Unit 4: Inference for numerical data

1. Bootstrap intervals

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

February 23, 2015

1. Housekeeping

2. Bootstrapping

3. Randomization testing for a single numerical variable

4. Bootstrapping for categorical data

Announcements



1. Housekeeping

2. Bootstrapping

Randomization testing for a single numerical variable

Bootstrapping for categorical data

Rotten horrors

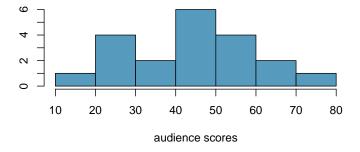


is a movie aggregator, where the audience is also able to review and score the movies. We want to estimate the average audience score of horror movies on RottenTomatoes.com. We start with a random sample of 20 horror movies.



	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

The dot plot 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?



► An alternative approach to constructing confidence intervals is *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.

- ► 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 impossible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.



▶ Bootstrapping works as follows:

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

- Bootstrapping works as follows:
 - 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

- Bootstrapping works as follows:
 - 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

- Bootstrapping works as follows:
 - 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

- Bootstrapping works as follows:
 - 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,

- 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,

OR

- 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,

OR

 $-\bar{x}_{boot} \pm z^{\star} SE_{boot}$

		1.1
	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

```
title audience score
 1
                          Vampire's Kiss
                  Phantasm IV: Oblivion
 2
                                                      41
 3
                  House of 1000 Corpses
                             Dying Breed
                                                      27
               Whoever Slew Auntie Roo?
                                                      56
 6
                            The Forsaken
                                                      44
            The Twilight Saga: New Moon
                                                      65
            The Twilight Saga: New Moon
                                                      65
               Whoever Slew Auntie Roo?
                                                      56
10
            The Twilight Saga: New Moon
                                                      65
11
                             The Mangler
                                                      23
                             Dving Breed
12
                                                      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:

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

```
title audience score
 1
                          Vampire's Kiss
 2
                   Phantasm IV: Oblivion
                                                       41
 3
                  House of 1000 Corpses
                                                       27
                             Dying Breed
               Whoever Slew Auntie Roo?
                                                       56
 6
                            The Forsaken
                                                       44
            The Twilight Saga: New Moon
                                                       65
            The Twilight Saga: New Moon
                                                       65
 9
               Whoever Slew Auntie Roo?
                                                       56
10
            The Twilight Saga: New Moon
                                                       65
                             The Mangler
                                                       23
11
12
                             Dving Breed
13
                             Creepshow 2
                                                       46
14
                  House of 1000 Corpses
                                                       65
                Whoever Slew Auntie Roo?
15
                                                       56
16
                               Tormented
                                                       34
17 Jason Lives: Friday the 13th Part VI
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 median = (48 + 56) / 2 = 52

(3) Record this value

	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
LO	All the Boys Love Mandy Lane	34
Ι1	The Twilight Saga: New Moon	65
12	All the Boys Love Mandy Lane	34
L3	Yellowbrickroad	28
۱4	Vampire's Kiss	48
۱5	Tormented	34
16	The Mangler	23
۱7	Phantasm IV: Oblivion	41
18	Patrick	52
19	House of 1000 Corpses	65
20	The Twilight Saga: New Moon	65

```
title audience score
1
                    Fright Night Part 2
2
                                  Carrie
                                                      73
 3
                            The Forsaken
                                                      44
                                                      23
                             The Mangler
5
                                  Primal
                                                      29
                                 Patrick
                                                      52
  Jason Lives: Friday the 13th Part VI
                                                      57
                                                      23
                             The Mangler
9
                         Vampire's Kiss
                                                      48
10
          All the Boys Love Mandy Lane
                                                      34
11
           The Twilight Saga: New Moon
                                                      65
           All the Boys Love Mandy Lane
12
13
                        Yellowbrickroad
                                                      28
14
                         Vampire's Kiss
                                                      48
15
                               Tormented
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:

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

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

(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

	title	audience_score
1	Tormented	34
2	The Witches of Eastwick	60
3	The Witches of Eastwick	60
4	The Witches of Eastwick	60
5	The Mangler	23
6	The Witches of Eastwick	60
7	Patrick	52
8	Phantasm IV: Oblivion	41
9	Yellowbrickroad	28
10	Jason Lives: Friday the 13th Part VI $$	57
11	Yellowbrickroad	28
12	Jason Lives: Friday the 13th Part VI	57
13	Fright Night Part 2	42
14	Primal	29
15	Fright Night Part 2	42
16	Whoever Slew Auntie Roo?	56
17	Fright Night Part 2	42
18	Fright Night Part 2	42
19	Under the Bed	12
20	Phantasm IV: Oblivion	41

```
title audience score
                               Tormented
1
                The Witches of Eastwick
                                                      60
                The Witches of Eastwick
                                                      60
                The Witches of Eastwick
                             The Mangler
6
                The Witches of Eastwick
                                                      60
                                 Patrick
                                                      52
8
                  Phantasm IV: Oblivion
                                                     41
                        Yellowbrickroad
                                                      28
10 Jason Lives: Friday the 13th Part VI
                                                     57
11
                        Yellowbrickroad
                                                      28
12 Jason Lives: Friday the 13th Part VI
13
                    Fright Night Part 2
14
                                  Primal
15
                    Fright Night Part 2
                                                     42
16
               Whoever Slew Auntie Roo?
17
                    Fright Night Part 2
18
                    Fright Night Part 2
                                                     42
19
                           Under the Red
                                                     12
20
                  Phantasm IV: Oblivion
                                                      41
```

(2) Calculate the median of the bootstrap sample:

```
title audience score
                               Tormented
1
                The Witches of Eastwick
                                                       60
                The Witches of Eastwick
                                                       60
                The Witches of Eastwick
                                                       60
                             The Mangler
 6
                The Witches of Eastwick
                                                       60
                                 Patrick
                                                      52
8
                  Phantasm IV: Oblivion
                                                      41
                         Yellowbrickroad
                                                      28
10 Jason Lives: Friday the 13th Part VI
                                                      57
11
                         Yellowbrickroad
                                                      28
12 Jason Lives: Friday the 13th Part VI
13
                    Fright Night Part 2
                                                      42
14
                                  Primal
15
                    Fright Night Part 2
               Whoever Slew Auntie Roo?
16
17
                    Fright Night Part 2
                    Fright Night Part 2
18
                                                      42
                           Under the Red
                                                      12
19
                                                      41
20
                  Phantasm IV: Oblivion
```

(2) Calculate the median of the bootstrap sample:

12, 23, 28, 28, 29, 34, 41, 41, 42, 42, 42, 42, 52, 56, 57, 57, 60, 60, 60 median = (42 + 42) / 2 = 42

```
title audience score
 1
                               Tormented
                 The Witches of Eastwick
                                                       60
                 The Witches of Eastwick
                                                       60
                The Witches of Eastwick
                                                       60
                             The Mangler
 6
                 The Witches of Eastwick
                                                       60
                                 Patrick
                                                       52
 8
                   Phantasm IV: Oblivion
                                                       41
                         Yellowbrickroad
                                                       28
10 Jason Lives: Friday the 13th Part VI
                                                       57
11
                         Yellowbrickroad
                                                       28
12 Jason Lives: Friday the 13th Part VI
13
                    Fright Night Part 2
                                                       42
14
                                   Primal
15
                     Fright Night Part 2
               Whoever Slew Auntie Roo?
16
17
                     Fright Night Part 2
                     Fright Night Part 2
18
                                                       42
                           Under the Red
                                                       12
19
20
                   Phantasm IV: Oblivion
                                                       41
```

(2) Calculate the median of the bootstrap sample:

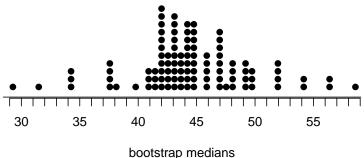
12, 23, 28, 28, 29, 34, 41, 41, 42, 42, 42, 42, 52, 56, 57, 57, 60, 60, 60 median = (42 + 42) / 2 = 42

(3) Record this value

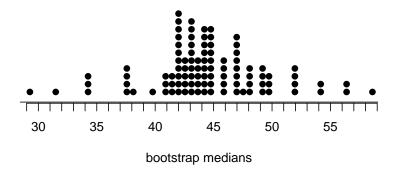
Many more bootstrap samples

... repeat

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



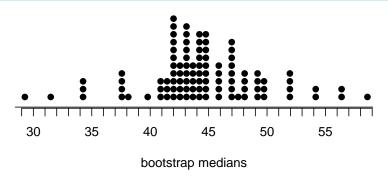
The dot plot below is the bootstrap distribution of medians constructed using 100 simulations. What does each dot on the dot plot represent?



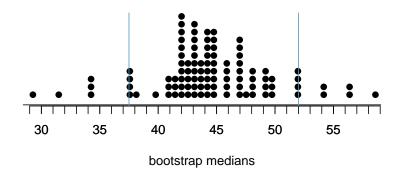
Each dot represents a bootstrap median from one simulation.

Bootstrap interval, percentile

The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median RottenTomatoes score of horror movies.

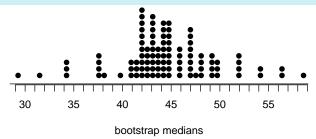


The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median RottenTomatoes score of horror movies.

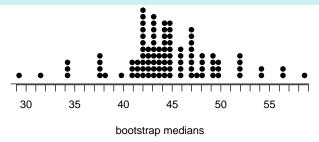


Botstrap interval, standard error

The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Given that the mean of the bootstrap distribution is 44.39, and the bootstrap standard error is 4.88, estimate the 90% bootstrap confidence interval for the median RottenTomatoes score of horror movies.



The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Given that the mean of the bootstrap distribution is 44.39, and the bootstrap standard error is 4.88, estimate the 90% bootstrap confidence interval for the median RottenTomatoes score of horror movies.



$$44.39 \pm (1.65 \times 4.88) = (36.34, 52.44)$$

Bootstrap vs. sampling distributions

Application exercise: 4.1 Bootstrap intervals

See the course webpage for details.

1. Housekeeping

2. Bootstrapping

3. Randomization testing for a single numerical variable

Bootstrapping for categorical data

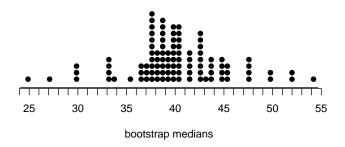
Randomization testing for a mean

➤ This is very similar to bootstrapping, i.e. we randomly sample with replacement from the sample, but this time we shift the bootstrap distribution to be centered at the null value.

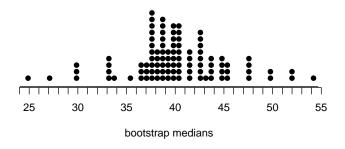
Randomization testing for a mean

- ➤ This is very similar to bootstrapping, i.e. we randomly sample with replacement from the sample, but this time we shift the bootstrap distribution to be centered at the null value.
- ➤ 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 observed sample statistic.

Do these data provide convincing evidence that the median audience score of horror movies is greater than 40?

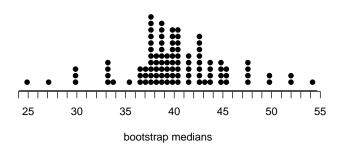


Do these data provide convincing evidence that the median audience score of horror movies is greater than 40?



 H_0 : median = 40 H_A : median > 40

Do these data provide convincing evidence that the median audience score of horror movies is greater than 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). $\rightarrow 20 / 100 = 0.20$

1. Housekeeping

2. Bootstrapping

Randomization testing for a single numerical variable

4. Bootstrapping for categorical data

Describe how you would construct a bootstrap interval for a proportion.