Bootstrap Methods and Applications

A data-based journey through a U.S. sitcom

Jacob Forbes

Sobes Research Group

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Table of Contents

Data
- Introduce the data
- View the data

Bootstrap methods with IMDB ratings
- Bootstrap confidence intervals
- Permutation tests for comparing groups
- Standard error approximation

Conclusion
- Pros and cons of bootstrap methods
- Bootstrapping in ATARI
- Questions

The sitcom

We're going to explore data from a well-known U.S. sitcom that aired from 2005 to 2013. Any guesses?

The sitcom

We're going to explore data from a well-known U.S. sitcom that aired from 2005 to 2013. Any guesses?

The data

Our data come from the The Office Episodes Data which is available on Kaggle. The data were read into R as a data frame. Let's take a look at the columns available for analysis.

The data

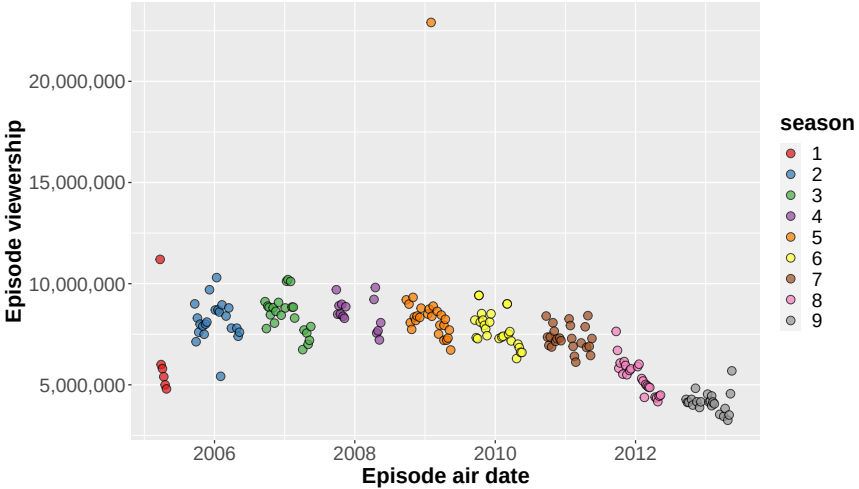
Our data come from the The Office Episodes Data which is available on Kaggle. The data were read into R as a data frame. Let's take a look at the columns available for analysis.

```
## 'data.frame':    188 obs. of  8 variables:
##  $ season     : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 ...
##  $ episode    : int  1 2 3 4 ...
##  $ title      : chr  "Pilot" ...
##  $ us_viewers : int  11200000 6000000 5800000 5400000 ...
##  $ air_date   : Date, format: "2005-03-24" ...
##  $ imdb_rating: num  7.4 8.3 7.7 8 ...
##  $ total_votes: int  7006 6902 5756 5579 ...
##  $ description: chr  "The premiere episode introduces the boss and staff of the Dunder-Mifflin Paper "| __truncated__ ...
```

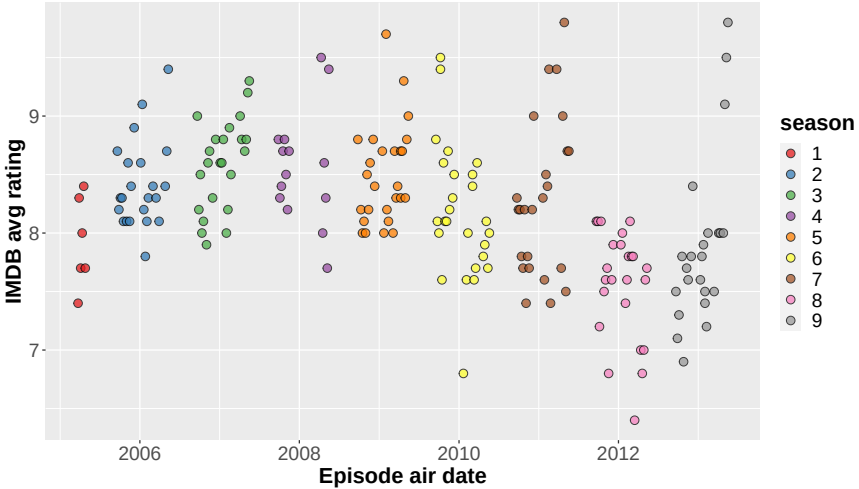
Data familiarity

It's appropriate to familiarize oneself with the data before embarking on any kind of analysis. As such, we would be wise to do some eyes-on data familiarization by watching a short clip from a representative episode of *The Office*.

U.S. viewership per episode

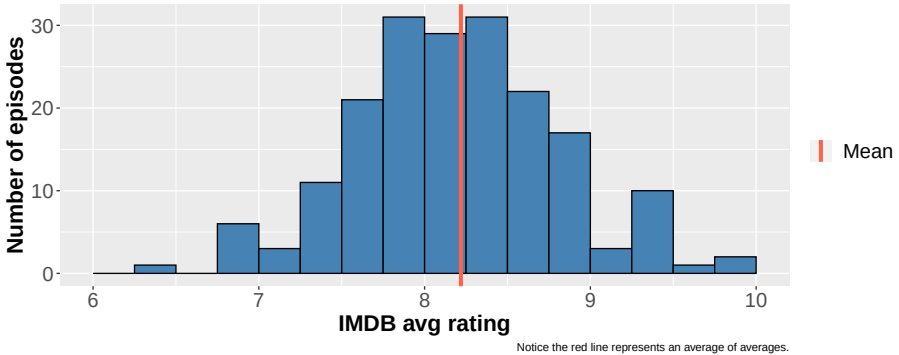


Average IMDB rating per episode



Exploring IMDB ratings further

Suppose IMDB avg ratings were our benchmark for episode quality or performance. We'll use these ratings as our variable of interest for today's examples. First, let's take a look at the distribution of IMDB avg ratings.



Some throat clearing

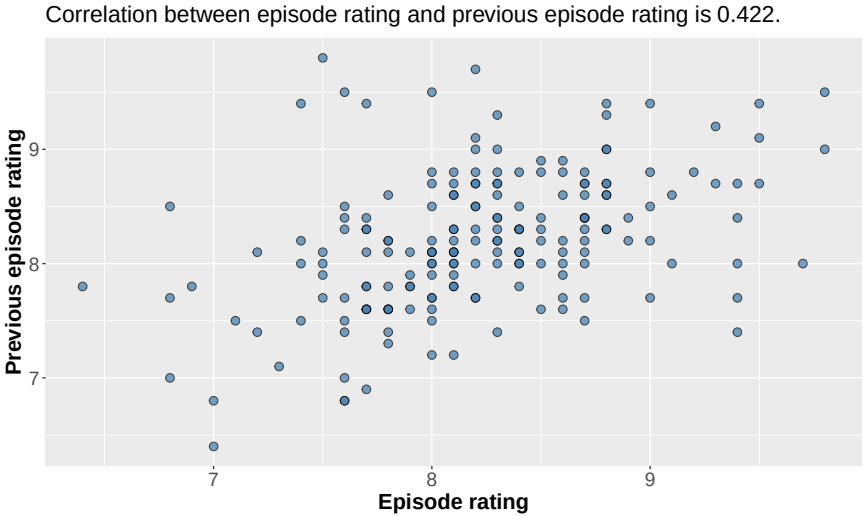
It's worth addressing a few peculiarities about doing statistical inference on the average IMDB ratings. Before I share my concerns, what concerns would you have about doing statistical inference on this data?

Some throat clearing

It's worth addressing a few peculiarities about doing statistical inference on the average IMDB ratings. Before I share my concerns, what concerns would you have about doing statistical inference on this data?

1. We're working with a population dataset.
   What value is there in doing inference with population data?
2. Each observation is itself an average.
   What would it mean to build a confidence interval for, say, the mean value of average IMDB ratings?
3. The observations are not independent.
   To what degree might this degrade our analysis?

Checking independence of observations



Confidence interval for the mean

Suppose we wish to build a 95% confidence interval for the mean value of our parameter of interest. Let's introduce some terms.

- Let $n = 188$ be the number of episodes.
- Let x_i be the average IMDB rating of episode i for $i = 1, ..., n$.
- Let $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the sample mean of the average IMDB ratings.
- Let μ_x be the population mean of the IMDB average ratings (think multiverse).

Then we wish to compute an interval $[b_{\text{lower}}, b_{\text{upper}}]$ such that

$$\mathrm{P}\left(b_{\text{lower}} \leq \mu_x \leq b_{\text{upper}}\right) = 0.95.$$

Computing it the traditional way

In your intro to statistics course, you learned how we can leverage the Central Limit Theorem to compute this confidence interval. Assuming we have

- Random samples
- Independent samples
- Sufficiently large sample size (relative to the population size)
- Sufficiently large sample size (relative to the population distribution)

Computing it the traditional way

In your intro to statistics course, you learned how we can leverage the Central Limit Theorem to compute this confidence interval. Assuming we have

- Random samples
- Independent samples
- Sufficiently large sample size (relative to the population size)
- Sufficiently large sample size (relative to the population distribution)

then $\bar{x}$ is distributed Normal $\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$, where σ_x is the population standard deviation of x. We can compute our 95% confidence interval for μ_x as follows.

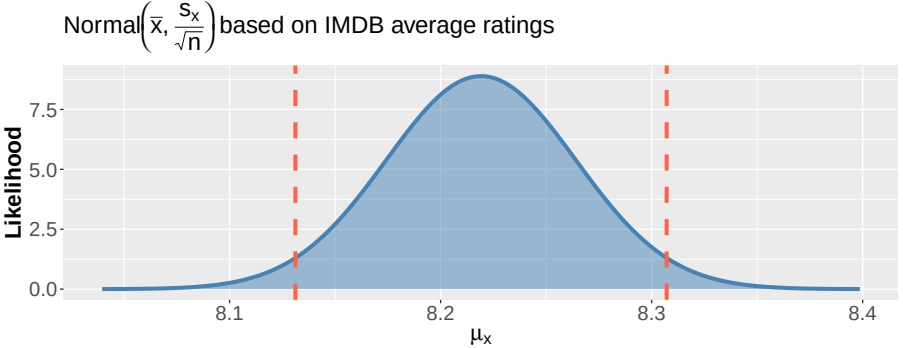
$$[b_{\text{lower}}, b_{\text{upper}}] \approx \bar{x} \pm z_{0.975} \frac{s_x}{\sqrt{n}}$$

where $z_{0.975}$ is the 97.5th quantile of a standard normal distribution, and s_x is the sample standard deviation of x.

Computing it the traditional way

According to our CLT-based calculations, we can be 95% confident that μ_x, the population mean of the IMDB average ratings, is contained in the interval

$$[b_{\text{lower}}, b_{\text{upper}}] \approx \bar{x} \pm z_{0.975} \frac{s_x}{\sqrt{n}} \approx [8.131, 8.307]\,.$$



Questions

What questions do you have?