CSE P 590

Building Data Analysis Pipelines

Fall 2024

Significance and confidence

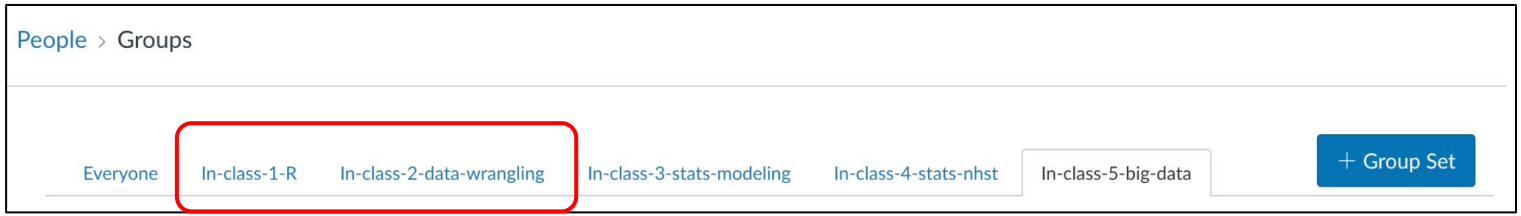
Today

- Housekeeping: Group work and grading
- Recap: Terminology
- Live demo: Statistical significance
- In-class exercise 4

Housekeeping

Canvas groups

How we see them



How you see them

Groups and group sets on Canvas

Self-assign to a group by Friday EOD

Grading

- Holistic grading – reasoning and justifications
- Fine-grained grading breakdown (now) on each assignment
  - Completion
  - Questions
  - Optional questions
- Reach out with questions/concerns

Recap: Terminology

Population vs. Sample vs. Sampling distribution

Population

- All possible individuals
- Parameters
  - **μ: mean, σ: standard deviation**

Sample

- A subset of the population
- Sample statistics
  - **$\bar{x}$: mean, s: standard deviation**

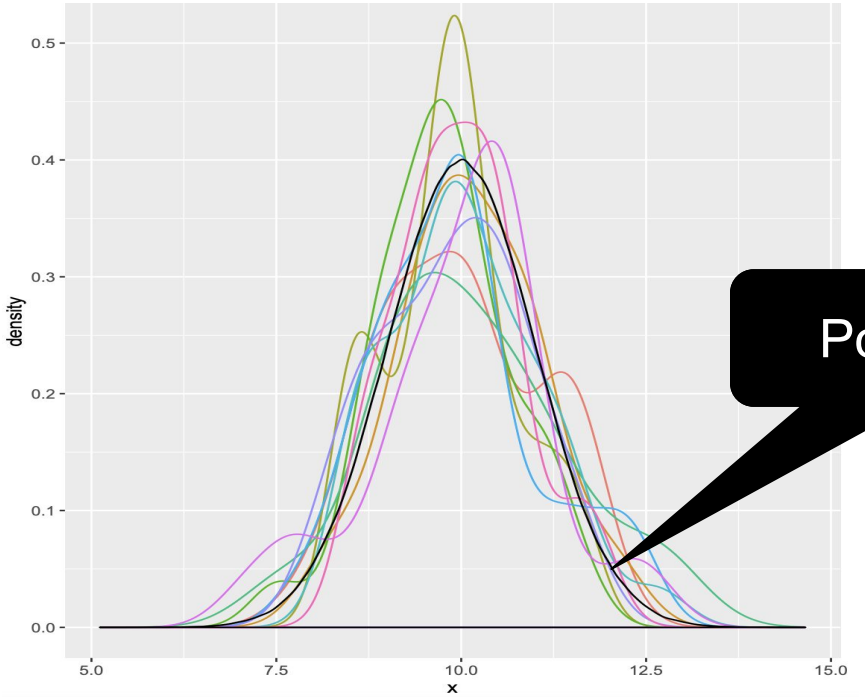
Sampling distribution

- Distribution of the sample statistic (e.g., mean)

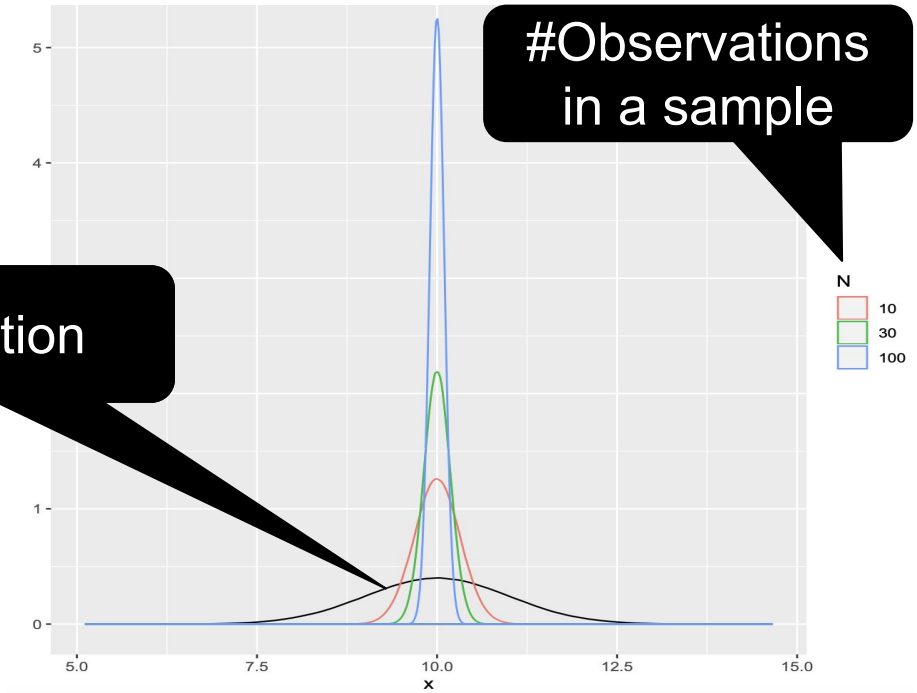
Population vs. Sample vs. Sampling distribution

Population and Samples

Sampling distribution



Population



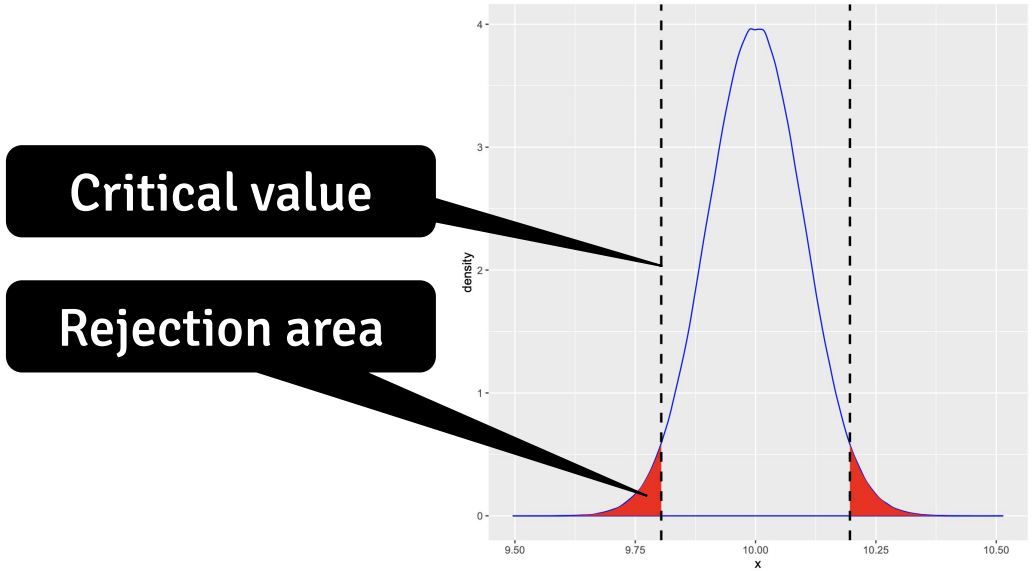
Given a sample mean, what is its p value?

P-value, critical value, and rejection area

Zooming in on the Sampling Distribution

Sampling distribution for ***N*(10, 1)** and **sample size 100**

- μ = 10, σ = 0.1 -> critical values (**α** = 0.05, two-tailed) = 9.804 and 10.196

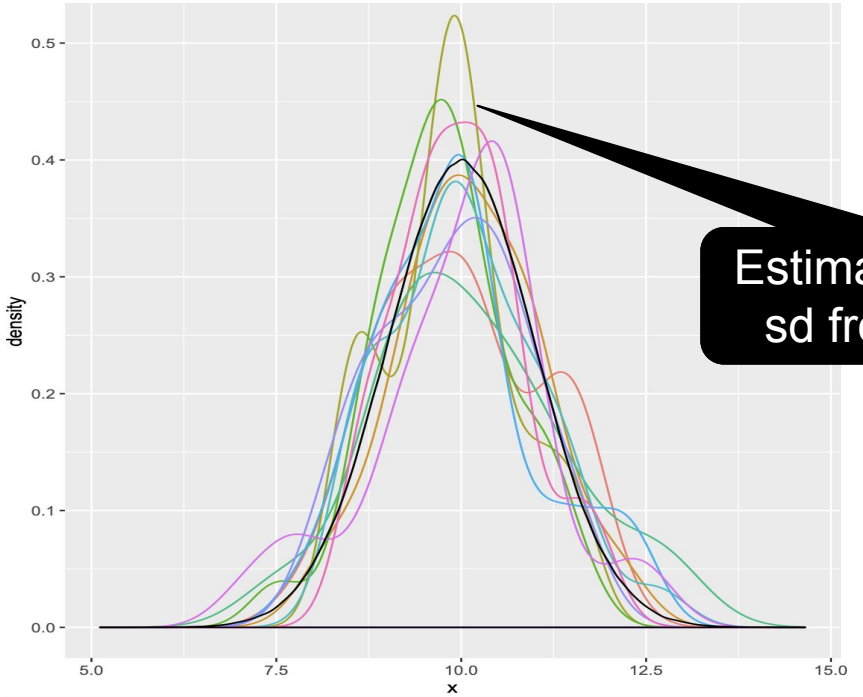


Great, but how do we know the sampling distribution?

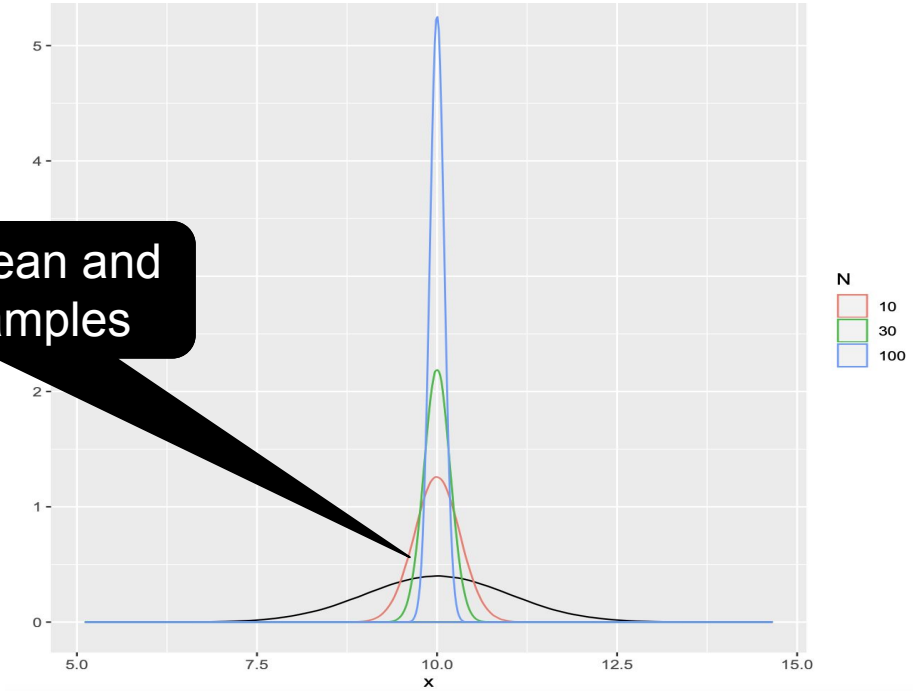
Estimating parameters from samples

Population and Samples

Sampling distribution

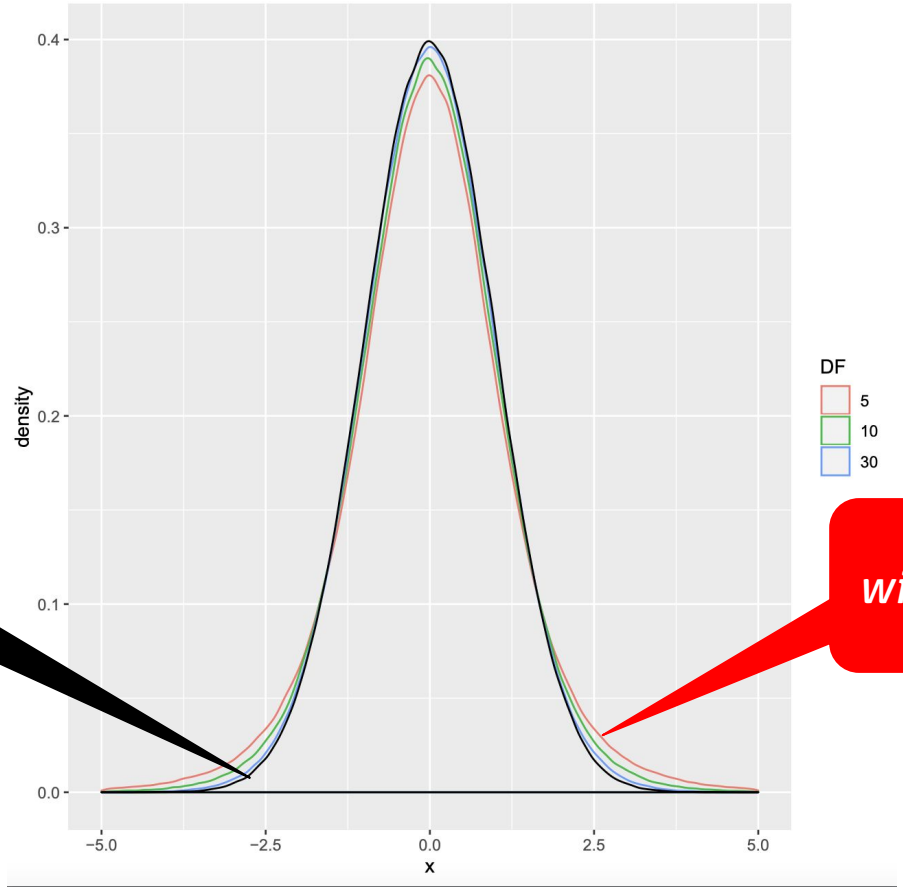


Estimate mean and sd from samples



How do we account for uncertainty in those estimates?

z distribution vs. t distribution (small samples)



z distribution:
standard normal
$N(0,1)$

t distribution
with 5 DF (degrees of freedom)

NHST: live demo

In-class exercise