

Bootstrapping

CS109, Stanford University



Okay, so what today?



A real difference?

	Learning in Context A	Learning in Context B	
18 students	4.44	2.15	23 students
	3.36	3.01	
	5.87	2.02	
	2.31	1.43	
	
	3.70	1.83	
	$\mu_1 = 3.1$	$\mu_2 = 2.4$	

Claim: Group 1 and Group 2 are samples from **different distributions** with a 0.7 difference of means.

How confident are you in this claim?

The Classic Science Test

Group 1	Group 2
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83

$\mu_1 = 3.1$ $\mu_2 = 2.4$

Claim: Group 1 and Group 2 are samples from **different distributions** with a 0.7 difference of means.

How confident are you in this claim?

Puzzle

Results of flipping a coin 20 times. Give your belief distribution of p :

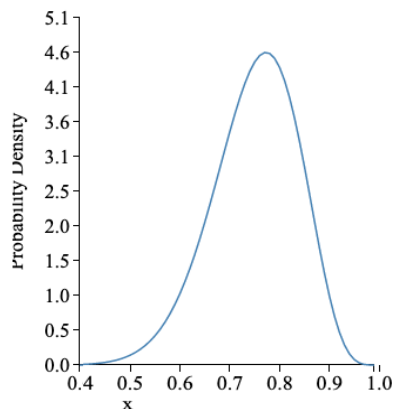
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

Let's use a prior belief


$$X \sim \text{Beta}(a = 18, b = 6)$$



How can you build
distribution for p
without using a prior?

Where are we in CS109?


You are here


Counting
Theory


Core
Probability

x_2
Random
Variables


Probabilistic
Models


Uncertainty
Theory


Machine
Learning

Uncertainty Theory

Beta
Distributions

Thompson
Sampling

Adding
Random Vars

Central Limit
Theorem

Sampling

Bootstrapping

Algorithmic
Analysis

<review>

Central Limit Theorem (Summation)

Consider n independent and identically distributed (i.i.d) variables X_1, X_2, \dots, X_n with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$.

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The **sum** of the variables is normally distributed

Central Limit Theorem (Average)

Consider n independent and identically distributed (i.i.d) variables X_1, X_2, \dots, X_n with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$.

$$\frac{1}{n} \sum_{i=1}^n X_i \underset{\text{As } n \rightarrow \infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The **average** of the variables is normally distributed

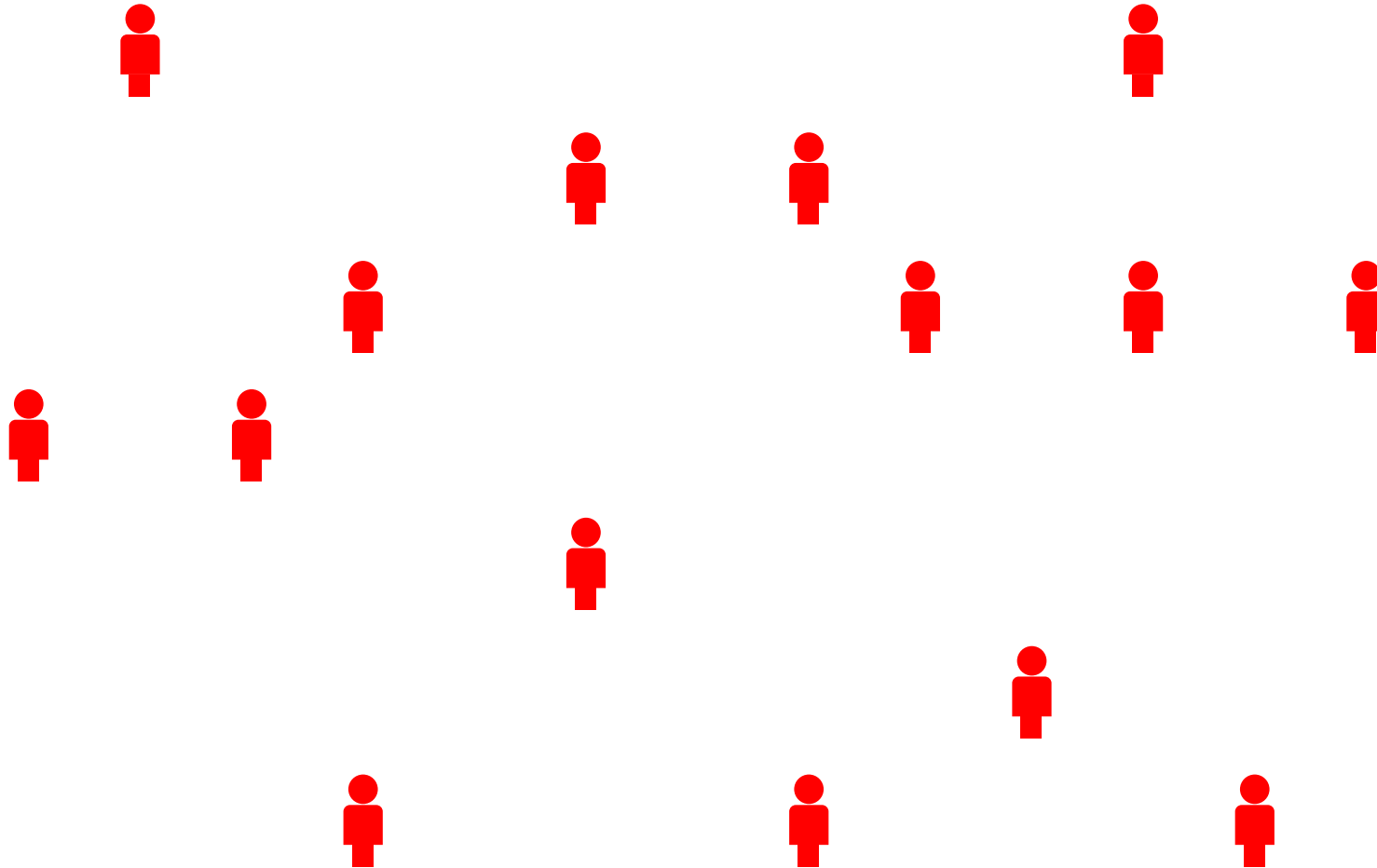
Population



Sample



Sample



Collect one (or more) numbers from each person

Sample



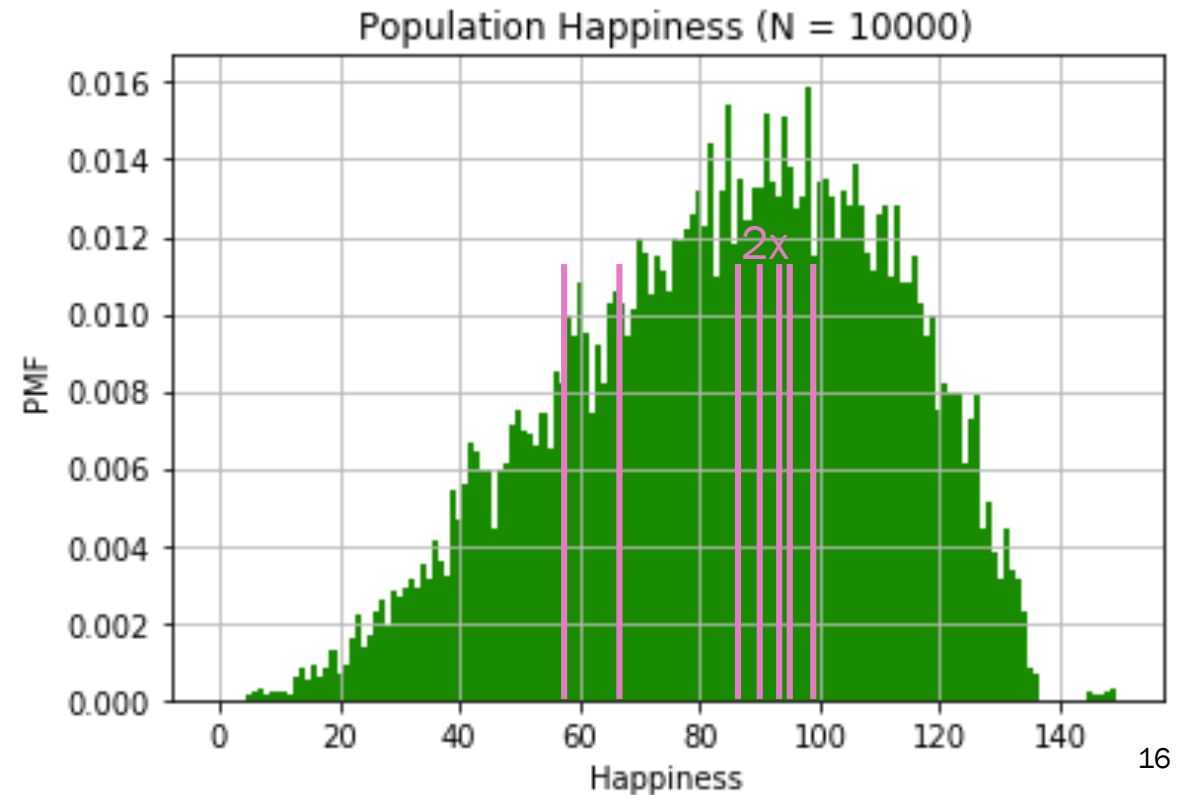
A sample, mathematically

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:

$(59, 87, 94, 99, 87, 78, 69, 91)$



Equations we used to get those values

sample
mean
estimate

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

Our best guess at the
true mean (unbiased)

sample
variance
estimate

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

sample mean



Our best guess at the
true variance
(unbiased)

Std error of
the mean
estimate

$$\text{Std}(\bar{X}) = \sqrt{\frac{S^2}{n}}$$

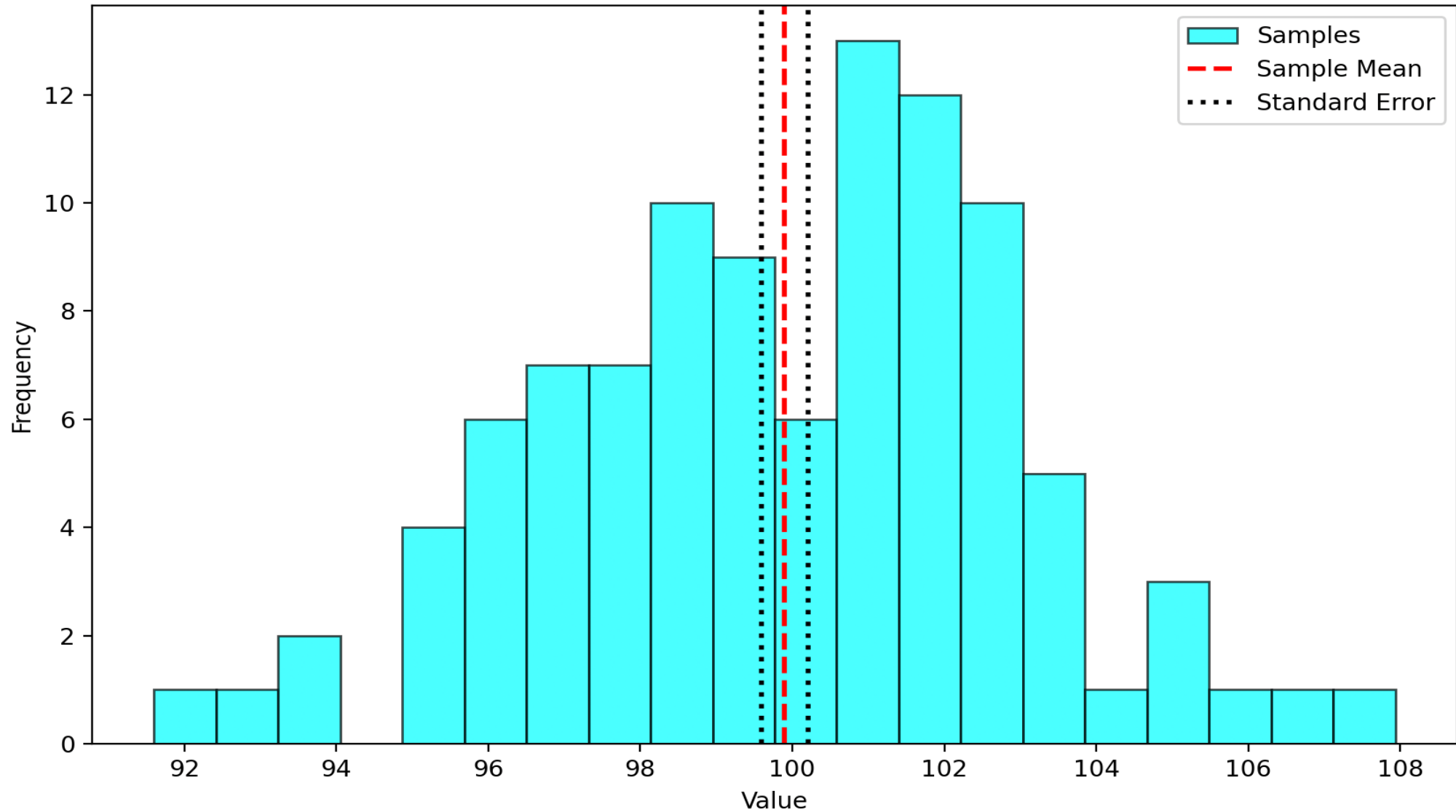
sample variance



How wrong do we
think our mean
estimate is?

Statistics Vs Distribution

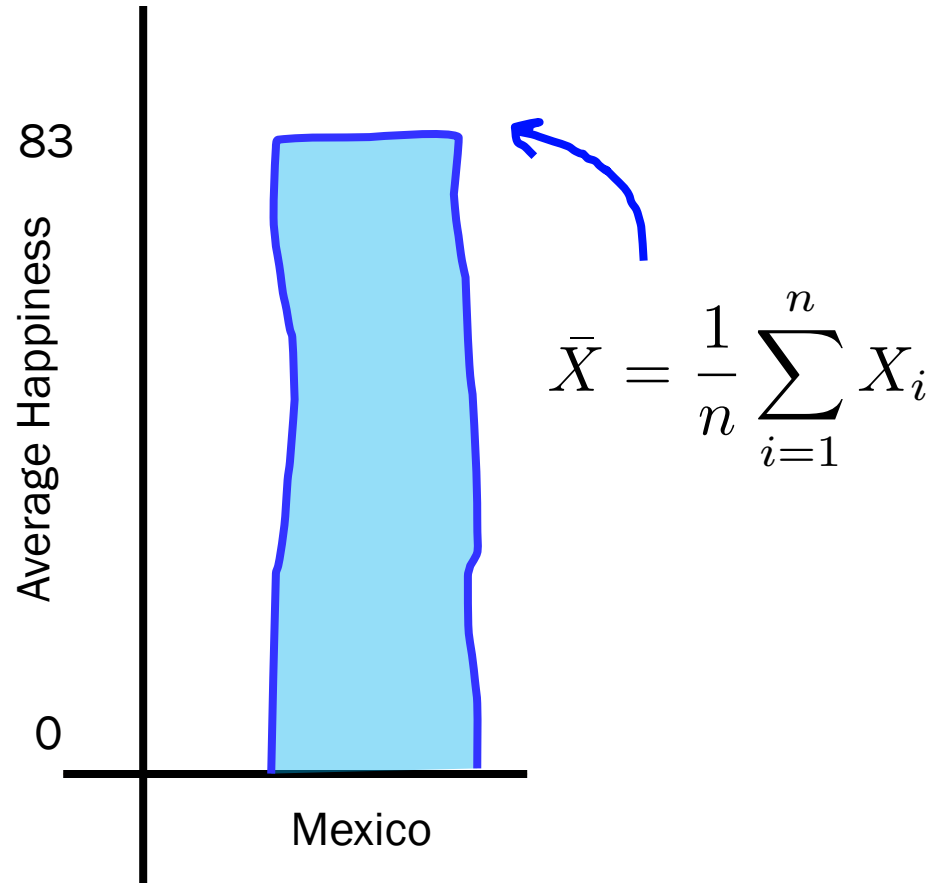
Distribution of Samples with Sample Mean and Standard Error Bars



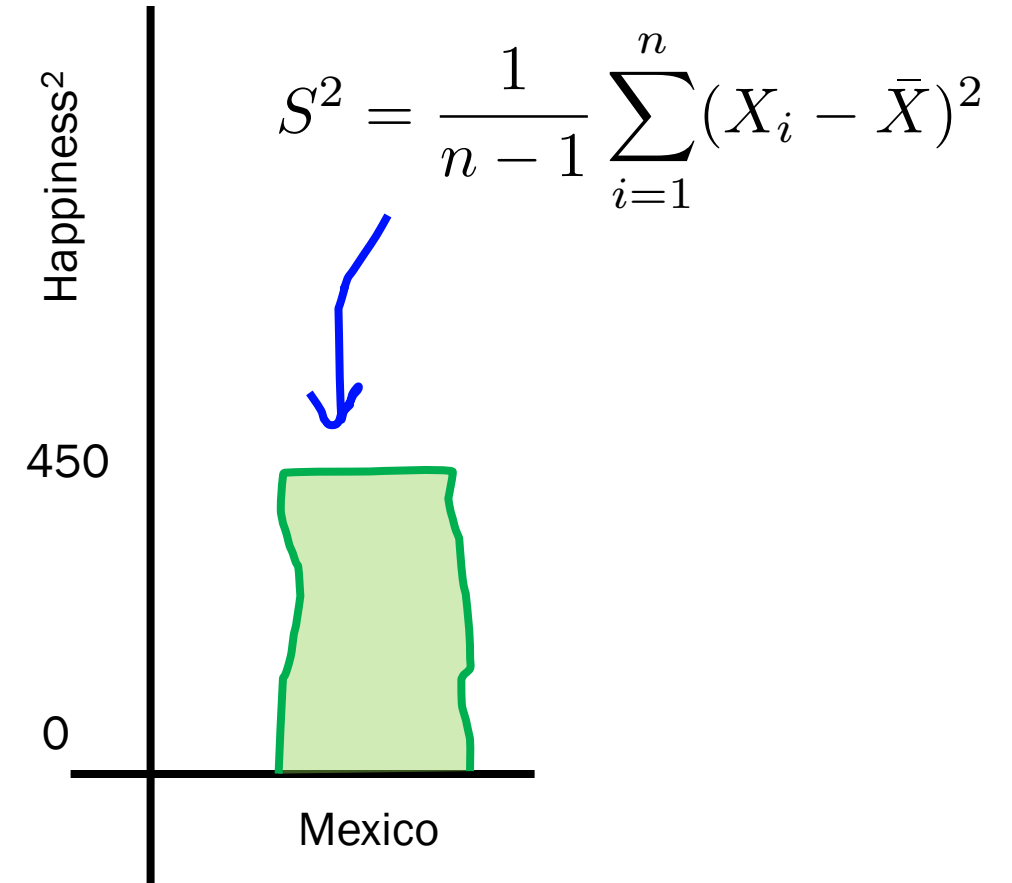
But what about Mexico?

Our Report to Mexican Government (after talking to 200 ppl)

Average Happiness



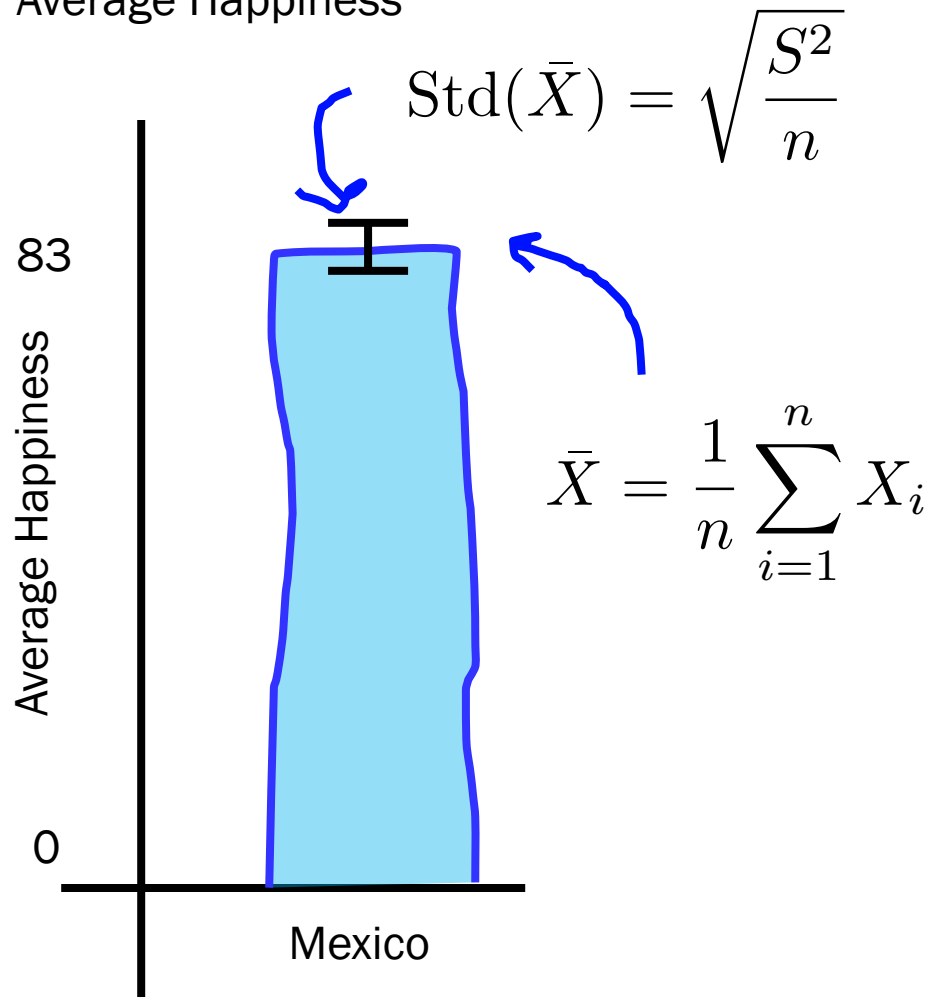
Variance of Happiness



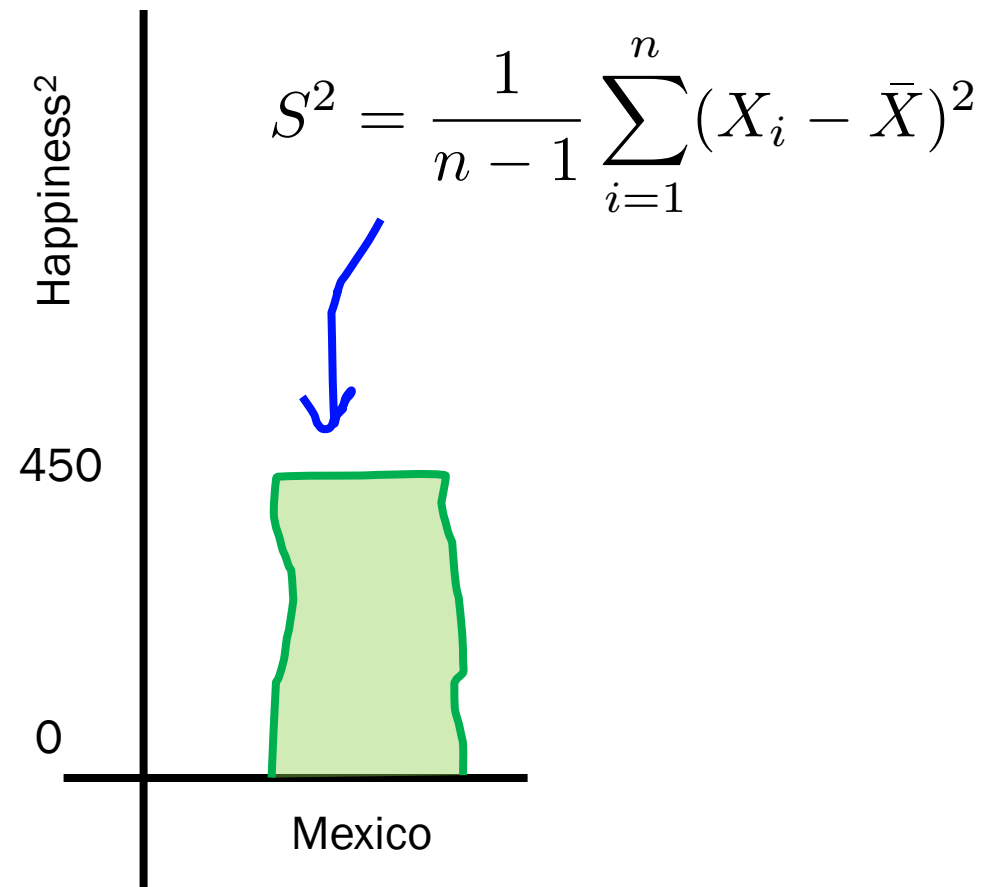
But what about **error bars**???

By CLT: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Average Happiness



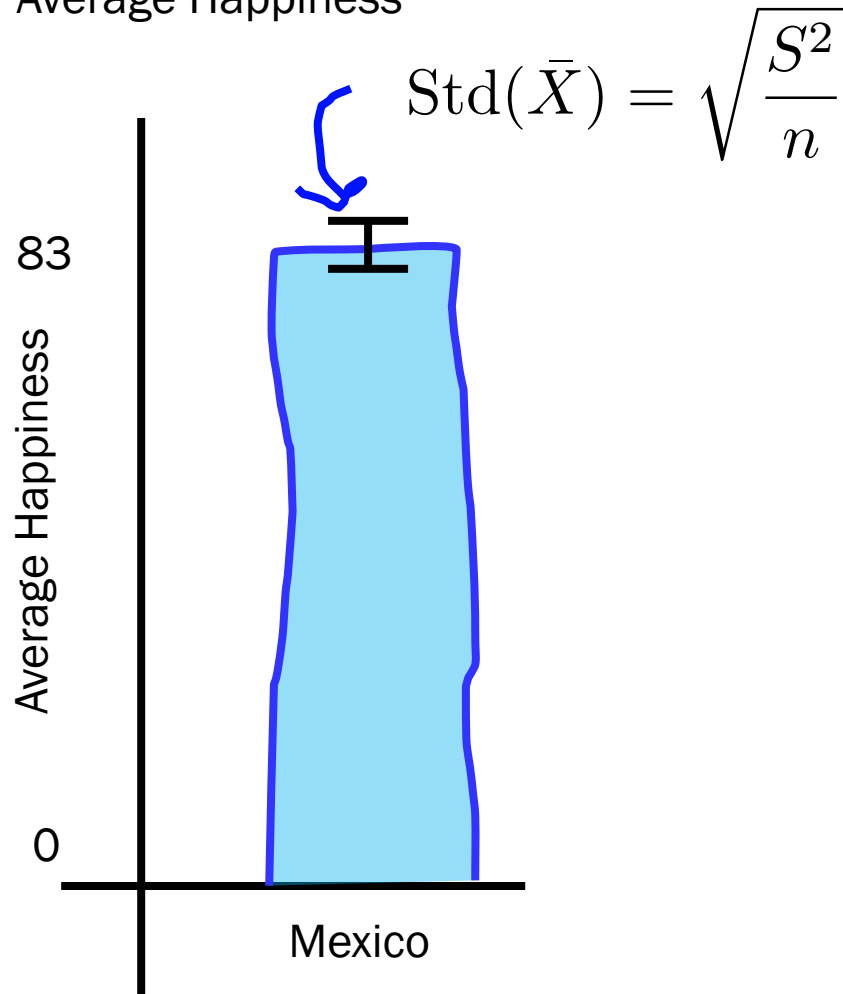
Variance of Happiness



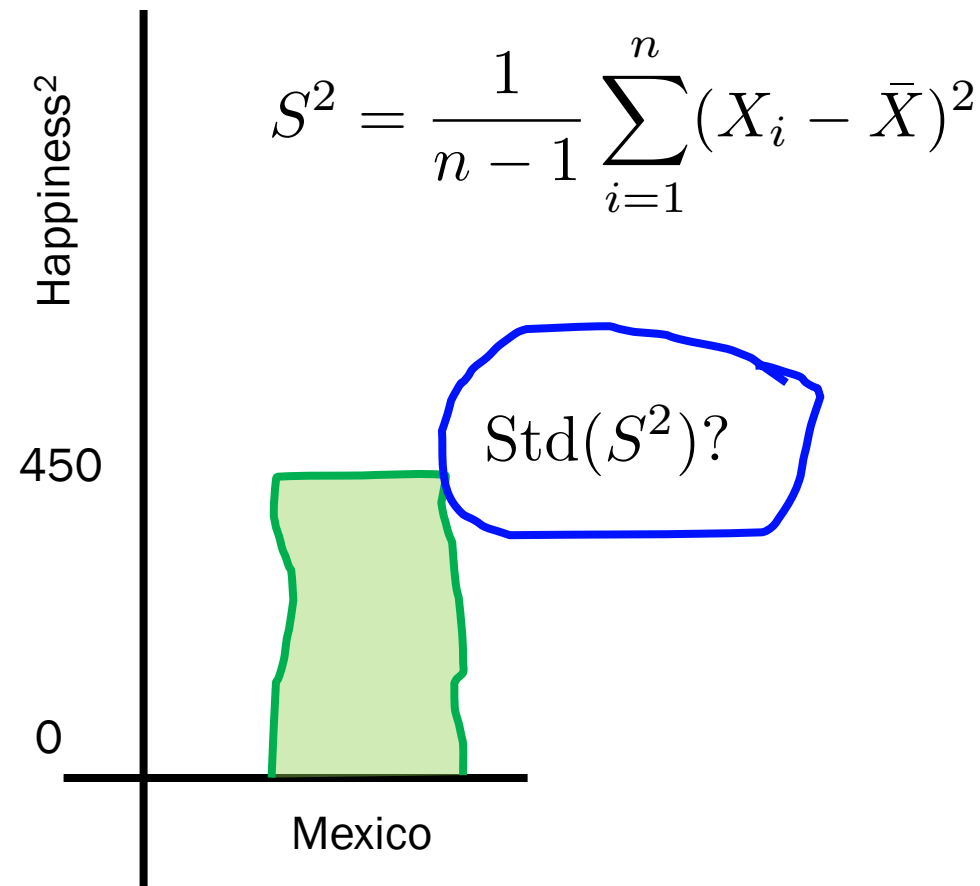
But what about **error bars**???

By CLT: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Average Happiness




Variance of Happiness




[suspense]

GPT-4 Tells about the Importance of Bootstrapping

✦ GPT-4

- >  Can you name 10 statistical tests that a datascientist should know?

-  If you had enough samples, which tests could be replaced with bootstrapping?

GPT-4 Tells about the Importance of Bootstrapping

2 / 2 >

PI Can you name 10 statistical tests that a datascientist should know?



Certainly! Here's a list of 10 statistical tests that are commonly used in data science, along with a brief description of when they might be used:



1. **Pearson's Chi-squared test:** Used for testing relationships between categorical variables.
2. **T-test (Independent and Paired):** The Independent samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Paired sample t-test is used to compare the means from the same group at different times.
3. **ANOVA (Analysis of Variance):** Compares the means among three or more groups to understand if at least one group mean is statistically different from the others.
4. **Linear Regression:** Used for predicting the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable, and the one you are using to predict the other variable's value is called the independent variable.
5. **Logistic Regression:** Used when the dependent variable is categorical, to predict the outcome of a categorical dependent variable based on one or more predictor variables.
6. **Wilcoxon Signed-Rank Test:** A non-parametric test that compares two paired groups. It's used as an alternative to the paired t-test when the data does not meet the assumptions required by the t-test.
7. **Mann-Whitney U Test:** A non-parametric test for assessing whether two independent samples of observations come from the same distribution. It is an alternative to the independent sample t-test when the data does not satisfy the assumptions of the t-test.
8. **Kruskal-Wallis H Test:** A non-parametric version of ANOVA, which is used when the assumptions of ANOVA are not met. It is used to compare more than two groups.
9. **Fisher's Exact Test:** Used for testing the association between two categorical variables in small sample sizes. It's an alternative to the chi-squared test when the sample size is too small to meet the chi-squared test's assumptions.
10. **Time Series Analysis (e.g., ARIMA):** Not a test per se, but a suite of methods for analyzing time series data, which can include tests for stationarity (ADF Test), seasonality, autocorrelation, and the ability to forecast future values from past data.

Regenerate

PI

If you had enough samples, which tests could be replaced with bootstrapping?



Bootstrapping is a powerful resampling method that can be used to estimate the distribution of a statistic (like the mean, median, or a regression coefficient) by sampling with replacement from the data. When you have a sufficiently large sample size, bootstrapping can be used to relax the assumptions of traditional parametric tests, like the normality assumption or the assumption of equal variances across groups.



In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

1. **T-test:** Both independent and paired t-tests can be replaced with bootstrap methods to estimate the difference in means without assuming normal distributions of the sample means.
2. **ANOVA:** Bootstrap methods can be used to assess the difference in means across groups without assuming normally distributed residuals or homogeneity of variance.
3. **Linear Regression:** Instead of assuming the normality of residuals, you can use bootstrapping to construct confidence intervals for the regression coefficients.
4. **Logistic Regression:** Similarly, bootstrapping can be used to obtain more robust confidence intervals for the coefficients in a logistic regression model.
5. **Wilcoxon Signed-Rank Test and Mann-Whitney U Test:** While these are non-parametric tests that do not assume normality, bootstrapping can still be used to obtain confidence intervals for the median difference or other statistics of interest.
6. **Kruskal-Wallis H Test:** Bootstrapping can be applied to obtain the distribution of the test statistic under the null hypothesis, and thus get a p-value without the assumption of identical distribution shapes across groups.
7. **Fisher's Exact Test:** When sample sizes are large, the exact calculations required for Fisher's Exact Test can become computationally intensive. Bootstrapping can be a practical alternative to obtain an approximation.

GPT-4 Tells about the Importance of Bootstrapping

2 / 2 >

PI Can you name 10 statistical tests that a datascientist should know?



Certainly! Here's a list of 10 statistical tests that are commonly used in data science, along with a brief description of when they might be used:



1. **Pearson's Chi-squared test:** Used for testing relationships between categorical variables.
2. **T-test (Independent and Paired):** The Independent samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Paired sample t-test is used to compare the means from the same group at different times.
3. **ANOVA (Analysis of Variance):** Compares the means among three or more groups to understand if at least one group mean is statistically different from the others.
4. **Linear Regression:** Used for predicting the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable, and the one you

PI

If you had enough samples, which tests could be replaced with bootstrapping?



Bootstrapping is a powerful resampling method that can be used to estimate the distribution of a statistic (like the mean, median, or a regression coefficient) by sampling with replacement from the data. When you have a sufficiently large sample size, bootstrapping can be used to relax the assumptions of traditional parametric tests, like the normality assumption or the assumption of equal variances across groups.



In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

1. **T-test:** Both independent and paired t-tests can be replaced with bootstrap methods to

In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

small sample sizes. It's an alternative to the chi-squared test when the sample size is too small to meet the chi-squared test's assumptions.

10. **Time Series Analysis (e.g., ARIMA):** Not a test per se, but a suite of methods for analyzing time series data, which can include tests for stationarity (ADF Test), seasonality, autocorrelation, and the ability to forecast future values from past data.

statistic under the null hypothesis, and thus get a p-value without the assumption of identical distribution shapes across groups.

7. **Fisher's Exact Test:** When sample sizes are large, the exact calculations required for Fisher's Exact Test can become computationally intensive. Bootstrapping can be a practical alternative to obtain an approximation.

Regenerate

Bootstrap: Probability for Computer Scientists

Bootstrapping allows you to:

- Know the **distribution of *statistics***
- Calculate **p values**
- **Using computers**
- You totally **could have invented it**

Hypothetical – You have the underlying distribution!

How wrong is an estimate of **sample variance**, calculated from 200 people?

Plot twist: I give you the *entire* underlying distribution



Hypothetical – You have the underlying distribution!

What is the **std** of the **sample variance**, calculated from 200 people?

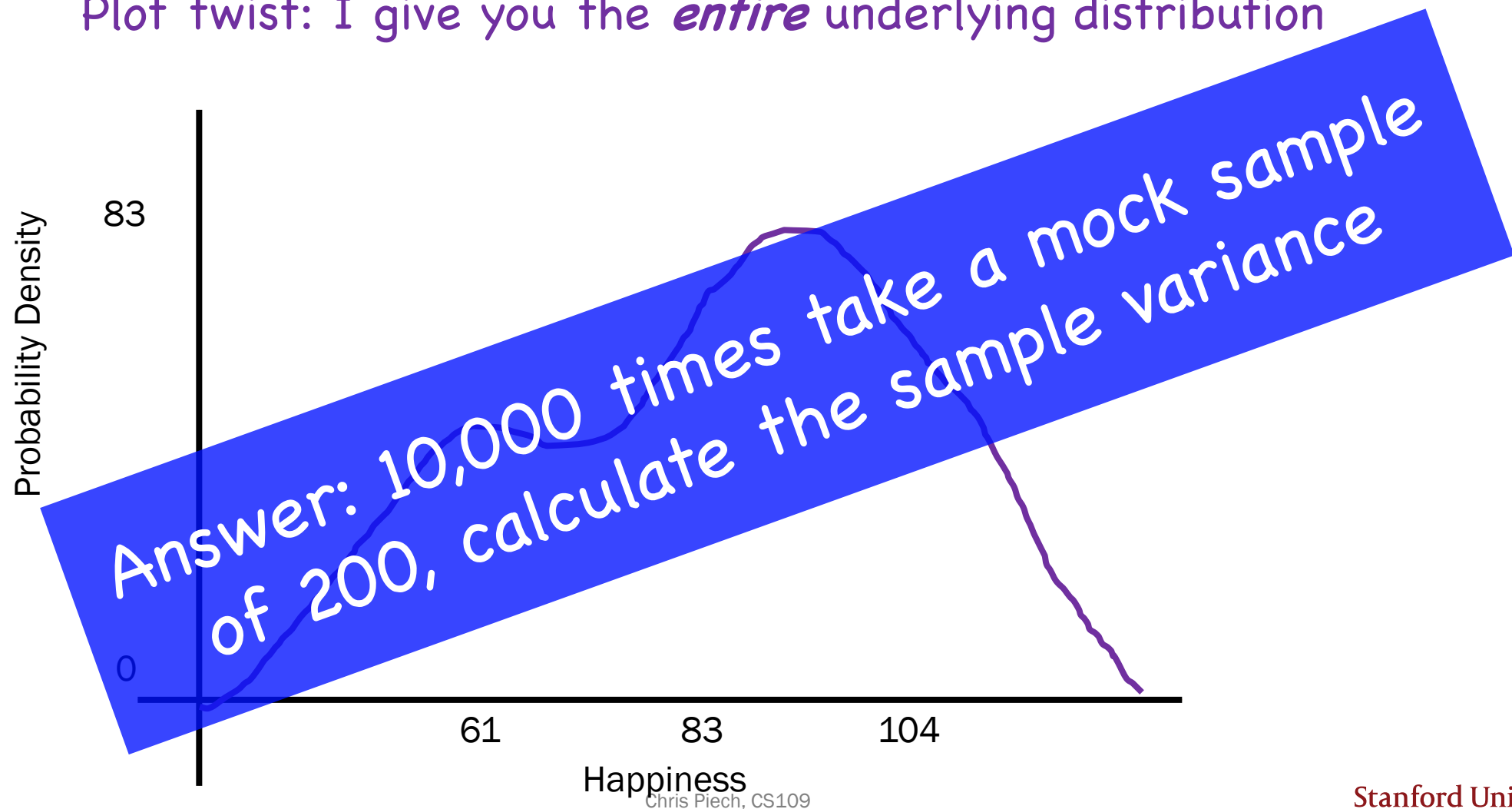
Plot twist: I give you the *entire* underlying distribution



Hypothetical – You have the underlying distribution!

What is the **std** of the **sample variance**, calculated from 200 people?

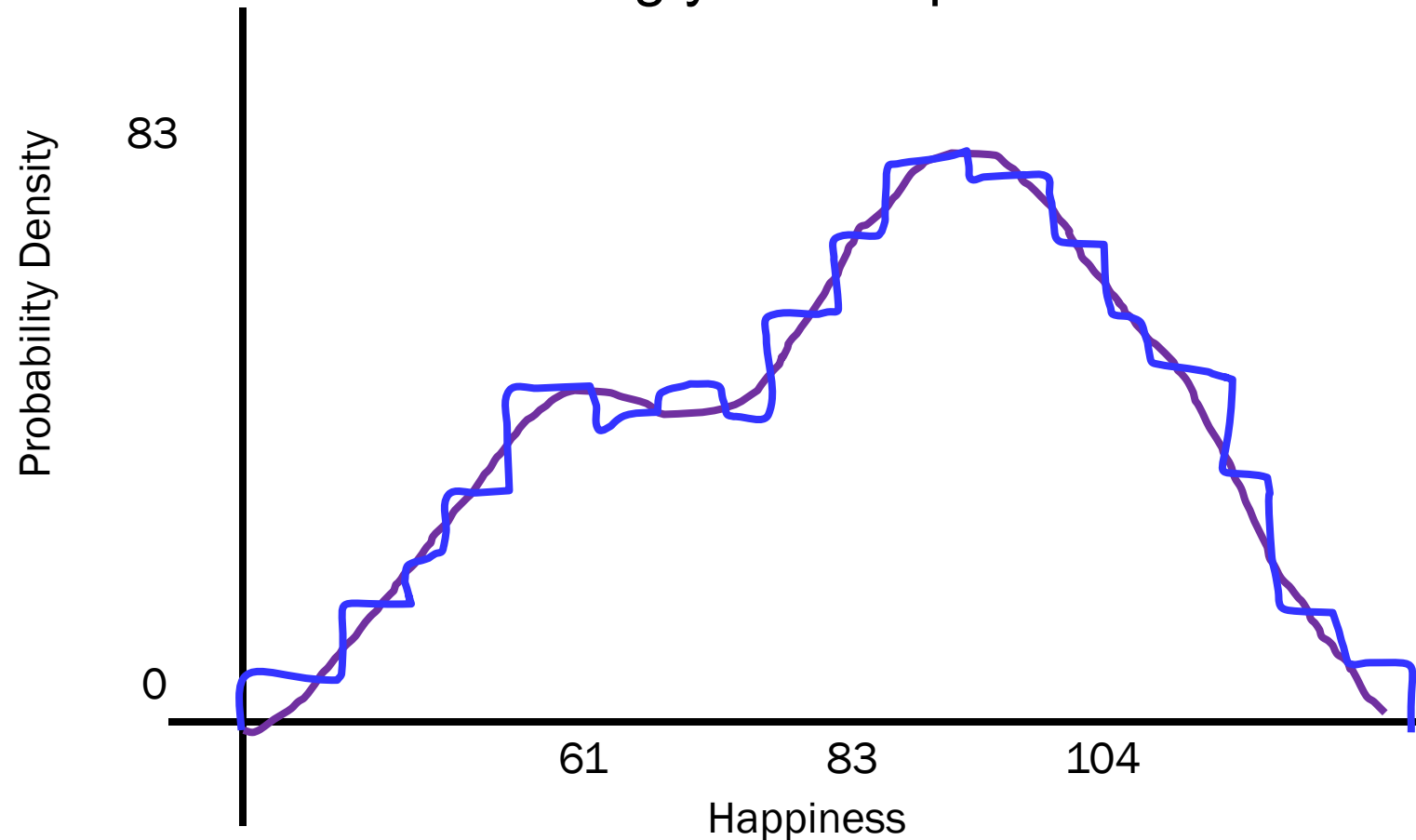
Plot twist: I give you the *entire* underlying distribution



Here comes the award winning idea....

But Wait – What If You Actually Have a Good Estimate?

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



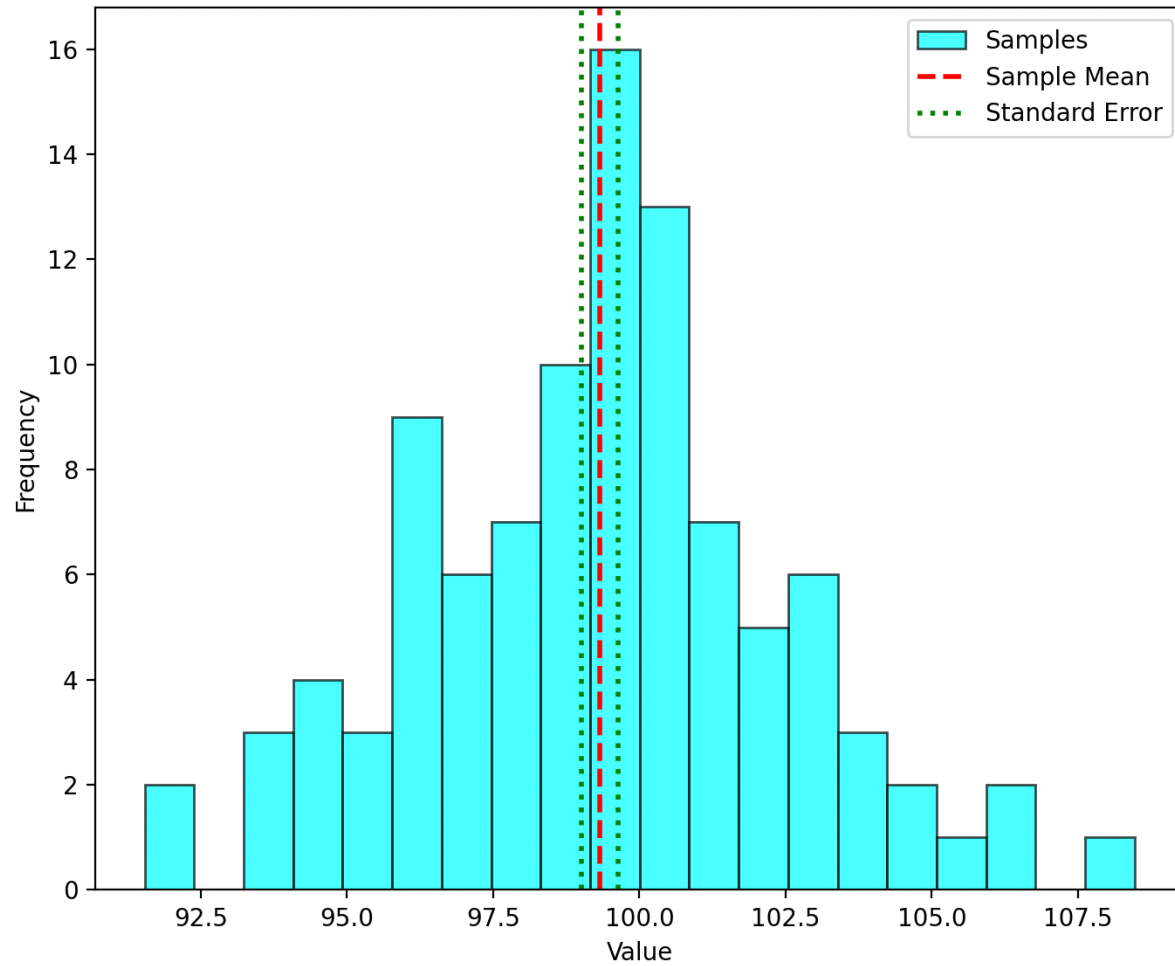
* This is just a histogram of your data!!

Chris Piech, CS109

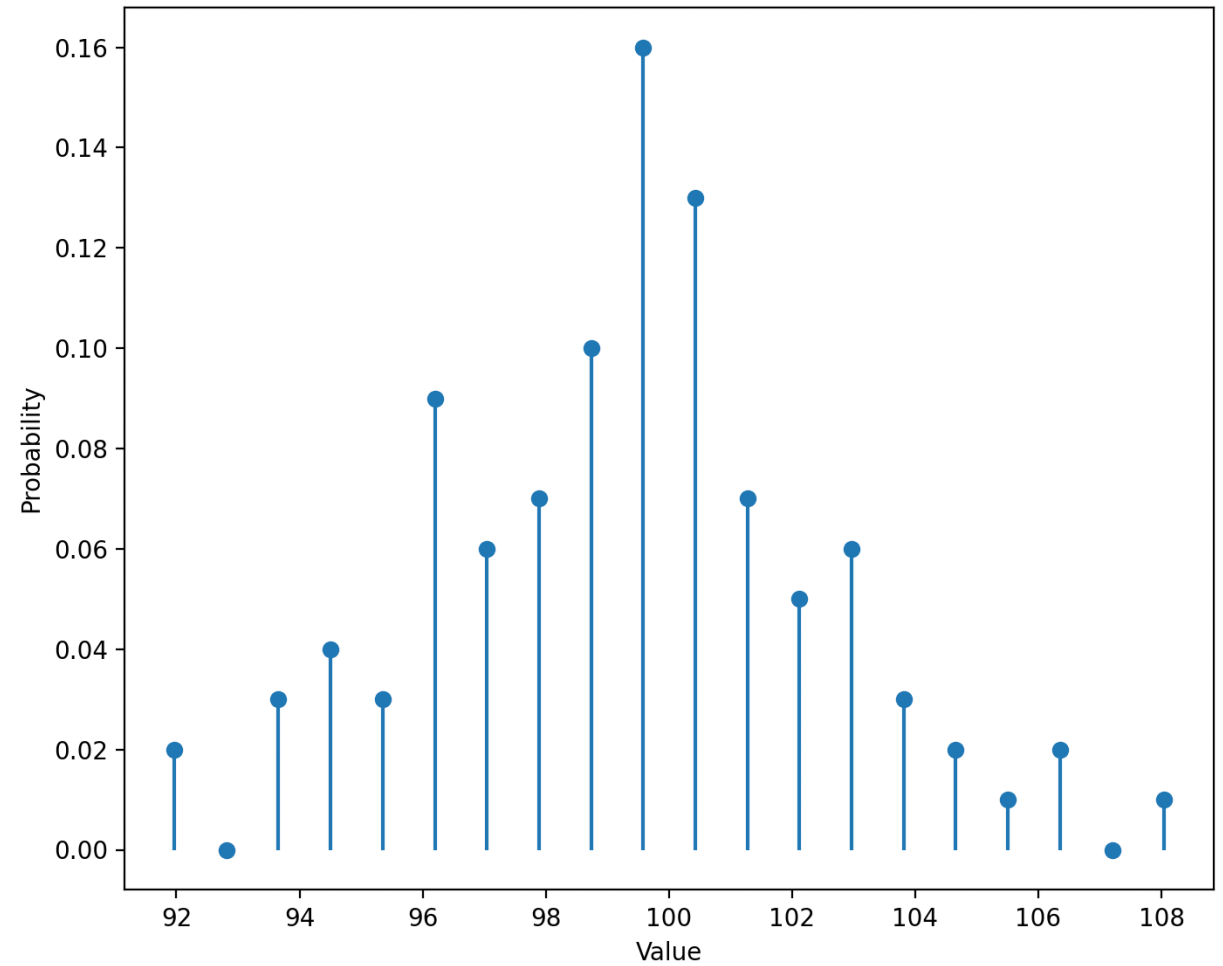
Stanford University

Key Insight

Distribution of Samples with Sample Mean and Standard Error



Probability Mass Function (PMF) of Samples



Bootstrapping Assumption

$$F \approx \hat{F}$$



The underlying
distribution



The sample
distribution

(aka the histogram of
your data)

Algorithm

Bootstrap Algorithm (sample):

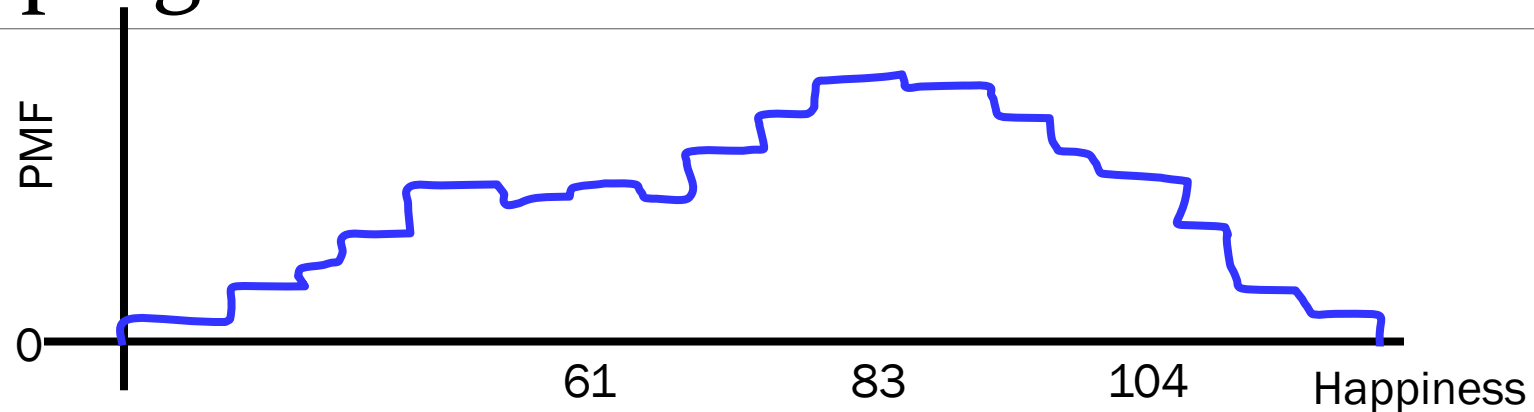
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Resample **len(sample)** from PMF
 - b. Recalculate the stat** on the resample
3. You now have a **distribution of your stat**

Bootstrapping of Means (we could do this with CLT)

Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the mean** on the resample
3. You now have a **distribution of your means**

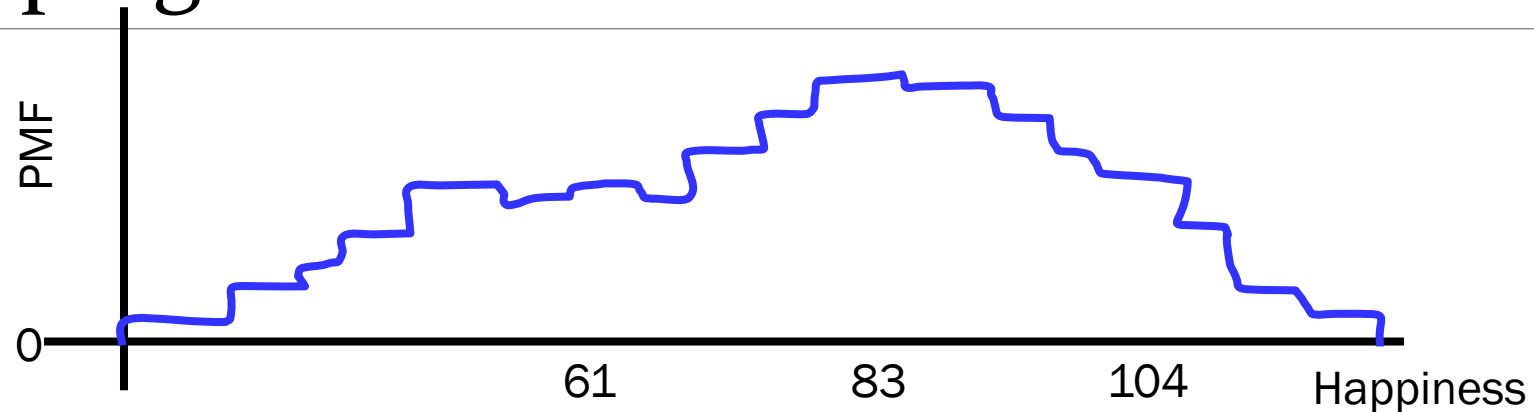
Bootstrapping of Means



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

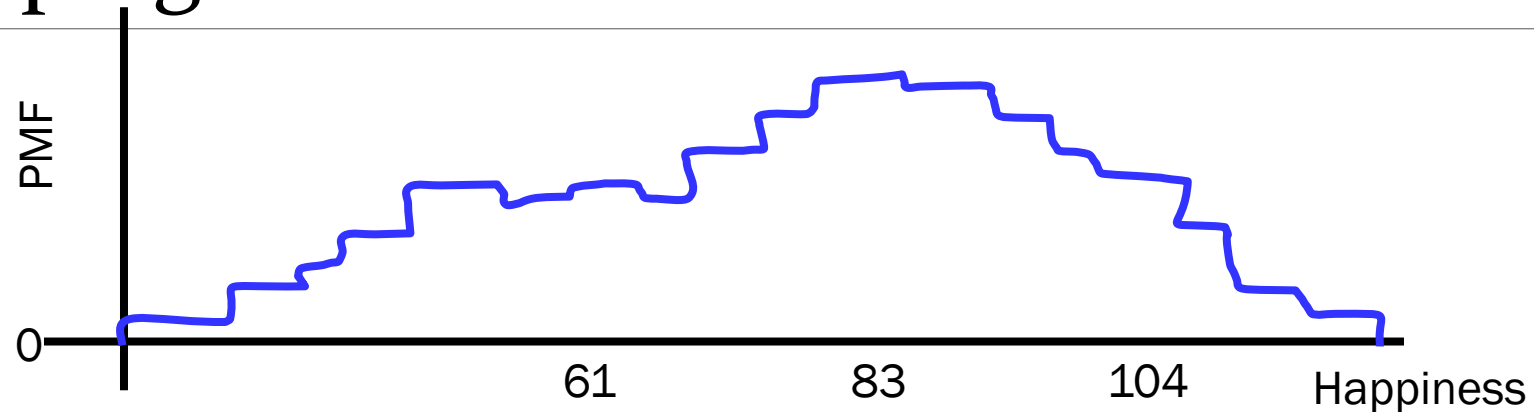
Bootstrapping of Means



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

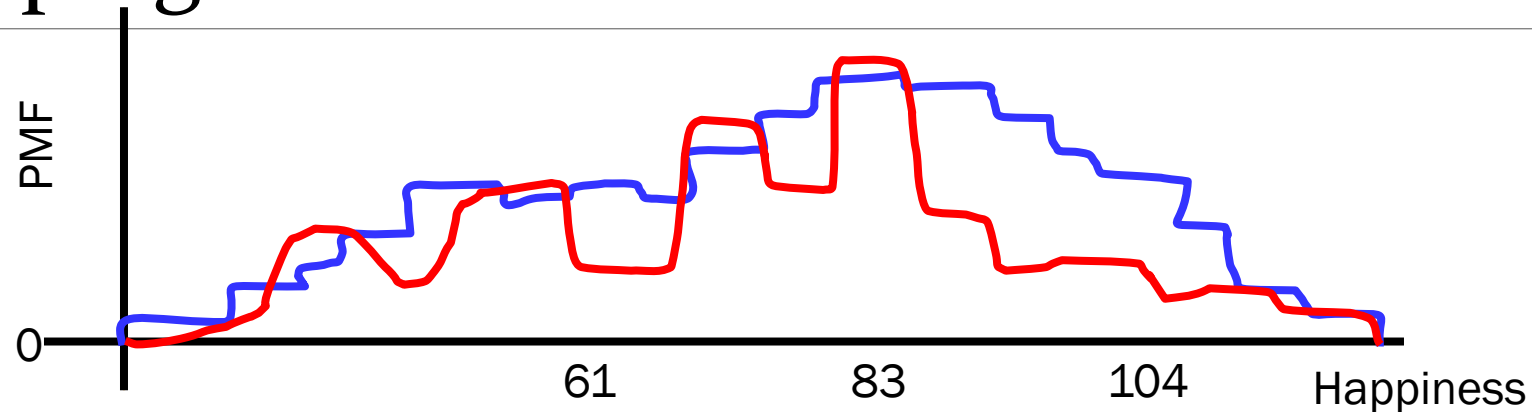
Bootstrapping of Means



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

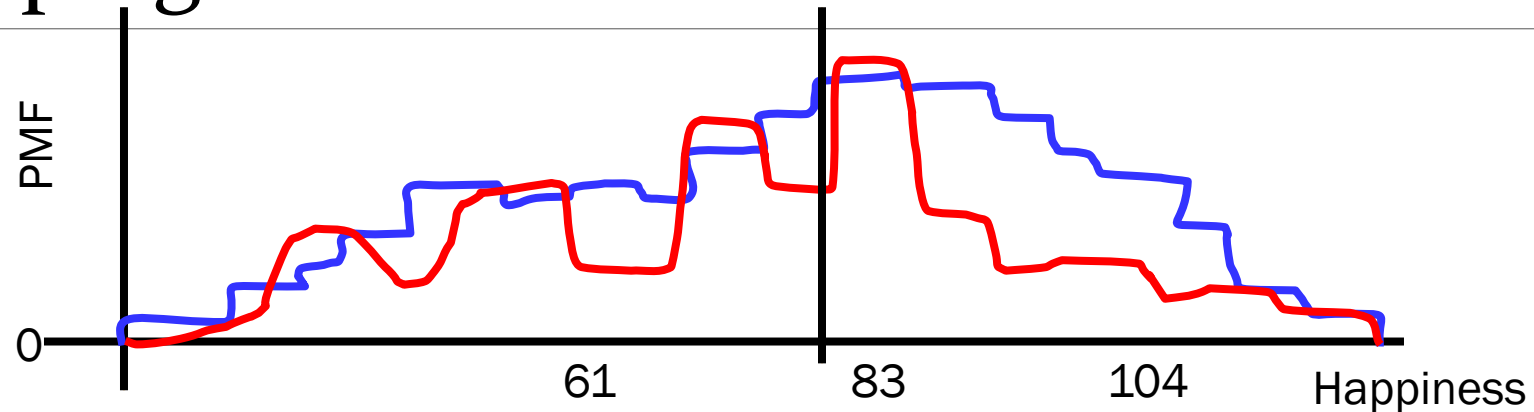
Bootstrapping of Means



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Bootstrapping of Means

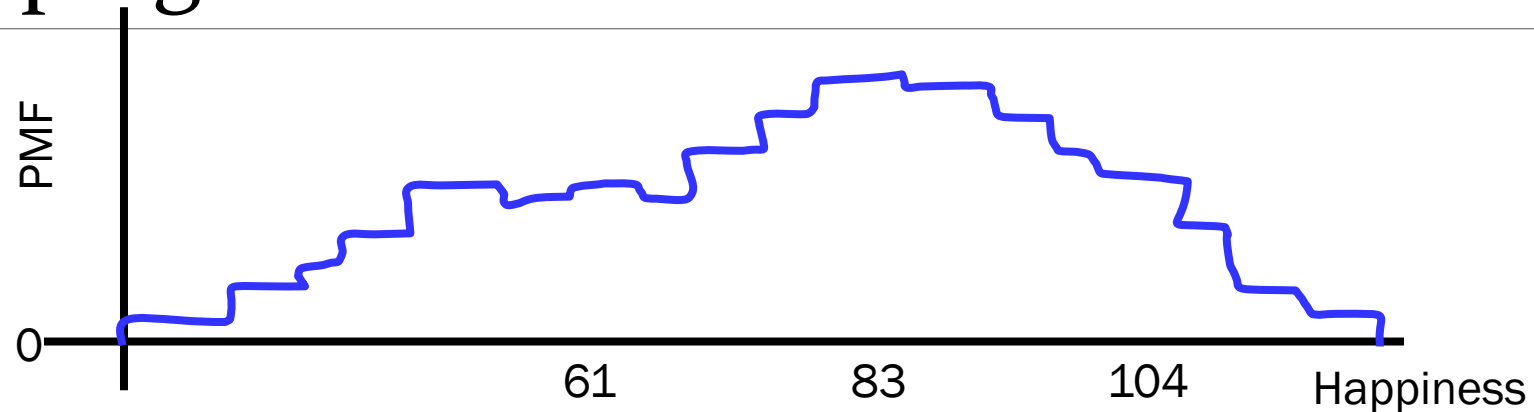


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means

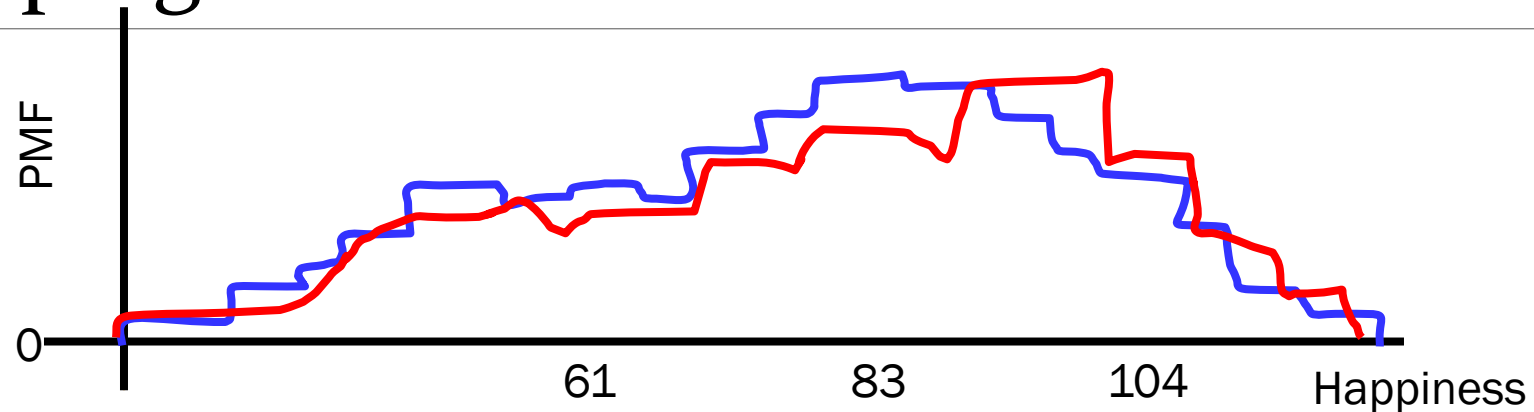


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means

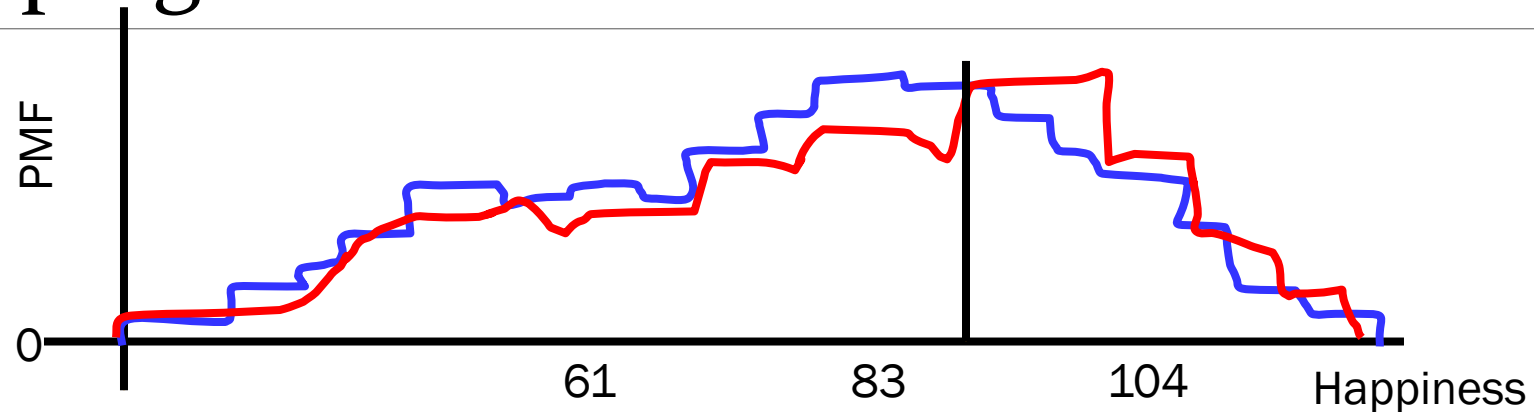


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means

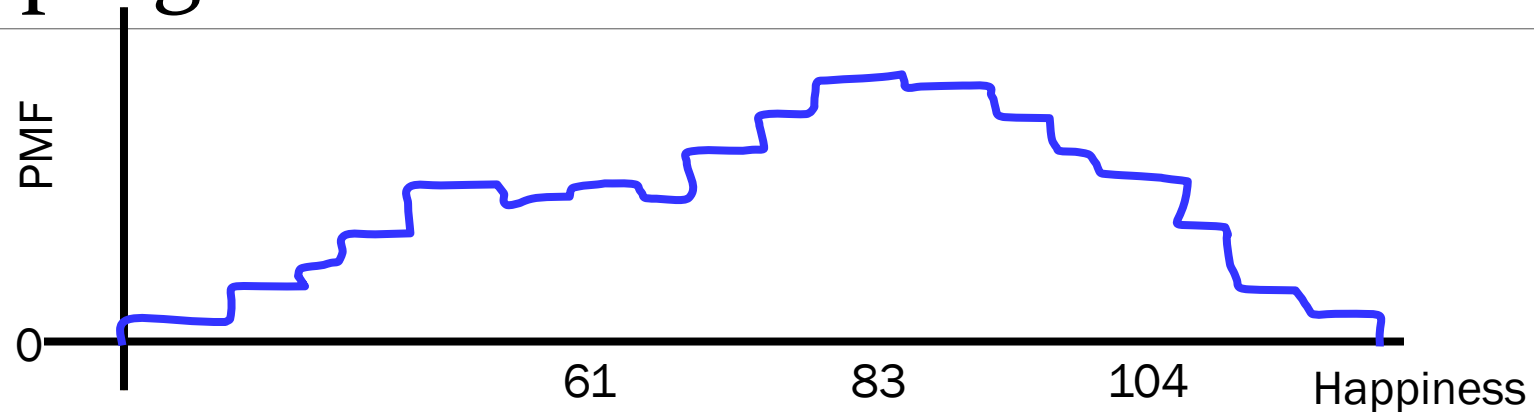


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7, 83.4]

Bootstrapping of Means

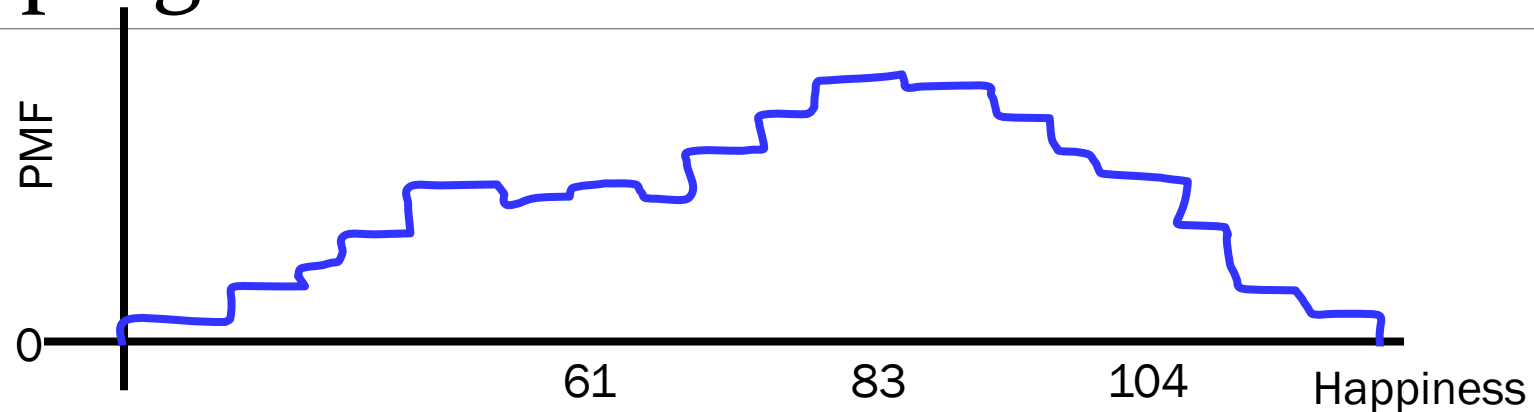


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7, 83.4]

Bootstrapping of Means



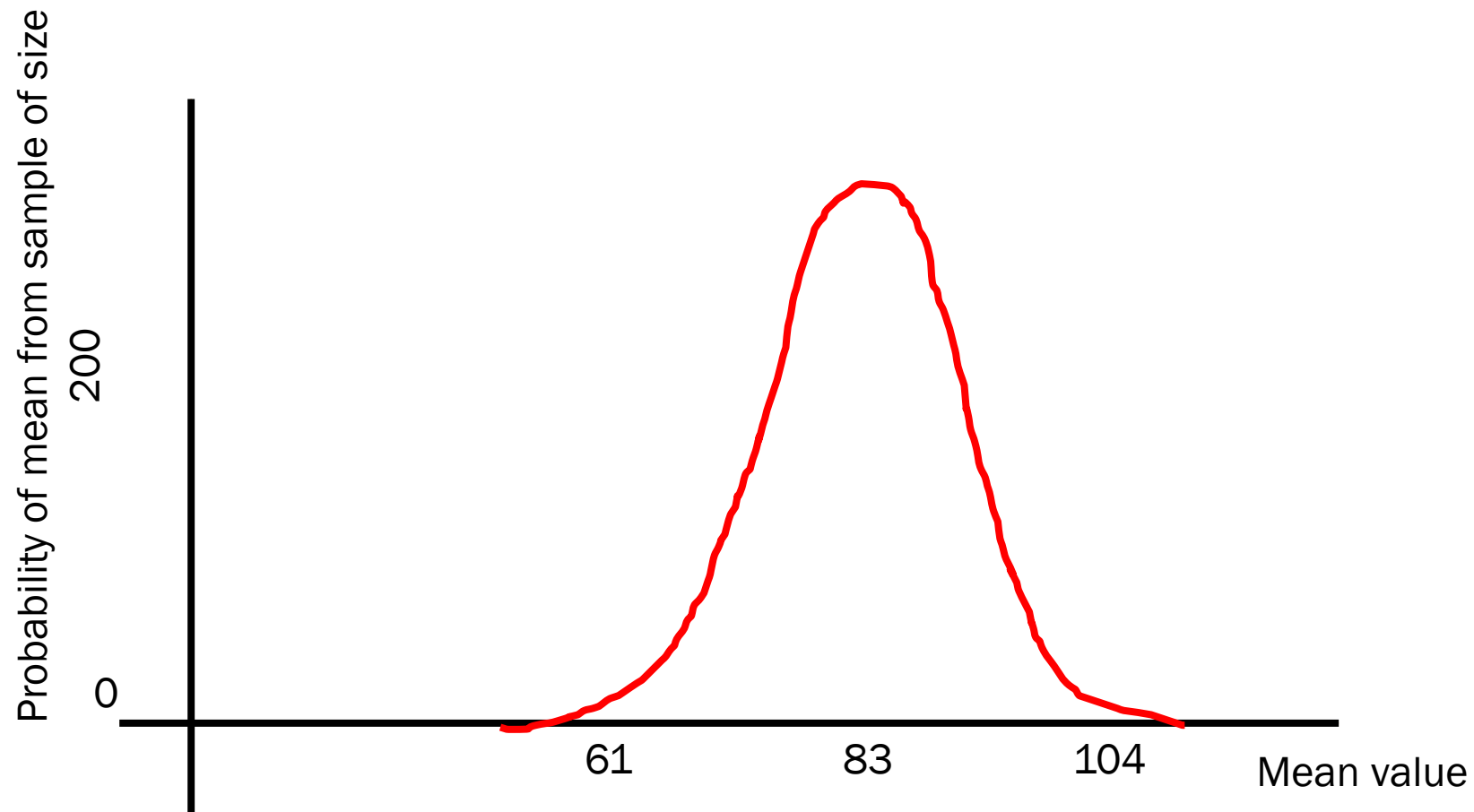
Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Means = [82.7, 83.4, 82.9, 91.4, 79.3, 82.1, ..., 81.7]

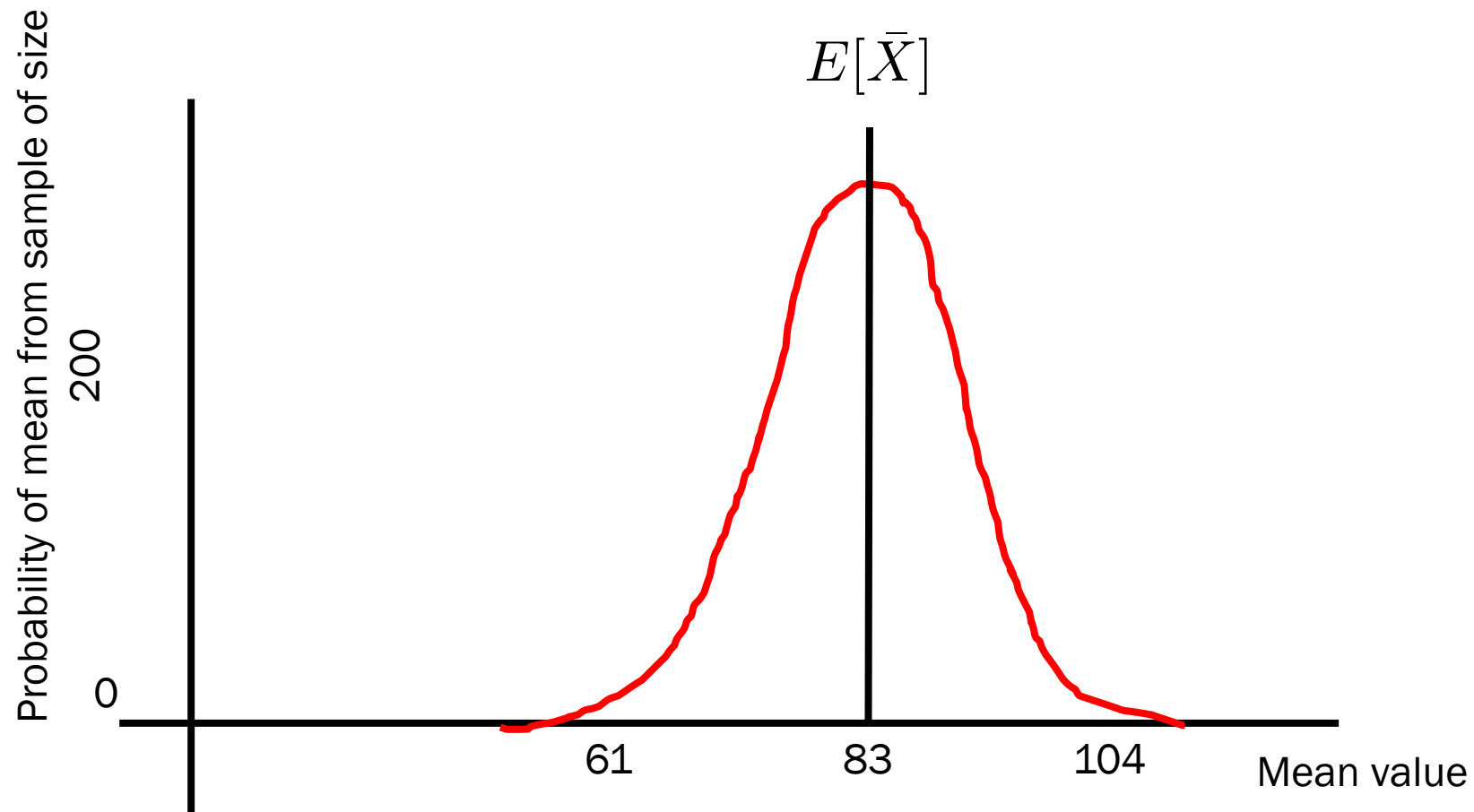
Bootstrapping of Means

Means = [82.7, 83.4, 82.9, 91.4, 79.3, 82.1, ..., 81.7]



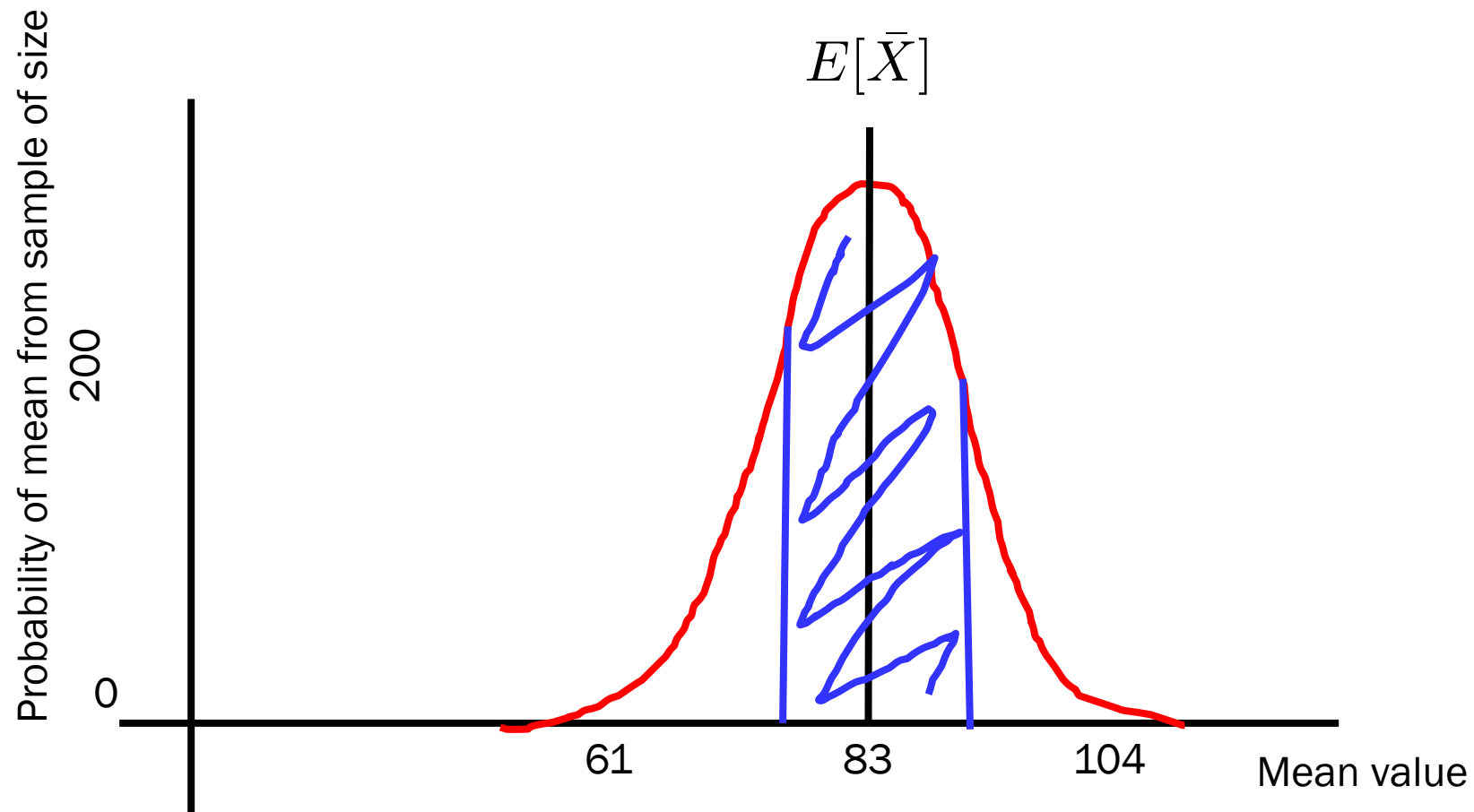
Bootstrapping of Means

Means = [82.7, 83.4, 82.9, 91.4, 79.3, 82.1, ..., 81.7]

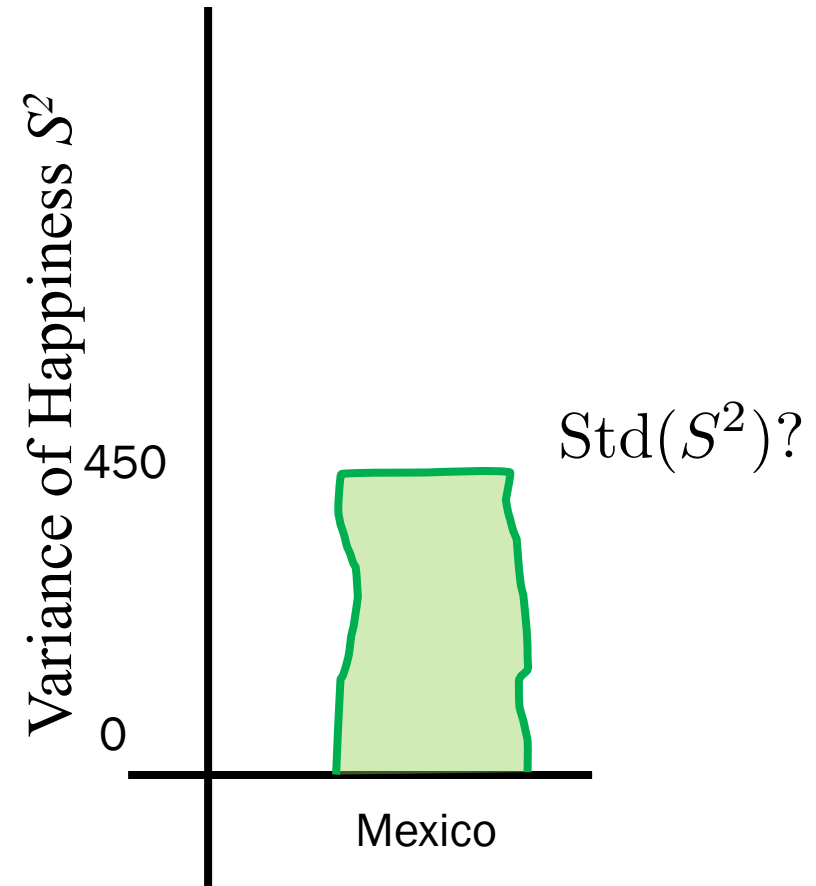
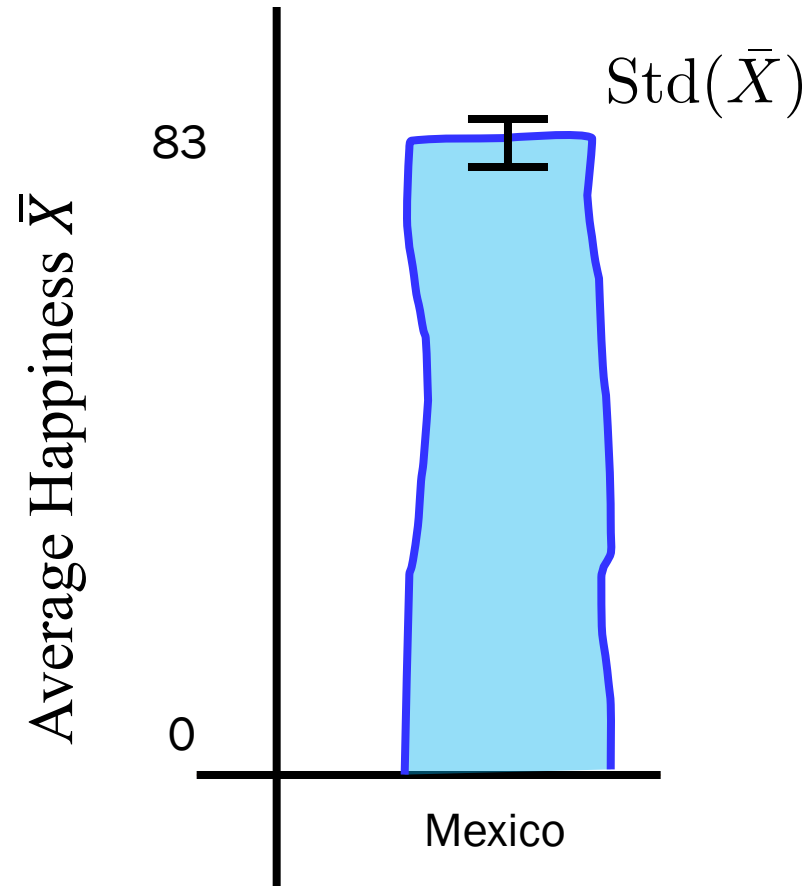


Bootstrapping of Means

What is the probability that the mean is in the range 81 to 85?



Our Report to Mexican Government



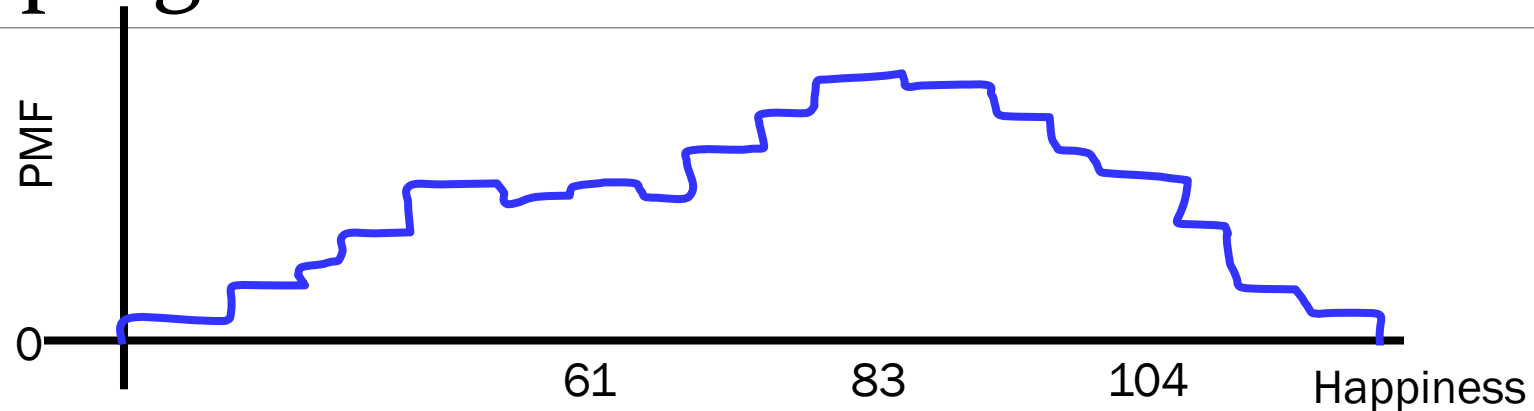
Claim: The average happiness of Mexico is 83 ± 2

Bootstrapping of Variance

Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the variance** on the resample
3. You have a **distribution of your variances**

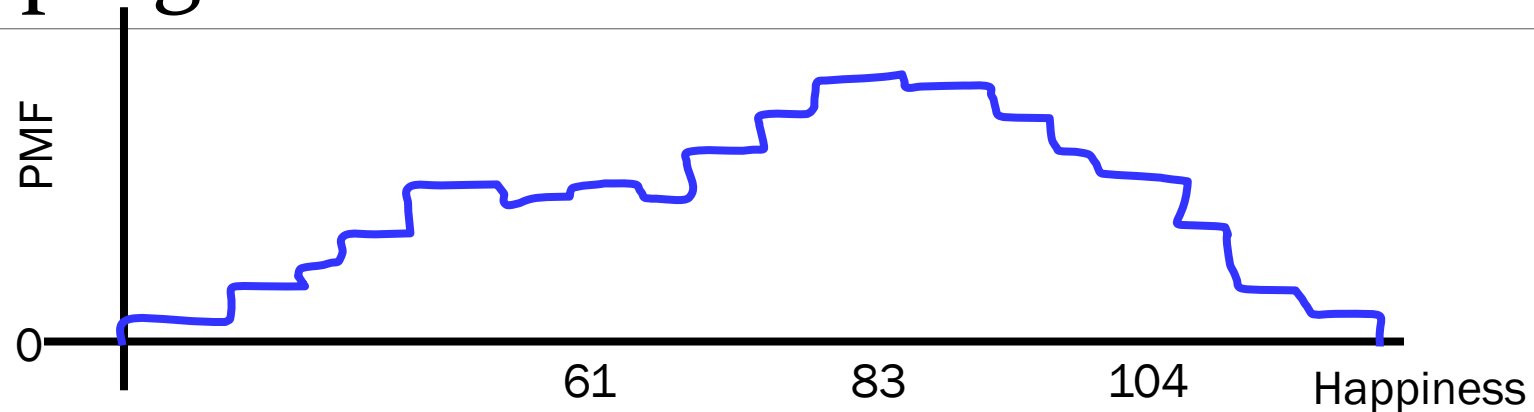
Bootstrapping of Variance



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the var** on the resample
3. You now have a **distribution of your vars**

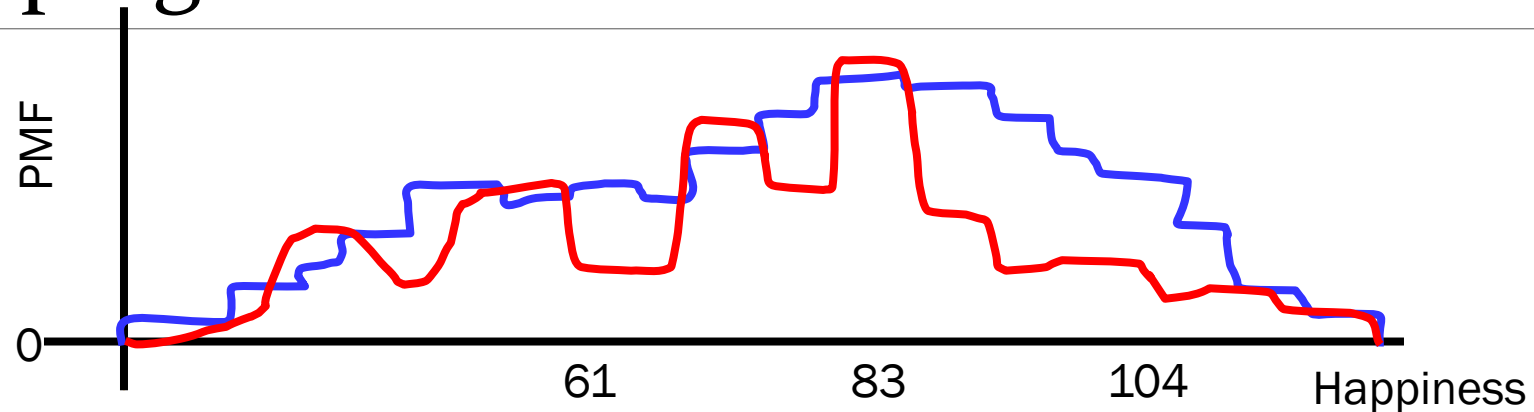
Bootstrapping of Variance



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the var** on the resample
3. You now have a **distribution of your vars**

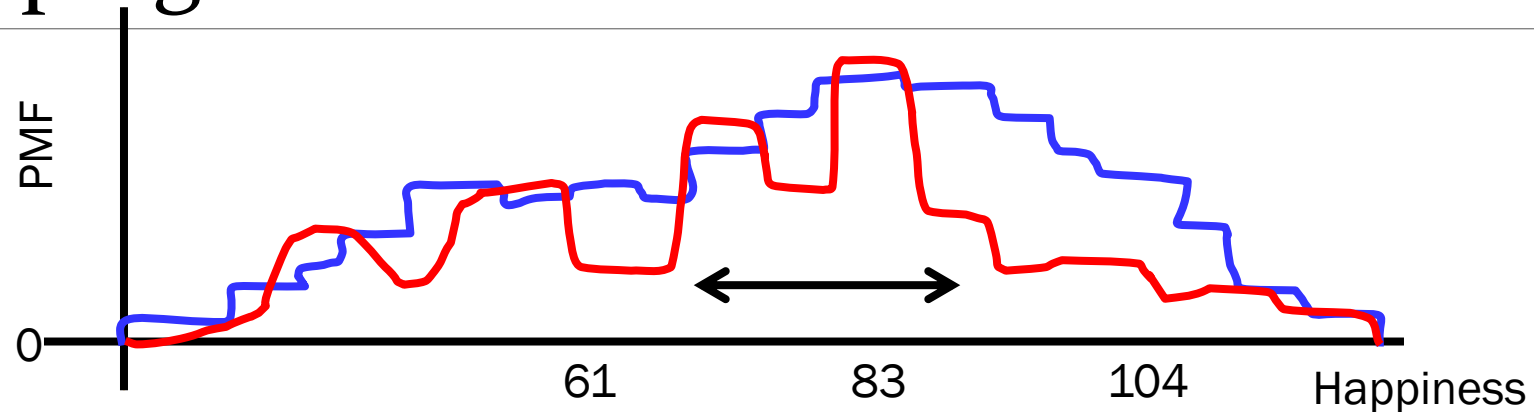
Bootstrapping of Variance



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the var** on the resample
3. You now have a **distribution of your vars**

Bootstrapping of Variance

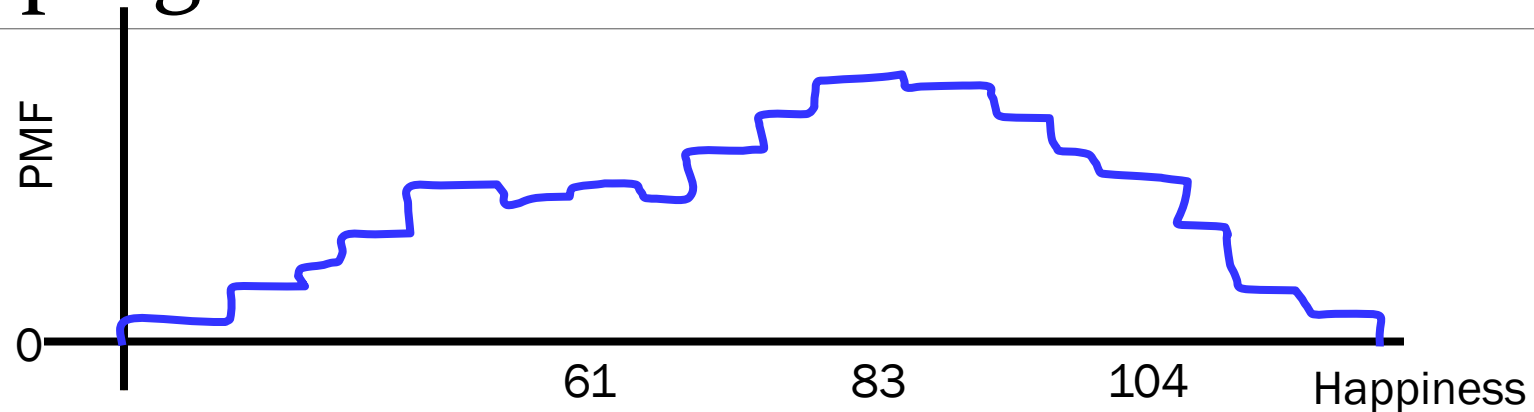


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the vars** on the resample
3. You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance

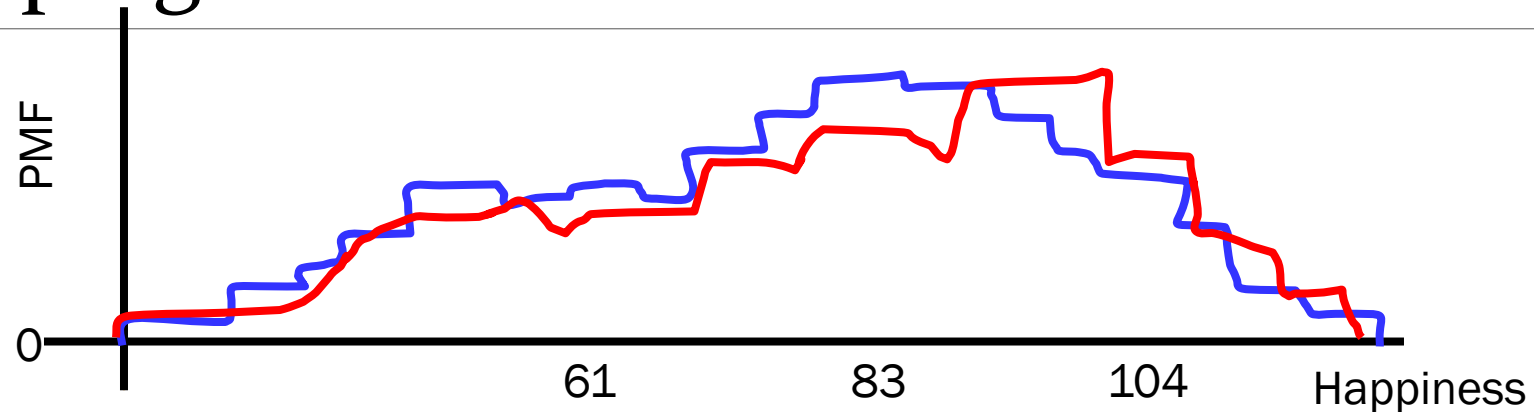


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the **var** on the resample
3. You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance

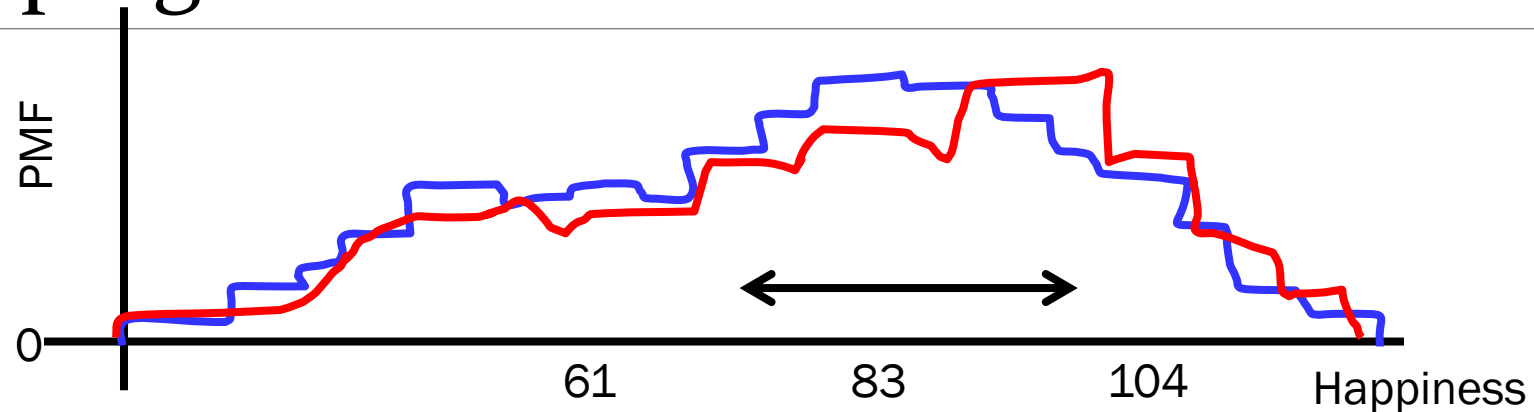


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the var** on the resample
3. You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance

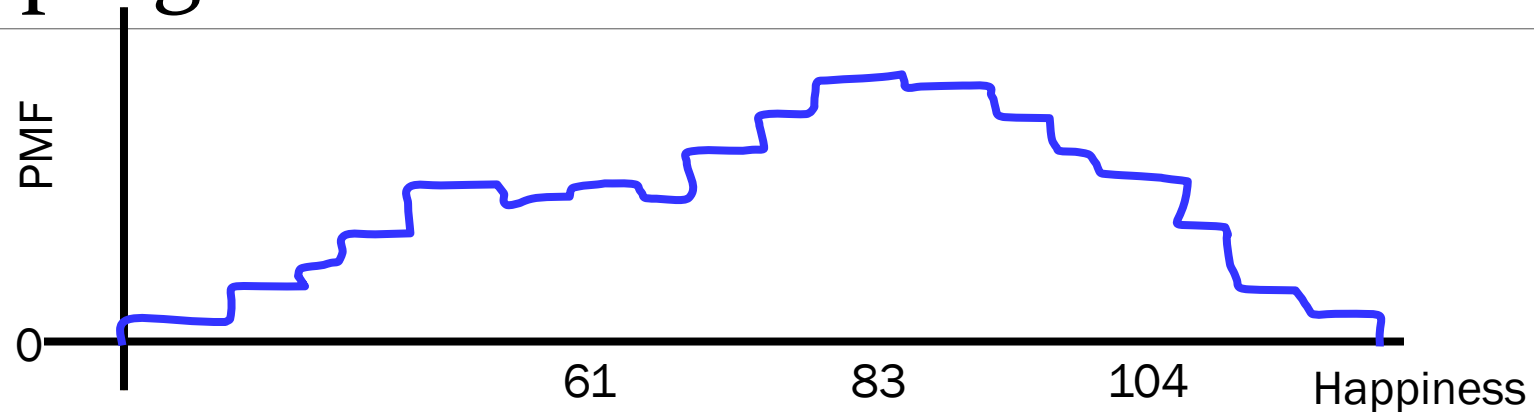


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. Recalculate the var** on the resample
3. You now have a **distribution of your vars**

Vars = [472.7, 478.4]

Bootstrapping of Variance

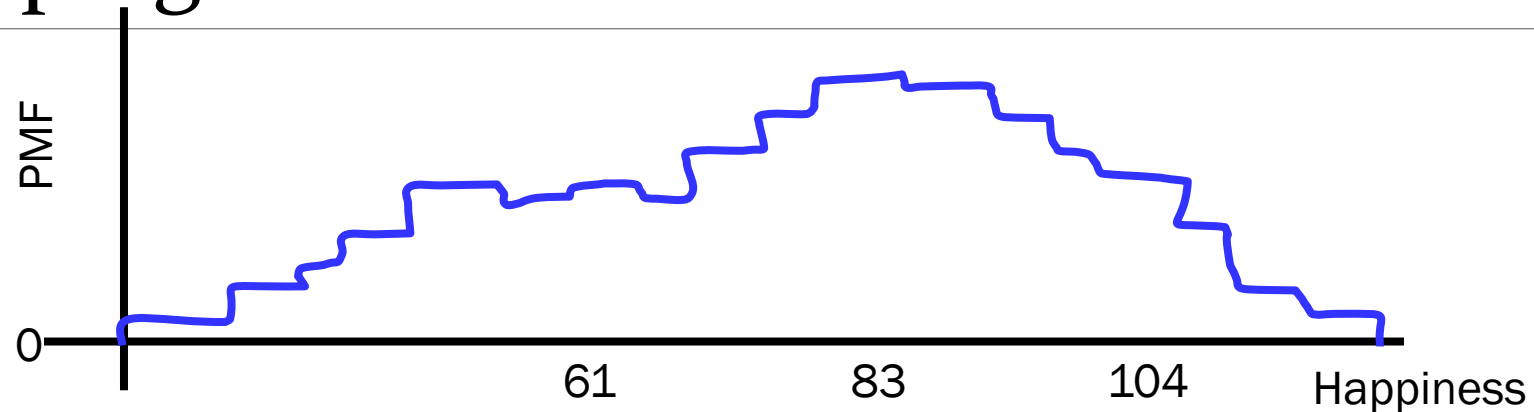


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the var** on the resample
3. You now have a **distribution of your vars**

Vars = [472.7, 478.4]

Bootstrapping of Variance



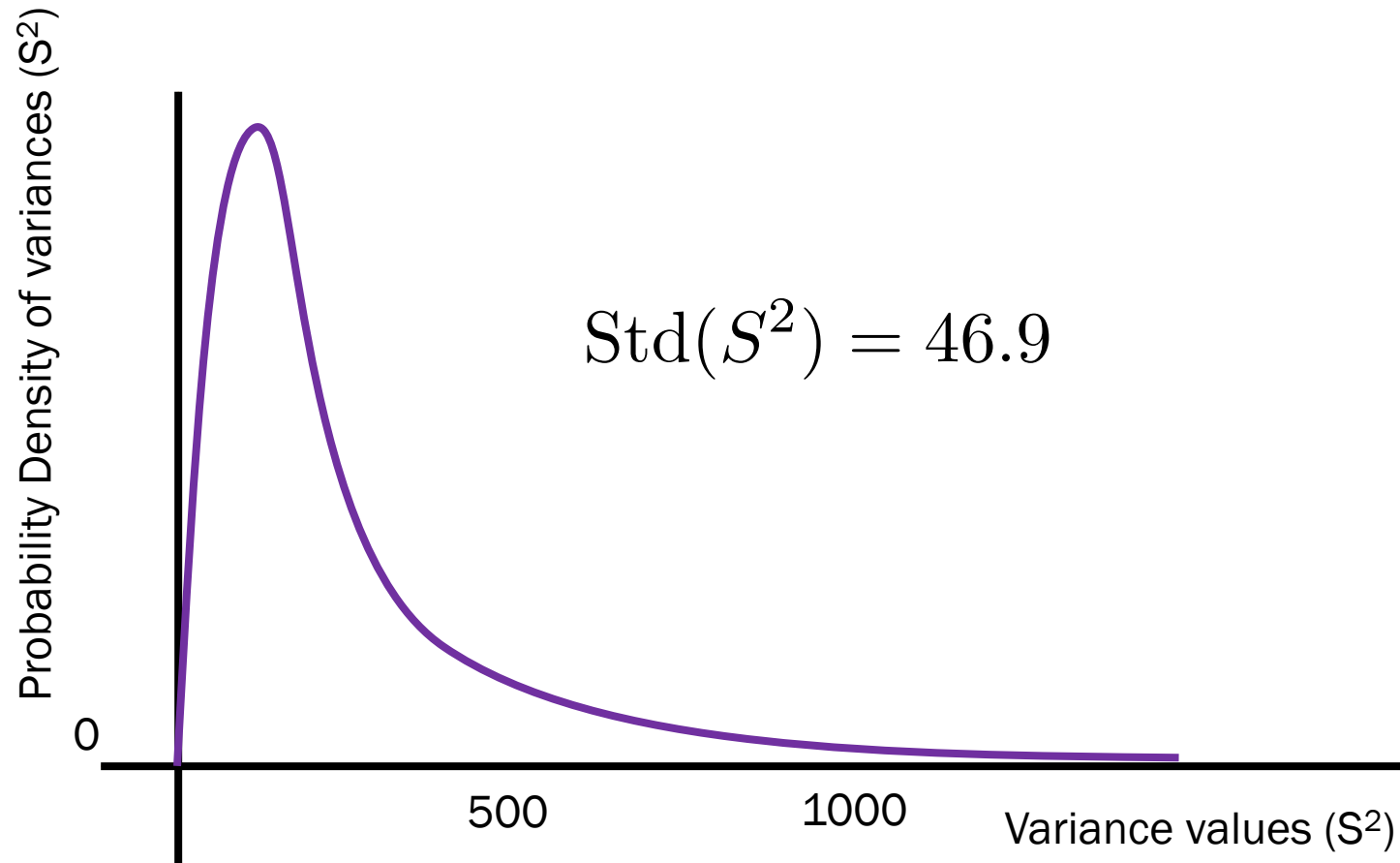
Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw **len(sample)** new samples from PMF
 - b. **Recalculate the var** on the resample
3. You now have a **distribution of your vars**

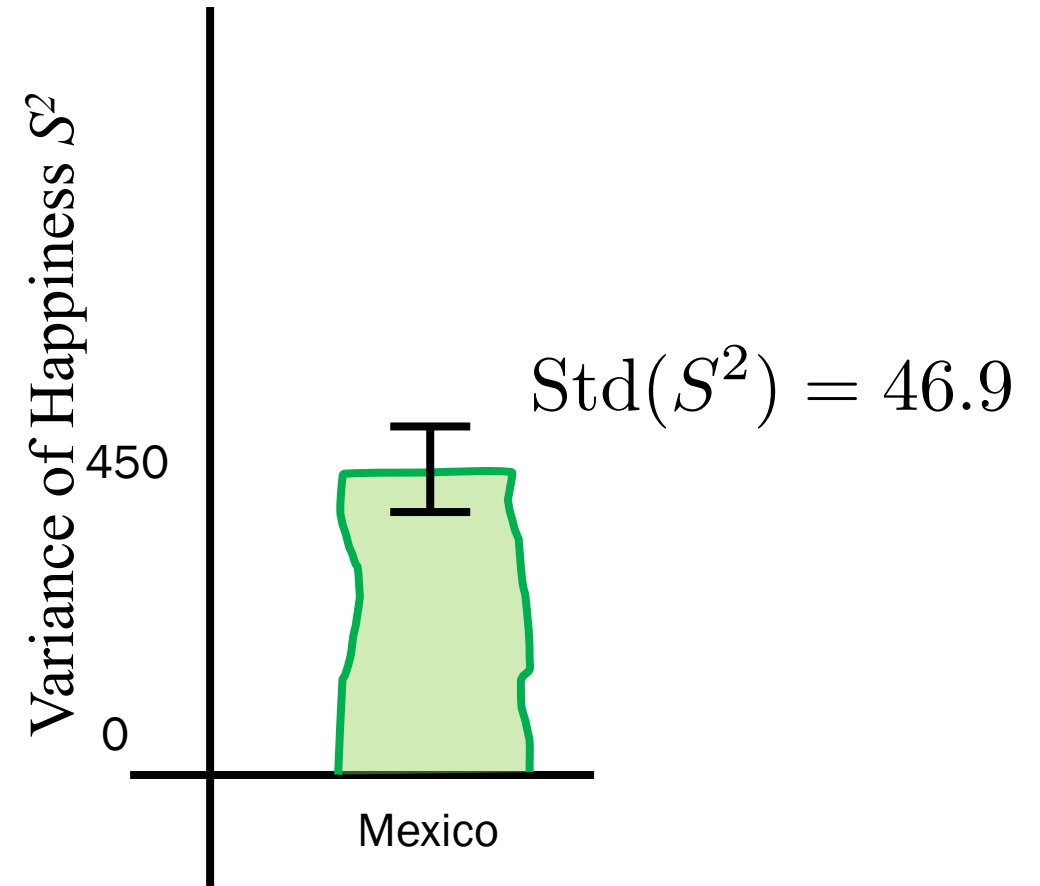
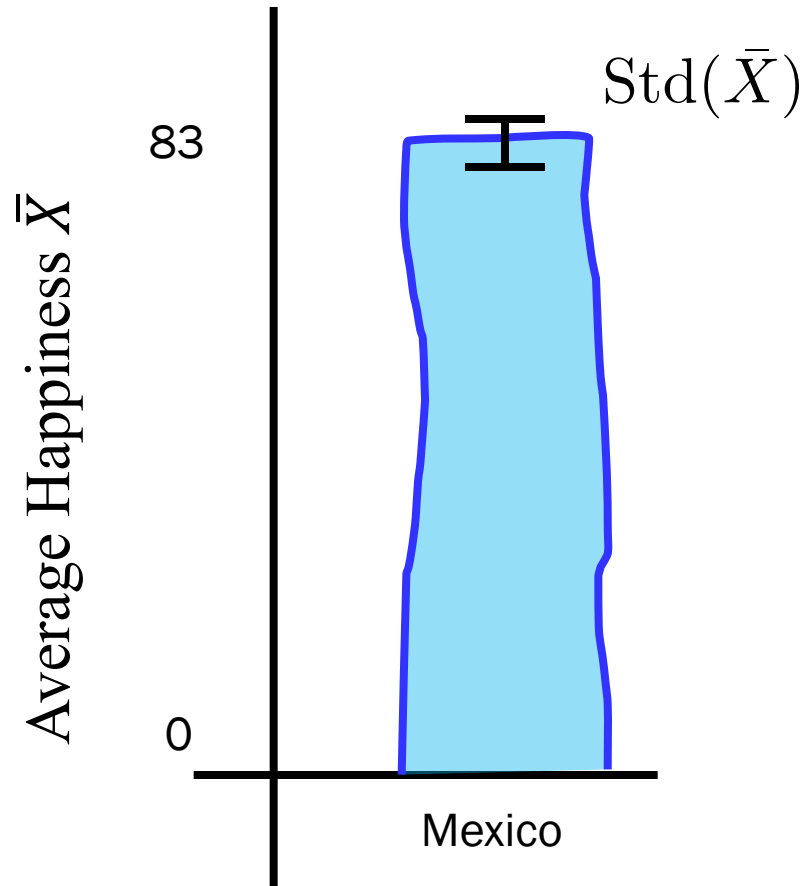
Vars = [472.7, 478.4, 469.2, ..., 476.2]

Bootstrapping of Variance

Sample Vars = [472.7, 478.4, 469.2, ..., 476.2]



Our Report to Mexico Government



Claim: The average happiness of Mexico is 83 ± 2

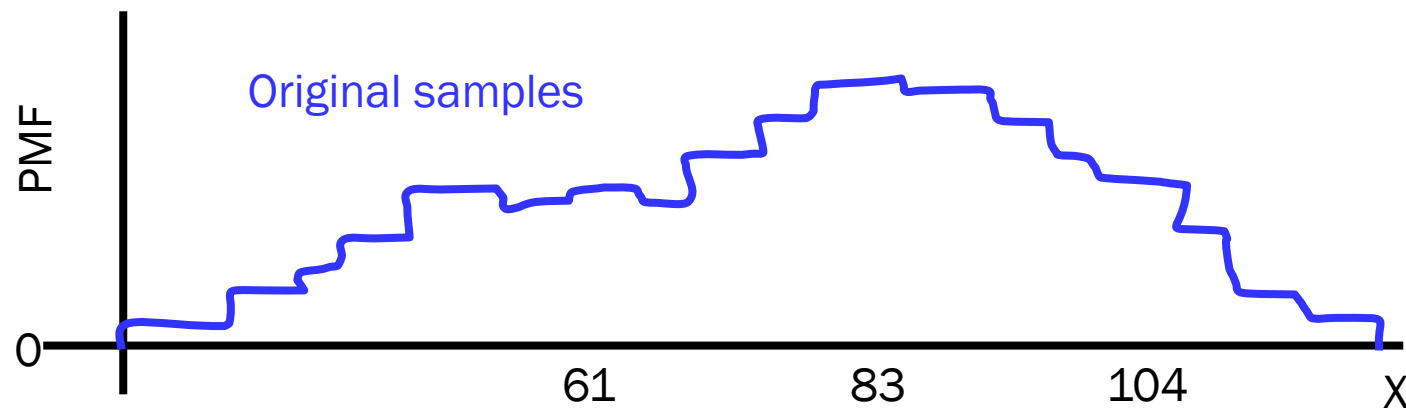
Pedagogical pause

Bootstrap Algorithm for $E[S^2]$ (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Draw `len(sample)` new samples from PMF
 - b. Recalculate the var** on the resample
3. You now have a **distribution of your vars**

Bootstrapping in Practice

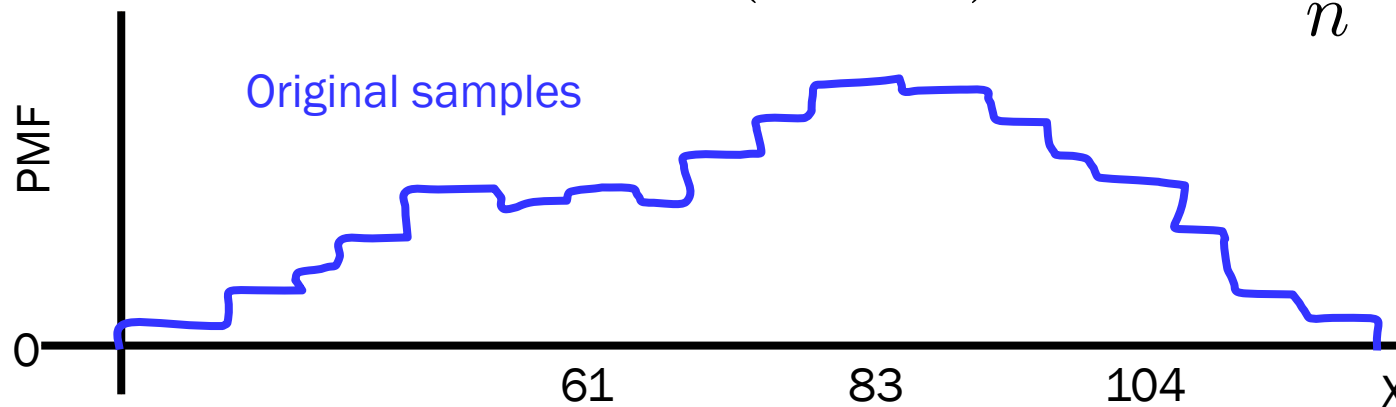
```
def resample(samples, K):  
    # Estimate the PMF using the samples  
    # Draw K new samples from the PMF
```



Bootstrapping in Practice

```
def resample(samples, K):  
    # Estimate the PMF using the samples  
    # Draw K new samples from the PMF  
    return np.random.choice(samples, K,  
                             replace = True)
```

$$P(X = k) = \frac{\text{count}(X = k)}{n}$$



OG Bootstrapping

Bootstrap Algorithm (sample):

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

Bootstrapping in Practice

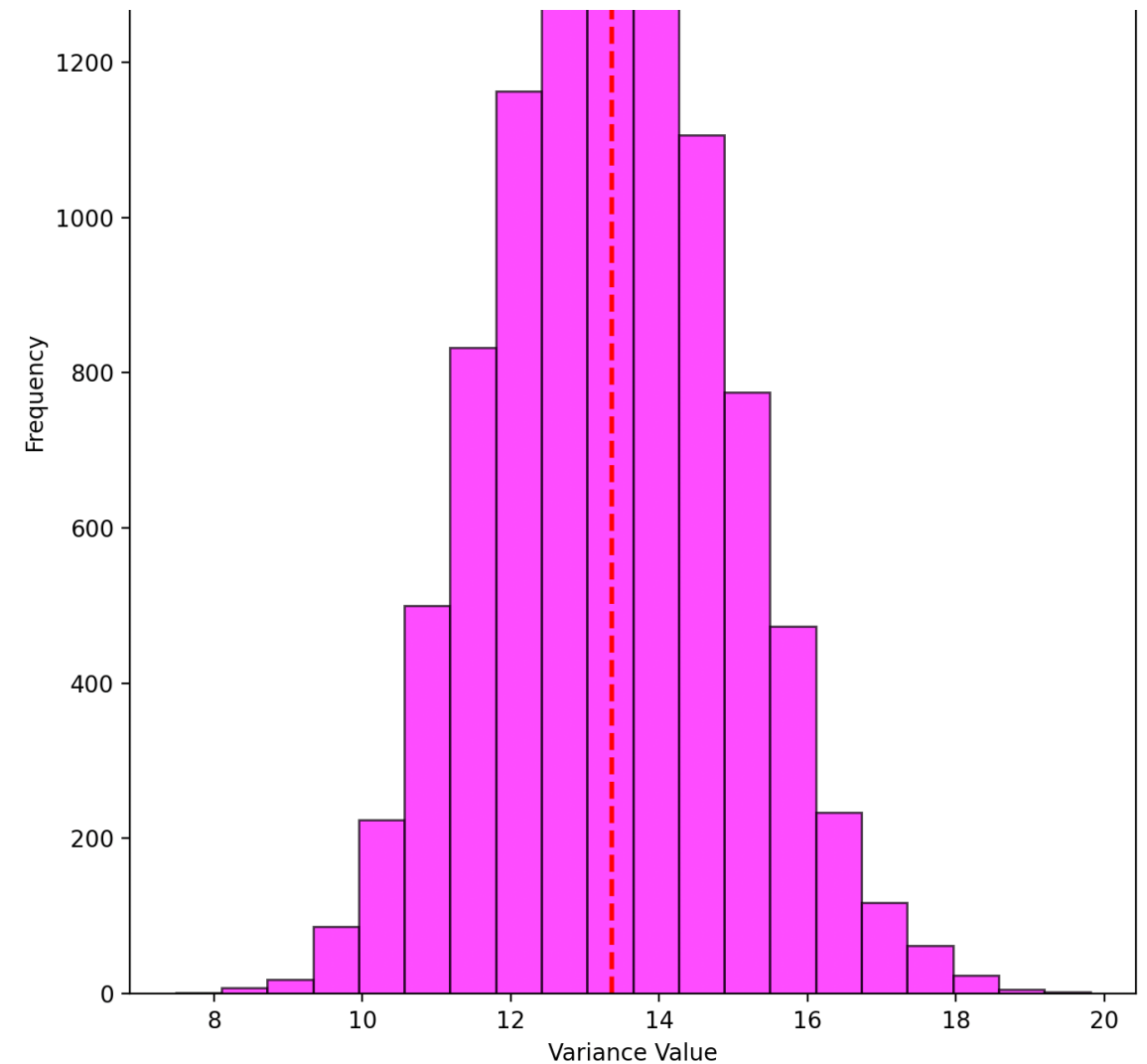
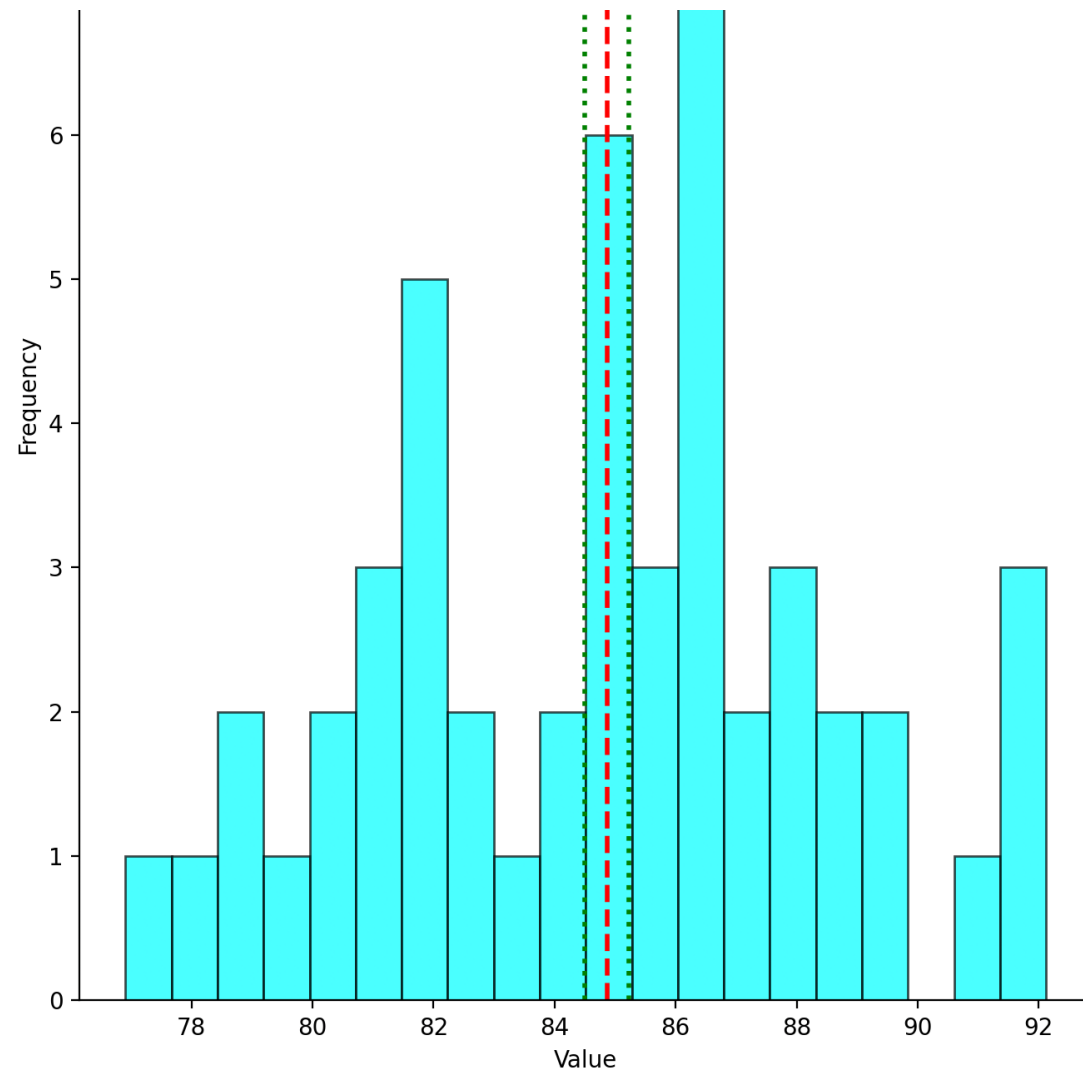
Bootstrap Algorithm (sample) :

1. Repeat **10,000** times:
 - a. Choose **len(sample)** elems from sample, **with replacement**
 - b. Recalculate the stat on the resample
2. You now have a **distribution of your stat**



To the code!

Our Distribution of the Sampling Variance





Bootstrap provides a way to calculate **probabilities of statistics** using code.

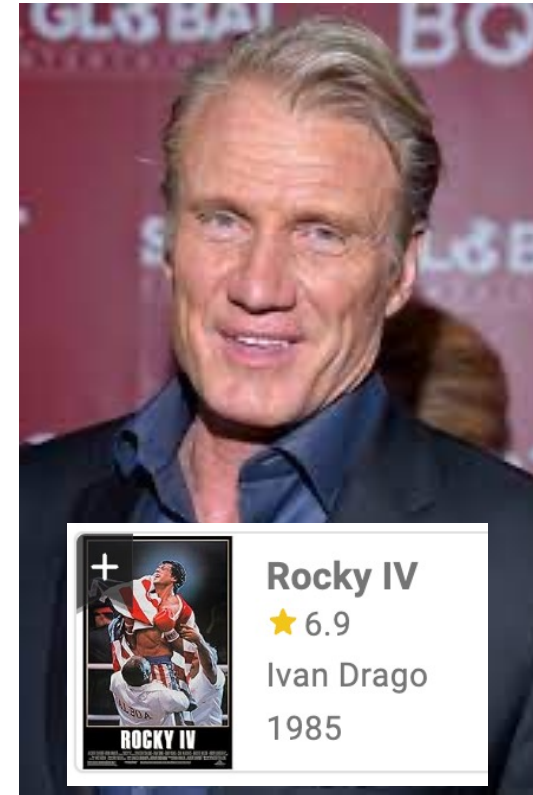
Bootstrap



Bradley Efron



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



According to starbyface.com:
Dolph Lundgren

Works for any statistic*

*as long as your samples are IID and the underlying distribution doesn't have a long tail

The Classic Science Test

Group 1	Group 2
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83

$\mu_1 = 3.1$ $\mu_2 = 2.4$

Claim: Group 1 and Group 2 are samples from **different distributions** with a 0.7 difference of means.

How confident are you in this claim?

A real difference?

	Learning in Context A	Learning in Context B	
18 students	4.44	2.15	23 students
	3.36	3.01	
	5.87	2.02	
	2.31	1.43	
	
	3.70	1.83	
	$\mu_1 = 3.1$	$\mu_2 = 2.4$	

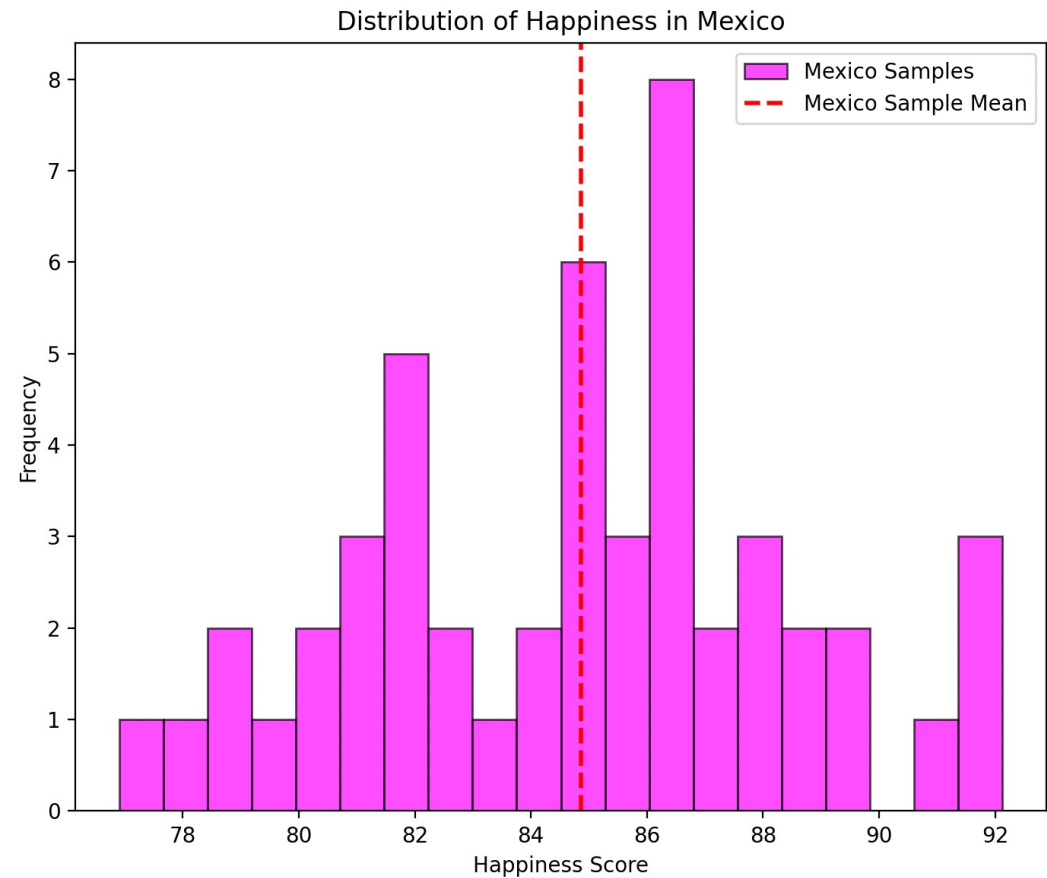
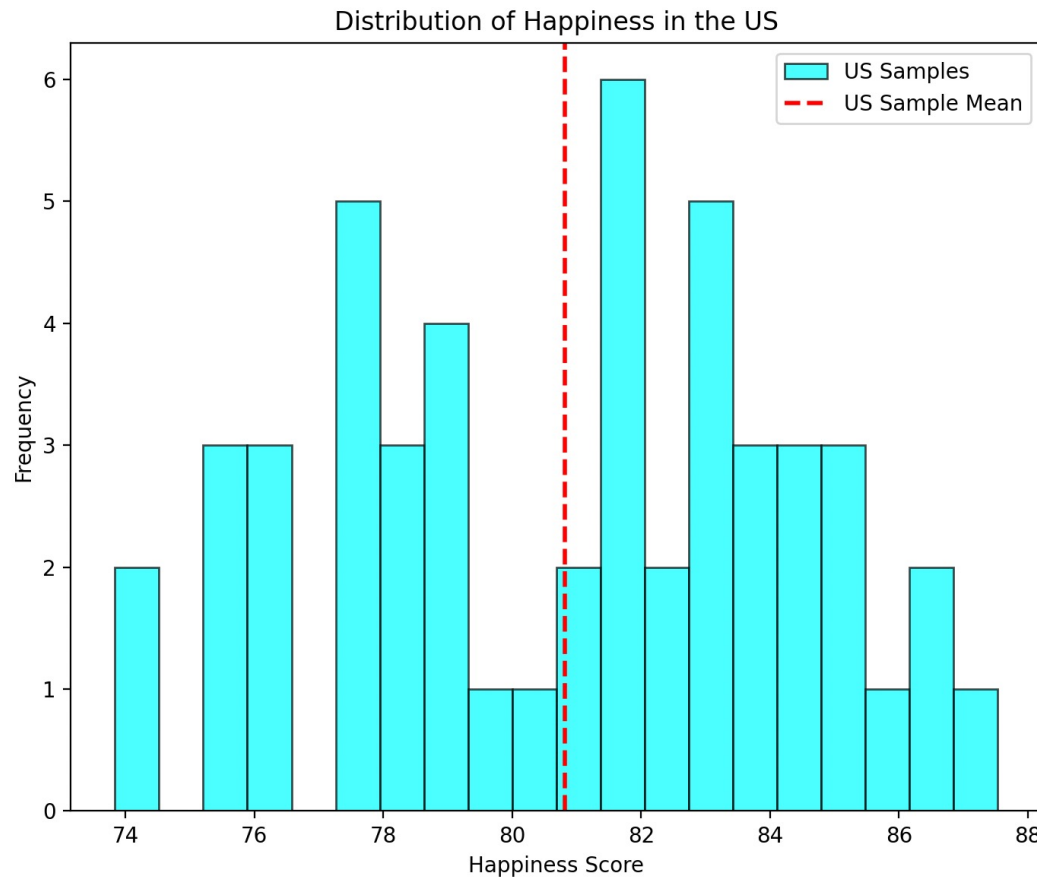
Claim: Group 1 and Group 2 are samples from **different distributions** with a 0.7 difference of means.

How confident are you in this claim?

A real difference? US vs Mexico

Mexico Sample Mean: 84.86

USA Sample Mean: 80.81

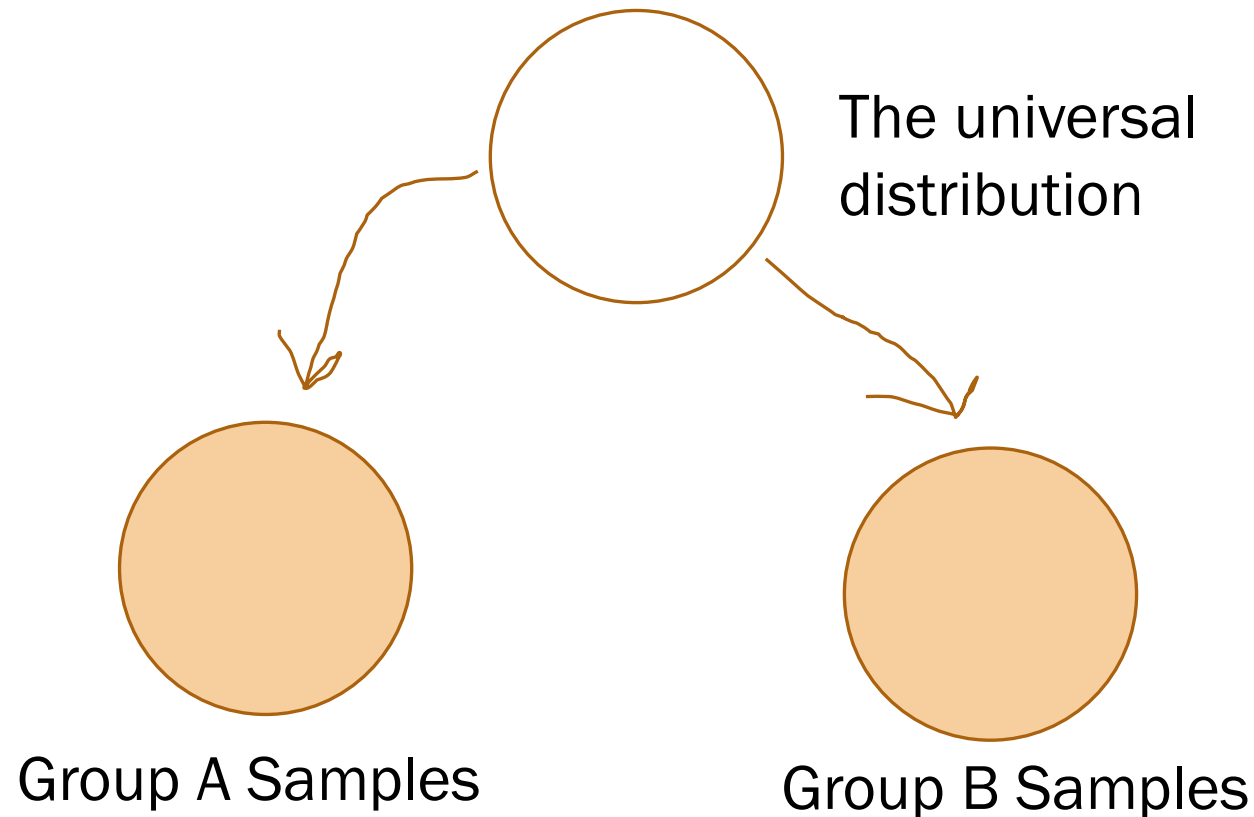


Claim: Group 1 and Group 2 are samples from **different distributions** with some 4.04 difference of means.

How confident are you in this claim?

The Null Hypothesis; Mexico and US no diff!

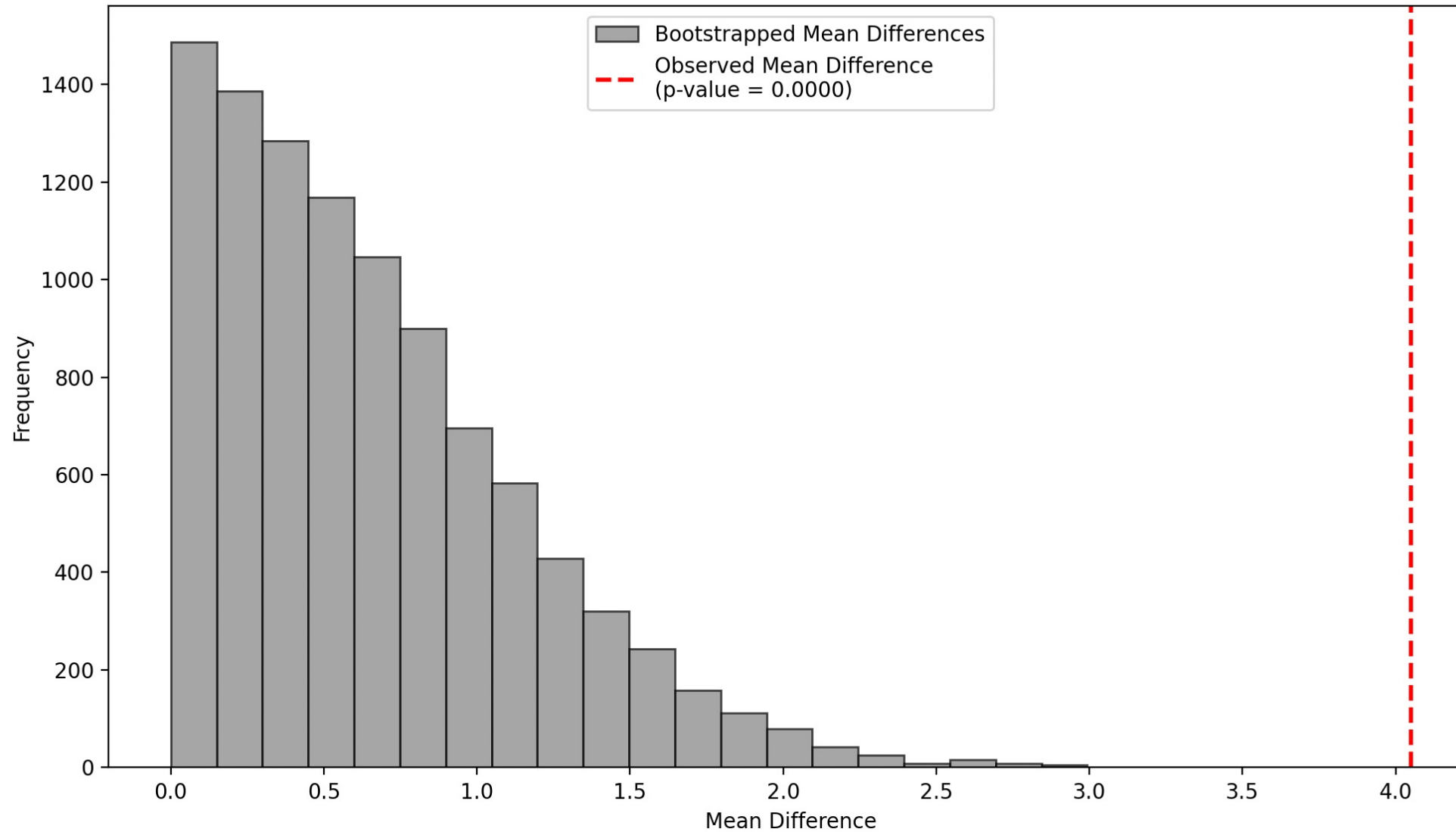
There is no difference between the two groups, so everyone is drawn from the same distribution. Any difference you observe is due to sampling error.



To the code!

Distribution of Mean Diffs under Null Hypothesis

Frequency of Observed Mean Differences Under Null Hypothesis



Food For Thought

Two Opinions on Distributions

Results of flipping a coin 20 times. Give your belief distribution of p :

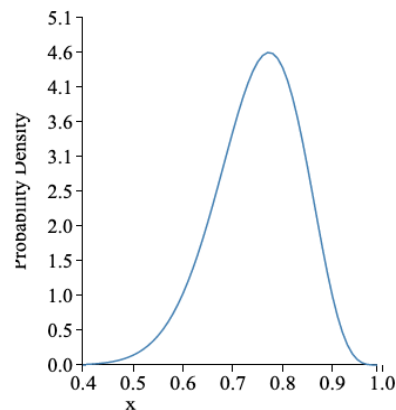
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

Let's use prior beliefs

$$X \sim \text{Beta}(a = 18, b = 6)$$



Frequentist:

Let's bootstrap

