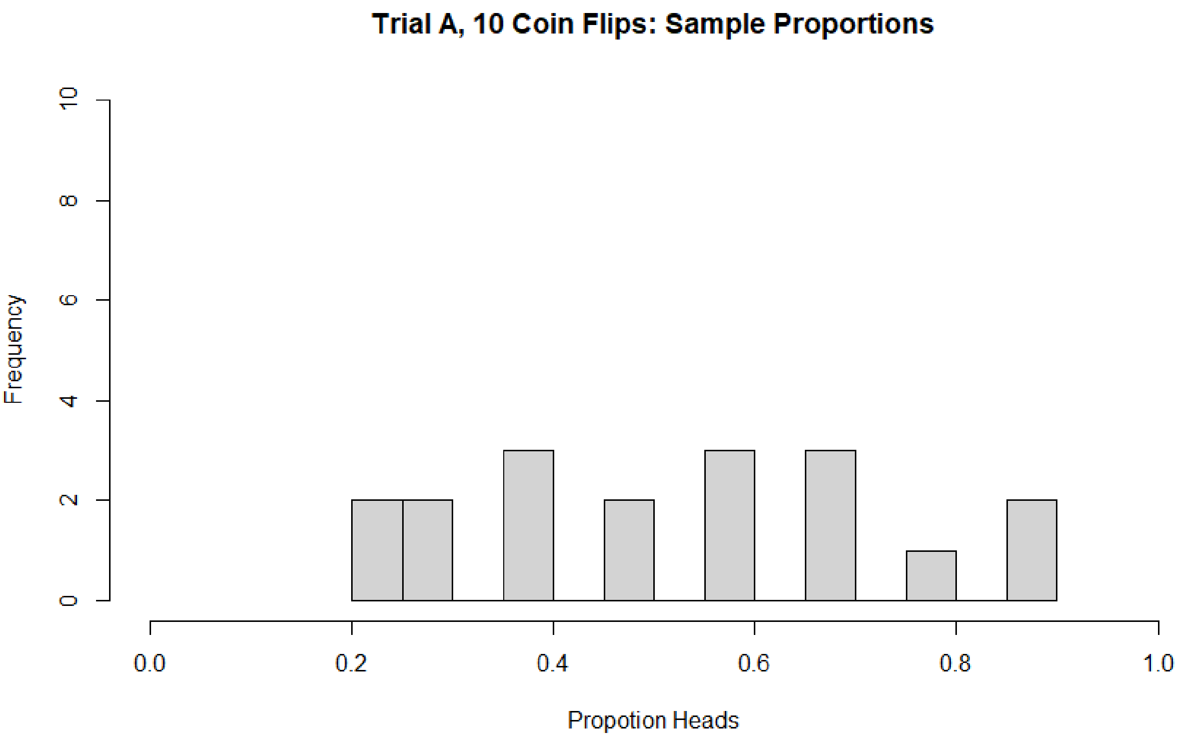
Can anyone describe in own words what this figure depicts, and some interpretations that it may help provide?

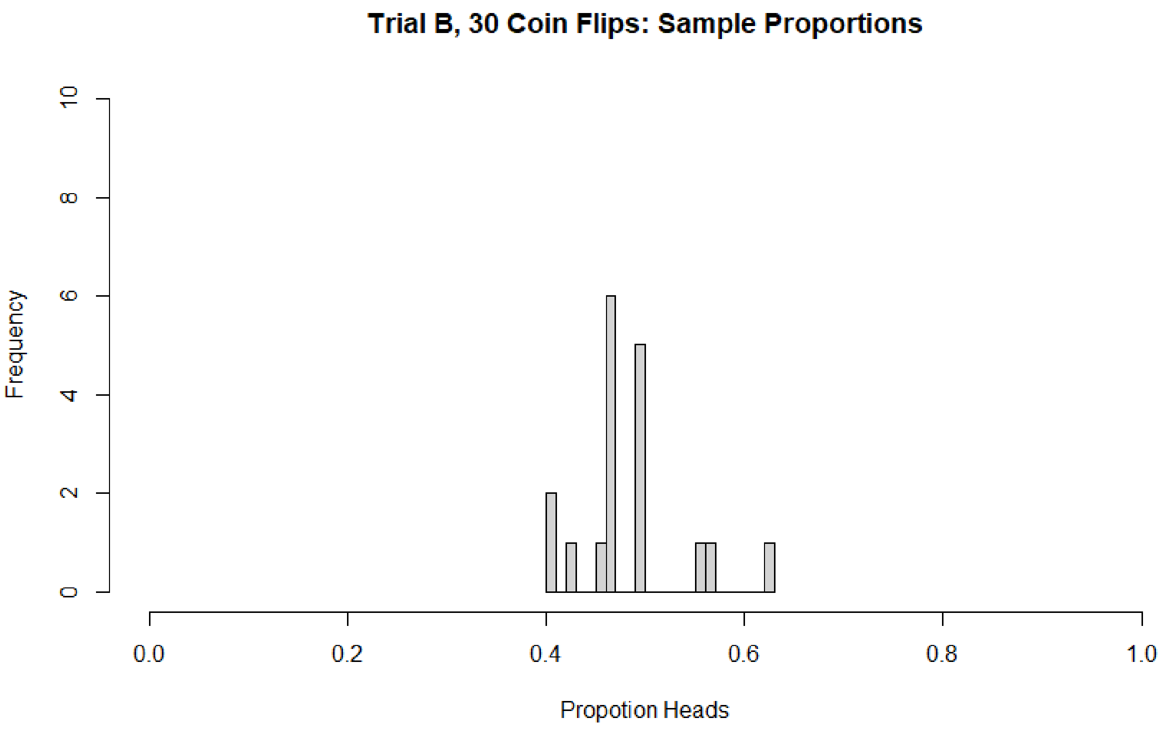


In what ways should we expect the distribution of sample proportions from Trial B to look compared to Trial A, and why?

RScript for this, and following figures, will be available on Canvas in Extra Examples module “Rscript_CoinToss_InClass”

Sample proportions

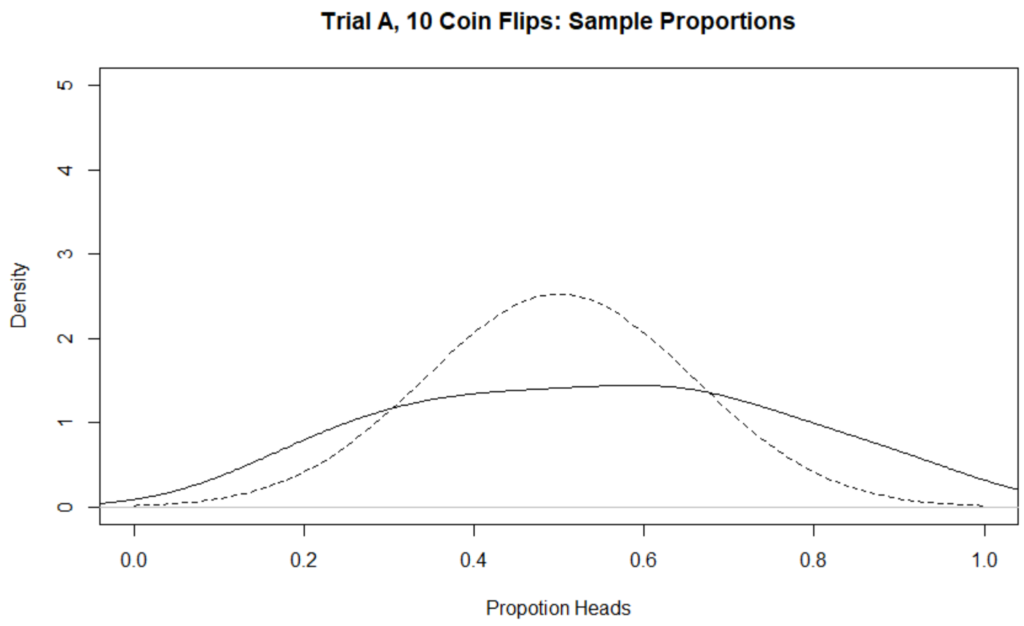


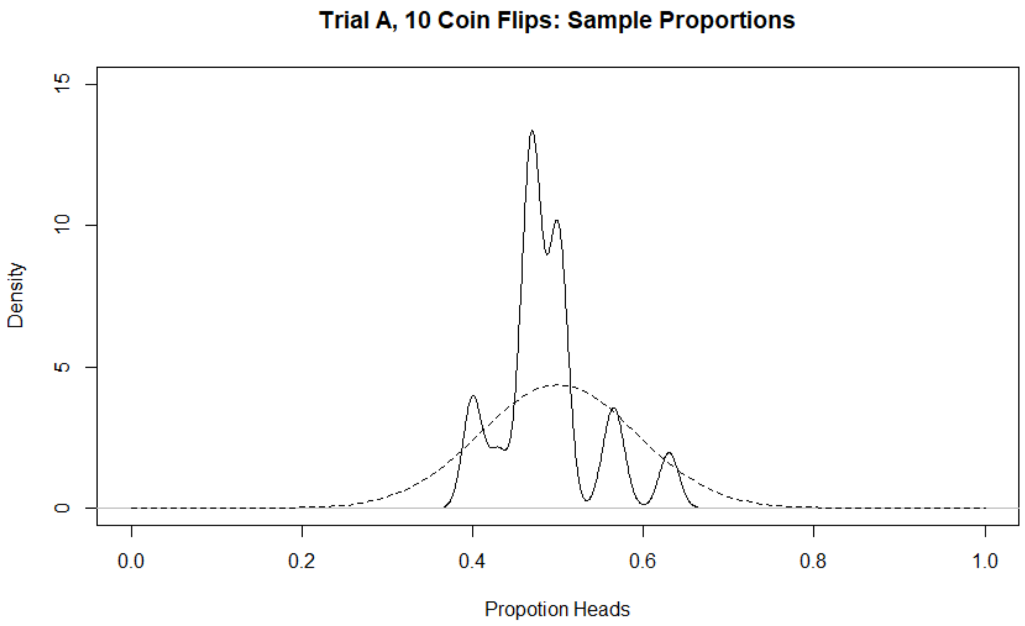


Does Trial B's sample distribution look as expected compared to A? How so, and why?

Let's take a closer look at how close these distributions approximate the theoretical standard normal distribution…

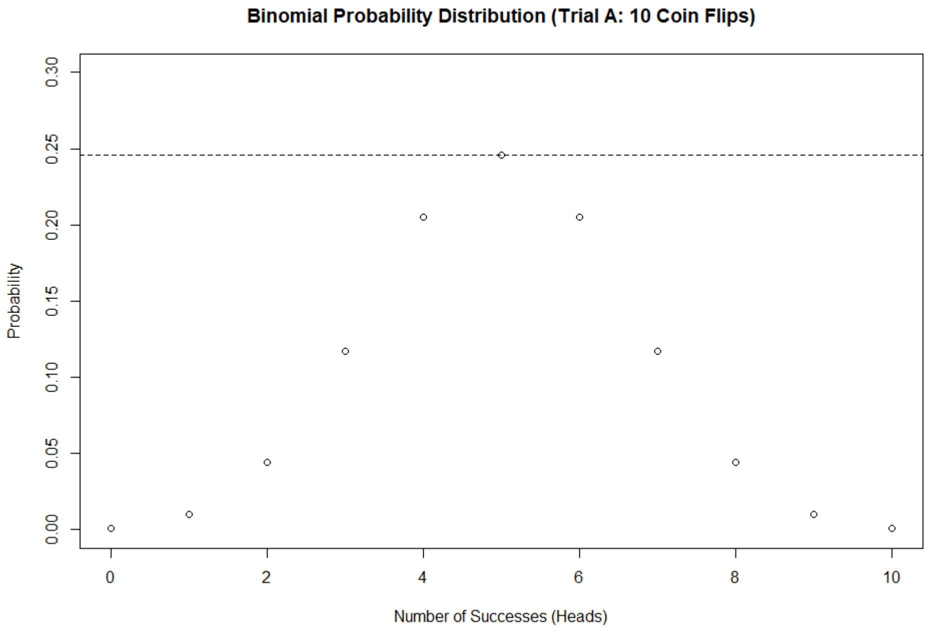
Sample proportions: approximation to normal

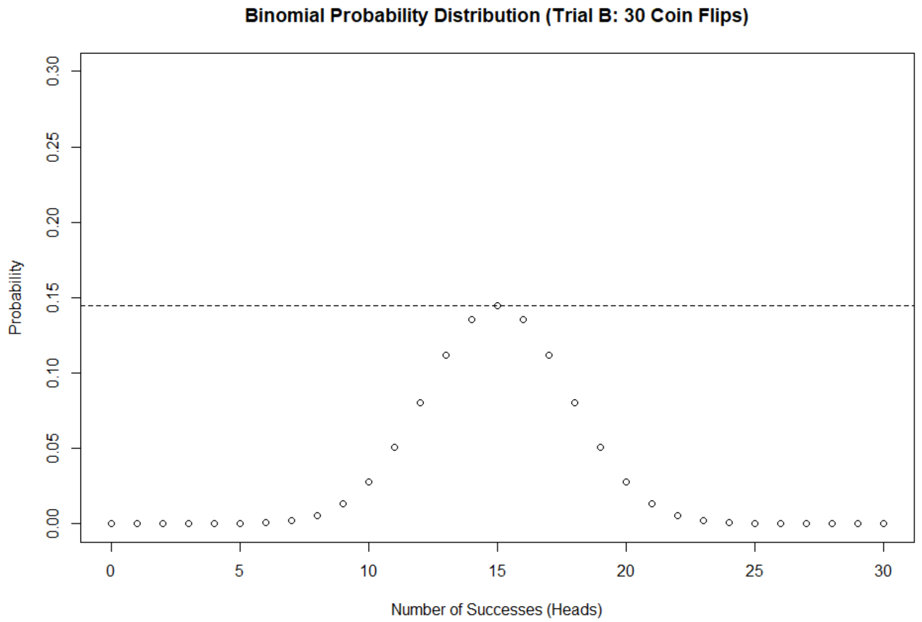




Considering the theoretical probability distributions may help us understand this a bit more…

Theoretical binomial probability distribution



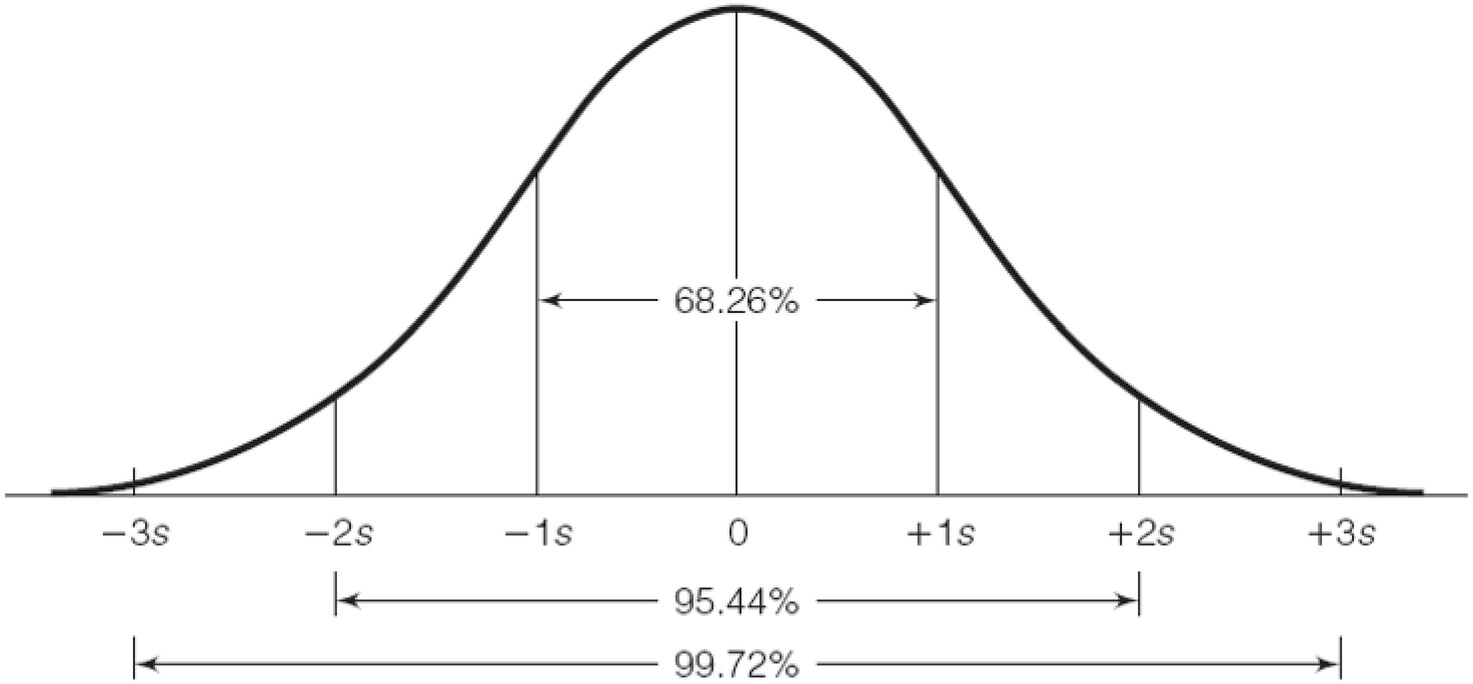


If we summed each probability for all possible successes, what would this equal and how do you know this?

The Central Limit Theorem allows us to connect sample statistics to probability…

Connecting theoretical normal distribution

- We can use this distribution’s special properties to describe a range of probabilities
  - see how standard deviation (sd) plays an important role in the normal distribution’s special properties



This involves some more math, because values will need to be standardized…

Probability connected to the normal distribution

- z-score can tell us how many standard deviations (sd) a sample proportion $(\hat{p}H_i)$ is away from the true parameter $(pH = 0.50)$
  - $z = \frac{\hat{p}H - pH}{SE}$

Also sample means $(\bar{X})$ vs population mean (μ), see PPT_7

z-score near 0 means the sample statistic is close to the population parameter

- Let's consider sample proportion i equals 0.40, then
- Trial A: $z = \frac{0.40 - 0.50}{0.158} = -0.63$
- Trial B: $z = \frac{0.40 - 0.50}{0.091} = -1.10$

The student's result of 4/10 heads is 0.63 sd below the expected proportion

The student's result of 12/30 heads is 1.10 sd below the expected proportion

Notice how the SE gets smaller, and the estimates get more precise, as the sample size increases

That's because the SE is a function of, depends on, the sample size (n): $SE = \sqrt{\frac{p(1-p)}{n}}$

Let's see if the normal distribution figure on the last slide makes any more sense now…

Review: practice quiz, preface

- If you missed, or feel uncertain about any questions on the practice quiz, did reviewing our coin toss results help in any certain way?
  - Conversely, did it illicit more confusion or raise any specific questions?
    - Pair-share…

- Let’s revisit our learning objectives before moving on to the practice quiz

- Please stop me and ask questions or expresss concerns at any point
  - This is critical so I can design fair exams and assignments that meet you where you’re at
    - Also, this helps inform the pace at which I should move forward with new material

Part 5

Learning objective: begin to understand that sample data have uncertainty due to chance, which must be addressed to make generalizable statements that can be applied to the broader population

recognize how:

probability theory underlies sampling

the Central Limit Theorem (CLT) connects probability and sampling

differences between a population and sample due to chance can be addressed

Takeaway: descriptive statistics help summarize sample data, but they cannot produce generalizable conclusions because they do not account for sampling variability

Can anyone describe any part of this in their own words? What seems to make most or least sense?