Bootstrapping

Alternative to Normal and t-distributions

Grinnell College

Review

We have seen how to make Confidence Intervals to estimate

- population mean (μ)
- difference in pop. means ($\mu_1 - \mu_2$)
- population proportion (p)
- difference in pop. proportions ($p_1 - p_2$)

Methods

- Normal distribution (p, $p_1 - p_2$)
- t-distribution (μ, $\mu_1 - \mu_2$)

Review

Sampling distribution

- Distribution of statistics from many samples
- Shows us variability in the statistics

CIs so far have been based on **CLT**

- For a large enough sample size, the sampling distribution for a sample mean (or proportion) looks like a Normal distribution

Outline

Motivation for today: What do we do when we want to estimate things other than means and proportions?

- CI for median, IQR, standard deviation?
- We can't use Normal/t because CLT doesn't work for these

We will see an alternative approach to estimate these with a CI.

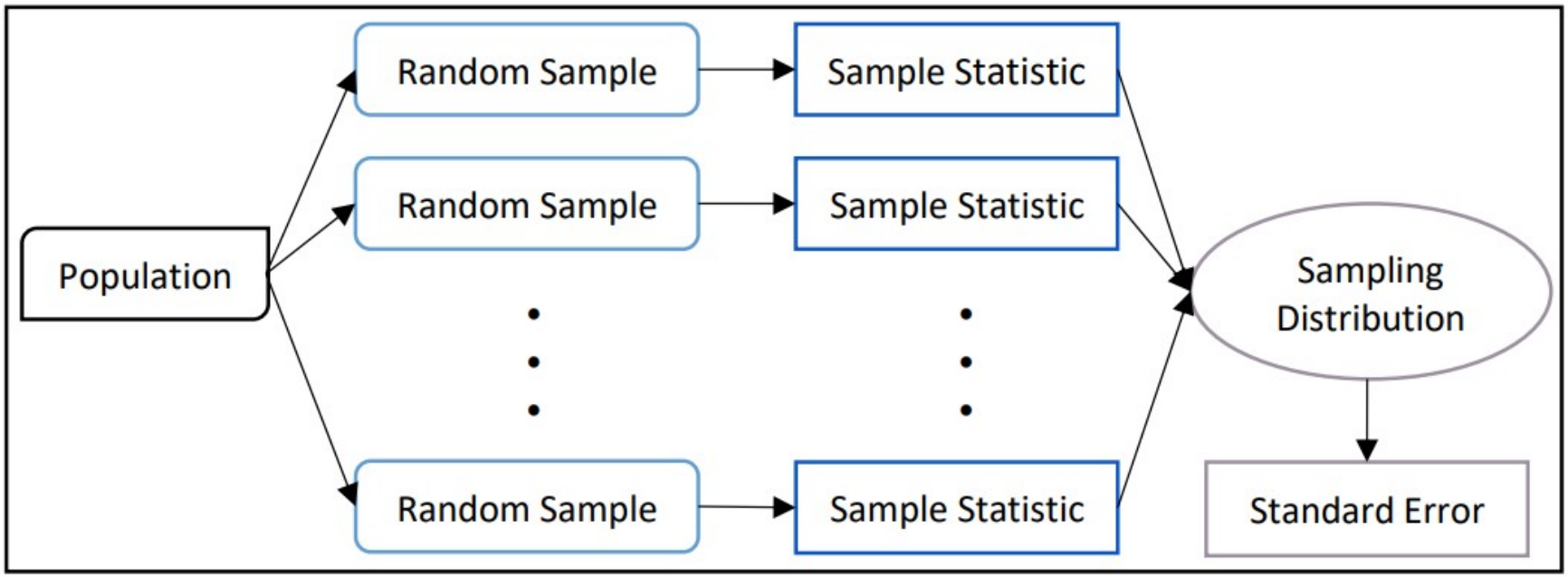
Repeated Samples

Confidence intervals we constructed had the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- Relied on assumptions about populations and CLT
- Examined what might happen if we could repeat sampling ad infinitum

Review – Sampling Distribution



How to construct?

1. Start with population.
2. Take a sample and compute the statistic of choice
3. Take another sample and compute the statistics
4. Continue to take more samples and compute statistic each time
5. Plot all of the statistics in a histogram or dotplot

Issues?

There are, naturally, some limitations:

- We are limited to collecting a single sample
- So... Can't make a sampling distribution in reality

Bootstrapping

Our solution is something called **"Bootstrapping"**

Bootstrapping:
Instead of taking a lot of samples from the population over and over...

- Bootstrapping simulates this process
- Create many "new samples" using the original sample we collect

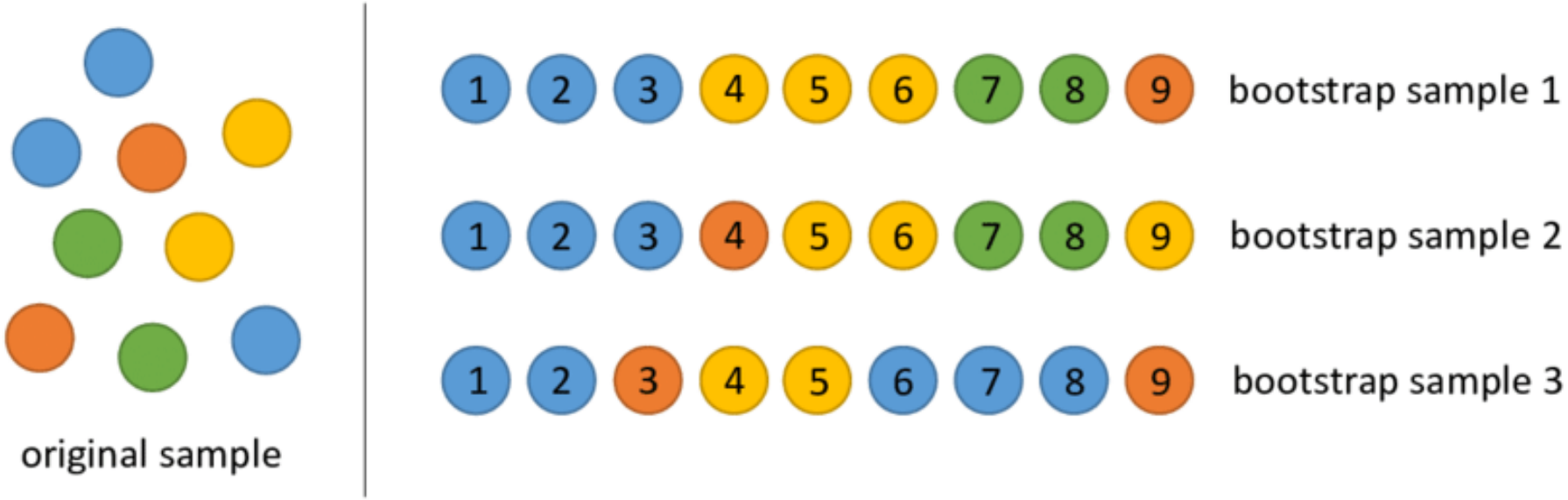
Logic:

1. If sample is randomly selected $\rightarrow$ representative
2. Make many copies of the sample $\rightarrow$ approximation of the population
3. Take samples from "new population" $\rightarrow$ approximates sampling dist.
4. Now we can make CI's

Bootstrapping

Method:

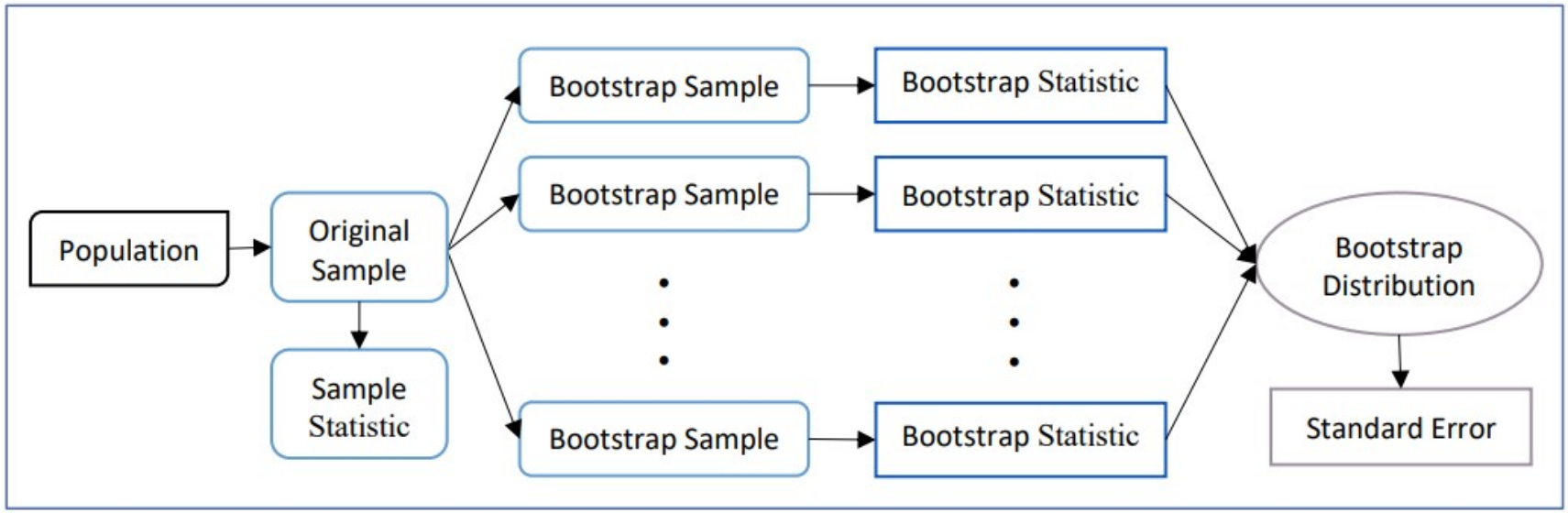
1. Random sample is representative of population
2. Use the sample as a proxy for the population
3. Draw new samples (with replacement) from the original sample
4. Sample size of new samples must match the original



Bootstrapping

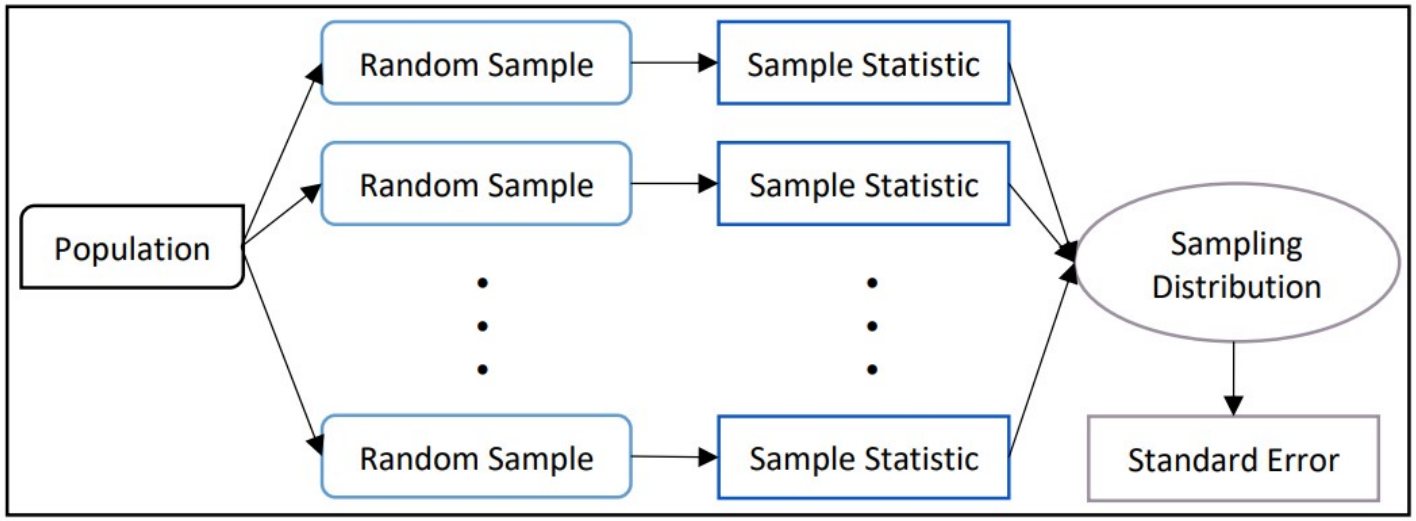
Method:

1. Random sample is representative of population
2. Use the sample as a proxy for the population
3. Draw new samples (with replacement) from the original sample
4. Sample size of new samples must match the original

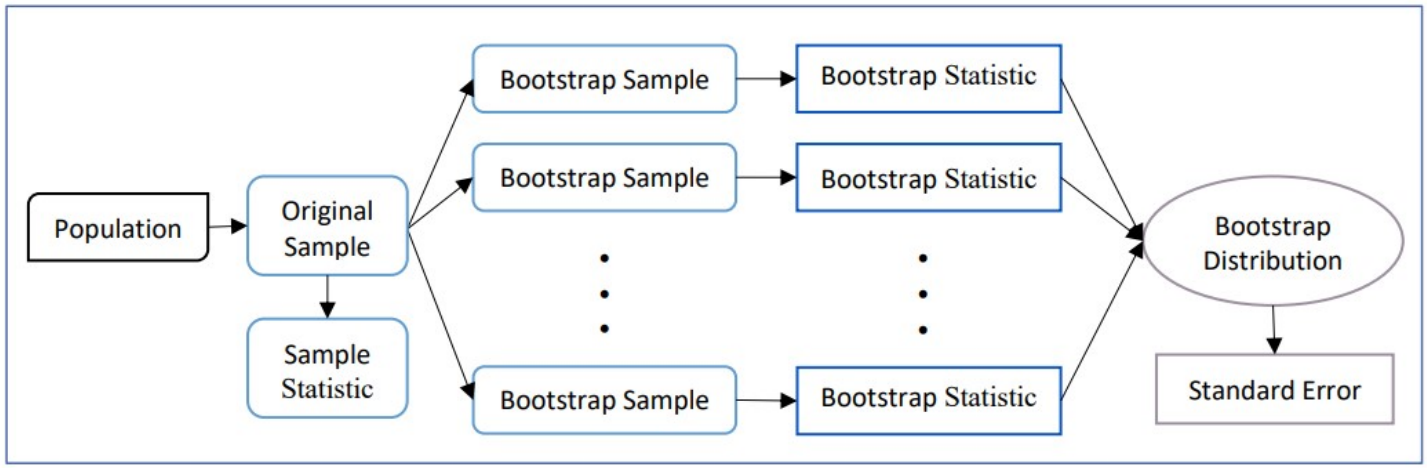


Comparison

Sampling Distribution: new samples from population

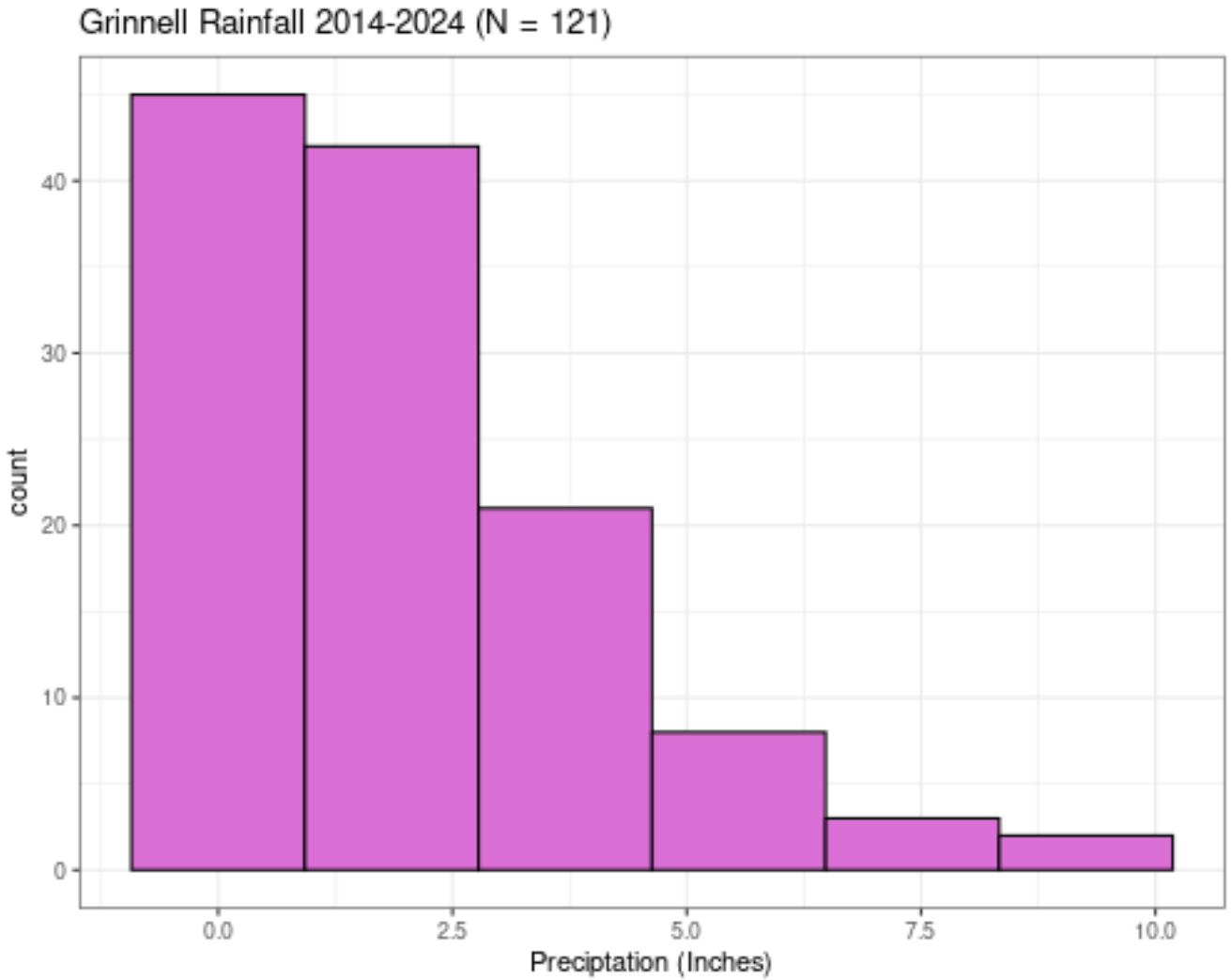


Bootstrap Distribution: new samples from the original sample



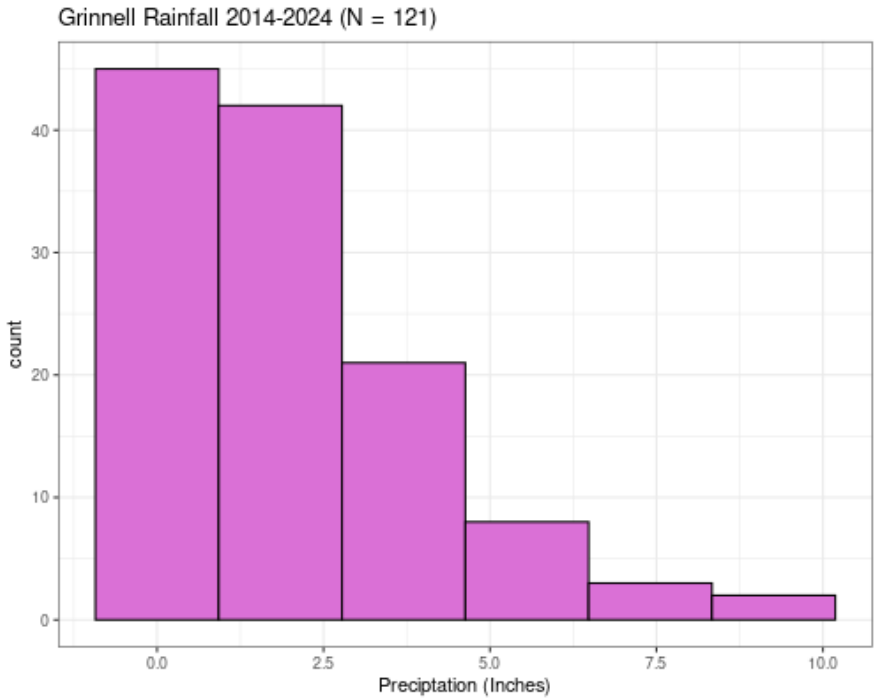
Rainfall Example

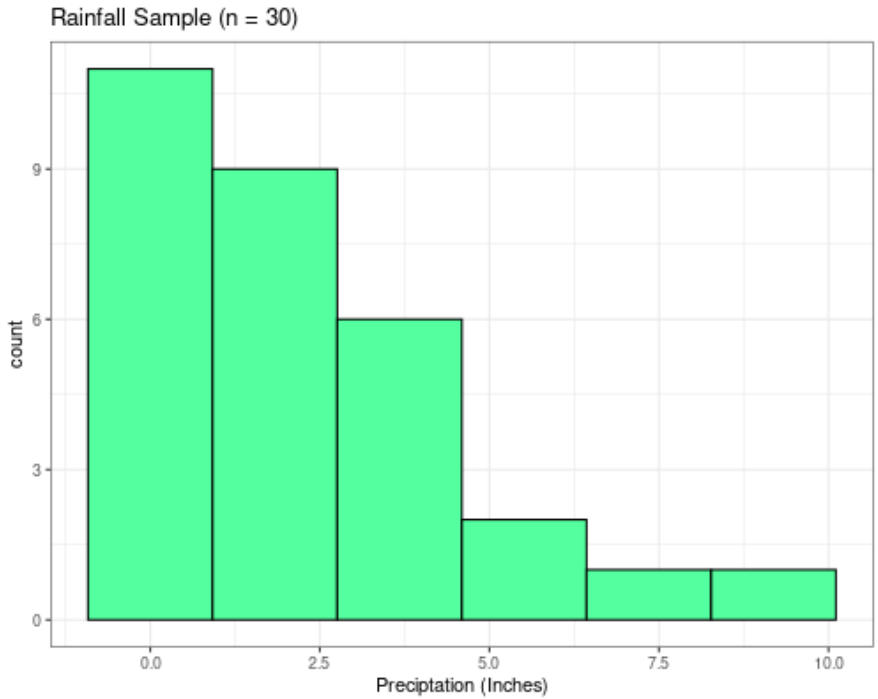
We have data collected on the amount of precipitation on 121 rainy days in Grinnell from 2014-2024 (courtesy of Professor Nolte)



Rainfall Example

Let's say we took a random sample of 30 rainy days...



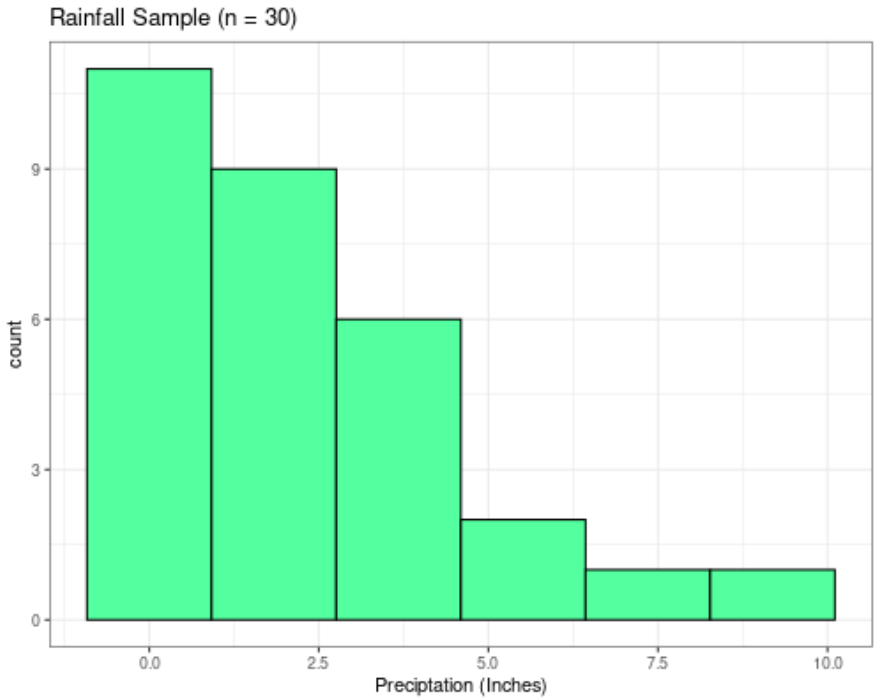


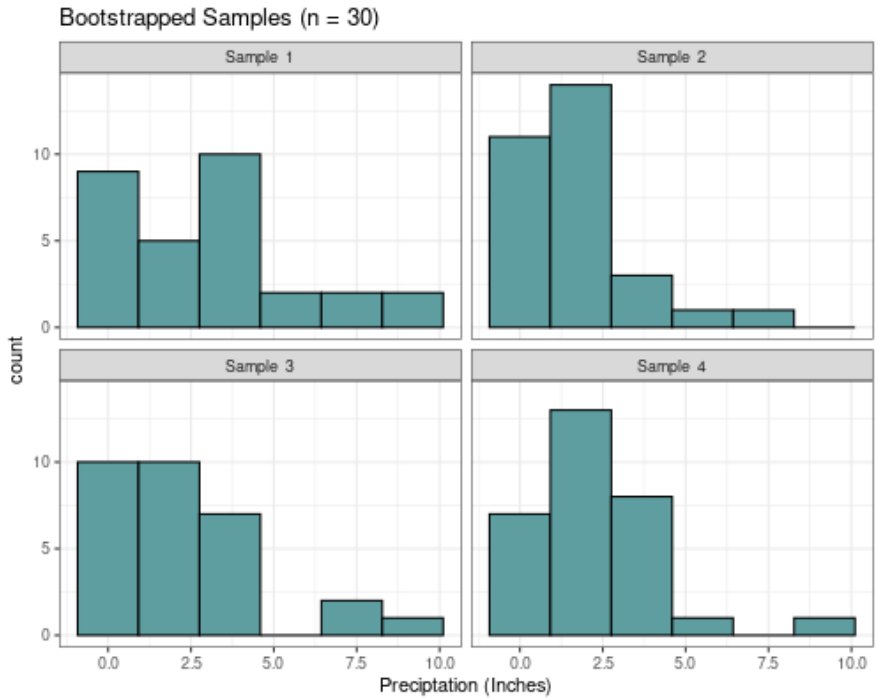
- Random sample $\rightarrow$ representative
- Same shape, very similar center and spread

Rainfall Example

Start the bootstrap process.

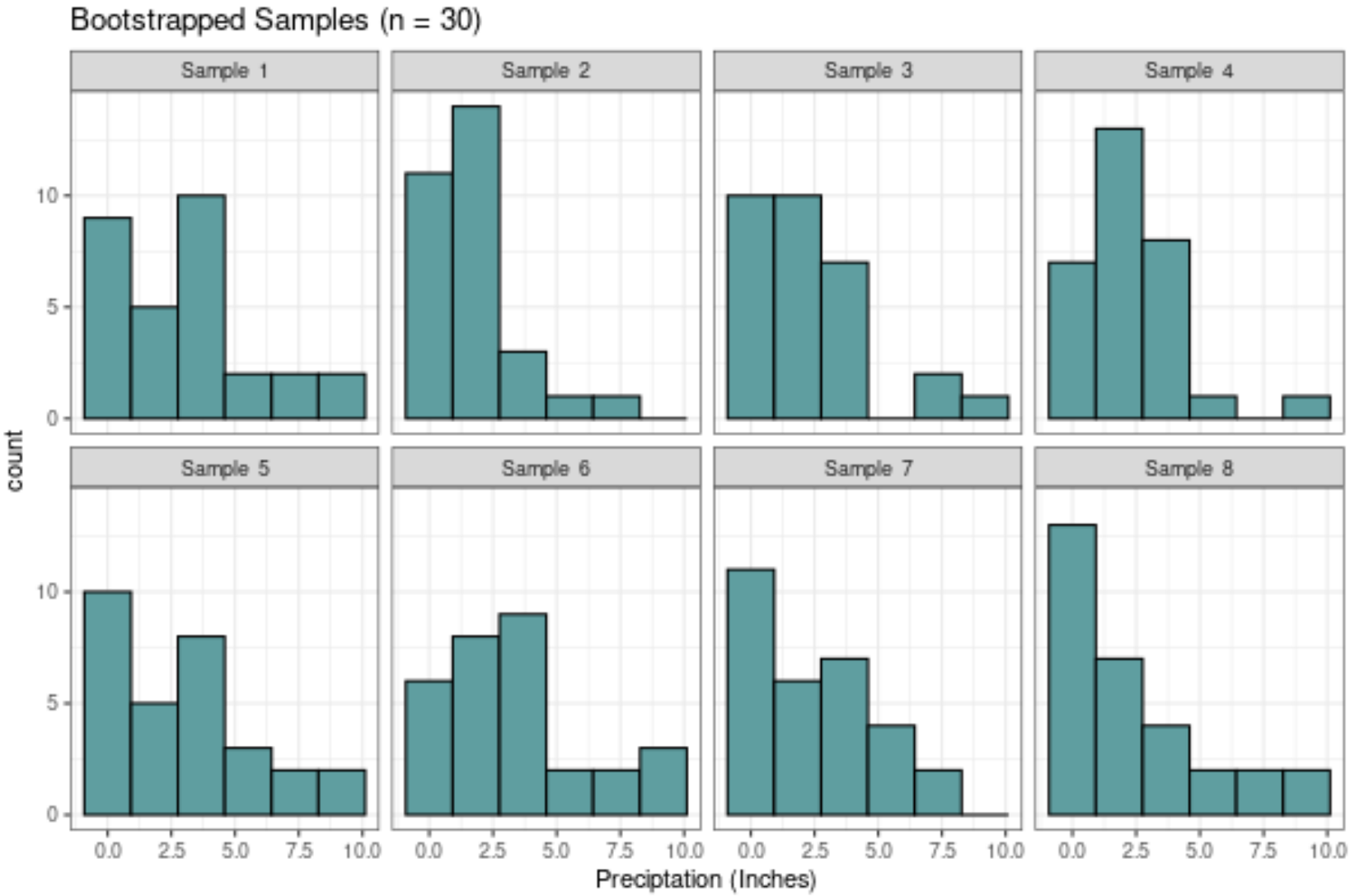
- Make some bootstrap samples of size 30
- Do they *kind of* look the same? Yeah!





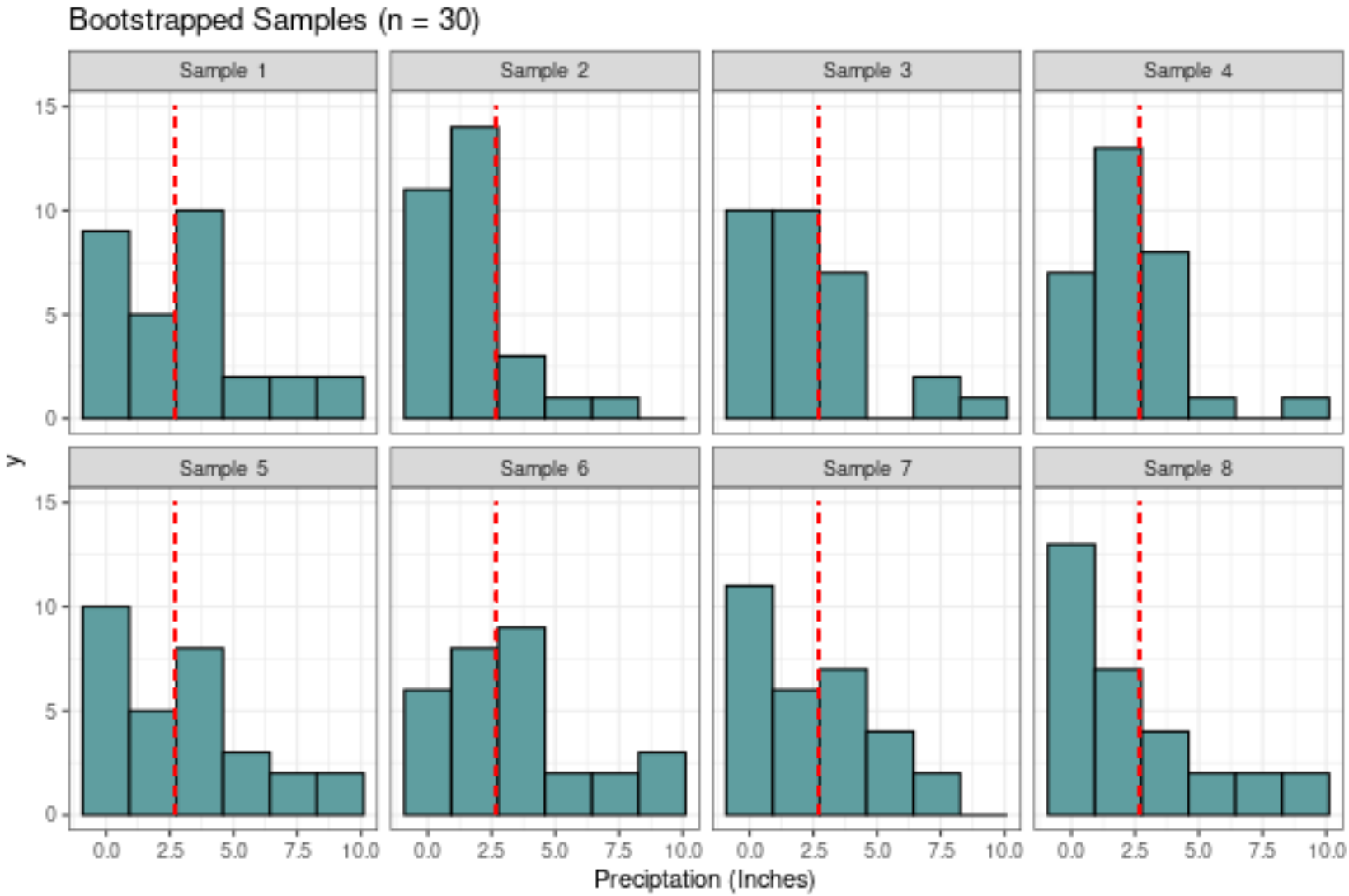
Rainfall Example

More bootstrap samples...



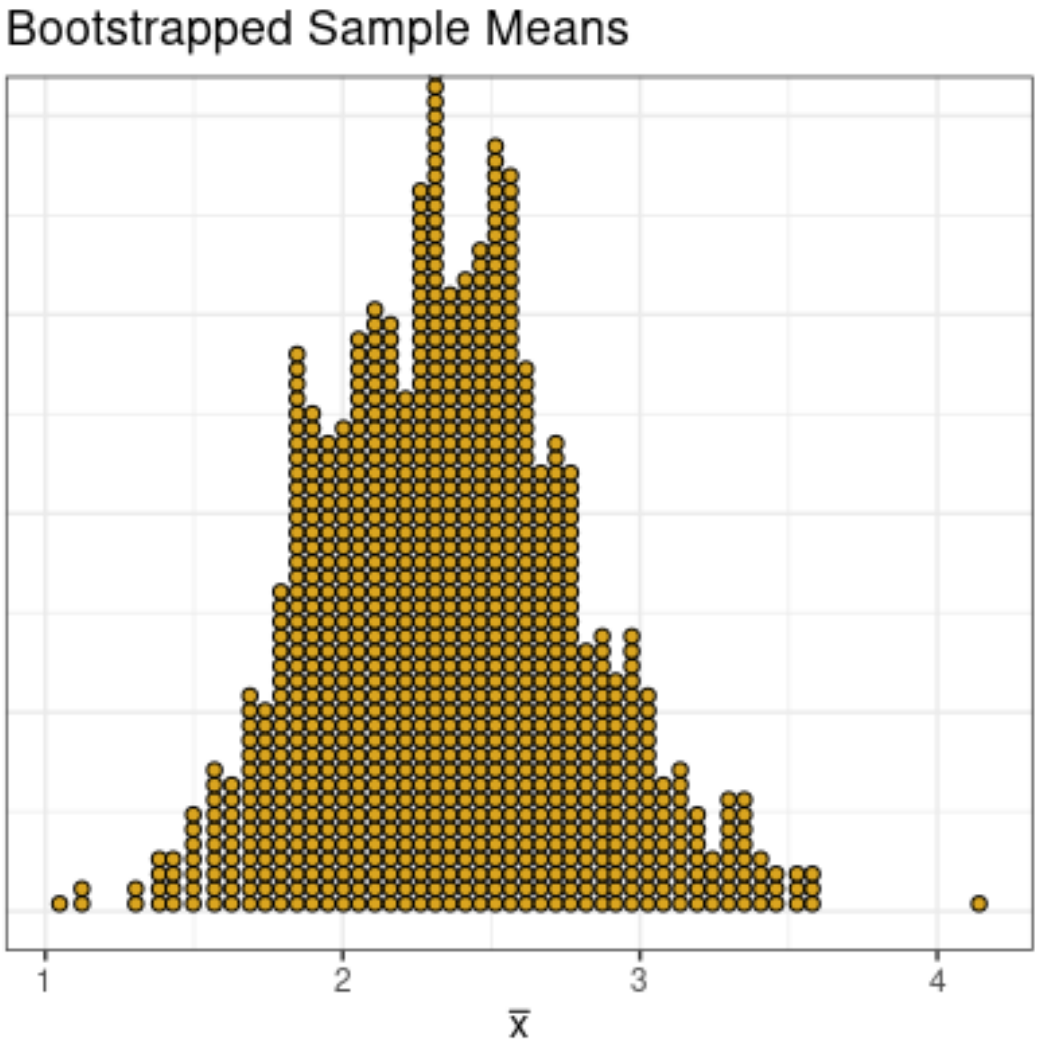
Rainfall Example

Compute the mean of each bootstrap sample...



Rainfall Example

Graph the means from the bootstrap samples $\rightarrow$ **Bootstrap Distribution**



Back to CI's

We can use the bootstrap distribution to make CI's without needing to go back to Normal or t-distribution stuff

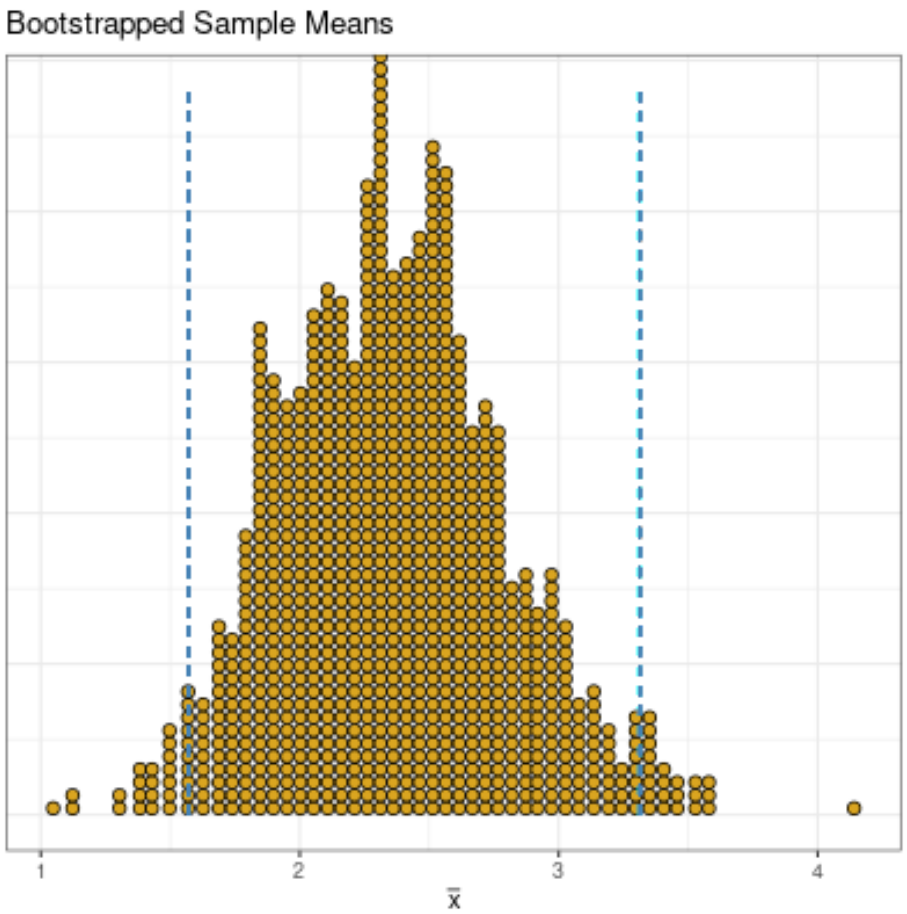
Percentiles

Remember percentiles?

- A value where some % of the distribution is below that value
- ex) median (50th percentile), Q1, Q3

Question: What % of *any* distribution is between the 97.5 percentile and the 2.5 percentile?

Bootstrap Percentiles



95% of bootstrap sample means are between the 2.5 percentile and the 97.5 percentile

- This constitutes a 95% confidence interval!

More on Bootstrapping

Other % CI's? Use different percentiles to make the CI

- 80% CI $\rightarrow$ 10 and 90 percentiles
- 90% CI $\rightarrow$ 5 and 95 percentiles
- 95% CI $\rightarrow$ 2.5 and 97.5 percentiles
- 99% CI $\rightarrow$ 0.5 and 99.5 percentiles

More on Bootstrapping

Benefits:

- Can use bootstrapping for things other than means or proportions
- Don't need to rely on Normal / t-distributions
  - use when Normal / t-distribution conditions aren't met

Downsides:

- Need access to computer to simulate bootstrap process
- bootstrap CIs are often wider than Normal/t-distribution intervals