Sta 111 - Summer II 2017
Probability and Statistical Inference

10. Bootstrap intervals

Lu Wang

Duke University, Department of Statistical Science

July 18, 2017

Outline

1. Bootstrapping = sampling with replacement from the observed sample

2. Bootstrap percentile intervals: middle XX% of the bootstrap distribution

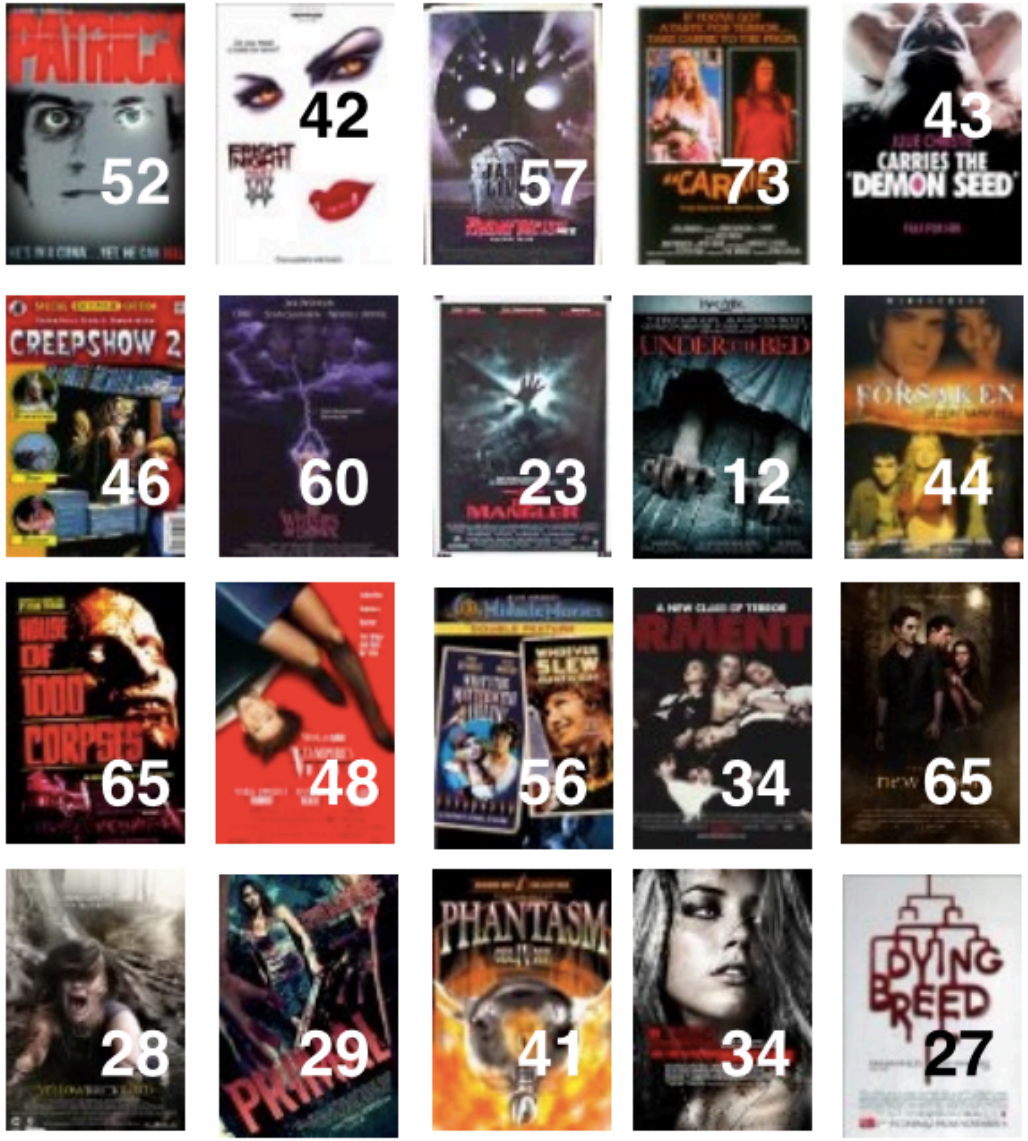
3. Bootstrap testing for a single statistic requires shifting the data to ensure that the null hypothesis is true

4. Summary

Rotten horrors

Rotten Tomatoes®

is a movie aggregator, where the audience is also able to review and score the movies. We want to estimate the median audience score of horror movies on RottenTomatoes.com. We start with a random sample of 20 horror movies.

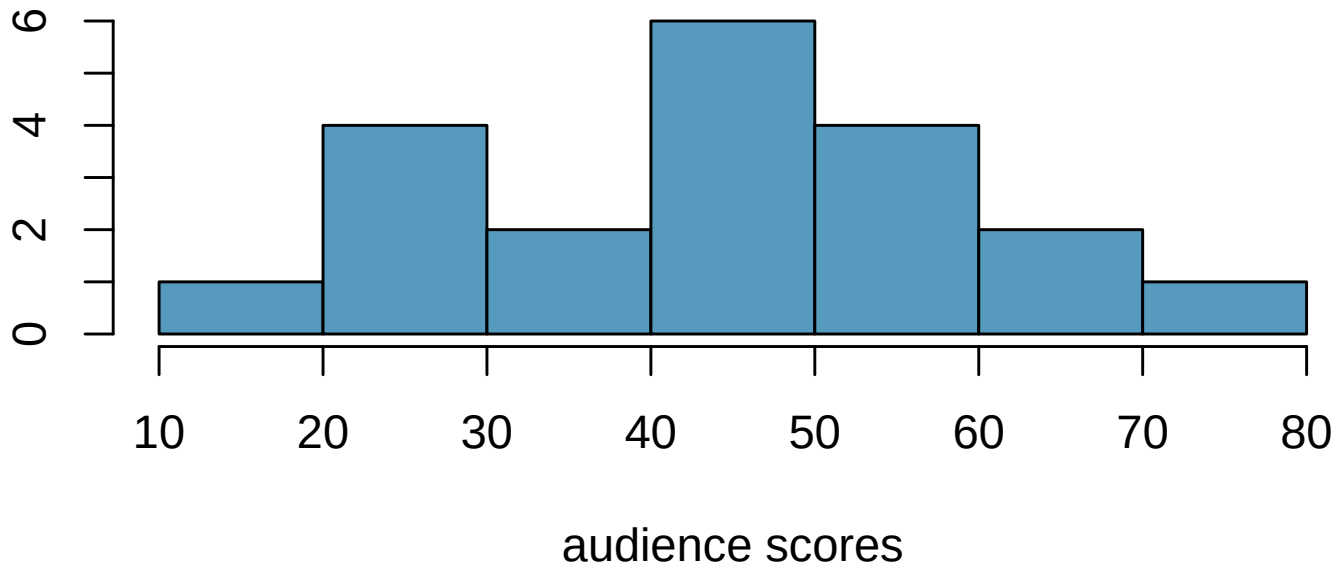


Data

	title	audience_score
1	Patrick	52
2	Demon Seed	43
3	Tormented	34
4	Under the Bed	12
5	Phantasm IV: Oblivion	41
6	Fright Night Part 2	42
7	House of 1000 Corpses	65
8	Creepshow 2	46
9	The Forsaken	44
10	All the Boys Love Mandy Lane	34
11	Jason Lives: Friday the 13th Part VI	57
12	Vampire's Kiss	48
13	The Witches of Eastwick	60
14	Yellowbrickroad	28
15	Dying Breed	27
16	Carrie	73
17	Whoever Slew Auntie Roo?	56
18	The Mangler	23
19	Primal	29
20	The Twilight Saga: New Moon	65

First look

The histogram below shows the distribution of the audience scores of these movies (ranging from 0 to 100). The median score in the sample is 43.5. Can we apply CLT based methods we have learned so far to construct a confidence interval for the median RottenTomatoes score of horror movies. Why or why not?



Bootstrapping

- An alternative approach to constructing confidence intervals is *bootstrapping*.
- This term comes from the phrase “pulling oneself up by one’s bootstraps”, which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the ~~im~~possible task is estimating a population parameter, and we’ll accomplish it using data from only the given sample.

Bootstrapping

- Bootstrapping works as follows:
  (1) take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample

  (2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples

  (3) repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

- The XX% bootstrap confidence interval can be estimated by the cutoff values for the middle XX% of the bootstrap distribution

Bootstrap sample 1

(1) Take a bootstrap sample:

	title	audience_score
1	Vampire's Kiss	48
2	Phantasm IV: Oblivion	41
3	House of 1000 Corpses	65
4	Dying Breed	27
5	Whoever Slew Auntie Roo?	56
6	The Forsaken	44
7	The Twilight Saga: New Moon	65
8	The Twilight Saga: New Moon	65
9	Whoever Slew Auntie Roo?	56
10	The Twilight Saga: New Moon	65
11	The Mangler	23
12	Dying Breed	27
13	Creepshow 2	46
14	House of 1000 Corpses	65
15	Whoever Slew Auntie Roo?	56
16	Tormented	34
17	Jason Lives: Friday the 13th Part VI	57
18	Vampire's Kiss	48
19	Primal	29
20	The Witches of Eastwick	60

(2) Calculate the median of the bootstrap sample:

23, 27, 27, 29, 34, 41, 44, 46, 48, *48, 56*, 56, 56, 57, 60, 65, 65, 65, 65, 65

median = (48 + 56) / 2 = 52

(3) Record this value

Bootstrap sample 2

(1) Take another bootstrap sample:

	title	audience_score
1	Fright Night Part 2	42
2	Carrie	73
3	The Forsaken	44
4	The Mangler	23
5	Primal	29
6	Patrick	52
7	Jason Lives: Friday the 13th Part VI	57
8	The Mangler	23
9	Vampire's Kiss	48
10	All the Boys Love Mandy Lane	34
11	The Twilight Saga: New Moon	65
12	All the Boys Love Mandy Lane	34
13	Yellowbrickroad	28
14	Vampire's Kiss	48
15	Tormented	34
16	The Mangler	23
17	Phantasm IV: Oblivion	41
18	Patrick	52
19	House of 1000 Corpses	65
20	The Twilight Saga: New Moon	65

(2) Calculate the median of the bootstrap sample:

23, 23, 23, 28, 29, 34, 34, 34, 41, *42, 44*, 48, 48, 52, 52, 57, 65, 65, 65, 73

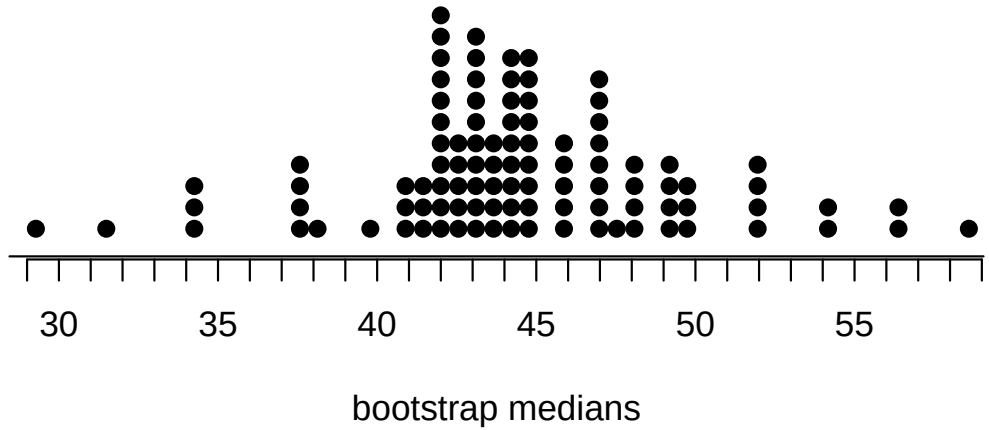
median = (42 + 44) / 2 = 43

(3) Record this value

Many more bootstrap samples

... repeat

The dot plot below is the bootstrap distribution of medians constructed using 100 simulations. What does each dot on the dot plot represent?



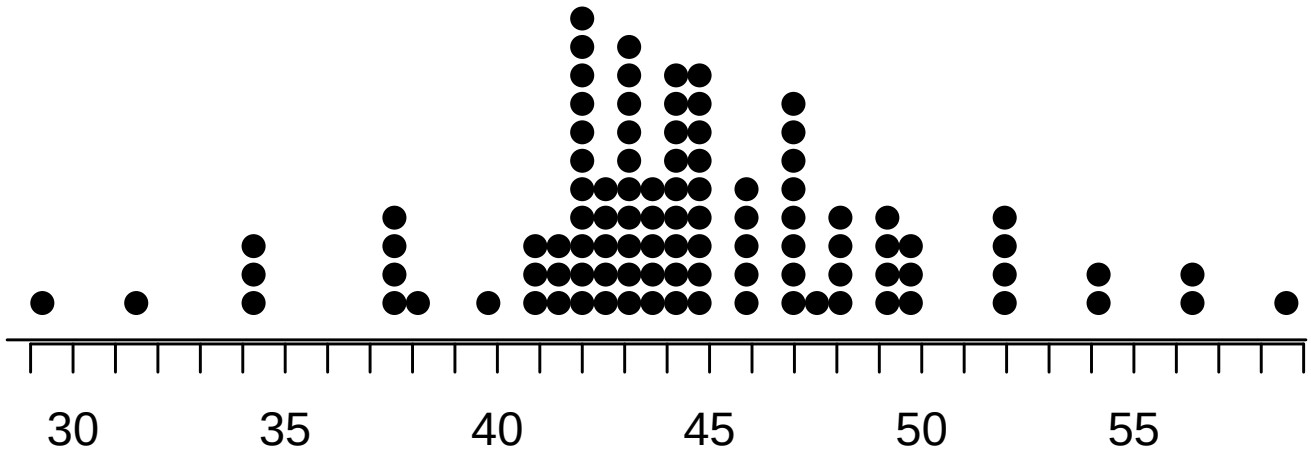
(a) Score of a horror movie in the original sample

(b) Score of a horror movie in the population

(c) Median from one bootstrap sample from the original sample

(d) Median from one sample from the population

The dot plot below shows the distribution of 100 bootstrap medians. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the percentile method.



(a) (29, 58.5)

(b) (34, 57)

(c) (37.5, 52)

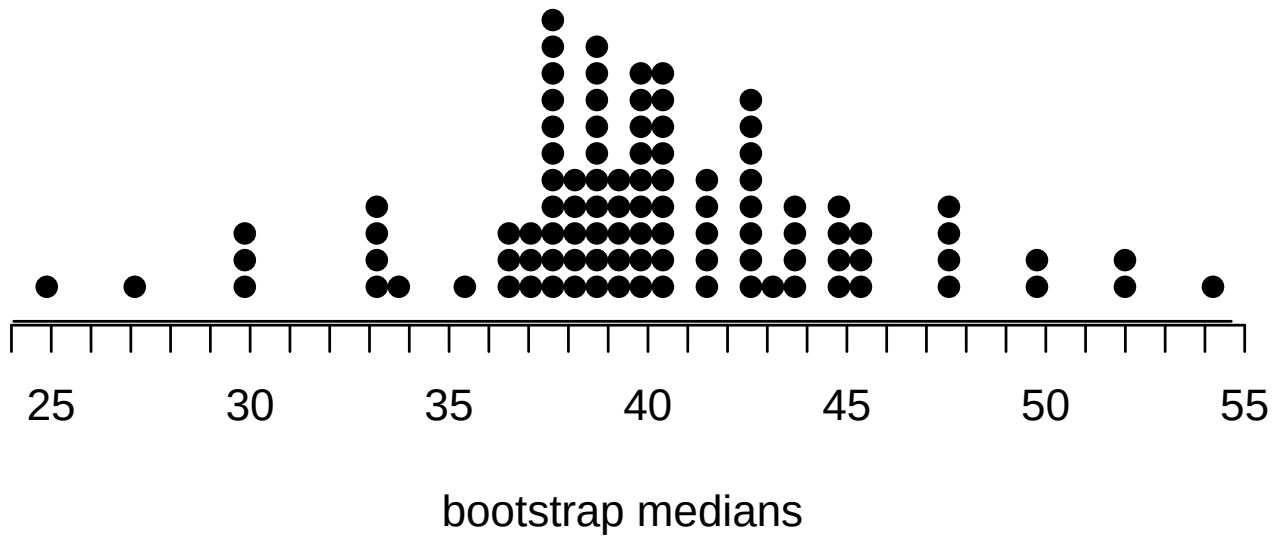
(d) (40, 49.5)

Bootstrap testing for a median

- This is very similar to bootstrapping, i.e. we randomly sample with replacement from our data, but this time we need to first shift the data to ensure that the null hypothesis is true.
  - To get p-value, we need to know the sampling distribution of our test statistic when the null hypothesis is true.
  - Normal distribution might not be a good approximation here. We instead perform a simulation under conditions in which we know the null hypothesis is true.
  - So we use our data to represent the population, but first we shift it over to ensure that the null hypothesis is true.
- The p-value is then defined as the proportion of simulations that yield a sample statistic at least as favorable to the alternative hypothesis as the *original observed* sample statistic.

Back to movie scores

The median of the original sample is 43.5. Do the data provide convincing evidence that the median audience score of horror movies is greater than 40? The dot plot below is the bootstrap distribution of medians constructed using 100 draws from the shifted data (the median is shifted to be 40).



H_0 : *median* $= 40$

H_A : *median* > 40

p-value: proportion of simulations where the simulated bootstrap sample median is at least as extreme as the one observed (43.5). $\rightarrow$ 20 / 100 = 0.20

Summary of Bootstrap method

1. Bootstrapping = sampling with replacement from the observed sample

2. Bootstrap percentile intervals: middle XX% of the bootstrap distribution

3. Bootstrap testing for a single statistic requires shifting the data to ensure that the null hypothesis is true