Sampling and Bootstrapping

Noah Arthurs
CS109, Stanford University

Announcements

- PS3 grades not out until Monday (solutions out today though)
- PS4 due Monday
- PS5 out today, covers up through Monday’s lecture.

Interpreting Midterm Score

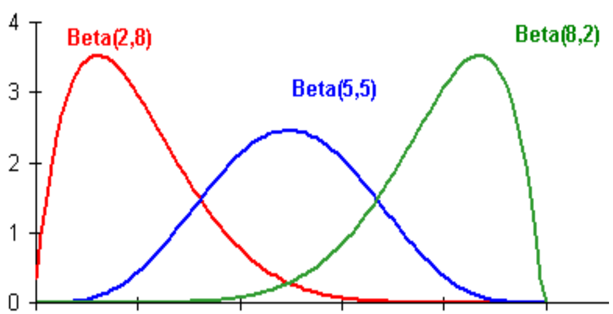
100-120: rock on, iron out details
80-100: keep practicing
< 80: work on fundamentals

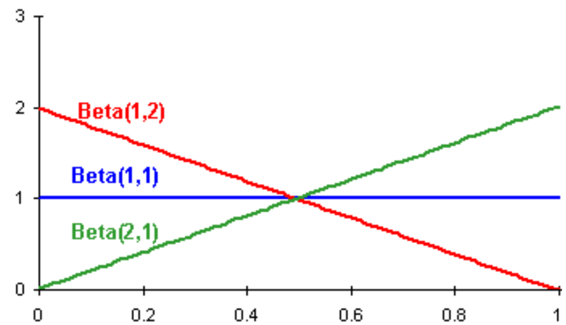
Review

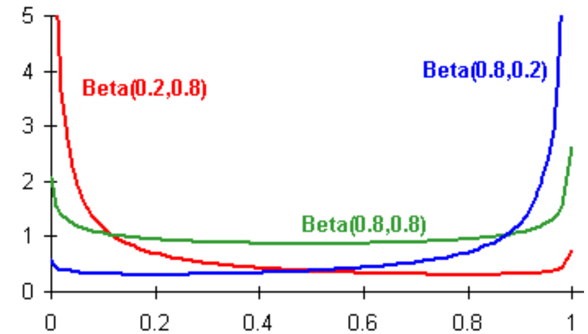
Beta Random Variable

- X is a **Beta Random Variable**: X ~ Beta(*a*, *b*)
  - Probability Density Function (PDF): (where *a*, *b* > 0)

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$







  - Symmetric when *a* = *b*
  - $E[X] = \frac{a}{a+b}$ $\quad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Beta is a distribution for probabilities

Beta Parameters *can* come from experiments:

a = "successes" + 1

b = "failures" + 1

If the Prior was a Beta...

X is our random variable for probability

If our **prior belief** about X was beta

$$f(X = x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

What is our **posterior belief** about X after observing *n* heads (and *m* tails)?

$$f(X = x | N = n) = ???$$

If the Prior was a Beta…

$$
\begin{aligned}
f(X = x|N = n) &= \frac{P(N = n|X = x)f(X = x)}{P(N = n)} \\
&= \frac{\binom{n+m}{n} x^n (1-x)^m \, f(X = x)}{P(N = n)} \\
&= \frac{\binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}}{P(N = n)} \\
&= K_1 \cdot \binom{n+m}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\
&= K_3 \cdot x^n (1-x)^m x^{a-1}(1-x)^{b-1} \\
&= K_3 \cdot x^{n+a-1}(1-x)^{m+b-1}
\end{aligned}
$$

$$X|N \sim \text{Beta}(n+a, m+b)$$

Understanding Beta

- If “Prior” distribution of X (before seeing flips) is Beta

- Then “Posterior” distribution of X (after flips) is Beta

- Beta is a **<u>conjugate</u>** distribution for Beta
  - Prior and posterior parametric forms are the same!
  - Practically, conjugate means easy update:
    - Add number of “heads” and “tails” seen to Beta parameters

Beta Example

Before being tested, a medicine is believed to “work” about 80% of the time. The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

Frequentist:

$$p \approx \frac{14}{20} = 0.7$$

Beta Example

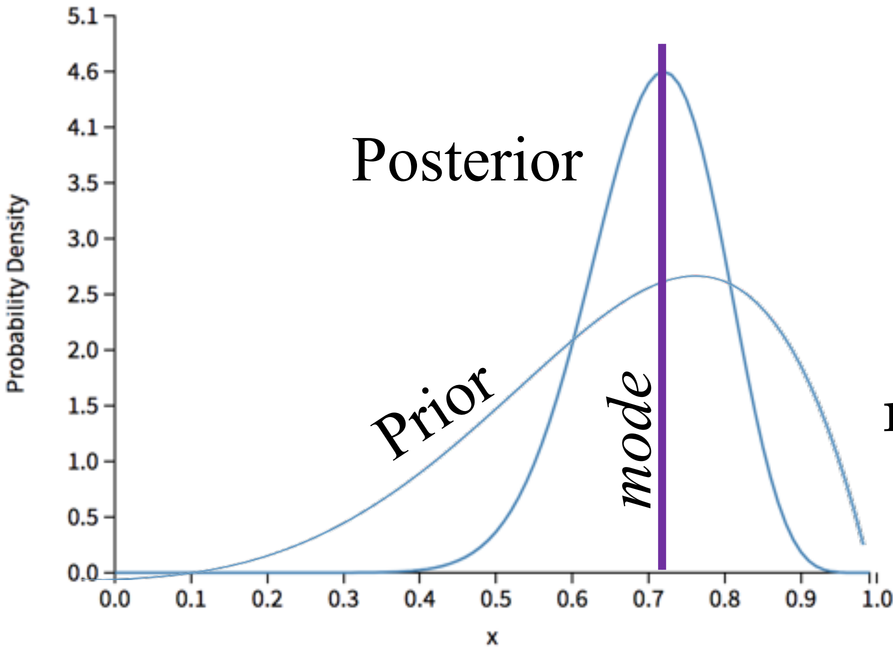
Before being tested, a medicine is believed to "work" about 80% of the time. The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

$$\text{Bayesian:} \quad X \sim \text{Beta}$$

$$\text{Prior:} \quad X \sim \text{Beta}(a = 5, b = 2)$$

$$\text{Posterior:} \quad X \sim \text{Beta}(a = 5 + 14, b = 2 + 6)$$

$$\sim \text{Beta}(a = 19, b = 8)$$



$$E[X] = \frac{a}{a+b} = \frac{19}{19+8} \approx 0.70$$

$$\text{mode}(X) = \frac{a-1}{a+b-2}$$

$$= \frac{19}{18+7} \approx 0.72$$

Multi Armed Bandit

Drug A

Drug B

Which one do you give to a patient?

Optimal Decision Making

You try drug B, 5 times. It is successful 2 times.
X is the probability of success.

$$X \sim \text{Beta}(a = 3, b = 4)$$

What is the probability that $X > 0.6$

$$P(X > 0.6) = 1 - P(X < 0.6) = 1 - F_X(0.6)$$

Wait! What's the Beta CDF??

```
stats.beta.cdf(x, a, b)
```

$$P(X > 0.6) = 1 - F_X(0.6) = 0.1792$$

How can we choose a drug?

We have a big problem! There is a tradeoff between:

Exploration – trying new drugs in order to find out if they work well

Exploitation – giving the drugs that have been proven to be successful.

Thompson Sampling addresses this issue by choosing drugs probabilistically based on your current beliefs (higher chance of choosing a drug the more strongly you believe it works well).

You will learn more on PS5, but check out the Wikipedia article if you're impatient!

As n approaches infinity, The sum of n independent, identically distributed variables:

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

Is normally distributed:

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

---

where $\mu = E[X_i]$

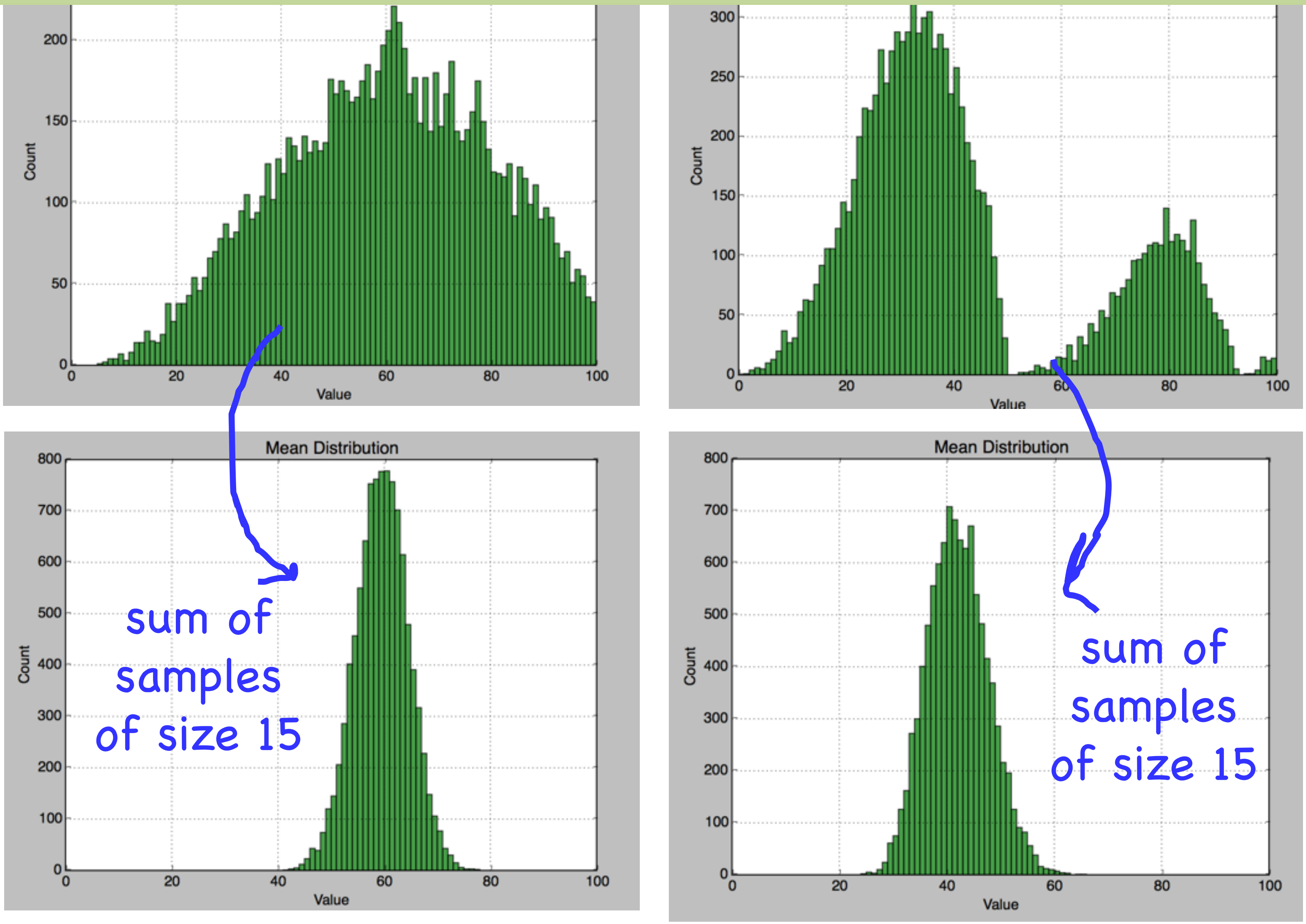
$$\sigma^2 = \mathrm{Var}(X_i)$$

IID Random Variables

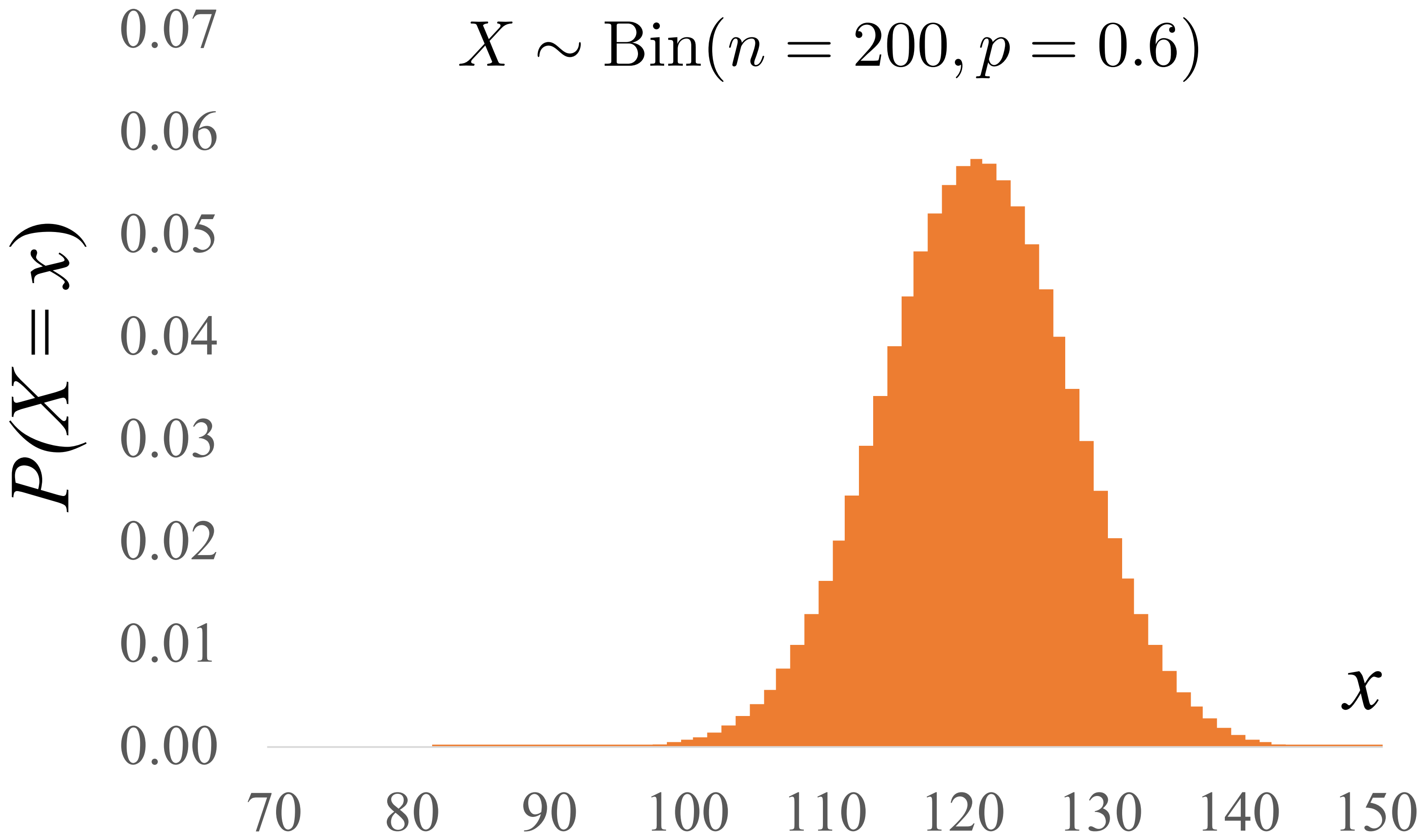
- Consider *n* random variables X_1, X_2, ... X_n
  - X_i are all independently and identically distributed (I.I.D.)
  - All have the same PMF (if discrete) or PDF (if continuous)
  - All have the same expectation
  - All have the same variance

IID iid

C.L.T. Explains This



C.L.T. Explains This



Since n is never actually infinite, the CLT is always an approximation. It’s a very good one though!

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

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

End Review

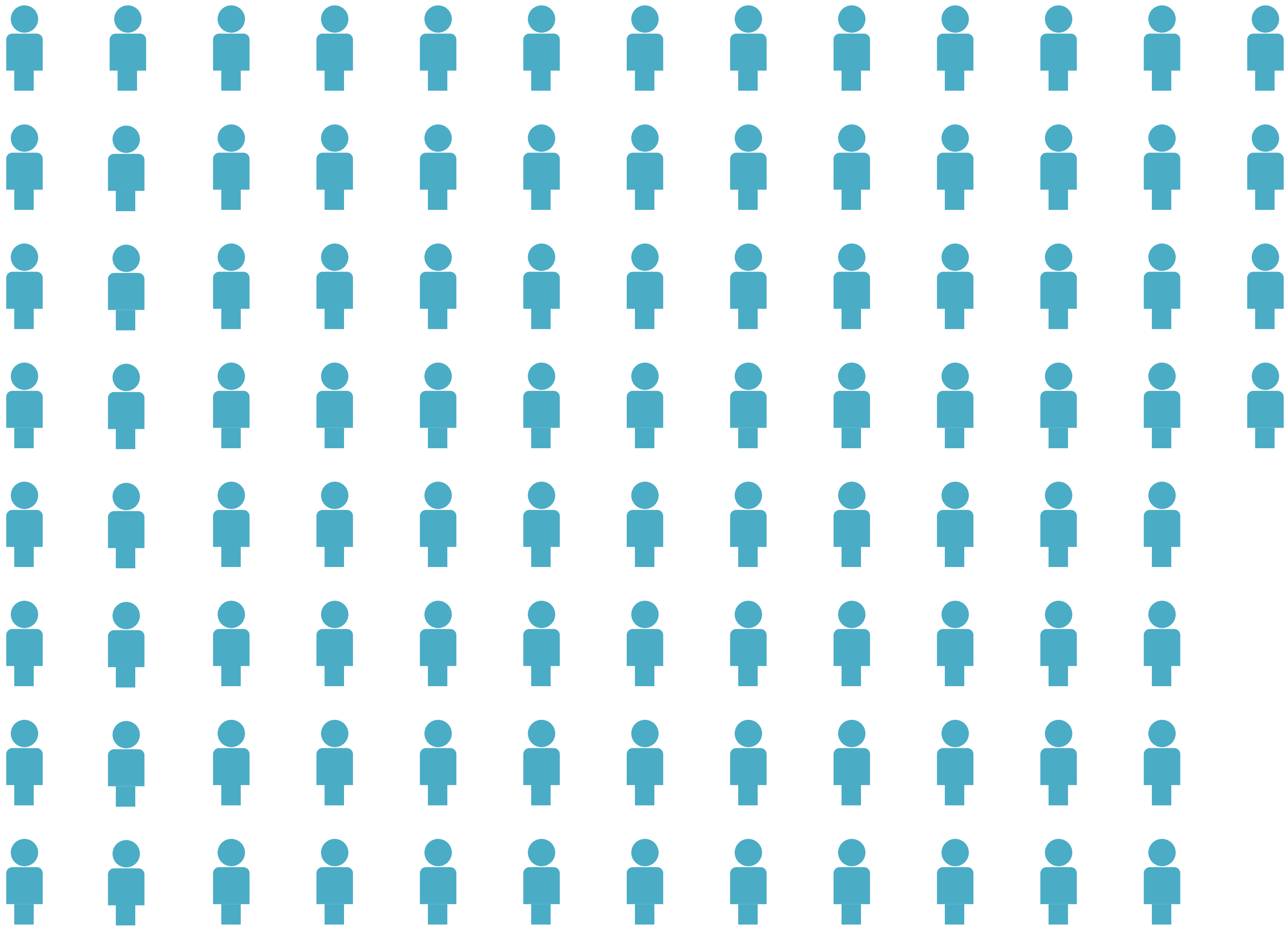
Motivating Example

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

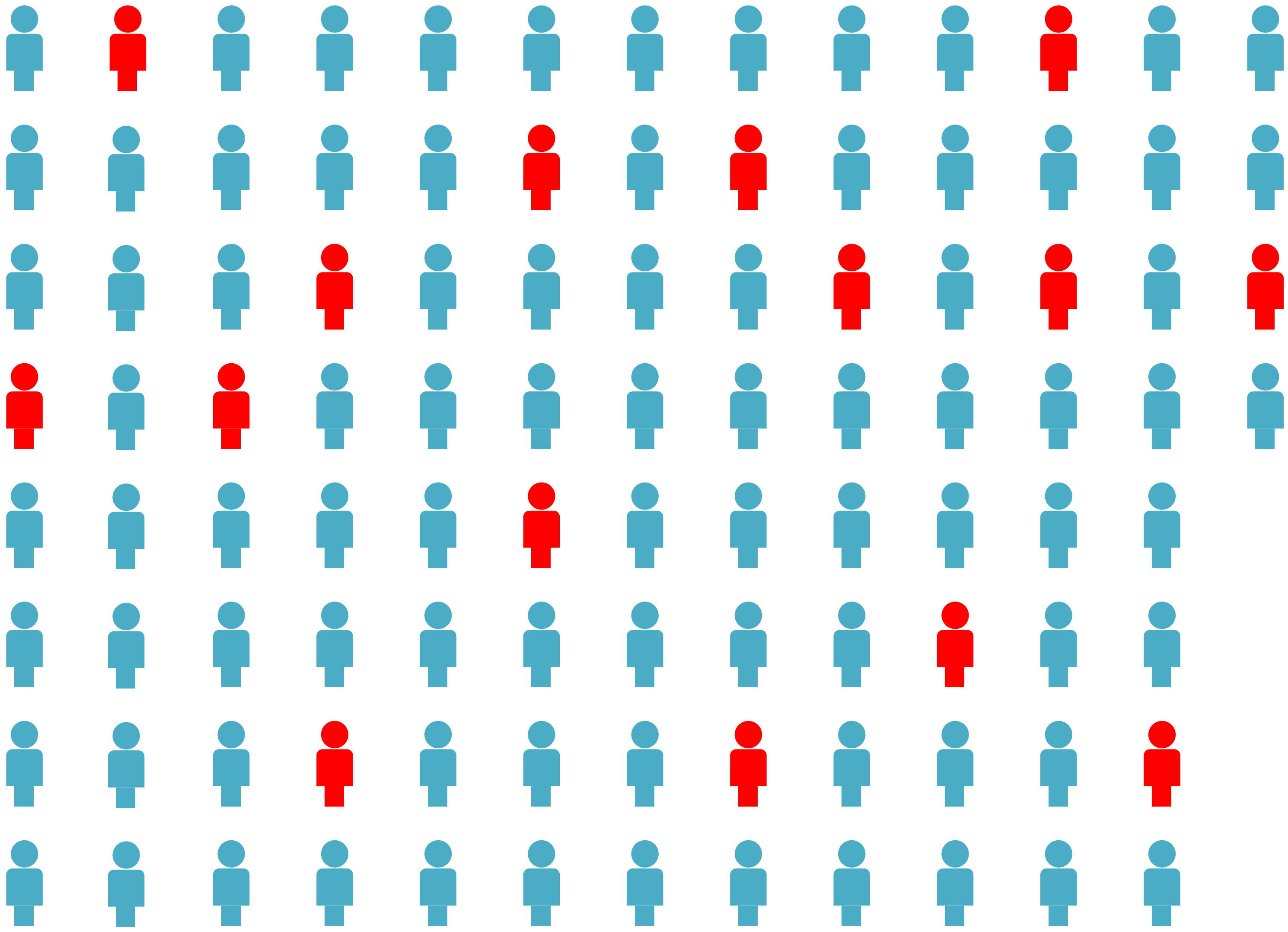
Happiness = {72, 85, 79, 91, 68, … , 71}

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

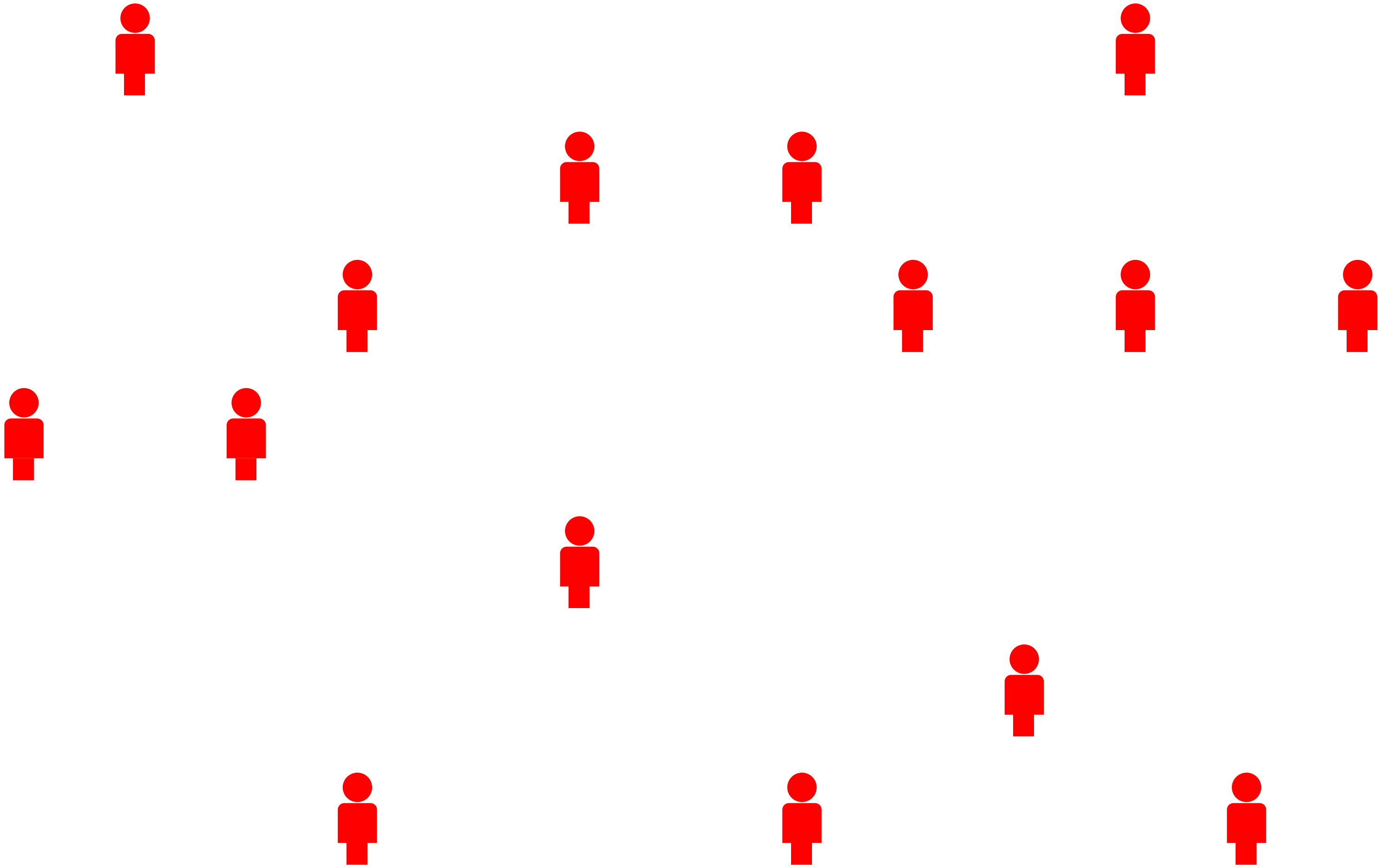
Population



Sample



Sample



Collect one (or more) numbers from each person

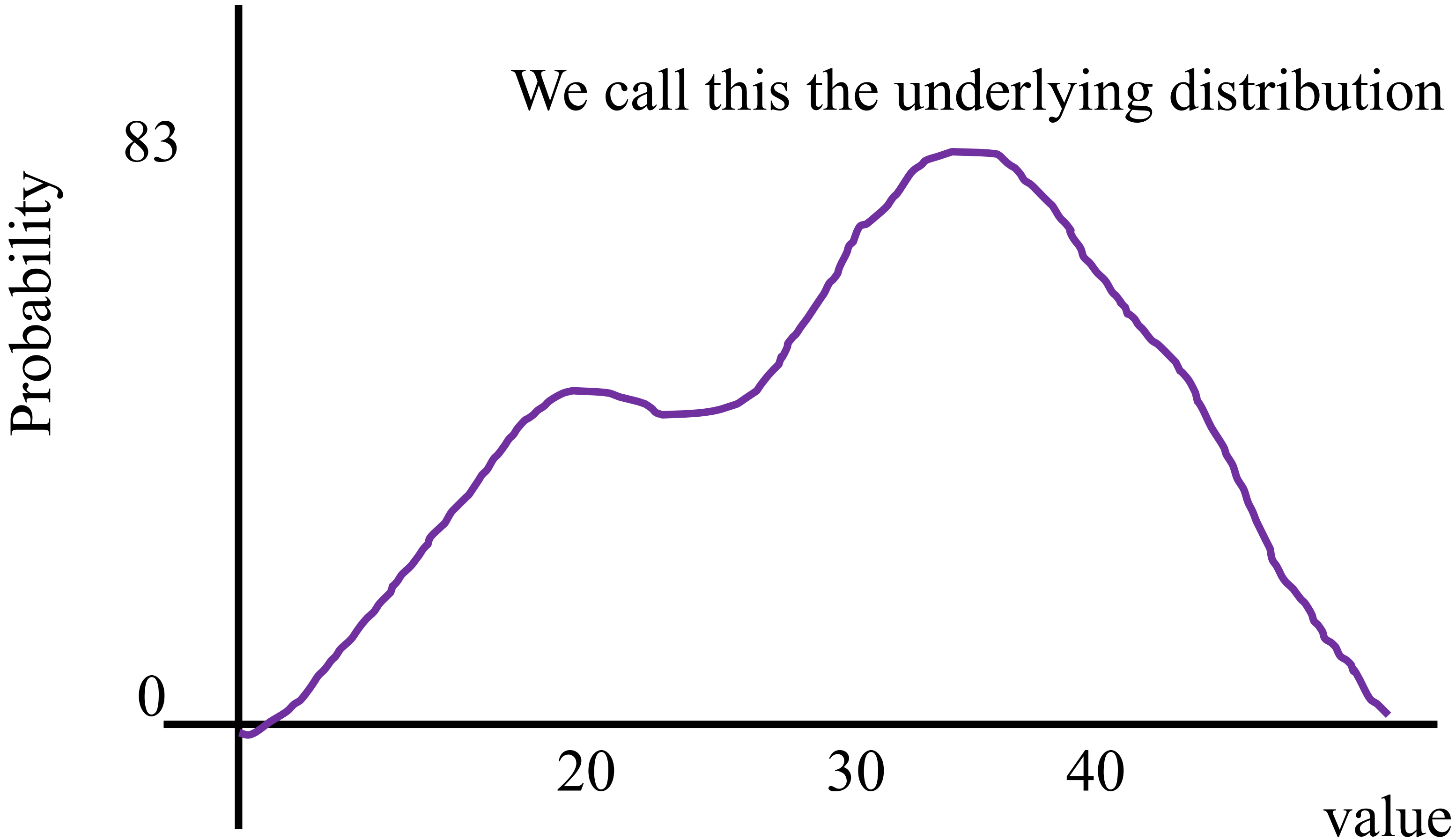
Sample

IID Samples

Consider n IID samples:

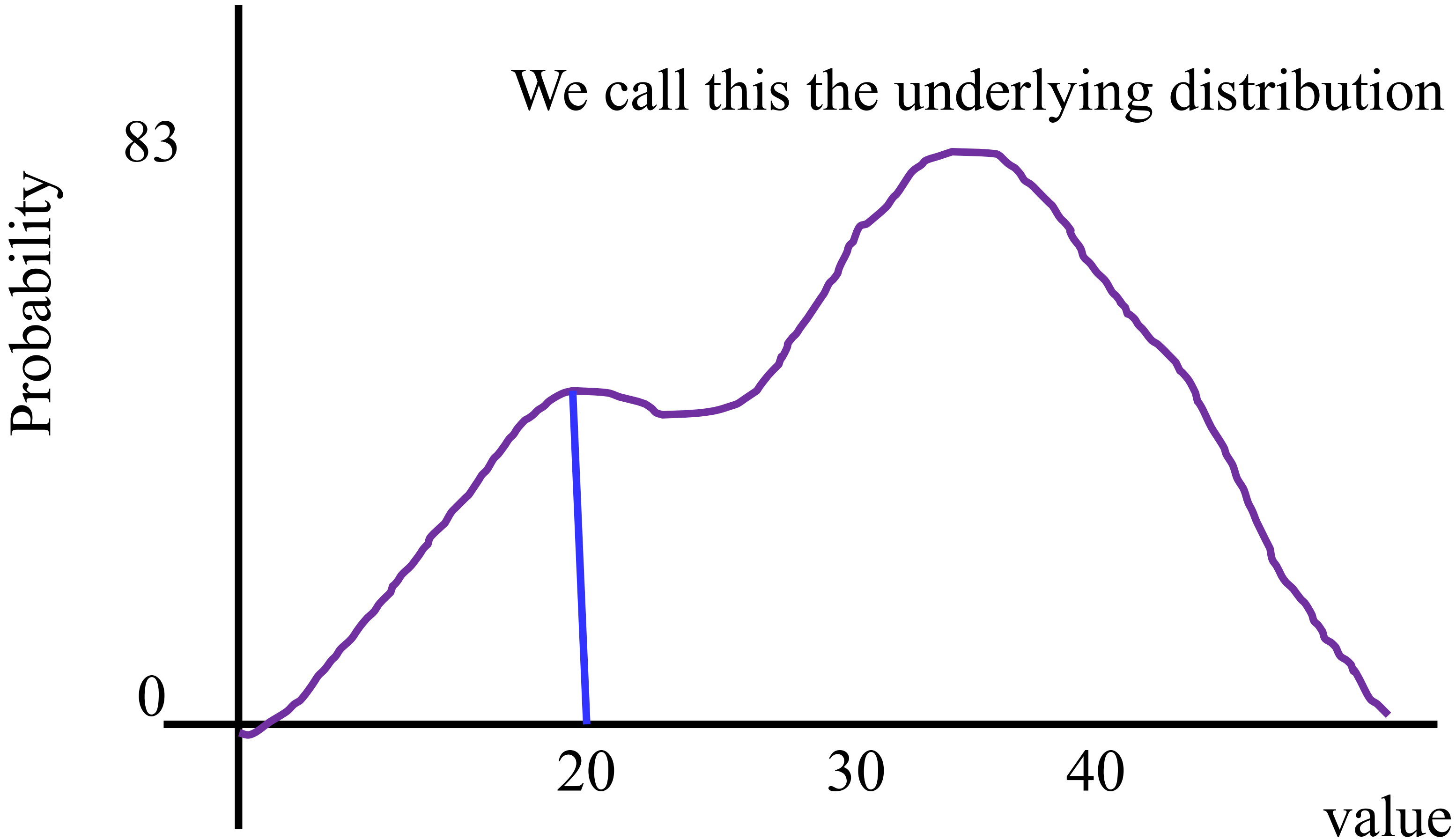
$$X_1, X_2, \ldots, X_n$$

IID Samples



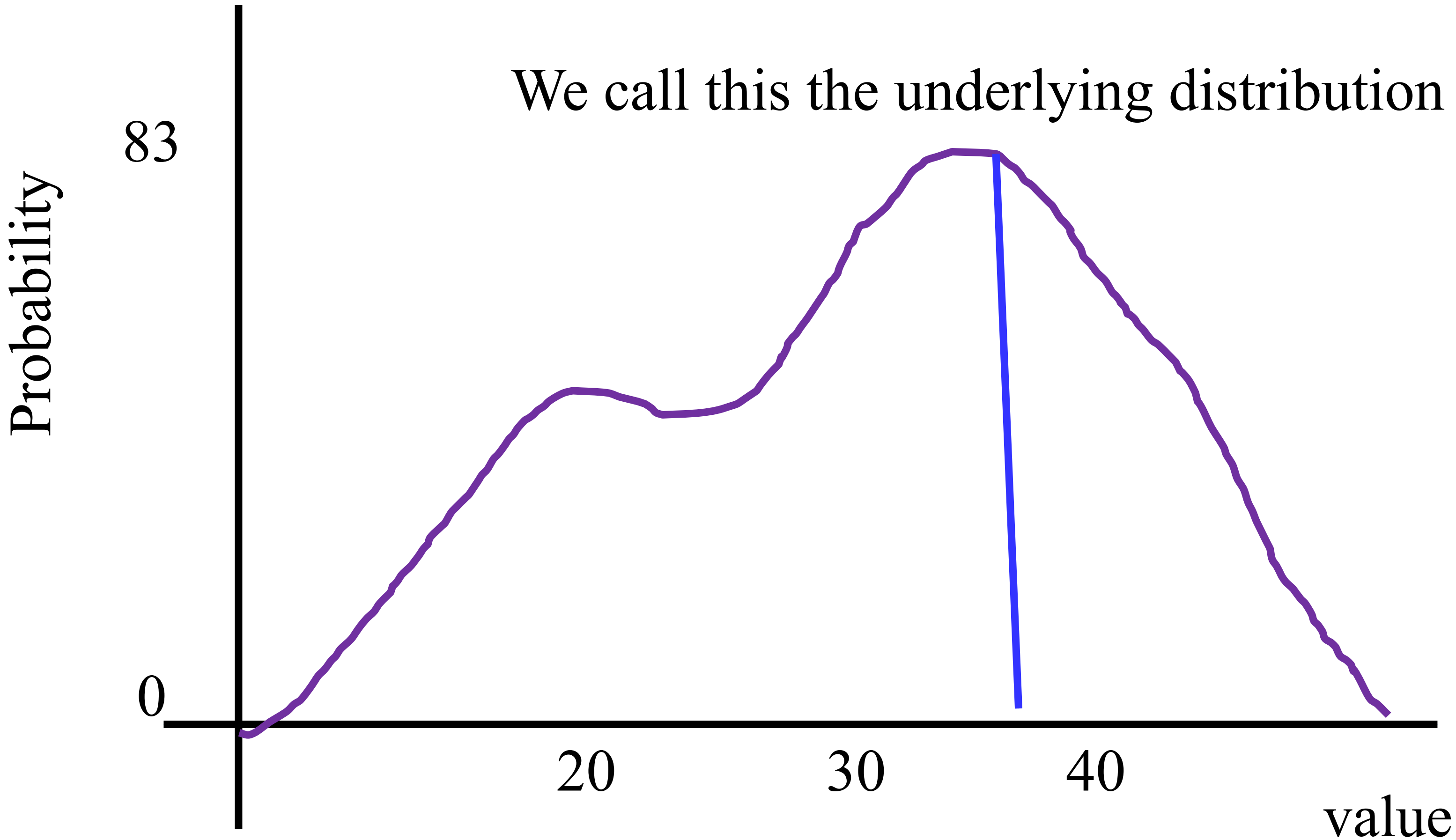
IID Samples = []

IID Samples



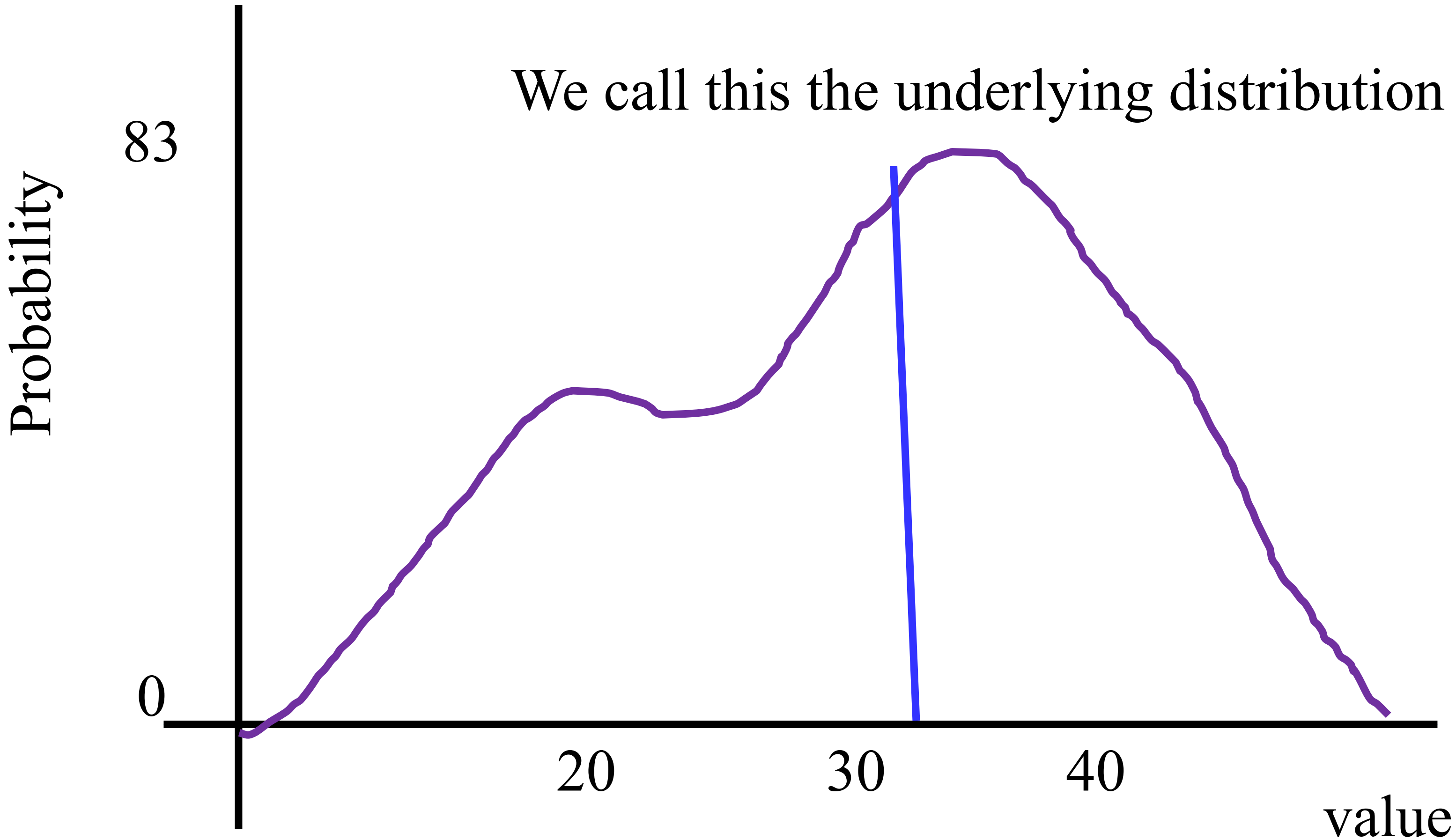
IID Samples = [20]

IID Samples



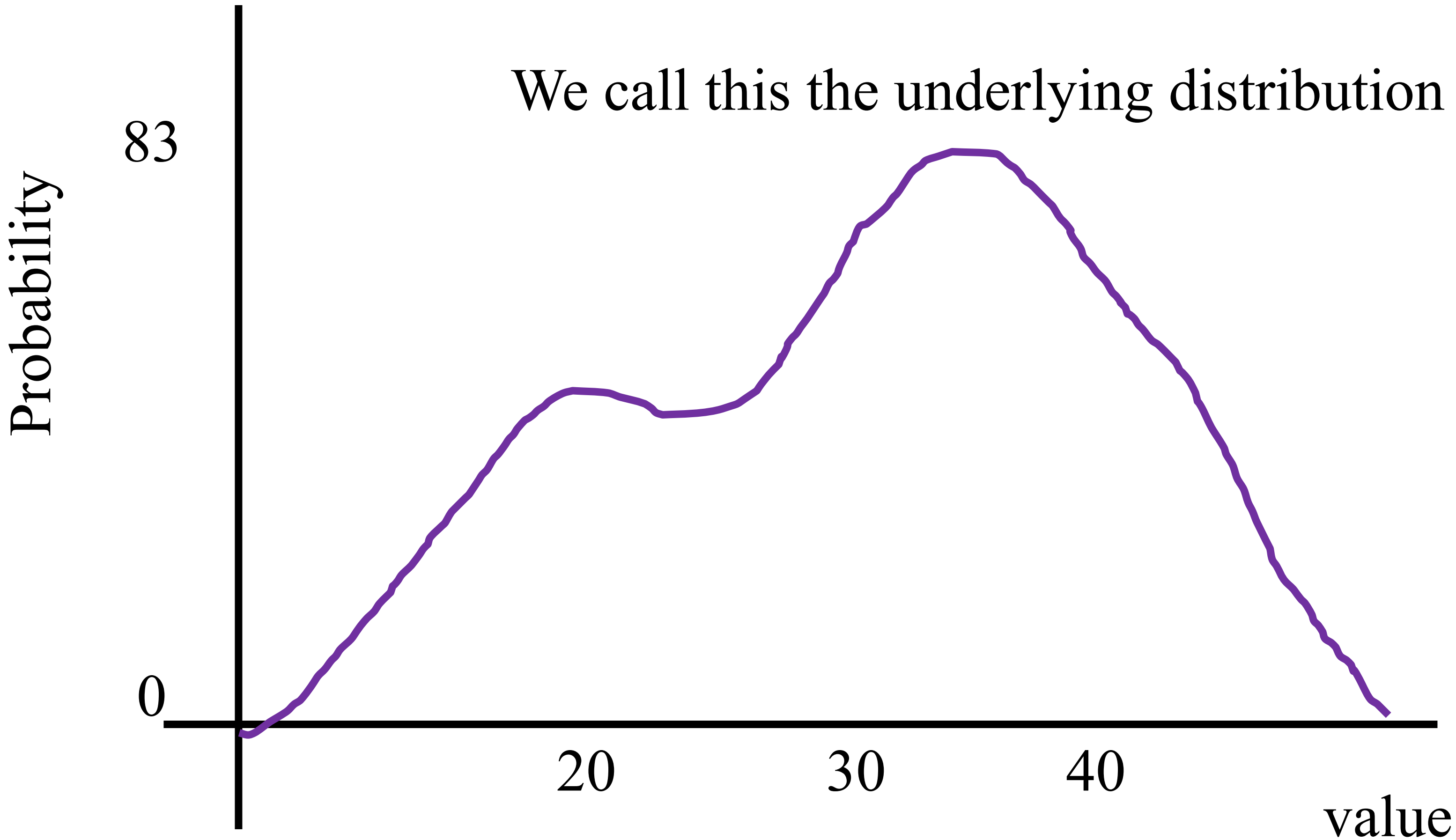
IID Samples = [20, 38]

IID Samples

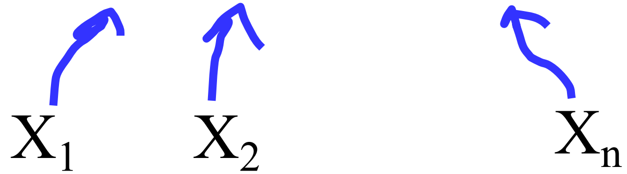


IID Samples = [20, 38, 32]

IID Samples



IID Samples = [20, 38, 32, …, 38]



Sample Mean

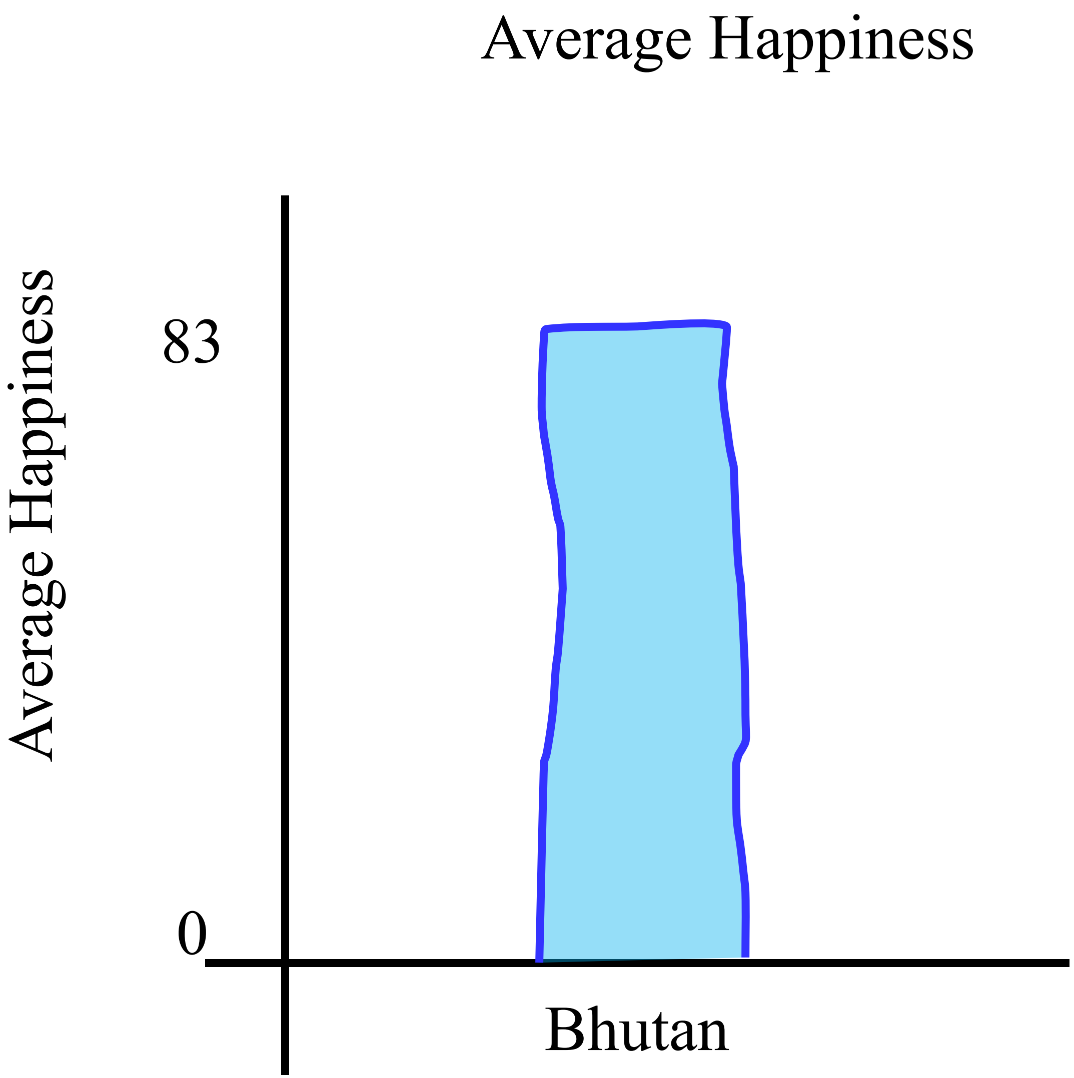
- Consider *n* random variables $X_1, X_2, \ldots X_n$
  - X_i are all independently and identically distributed (I.I.D.)
  - Have same distribution function *F* and $E[X_i] = \mu$
  - We call sequence of X_i a **<u>sample</u>** from distribution *F*
  - Sample mean: $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$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

ith sample

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

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: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
  - Sample deviation: $\bar{X} - X_i$ for i = 1, 2, ..., 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

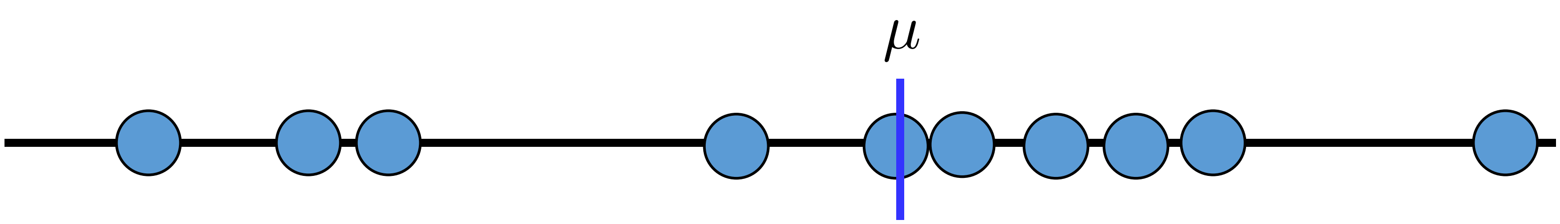
$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

This is the actual mean

Unbiased sample variance

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

This is the sample mean

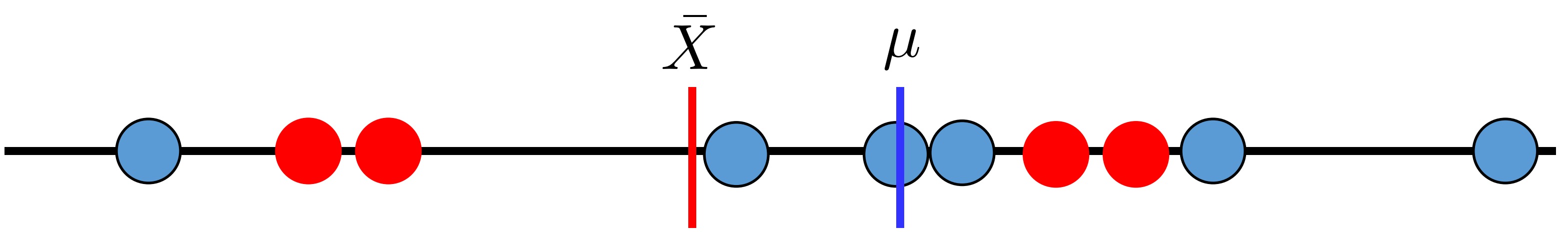


Intution that $E[S^2] = \sigma^2$

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



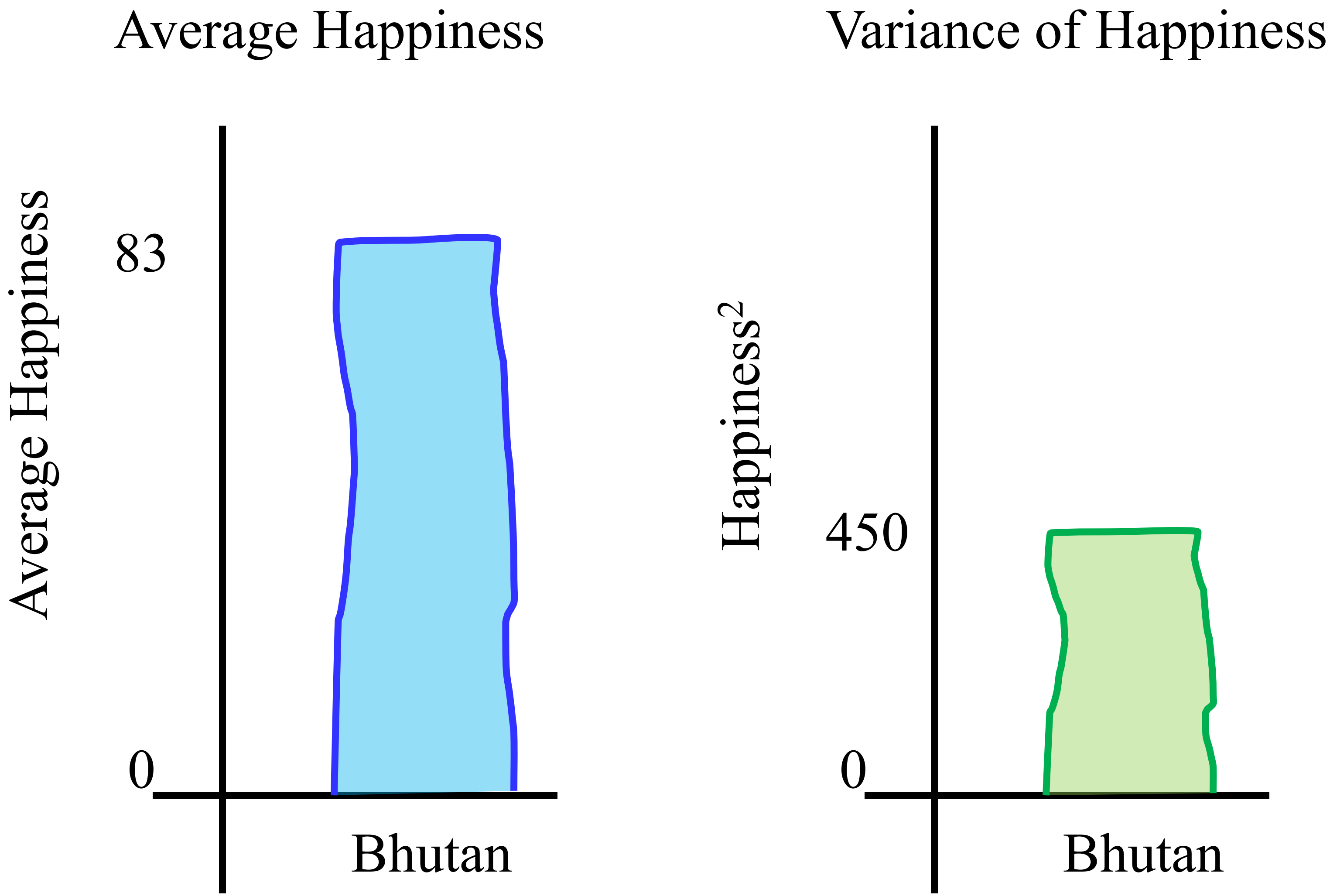
Proof that E[S^2] = σ^2 (just for reference)

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

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

- So, E[S^2] = σ^2

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"

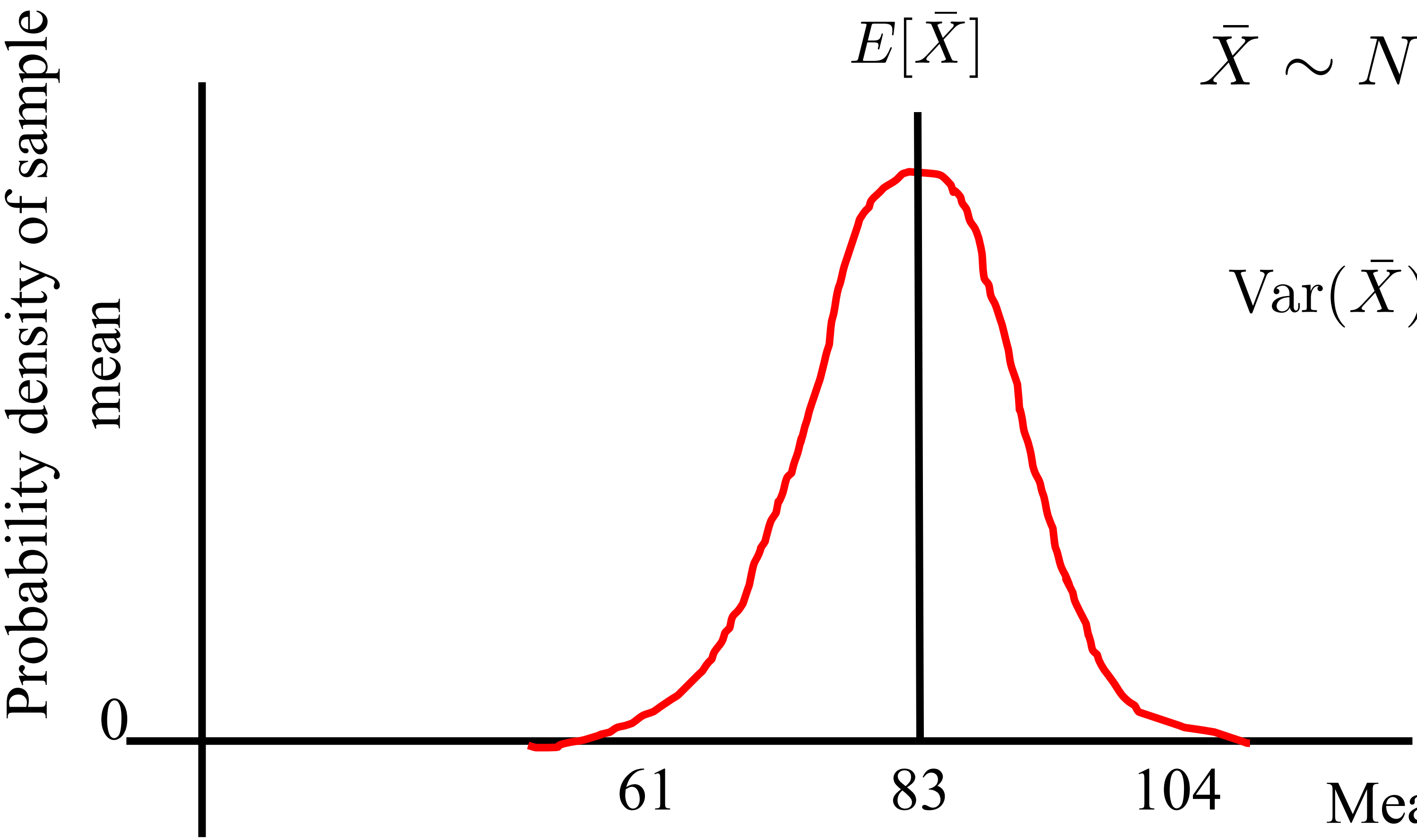
No Error Bars ☹

Variance of Sample Mean

By central limit theorem:

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

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



Variance of Sample Mean

- 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: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
  - What is $\mathrm{Var}(\bar{X})$?

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

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

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

Standard Error of the Mean

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

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

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

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

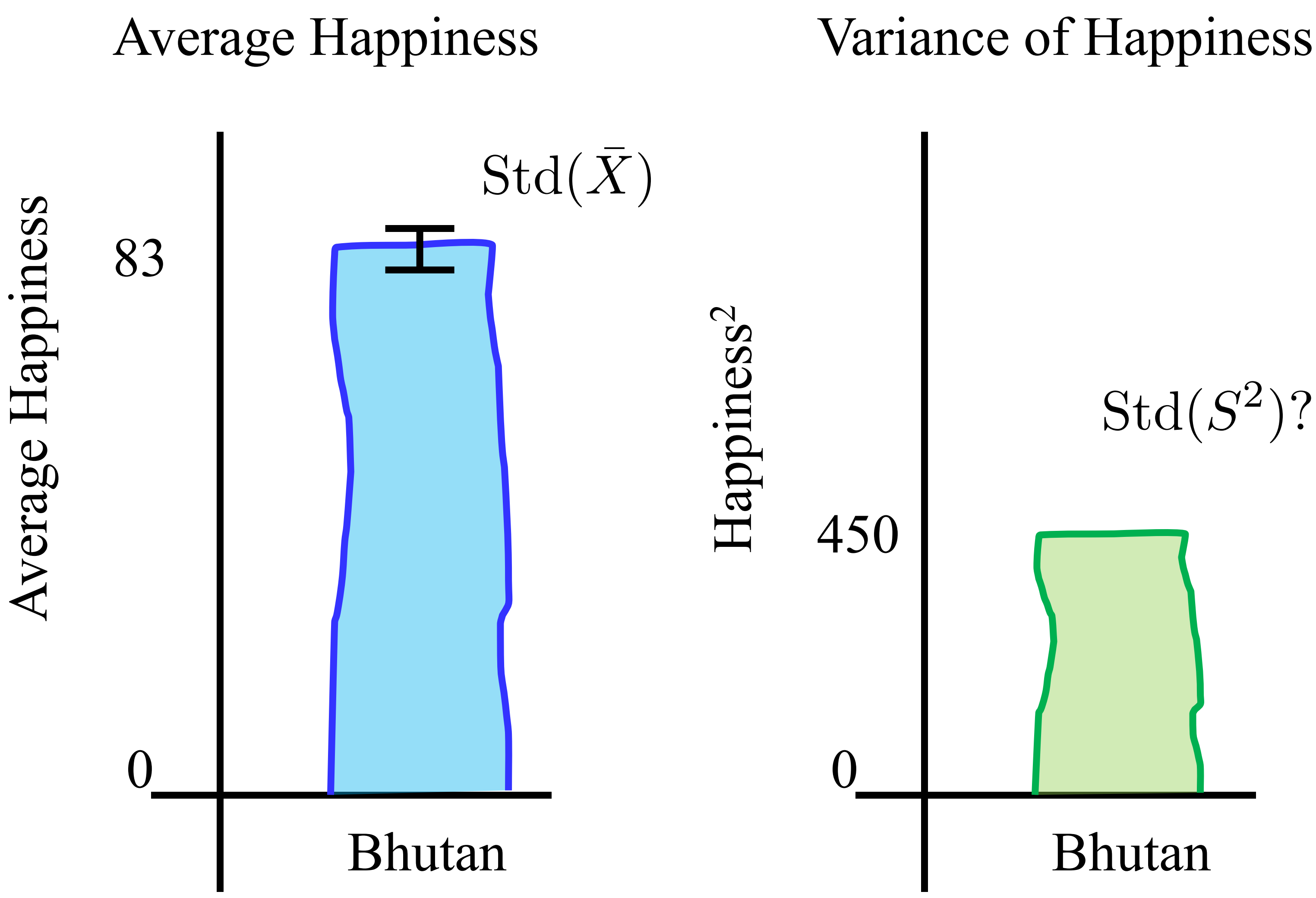
$$= \sqrt{\frac{450}{200}}$$

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

Stretch!

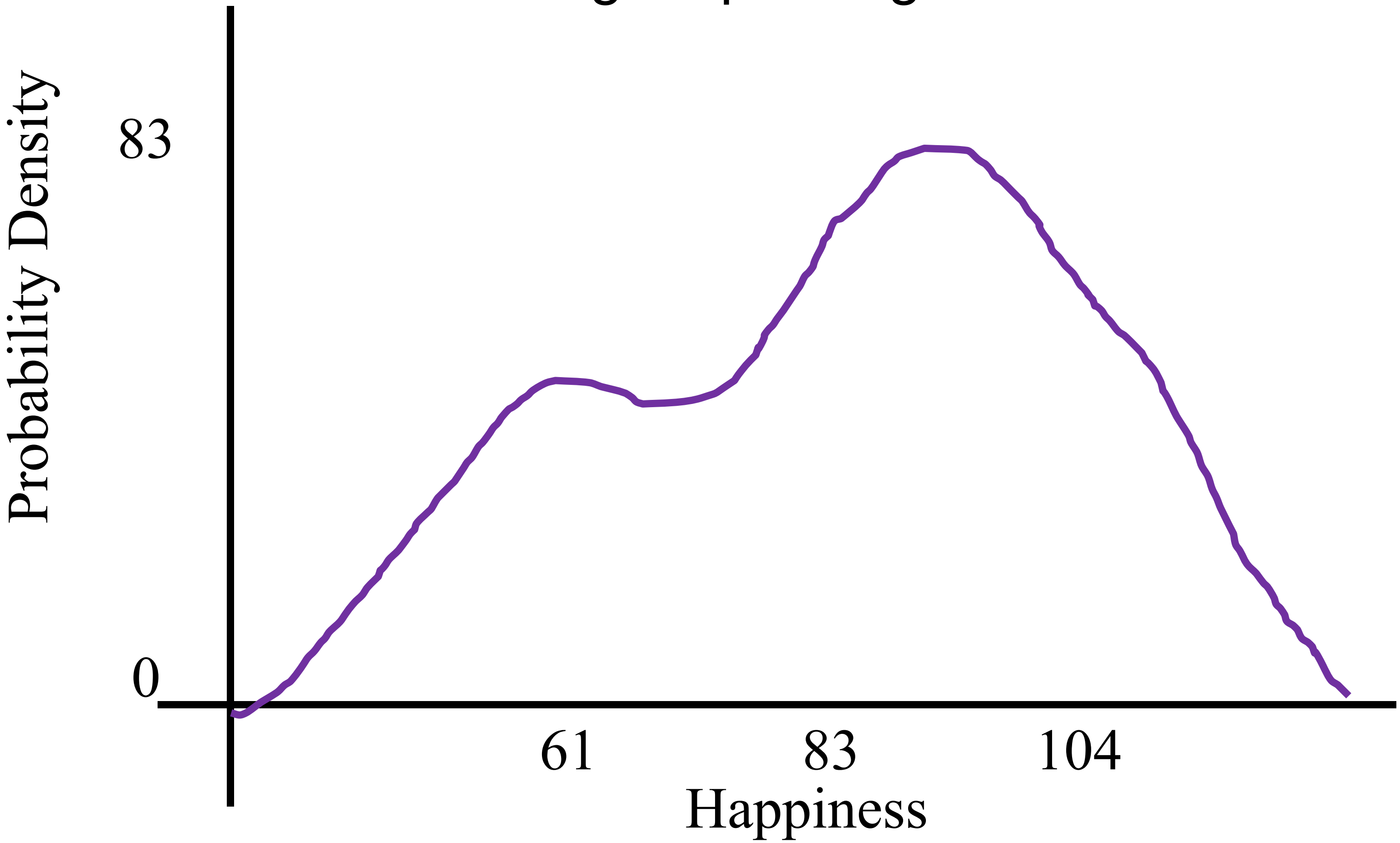
Bootstrapping:
Probability for Computer Scientists

Bootstrapping allows you to:

- Know the distribution of statistics
- Calculate p values

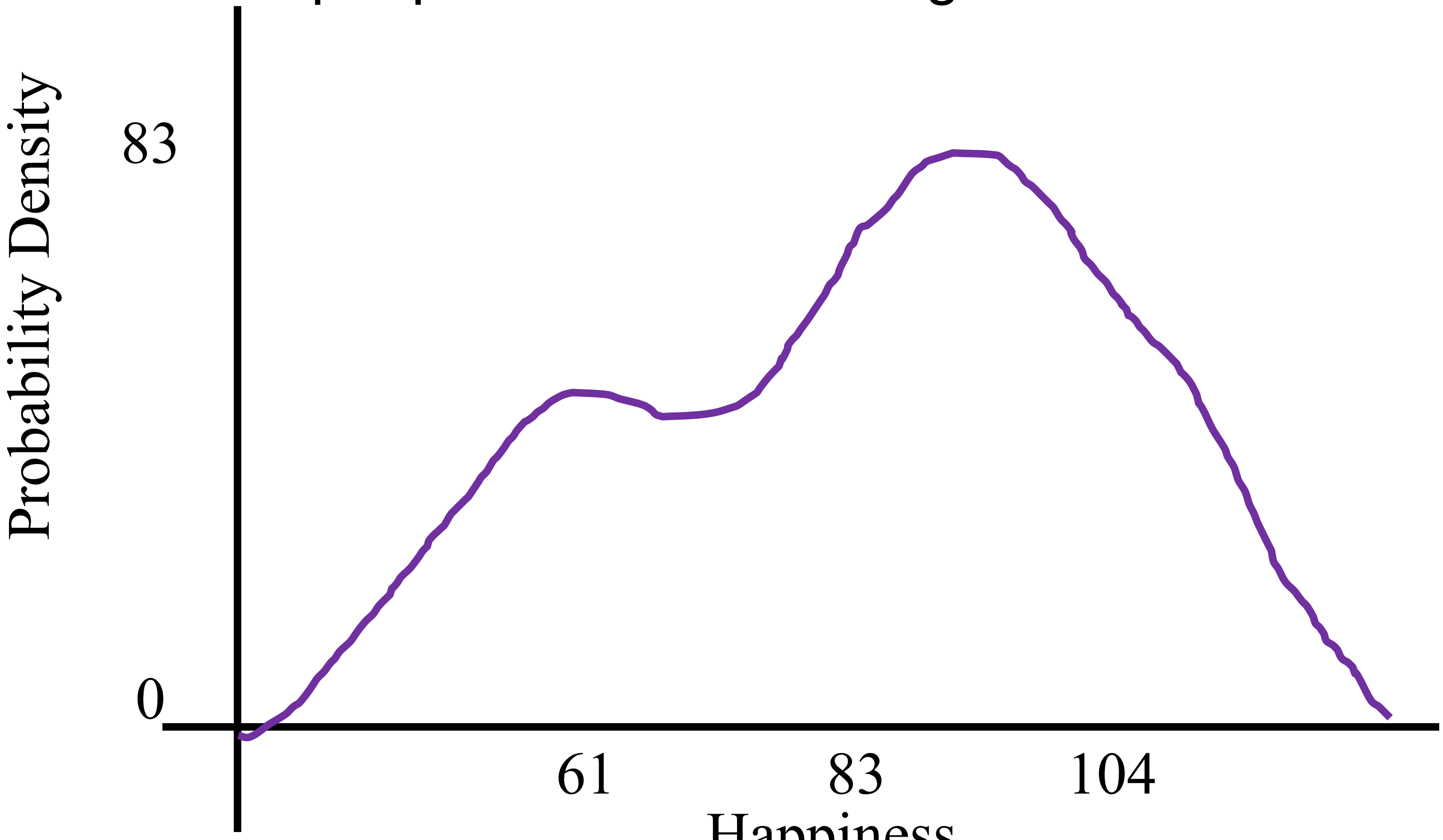
Hypothetical

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



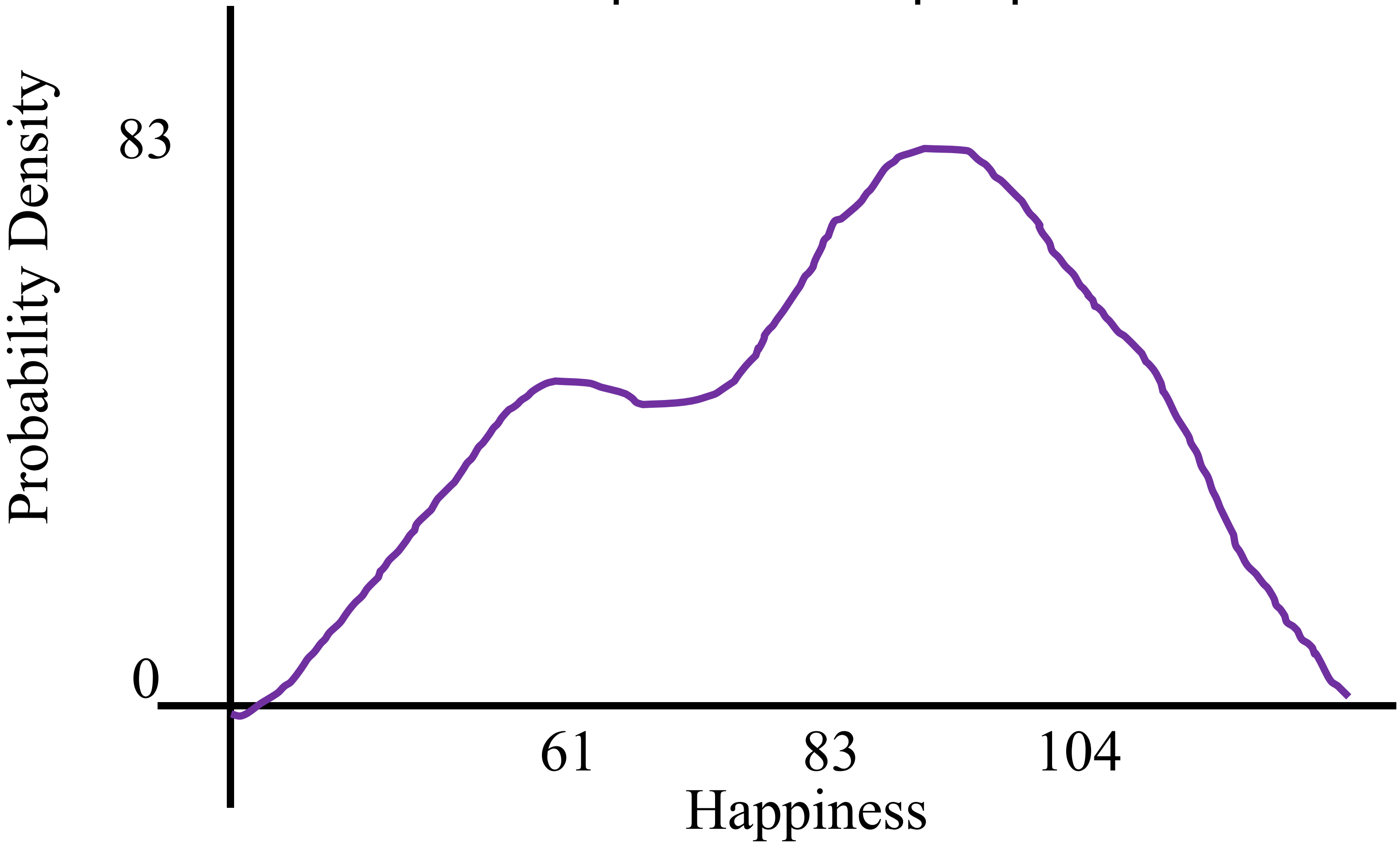
Hypothetical

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



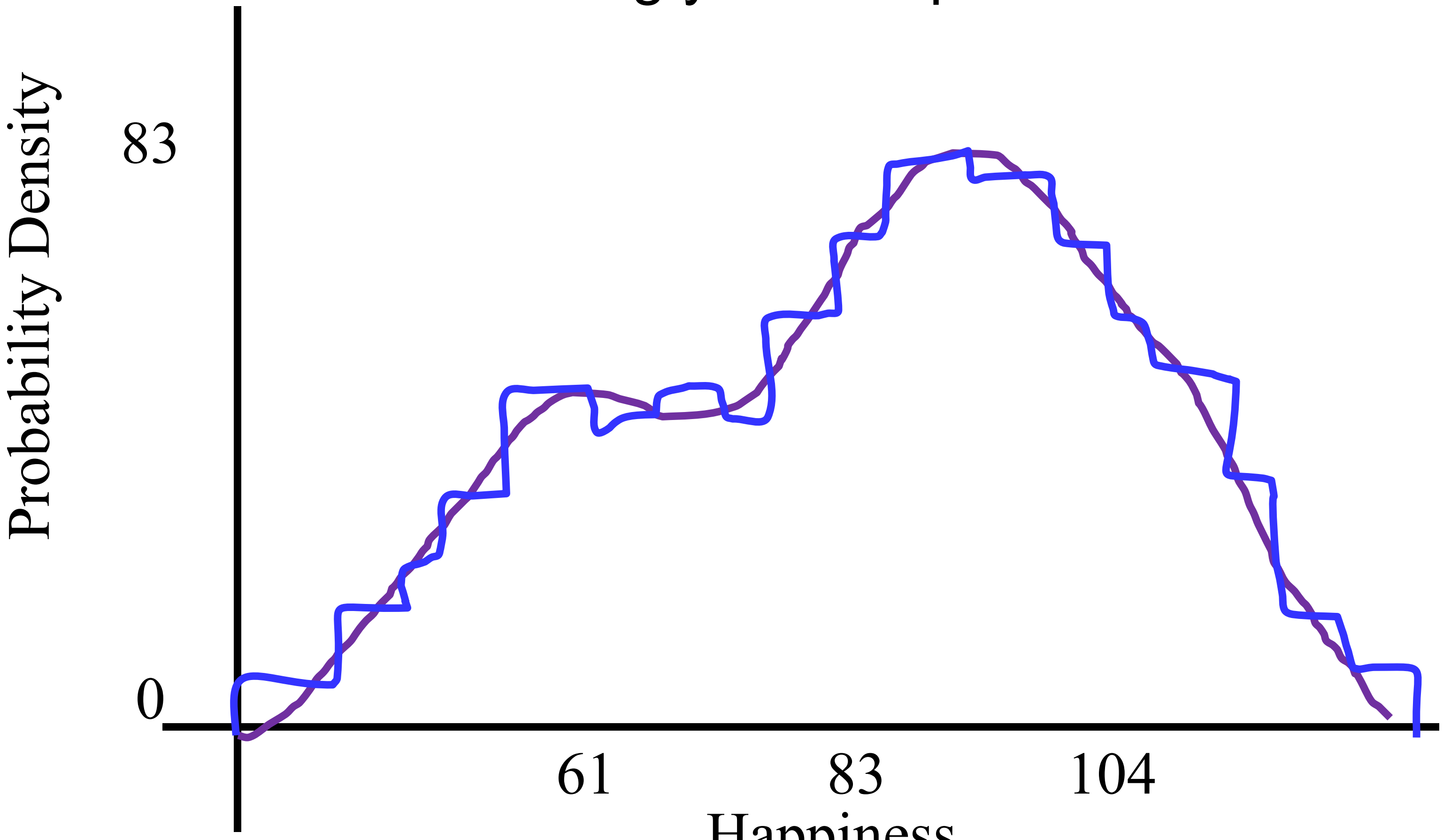
Hypothetical

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.*



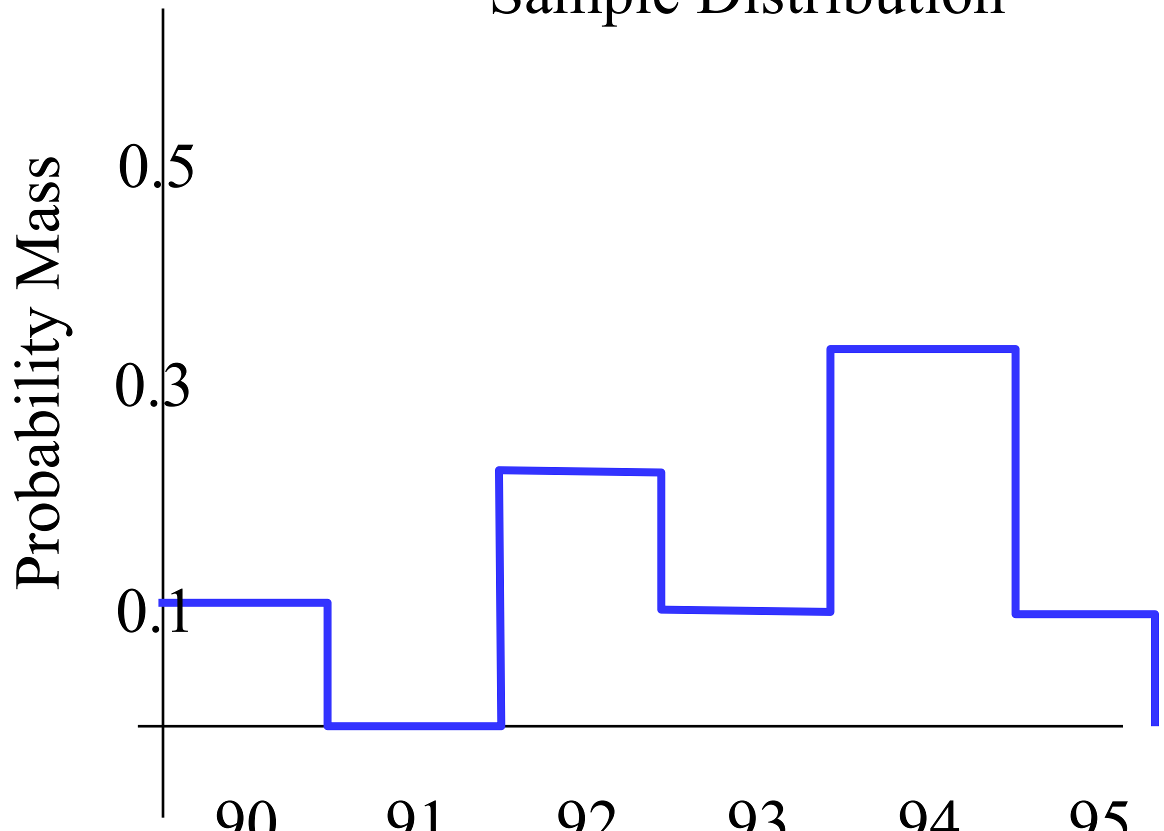
* This is just a histogram of your data!!

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):

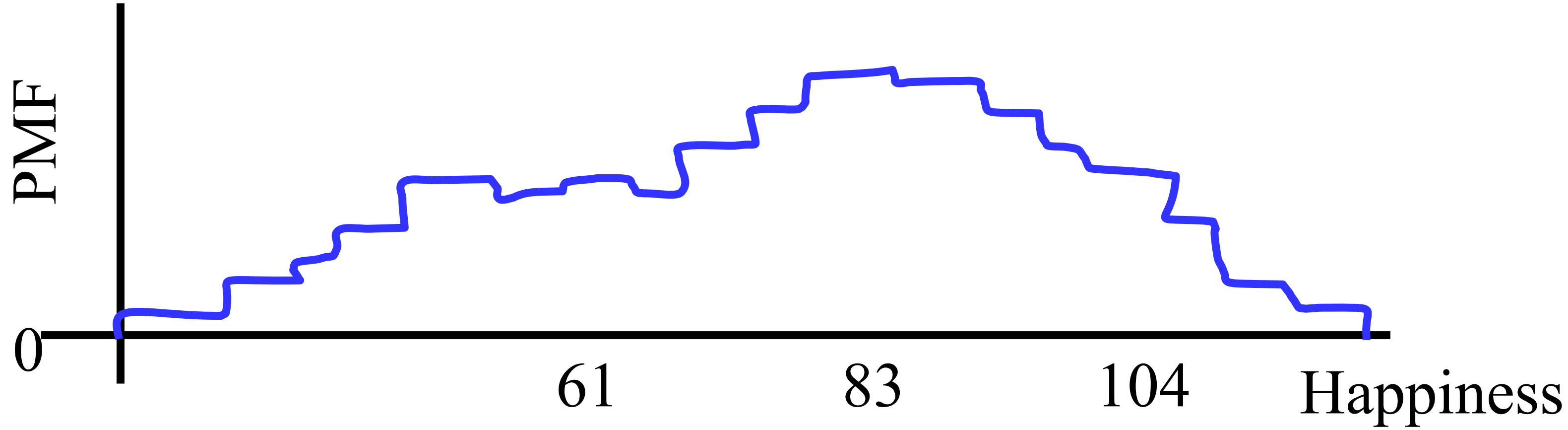
1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Resample **sample.size**() from PMF
   b. Recalculate the stat on the resample
3. You now have a **distribution of your stat**

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**

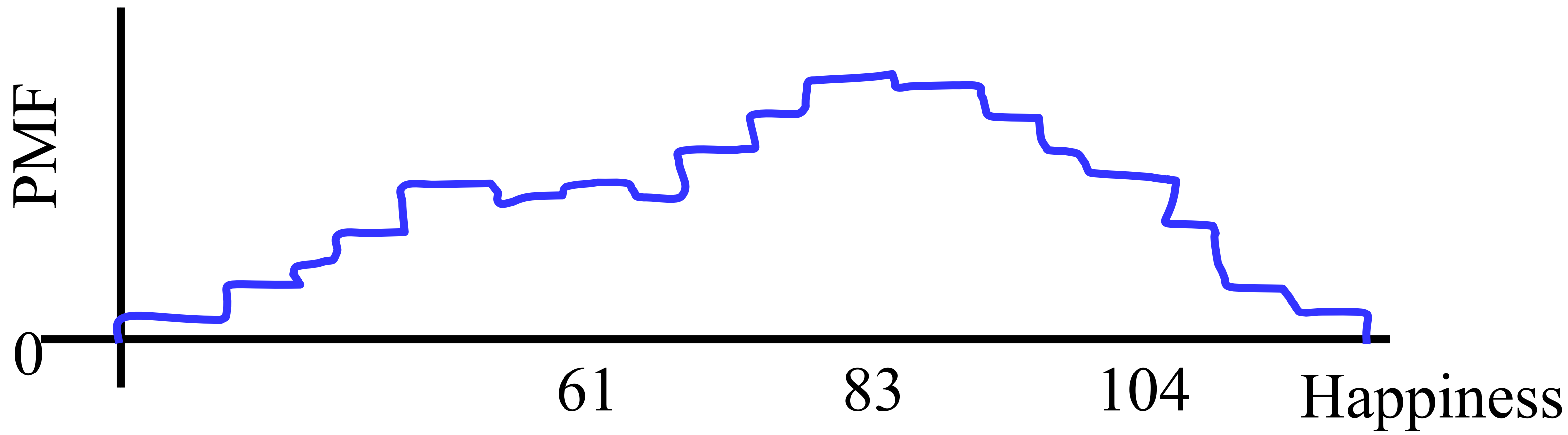
Bootstrap of Means



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**

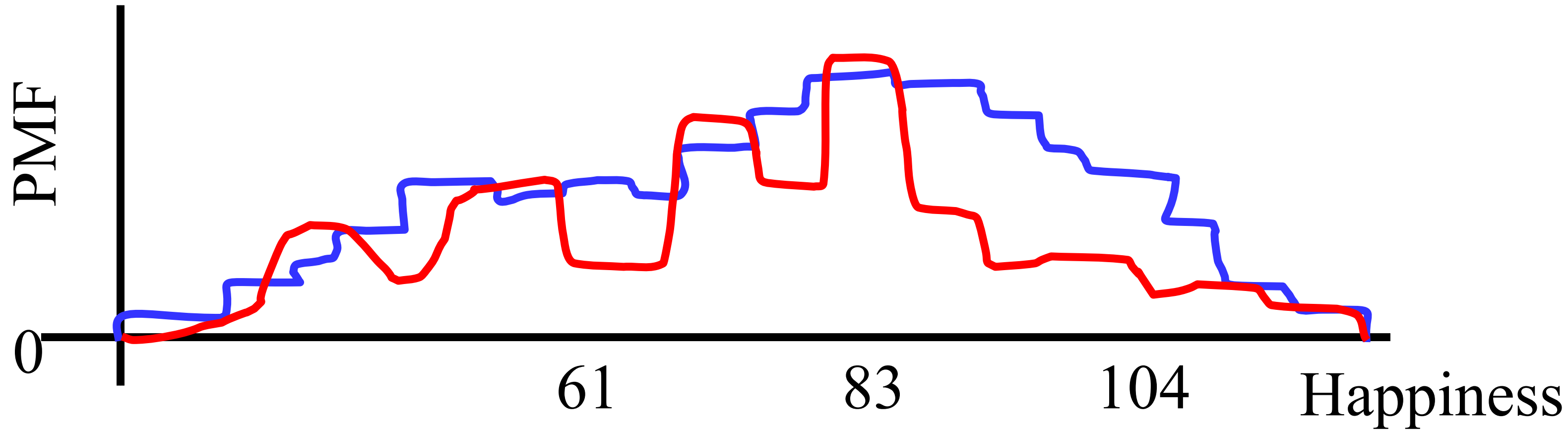
Bootstrap of Means



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**

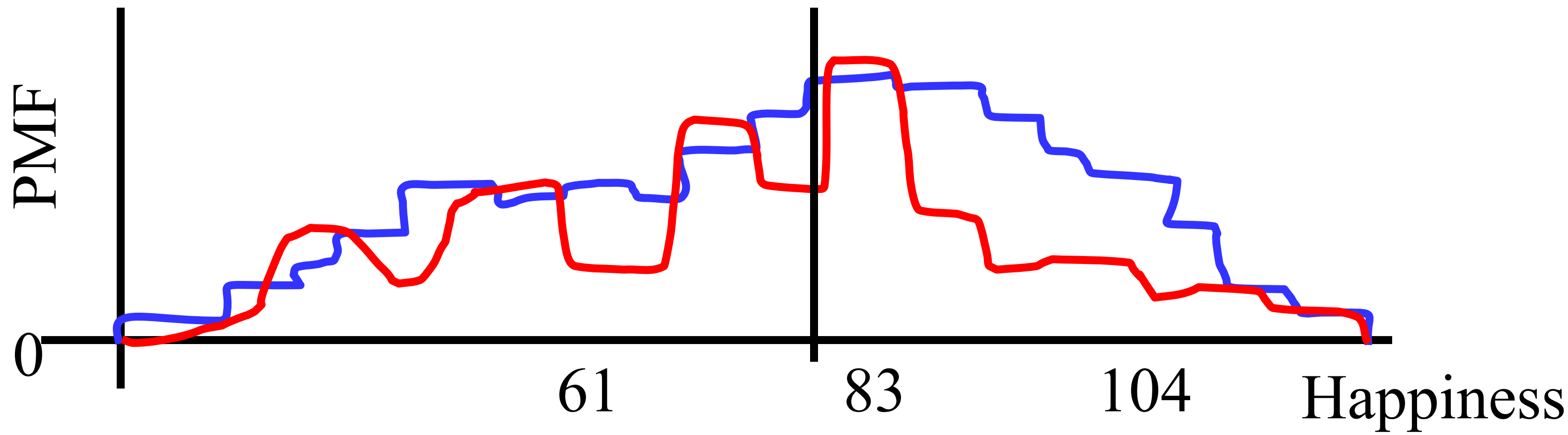
Bootstrap of Means



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

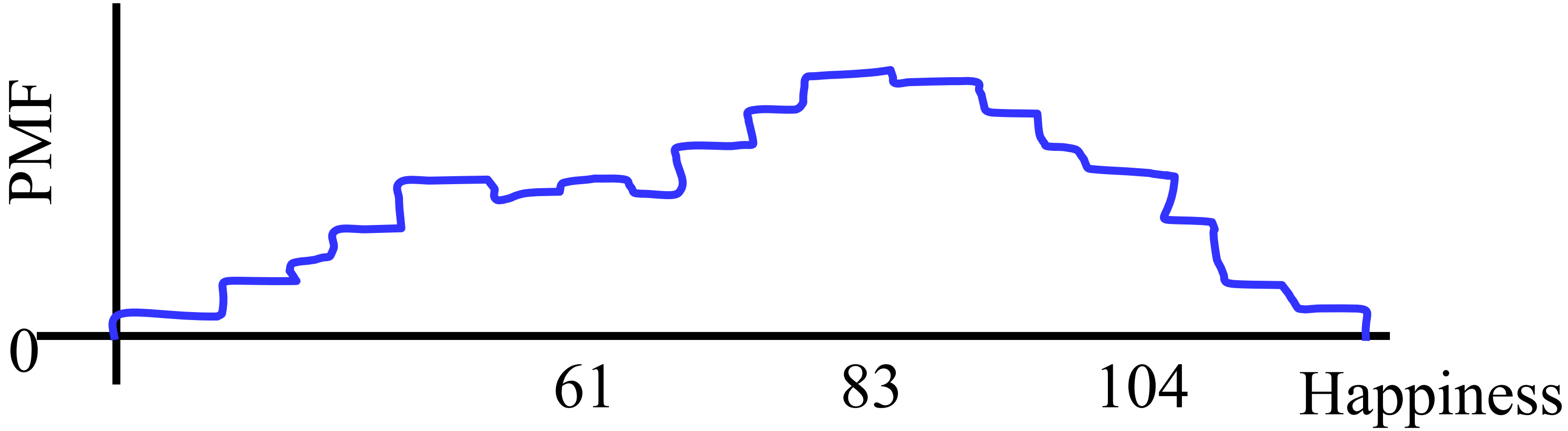


Bootstrap Algorithm (sample):

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

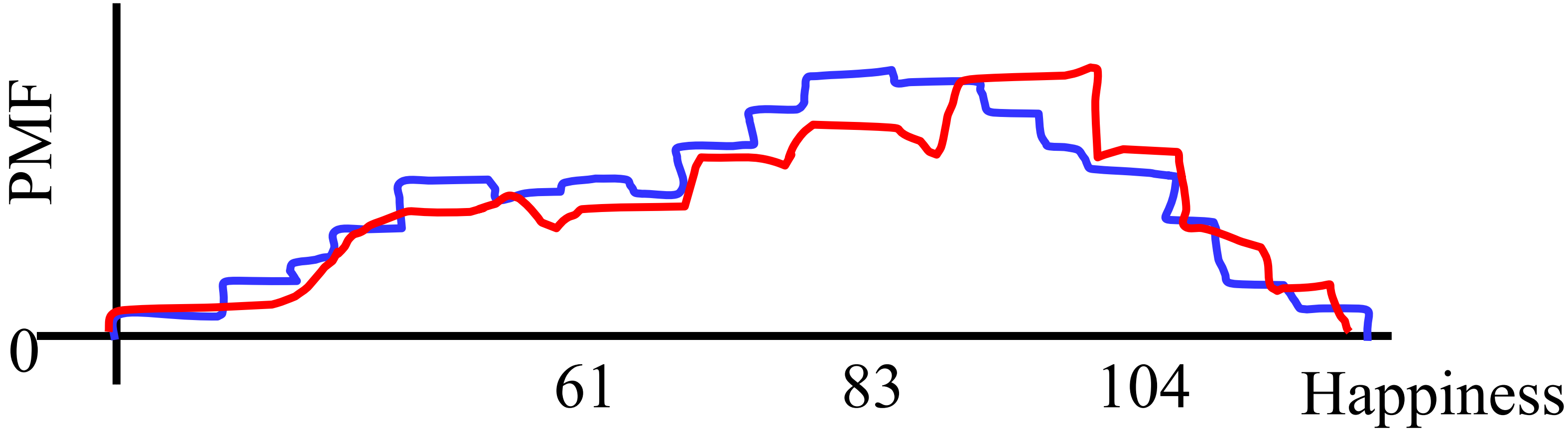


Bootstrap Algorithm (sample):

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

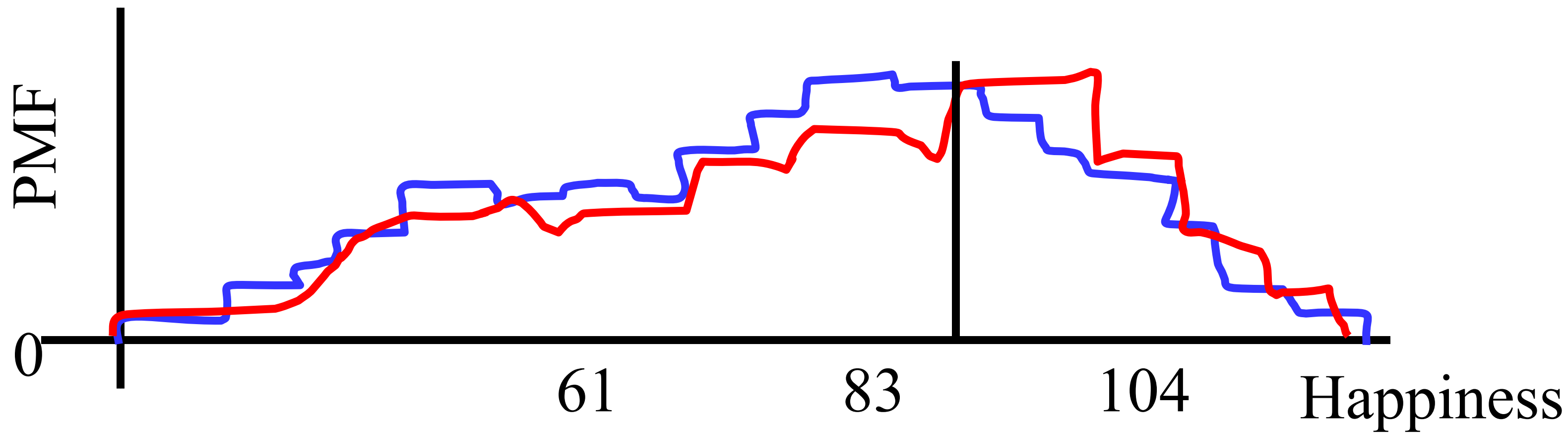


Bootstrap Algorithm (sample):

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

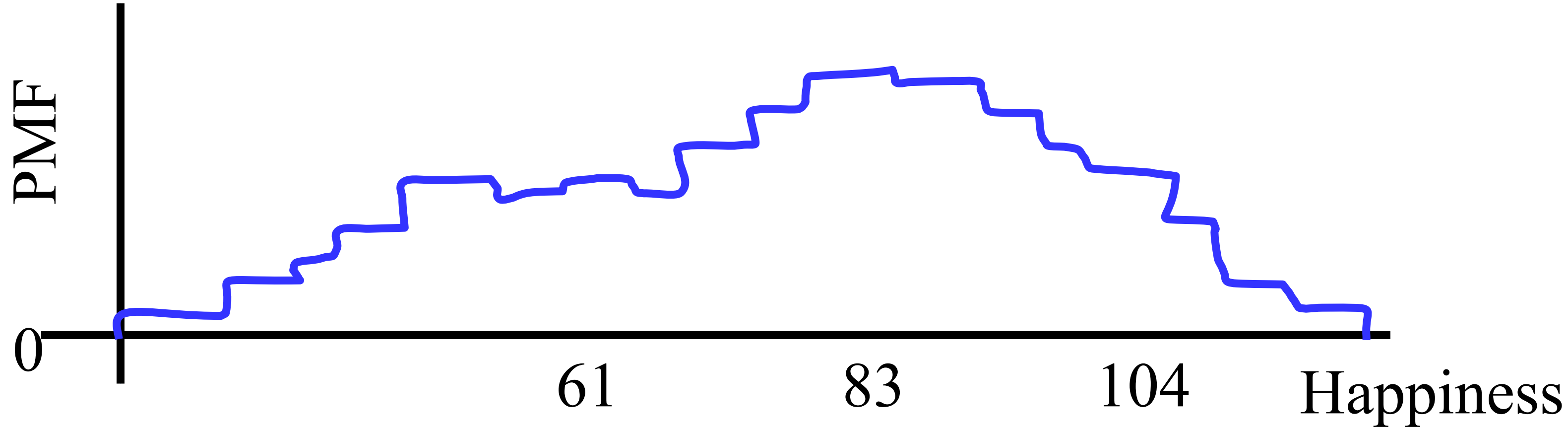


Bootstrap Algorithm (sample):

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

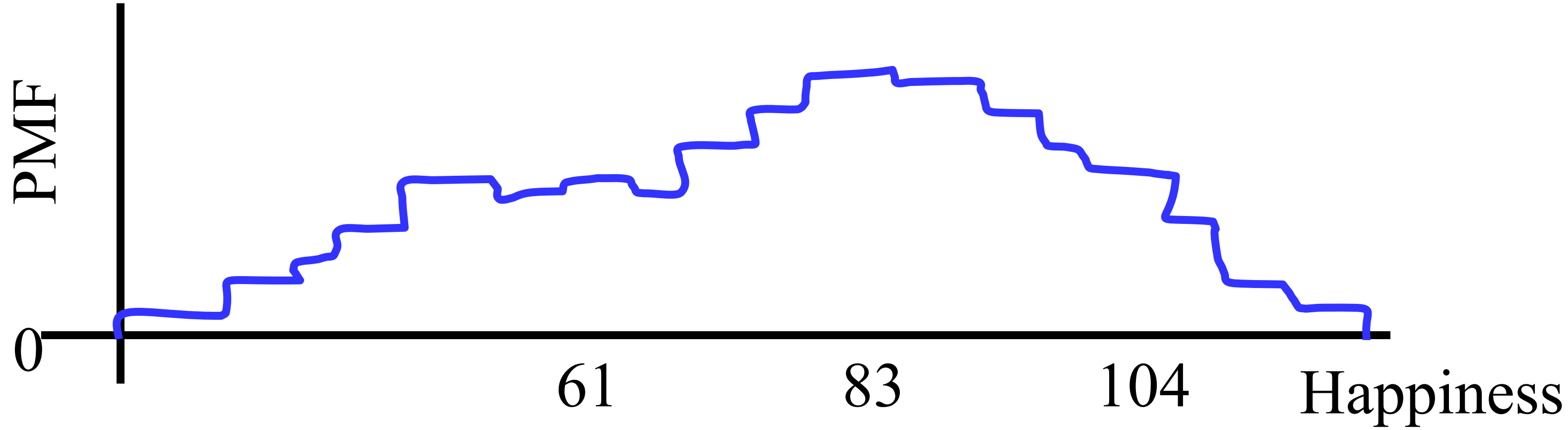


Bootstrap Algorithm (sample):

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



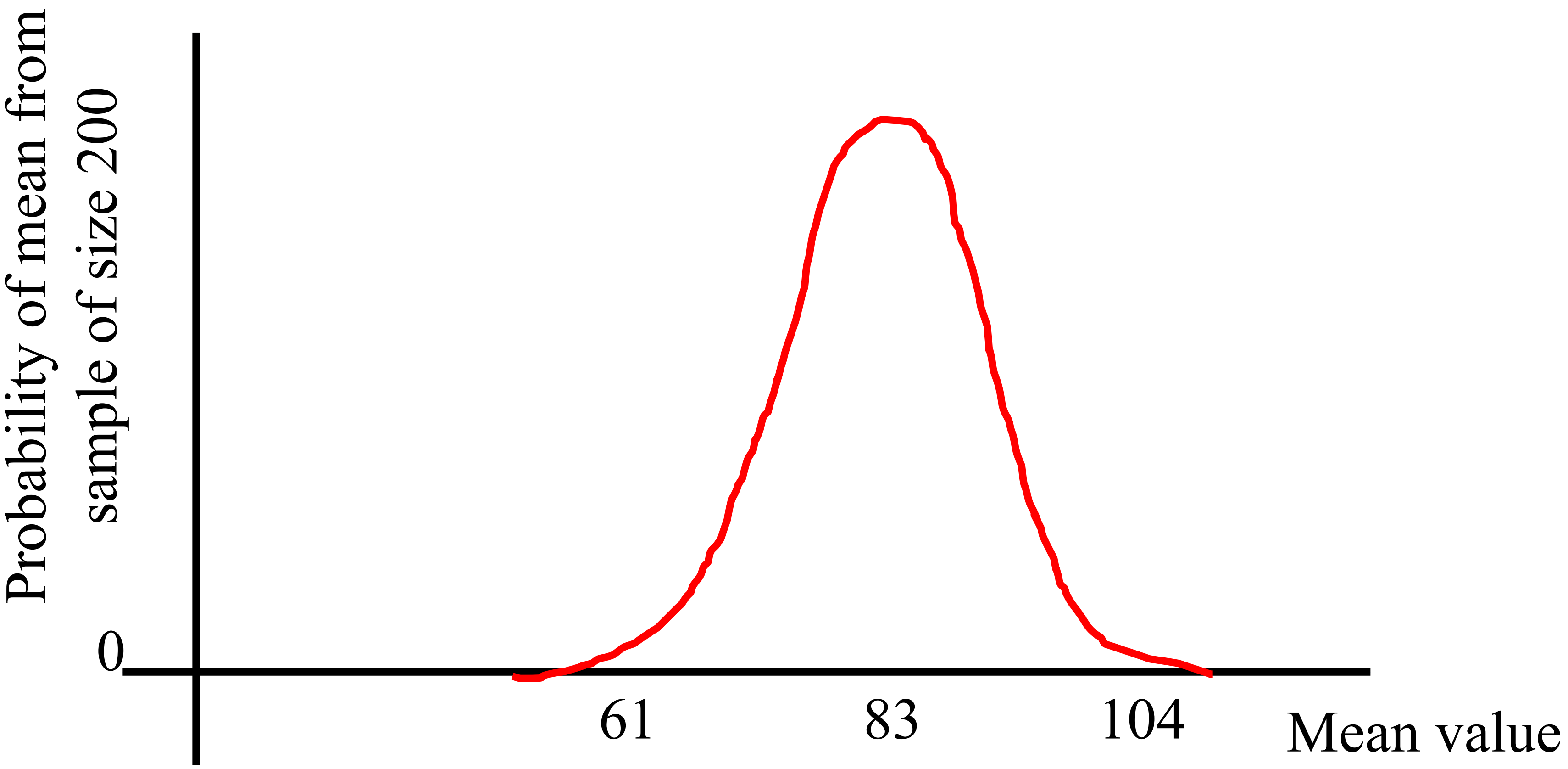
Bootstrap Algorithm (sample):

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]

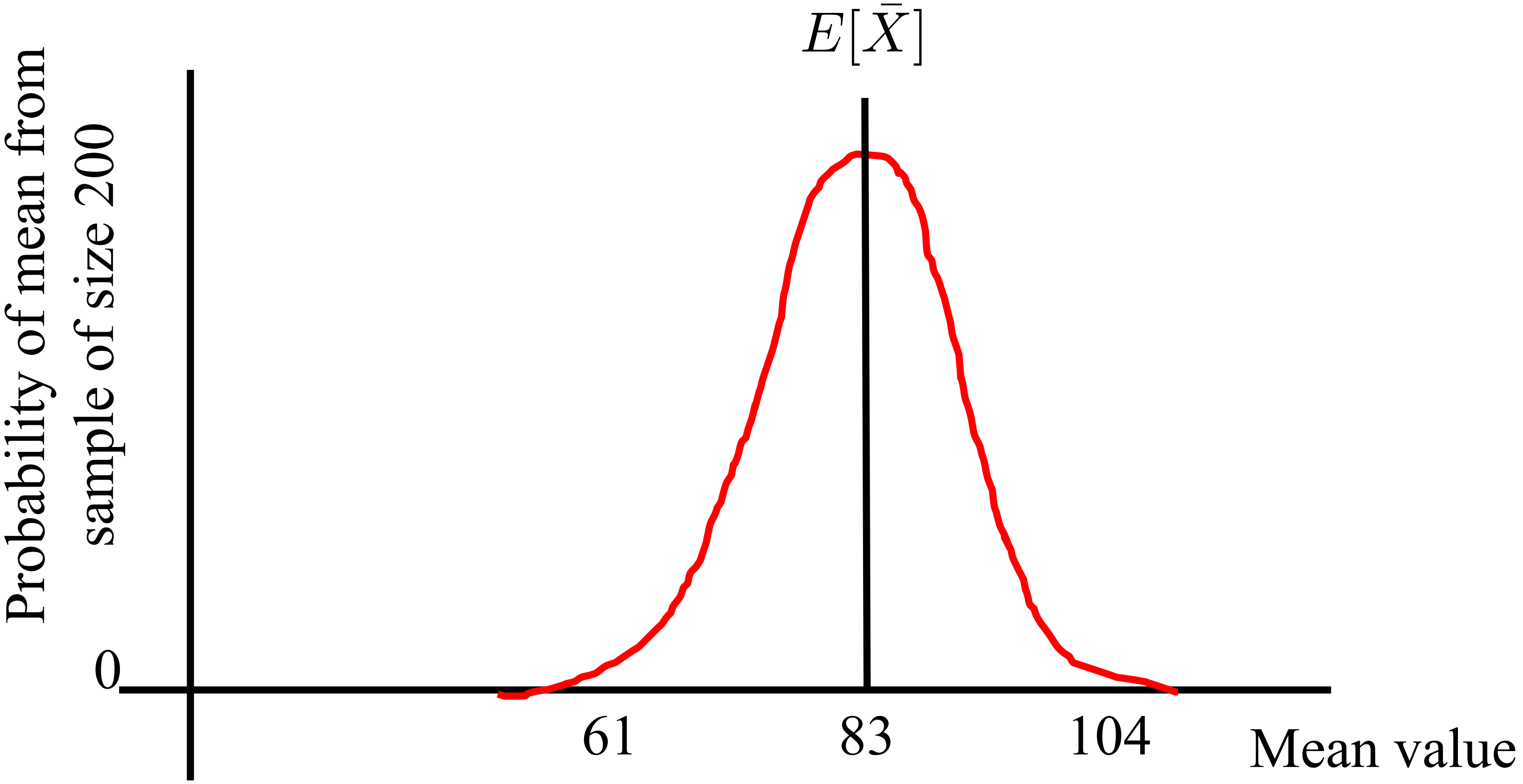
Bootstrap of Means

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



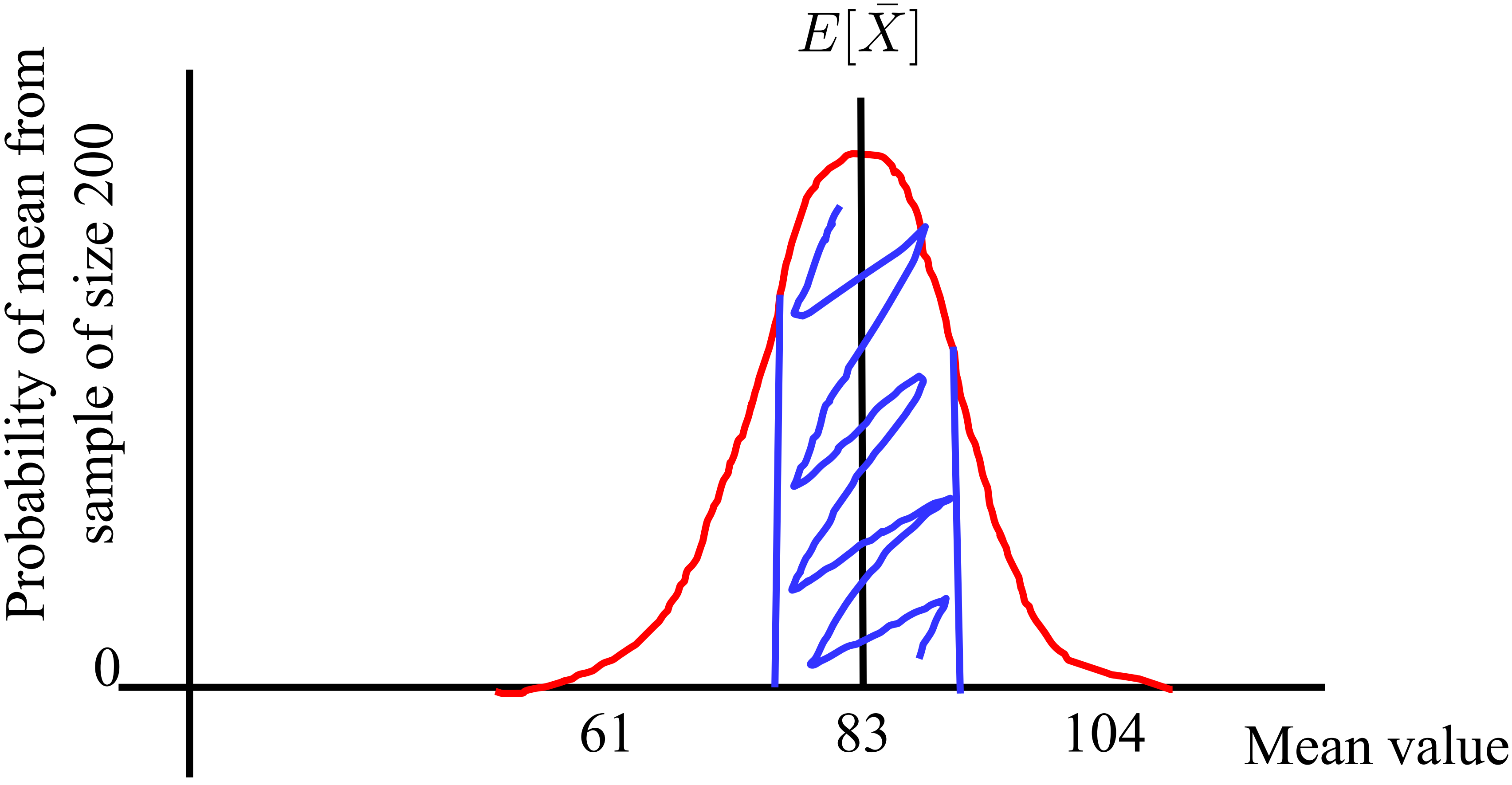
Bootstrap of Means

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

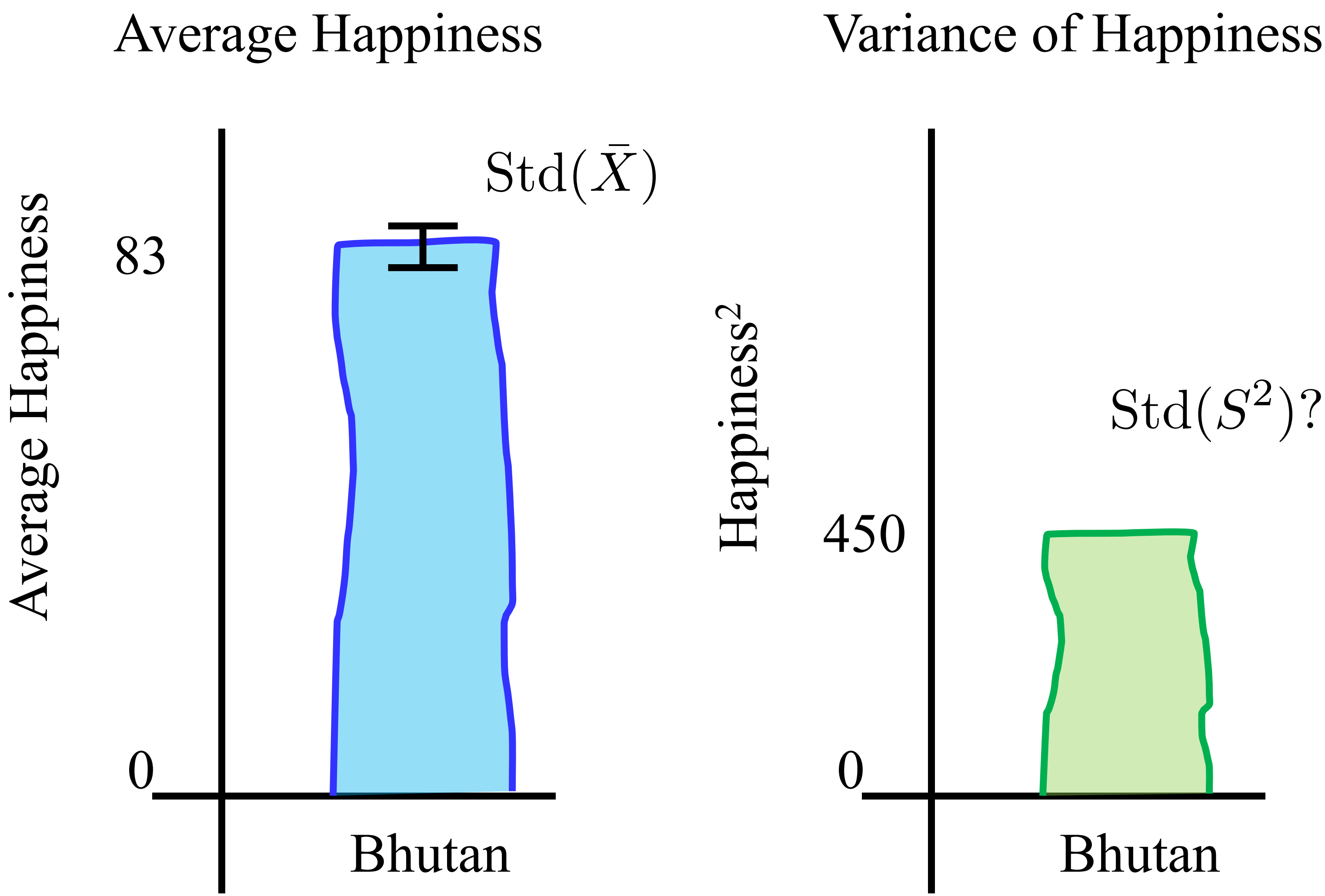


Bootstrap of Means

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



Sample Mean



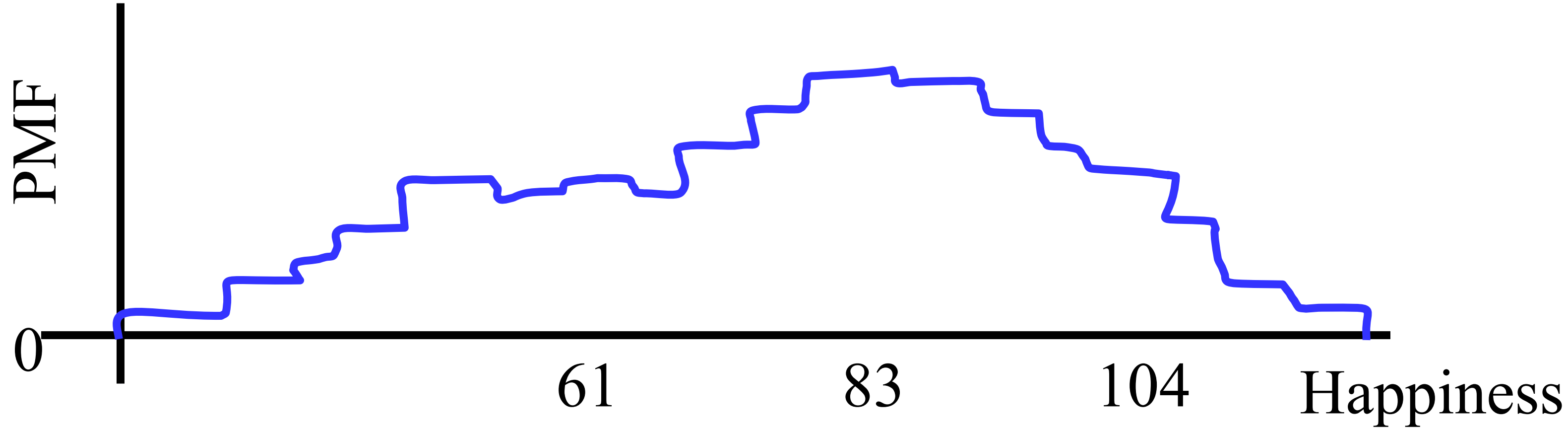
Claim: The average happiness of Bhutan is 83 ± 2

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**

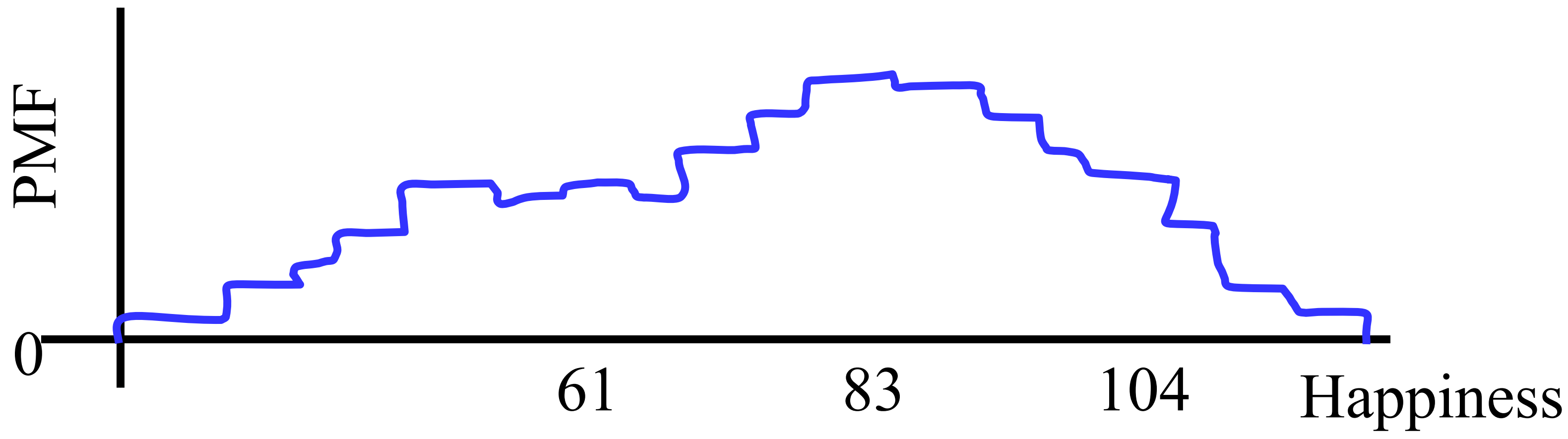
Bootstrap of Variance



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**

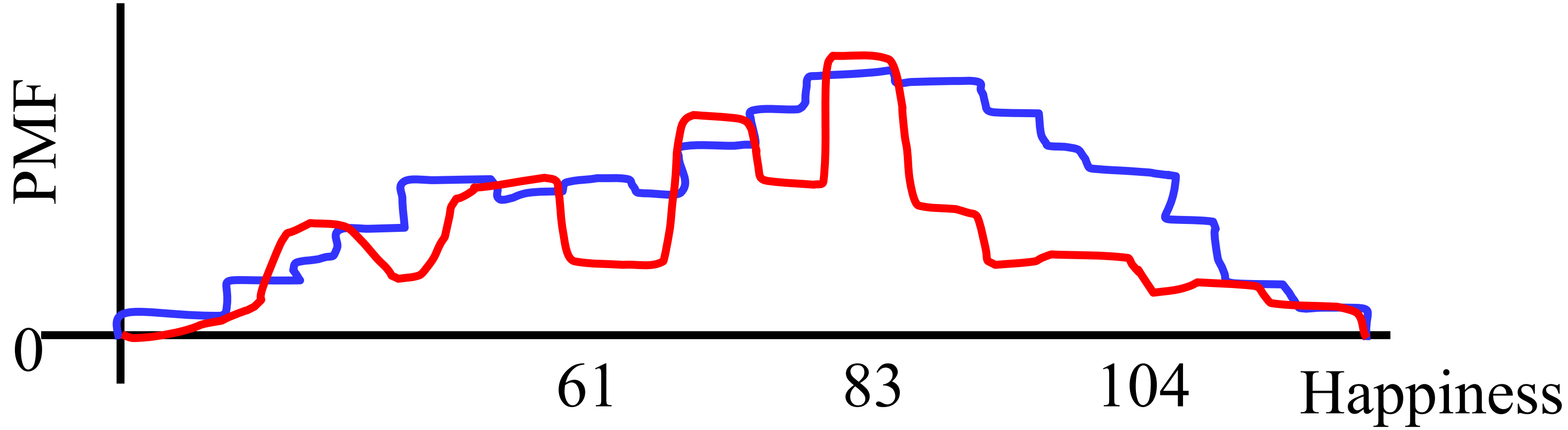
Bootstrap of Variance



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**

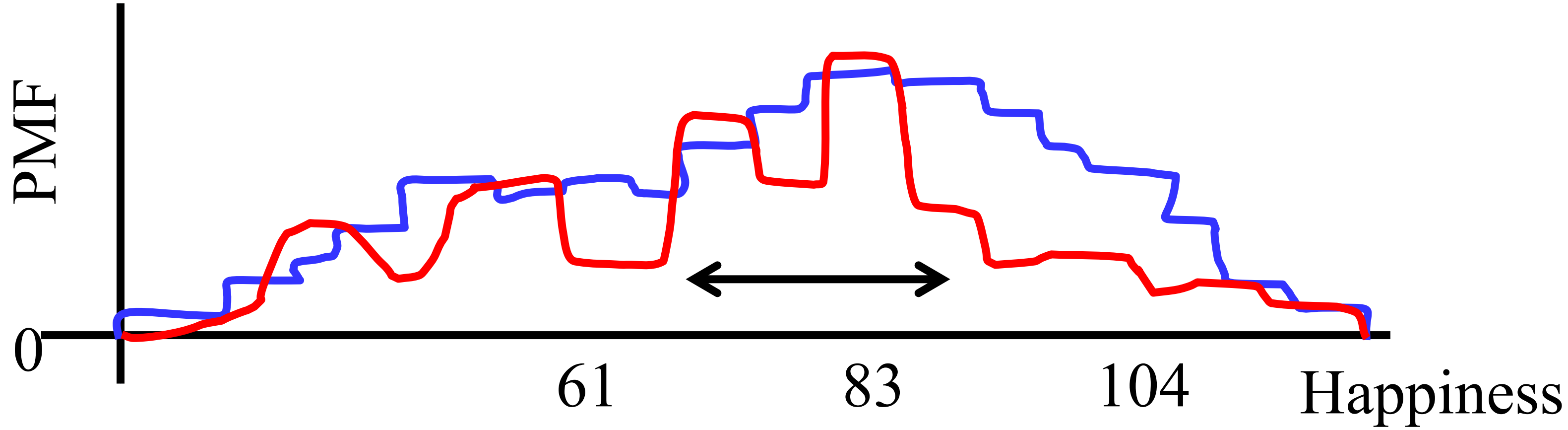
Bootstrap of Variance



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

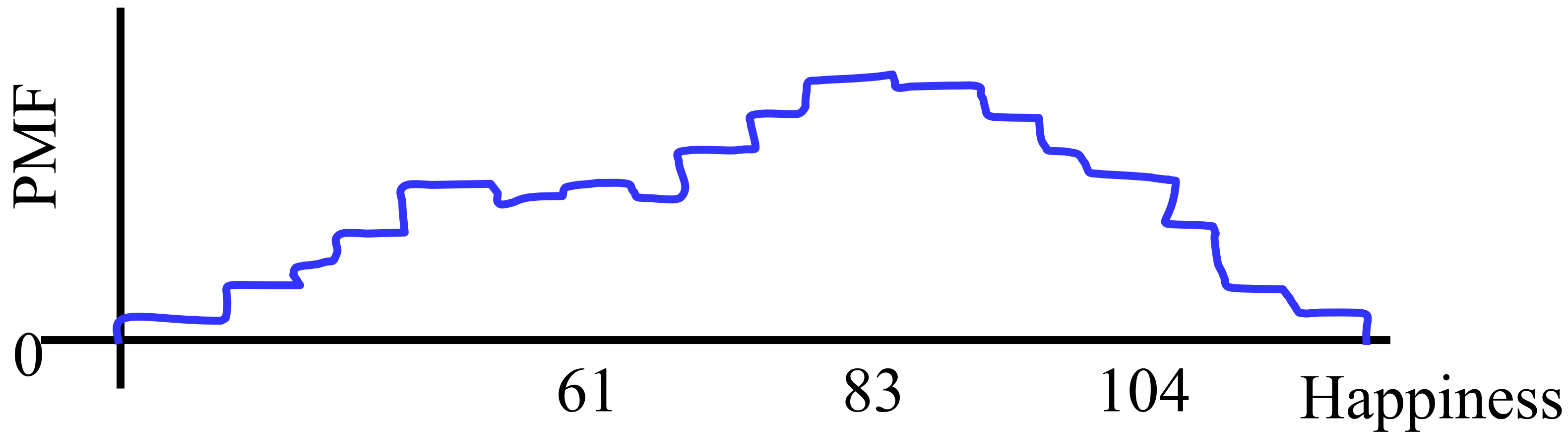


Bootstrap Algorithm (sample):

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

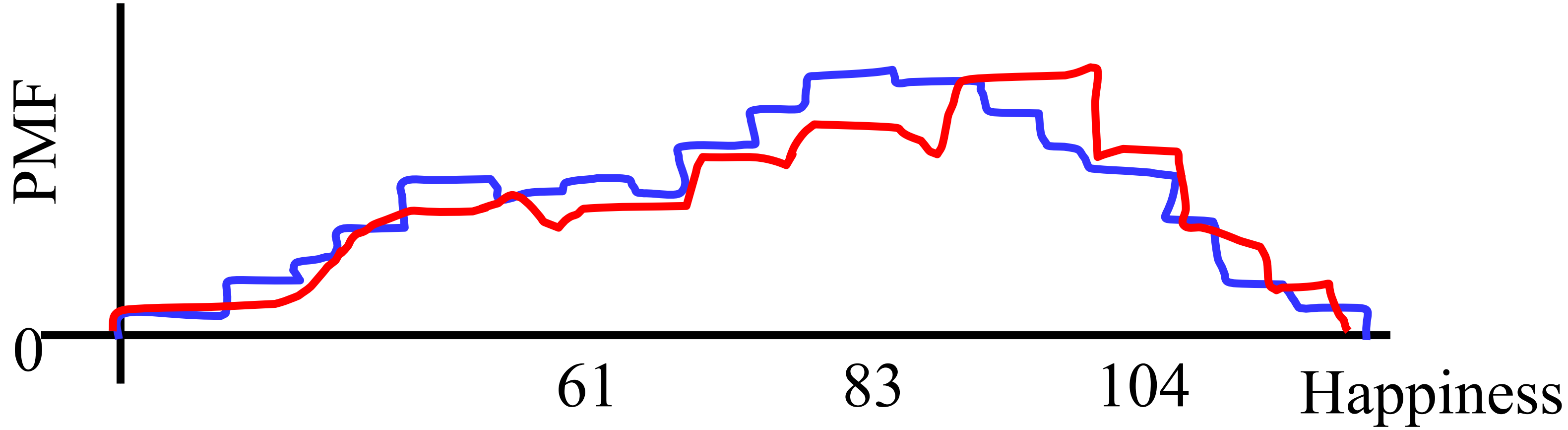


Bootstrap Algorithm (sample):

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

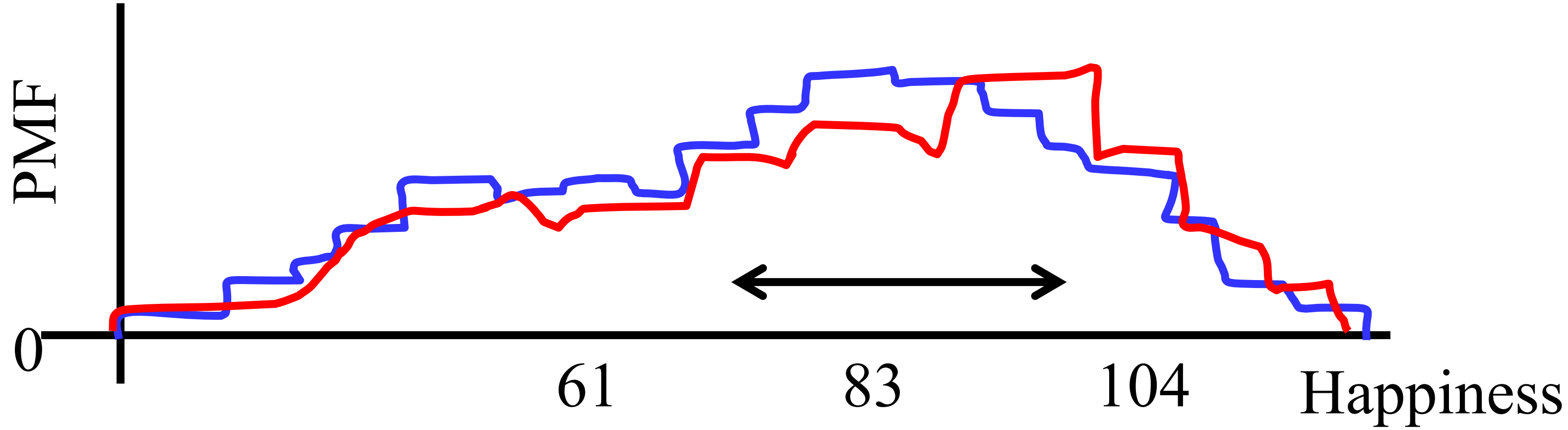


Bootstrap Algorithm (sample):

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

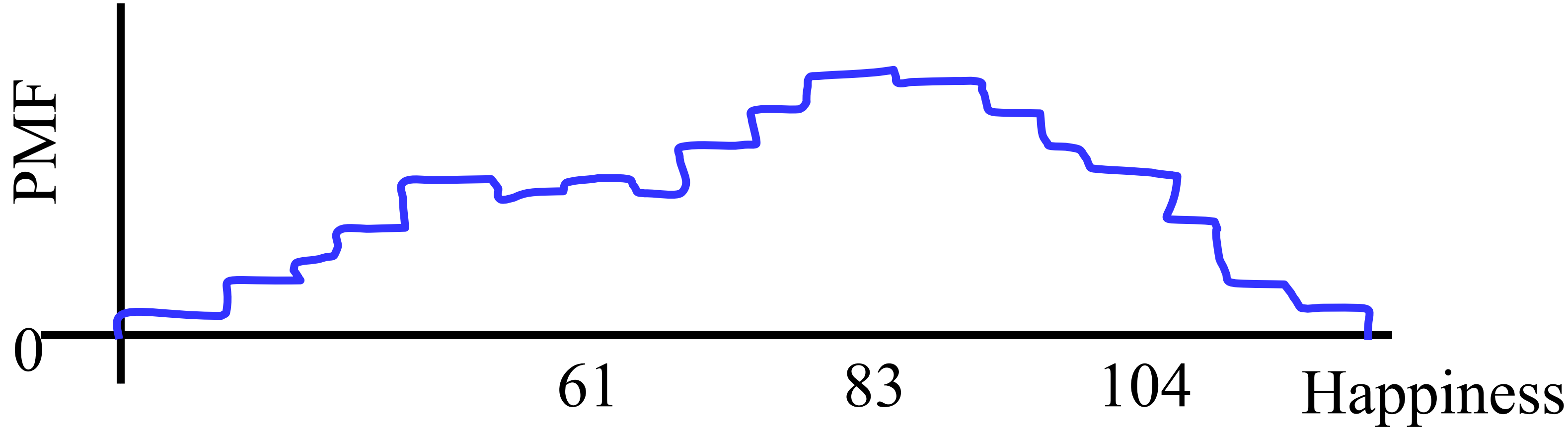


Bootstrap Algorithm (sample):

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

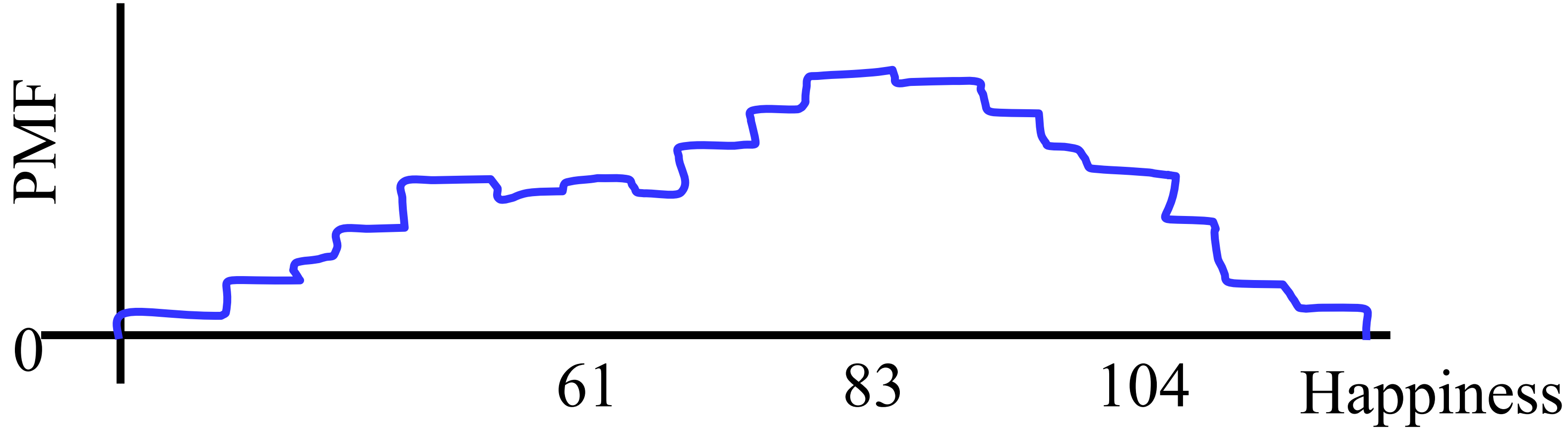


Bootstrap Algorithm (sample):

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



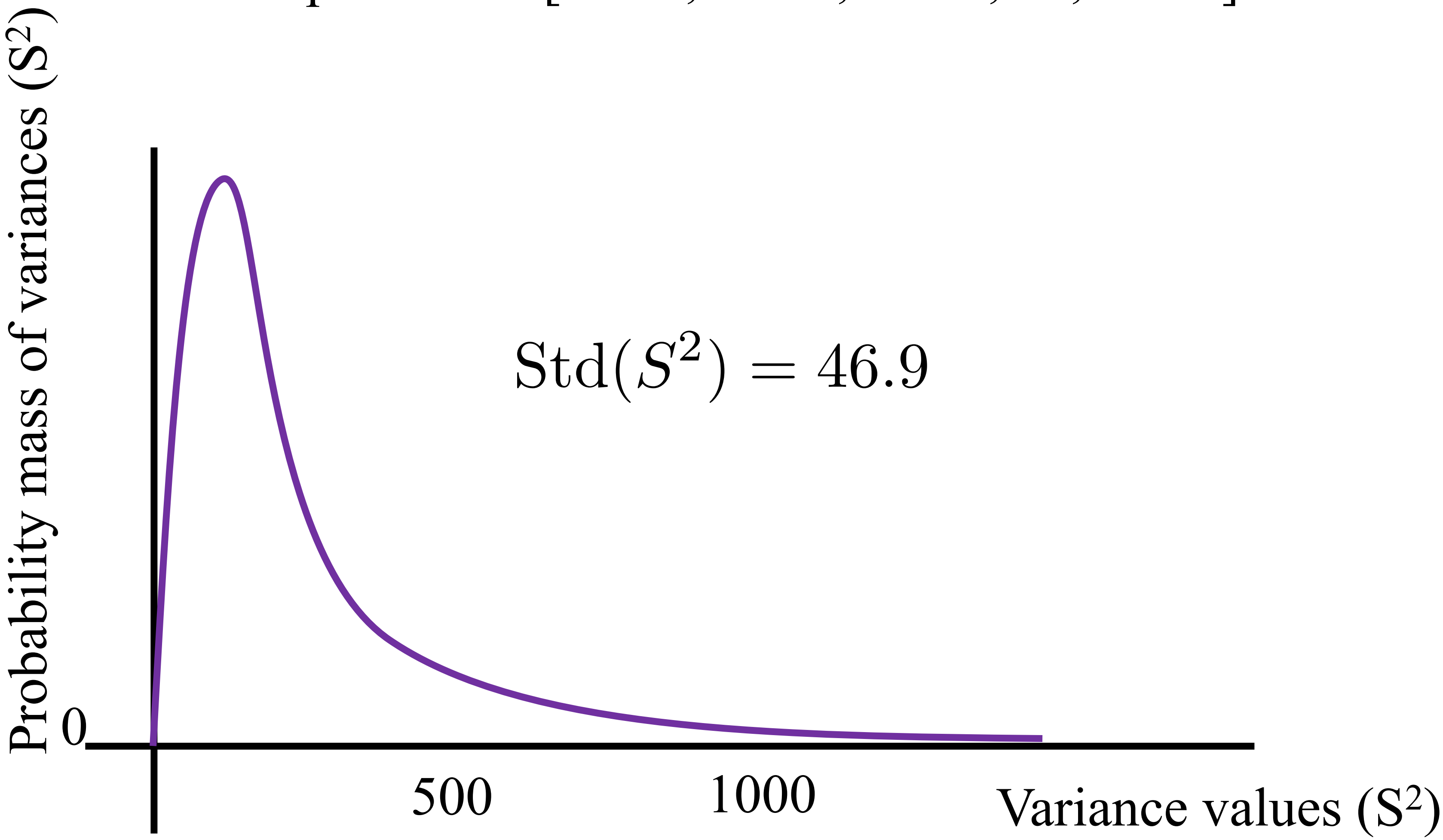
Bootstrap Algorithm (sample):

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]

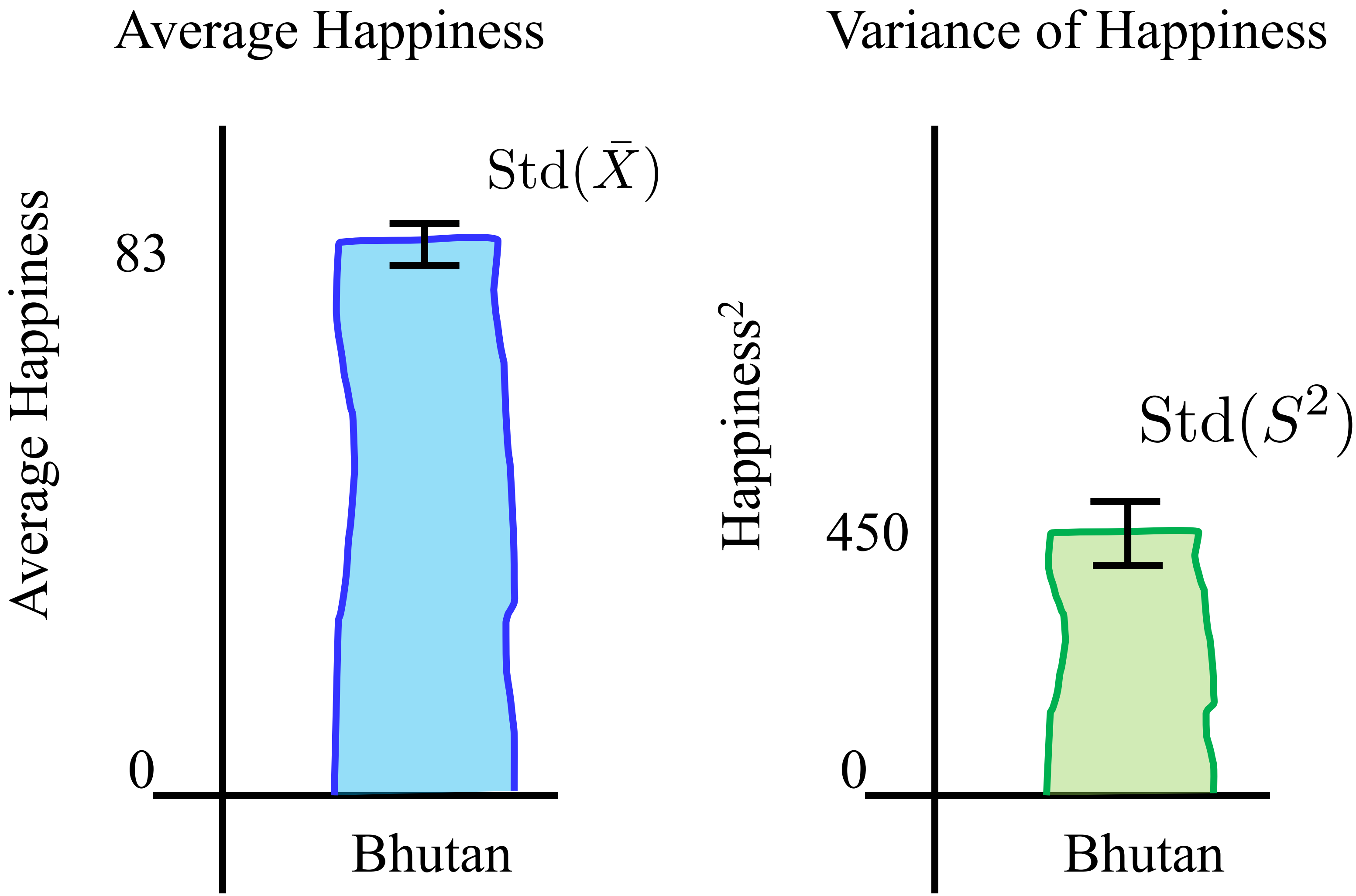
Bootstrap of Variance

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



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

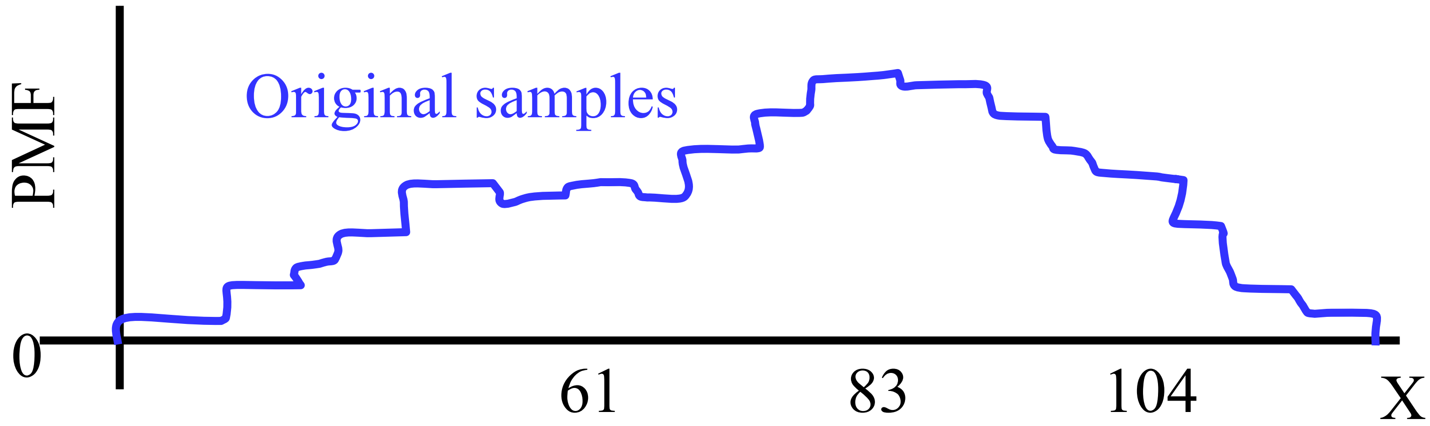
Sample Mean



Claim: The average happiness of Bhutan is 83 ± 2

Algorithm in Practice

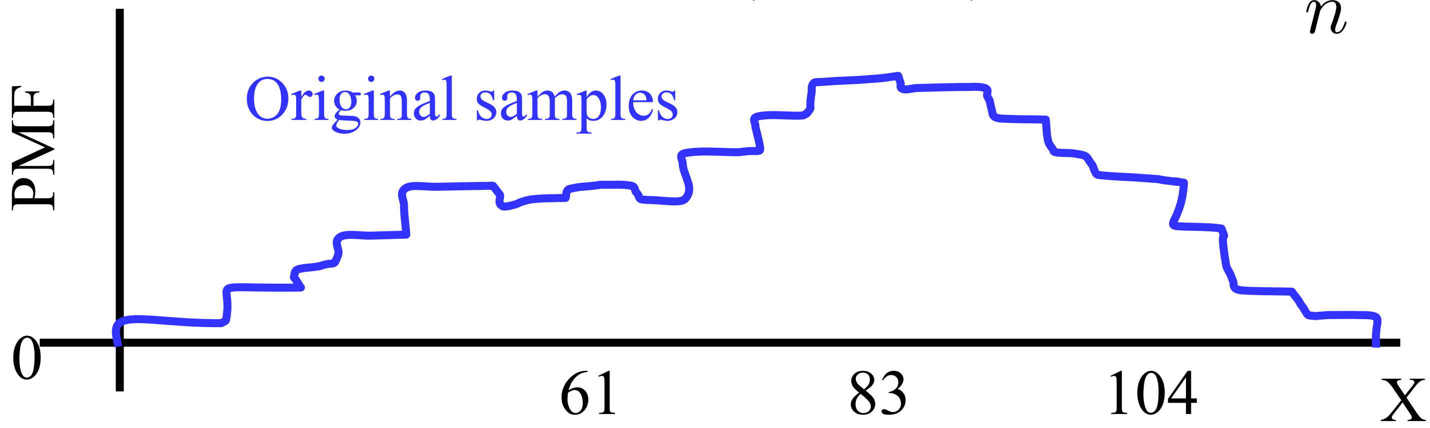
```
def resample(samples):
    # Estimate the PMF using the samples
    # Draw K new samples from the PMF
```



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

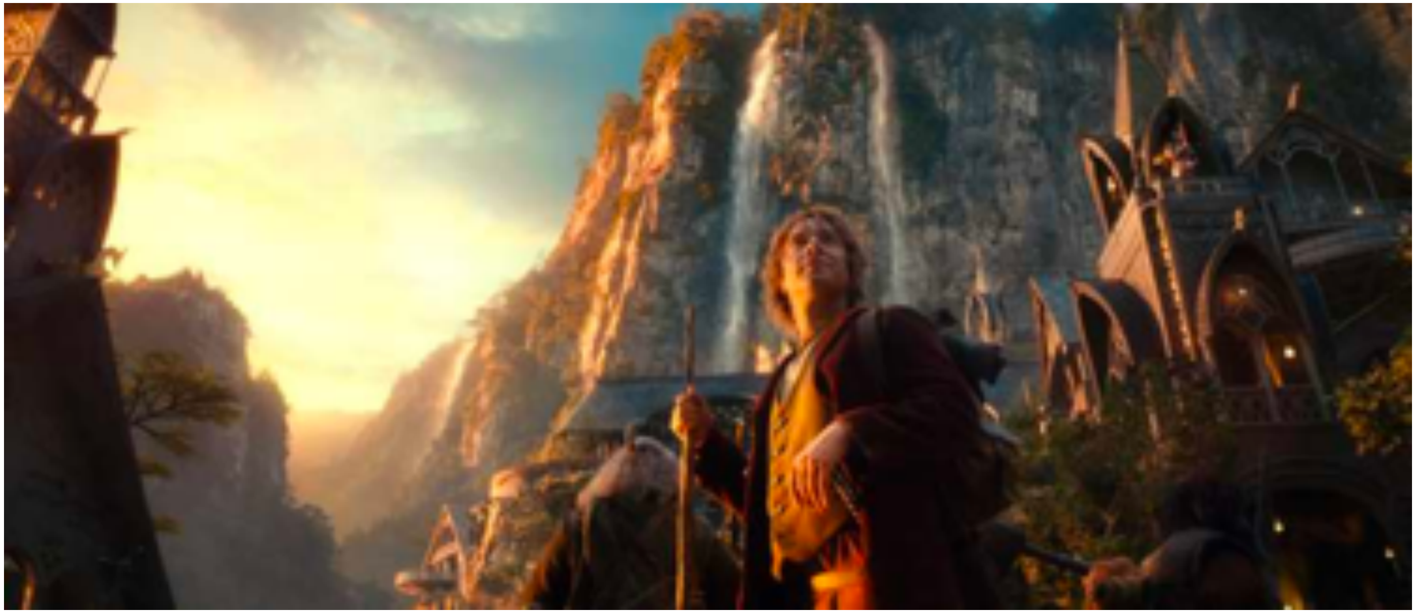
Bootstrap Algorithm (sample):

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!

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


Bootstrap

“The very language we use to describe the self-made ideal has these fault lines embedded within it: To ‘pull yourself up by your bootstraps’ is to succeed by dint of your own efforts. But that’s a modern corruption of the phrase’s original meaning. It used to describe a quixotic attempt to achieve an impossibility, not a feat of self-reliance. You can’t pull yourself up by your bootstraps, anymore than you can by your shoelaces. (Try it.) The phrase’s first known usage comes from a sarcastic 1834 account of a crackpot inventor’s attempt to build a perpetual motion machine.”

- John Swansburg

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
4.44
3.36
5.87
2.31
...
3.70

Population 2
2.15
3.01
2.02
1.43
...
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

Null Hypothesis Test

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

We want to know the probability of making this observation given the **Null Hypothesis**, which says that the populations come from the same distribution. This probability is the **p-value**.

Intuitively: if the p-value is high, then we would have been likely to make this observation even if there were no difference in the underlying distributions of the populations.

We can calculate this p-value with bootstrapping!

1. Combine the populations (Null Hypothesis)
2. Repeatedly resample Pop1, Pop2 with replacement.
3. See how often we get our observation (difference in means > 0.7)

Introduction to Likelihood

A snippet of my research ☺

Grades are not Normal: Improving Exam Score Models Using the Logit-Normal Distribution

NOAH ARTHURS
NARTHURS@CS.STANFORD.EDU

BEN STENHAUG
STENHAUG@STANFORD.EDU

SERGEY KARAYEV
SERGEYK@GRADESCOPE.COM

CHRIS PIECH
PIECH@CS.STANFORD.EDU

Stakes

- Models need priors over assignment scores.
- The “bell curve” deeply informs how we understand and interpret student performance.
- IRT has not been tuned to highly polytomous data.

Gradescope Data

- 4115 exams with 75 or more students
- Highly polytomous

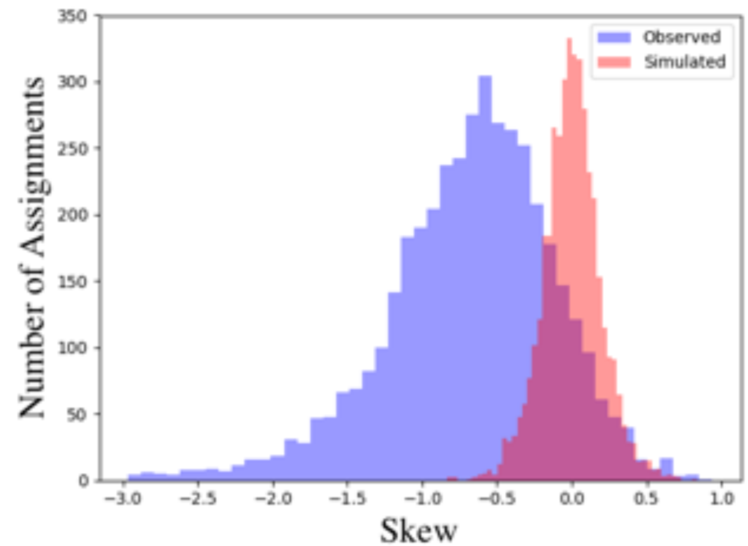
	Question 1	Question 2	Question 3	…
Student 1	4/7	6/10	7/12	…
Student 2	7/7	8/10	11/12	…
…	…	…	…	…

Gradescope will make data available to researchers.

Why not Normal?

1. Normal is unbounded

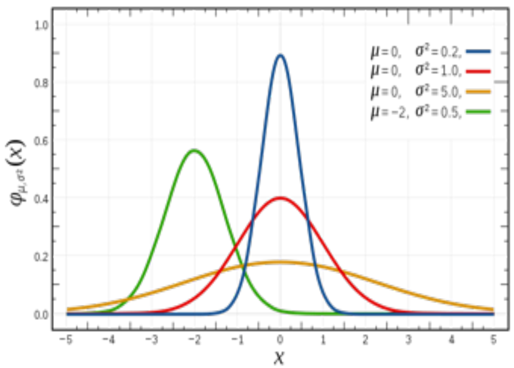
2. Normal is symmetric



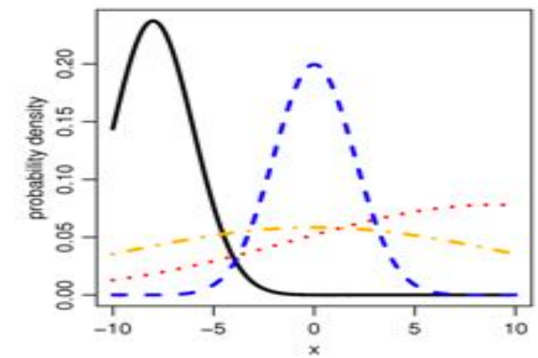
3. Danger of the “bell curve mindset”

Distribution Candidates (Two Parameters)

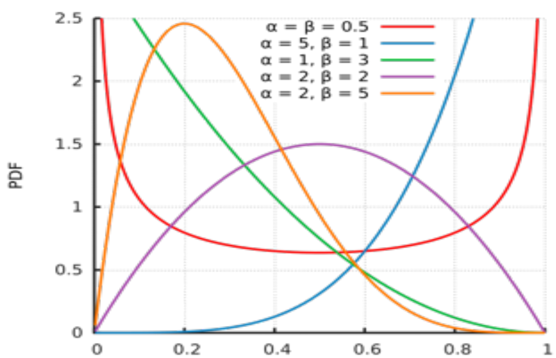
Normal



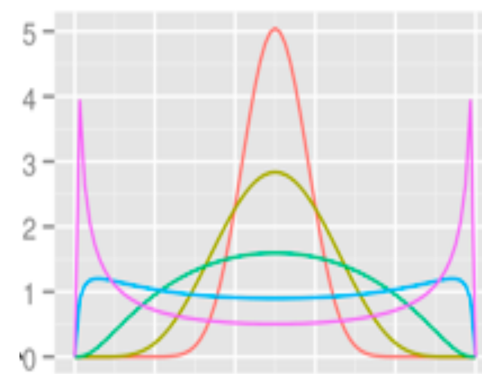
Truncated Normal



Beta



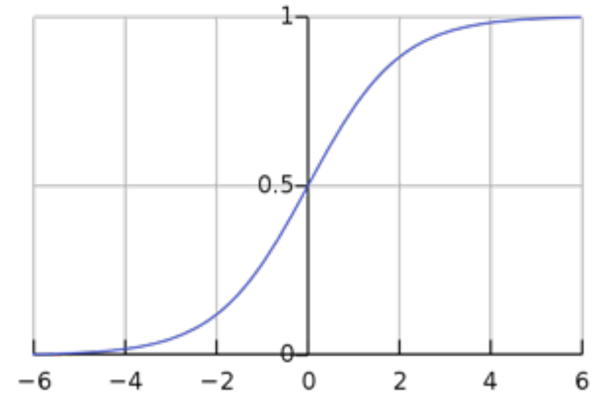
Logit-Normal



The Logit Normal

The result of applying a sigmoid to normally distributed data:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Parameters are μ, σ, the parameters of the underlying normal.

Distribution Methodology and Results

Distributions were fit to assignments using MLE and evaluated using Log Likelihood:

$$LL(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log f(x_i|\theta)$$

	Normal	Trunc	Beta	Logit
Beats Normal	-	100%	92%	87%
Beats Trunc	0	-	67%	75%
Beats Beta	8%	33%	-	68%
Beats Logit	13%	25%	32%	-
Average LL	**0.272**	**0.333**	**0.336**	**0.353**

Distribution Results