

Sampling and Bootstrapping

Chris Piech

CS109, Stanford University

<review>

IID Random Variables

- Consider n random variables X_1, X_2, \dots, X_n
 - X_i are all independently and identically distributed (I.I.D.)
 - All have the same PMF (if discrete) or PDF (if continuous)
 - All have the same expectation
 - All have the same variance

IID

iid

The sum of independent, identically distributed variables:

$$Y = \sum_{i=0}^n X_i$$

Is normally distributed:

$$Y \sim N(n\mu, n\sigma^2)$$

where $\mu = E[X_i]$

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





By the Central Limit Theorem, the sample mean of IID variables are distributed normally.

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



</review>

Motivating Example

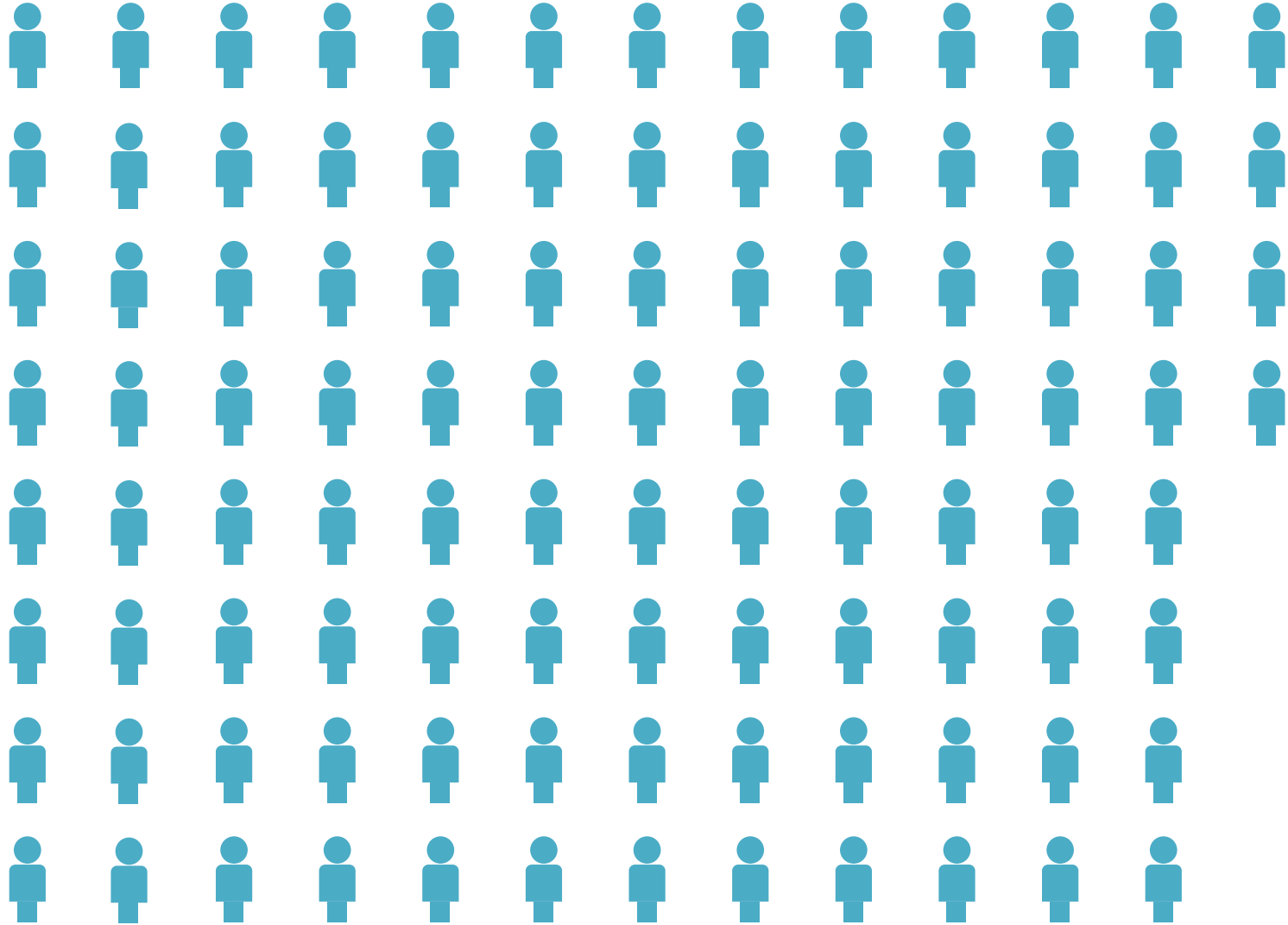
- You want to know the true mean and variance of happiness in Buthan
 - But you can't ask everyone.
 - Randomly sample 200 people.
 - Your data looks like this:



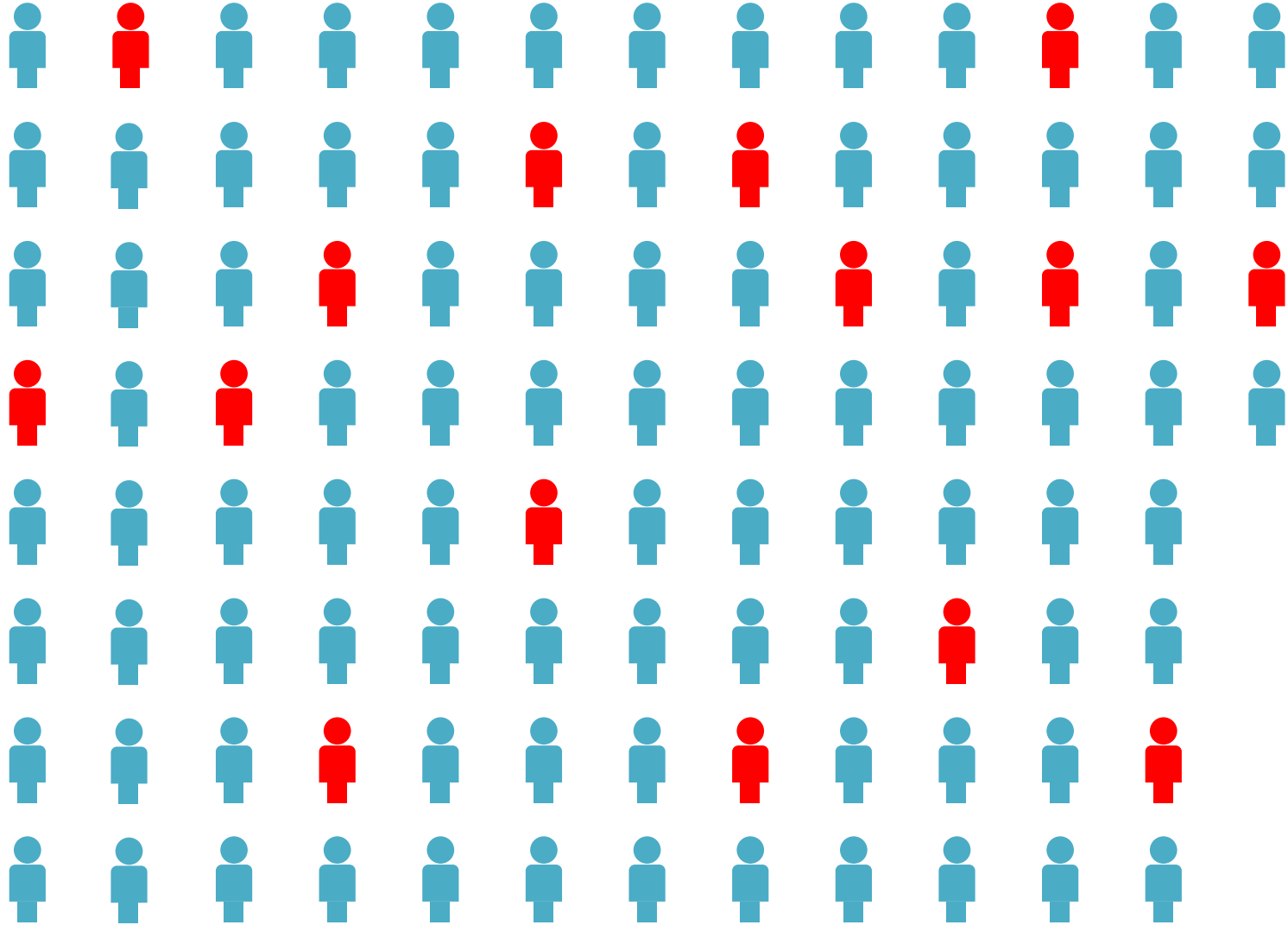
Happiness = {72, 85, 79, 91, 68, ... , 71}

- The mean of all of those numbers is 83. Is that the true average happiness of Bhutanese people?

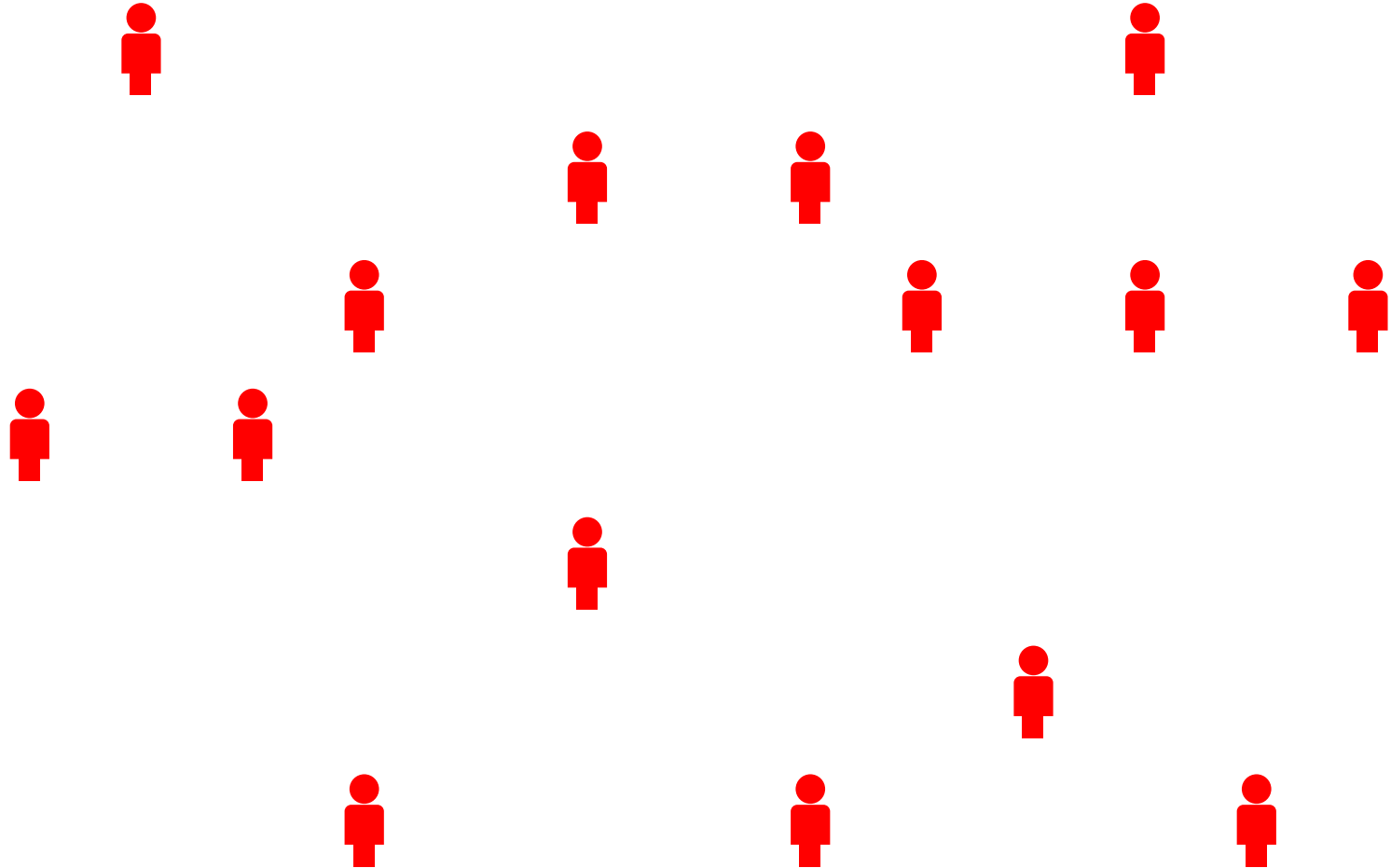
Population



Sample



Sample



Collect one (or more) numbers from each person

Sample

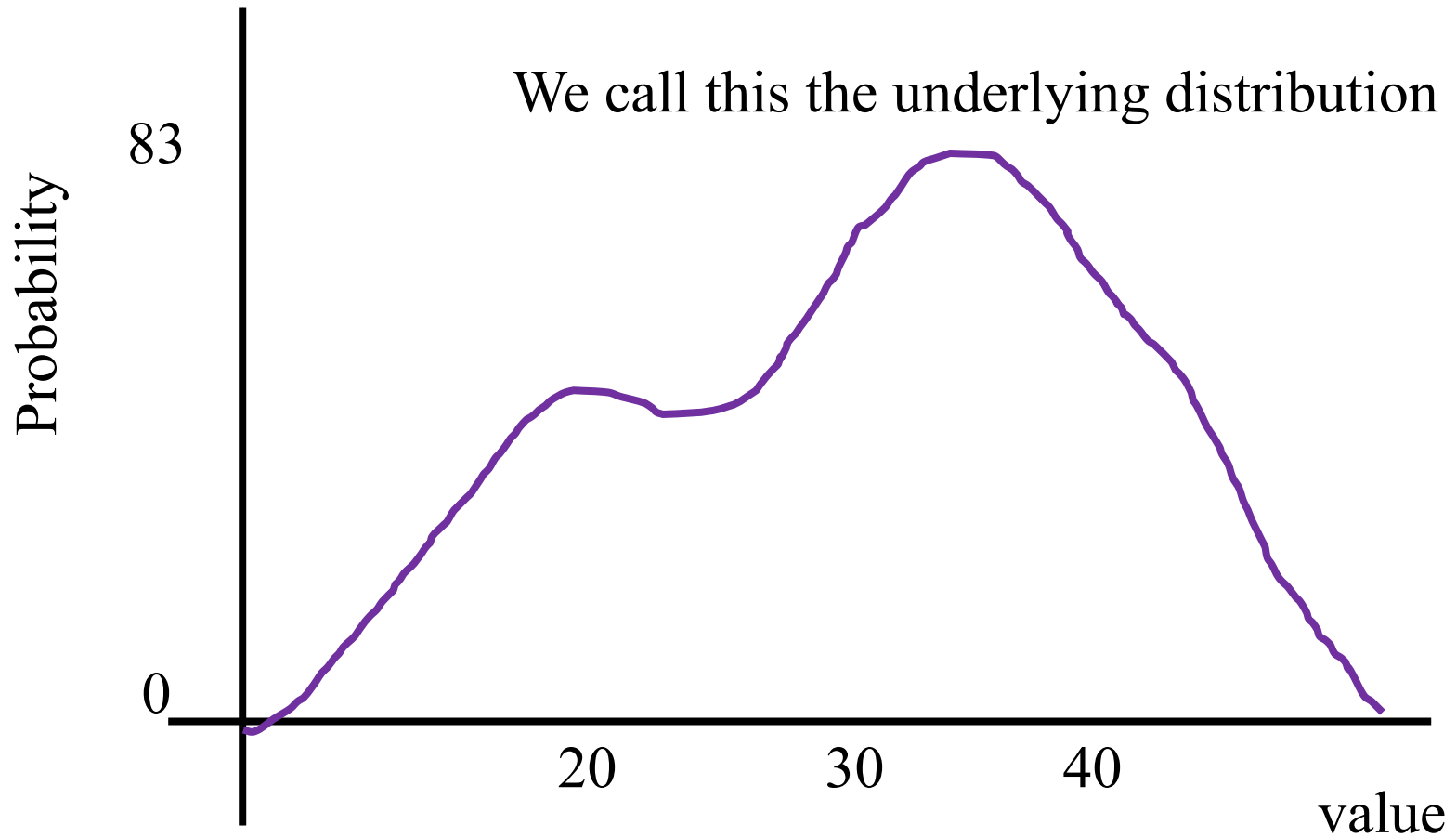


IID Samples

Consider n IID samples:

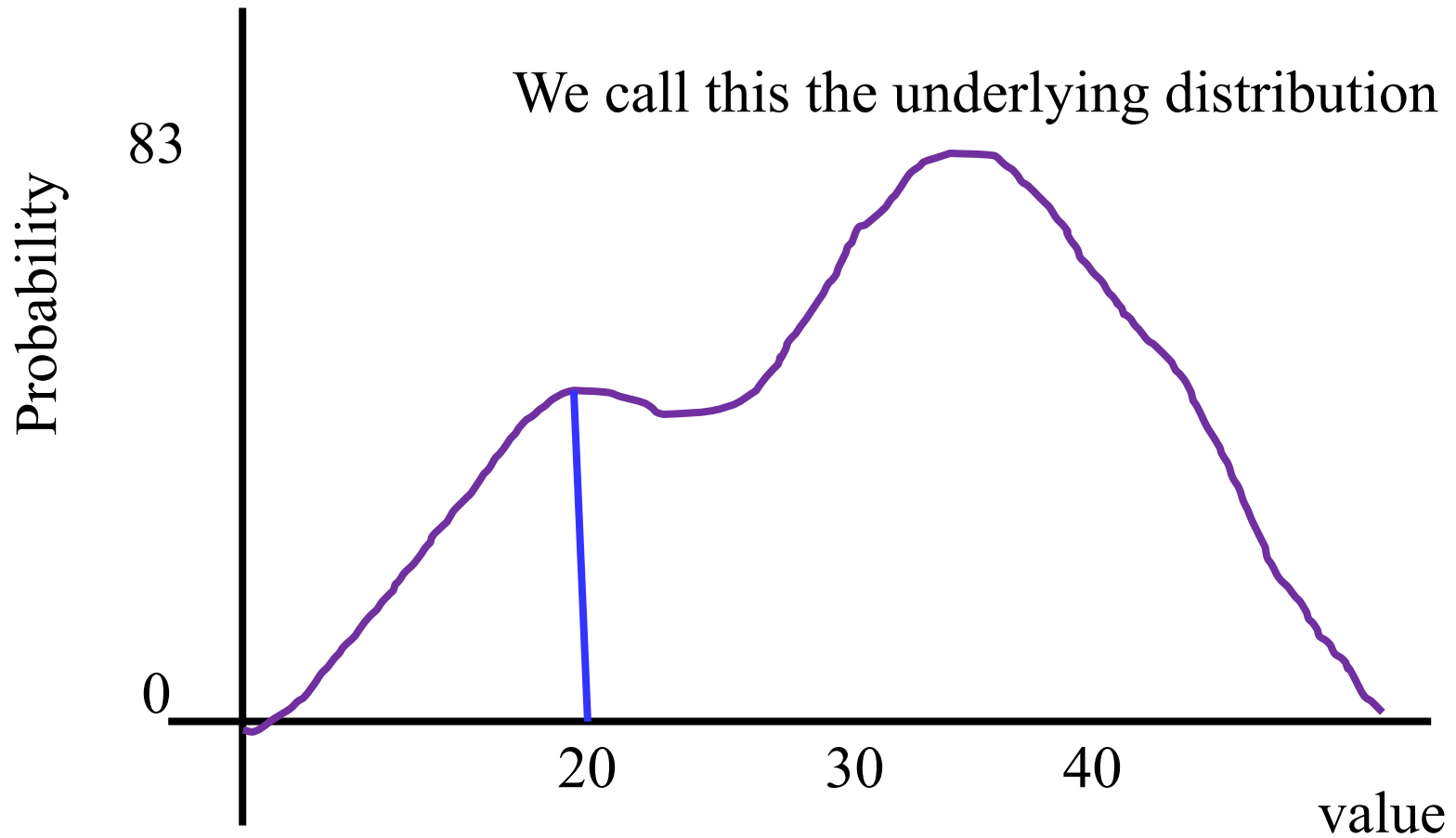
$$X_1, X_2, \dots, X_n$$

IID Samples



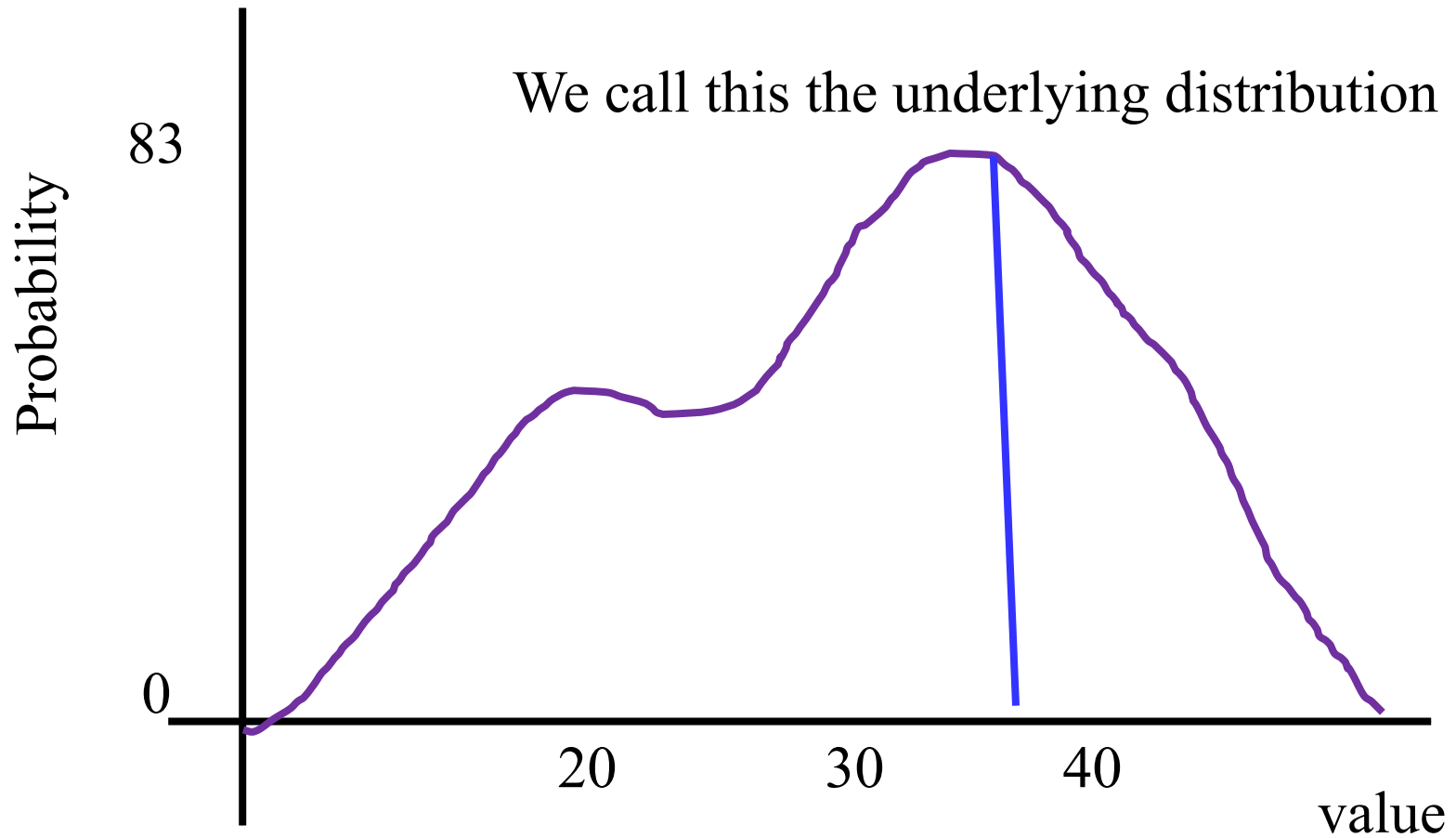
IID Samples = []

IID Samples



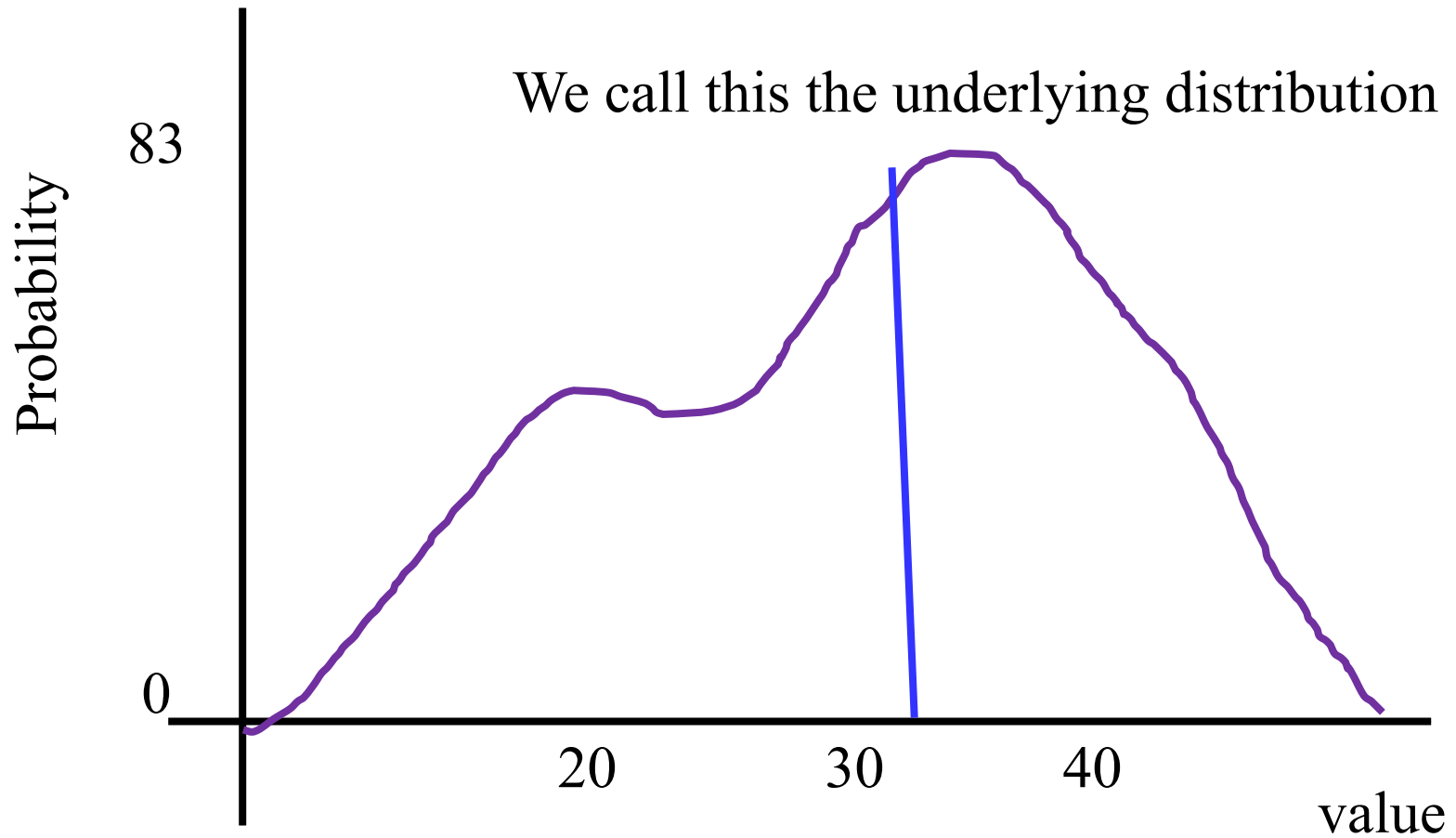
IID Samples = [20]

IID Samples



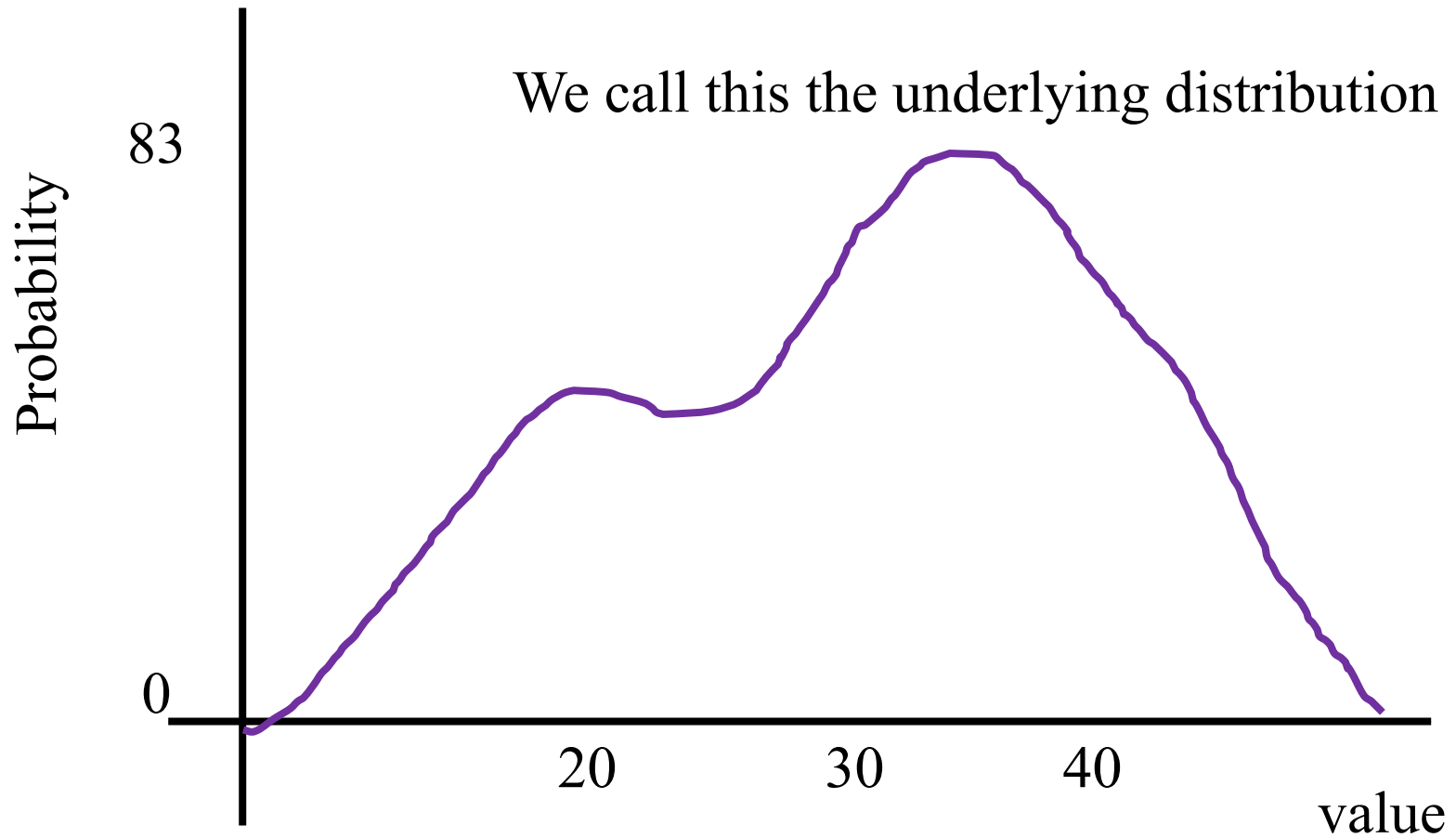
IID Samples = [20, 38]

IID Samples



IID Samples = [20, 38, 32]

IID Samples



IID Samples = [20, 38, 32, ..., 38]

X_1

X_2

X_n

Sample Mean

- Consider n random variables X_1, X_2, \dots, X_n
 - X_i are all independently and identically distributed (I.I.D.)
 - Have same distribution function F and $E[X_i] = \mu$
 - We call sequence of X_i a **sample** from distribution F

- Sample mean: $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$

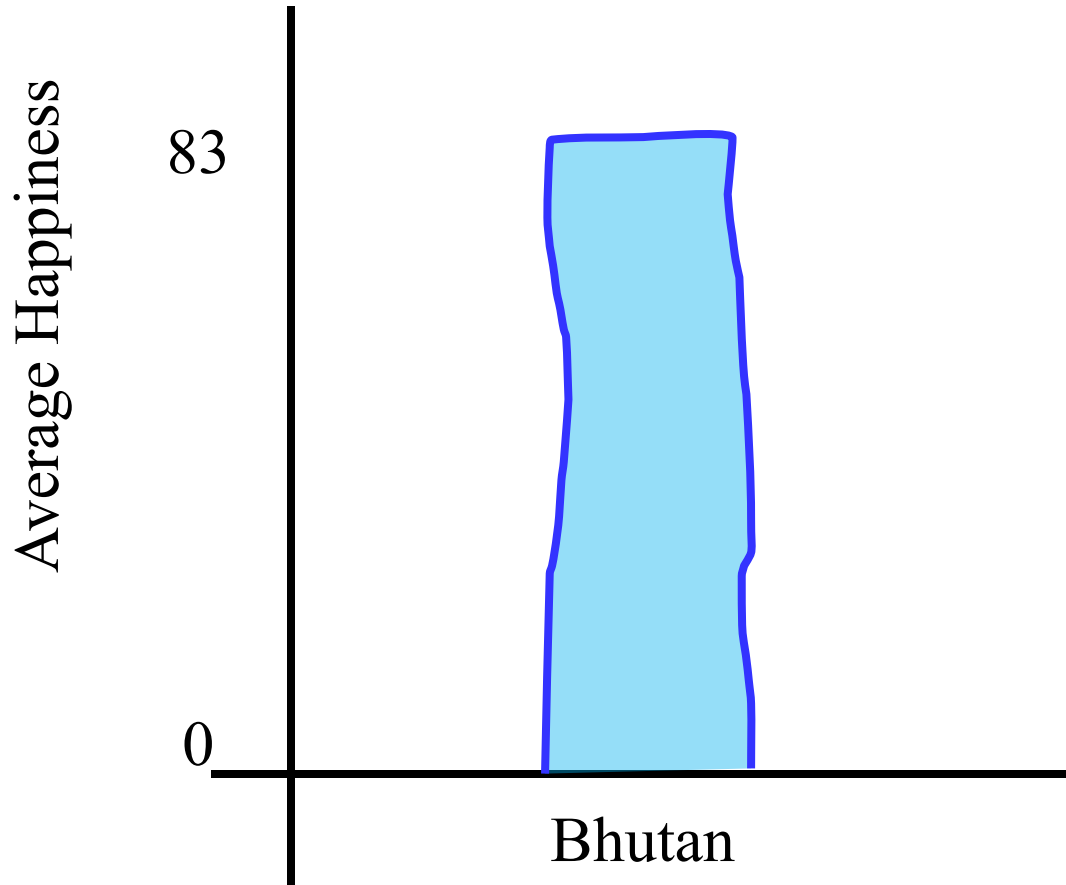
- Compute $E[\bar{X}]$

$$\begin{aligned} E[\bar{X}] &= E\left[\sum_{i=1}^n \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu \end{aligned}$$

- \bar{X} is “unbiased” estimate of μ ($E[\bar{X}] = \mu$)

Sample Mean

Average Happiness





Sample Mean:

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

ith sample

Size of the sample



Sample Variance

- Consider n I.I.D. random variables X_1, X_2, \dots, X_n
 - X_i have distribution F with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$
 - We call sequence of X_i a **sample** from distribution F
 - Recall sample mean: $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
 - Sample deviation: $\bar{X} - X_i$ for $i = 1, 2, \dots, n$
 - Sample variance: $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$
 - What is $E[S^2]$?
 - $E[S^2] = \sigma^2$
 - We say S^2 is “unbiased estimate” of σ^2

I Believe What I See

Intuition that $E[S^2] = \sigma^2$

Population
variance

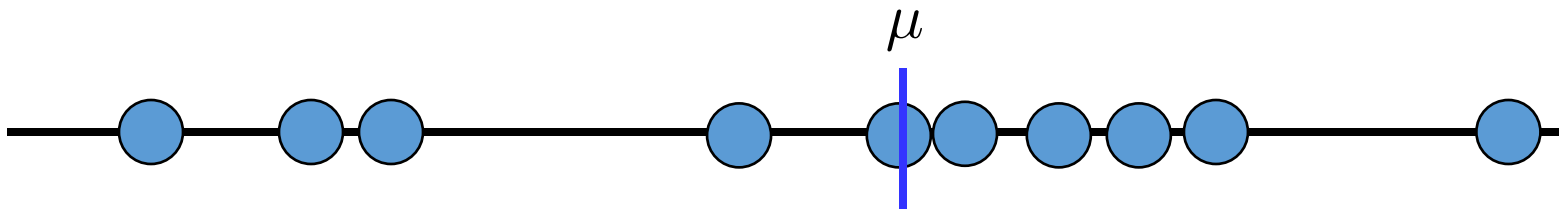
$$\sigma^2 = \sum_{i=1}^N \frac{(X_i - \mu)^2}{N}$$

This is the
actual mean

Unbiased
sample variance

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

This is the
sample mean



Intuition that $E[S^2] = \sigma^2$

Population variance

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

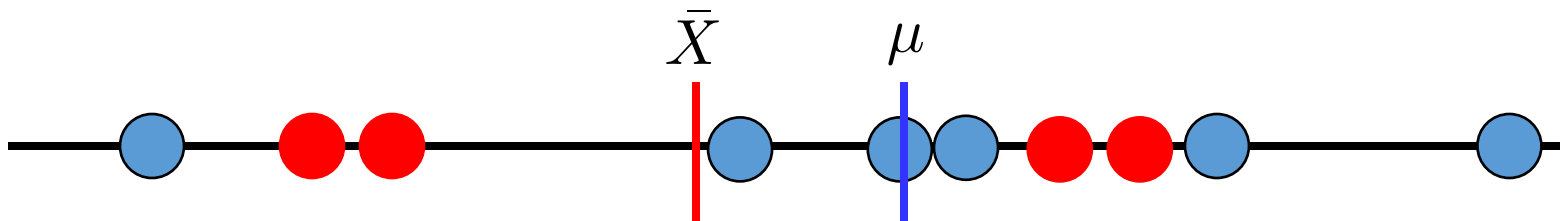
This is the actual mean

Unbiased sample variance

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

This is the sample mean

The variance of the sample mean? Related to population variance



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

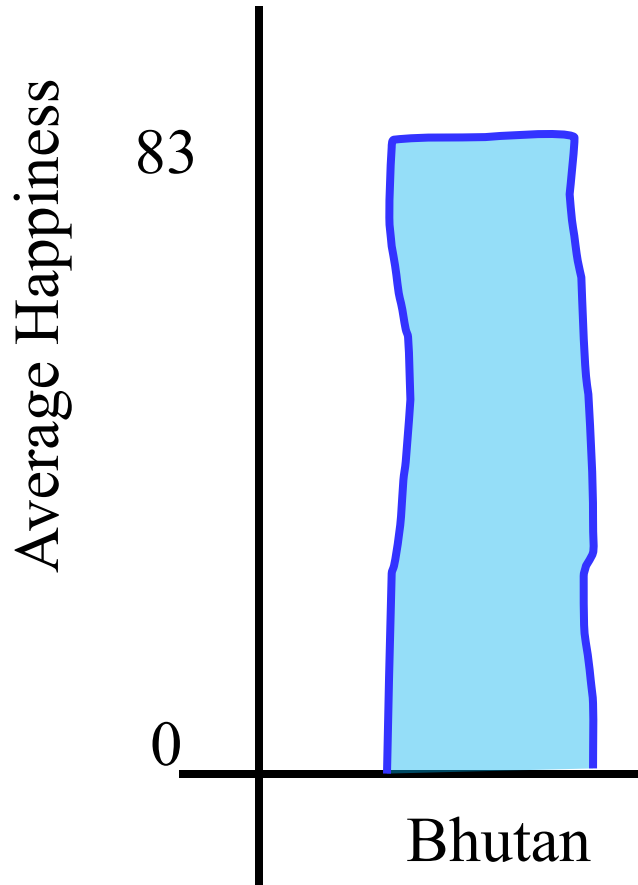
$$E[S^2] = E\left[\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}\right] \Rightarrow (n-1)E[S^2] = E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]$$

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^n ((X_i - \mu) + (\mu - \bar{X}))^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\mu - \bar{X})^2 + 2\sum_{i=1}^n (X_i - \mu)(\mu - \bar{X})\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})\sum_{i=1}^n (X_i - \mu)\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})n(\bar{X} - \mu)\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\mu - \bar{X})^2\right] = \sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\mu - \bar{X})^2] \\ &= n\sigma^2 - n\text{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2\end{aligned}$$

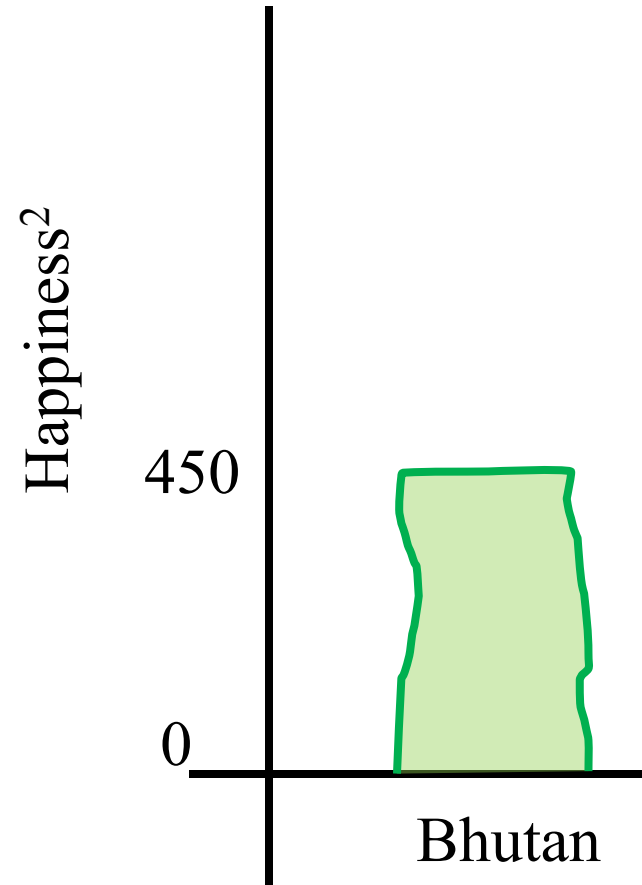
- So, $E[S^2] = \sigma^2$

Sample Mean

Average Happiness



Variance of Happiness





Sample Variance:

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

Sample mean

Makes it "unbiased"



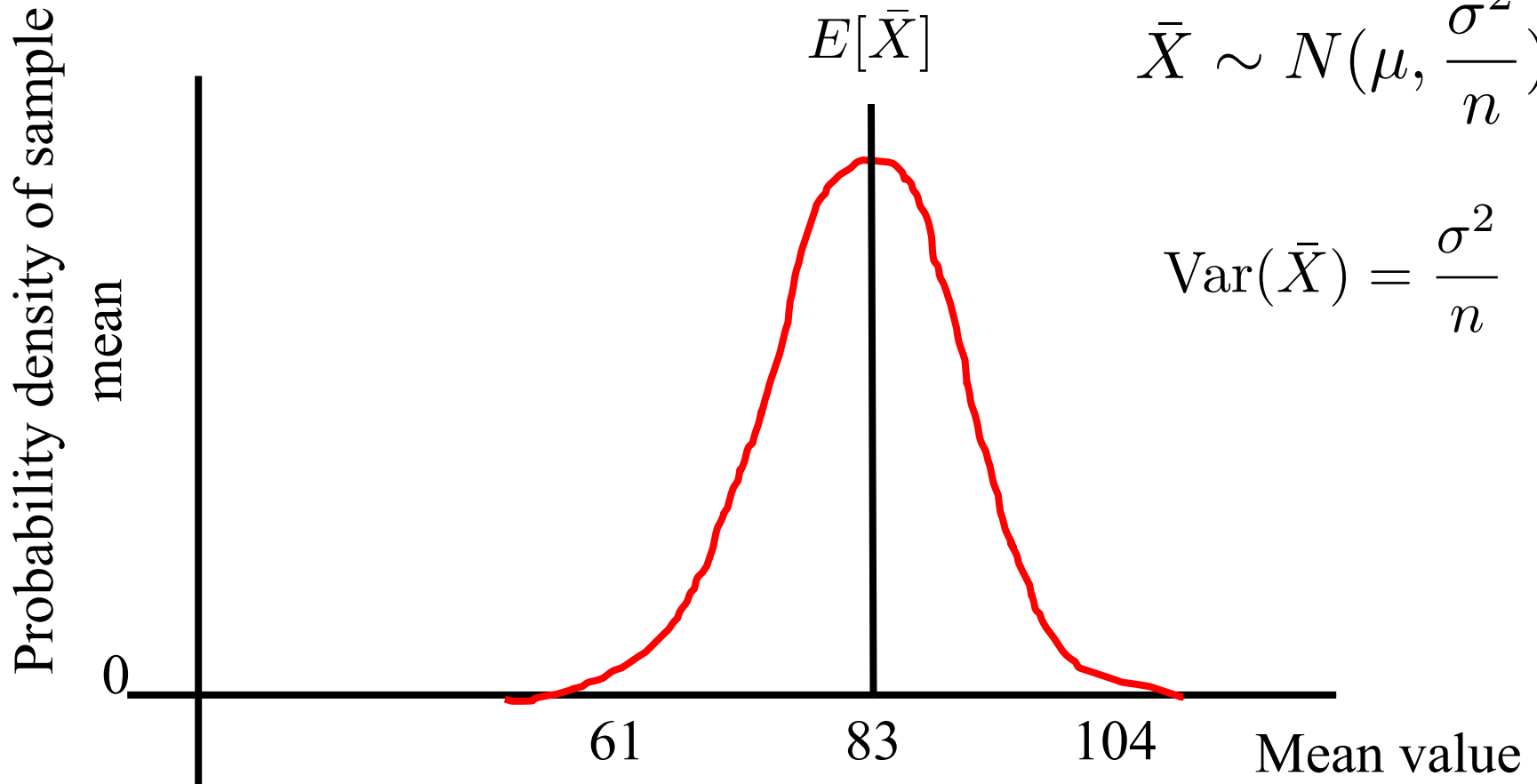
No Error Bars ☹️

Variance of Sample Mean

By central limit theorem:

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

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



Variance of Sample Mean

- Consider n **I.I.D.** random variables X_1, X_2, \dots, X_n
 - X_i have distribution F with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$
 - We call sequence of X_i a **sample** from distribution F
 - Recall sample mean: $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
 - What is $\text{Var}(\bar{X})$?

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\sum_{i=1}^n \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \text{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \sigma^2 = \left(\frac{1}{n}\right)^2 n \sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Standard Error of the Mean

$$\text{Var}(\bar{X}) = \text{Var}\left(\sum_{i=1}^n \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{\sigma^2}{n}$$

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

$$= \frac{S^2}{n}$$

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

$$= \sqrt{\frac{450}{200}}$$

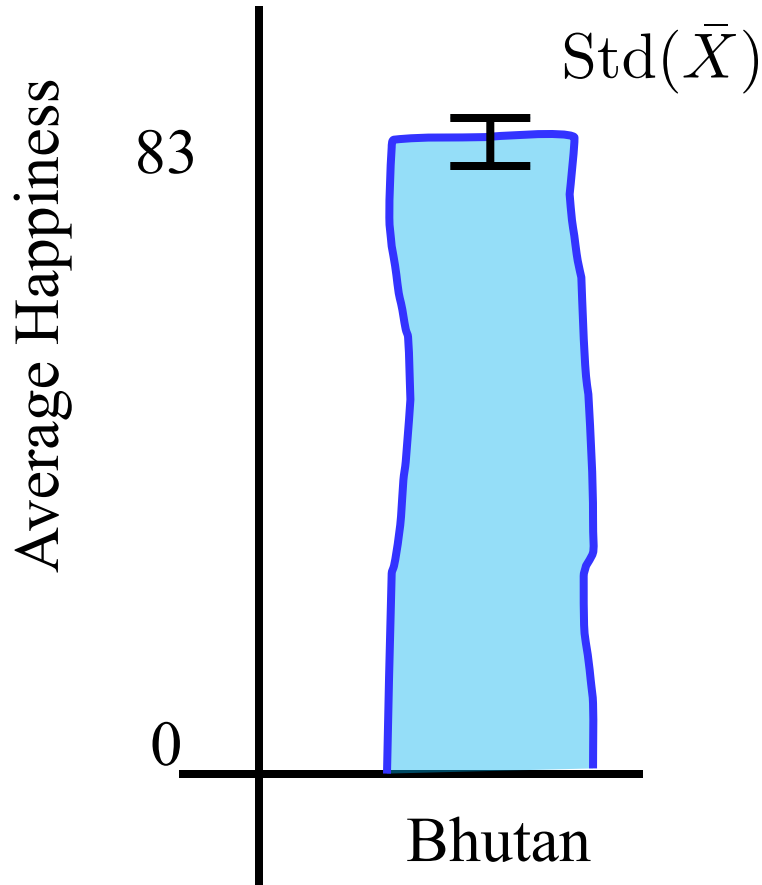
The numbers for our Bhutanesse poll

$$= 1.5$$

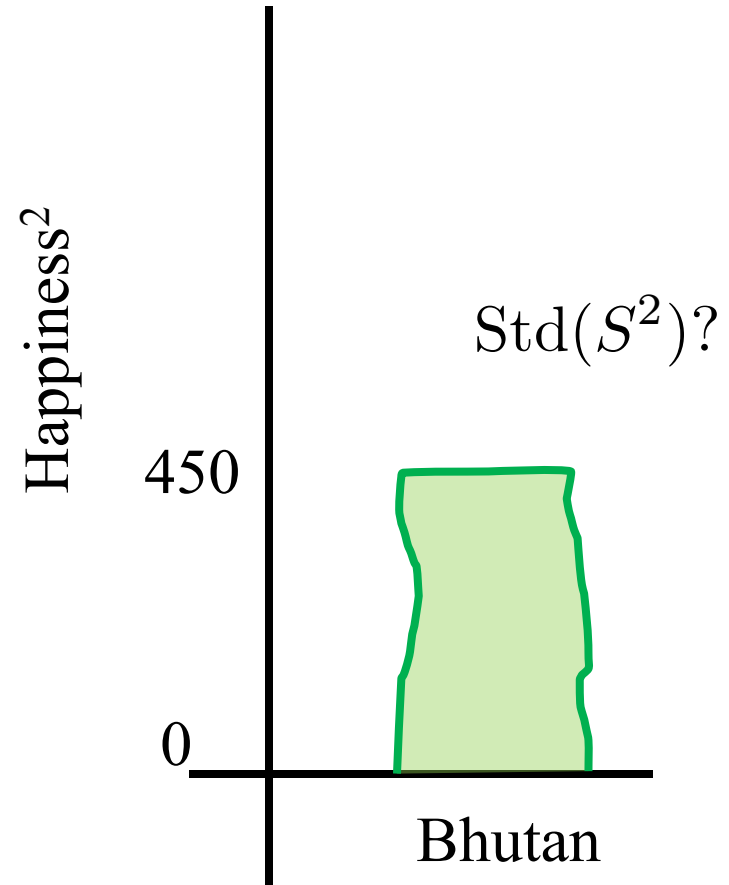
Bhutanesse standard error of the mean

Sample Mean

Average Happiness



Variance of Happiness



Claim: The average happiness of Bhutan is 83 ± 2

Bootstrap: Probability for Computer Scientists

Bootstrapping allows you to:

- Know the distribution of statistics
- Calculate p values

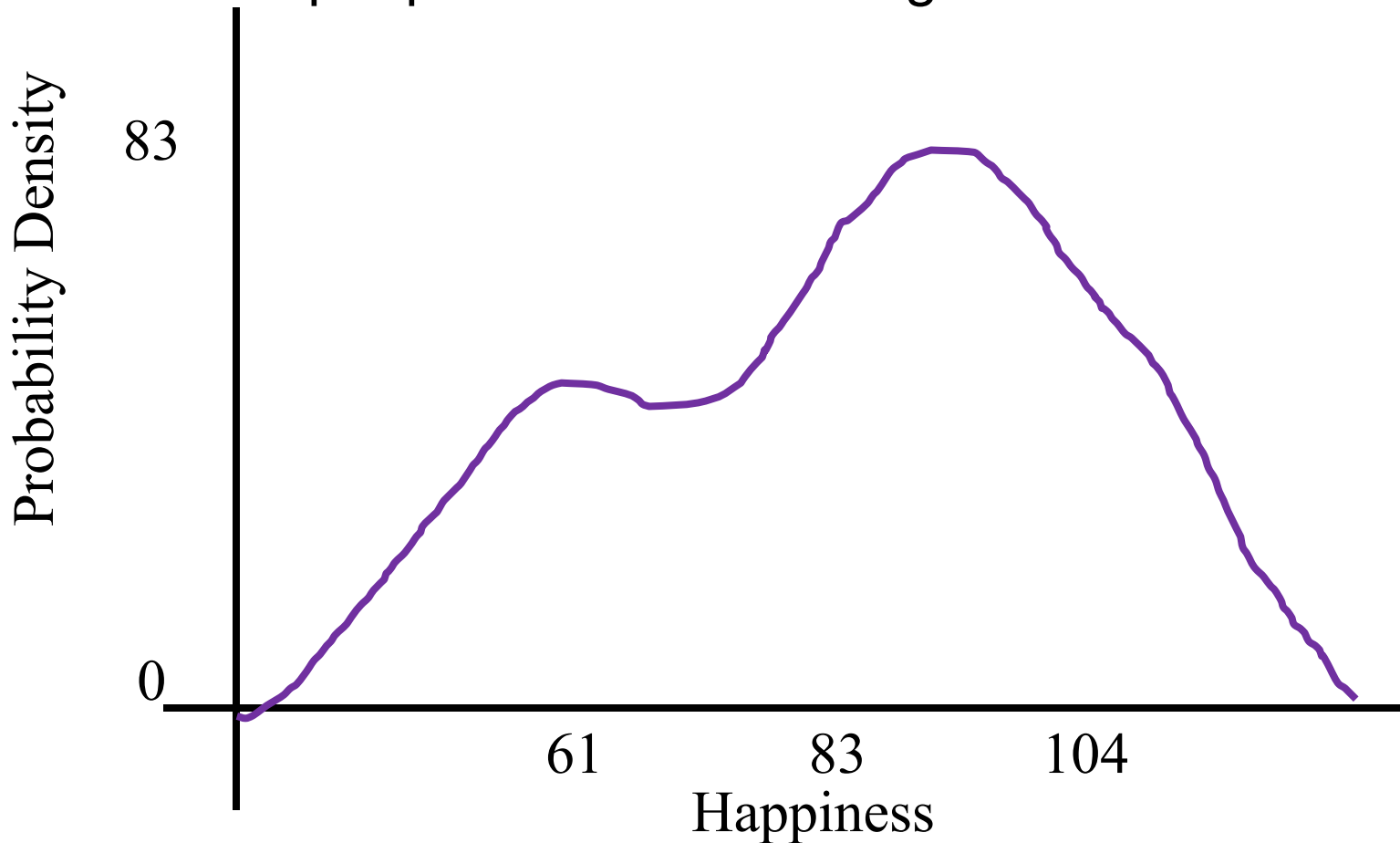
Hypothetical

What is the probability that a Bhutanese peep is just straight up loving life?



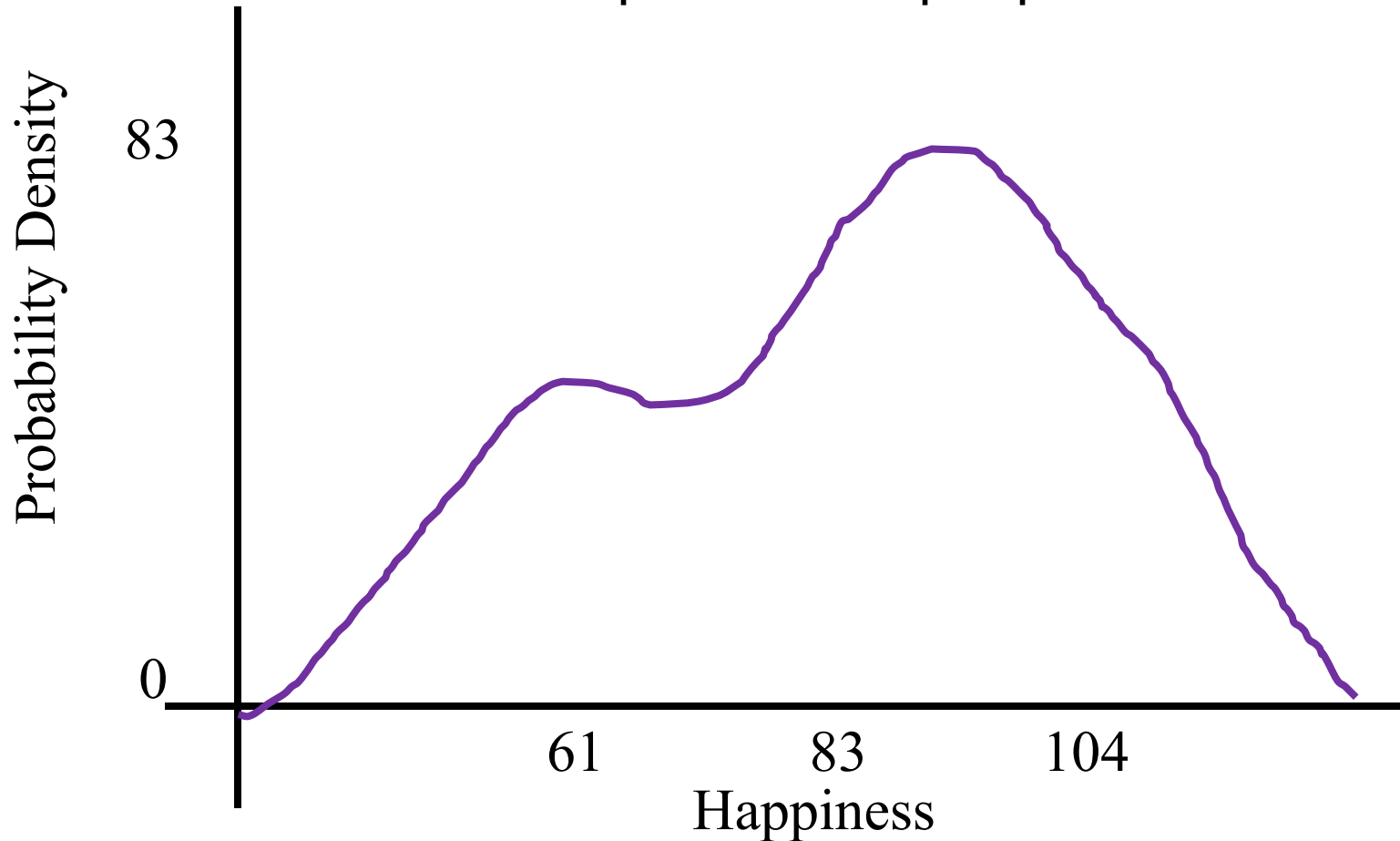
Hypothetical

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



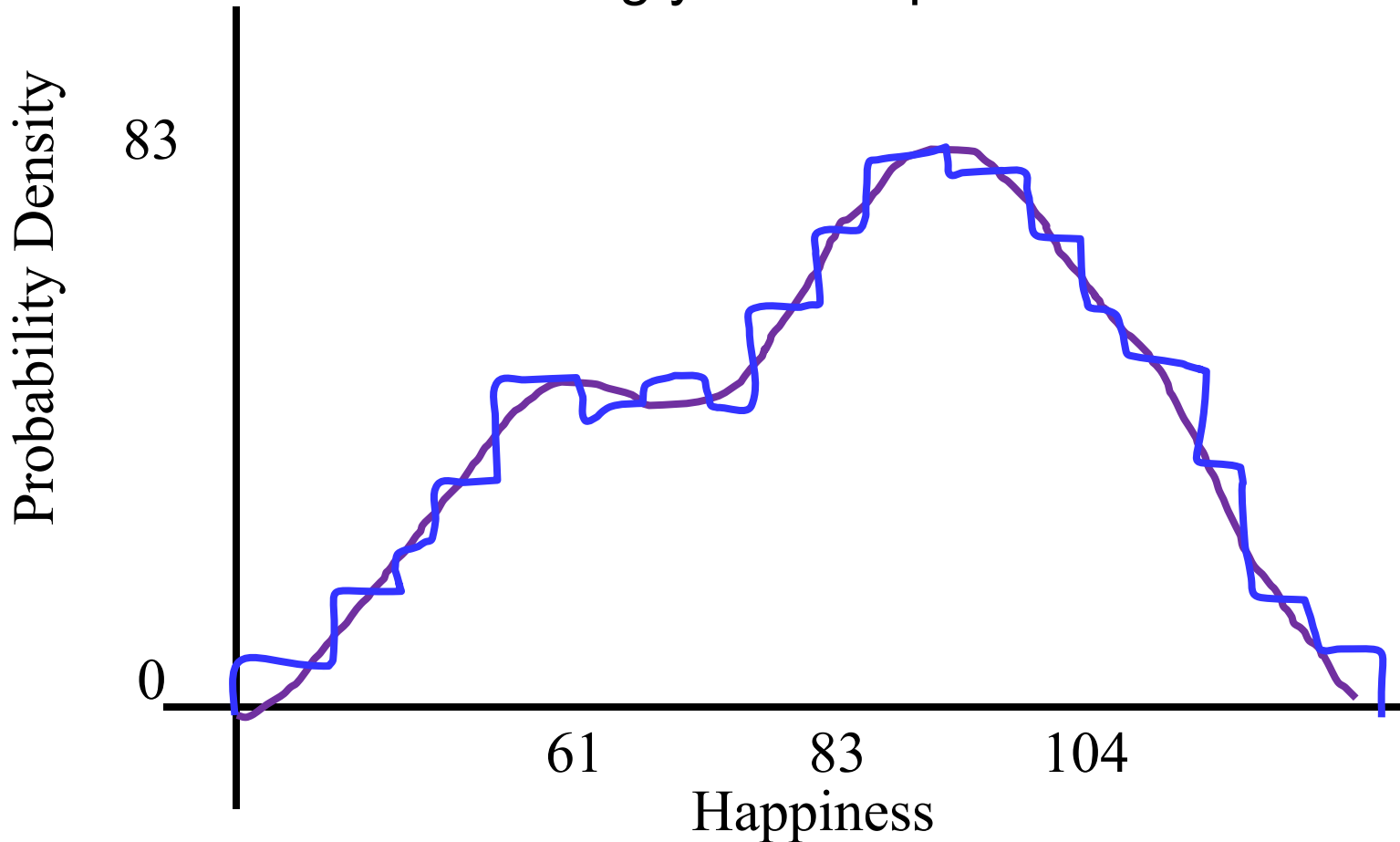
Hypothetical

What is the variance of the sample variance of subsamples of 200 people?



Key Insight

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



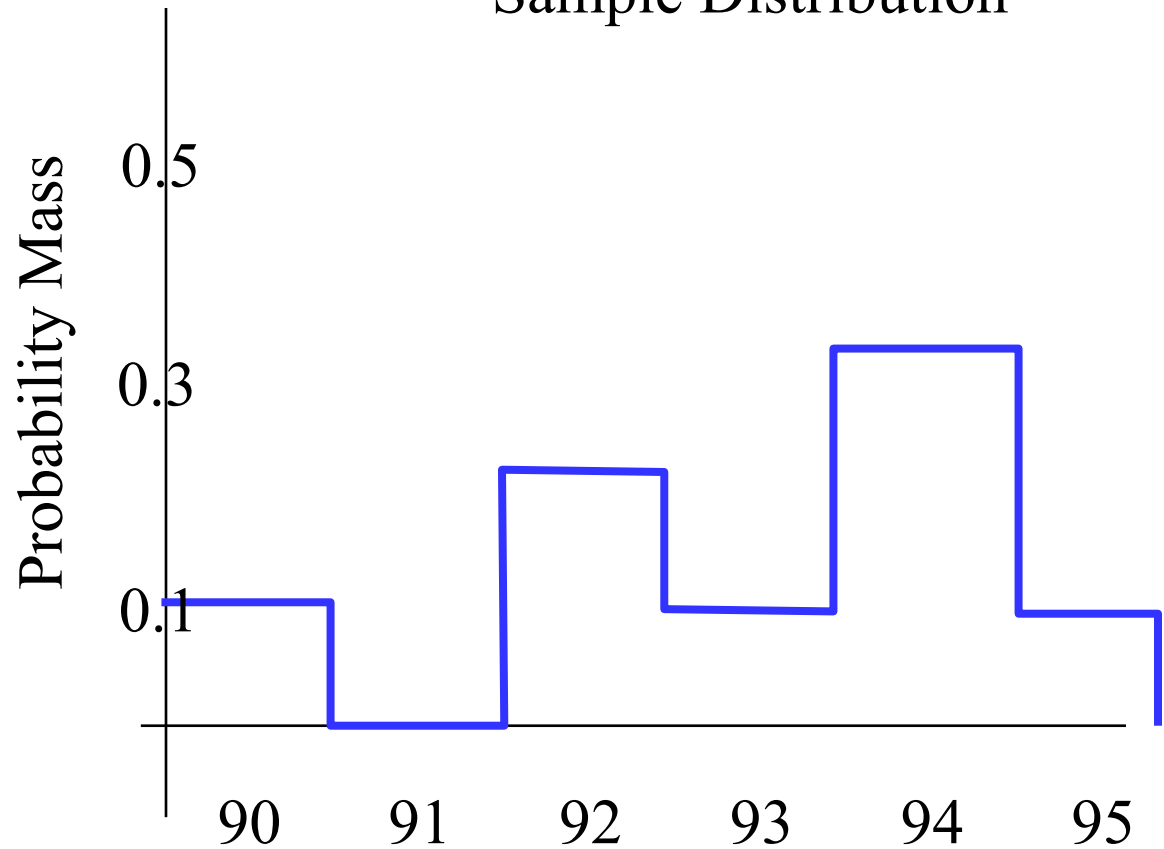
* This is just a histogram of your data!!

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):

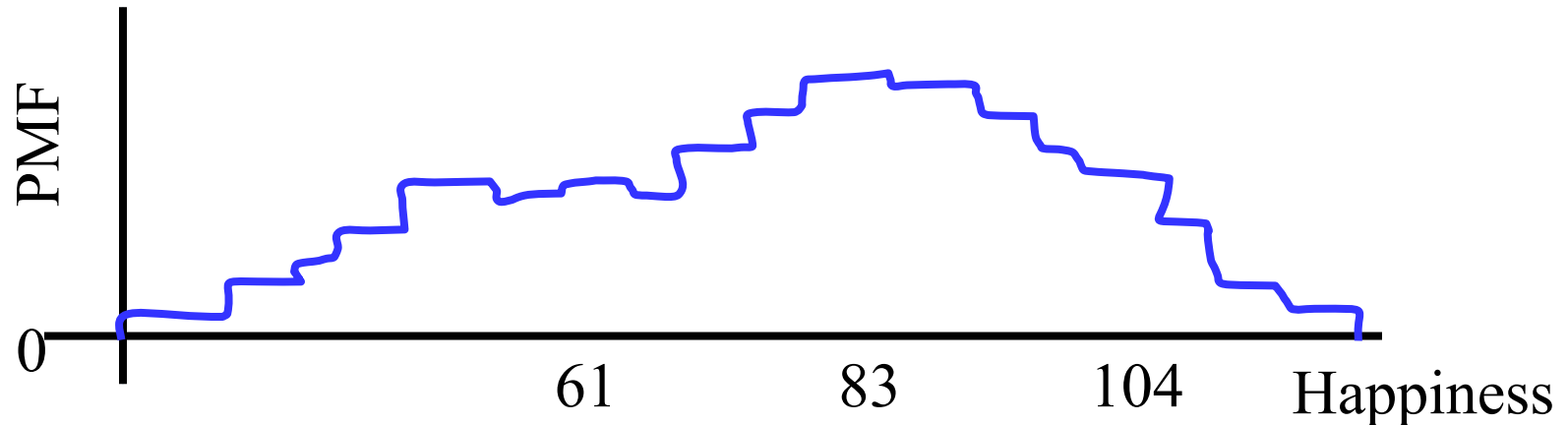
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
 - a. Resample **sample.size()** from PMF
 - b. Recalculate the stat** on the resample
3. You now have a **distribution of your stat**

Bootstrap 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**

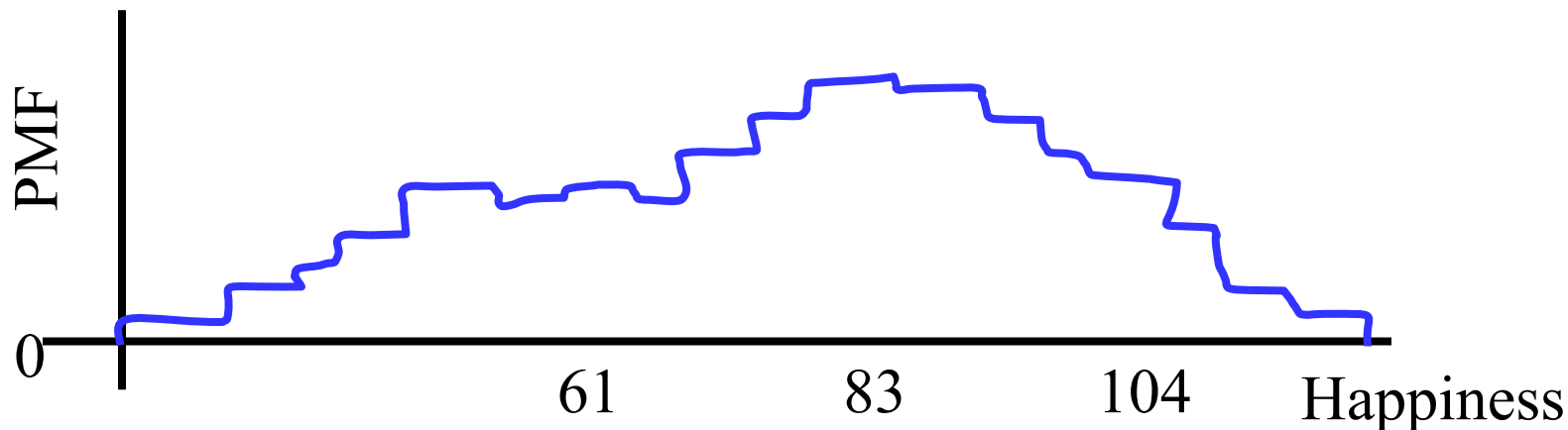
Bootstrap of Means



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**

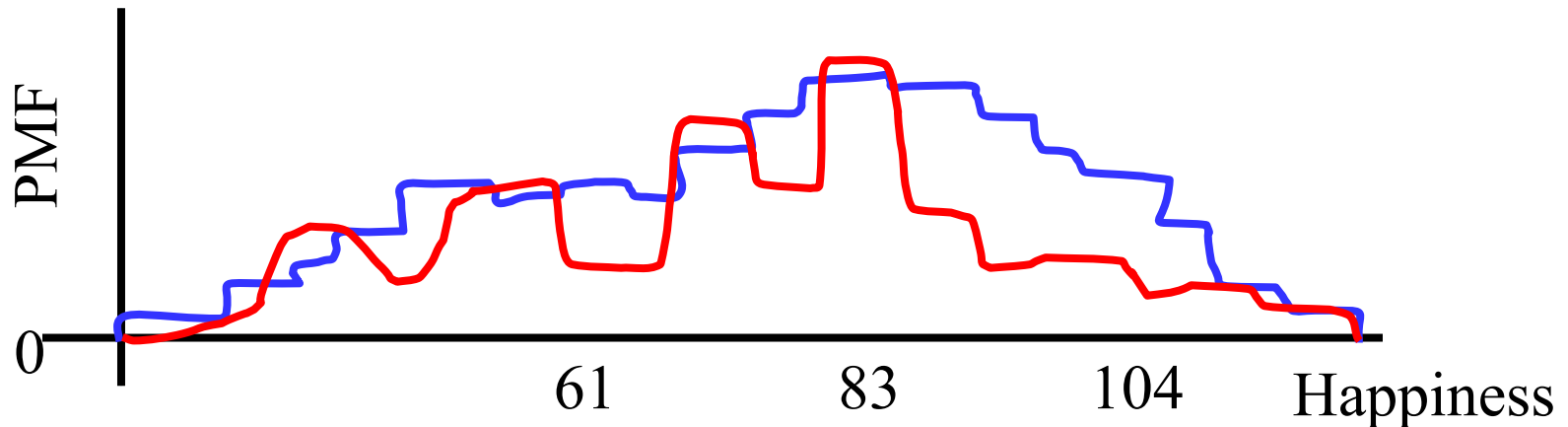
Bootstrap of Means



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**

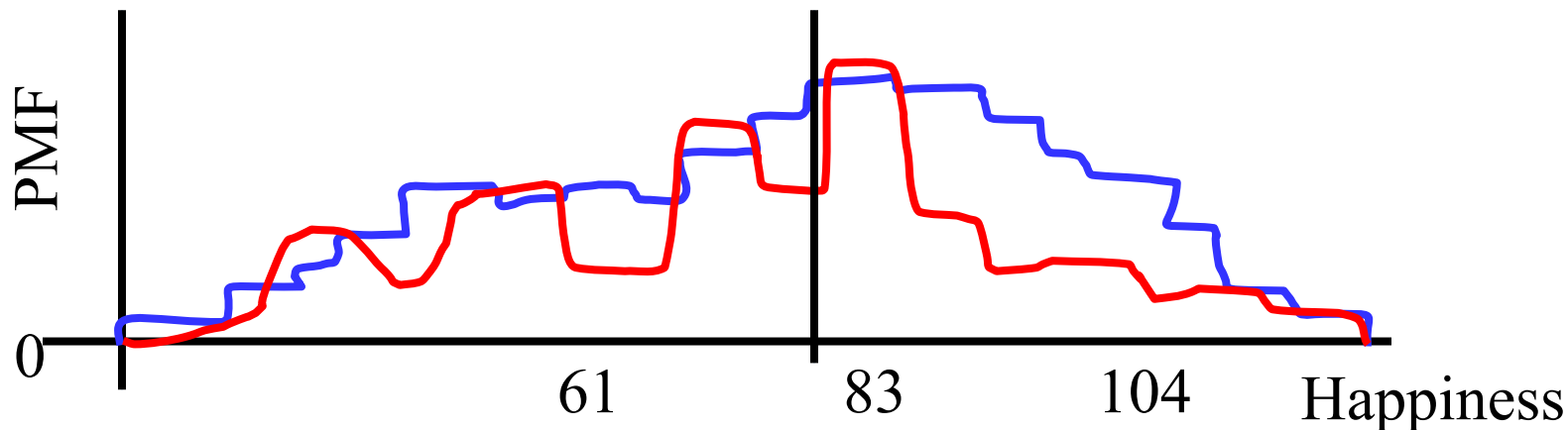
Bootstrap of Means



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**

Bootstrap of Means

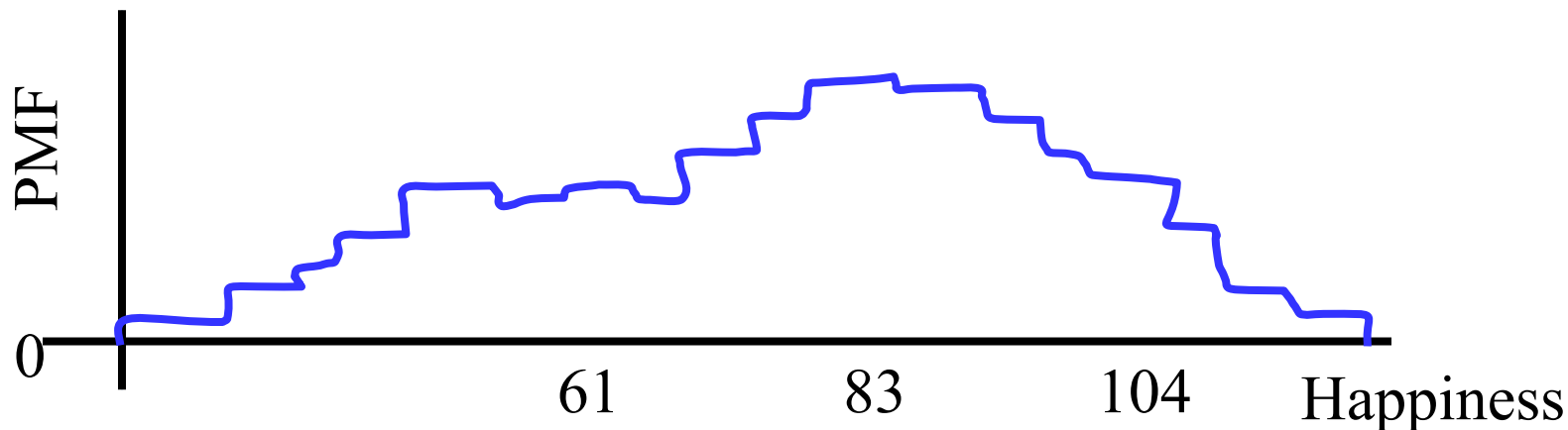


Bootstrap Algorithm (sample):

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]

Bootstrap of Means

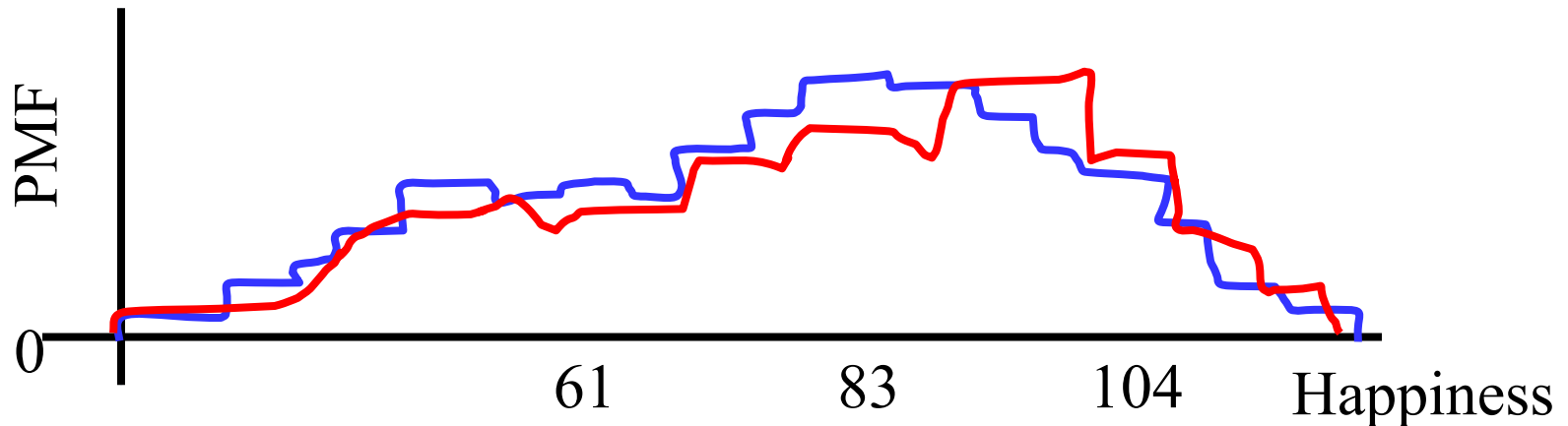


Bootstrap Algorithm (sample):

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]

Bootstrap of Means

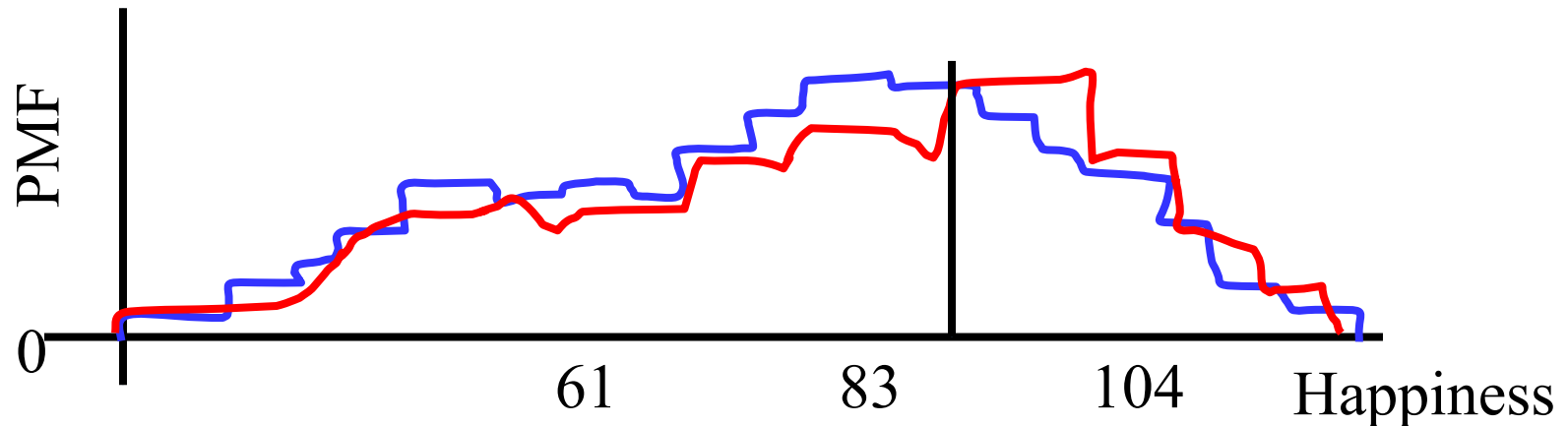


Bootstrap Algorithm (sample):

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]

Bootstrap of Means

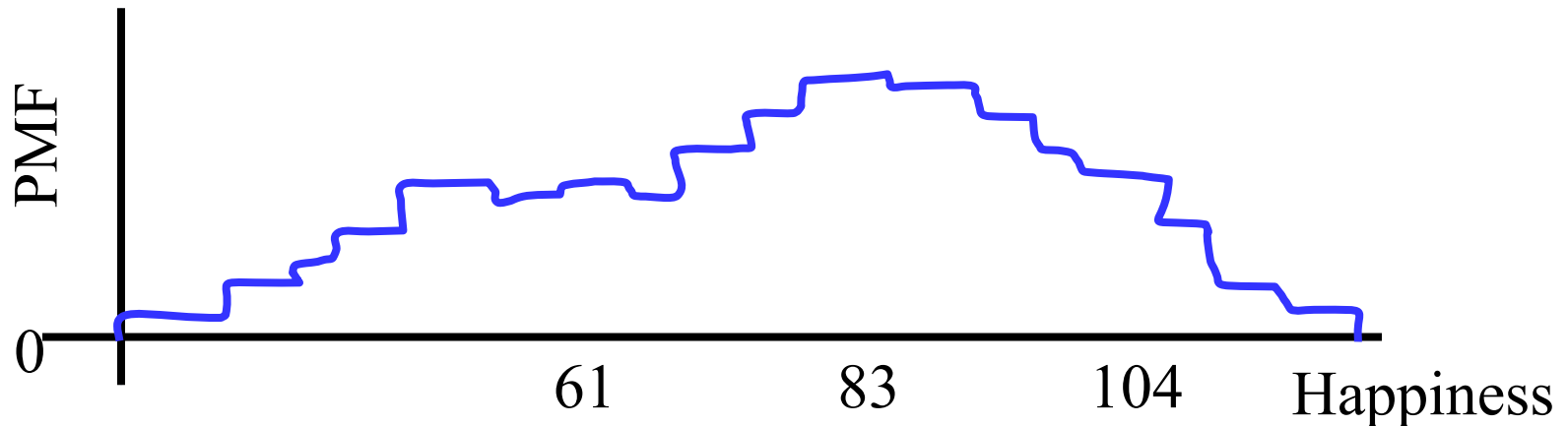


Bootstrap Algorithm (sample):

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]

Bootstrap of Means

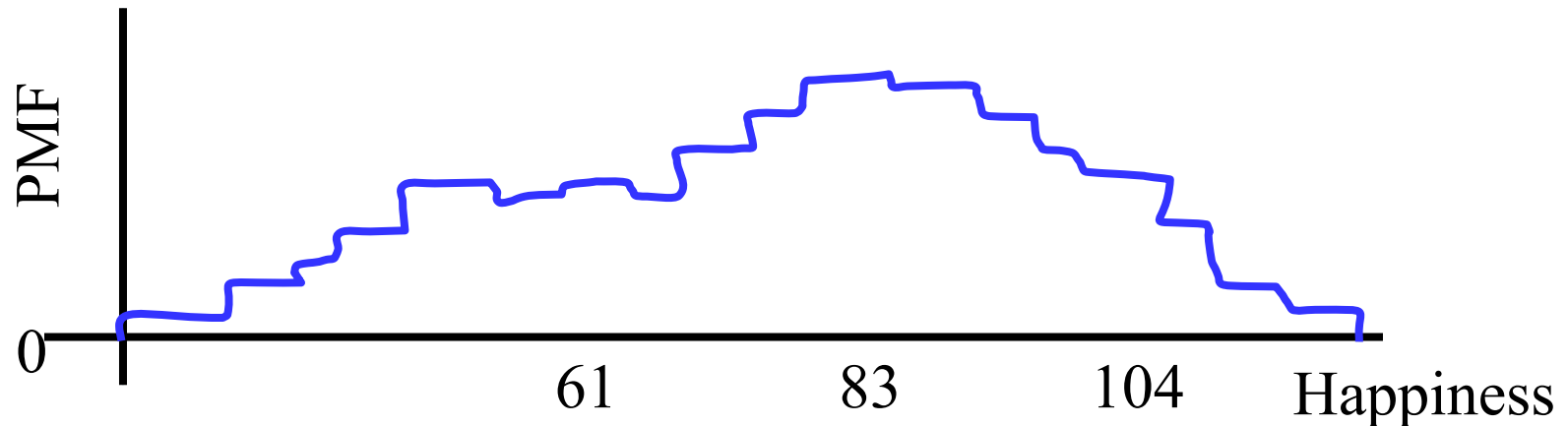


Bootstrap Algorithm (sample):

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]

Bootstrap of Means



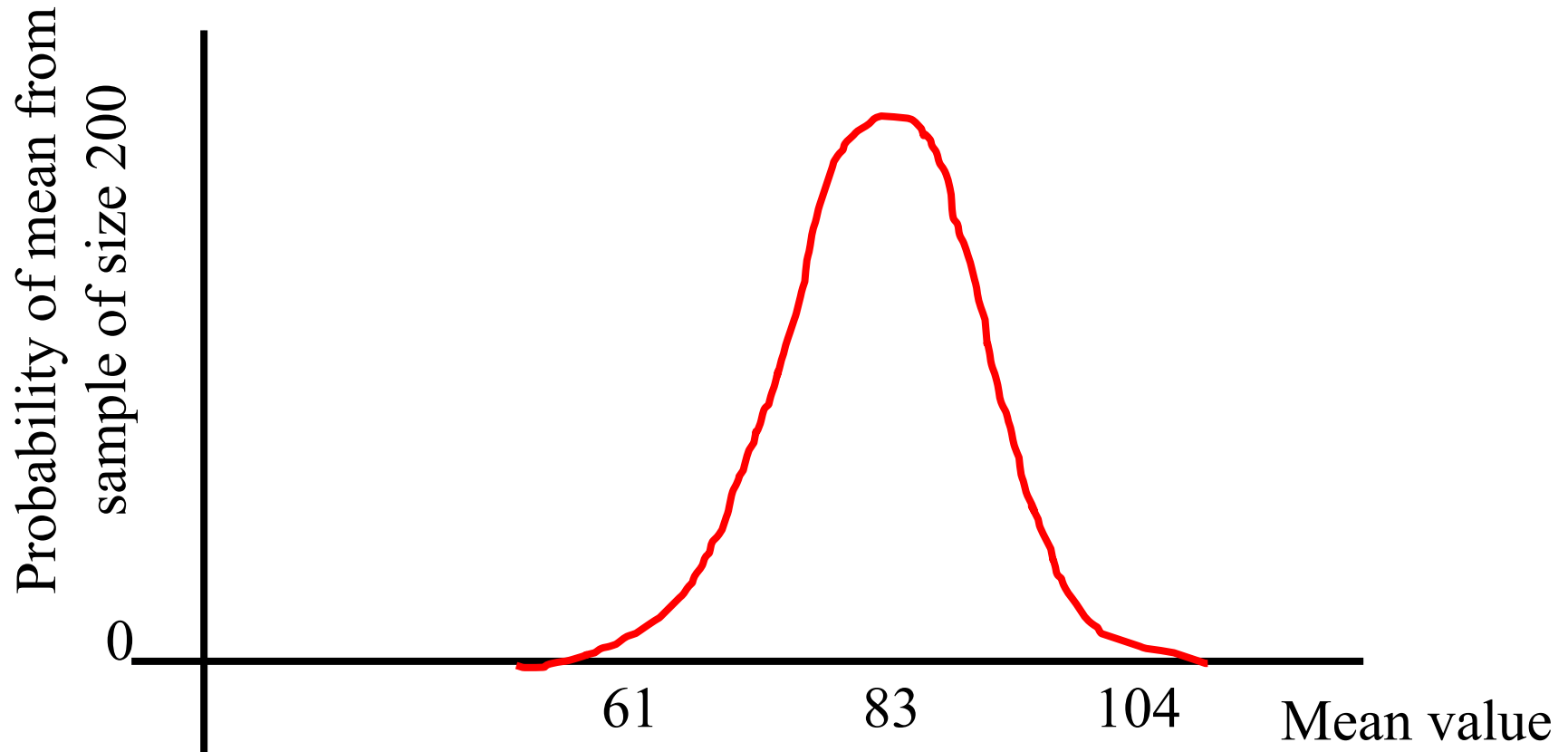
Bootstrap Algorithm (sample):

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]

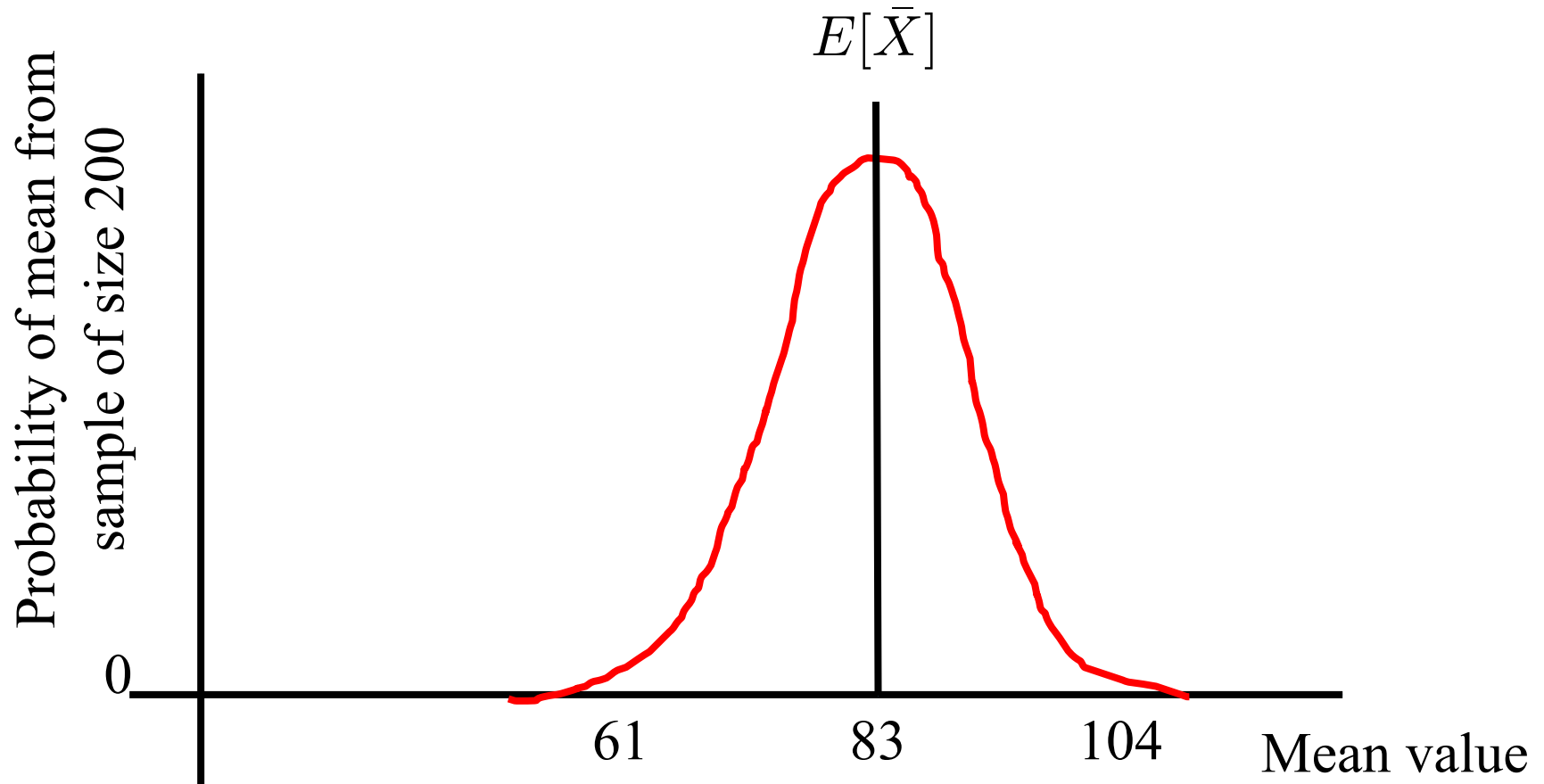
Bootstrap of Means

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



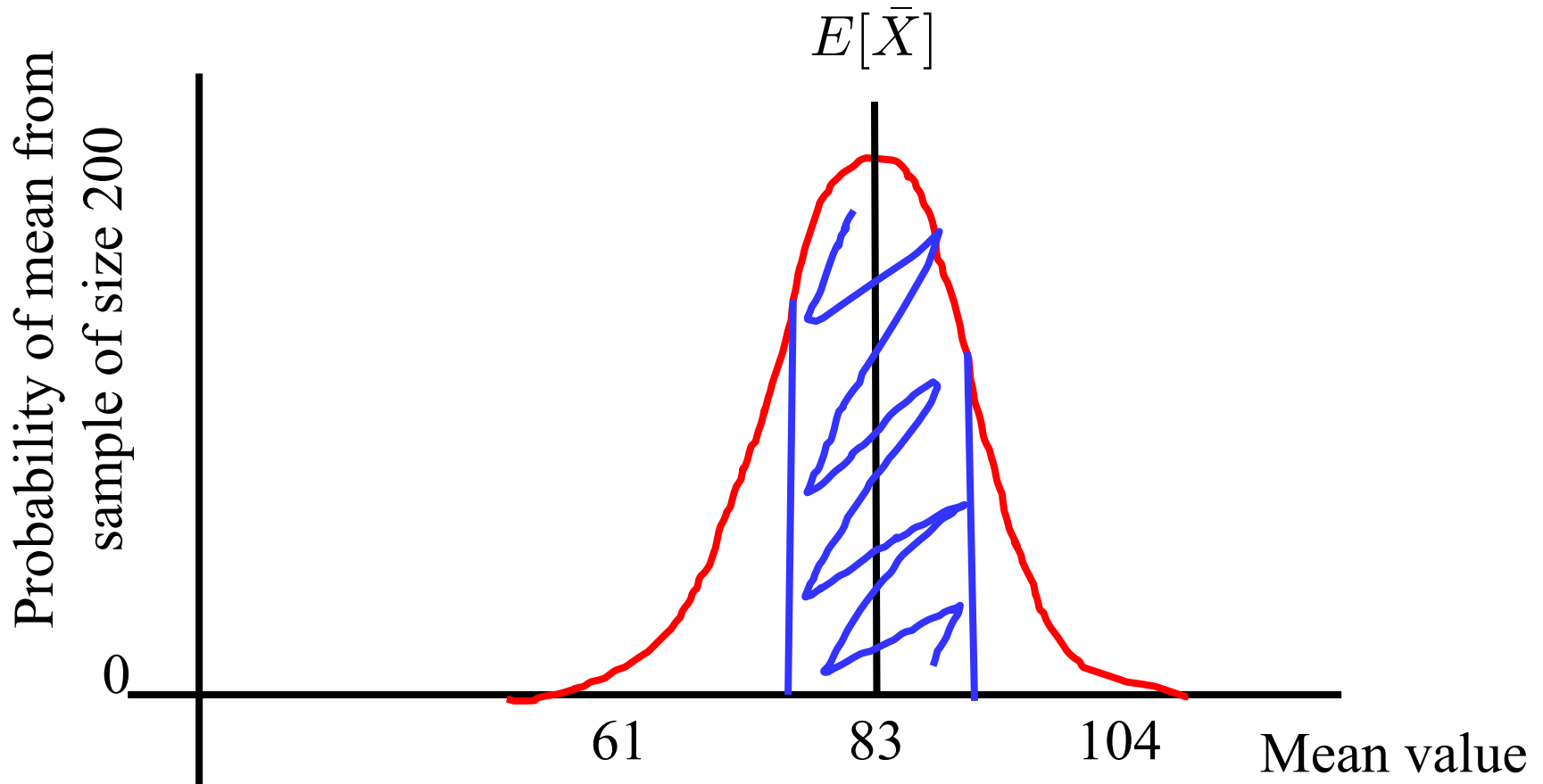
Bootstrap of Means

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



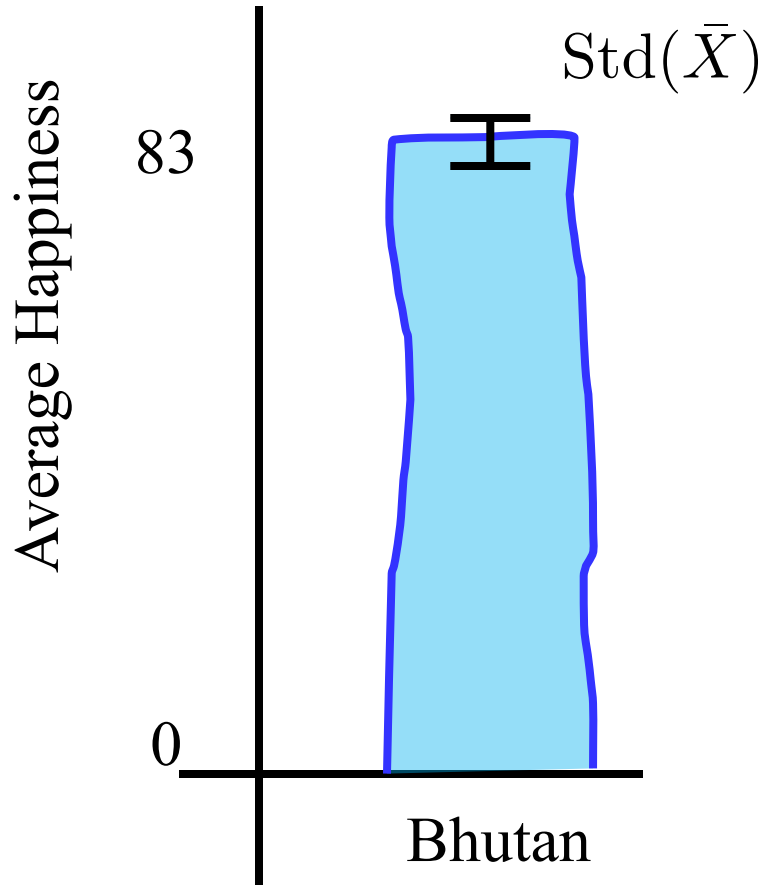
Bootstrap of Means

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

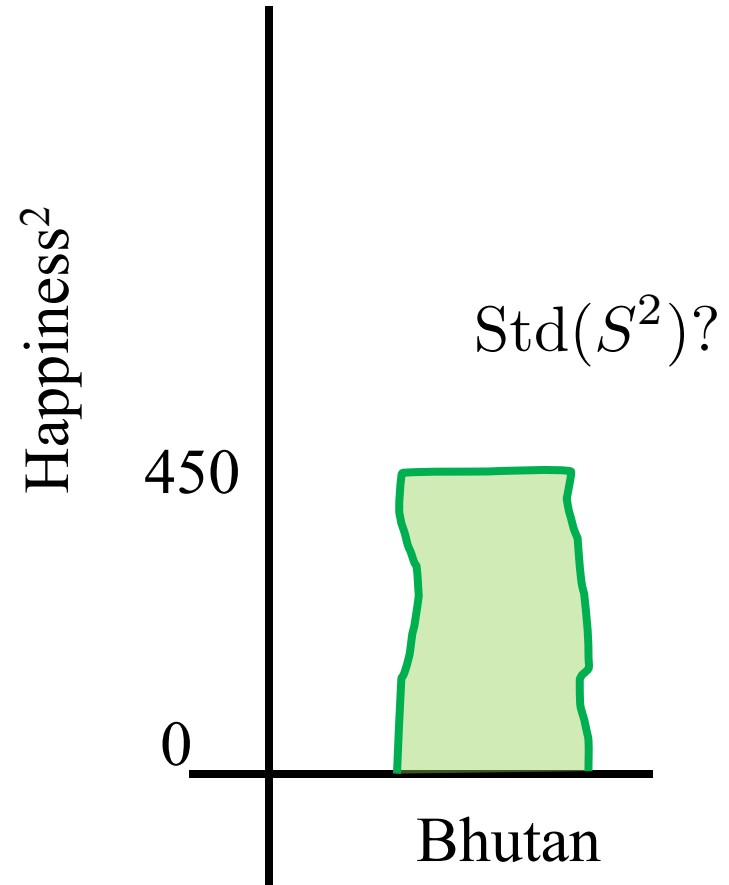


Sample Mean

Average Happiness



Variance of Happiness



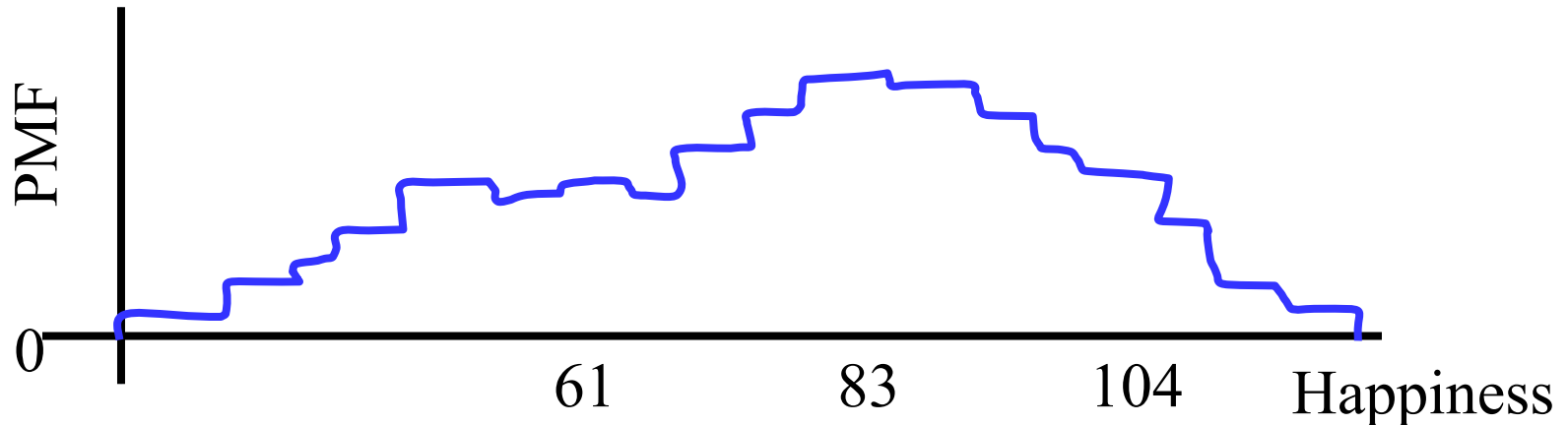
Claim: The average happiness of Bhutan is 83 ± 2

Bootstrap 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**

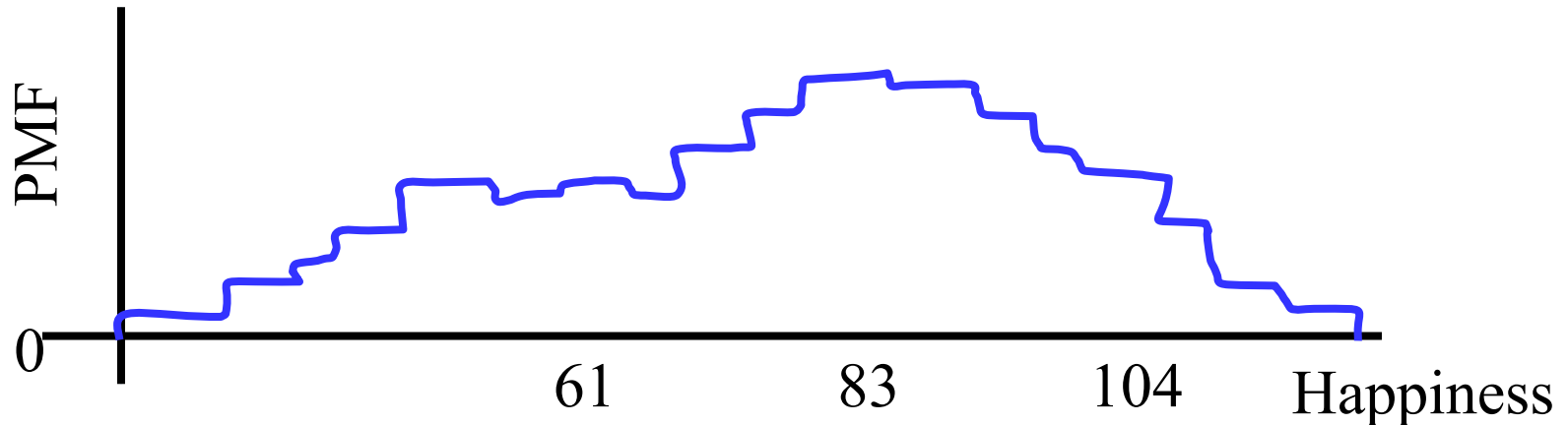
Bootstrap of Variance



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**

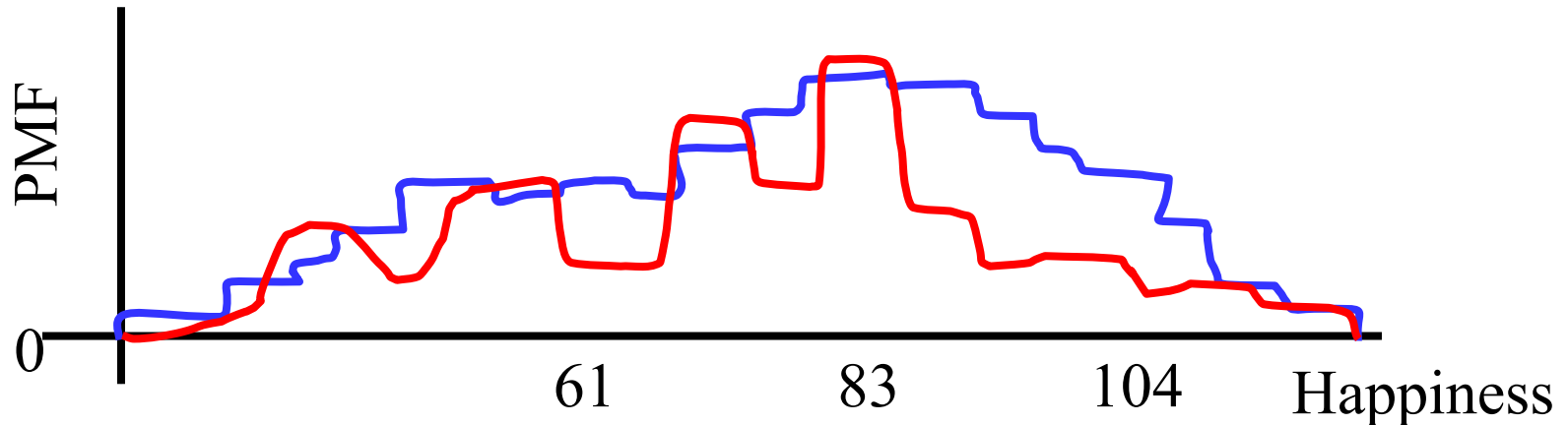
Bootstrap of Variance



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**

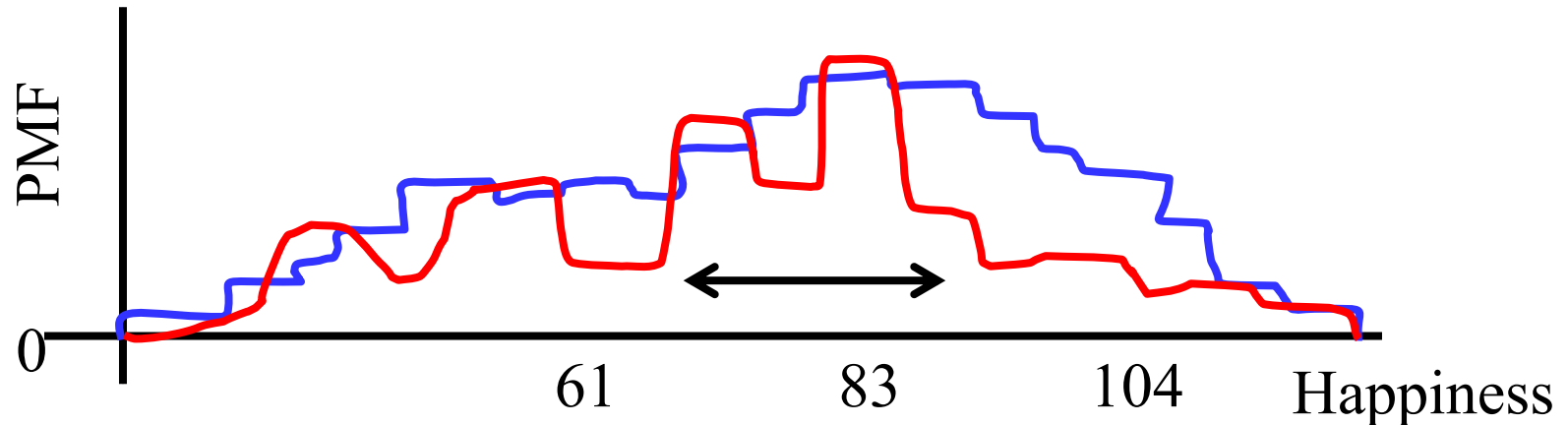
Bootstrap of Variance



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**

Bootstrap of Variance

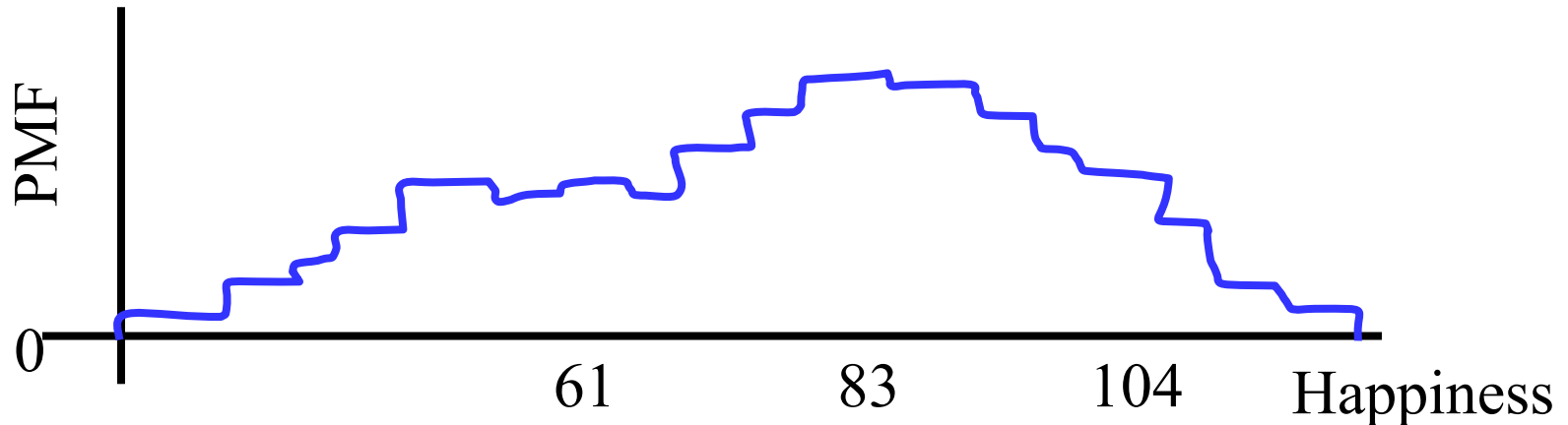


Bootstrap Algorithm (sample):

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]

Bootstrap of Variance

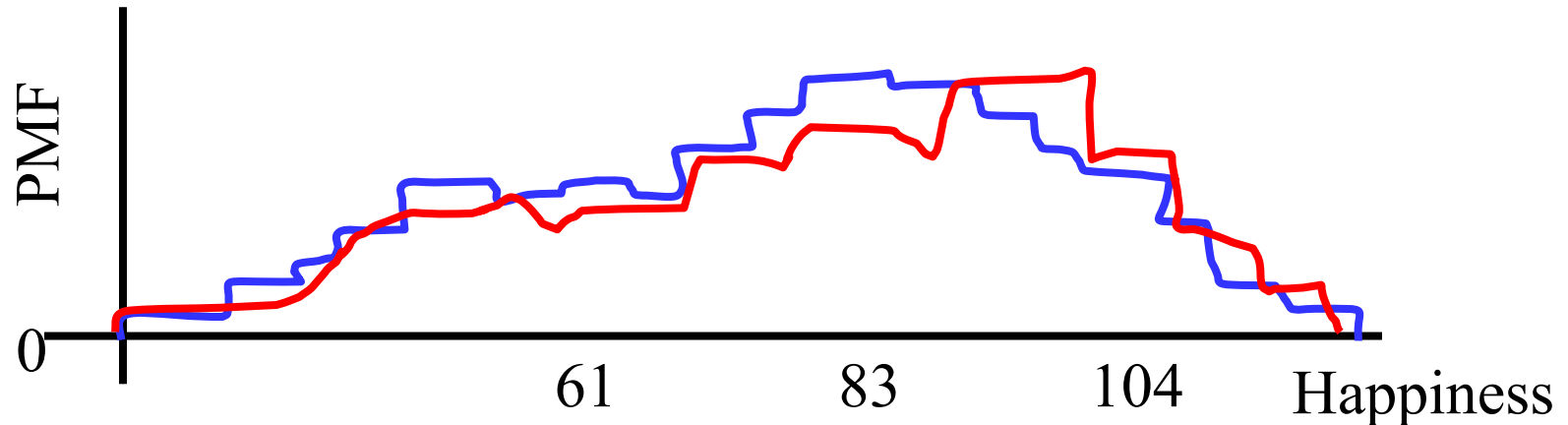


Bootstrap Algorithm (sample):

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]

Bootstrap of Variance

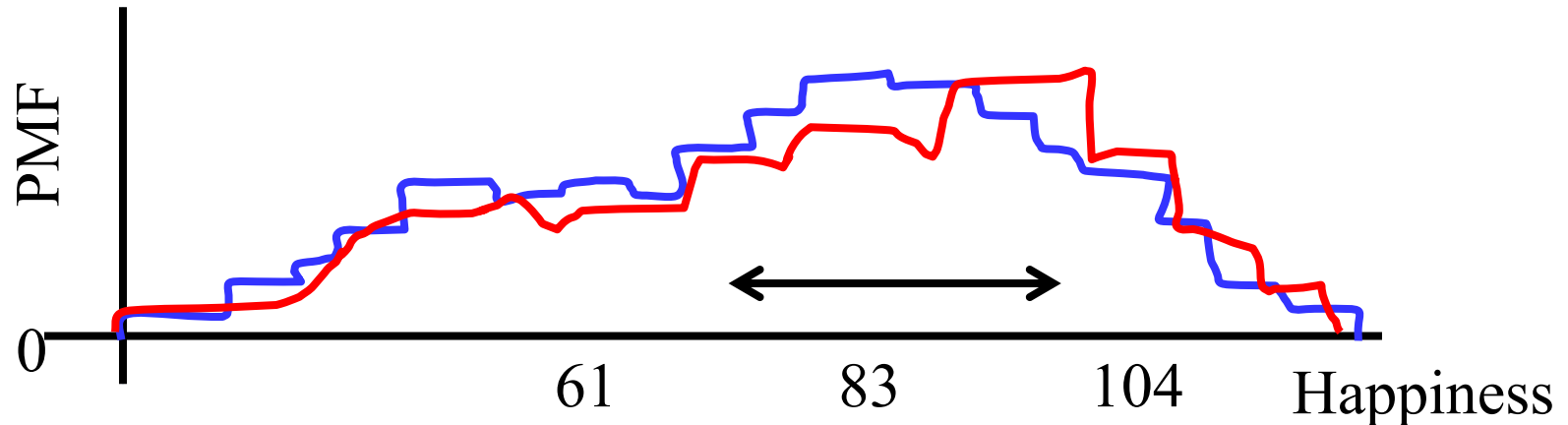


Bootstrap Algorithm (sample):

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]

Bootstrap of Variance

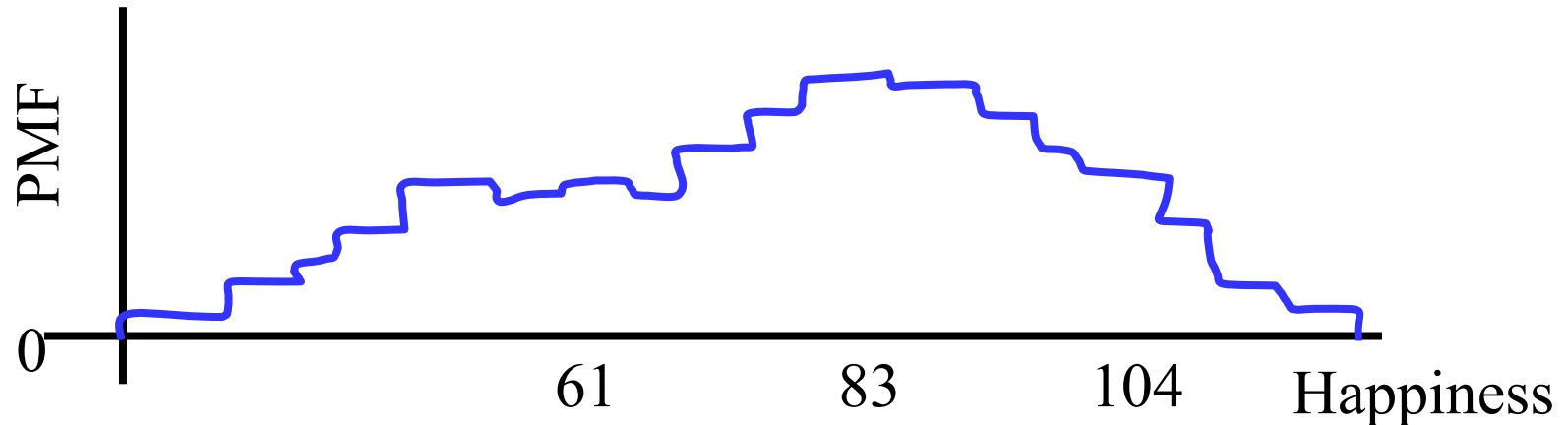


Bootstrap Algorithm (sample):

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]

Bootstrap of Variance

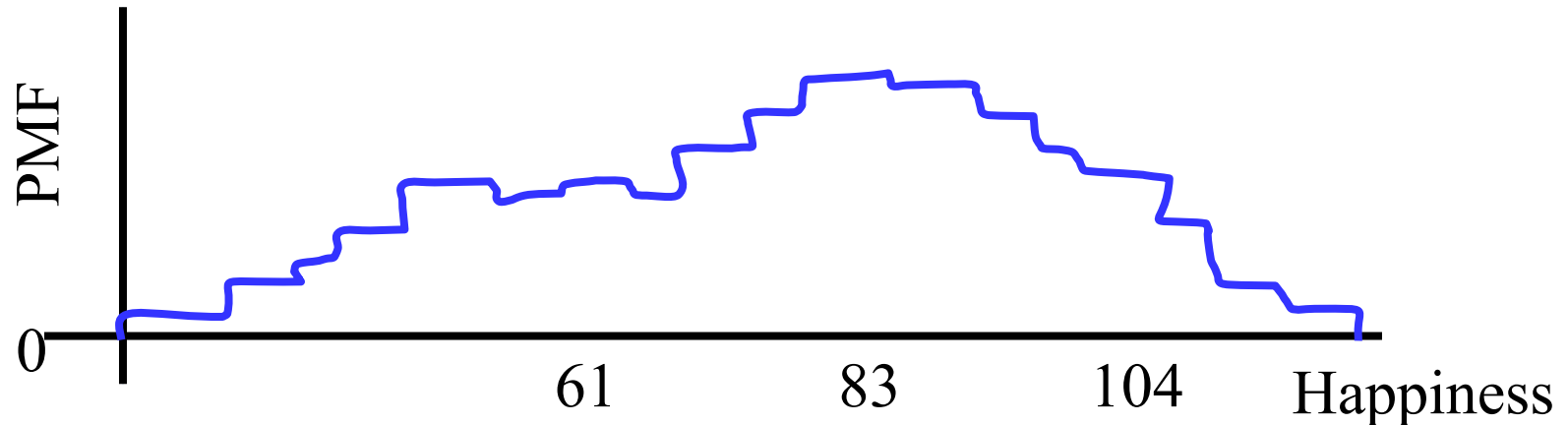


Bootstrap Algorithm (sample):

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]

Bootstrap of Variance



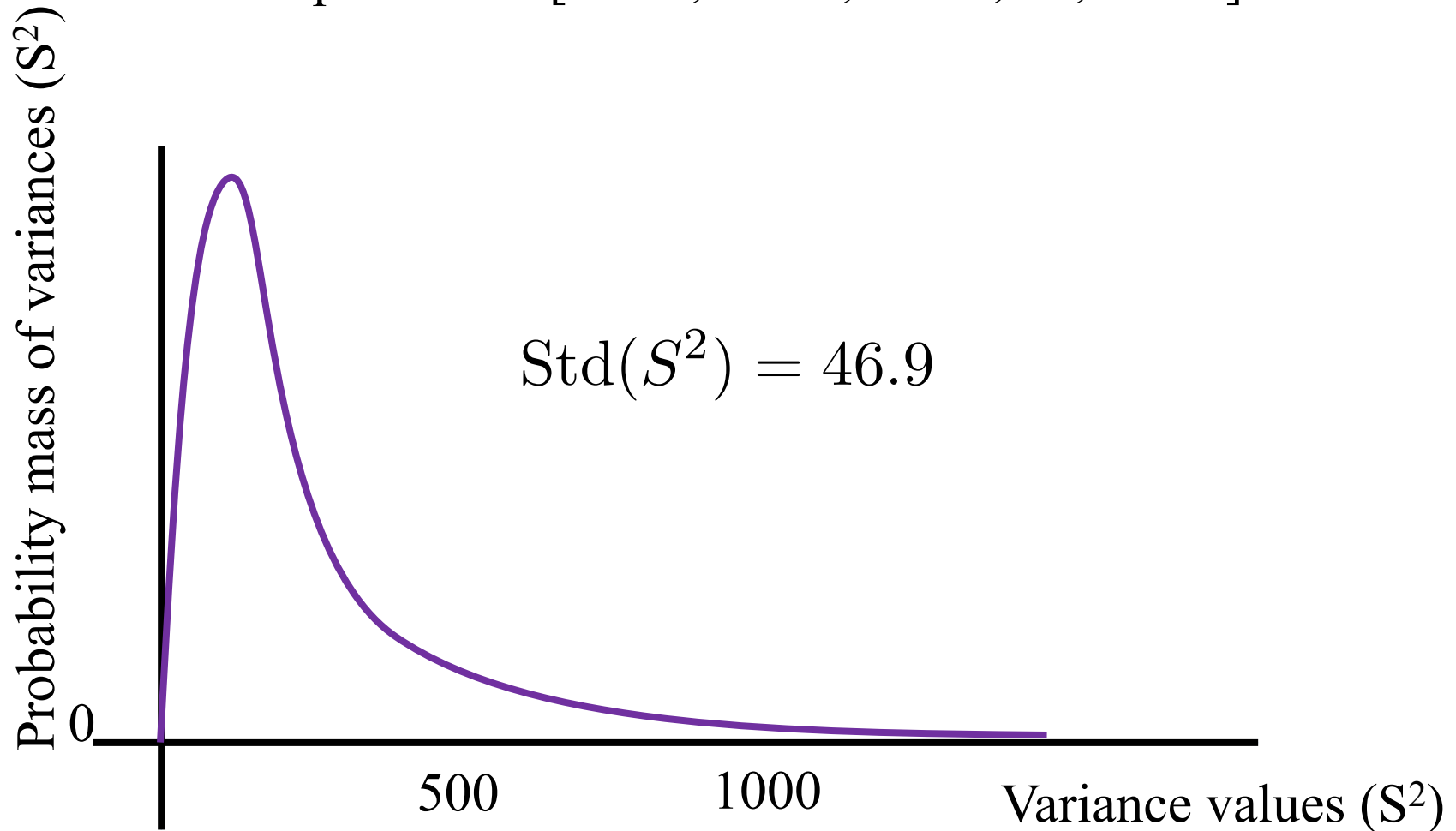
Bootstrap Algorithm (sample):

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]

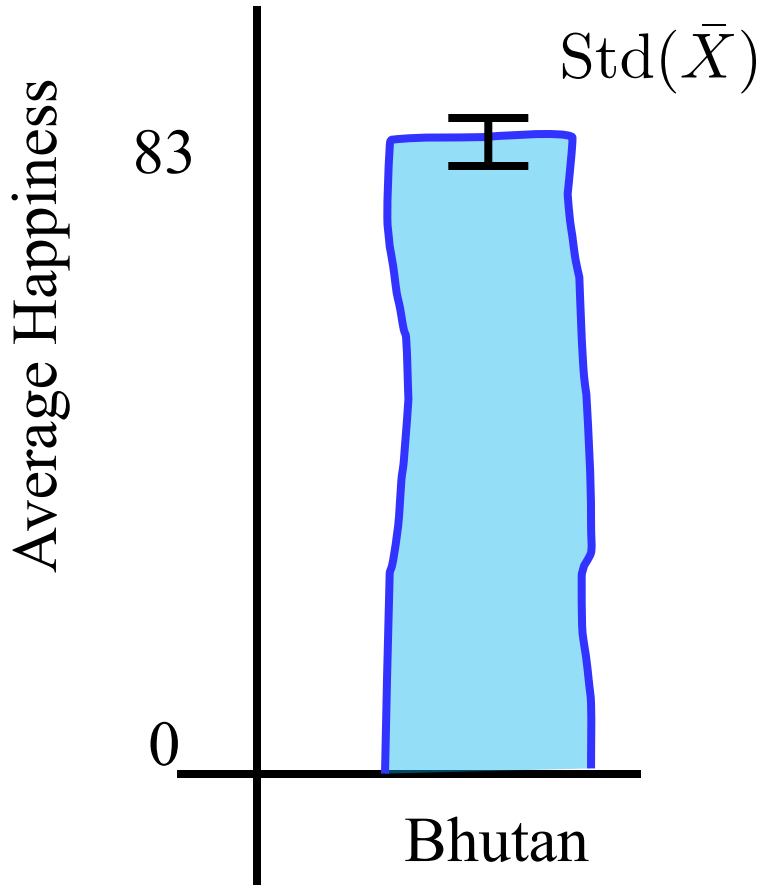
Bootstrap of Variance

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

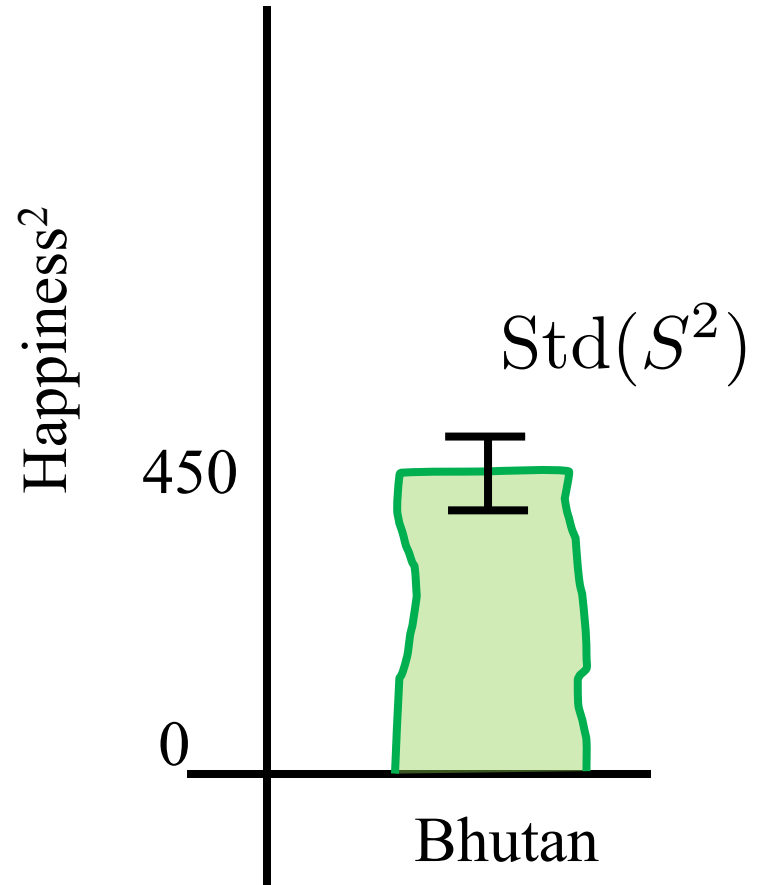


Sample Mean

Average Happiness



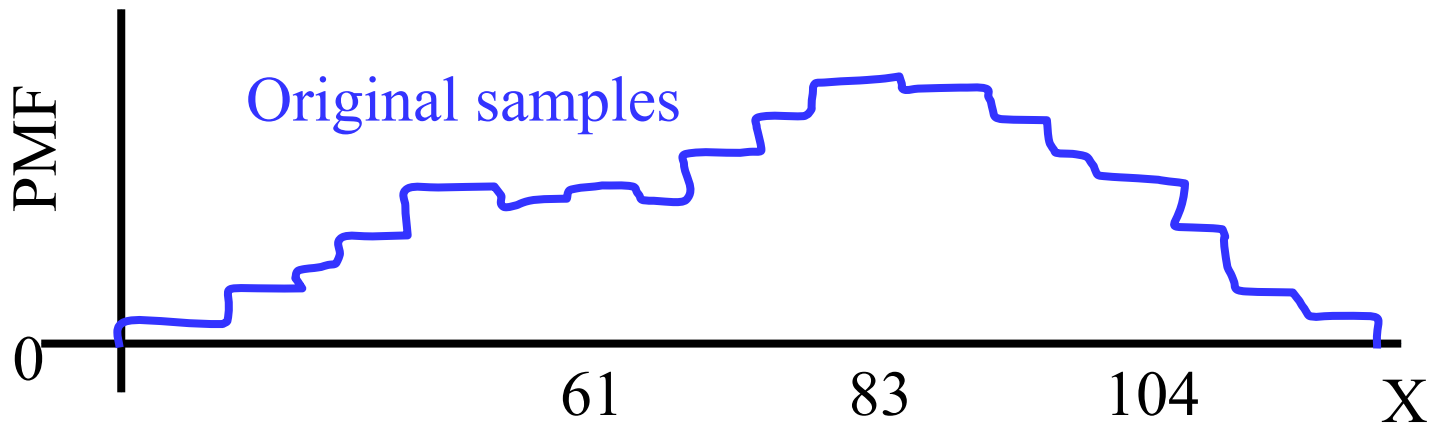
Variance of Happiness



Claim: The average happiness of Bhutan is 83 ± 2

Algorithm in Practice

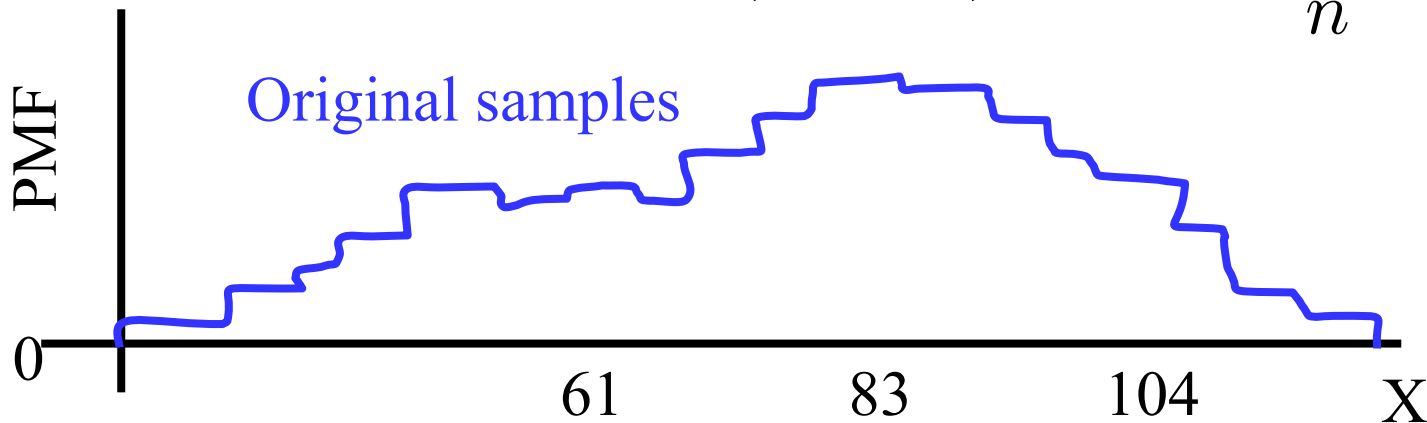
```
def resample(samples):  
    # Estimate the PMF using the samples  
    # Draw K new samples from the PMF
```



Algorithm 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)
```

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



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**

Algorithm 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!



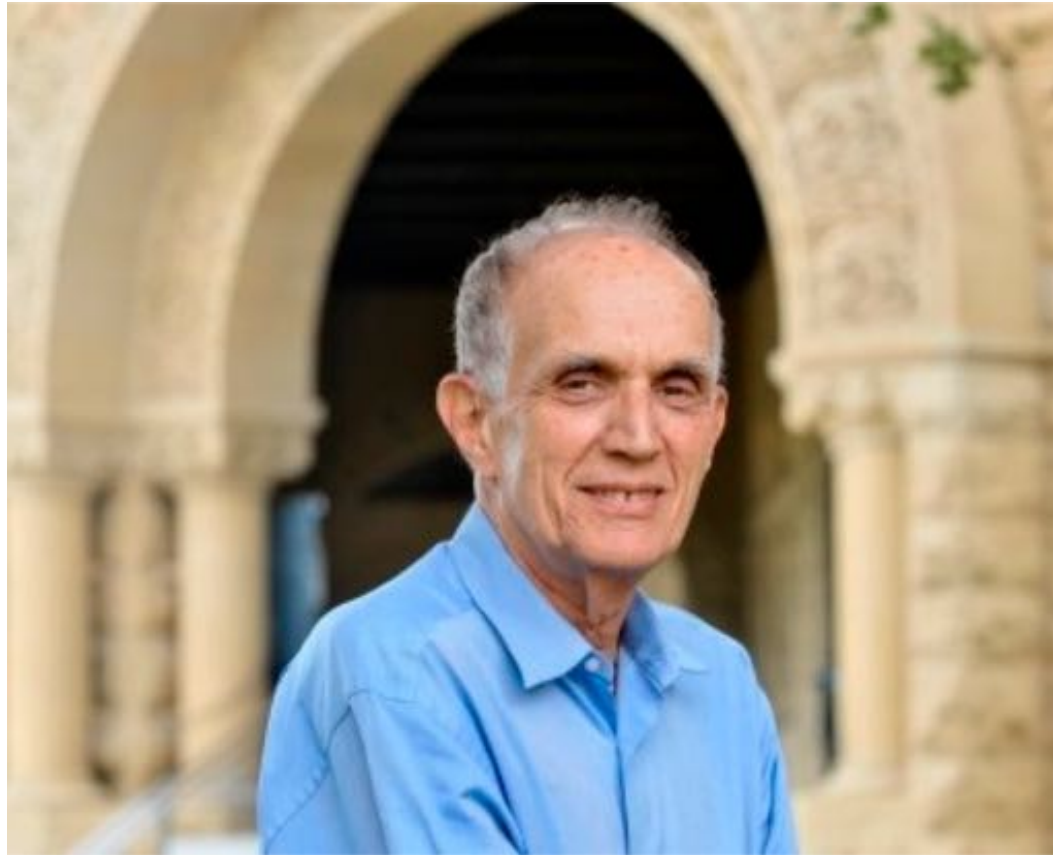
Bootstrap provides a way to calculate probabilities of statistics using code.



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

Null Hypothesis Test

Population 1	Population 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: Population 1 and population 2 are different distributions with a 0.7 difference of means

Midterm

Midterm (part 1)

1



2



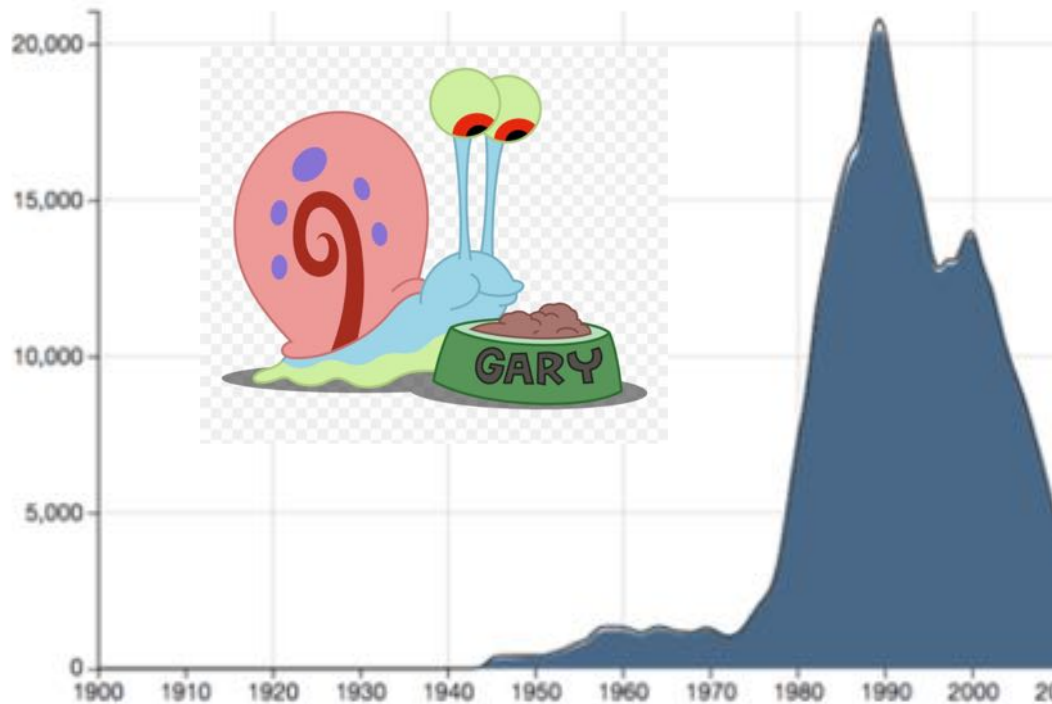
3



Midterm (part 2)

4

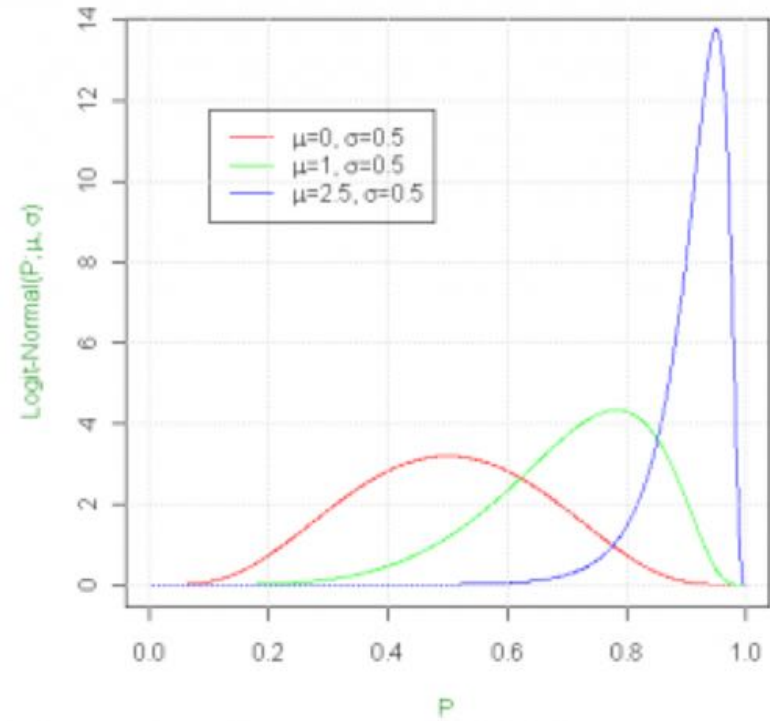
Birth years of American girls named Lauren



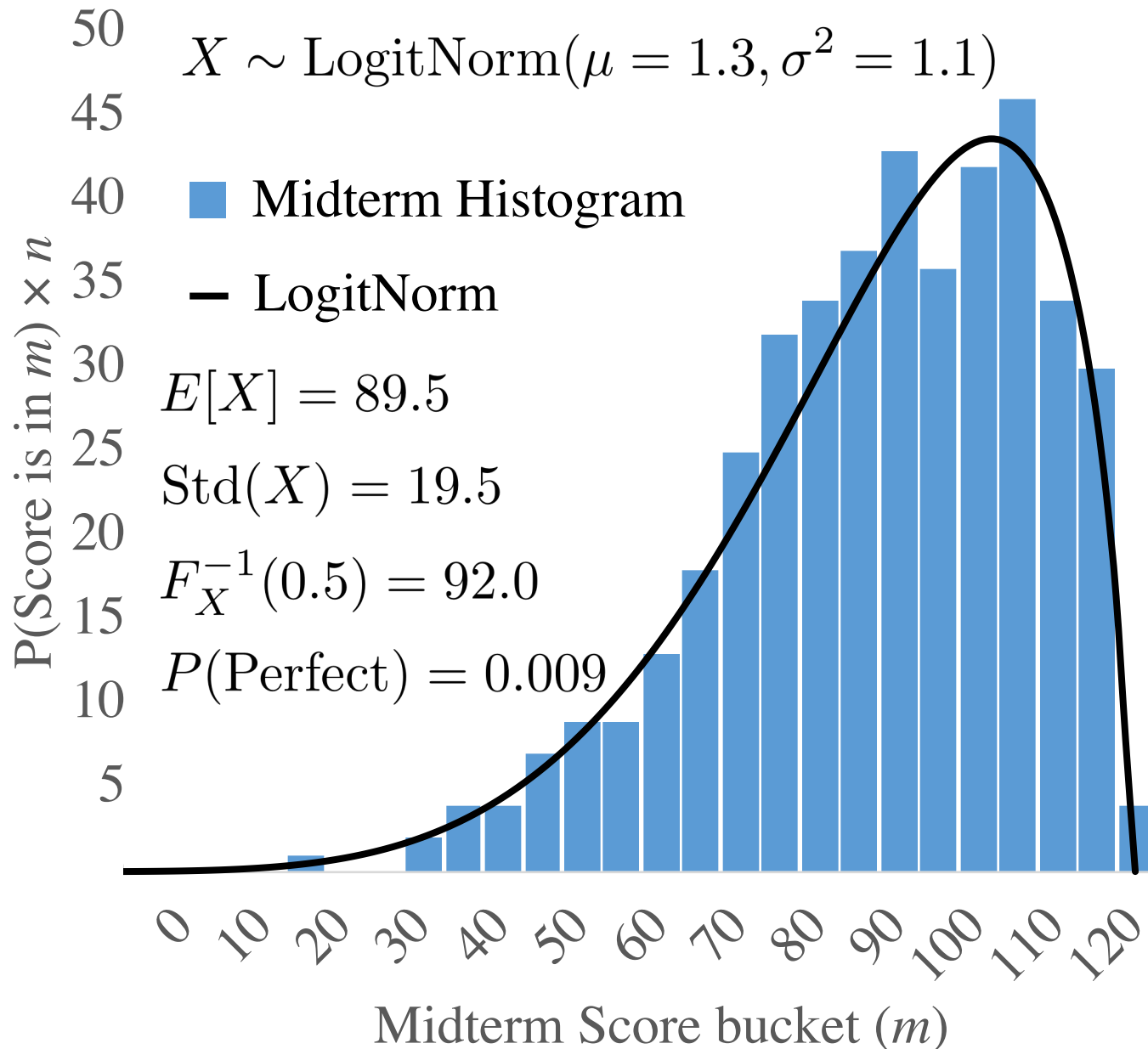
The median living girl named Lauren was born around 1992 and ranges from 16 to 29 years old.

5

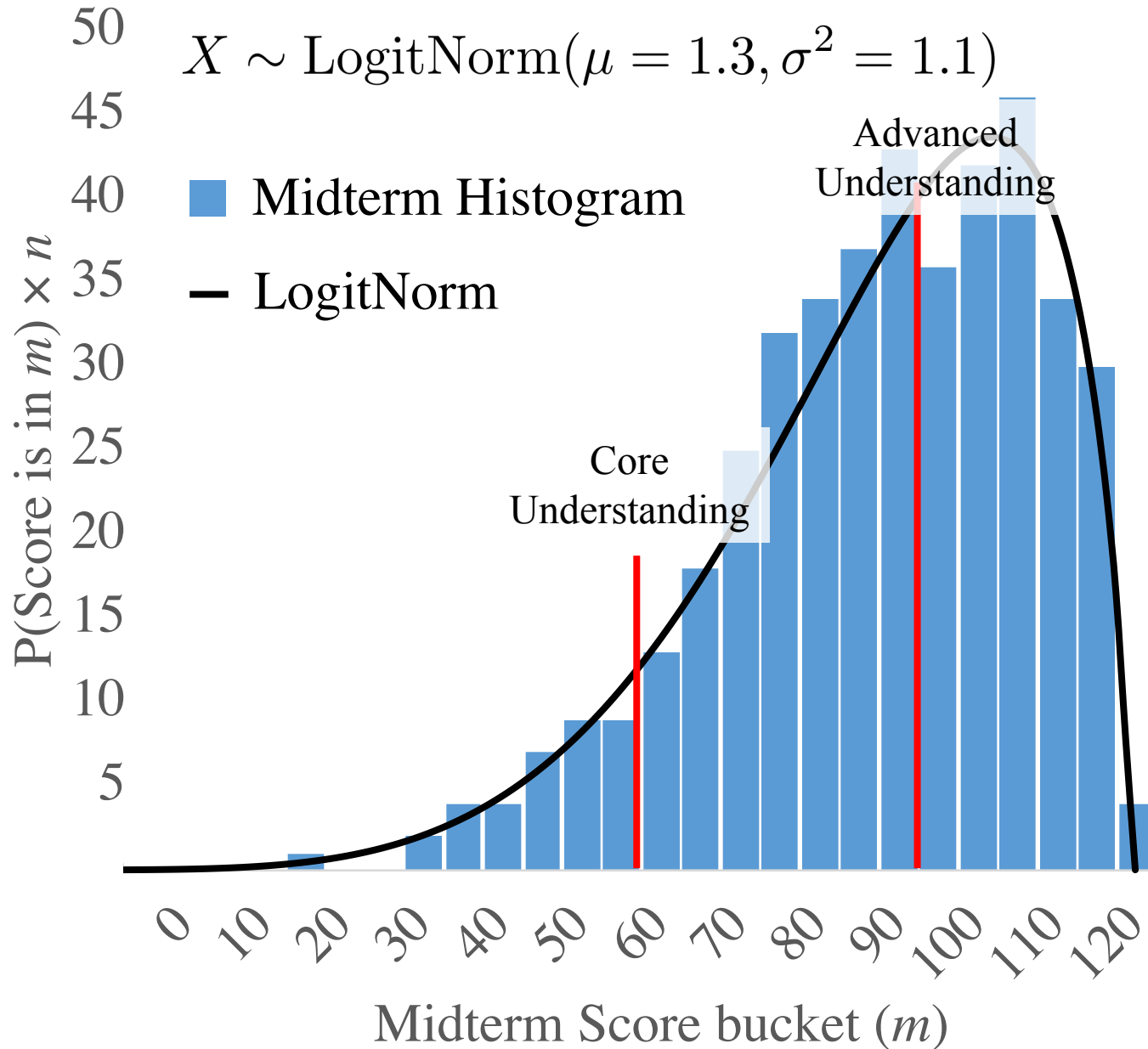
Examples of the Logit-Normal Distribution



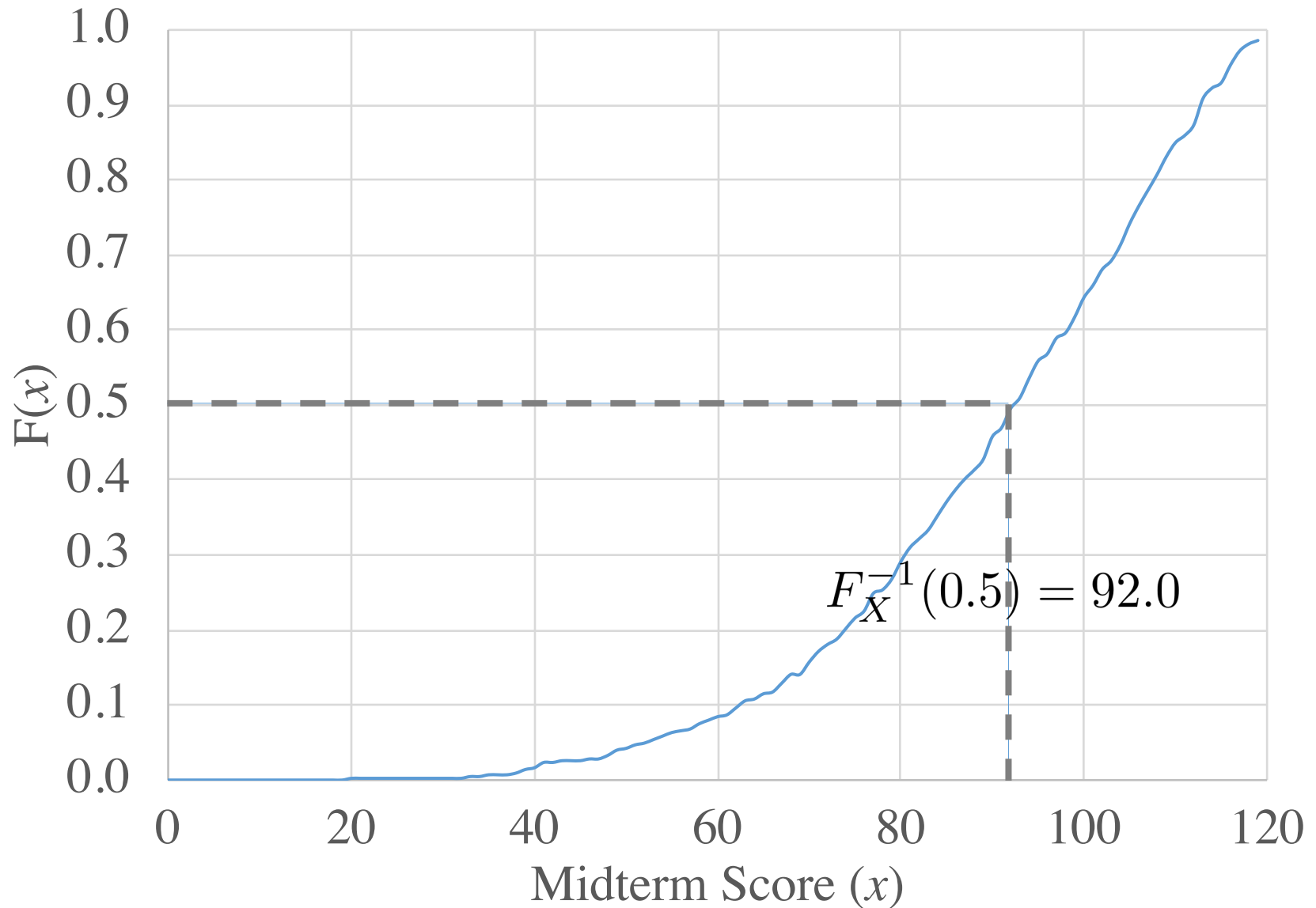
Midterm Distribution



Midterm Distribution



Midterm Cumulative Density



Midterm Distribution

