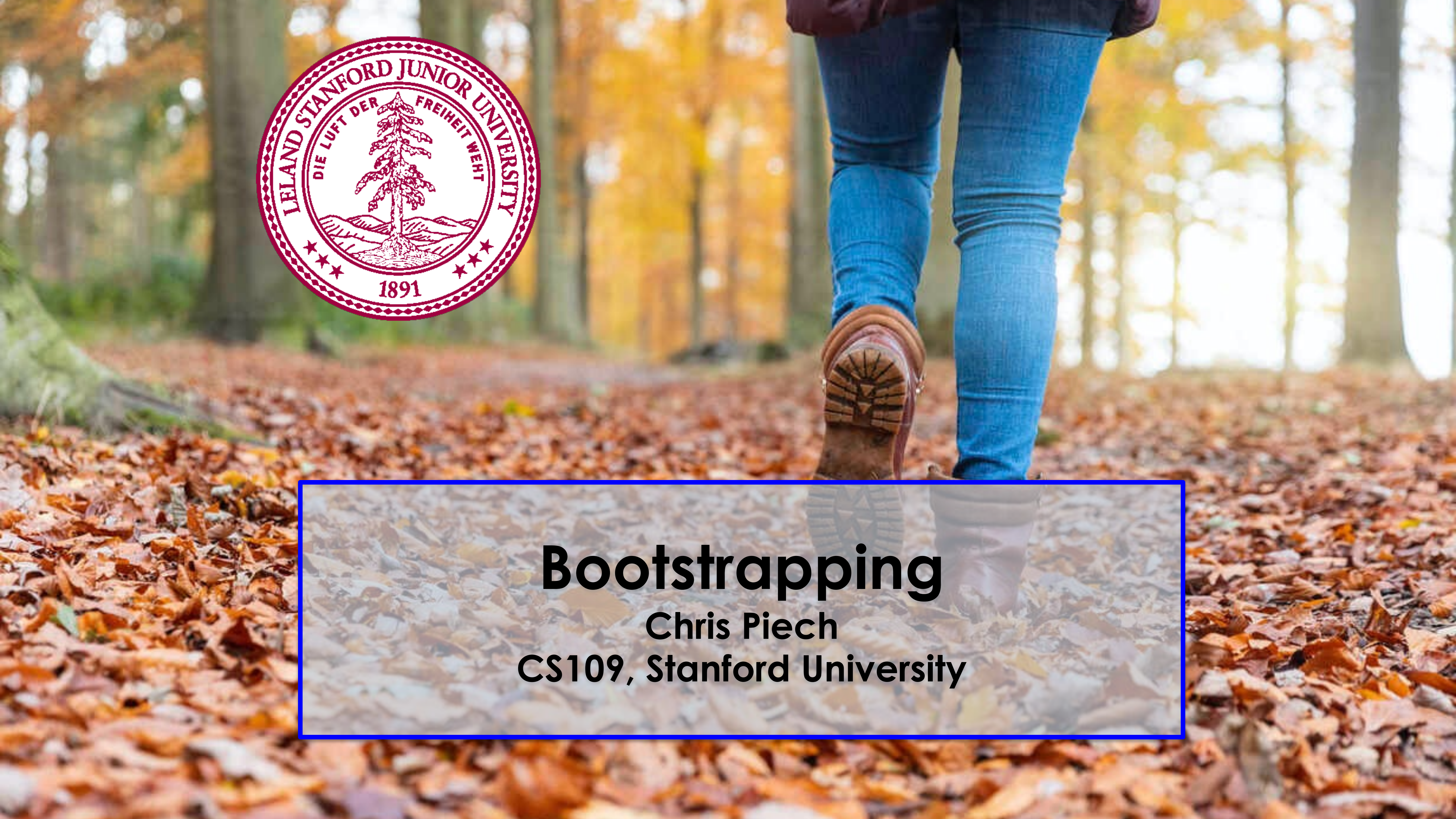


Bootstrapping

Chris Piech

CS109, Stanford University

A real difference?

Learning in Context A
4.44
3.36
5.87
2.31
…
3.70

18 students

$$\mu_1 = 3.1$$

Learning in Context B
2.15
3.01
2.02
1.43
…
1.83

23 students

$$\mu_2 = 2.4$$

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
4.44
3.36
5.87
2.31
…
3.70

Group 2
2.15
3.01
2.02
1.43
…
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?

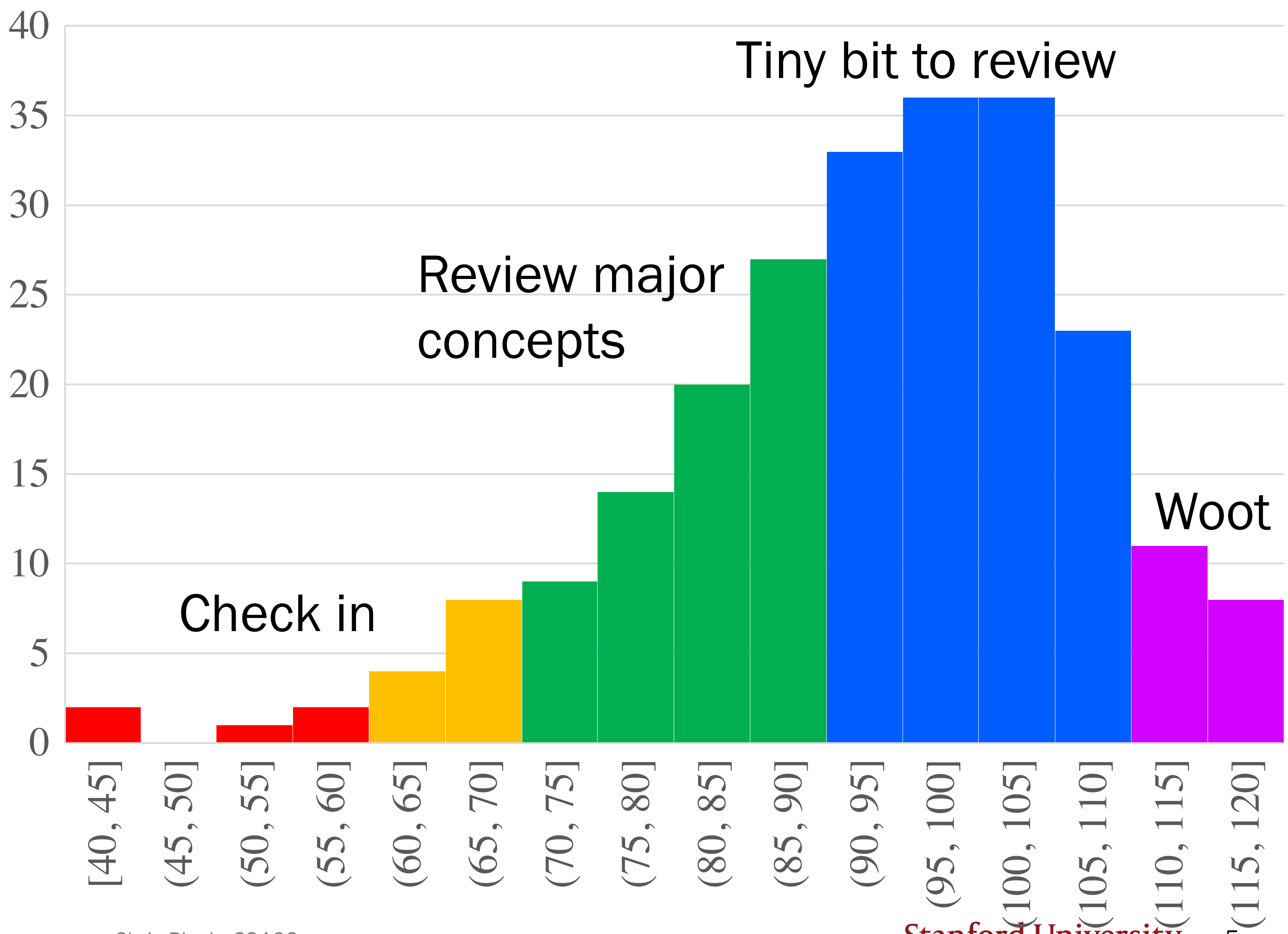
<review>

Logistics

Regrade requests until next Monday

This was a diagnostic. You can show what you know in other ways:

1. Improvement between midterm and final
2. Challenge!



NOV
29TH

Where are we in CS109?

You are here

Counting Theory

Core Probability

Random Variables

Probabilistic Models

Uncertainty Theory

Machine Learning

Uncertainty Theory

- Beta Distributions
- Thompson Sampling
- Adding Random Vars
- Central Limit Theorem
- Sampling
- Bootstrapping
- Algorithmic Analysis

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 $\text{Var}(X_i) = \sigma^2$.

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 $\mathrm{Var}(X_i) = \sigma^2$.

As $n \to \infty$

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

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

Sampling definitions

Motivating example

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

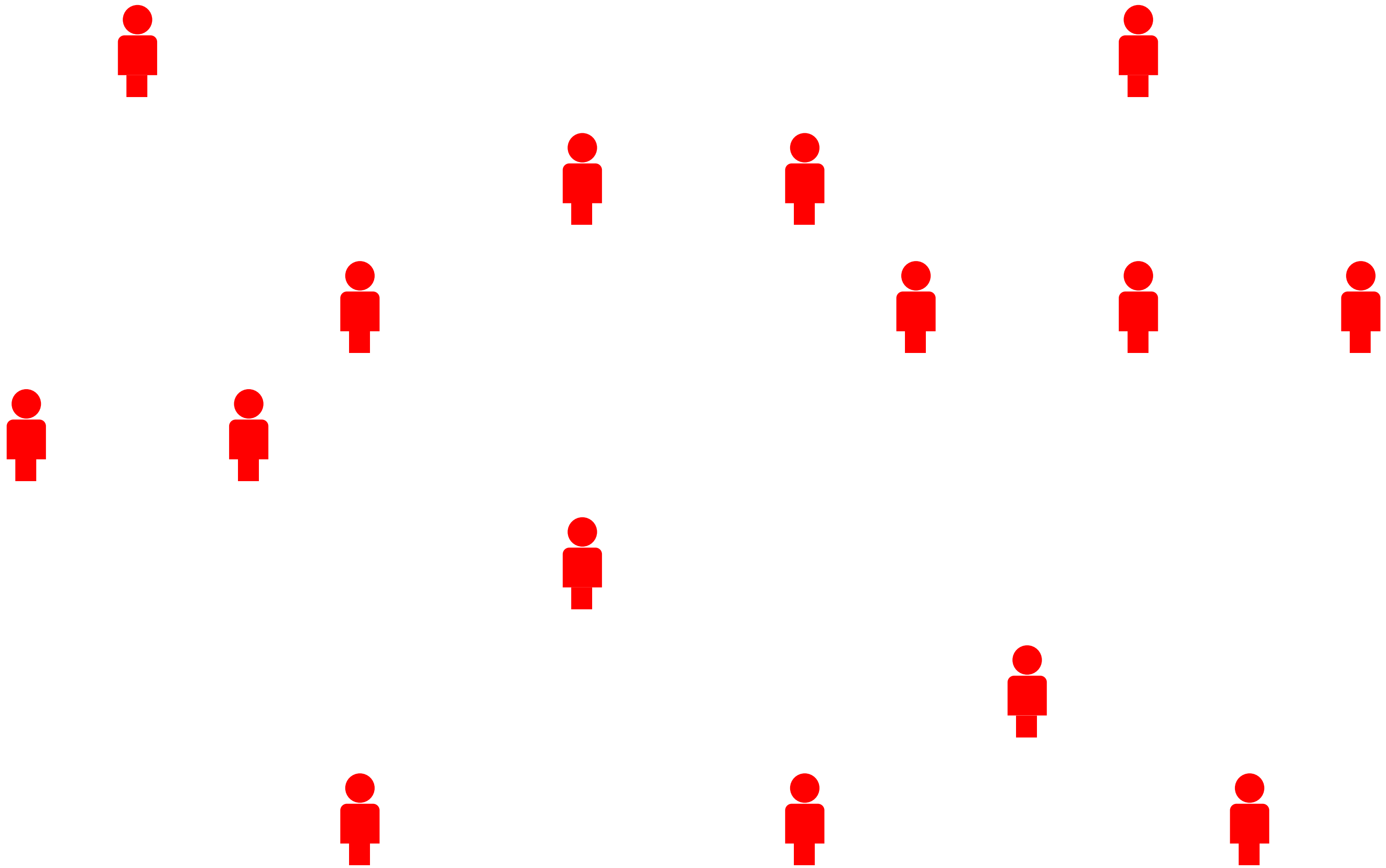
- But you can't ask everyone.
- You poll 200 random people.
- Your data looks like this:

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

Population

Sample

Sample



Collect one (or more) numbers from each person

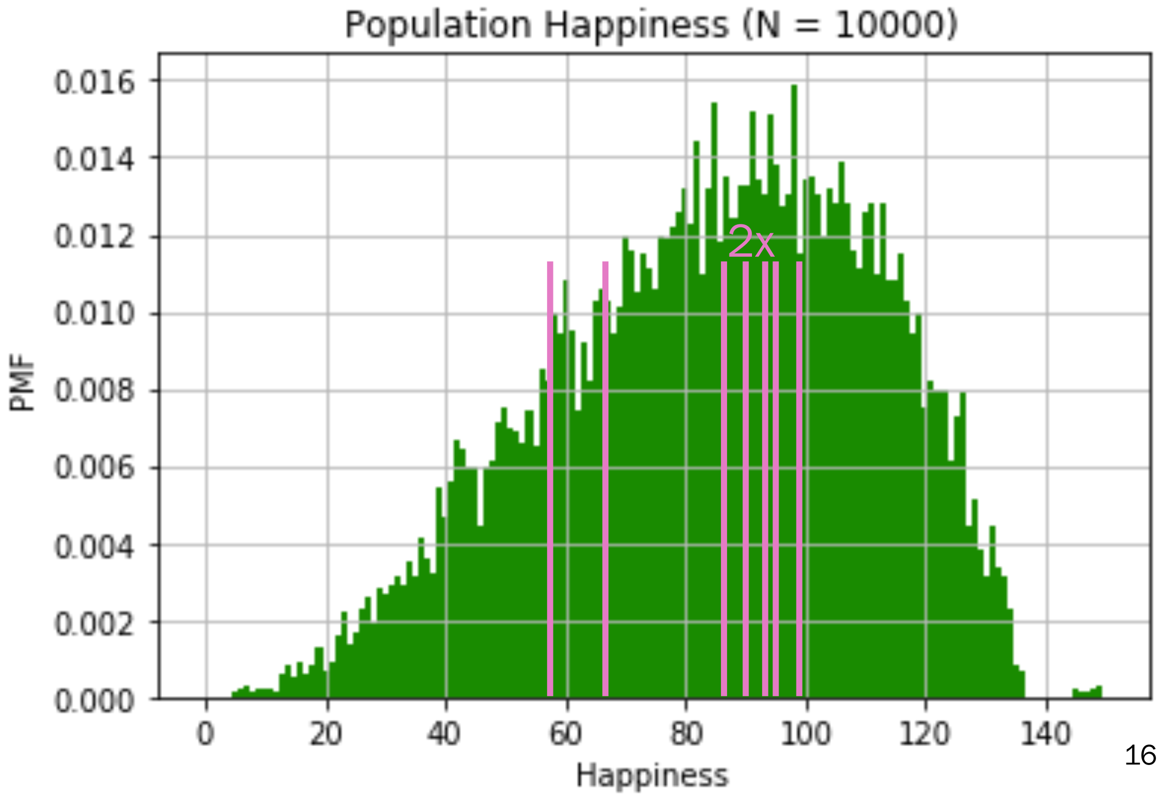
A sample, mathematically

A sample of **sample size** 8:

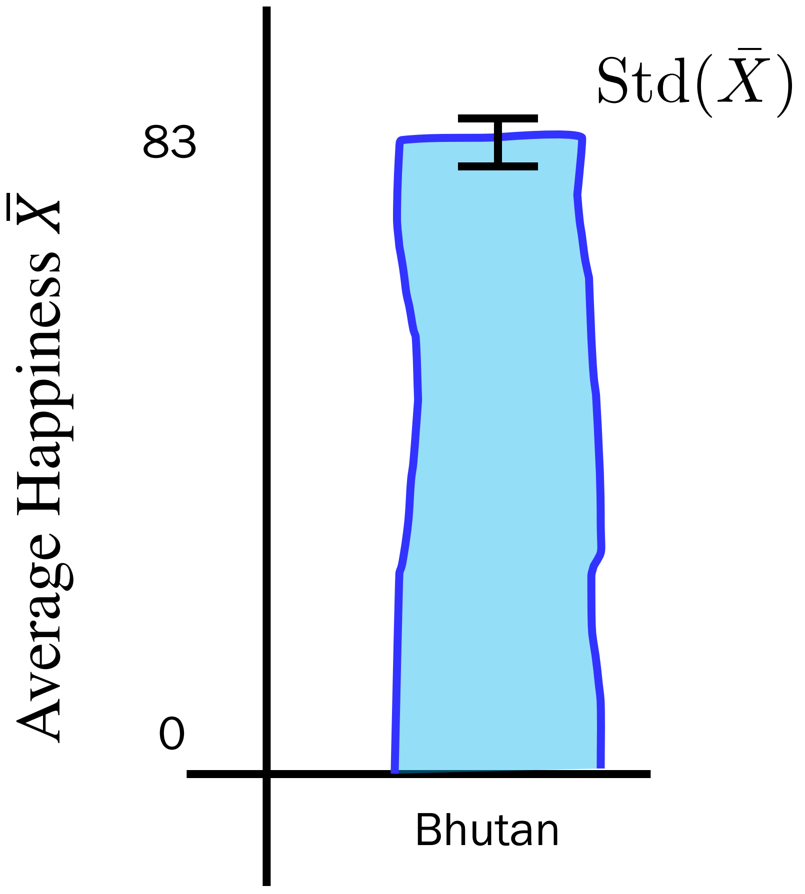
$$(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

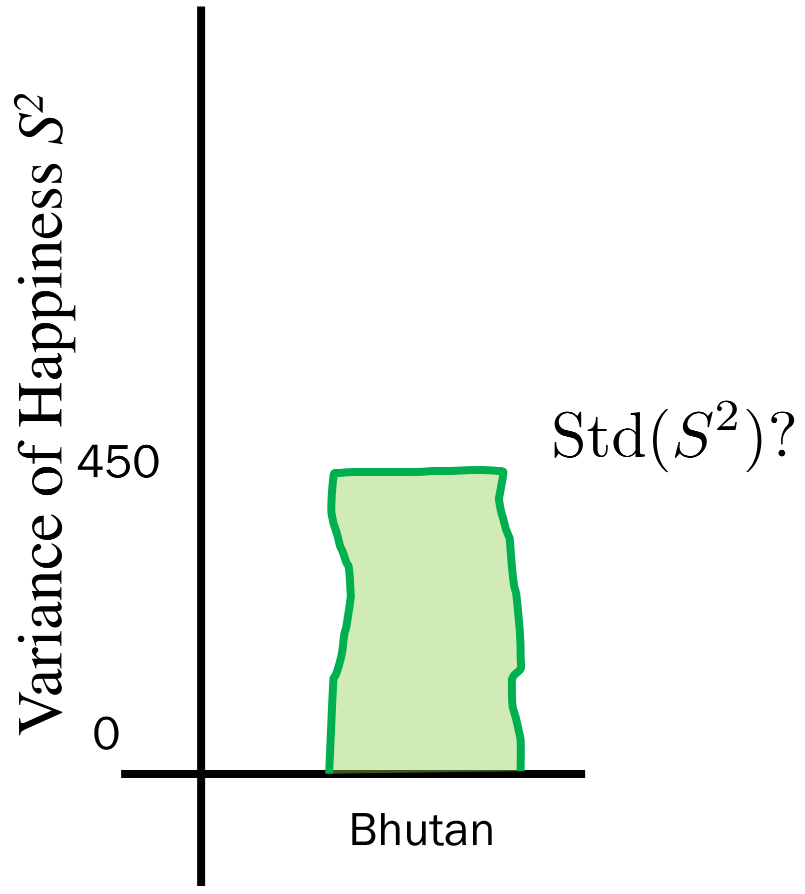
A **realization** of a sample of size 8:

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



Our Report to Bhutan Government





Claim: The average happiness of Bhutan is 83 ± 2

Equations we used to get those values

sample mean

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

Our best guess at the true mean

sample variance

sample mean

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

Our best guess at the true variance

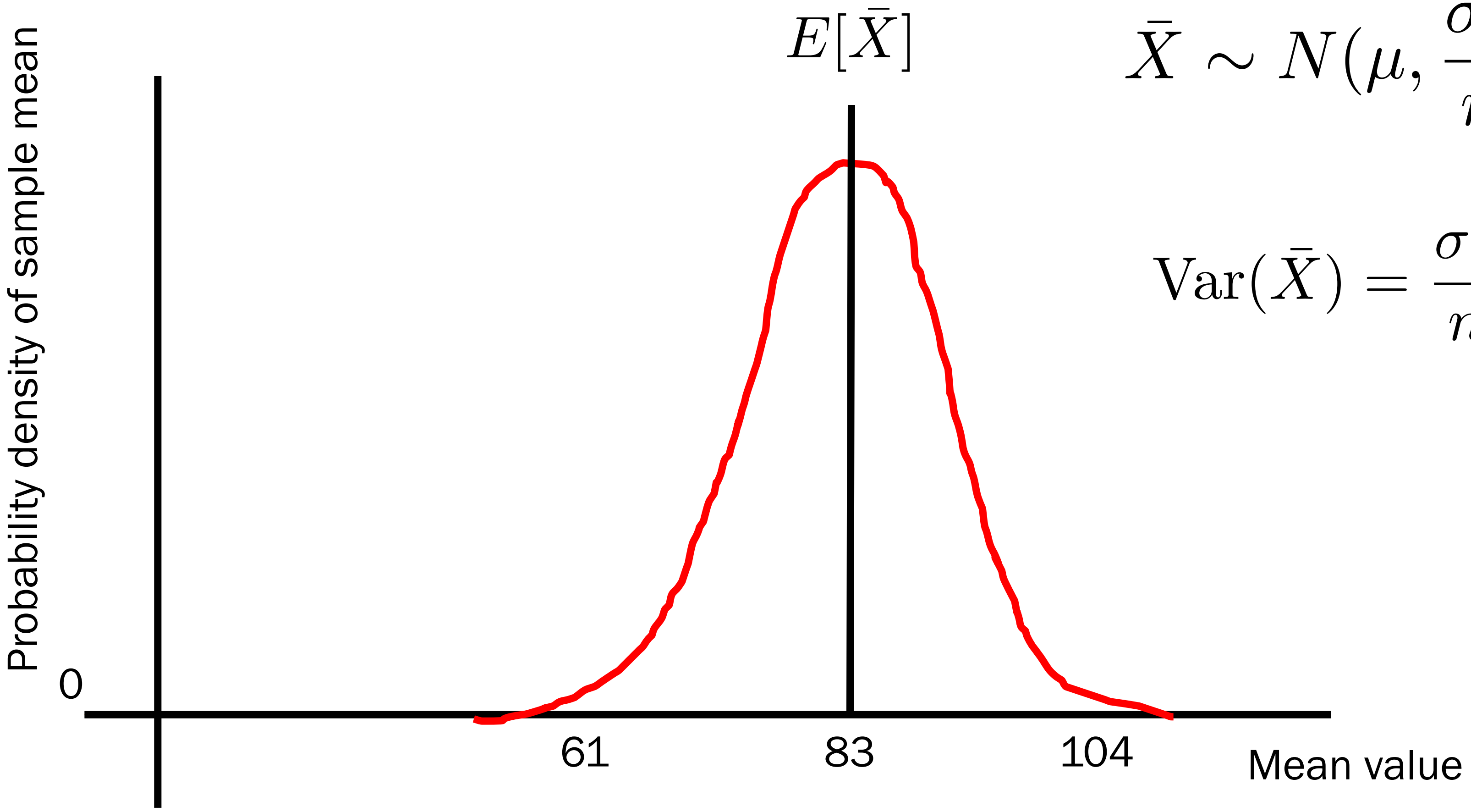
Std error of the mean

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

sample variance

How wrong do we think our mean estimate is?

Insight: Sample Mean is an RV with known Var

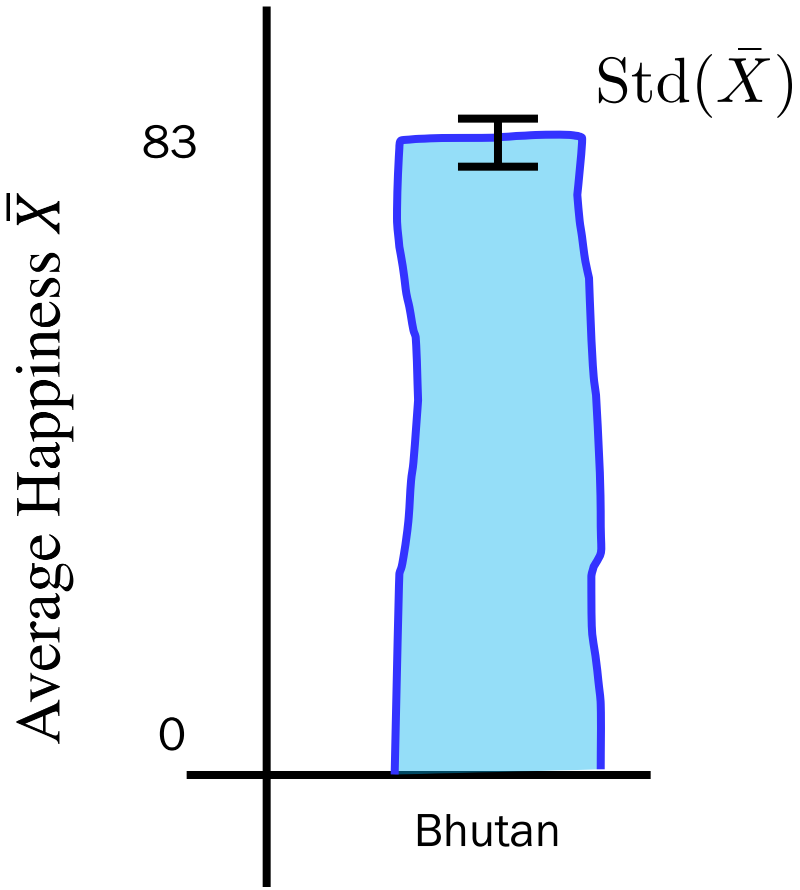


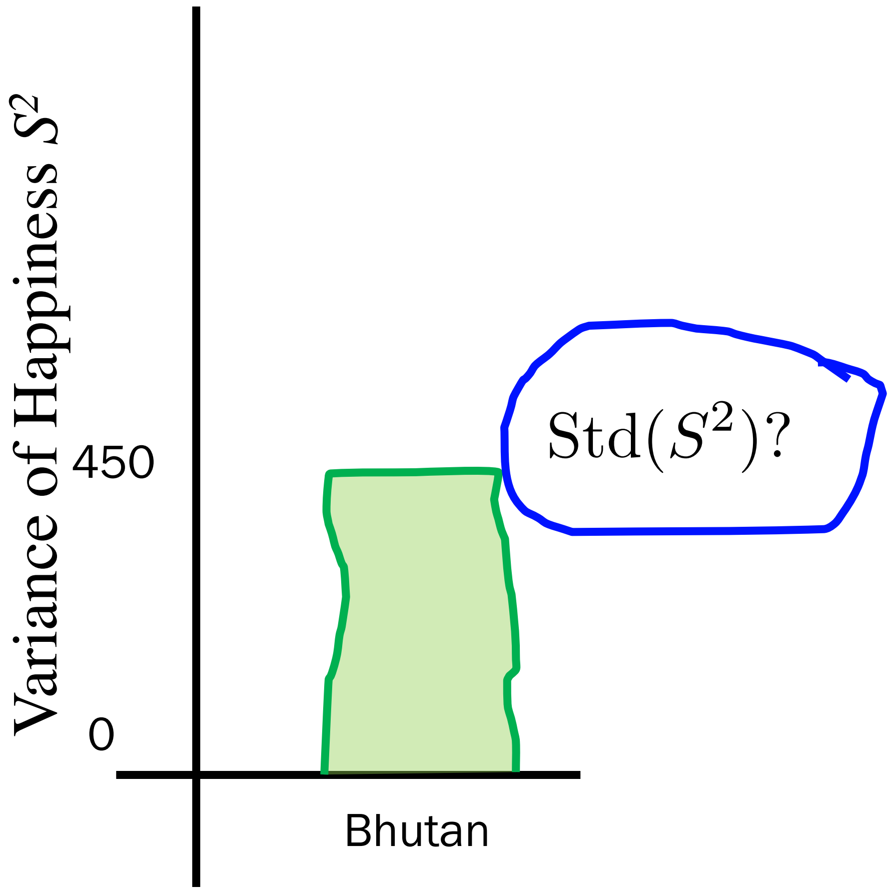
By central limit theorem:

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

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

Our Report to Bhutan Government





Claim: The average happiness of Bhutan is 83 ± 2

<end review>

Bootstraping

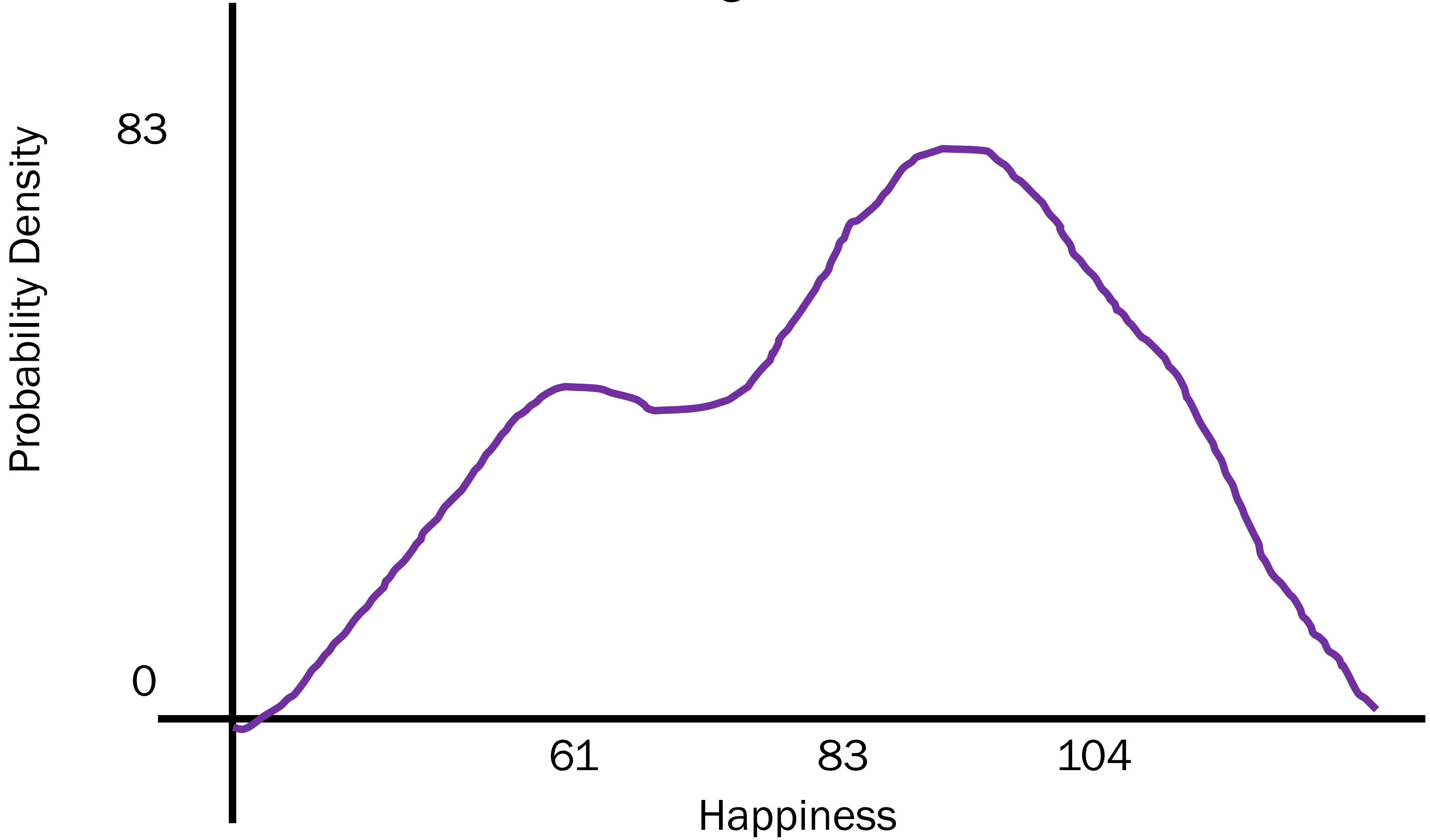
Bootstrap:
Probability for Computer Scientists

Bootstraping allows you to:

- Know the **distribution of statistics**
- Calculate **p values**

Hypothetical

What is the probability that the **mean** of a sample of 200 people is within the range 81 to 85?

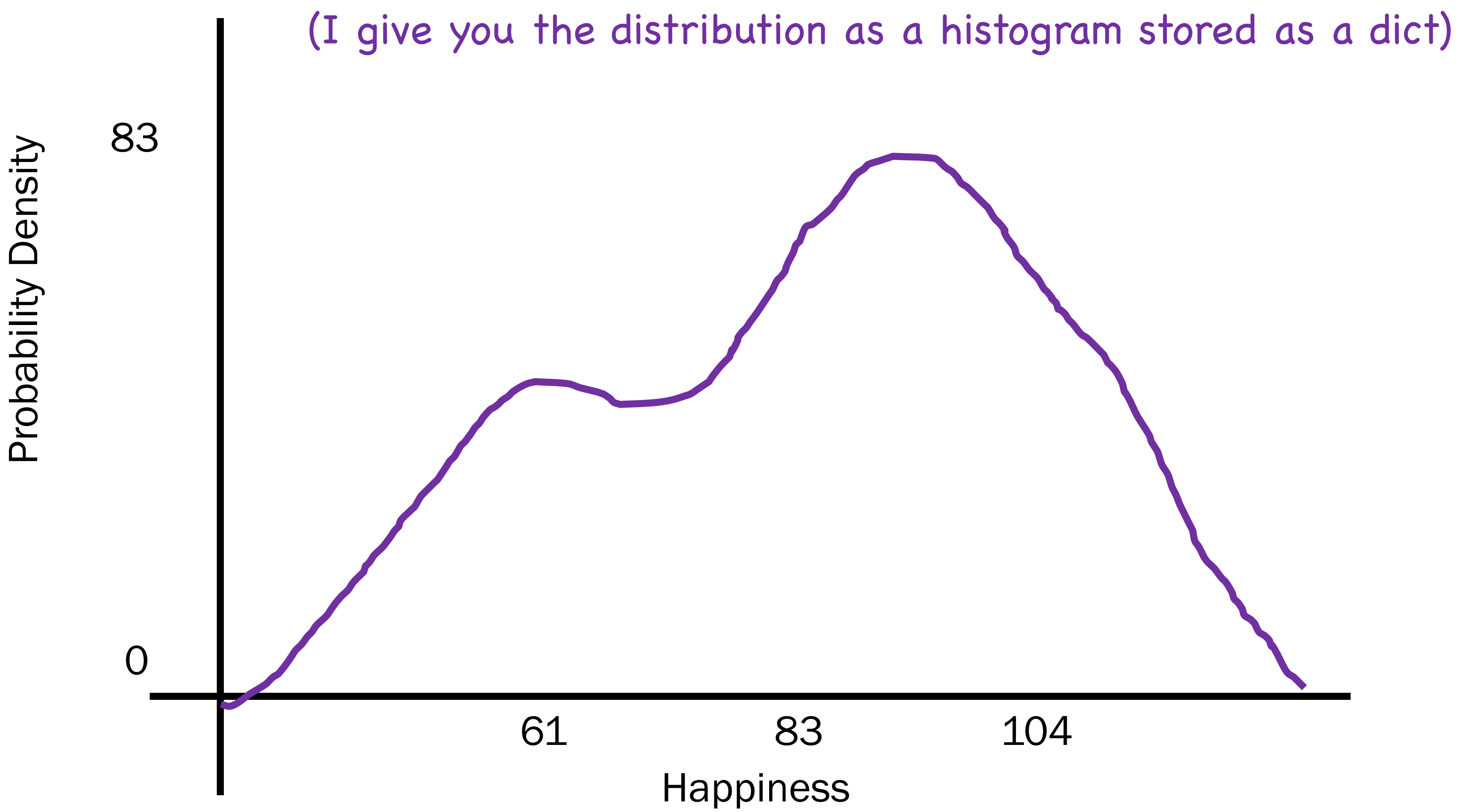


Hypothetical

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

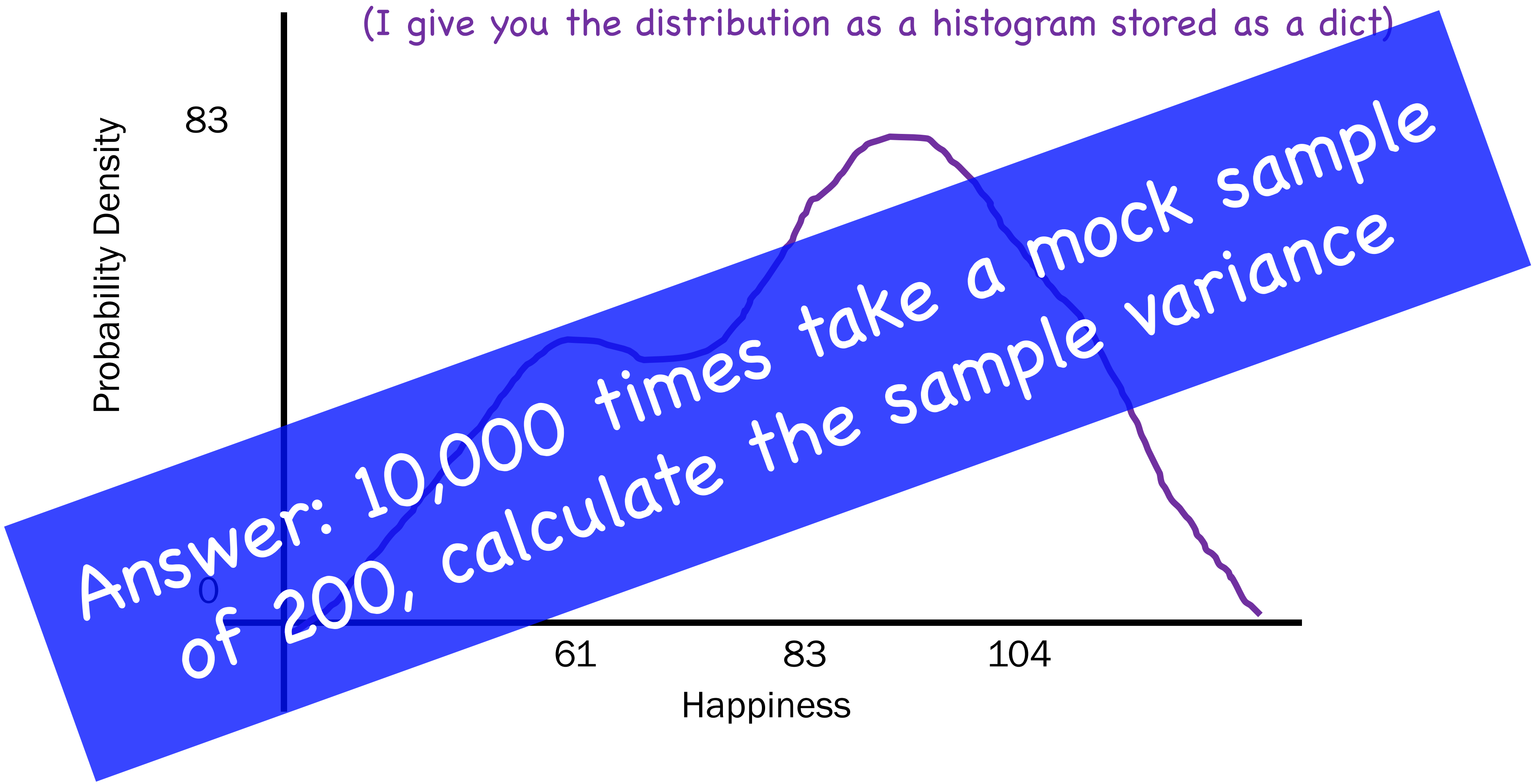
If I Gave You the True Distribution, what would you do?

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



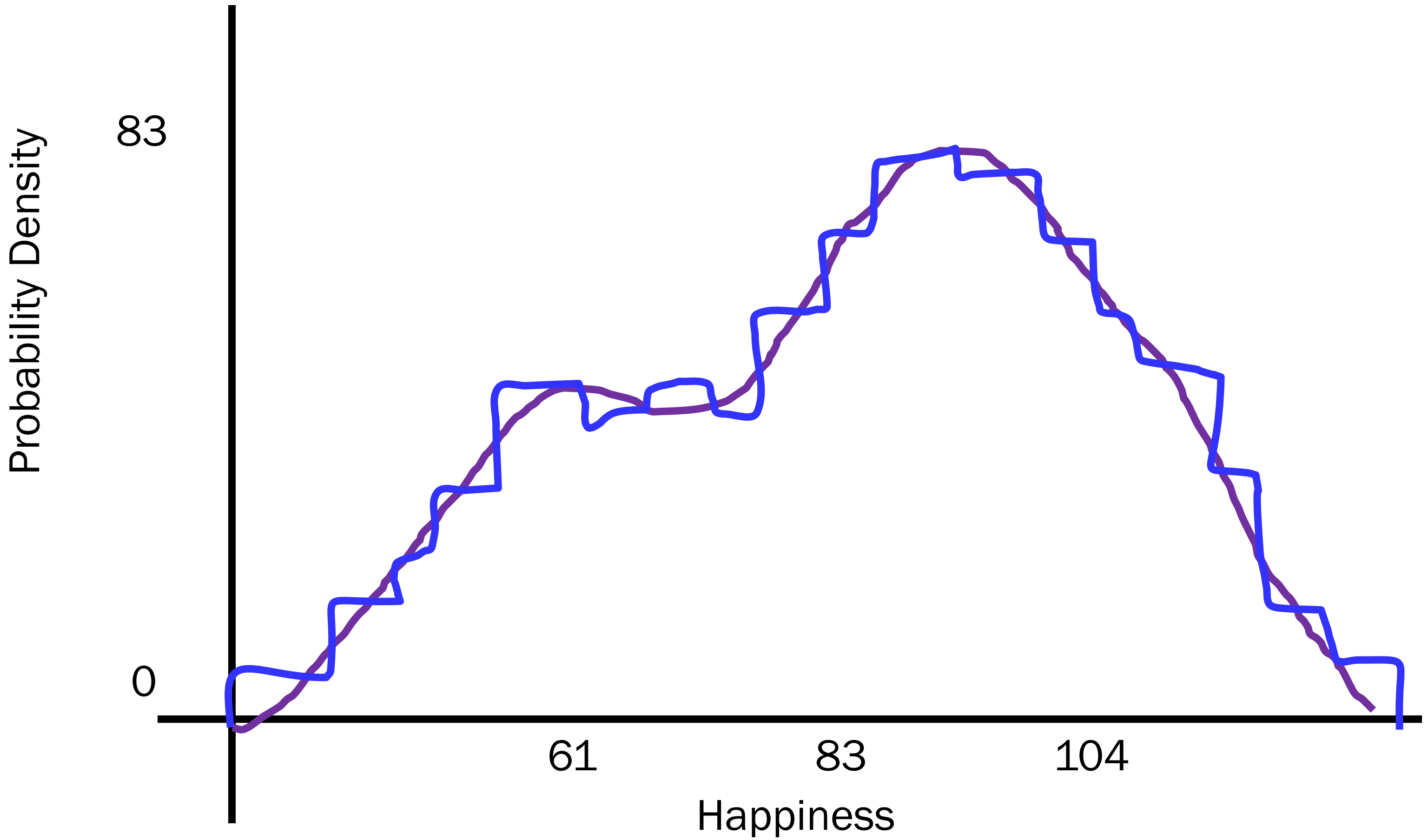
If I Gave You the True Distribution, what would you do?

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



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

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



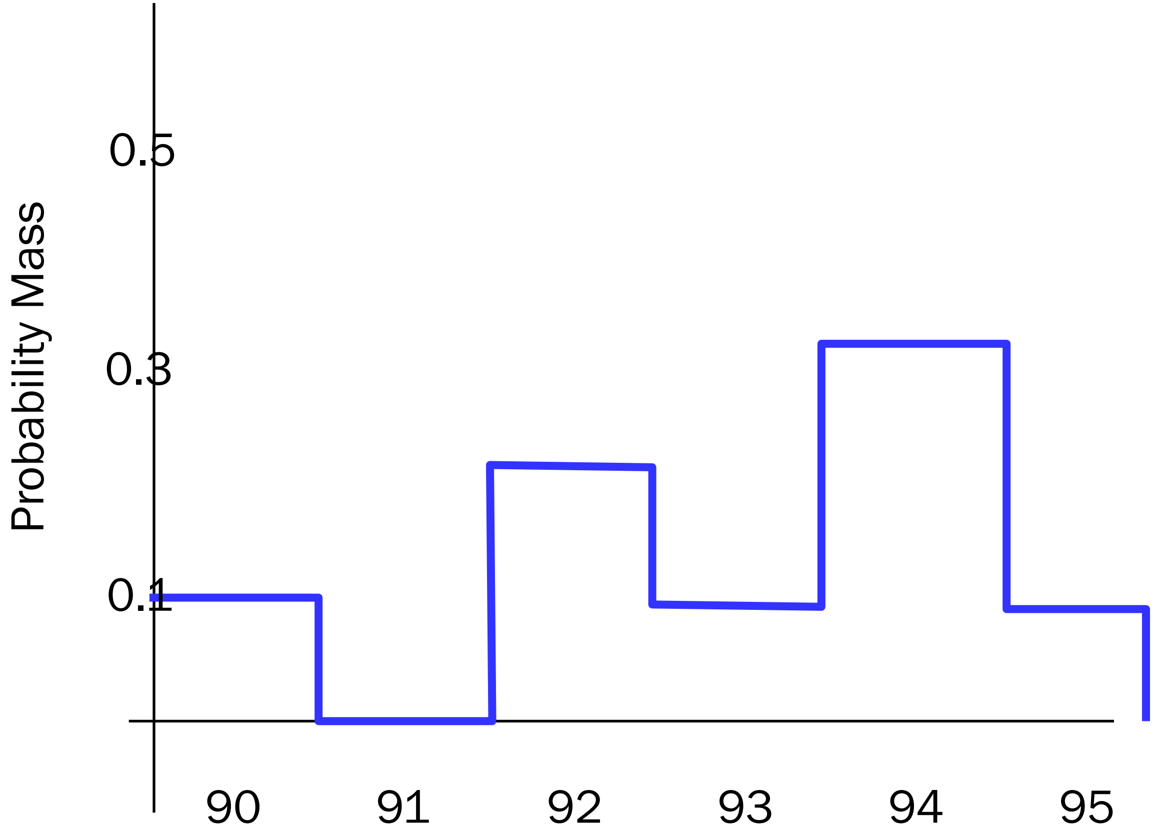
* This is just a histogram of your data!!

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)

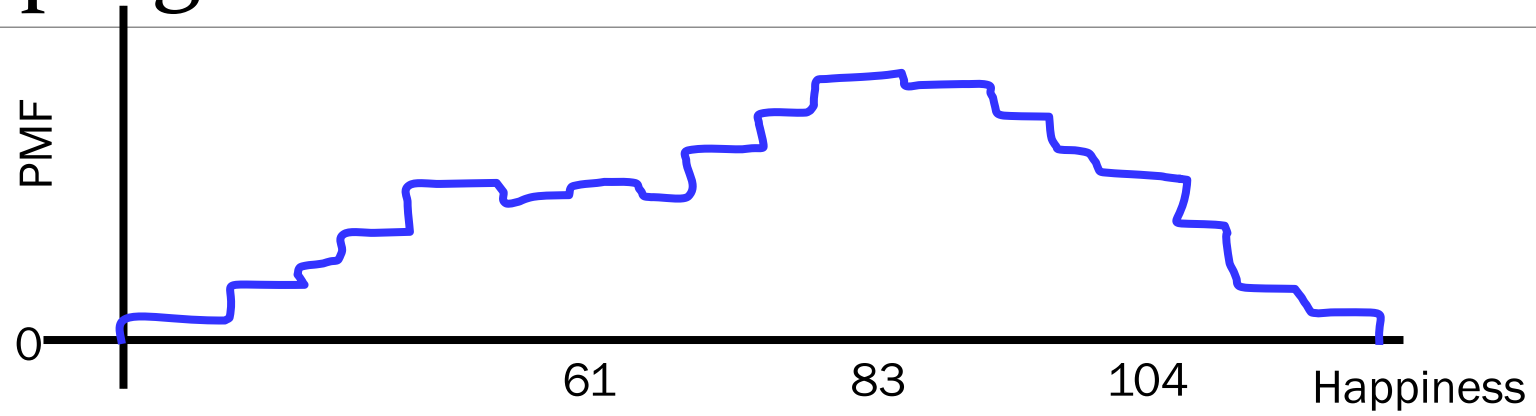
Algorithm

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

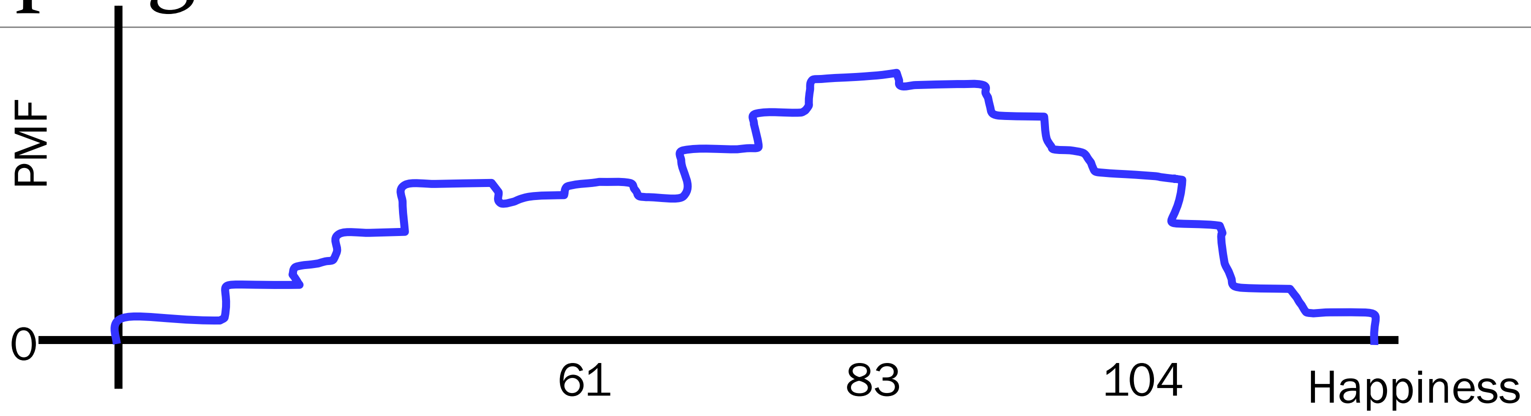
Bootstrapping of Means



Bootstrap Algorithm (sample):

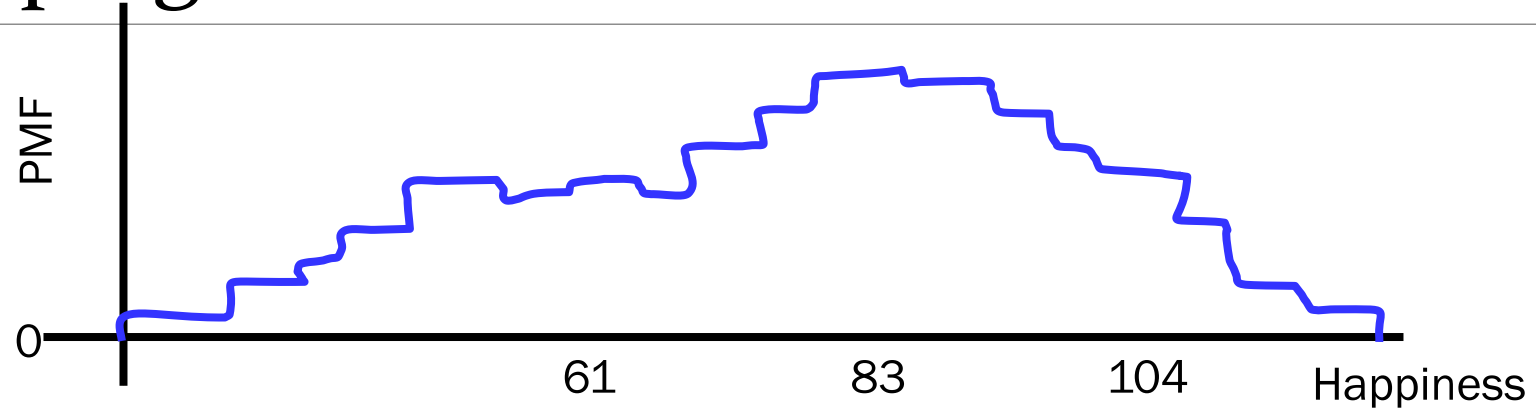
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() new samples from PMF
   b. Recalculate the mean on the resample
3. You now have a **distribution of your means**

Bootstrapping of Means




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

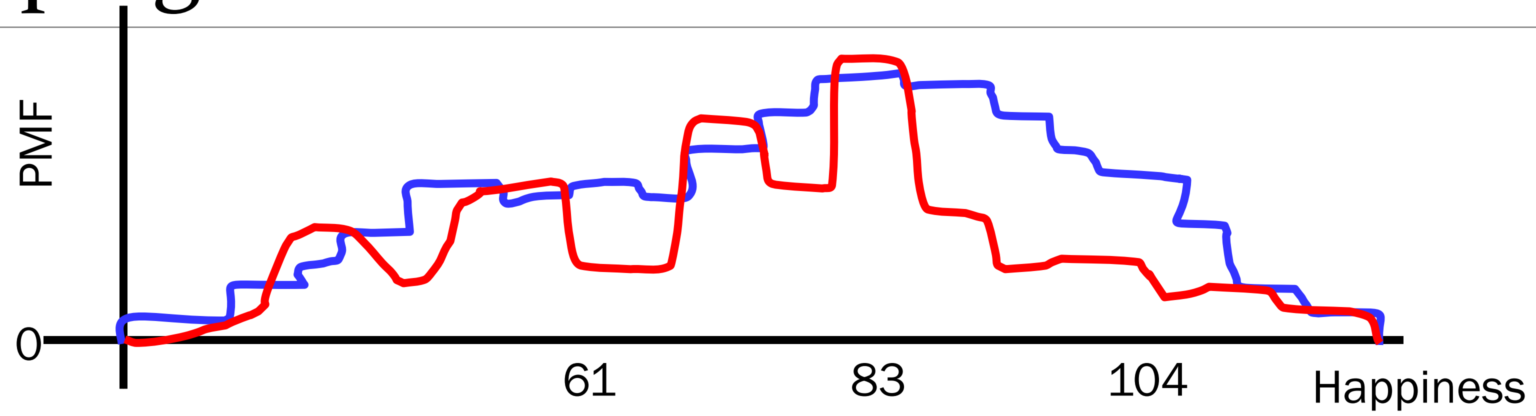
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

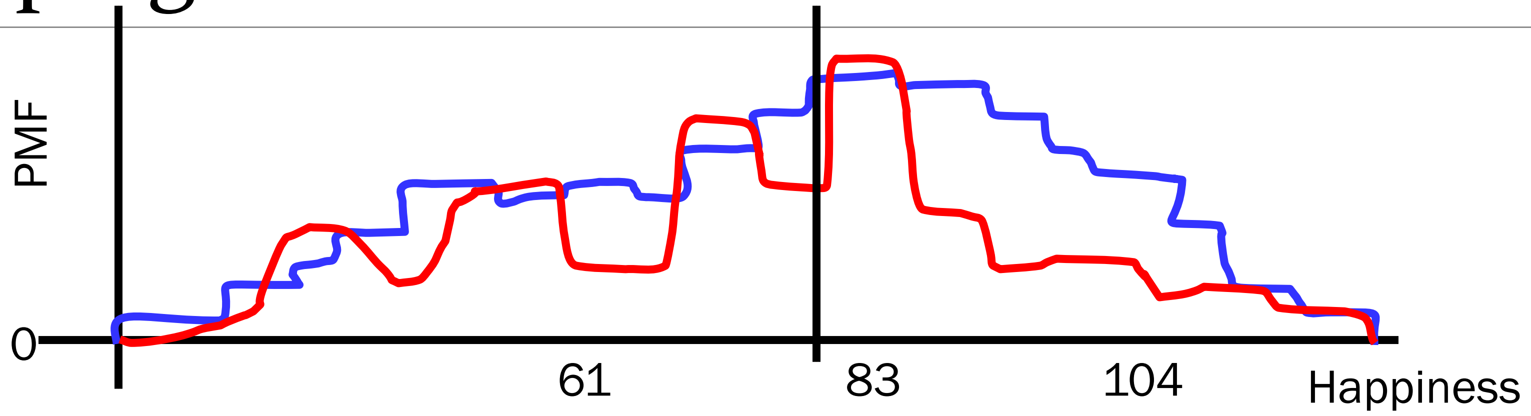
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

Bootstrapping of Means

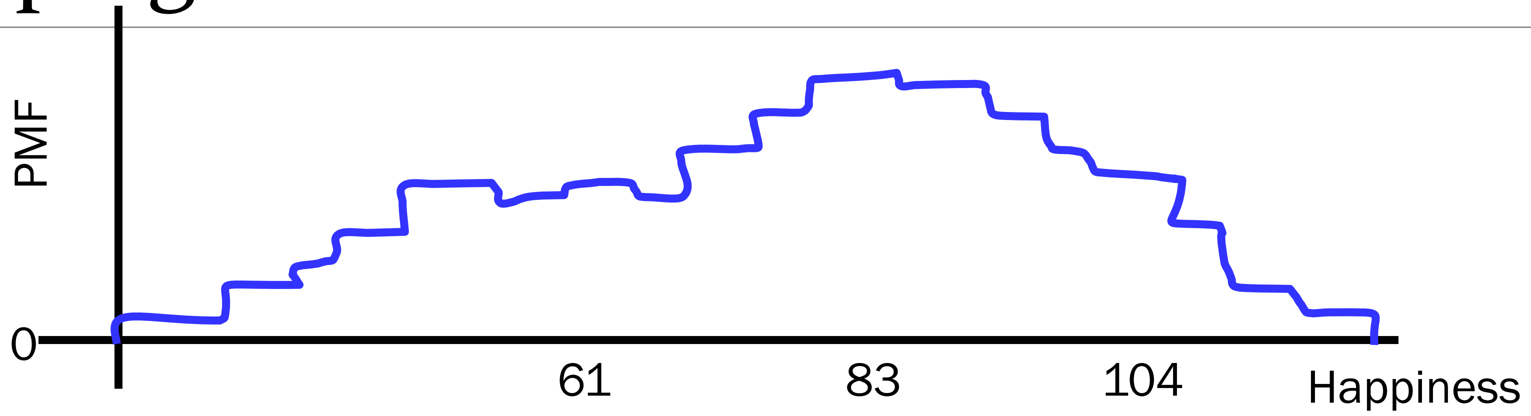


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

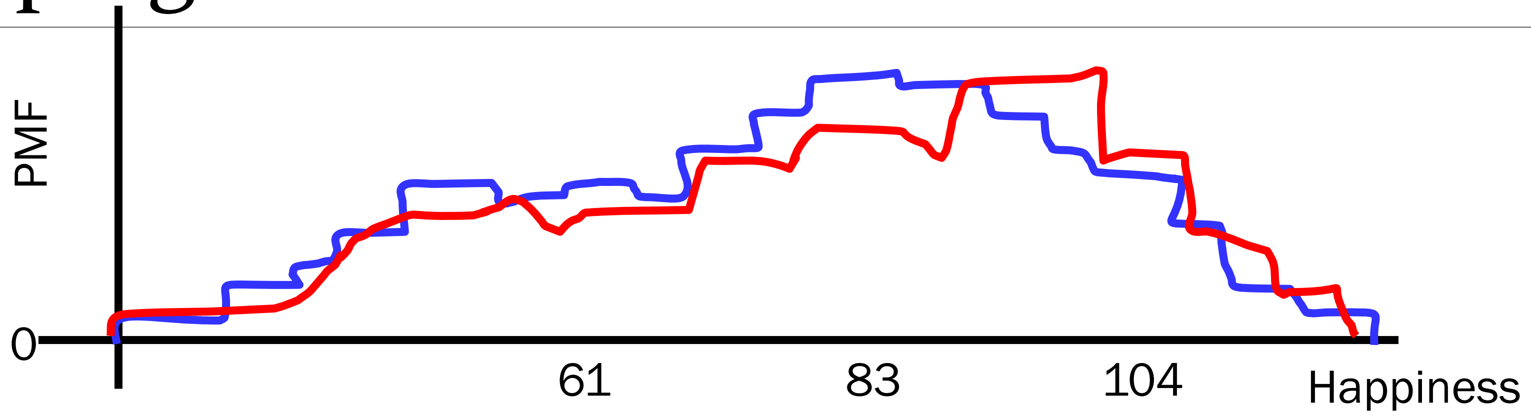


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

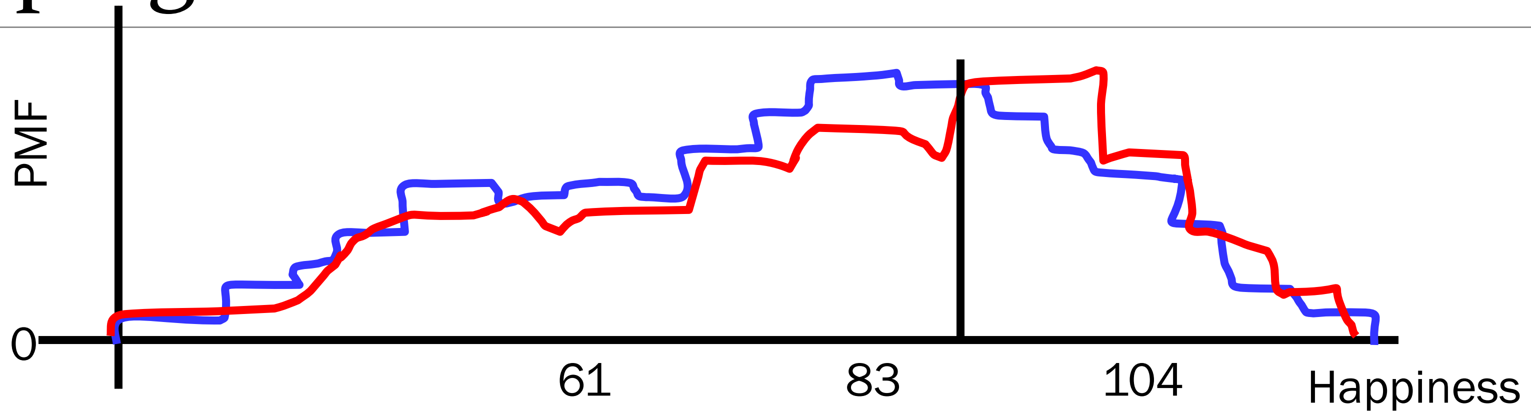


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

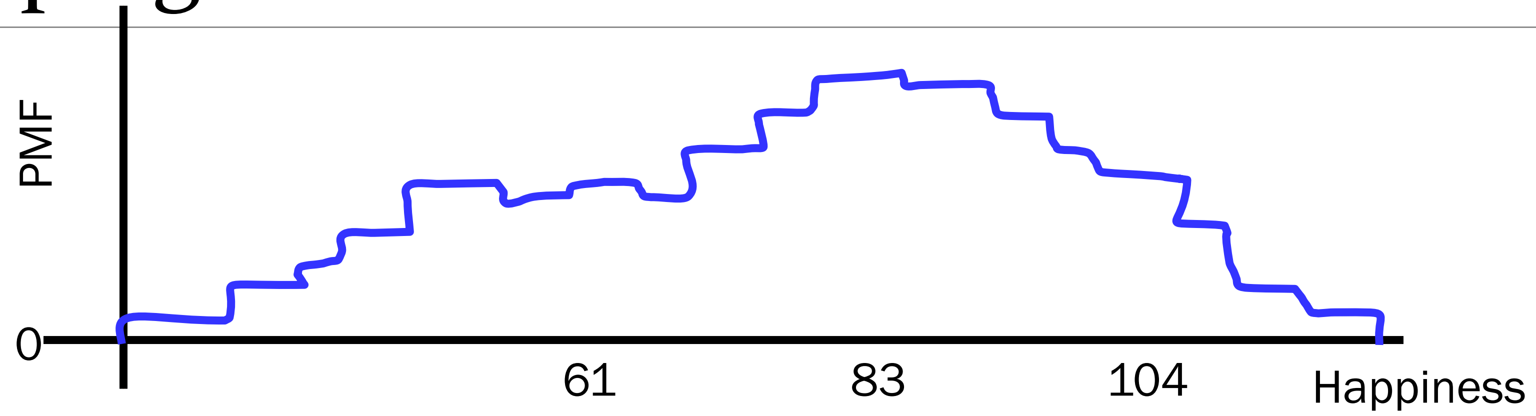


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

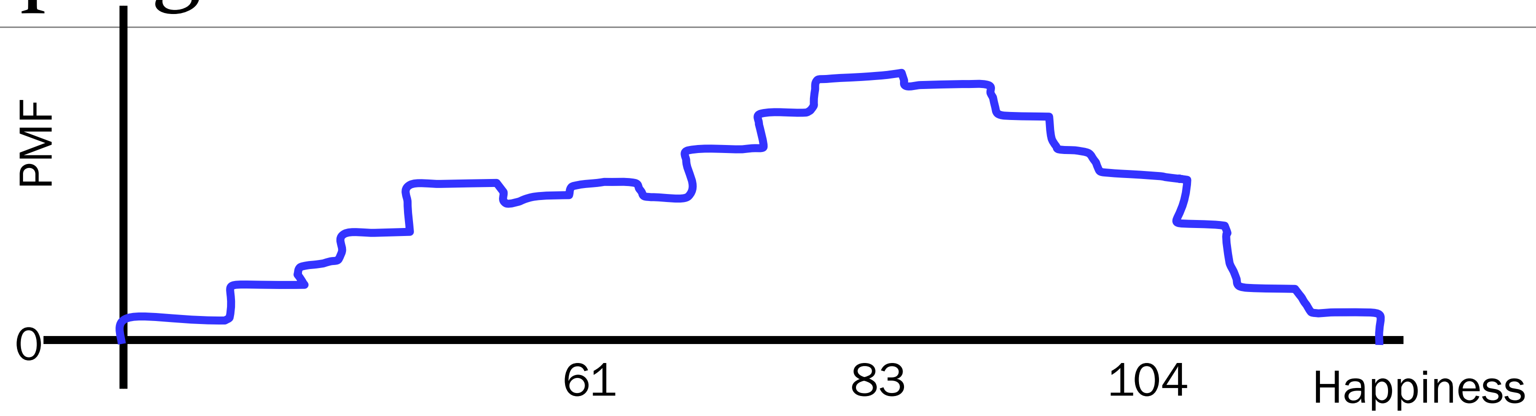


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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



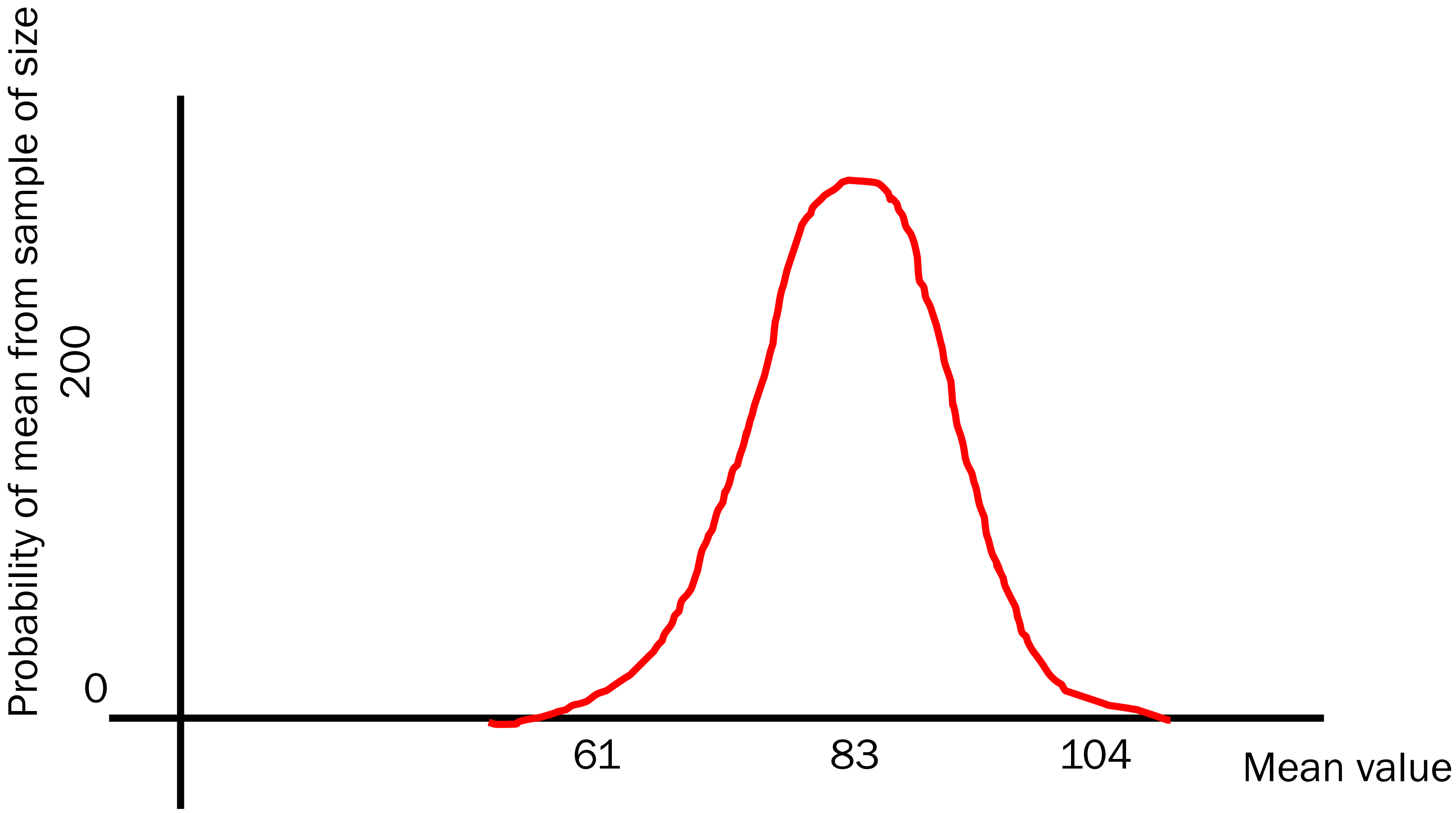
Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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]

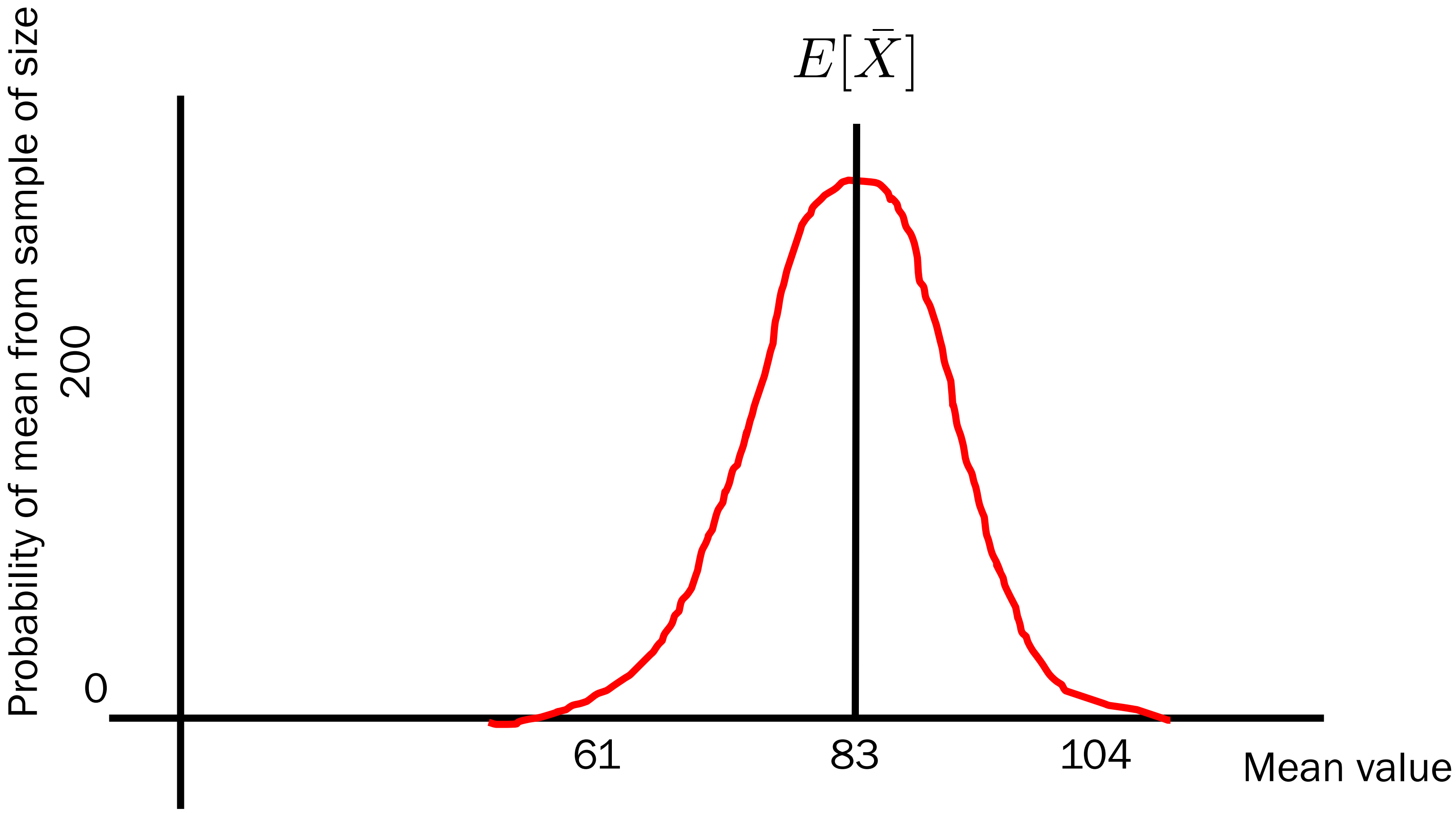
Bootstrapping of Means

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



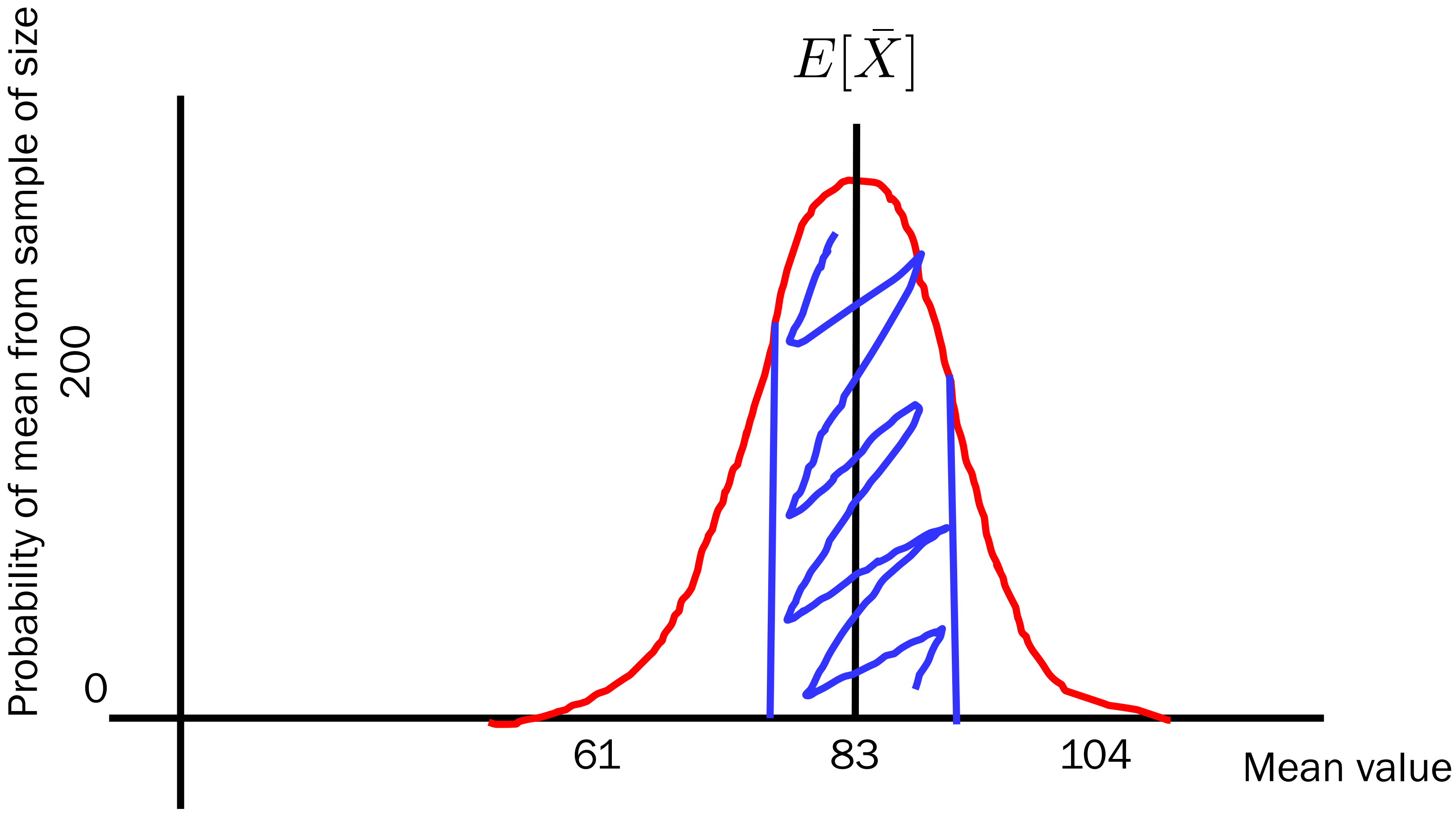
Bootstrapping of Means

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

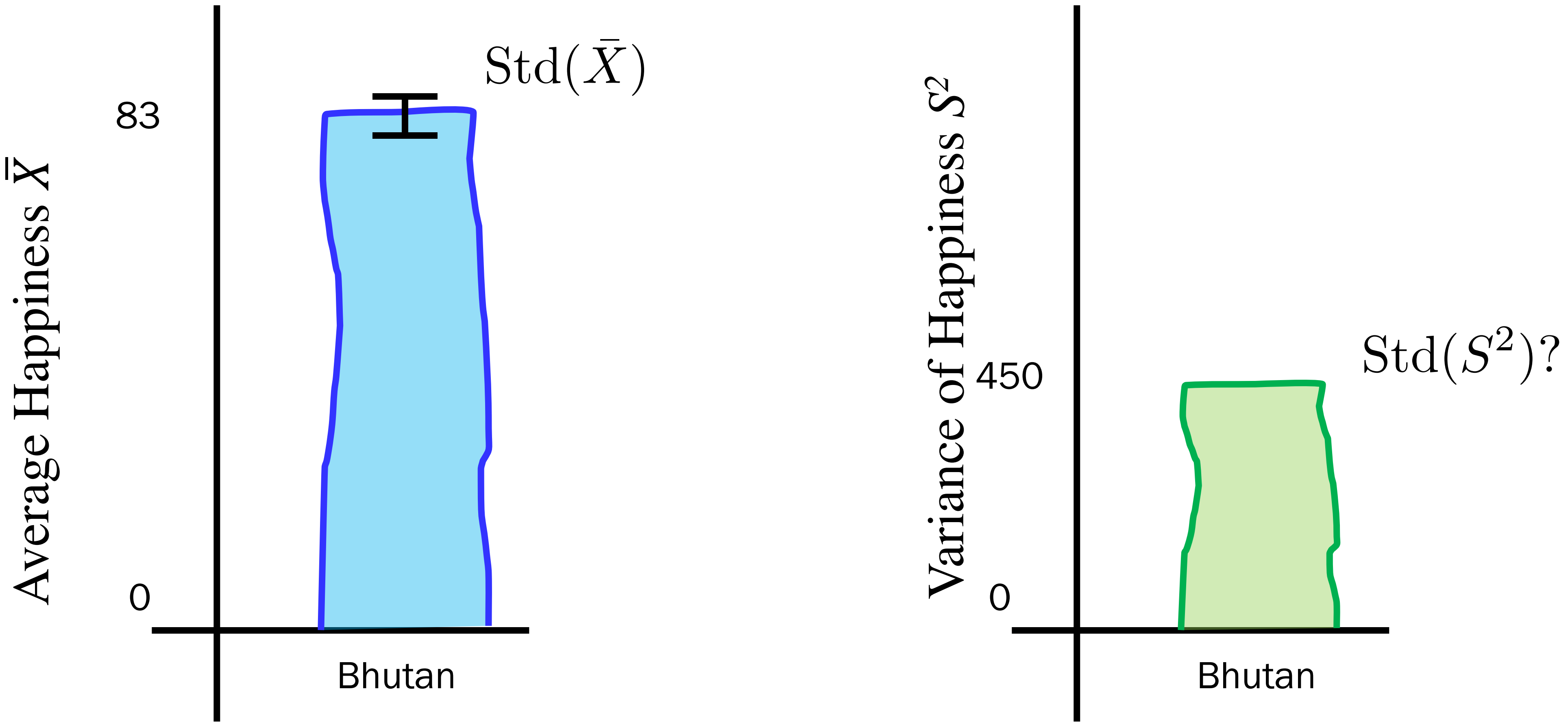


Bootstrapping of Means

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



Our Report to Bhutan Government

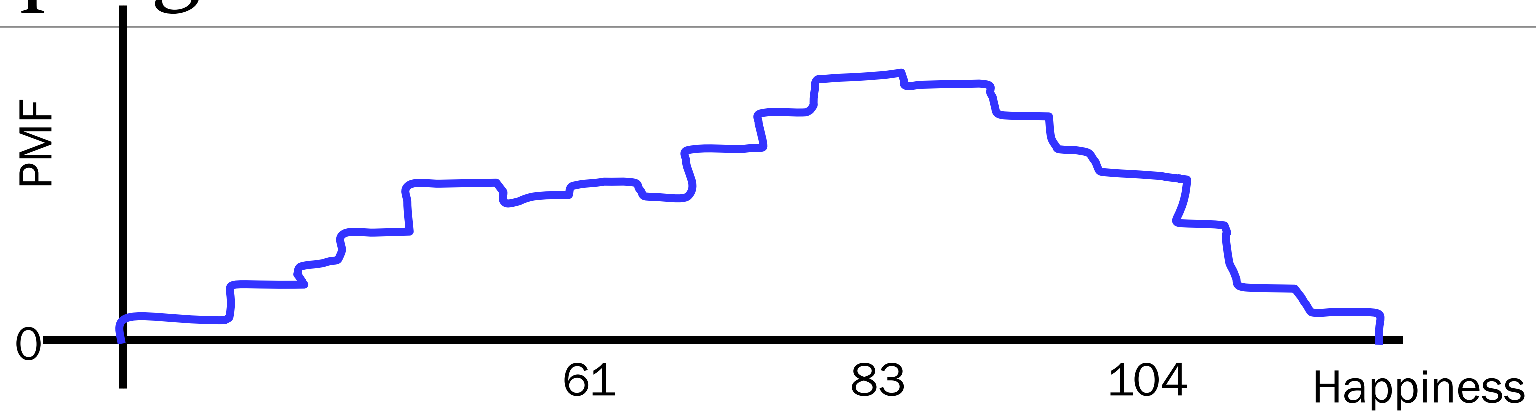


Claim: The average happiness of Bhutan is 83 ± 2

Bootstrapping of Variance

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

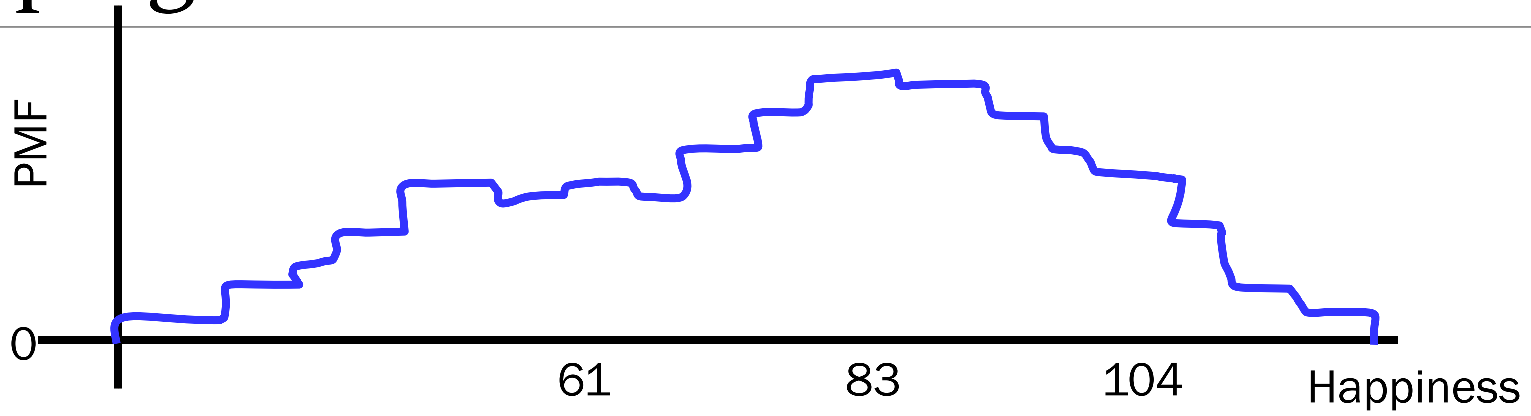
Bootstrapping of Variance



Bootstrap Algorithm (sample):

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

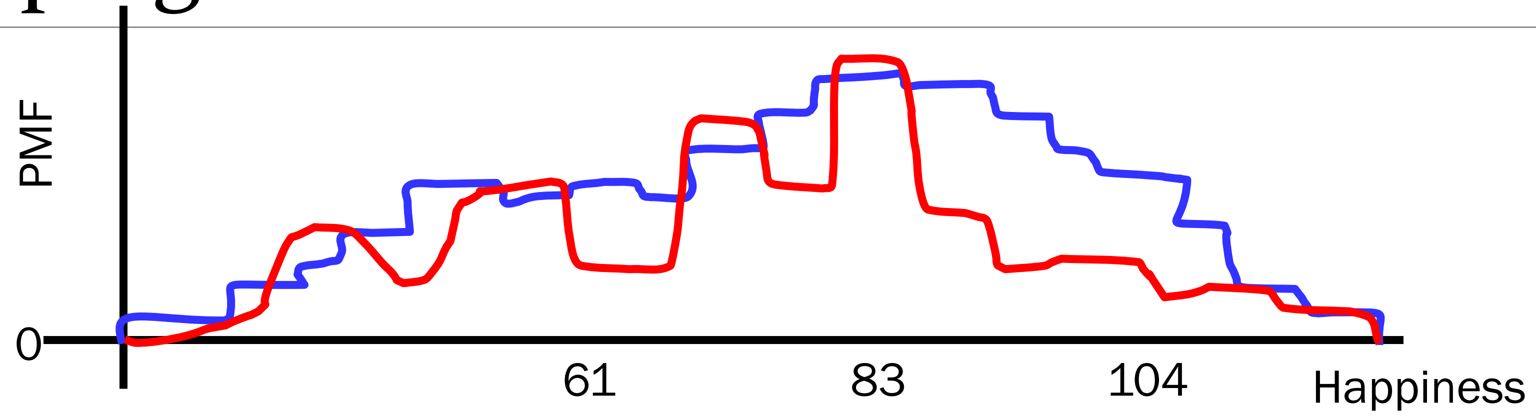
Bootstrapping of Variance




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


Stanford University

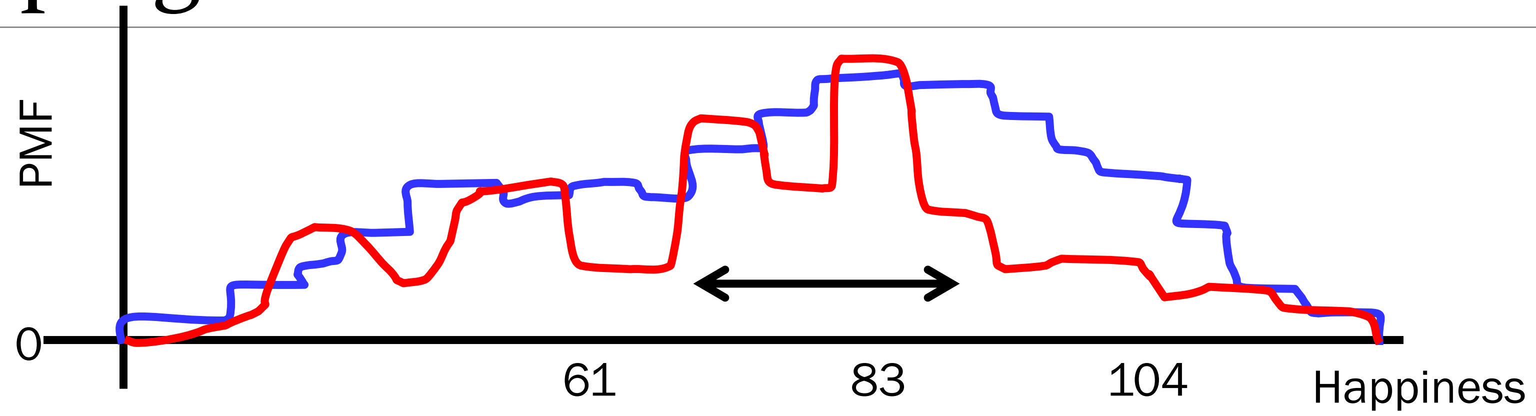
Bootstrapping of Variance



Bootstrap Algorithm (sample):

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

Bootstrapping of Variance

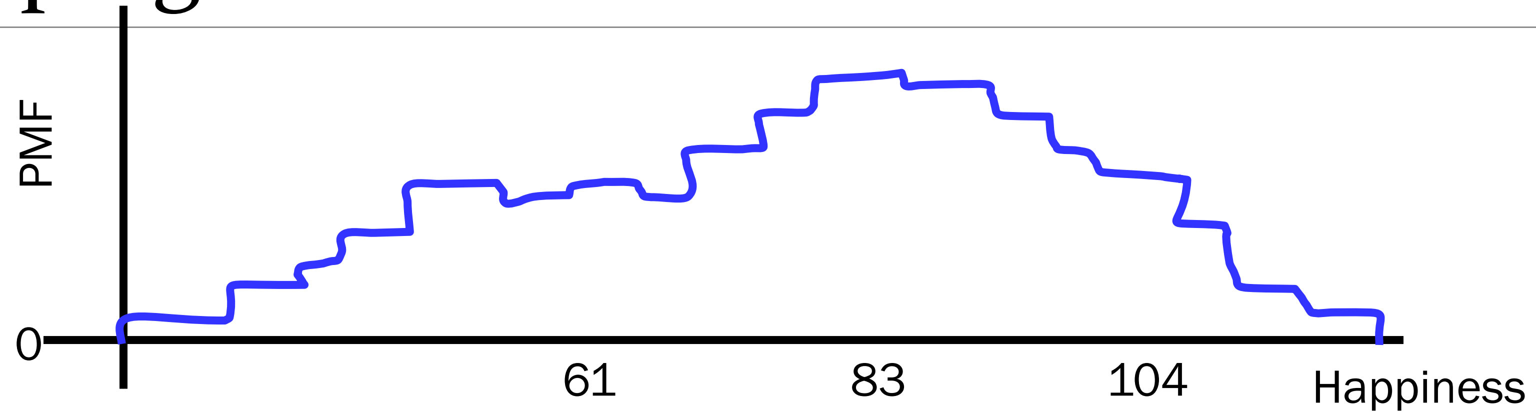


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

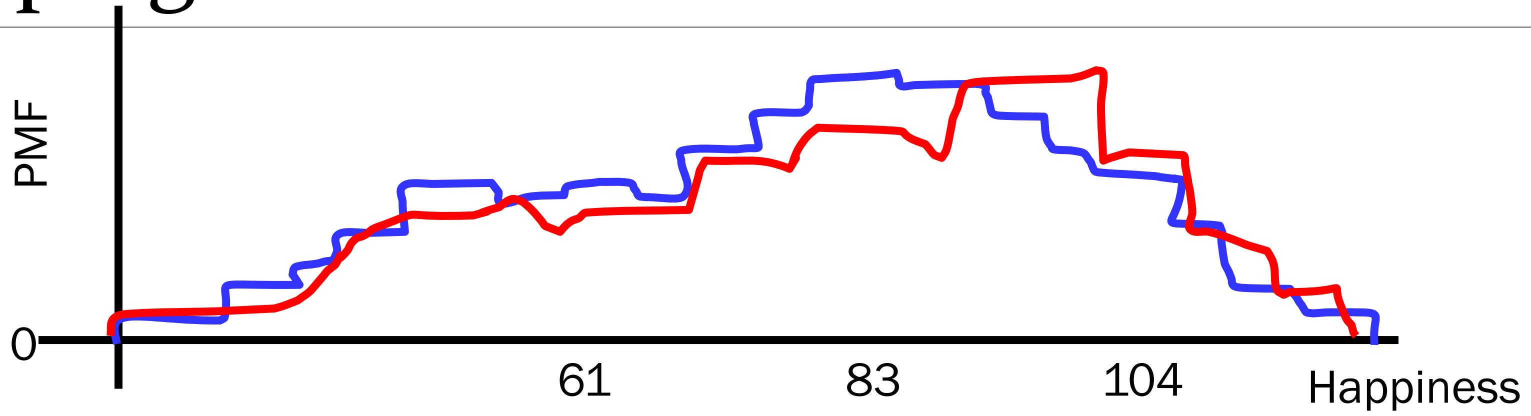


Bootstrap Algorithm (sample):

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

Vars = [472.7]

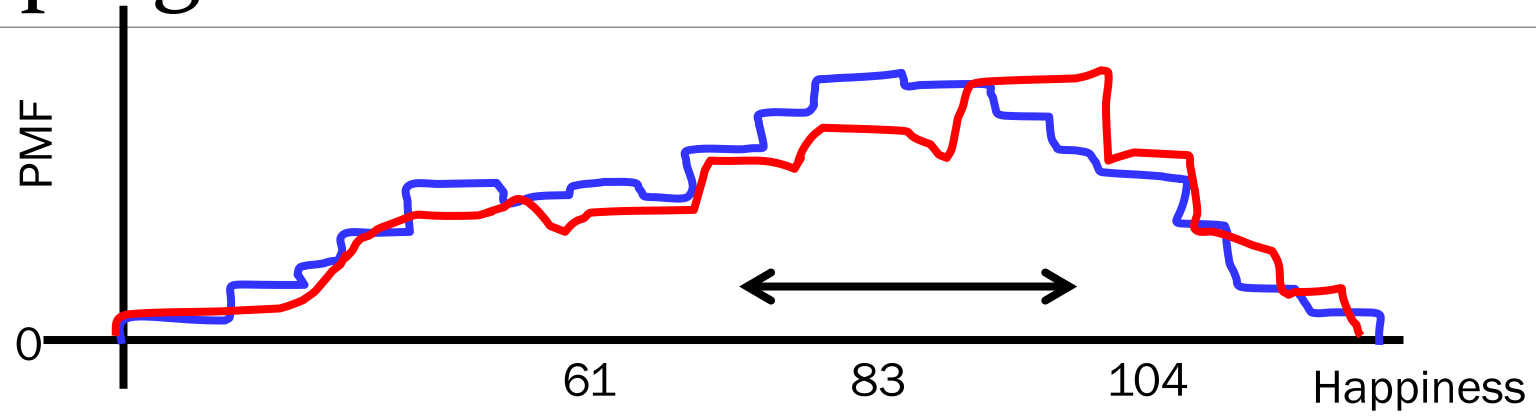
Bootstrapping of Variance




```
Bootstrap Algorithm (sample):
  1.   Estimate the PMF using the sample
  2.   Repeat 10,000 times:
    a.  Draw sample.size() 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

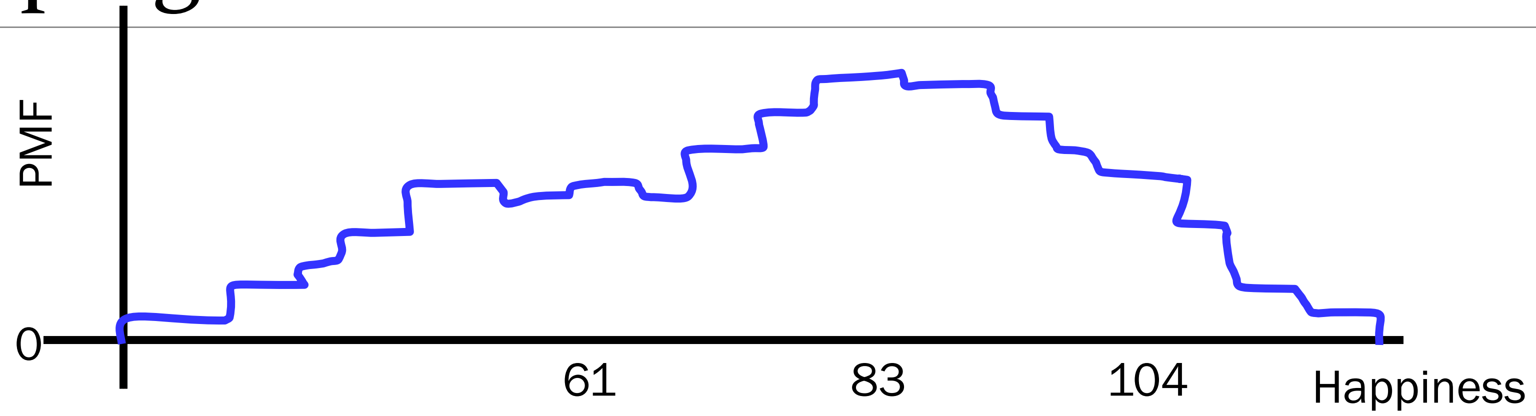


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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

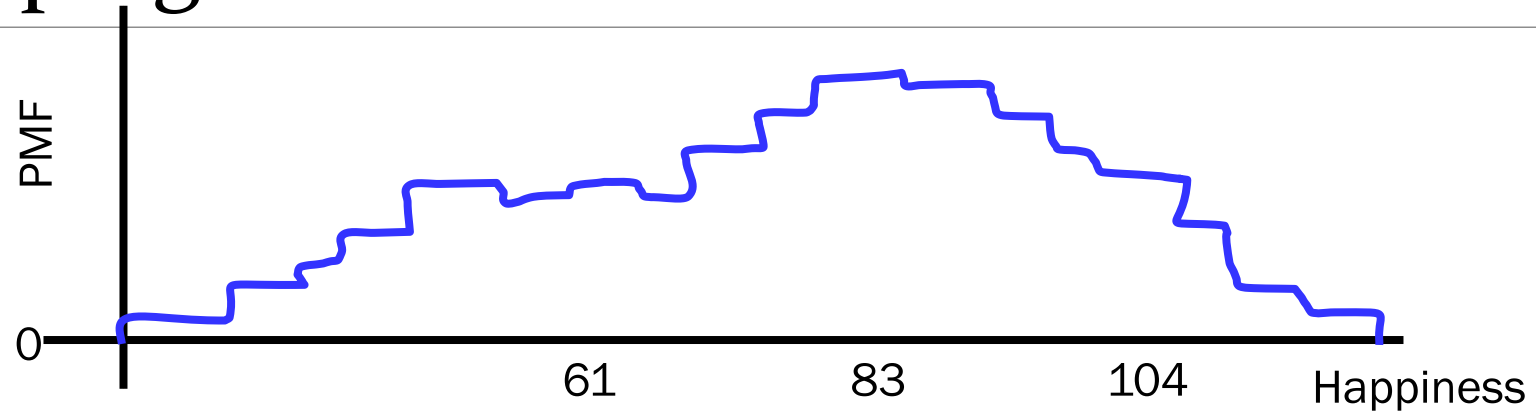


Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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



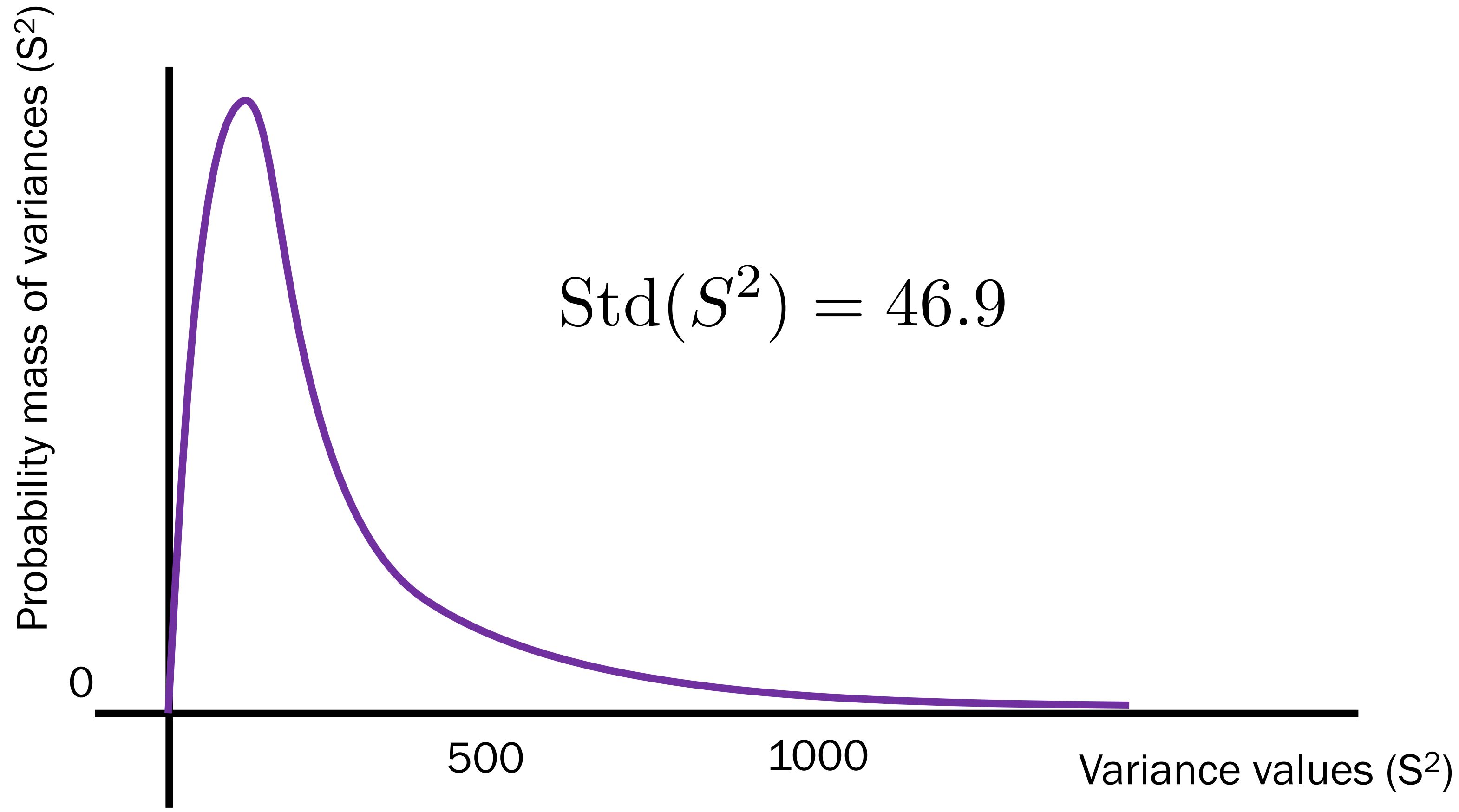
Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size**() 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]

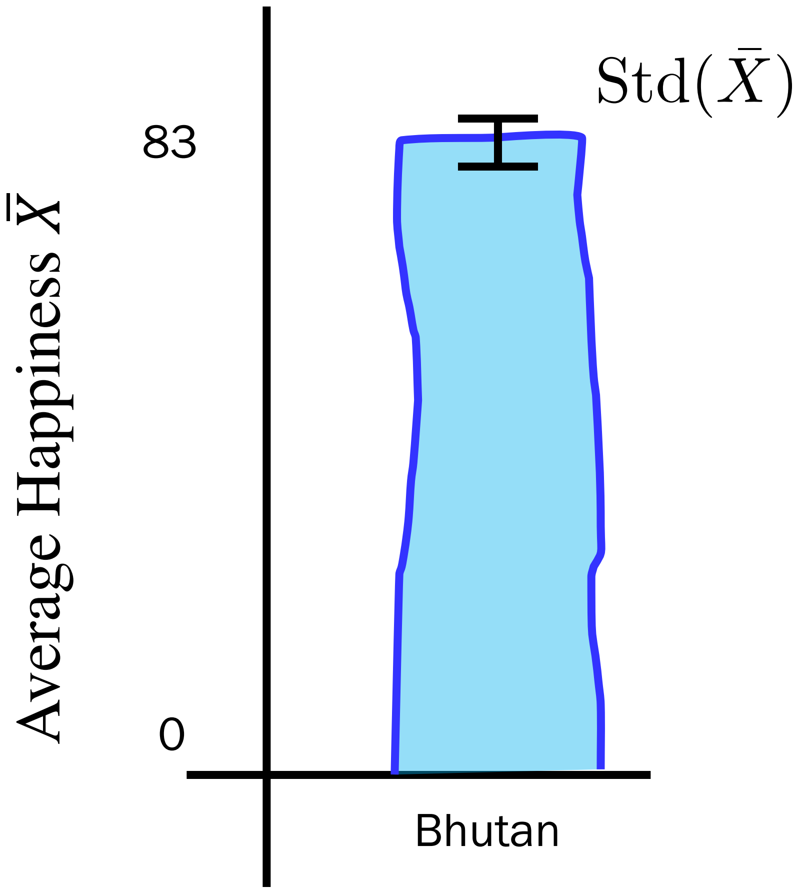
Bootstrapping of Variance

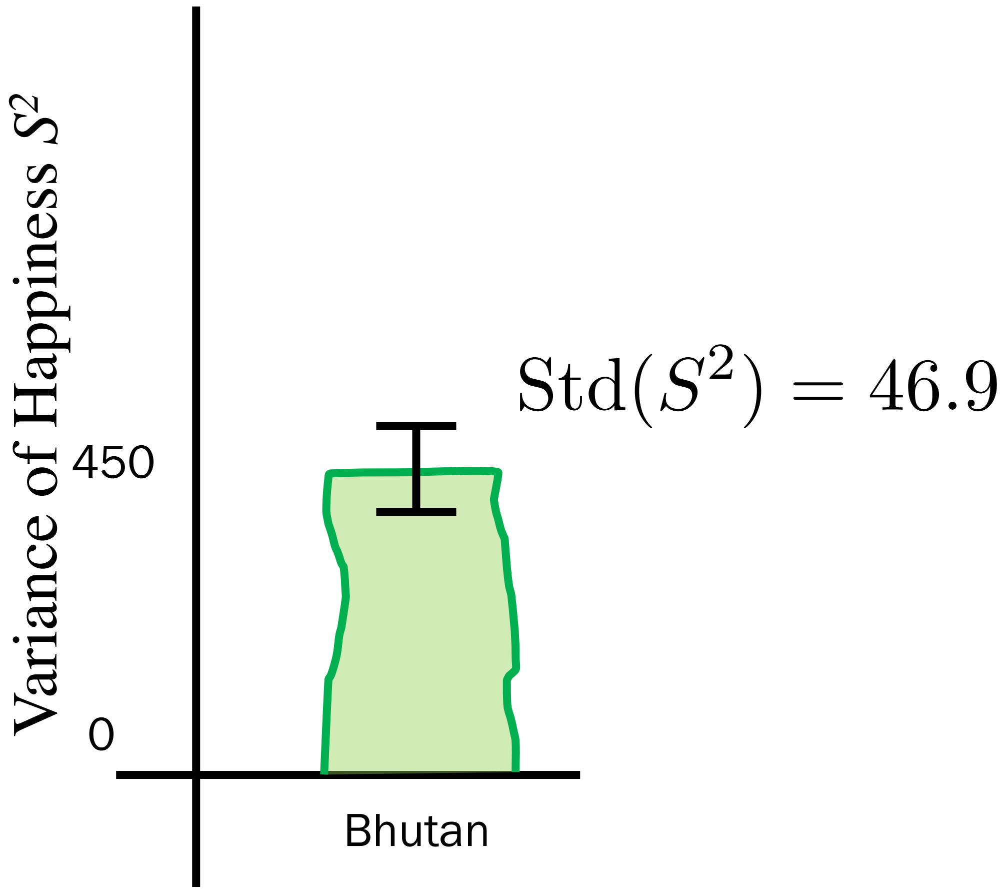
Sample Vars = [472.7, 478.4, 469.2, ..., 476.2]



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

Our Report to Bhutan Government

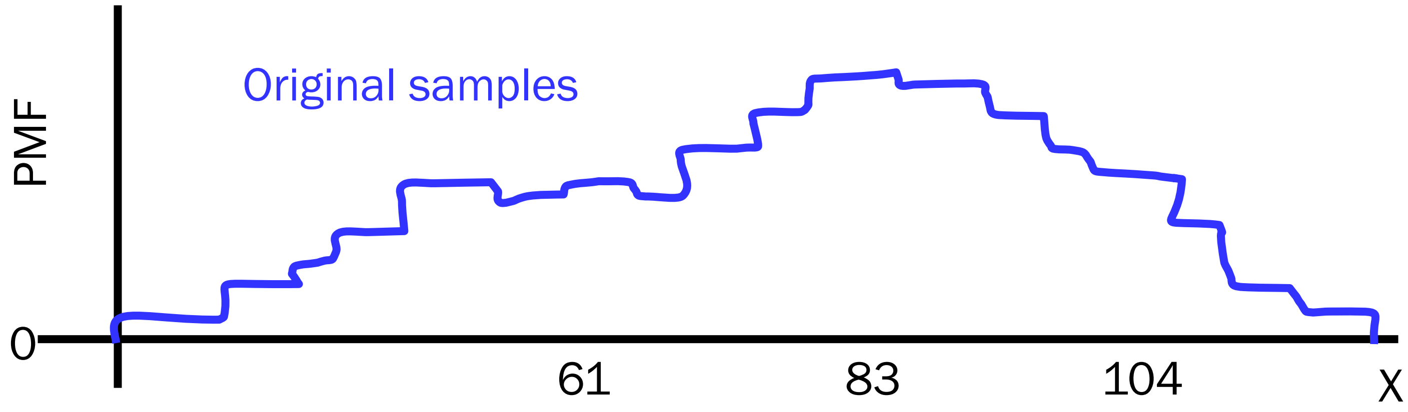




Claim: The average happiness of Bhutan is 83 ± 2

Bootstrapping in Practice

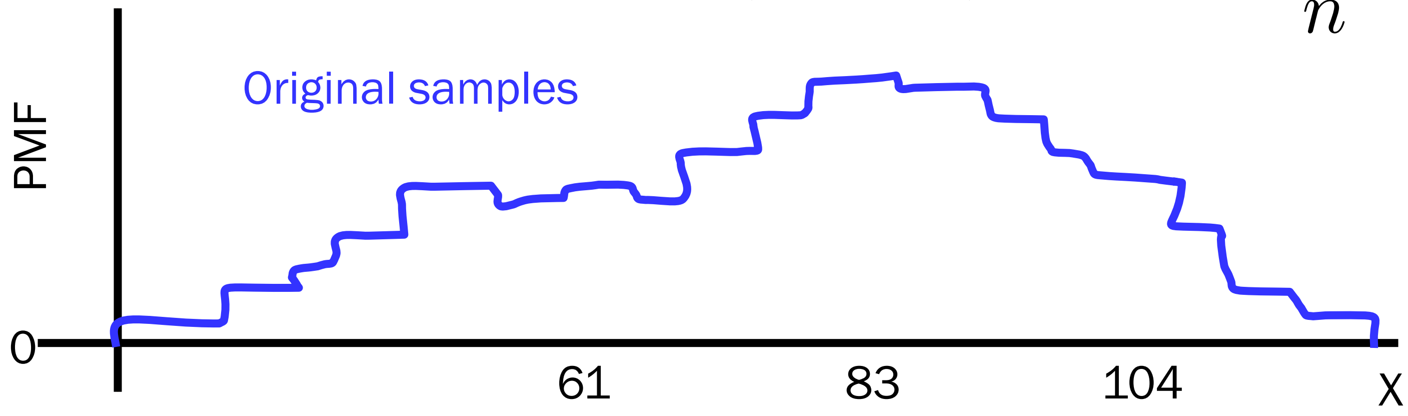
```
def resample(samples):
    # Estimate the PMF using the samples
    # Draw K new samples from the PMF
```



Bootstrapping in Practice

```
def resample(samples):
    # 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 sample.size() from PMF
    b.  Recalculate the stat on the resample
  3.   You now have a distribution of your stat
```

Bootstrapping in Practice

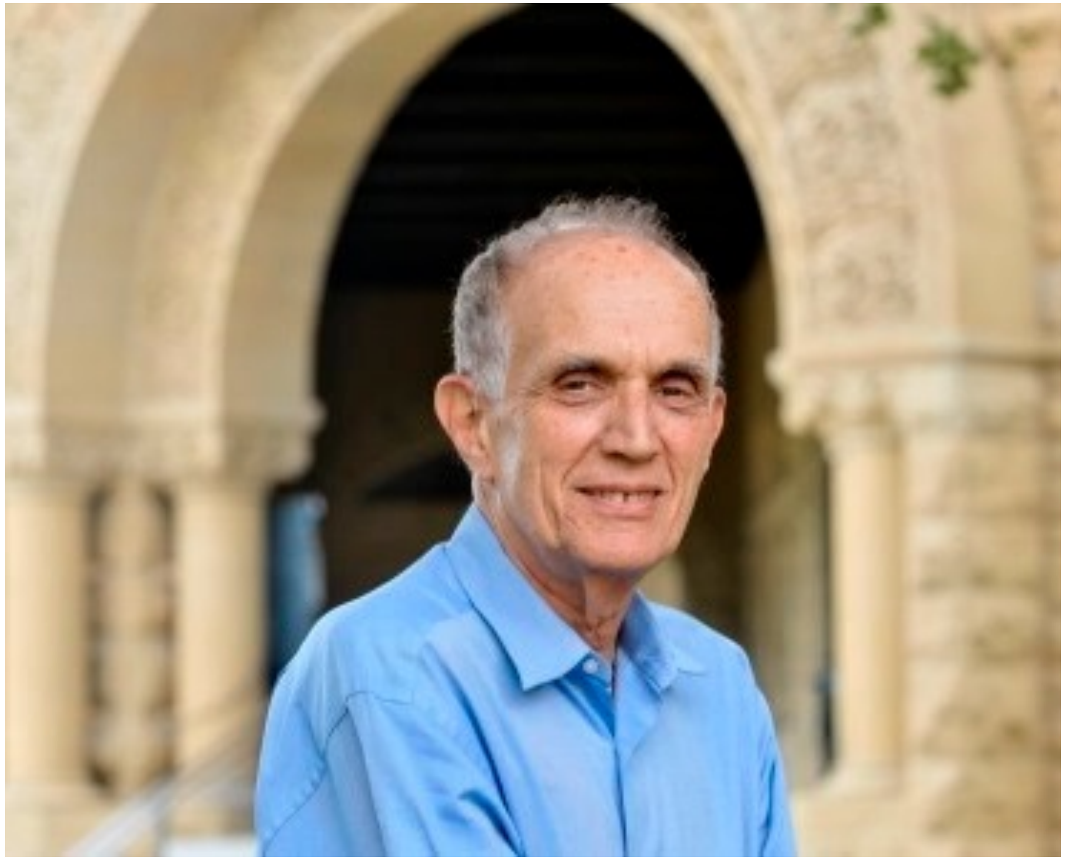
```
Bootstrap Algorithm (sample):
  1.   Repeat 10,000 times:
    a.  Choose sample.size elems from sample, with replacement
    b.  Recalculate the stat on the resample
  2.   You now have a distribution of your stat
```

To the code!

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

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
4.44
3.36
5.87
2.31
...
3.70

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

18 students (Context A); 23 students (Context B)

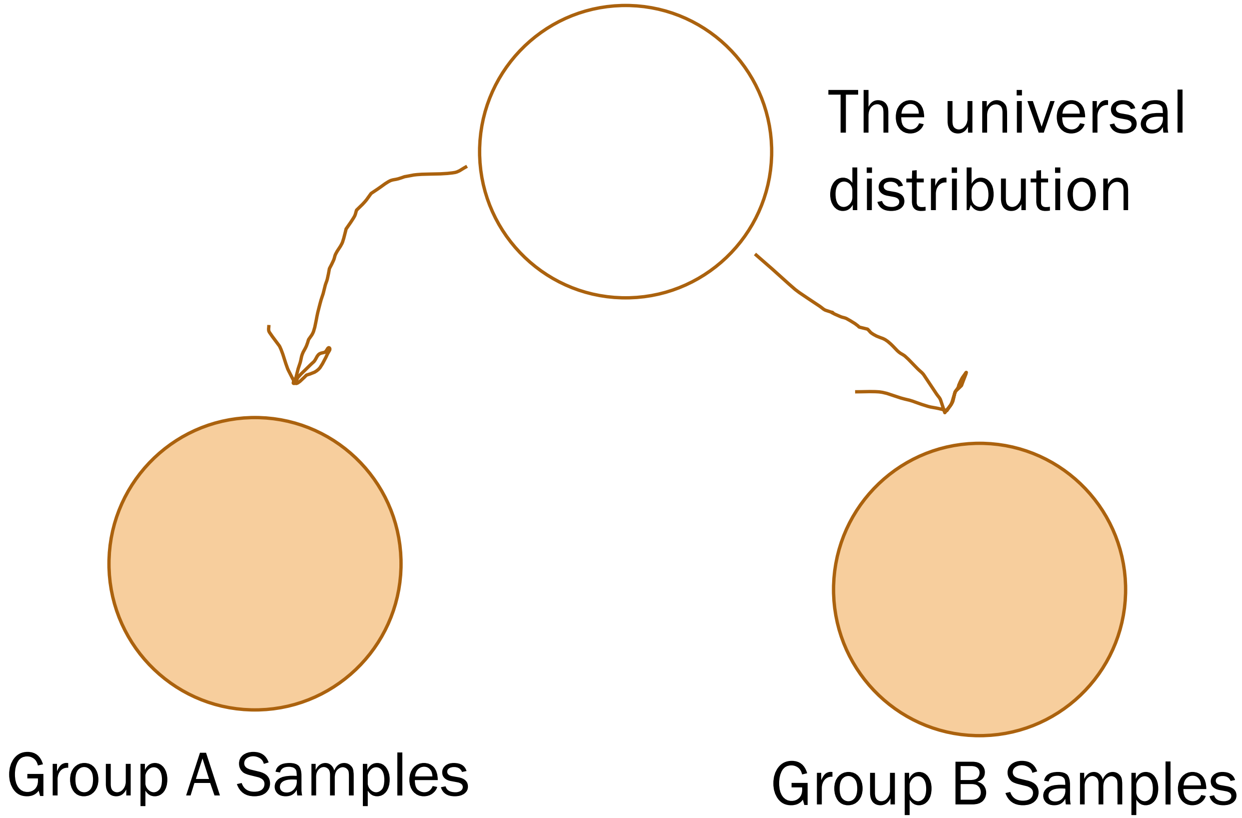
$\mu_1 = 3.1$ $\mu_2 = 2.4$

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 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.



To the code!