19: Sampling and the Bootstrap

Jerry Cain
May 13th, 2024

Lecture Discussion on Ed
Sampling definitions
Motivating example

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

• You can’t ask everyone very easily, so..
• You poll 200 random people.
• Your poll data looks like this:

Happiness = {72, 85, 79, 91, 68, ..., 71}

• The mean of all these numbers is 83.

Is this the **true average happiness** of all Bhutanese people?
Population

This is a population.
A **sample** is selected from a population.
Sample

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 infer by polling a sample drawn from that population?

2. How large does your sample need to be before your conclusions become trustworthy, and how do we express our confidence in those conclusions.

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 (iid)
- \( X_i \) 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:
\((X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)\)

The 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—or rather, a sample.

Assuming we have just 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 a sample. How do we report confidence intervals on these estimates? How do we perform something called hypothesis testing? Oh, and what is it?
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 a sample of happiness scores from 200 people.

These population-level statistics are unknown:
- $\mu$, the population mean
- $\sigma^2$, the population variance
If we had a distribution $F$ of our entire population, we could compute exact statistics about happiness. But we only have a sample of happiness scores from 200 people.

- Using this sample, what is our best guess estimate of the population mean and the population variance?
- How exactly do we define best guess 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, \( \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \)

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 of 83, which is an unbiased estimate of $\mu$. 

$X_i \sim F$

\[ \bar{X} \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{n} \right) \]
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: $(X_1, X_2, \ldots, X_n)$:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$
Calculating population statistics **exactly** requires us knowing all \( N \) datapoints.

Intuition about the sample variance, \( S^2 \)

### Actual, \( \sigma^2 \)

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

- **population variance**
- **population mean**
- \( x_i \) is an individual data point
- \( \mu \) is the population mean

0
Happiness

Population size, \( N \)

Calculated sample variance, \( S^2 \)
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

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

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2024

Happiness

Population size, $N$
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**

**Estimate, $S^2$**

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

**Sample variance**

- Population mean: $\mu$
- Sample mean: $\bar{X}$

Population size, $N$

Happiness

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2024
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
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Sample variance is an estimate computed using another estimate, so the denominator needs to be slightly smaller to compensate.
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)

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

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

\begin{align*}
&= 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
\end{align*}

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

1. Collect a sample, $X_1, X_2, \ldots, X_n$.  
   \[ (72, 85, 79, 79, 91, 68, \ldots, 71) \]
   \[ n = 200 \]

2. Compute sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.  
   \[ \bar{X} = 83 \]

3. Compute sample deviation, $X_i - \bar{X}$.  
   \[ (-11, 2, -4, -4, 8, -15, \ldots, -12) \]

4. Compute sample variance, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.  
   \[ S^2 = 793 \]

How close are our estimates $\bar{X}$ and $S^2$ to the true $\mu$ and $\sigma^2$?
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 $\overline{X}$ to $\mu$?

\[ E[\overline{X}] = \mu \]

\[ \text{Var}(\overline{X}) = \frac{\sigma^2}{n} \]

We want to estimate this

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

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

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

error bars

These 2 statistics give a sense of how \( \bar{X} \)—that is, the sample mean random variable—behaves.
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

⚠️ 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 $n = 200$)

$\frac{\sigma}{\sqrt{n}} = 1.869$

Population distribution
(we don’t have this)

Sample distribution
(we do have this)

Exact statistic
(we don’t have this)

Simulated statistic
(we don’t have this)

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.*

*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

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

Estimate the true PMF using our "PMF" (histogram) of our sample.

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

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

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

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

$\sigma \over \sqrt{n} = 1.869$

Exact statistic (we don’t have this)

Simulated statistic (we don’t have this)

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

Estimated statistic, by formula, standard error

$2.003$

Simulated estimated statistic, bootstrapped standard error
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 with a demo!
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 without closed forms, we estimate statistics of the 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?
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 \texttt{sample.size()} from PMF
   b. Recalculate the \texttt{sample variance} on the resample
3. You now have a distribution of your sample variance

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**

   \[
   \text{variances} = [827.4]
   \]
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
   \[ \text{variances} = [827.4] \]
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

   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**

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

2. Variance of happiness:

Claim: The variance of happiness of Bhutan is 793, with a \textbf{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.

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

Inventor of Efron’s dice: 4 dice $A, B, C, D$ where:

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