Unit 4: Inference for numerical data

1. Decision errors, significance levels, sample size & power

Sta 104 - Summer 2015

Duke University, Department of Statistical Science

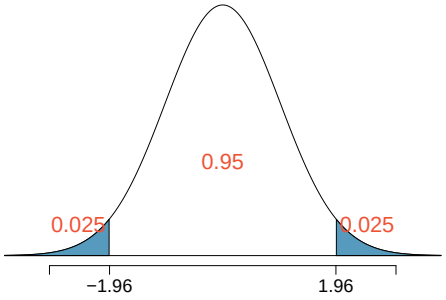
June 1, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bit.ly/sta104su15

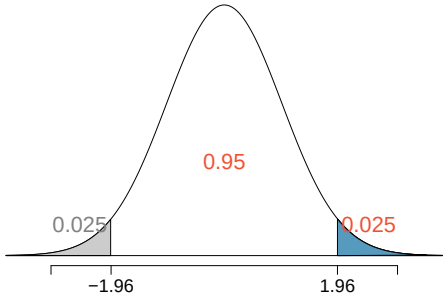
- PS3 due tonight
- Project proposals due Thursday night
- MT corrections extra credit: Work **as a team** to write up a collective exam corrections document that discusses all questions missed by any member of the team. Your corrections should show full work and explain reasoning, even for the multiple choice questions. Due by the end of the day on Wednesday, June 3. **Extra credit:** +2 points on the exam.

Two sided



95% confidence level is equivalent to two sided HT with $\alpha = 0.05$

One sided



95% confidence level is equivalent to one sided HT with $\alpha = 0.025$

Clicker question

What is the significance level of a two-sided hypothesis test that is equivalent to a 90% confidence interval? *Hint: Draw a picture and mark the confidence level in the center.*

(a) 0.001
(b) 0.01
(c) 0.025
(d) 0.05
(e) 0.10

Clicker question

What is the significance level of a one-sided hypothesis test that is equivalent to a 90% confidence interval? *Hint: Draw a picture and mark the confidence level in the center.*

(a) 0.001
(b) 0.01
(c) 0.025
(d) 0.05
(e) 0.10

Clicker question

What is the confidence level of a confidence interval that is equivalent to a two-sided hypothesis test with $\alpha = 0.01$. *Hint: Draw a picture and mark the confidence level in the center.*

(a) 0.80
(b) 0.90
(c) 0.95
(d) 0.98
(e) 0.99

Clicker question

What is the confidence level of a confidence interval that is equivalent to a one-sided hypothesis test with $\alpha = 0.01$. *Hint: Draw a picture and mark the confidence level in the center.*

(a) 0.80
(b) 0.90
(c) 0.95
(d) 0.98
(e) 0.99

Clicker question

A 95% confidence interval for the average normal body temperature of humans is found to be (98.1 F, 98.4 F). Which of the following is true?

(a) The hypothesis $H_0 : \mu = 98.2$ would be rejected at $\alpha = 0.05$ in favor of $H_A : \mu \neq 98.2$.
(b) The hypothesis $H_0 : \mu = 98.2$ would be rejected at $\alpha = 0.025$ in favor of $H_A : \mu > 98.2$.
(c) The hypothesis $H_0 : \mu = 98$ would be rejected using a 90% confidence interval.
(d) The hypothesis $H_0 : \mu = 98.2$ would be rejected using a 99% confidence interval.

2. Results that are statistically significant are not necessarily practically significant

Clicker question

All else held equal, will p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$
(b) $n = 10,000$

3. Calculate the sample size *a priori* to achieve desired margin of error

Application exercise: 4.1 Sample size

See course website for details.

4. Hypothesis tests are prone to decision errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true: α
  - For those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times
  - Increasing α increases the Type 1 error rate, hence we prefer to small values of α
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true: β
- *Power* is the probability of correctly rejecting H_0, and hence the complement of the probability of a Type 2 Error: $1 - \beta$

5. Power depends on the n, a, α, effect size

Power can be increased (and hence Type 2 error rate can be decreased) by

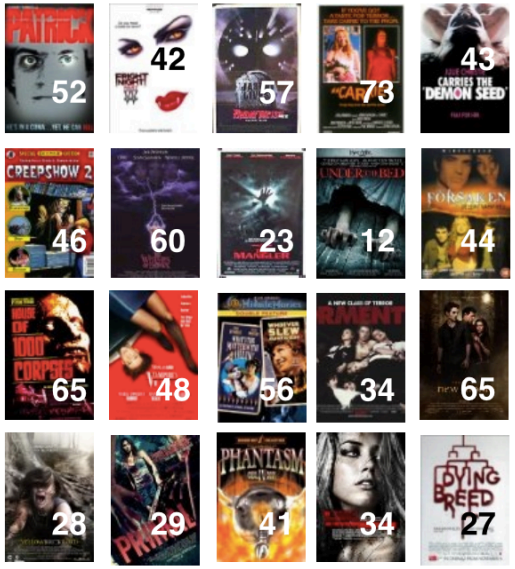
- increasing the sample size
- decreasing the standard deviation of the sample (difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help)
- increasing α
- increasing the *effect size*

Summary of main ideas

1. Hypothesis tests and confidence intervals at equivalent significance/confidence levels should agree
2. Results that are statistically significant are not necessarily practically significant
3. Calculate the sample size a priori to achieve desired margin of error
4. Hypothesis tests are prone to decision errors
5. Power depends on the effect size, α, n, and s

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 average audience score of horror movies on RottenTomatoes.com. We start with a random sample of 20 horror movies.

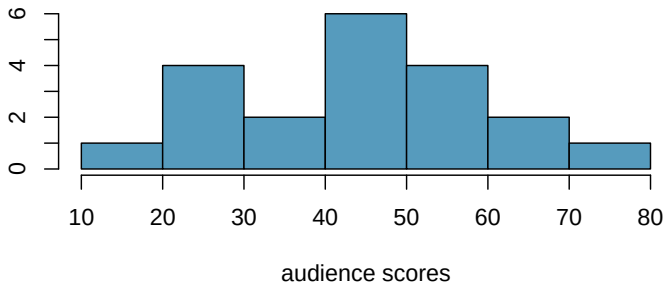


Data

	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?



- 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 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

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

(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

(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

Bootstrap sample 3

(1) Take another bootstrap sample:

	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

(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, 60
median = (42 + 42) / 2 = 42

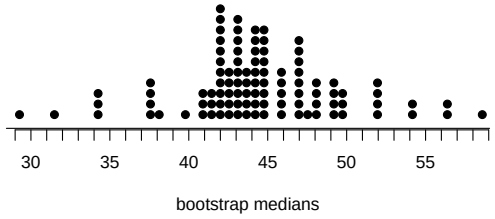
(3) Record this value

Many more bootstrap samples

... repeat

Clicker question

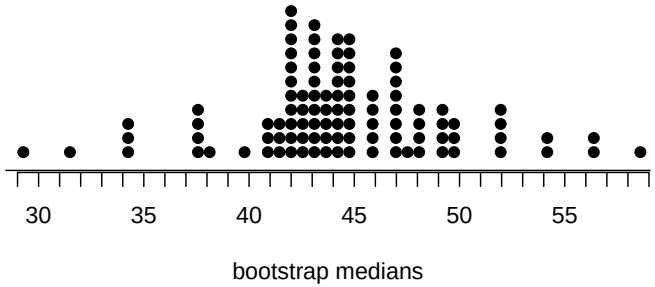
The dot plot below is the bootstrap distribution of medians constructed using 100 simulations. What does each dot on the dot plot represent?



(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

Clicker question

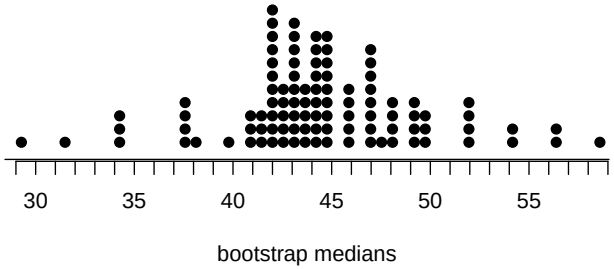
The dot plot below shows the distribution of 100 bootstrap medians. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the percentile method.



(a) (29, 58.5)
(b) (34, 57)
(c) (37.5, 52)
(d) (40, 49.5)

Botstrap interval, standard error

The dot plot below shows the distribution of 100 bootstrap medians. The median of the original sample is 43.5 and the bootstrap standard error is 4.88. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the standard error method.



Bootstrap vs. sampling distributions

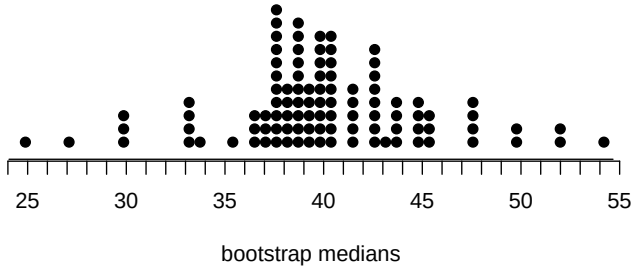
Application exercise: 4.2 Bootstrap intervals

See the course webpage for details.

Bootstrap 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? Remember that the median of the original sample was 43.5.



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

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