Bootstrapping

Chris Piech
CS109, Stanford University

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

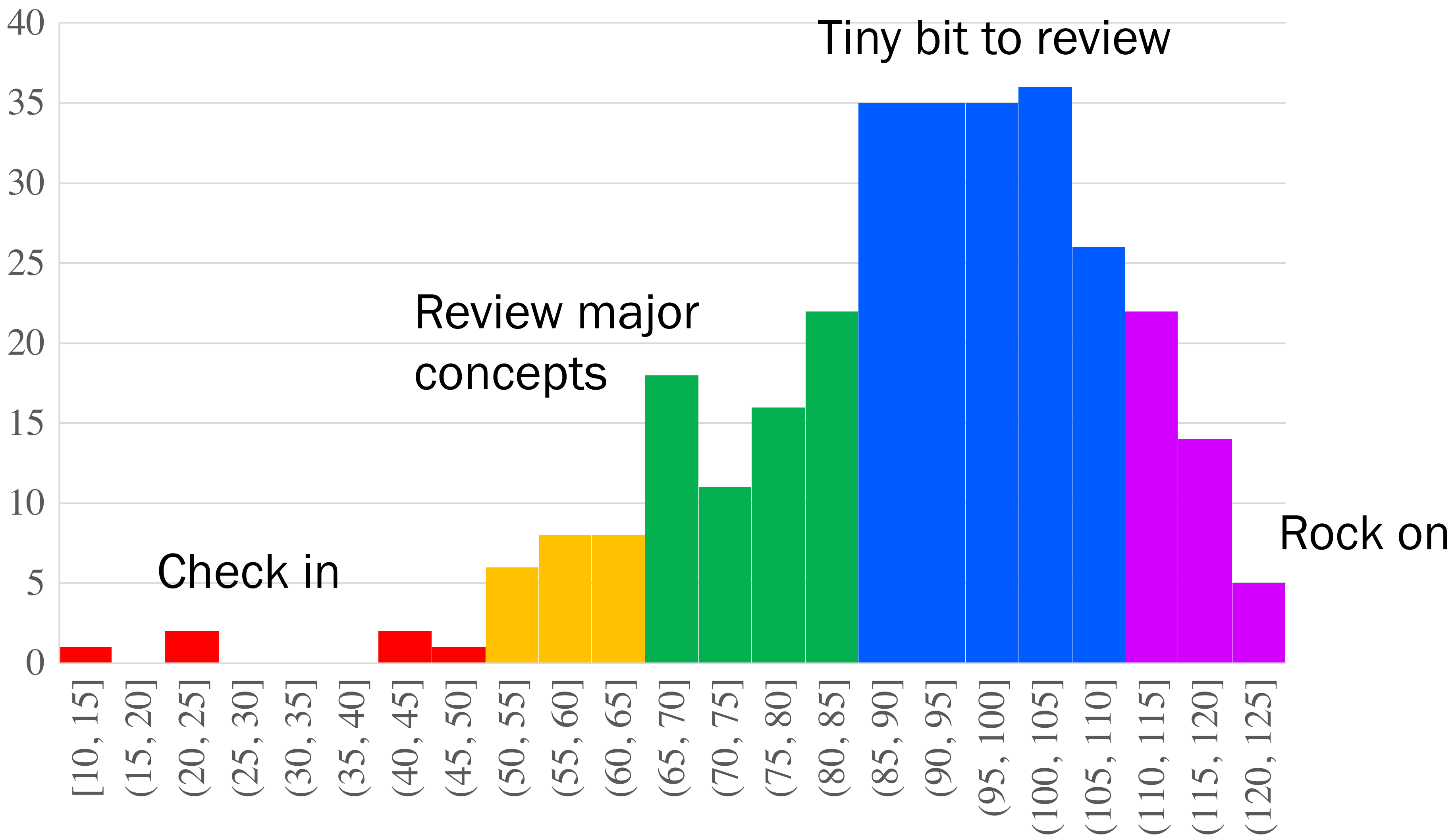
Central Limit Theorem

Sampling

Bootstrapping

Algorithmic Analysis

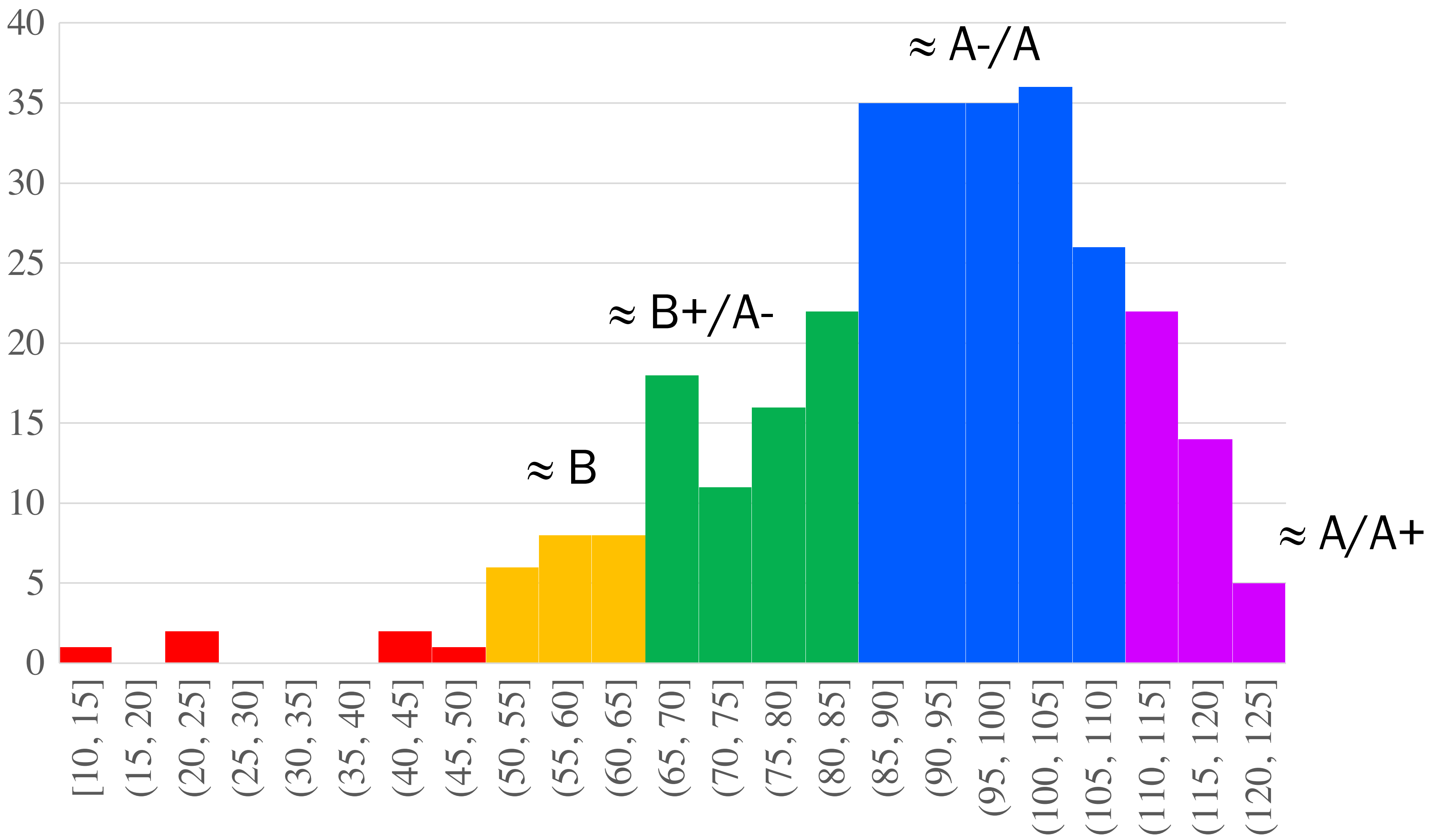
Grade Distribution



$$\mu = \frac{91}{120}$$

$$\sigma = \frac{19}{120}$$

Grade Distribution



$$\mu = \frac{91}{120}$$

$$\sigma = \frac{19}{120}$$

How should you normalize exam scores?

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

Noah Arthurs
Stanford University
narthurs@cs.stanford.edu

Ben Stenhaug
Stanford University
stenhaug@stanford.edu

Sergey Karayev
Gradescope
sergeyk@gradescope.com

Chris Piech
Stanford University
piech@cs.stanford.edu

ABSTRACT

Understanding exam score distributions has implications for item response theory (IRT), grade curving, and downstream modeling tasks such as peer grading. Historically, grades have been assumed to be normally distributed, and to this day the normal is the ubiquitous choice for modeling exam scores. While this is a good assumption for tests comprised of equally-weighted dichotomous items, it breaks down on the highly polytomous domain of undergraduate-level exams. The logit-normal is a natural alternative because it is has a bounded range, can represent asymmetric distributions, and lines up with IRT models that perform logistic transformations on normally distributed abilities. To tackle this question, we analyze an anonymized dataset from Gradescope consisting of over 4000 highly polytomous undergraduate exams. We show that the logit-normal better models this data without having more parameters than the normal. In addition, we propose a new continuous polytomous IRT model that reduces the number of item-parameters by using a logit-normal assumption at the item level.

1. INTRODUCTION

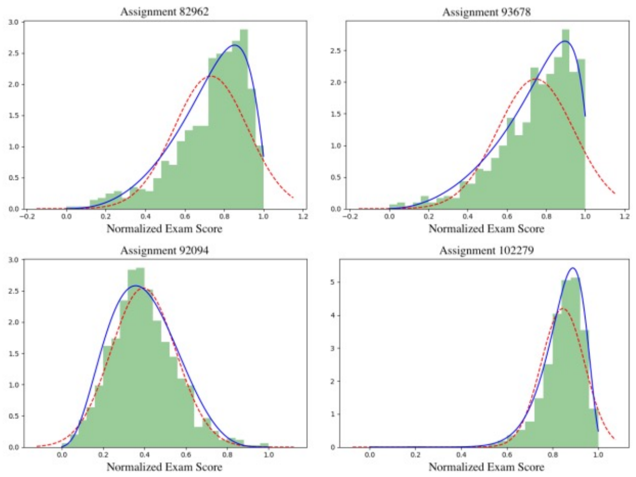
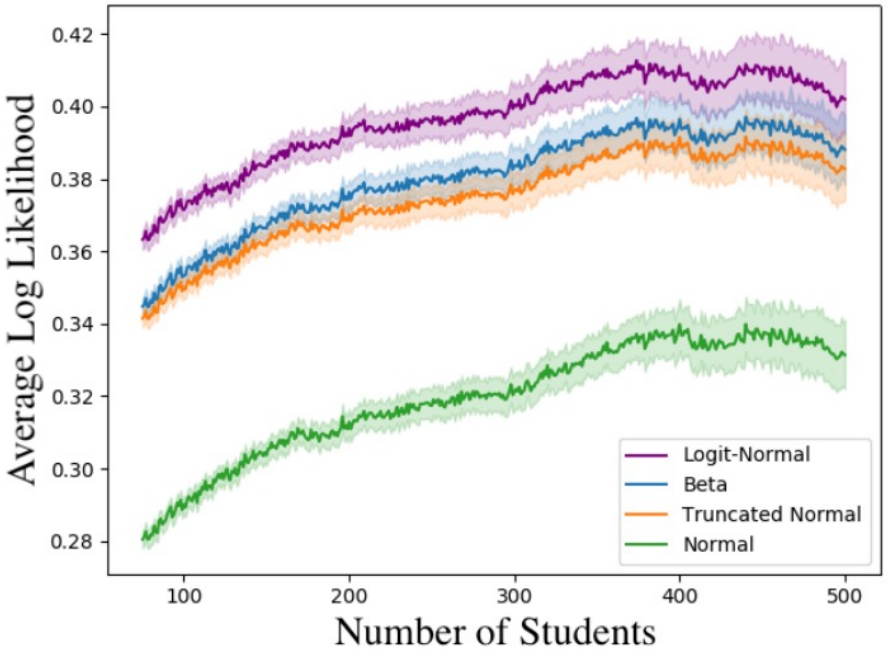


Figure 1: Score histograms of four assignments, along with the PDFs of the best-fit normals (dashed red) and best-fit logit-normals (solid blue).



Our Goal in Grading

Let G be the grade that you get in the class
Let S be the score that you get on a midterm
Let D be the difficulty of the midterm

$$P(G = g) = P(G = g | D = d)$$

For any value g and for any value d

Improvement Between Midterm and Final

Bad midterm? The final can show me you have learned

DEC 5TH

A real difference?

18 students

Learning in Context A
4.44
3.36
5.87
2.31
…
3.70

$\mu_1 = 3.1$

23 students

Learning in Context B
2.15
3.01
2.02
1.43
…
1.83

$\mu_2 = 2.4$

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>

Central Limit Theorem (Summation)

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

As $n \to \infty$

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

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

Central Limit Theorem (Average)

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

As $n \to \infty$

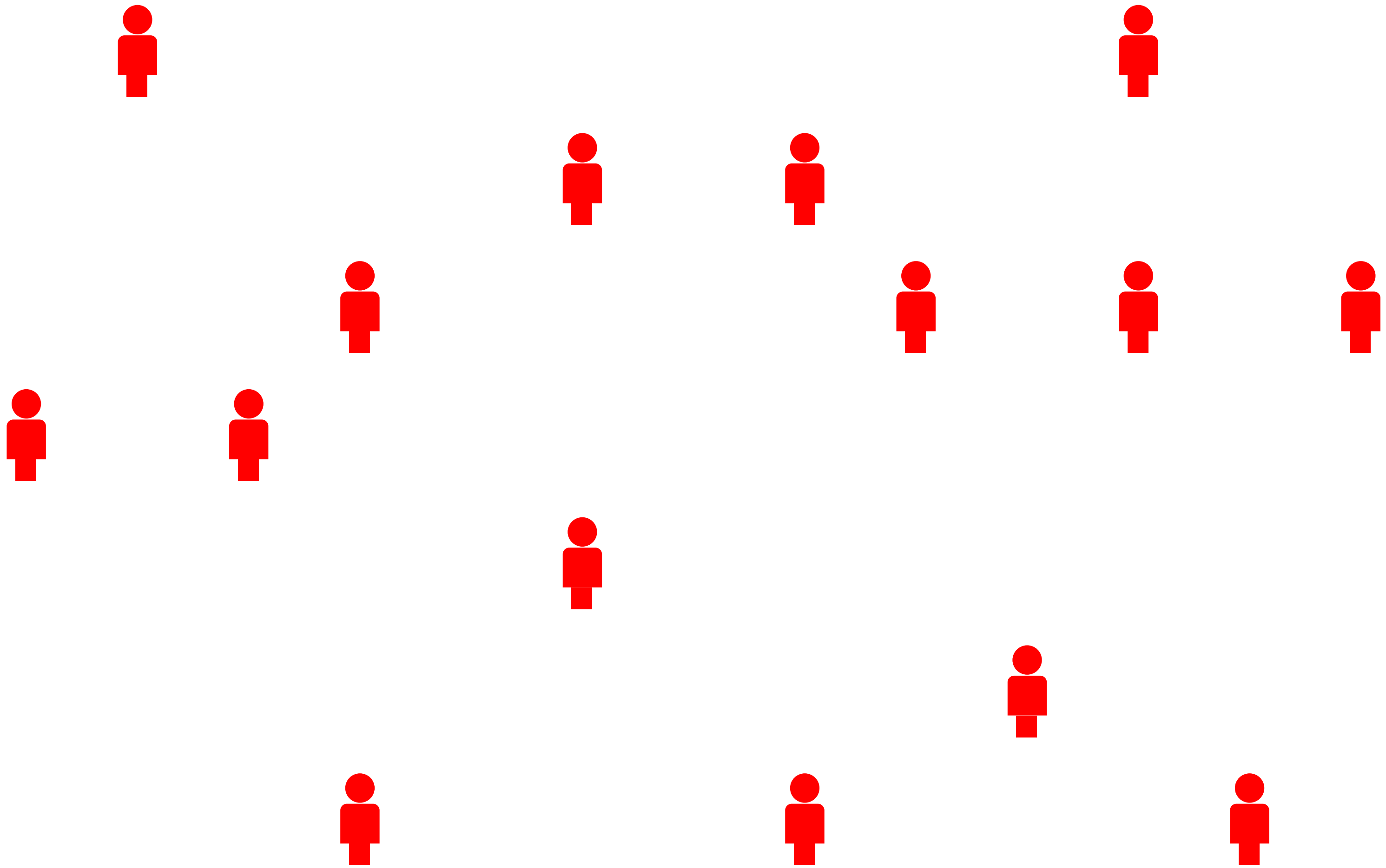
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

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

Population

Sample

Sample



Collect one (or more) numbers from each person

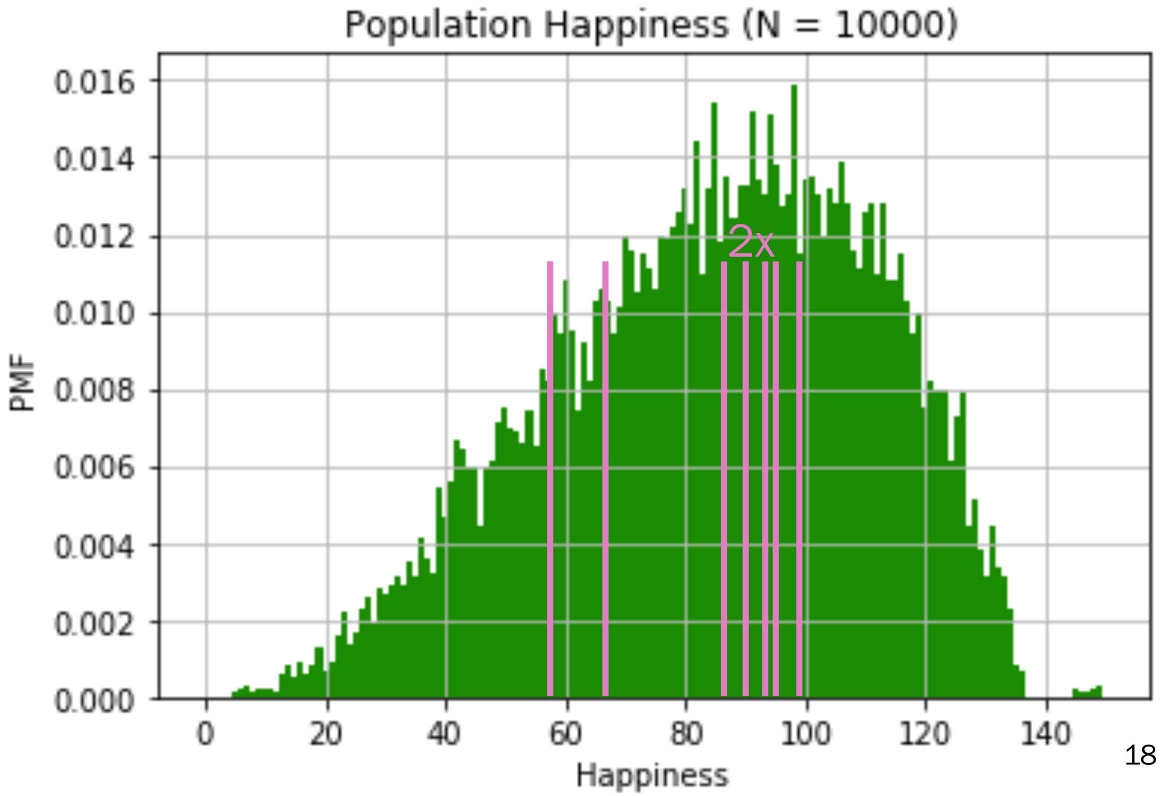
A sample, mathematically

A sample of **sample size** 8:

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

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

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



Equations we used to get those values

sample mean estimate

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

Our best guess at the true mean

sample mean

sample variance estimate

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

Our best guess at the true variance

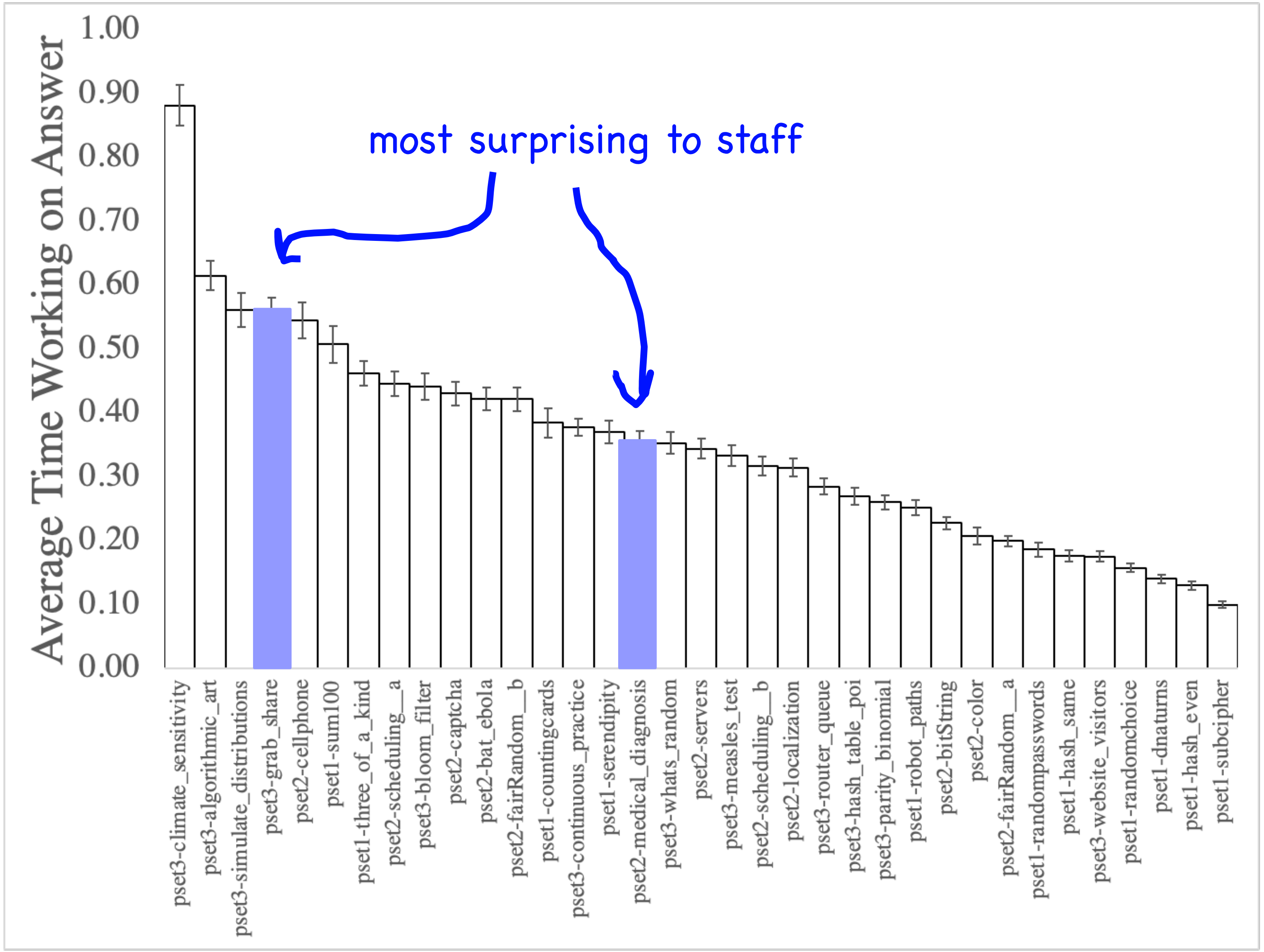
Std error of the mean estimate

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

sample variance

How wrong do we think our mean estimate is?

Sample Mean and Standard Error for PSets



Error bars are standard error of the mean

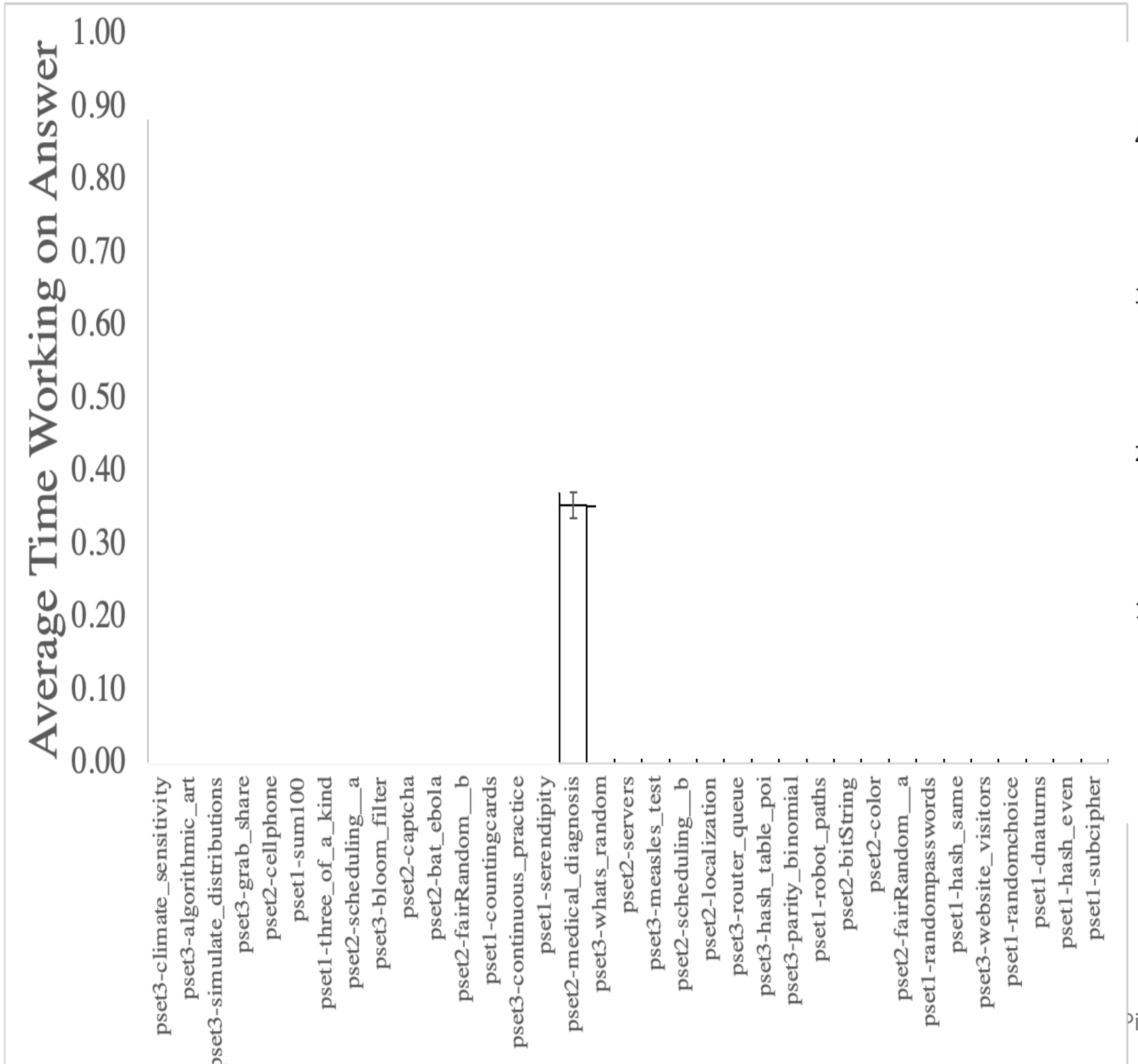
Expectation of the sum of problems is sum of expectations:

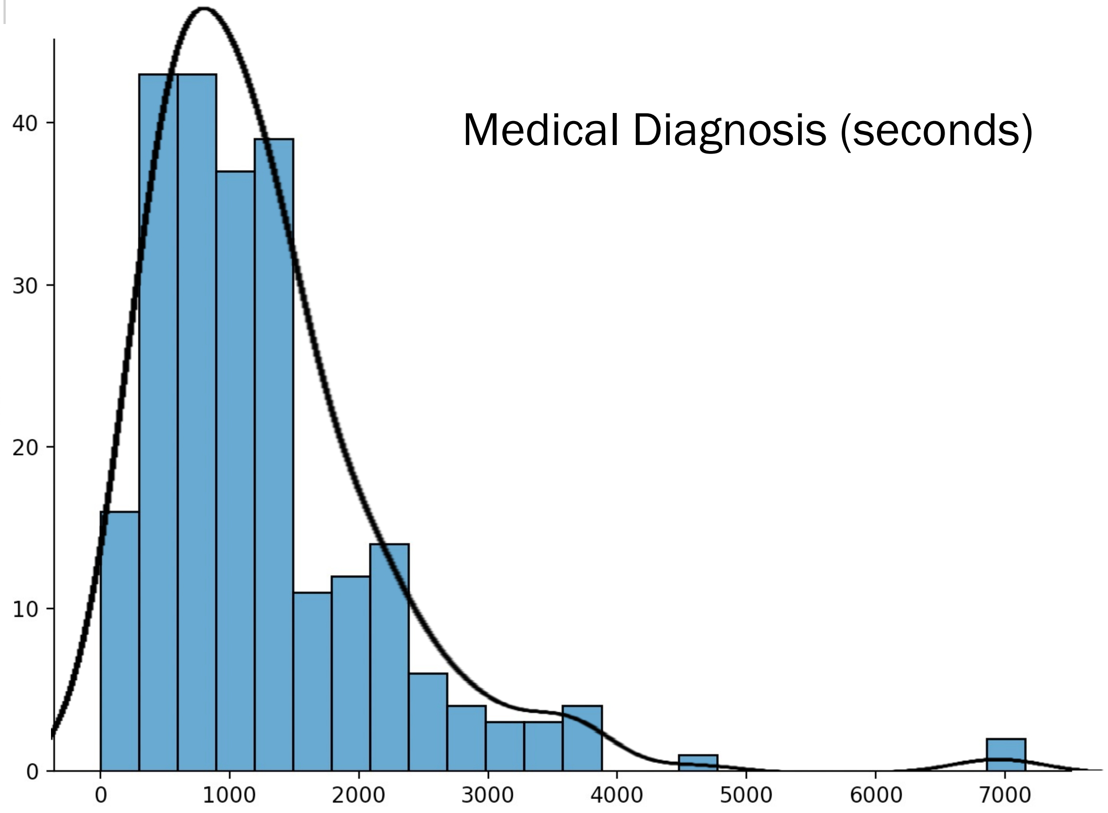
pset1: 2.87 hours on answers
pset2: 4.23 hours on answers
pset3: 5.11 hours on answers

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

Statistics Vs Distribution

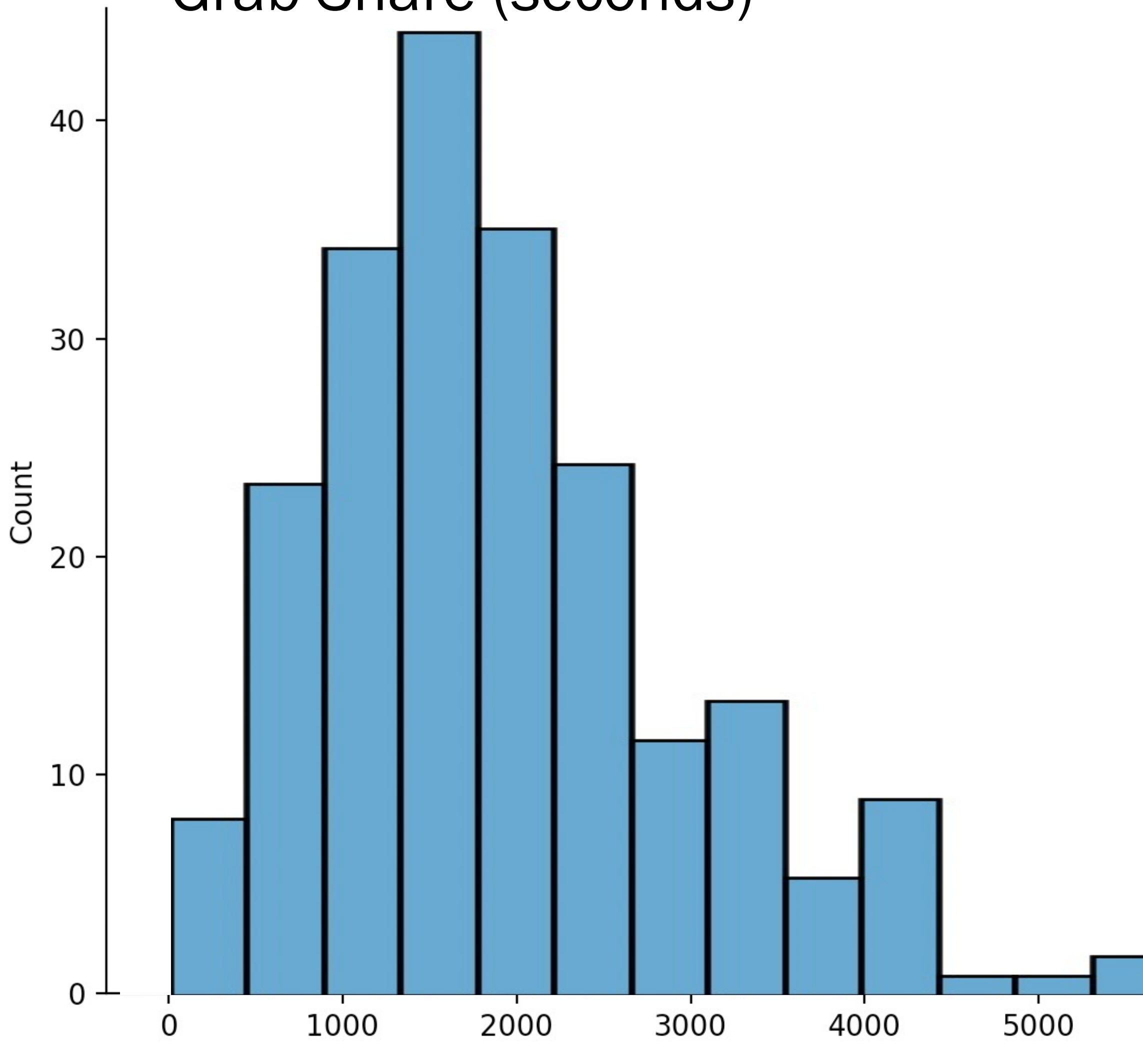
Sampling statistics vs Sampling distribution





[Aside] Distribution of PSet Completion Times

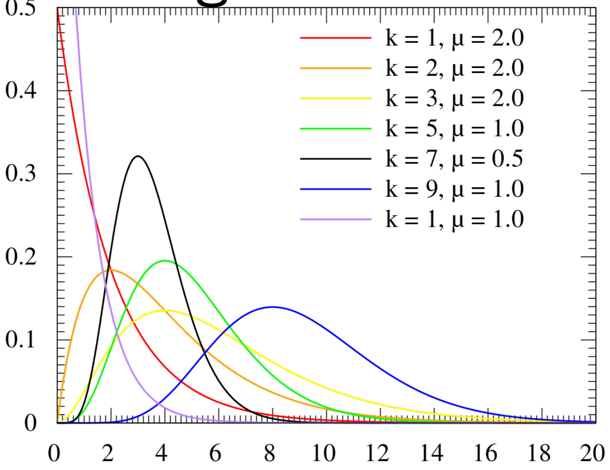
Grab Share (seconds)



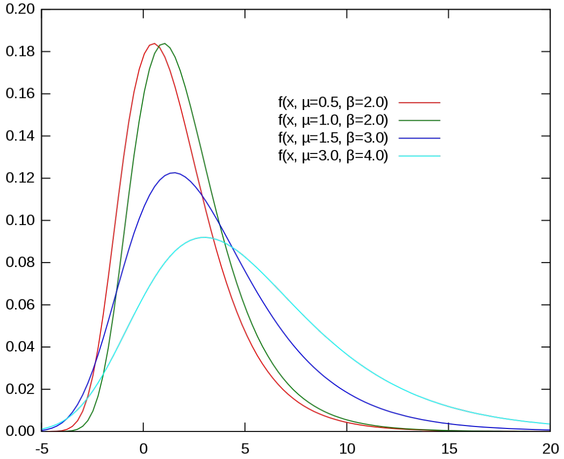
Aside: could be Erlang, could be Gumbel

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

Erlang



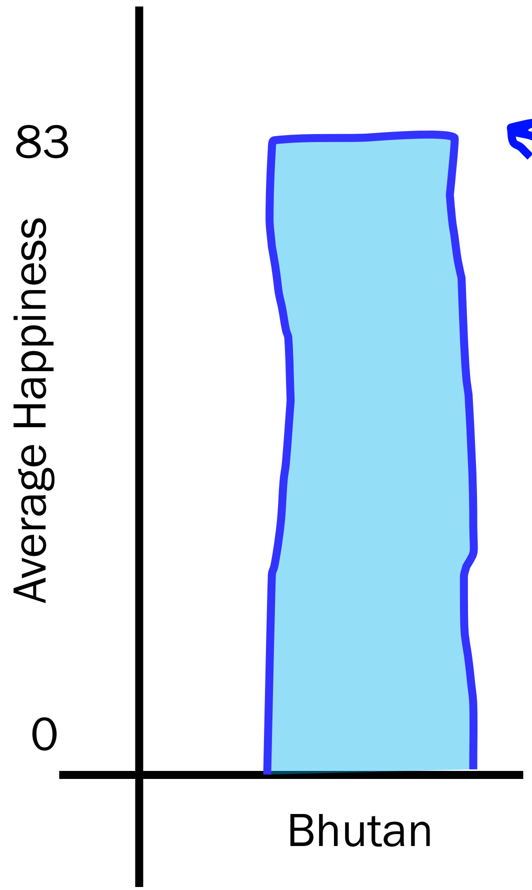
Gumbel



But what about Bhutan?

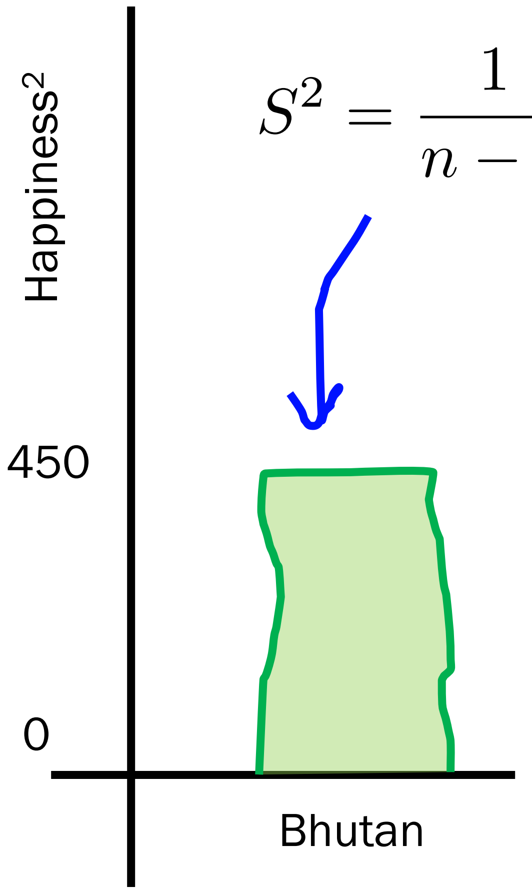
Our Report to Bhutan Government (after talking to 200 ppl)

Average Happiness



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

Variance of Happiness

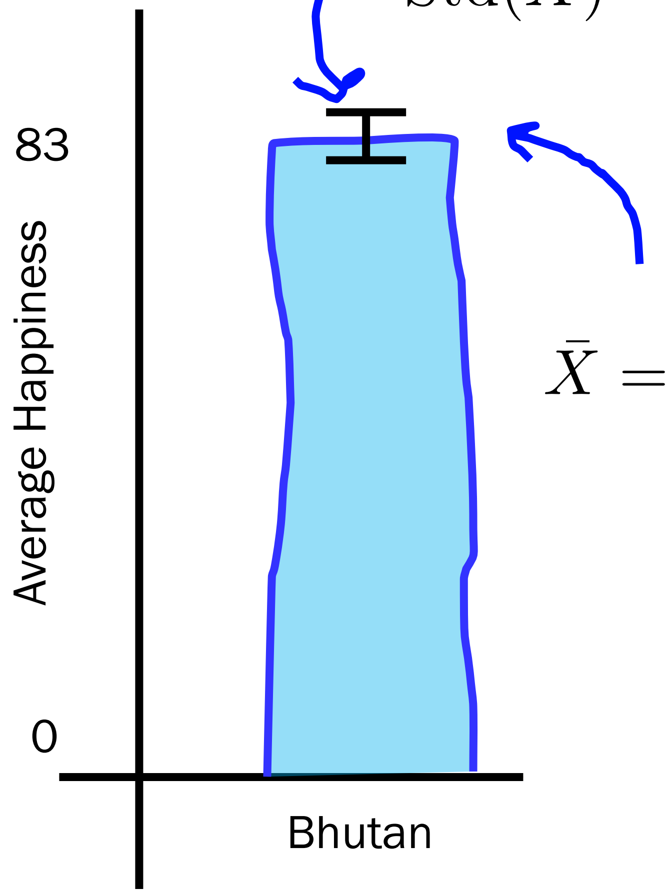


$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

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

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

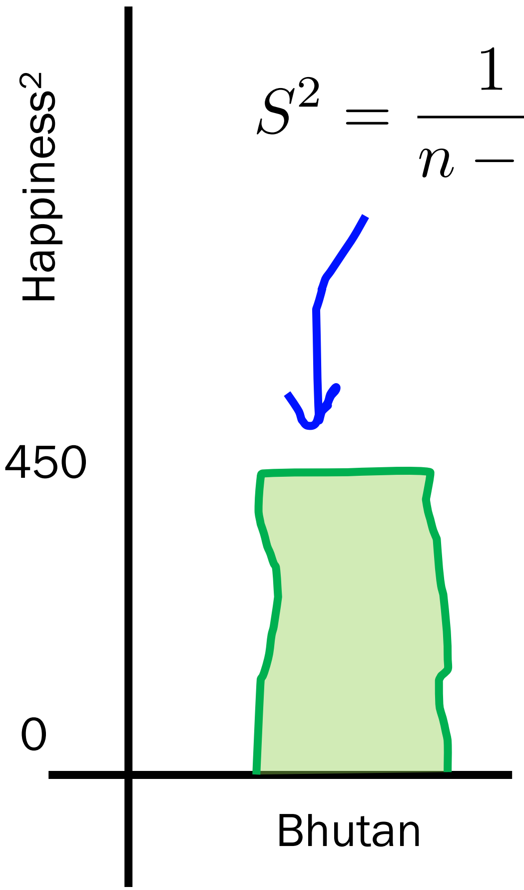
Average Happiness



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

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

Variance of Happiness



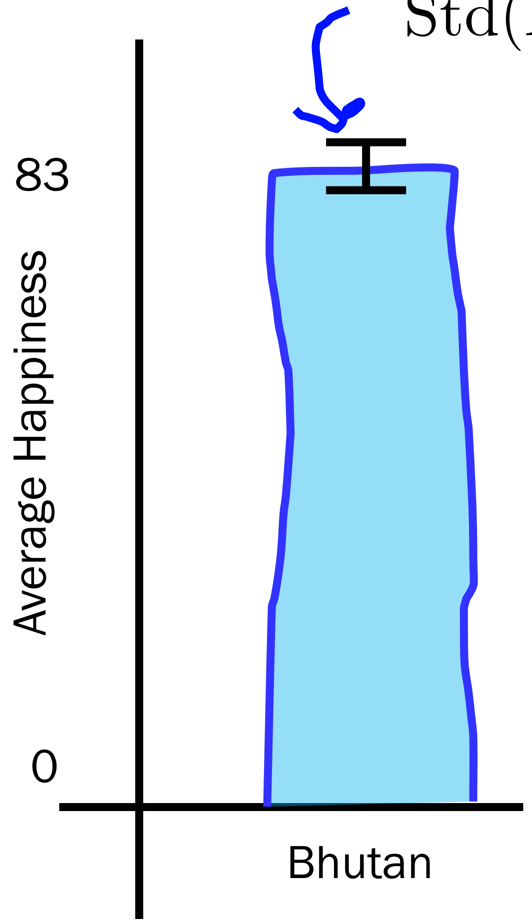
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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

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

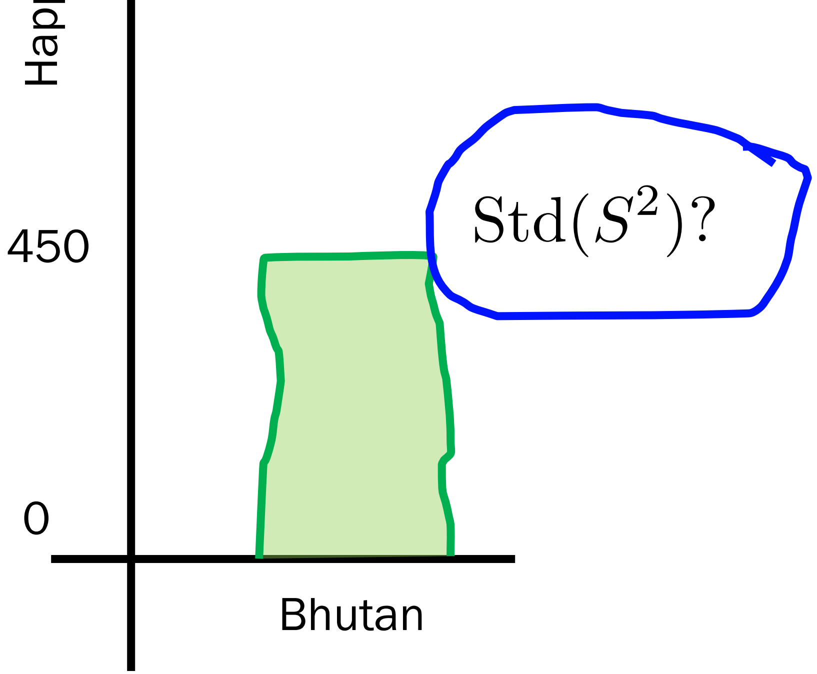
Average Happiness

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



Variance of Happiness

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



[suspense]

Bootstrap:
Probability for Computer Scientists

Bootstraping allows you to:

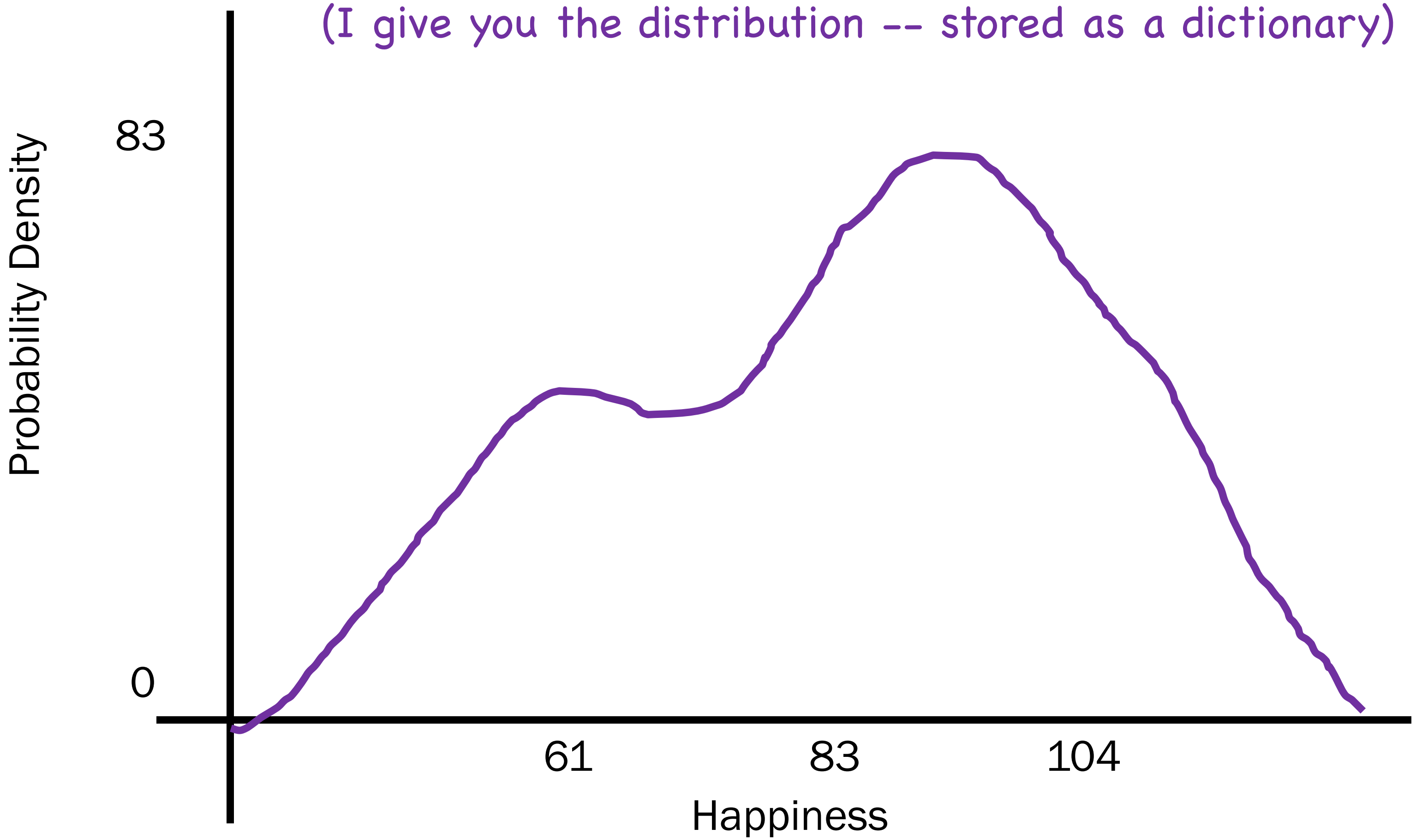
- Know the **distribution of *statistics***
- Calculate **p values**
- **Using computers**

Hypothetical

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

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

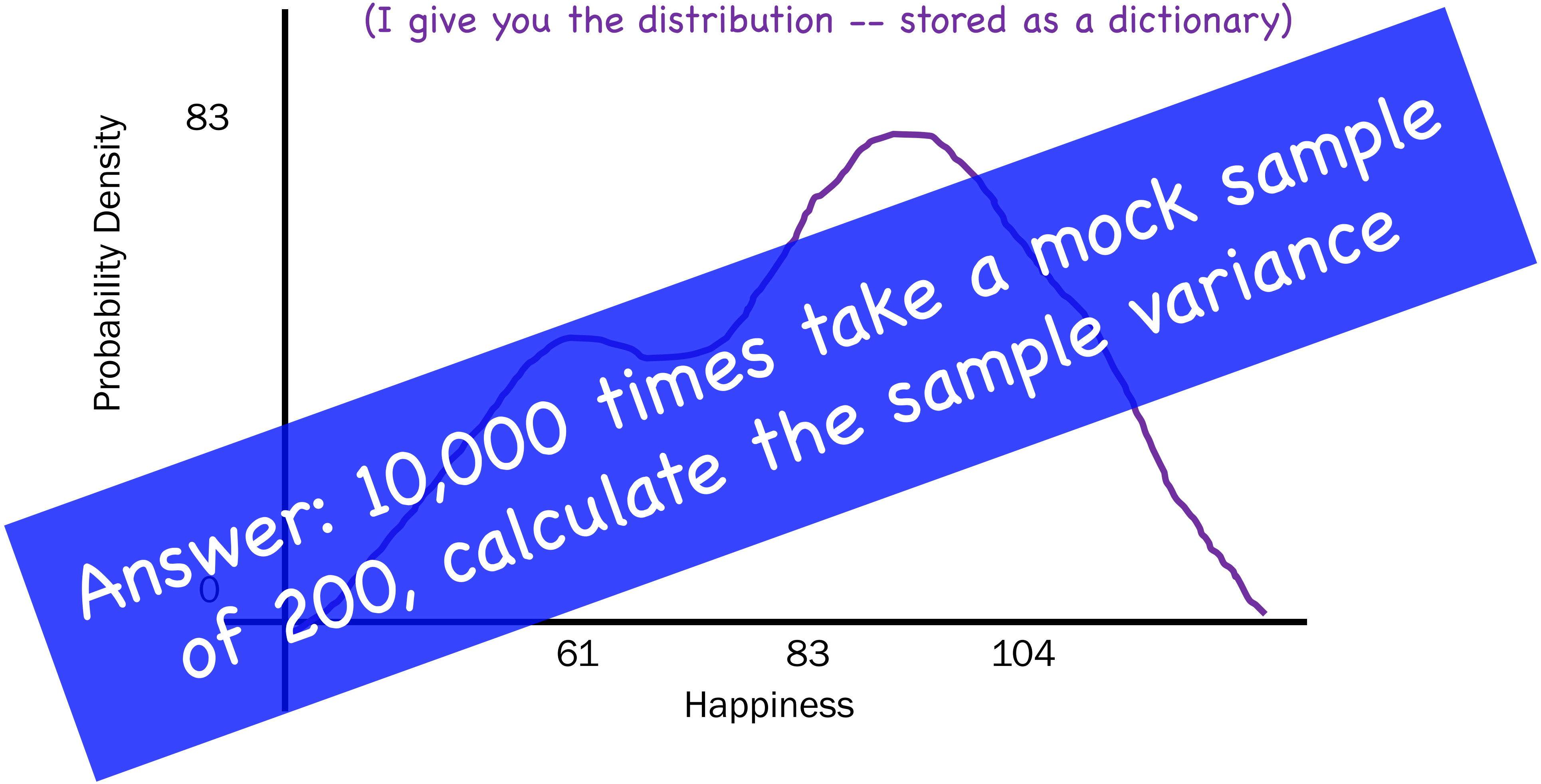
What is the **std** of the **sample variance**, calculated from 200 people?



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

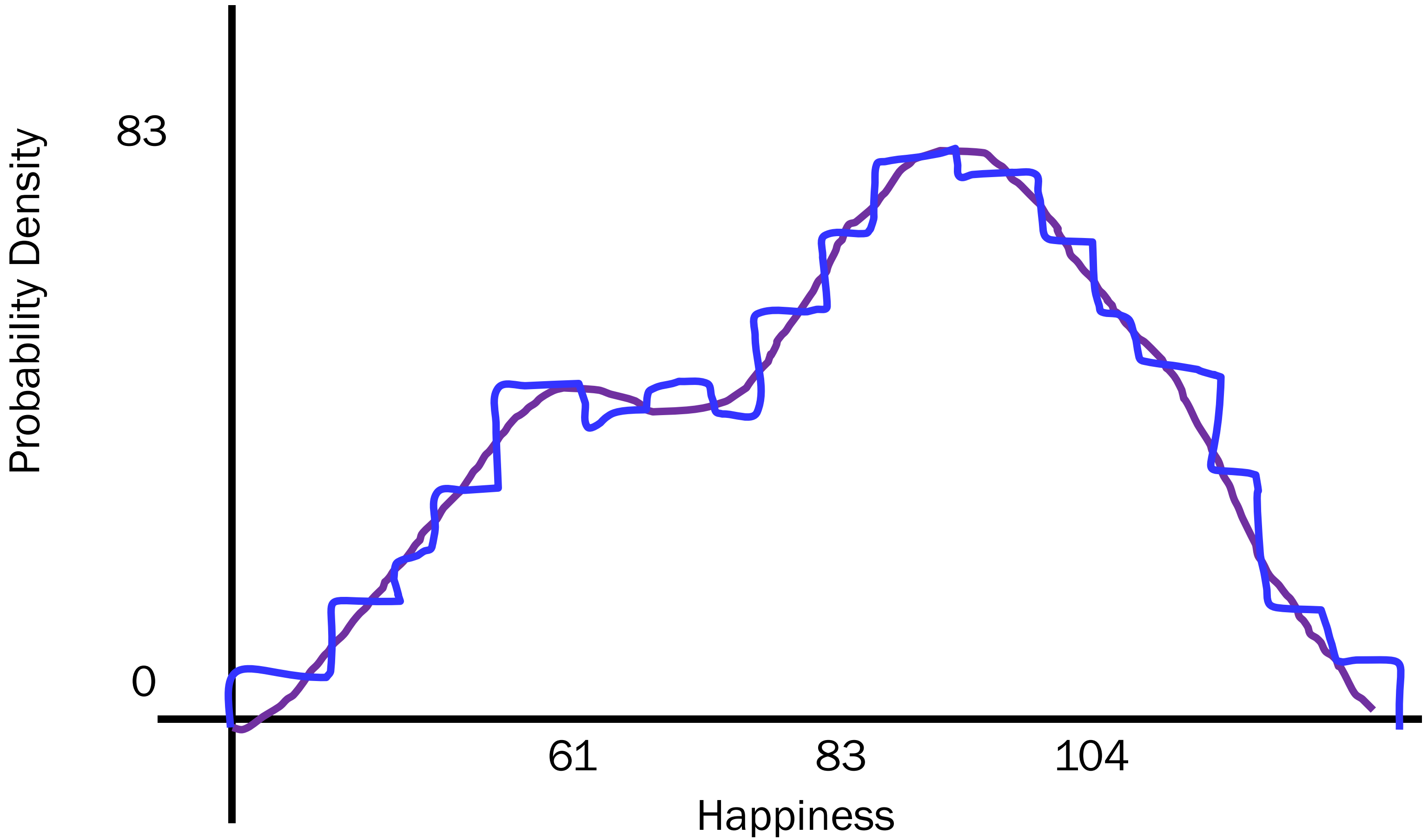
What is the **std** of the **sample variance**, calculated from 200 people?

(I give you the distribution -- stored as a dictionary)



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

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



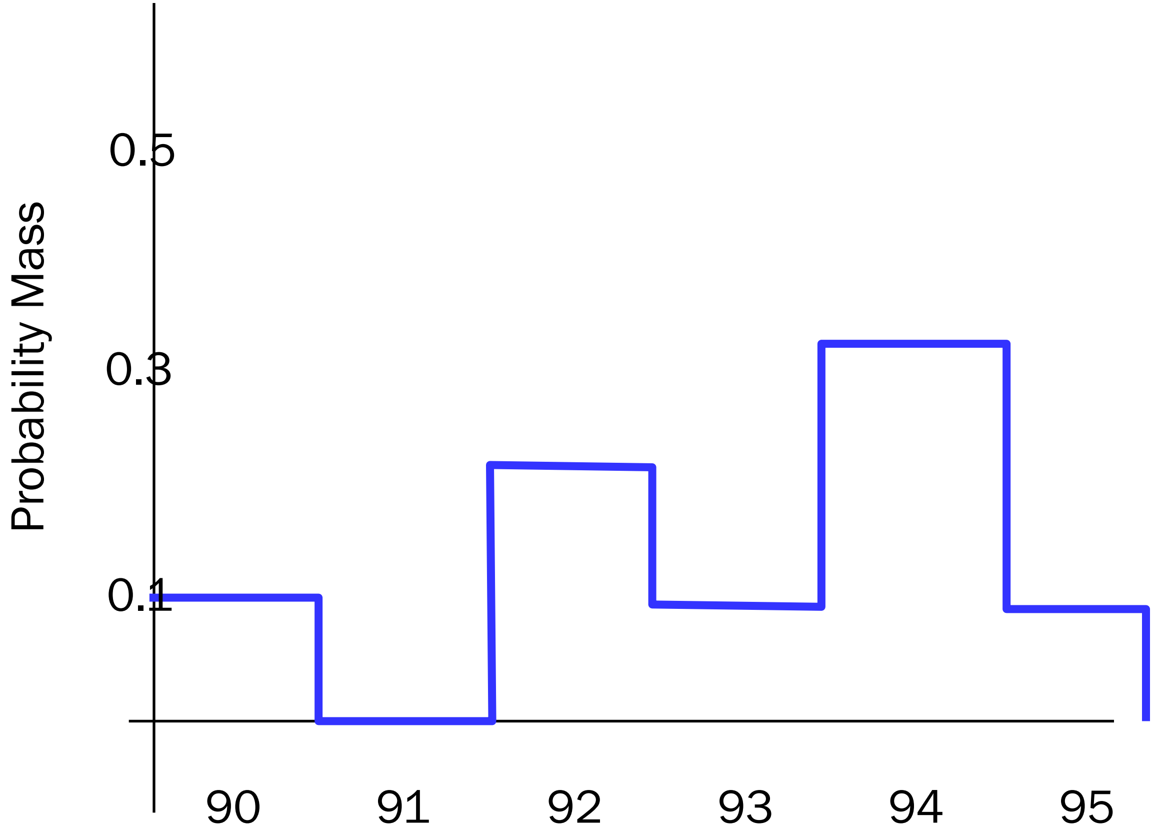
* This is just a histogram of your data!!

Key Insight

IID Samples

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

Sample Distribution



Bootstrapping Assumption

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

The underlying distribution

The sample distribution

(aka the histogram of your data)

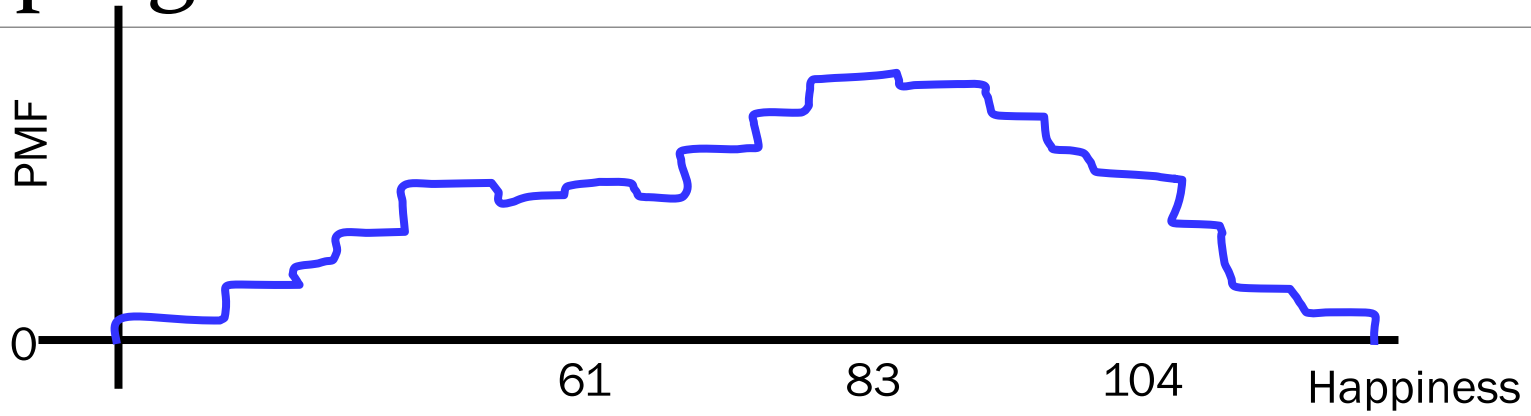
Algorithm

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

Bootstrapping of Means (we could do this with CLT)

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

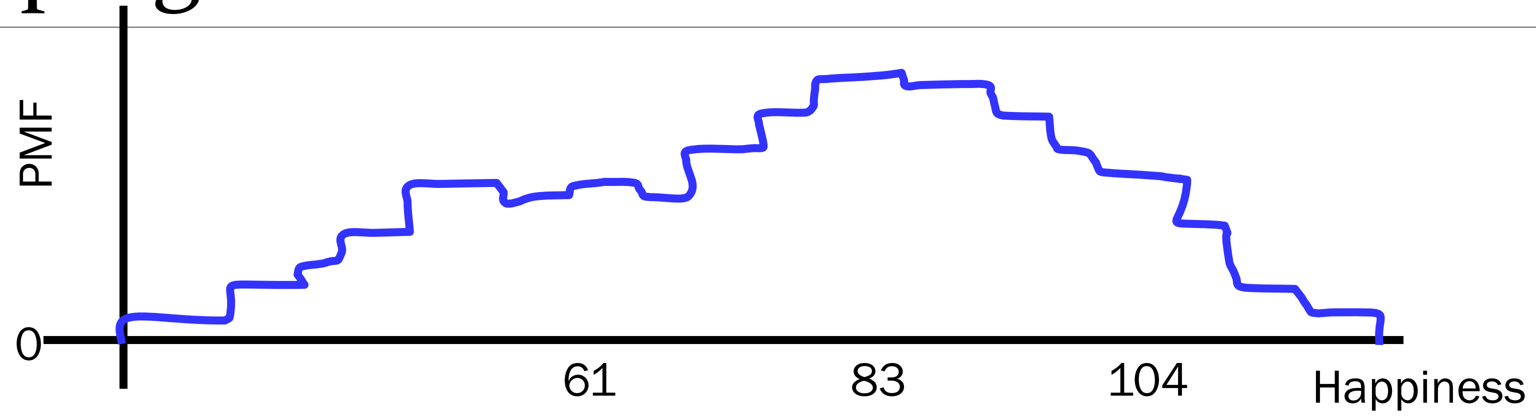
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

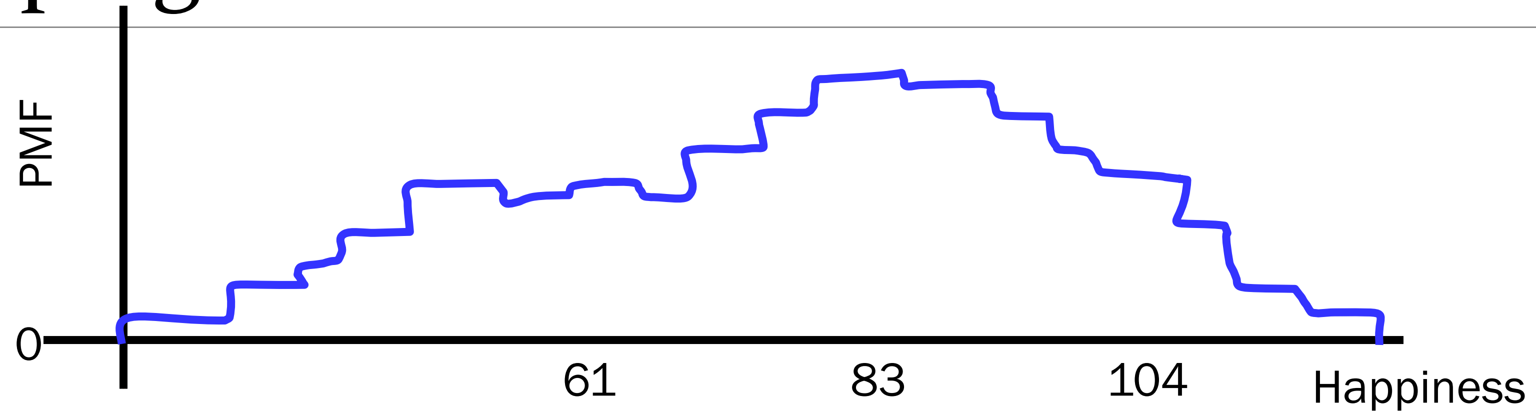
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

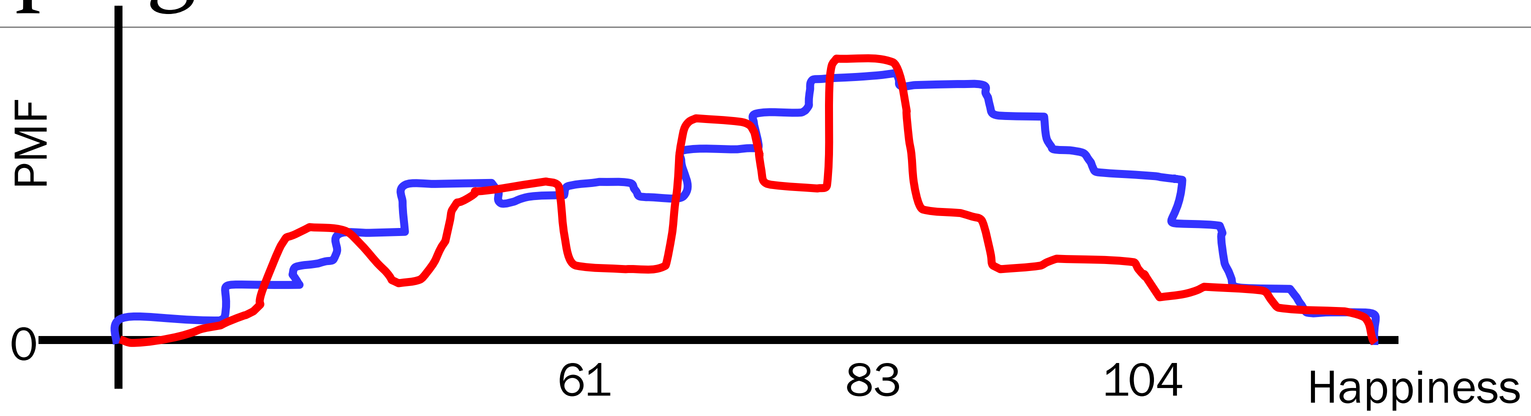
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

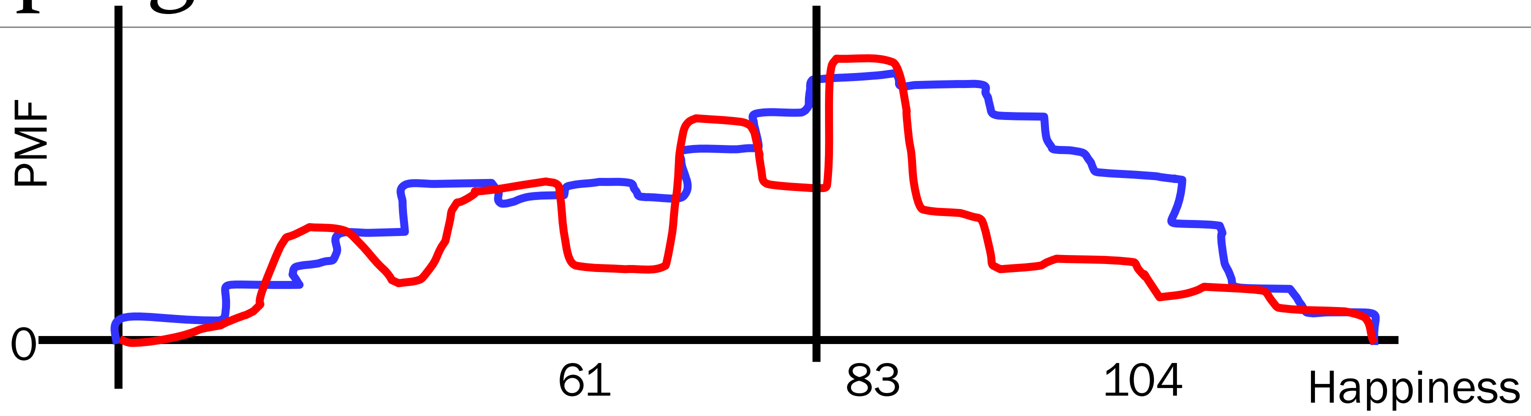
Bootstrapping of Means



Bootstrap Algorithm (sample):

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

Bootstrapping of Means

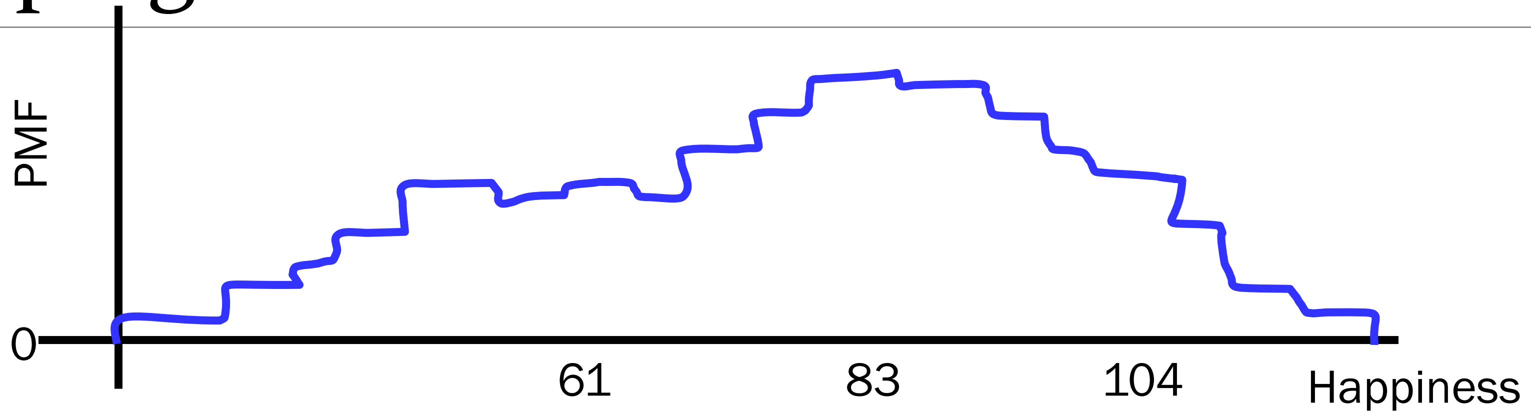


Bootstrap Algorithm (sample):

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

Means = [82.7]

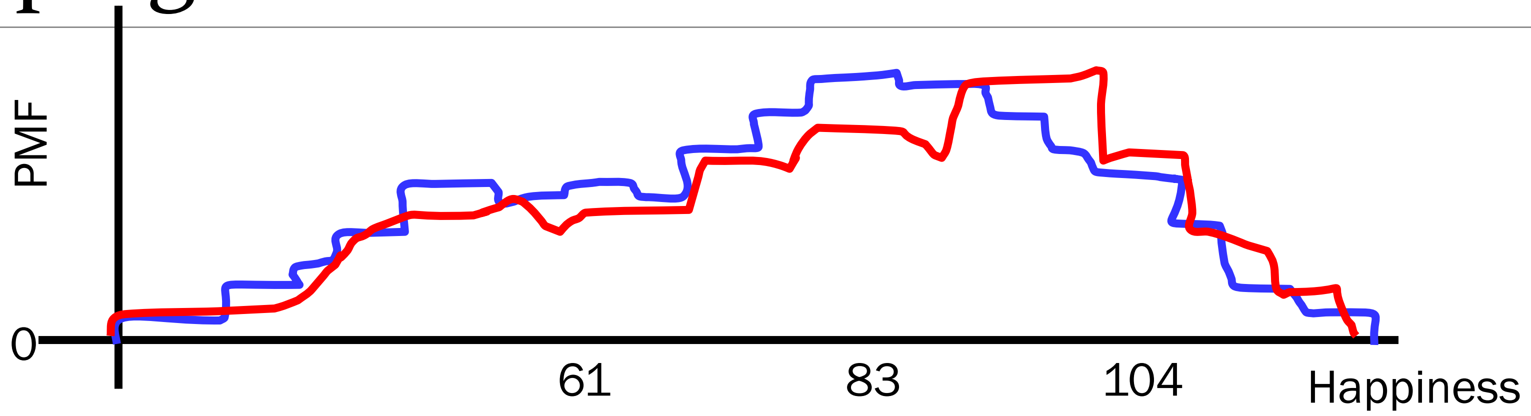
Bootstrapping of Means




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

Means = [82.7]

Bootstrapping of Means

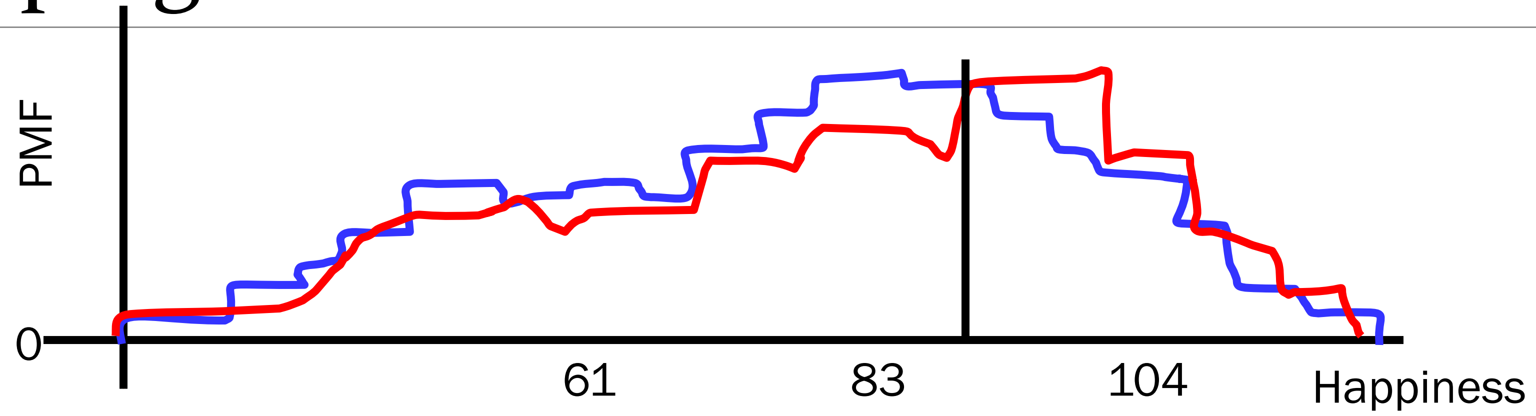


Bootstrap Algorithm (sample):

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

Means = [82.7]

Bootstrapping of Means

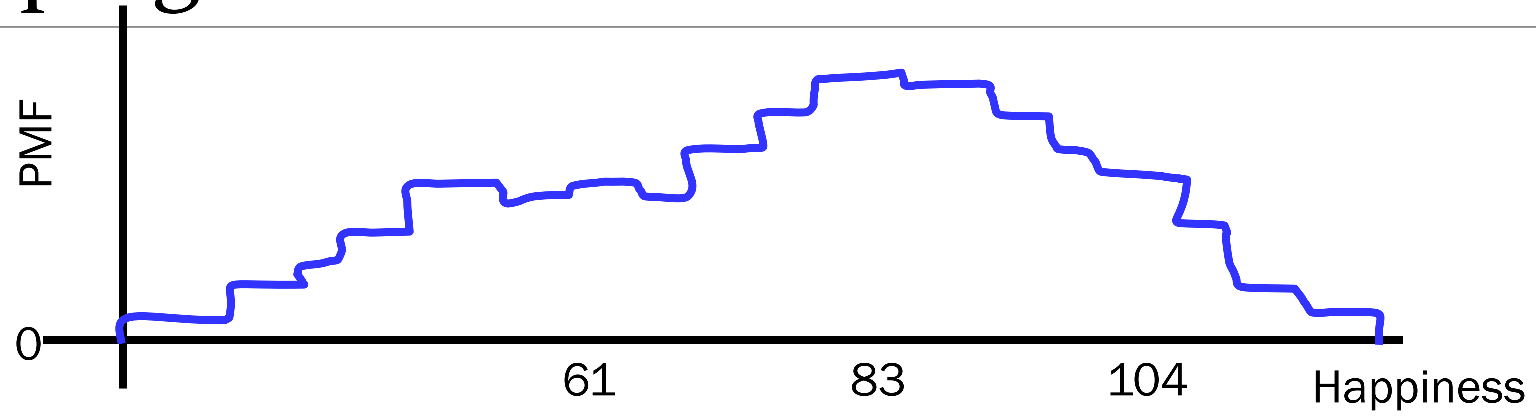


Bootstrap Algorithm (sample):

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

Means = [82.7, 83.4]

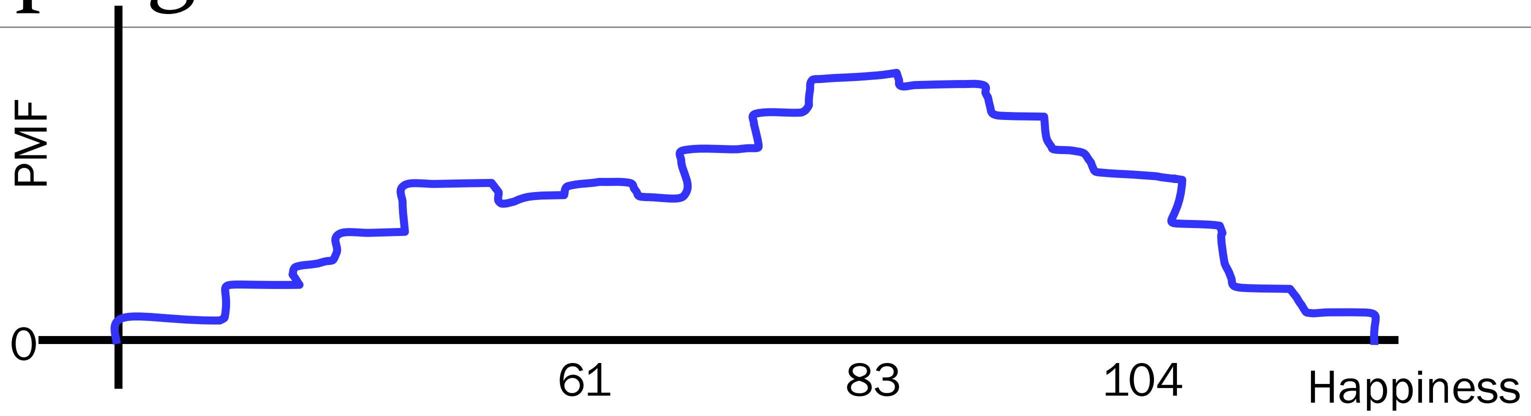
Bootstrapping of Means




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

Means = [82.7, 83.4]

Bootstrapping of Means



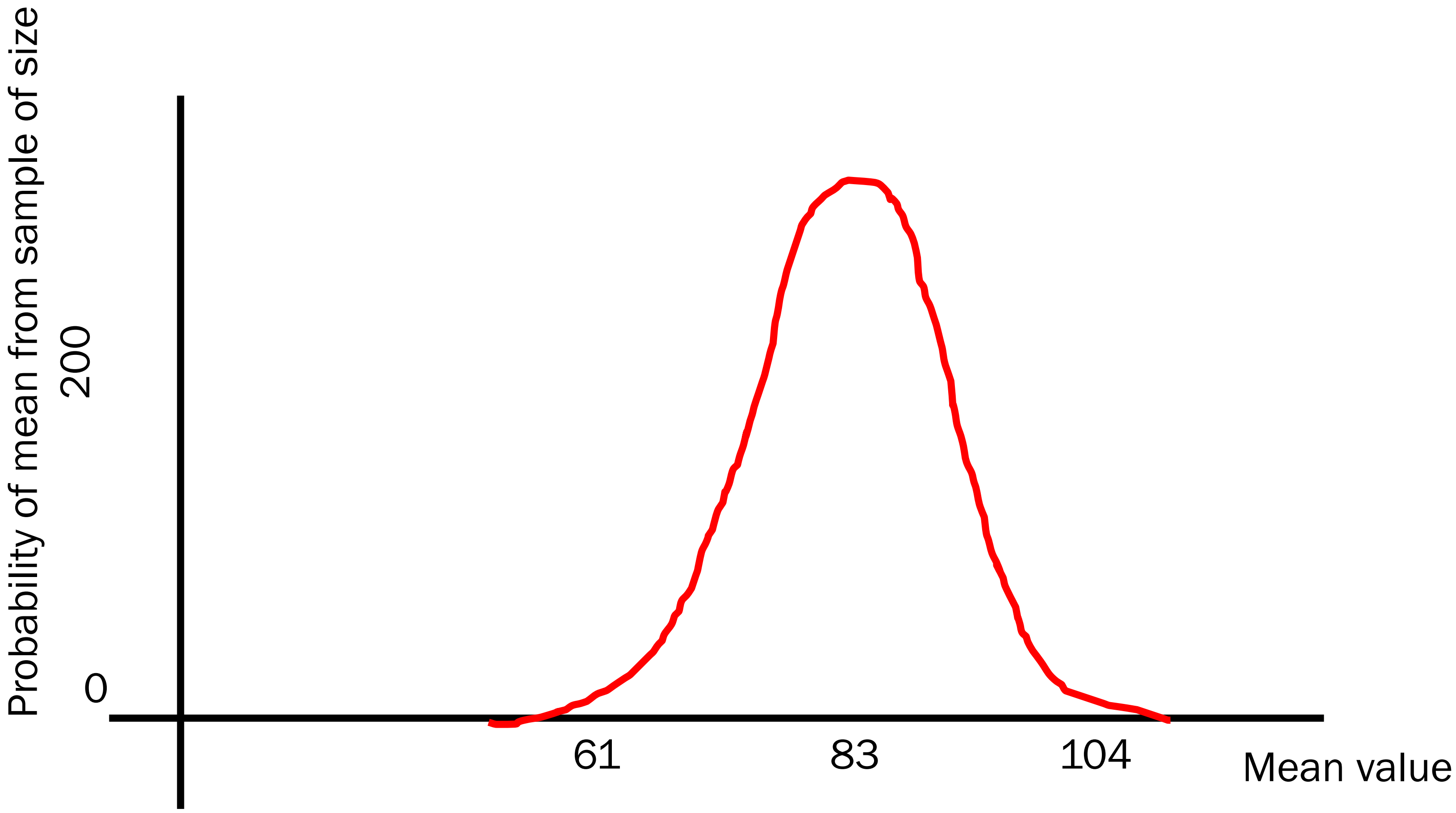
Bootstrap Algorithm (sample):

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

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

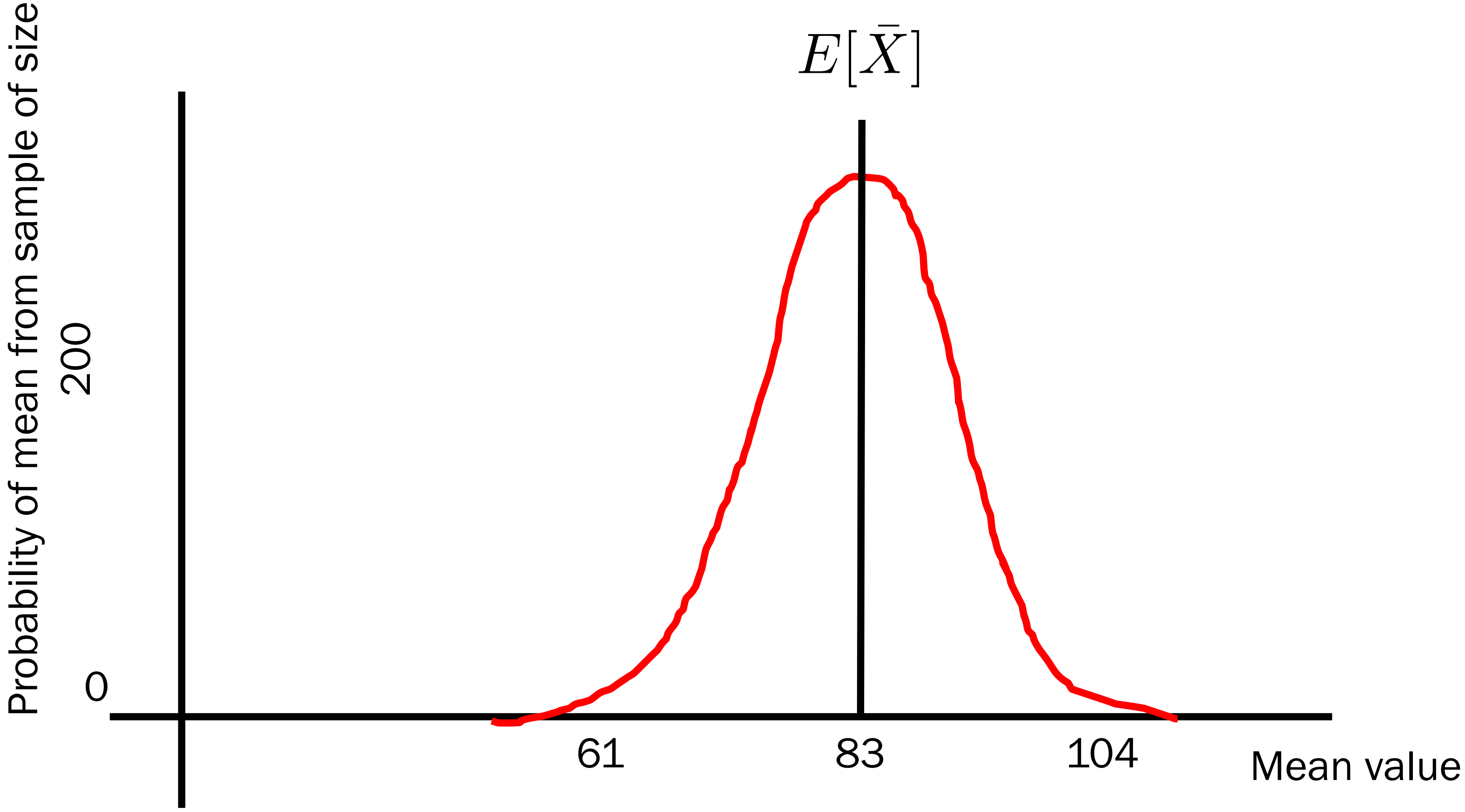
Bootstrapping of Means

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



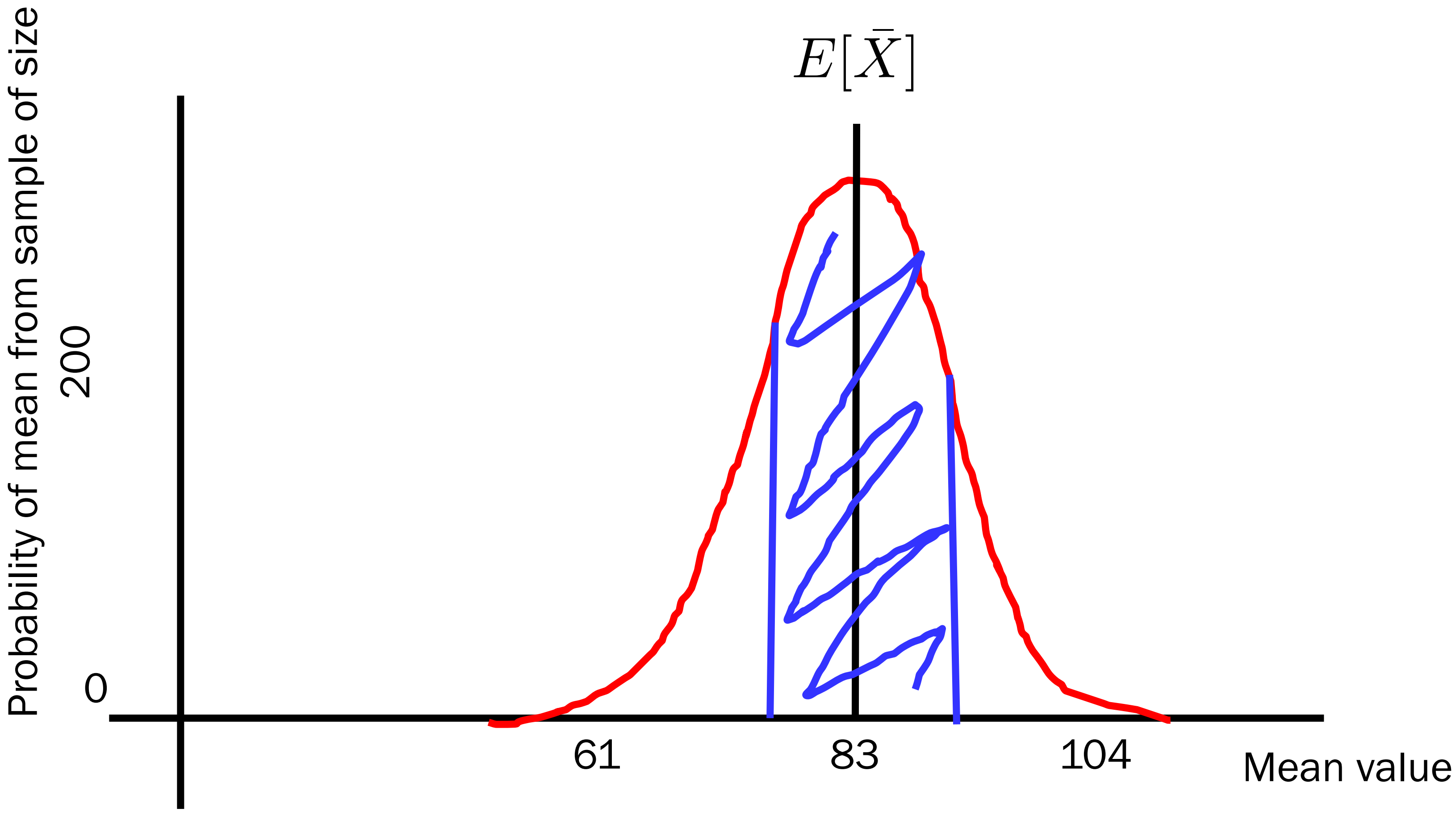
Bootstrapping of Means

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

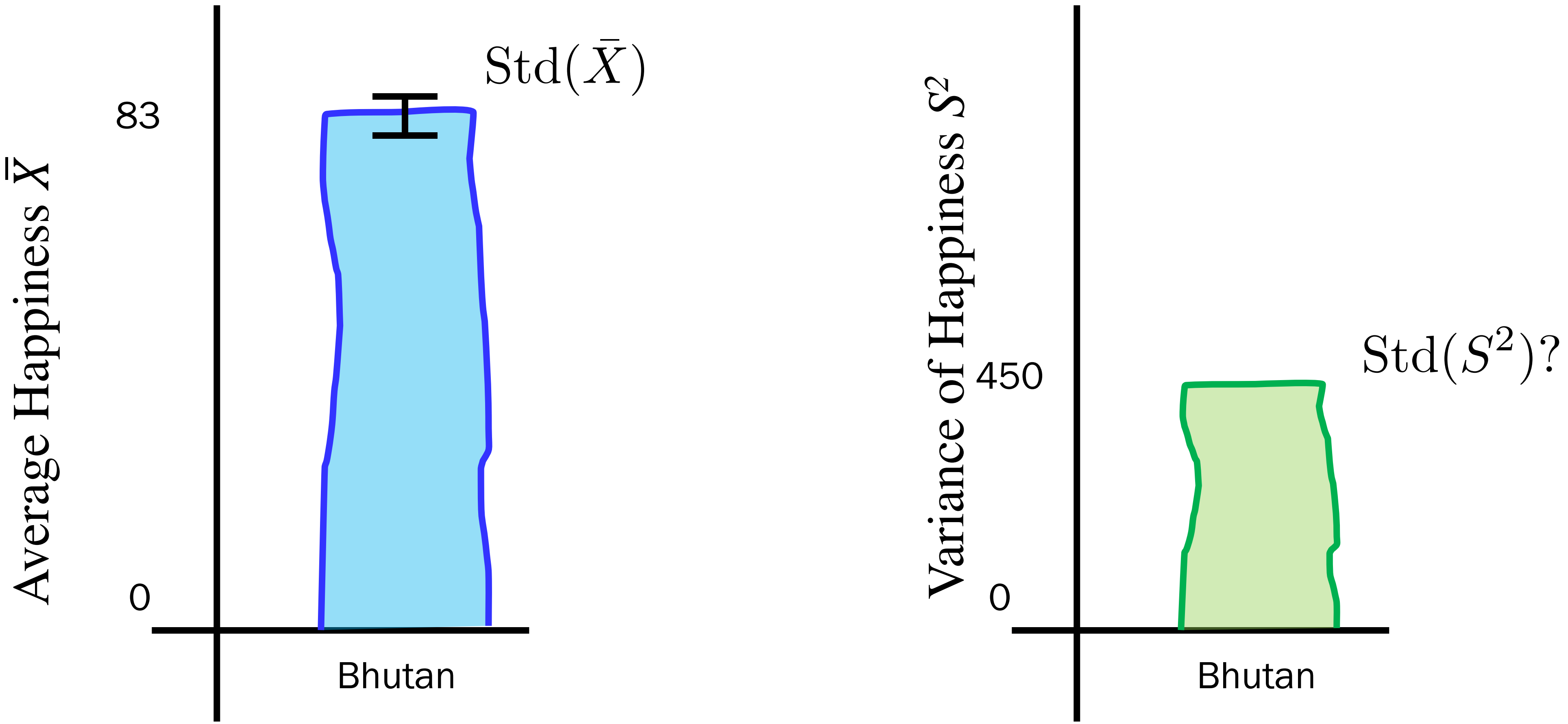


Bootstrapping of Means

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



Our Report to Bhutan Government

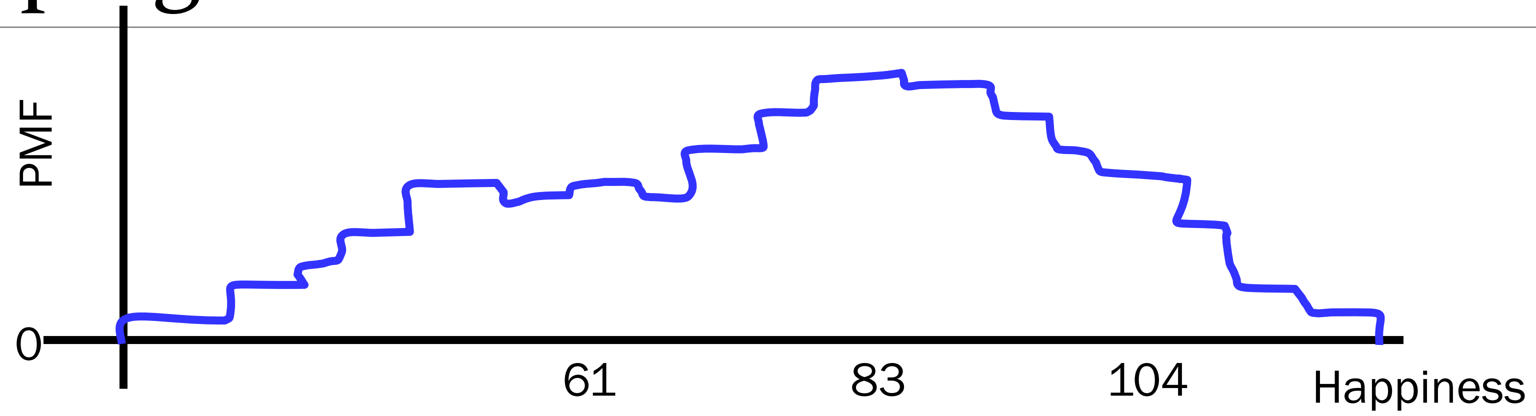


Claim: The average happiness of Bhutan is 83 ± 2

Bootstrapping of Variance

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

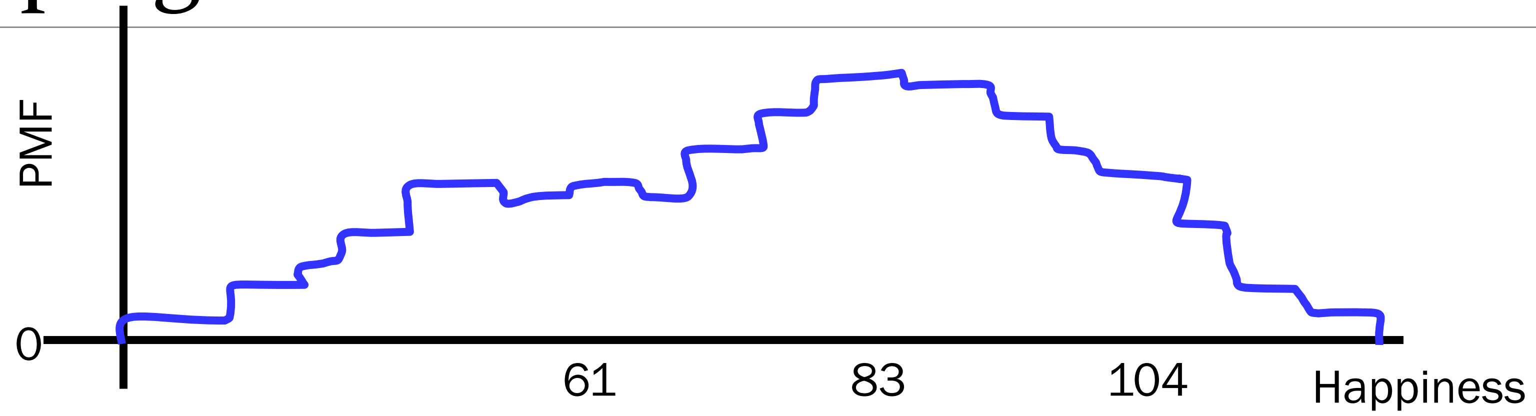
Bootstrapping of Variance



Bootstrap Algorithm (sample):

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

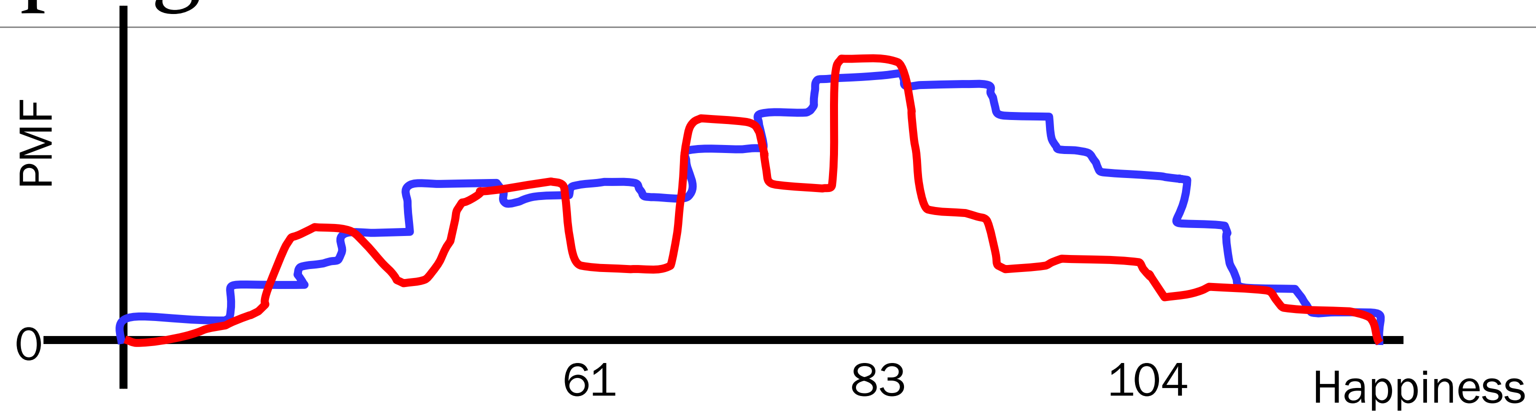
Bootstrapping of Variance



Bootstrap Algorithm (sample):

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

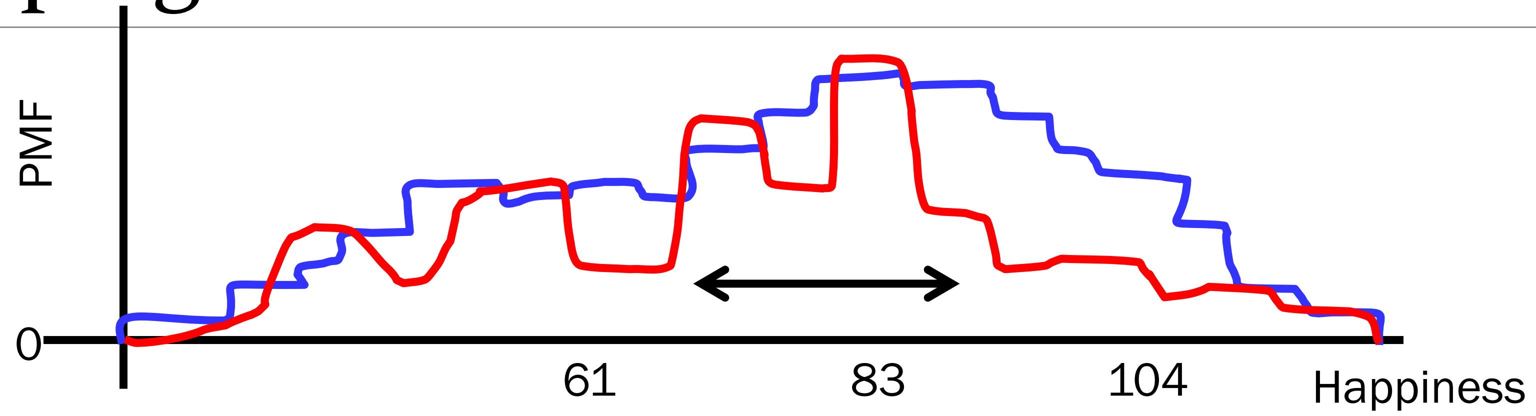
Bootstrapping of Variance



Bootstrap Algorithm (sample):

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

Bootstrapping of Variance

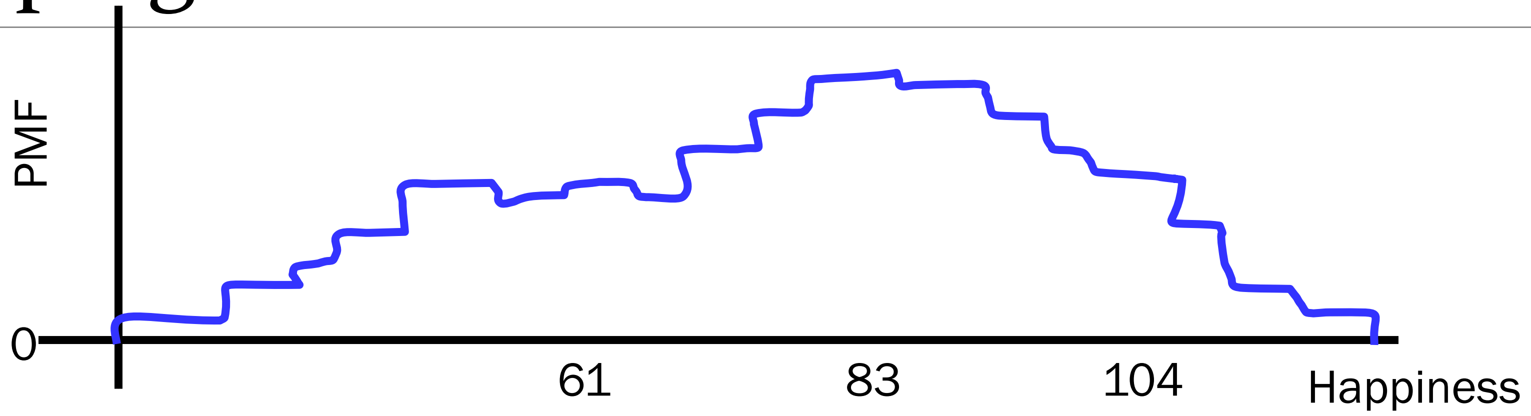


Bootstrap Algorithm (sample):

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

Vars = [472.7]

Bootstrapping of Variance

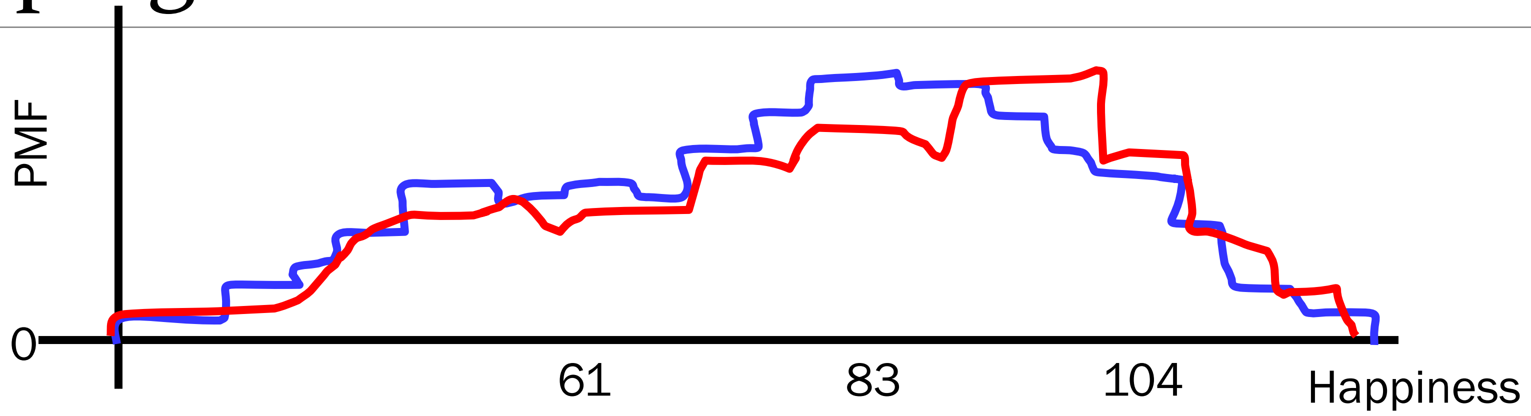


Bootstrap Algorithm (sample):

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

Vars = [472.7]

Bootstrapping of Variance

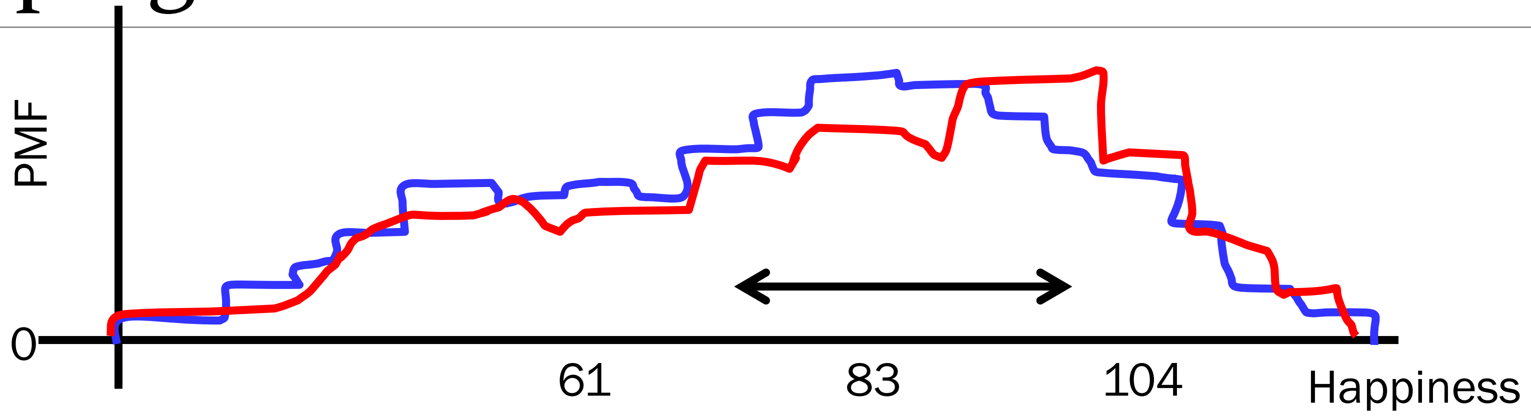


Bootstrap Algorithm (sample):

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

Vars = [472.7]

Bootstrapping of Variance

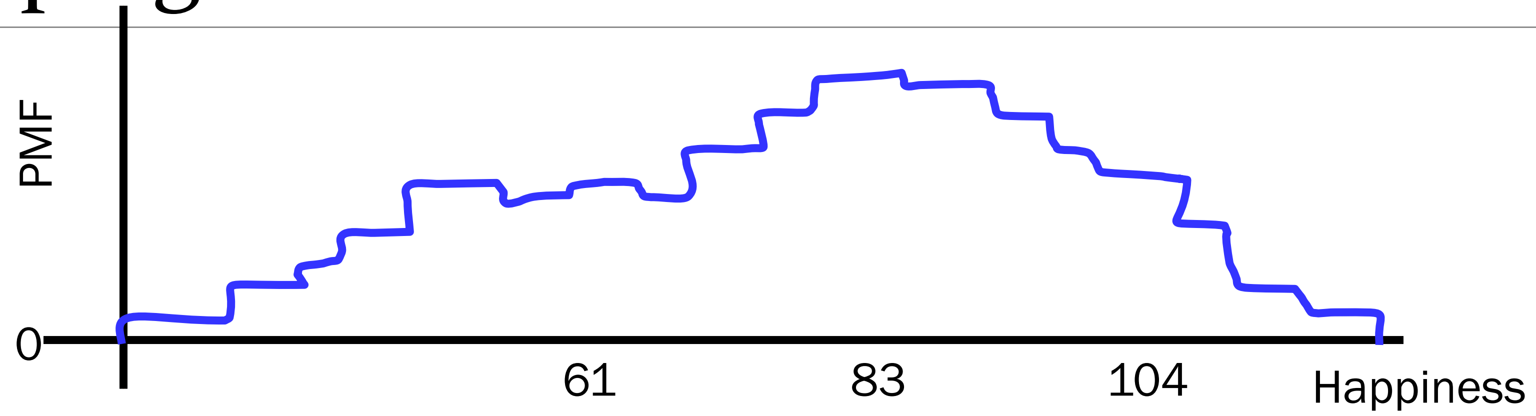


Bootstrap Algorithm (sample):

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

Vars = [472.7, 478.4]

Bootstrapping of Variance

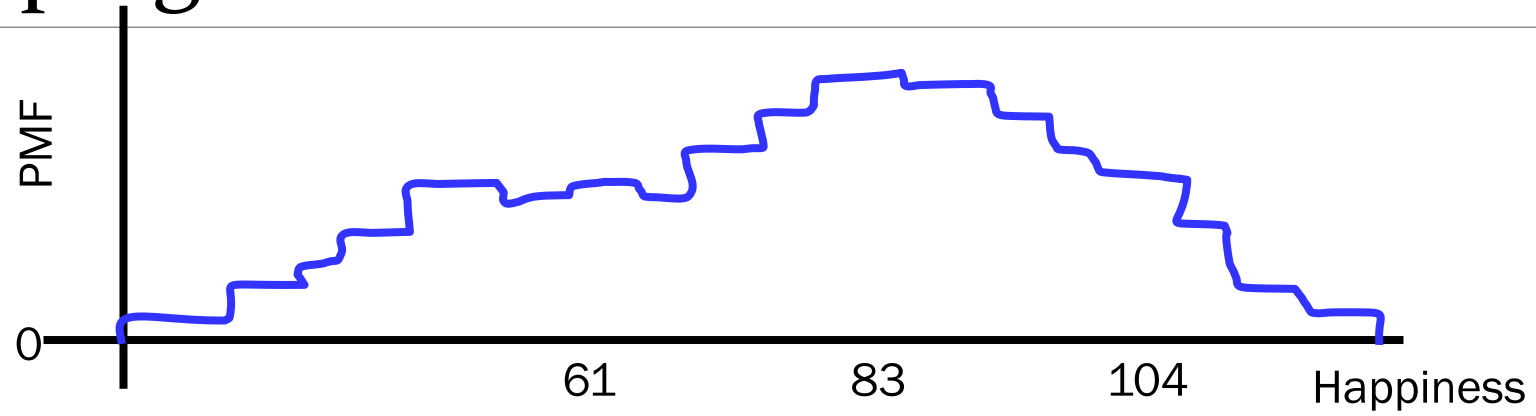


Bootstrap Algorithm (sample):

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

Vars = [472.7, 478.4]

Bootstrapping of Variance



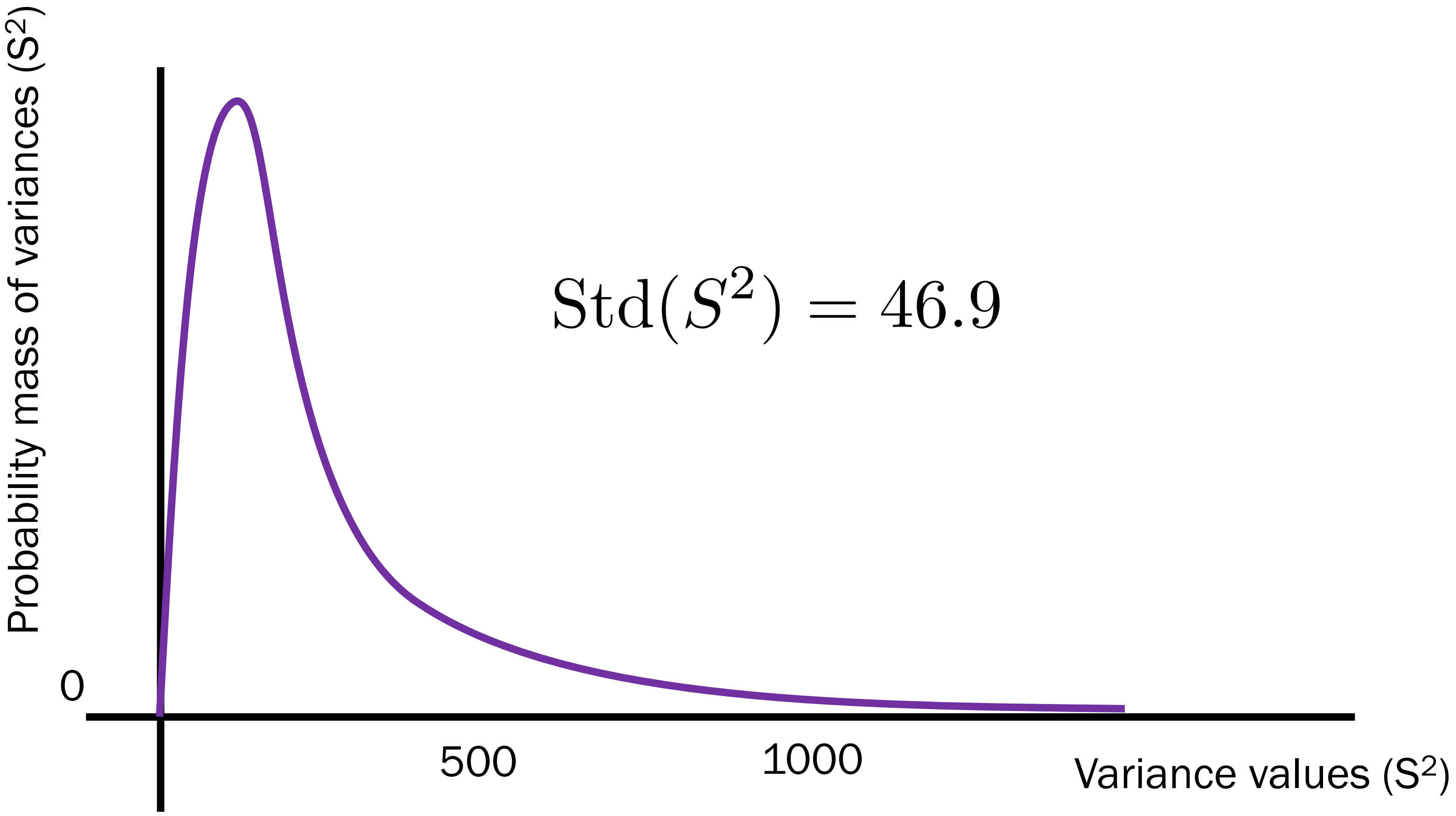
Bootstrap Algorithm (sample):

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

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

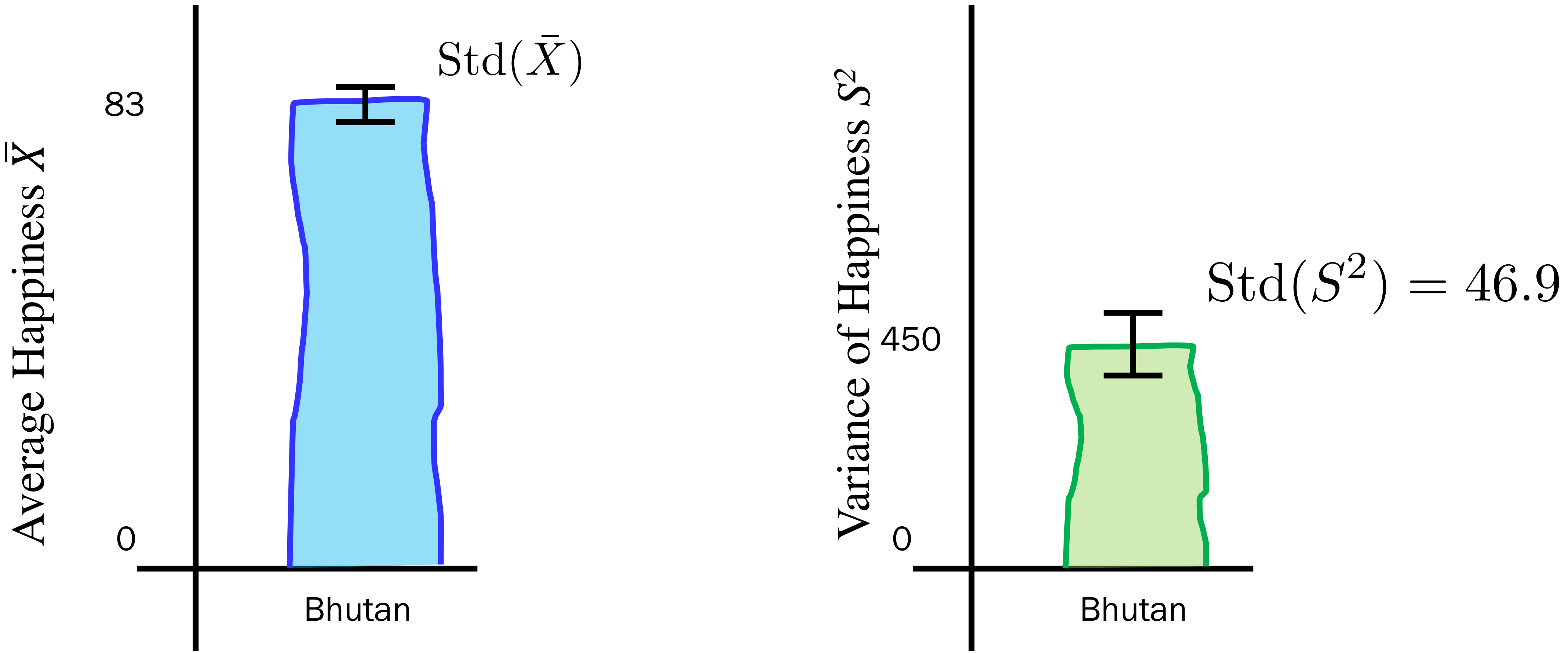
Bootstrapping of Variance

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



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

Our Report to Bhutan Government



Claim: The average happiness of Bhutan is 83 ± 2

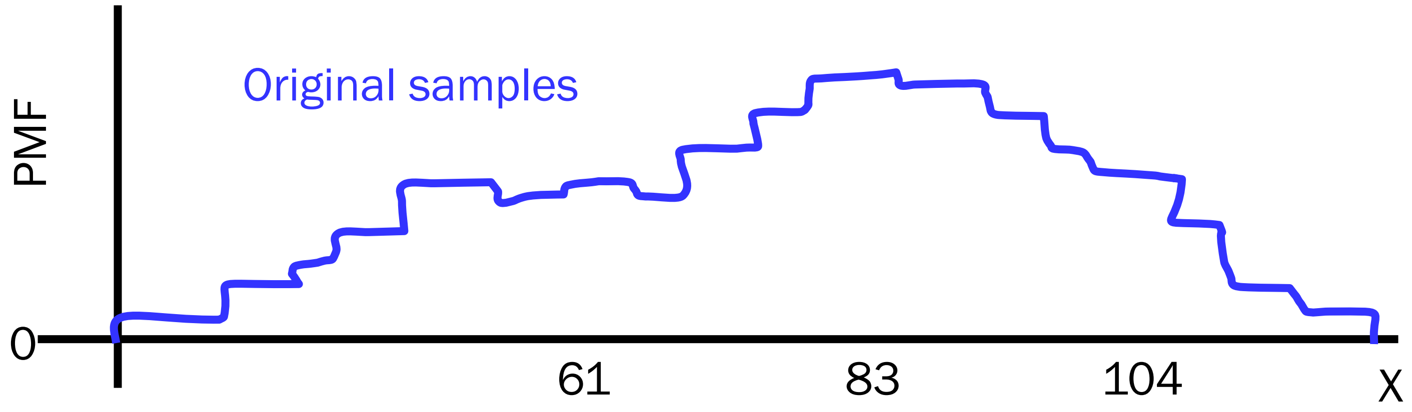
Pedagogical pause

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

Warmup: what is the relationship between a histogram and a PMF?

Bootstrapping in Practice

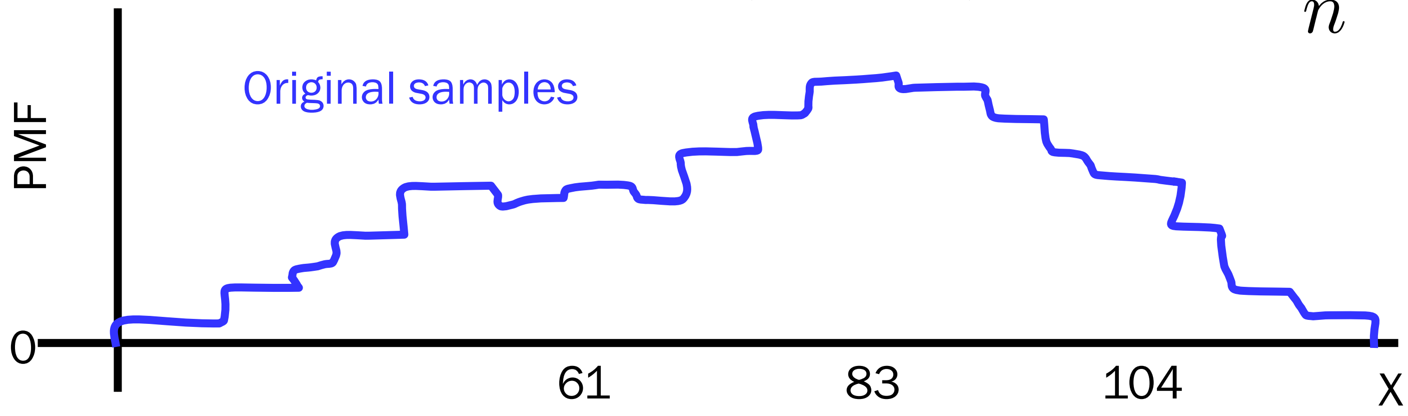
```
def resample(samples, K):
    # Estimate the PMF using the samples
    # Draw K new samples from the PMF
```



Bootstrapping in Practice

```
def resample(samples, K):
    # Estimate the PMF using the samples
    # Draw K new samples from the PMF
    return np.random.choice(samples, K,
                            replace = True)
```

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



OG Bootstrapping

Bootstrap Algorithm (sample):

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

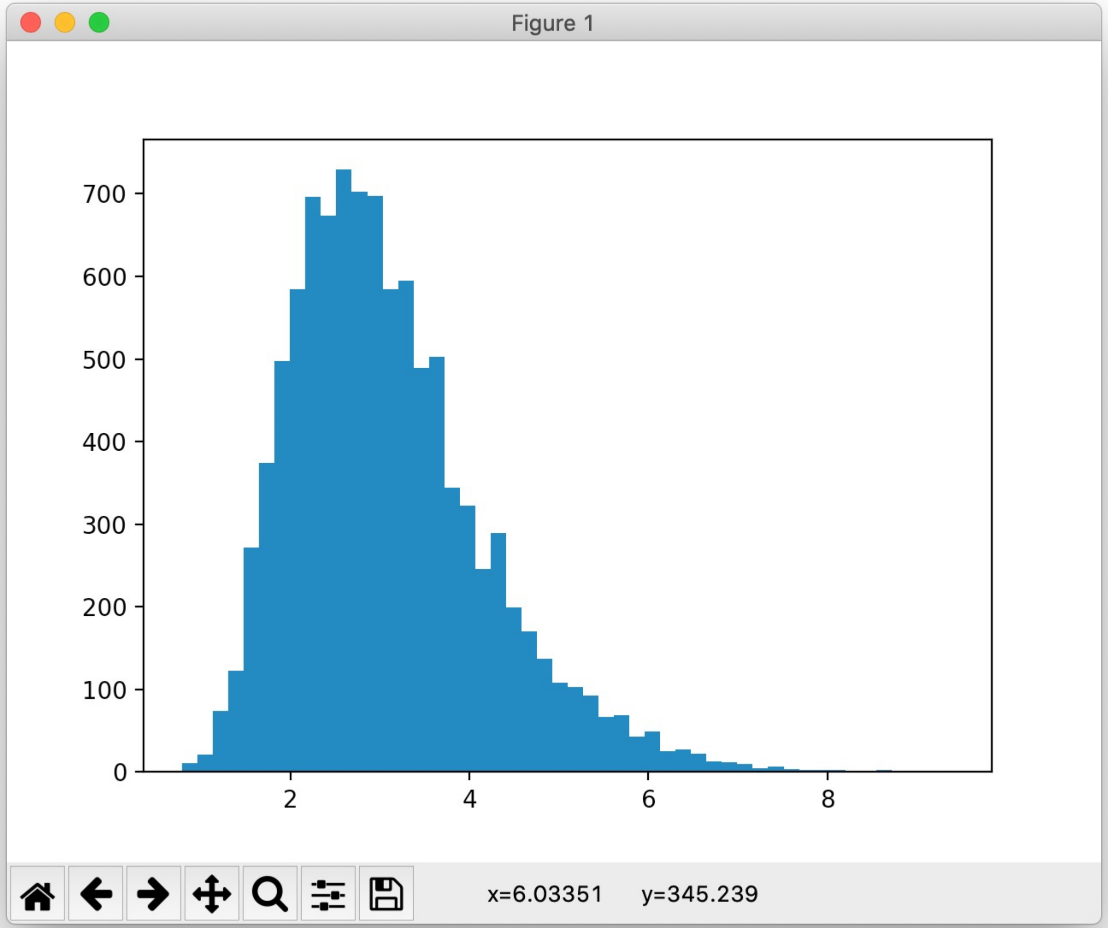
Bootstrapping in Practice

Bootstrap Algorithm (sample):

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

To the code!

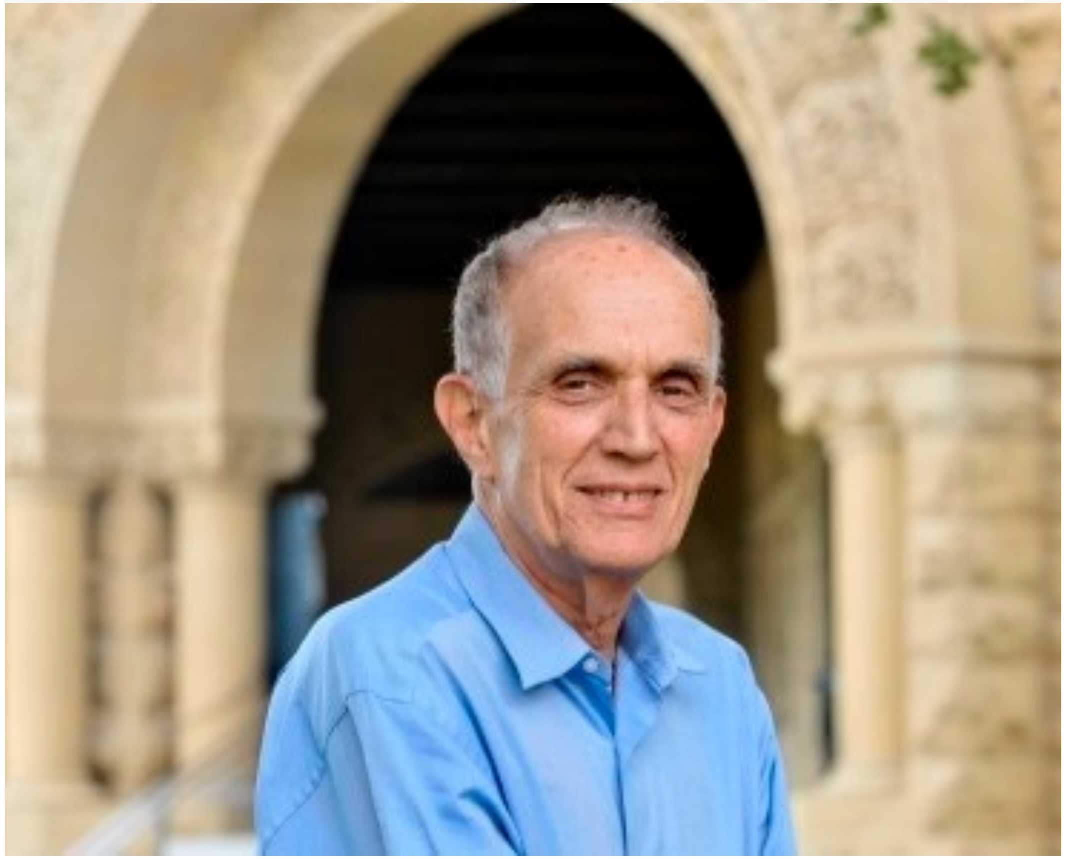
The Distribution of the Sampling Variance



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

Bootstrap

Bradley Efron



Invented bootstrapping in 1979

Still a professor at Stanford

Won a National Science Medal

According to starbyface.com:
Dolph Lundgren

Works for any statistic*

*as long as your samples are IID and the underlying distribution doesn’t have a long tail

The Classic Science Test

Group 1	Group 2
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
…	…
3.70	1.83

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

Claim: Group 1 and Group 2 are samples from **different distributions** with a 0.7 difference of means.

How confident are you in this claim?

A real difference?

18 students

Learning in Context A
4.44
3.36
5.87
2.31
…
3.70

$\mu_1 = 3.1$

Learning in Context B
2.15
3.01
2.02
1.43
…
1.83

$\mu_2 = 2.4$

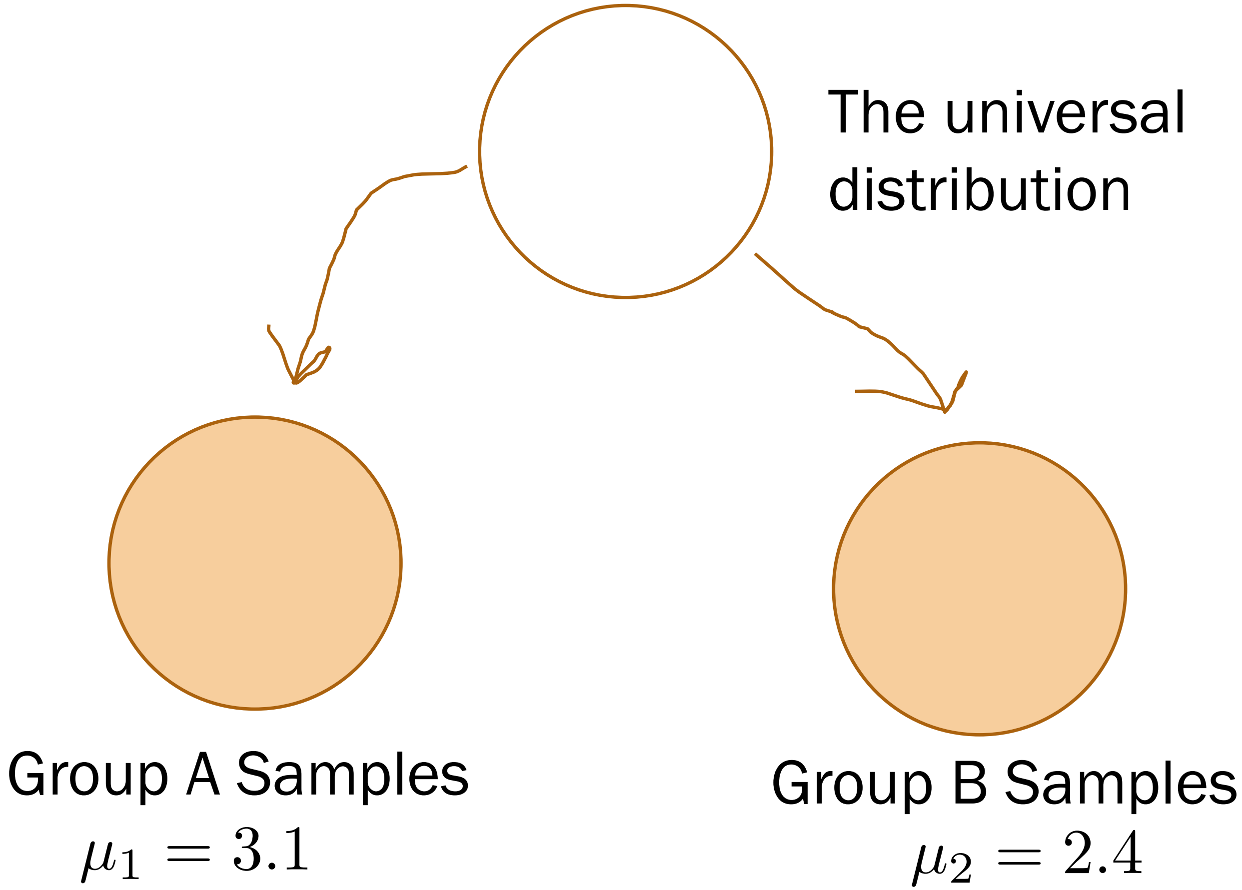
23 students

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?

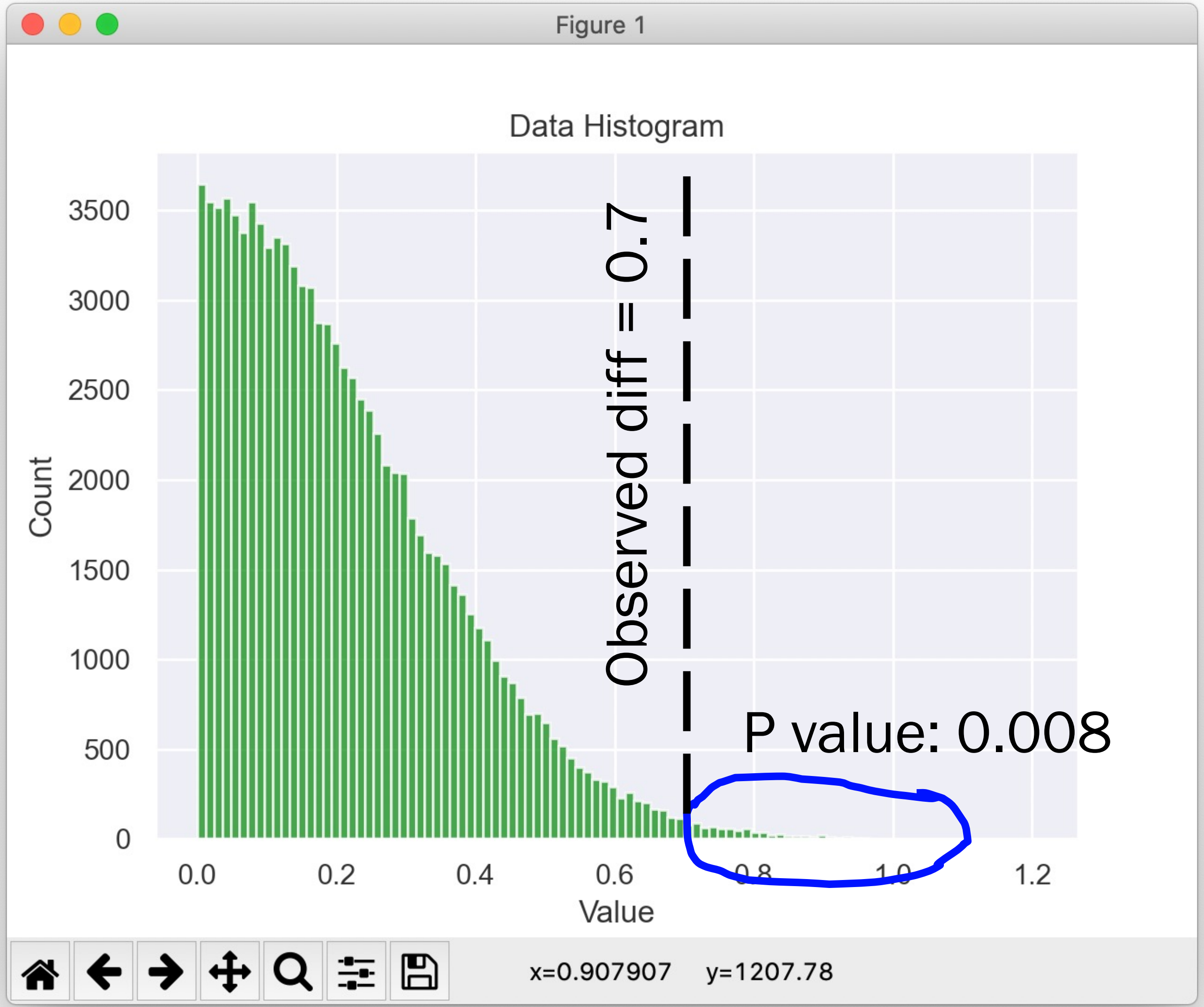
The Null Hypothesis

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!

Distribution of Mean Diffs under Null Hypothesis



Food For Thought

Two Opinions on Distributions

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

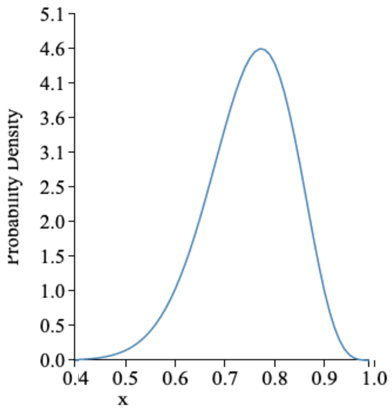
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

Let's use Laplace prior

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



Frequentist:

Let's bootstrap