Confidence Intervals

Part 2 - Bootstrapping

Ryan Miller

Grinnell College
Statistics

Introduction

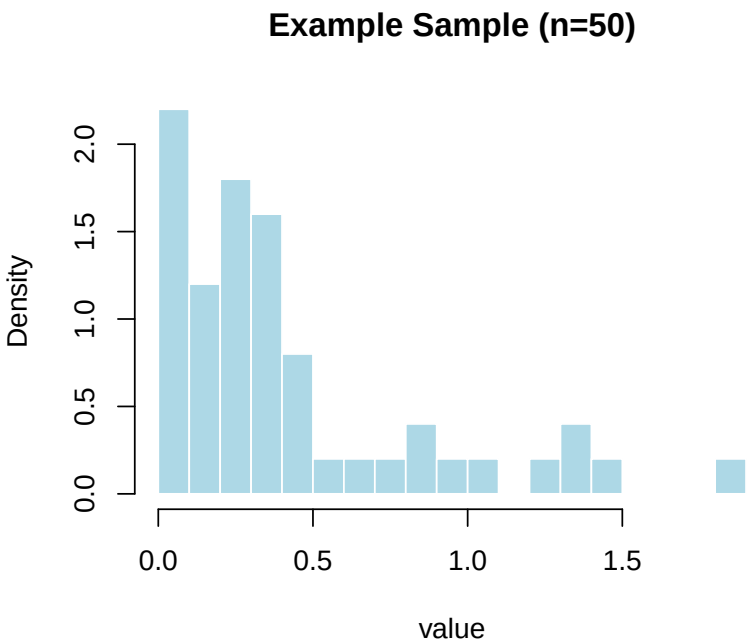
So far, we've created confidence intervals using the formula:

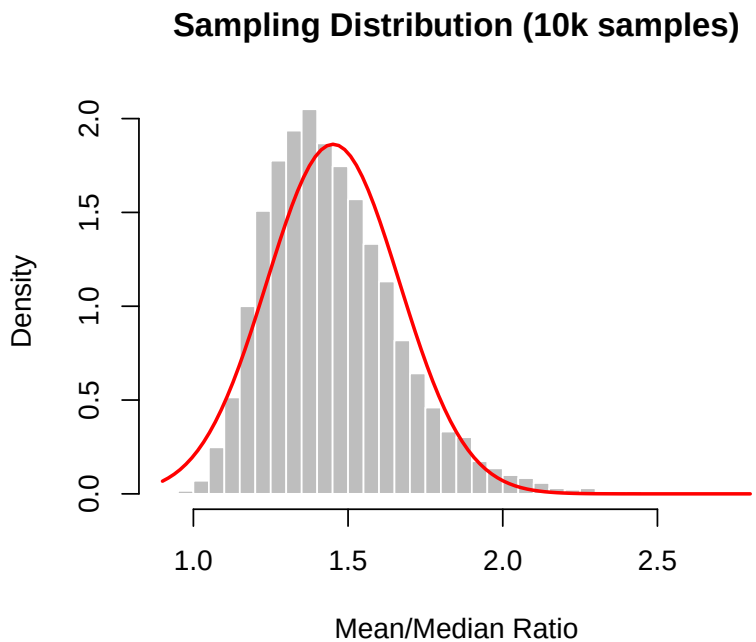
$$\text{point estimate} \pm c \cdot SE$$

These intervals rely upon using a probability model (ie: Normal distribution) to approximate the **sampling distribution**, or the distribution of a sample statistic (such as the sample mean) across a large number of different samples.

Introduction (cont.)

- For some descriptive statistics, such as means and proportions, we can rely upon statistical theory to inform the probability model used by the confidence interval
  - For other statistics this is very difficult, an example is the mean to median ratio:



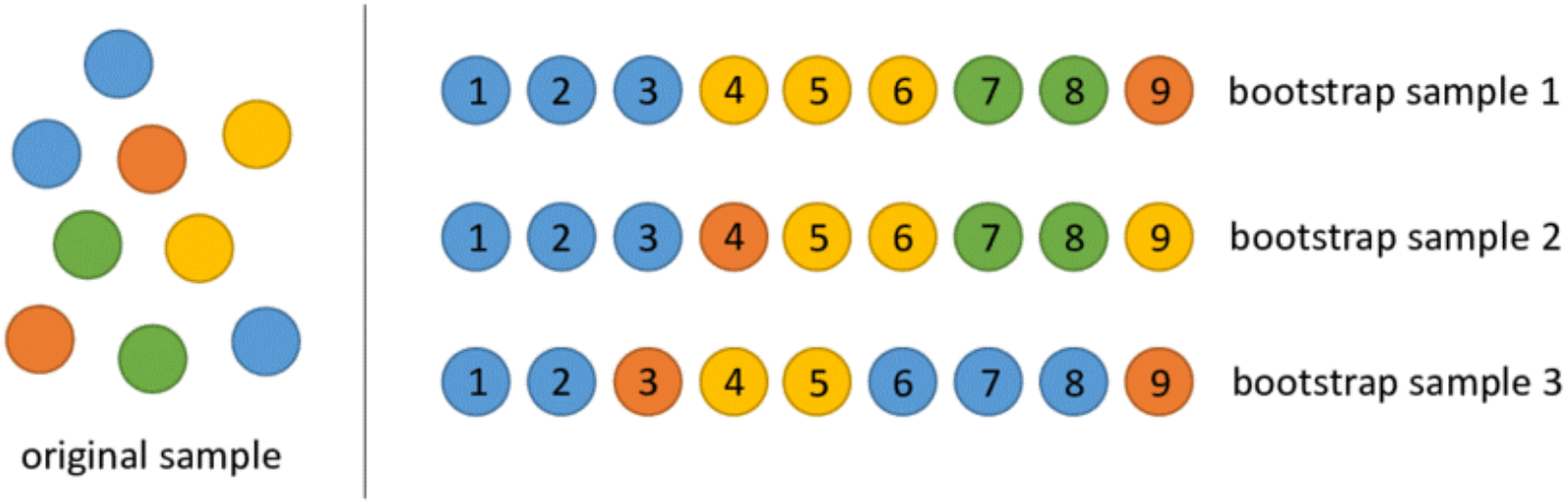


Confidence Interval Validity

- When the probability model used to find c and SE in our confidence interval formula is a poor approximation of the sampling distribution the resulting confidence intervals are likely to be *invalid*
- Today we will learn about **bootstrapping**, which can be used to create confidence intervals without assuming any underlying probability model
  - Bootstrapping re-samples cases from the original sample with replacement to mimic the sampling variability that arises when sampling from a population

Bootstrapping

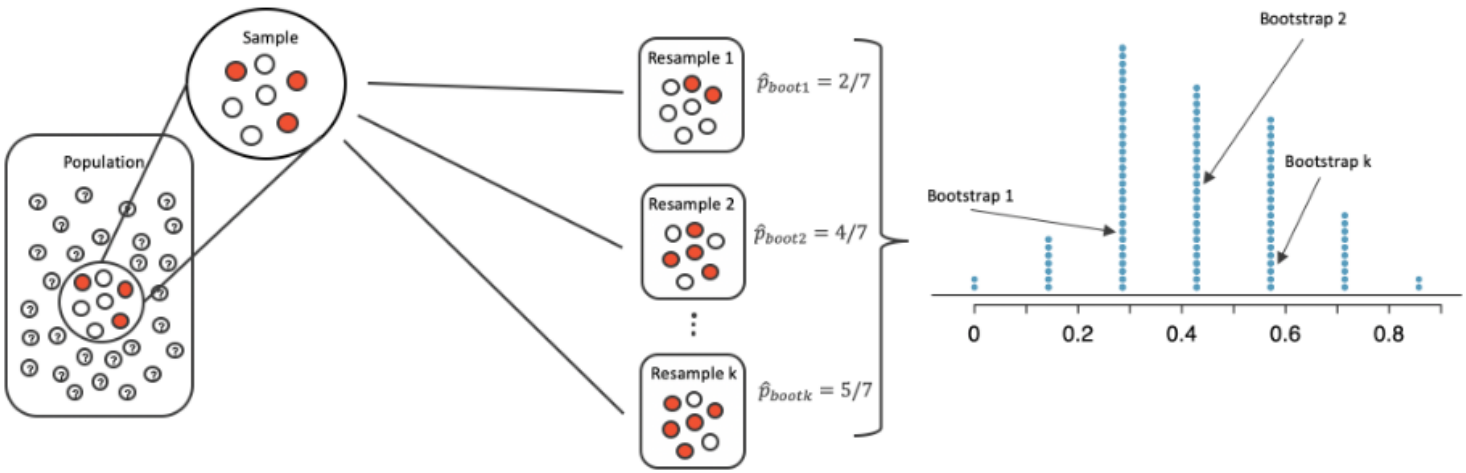
Re-sampling with replacement means that some cases in the original sample will appear more than once in a bootstrap sample:



This is necessary to generate synthetic samples of size n (you should recall that n is an important contributor to *sampling variability*)

Bootstrapping

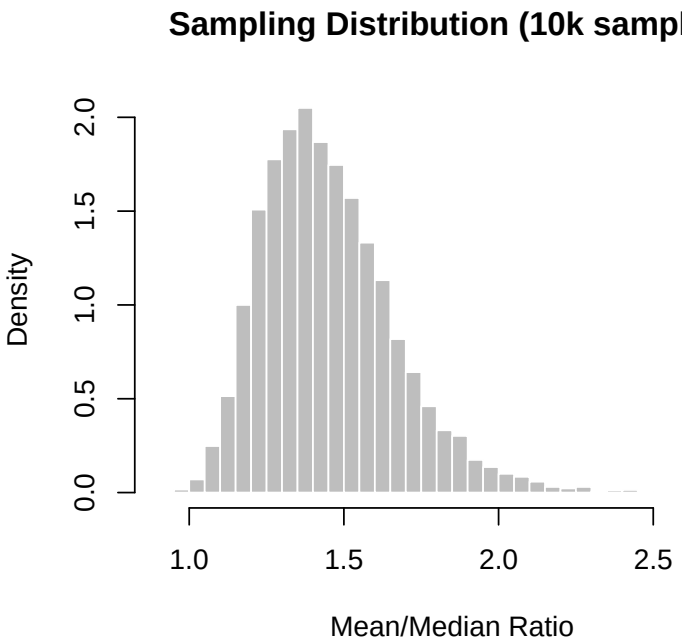
The purpose of bootstrapping is to use your original sample to approximate the sampling distribution:

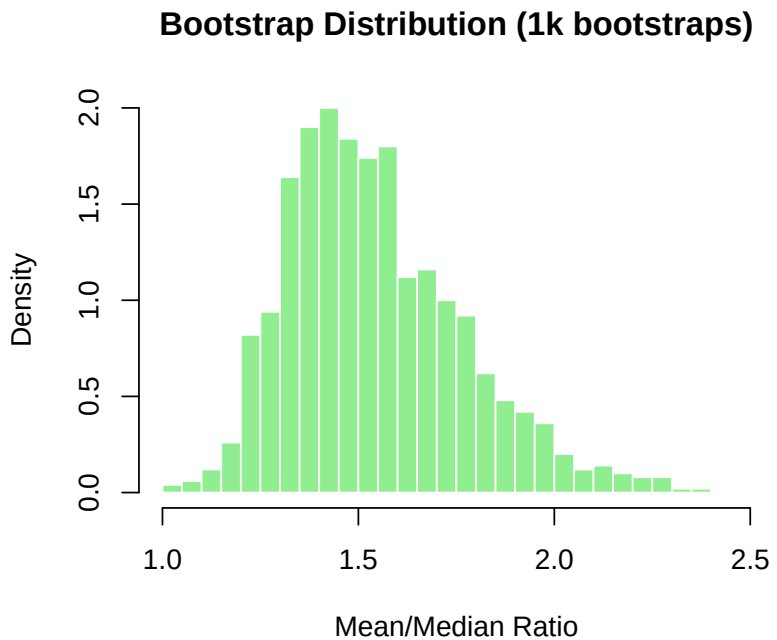


Each *bootstrap sample* produces a *bootstrap estimate*; together, these form the **bootstrap distribution**, which is similar to the sampling distribution

Bootstrap vs. Sampling Distributions

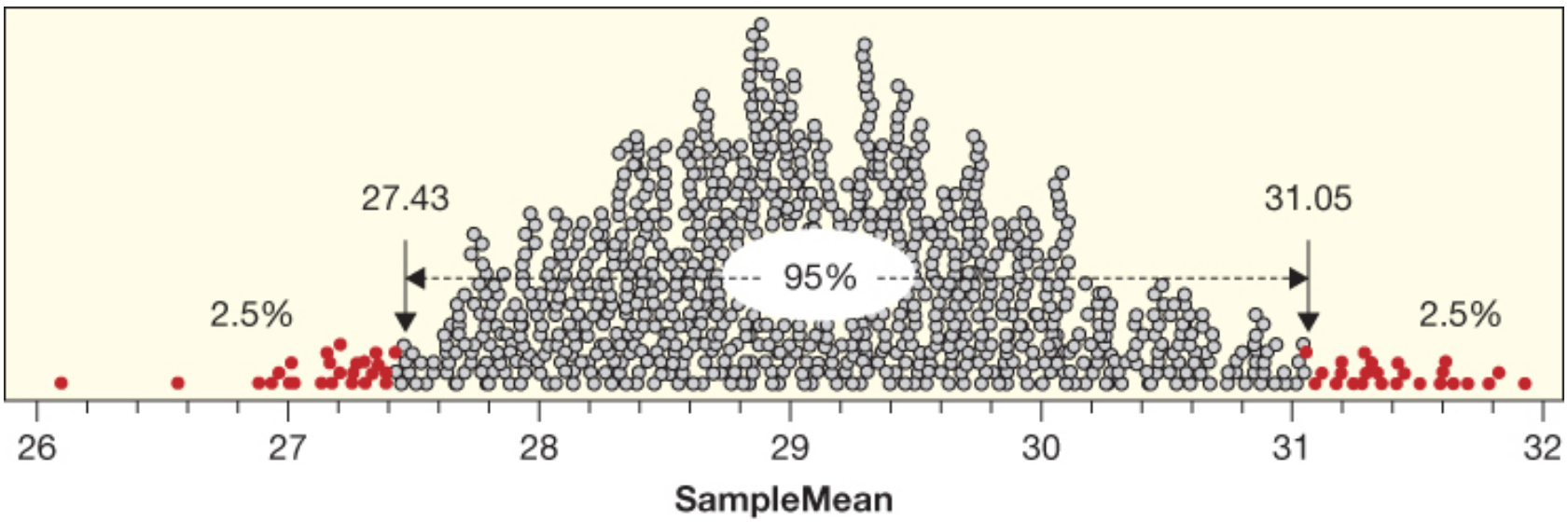
Below is the sampling distribution of the mean to median ratio we previous saw next to the bootstrap distribution created by re-sampling from the single sample of $n = 50$ that was previously shown:





Percentile Bootstrap Confidence Intervals

- When relying upon a probability model to create our confidence interval we focus on the middle $P\%$ of values, where P is the desired confidence level
  - Similarly, we can find a confidence interval using the bootstrap distribution by finding the middle $P\%$ of bootstrapped estimates:



Practice

The data below were recorded by a commuter in the greater Toronto area:

https://remiller1450.github.io/data/CommuteTracker.csv

1. Use StatKey to find a 99% percentile bootstrap interval for the correlation between average moving speed and total commute time.
2. Considering the shape of the bootstrap distribution, why might bootstrapping be a good idea here?
3. Use `R` and `cor.test()` to find a 99% confidence interval for this scenario. How does it compare to the bootstrapped interval?

Practice (solution)

1. The 99% percentile bootstrap confidence interval is approximately (-0.929, -0.760).
2. The bootstrap distribution shows some amount of right-skew, suggesting the sampling distribution isn't symmetric, so a Normal or t-distribution wouldn't be a good model.
3. See `cor.test()` results below. The interval is substantially narrower, suggesting it might not be valid.

```
##
##  Pearson's product-moment correlation
##
## data:  cm$AvgMovingSpeed and cm$TotalTime
## t = -23.693, df = 203, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8982567 -0.8007322
## sample estimates:
##       cor
## -0.8569861
```

Conclusion

- Our previous methods of creating confidence interval estimates are only valid when the underlying probability model provides an accurate approximation of the sampling distribution
- Bootstrapping gives us a flexible method that can be used to create valid confidence intervals in scenarios where we aren't sure if the sampling distribution can be reasonably approximated by a known distribution