Lecture 15: Confidence Intervals: Two Groups

- Statements on confidence intervals
- Two groups
- Bootstrap intervals

Introduction

– Last time we talked about confidence intervals for one group problems.

– Today, we will start with reviewing some *statements about confidence intervals*[1], to gain the better understanding on it.

– We will also introduce confidence intervals for two groups problems (e.g. difference in mean or proportion).

– Lastly, we will briefly introduce the bootstrap method.

[1] From STA210 (Fall 2017) by Prof. Reiter: review problems on confidence intervals

Statements about Confidence Interval

Decide whether the following statements are true or false. Explain your reasoning.

– For a given standard error, lower confidence levels produce wider confidence intervals.
False. To get higher confidence, we need to make the interval wider interval. This is evident in the multiplier, which increases with confidence level.

– The statement, “the 95% confidence interval for the population mean is (350, 400)", is equivalent to the statement, “there is a 95% probability that the population mean is between 350 and 400".
False. 95% confidence means that we used a procedure that works 95% of the time to get this interval. That is, 95% of all intervals produced by the procedure will contain their corresponding parameters. For any one particular interval, the true population average is either inside the interval or outside the interval. In this case, it is either in between 350 and 400, or it is not in between 350 and 400. Hence, the probability that the population average is in between those two exact numbers is either zero or one.

– If you increase sample size, the width of confidence intervals will increase.
False. Increasing the sample size decreases the width of confidence intervals, because it decreases the standard error.

– To reduce the width of a confidence interval by a factor of two (i.e., in half), you have to quadruple the sample size.
True, as long as we're talking about a CI for a population percentage. The standard error for a population percentage has the square root of the sample size in the denominator. Hence, increasing the sample size by a factor of 4 (i.e., multiplying it by 4) is equivalent to multiplying the standard error by 1/2. Hence, the interval will be half as wide. This also works approximately for population averages as long as the multiplier from the t-curve doesn't change much when increasing the sample size (which it won't if the original sample size is large).

– The statement, "the 95% confidence interval for the population mean is (350, 400)" means that 95% of the population values are between 350 and 400.
False. The confidence interval is a range of plausible values for the population average. It does not provide a range for 95% of the data values from the population. To find the percentage of values in the population between 350 and 400, we need to look at a histogram of the data values and determine what percentage of observations are between 350 and 400.

– If you take large random samples over and over again from the same population, and make 95% confidence intervals for the population average, about 95% of the intervals should contain the population average.
True. This is the definition of confidence intervals.

– If you take large random samples over and over again from the same population, and make 95% confidence intervals for the population average, about 95% of the intervals should contain the sample average.
False. The confidence interval is a range for the population average, not for the sample average. In fact, every confidence interval contains its corresponding sample average, because CIs are of the form: sample avg. +/- multiplier SE. So, the sample average is right in the middle of the CI.

– It is necessary that the distribution of the variable of interest follows a normal curve.
False. It is necessary that the distribution of the sample average follows a normal curve. The data values of the variable, however, need not follow a normal curve, because if the sample size is large enough the central limit theorem for the sample average will apply.

– A 95% confidence interval obtained from a random sample of 1000 people has a better chance of containing the population percentage than a 95% confidence interval obtained from a random sample of 500 people.
False. All 95% confidence intervals have the property that they come from a procedure that has a 95% chance of yielding an interval that contains the true value. The confidence interval method automatically accounts for sample size in the standard error. A 95% CI with n=1000 will be narrower than a 95% CI with n=500, but both CIs will have 95% confidence of containing the population percentage.

– If you make go through life making 99% confidence intervals for all sorts of population means, about 1% of the time the intervals won't cover their respective population means.
True. Since 99% of the intervals should contain the corresponding population mean, 1% of them will not.

Normal, known σ^2

Suppose we have observations from two groups

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma_1^2)$$
$$Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma_2^2)$$

and assume that $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ are independent. Suppose further that σ_1^2 and σ_2^2 are **known** and we want to find a confidence interval for the difference in means $\mu_1 - \mu_2$. By properties of Normals, we have

$$\overline{X}_n - \overline{Y}_m \sim \text{Normal}(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$$

The setup looks awkward, but it is actually the same as in the interval for μ with one group and a known variance σ^2 (**why?**), so a $(1-\alpha)$ confidence interval is

$$(\overline{X}_n - \overline{Y}_m) \pm z_{\alpha/2} \sqrt{\sigma_1^2/n + \sigma_2^2/m}$$

Normal, unknown σ^2

Suppose we have observations from two groups

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma^2)$$

$$Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma^2)$$

with $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ independent. The two groups have the same variance. The population variance σ^2 is **unknown** and estimated from the data using a weighted average of sample variances:

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

where s_1^2 is the sample variance in group 1 (the group with the Xs) and s_2^2 is the sample variance in group 2 (the group with the Ys). Some textbook called s^2 the pooled sample variance. A $(1-\alpha)$ confidence interval is

$$(\overline{X}_n - \overline{Y}_m) \pm t_{n+m-2,\alpha/2}\, s \sqrt{1/n + 1/m}$$

What can we do if the population variances of the groups are not equal?

- Well, if the sample sizes n and m are big enough, one option is (as usual) approximating σ_1^2 and σ_2^2 by s_1^2 and s_2^2 and reporting the interval

$$(\overline{X}_n - \overline{Y}_m) \pm z_{\alpha/2} \sqrt{\sigma_1^2/n + \sigma_2^2/m}$$

  as an approximate 95% confidence interval for $\mu_1 - \mu_2$.
- If the sample sizes aren't big enough, most statistical software packages have implementations of appropriate intervals. The formulas are awkward and we won't cover them here, but you should know that intervals exist if you ever need them.

Example 1: Deception in Dating

In 2007, a paper called Deception in Dating was released. The paper included a study of 40 men and 40 women who were using online dating websites. According to the paper, "On average, the men under-reported their weight by 1.94 pounds, with a standard deviation of 10.34*lb*. Women on average under-reported their weight by 8.48 lb, with a standard deviation of 8.78 lb."

Let's find a 95% confidence interval for the difference "actual − reported weight" between men and women (so this is a confidence interval for the average difference of differences! I hope this doesn't confuse you much...)

First, let's assume

- the outcome "actual - reported weight" follows a Normal distribution in both groups (men and women)
- the population variance in the two groups is the same
- all the observations are independent

The pooled variance is

$$s^2 = \frac{39 \times 10.34^2 + 39 * 8.87^2}{(39 + 39)} \approx 92.796$$

In case $t_{78,0.025}$ is not available in a Student-t table, use its closest value $t_{80,0.025} = 1.990$. Our confidence interval is

$$(1.94 - 8.48) \pm 1.990 \sqrt{92.796(2/40)} = [-10.826, -2.254]$$

Now let's find the approximate interval (using CLT) that doesn't assume that the population variances are the same:

$$(1.94 - 8.48) \pm 1.96 \sqrt{10.34^2/40 + 8.87^2/40} = [-10.762, -2.318]$$

As you can see, the interval are almost identical and the conclusions don't change much.

There is a similar problem in HW5.

Example2: Aluminium Cans

Two factories manufacture aluminium cans. We want to know whether the average weight of a can is different between them. So we want a 95% two-sided confidence interval on $\mu_1 - \mu_2$.

A sample of 100 cans from Factory 1 has mean 16g and sample standard deviation 1g. A sample of 64 cans from Factory 2 has mean 16.5g and sample standard deviation 2g.

By the CLT, we know that

$$\overline{X}_1 \dot{\sim} N(\mu_1, \frac{\sigma_1}{10}) \quad \text{and} \quad \overline{X}_2 \dot{\sim} N(\mu_2, \frac{\sigma_2}{8}).$$

From the properties of linear combinations of normal random variables,

$$Y = \bar{X_1} - \bar{X_2} \dot{\sim} N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{100} + \frac{\sigma_2^2}{64}}).$$

So a 95% confidence interval on the difference is equivalent to a 95% confidence interval on Y which is

$$\begin{aligned} L, U &= (\overline{X}_1 - \overline{X}_2) \pm z_C \sqrt{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)} \\ &= (16 - 16.5) \pm 1.96 \sqrt{\left(\frac{1}{100} + \frac{4}{64}\right)} \\ &= -0.5 \pm 0.5277. \end{aligned}$$

The 95% confidence interval is [-1.028, 0.028]. From this, we cannot be certain that the difference in means is not zero. The two factories may have the same average weight.

Proportions

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p_1)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Bernoulli}(p_2)$. As usual, assume that $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ are independent. Let $\widehat{p}_1 = \overline{X}_n$ and $\widehat{p}_2 = \overline{Y}_m$ be the sample proportions in the two groups. By CLT, we have

$$\widehat{p}_1 - \widehat{p}_2 \approx \text{Normal}(p_1 - p_2, p_1(1-p_1)/n + p_2(1-p_2)/m)$$

so an approximate $(1-\alpha)$ confidence interval for the difference $p_1 - p_2$ is

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2} \sqrt{\widehat{p}_1(1-\widehat{p}_1)/n + \widehat{p}_2(1-\widehat{p}_2)/m}$$

We could use the same idea for Poisson and find an interval for the difference of λs as needed.

Motivation

A lot of theoretical statistics has focused on developing methods for setting confidence intervals and testing hypotheses. A key tool for doing this is the **Central Limit Theorem**, which says that for large samples, the average is **approximately normally distributed**.

With some work, the CLT allows confidence intervals on the mean, the proportion, the sum, the difference of means, and the difference of proportions. But what can we do if we want to set confidence intervals on a correlation or an sd or a ratio?

Motivation

For many years, statisticians could not set confidence intervals on many parameters of interest without having to make **strong and often unrealistic assumptions** about the distribution from which the data were obtained.

For example, there is theory that tells how one can set a confidence interval on the sd, provided the data come from a normal distribution. But if one is interested on the sd of income in the U.S., we know from the histogram that there is **a very long right tail**. Income is not normally distributed, but economists still need to estimate the sd.

Similarly, there is theory on how to estimate confidence intervals for ratios, provided that both the numerator and the denominator are independent normal random variables. But for many applications, this is untrue—income per hour worked is an example.

Motivation

In 1979, Brad Efron invented the bootstrap. This is a computer-intensive procedure that substitutes fast computation for theoretical math.

The main benefit of the procedure is that it allows statisticians to set confidence intervals on parameters without having to make unreasonable assumptions. It was a revolution.

This was one of the first of many breakthroughs in computational statistics, which is the way that nearly all work is done now.

Bootstrapping

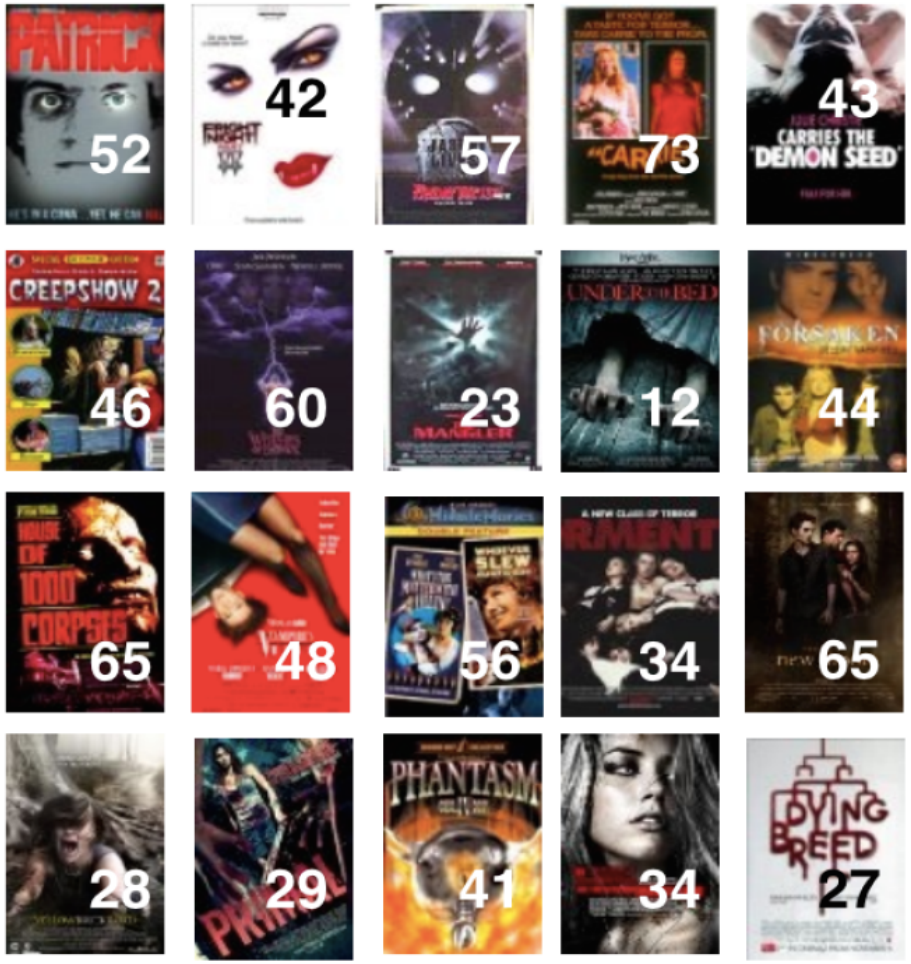
- This term comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any additional help
- In this case the impossible task is estimating a population parameter, and we'll accomplish it using data from only the given sample

Bootstrap Procedures

- 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 means, 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

Example: 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.

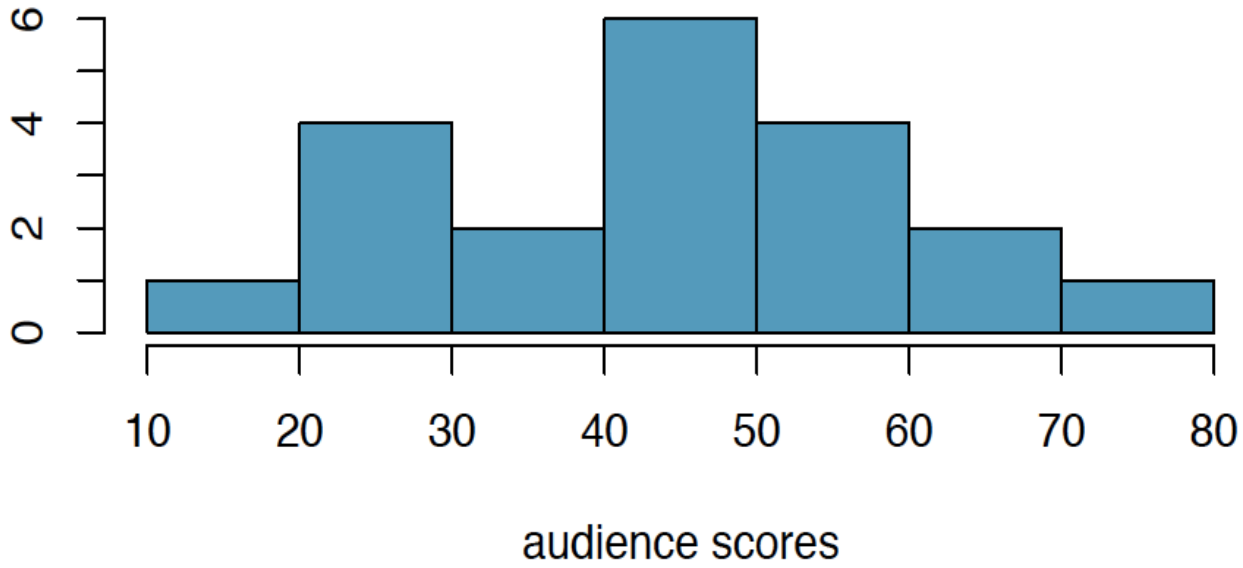


Data

	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?



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

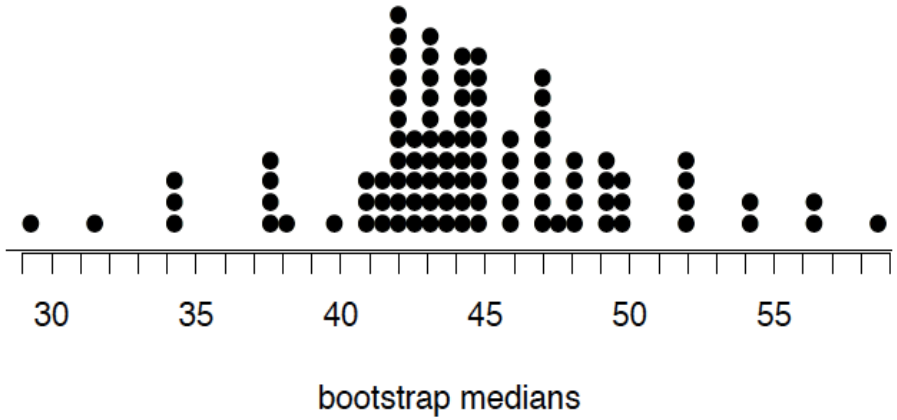
median = (42 + 44) / 2 = 43

(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?



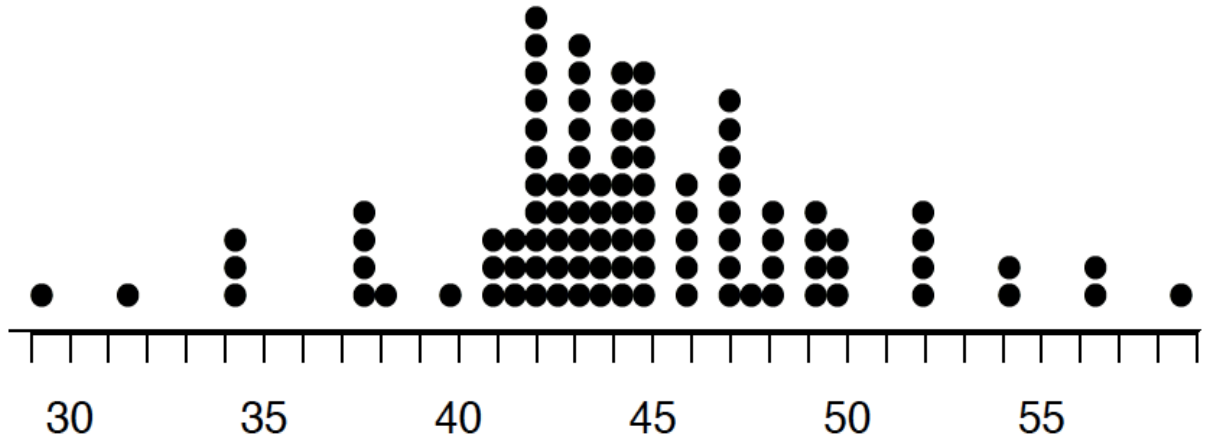
(a) Score of a horror movie in the original sample

(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)

How the Bootstrap Works

If one samples from a population without replacement and makes a histogram of the results, **then as the sample size increases, the histogram of results converges to the probability histogram for that population**.

Thus if one draws 10^7 people at random and makes a histogram of their incomes, one can use this to approximate, with pretty good accuracy, the probability that the next draw will be, say, a millionaire.

How the Bootstrap Works

Let n be the sample size, and suppose one observes a random sample $X_1, \ldots, X_n$. One can form the histogram of the data by putting rectangles of size 1/n at each of the X_i values.

As the sample size increases one can let the width of the rectangle go to zero, and in the limit, by the convergence, one gets the probability histogram.

Note that this ensures that the total area under the histogram is 1, as required. If two of the observations are identical, then one gets a bar of twice the height, suggesting that the observation is more common.

How the Bootstrap Works

Suppose we use a computer to draw 1000 bootstrap samples of size n. For each such sample, we can calculate a new estimate of the parameter of interest.

Rank these estimates from least to largest. We denote these ordered bootstrap estimates by

$$\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(1000)}$$

where the number in parentheses shows the order in terms of size. Thus $\hat{\theta}^*_{(1)}$ is the smallest estimate of the sd found in one of the 1000 bootstrap samples, and $\hat{\theta}^*_{(1000)}$ is the largest.

The spread in these bootstrap estimates tells us how large the effect of chance error is on the estimate $\hat{\theta}$ that we got in our original sample.

How the Bootstrap Works

Suppose we want to set a 95% confidence interval on θ, the true parameter value for the real population F. And suppose we take 1000 bootstrap samples. The bootstrap method suggests that about 95% of the time, the true parameter value for $\hat{F}_n$ falls the 25th largest observation to the 975th largest observation.

Since $\hat{F}_n$ converges to F, the correct confidence interval for the true parameter on $\hat{F}_n$ should converge to the correct confidence interval on the parameter for F.
This logic gives the 95% **percentile confidence interval**, or:

$$L = \hat{\theta}^*_{(.025)} \qquad U = \hat{\theta}^*_{(.975)}.$$

But this does not take full account of the difference between θ for F and $\hat{\theta}$, the true value for $\hat{F}_n$. We can do a bit better.

How the Bootstrap Works

The **pivot confidence interval** argues that the behavior of $\theta - \hat{\theta}$ is approximately the same as the behavior of $\hat{\theta} - \hat{\theta}^*$. Thus

$$\begin{aligned}
0.95 &\approx \mathrm{P}[\hat{\theta}^*_{(.025)} \leq \hat{\theta}^* \leq \hat{\theta}^*_{(.975)}] \\
&= \mathrm{P}[\hat{\theta}^*_{(.025)} - \hat{\theta} \leq \hat{\theta}^* - \hat{\theta} \leq \hat{\theta}^*_{(.975)} - \hat{\theta}] \\
&= \mathrm{P}[\hat{\theta} - \hat{\theta}^*_{(.025)} \geq \hat{\theta} - \hat{\theta}^* \geq \hat{\theta} - \hat{\theta}^*_{(.975)}] \\
&\approx \mathrm{P}[\hat{\theta} - \hat{\theta}^*_{(.025)} \geq \theta - \hat{\theta} \geq \hat{\theta} - \hat{\theta}^*_{(.975)}] \\
&= \mathrm{P}[2\hat{\theta} - \hat{\theta}^*_{.025} \geq \theta \geq 2\hat{\theta} - \hat{\theta}^*_{(.975)}]
\end{aligned}$$

So

$$L = 2\hat{\theta} - \hat{\theta}^*_{(.975)} \quad U = 2\hat{\theta} - \hat{\theta}^*_{(.025)}.$$

How the Bootstrap Works

$\hat{F}_n$ converges to F. It is not obvious, but one can show that this implies that the chance error in estimating θ for F converges to the chance error in estimating θ^* for $\hat{F}_n$.

In practice, one has to be able to draw many samples from the box model, and calculate an estimate for each. This can be time consuming, and for realistic examples on usually needs the computer.

Before the bootstrap, statisticians had to write all estimates as special kinds of averages and use the Central Limit Theorem to set approximate confidence intervals. But one can show that, as n gets large, the bootstrap is never worse than the Central Limit Theorem approximation and for many parameters it can be much better.

Example

Suppose one wants to estimate the sd in the number of hours that people work in a week. One draws a random sample of size 8, and finds

$$40, 35, 40, 0, 0, 40, 50, 10$$

The point estimate for the sd is easy. It is just the sd of the sample, or

$$\sqrt{\frac{1}{8}(40^2 + \ldots + 10^2) - 26.875^2} = 18.864.$$

Example

The bootstrap trick tells us how to put a confidence interval on this estimate.

Suppose we draw 500 bootstrap samples. We might get samples like the following:

Sample Number	Sample	Estimate
1	0, 40, 40, 10, 10, 10, 0, 0	15.762
2	50, 10, 0, 0, 0, 40, 40, 40	20.463
3	0, 10, 40, 35, 0, 0, 10, 0	15.398
4	40, 40, 40, 40, 40, 40, 40, 40	0
5	0, 0, 50, 50, 0, 0, 50, 50	25
etc.		

Note that the largest possible estimate is 25, and the smallest possible estimate is 0.

Example

Suppose we want to use the 500 bootstrap samples to form a 90% confidence interval on the true sd of the number of hours that people work. We shall need to find the 25th largest and the 475th largest values from the previous table, extended to have 500 samples.

Normally we would use a computer. But for tutelary purposes, suppose the 25th largest value was 14.28 and the 475th largest value was 21.62.

Then the percentile confidence interval is (14.28, 21.62). And the pivot confidence interval, which is better, is:

$$L = 2\hat{\theta} - \hat{\theta}^*_{(475)} = 2 * 18.864 - 21.62 = 16.108$$
$$U = 2\hat{\theta} - \hat{\theta}^*_{(25)} = 2 * 18.864 - 14.28 = 23.448.$$

Recap

Today we covered:

- Some common misinterpretations of confidence intervals
- We also covered some special confidence intervals (two group)
- We talked about the history, importance, and some uses of the bootstrap method.

We will learn even more about the bootstrap in the next lab.

Suggested reading:

- D.S. Sec. 8.5, 12.6
- OpenIntro3: Sec. 4.2