Bootstrap Methods and Applications

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

Jacob Forbes

University of Tennessee, Knoxville

SOBES RESEARCH GROUP

Table of Contents

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

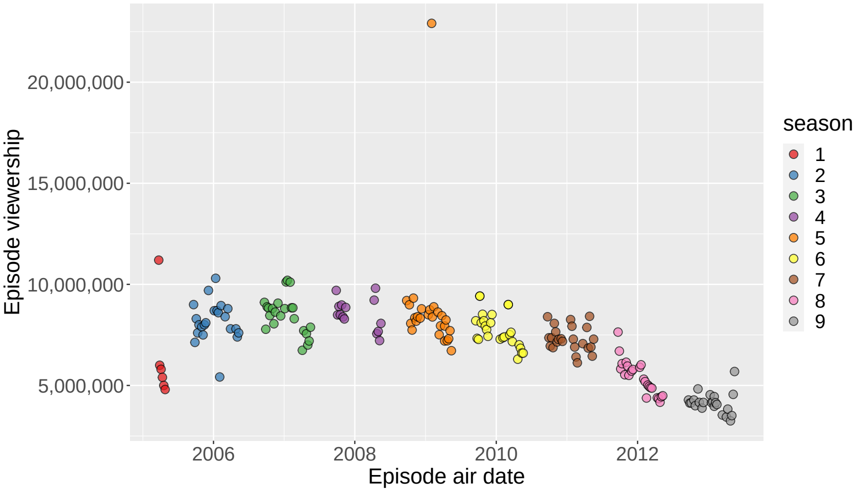
Our data come from the The Office Episodes Data which is available on Kaggle. The data were read into R as a data frame. Let's take a look at the columns available for analysis.

```
## '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__ ...
```

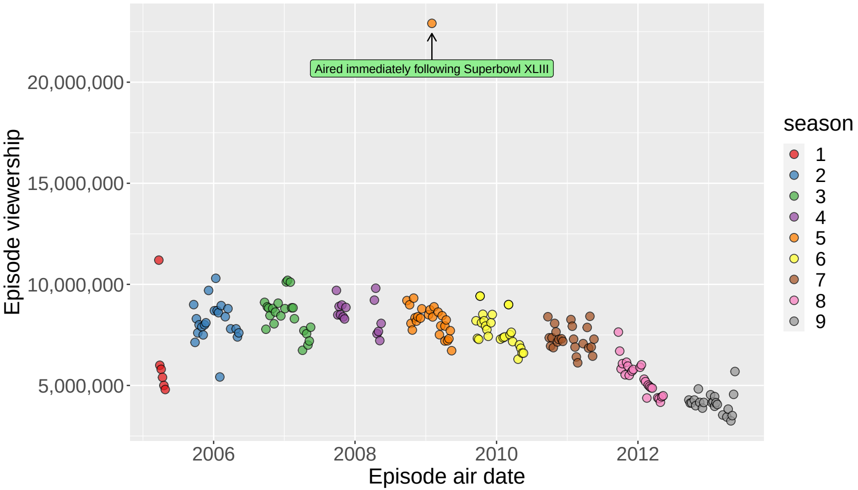
Data familiarity

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*.

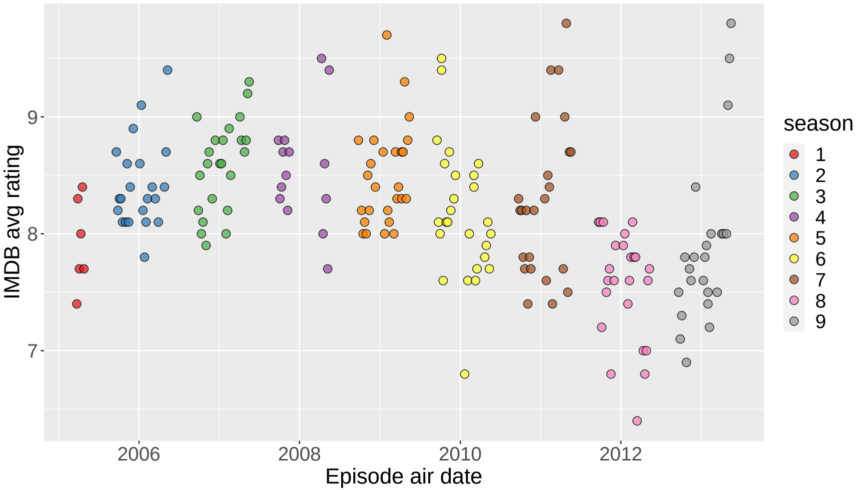
U.S. viewership per episode



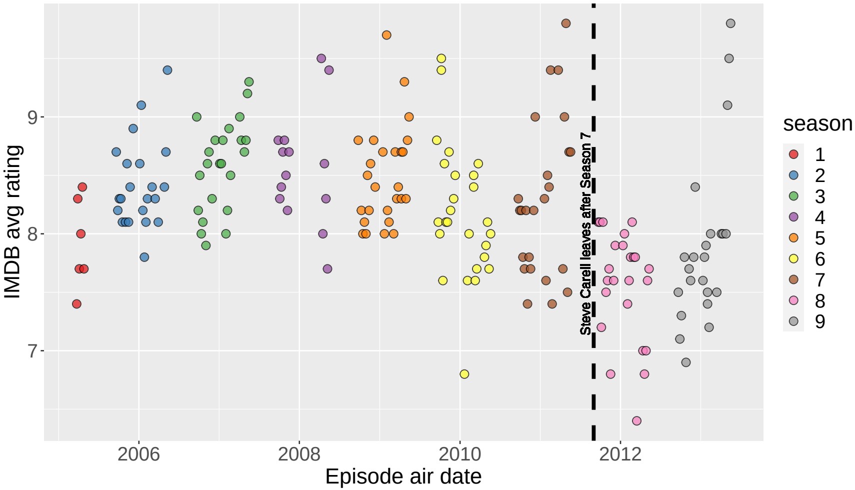
U.S. viewership per episode



Average IMDB rating per episode

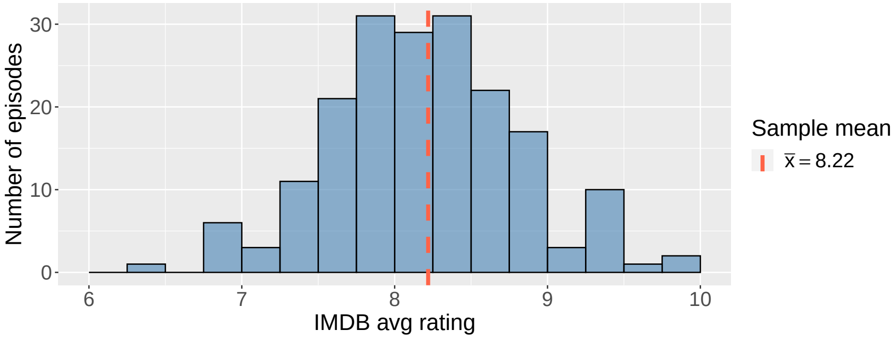


Average IMDB rating per episode



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.



Notice the red line represents an average of averages.

Some throat clearing

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

Some throat clearing

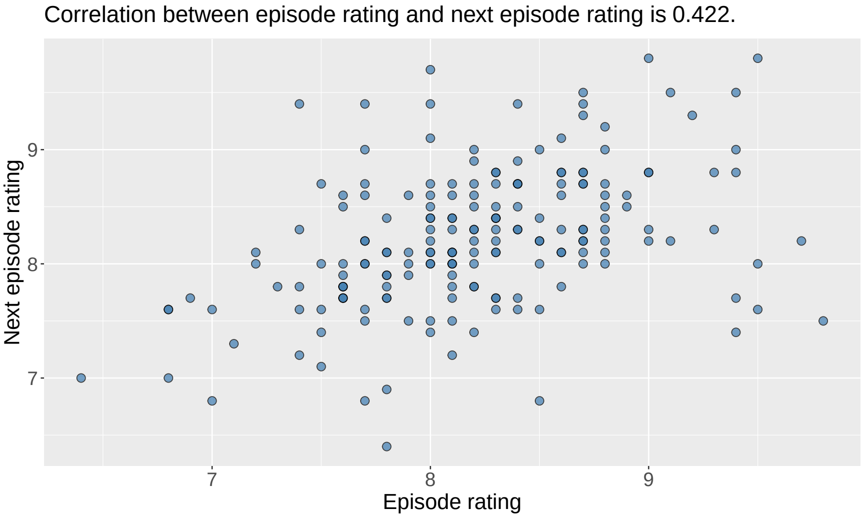
It's worth addressing a few peculiarities about doing statistical inference on the IMDB average ratings. Before I share my concerns, what concerns would you have about doing statistical inference on this data?

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 IMDB average ratings?

3. The observations are not independent.
   To what degree might this degrade our analysis?

Checking independence of observations



The Michael Scott effect

We saw previously that ratings dropped when Steve Carell left *The Office*. Suppose we were to partition our data into

1. Episodes whose episode descriptions do reference Michael
2. Episodes whose episode descriptions do NOT reference Michael

The Michael Scott effect

We saw previously that ratings dropped when Steve Carell left *The Office*. Suppose we were to partition our data into

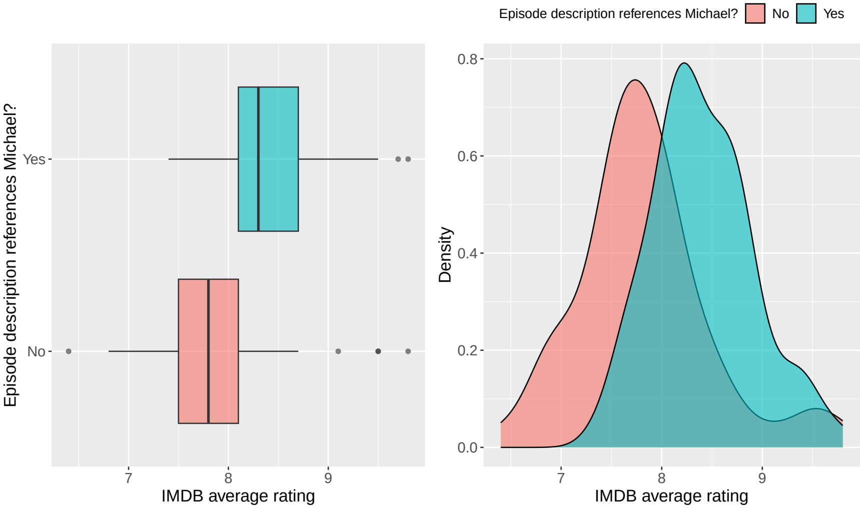
1. Episodes whose episode descriptions do reference Michael
2. Episodes whose episode descriptions do NOT reference Michael

Here's an episode description of an episode which would belong to the first group.

> *Dwight's too-realistic fire alarm gives Stanley a heart attack. When he returns, Michael learns that he is the cause of Stanley's stress. To remedy the situation, he forces the office to throw a roast for him.*

Would you expect to see a noticeable difference in IMDB average ratings between the two groups?

The Michael Scott effect



Note: 128 of the 188 episodes have descriptions which make reference to Michael.

The Michael Scott effect

We might want to know

> *Are these differences in ratings statistically significant? Or, could they just be attributed to just "noise" in the data?*

Based on your knowledge of statistics, how would you answer these questions?

The traditional approach

Let μ_M be the population mean of IMDB average ratings for episodes whose episode descriptions refer to Michael. Let μ_N be the population mean of IMDB average ratings for all other episodes. Then we can test the following hypotheses

$$H_0: \ \mu_M = \mu_N$$
$$H_A: \ \mu_M > \mu_N$$

using a t-test for difference of means.

The traditional approach

What are the mechanics of the t-test for difference of means?

The traditional approach

What are the mechanics of the t-test for difference of means?

1. Assume the observations are independent, normally distributed random variables with means μ_k and roughly equal variances σ_k^2 for groups $k \in \{M, N\}$. That is, assume each observation, x_i, has the following probability density function.

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}}$$

The traditional approach

What are the mechanics of the t-test for difference of means?

1. Assume the observations are independent, normally distributed random variables with means μ_k and roughly equal variances σ_k^2 for groups $k \in \{M, N\}$. That is, assume each observation, x_i, has the following probability density function.

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}}$$

2. Compute point estimate $\bar{x}_M - \bar{x}_N$, where $\bar{x}_k$ is the sample mean of group k.

The traditional approach

What are the mechanics of the t-test for difference of means?

1. Assume the observations are independent, normally distributed random variables with means μ_k and roughly equal variances σ_k^2 for groups $k \in \{M, N\}$. That is, assume each observation, x_i, has the following probability density function.

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}}$$

2. Compute point estimate $\bar{x}_M - \bar{x}_N$, where $\bar{x}_k$ is the sample mean of group k.
3. Compute the standard error of the point estimate,

$$\mathrm{SE}_{\bar{x}_M - \bar{x}_N} = \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_N^2}{n_N}} \approx \sqrt{\frac{s_M^2}{n_M} + \frac{s_N^2}{n_N}}$$

where s_k^2 is the sample variance and n_k is the sample size of group k.

The traditional approach

What are the mechanics of the t-test for difference of means? (continued)

4. Compute the test statistic,

$$t^* = \frac{\bar{x}_M - \bar{x}_N}{\mathrm{SE}_{\bar{x}_M - \bar{x}_N}}.$$

The traditional approach

What are the mechanics of the t-test for difference of means? (continued)

4. Compute the test statistic,

$$t^* = \frac{\bar{x}_M - \bar{x}_N}{\mathrm{SE}_{\bar{x}_M - \bar{x}_N}}.$$

5. Note that given our assumptions, t^* follows a student-t distribution with

$$v = \frac{\left(\frac{s_M^2}{n_M} + \frac{s_N^2}{n_N}\right)^2}{\frac{s_M^4}{n_M^2(n_M-1)} + \frac{s_N^4}{n_N^2(n_N-1)}}$$

degrees of freedom.

The traditional approach

What are the mechanics of the t-test for difference of means? (continued)

5. ... That is, given our assumptions, our test statistic will have the following probability density function.

$$f(t^*) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{{t^*}^2}{v}\right)^{-\frac{v+1}{2}}$$

The traditional approach

What are the mechanics of the t-test for difference of means? (continued)

5. ... That is, given our assumptions, our test statistic will have the following probability density function.

$$f(t^*) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{{t^*}^2}{v}\right)^{-\frac{v+1}{2}}$$

6. Compute p-value $= \mathrm{P}\left(t > t^*\right) = \int_{t^*}^{\infty} f(t)dt$ where $f(\cdot)$ is the probability density function for a student-t distribution with v degrees of freedom.

The traditional approach

What are the mechanics of the t-test for difference of means? (continued)

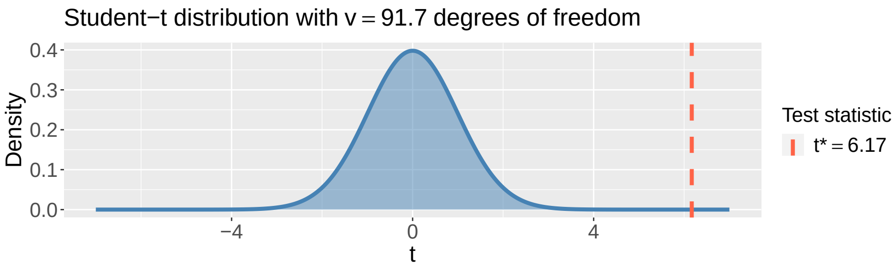
5. ... That is, given our assumptions, our test statistic will have the following probability density function.

$$f(t^*) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{{t^*}^2}{v}\right)^{-\frac{v+1}{2}}$$

6. Compute p-value $= \mathrm{P}\left(t > t^*\right) = \int_{t^*}^{\infty} f(t)dt$ where $f(\cdot)$ is the probability density function for a student-t distribution with v degrees of freedom.

7. Use the p-value to evaluate H_0 vs. H_A. This is often done by comparing the p-value to some previously agreed upon significance level, α.

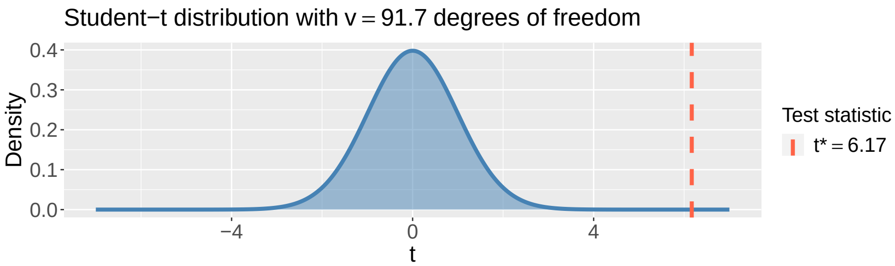
The traditional approach



Using the traditional approach, we compute

$$\text{p-value} = \mathrm{P}\left(t > t^*\right) = \int_{t^*}^{\infty} f(t)dt \approx 9.199 \times 10^{-9} \approx 0.$$

The traditional approach



Using the traditional approach, we compute

$$\text{p-value} = \mathrm{P}\left(t > t^*\right) = \int_{t^*}^{\infty} f(t)dt \approx 9.199 \times 10^{-9} \approx 0.$$

That is, given all the aforementioned assumptions, the probability we would observe a difference in sample means $\bar{x}_M - \bar{x}_N$ as large or larger than what we actually observed if $H_0: \ \mu_M = \mu_N$ were in fact true is nearly 0. There is statistically significant evidence for the "Michael Scott effect".

A computational approach: Permutation Test

A different approach and the logic behind it:

A computational approach: Permutation Test

A different approach and the logic behind it:

- Suppose there really were no “Michael Scott effect.” That is, suppose that the distribution of IMDB average ratings were the same between the two groups, and the observed difference is just due to chance.

A computational approach: Permutation Test

A different approach and the logic behind it:

- Suppose there really were no "Michael Scott effect." That is, suppose that the distribution of IMDB average ratings were the same between the two groups, and the observed difference is just due to chance.
- Under this assumption, any partition of the 188 IMDB average ratings into two groups of size $n_M = 128$ and $n_N = 60$ would be equally probable to any other such partition.

A computational approach: Permutation Test

A different approach and the logic behind it:

- Suppose there really were no "Michael Scott effect." That is, suppose that the distribution of IMDB average ratings were the same between the two groups, and the observed difference is just due to chance.
- Under this assumption, any partition of the 188 IMDB average ratings into two groups of size $n_M = 128$ and $n_N = 60$ would be equally probable to any other such partition.
- Theoretically, we could compute the difference in sample means of IMDB average ratings we *would* observe under each of these partitions. This would give us a distribution of all the possible differences in means one *would* expect to see *if* there really were no "Michael Scott effect."

A computational approach: Permutation Test

A different approach and the logic behind it:

- Suppose there really were no "Michael Scott effect." That is, suppose that the distribution of IMDB average ratings were the same between the two groups, and the observed difference is just due to chance.
- Under this assumption, any partition of the 188 IMDB average ratings into two groups of size $n_M = 128$ and $n_N = 60$ would be equally probable to any other such partition.
- Theoretically, we could compute the difference in sample means of IMDB average ratings we *would* observe under each of these partitions. This would give us a distribution of all the possible differences in means one *would* expect to see *if* there really were no "Michael Scott effect."
- Now, if the real, observed difference is unusually large (i.e., highly improbable) relative to the distribution of possible differences, then we have evidence against the hypothesis that there is no "Michael Scott effect."

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.
- $n_E = n_M + n_N = 128 + 60 = 188$ be the total number of episodes.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.
- $n_E = n_M + n_N = 128 + 60 = 188$ be the total number of episodes.
- $\bar{x}_M = \frac{1}{n_M} \sum_{i \in M} x_i$ and $\bar{x}_N = \frac{1}{n_N} \sum_{i \in N} x_i$ be the sample means of IMDB average ratings for the two groups of interest.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, \ldots, 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.
- $n_E = n_M + n_N = 128 + 60 = 188$ be the total number of episodes.
- $\bar{x}_M = \frac{1}{n_M} \sum_{i \in M} x_i$ and $\bar{x}_N = \frac{1}{n_N} \sum_{i \in N} x_i$ be the sample means of IMDB average ratings for the two groups of interest.
- $\bar{d} = \bar{x}_M - \bar{x}_N$.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.
- $n_E = n_M + n_N = 128 + 60 = 188$ be the total number of episodes.
- $\bar{x}_M = \frac{1}{n_M} \sum_{i \in M} x_i$ and $\bar{x}_N = \frac{1}{n_N} \sum_{i \in N} x_i$ be the sample means of IMDB average ratings for the two groups of interest.
- $\bar{d} = \bar{x}_M - \bar{x}_N$.
- $n_B = \binom{n_E}{n_M} = \frac{n_E!}{n_M! n_N!} = \frac{188!}{128!60!} = 8.401164 \times 10^{49}$ be the number of ways you can partition the elements of E into two subsets of size n_M and n_N.

A computational approach: Permutation Test

Let's formalize this by defining and reviewing some terms. Let

- x_i be the IMDB average rating of episode i.
- $E = \{1, 2, ..., 188\}$ be the set of all episode numbers.
- M and N be the subsets of episode numbers with and without descriptions which refer to Michael, respectively.
- $n_E = n_M + n_N = 128 + 60 = 188$ be the total number of episodes.
- $\bar{x}_M = \frac{1}{n_M} \sum_{i \in M} x_i$ and $\bar{x}_N = \frac{1}{n_N} \sum_{i \in N} x_i$ be the sample means of IMDB average ratings for the two groups of interest.
- $\bar{d} = \bar{x}_M - \bar{x}_N$.
- $n_B = \binom{n_E}{n_M} = \frac{n_E!}{n_M! n_N!} = \frac{188!}{128!60!} = 8.401164 \times 10^{49}$ be the number of ways you can partition the elements of E into two subsets of size n_M and n_N.
- B be the set of all n_B possible partitions of the elements of E into two subsets of size n_M and n_N. The jth element of B is $\{M^*_{(j)}, N^*_{(j)}\}$, the subsets to which the elements of E are partitioned.

A computational approach: Permutation Test

The procedure for the permutation test is as follows.

for $j = 1 \ldots n_B$ **do**

$\bar{x}^*_{M,j} = \frac{1}{n_M} \sum_{i \in M^*_{(j)}} x_i$

$\bar{x}^*_{N,j} = \frac{1}{n_N} \sum_{i \in N^*_{(j)}} x_i$

$\bar{d}^*_j = \bar{x}^*_{M,j} - \bar{x}^*_{N,j}$

end for

$\texttt{p-value} = \frac{1}{n_B} \sum_{j=1}^{n_B} I_{\bar{d}^*_j > \bar{d}}$

Translation: Loop through every possible partition of the episodes into $n_M = 128$ and $n_N = 60$ episodes, and compute the difference in sample means. Compare the true difference in sample means to the distribution of possible differences under the no-Michael-Scott-effect assumption.

A computational approach: Permutation Test

The procedure for the permutation test is as follows.

for $j = 1 \ldots n_B$ **do**

$\bar{x}^*_{M,j} = \frac{1}{n_M} \sum_{i \in M^*_{(j)}} x_i$

$\bar{x}^*_{N,j} = \frac{1}{n_N} \sum_{i \in N^*_{(j)}} x_i$

$\bar{d}^*_j = \bar{x}^*_{M,j} - \bar{x}^*_{N,j}$

end for

$\texttt{p-value} = \frac{1}{n_B} \sum_{j=1}^{n_B} I_{\bar{d}^*_j > \bar{d}}$

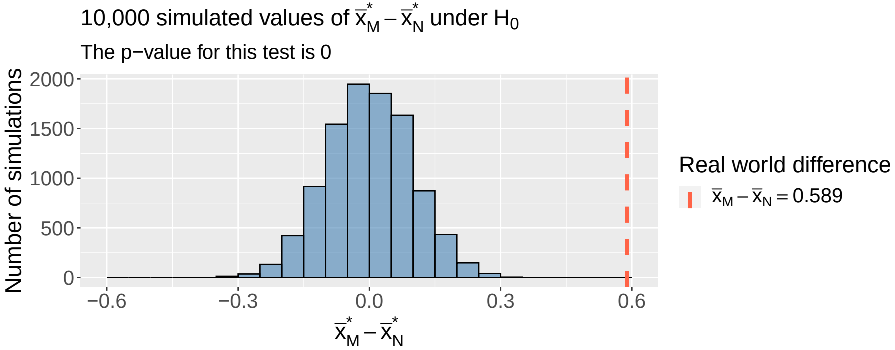
Translation: Loop through every possible partition of the episodes into $n_M = 128$ and $n_N = 60$ episodes, and compute the difference in sample means. Compare the true difference in sample means to the distribution of possible differences under the no-Michael-Scott-effect assumption. **Do you see any potential technical problems with this approach?**

A computational approach: Permutation Test

In practice, we don't loop through all possible partitions. Instead, we build the reference distribution with some sufficiently large number of partitions. For example,

A computational approach: Permutation Test

In practice, we don't loop through all possible partitions. Instead, we build the reference distribution with some sufficiently large number of partitions. For example,



Note: Each of the 10,000 partitions is formed by random sampling $n_M = 128$ of the IMDB avg ratings without replacement. Those selected form the first group; the remaining $n_N = 60$ ratings form the second.

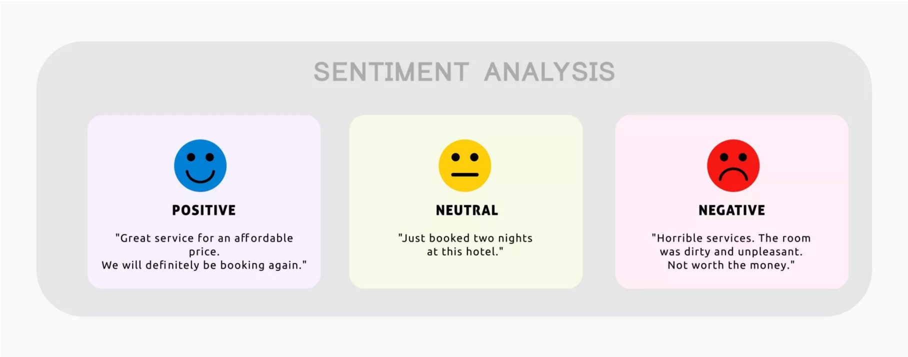
Sentiment analysis

We've found statistically significant evidence for the "Michael Scott effect." What other information might we use to explain the variance in the IMDB average ratings?

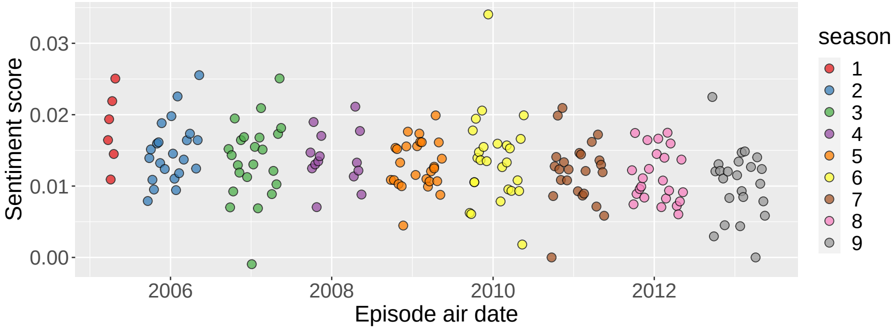
Here I thought it would be interesting to explore a sentiment analysis of the episode scripts. Do more positive or negative episode scripts correlate to IMDB average ratings?

This is not a presentation about sentiment analysis. Real quickly, has anyone here heard of it?

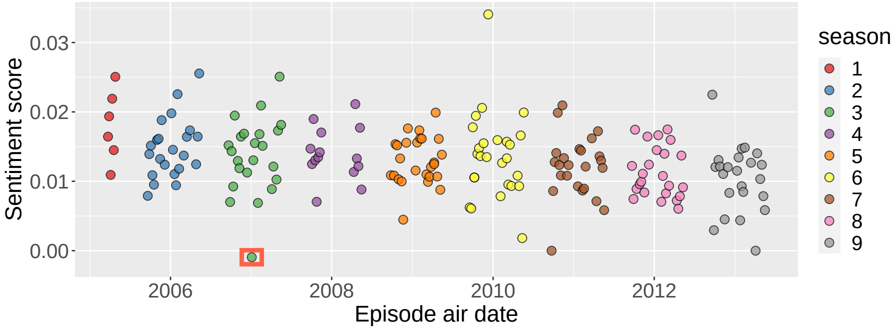
Sentiment analysis



Sentiment analysis

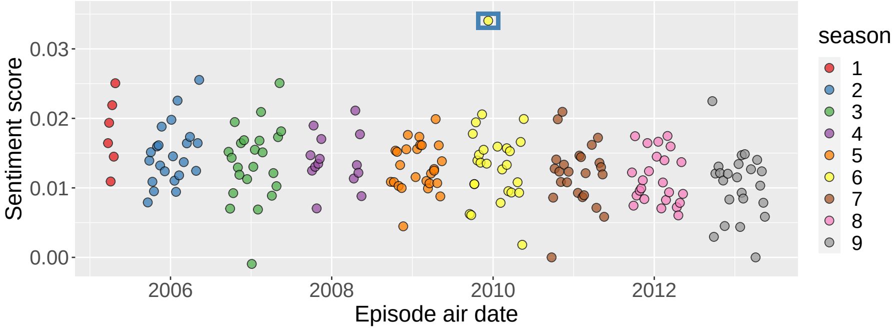


Sentiment analysis



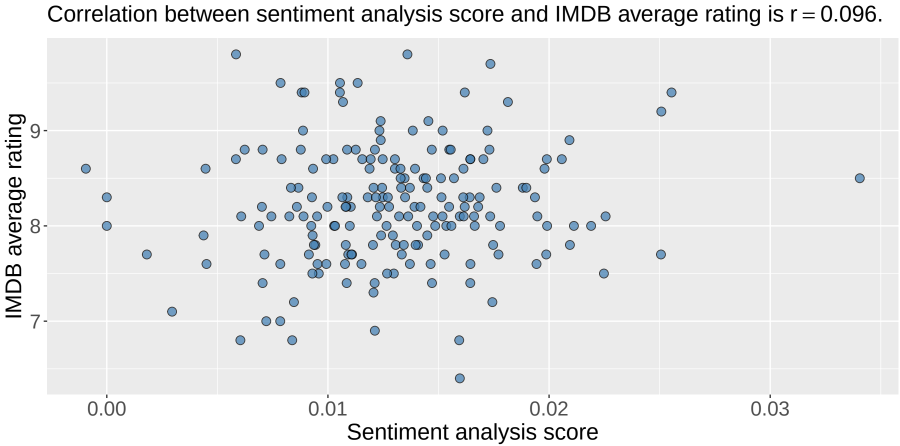
Season 3, Episode 11: *Back from Vacation*. "Michael returns from his Jamaican vacation healthy and revitalized, but it is short lived as a saucy photograph from his vacation begins circulating around the office. Meanwhile, Jim and Karen have an argument and Pam is caught right in the middle of it."

Sentiment analysis



Season 6, Episode 13: *Secret Santa*. "Michael is outraged when Jim allows Phyllis to be Santa at the office Christmas party. Jim and Dwight try to get everyone into the holiday spirit despite the uncertainty with Dunder Mifflin. Meanwhile, Oscar has a secret crush."

Sentiment analysis



We might want to know: how stable is r? I.e., what is the standard error of r?

Estimating standard error via the bootstrap

There are several proposed formulas for SE_r depending on what kind of distributional assumptions one is willing to make about x and y. But why bother with all that? This situation is the bread and butter of bootstrap methods. Here's the procedure:

Estimating standard error via the bootstrap

There are several proposed formulas for SE_r depending on what kind of distributional assumptions one is willing to make about x and y. But why bother with all that? This situation is the bread and butter of bootstrap methods. Here's the procedure:

1. Let (x_i, y_i) be the IMDB average rating and sentiment score for episode i.

Estimating standard error via the bootstrap

There are several proposed formulas for SE_r depending on what kind of distributional assumptions one is willing to make about x and y. But why bother with all that? This situation is the bread and butter of bootstrap methods. Here's the procedure:

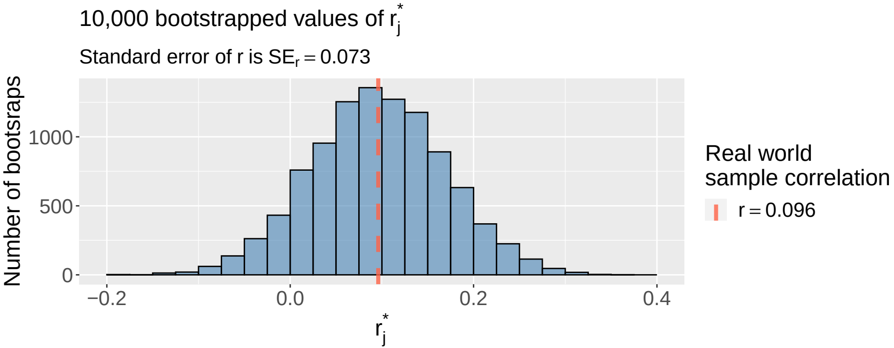
1. Let (x_i, y_i) be the IMDB average rating and sentiment score for episode i.
2. For some sufficiently large number of times, n_B, do the following.
   2.1 Randomly sample $n_E = 188$ values from $E = \{1, ..., 188\}$ *with replacement*. Define set $B^*_{(j)}$ as the set of sampled values of simulation iteration j.
   2.2 Compute the sample correlation, r^*_j, of the points $(x_i, y_i)\ \forall i \in B^*_{(j)}$.

Estimating standard error via the bootstrap

There are several proposed formulas for SE_r depending on what kind of distributional assumptions one is willing to make about x and y. But why bother with all that? This situation is the bread and butter of bootstrap methods. Here's the procedure:

1. Let (x_i, y_i) be the IMDB average rating and sentiment score for episode i.
2. For some sufficiently large number of times, n_B, do the following.
   2.1 Randomly sample $n_E = 188$ values from $E = \{1, ..., 188\}$ *with replacement*. Define set $B^*_{(j)}$ as the set of sampled values of simulation iteration j.
   2.2 Compute the sample correlation, r^*_j, of the points $(x_i, y_i)\ \forall i \in B^*_{(j)}$.
3. Approximate $\mathrm{SE}_r \approx \sqrt{\frac{\sum_{j=1}^{n_B} \left(r^*_j - \bar{r}^*\right)^2}{n_B - 1}}$. I.e., compute the sample standard deviation of the bootstrapped r^*_j values.

Estimating standard error via the bootstrap



This clarifies how tenuous the relationship between sentiment score and IMDB average rating is. Values of $r \leq 0$ are not highly improbable based on the sampling distribution of r.

Pros and cons of bootstrap methods

Pros

- **We can compute standard errors, build confidence intervals, and perform hypothesis tests for any statistic of the data.** When the only tool you have is a hammer, every problem begins to look like a nail. This is an issue for data analysts who limit themselves to traditional statistics. If traditional statistics is one's only tool, one limits themselves only to inference on means, variances, and linear regression coefficients for which the pioneers of math have analytically derived the distributional forms. We are not so limited. Using bootstrap methods, we could build a confidence interval for, say, $e^{\cos(x_{\text{median}})}$ if the problem required it.

Pros and cons of bootstrap methods

Pros

- **We can compute standard errors, build confidence intervals, and perform hypothesis tests for any statistic of the data.** When the only tool you have is a hammer, every problem begins to look like a nail. This is an issue for data analysts who limit themselves to traditional statistics. If traditional statistics is one's only tool, one limits themselves only to inference on means, variances, and linear regression coefficients for which the pioneers of math have analytically derived the distributional forms. We are not so limited. Using bootstrap methods, we could build a confidence interval for, say, $e^{\cos(x_{\text{median}})}$ if the problem required it.
- **No distributional assumptions.** We need not rely on the normality assumptions which are often required for traditional statistics. Many traditional methods are robust to these assumptions, but still, why needlessly tie ourselves to assumptional requirements?

Pros and cons of bootstrap methods

Pros (continued)

- **Simple procedures.** As anyone who has taken a mathematical statistics course knows, the logic behind traditional methods is not intuitive for most mere academic mortals. We compute seemingly arbitrary test statistics of the data because they've been proven to follow some complex distributional form. The first year student usually doesn't have the requisite mathematical foundation for understanding these proofs. They're expected to hit the "I believe" button and trust the procedure works. Computational methods are based on more intuitive principles. As such, they're relatively easy to implement in code.

Pros and cons of bootstrap methods

Cons

- **Can be computationally expensive.** Computing $n_B = 10000$ correlation coefficients or sample mean differences is easy work for modern computers. Some real world problems require more computationally demanding models, and nesting these models in a bootstrap procedure may generate extremely long runtimes.

Pros and cons of bootstrap methods

Cons

- **Can be computationally expensive.** Computing $n_B = 10000$ correlation coefficients or sample mean differences is easy work for modern computers. Some real world problems require more computationally demanding models, and nesting these models in a bootstrap procedure may generate extremely long runtimes.
- **May lose statistical power.** In the rare instances when normality assumptions actually hold, the traditional methods may be more powerful. I.e., they may be able to detect smaller differences in means or build tighter confidence intervals than bootstrap methods could achieve with the same sample sizes and confidence levels.

Bootstrapping in ATARI

- One of the main projects of ATARI is to fit models to experimental nuclear resonance data. Given experimental data, can we accurately identify the locations (in energy) of the resonances, and can we specify the parameters which describe their amplitudes and widths? These are the questions the model builders seek to answer.

Bootstrapping in ATARI

- One of the main projects of ATARI is to fit models to experimental nuclear resonance data. Given experimental data, can we accurately identify the locations (in energy) of the resonances, and can we specify the parameters which describe their amplitudes and widths? These are the questions the model builders seek to answer.
- An additional problem is to quantify the uncertainty of the model prediction at each energy level. What is the variance of the cross section estimate of the model at a given energy level? This is where bootstrapping methods may be useful.

Bootstrapping in ATARI

- One of the main projects of ATARI is to fit models to experimental nuclear resonance data. Given experimental data, can we accurately identify the locations (in energy) of the resonances, and can we specify the parameters which describe their amplitudes and widths? These are the questions the model builders seek to answer.
- An additional problem is to quantify the uncertainty of the model prediction at each energy level. What is the variance of the cross section estimate of the model at a given energy level? This is where bootstrapping methods may be useful.
- Additionally, we will use Noah's synthetic data generation to validate the bootstrapping approach.

Bootstrapping in ATARI

INNER STEP: Fit the model ATARI team is exploring different approaches.

Bootstrapping in ATARI

MIDDLE STEP: Uncertainty quantification. For some sufficiently large number of times, use bootstrap resampling procedure to quantify uncertainty.

INNER STEP: Fit the model ATARI team is exploring different approaches.

End MIDDLE STEP.

Bootstrapping in ATARI

OUTER STEP: Validate the uncertainty quantification. For some sufficiently large number of times, feed a unique synthetic data set into middle and inner steps. Evaluate the accuracy of the uncertainty quantification.

MIDDLE STEP: Uncertainty quantification. For some sufficiently large number of times, use bootstrap resampling procedure to quantify uncertainty.

INNER STEP: Fit the model ATARI team is exploring different approaches.

End MIDDLE STEP.

End OUTER STEP.

Wrap up

Today we

- Explored data about the show *The Office*
- Dabbled in textual analysis
- Compared traditional vs. bootstrap methods for hypothesis testing and standard error calculation
- Considered the pros and cons of bootstrap procedures
- Discussed how bootstrap methods will be deployed in the ATARI problem set

Questions

What questions do you have?