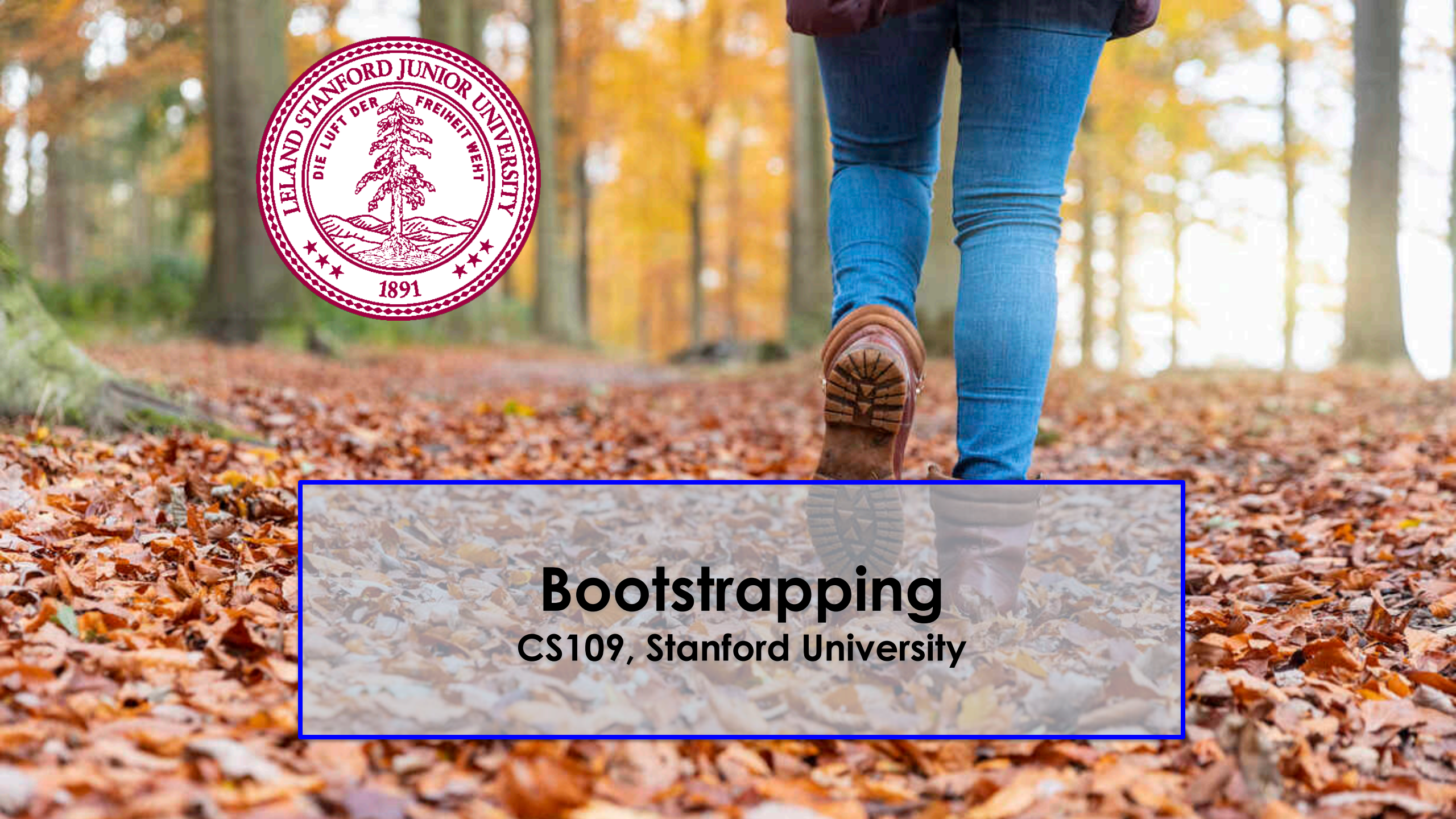


Bootstrapping

CS109, Stanford University

Where are we in CS109?

Uncertainty Theory

Beta Distributions

Thompson Sampling

Adding Random Vars

Central Limit Theorem

Sampling

Bootstrapping

Algorithmic Analysis

A real difference?

Learning in Context A	Learning in Context B
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83

18 students

23 students

$\mu_1 = 3.1$

$\mu_2 = 2.4$

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?

<review>

Central Limit Theorem (Summation)

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

As $n \to \infty$

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

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

Central Limit Theorem (Average)

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

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

The average of the variables is normally distributed

Simulations and Indicator Random Variables

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

Consider an event E. An Indicator Random Variable for E is defined as:

$$I_E = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Fact 1: $\mathbb{E}[I_E] = P(E)$

$$\begin{aligned} \mathbb{E}[I_E] &= 1 \cdot P(I_E = 1) + 0 \cdot P(I_E = 0) \\ &= 1 \cdot P(E) + 0 \cdot P(E^C) = P(E) \end{aligned}$$

Fact 2: $Var(I_E) = P(E)\big(1 - P(E)\big) < 1$

$$\frac{\#\text{ of times } E \text{ occurred}}{\#\text{ of Trials}} = \frac{\sum_{i=0}^{n} I_{E,i}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{E,i} \sim \mathcal{N}\left(\mathbb{E}[I_E], \frac{Var(I_E)}{n}\right) = \mathcal{N}\left(P(E), \frac{Var(I_E)}{n}\right)$$

Simulations and Indicator Random Variables

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

Consider an event E. An Indicator Random Variable for E is defined as:

$$I_E = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Fact 1: $\mathbb{E}[I_E] = P(E)$

$$\begin{aligned} \mathbb{E}[I_E] &= 1 \cdot P(I_E = 1) + 0 \cdot P(I_E = 0) \\ &= 1 \cdot P(E) + 0 \cdot P(E^C) = P(E) \end{aligned}$$

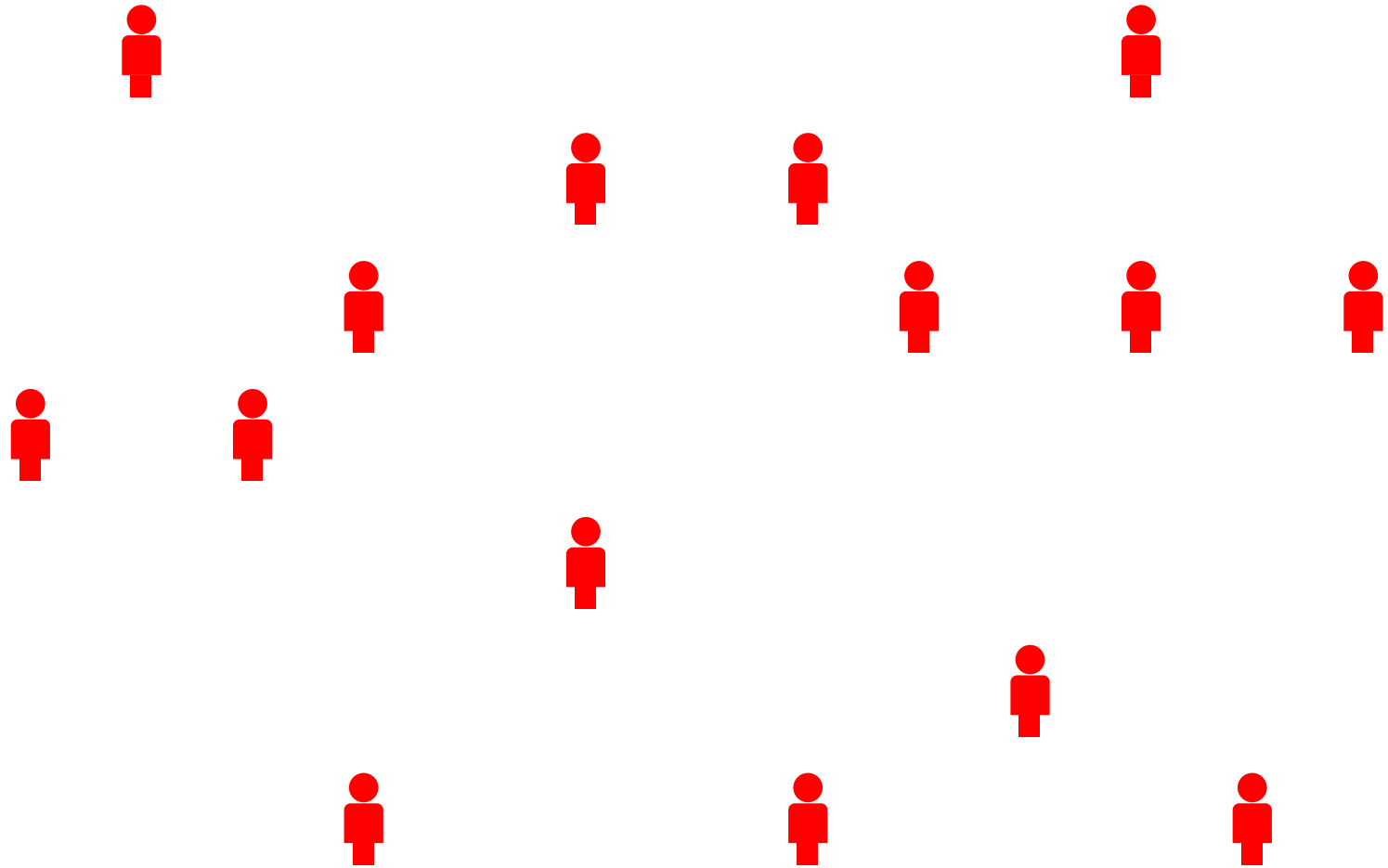
Fact 2: $Var(I_E) = P(E)\big(1 - P(E)\big) < 1$

$$\frac{\#\text{ of times } E \text{ occurred}}{\#\text{ of Trials}} \sim \mathcal{N}\left(P(E), \frac{Var(I_E)}{n}\right)$$

Population

Sample

Sample



Collect one (or more) numbers from each person

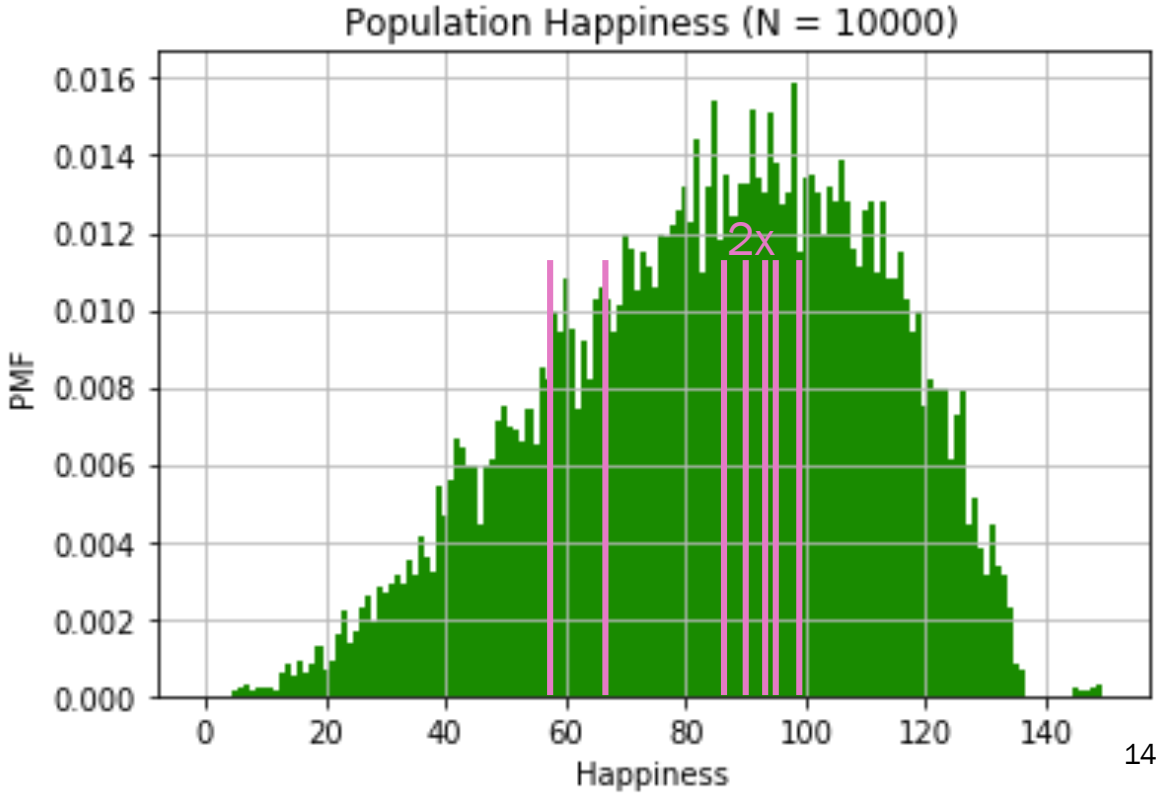
A sample, mathematically

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

sample mean

sample variance estimate

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

Our best guess at the true variance

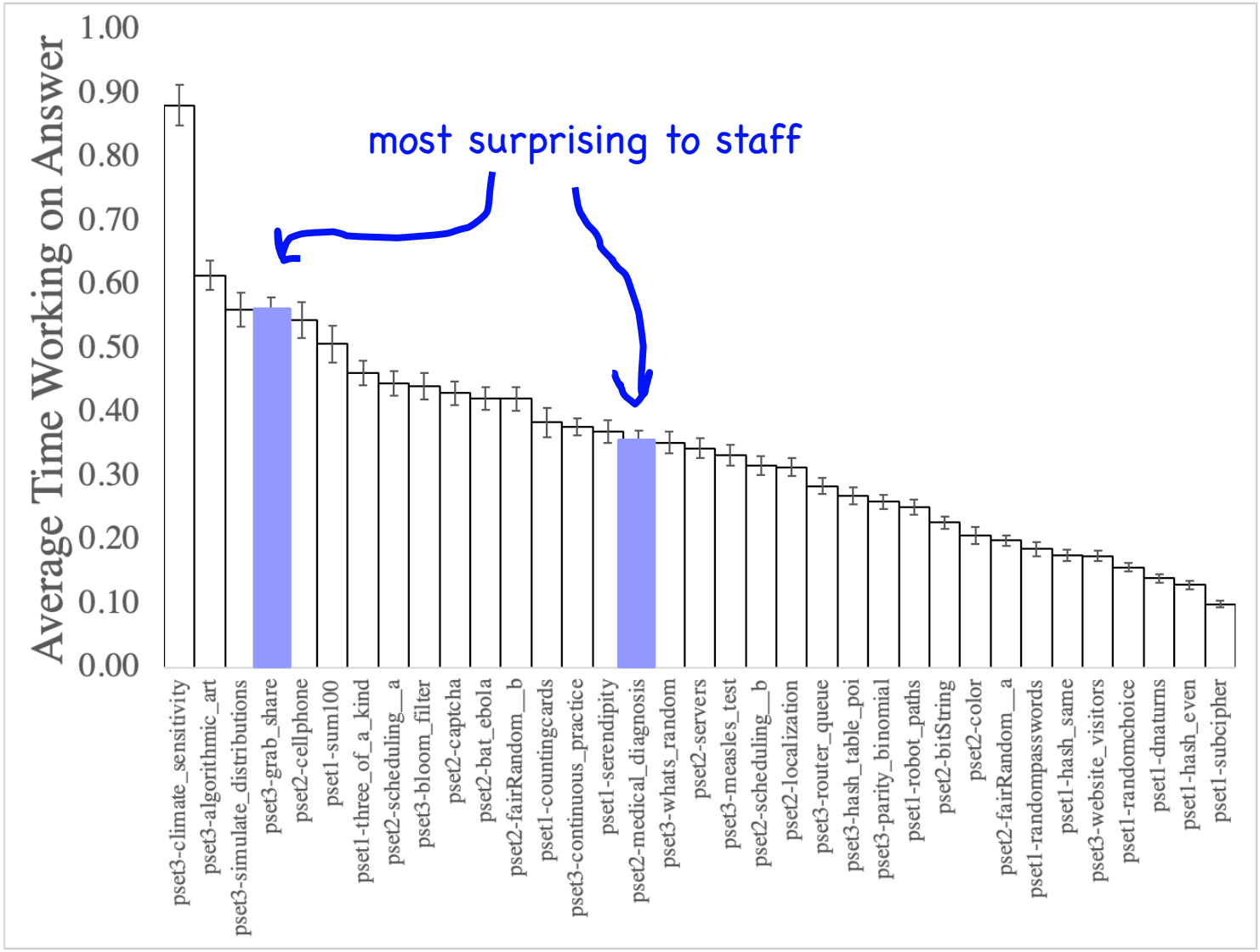
Std error of the mean estimate

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

sample variance

How wrong do we think our mean estimate could be?

Sample Mean and Standard Error for PSets



Error bars are standard error of the mean

Expectation of the sum of problems is sum of expectations:

pset1: 2.87 hours on answers
pset2: 4.23 hours on answers
pset3: 5.11 hours on answers

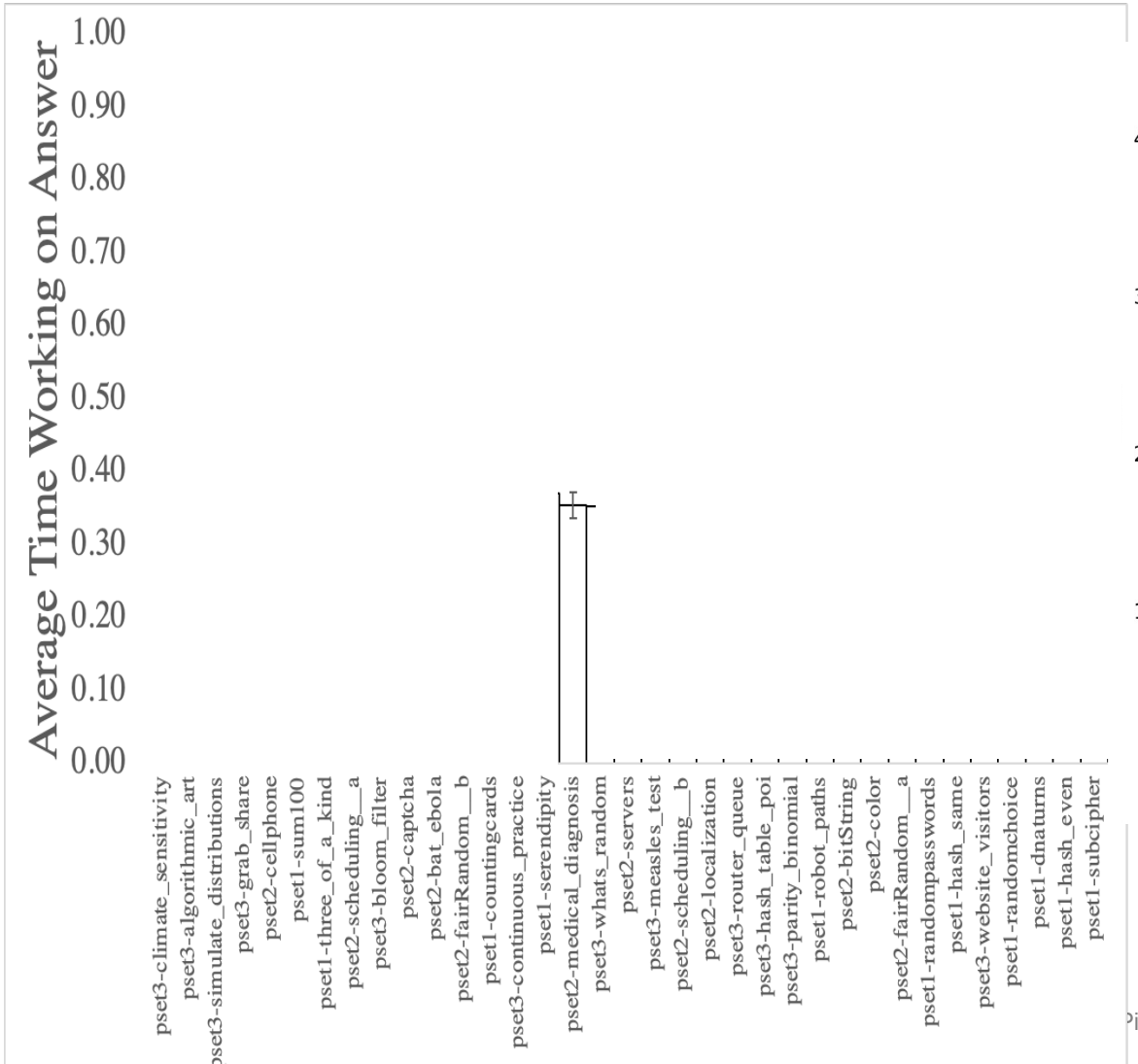
Total: 12.1 hours on answers
Budget: 50 hours for psets

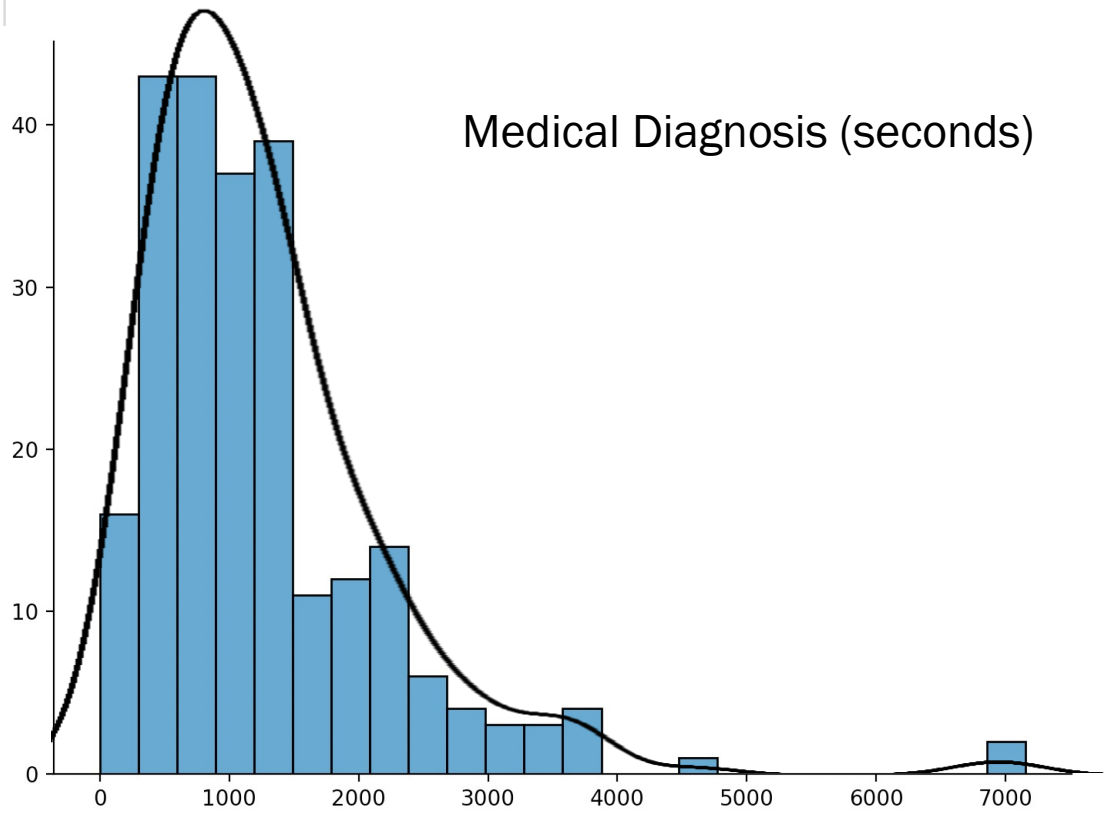
Statistics Vs Distribution

Sampling statistics

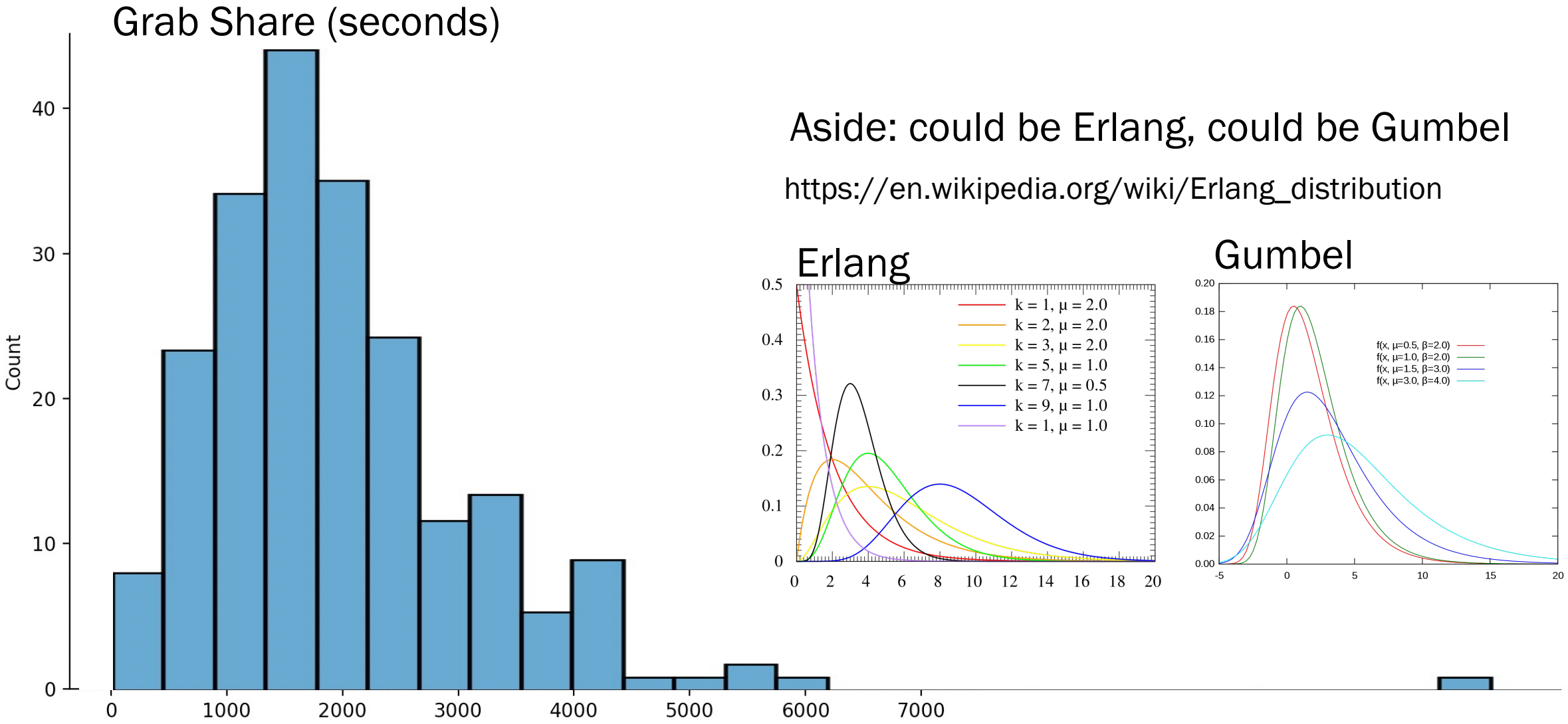
vs

Sampling distribution





[Aside] Distribution of PSet Completion Times



Aside: could be Erlang, could be Gumbel

https://en.wikipedia.org/wiki/Erlang_distribution

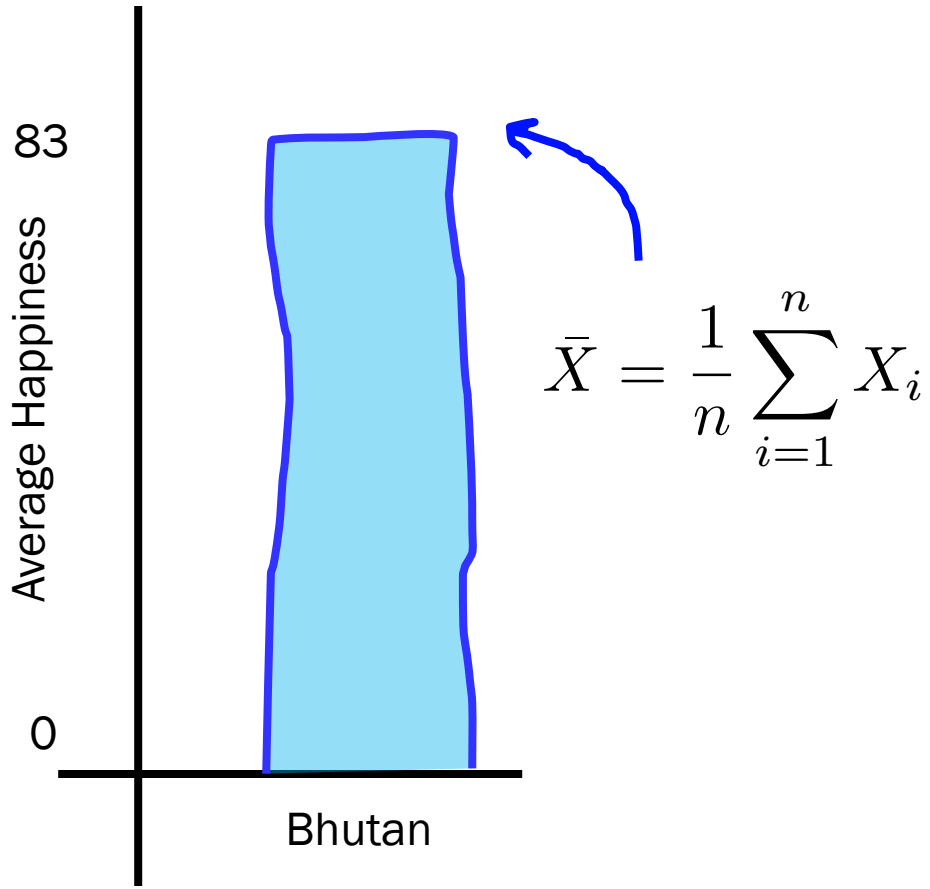
Erlang

Gumbel

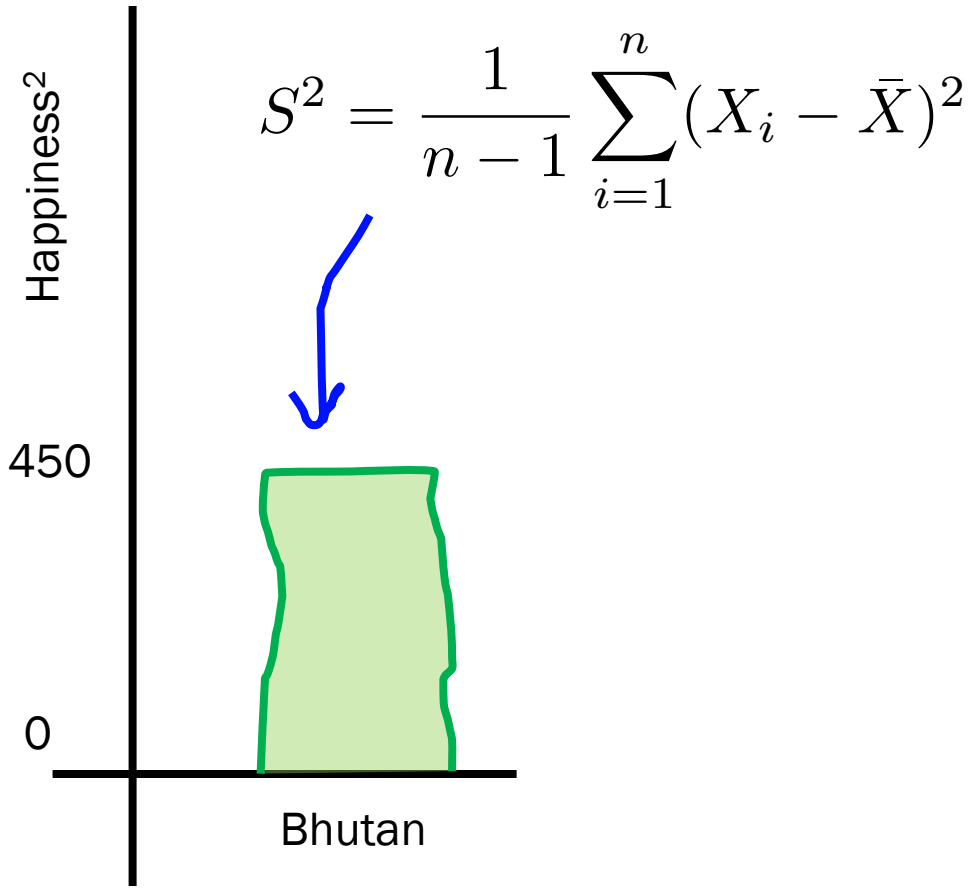
But what about Bhutan?

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

Average Happiness



Variance of Happiness

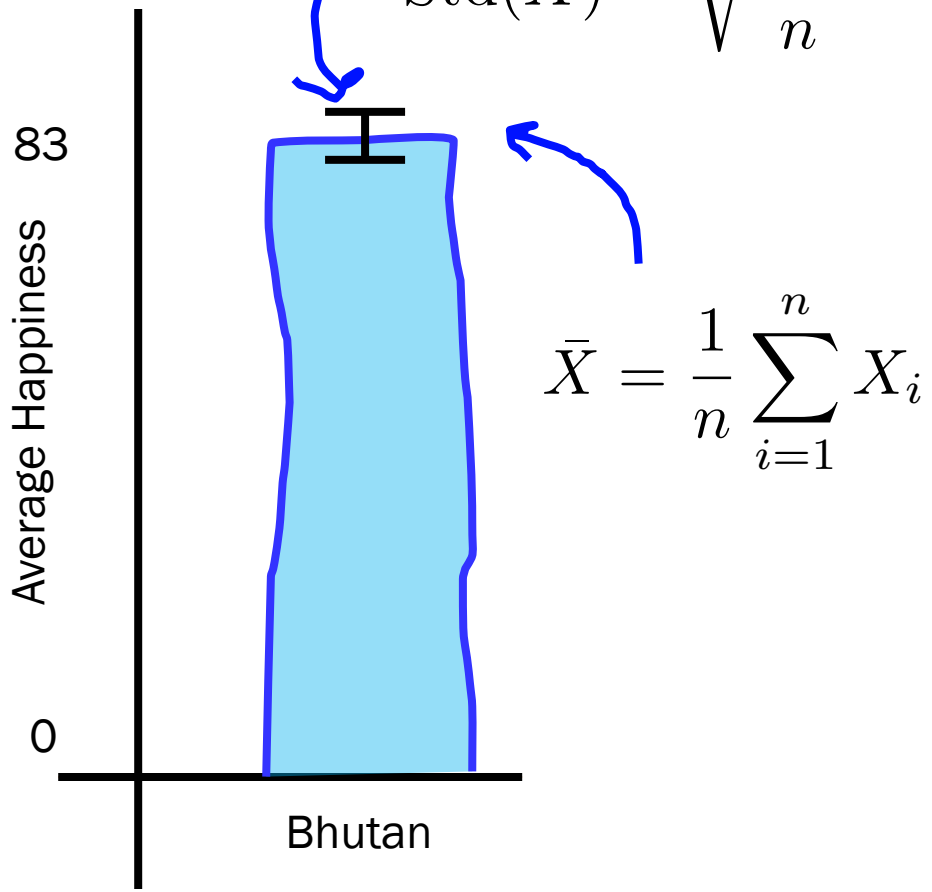


But what about **error bars**???

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

Average Happiness

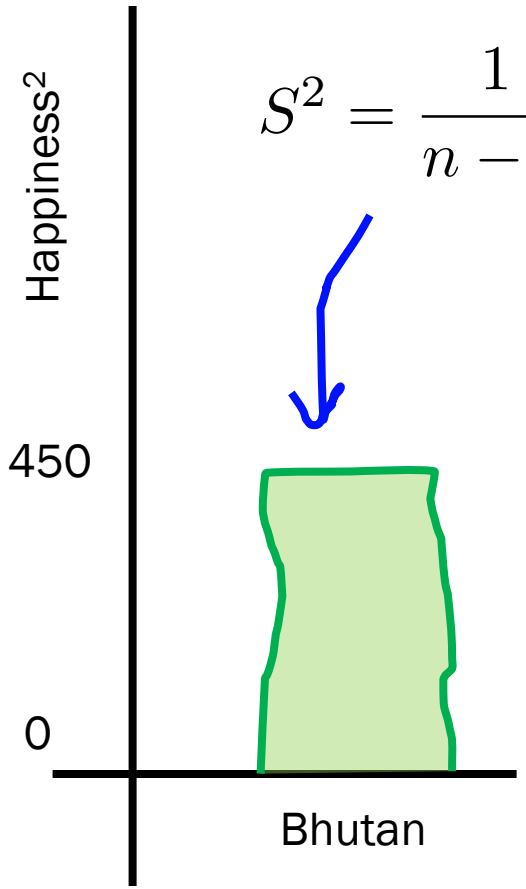
$$\text{Std}(\bar{X}) = \sqrt{\frac{S^2}{n}}$$



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

Variance of Happiness

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

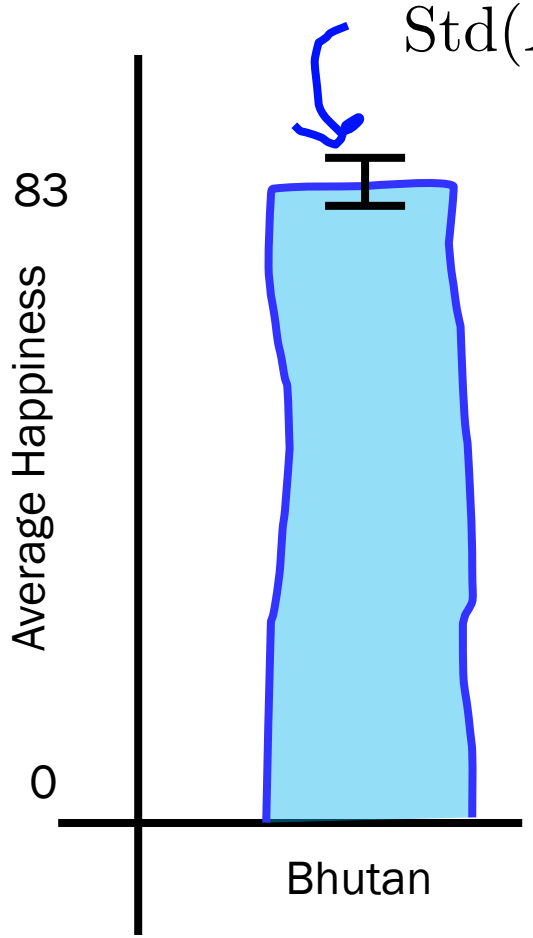


But what about **error bars**???

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

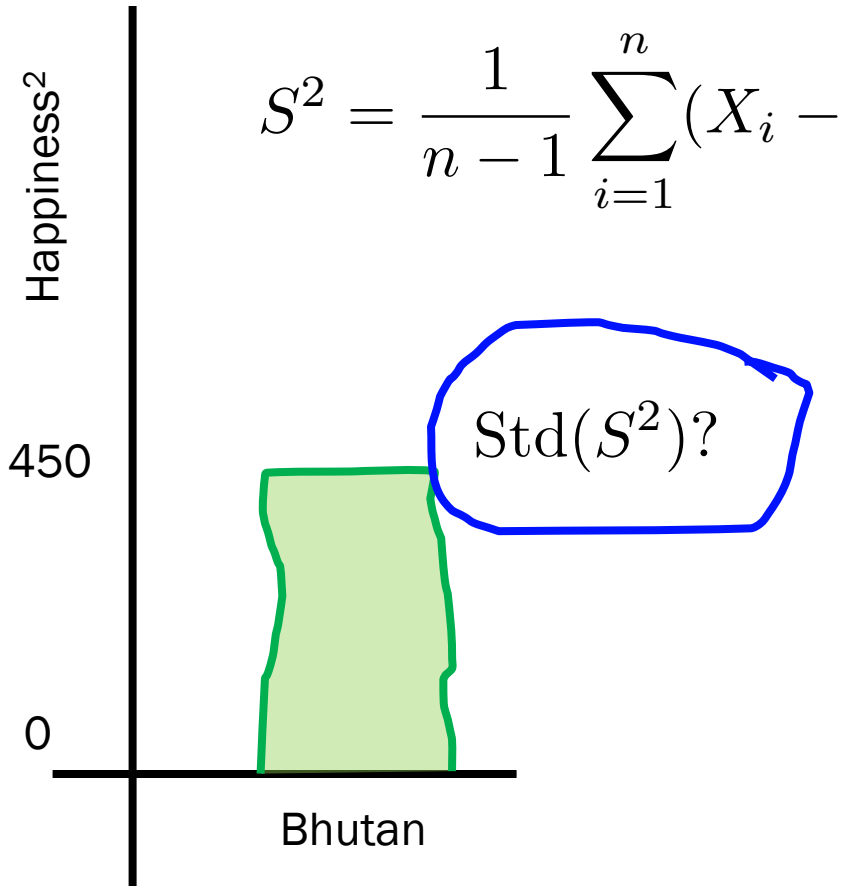
Average Happiness

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



Variance of Happiness

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



[suspense]

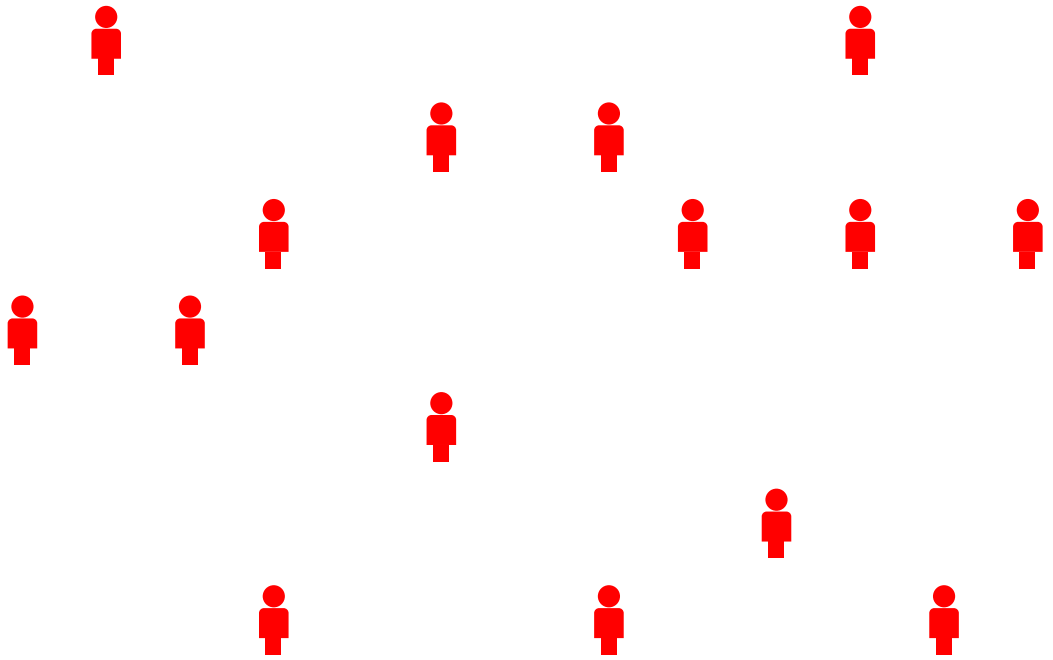
Bootstrap:
Probability for Computer Scientists

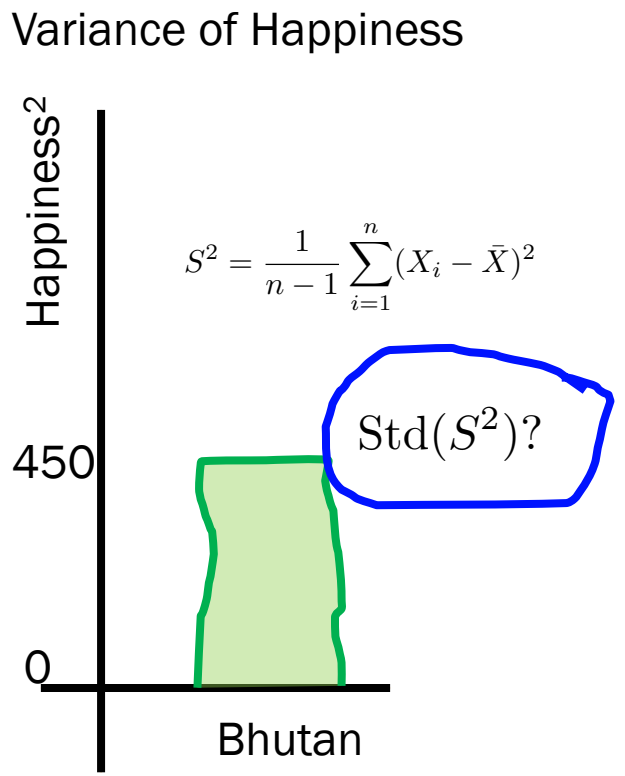
Bootstraping allows you to:

- Measure “accuracy” of **any statistic**
- Calculate **p values**
- By approximating the **distribution of statistic using computers**

Hypothetical

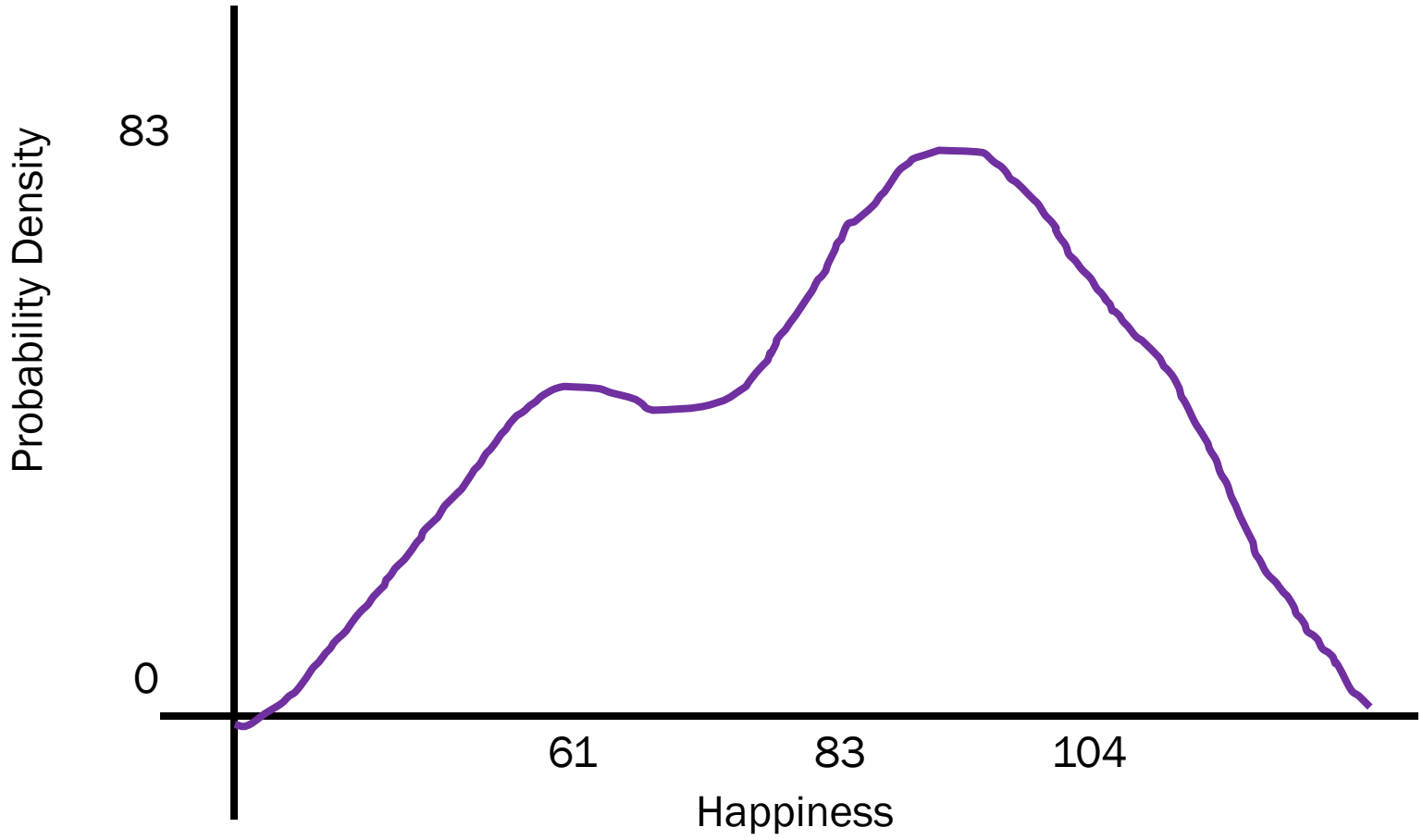
What is the **std** of the **sample variance**, calculated from 200 people?





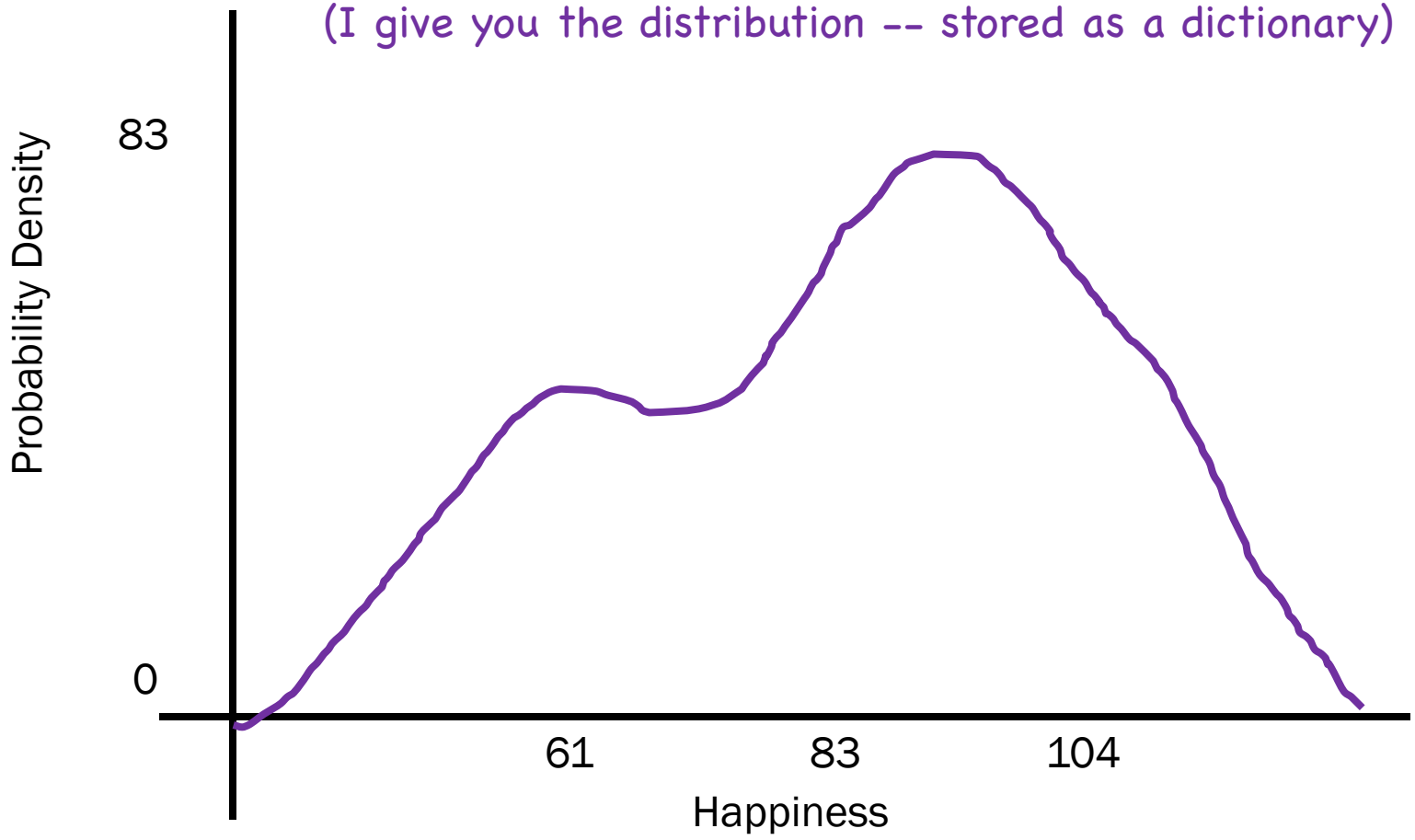
Hypothetical

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



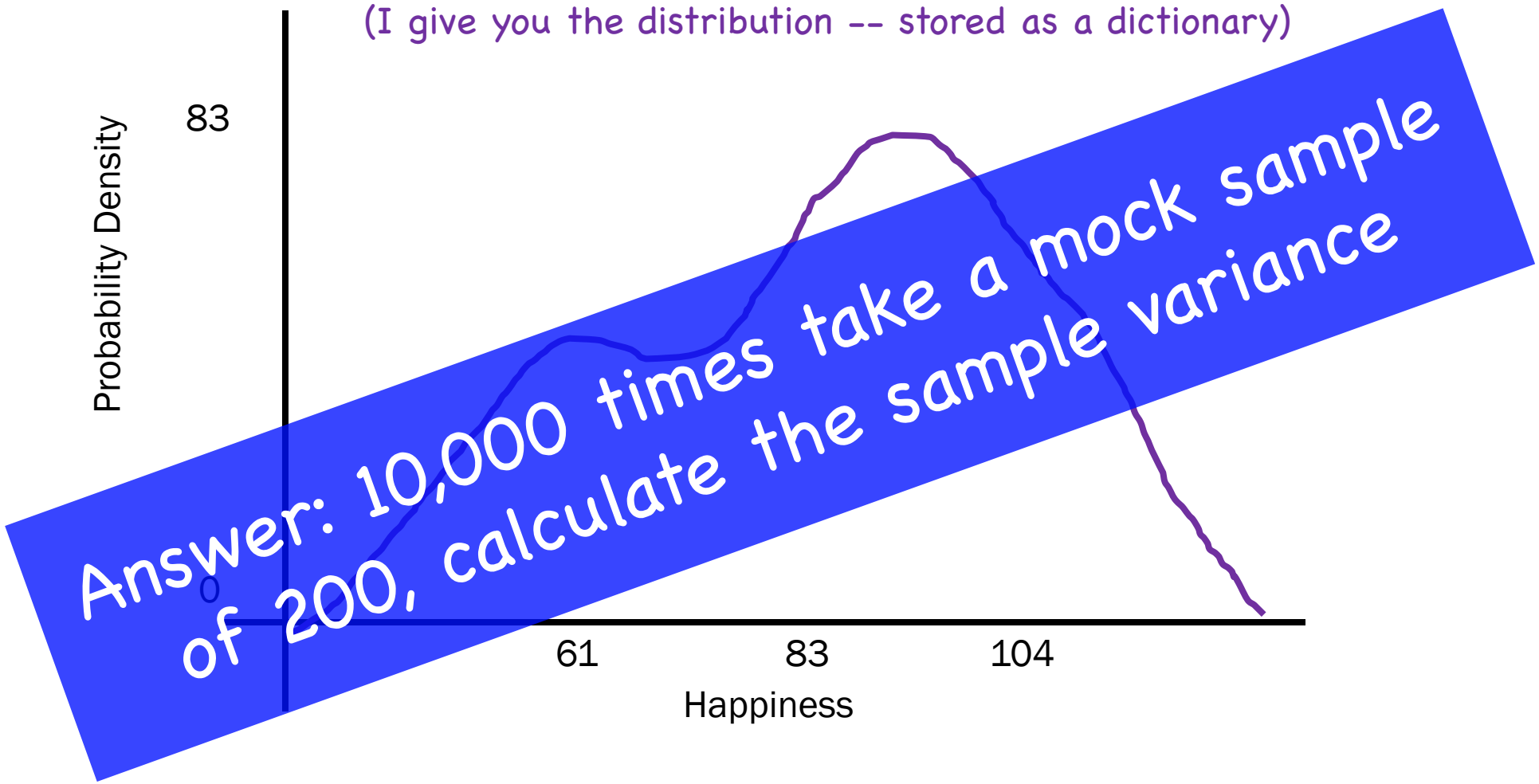
If I Gave You the True Distribution, what would you do?

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



If I Gave You the True Distribution, what would you do?

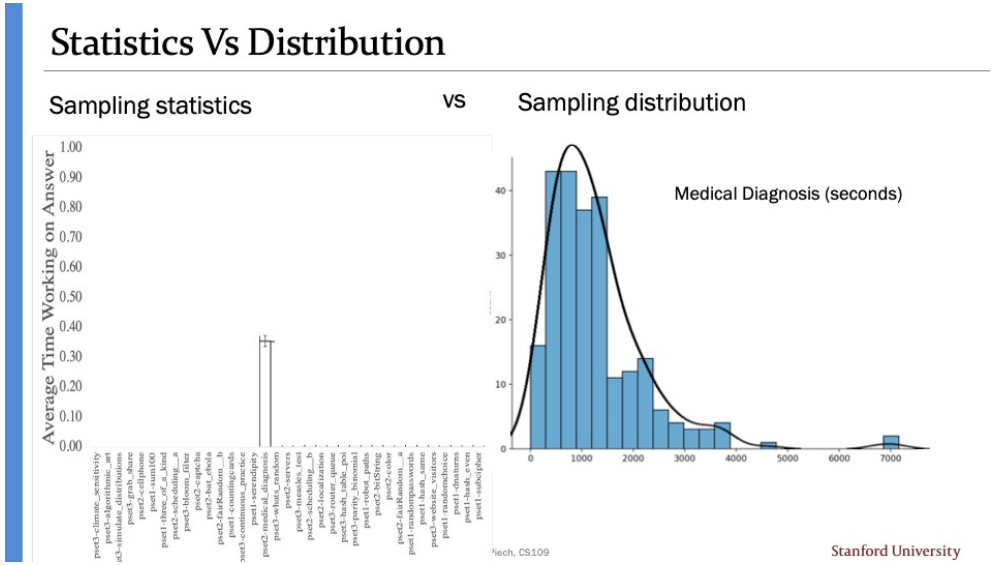
What is the **std** of the **sample variance**, calculated from 200 people?



Algorithm

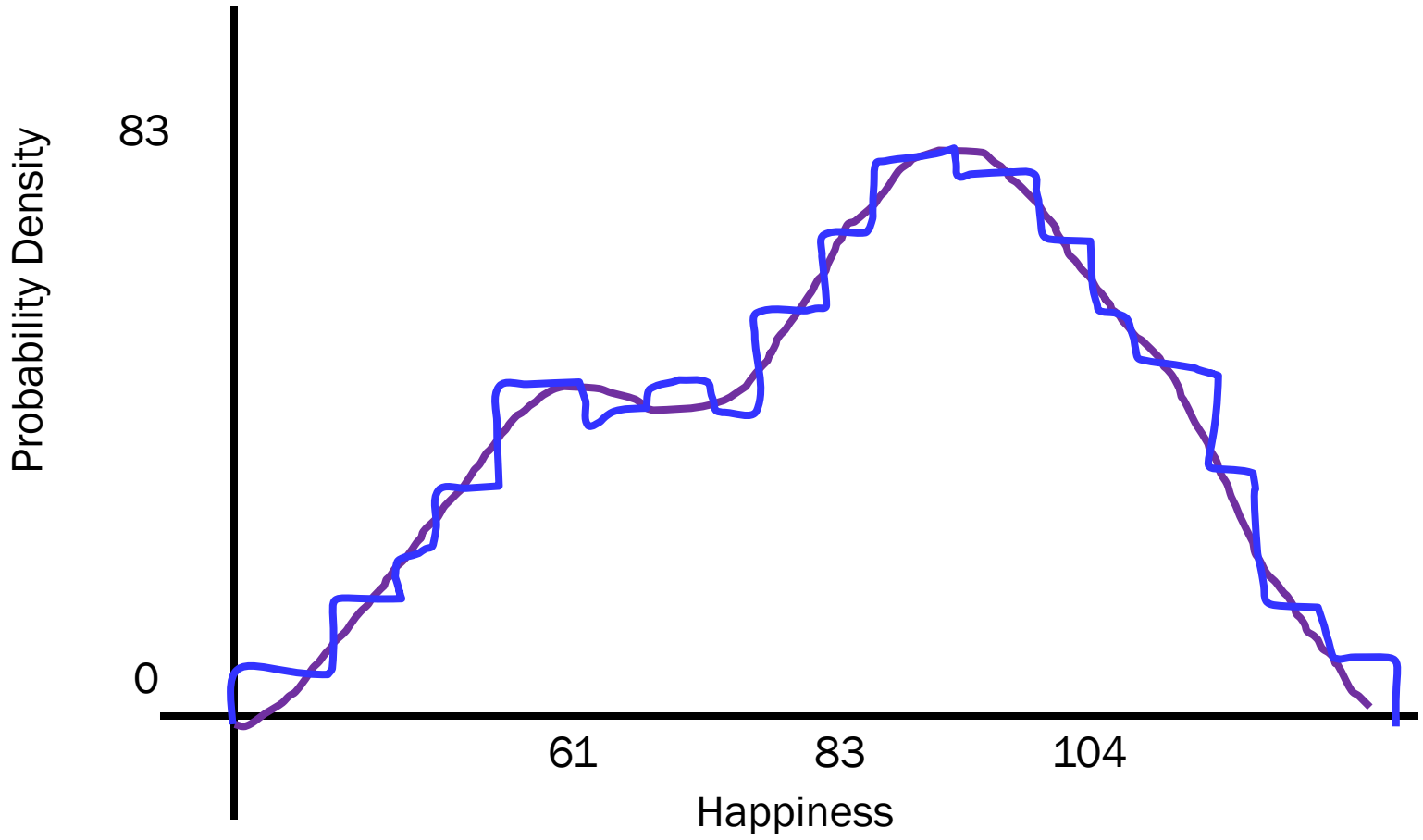
Estimating std of sample variance (sample):

1. Take the **PMF** of population
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**



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

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



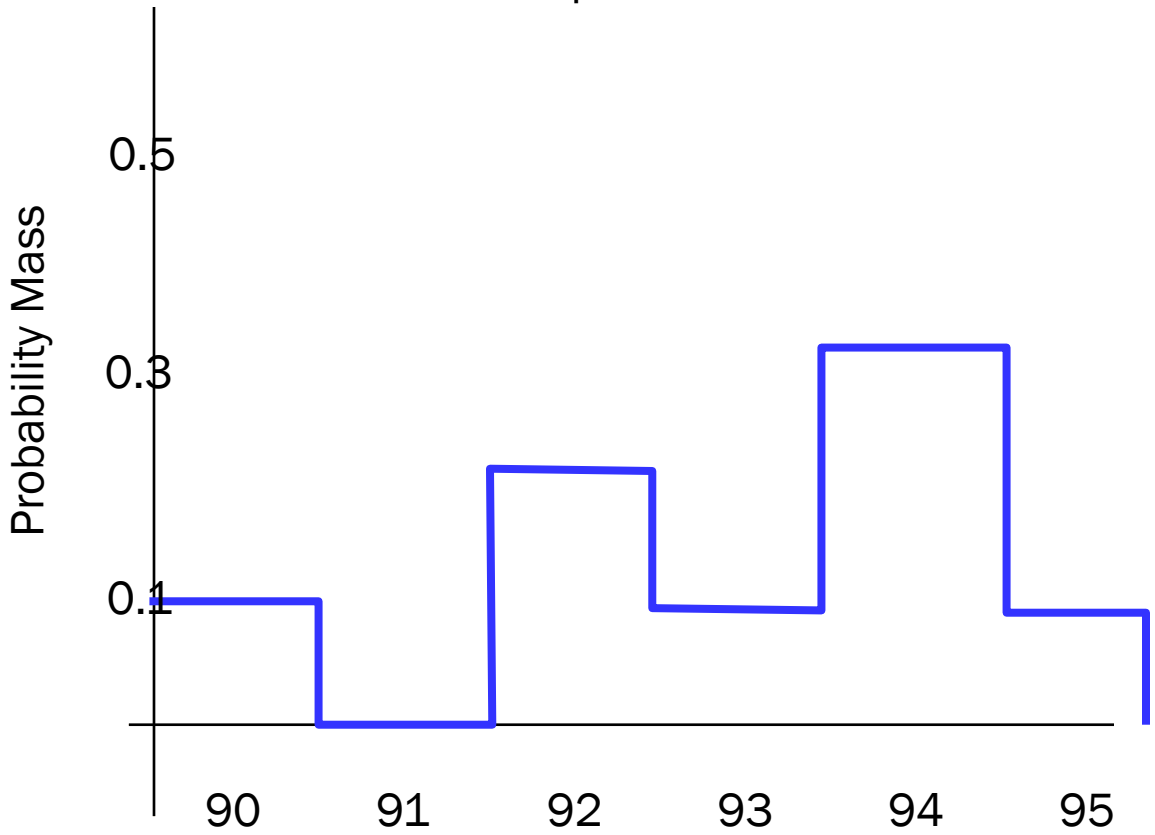
* This is just a histogram of your data!!

Key Insight

IID Samples

90,
92,
92,
93,
94,
94,
94,
95,

Sample Distribution



Bootstrapping Assumption

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

The underlying distribution

The sample distribution

(aka the histogram of your data)

Algorithm

Bootstrap Algorithm (sample):

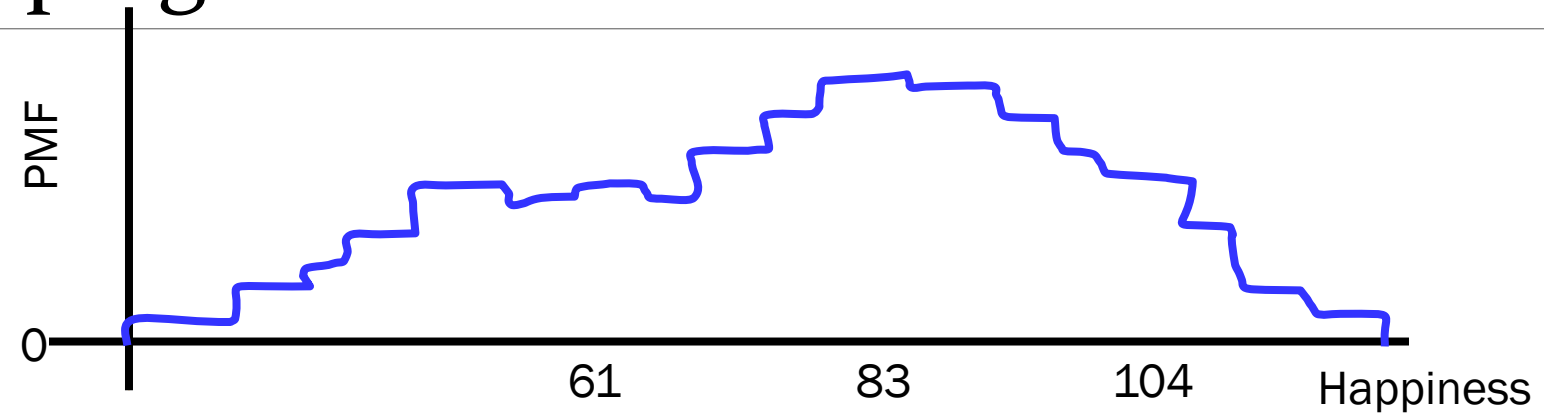
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Resample **len(sample)** from PMF
   b. **Recalculate the stat** on the resample
3. You now have a **distribution of your stat**

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**

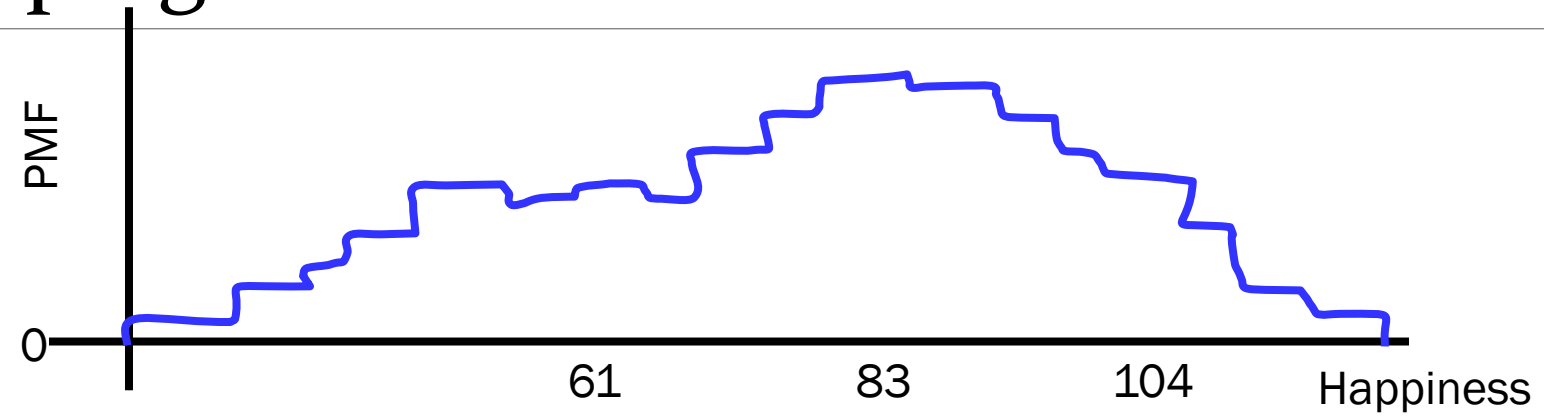
Bootstrapping of Means



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**

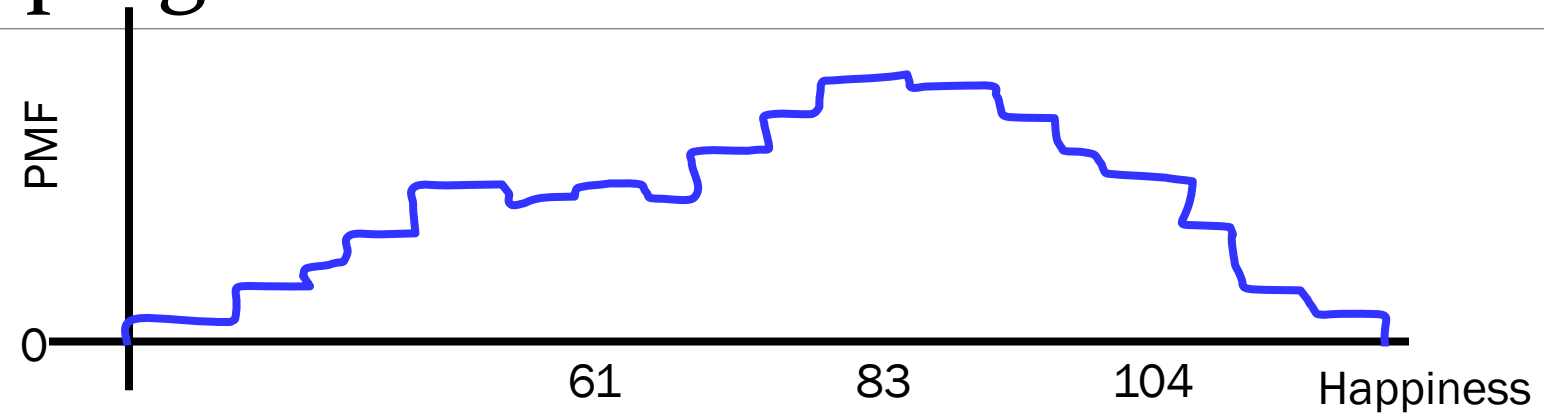
Bootstrapping of Means



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**

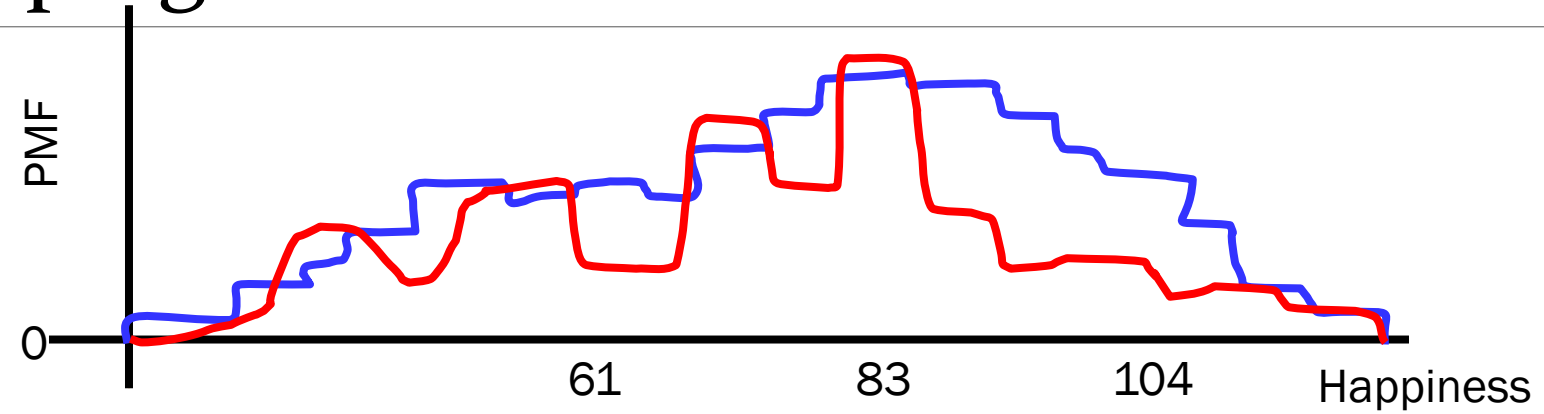
Bootstrapping of Means



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**

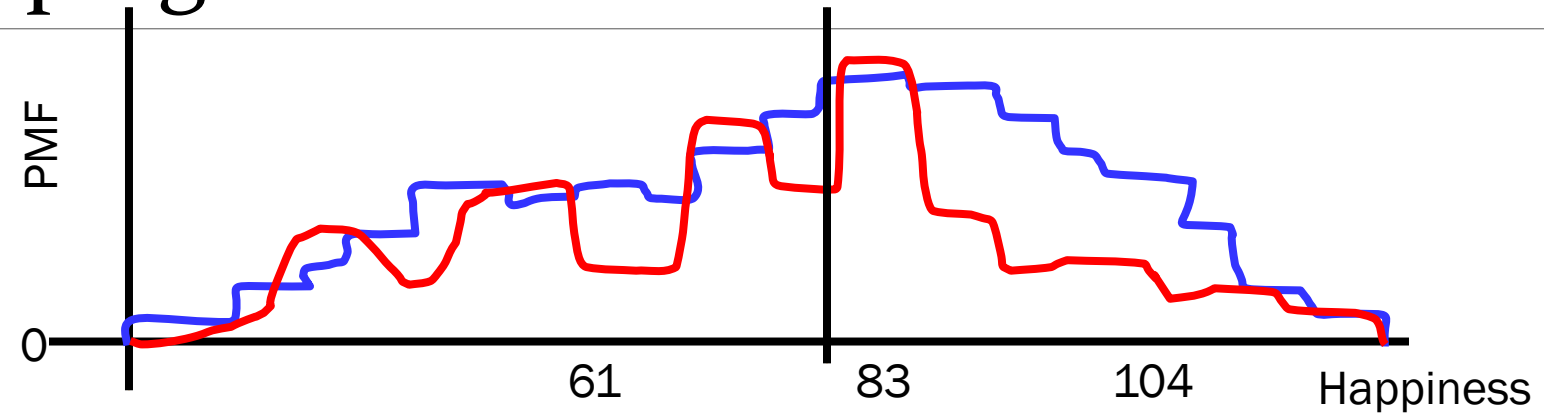
Bootstrapping of Means



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

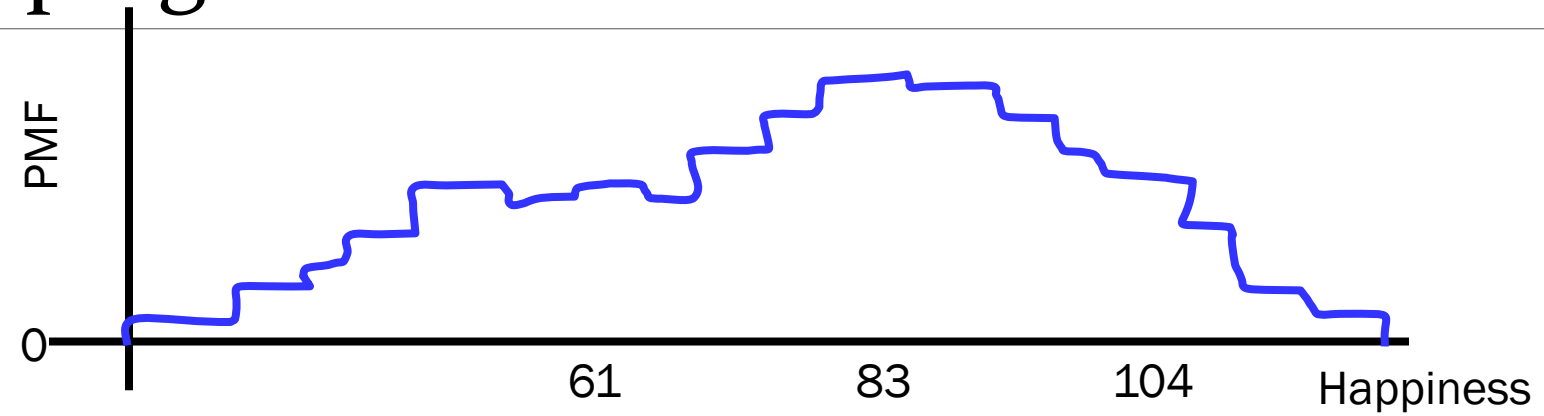


Bootstrap Algorithm (sample):

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

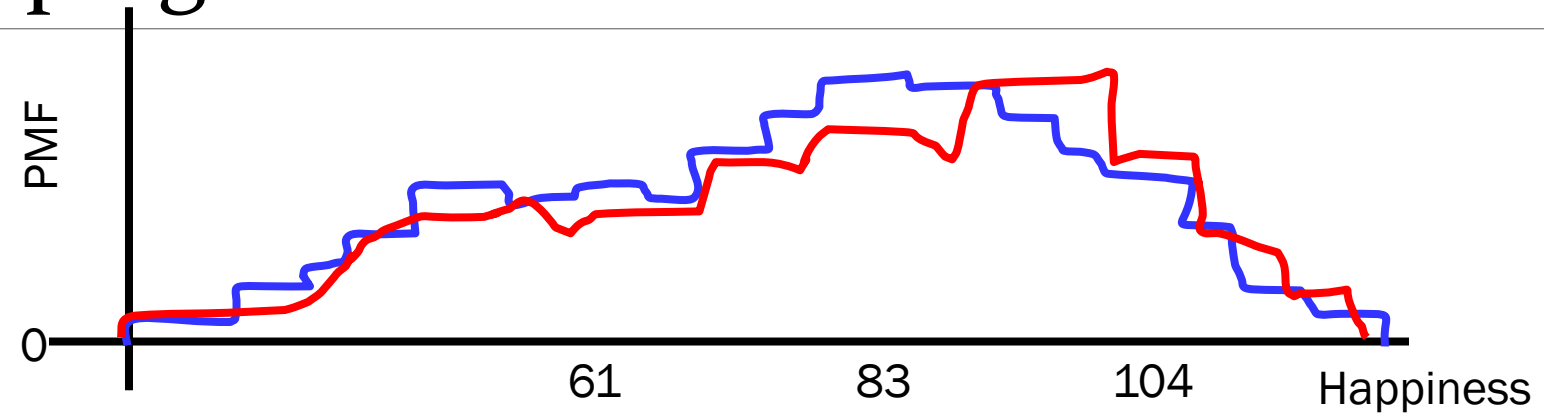


Bootstrap Algorithm (sample):

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

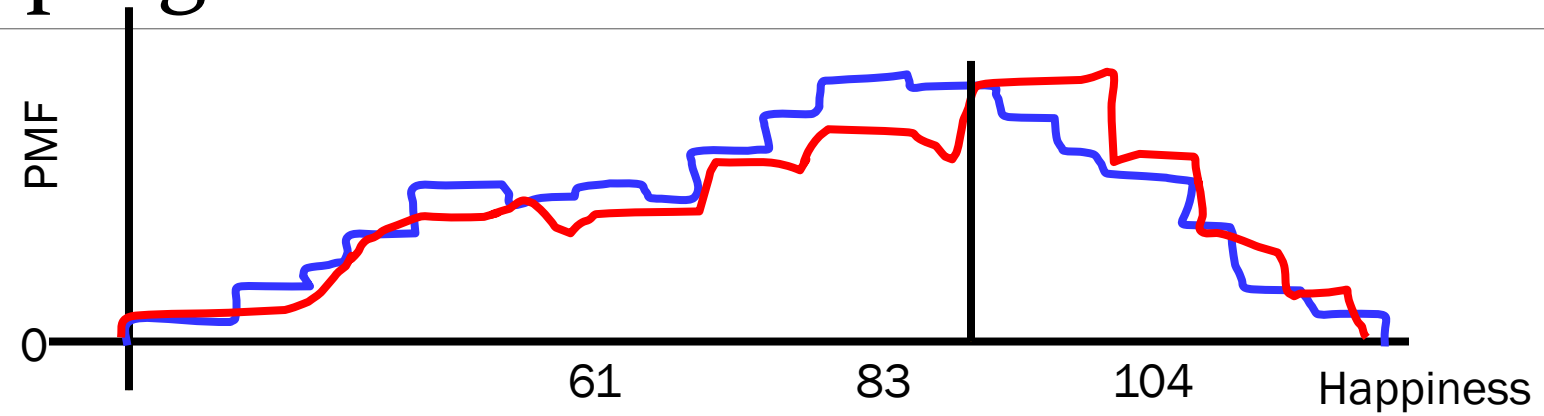


Bootstrap Algorithm (sample):

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

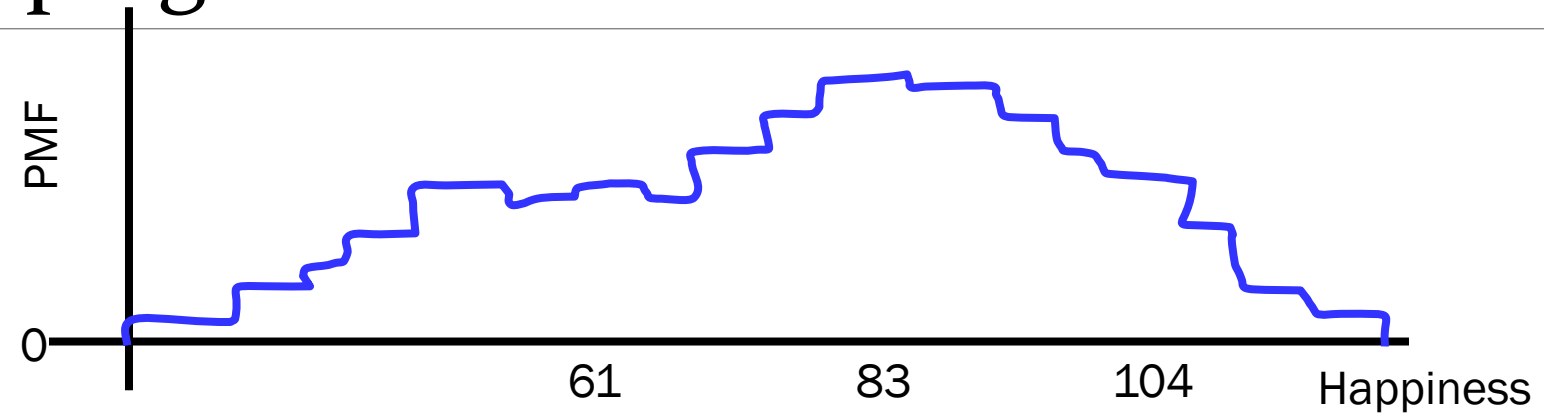


Bootstrap Algorithm (sample):

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

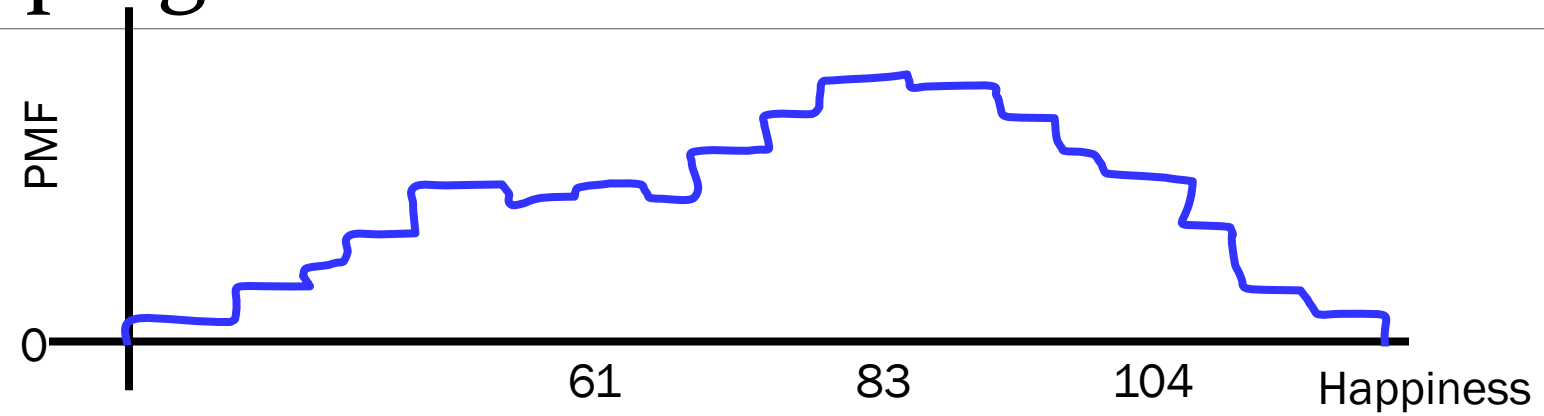


Bootstrap Algorithm (sample):

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



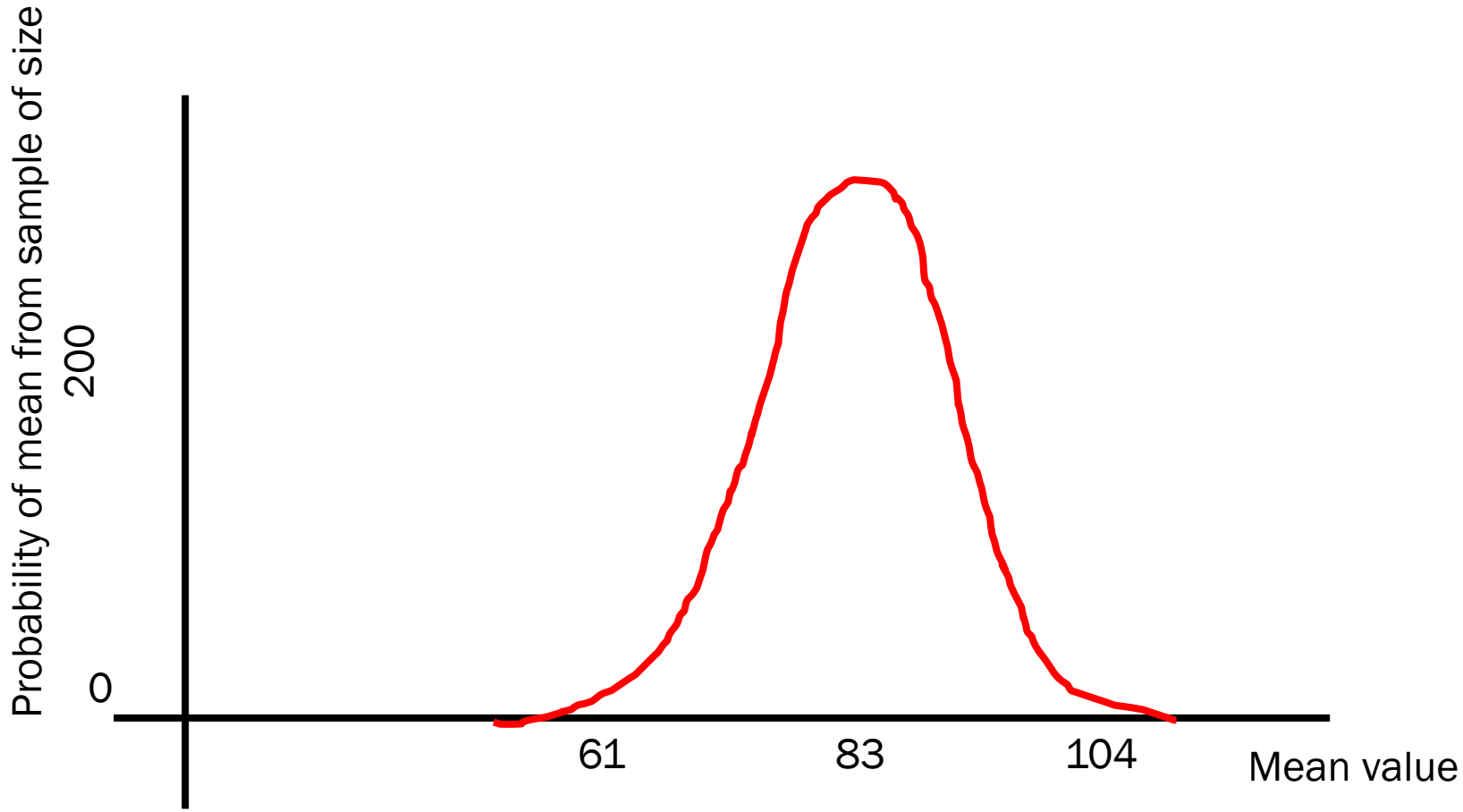
Bootstrap Algorithm (sample):

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]

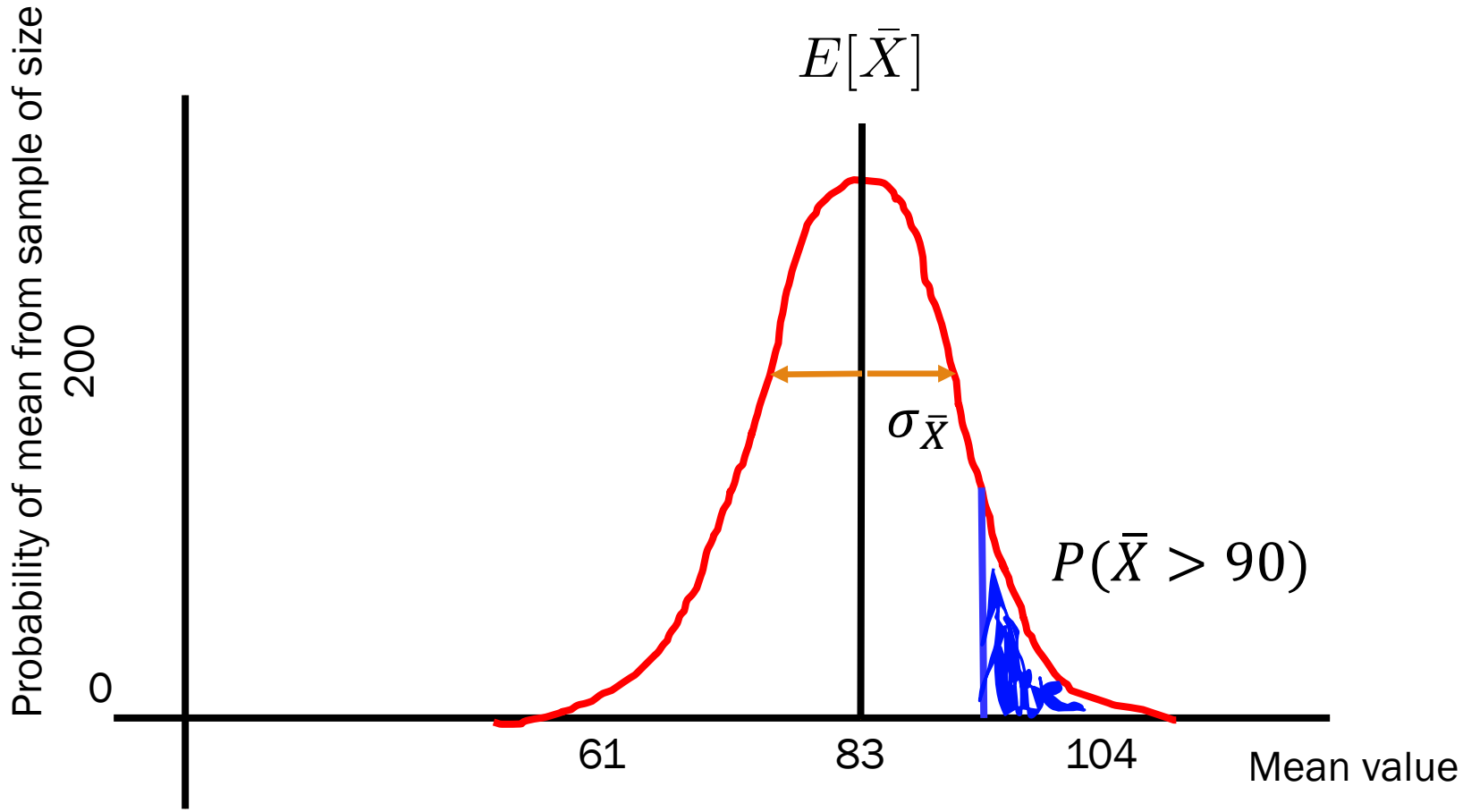
Bootstrapping of Means

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

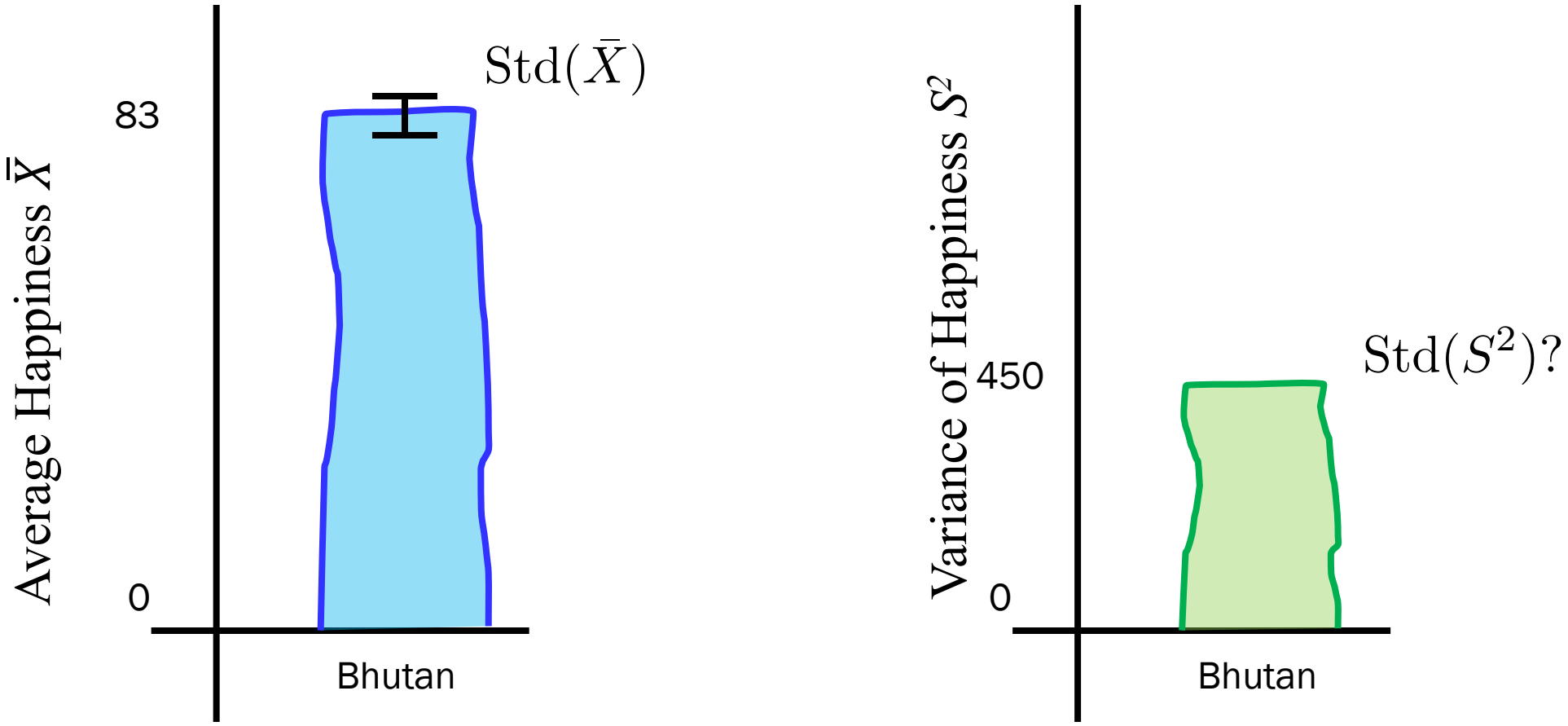


Bootstrapping of Means

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



Our Report to Bhutan Government



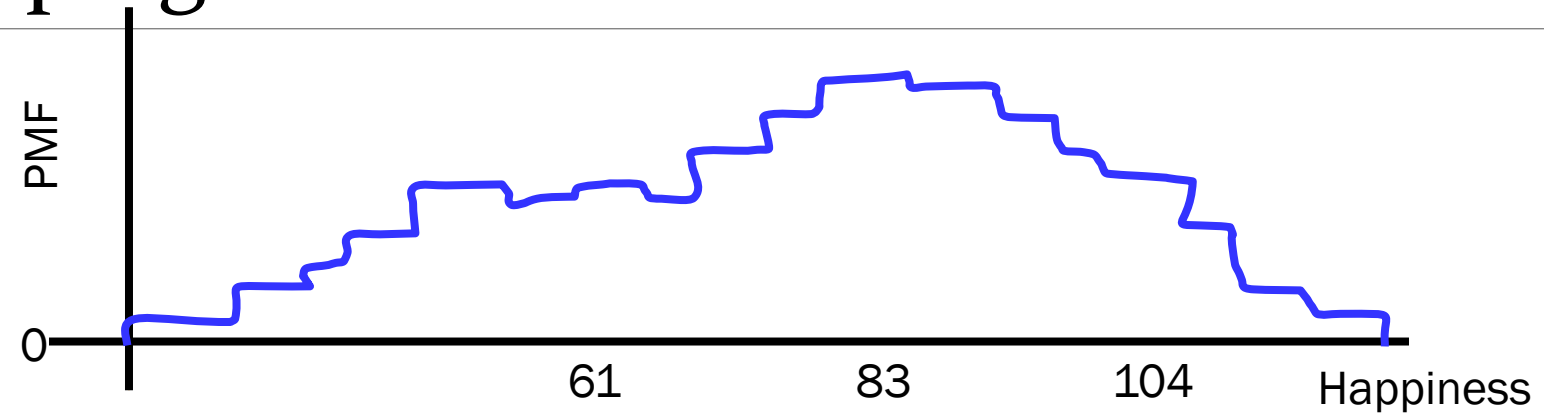
Claim: The average happiness of Bhutan is 83 ± 2

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**

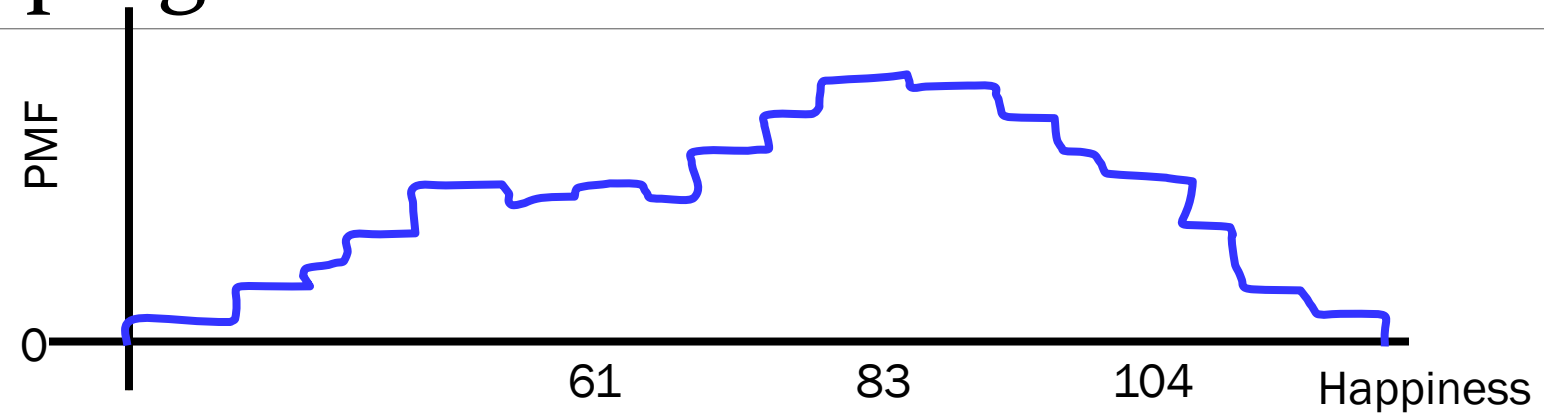
Bootstrapping of Variance



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**

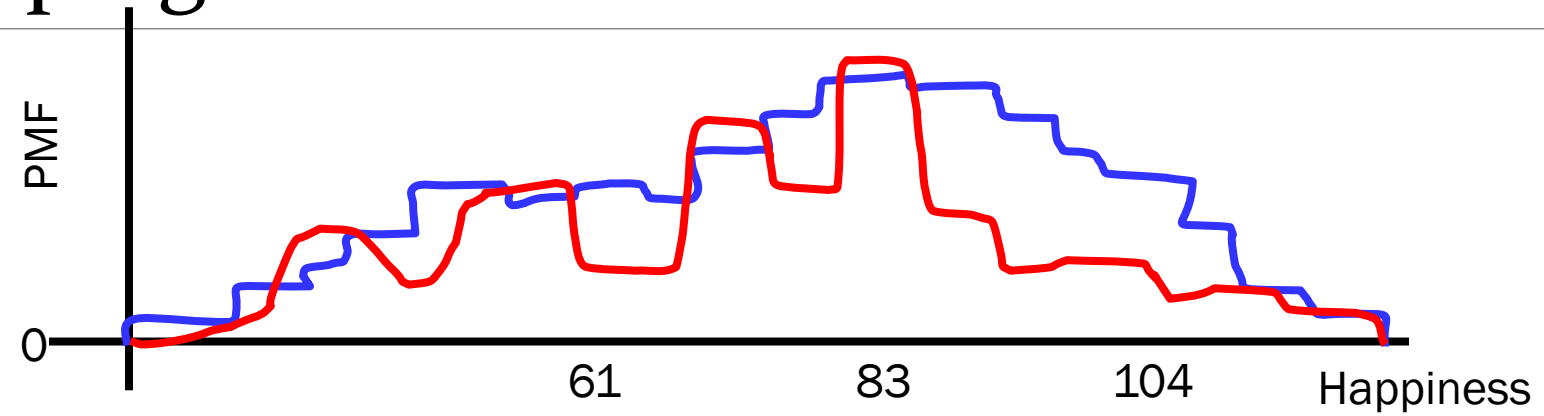
Bootstrapping of Variance



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**

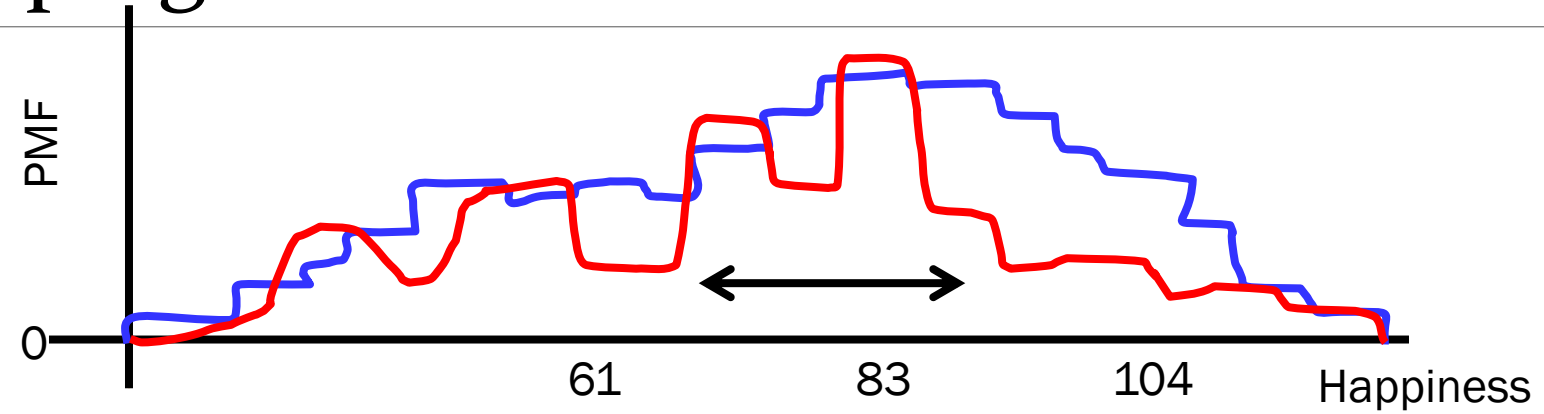
Bootstrapping of Variance



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

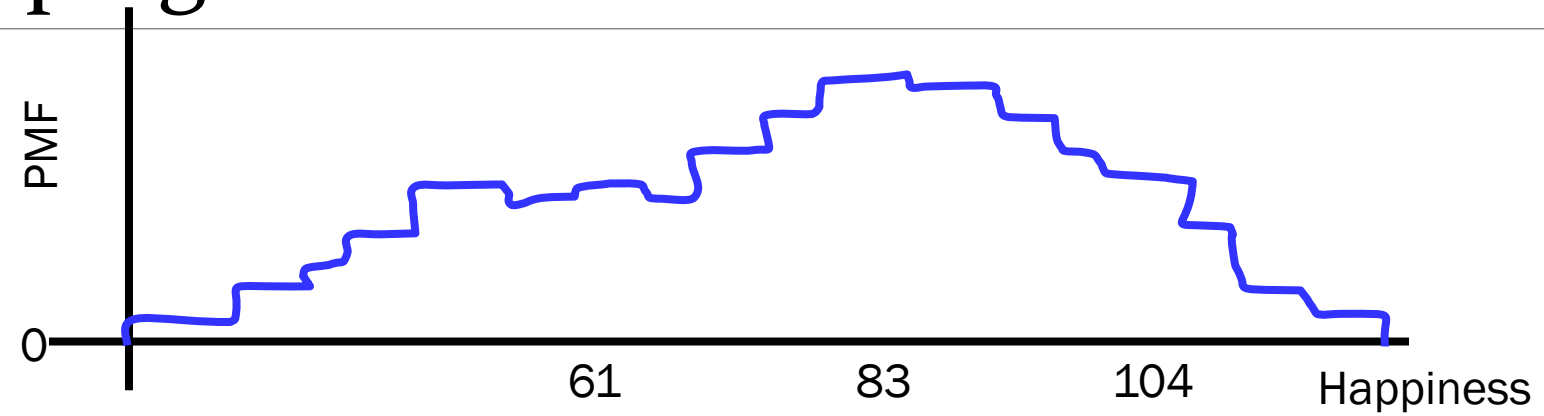


Bootstrap Algorithm (sample):

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

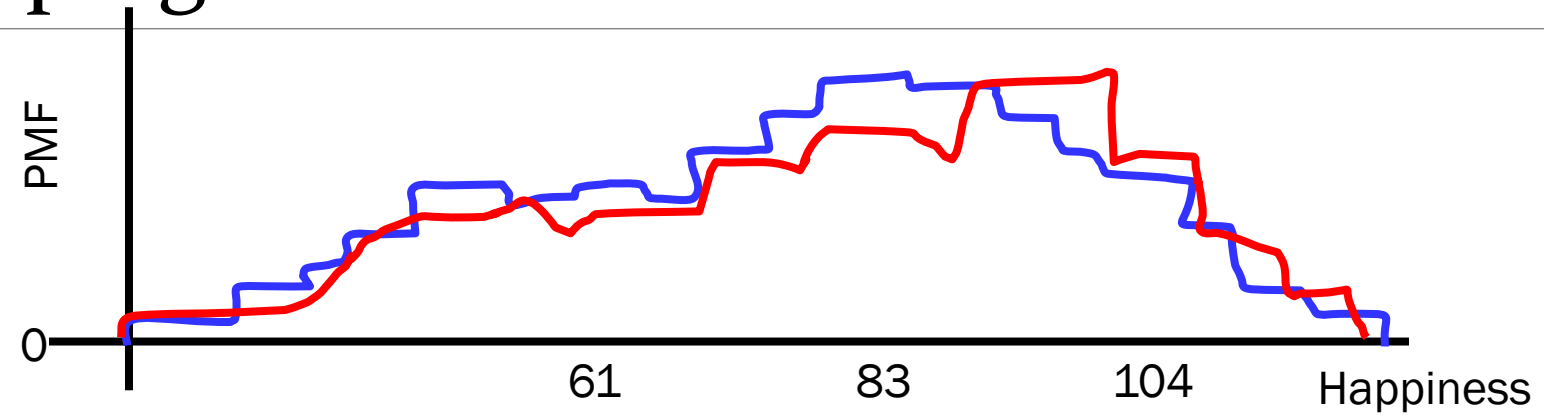


Bootstrap Algorithm (sample):

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

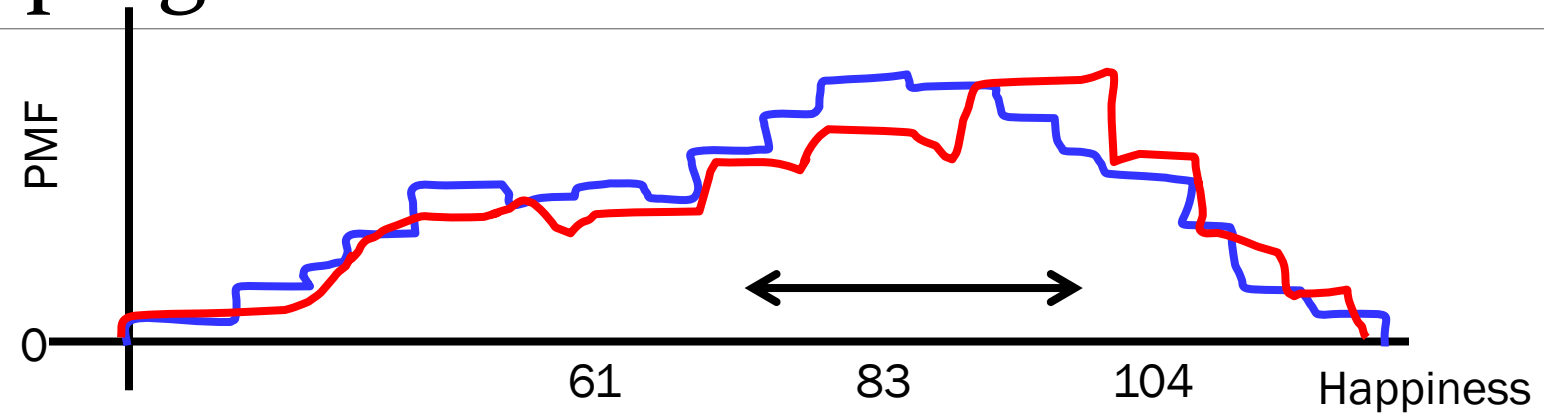


Bootstrap Algorithm (sample):

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

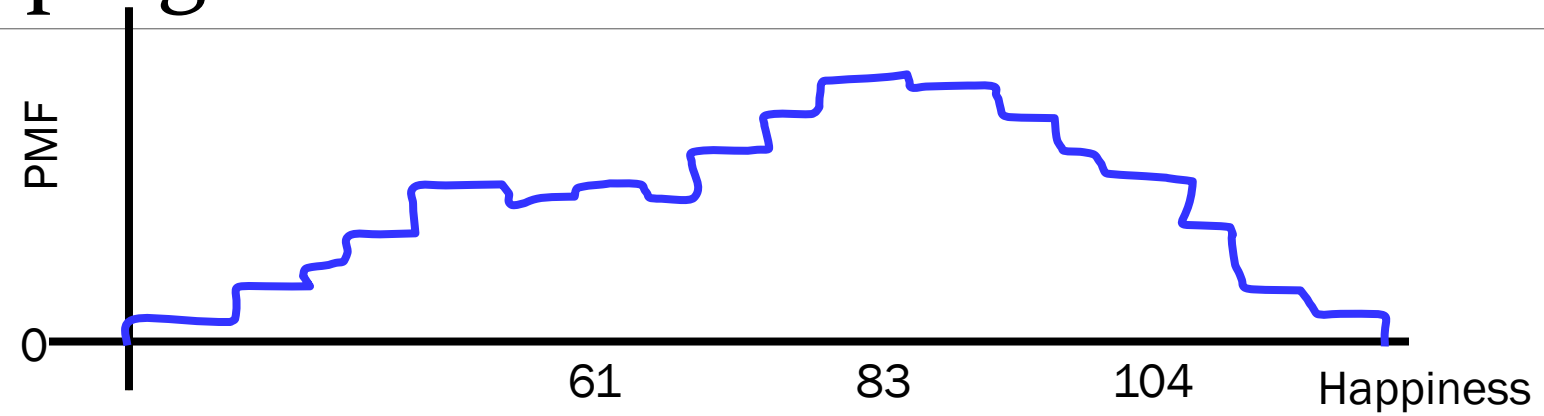


Bootstrap Algorithm (sample):

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

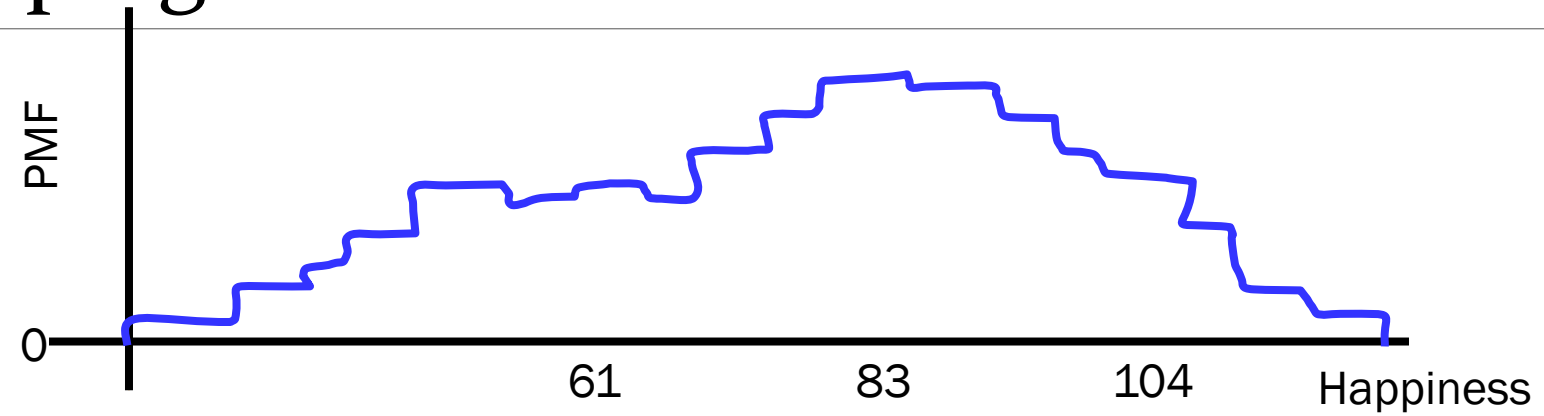


Bootstrap Algorithm (sample):

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



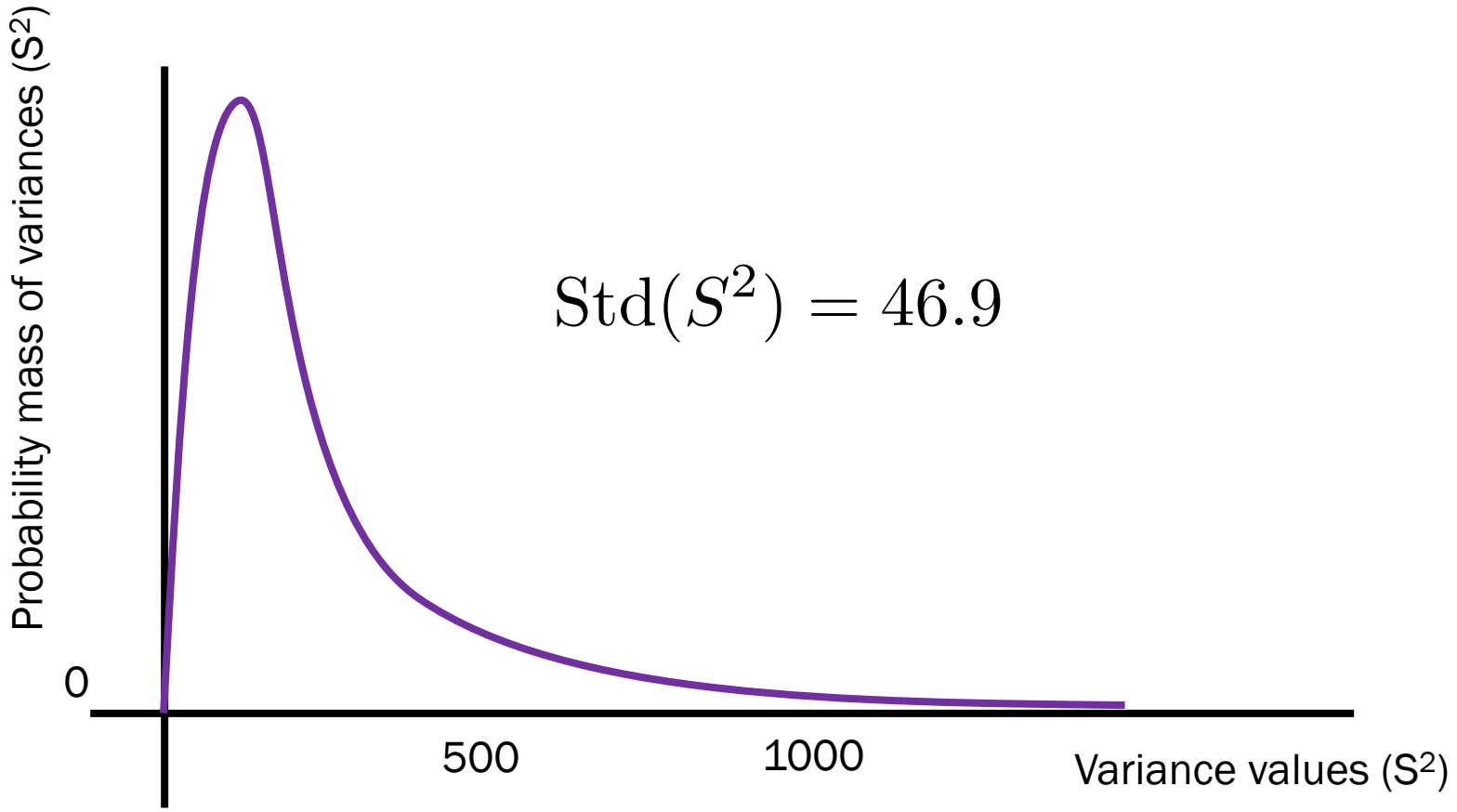
Bootstrap Algorithm (sample):

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]

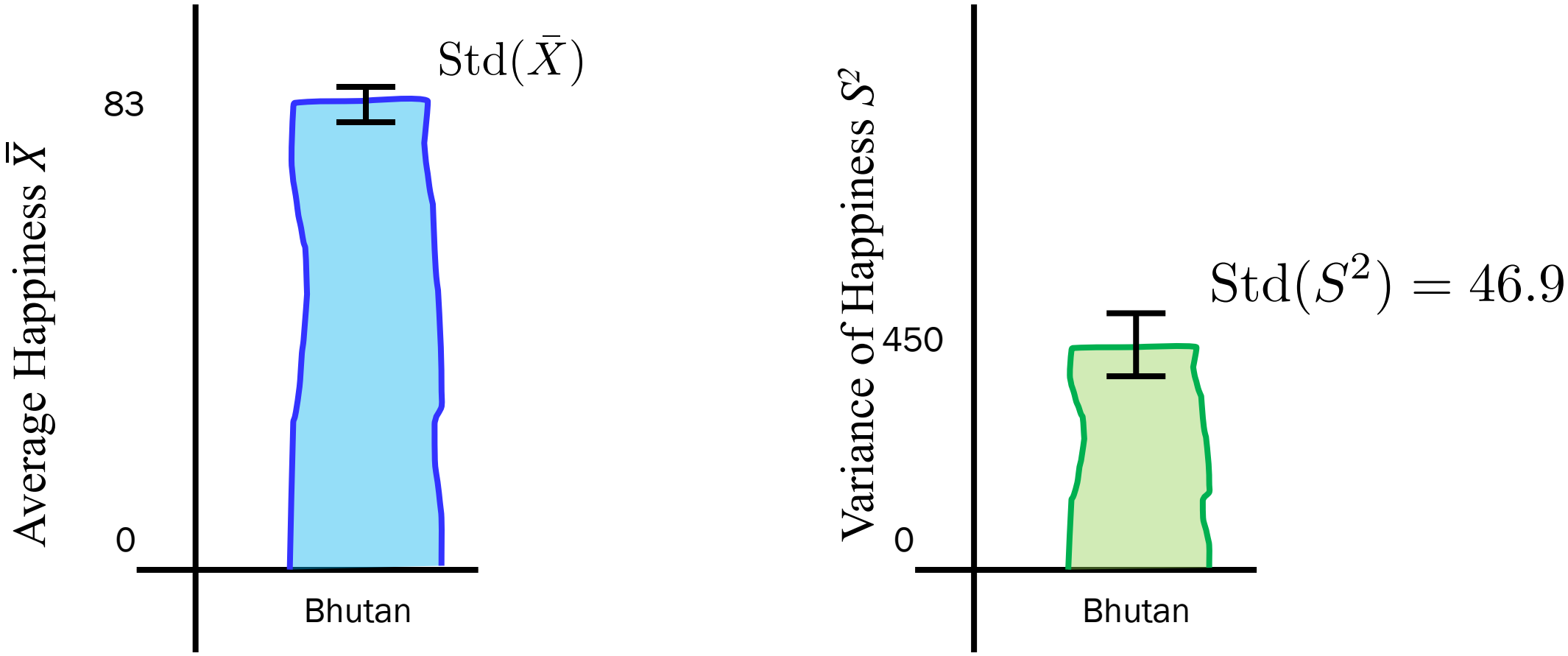
Bootstrapping of Variance

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



$$\text{Std}(S^2) = 46.9$$

Our Report to Bhutan Government



Claim: The average happiness of Bhutan is 83 ± 2

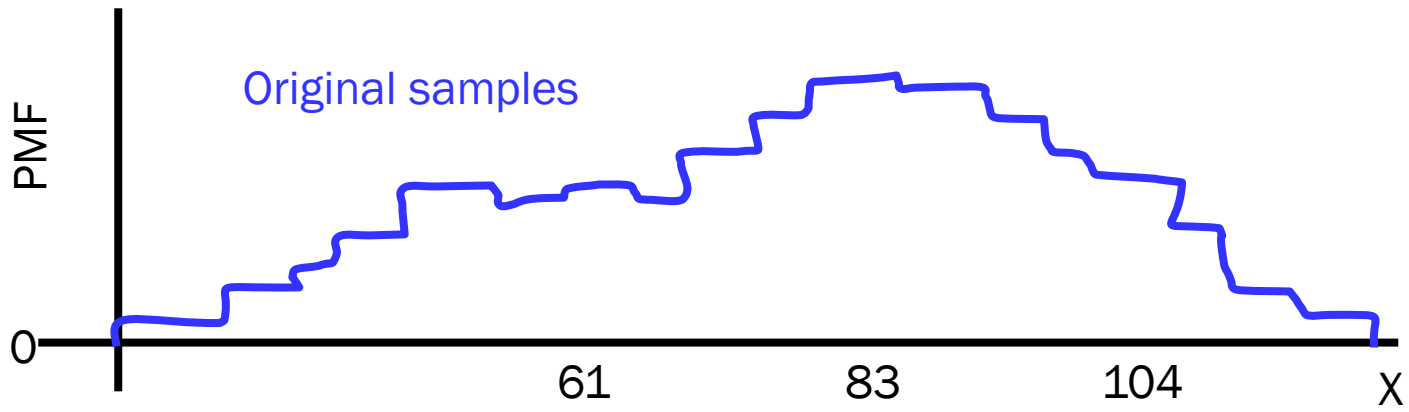
Pedagogical pause

Bootstrap Algorithm for 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

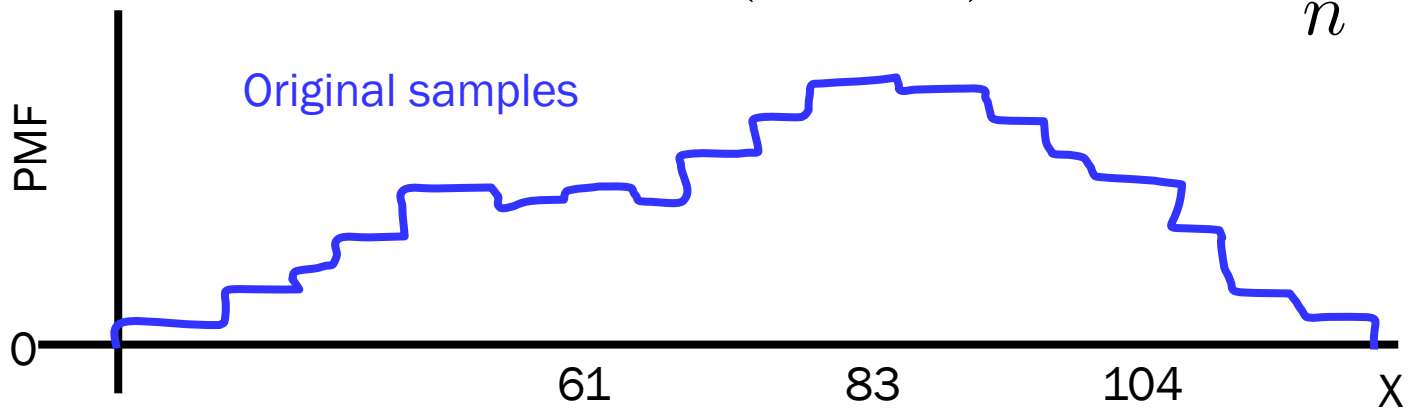
```
def resample(samples, K):
    # Estimate the PMF using the samples
    # Draw K new samples from the PMF
```



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**

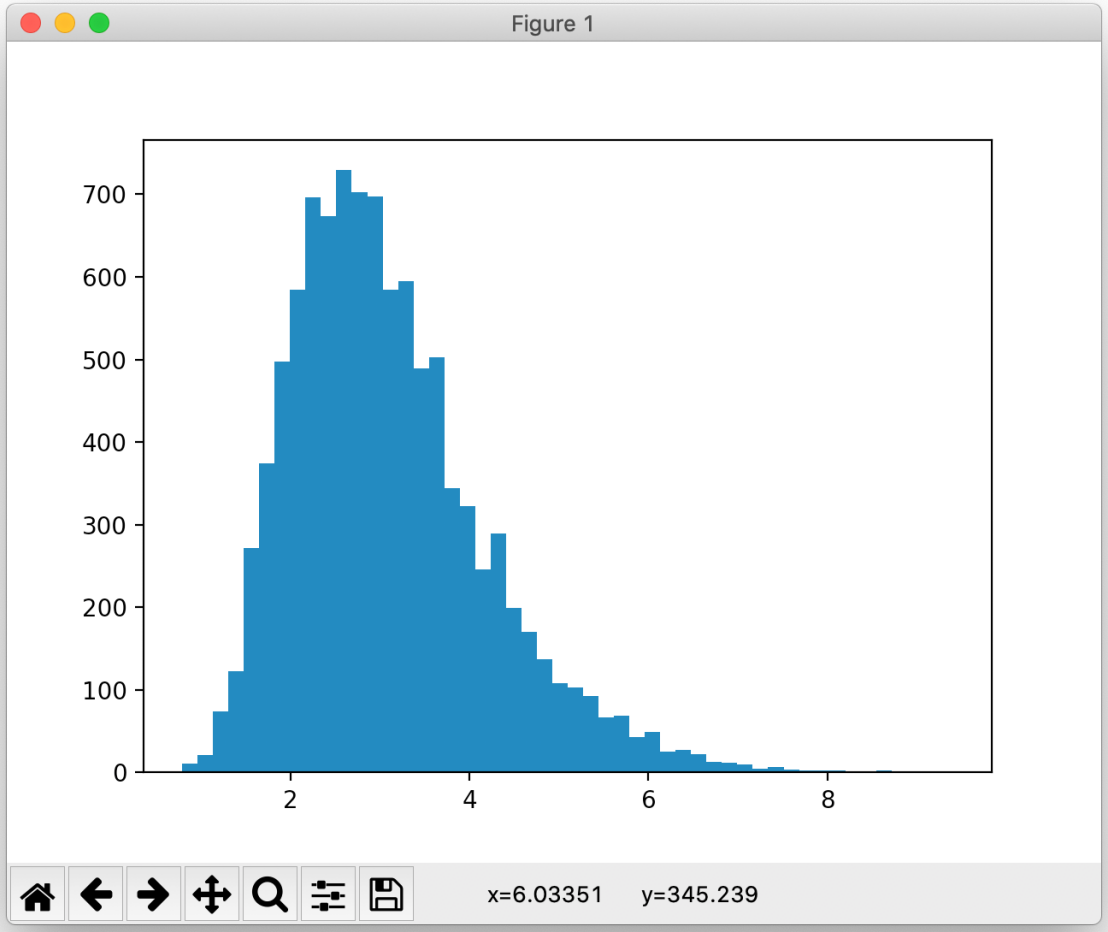
Bootstrapping in Practice

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!

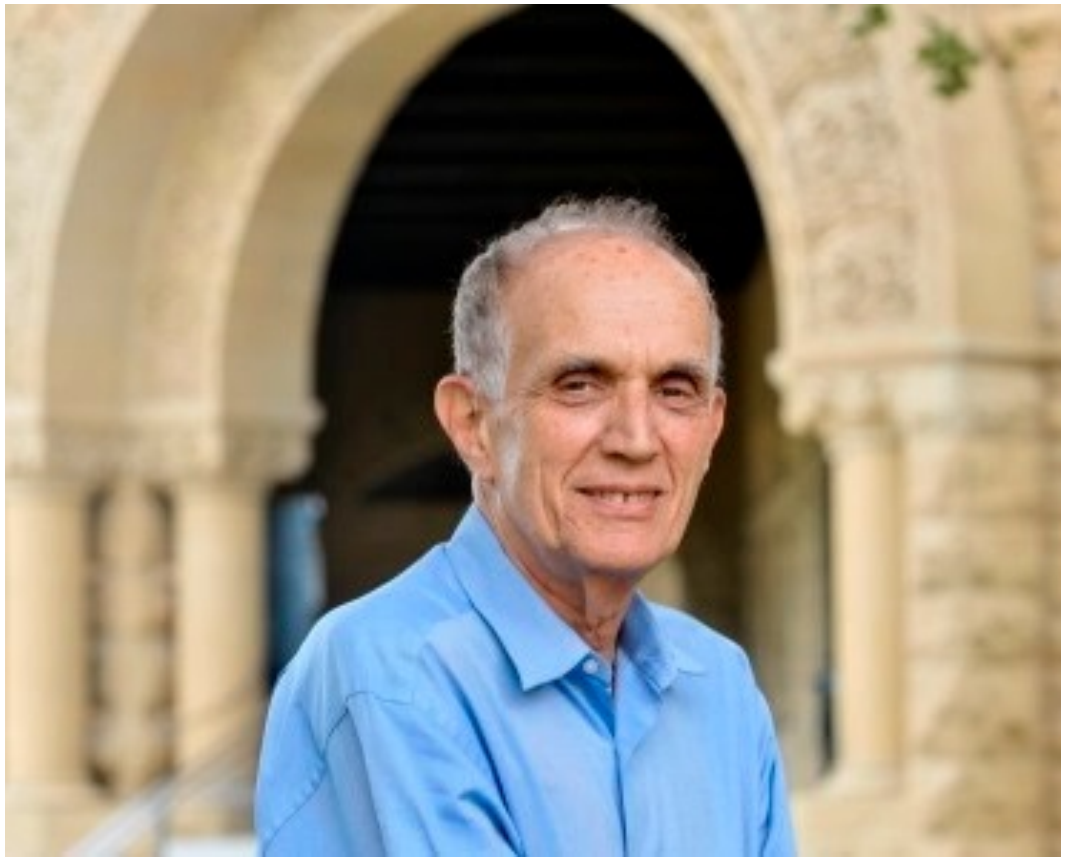
The Distribution of the Sampling Variance



Bootstrap provides a way to estimate probabilities of statistics using code.

Bootstrap

Bradley Efron



Invented bootstrapping in 1979

Still a professor at Stanford

Won a National Science Medal

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

How confident are you in this claim?

A real difference?

Learning in Context A	Learning in Context B
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83

18 students

23 students

$\mu_1 = 3.1$

$\mu_2 = 2.4$

Claim: Group 1 and Group 2 are samples from different distributions

How confident are you in this claim?

The Null Hypothesis

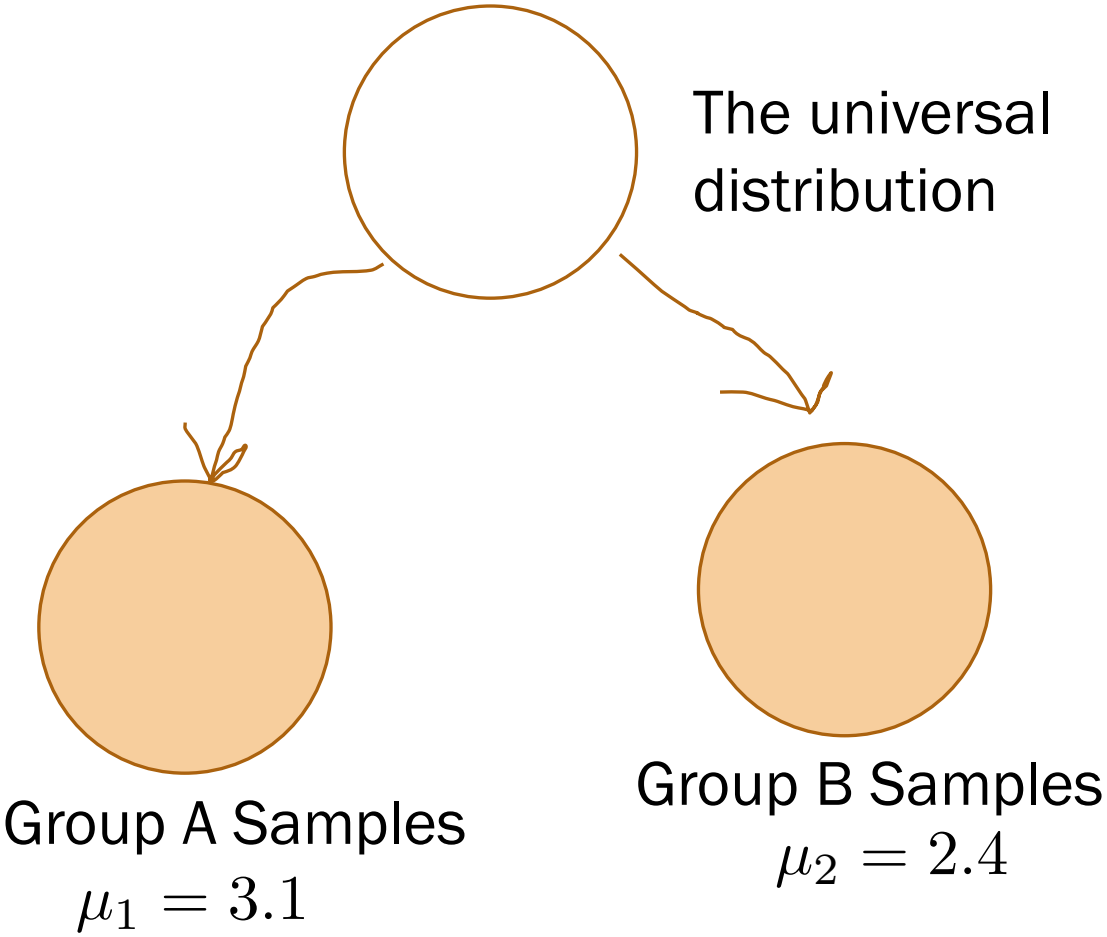
There is no difference between the two groups, so everyone is drawn from the same distribution. Any difference you observe is due to sampling error.

H_0 (Null): From the same distribution
H_a (Alternative): Not from the same distribution

p-value: What's the probability of seeing a difference this big under the null hypothesis?

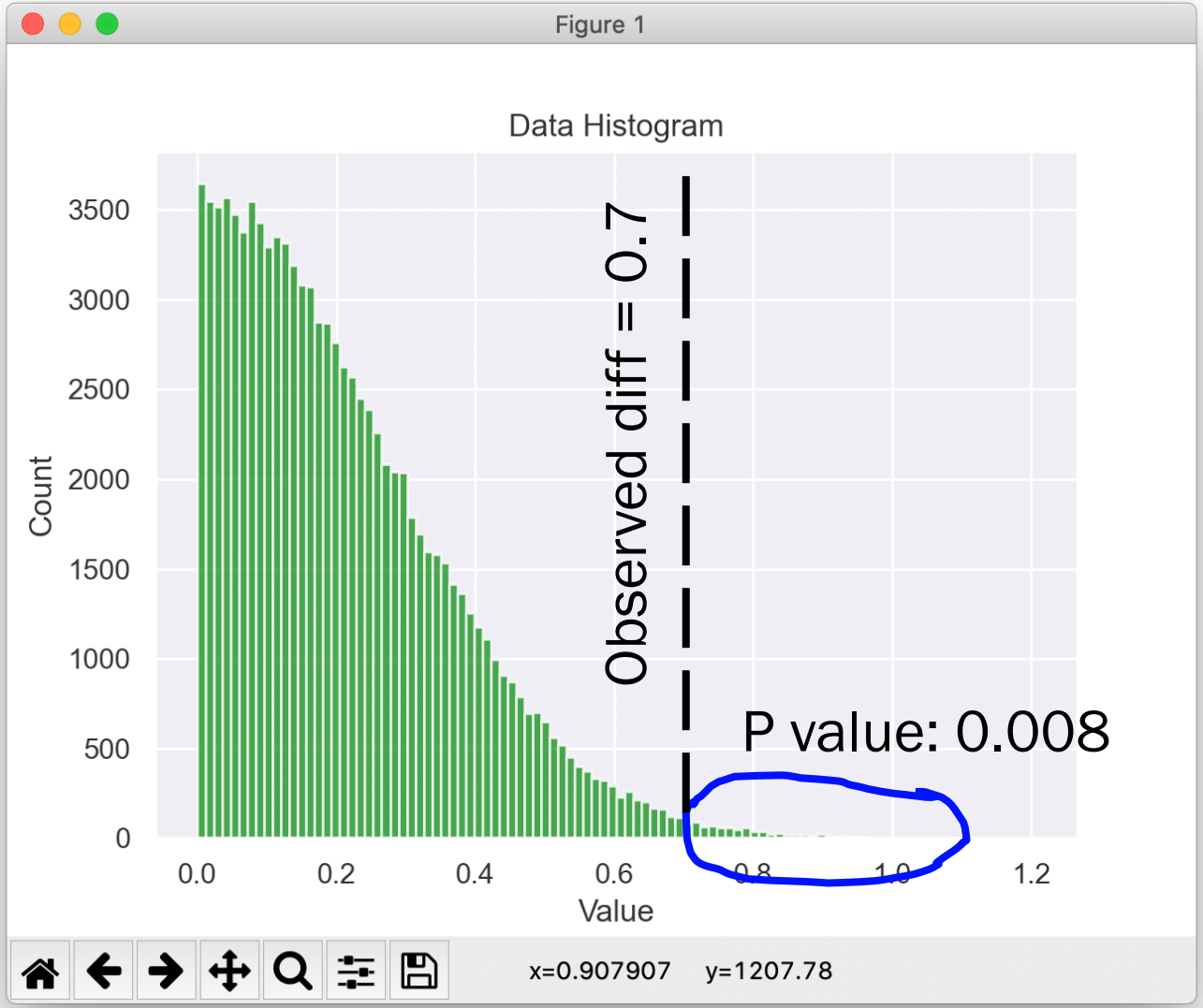
$$P(|\mu_1 - \mu_2| > 0.7) \text{ if } H_0 \text{ is True?}$$

If this value is very small, we "reject" the null



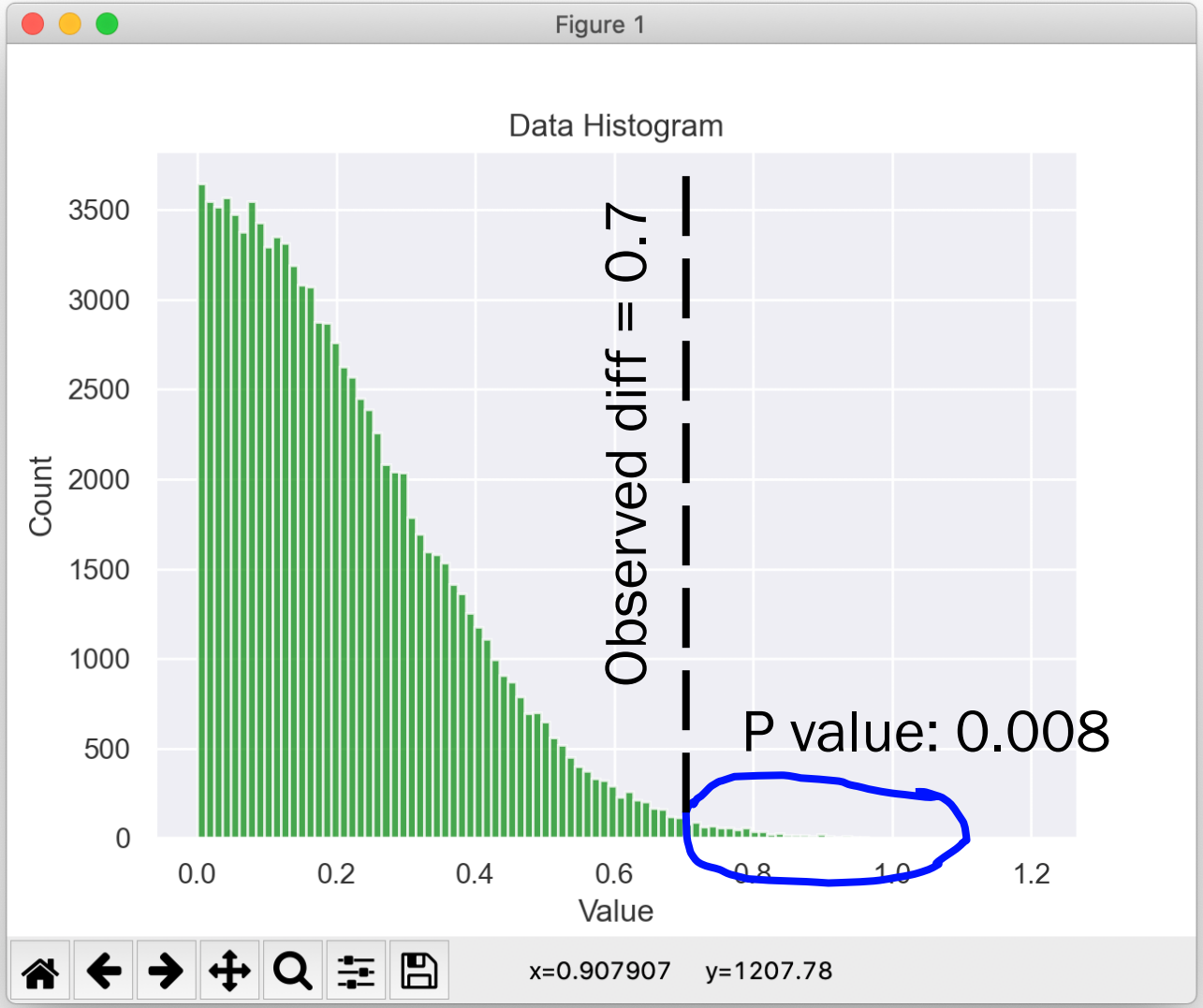
To the code!

Distribution of Mean Diffs under Null Hypothesis



Every* Science Result needs a p-value!

* almost



Food For Thought

Two Opinions on Distributions

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

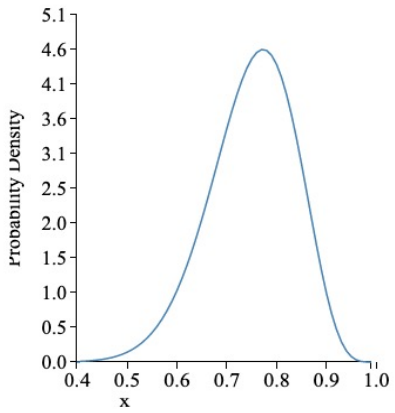
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

Let's use Laplace prior

X ~ Beta(a = 18, b = 6)



Frequentist:

Let's bootstrap