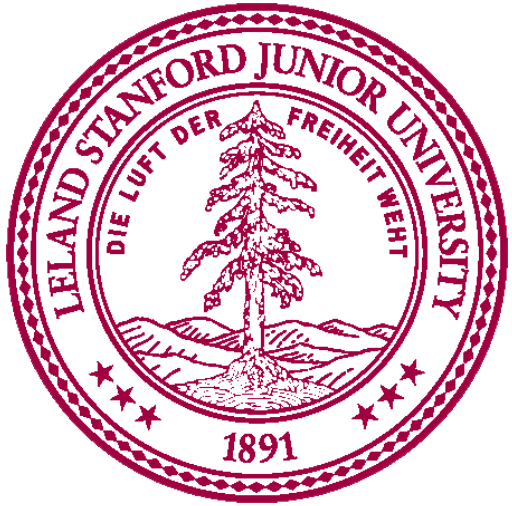
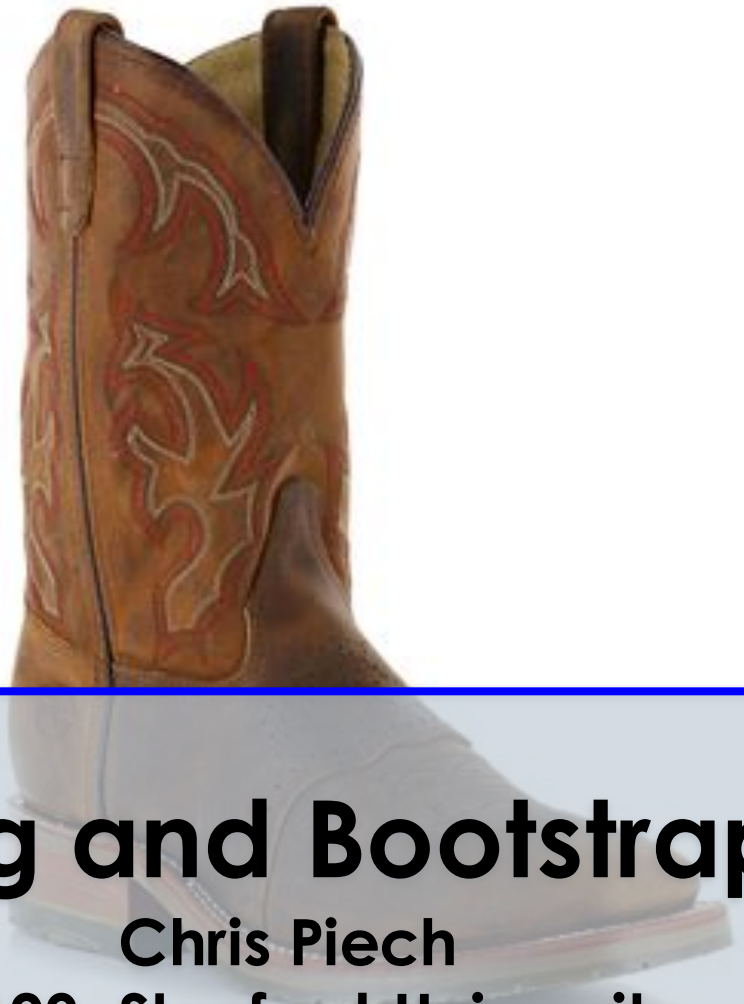


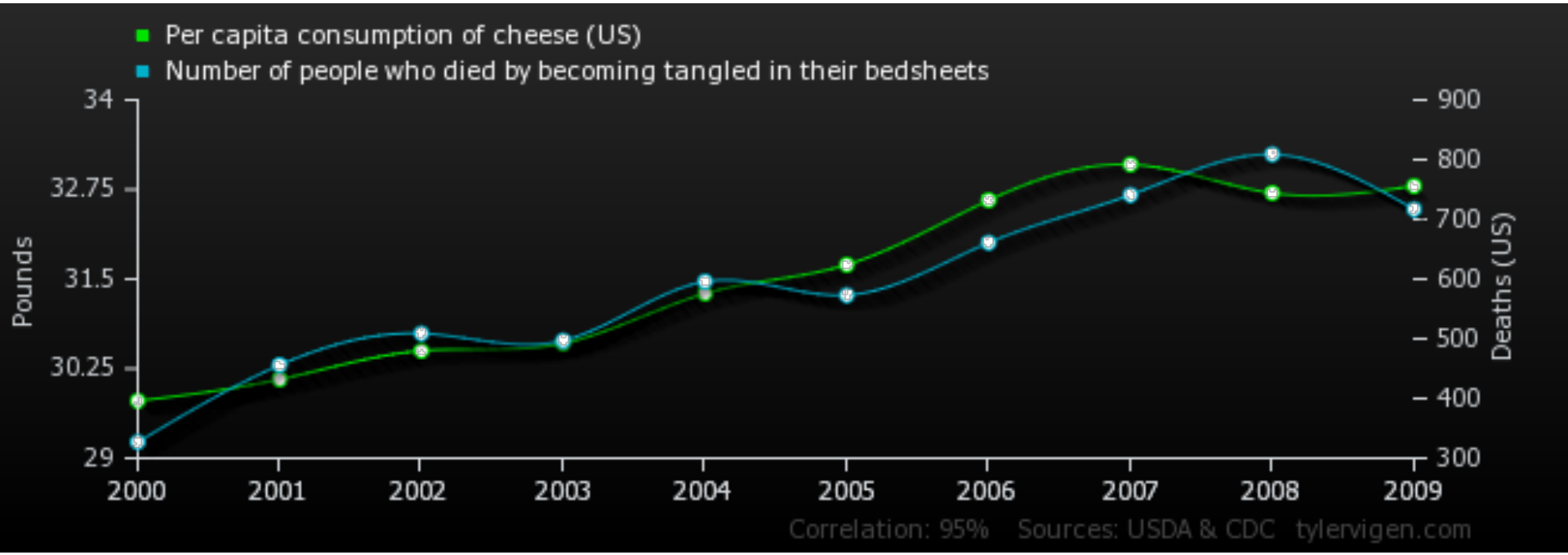


Sampling and Bootstrapping

Chris Piech
CS109, Stanford University

Review

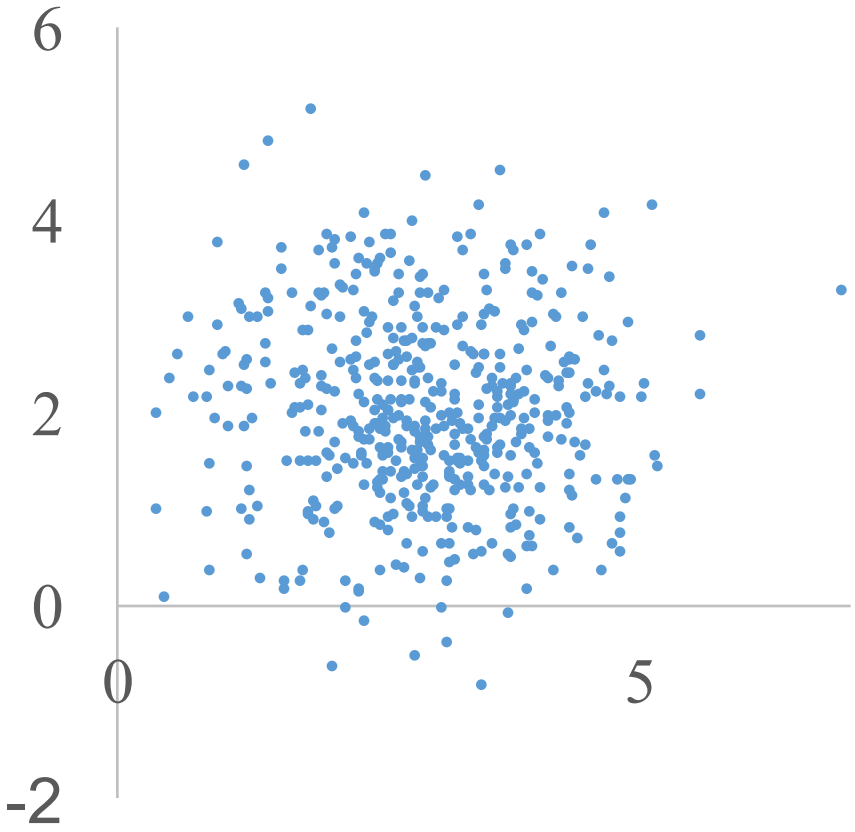
Tell your friends!

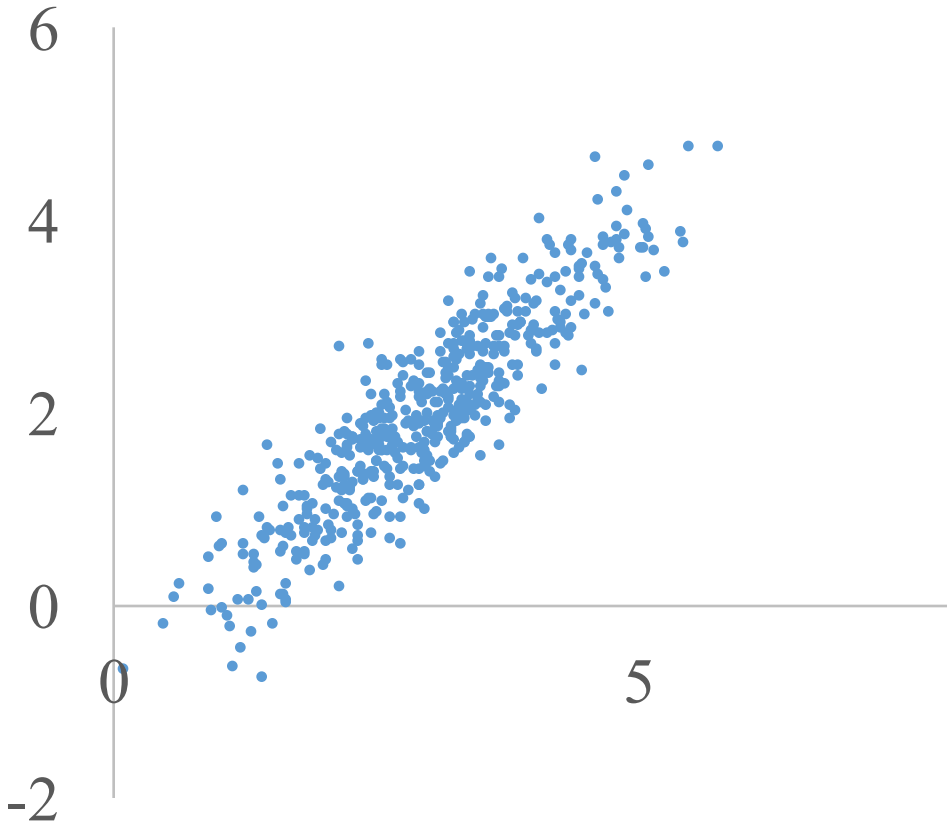


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

Spot The Difference





The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- Equivalently:

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

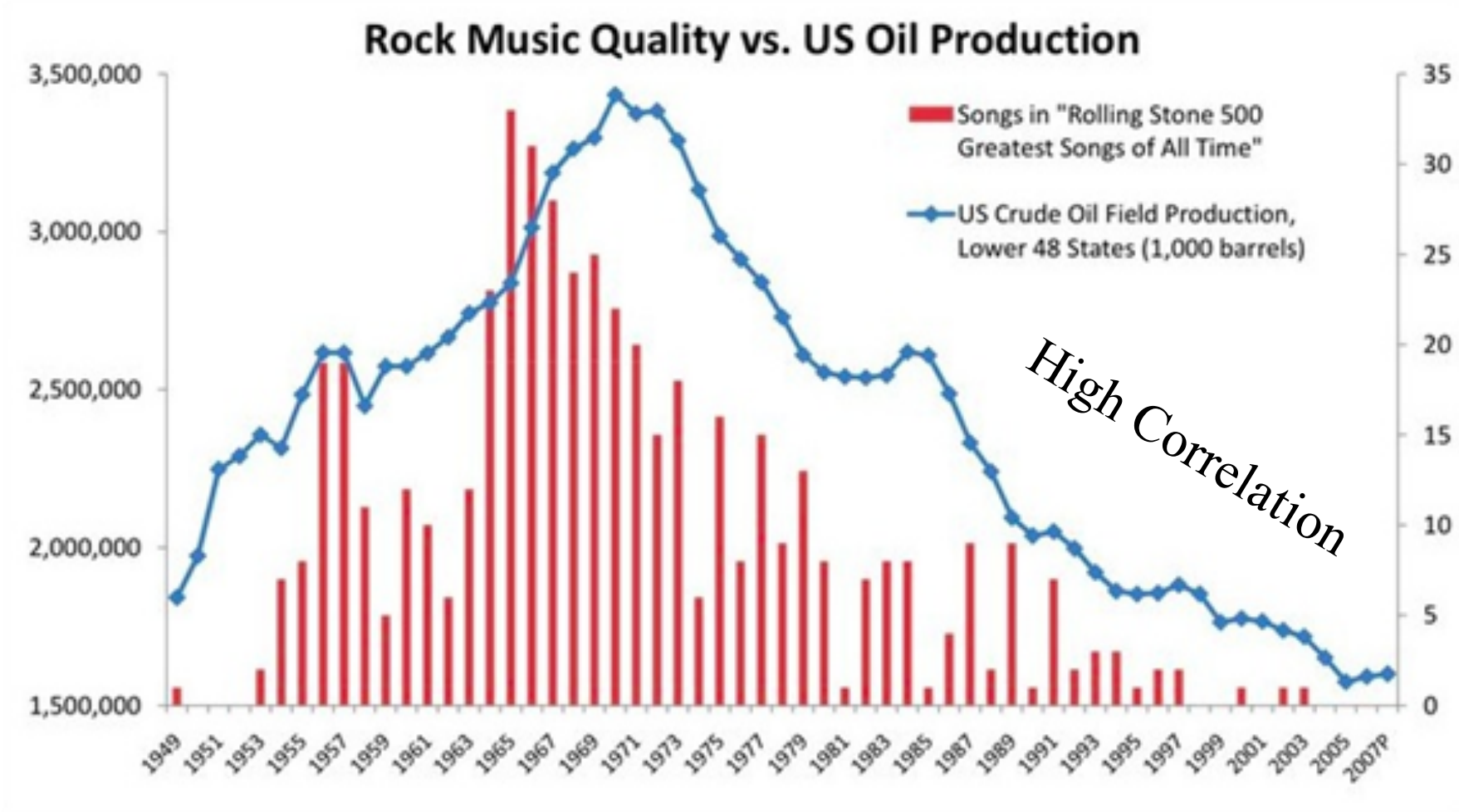
Viva La Correlatión

- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var(X)Var(Y)}}}$$

Philosophy

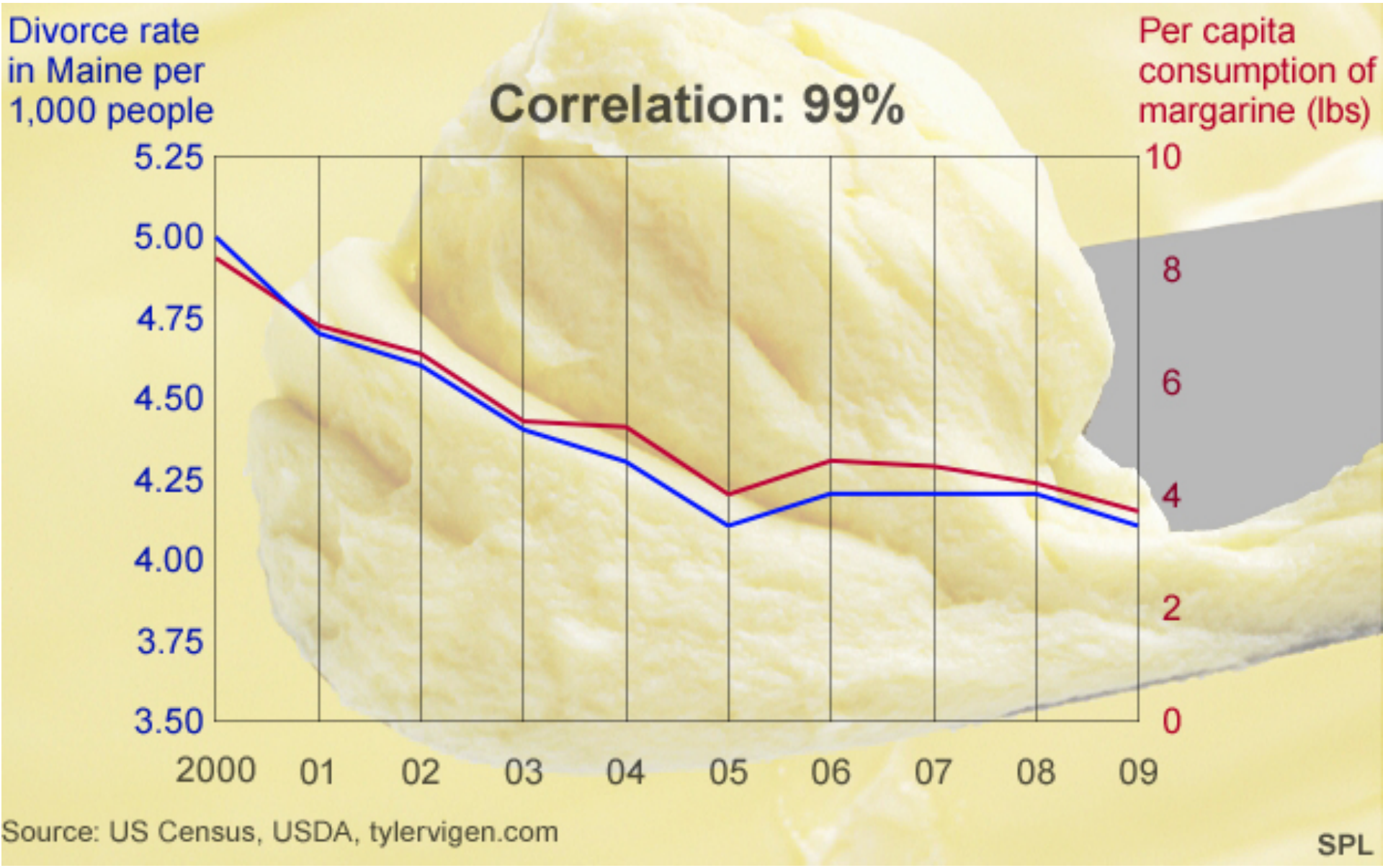
Rock Music Vs Oil?



Hubbert Peak Theory

http://www.aei.org/publication/blog/

Divorce Vs Butter?



http://www.bbc.com/news/magazine-27537142

Science

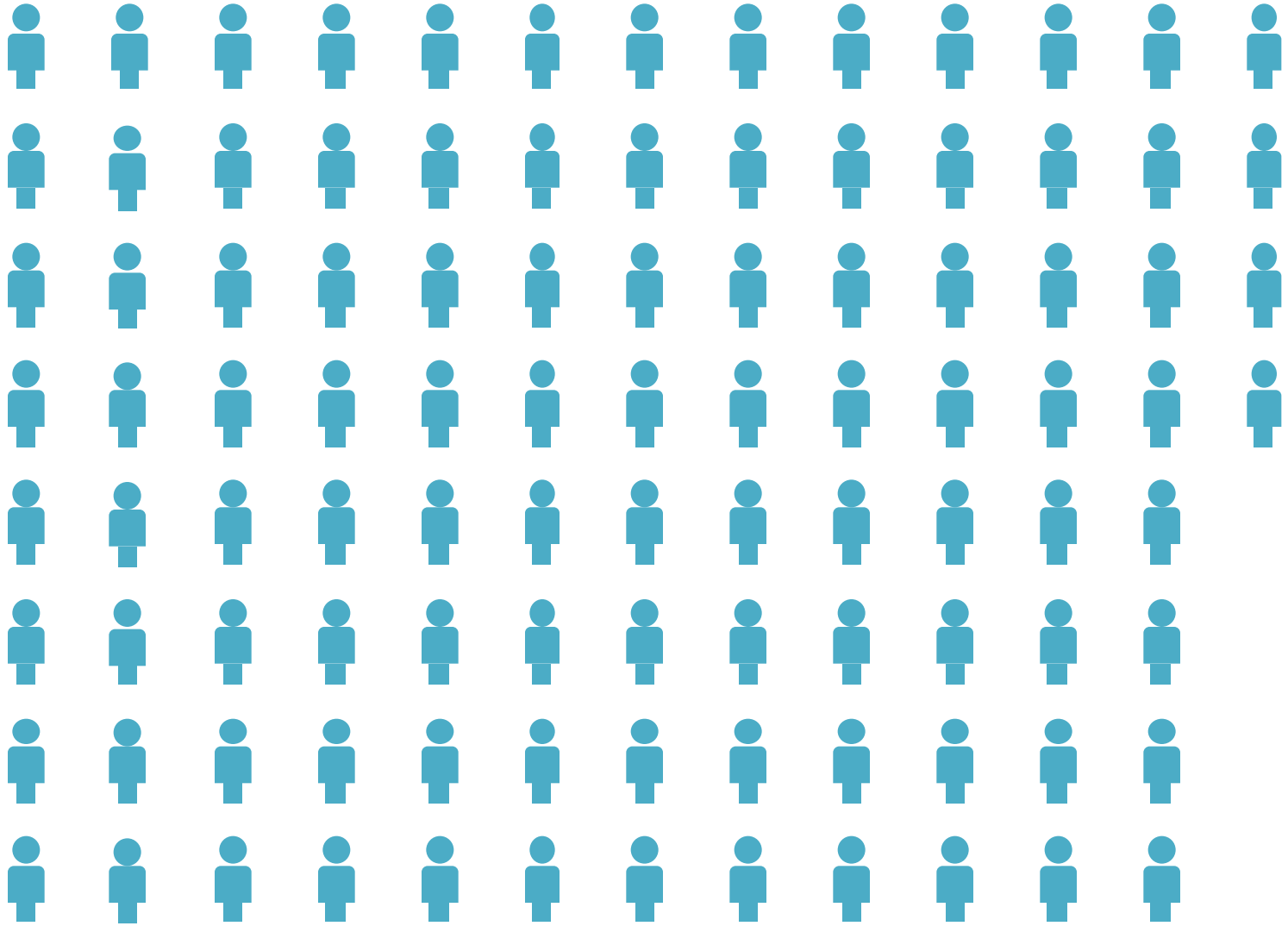
Machine Learning Example

- 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:

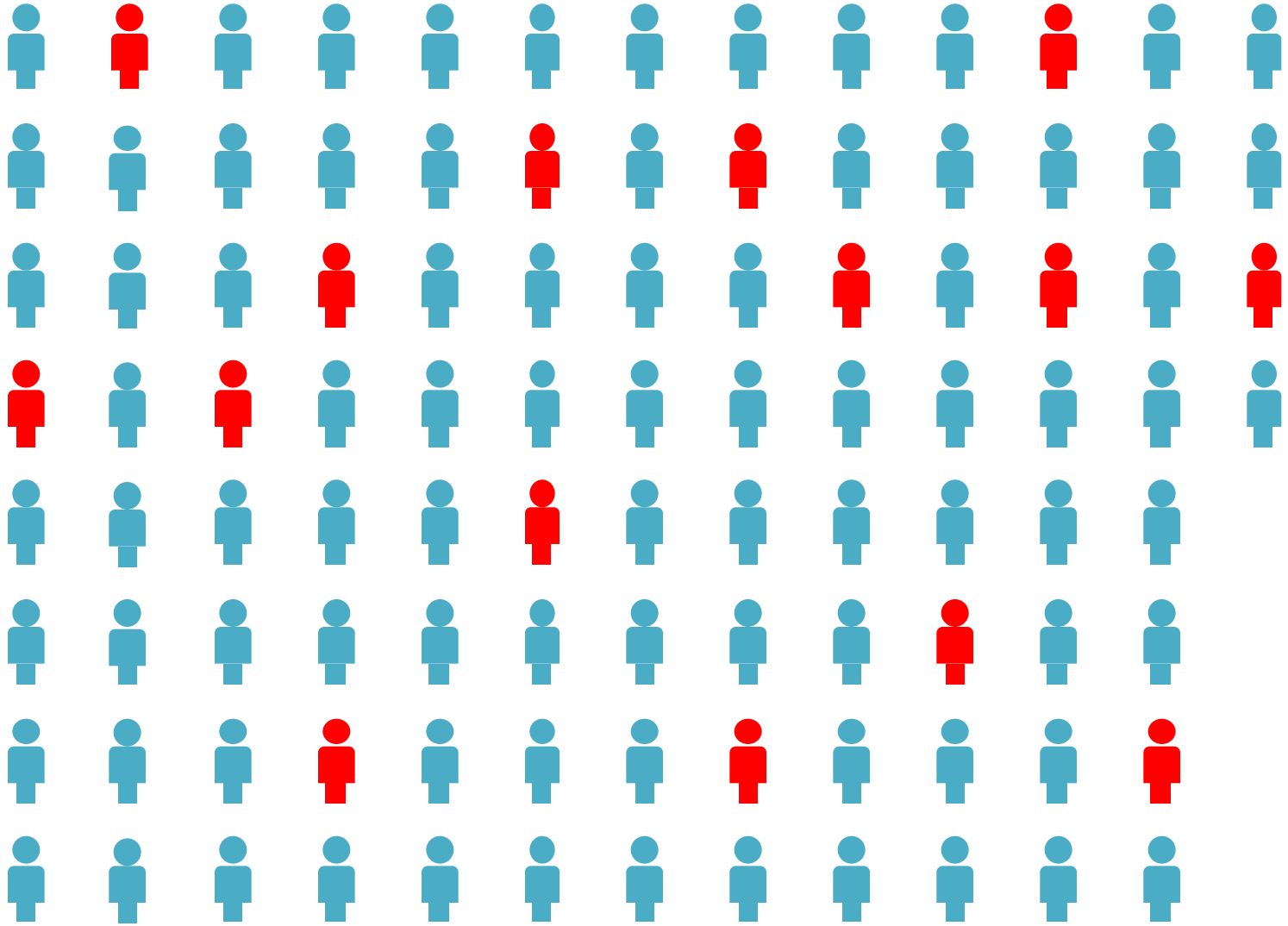
Happiness = {72, 85, 79, 91, 68, … , 71}

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

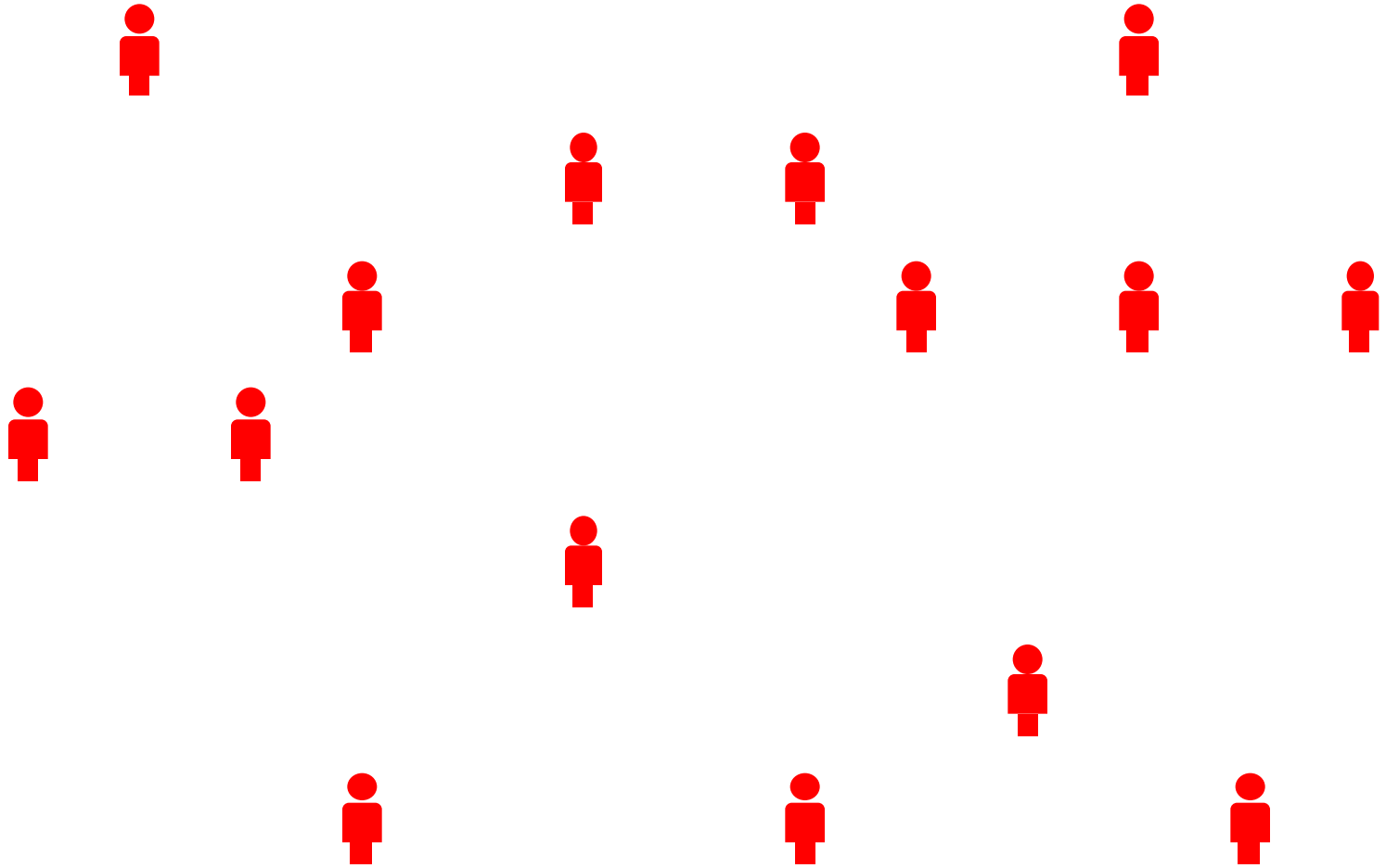
Population



Sample



Sample



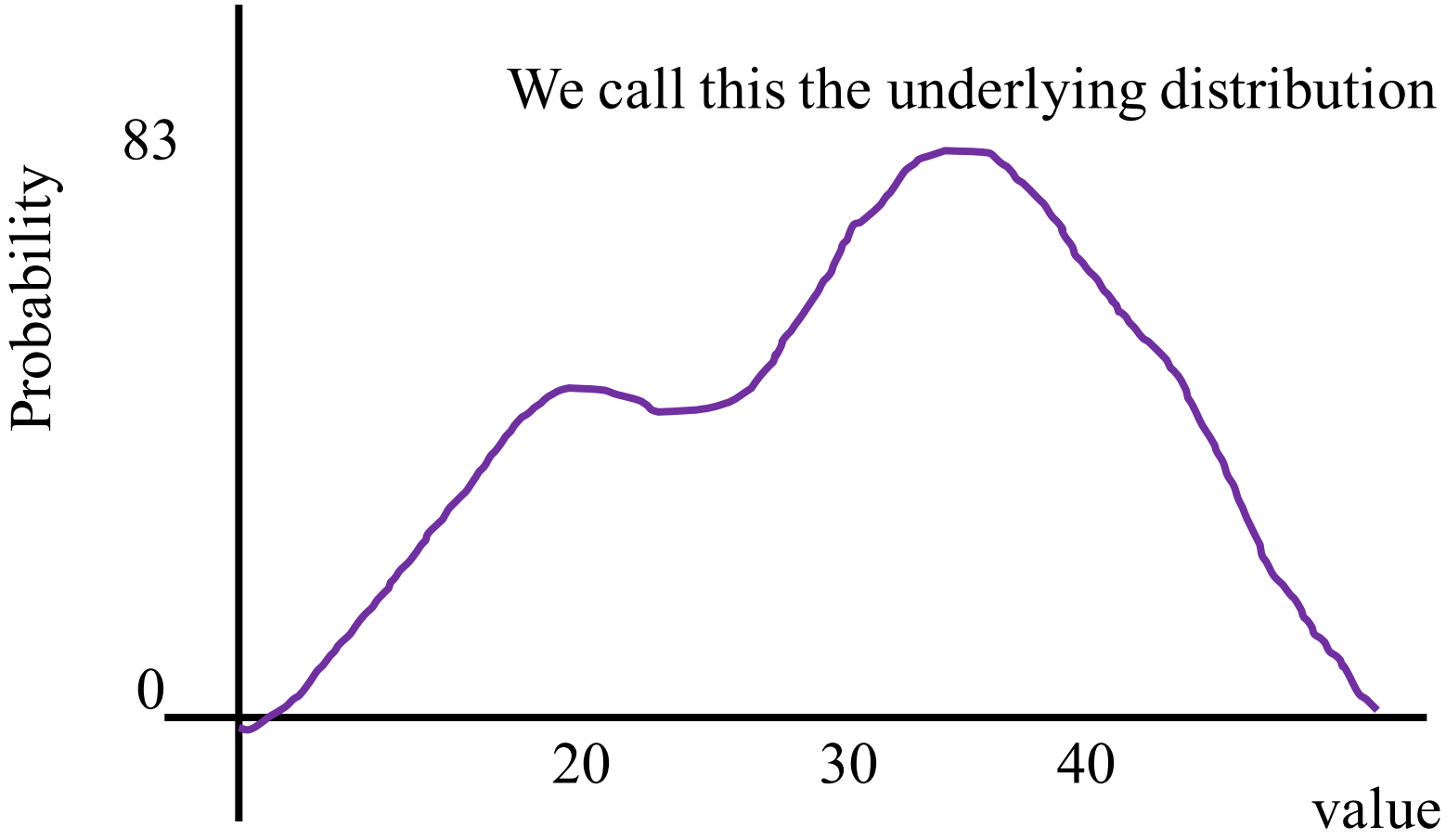
Collect one (or more) numbers from each person

Sample

Sample Mean

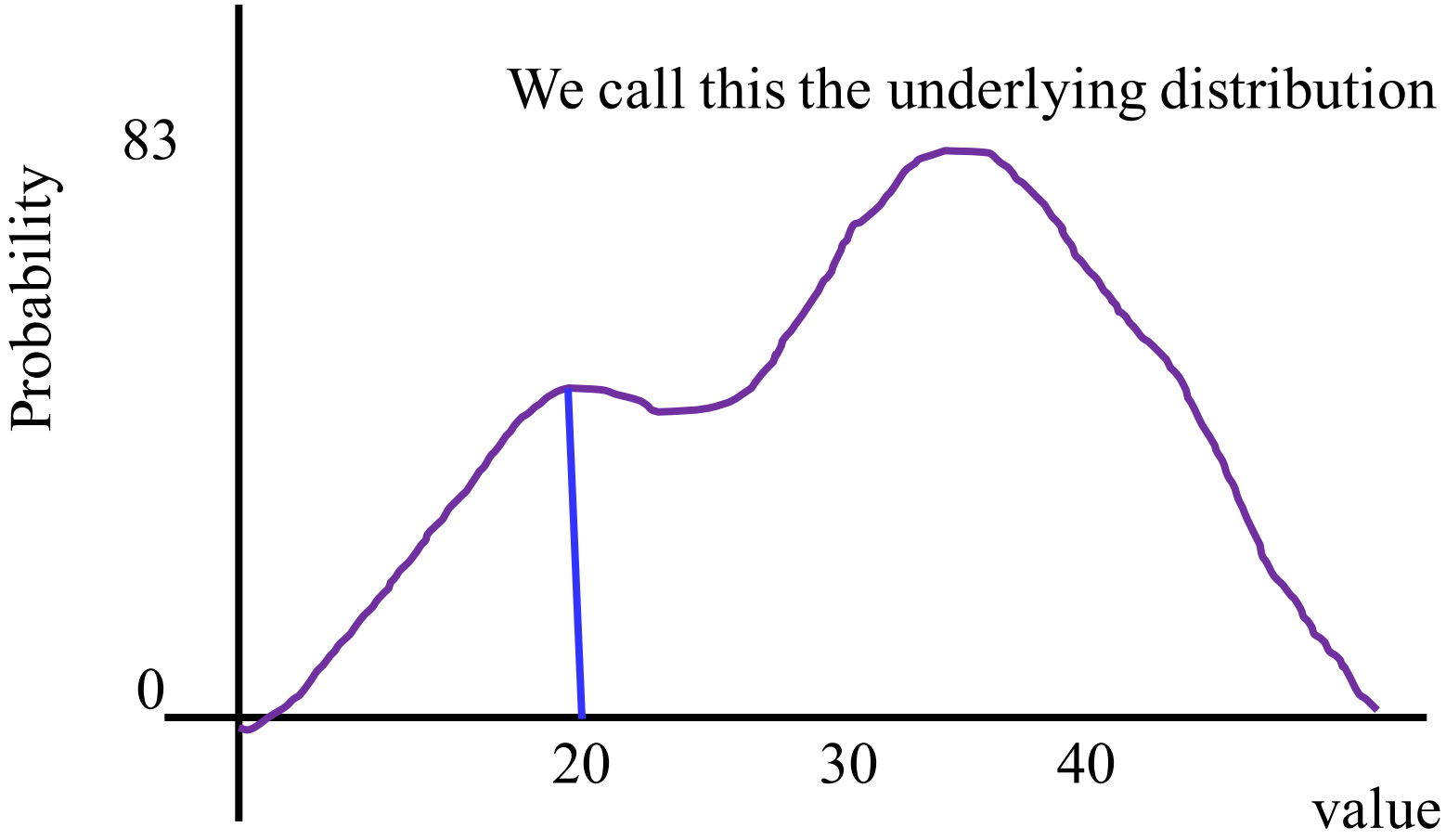
- Consider *n* random variables $X_1, X_2, \ldots X_n$
  - X_i are all independently and identically distributed (I.I.D.)
  - Have same distribution function *F* and $E[X_i] = \mu$
  - We call sequence of X_i a **sample** from distribution *F*

IID Samples



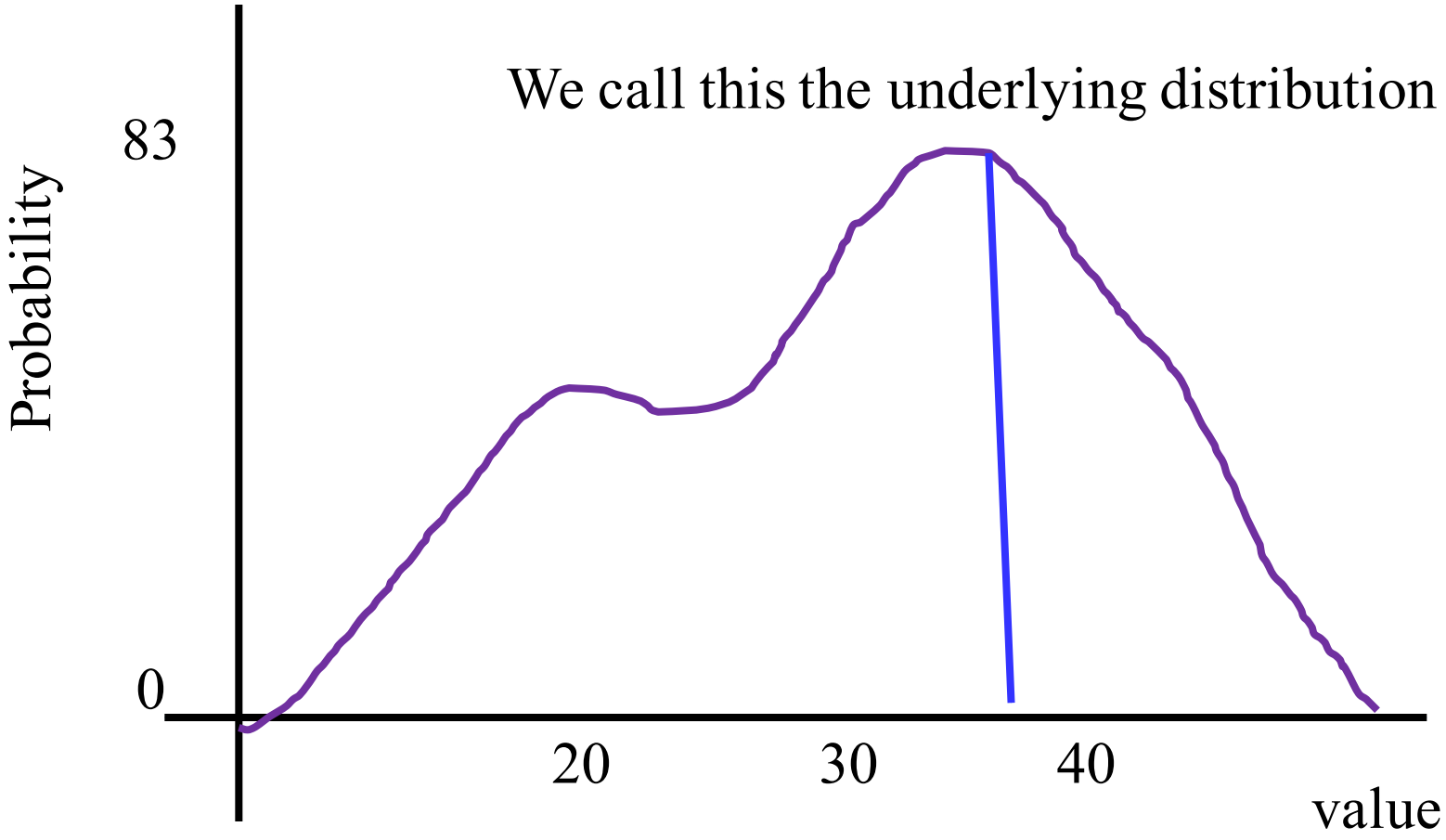
IID Samples = []

IID Samples



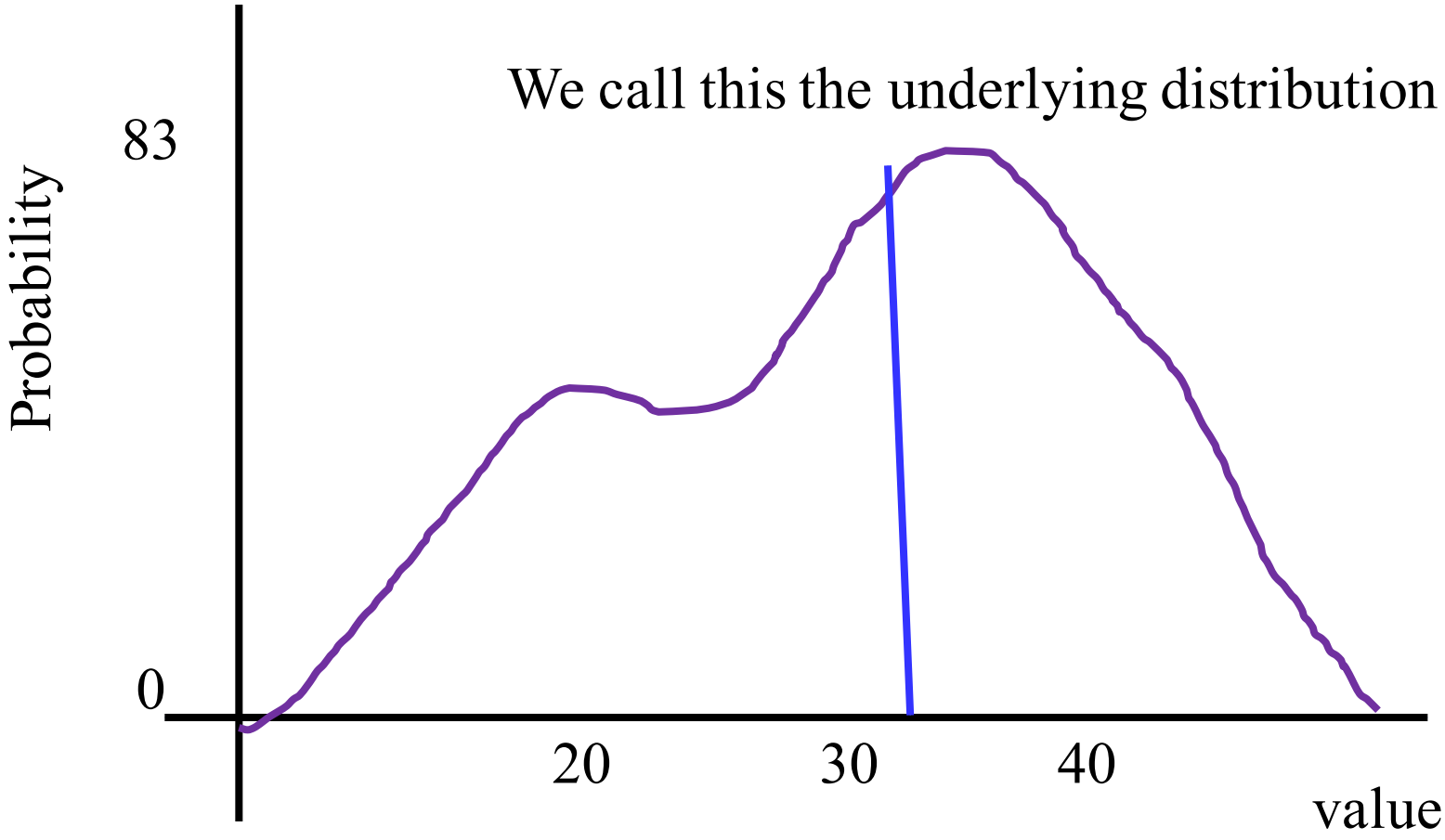
IID Samples = [20]

IID Samples



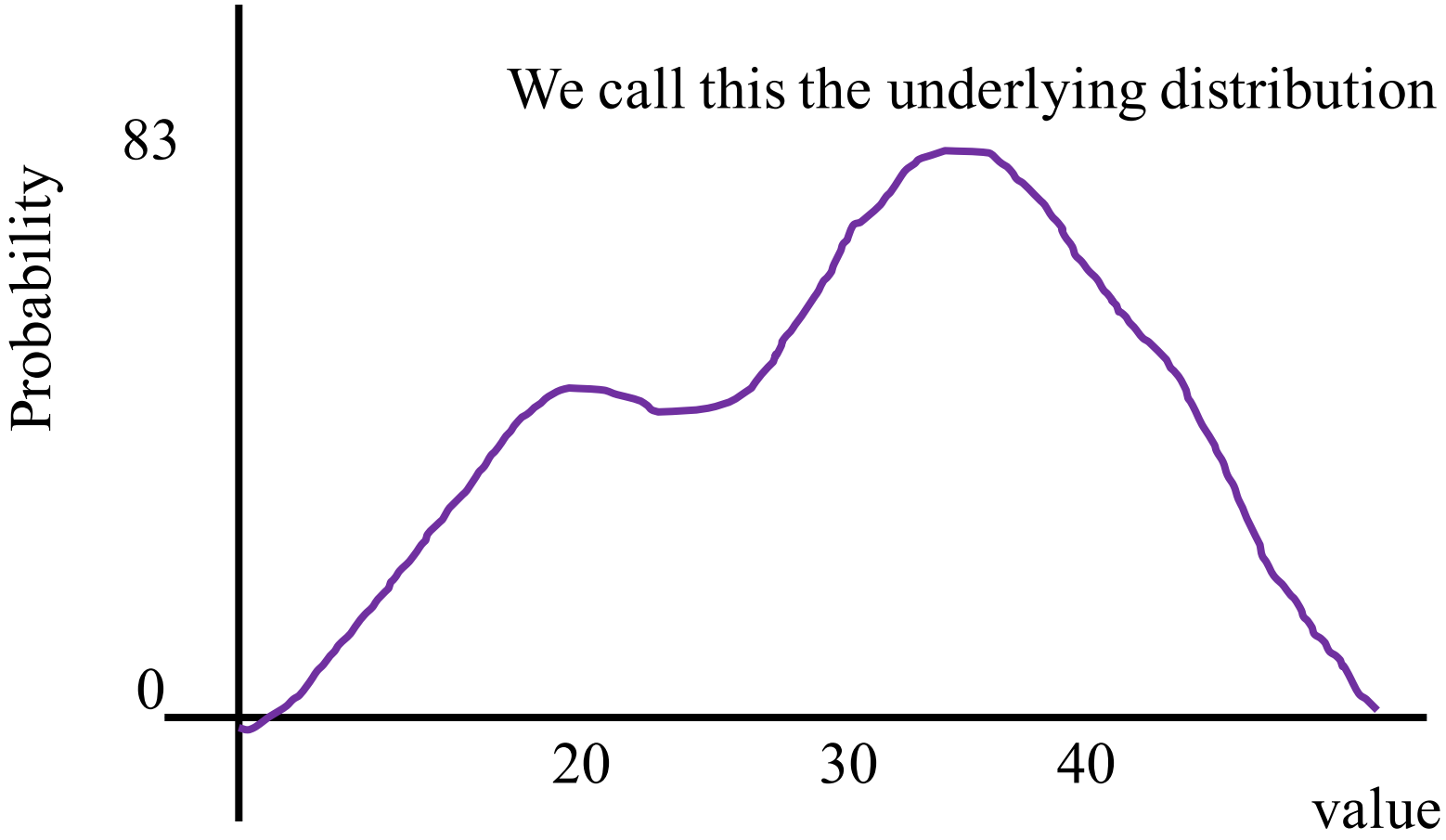
IID Samples = [20, 38]

IID Samples



IID Samples = [20, 38, 32]

IID Samples



IID Samples = [20, 38, 32, …, 38]

X_1 X_2 X_n

Sample Mean

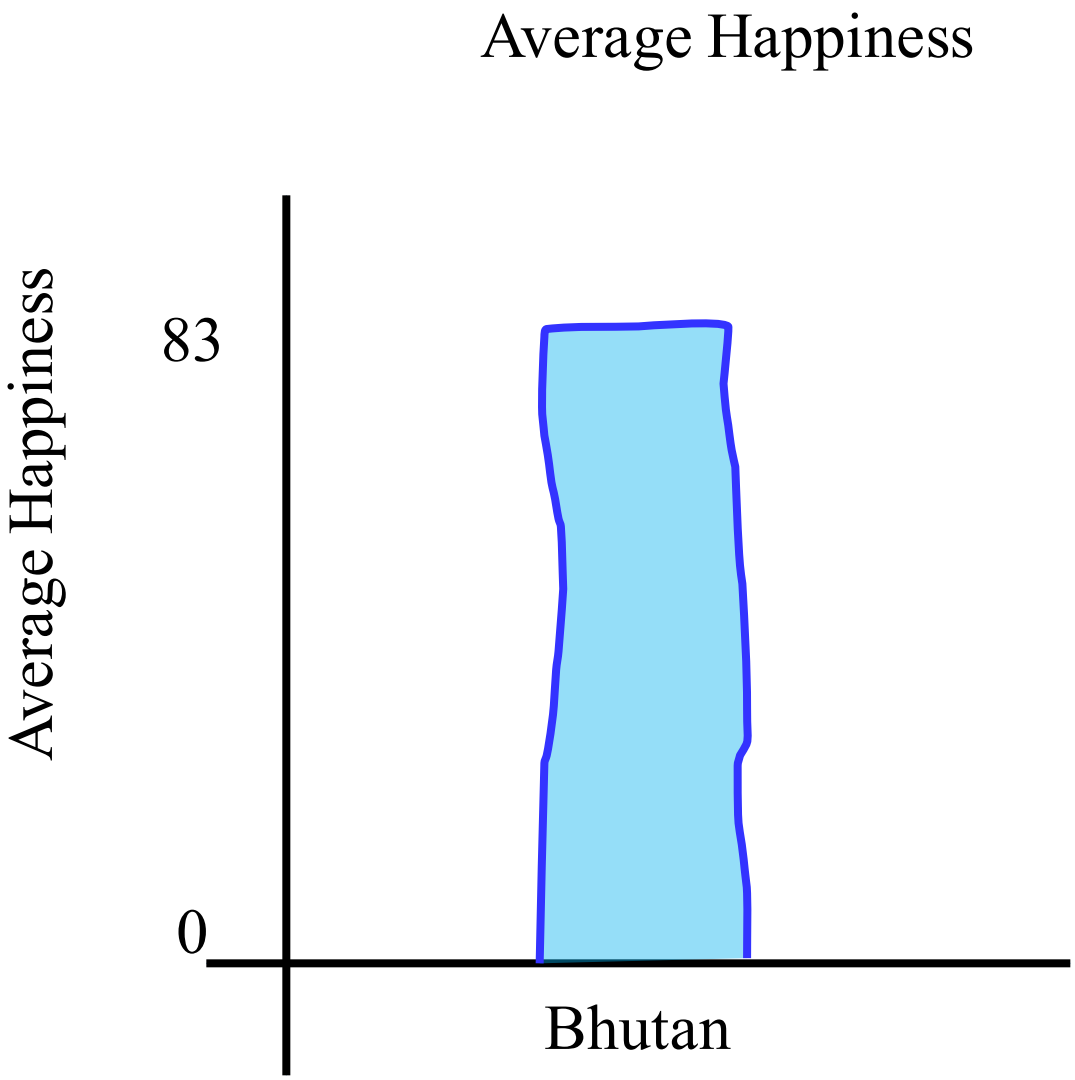
- Consider *n* random variables $X_1, X_2, \ldots X_n$
  - X_i are all independently and identically distributed (I.I.D.)
  - Have same distribution function *F* and $E[X_i] = \mu$
  - We call sequence of X_i a **sample** from distribution *F*
  - Sample mean: $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$
  - Compute $E[\overline{X}]$

$$E[\overline{X}] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

  - $\overline{X}$ is "unbiased" estimate of μ $(E[\overline{X}] = \mu)$

Sample Mean



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

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

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

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

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

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

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

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

$$= n\sigma^2 - n\mathrm{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

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

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

Population variance

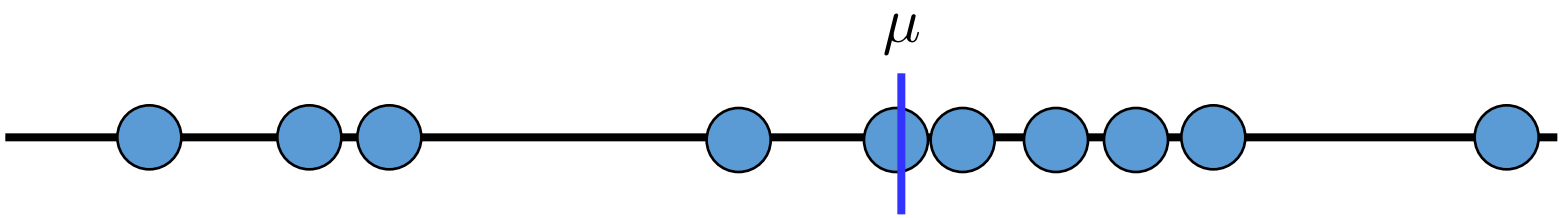
$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

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



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

Population variance

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

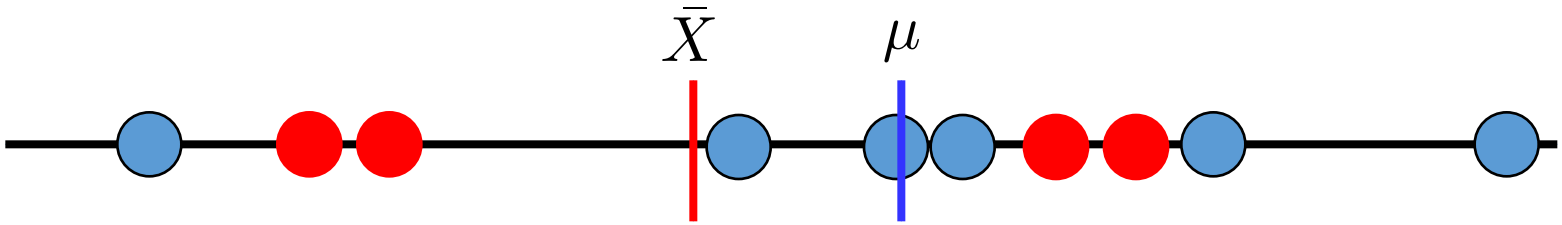
This is the actual mean

Unbiased sample variance

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

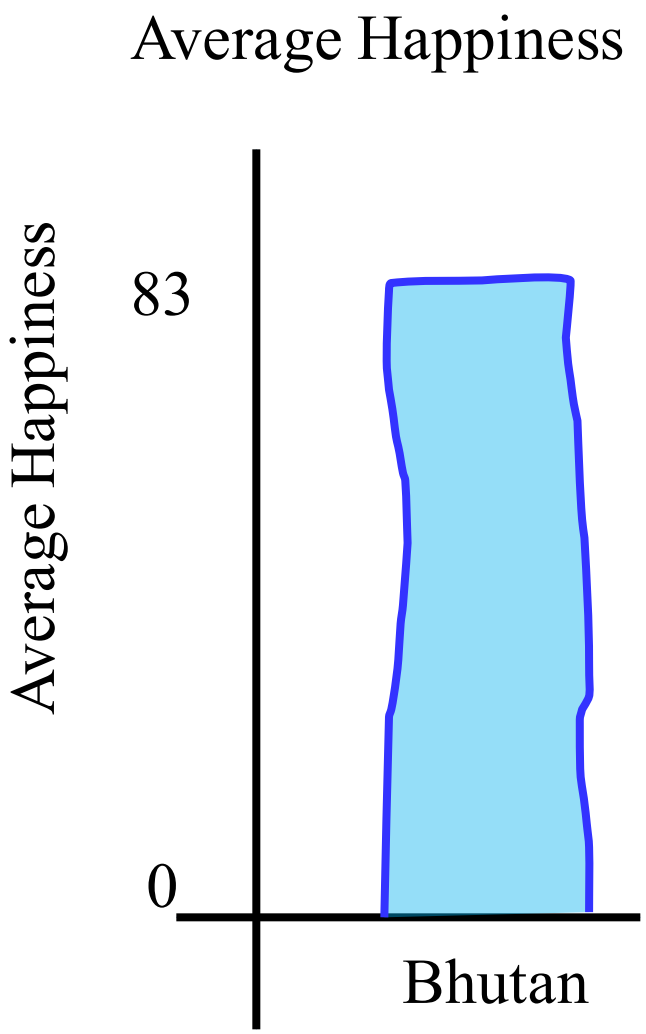
This is the sample mean

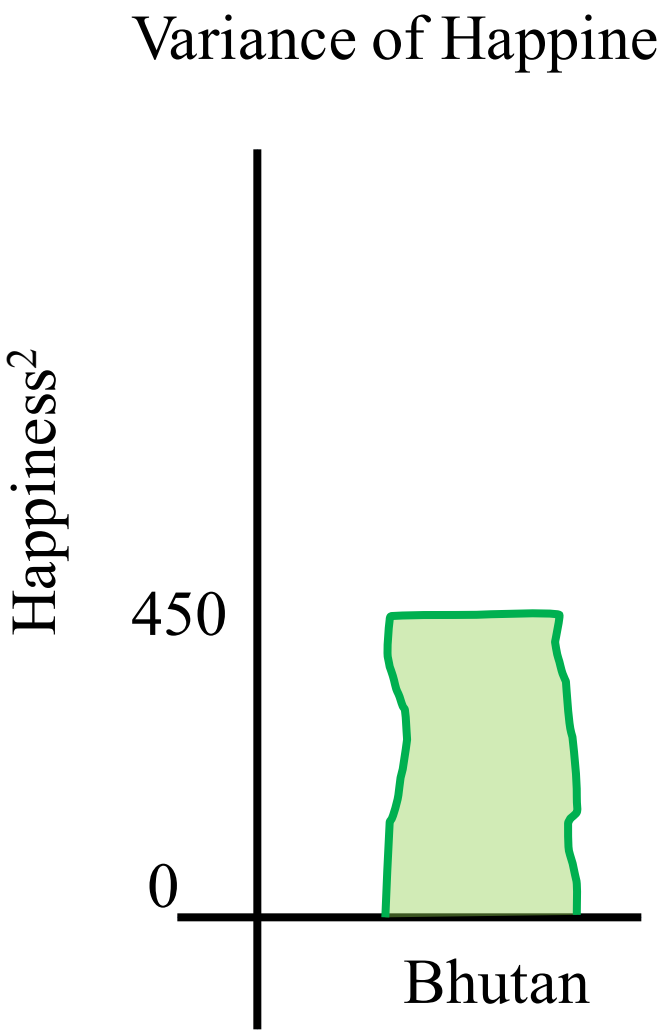
The variance of the sample mean? Related to population variance



I Believe What I See

Sample Mean


Average Happiness
Average Happiness
83
0
Bhutan


Variance of Happiness
Happiness^2
450
0
Bhutan

Sample Variance:

Sample mean

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

Makes it "unbiased"

No Error Bars ☹

Variance of Sample Mean

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

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

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

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

Standard Error of the Mean

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

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

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

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

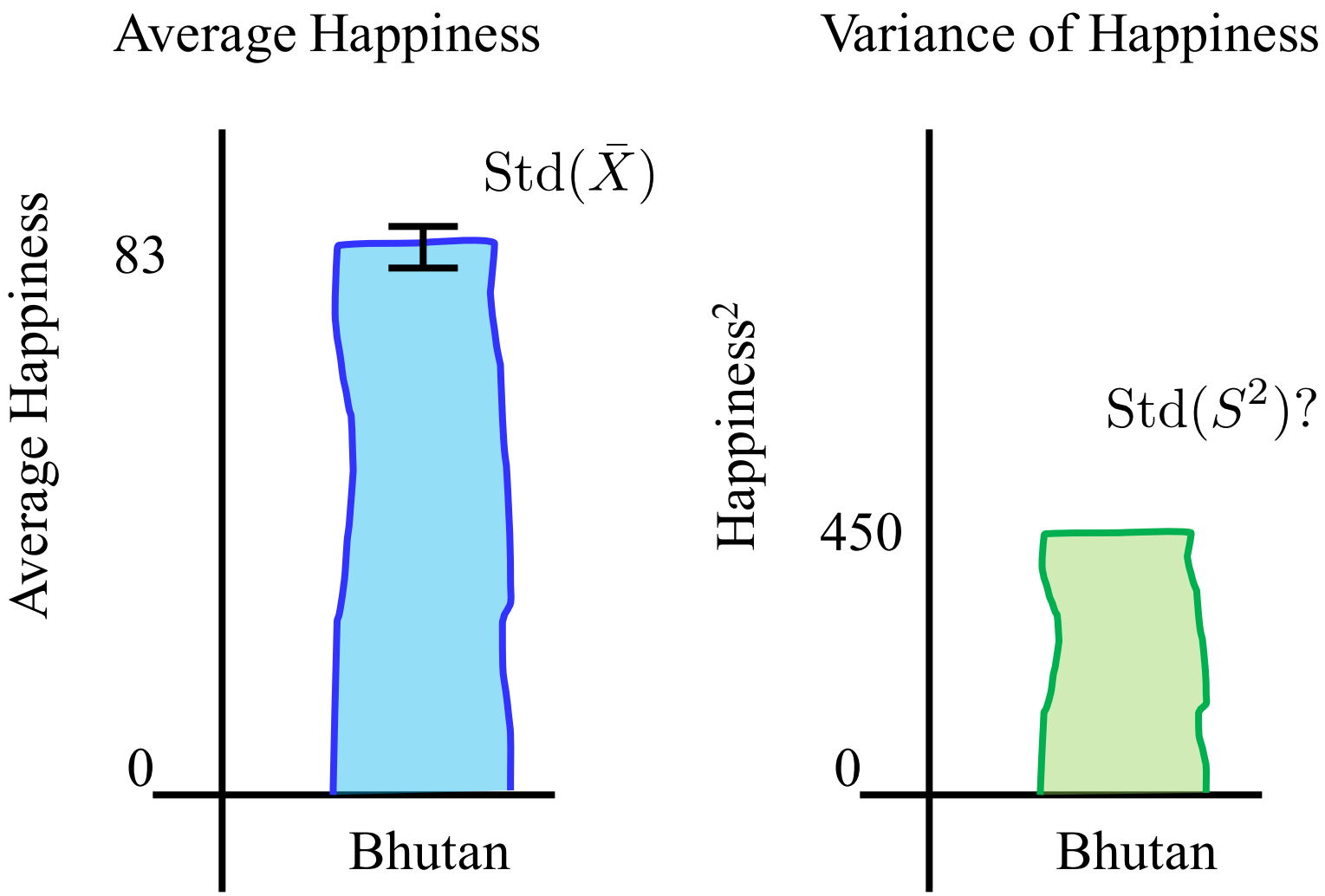
$$= \sqrt{\frac{450}{200}}$$

The numbers for our Bhutanese poll

$$= 1.5$$

Bhutanese standard error of the mean

Sample Mean



Claim: The average happiness of Bhutan is 83 ± 2

Bootstrap:
Probability for Computer Scientists

Bootstrap

Bootstraping 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

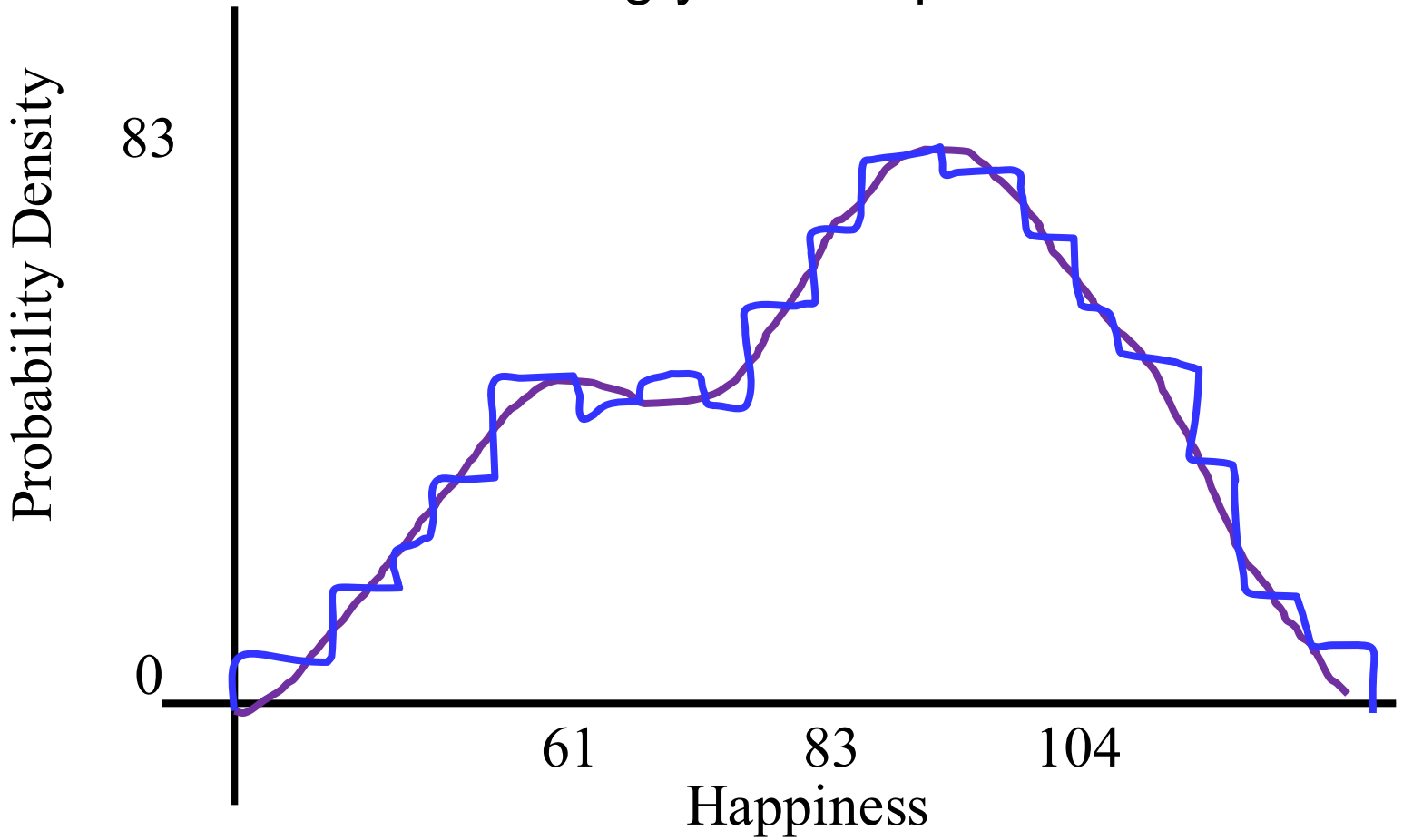
What is the probability that the mean of a sample of 200 people is within the range 81 to 85?

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.



Bootstrapping Assumption

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

The underlying distribution

The sample distribution

Algorithm

Bootstrap Algorithm (sample):

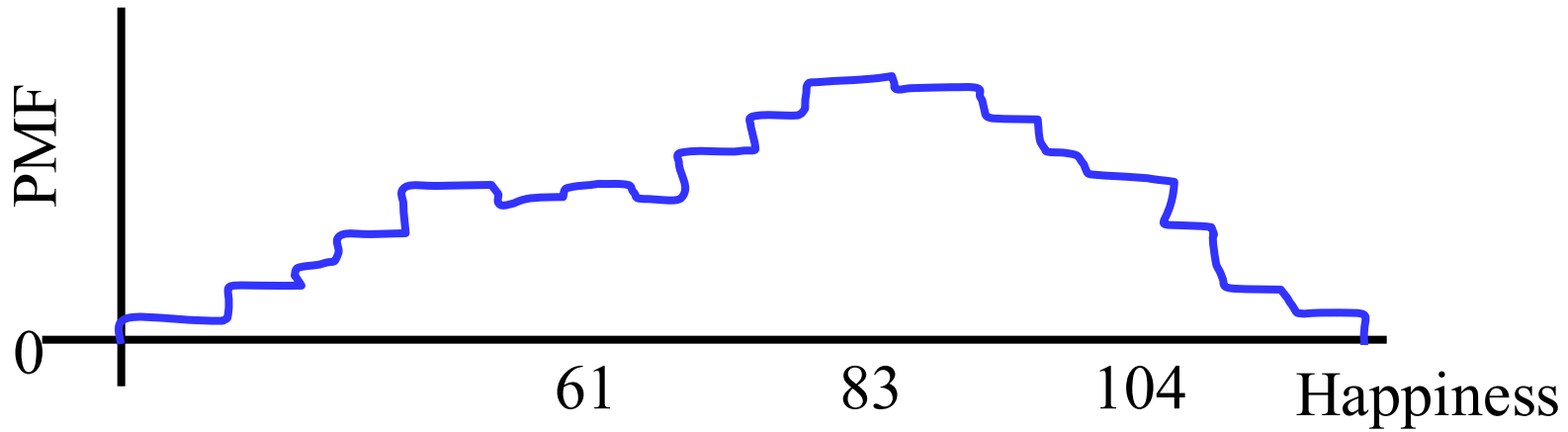
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Resample **sample.size()** from PMF
   b. Recalculate the stat on the resample
3. You now have a **distribution of your stat**

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**

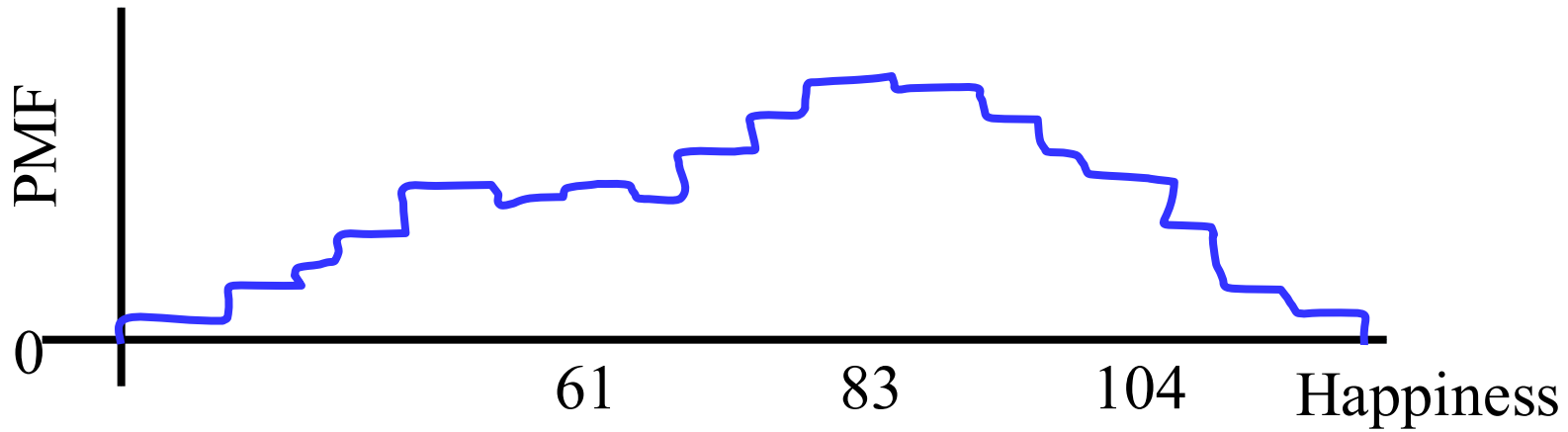
Bootstrap of Means



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**

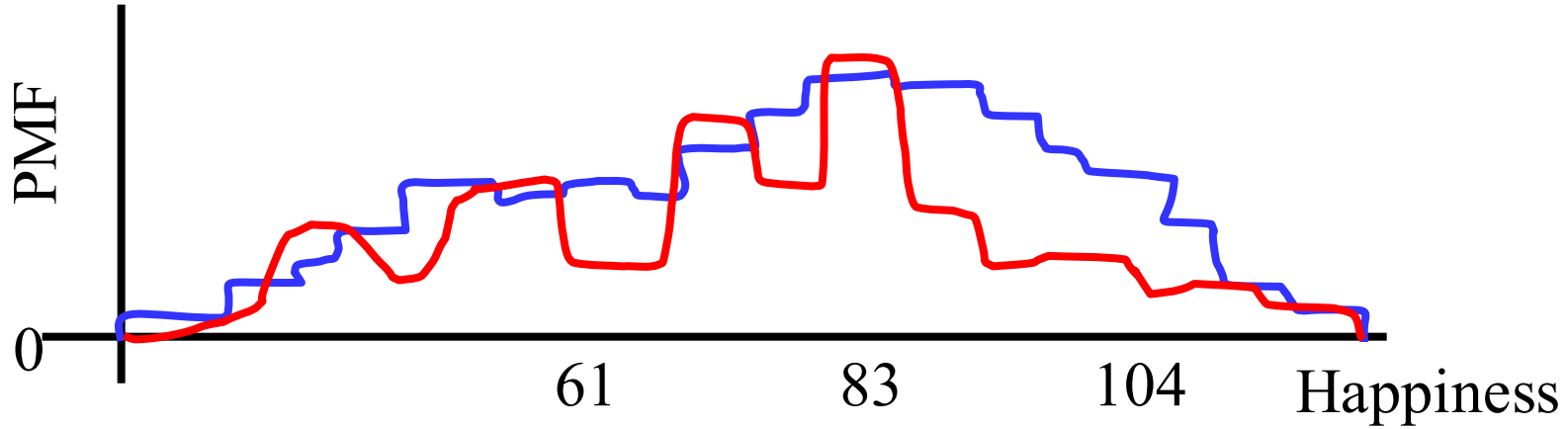
Bootstrap of Means



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**

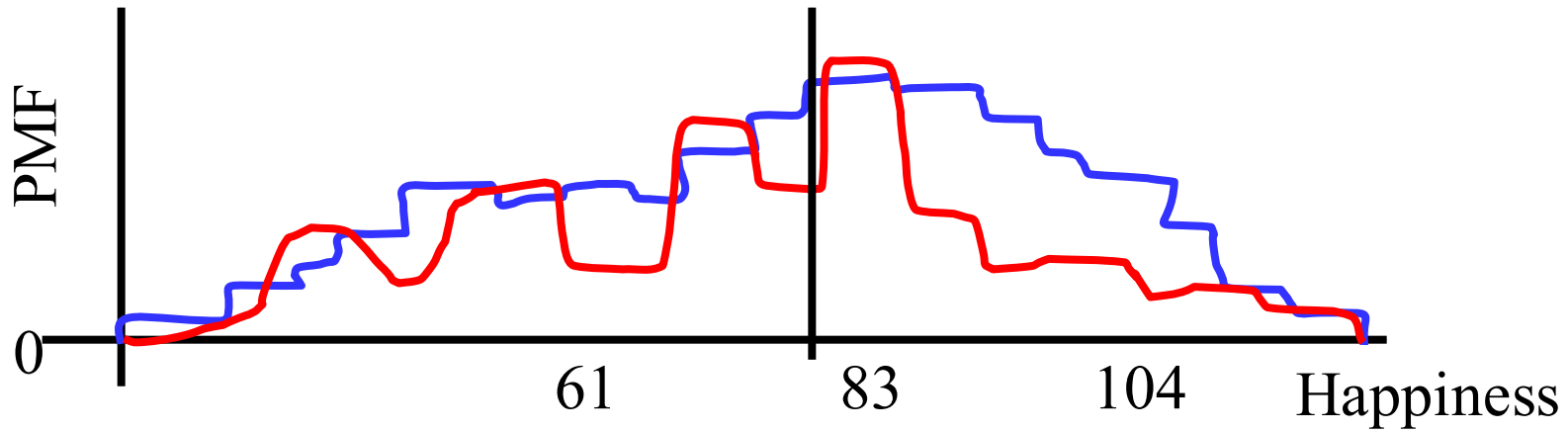
Bootstrap of Means



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

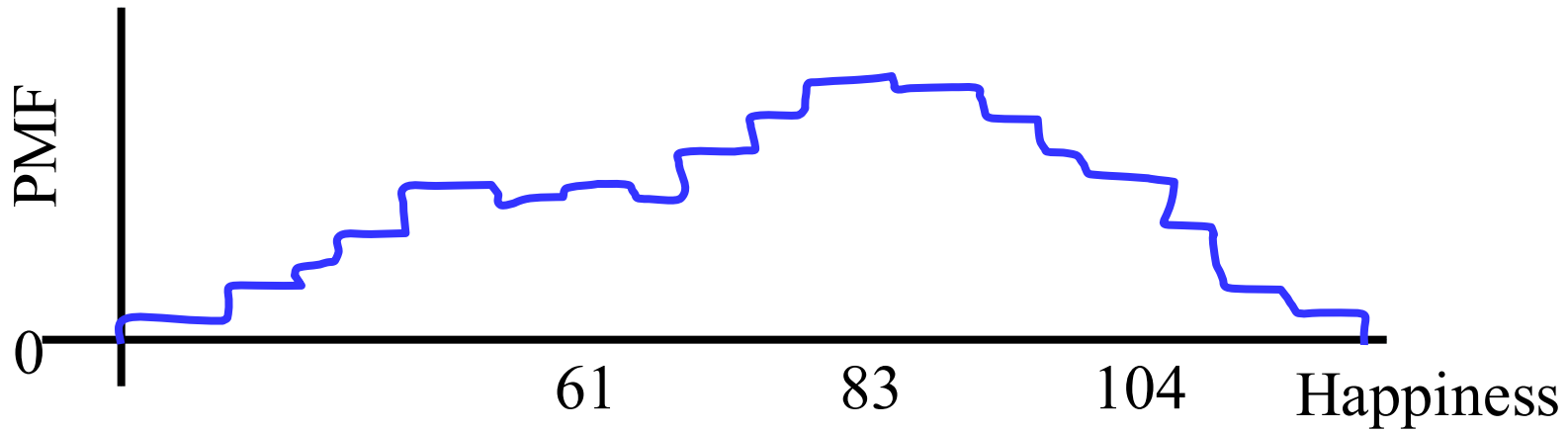


Bootstrap Algorithm (sample):

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

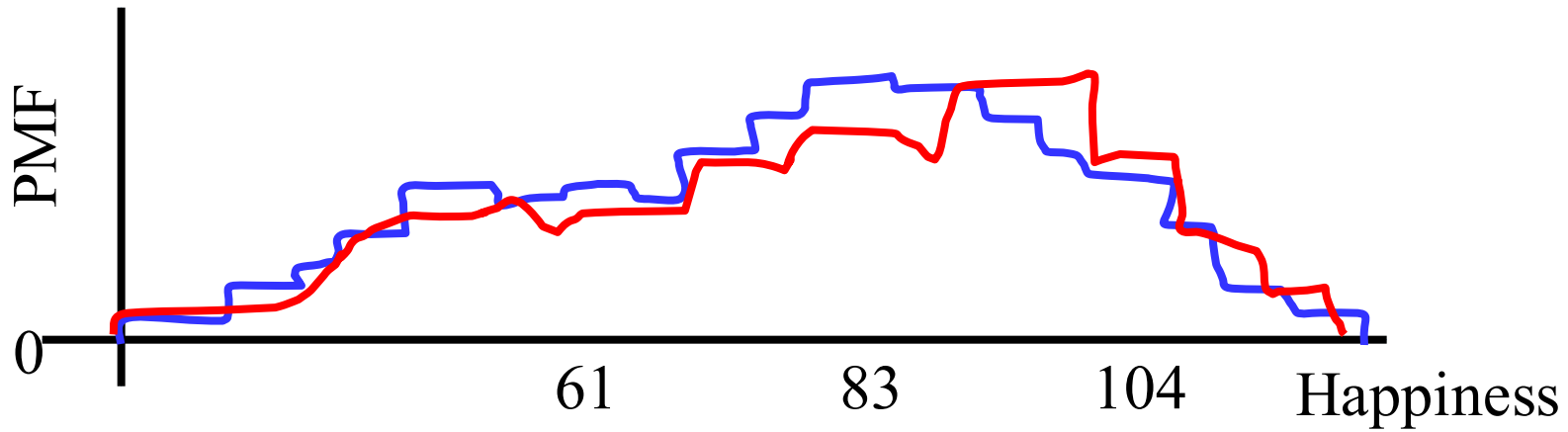


Bootstrap Algorithm (sample):

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

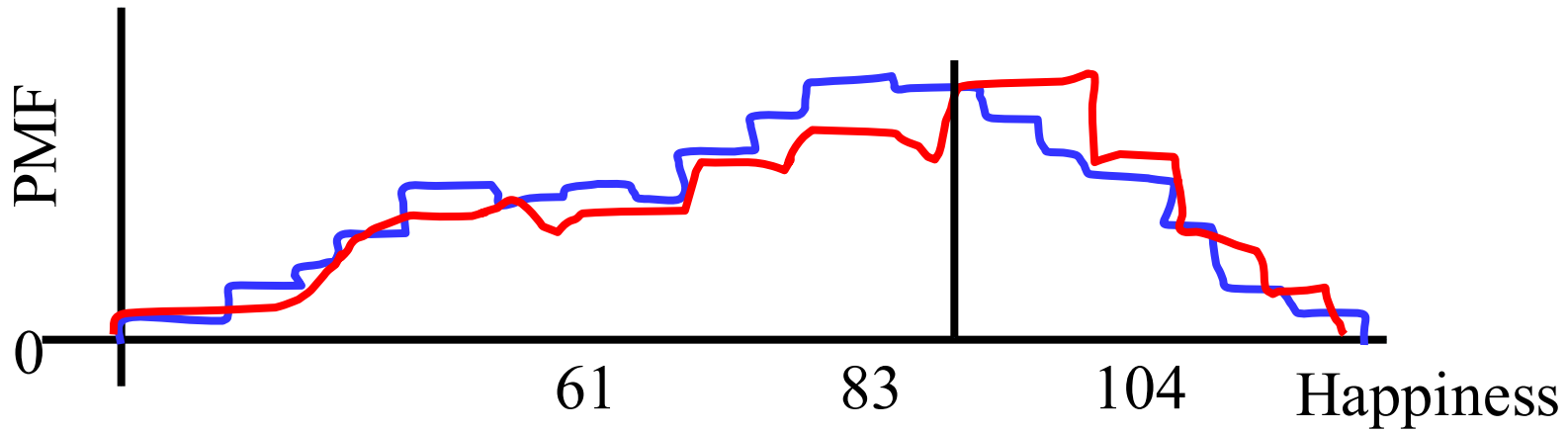


Bootstrap Algorithm (sample):

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

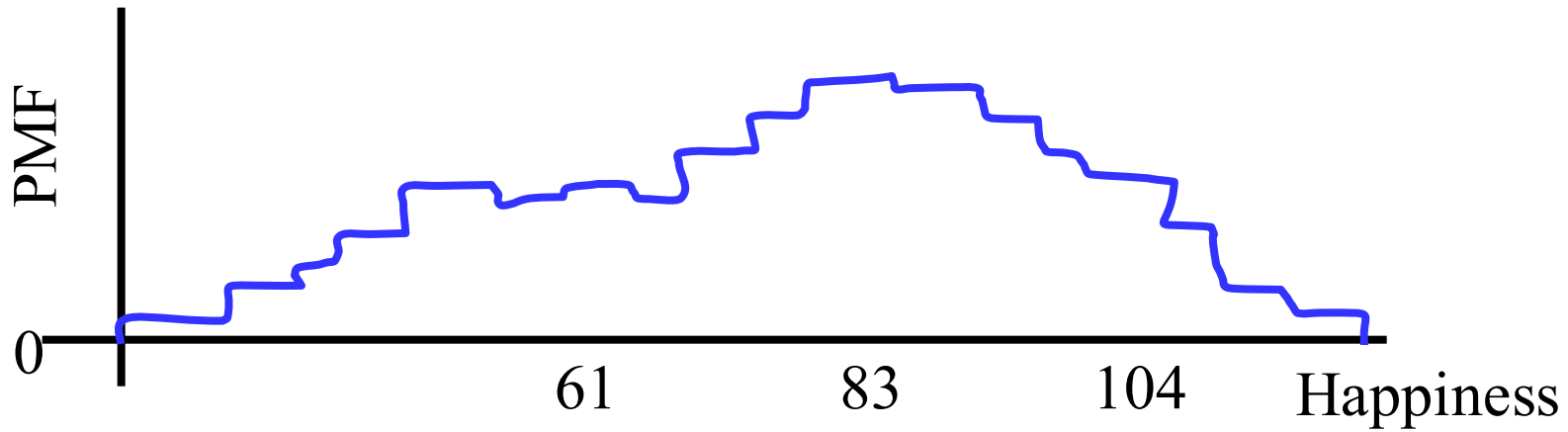


Bootstrap Algorithm (sample):

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

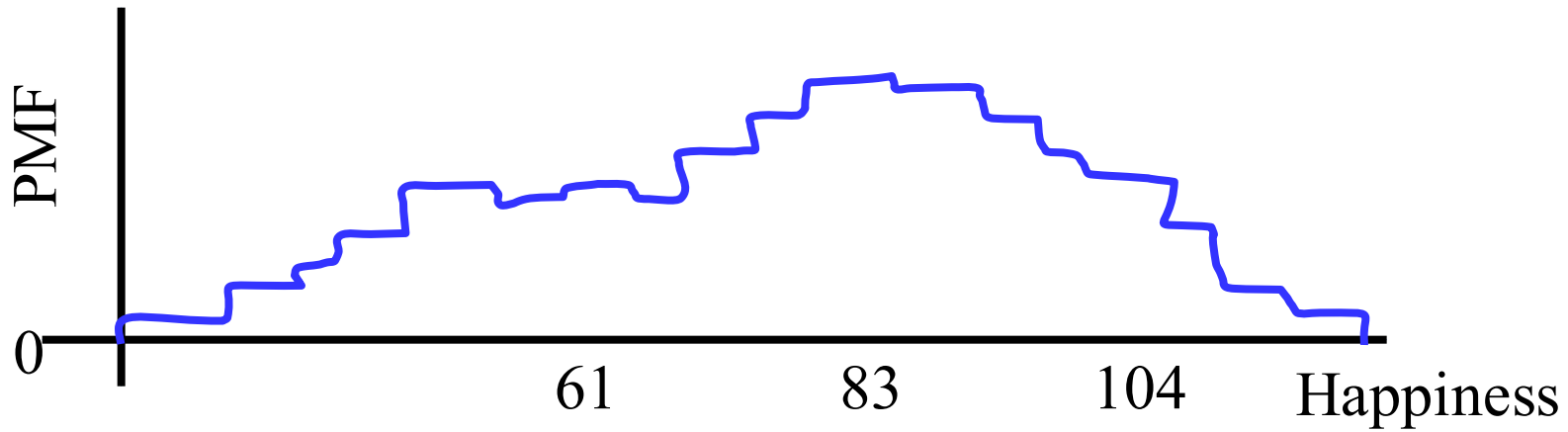


Bootstrap Algorithm (sample):

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



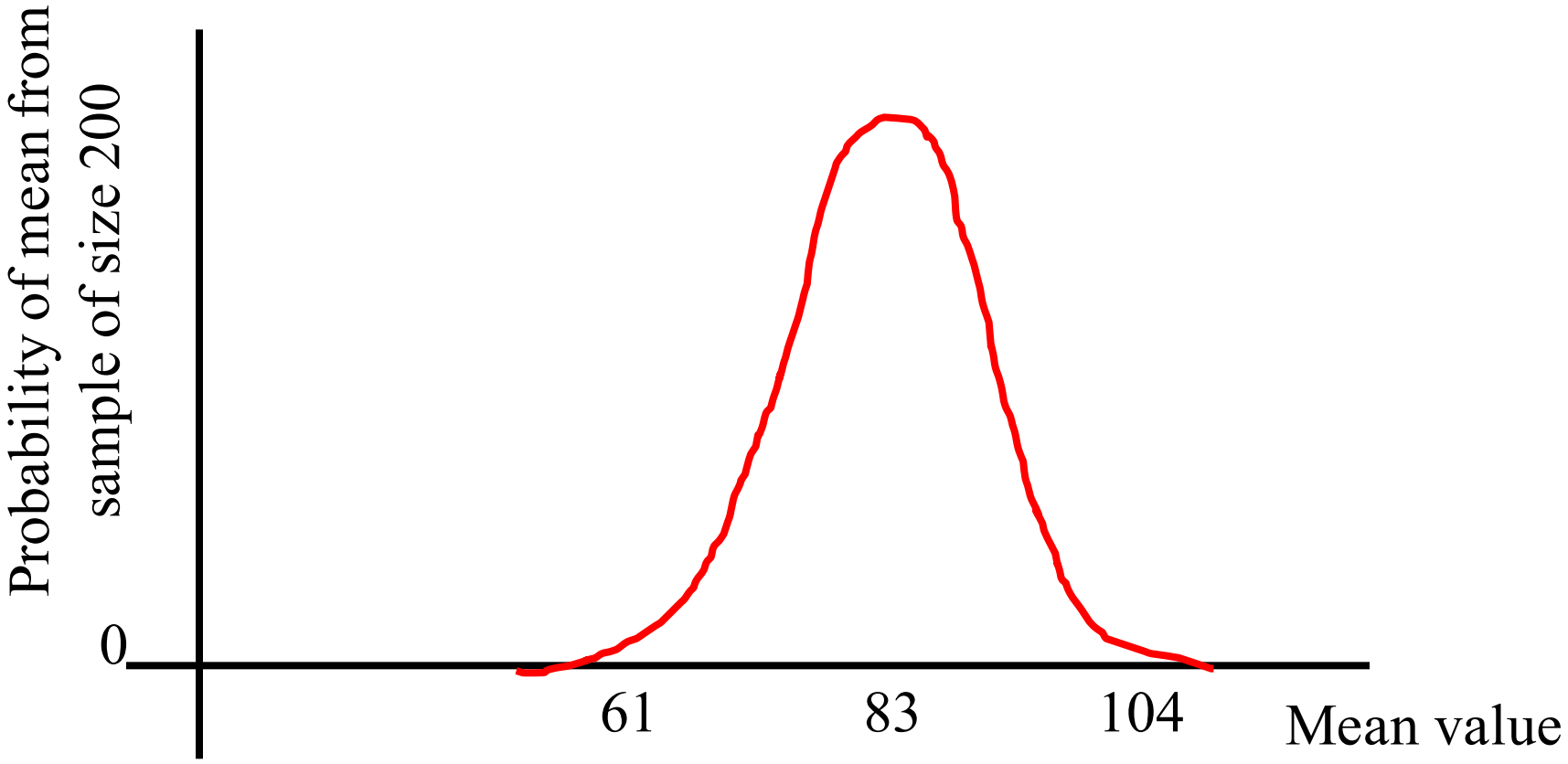
Bootstrap Algorithm (sample):

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]

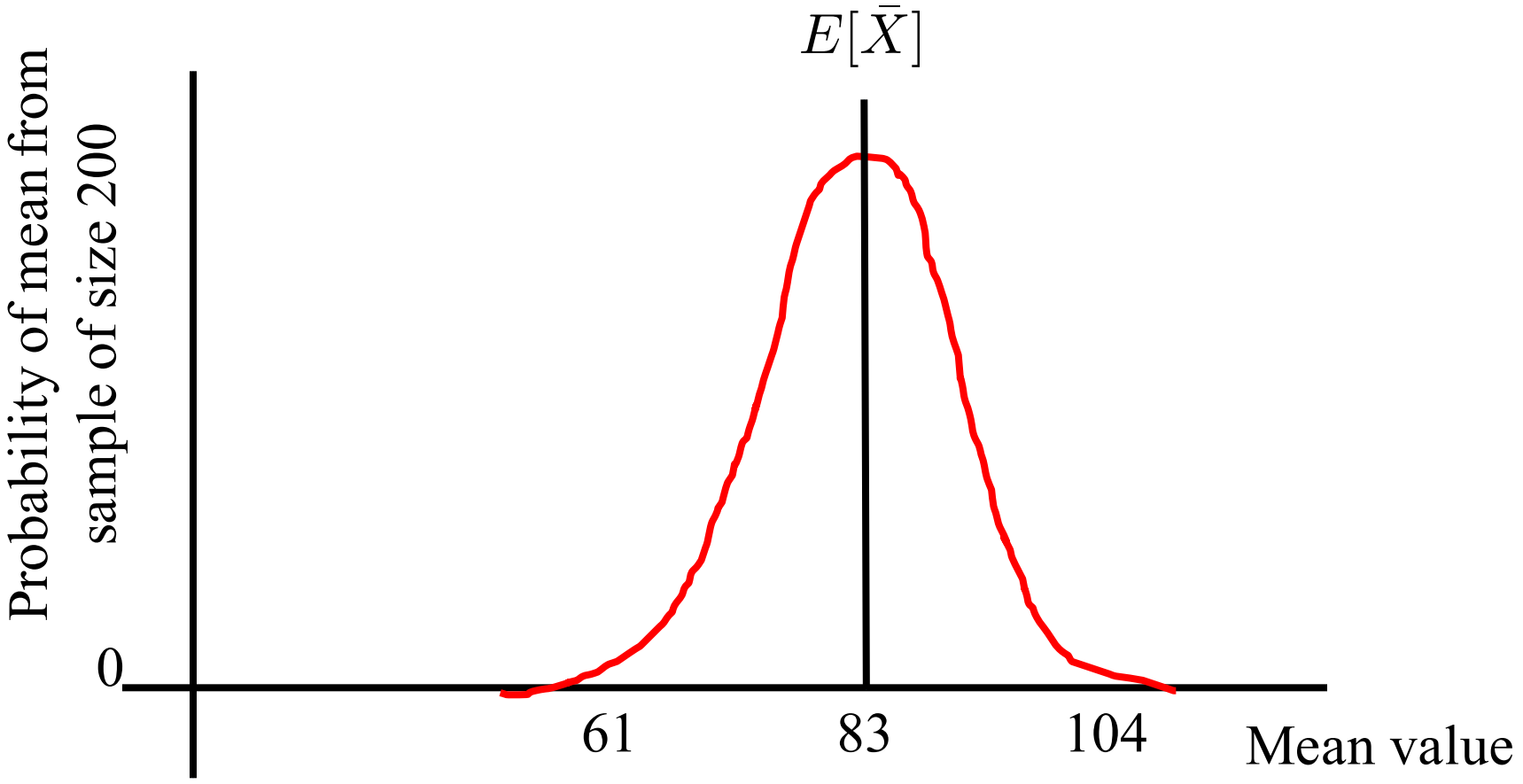
Bootstrap of Means

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



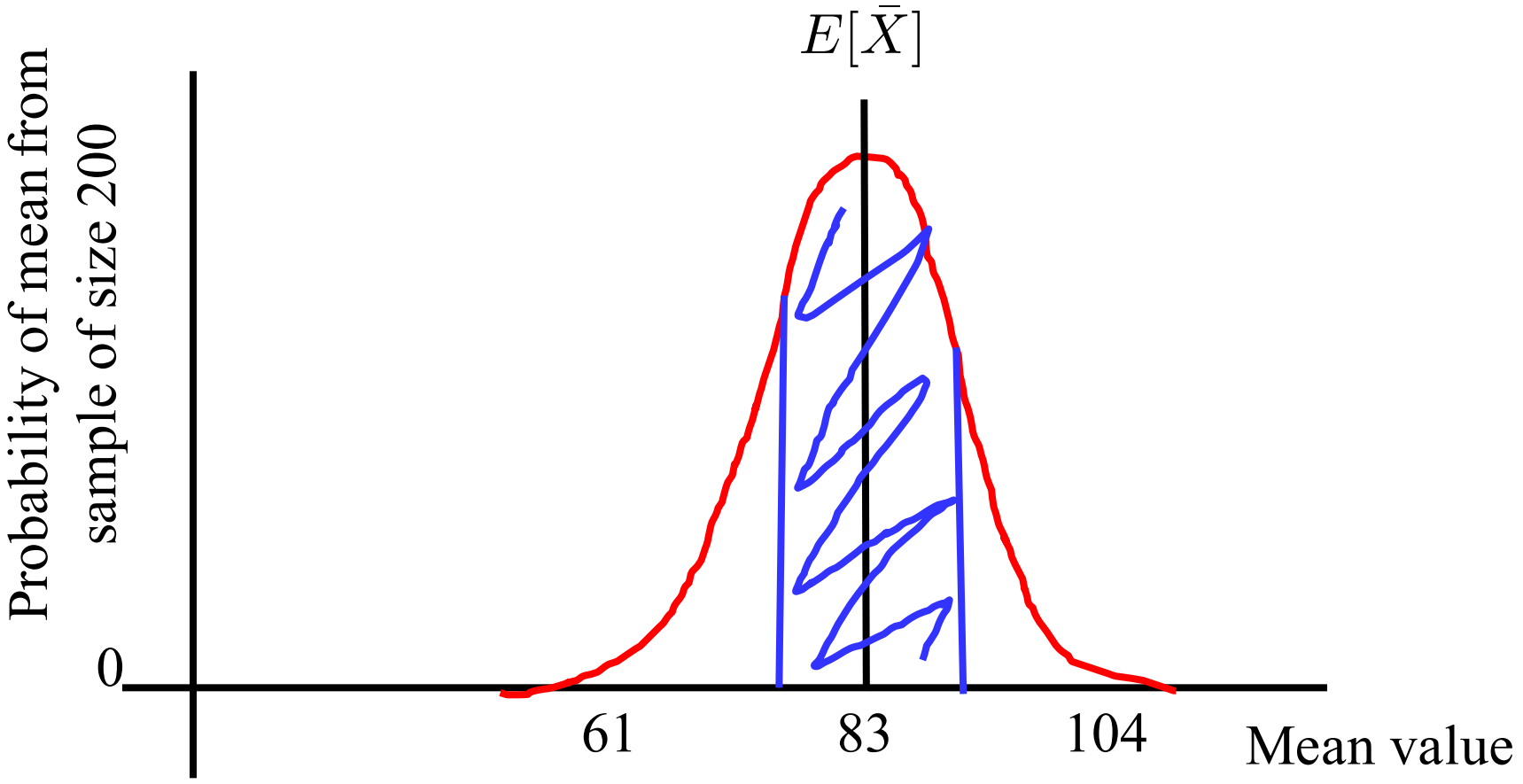
Bootstrap of Means

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?



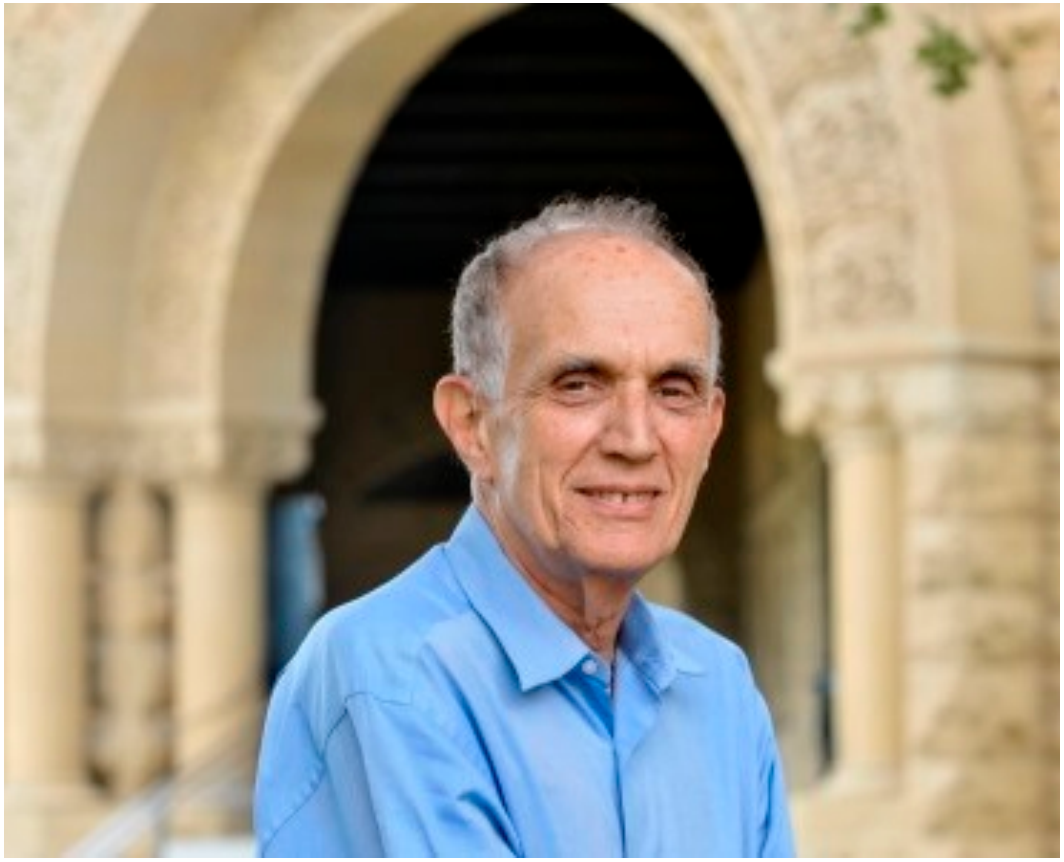
Algorithm in Practice

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**

Bootstrap provides a way to calculate probabilities using programs.

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