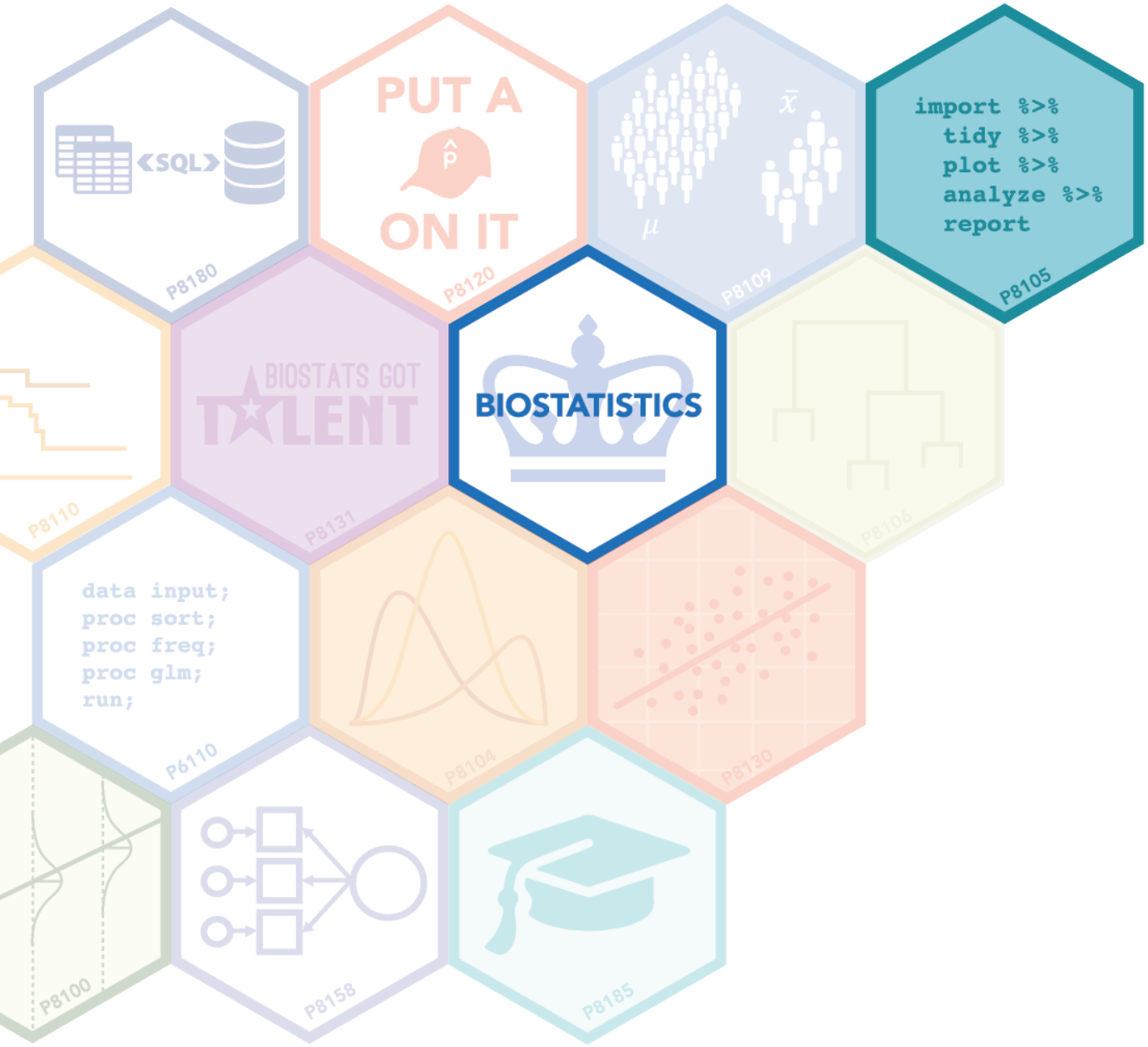

SIMULATIONS AND BOOTSTRAPPING

Jeff Goldsmith, PhD

Department of Biostatistics

Repeated sampling

- "Repeated sampling" is a conceptual framework that underlies almost all of statistics
  - Repeatedly draw random samples of the same size from a population
  - For each sample, compute the mean
  - The distribution of the sample mean converges to a Normal distribution

- Repeated sampling doesn't happen in reality
  - Data are difficult and expensive to collect
  - You get your data, and that's pretty much it

- Repeated sampling can happen on a computer

Simulation

- Hard to overstate how important and useful simulations are in statistics

- Basic idea is to generate repeated samples under a process you design
  - Define a data generating mechanism (e.g. a Normal distribution)
  - Draw a random sample from that data generating mechanism
  - Analyze the sample (e.g. compute the sample mean)
  - Repeat
  - **Understand the analysis approach under repeated sampling**

Simulation

- Hard to overstate how important and useful simulations are in statistics

- Basic idea is to generate repeated samples under a process you design
  - Define a data generating mechanism (e.g. a Normal distribution)
  - Draw a random sample from that data generating mechanism
  - Analyze the sample (e.g. compute the sample mean)
  - Repeat
  - **Understand the analysis approach under repeated sampling**

- Might vary the underlying process to inspect changes
  - Different sample size
  - Different covariate effect

Coding a simulation

- Simulations are natural in the context of iteration

- Write a function (or functions) to:
  - Define data generating mechanism
  - Draw a sample
  - Analyze the sample
  - Return object of interest
- Use a loop / loop function to repeat many times

- Inspect the properties of your analysis …

Coding a simulation

- Simulations are natural in the context of iteration

- Write a function (or functions) to:
  - Define data generating mechanism
  - Draw a sample
  - Analyze the sample
  - Return object of interest
- Use a loop / loop function to repeat many times

- Inspect the properties of your analysis …

…under repeated sampling!!!

Coding a simulation

- Simulations are natural in the context of iteration

- Write a function (or functions) to:
  - Define data generating mechanism
  - Draw a sample
  - Analyze the sample
  - Return object of interest
- Use a loop / loop function to repeat many times

- Inspect the properties of your analysis …

…under repeated sampling!!!

Bootstrapping

- Hard to overstate how important and useful bootstrapping is in statistics

- Basic idea is to mimic repeated sampling with the one sample you have
  - That sample is draw at random from your population
  - You'd like to draw more samples, but you can't
  - So you draw a bootstrap sample from the one sample you have
  - The bootstrap sample has the same size as the original sample, and is drawn with replacement
  - Repeat

Why bootstrap?

- The repeated sampling framework often provides useful theoretical results under certain assumptions or asymptotics
  - Sample means follow a known distribution
  - Regression coefficients follow a known distribution
  - Odds ratios follow a known distribution

- If your assumptions aren't met, or your sample isn't large enough for asymptotics, you can't use the "known distribution"

- Bootstrapping gets you back to repeated sampling, and uses an empirical rather than a theoretical distribution for your statistic of interest

Coding the bootstrap

- Bootstrapping is natural in the context of iteration

- Write a function (or functions) to:
  - Draw a sample with replacement
  - Analyze the sample
  - Return object of interest
- Repeat this process many times

- Keeping track of the bootstrap samples, analyses, and results in a single data frame organizes the process and prevents mistakes

Coding the bootstrap

- Bootstrapping is natural in the context of iteration

- Write a function (or functions) to:
  - Draw a sample with replacement
  - Analyze the sample
  - Return object of interest
- Repeat this process many times

- Keeping track of the bootstrap samples, analyses, and results in a single data frame organizes the process and prevents mistakes

- That's why you use **LIST COLUMNS!!**

Coding the bootstrap

- Bootstrapping is natural in the context of iteration

- Write a function (or functions) to:
  - Draw a sample with replacement
  - Analyze the sample
  - Return object of interest
- Repeat this process many times

- Keeping track of the bootstrap samples, analyses, and results in a single data frame organizes the process and prevents mistakes

- That's why you use **LIST COLUMNS!!**