19: Sampling and the Bootstrap

Jerry Cain
May 9, 2022

Table of Contents

2 Sampling Definitions
11 Unbiased Estimators
23 Standard Error
29 Bootstrap: Sample Mean
40 Bootstrap: Sample Variance
57 Bootstrap: p-values
Sampling definitions
Motivating example

You want to know the true mean and variance of happiness in Bhutan.

- But you can’t ask everyone.
- You poll 200 random people.
- Your data looks like this:

\[
\text{Happiness} = \{72, 85, 79, 91, 68, \ldots, 71\}
\]

- The mean of all these numbers is 83.

Is this the true mean happiness of Bhutanese people?
Population

This is a population.
A sample is selected from a population.
A **sample** is selected from a population.
Reasonable Questions Starting Out

1. In situations where we can’t observe the entire population, what can we safely conclude by polling a sample drawn from that population?
2. How large does your sample need to be before your conclusions are trustworthy, and how do we express confidence with any conclusions we draw?
3. Are there alternative ways to infer population statistics without polling entire populations?
A sample, mathematically

Consider $n$ random variables $X_1, X_2, \ldots, X_n$.

The sequence $X_1, X_2, \ldots, X_n$ is a **sample** from distribution $F$ if:

- $X_i$ are all independent and identically distributed (i.i.d.)
- $X_i$ all have same distribution function $F$ (the **underlying distribution**), where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$
A sample, mathematically

A sample of size 8:
\((X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)\)

A realization of a sample of size 8:
\((59, 87, 94, 99, 87, 78, 69, 91)\)
If we had a distribution $F$ of our entire population, we could compute exact statistics about happiness.

But we only have 200 people (a sample).

Today: If we only have a single sample,

• How do we report estimated statistics?
  ◦ We’re careful to call them estimated mean and estimated variance, since they’re based on samples (i.e. experiments)
• How do we report estimated error of these estimates?
• How do we perform something called hypothesis testing?
Unbiased estimators
A single sample

If we had a distribution $F$ of our entire population, we could compute exact statistics about happiness.

But we only have 200 people (a sample).

These population-level statistics are unknown:

- $\mu$, the population mean
- $\sigma^2$, the population variance
A single sample

If we had a distribution $F$ of our entire population, we could compute exact statistics about happiness.

But we only have 200 people (a sample).

- From these 200 people, what is our best estimate of the population mean and the population variance?
- How exactly do we define best estimate?
Estimating the population mean

1. What is our best estimate of $\mu$, the mean happiness of Bhutanese people?

If we only have $(X_1, X_2, \ldots, X_n)$:

The best estimate of $\mu$ is the sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$\bar{X}$ is an unbiased estimator of the population mean $\mu$. $E[\bar{X}] = \mu$

Intuition: By the CLT, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ If we could take multiple samples of size $n$:

1. For each sample, compute sample mean
2. On average, we would get the population mean
Even if we can’t report $\mu$, we can report our sample mean 83.03, which is an unbiased estimate of $\mu$. 

\[
\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})
\]
Estimating the population variance

2. What is $\sigma^2$, the variance of happiness of Bhutanese people?

If we knew the entire population $(x_1, x_2, \ldots, x_N)$:

\[
\sigma^2 = E[(X - \mu)^2] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]

If we only have one sample: $(X_1, X_2, \ldots, X_n)$:

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
Intuition about the sample variance, $S^2$

Actual, $\sigma^2$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Population variance

Population mean

$x_i - \mu$

Calculating population statistics **exactly** requires us knowing all $N$ datapoints.
Intuition about the sample variance, $S^2$

**Actual, $\sigma^2$**

Population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Population mean

**Estimate, $S^2$**

Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Sample mean

---

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2022
Intuition about the sample variance, $S^2$

**Actual, $\sigma^2$**

Population variance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

**Estimate, $S^2$**

Sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Population size, $N$

Happiness

0

$\bar{X}$

$\mu$

150
Intuition about the sample variance, $S^2$

**Actual, $\sigma^2$**

Population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Population mean

**Estimate, $S^2$**

Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Sample mean

Sample variance is an estimate using an estimate, so it needs additional scaling.

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2022

Stanford University
Estimating the population variance

2. What is \( \sigma^2 \), the \textbf{variance of happiness} of Bhutanese people?

If we only have a sample, \((X_1, X_2, \ldots, X_n)\):

The best estimate of \( \sigma^2 \) is the \textbf{sample variance}:

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

\( S^2 \) is an \textbf{unbiased estimator} of the population variance, \( \sigma^2 \).

\[
E[S^2] = \sigma^2
\]
Proof that $S^2$ is unbiased  

$$E[S^2] = E \left[ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] \Rightarrow (n-1)E[S^2] = E \left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right]$$

$$(n-1)E[S^2] = E \left[ \sum_{i=1}^{n} ((X_i - \mu) + (\mu - \bar{X}))^2 \right] \quad \text{(introduce } \mu - \mu)$$

$$= E \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\mu - \bar{X})^2 + 2 \sum_{i=1}^{n} (X_i - \mu)(\mu - \bar{X}) \right]$$

$$= E \left[ \sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 - 2n(\mu - \bar{X})^2 \right]$$

$$= E \left[ \sum_{i=1}^{n} (X_i - \mu)^2 - n(\mu - \bar{X})^2 \right] = \sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2]$$

$$= n\sigma^2 - n\text{Var}(\bar{X}) = n\sigma^2 - n \frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n - 1)\sigma^2$$

Therefore $E[S^2] = \sigma^2$
Standard error
Estimating population statistics

1. Collect a sample, $X_1, X_2, ..., X_n$. 

2. Compute sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. 

3. Compute sample deviation, $X_i - \bar{X}$. 

4. Compute sample variance, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. 

How close are our estimates $\bar{X}$ and $S^2$?

A particular outcome

$(72, 85, 79, 79, 91, 68, \ldots, 71)$

$n = 200$

$\bar{X} = 83$

$S^2 = 793$
Sample mean

• $\text{Var}(\bar{X})$ is a measure of how close $\bar{X}$ is to $\mu$.
• **How do we estimate $\text{Var}(\bar{X})$?**
How close is our estimate $\bar{X}$ to $\mu$?

$$E[\bar{X}] = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

We want to estimate this

**def** The **standard error** of the mean is an estimate of the standard deviation of $\bar{X}$.

Intuition:

- $S^2$ is an unbiased estimate of $\sigma^2$
- $S^2/n$ is an unbiased estimate of $\sigma^2/n = \text{Var}(\bar{X})$
- $\sqrt{S^2/n}$ can estimate $\sqrt{\text{Var}(\bar{X})}$

More info on bias of standard error: [wikipedia](https://en.wikipedia.org/wiki/Standard_error)
Standard error of the mean

1. Mean happiness:

Claim: The average happiness of Bhutan is 83, with a standard error of 1.99.

Closed form: \( SE = \sqrt{\frac{S^2}{n}} \)

These 2 statistics give a sense of how the sample mean random variable \( \bar{X} \) behaves.
Standard error of variance?

1. Mean happiness:

Claim: The average happiness of Bhutan is 83, with a standard error of 1.99.

Closed form: \( SE = \frac{S^2}{n} \)

2. Variance of happiness:

Claim: The variance of happiness of Bhutan is 793.

Closed form: Not covered in CS109

But how close are we?

⚠️ this is our best estimate of \( \sigma^2 \)

Up next: Compute Statistics with code!
Bootstrap: Sample mean
Bootstrap

The Bootstrap:

Probability for Computer Scientists
Computing statistic of sample mean

What is the standard deviation of the sample mean $\bar{X}$? (sample size $n = 200$)

$$\sigma \sqrt{n} = 1.886$$

Exact statistic
(we don’t have this)

$1.869$

Simulated statistic
(we don’t have this)

$$SE = \frac{S}{\sqrt{n}} = 1.992$$

Estimated statistic, by formula, standard error

Simulated estimated statistic

Note: We don’t have access to the population. But Doris is sharing the exact statistic with you.
Bootstrap insight 1: Estimate the true distribution
Bootstrap insight 1: Estimate the true distribution

You can estimate the PMF of the underlying distribution, using your sample.*

\[ \approx \]

The underlying distribution \( F \) \( \approx \) \( \hat{F} \)

\( \approx \)

the sample distribution (aka the histogram of your data)

*This is just a histogram of your data!
Bootstrap insight 2: Simulate a distribution

Approximate the procedure of simulating a distribution of a statistic, e.g., $\bar{X}$.

Population distribution (we don’t have this) $\approx$ Sample distribution (we do have this)

Simulated distribution of sample means

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2022
Estimate the true PMF using our "PMF" (histogram) of our sample.

...generate a whole bunch of sample means of this estimated distribution...

...and compute the standard deviation of this distribution.

\[
\text{means} = [84.7, 83.9, 80.6, 79.8, 90.3, \ldots, 85.2]
\]

\[
\text{np.std(means)} = 2.003
\]
Computing statistic of sample mean

What is the standard deviation of the sample mean $\bar{X}$? (sample size $n = 200$)

Population distribution (we don’t have this)

$\frac{\sigma}{\sqrt{n}} = 1.886$

Exact statistic
(we don’t have this)

Sample distribution (we do have this)

$SE = \frac{S}{\sqrt{n}} = 1.992$

Estimated statistic, by formula, standard error

Simulated estimated statistic, bootstrapped standard error

$\mu = 1.869$

$\sigma = 1.886$

$S = 2.003$
Bootstrap algorithm

Bootstrap Algorithm (sample):
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the sample mean on the resample
3. You now have a distribution of your sample mean

What is the distribution of your sample mean?

We’ll talk about this algorithm in detail either today or Wednesday!
Bootstrap Algorithm (sample):
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample sample.size() from PMF
   b. Recalculate the statistic on the resample
3. You now have a distribution of your statistic

What is the distribution of your statistic?
Bootstrapped sample variance

**Bootstrap Algorithm (sample):**
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the sample variance on the resample
3. You now have a distribution of your sample variance

What is the distribution of your sample variance?

Even if we don’t have a closed form equation, we estimate statistics of sample variance with bootstrapping!
Bootstrap: Sample variance
Bootstrapped sample variance

**Bootstrap Algorithm (sample):**
1. Estimate the PMF using the sample
2. Repeat **10,000** times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the **sample variance** on the resample
3. You now have a **distribution of your sample variance**

**Goal**  What is the distribution of your **sample variance**?
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the sample variance on the resample
3. You now have a distribution of your sample variance
Bootstrapped variance

1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the `sample variance` on the resample
3. You now have a distribution of your sample variance
1. Estimate the **PMF** using the sample

2. Repeat **10,000** times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the **sample variance** on the resample

3. You now have a **distribution of your**

   ![Histogram of sample happiness](image)

   ![Histogram of resampled sample variance](image)

Why are these samples different?

This resampled sample is generated **with replacement**.
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the sample variance on the resample
3. You now have a distribution of your sample variance

```
variances = [827.4]
```
1. Estimate the PMF using the sample

2. Repeat **10,000** times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the `sample variance` on the resample

3. You now have a distribution of your sample variance

   \[
   \text{variances} = [827.4]\]
1. Estimate the PMF using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the sample variance on the resample
3. You now have a distribution of your sample variance
   \[
   \text{variances} = [827.4]
   \]
1. Estimate the PMF using the sample

2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the `sample variance` on the resample

3. You now have a distribution of your sample variance

   \[ \text{variances} = [827.4, 846.1] \]
1. Estimate the PMF using the sample

2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the `sample variance` on the resample

3. You now have a distribution of your sample variance

   variances = [827.4, 846.1]
Bootstrapped variance

1. Estimate the PMF using the

2. Repeat 10,000 times:
   a. Resample sample.size()
   b. Recalculate the sample variance

3. You now have a distribution of your sample variance

   variances = [827.4, 846.1, 726.0, ..., 860.7]
Bootstrapped variance

3. You now have a distribution of your sample variance

variances = [827.4, 846.1, 726.0, ..., 860.7]

What is the bootstrapped standard error?

np.std(variances)

Bootstrapped standard error: 66.16
Standard error

1. Mean happiness:

Claim: The average happiness of Bhutan is 83, with a standard error of 1.99.

Closed form: \[ SE = \sqrt{\frac{S^2}{n}} \]

2. Variance of happiness:

Claim: The variance of happiness of Bhutan is 793, with a bootstrapped standard error of 66.16.
Algorithm in practice: Resampling

1. Estimate the **PMF** using the sample
2. Repeat 10,000 times:
   a. Resample `sample.size()` from PMF
   b. Recalculate the statistic on the resample
3. You now have a distribution of your statistic

\[ P(X = k) = \frac{\text{# values in sample equal to } k}{n} \]
Algorithm in practice: Resampling

```python
def resample(sample, n):
    # estimate the PMF using the sample
    # draw n new samples from the PMF
    return np.random.choice(sample, n, replace=True)
```

This resampled sample is generated with replacement.

\[ P(X = k) = \frac{\text{# values in sample equal to } k}{n} \]
Bootstrap provides a way to calculate probabilities of statistics using code.

Bootstrapping works for any statistic*

*as long as your sample is i.i.d. and the underlying distribution does not have a long tail

Google colab notebook link
Bradley Efron

- Invented bootstrapping in 1979
- Still a professor at Stanford
- Won a National Science Medal

Efron’s dice: 4 dice $A, B, C, D$ such that

$$P(A > B) = P(B > C) = P(C > D) = P(D > A) = \frac{2}{3}$$
Bootstrap: p-value
Null hypothesis test

<table>
<thead>
<tr>
<th>Nepal Happiness</th>
<th>Bhutan Happiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.45</td>
<td>0.91</td>
</tr>
<tr>
<td>2.45</td>
<td>0.34</td>
</tr>
<tr>
<td>6.37</td>
<td>1.91</td>
</tr>
<tr>
<td>2.07</td>
<td>1.61</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1.63</td>
<td>1.08</td>
</tr>
</tbody>
</table>

\[ \bar{X}_1 = 3.1 \quad \bar{X}_2 = 2.4 \]

Claim: The difference in mean happiness between Nepal and Bhutan is 0.7 happiness points, and this is statistically significant.
Null hypothesis test

**def null hypothesis** – Even if there is no pattern (i.e., the two samples are from identical distributions), your claim might have arisen by chance.

**def p-value** – What is the probability that the observed difference occurs under the null hypothesis?

**Example:**
- Flip some coin 100 times.
- Flip the same coin another 150 times.
- Compute fraction of heads in both groups.
- There is a possibility we’ll see the observed difference in these fractions even if we used the same coin.

A **significant** p-value (< 0.05) means we **reject** the null hypothesis.
Universal sample (this is what the null hypothesis assumes)

Want \textbf{p-value}: probability $|\bar{X}_1 - \bar{X}_2| = |3.1 - 2.4|$ happens under null hypothesis

\( \bar{X}_1 = 3.1 \)

\( \bar{X}_2 = 2.4 \)
Bootstrap for p-values

1. Create a universal sample using your two samples

i.e., recreate the null hypothesis
Bootstrap for p-values

1. Create a **universal sample** using your two samples

2. Repeat **10,000** times:
   a. Resample **both samples**
   b. Recalculate the **mean difference** between the resamples

3. p-value = \( \frac{\#(\text{mean diffs} \geq \text{observed diff})}{n} \)

Probability that observed difference arose by chance
Bootstrap for p-values

```python
def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = |mean of bhutan_sample - mean of nepal_sample|

    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample
        nepal_resample = draw M resamples from the uni_sample
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = |muNepal - muBhutan|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
```
Bootstrap for p-values

```python
def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = |mean of bhutan_sample - mean of nepal_sample|
    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample
        nepal_resample = draw M resamples from the uni_sample
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = |muNepal - muBhutan|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
```

1. Create a universal sample using your two samples
Bootstrap for p-values

```
def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = |mean of bhutan_sample – mean of nepal_sample|

    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample
        nepal_resample = draw M resamples from the uni_sample
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = |muNepal – muBhutan|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
```
Bootstrap for p-values

```python
def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = |mean of bhutan_sample – mean of nepal_sample|

    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample
        nepal_resample = draw M resamples from the uni_sample
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = |muNepal – muBhutan|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
```

2. b. Recalculate the mean difference b/t resamples
Bootstrap for p-values

3. \[ p-value = \frac{\text{# (mean diffs} > \text{observed diff)}}{n} \]

```python
def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = \|\text{mean of bhutan_sample} - \text{mean of nepal_sample}\|

    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample
        nepal_resample = draw M resamples from the uni_sample
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = \|\muNepal - \muBhutan\|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
```
Bootstrap for p-values

def pvalue_boot(bhutan_sample, nepal_sample):
    N = size of the bhutan_sample
    M = size of the nepal_sample
    observed_diff = |mean of bhutan_sample – mean of nepal_sample|

    uni_sample = combine bhutan_sample and nepal_sample
    count = 0

    repeat 10,000 times:
        bhutan_resample = draw N resamples from the uni_sample with replacement!
        nepal_resample = draw M resamples from the uni_sample with replacement!
        muBhutan = sample mean of the bhutan_resample
        muNepal = sample mean of the nepal_resample
        diff = |muNepal – muBhutan|
        if diff >= observed_diff:
            count += 1

    pValue = count / 10,000
Bootstrap

Let’s try it!

Google colab notebook [link]
Null hypothesis test

<table>
<thead>
<tr>
<th>Nepal Happiness</th>
<th>Bhutan Happiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.45</td>
<td>0.91</td>
</tr>
<tr>
<td>2.45</td>
<td>0.34</td>
</tr>
<tr>
<td>6.37</td>
<td>1.91</td>
</tr>
<tr>
<td>2.07</td>
<td>1.61</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1.63</td>
<td>1.08</td>
</tr>
</tbody>
</table>

$\bar{X}_1 = 3.1$

$\bar{X}_2 = 2.4$

Claim: The happiness of Nepal and Bhutan have a 0.7 difference of means, and this is statistically significant ($p < 0.05$).