

Bootstrapping

Chris Piech

CS109, Stanford University

Today, we do science!



A real difference?

	Learning in Context A	Learning in Context B	
18 students	4.44	2.15	23 students
	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?

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?

GPT-4 Tells about the Importance of Bootstrapping

✦ GPT-4

>  Can you name 10 statistical tests that a datascientist should know?

 If you had enough samples, which tests could be replaced with bootstrapping?

GPT-4 Tells about the Importance of Bootstrapping

2 / 2 >



Can you name 10 statistical tests that a datascientist should know?



Certainly! Here's a list of 10 statistical tests that are commonly used in data science, along with a brief description of when they might be used:



1. **Pearson's Chi-squared test:** Used for testing relationships between categorical variables.
2. **T-test (Independent and Paired):** The Independent samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Paired sample t-test is used to compare the means from the same group at different times.
3. **ANOVA (Analysis of Variance):** Compares the means among three or more groups to understand if at least one group mean is statistically different from the others.
4. **Linear Regression:** Used for predicting the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable, and the one you are using to predict the other variable's value is called the independent variable.
5. **Logistic Regression:** Used when the dependent variable is categorical, to predict the outcome of a categorical dependent variable based on one or more predictor variables.
6. **Wilcoxon Signed-Rank Test:** A non-parametric test that compares two paired groups. It's used as an alternative to the paired t-test when the data does not meet the assumptions required by the t-test.
7. **Mann-Whitney U Test:** A non-parametric test for assessing whether two independent samples of observations come from the same distribution. It is an alternative to the independent sample t-test when the data does not satisfy the assumptions of the t-test.
8. **Kruskal-Wallis H Test:** A non-parametric version of ANOVA, which is used when the assumptions of ANOVA are not met. It is used to compare more than two groups.
9. **Fisher's Exact Test:** Used for testing the association between two categorical variables in small sample sizes. It's an alternative to the chi-squared test when the sample size is too small to meet the chi-squared test's assumptions.
10. **Time Series Analysis (e.g., ARIMA):** Not a test per se, but a suite of methods for analyzing time series data, which can include tests for stationarity (ADF Test), seasonality, autocorrelation, and the ability to forecast future values from past data.

Regenerate



If you had enough samples, which tests could be replaced with bootstrapping?



Bootstrapping is a powerful resampling method that can be used to estimate the distribution of a statistic (like the mean, median, or a regression coefficient) by sampling with replacement from the data. When you have a sufficiently large sample size, bootstrapping can be used to relax the assumptions of traditional parametric tests, like the normality assumption or the assumption of equal variances across groups.



In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

1. **T-test:** Both independent and paired t-tests can be replaced with bootstrap methods to estimate the difference in means without assuming normal distributions of the sample means.
2. **ANOVA:** Bootstrap methods can be used to assess the difference in means across groups without assuming normally distributed residuals or homogeneity of variance.
3. **Linear Regression:** Instead of assuming the normality of residuals, you can use bootstrapping to construct confidence intervals for the regression coefficients.
4. **Logistic Regression:** Similarly, bootstrapping can be used to obtain more robust confidence intervals for the coefficients in a logistic regression model.
5. **Wilcoxon Signed-Rank Test and Mann-Whitney U Test:** While these are non-parametric tests that do not assume normality, bootstrapping can still be used to obtain confidence intervals for the median difference or other statistics of interest.
6. **Kruskal-Wallis H Test:** Bootstrapping can be applied to obtain the distribution of the test statistic under the null hypothesis, and thus get a p-value without the assumption of identical distribution shapes across groups.
7. **Fisher's Exact Test:** When sample sizes are large, the exact calculations required for Fisher's Exact Test can become computationally intensive. Bootstrapping can be a practical alternative to obtain an approximation.

GPT-4 Tells about the Importance of Bootstrapping

2 / 2 >



Can you name 10 statistical tests that a datascientist should know?



Certainly! Here's a list of 10 statistical tests that are commonly used in data science, along with a brief description of when they might be used:



1. **Pearson's Chi-squared test:** Used for testing relationships between categorical variables.
2. **T-test (Independent and Paired):** The Independent samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Paired sample t-test is used to compare the means from the same group at different times.
3. **ANOVA (Analysis of Variance):** Compares the means among three or more groups to understand if at least one group mean is statistically different from the others.
4. **Linear Regression:** Used for predicting the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable, and the one you



If you had enough samples, which tests could be replaced with bootstrapping?



Bootstrapping is a powerful resampling method that can be used to estimate the distribution of a statistic (like the mean, median, or a regression coefficient) by sampling with replacement from the data. When you have a sufficiently large sample size, bootstrapping can be used to relax the assumptions of traditional parametric tests, like the normality assumption or the assumption of equal variances across groups.



In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

1. **T-test:** Both independent and paired t-tests can be replaced with bootstrap methods to

In principle, any statistical test could be replaced with a bootstrap equivalent if you're interested in hypothesis testing or constructing confidence intervals, especially if the assumptions of the traditional test are questionable. However, it's particularly common to use bootstrapping for the following tests:

small sample sizes. It's an alternative to the chi-squared test when the sample size is too small to meet the chi-squared test's assumptions.

10. **Time Series Analysis (e.g., ARIMA):** Not a test per se, but a suite of methods for analyzing time series data, which can include tests for stationarity (ADF Test), seasonality, autocorrelation, and the ability to forecast future values from past data.

statistic under the null hypothesis, and thus get a p-value without the assumption of identical distribution shapes across groups.

7. **Fisher's Exact Test:** When sample sizes are large, the exact calculations required for Fisher's Exact Test can become computationally intensive. Bootstrapping can be a practical alternative to obtain an approximation.

Regenerate

In other words.

Datascientists **Hate Him!**



He bongclouded from
level 400 data scientist to
a level 3000 in hours

using this
one ~~weird~~ trick!!!

(Computer science)

In other words.

Datascientists



Carissa Yip reposted



Hikaru Nakamura

@GMHikaru

Follow

Hilarious



Carissa Yip @carissayipchess · Nov 26, 2023

just saw this in my stanford cs class's lecture slides

In other words.

Datascientists **Hate Him!**



He bongclouded from **420** bulet rating to **3000** bulet rating in hours to save your loved ones using this **one weird trick!!!**

32

Stanford University

1:49 PM · Nov 26, 2023 · 250K Views



16



44



2.4K



26



Where are we in CS109?

You are here



Core
Probability

x_2

Random
Variables



Probabilistic
Models



Uncertainty
Theory



Machine
Learning



Uncertainty Theory

Beta
Distributions

Thompson
Sampling

Adding
Random Vars

Central Limit
Theorem

Sampling

Bootstrapping

Algorithmic
Analysis

Information
Theory +
Divergence

As requested by AI faculty



<announcements>

Pset 5 is Out

The image displays three browser windows from the course website, each showing a different problem set (PS6).

- Window 1: "Which Learning Activity is Better?"**
 - Text: "In this question you are going to learn how to calculate p-values for experiments that are called 'a/b tests'. These experiments are ubiquitous. They are a staple of both scientific experiments and user interaction design."
 - Diagram: Two laptops labeled A and B. Laptop A shows a 23% increase and is labeled "CONTROL". Laptop B shows a 37% increase and is labeled "VARIATION".
 - Text: "Massive online classes have allowed for distributed experimentation into what practices optimize students learning - and promise to be able to scale more personalized educational experiences. Coursera, a free online education platform that started at Stanford, is testing out a set of ways of teaching a concept in probability. They have two different learning activities activity1 and activity2 and they want to figure out which activity leads to better learning outcomes. After interacting with a learning activity Coursera evaluates a student's learning outcome by asking them to solve a set of questions."
 - Text: "Over a two-week period, Coursera randomly assigns each student to either be given activity1 (group A), or activity2 (group B). The activity that is shown to each student and the student's measured learning outcomes can be found in the file: learningOutcomes.csv in the pset6 data zip: [pset6.zip](#)"
 - Text: "You should answer questions a) and b). The numeric answer will check the p-value you calculate in part b. Even though many calculations will require you to use code, we want you to show your work by using words to succinctly describe what your code does (don't paste your code into the app). This is similar to how you would present a result like this in a scientific paper."
 - Section: "a) Observed Difference"
 - Text: "What is the difference in sample means of learning outcomes between"
- Window 2: "Better Peer Grading"**
 - Text: "Stanford's HCI class runs a massive online class that was taken by ten thousand students. The class used peer assessment to evaluate student's work. We are going to use their data to learn more about peer graders. In the class, each student has their work evaluated by 5 peers and every student is asked to evaluate 6 assignments: five peers and the control assignment (the graders were un-aware of which assignment was the control). All 10,000 students evaluated the same control assignment and the scores they gave are in the file peerGrades.csv in the pset6 data zip: [pset6.zip](#)"
 - Text: "Would you use the **mean** or the **median** of 5 peer grades to assign scores in the online version of Stanford's HCI class? **Explain why.** You may use simulations to solve any part of this question. Hint: it might help to visualize the scores in peerGrades.csv. In order to make your decision compute the statistics in part a) and b)."
 - Diagram: A network graph visualization with many blue nodes and several red nodes connected by red lines.
 - Caption: "Coursera HCI Peer Grading: A visualization of a peer grading network of 63K peer grades for 7K students in an online course. A single student is"
- Window 3: "Learning While Helping"**
 - Text: "You are designing a randomized algorithm that delivers one of two new drugs (which we call drug A and drug B) to patients who come to your clinic. Each patient can only receive one of the drugs. Initially you know nothing about the effectiveness of the two drugs. You are simultaneously trying to learn which drug is the best and, at the same time, cure the maximum number of people. To do so we will use the Thompson Sampling Algorithm."
 - Text: "Your job is to implement the `thompson_sampling` function which will decide whether to give drug A or drug B, based on a limited history of observations."
 - Section: "Thompson Sampling Algorithm:"
 - Text: "For each drug we maintain a Beta distribution to represent the drug's probability of being successful. Our initial belief in the probability of success is uniform for both drug A and drug B: $\theta_i \sim \text{Beta}(1, 1)$."
 - Text: "When choosing which drug to give to the next patient we sample a value from the Beta representing drug A, and we sample a value from the Beta representing drug B. We select the drug with the largest sampled value. We administer the drug, observe if the patient was cured, and update the Beta that represents our belief about the probability of the drug being successful."
 - Diagram: A cartoon illustration of a robot character standing between two buildings. One building is labeled "The Usual Place" and the other is labeled "GRAND OPENING!". A question mark is next to the robot.
 - Text: "Thompson sampling helps navigate the tradeoff between exploring an option we don't know much about and an option we already know works well"
 - Section: "History"
 - Text: "Each time your function is called you will be given data on how many times you have administered each drug, and how many times that drug successfully cured the patient. For example the following history"

On the right side of the third window, there is an "Answer Editor" section with a "Solution" icon and a code editor showing:

```
Agent:
1 def thompson_sampling(history):
2     return 'A'
```

Below the code editor are buttons for "Run One Game" and "Test Agent".

Trust in the System



Thoughts

Integrity of the class is communal

The only way to really mess up CS109 is to cheat, especially on a test

We will likely find this person

See something, say something

Tell us before we tell you

Sometimes fizz is bs

Hard Midterm? Who knows until Monday.

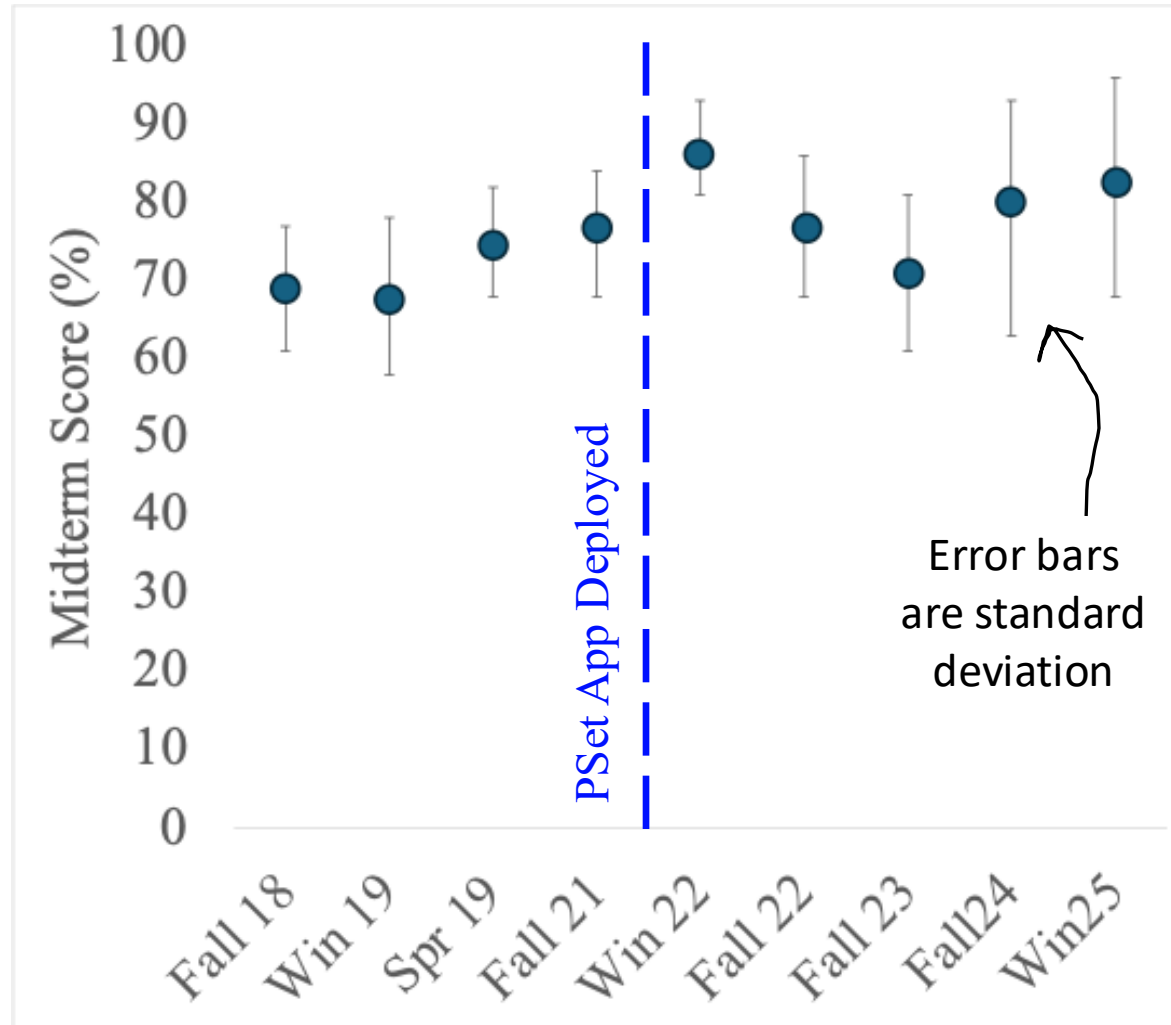


We are here to make you hit heights you didn't think you could -- not to judge you.

Be easy on yourself too. But don't sell yourself short.

Midterm is a diagnostic instead of just summative assessment.

History of Midterms



Course grades
are
independent
of difficulty

Partial credit is
a thing!

DEC
3RD



<end>

<review>

Central Limit Theorem (Summation)

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

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

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, \dots, X_n with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$.

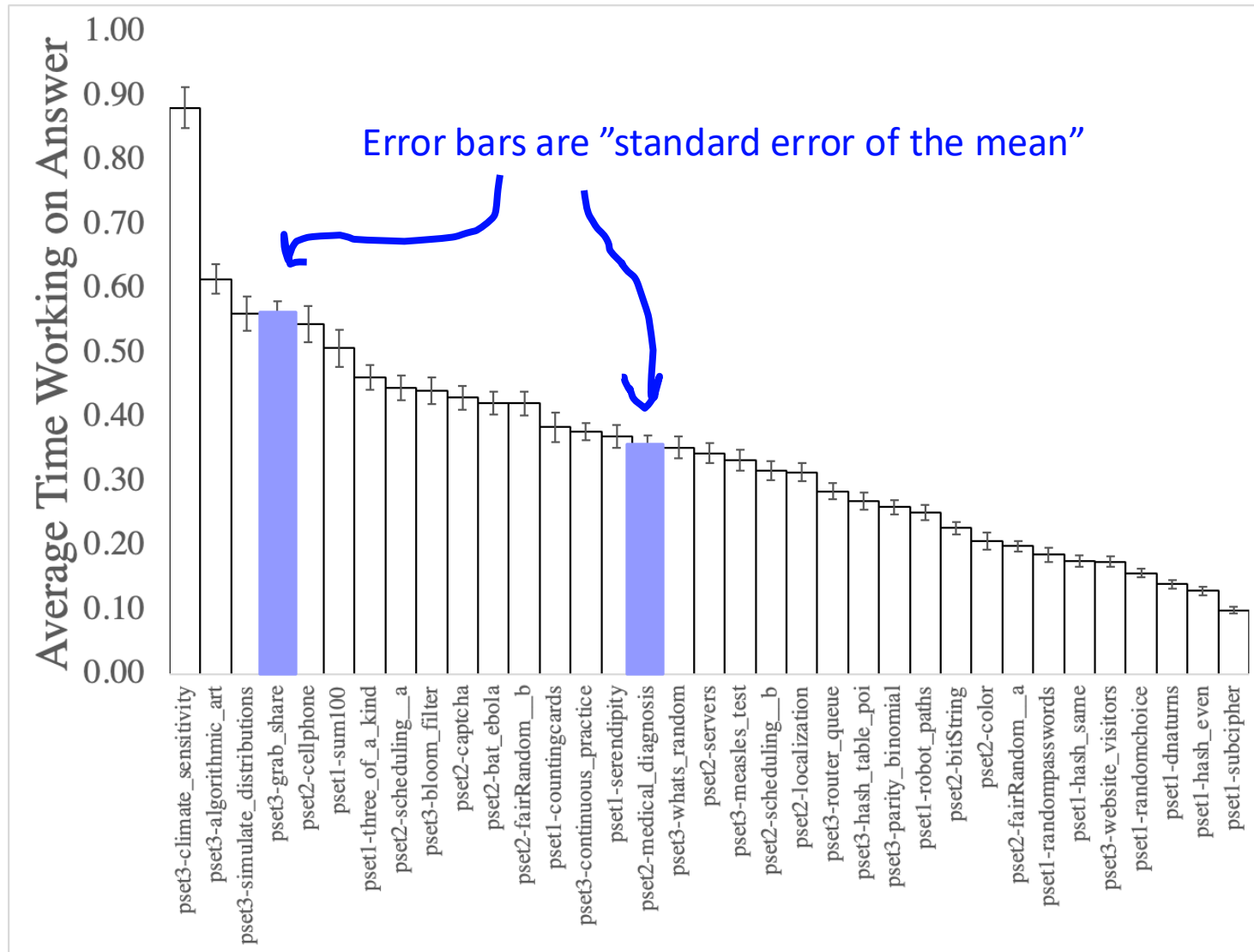
$$\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{As } n \rightarrow \infty$$

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

Sample Mean and Standard Error for Song Averages

#	Song	Sample Mean PDF	Votes	NumVotes	SampleMean	SEOM	SongId	Pr(Top16)	Pr(Best)
1	Get Lucky - Daft Punk			45	3.82	0.18	117	0.851	0.500
2	Life is a Highway - Rascal Flatts			36	3.78	0.24	67	0.735	0.447
3	Let It Be - The Beatles			40	3.78	0.19	150	0.782	0.439
4	Upside Down - Jack Johnson			92	3.66	0.14	137	0.581	0.241
5	September - Earth, Wind & Fire			55	3.6	0.16	15	0.429	0.180
6	Time of Our Lives - Pitbull			24	3.58	0.26	39	0.424	0.224
7	Vienna - Billy Joel			24	3.58	0.28	140	0.427	0.235
8	Just the Two of Us (feat. Bill Withers) - Grover Washing			25	3.56	0.22	78	0.377	0.180
9	Voulez-Vous - ABBA			20	3.55	0.27	69	0.386	0.203
10	Let it Happen - Tame			22	3.55	0.29	73	0.392	0.214
11	Careless Whisper - George Michael			24	3.54	0.24	87	0.357	0.175
12	Take Five - Dave Brubeck			18	3.5	0.33	37	0.346	0.197
13	Clairo - Juna			18	3.5	0.28	57	0.324	0.168
14	We Are The Champions - Queen			22	3.5	0.24	77	0.294	0.143
15	All Star - Smash Mouth			17	3.41	0.37	0	0.273	0.160
16	Feel it Still - Portugal. The Man			17	3.41	0.32	68	0.247	0.132
17	Magnetic - ILLIT			15	3.4	0.38	16	0.272	0.159
18	The Dock of the Bay - Otis Redding			13	3.38	0.38	107	0.254	0.148

Sample Mean and Standard Error for Pset Time Working



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

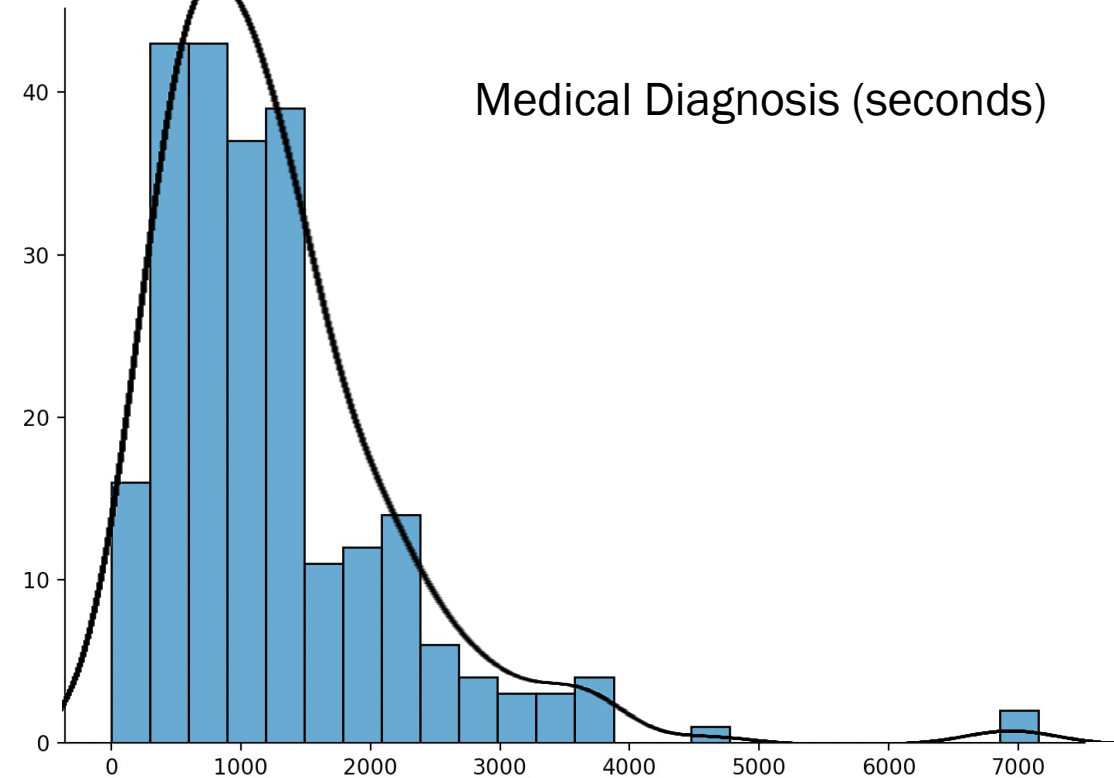
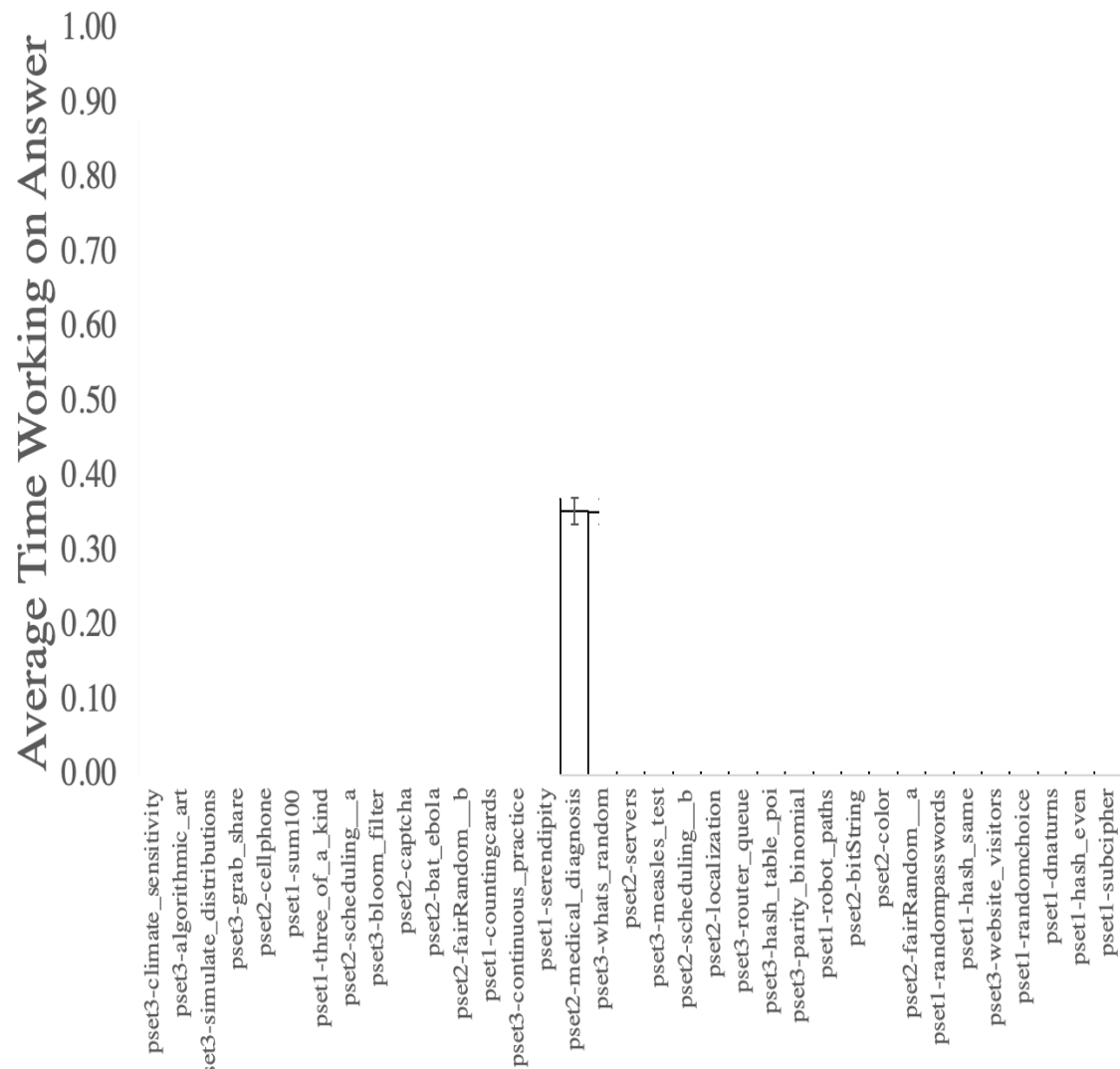
Total: 12.1 hours on answers
Budget: 50 hours for psets

Statistics Vs Distribution

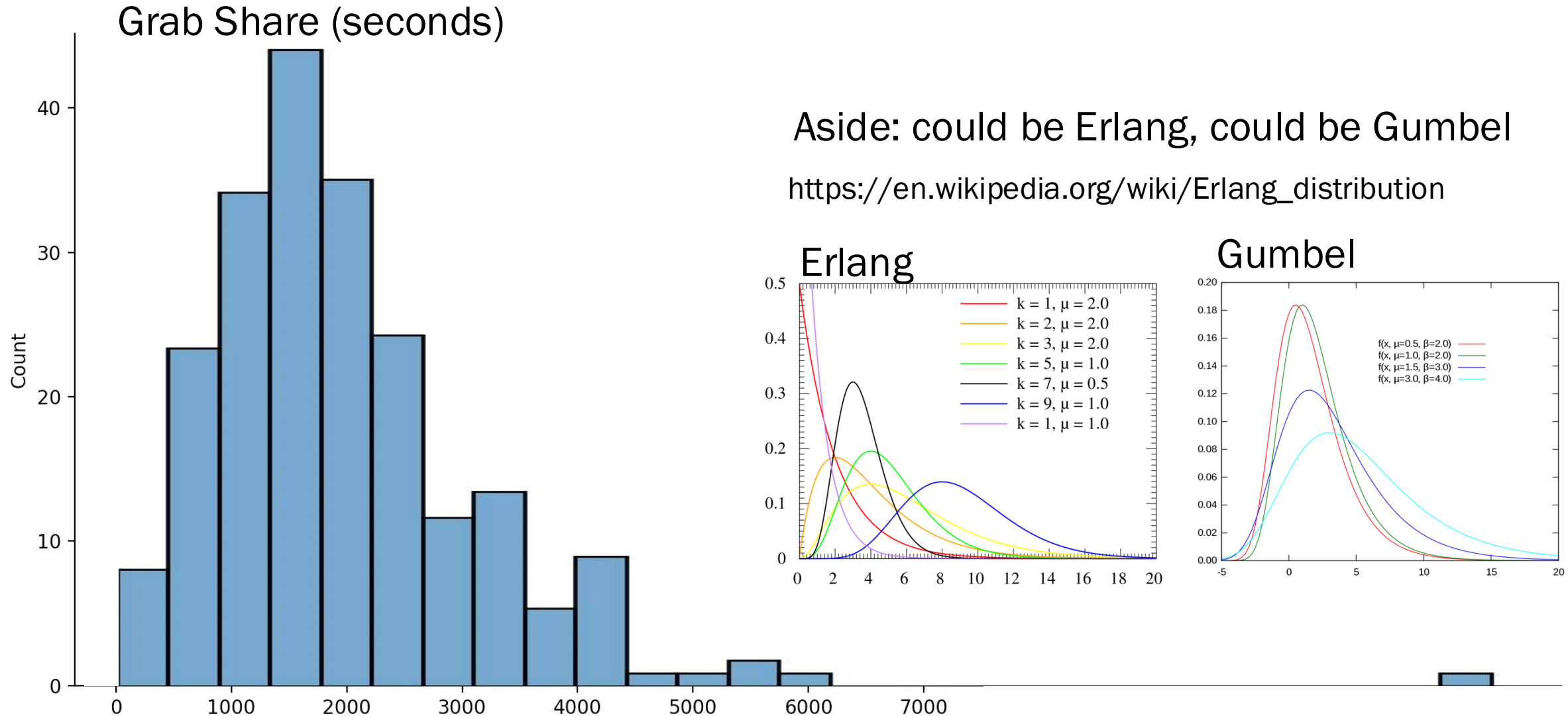
Sampling statistics

vs

Sampling distribution



[Aside] Distribution of PSet Completion Times



I don't think its exponential. Must not be a poisson process!

Happiness in Bhutan

Motivating example

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

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

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

- The mean of all these numbers is 83.

Is this the **true mean happiness** of Bhutanese people?



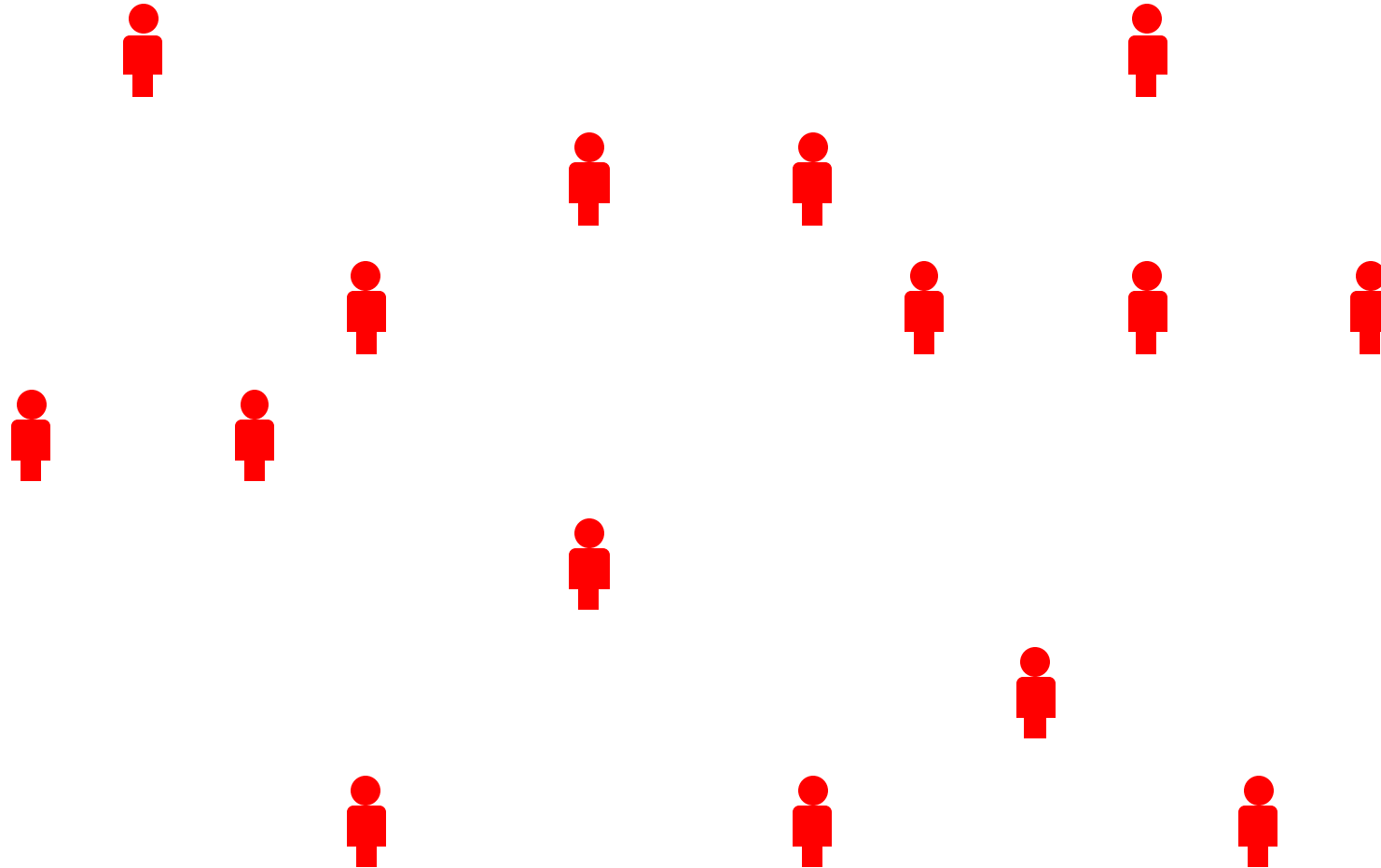
Population



Sample



Sample

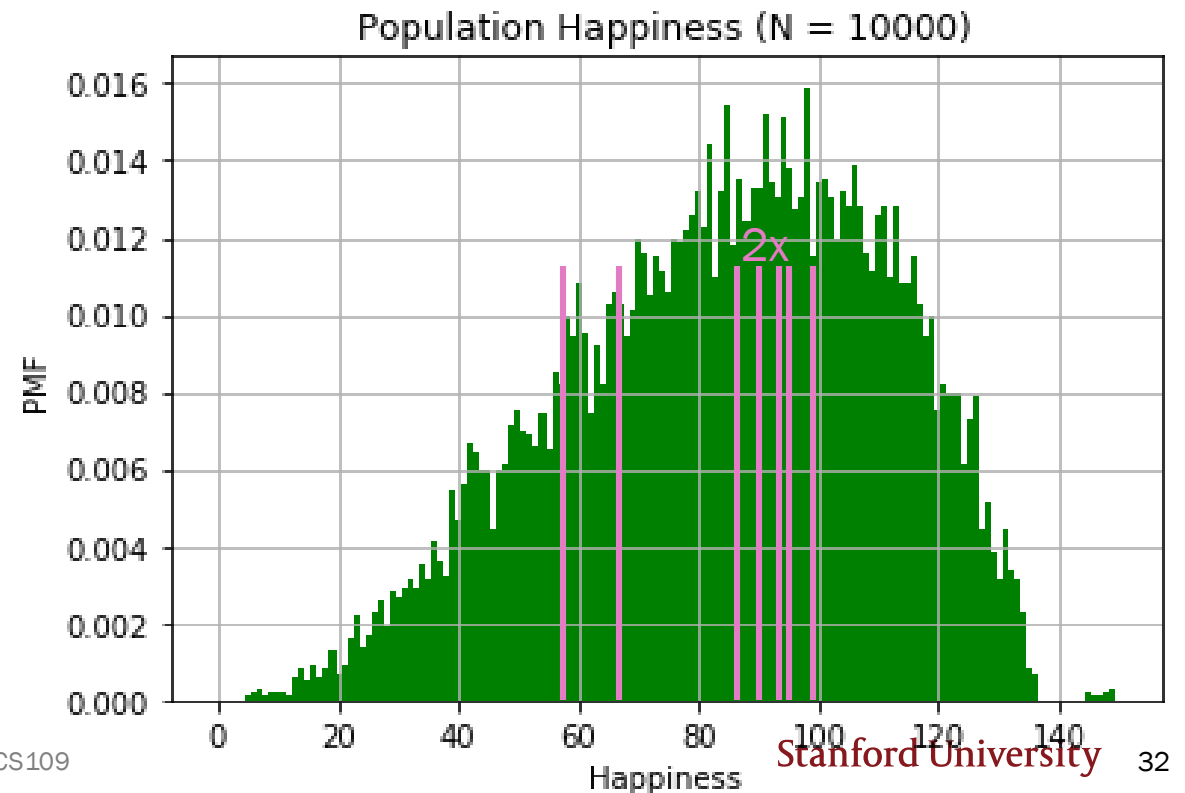


Collect one (or more) numbers from each person

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



A single sample



A happy
person

If we had a distribution F of our entire population, we could compute exact statistics about about happiness.

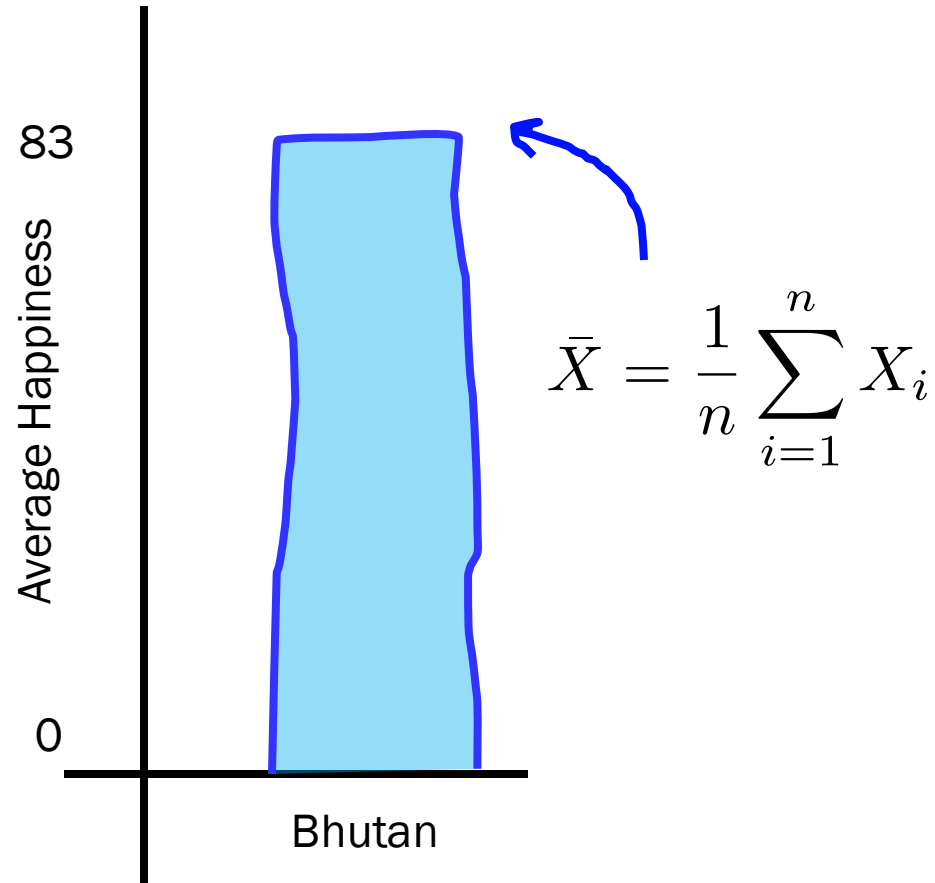
But we only have 200 people (a sample).

Today: If we only have a sample,

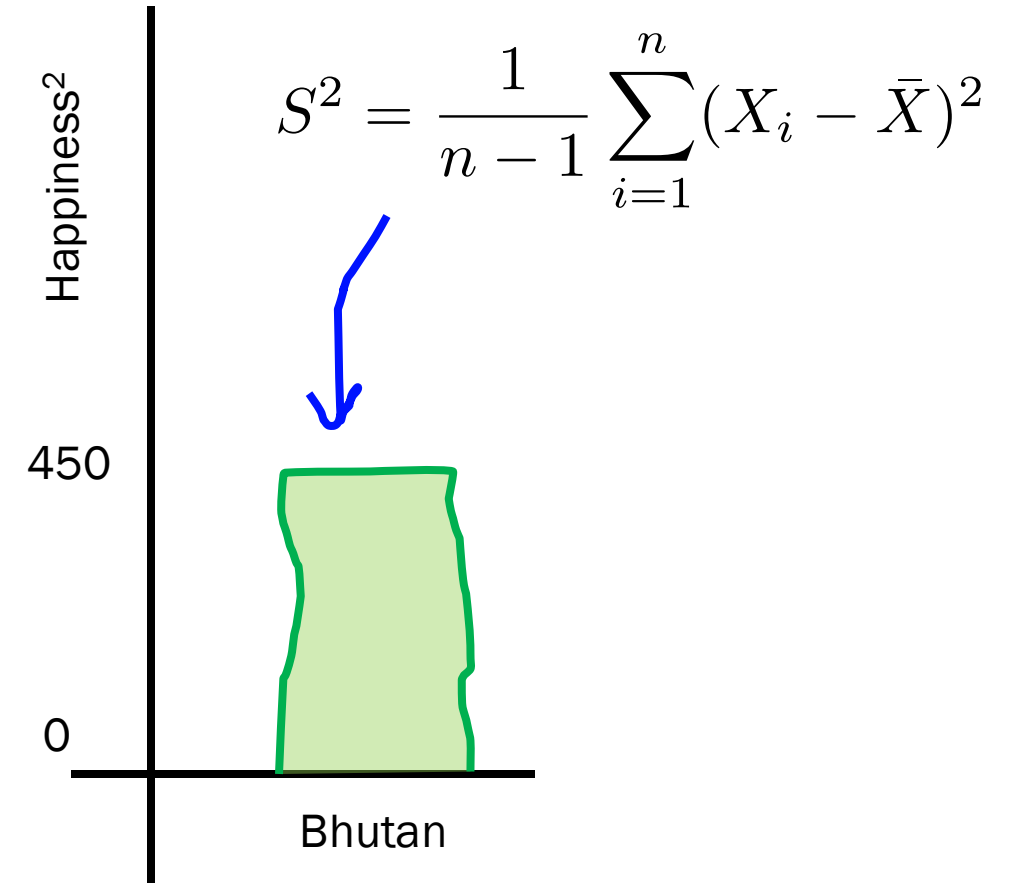
- How do we report *estimated* statistics?
- How do we report estimated error of these estimates?
- How do we perform hypothesis testing?

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

Average Happiness



Variance of Happiness



Side quest: sample variance

Estimating the population variance



What is σ^2 , the **variance of happiness** of Bhutanese people?

If we knew the entire population (x_1, x_2, \dots, x_N) :

population variance

$$\sigma^2 = E[(X - \mu)^2] = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

The diagram shows the text "population mean" in purple above the symbol μ in the equation. A purple arrow points downwards from "population mean" to μ . The text "population variance" is to the left of the equation.

If we only have a sample, (X_1, X_2, \dots, X_n) :

sample variance

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

sample mean

The diagram shows the text "sample mean" in purple above the symbol \bar{X} in the equation. A purple arrow points downwards from "sample mean" to \bar{X} . The text "sample variance" is to the left of the equation.



Intuition about the sample variance, S^2

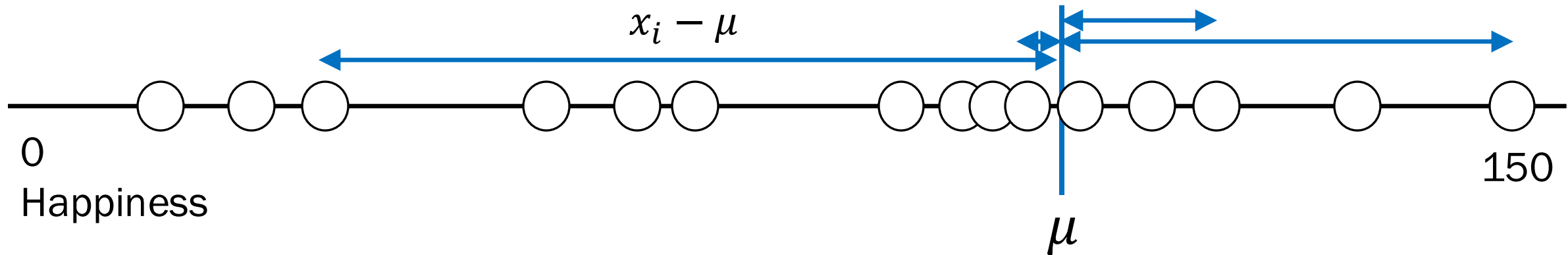


Actual, σ^2

population mean

population variance

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



Population size, N

Calculating population statistics exactly requires us knowing all N datapoints.

Intuition about the sample variance, S^2



Actual, σ^2

Estimate, S^2

population
variance

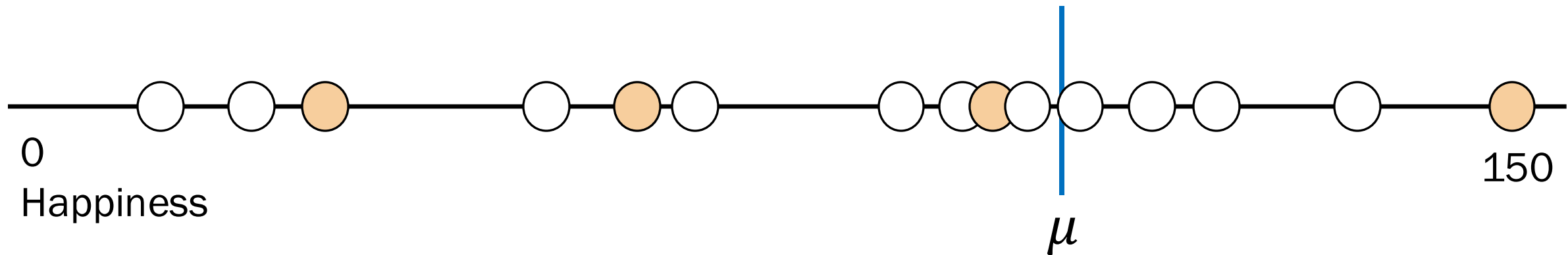
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

sample
variance

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

sample mean



Population size, N

Intuition about the sample variance, S^2



Actual, σ^2

Estimate, S^2

population
variance

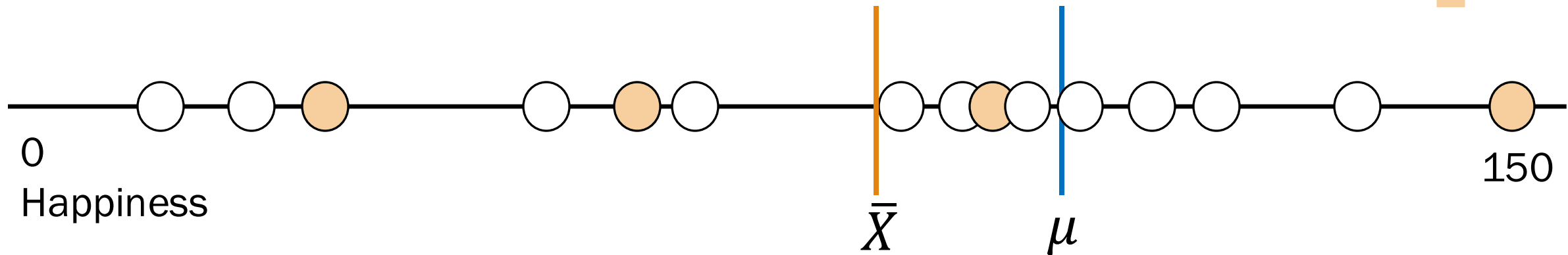
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

sample
variance

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

sample mean



Population size, N

Intuition about the sample variance, S^2



Actual, σ^2

Estimate, S^2

population variance

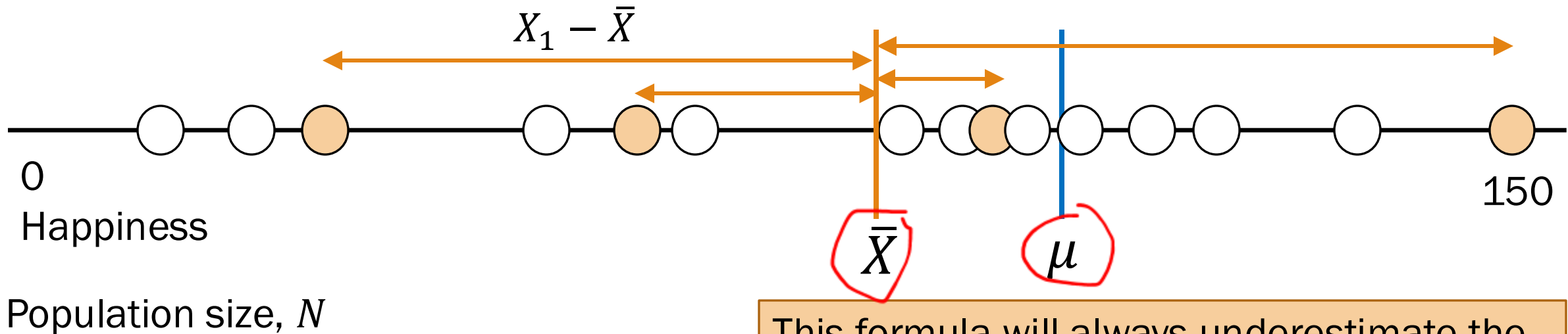
population mean

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

sample variance

sample mean

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



This formula will always underestimate the variance...

Ahhh! We are always underestimating!
What should we do?

Estimating the population variance



2. What is σ^2 , the **variance of happiness** of Bhutanese people?

If we knew the entire population (x_1, x_2, \dots, x_N) :

population variance

$$\sigma^2 = E[(X - \mu)^2] = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

Bug!



If we only have a sample, (X_1, X_2, \dots, X_n) :

sample variance

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

sample mean

Estimating the population variance



2. What is σ^2 , the **variance of happiness** of Bhutanese people?

If we knew the entire population (x_1, x_2, \dots, x_N) :

population variance

$$\sigma^2 = E[(X - \mu)^2] = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

An orange arrow points from the text "population mean" to the Greek letter μ in the equation.

If we only have a sample, (X_1, X_2, \dots, X_n) :

sample variance

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

sample mean

An orange arrow points from the text "sample mean" to the \bar{X} in the equation. The denominator $n-1$ is circled in red.

Proof that S^2 is unbiased (just for reference)

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

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

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

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

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

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

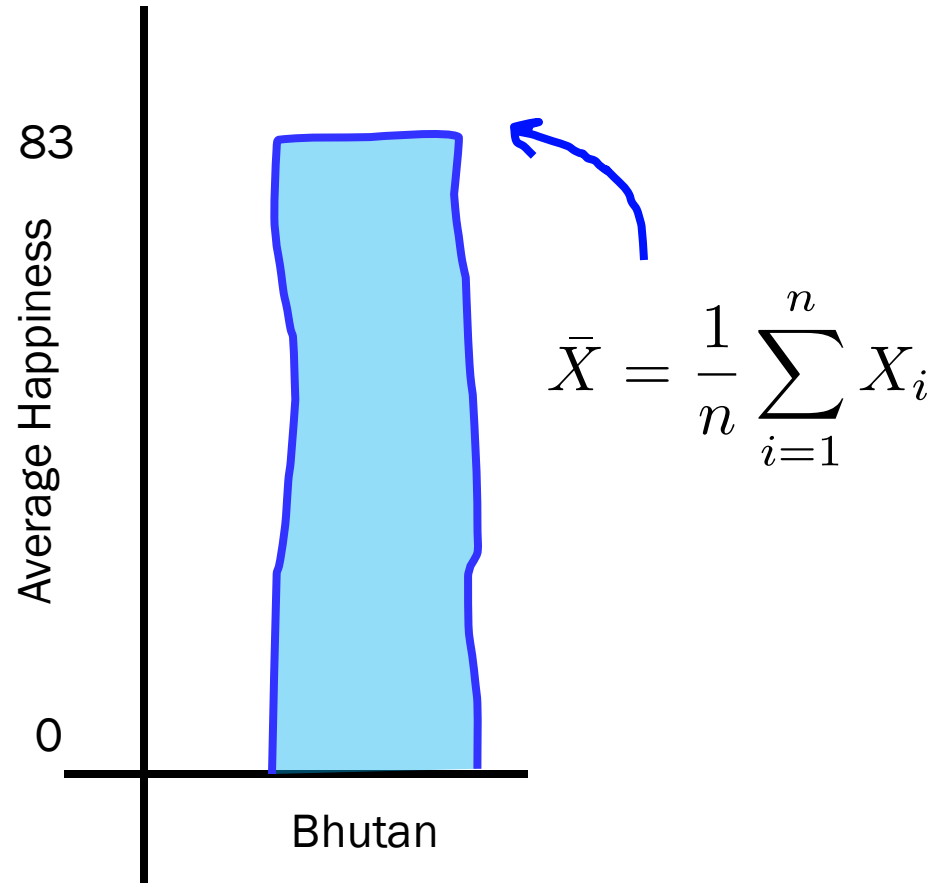
$$= n\sigma^2 - n\text{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Therefore $E[S^2] = \sigma^2$

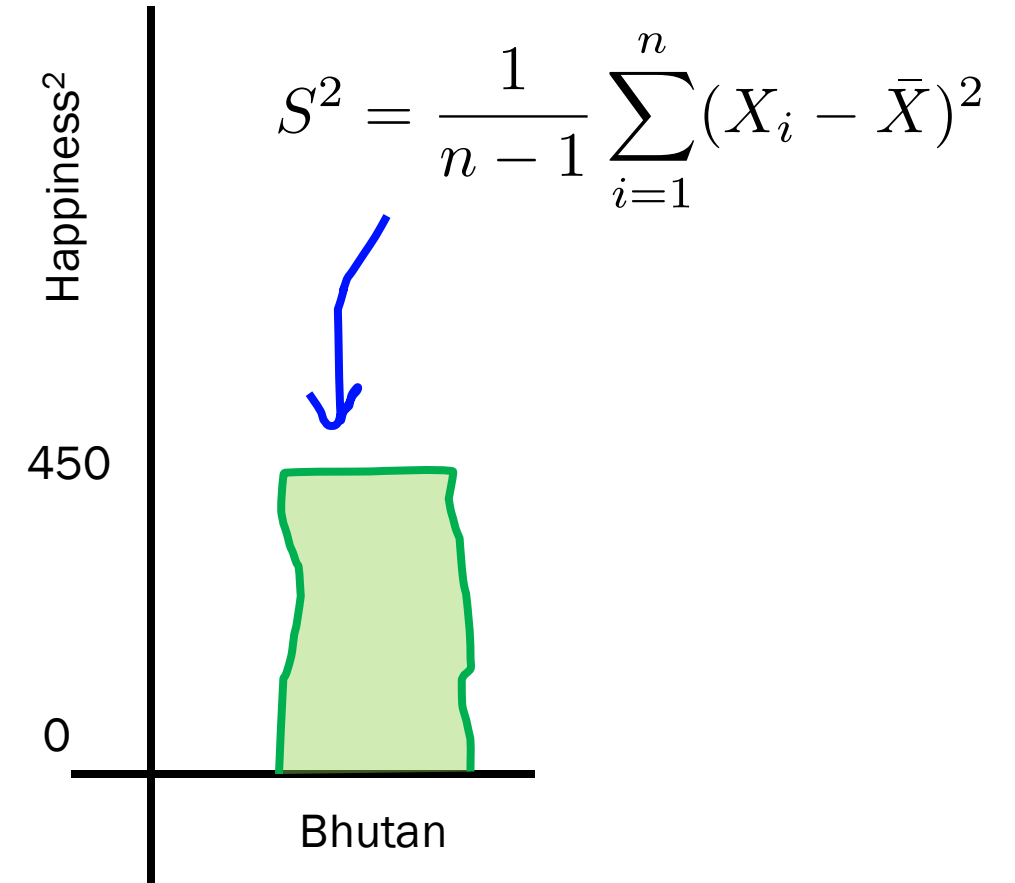
End Side Quest

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

Average Happiness



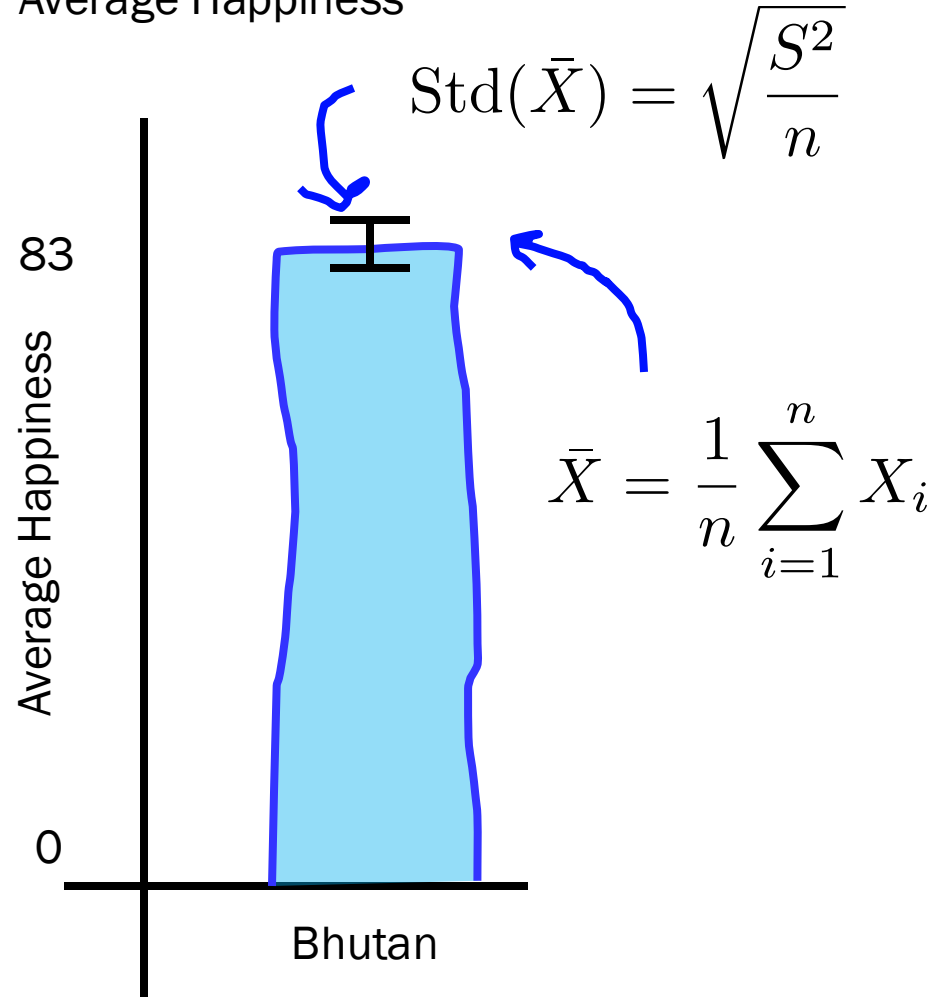
Variance of Happiness



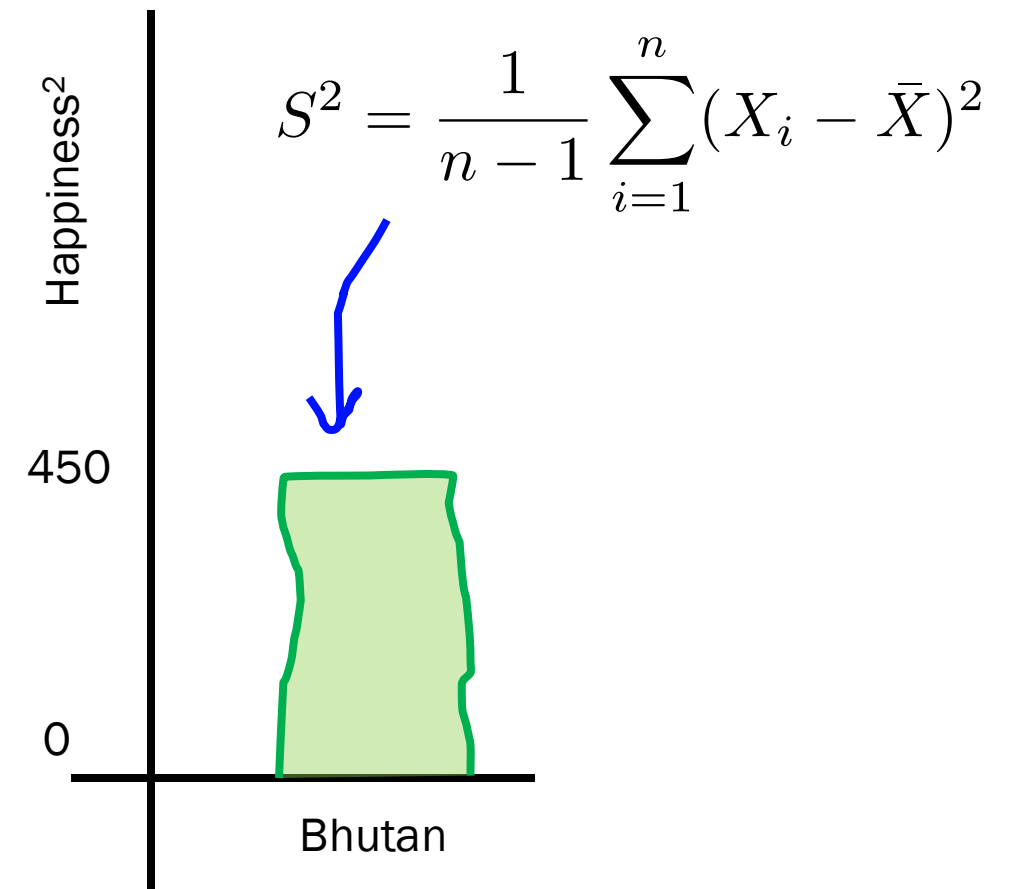
But what about **error bars**???

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

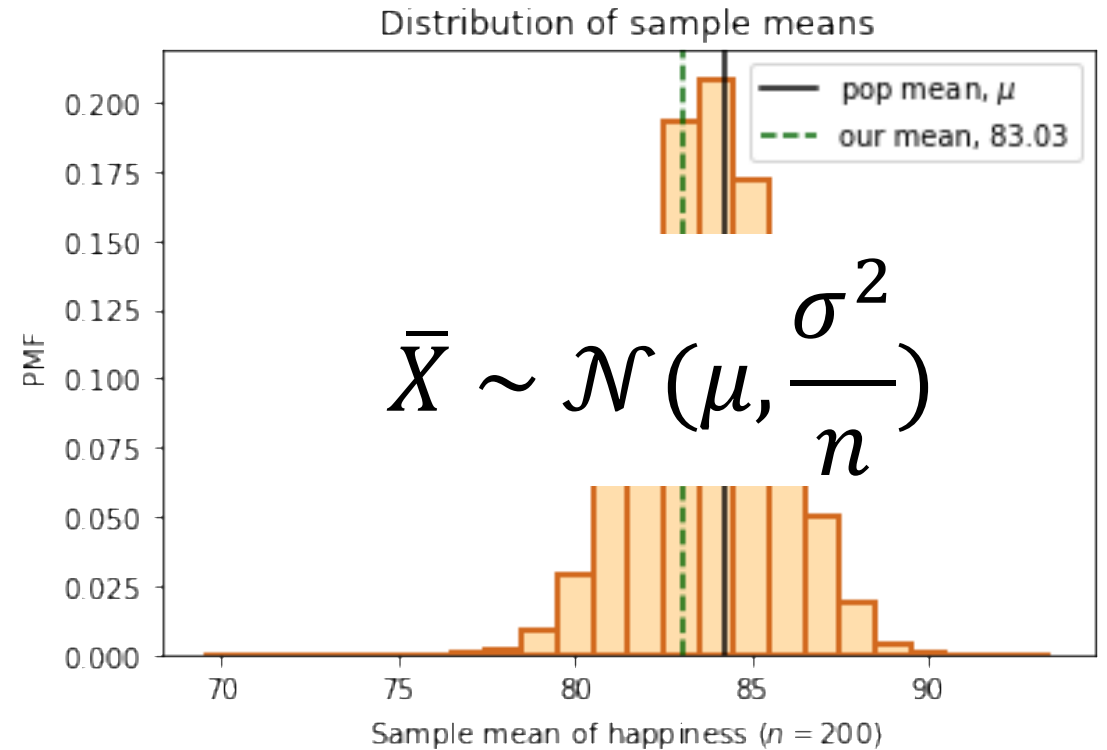
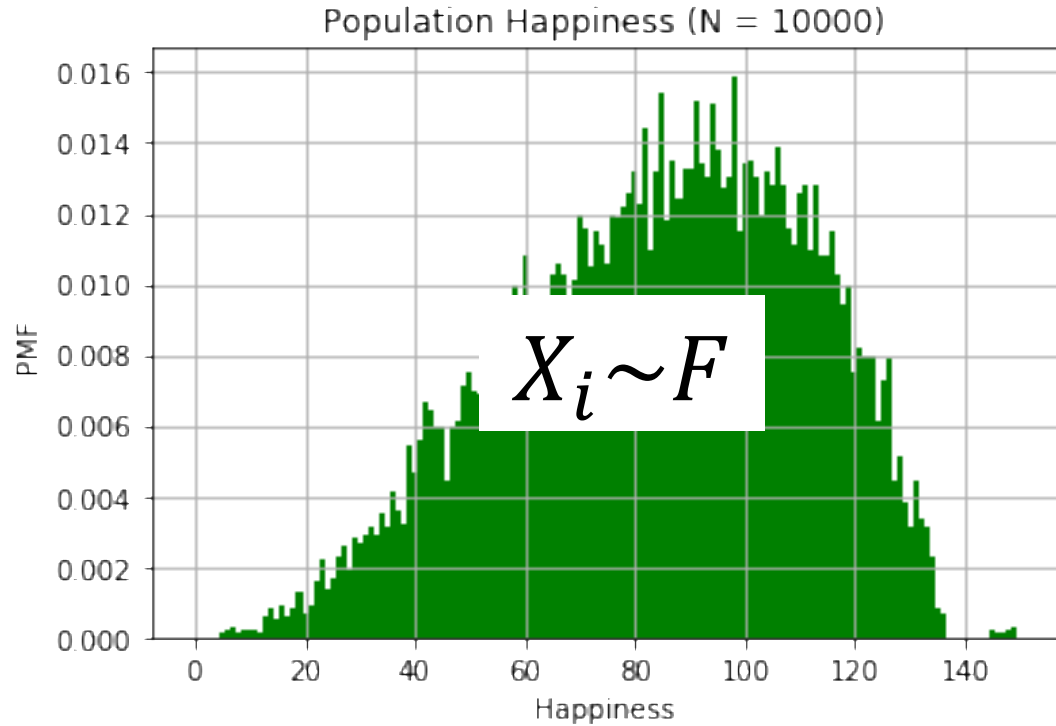
Average Happiness



Variance of Happiness



Sample mean by the CLT



Even if we can't report μ , we can report our sample mean 83.03, which is an unbiased estimate of μ .

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
variance
estimate

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

sample mean

Our best guess at
the true variance

Std error of
the mean
estimate

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

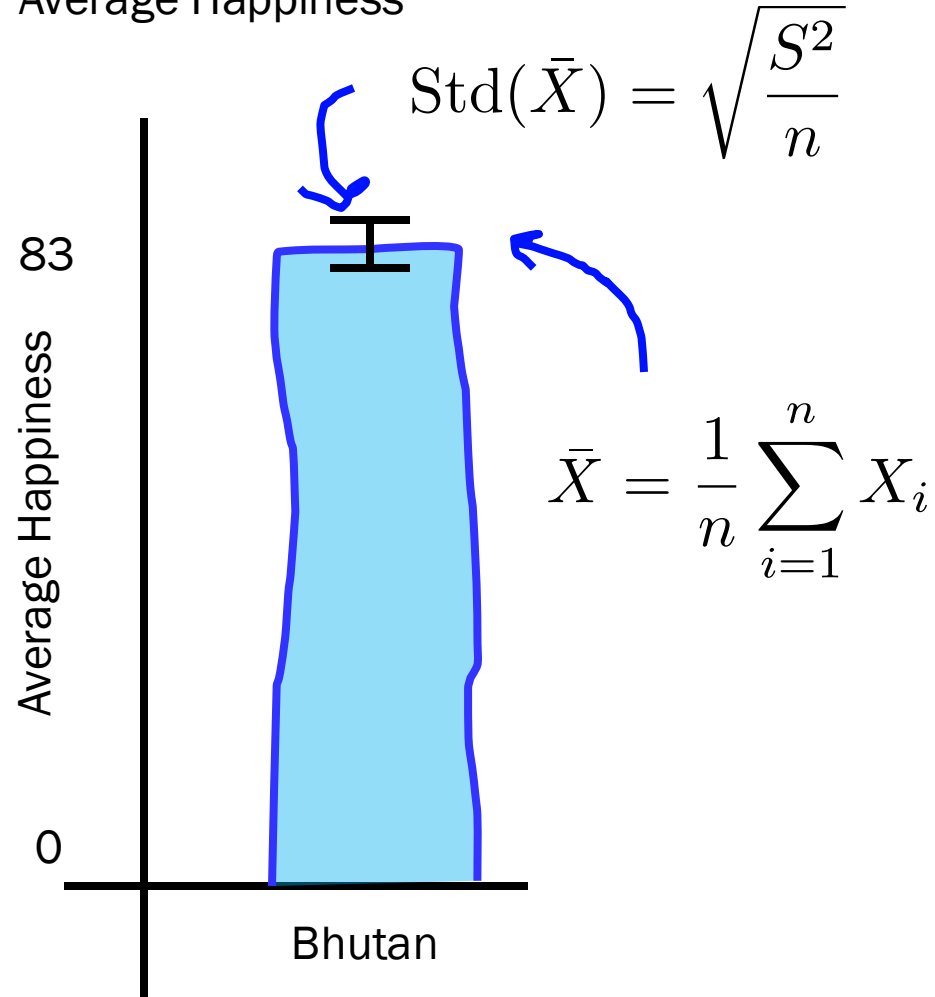
sample variance

How wrong do we
think our mean
estimate is?

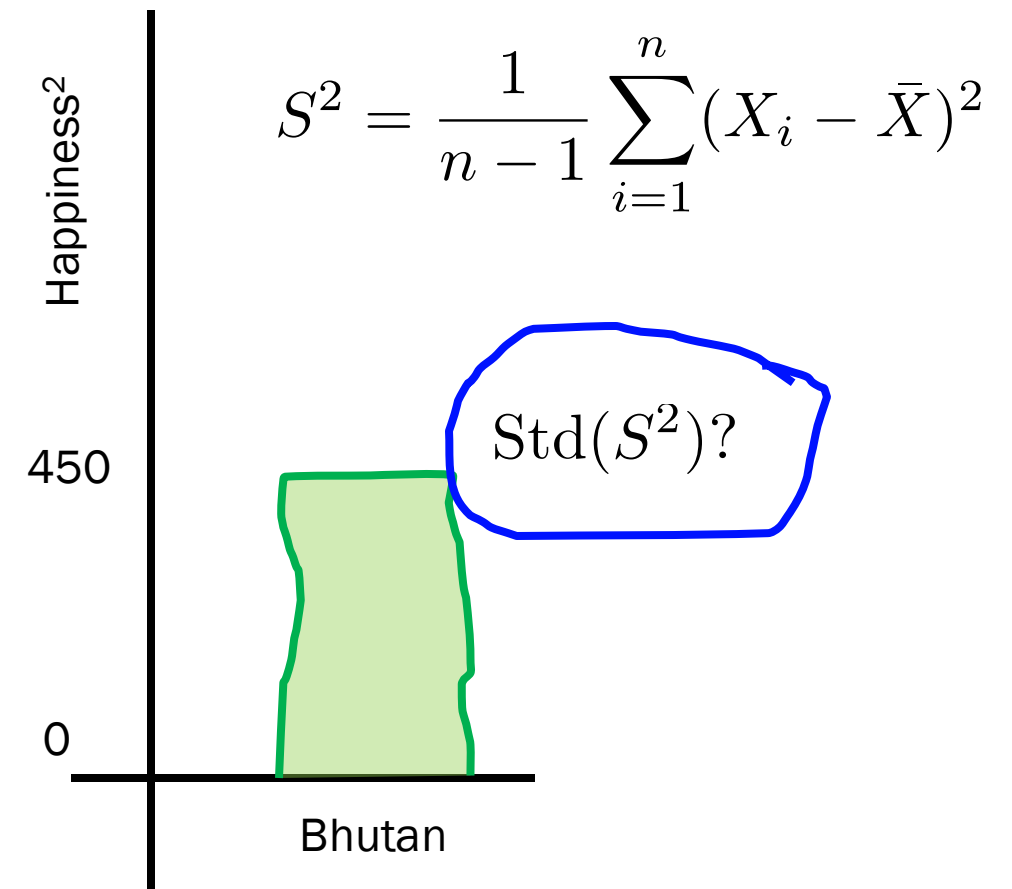
But what about **error bars**???

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

Average Happiness



Variance of Happiness



Hypothetical – You have the underlying distribution!

How wrong is an estimate of **sample variance**, calculated from 200 people?

Plot twist: I give you the *entire* underlying distribution



Hypothetical – You have the underlying distribution!

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

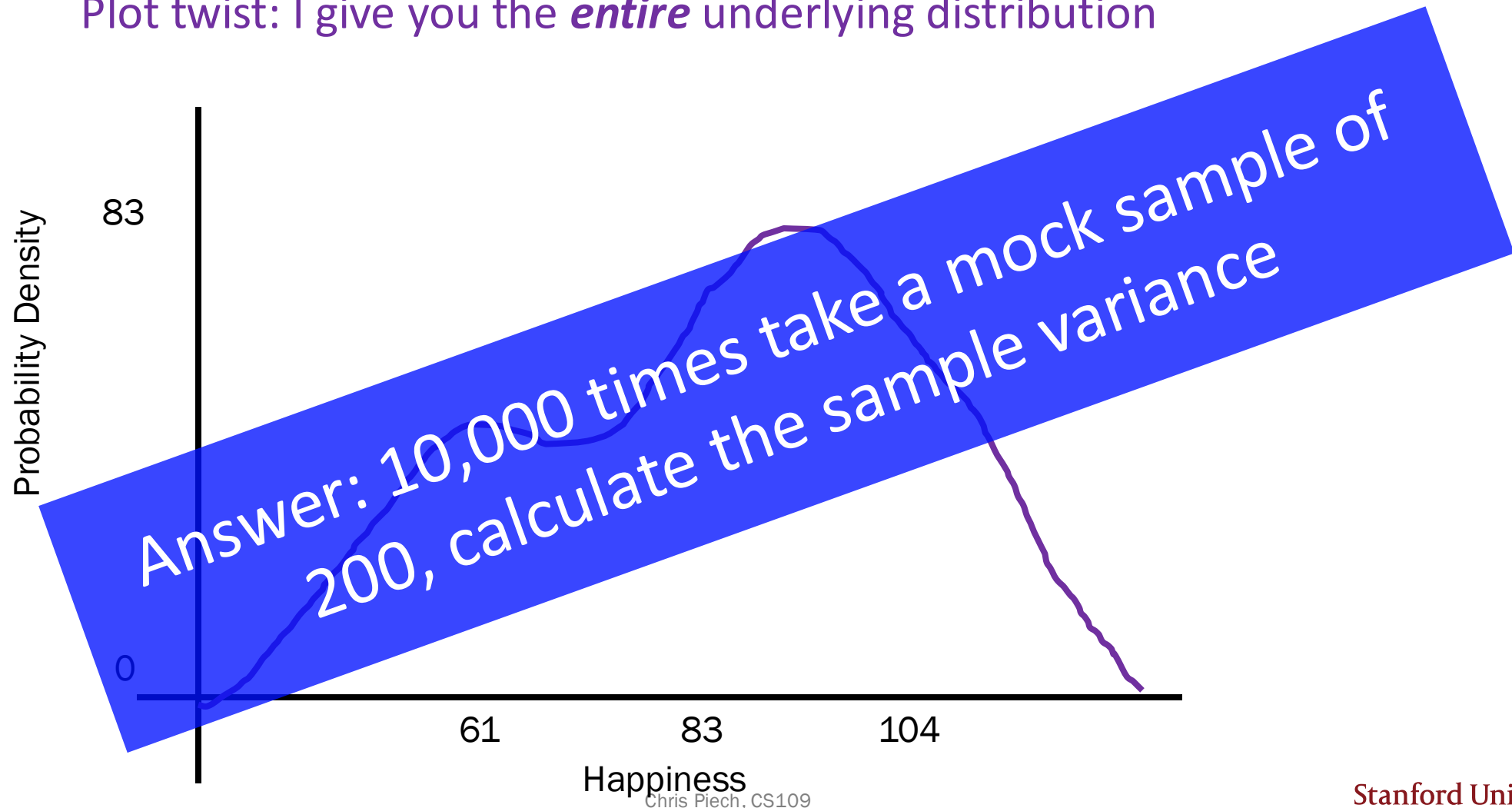
Plot twist: I give you the *entire* underlying distribution



Hypothetical – You have the underlying distribution!

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

Plot twist: I give you the *entire* underlying distribution



Hypothetical – You have the underlying distribution!

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

```
brute_force_algorithm():
```

```
# Estimate distribution of Sample Var with  
# infinite resources
```

```
sample_vars = []
```

```
Repeat 10,000 times:
```

```
    new_samples = collect_new_samples(n=200)
```

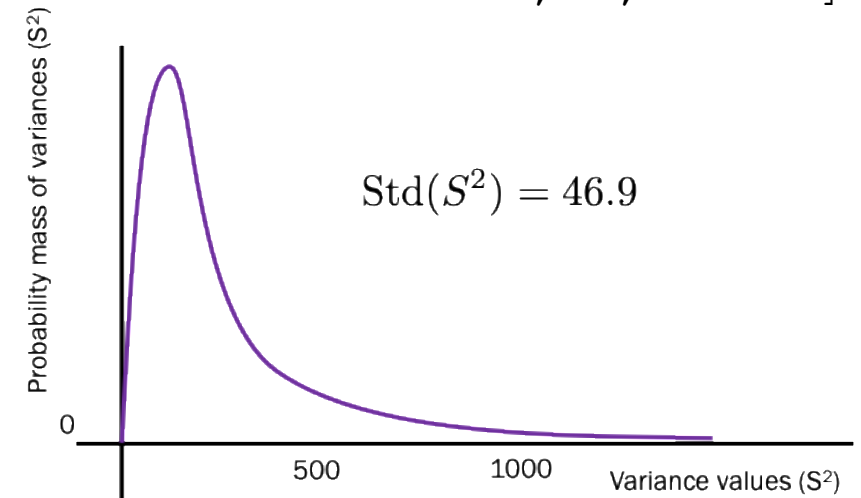
```
    sample_var = calculate_sample_var(new_samples)
```

```
    sample_vars.append(sample_var)
```

```
# You now have a distribution of sample vars
```



```
sample_vars = [472.7, 478.4,  
              469.2, ..., 476.2]
```



[suspense]

Bootstrap: Probability for Computer Scientists

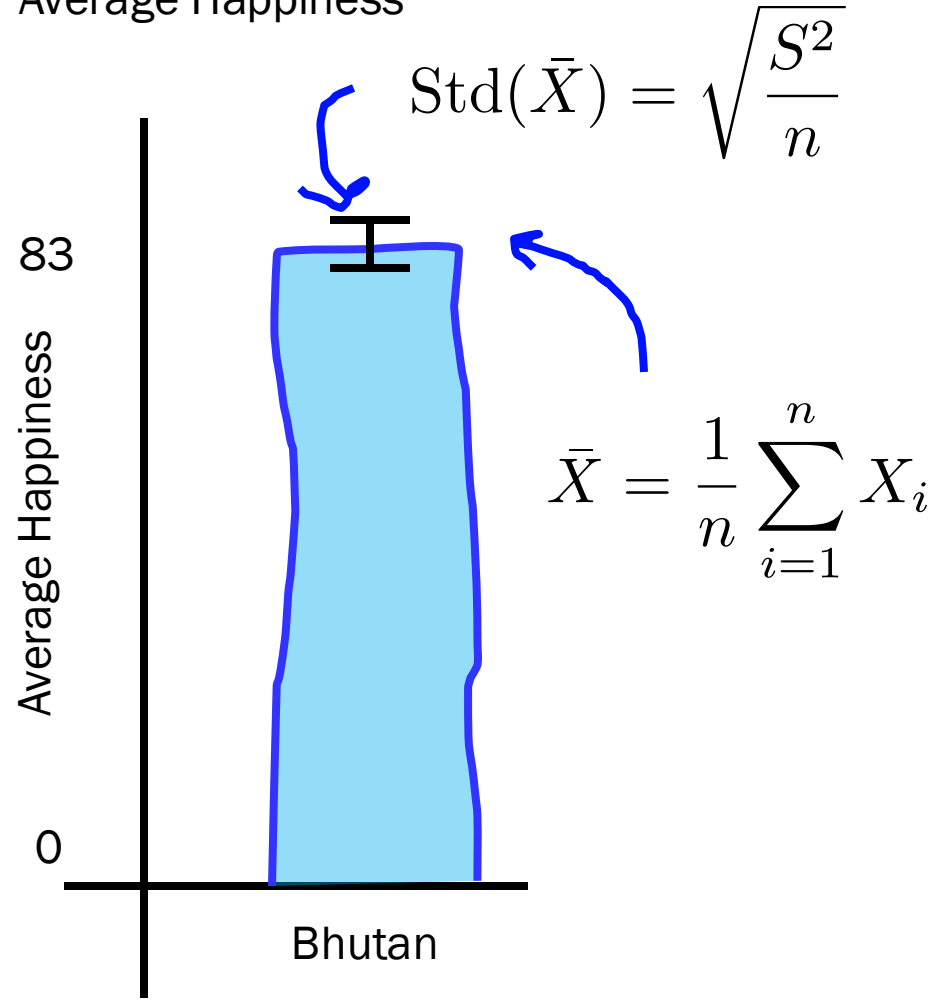
Bootstrapping allows you to:

- Know the **distribution of *statistics***
- Calculate **p values**
- **Using computers**
- You totally **could have invented it**

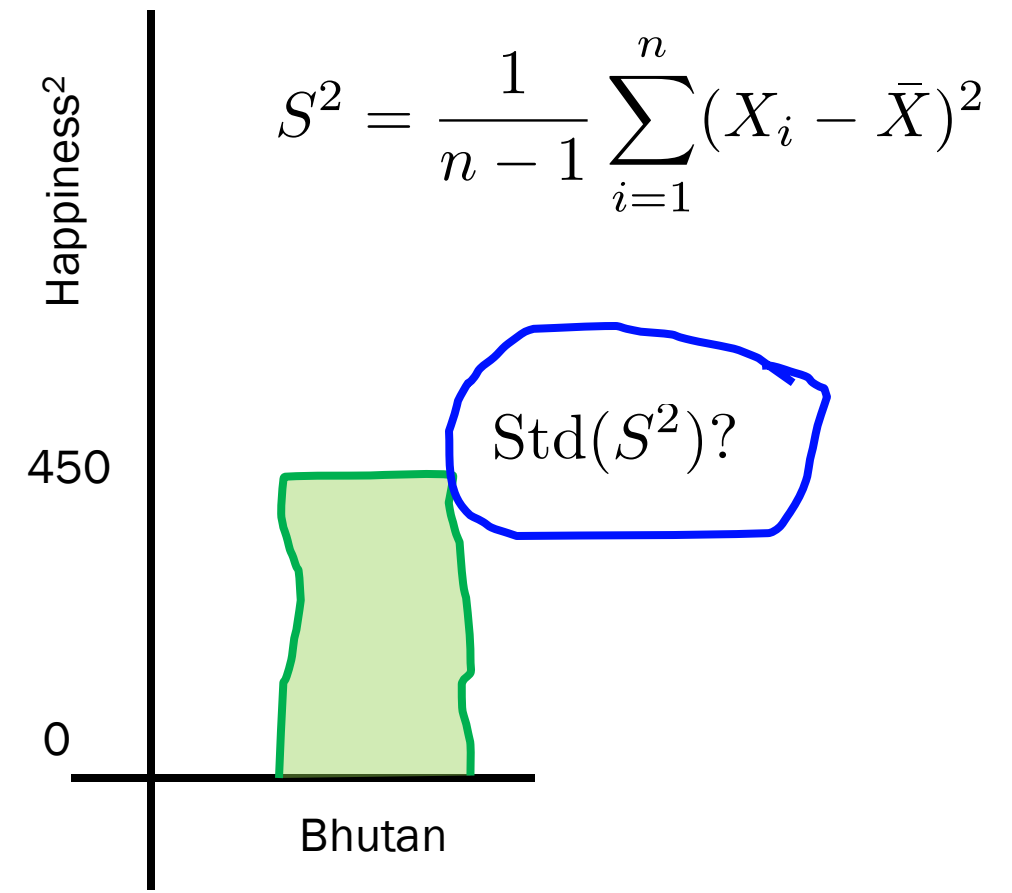
But what about **error bars**???

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

Average Happiness



Variance of Happiness



Hypothetical – You have the underlying distribution!

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

```
brute_force_algorithm():
```

```
# Estimate distribution of Sample Var with  
# infinite resources
```

```
sample_vars = []
```

```
Repeat 10,000 times:
```

```
    new_samples = collect_new_samples(n=200)
```

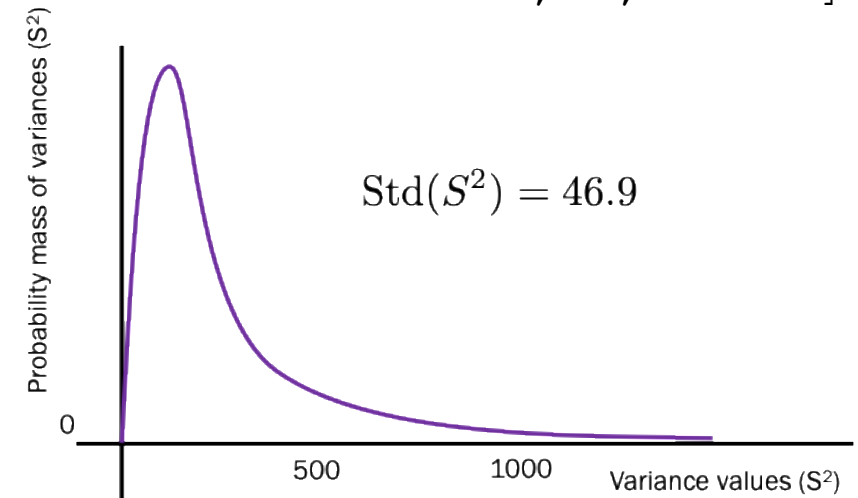
```
    sample_var = calculate_sample_var(new_samples)
```

```
    sample_vars.append(sample_var)
```

```
# You now have a distribution of sample vars
```



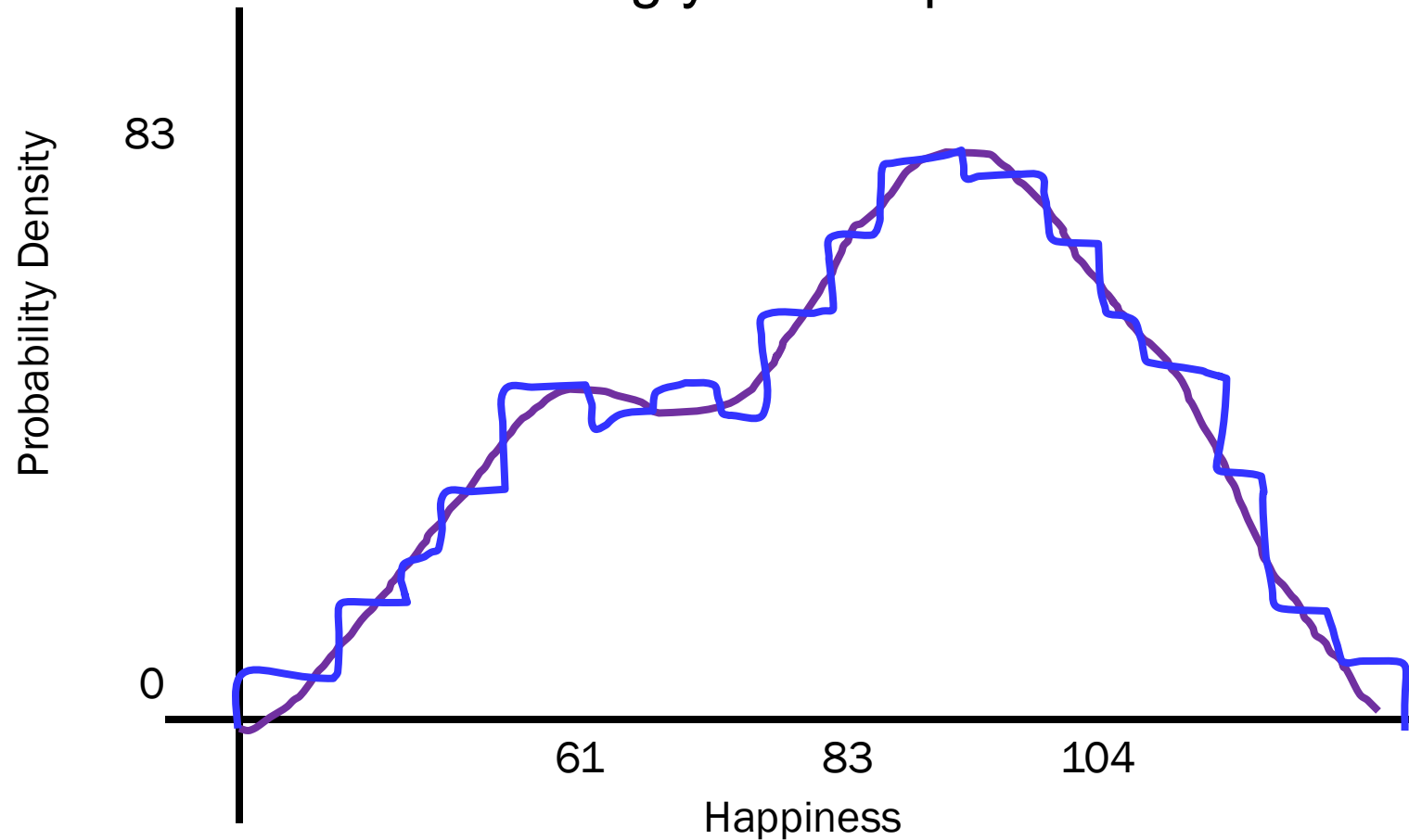
```
sample_vars = [472.7, 478.4,  
              469.2, ..., 476.2]
```



Here comes the award winning idea....

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!!

Chris Piech, CS109

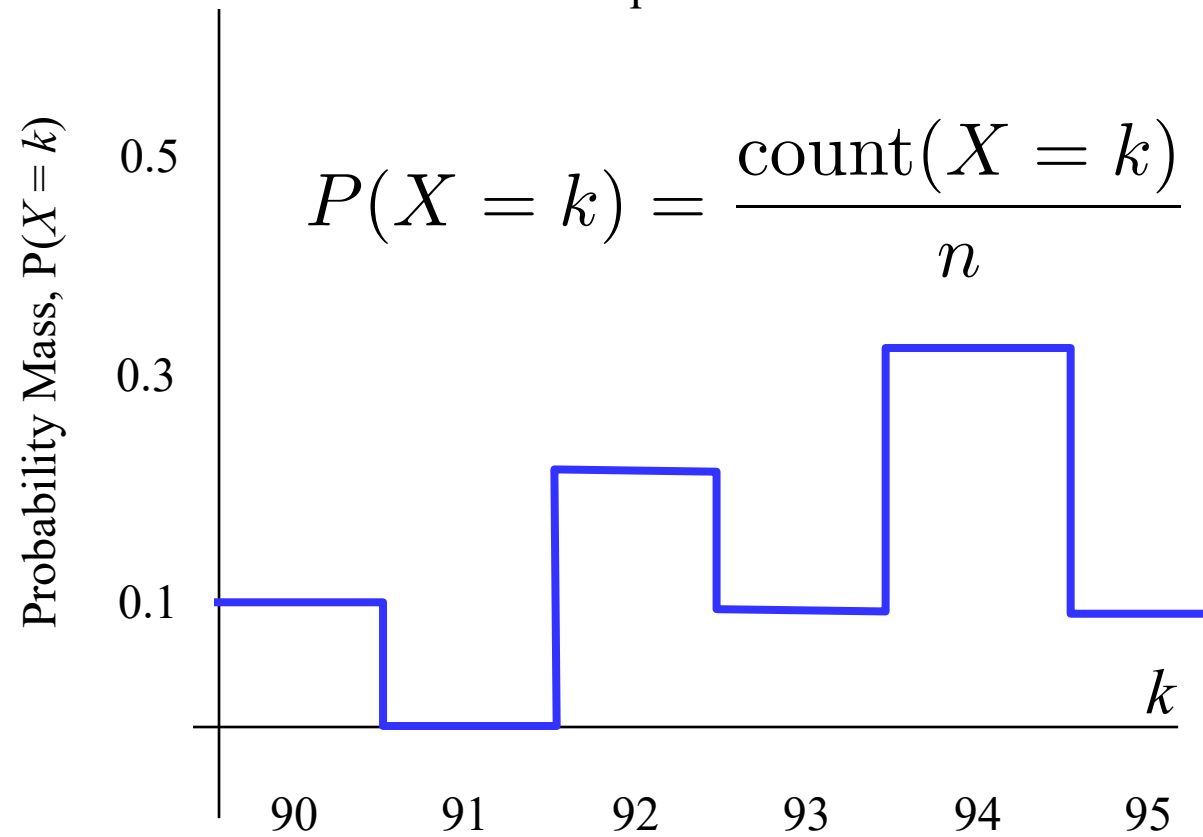
Stanford University

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):

Estimate the **PMF** using the sample

Repeat **10,000** times:

- a. Draw **len(sample)** new samples from PMF
- b. Recalculate the stat** on the resample

You now have a **distribution of your stat**

Bootstrapping of Variance

Bootstrap Algorithm (sample):

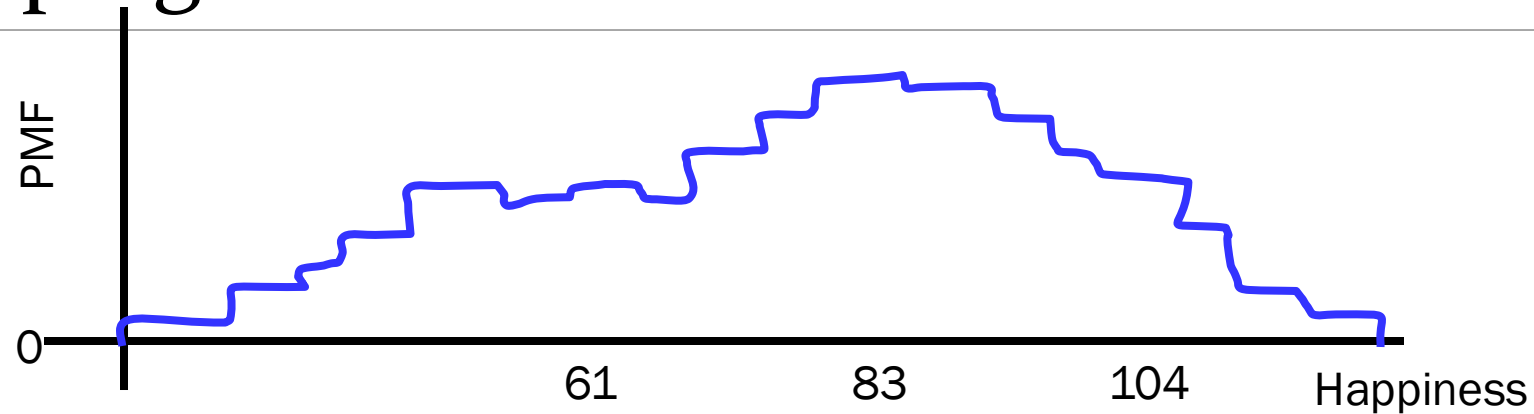
Estimate the **PMF** using the sample

Repeat **10,000** times:

- a. Draw **len(sample)** new samples from PMF
- b. Recalculate the variance** on the resample

You now have a **distribution of your variances**

Bootstrapping of Variance



Bootstrap Algorithm (sample):

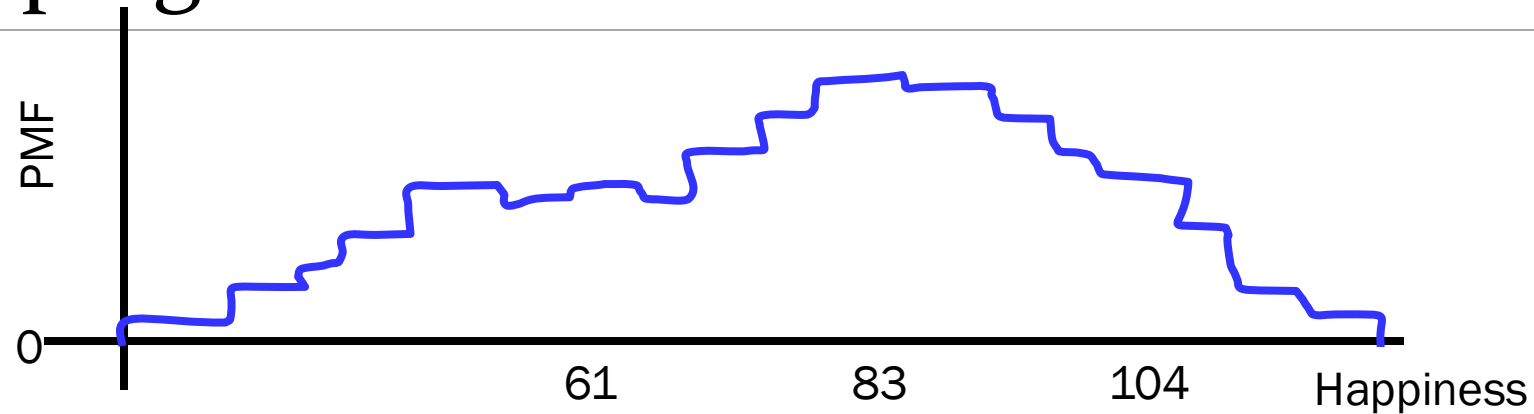
Estimate the **PMF** using the sample

Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the var** on the resample

You now have a **distribution of your vars**

Bootstrapping of Variance



Bootstrap Algorithm (sample):

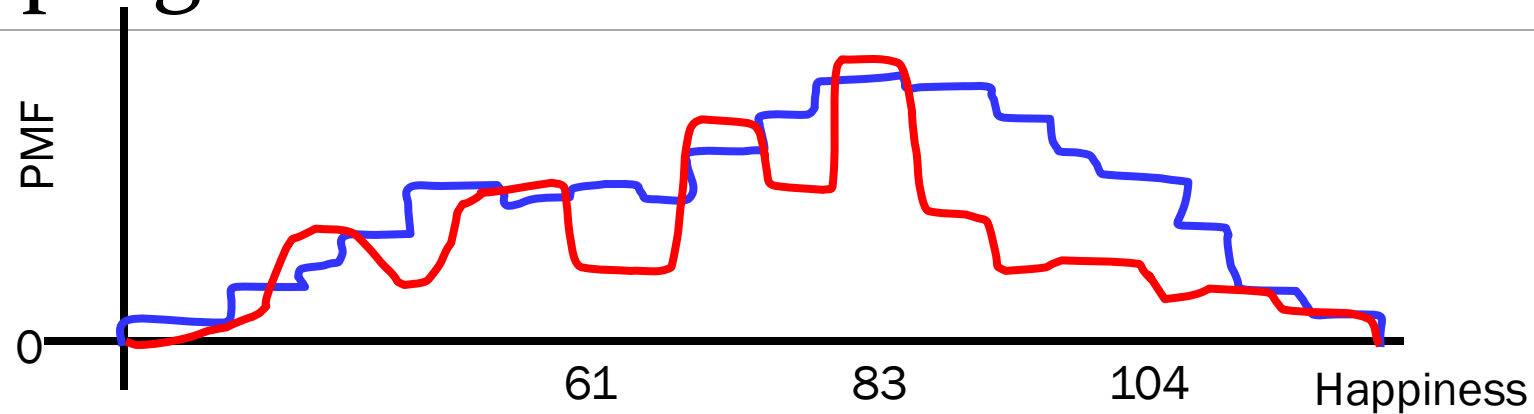
Estimate the **PMF** using the sample

Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the var** on the resample

You now have a **distribution of your vars**

Bootstrapping of Variance



Bootstrap Algorithm (sample):

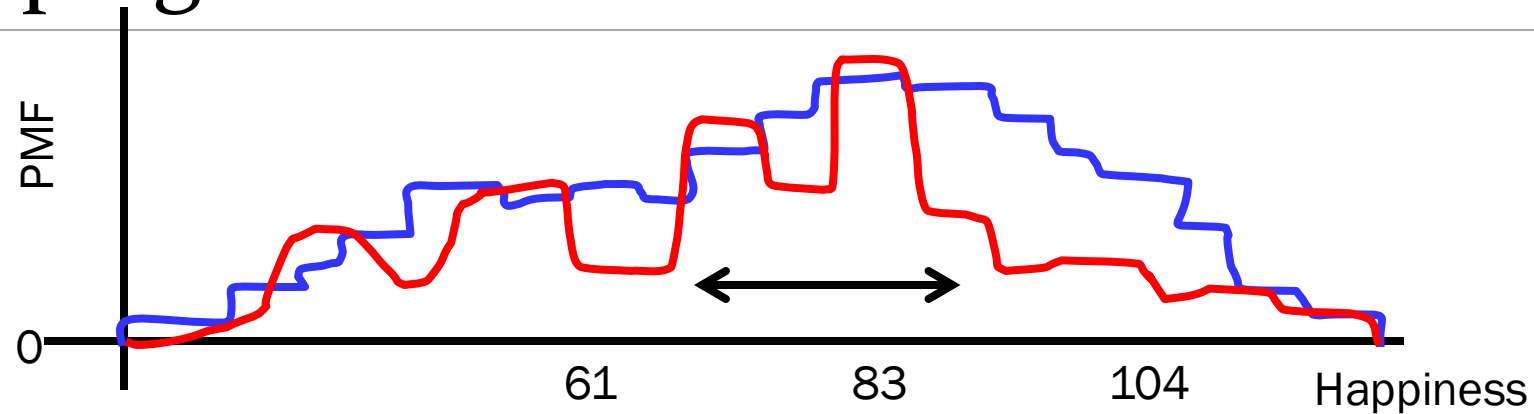
Estimate the **PMF** using the sample

Repeat **10,000** times:

- a. Draw **len(sample)** new samples from PMF
- b. Recalculate the **var** on the resample

You now have a **distribution of your vars**

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

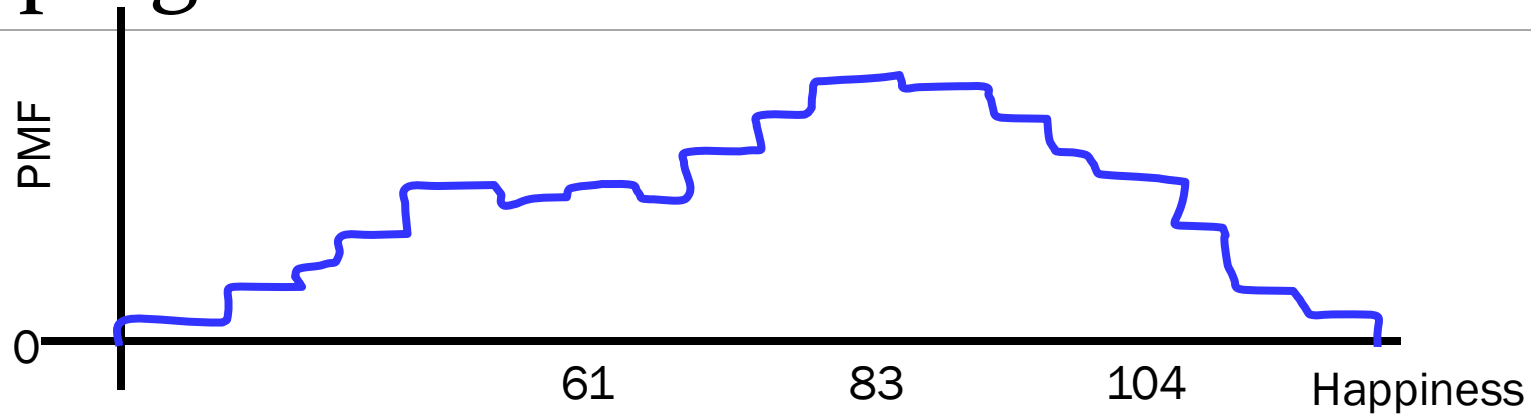
a. Draw `len(sample)` new samples from PMF

b. Recalculate the **vars** on the resample

You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

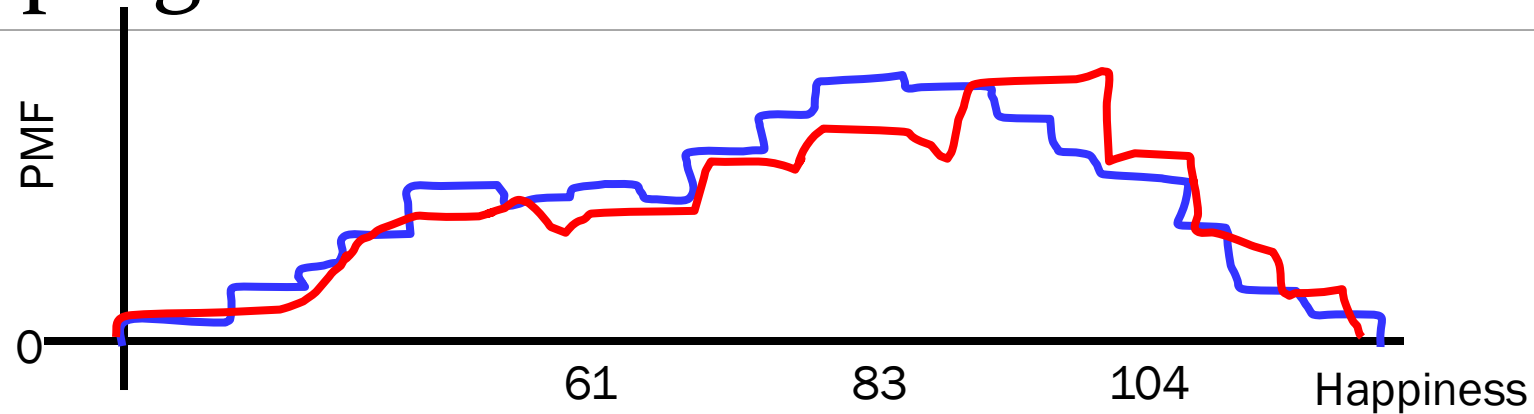
Repeat ~~10,000~~ times:

- a. Draw `len(sample)` new samples from PMF
- b. Recalculate the **var** on the resample

You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

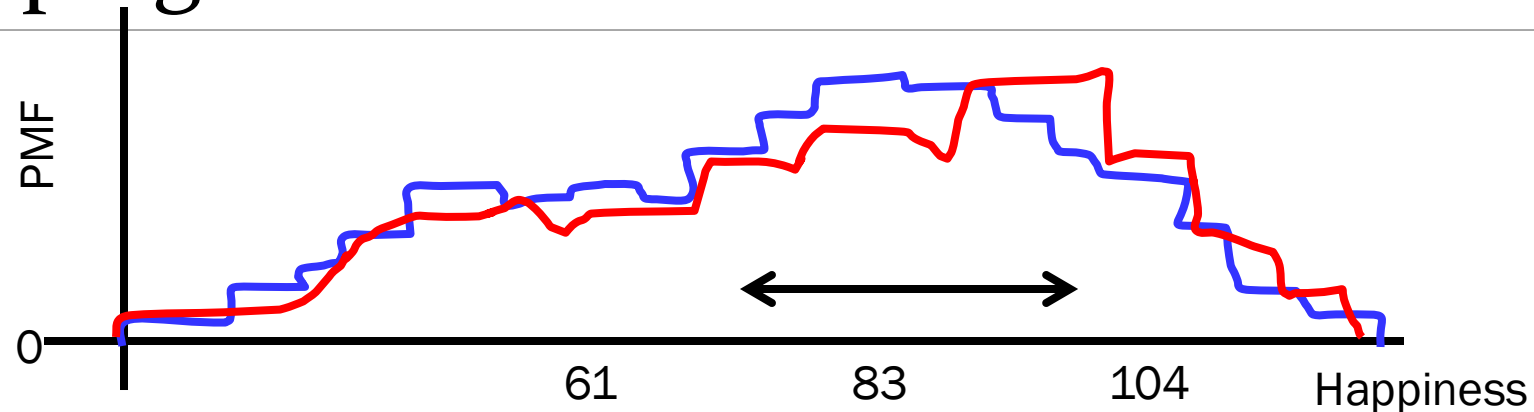
a. Draw `len(sample)` new samples from PMF

b. Recalculate the **var** on the resample

You now have a **distribution of your vars**

Vars = [472.7]

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

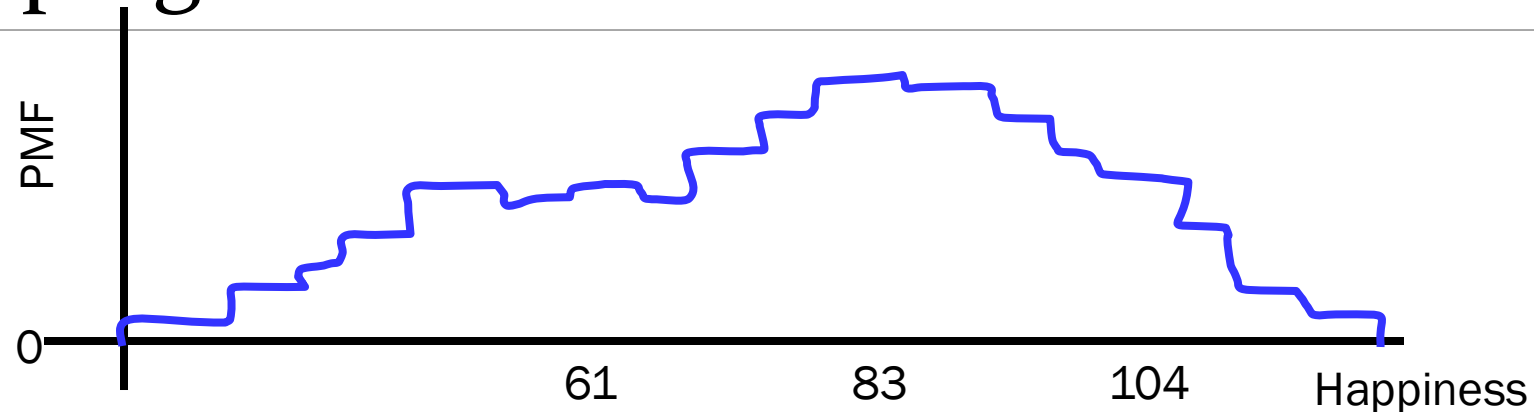
a. Draw **len(sample)** new samples from PMF

b. Recalculate the **var** on the resample

You now have a **distribution of your vars**

Vars = [472.7, 478.4]

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

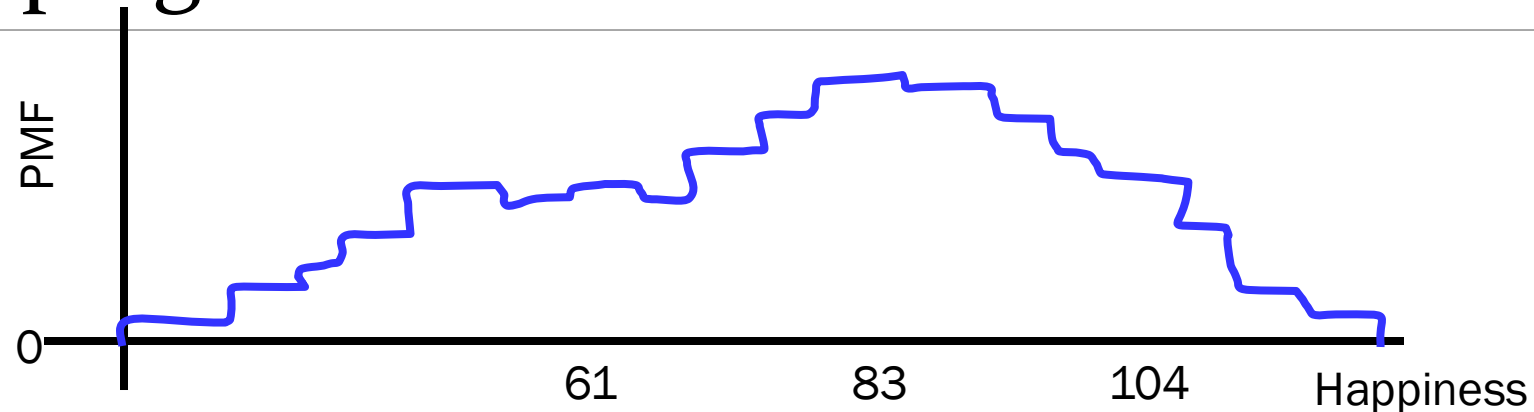
Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the var** on the resample

You now have a **distribution of your vars**

Vars = [472.7, 478.4]

Bootstrapping of Variance



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

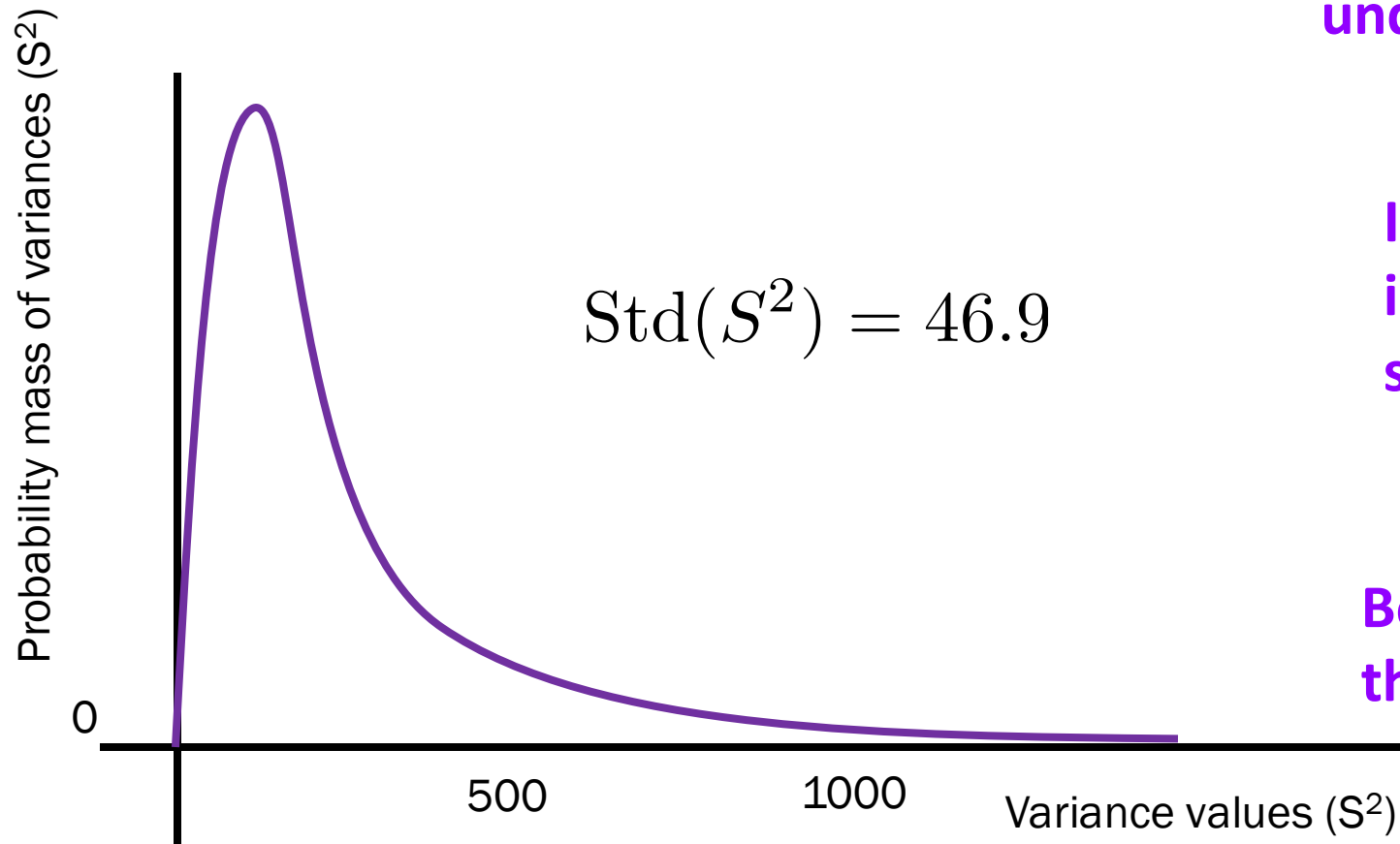
- a. Draw **len(sample)** new samples from PMF
- b. **Recalculate the var** on the resample

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]



Aside: the distribution of variance depends on the underlying distribution

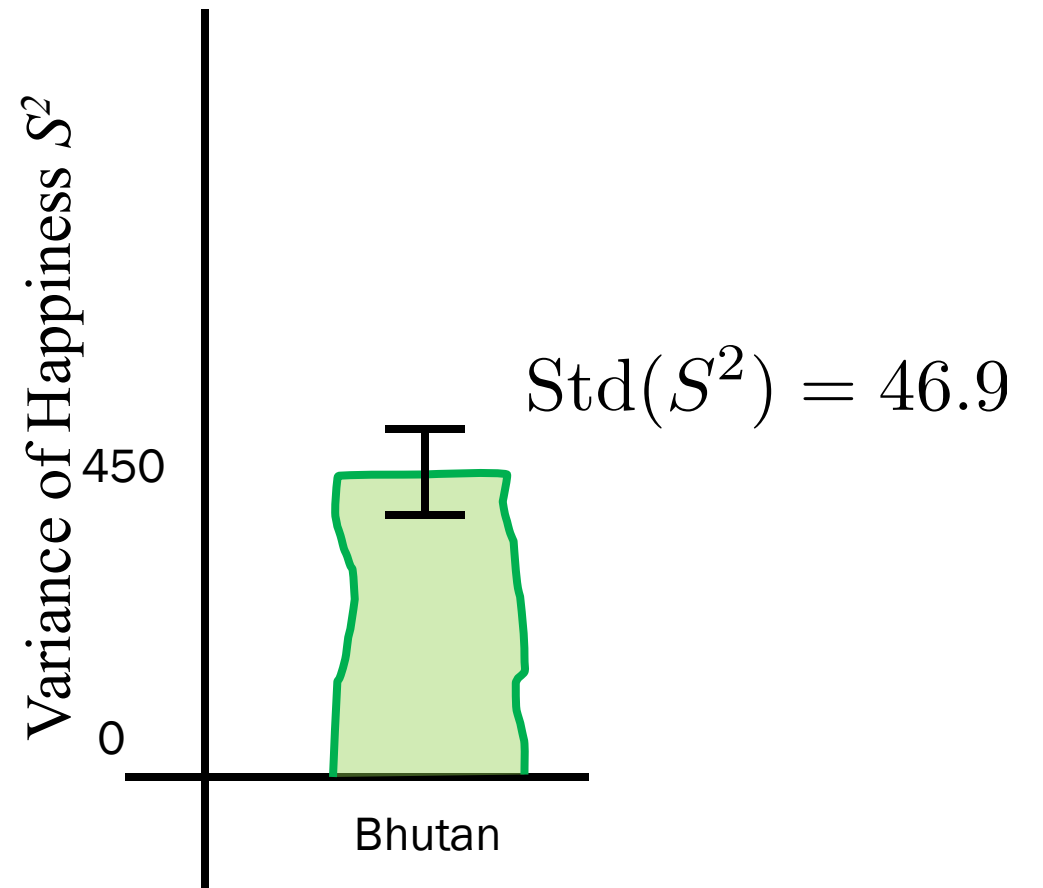
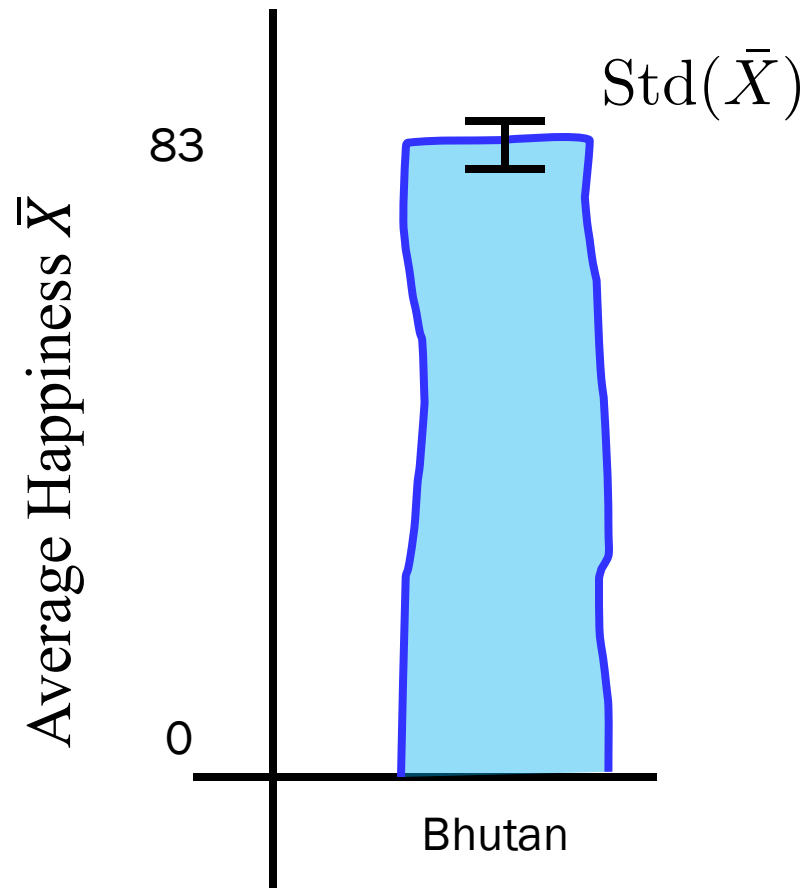


If the underlying distribution is Gaussian, variance is “chi-squared”



Bootstrapping doesn't need to know that...

Our Report to Bhutan Government



Claim: The average happiness of Bhutan is 83 ± 2

Validation with Sample Mean

Bootstrapping of Means (we could do this with CLT)

Bootstrap Algorithm (sample):

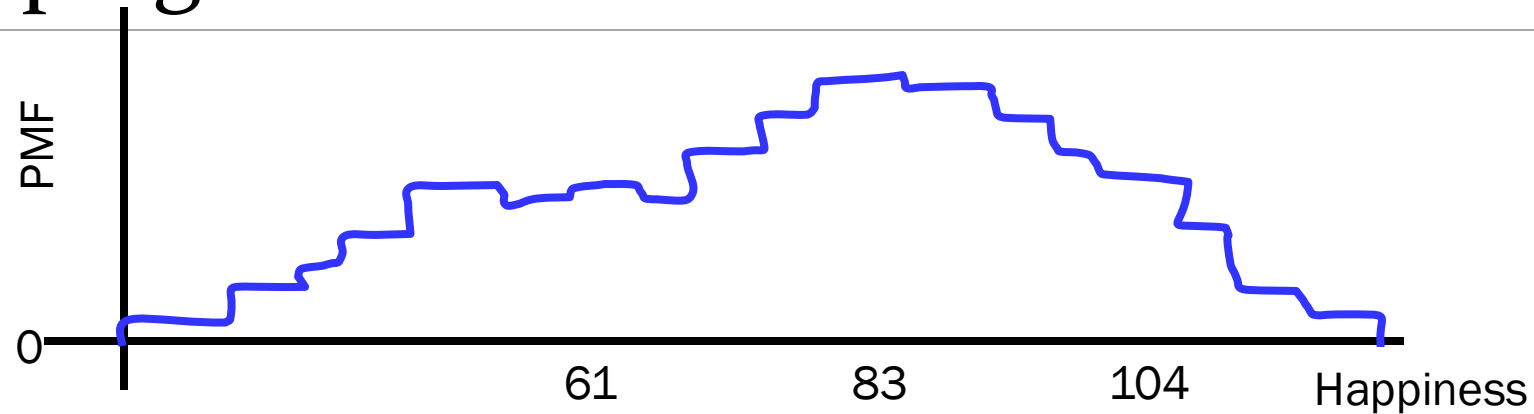
Estimate the **PMF** using the sample

Repeat **10,000** times:

- a. Draw **len(sample)** new samples from PMF
- b. Recalculate the mean** on the resample

You now have a **distribution of your means**

Bootstrapping of Means



Bootstrap Algorithm (sample):

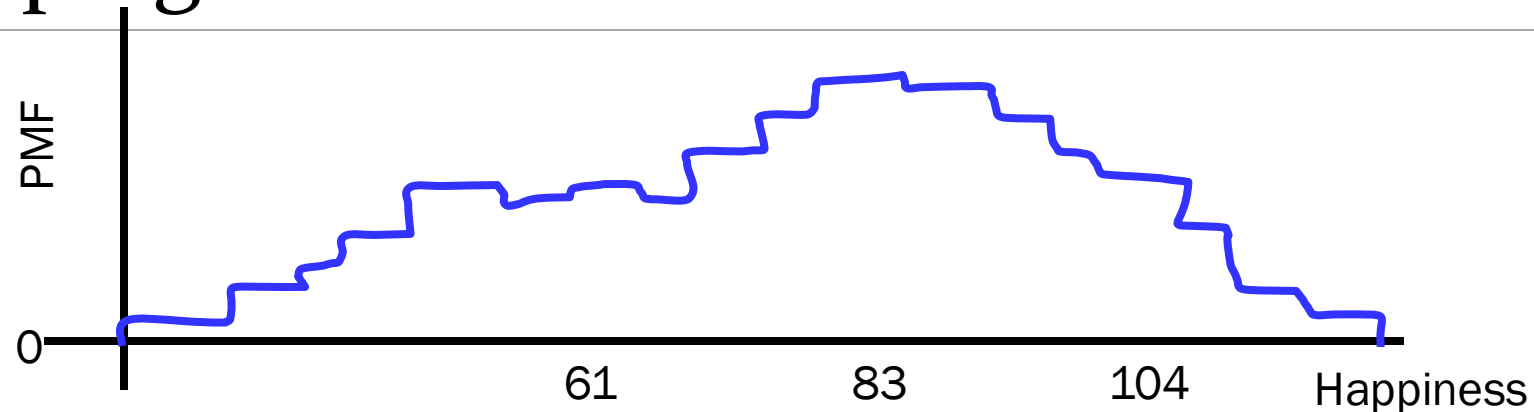
Estimate the **PMF** using the sample

Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the mean** on the resample

You now have a **distribution of your means**

Bootstrapping of Means



Bootstrap Algorithm (sample):

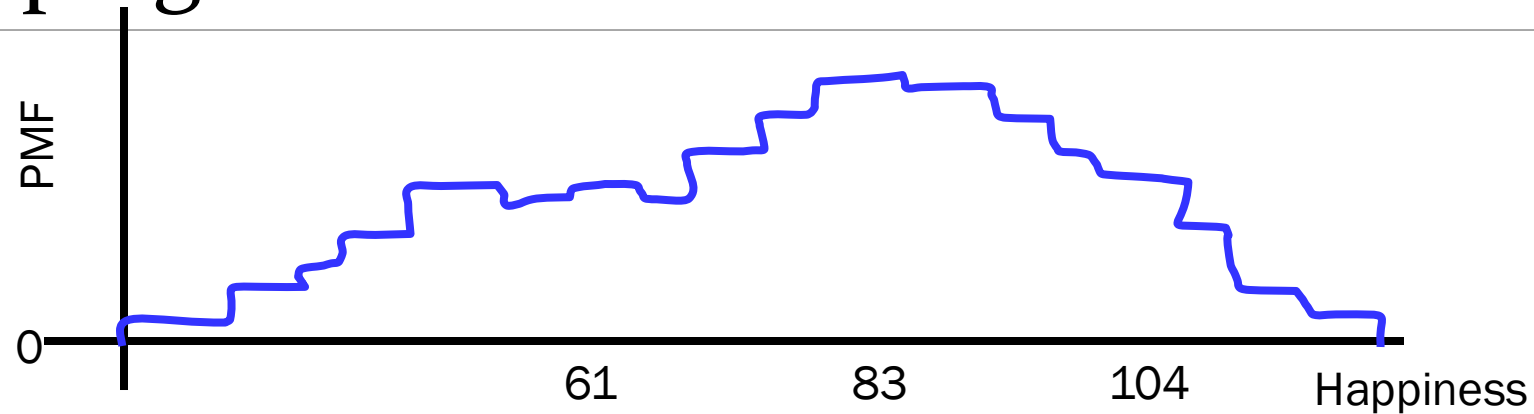
Estimate the **PMF** using the sample

Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the mean** on the resample

You now have a **distribution of your means**

Bootstrapping of Means



Bootstrap Algorithm (sample):

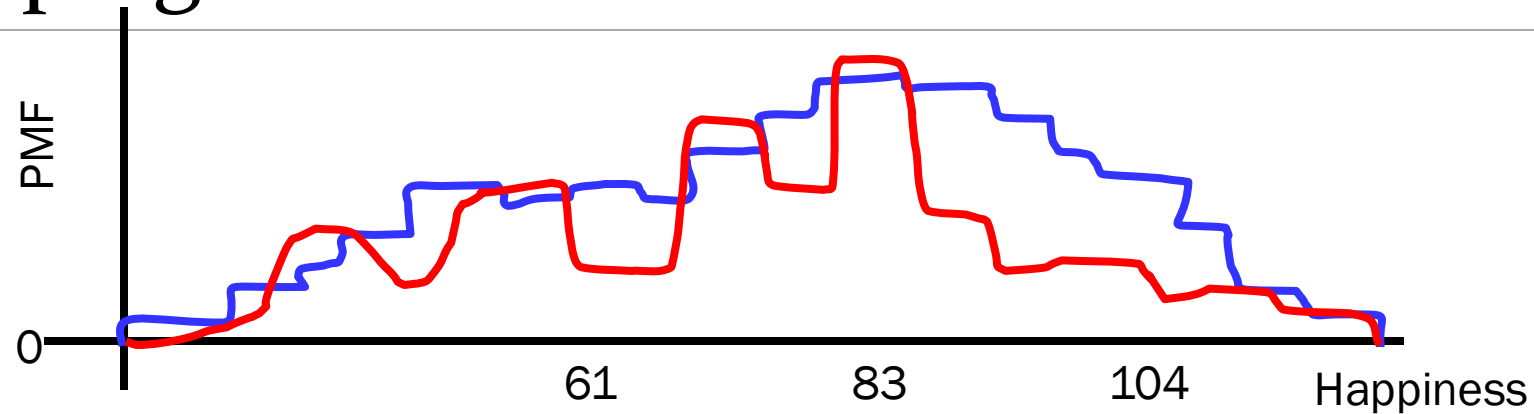
Estimate the **PMF** using the sample

Repeat **10,000** times:

- a. Draw **len(sample)** new samples from PMF
- b. Recalculate the **mean** on the resample

You now have a **distribution of your means**

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

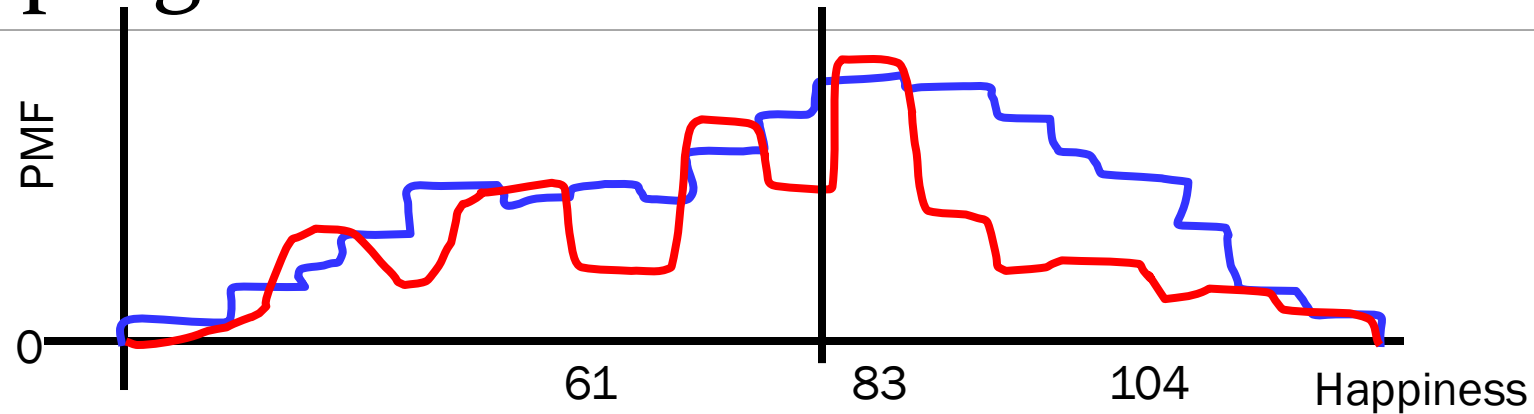
Repeat **10,000** times:

a. Draw **len(sample)** new samples from PMF

b. Recalculate the mean on the resample

You now have a **distribution of your means**

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

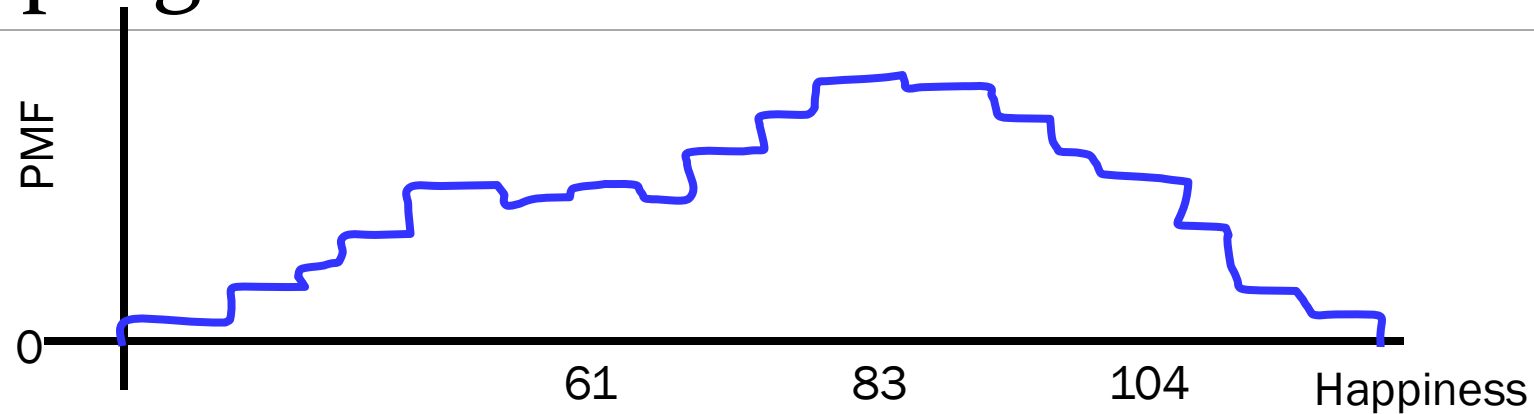
a. Draw **len(sample)** new samples from PMF

b. Recalculate the mean on the resample

You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

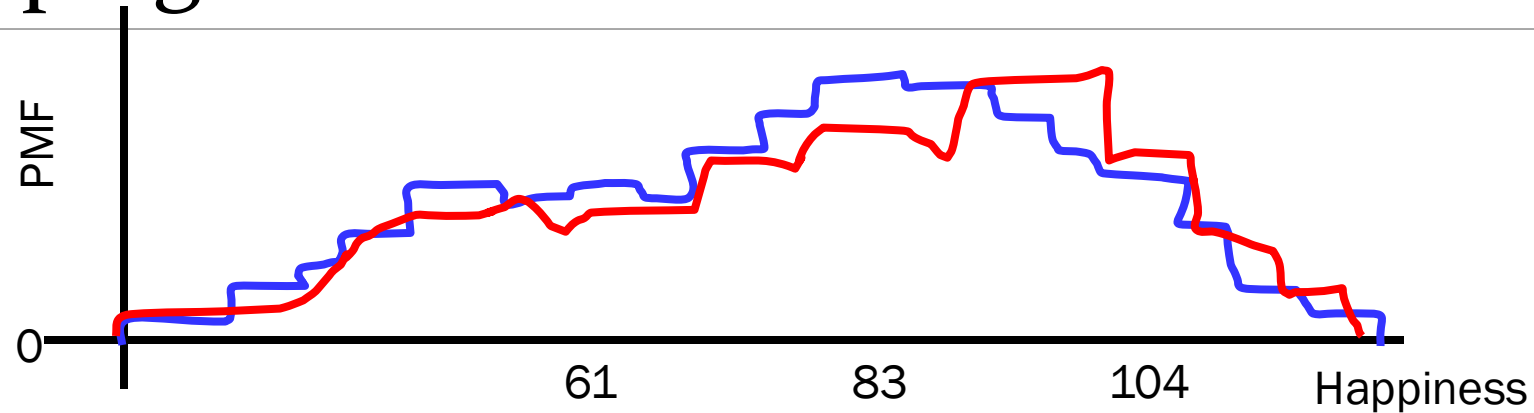
Repeat **10,000** times:

- a. Draw `len(sample)` new samples from PMF
- b. Recalculate the **mean** on the resample

You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

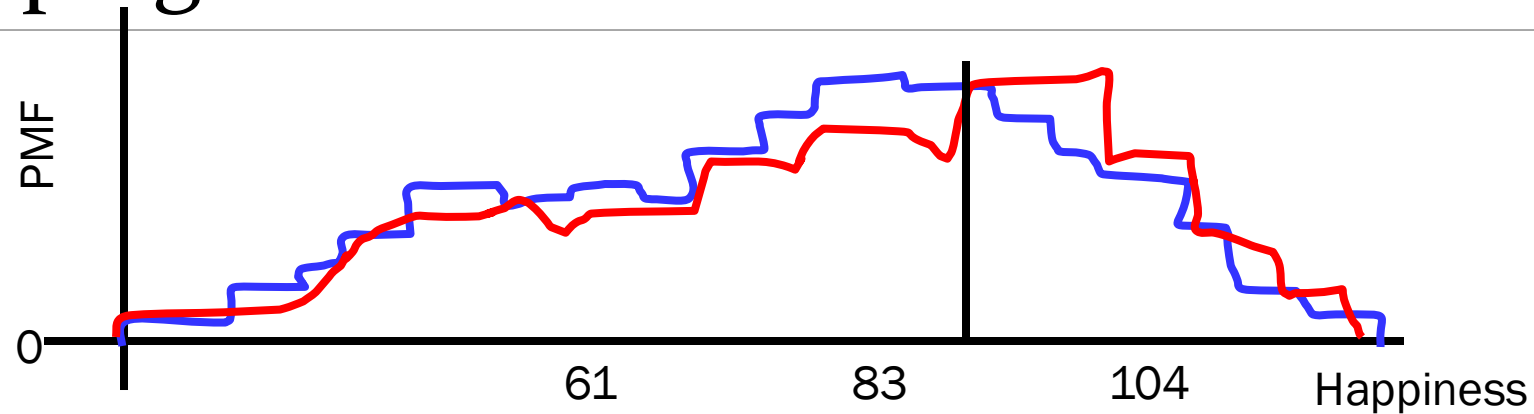
a. Draw **len(sample)** new samples from PMF

b. Recalculate the mean on the resample

You now have a **distribution of your means**

Means = [82.7]

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

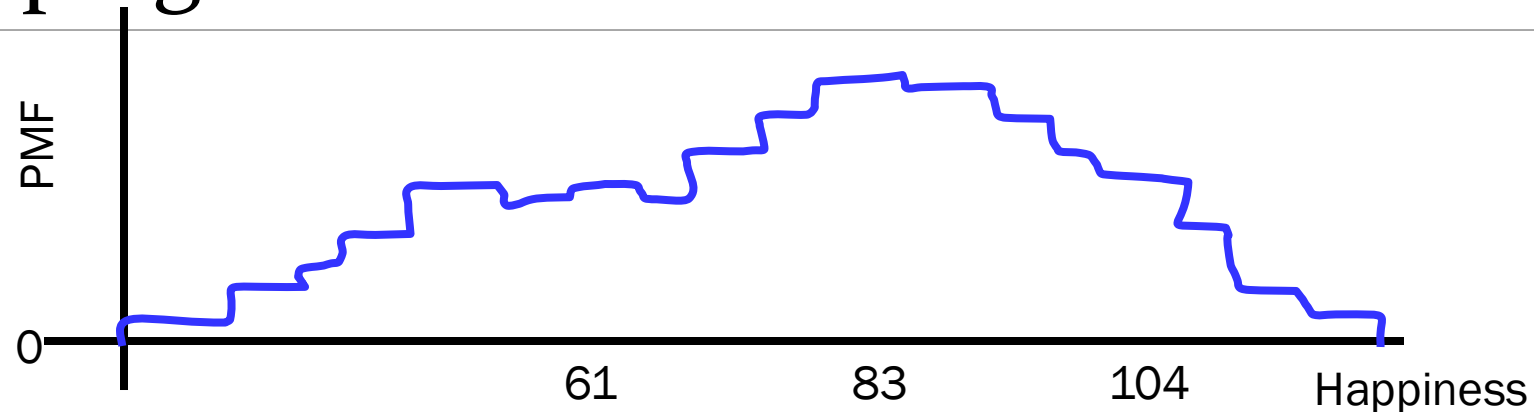
a. Draw **len(sample)** new samples from PMF

b. Recalculate the mean on the resample

You now have a **distribution of your means**

Means = [82.7, 83.4]

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

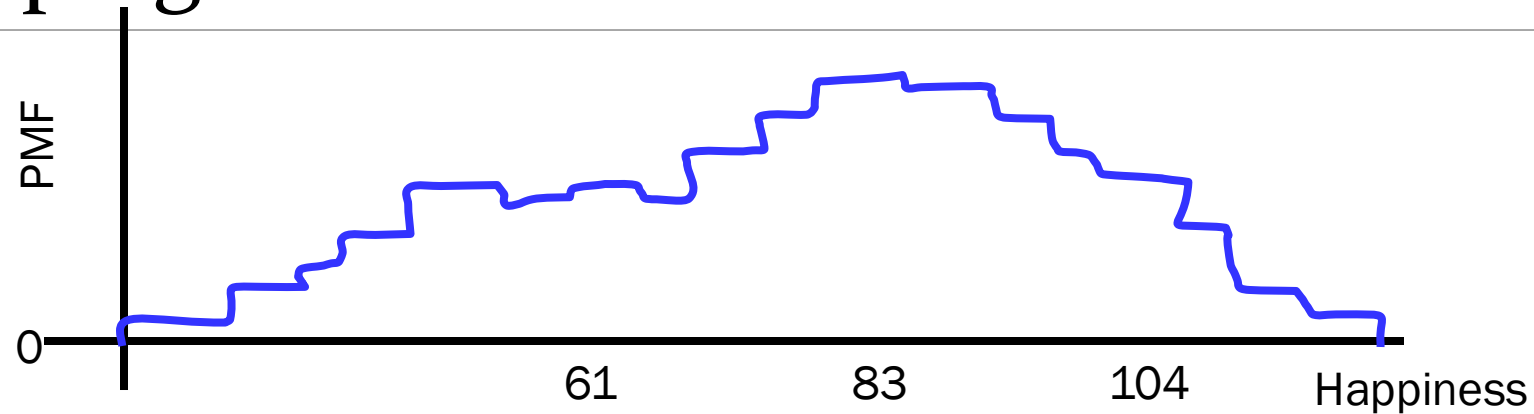
Repeat **10,000** times:

- Draw **len(sample)** new samples from PMF
- Recalculate the mean** on the resample

You now have a **distribution of your means**

Means = [82.7, 83.4]

Bootstrapping of Means



Bootstrap Algorithm (sample):

Estimate the **PMF** using the sample

Repeat **10,000** times:

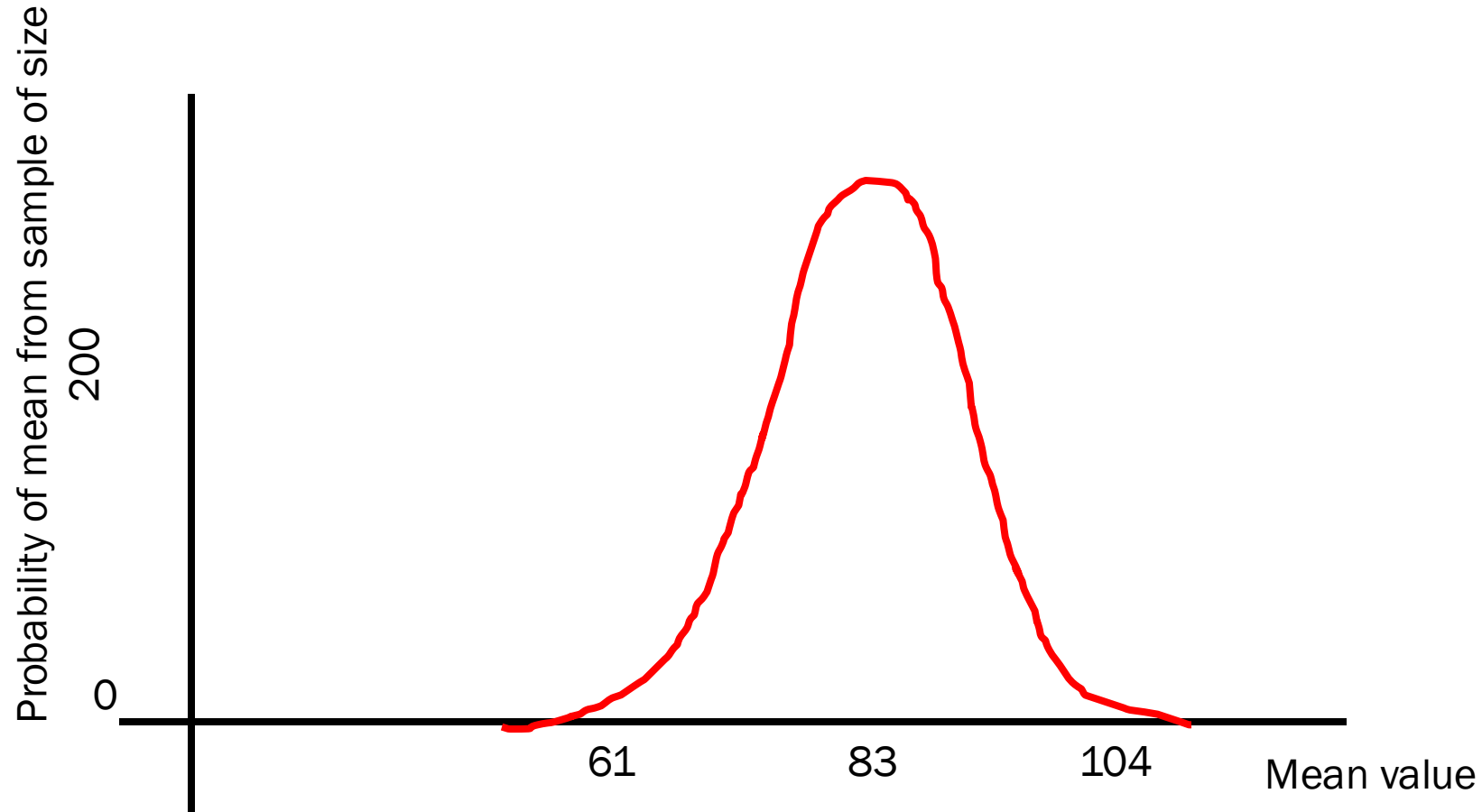
- a. Draw **len(sample)** new samples from PMF
- b. **Recalculate the mean** on the resample

You now have a **distribution of your means**

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

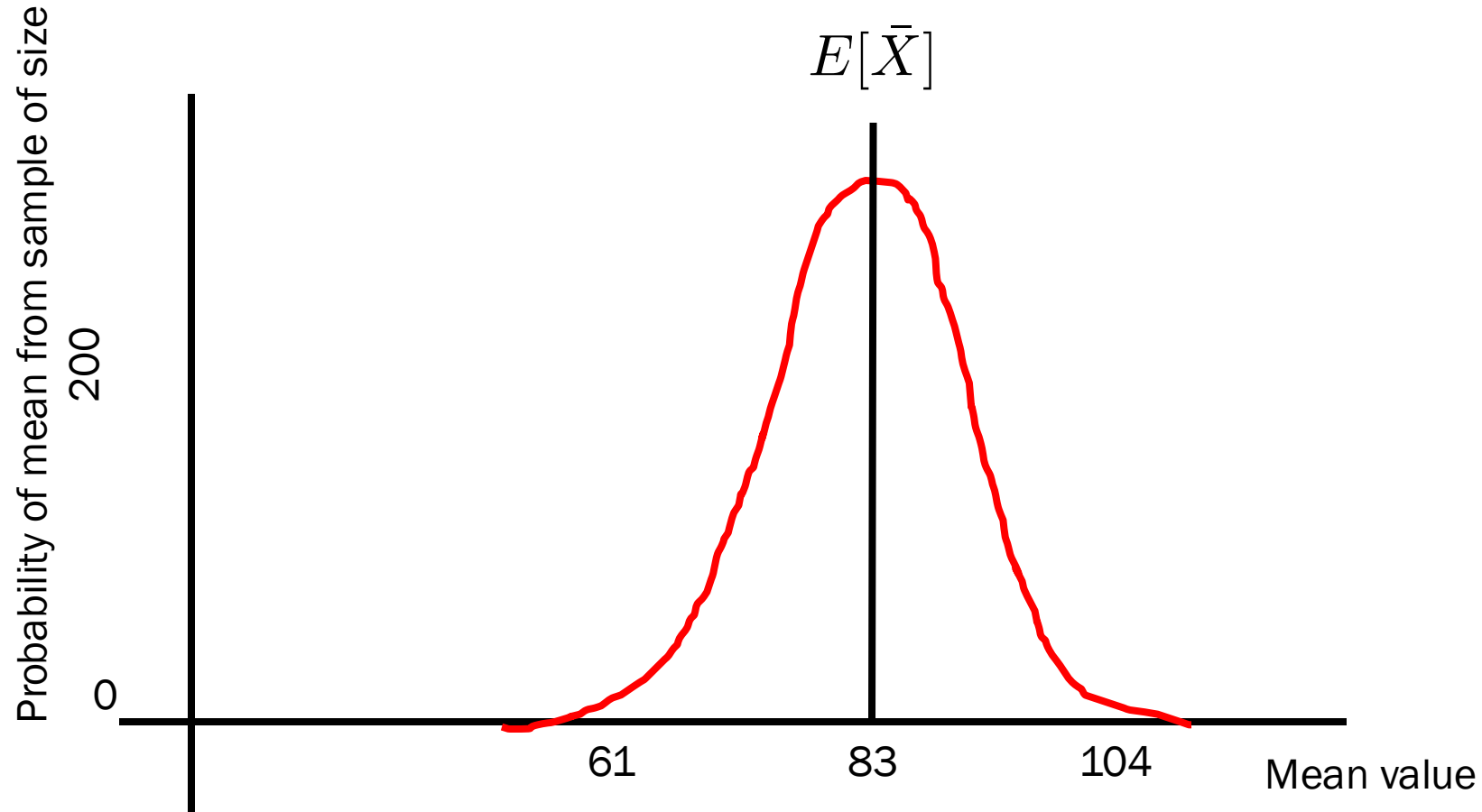
Bootstrapping of Means

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



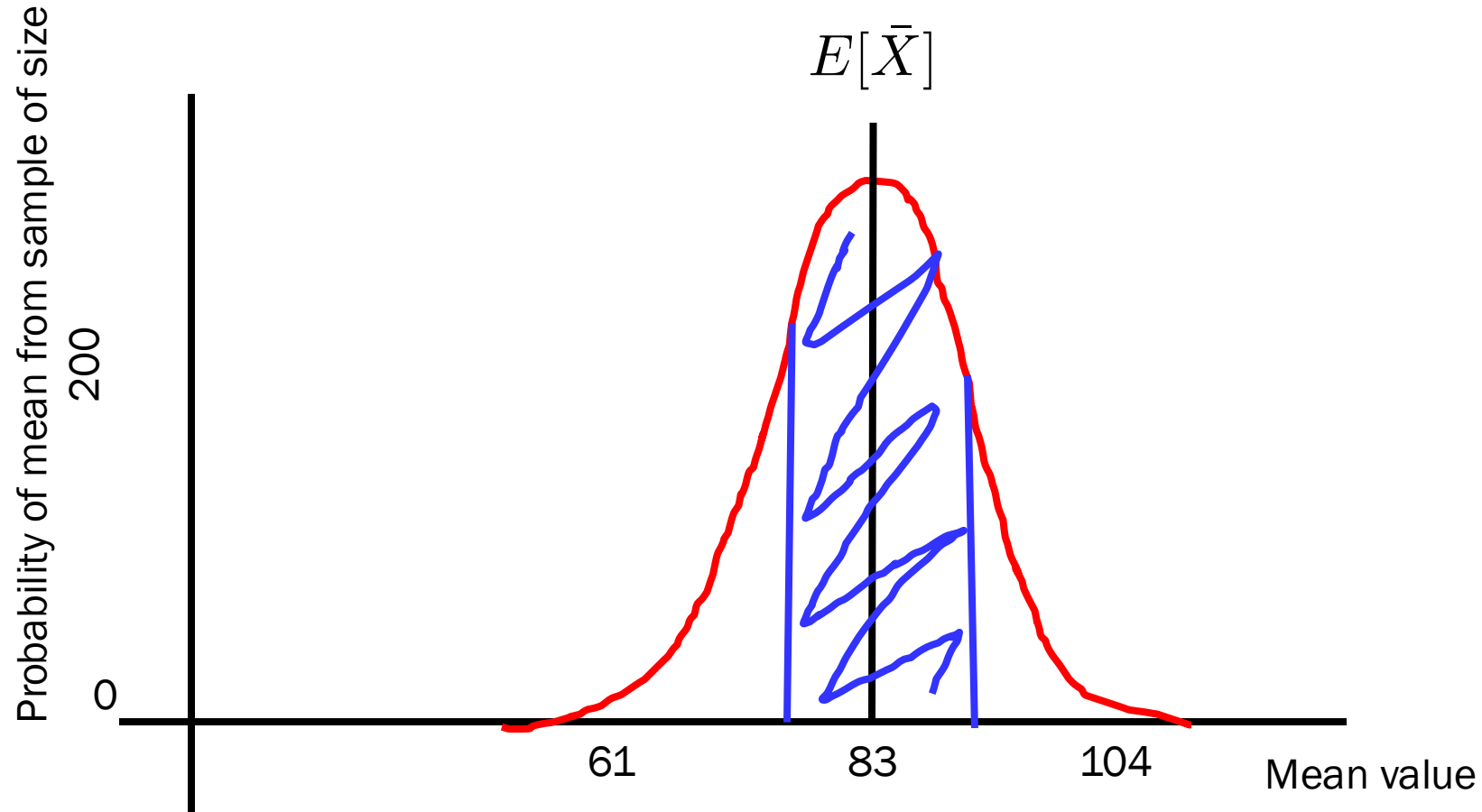
Bootstrapping of Means

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?

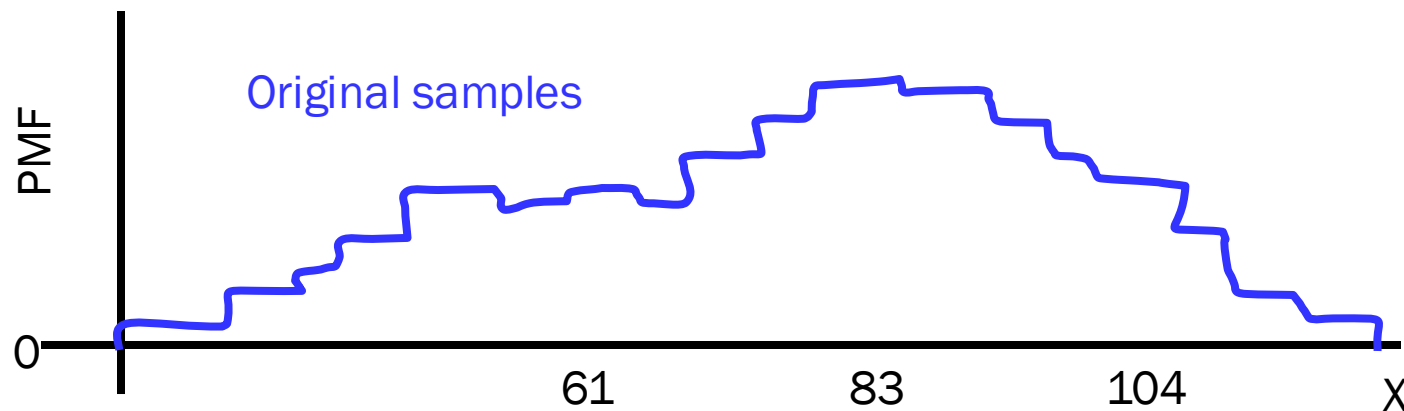


Contrast with Central Limit Theorem

Ok Good!

Bootstrapping in Practice

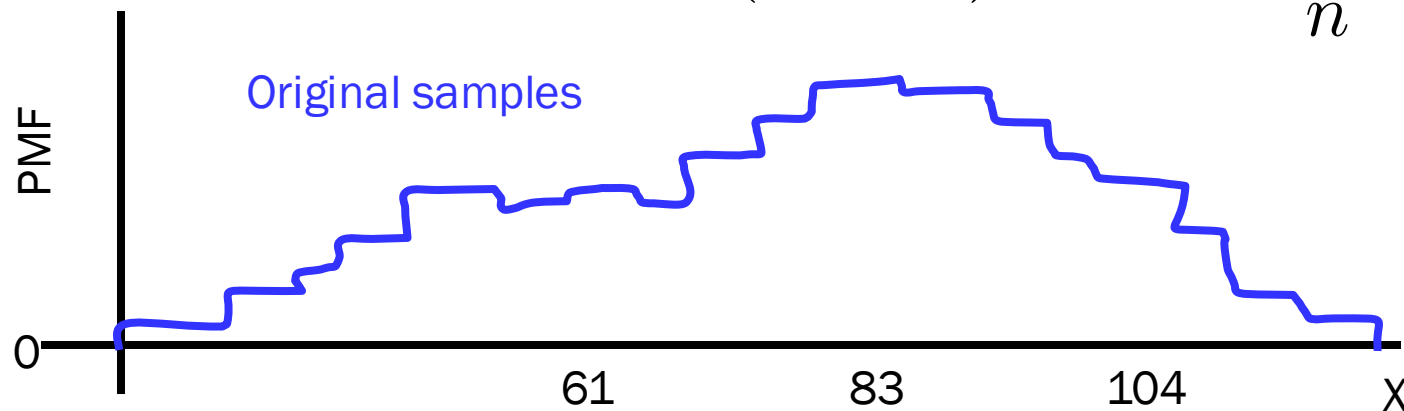
```
def resample(samples, K):  
    # Estimate the PMF using the samples  
    # Draw K new samples from the PMF
```



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)
```

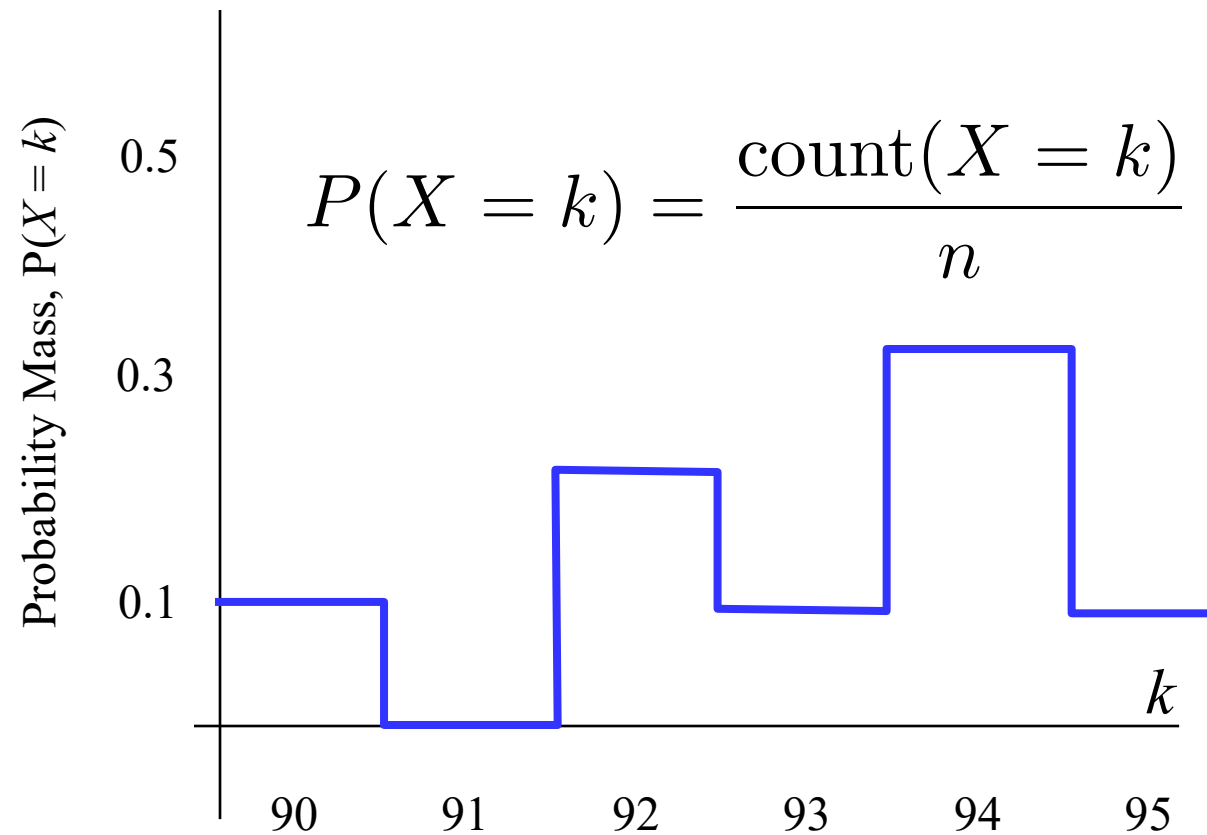
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$



`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

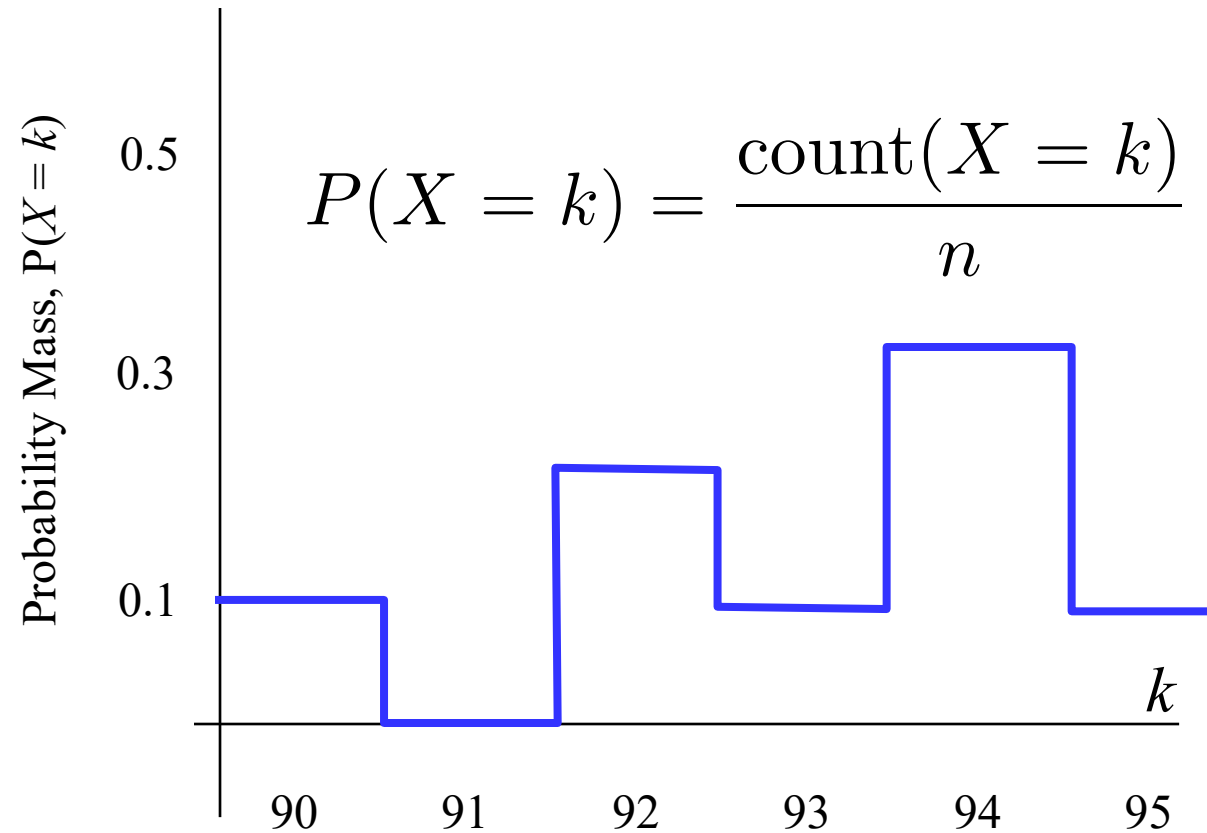
Resample:



`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

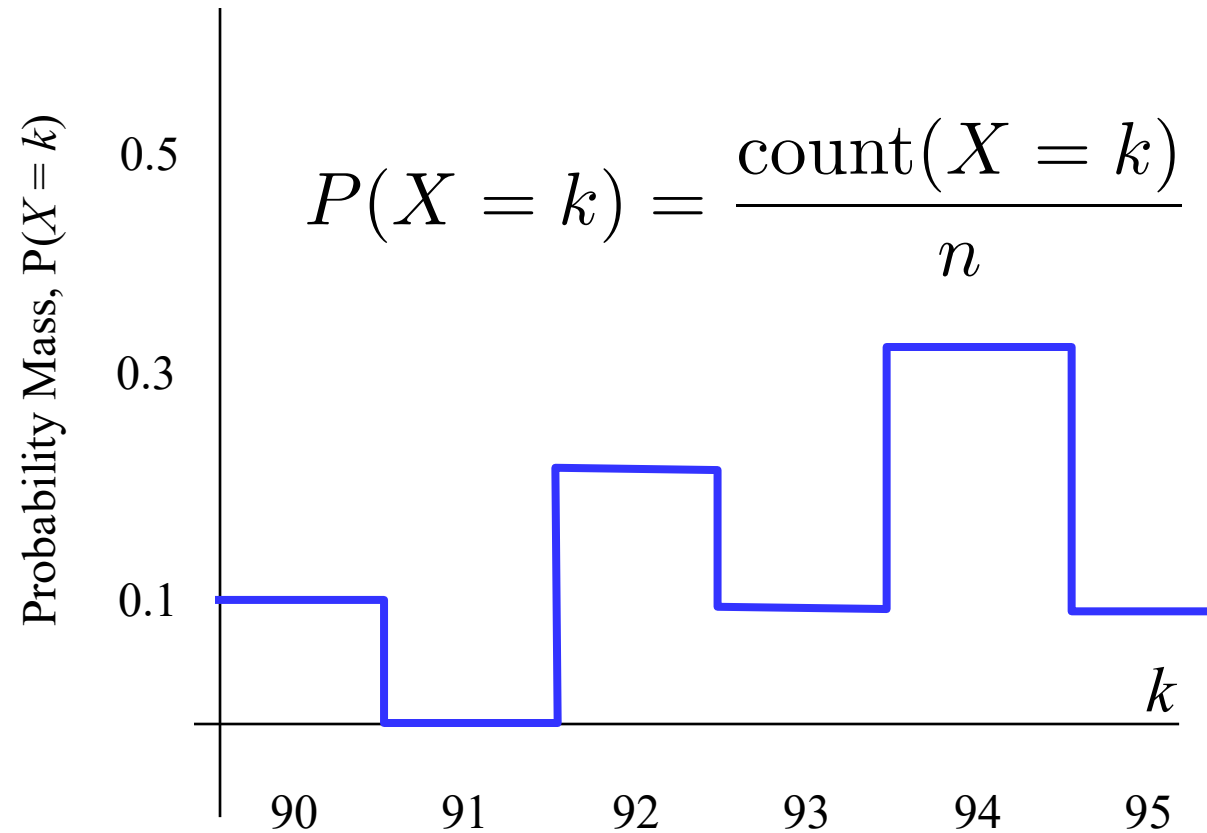


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, **94**, 94, 94, 95]

Resample:

[94]

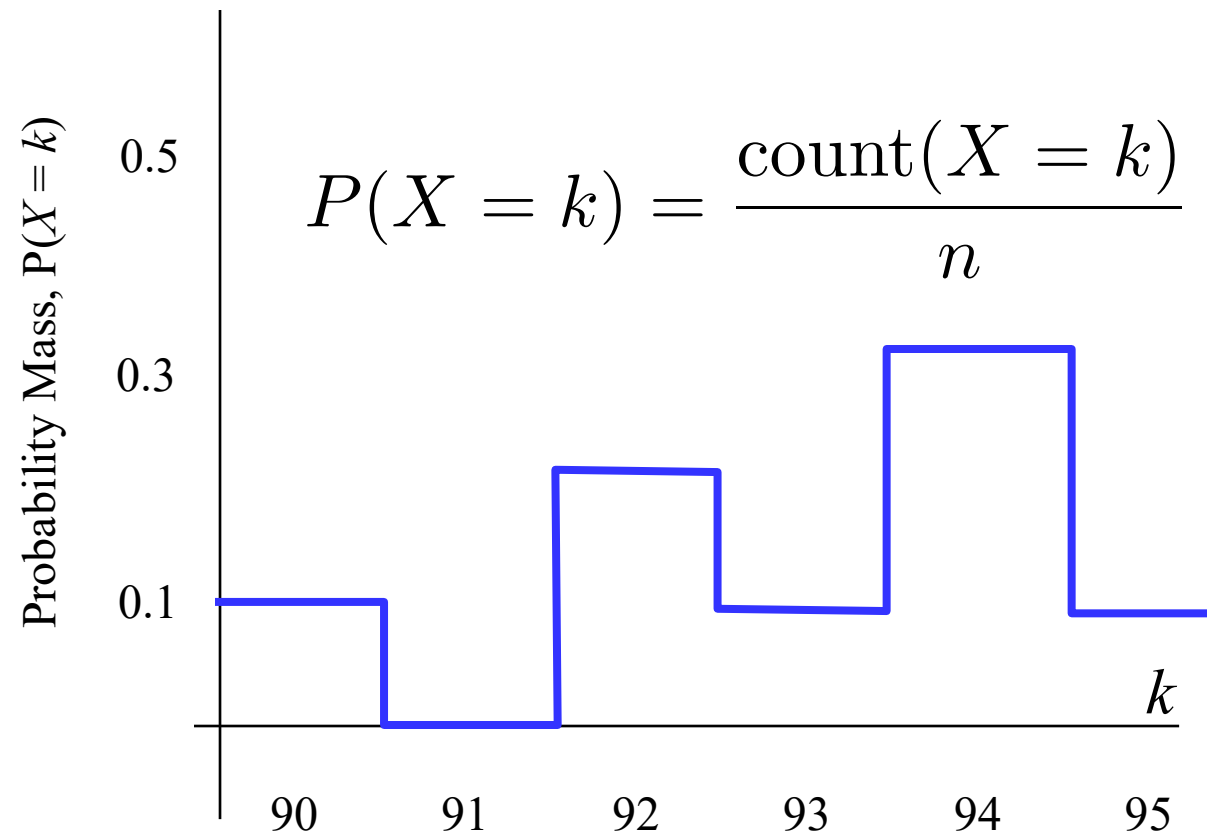


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94]

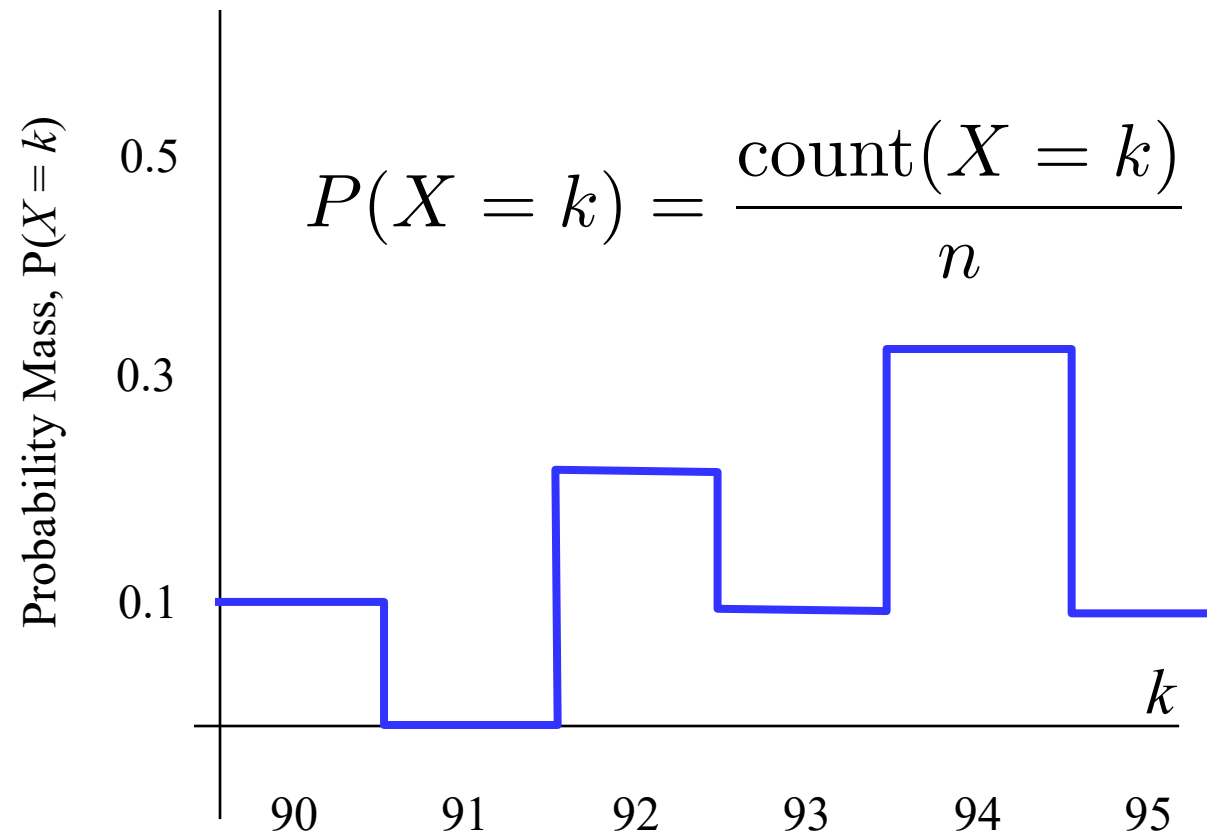


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94]

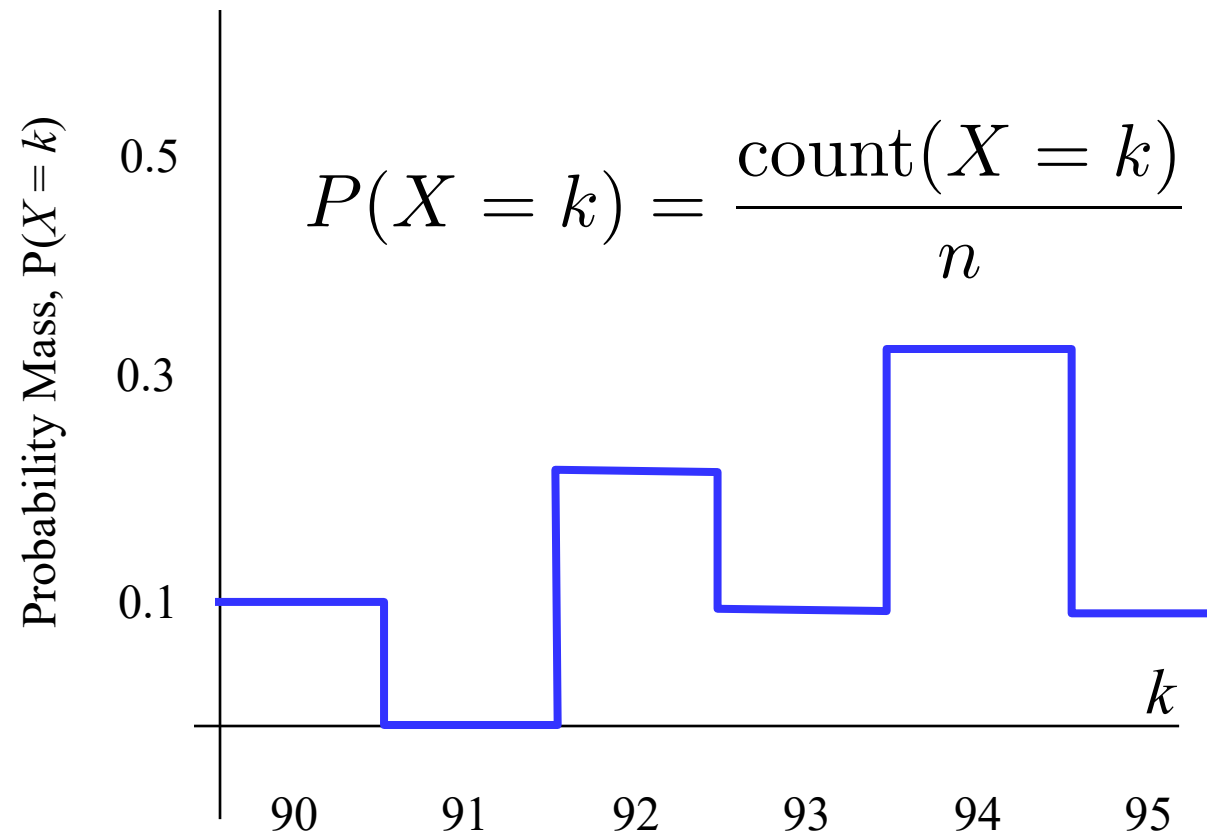


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94, 90]

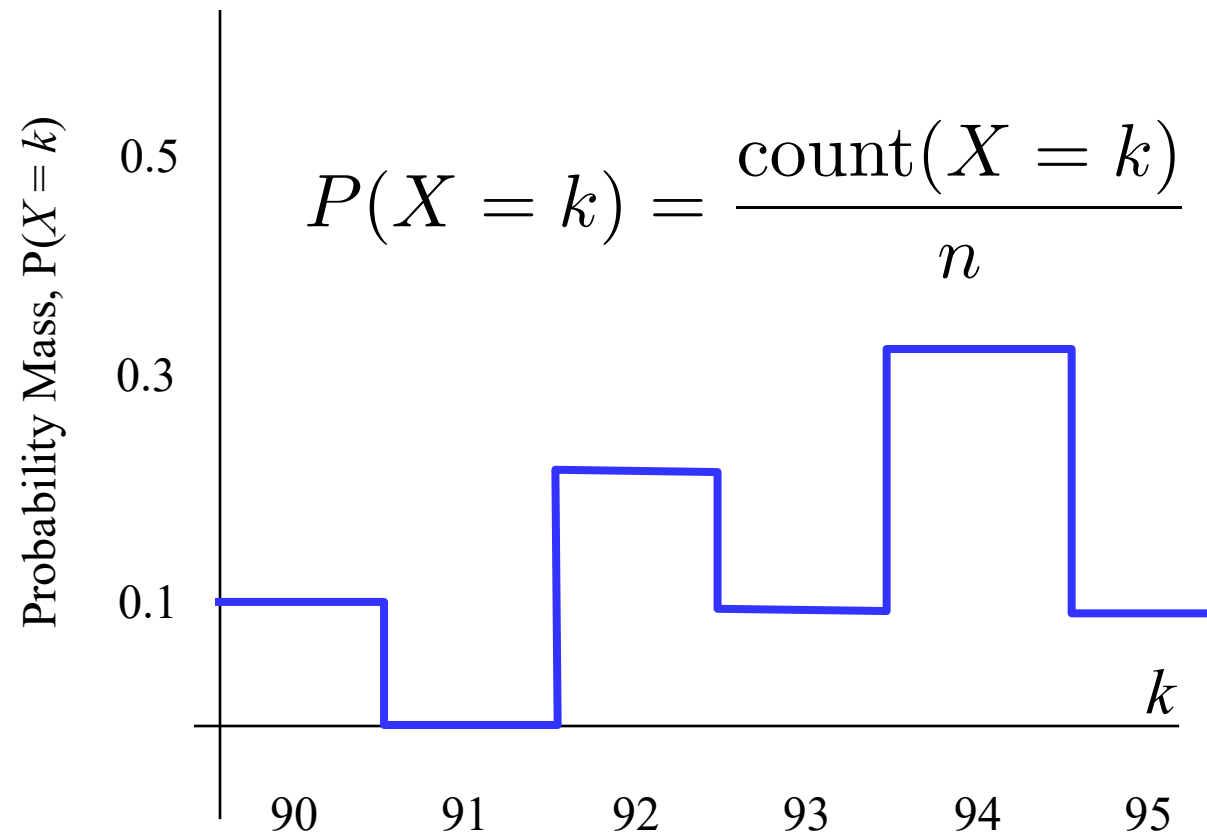


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94, 90]

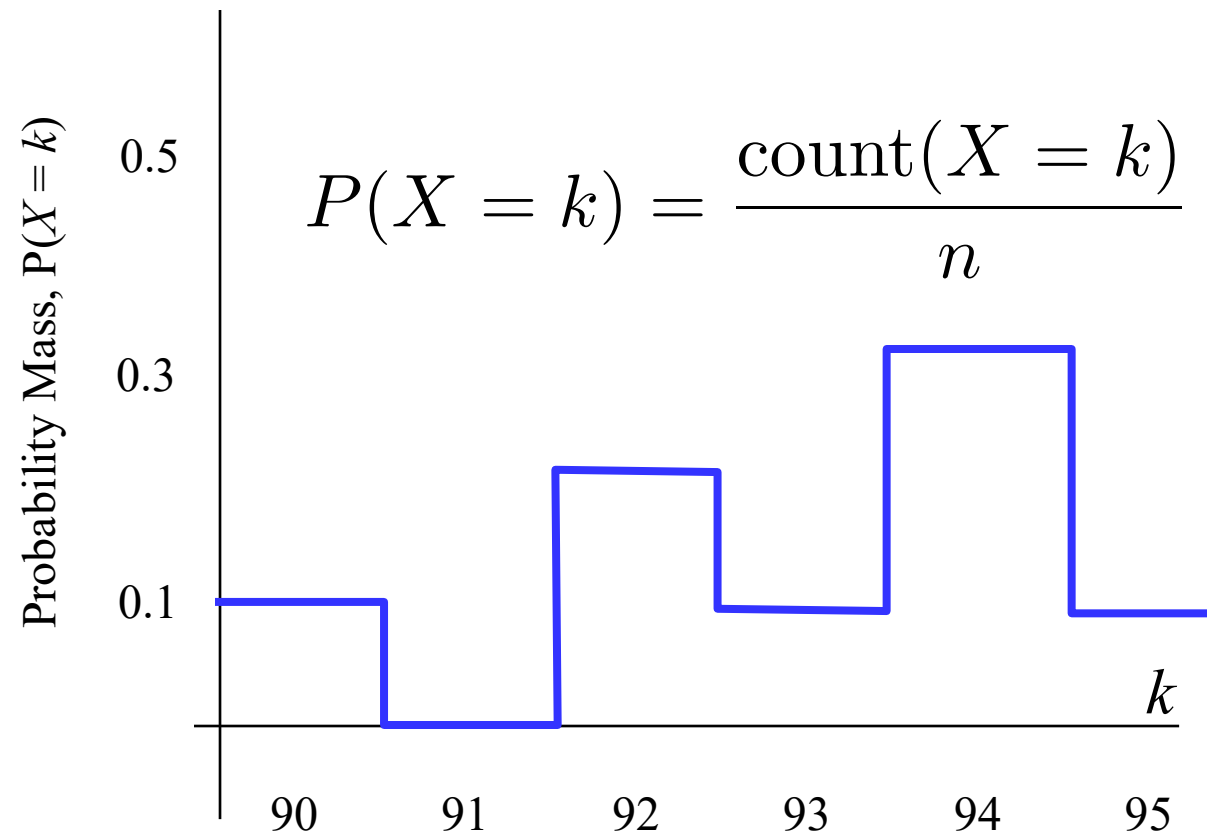


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94, 90]

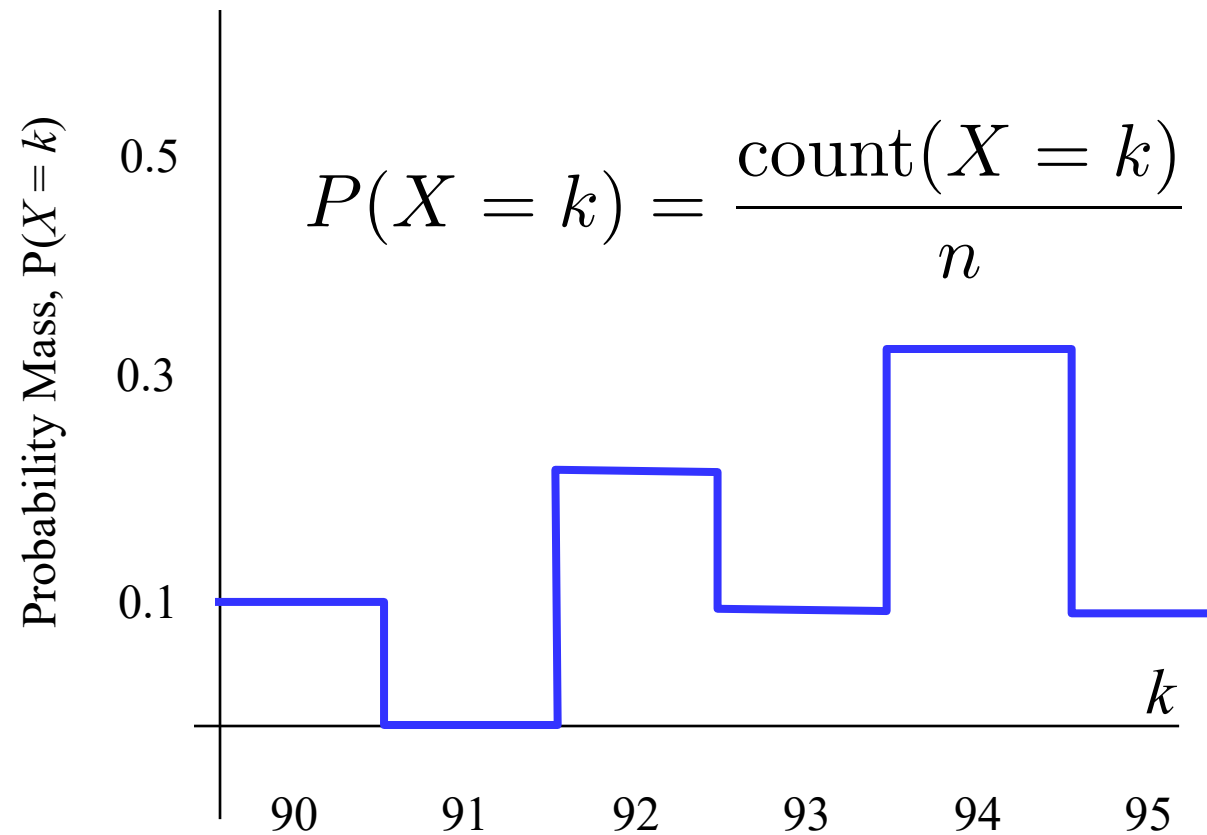


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94, 90, 90]

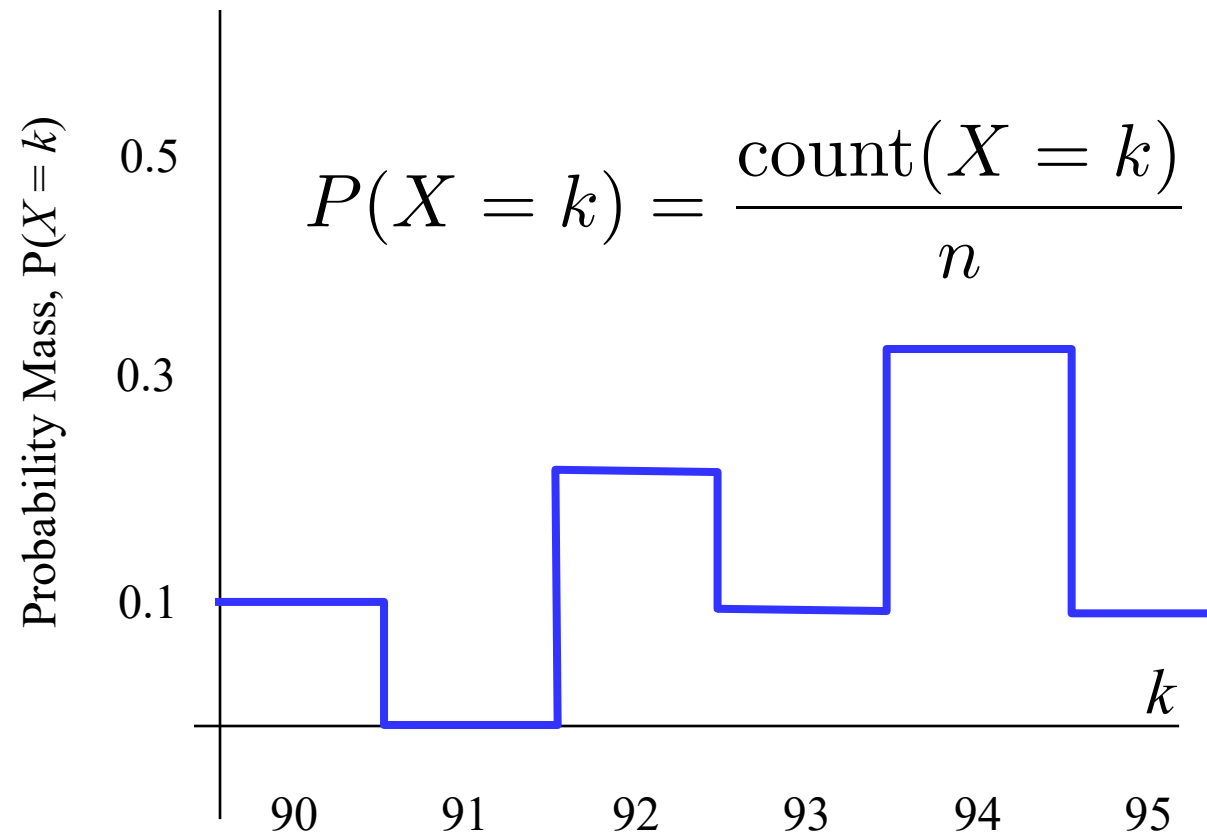


`np.random.choice(samples, K, replace = True)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

[94, 90, 90]

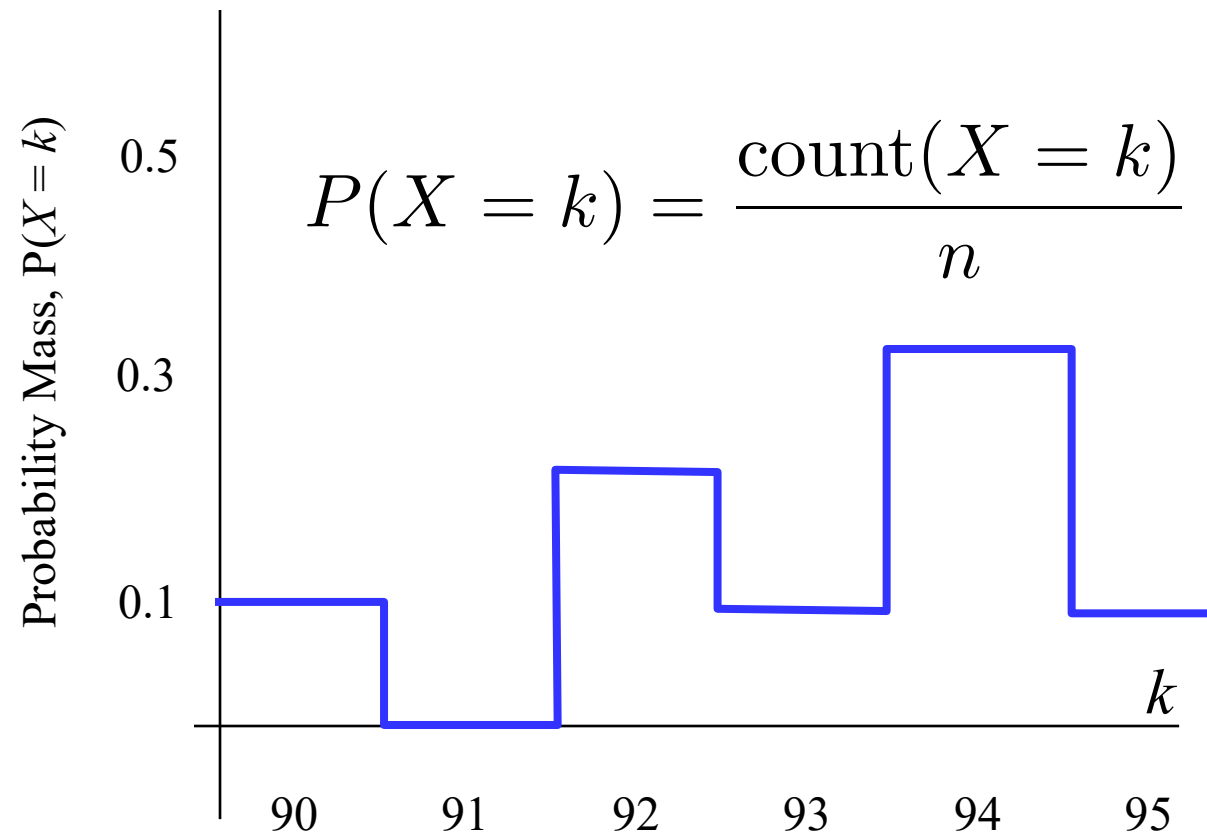


Now with `replace = False`

`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

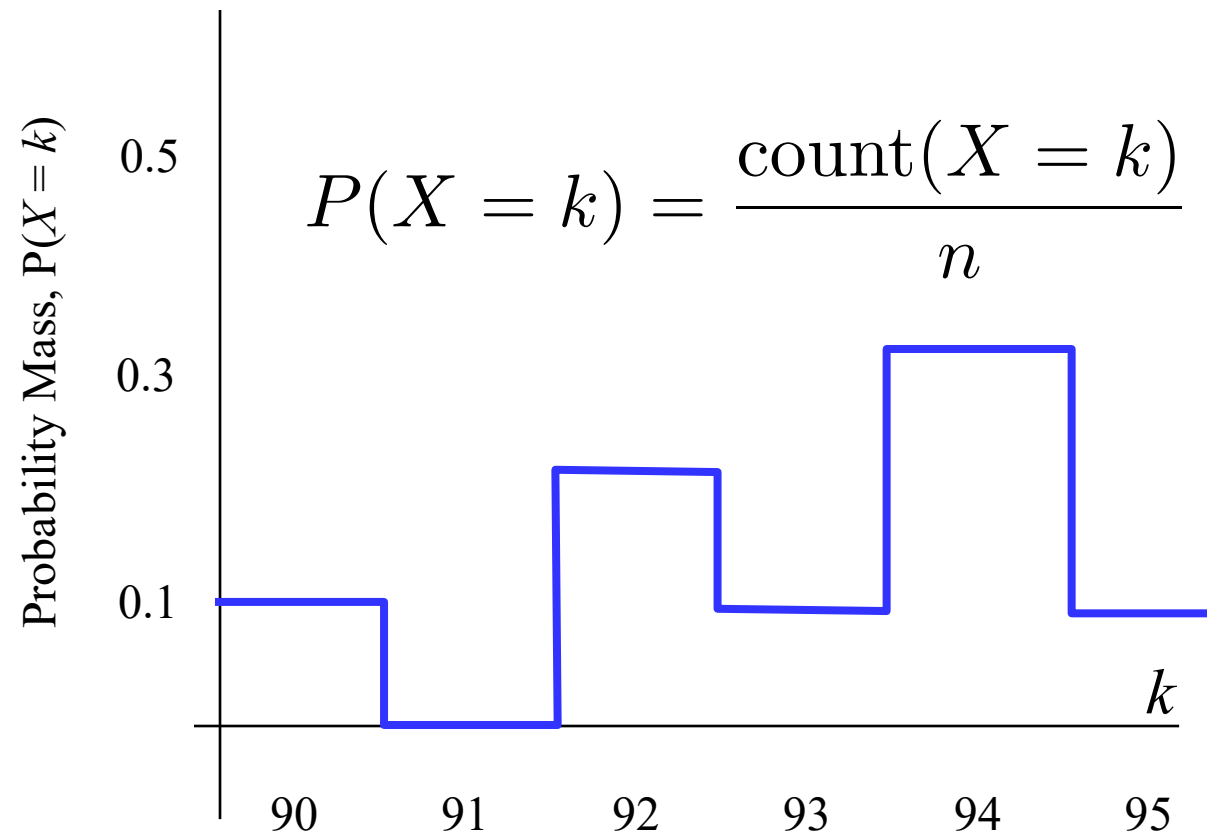
Resample:



`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 94, 95]

Resample:

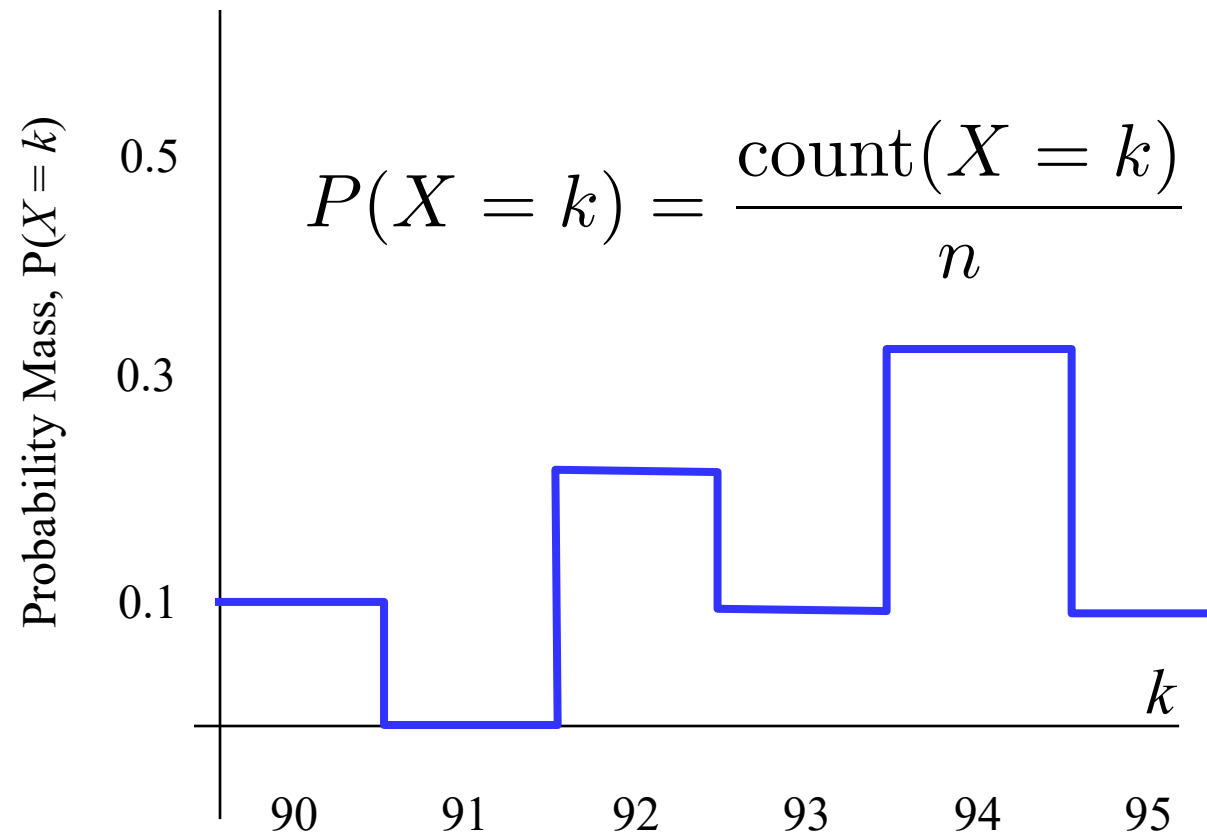


`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, **94**, 94, 94, 95]

Resample:

[94]



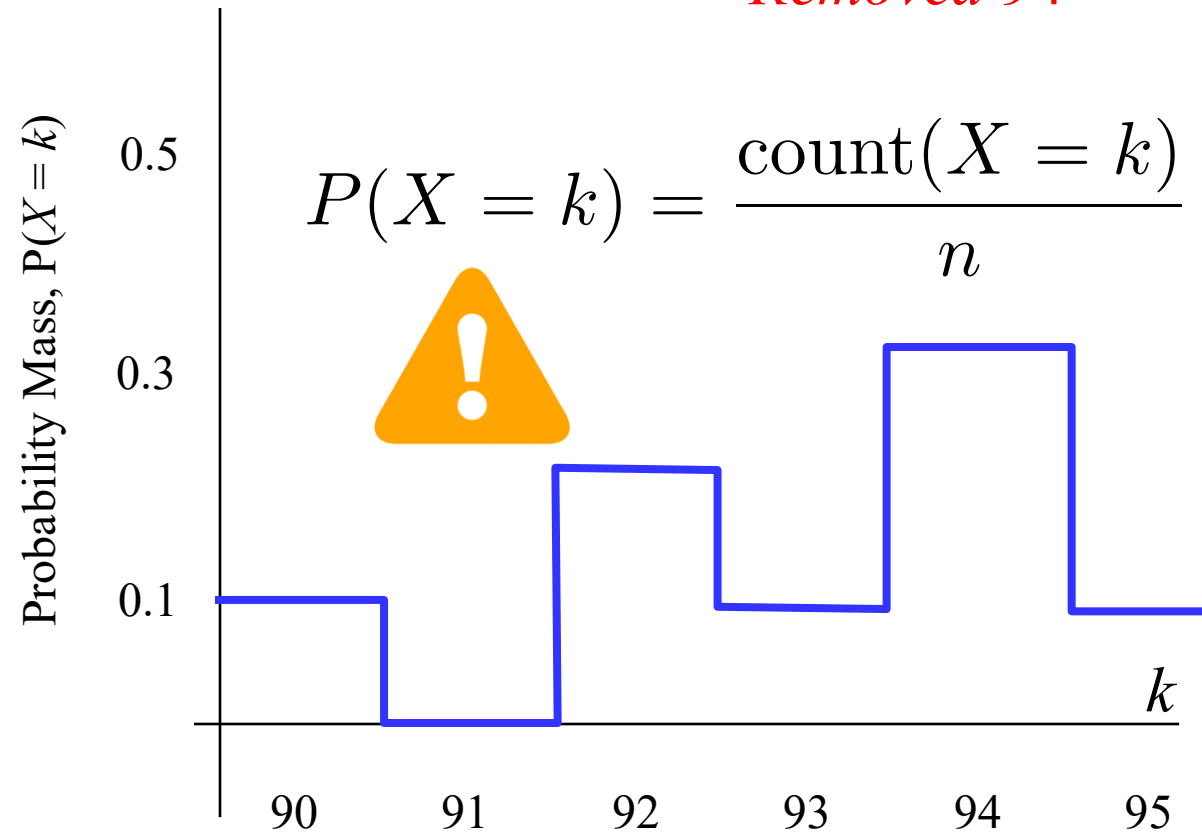
`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 95]

Resample:

[94]

Removed 94

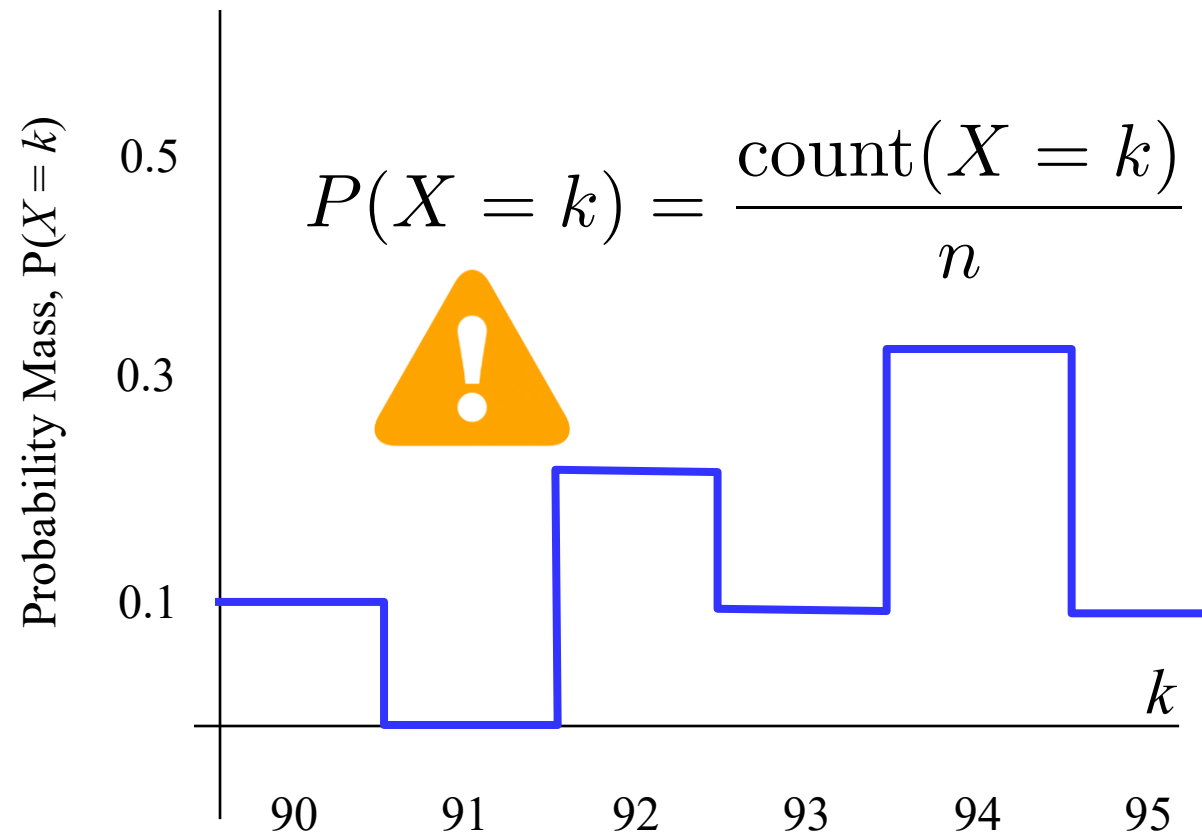


`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 95]

Resample:

[94]

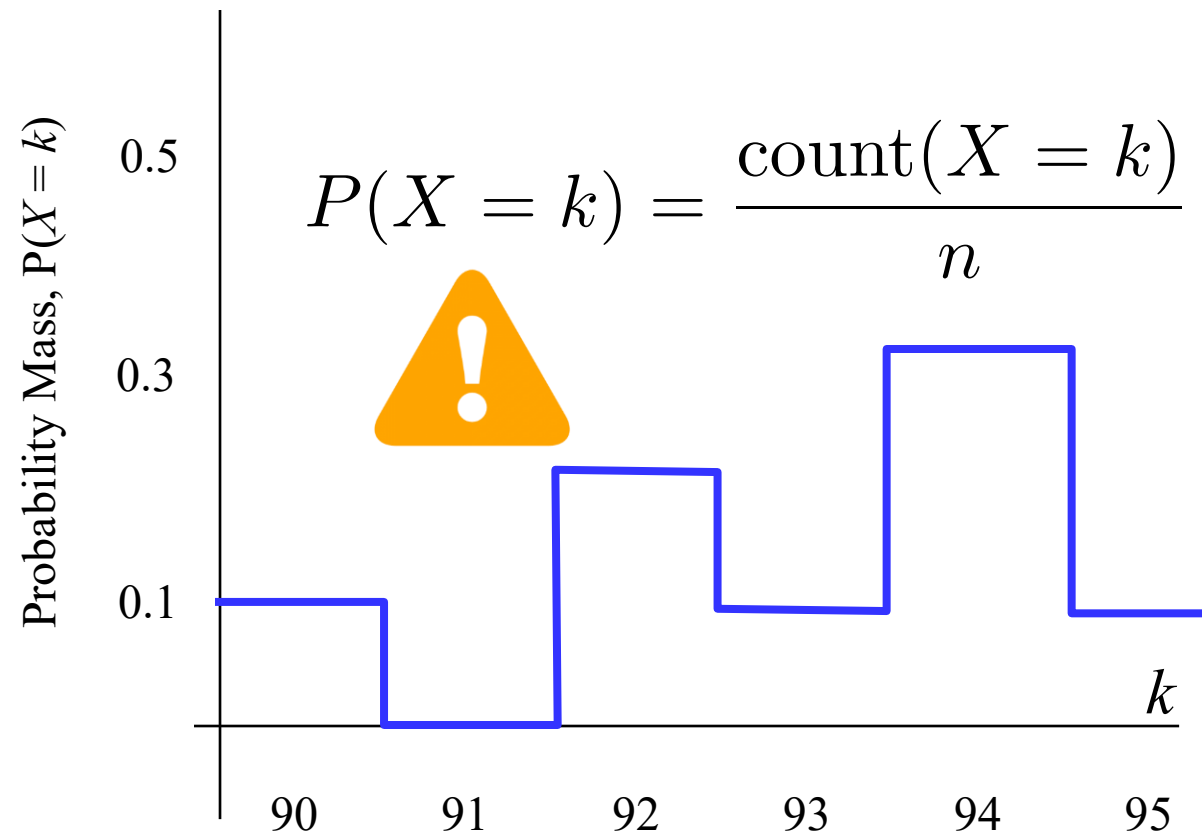


`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 95]

Resample:

[94]

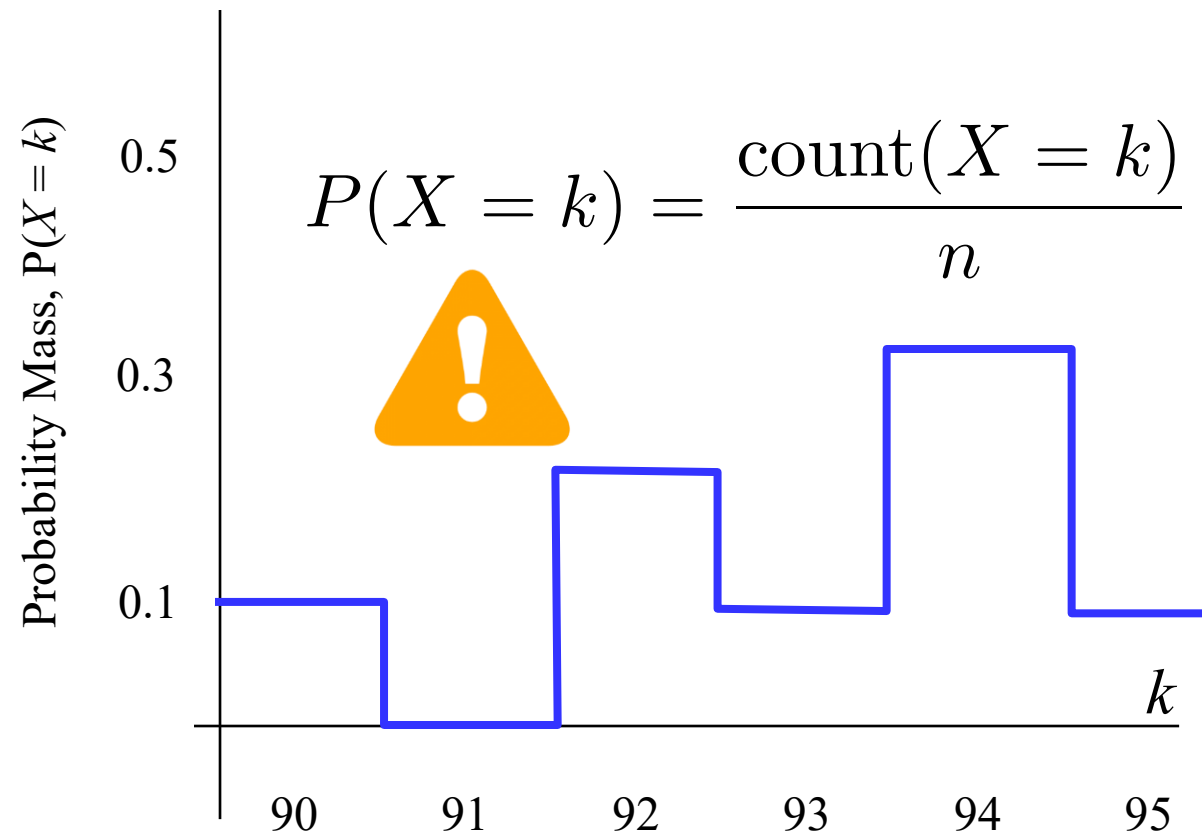


`np.random.choice(samples, K, replace = False)`

Original Samples: [90, 92, 92, 93, 94, 94, 95]

Resample:

[94, 90]



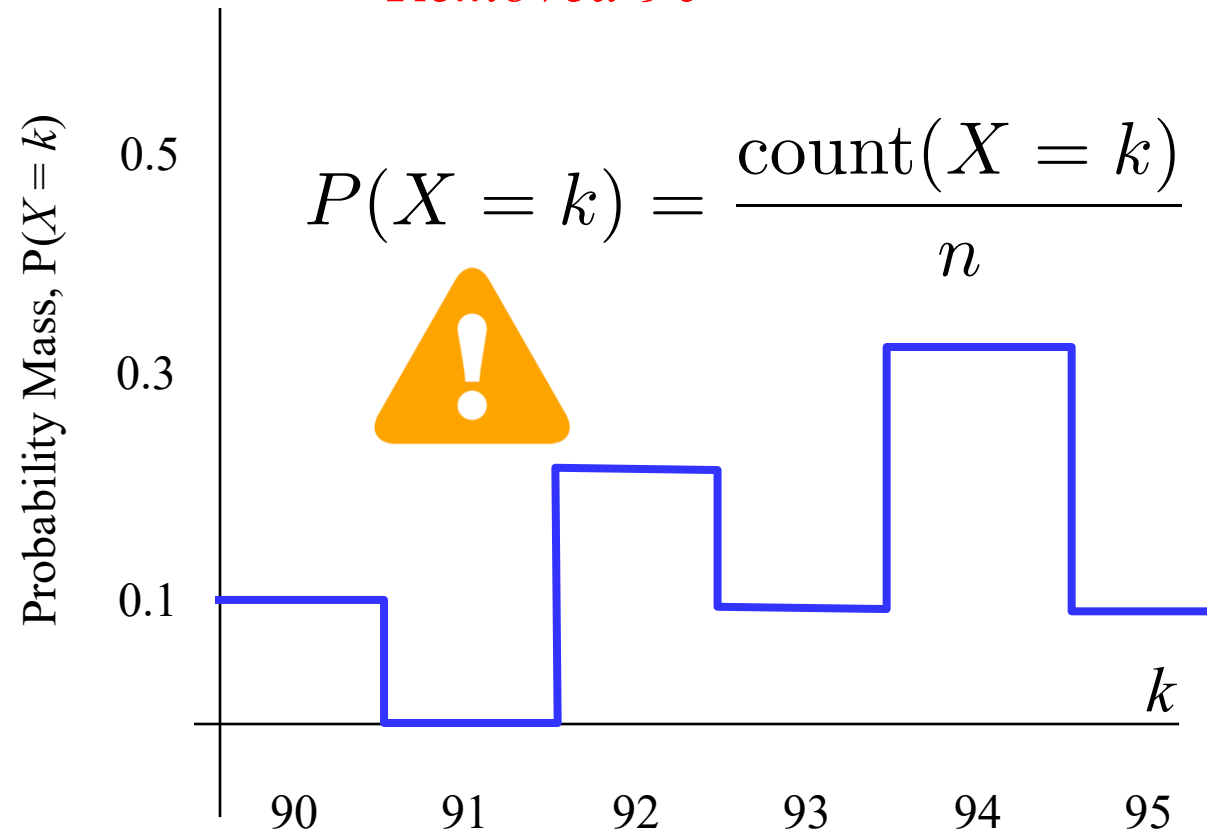
`np.random.choice(samples, K, replace = False)`

Original Samples: [92, 92, 93, 94, 94, 95]

Resample:

[94, 90]

Removed 90



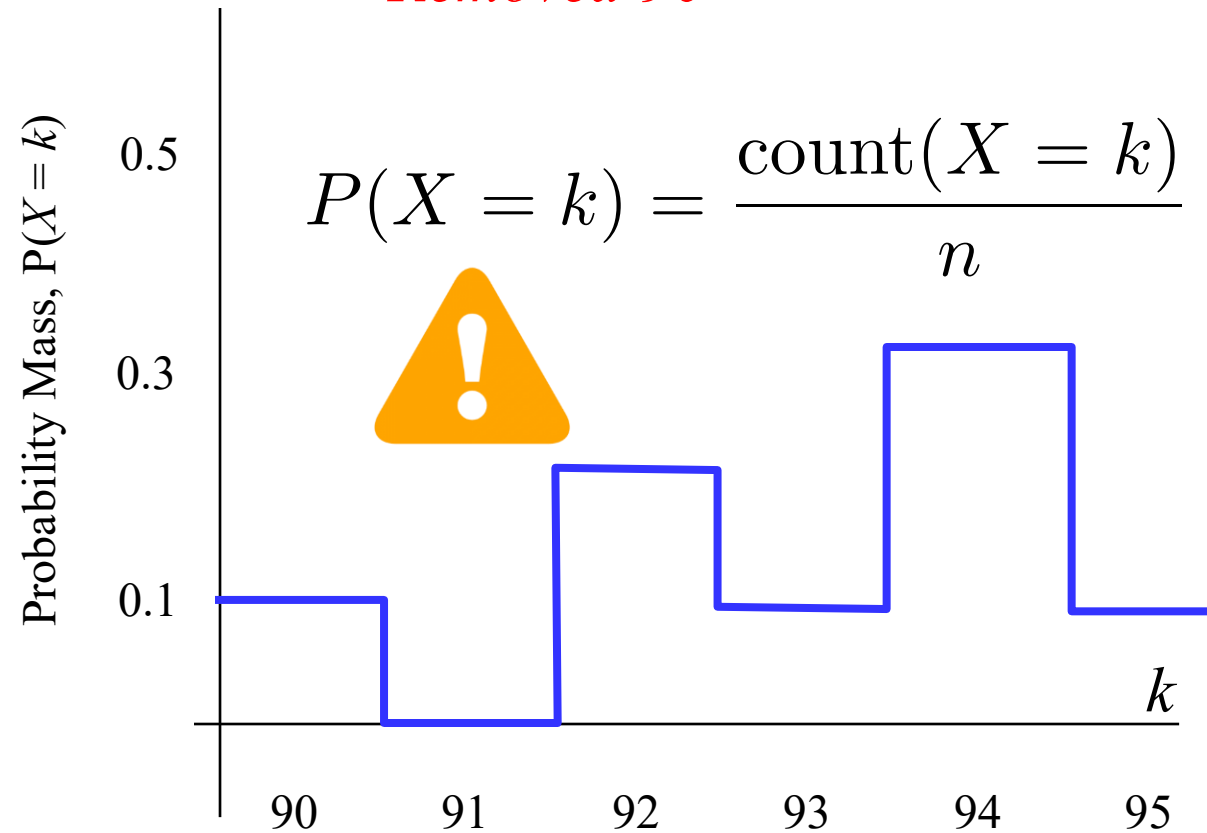
`np.random.choice(samples, K, replace = False)`

Original Samples: [92, 92, 93, 94, 94, 95]

Resample:

[94, 90]

Removed 90



The probability of sampling a 90
is no longer 0.1

The probability of sampling 94 is
no longer 0.3

OG Bootstrapping

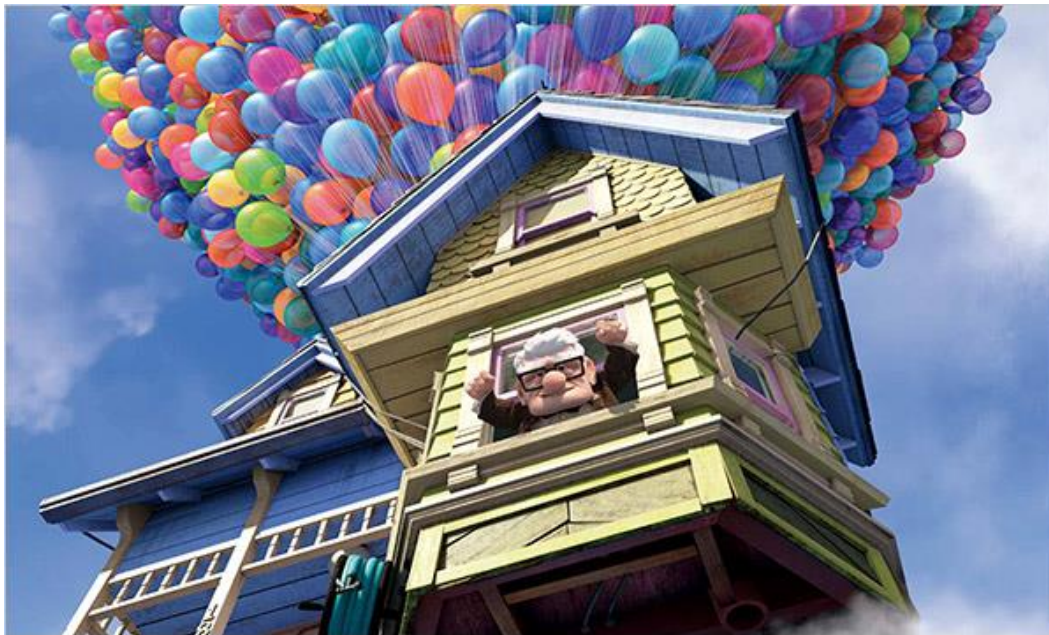
Bootstrap Algorithm (sample):

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

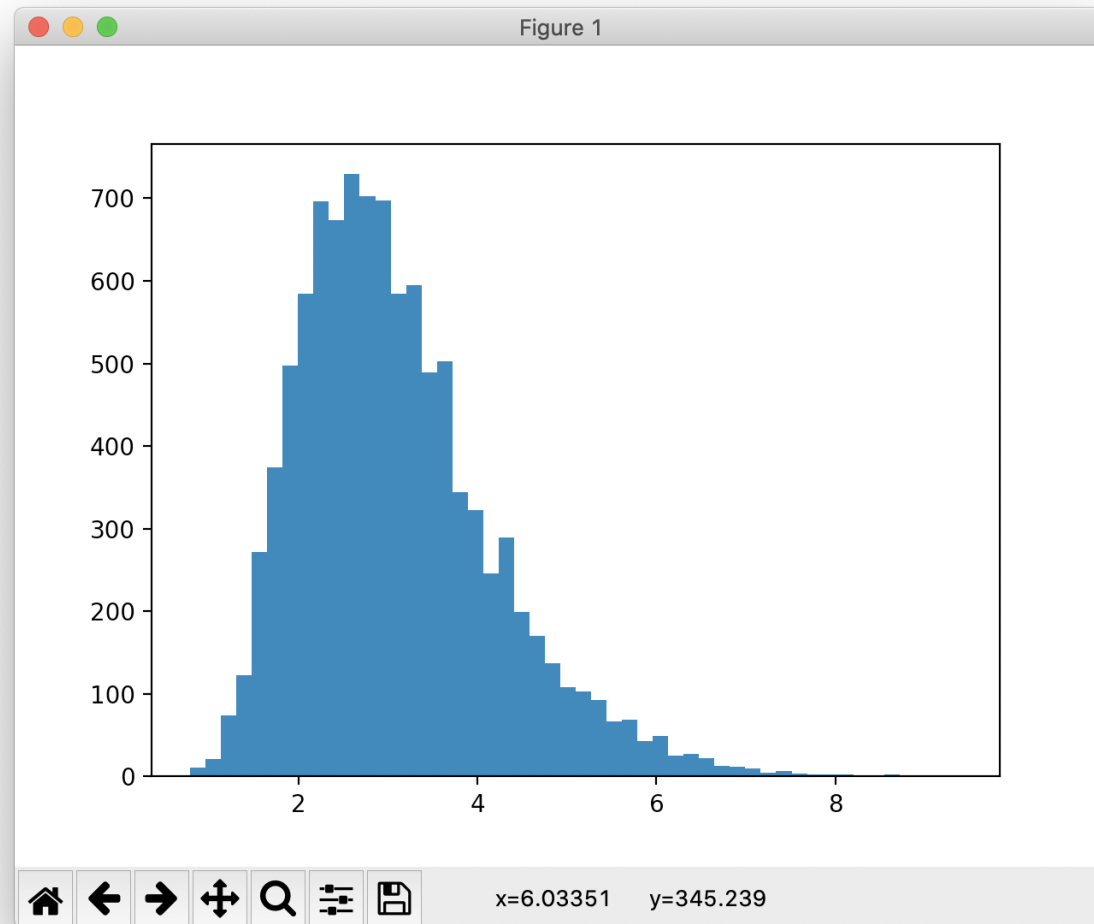
Bootstrap Algorithm (sample):

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



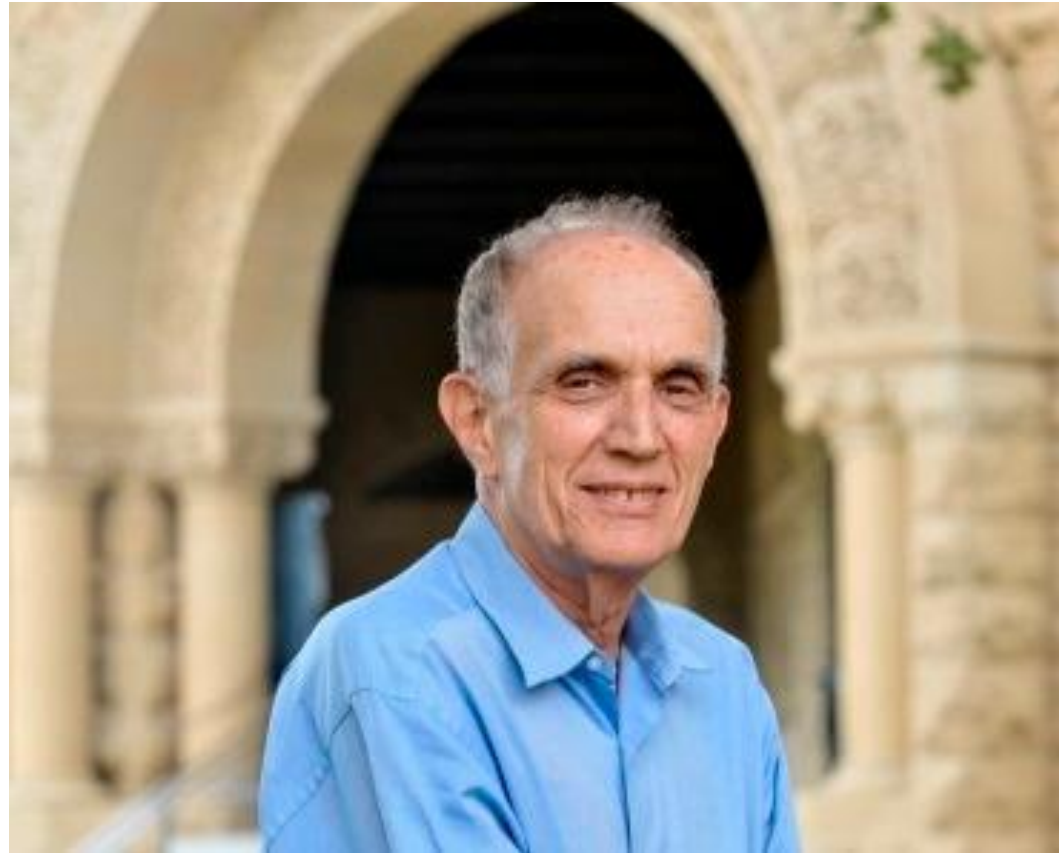


Bootstrap provides a way to calculate **probabilities of statistics** using code.

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?

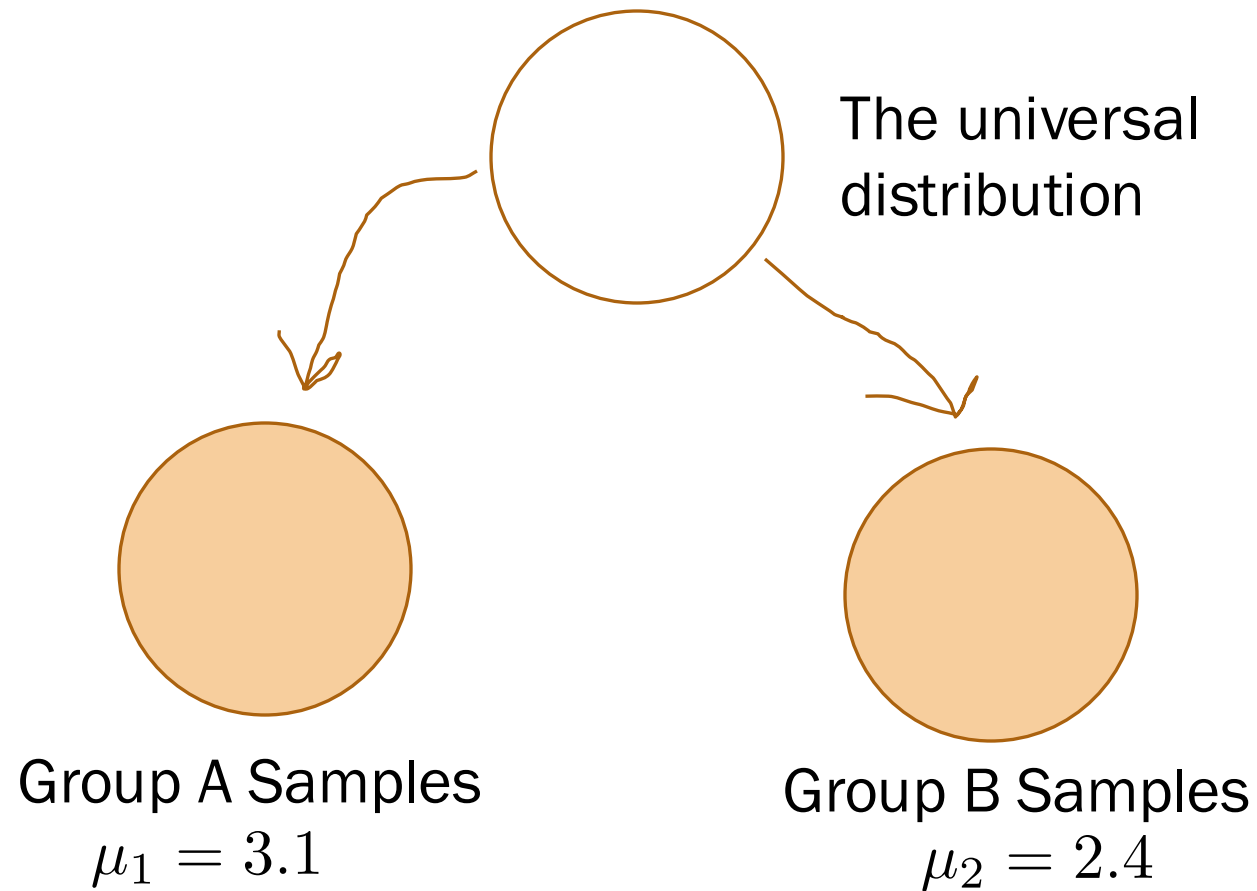
	Learning in Context A	Learning in Context B	
18 students	4.44	2.15	23 students
	3.36	3.01	
	5.87	2.02	
	2.31	1.43	
	
	3.70	1.83	
	$\mu_1 = 3.1$	$\mu_2 = 2.4$	

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?

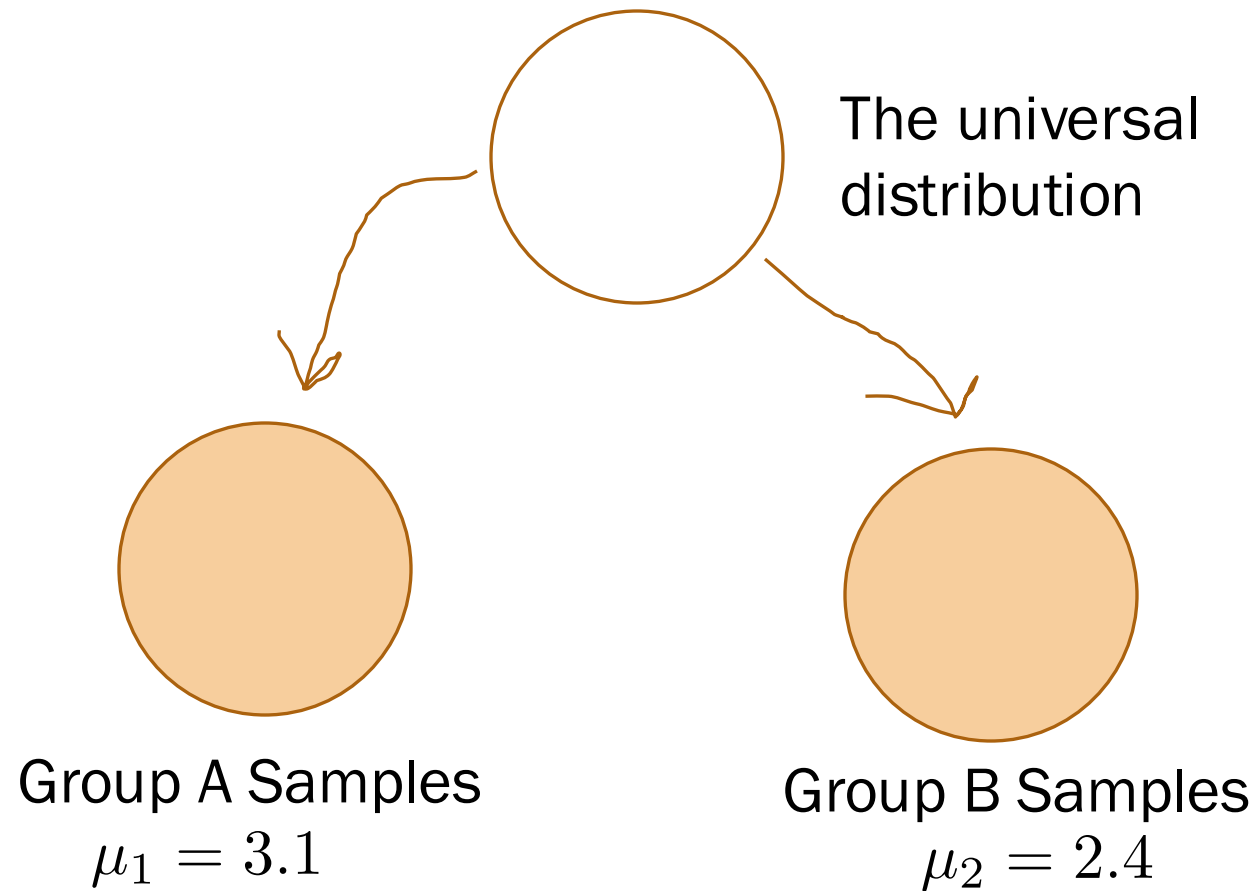
The Null Hypothesis

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.



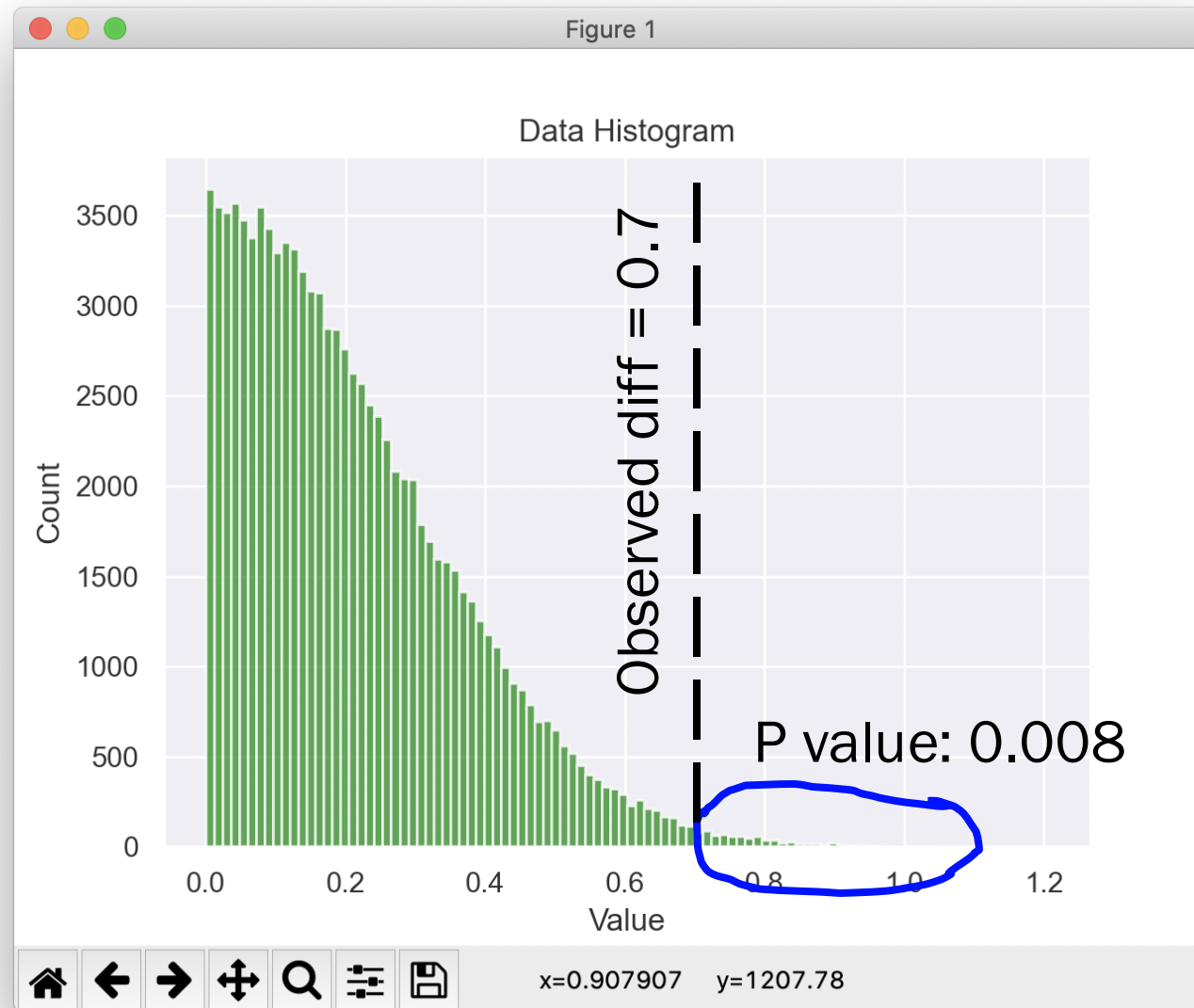
P-Value

The probability of obtaining test results **at least as extreme** as the result actually observed, if the null hypothesis is correct



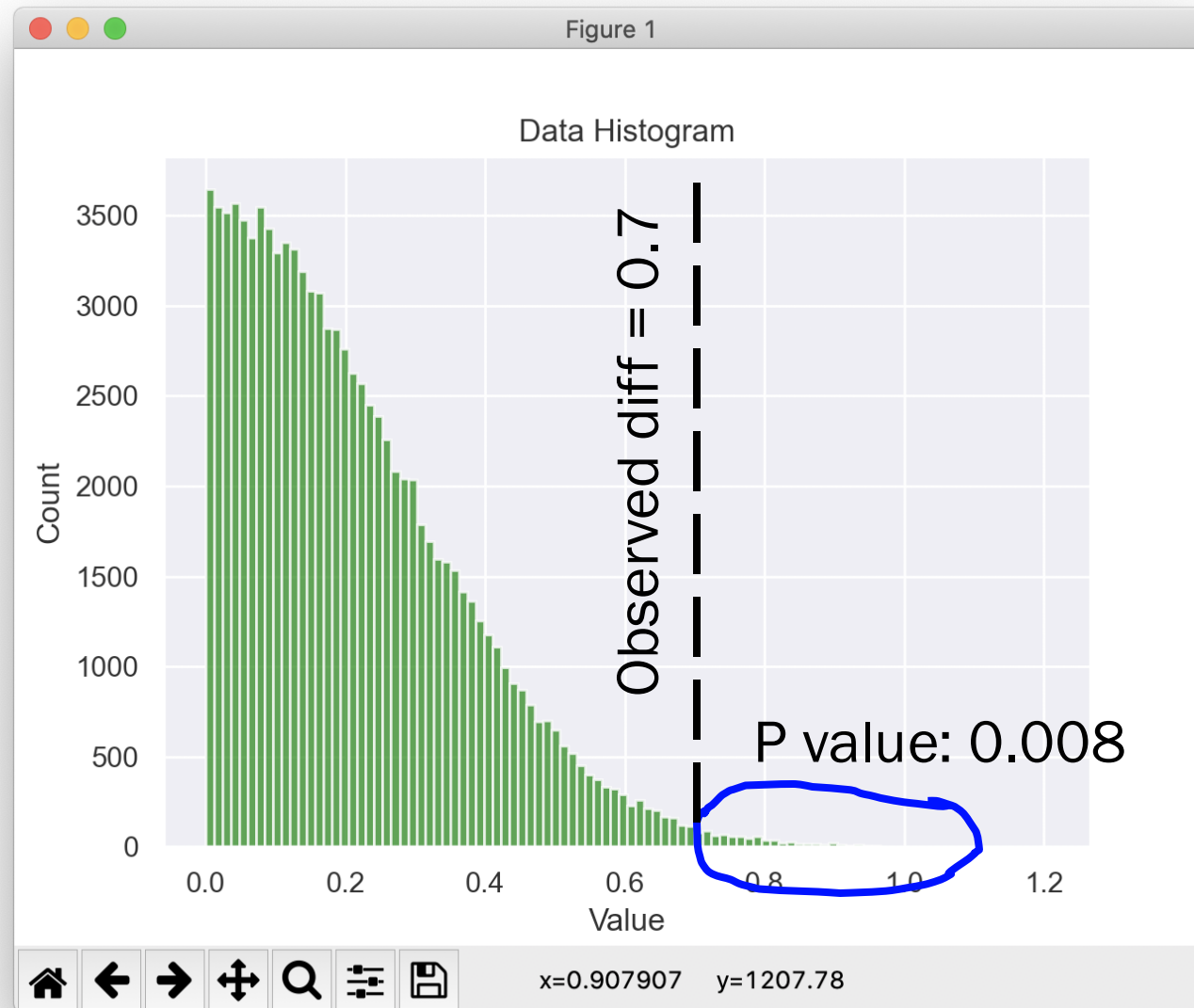
To the code!

Distribution of Mean Diffs under Null Hypothesis



Every* Science Result needs a p-value!

* almost



Food For Thought
(if extra time)

Puzzle

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

How can you build
distribution for p without
using a prior?

Two Opinions on Distributions

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

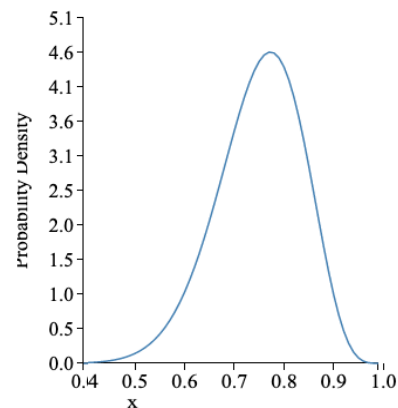
H, H, H, T, H, T, H, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

Let's use Laplace prior $X \sim \text{Beta}(2, 2)$

$X \sim \text{Beta}(a = 18, b = 6)$



Frequentist:

Let's bootstrap

