# PSY 503: Foundations of Statistical Methods in Psychological Science

## Bootstrapping

Suyog Chandramouli

311 PSH (Princeton University)

12th November, 2025

# The statistical inference problem

- We want to estimate population parameters.
- But we only have data from a sample
- This introduces uncertainty
- And we try to estimate uncertainty

# Think like a frequentist...

- Take many samples from the population
- Calculate your statistic each time
- See how much it varies - that's your sampling distribution
- But wait - we only have ONE sample...

- **Theory to the rescue**
  - theory gives us formulas for standard errors
  - SE = SD/$\sqrt{n}$
  - We can build confidence intervals from this

# Think like a frequentist (Confidence Intervals)

- If we repeated our study many times…
  - For each experiment we calculate
    - Mean (estimated mean of population)
    - Confidence interval around the mean estimate
      - *This interval may or may not contain the true mean*
  - Across experiments, 95% of these intervals contain the true mean
- CI for means: : $\bar{X} \pm t^* \times (SD/\sqrt{n})$
- Similarly for,
  - Difference in means, regression coefficients, etc.
  - Works by making sampling distributions normal
  - Sampling variation has a direct connection to uncertainty in statistic

# But for some statistics it’s not as straightforward

- Median
  - Requires sorting data
  - Variation depends on density around median
    - We only have one sample
    - Very different behaviors for different population distributions
      - (Uniform vs. Bimodal vs. Normal)
- Correlation
  - Requires multiplying deviations from mean (products, not sums)
  - Variation depends on true correlation
    - Very different behavior for different population distributions
    - (Height & IQ) vs (Test vs Re-test)
- Ratio..

# But for some statistics it's not as straightforward

- Median
  - Requires sorting data
  - Variation depends on density around median
    - We only have one sample
    - Very different behaviors for different population distributions
      - (Uniform vs. Bimodal vs. Normal)
- Correlation
  - Requires multiplying deviations from mean (products, not sums)
  - Variation depends on true correlation
    - Very different behavior for different population distributions
    - (Height & IQ) vs (Test vs Re-test)
- Ratio..

**NOTE: CLT still holds**

- **But we don't know the spread of this Gaussian**
- **Formulas exist (for for very large N)**

# The Bootstrap idea

- Forget formulas
- Your sample has some information, and is your best guess
  - Shape
  - Variability
  - Relationship
- Your sample is like a mini-population
  - Take samples of this sample
  - Construction a ‘sampling distribution’ of the ***statistic***
  - Calculate the variation of this distribution
    - SD
    - Percentiles