19: Sampling and the Bootstrap

Jerry Cain
May 10, 2021
## Quick slide reference

<table>
<thead>
<tr>
<th>Page</th>
<th>Topic</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Sampling definitions</td>
<td>19a_intro</td>
</tr>
<tr>
<td>11</td>
<td>Unbiased estimators</td>
<td>19b_sample_stats</td>
</tr>
<tr>
<td>23</td>
<td>Reporting estimation error</td>
<td>19c_statistical_error</td>
</tr>
<tr>
<td>29</td>
<td>Bootstrap: Sample mean</td>
<td>19d_bootstrap_mean</td>
</tr>
<tr>
<td>40</td>
<td>Bootstrap: Sample variance</td>
<td>LIVE</td>
</tr>
<tr>
<td>*</td>
<td>Bootstrap: Hypothesis testing</td>
<td>LIVE</td>
</tr>
</tbody>
</table>
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:

   Happiness = {72, 85, 79, 91, 68, ..., 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.
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 **sample 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: 
\((59, 87, 94, 99, 87, 78, 69, 91)\)
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).

Today: If we only have a single sample,
- How do we report estimated statistics?
- How do we report estimated error of these estimates?
- How do we perform hypothesis testing?
Unbiased estimators
A single sample

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

But we only have 200 people (a sample).

So these population 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 population mean and population variance?
- How 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 a sample, $(X_1, X_2, \ldots, X_n)$:

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
Sample mean

Even if we can’t report $\mu$, we can report our sample mean 83.03, which is an unbiased estimate of $\mu$. 
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 a sample, \((X_1, X_2, \ldots, X_n)\):

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
Calculating population statistics exactly requires us knowing all \( N \) datapoints.
Intuition about the sample variance, $S^2$

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

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

**Estimate, $S^2$**

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

Population size, $N$

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
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
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Diagram:
- Happiness
- Population size, $N$
- $\mu$
- $\bar{X}$
Intuition about the sample variance, \( S^2 \)

Actual, \( \sigma^2 \):
\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]

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

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

2. What is \( \sigma^2 \), the 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 sample variance:

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

\( S^2 \) is an unbiased estimator of the population variance, \( \sigma^2 \).  
\[ E[S^2] = \sigma^2 \]
Proof that $S^2$ is unbiased  

(just for reference) \[ E[S^2] = \sigma^2 \]

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

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

(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, $X_1, X_2, \ldots, X_n$.

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

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

$n = 200$

$\bar{X} = 83$

$(-11, 2, -4, -4, 8, -15, \ldots, -12)$

$S^2 = 793$
Sample mean

\[ X_i \sim F \]

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

\[ \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \]
How “close” is our estimate $\bar{X}$ to $\mu$?

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

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

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

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 \( n = 200 \))

\[
\frac{\sigma}{\sqrt{n}} = 1.886
\]

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 Lisa 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.*

The underlying distribution $F$ $\approx$ $\hat{F}$

*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)

≈

Sample distribution (we do have this)
Bootstrapped sample means

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.

![Histogram of Happiness](Image)

![Histogram of Sample Means](Image)

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

\[
\text{np.std(means)} = 2.003
\]
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)
  - $\sigma / \sqrt{n} = 1.886$
  - Exact statistic (we don’t have this)

- **Sample distribution** (we do have this)
  - $SE = S / \sqrt{n} = 1.992$
  - Estimated statistic, by formula, standard error
  - Simulated estimated statistic, bootstrapped standard error

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
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 during live lecture!
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 **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!
19: Sampling and the Bootstrap (live)

Jerry Cain
May 10, 2021
Think Slide 42 has a question to go over by yourself.

Post any clarifications here or in Zoom chat! https://edstem.org/us/courses/5090/discussion/428950

Think by yourself: 2 min
Quick check

1. \( \mu \), the population mean

2. \( (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \), a sample

3. \( \sigma^2 \), the population variance

4. \( \bar{X} \), the sample mean

5. \( \bar{X} = 83 \)

6. \( (X_1 = 59, X_2 = 87, X_3 = 94, X_4 = 99, X_5 = 87, X_6 = 78, X_7 = 69, X_8 = 91) \)
Quick check

1. $\mu$, the population mean

2. $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$, a sample

3. $\sigma^2$, the population variance

4. $\bar{X}$, the sample mean

5. $\bar{X} = 83$

6. $(X_1 = 59, X_2 = 87, X_3 = 94, X_4 = 99, X_5 = 87, X_6 = 78, X_7 = 69, X_8 = 91)$

These are outcomes from your collected data.
Today: Crash course on (bootstrapped) statistics

If we only have a single sample of RVs generated i.i.d. from the same unknown distribution, how can we perform statistical analysis?

- What is the probability that a Bhutanese peep is just straight up loving life?
- What is a good estimate of the population mean (and how “close” is the estimate)?
- What is a good estimate of the population variance (and how “close” is the estimate)?
1. Mean happiness:

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

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

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

Verified via bootstrap: \[ \text{np.std(means)} = 2.003 \]

\[ 83 \]

\[ 0 \]

Bhutan

Average Happiness

SE

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
1. Mean happiness:

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

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

2. Variance of happiness:

Claim: The variance of happiness of Bhutan is 793.

Closed form: Not covered in CS109

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

We can bootstrap for standard error of sample variance—a statistic of a statistic.
Bootstrap

The Bootstrap:

Probability for Computer Scientists

Allows you to do the following:

• Calculate distributions over statistics
• Calculate p values
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**?
**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**
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

This resampled sample is generated with replacement.

Why are these samples different?
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]
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

   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
   
   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

   \[ \text{variances} = [827.4, 846.1] \]
Bootstrapped variance

1. Estimate the PMF using the sample

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

3. You now have a distribution of your sample variance

   \[
   \text{variances} = [827.4, 846.1, 726.0, \ldots, 860.7]
   \]
3. You now have a **distribution of your sample variance**

$$\text{variances} = [827.4, 846.1, 726.0, ..., 860.7]$$

What is the bootstrapped standard error?

```
np.std(variances)
```

**Bootstrapped standard error: 66.16**

- Simulate a distribution of sample variances
- Compute standard deviation
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.

\( S^2 \) is our best estimate of \( \sigma^2 \)

this is how close we are, calculated by bootstrapping
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

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

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
**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.

$$P(X = k) = \frac{\text{# values in sample equal to } k}{n}$$

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
To the code!

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](#)
(we will use this in Breakout rooms)
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

$$P(A > B) = P(B > C) = P(C > D) = P(D > A) = \frac{2}{3}$$
Interlude for announcements
Announcements

Problem Set 5
Out: now
Due: Friday 5/21 10:00am
Covers: Up to and including today

Quiz #2
Time frame: This Wednesday 5/12 11:00am – Friday 5/14 10:00am PT
Covers: Up to end of Week 5 (including Lecture 15). PS3+PS4
Emma’s Review session: Tonight at 7pm PT (and will be recorded)
Info and practice: http://web.stanford.edu/class/cs109/quizzes/
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 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)

\[ X_1 = 3.1 \]

\[ X_2 = 2.4 \]

Want **p-value**: probability \(|X_1 - X_2| = |3.1 - 2.4|\) happens under null hypothesis
Bootstrap for p-values

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

    i.e., recreate the null hypothesis

---

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
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. \[ p\text{-value} = \frac{\# (\text{mean diffs} \geq \text{observed diff})}{n} \]
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

```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. a. Resample both samples

2. b. Repeat this process 10,000 times to get a p-value.

2. c. Calculate the proportion of times the absolute difference between the resampled means is greater than or equal to the observed difference.
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 = |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
```

P. Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
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

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>

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