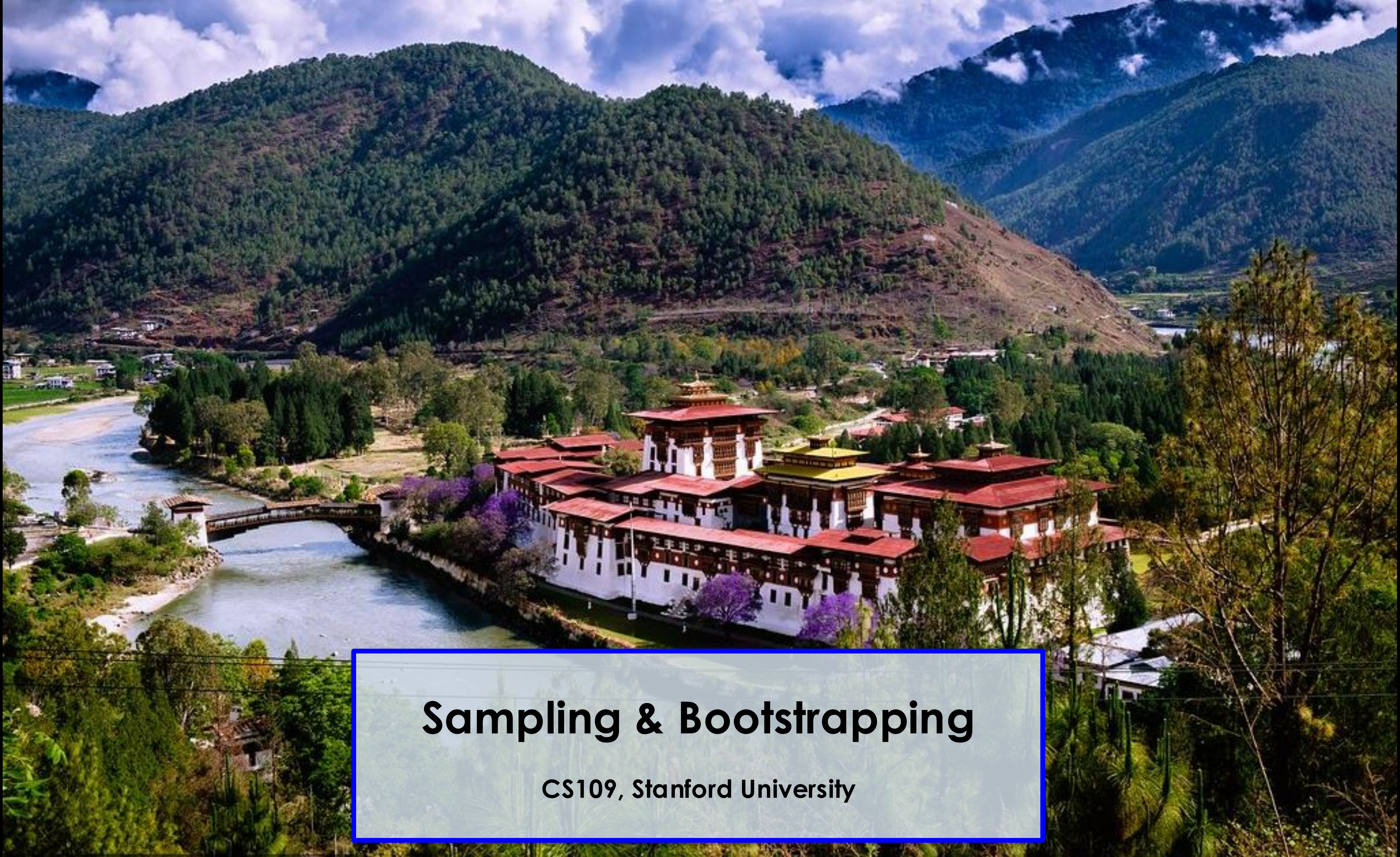
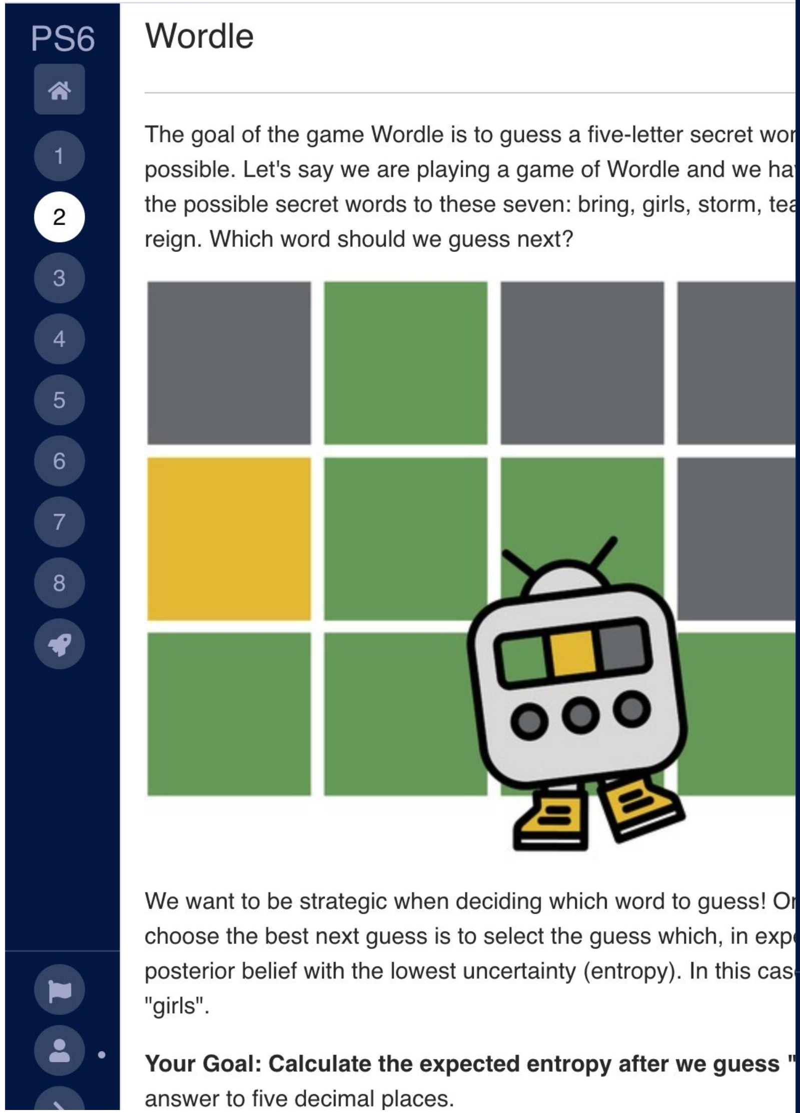


Sampling & Bootstrapping

CS109, Stanford University

Pset 6 is Out

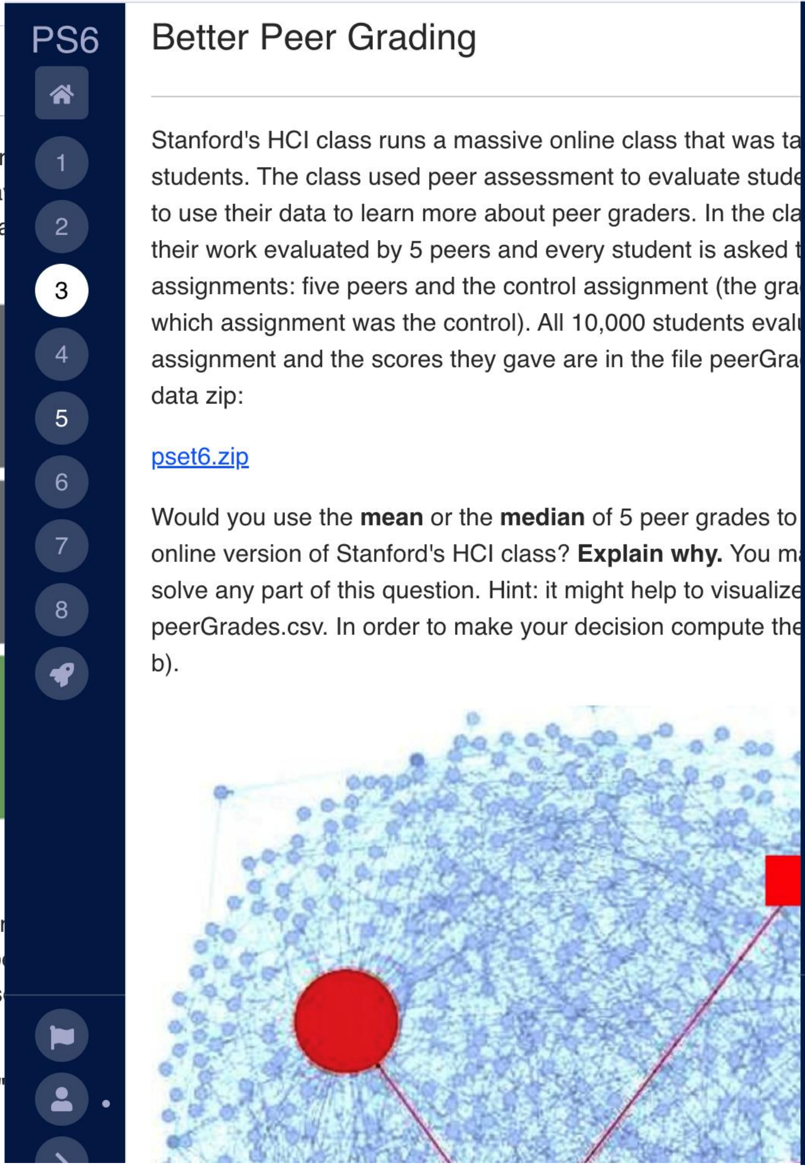


Wordle

The goal of the game Wordle is to guess a five-letter secret wor
possible. Let's say we are playing a game of Wordle and we ha
the possible secret words to these seven: bring, girls, storm, tea
reign. Which word should we guess next?

We want to be strategic when deciding which word to guess! O
choose the best next guess is to select the guess which, in exp
posterior belief with the lowest uncertainty (entropy). In this cas
"girls".

Your Goal: Calculate the expected entropy after we guess "
answer to five decimal places.

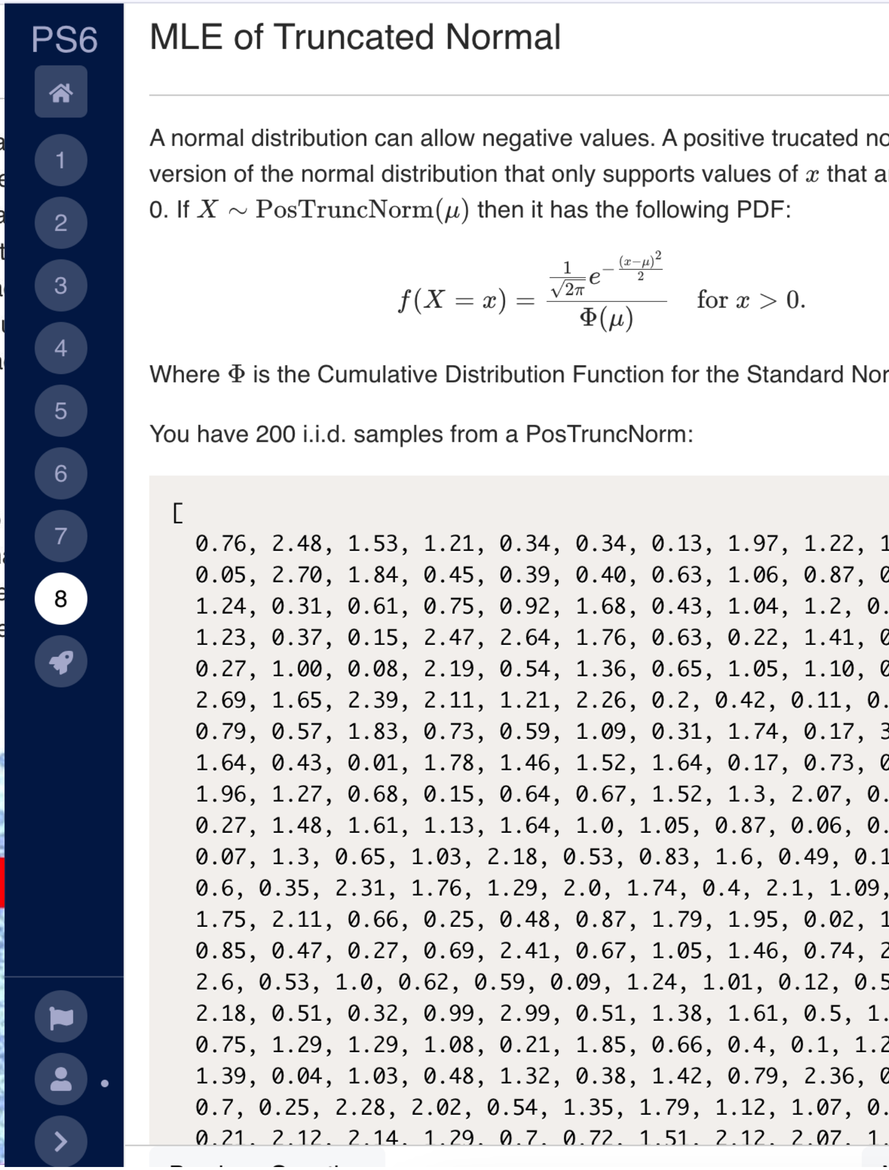


Better Peer Grading

Stanford's HCI class runs a massive online class that was ta
students. The class used peer assessment to evaluate stude
to use their data to learn more about peer graders. In the cla
their work evaluated by 5 peers and every student is asked t
assignments: five peers and the control assignment (the gra
which assignment was the control). All 10,000 students evalu
assignment and the scores they gave are in the file peerGra
data zip:

pset6.zip

Would you use the **mean** or the **median** of 5 peer grades to
online version of Stanford's HCI class? **Explain why.** You m
solve any part of this question. Hint: it might help to visualize
peerGrades.csv. In order to make your decision compute the
b).



MLE of Truncated Normal

A normal distribution can allow negative values. A positive trucated no
version of the normal distribution that only supports values of x that a
0. If $X \sim \text{PosTruncNorm}(\mu)$ then it has the following PDF:

$$f(X = x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}}{\Phi(\mu)} \quad \text{for } x > 0.$$

Where Φ is the Cumulative Distribution Function for the Standard Nor

You have 200 i.i.d. samples from a PosTruncNorm:

```
[
  0.76, 2.48, 1.53, 1.21, 0.34, 0.34, 0.13, 1.97, 1.22, 1
  0.05, 2.70, 1.84, 0.45, 0.39, 0.40, 0.63, 1.06, 0.87, 0
  1.24, 0.31, 0.61, 0.75, 0.92, 1.68, 0.43, 1.04, 1.2, 0.
  1.23, 0.37, 0.15, 2.47, 2.64, 1.76, 0.63, 0.22, 1.41, 0
  0.27, 1.00, 0.08, 2.19, 0.54, 1.36, 0.65, 1.05, 1.10, 0
  2.69, 1.65, 2.39, 2.11, 1.21, 2.26, 0.2, 0.42, 0.11, 0.
  0.79, 0.57, 1.83, 0.73, 0.59, 1.09, 0.31, 1.74, 0.17, 3
  1.64, 0.43, 0.01, 1.78, 1.46, 1.52, 1.64, 0.17, 0.73, 0
  1.96, 1.27, 0.68, 0.15, 0.64, 0.67, 1.52, 1.3, 2.07, 0.
  0.27, 1.48, 1.61, 1.13, 1.64, 1.0, 1.05, 0.87, 0.06, 0.
  0.07, 1.3, 0.65, 1.03, 2.18, 0.53, 0.83, 1.6, 0.49, 0.1
  0.6, 0.35, 2.31, 1.76, 1.29, 2.0, 1.74, 0.4, 2.1, 1.09,
  1.75, 2.11, 0.66, 0.25, 0.48, 0.87, 1.79, 1.95, 0.02, 1
  0.85, 0.47, 0.27, 0.69, 2.41, 0.67, 1.05, 1.46, 0.74, 2
  2.6, 0.53, 1.0, 0.62, 0.59, 0.09, 1.24, 1.01, 0.12, 0.5
  2.18, 0.51, 0.32, 0.99, 2.99, 0.51, 1.38, 1.61, 0.5, 1.
  0.75, 1.29, 1.29, 1.08, 0.21, 1.85, 0.66, 0.4, 0.1, 1.2
  1.39, 0.04, 1.03, 0.48, 1.32, 0.38, 1.42, 0.79, 2.36, 0
  0.7, 0.25, 2.28, 2.02, 0.54, 1.35, 1.79, 1.12, 1.07, 0.
  0.21, 2.12, 2.14, 1.29, 0.7, 0.72, 1.51, 2.12, 2.07, 1.
```

Previously on CS109

Where are we in CS109?

You are here

Counting Theory

Core Probability

Random Variables

Probabilistic Models

Uncertainty Theory

Machine Learning

Uncertainty Theory

Beta Distributions

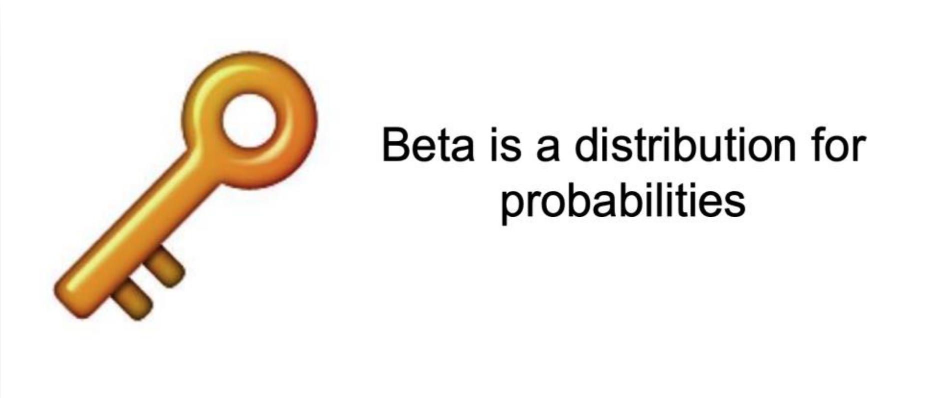
Adding Random Vars

Central Limit Theorem

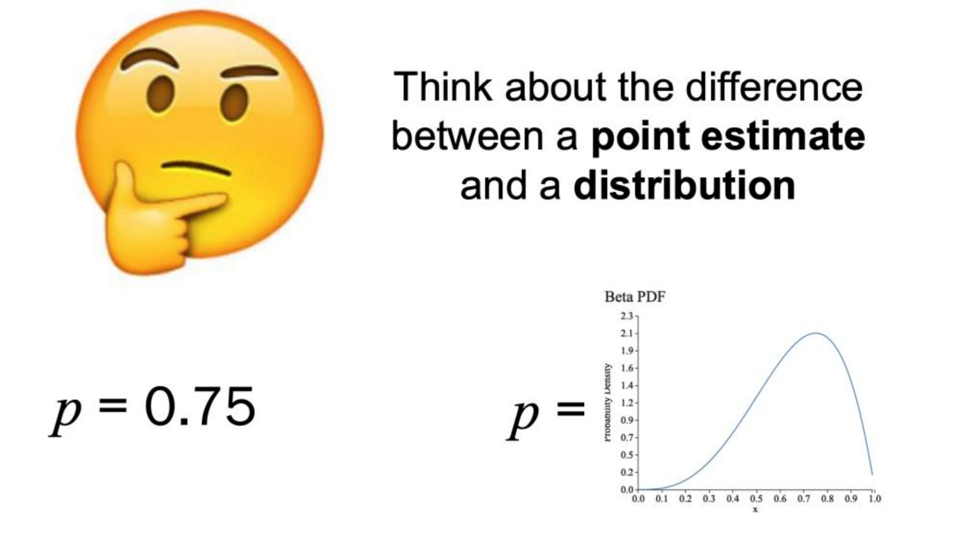
Algorithmic Analysis

Information Theory

Beta Distributions



Beta is a distribution for probabilities

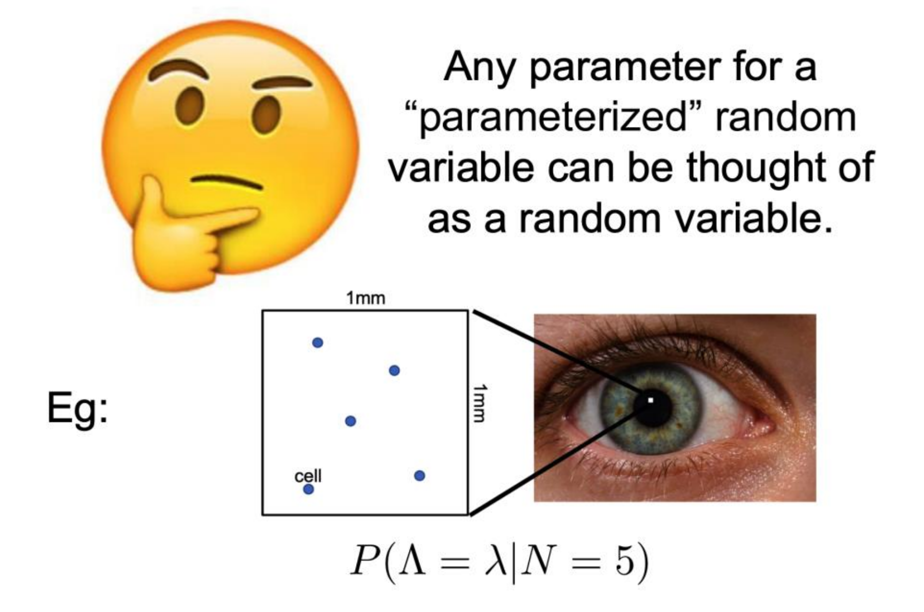


Think about the difference between a **point estimate** and a **distribution**

$p = 0.75$

$p =$

Beta PDF
2.3
2.1
1.9
1.6
1.4
1.2
0.9
0.7
0.5
0.2
0.0
Probability Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
x



Any parameter for a "parameterized" random variable can be thought of as a random variable.

Eg:

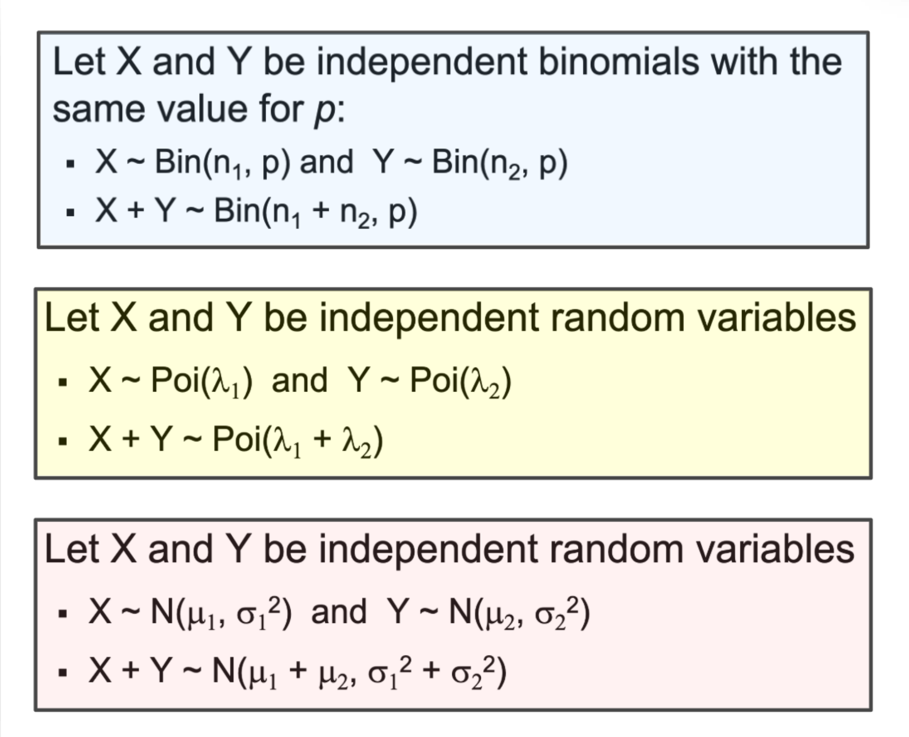
1mm
1mm
cell

$$P(\Lambda = \lambda | N = 5)$$

Adding Random Vars

The sum of two random variables is another random variable

Z = X + Y



Let X and Y be independent binomials with the same value for *p*:

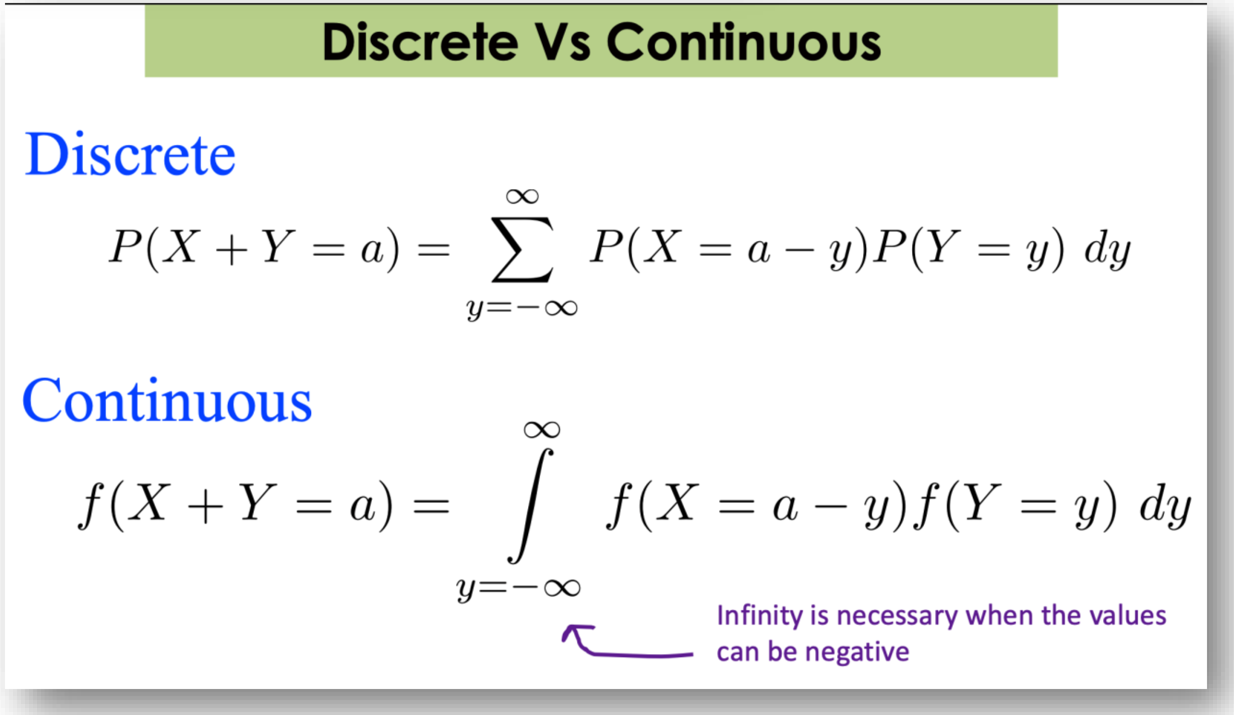
- $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$
- $X + Y \sim \text{Bin}(n_1 + n_2, p)$

Let X and Y be independent random variables

- $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$
- $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$

Let X and Y be independent random variables

- $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$
- $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$



Discrete Vs Continuous

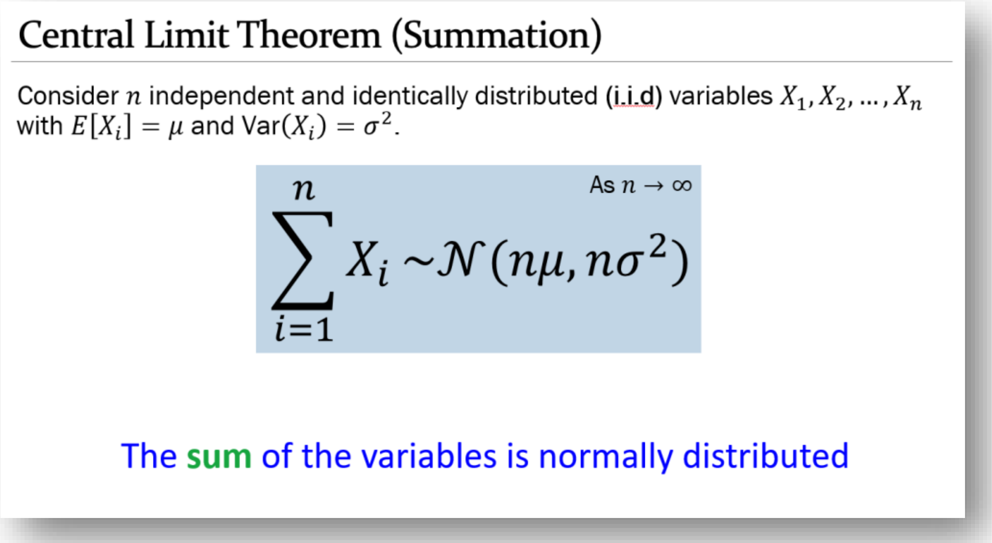
Discrete

$$P(X + Y = a) = \sum_{y=-\infty}^{\infty} P(X = a - y)P(Y = y)\ dy$$

Continuous

$$f(X + Y = a) = \int_{y=-\infty}^{\infty} f(X = a - y)f(Y = y)\ dy$$

Infinity is necessary when the values can be negative



Central Limit Theorem (Summation)

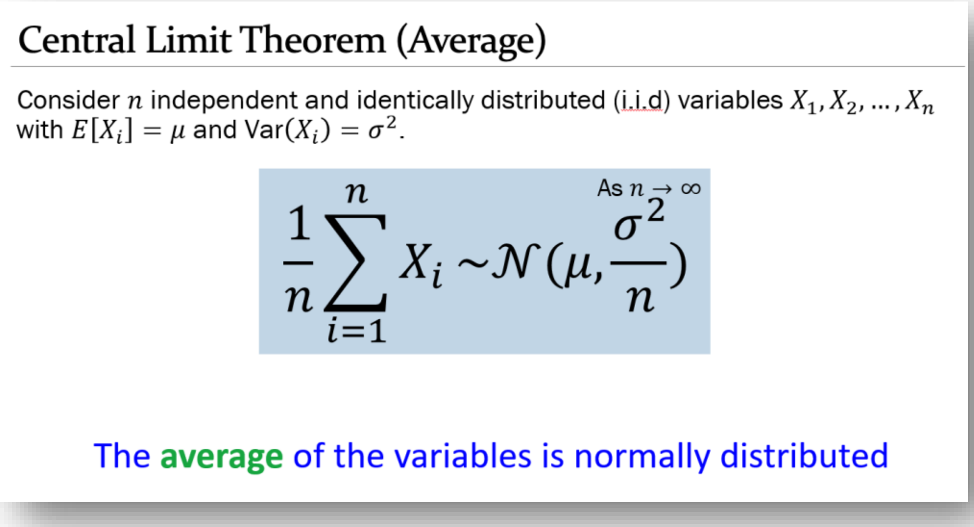
Consider n independent and identically distributed (i.i.d) variables $X_1, X_2, \dots, X_n$ with $E[X_i] = \mu$ and Var$(X_i) = \sigma^2$.

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



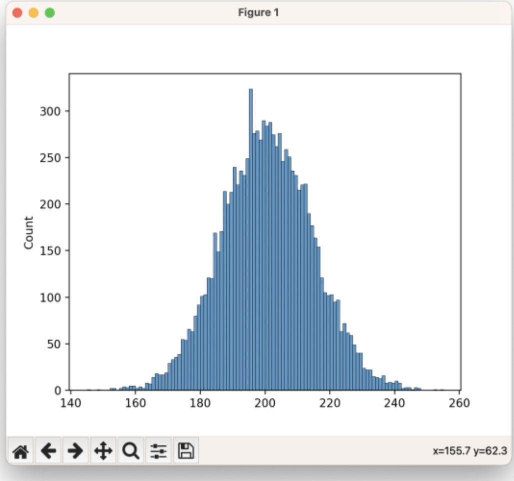
Central Limit Theorem (Average)

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

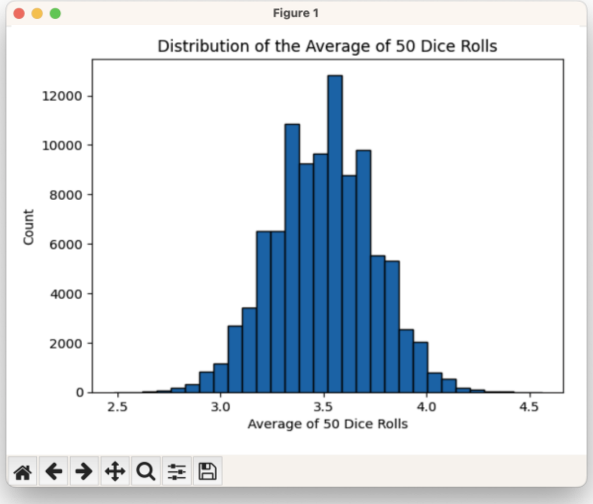
As $n \to \infty$

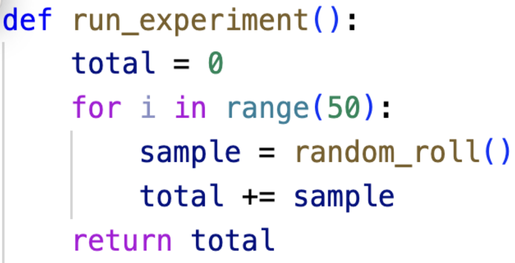
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

The average of the variables is normally distributed

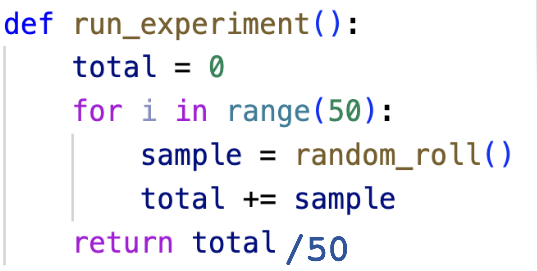


Repeat experiment many many times



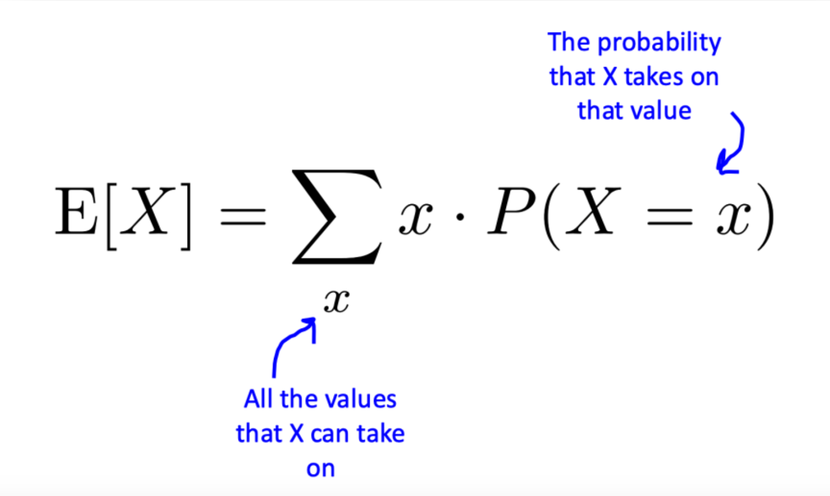


```
def run_experiment():
    total = 0
    for i in range(50):
        sample = random_roll()
        total += sample
    return total
```



```
def run_experiment():
    total = 0
    for i in range(50):
        sample = random_roll()
        total += sample
    return total /50
```

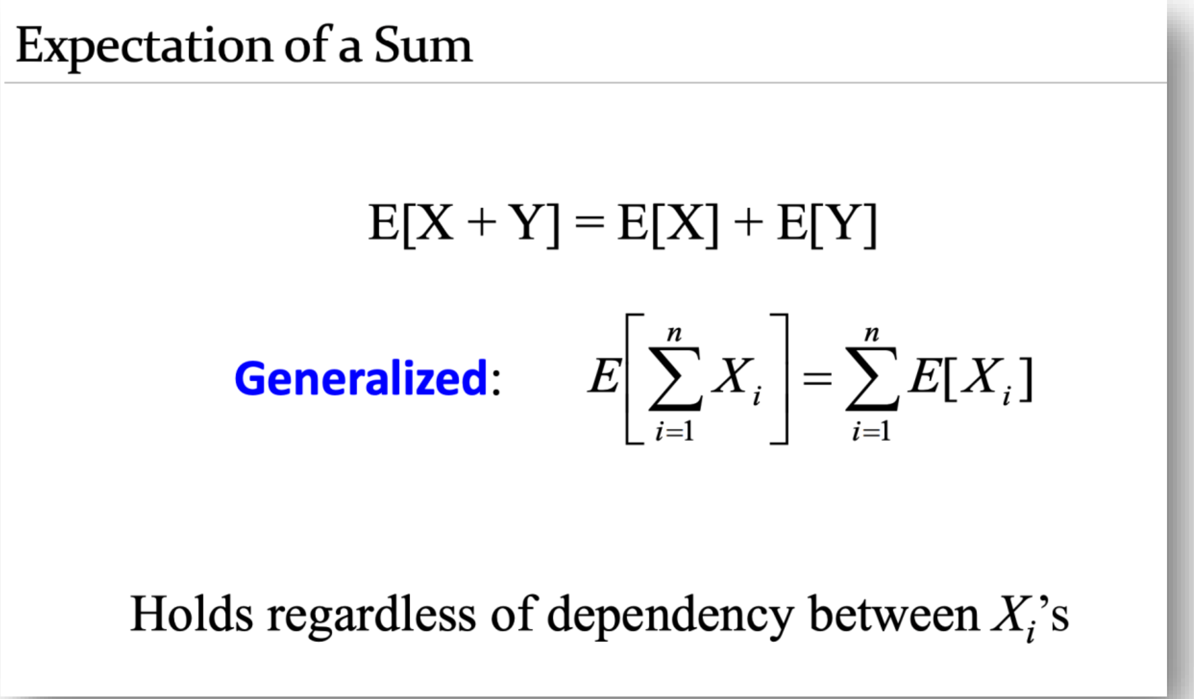
Algorithmic Analysis



The probability that X takes on that value

$$\mathrm{E}[X] = \sum_x x \cdot P(X = x)$$

All the values that X can take on

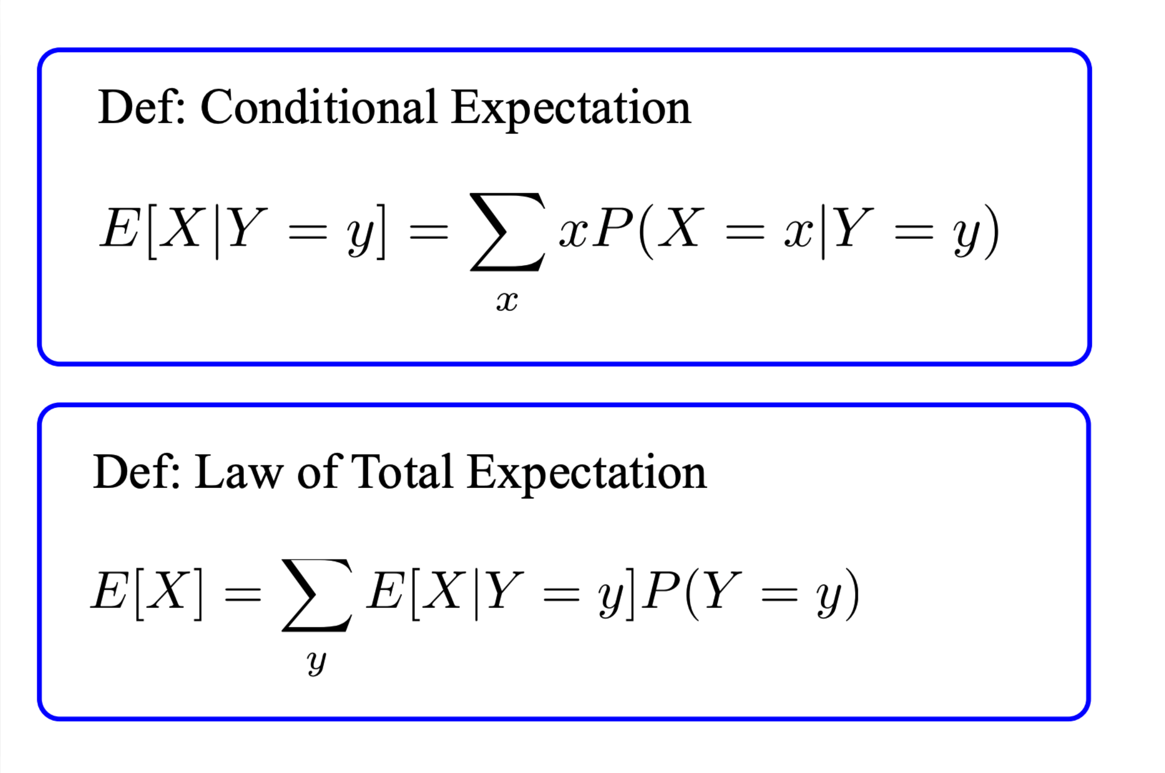


Expectation of a Sum

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

Generalized: $$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Holds regardless of dependency between X_i's



Def: Conditional Expectation

$$E[X|Y = y] = \sum_x xP(X = x|Y = y)$$

Def: Law of Total Expectation

$$E[X] = \sum_y E[X|Y = y]P(Y = y)$$

Information Theory

Uncertainty of a Random Variable (Entropy)

Let X be any random variable. We can calculate a statistic, “**Uncertainty**” to express how much we don’t know about X

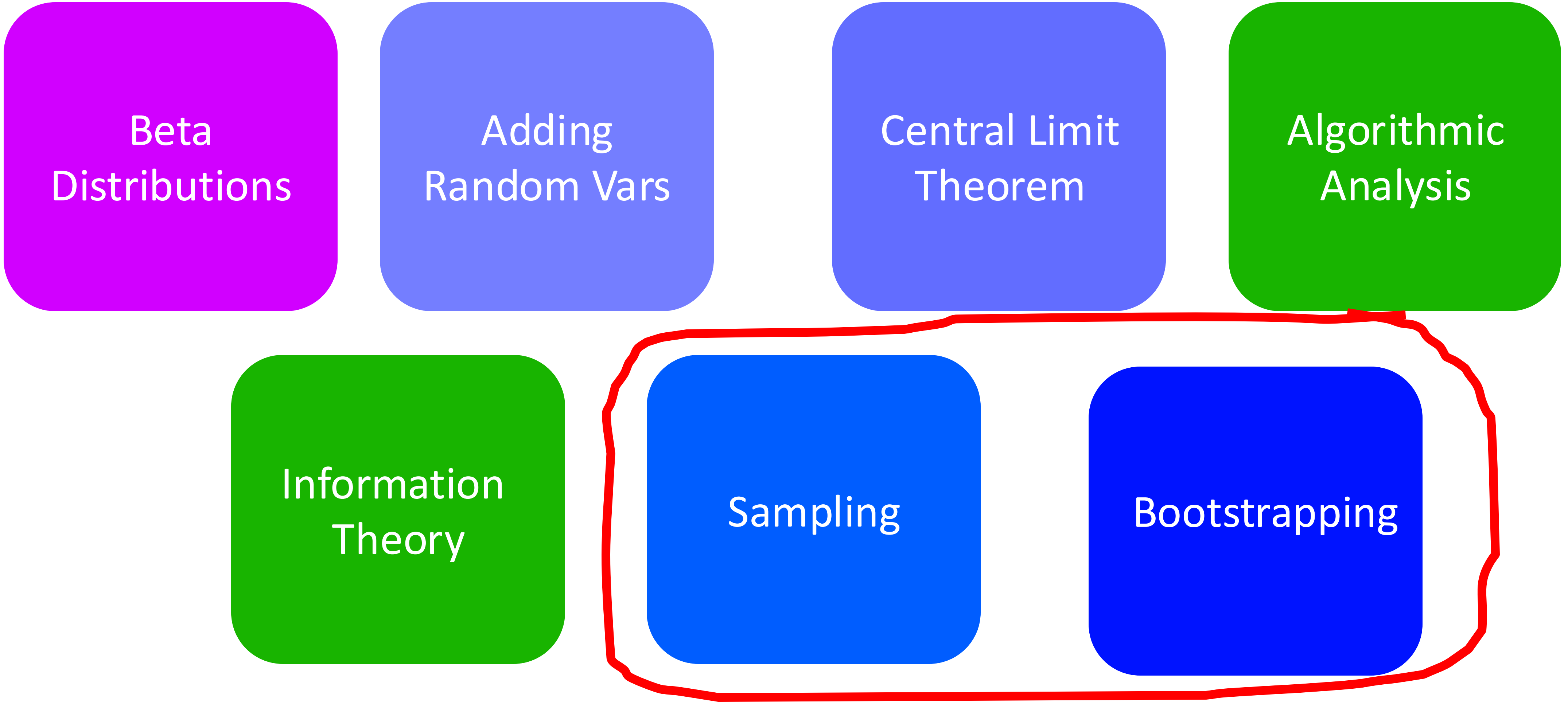
Calculates expected surprise

$$\overset{H}{\cancel{\text{Uncertainty}}}(X) = \sum_{x \in X} \overbrace{\log_2 \frac{1}{P(X = x)}} \cdot P(X = x)$$

My preferred name for ”entropy” aka H(X)

$$\text{Surprise}(X = x) = \log_2 \frac{1}{P(X = x})$$

Uncertainty Theory



Beta Distributions

Adding Random Vars

Central Limit Theorem

Algorithmic Analysis

Information Theory

Sampling

Bootstrapping

<end review>

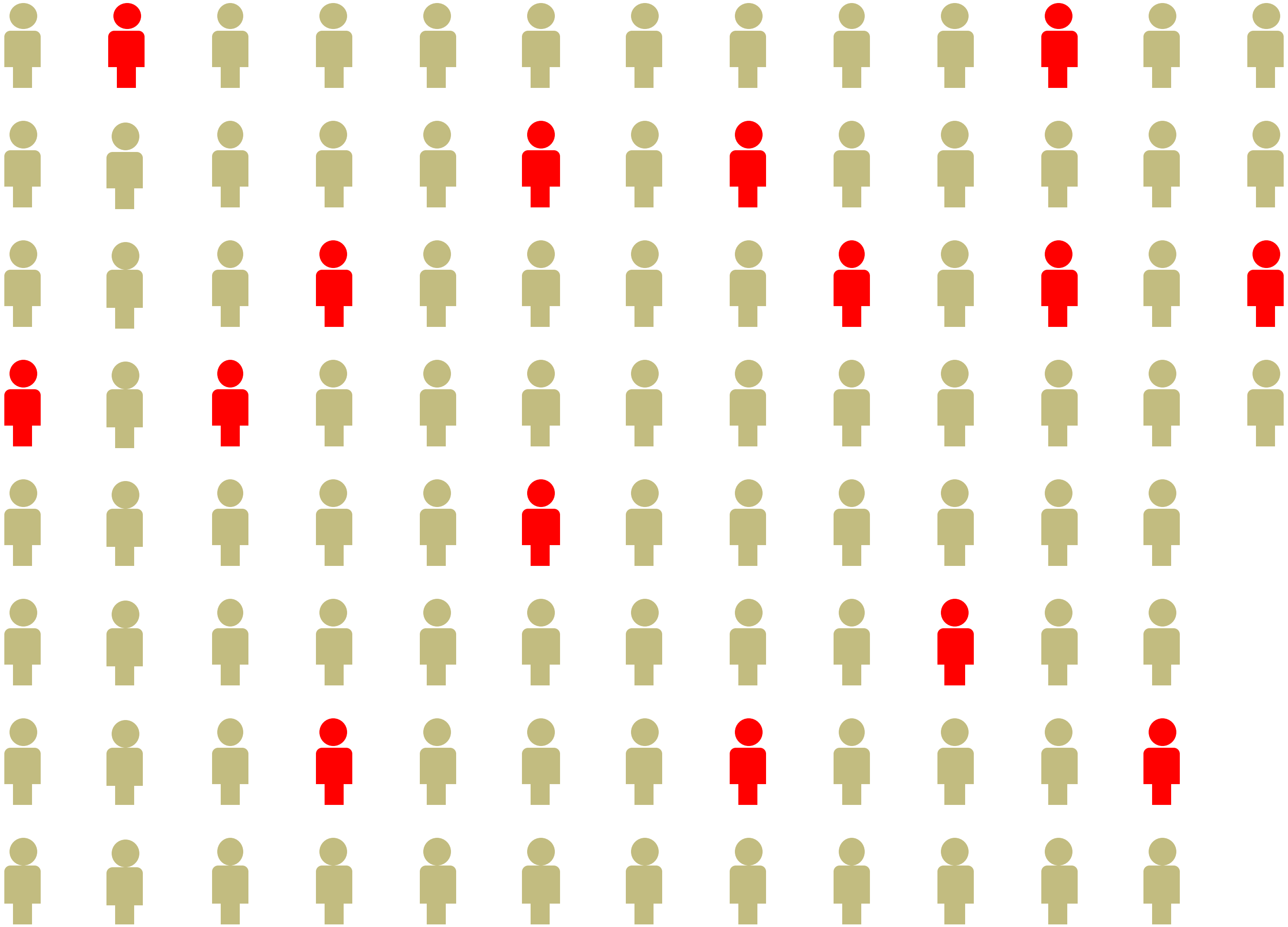
Motivating example

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

- But you can't ask everyone.
- You poll 200 random people.

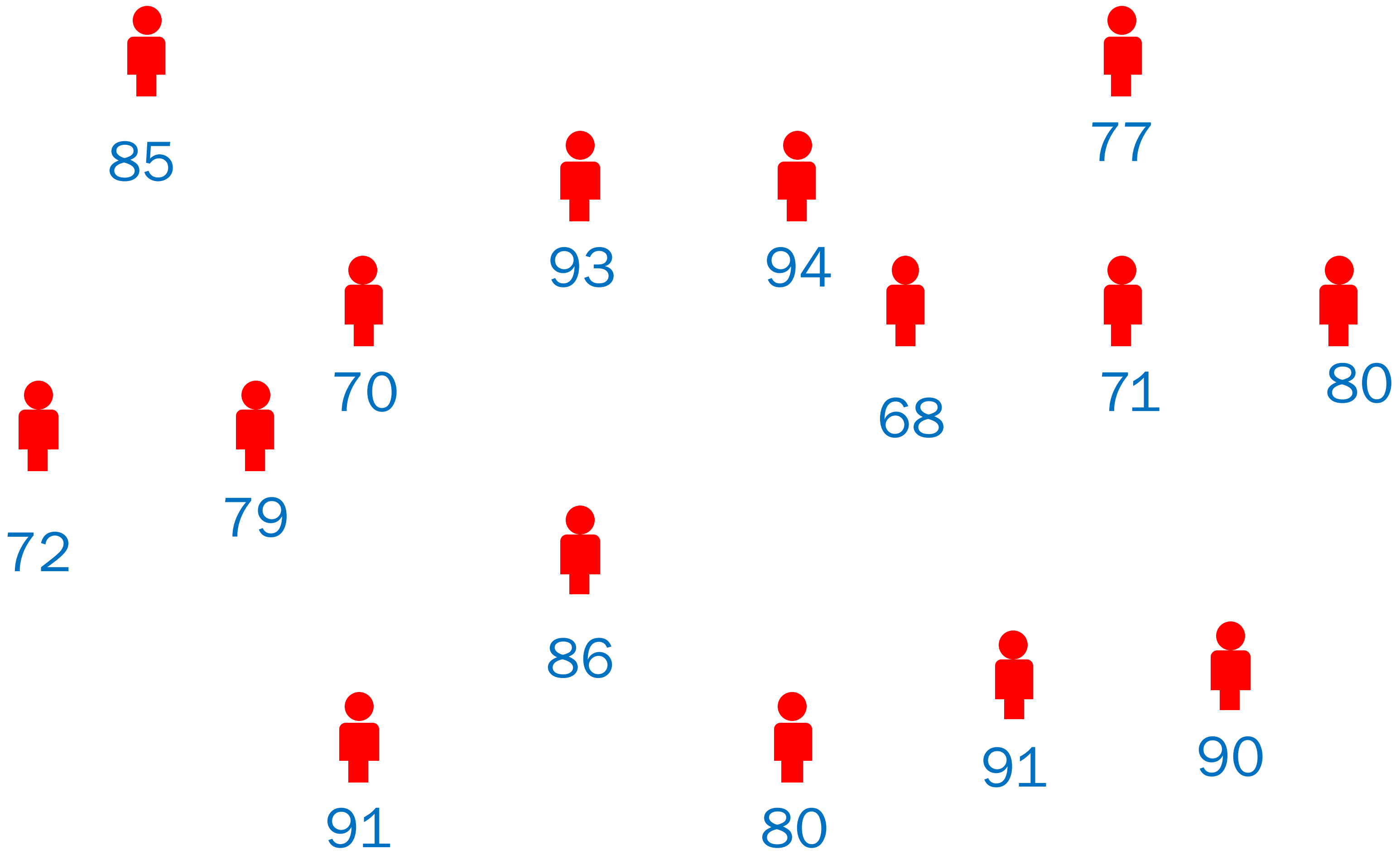
Population

Sample



Sample

sample = [72, 85, 79, 91, 68, …, 71]



Collect one (or more) measurements from each person

Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

= 83

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

Sample Variance

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

= 40

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

Population

Unknown
(need to estimate)

Population Mean

true mean

$$\mu = \frac{\sum x_i}{N}$$

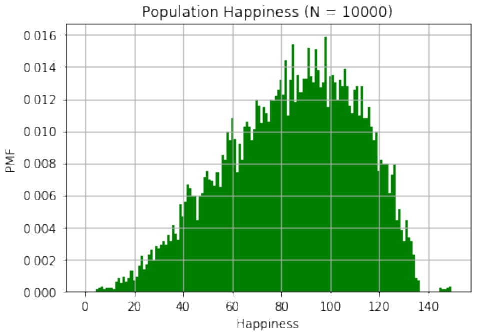
Population Variance

true variance

population mean

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

Population Distribution

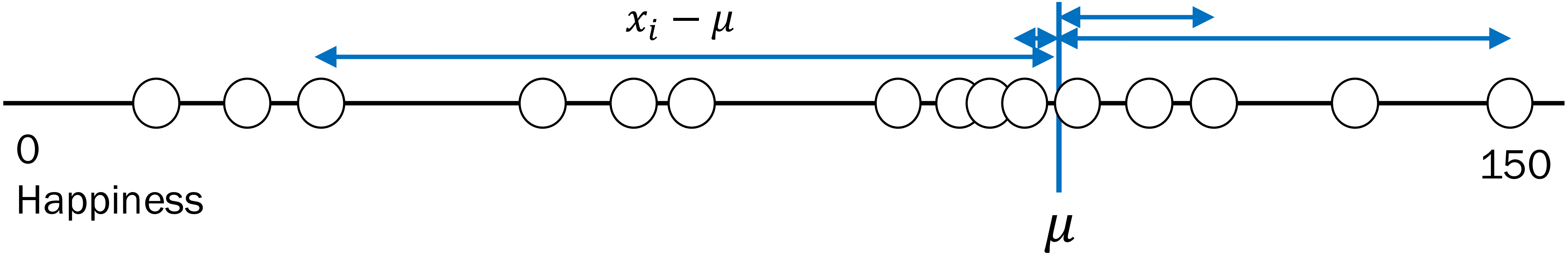


Population Statistics

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.

A single sample

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

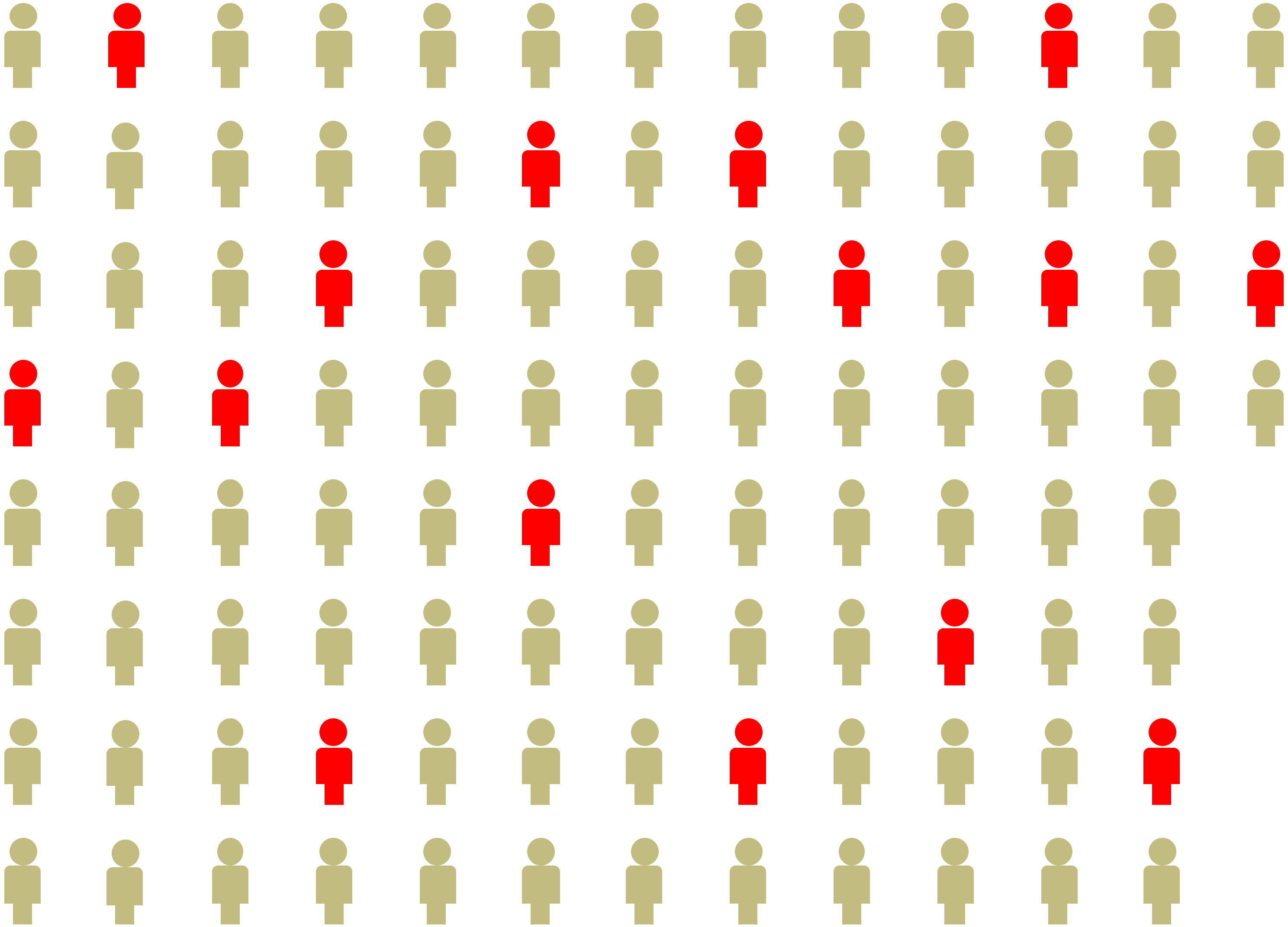
But we only have 200 people (a sample).

So these population statistics are unknown:

- μ, the population mean
- σ^2, the population variance

Estimating Population Statistics (Mean + Var)

Sample

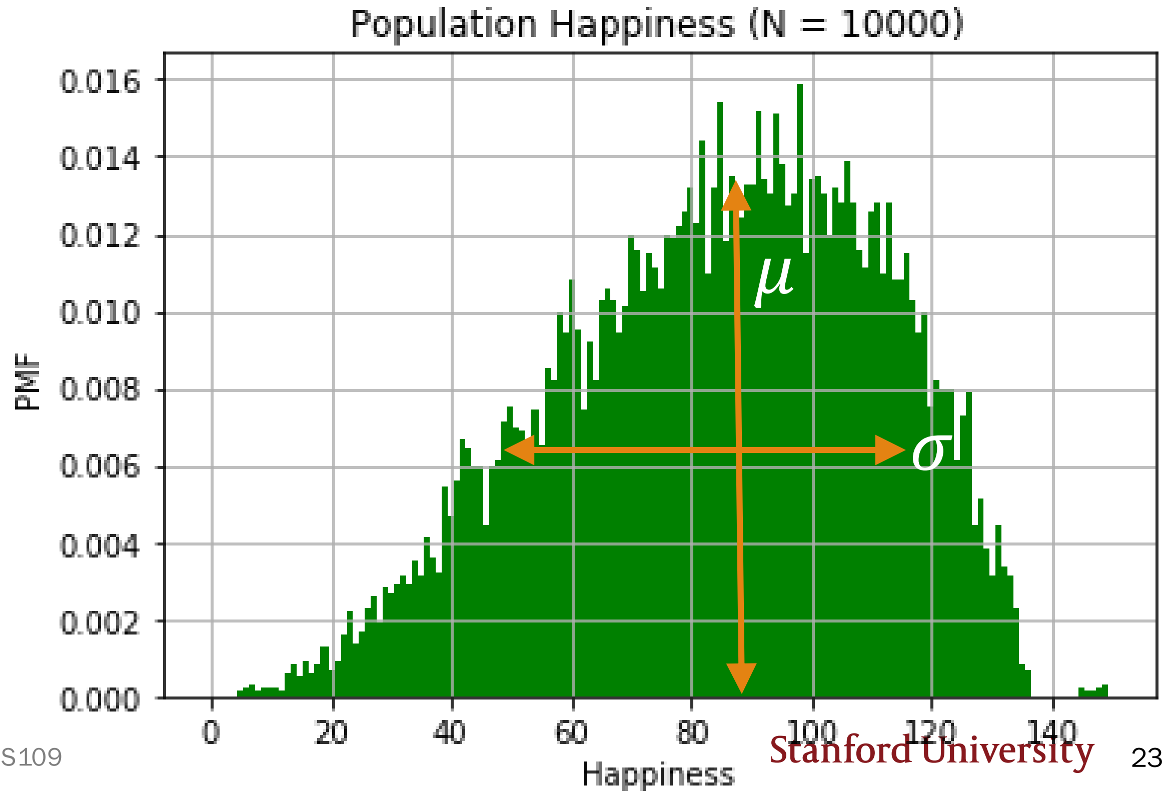


A sample, mathematically

Consider n random variables $X_1, X_2, \ldots, X_n$.

The sequence $X_1, X_2, \ldots, X_n$ is a **sample** from distribution F if:

- X_i are all independent and identically distributed (i.i.d.)
- X_i all have same distribution function F (the **underlying distribution**), where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$



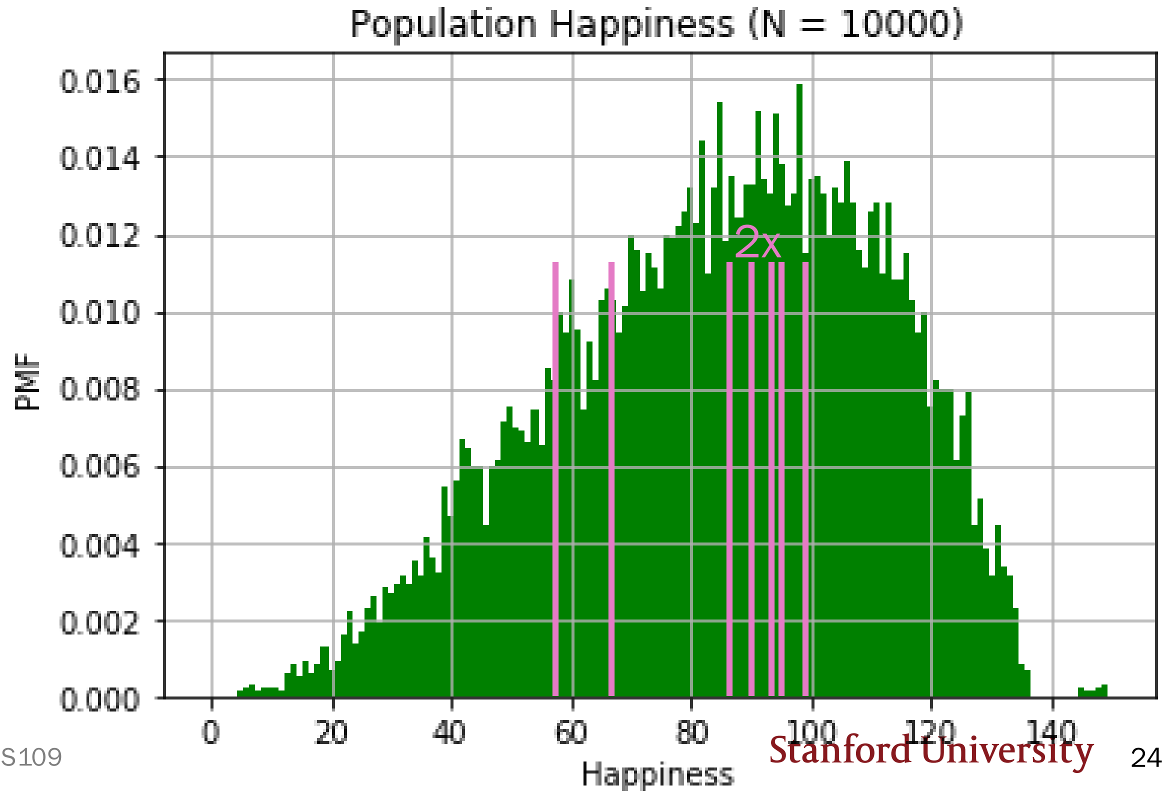
A sample, mathematically

A sample of sample size 8:

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

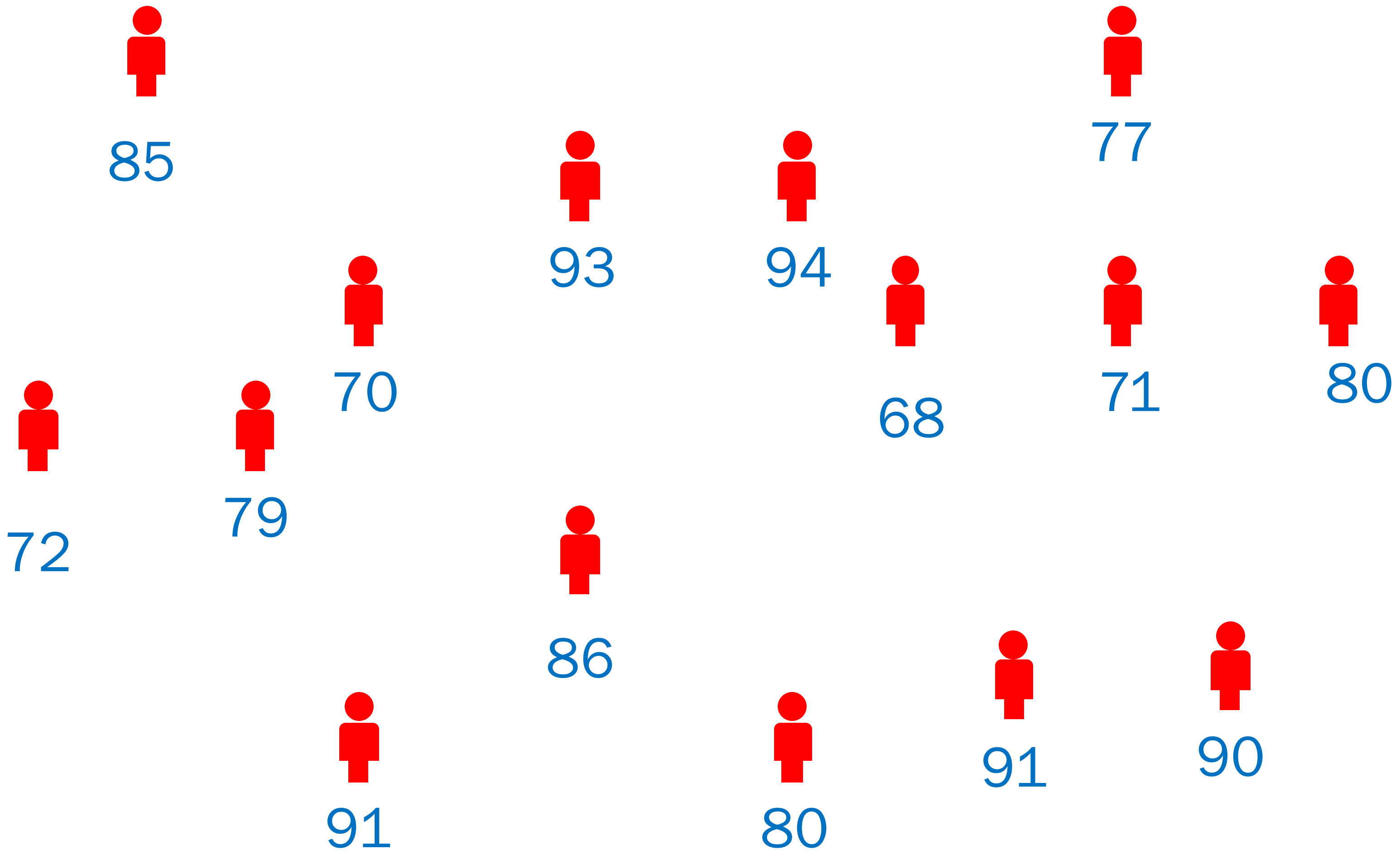
A realization of a sample of size 8:

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



Sample

sample = [72, 85, 79, 91, 68, …, 71]



Sample Mean

Unbiased estimator of population mean, μ

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

= 83

Sample Variance

Unbiased estimator of population variance, σ^2

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

= 40

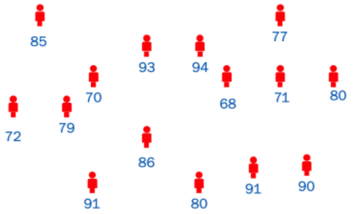
Sample Mean

Is that estimate any good? Biased vs Unbiased Estimators

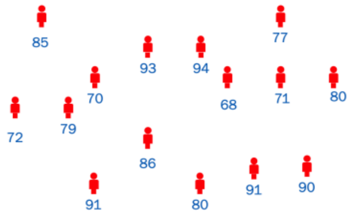
Suppose that we knew the population mean happiness in Bhutan. How can we evaluate whether the sample mean estimates the population mean well enough?

$\mu = 80$

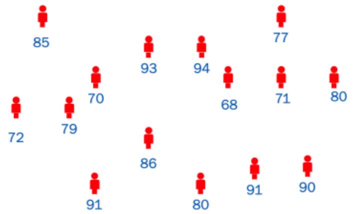
Test the estimator and keep track of the sample means:



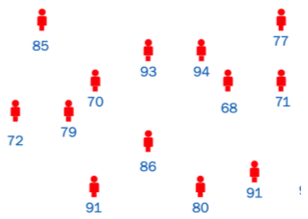
$\bar{x} = 83$



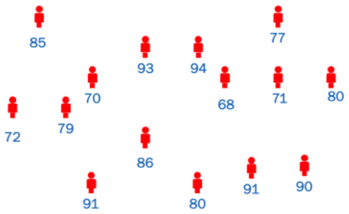
$\bar{x} = 90$



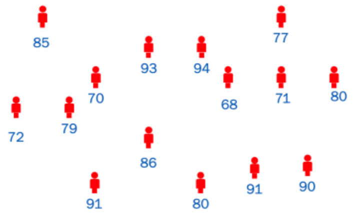
$\bar{x} = 82$



$\bar{x} = 70$



$\bar{x} = 65$



$\bar{x} = 82$

Take many, many more samples and compute the sample means

Sample mean is an unbiased estimator of population mean

If we take many samples of size *n*, the average of the sample means will be the population mean.

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

$$E[\bar{X}] = \mu$$

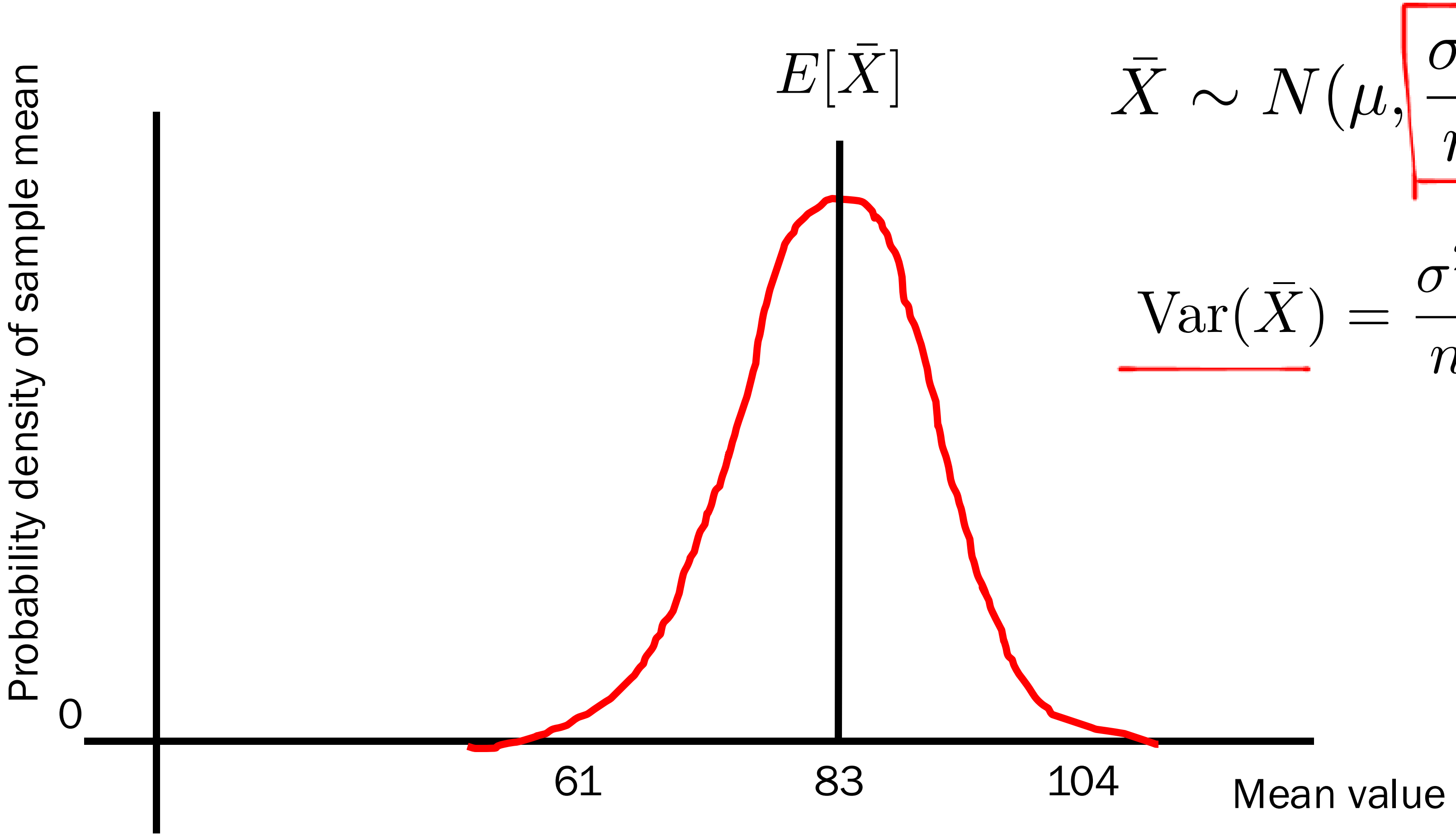
$$= \frac{1}{n} \sum_{i=1}^{n} E[X_i]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu$$

$$= \frac{1}{n} n\mu$$

$$= \mu$$

Insight: Sample Mean is an RV with known Distribution



By central limit theorem:

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

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

Sample mean is an unbiased estimator of population mean

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

The best estimate of μ is the **sample mean**:

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

$\bar{X}$ is an **unbiased estimator** of the population mean μ.

$$E[\bar{X}] = \mu$$

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

If we could take *multiple* samples of size n:

1. For each sample, compute sample mean
2. On average, we would get the population mean

Sample Mean:

ith sample

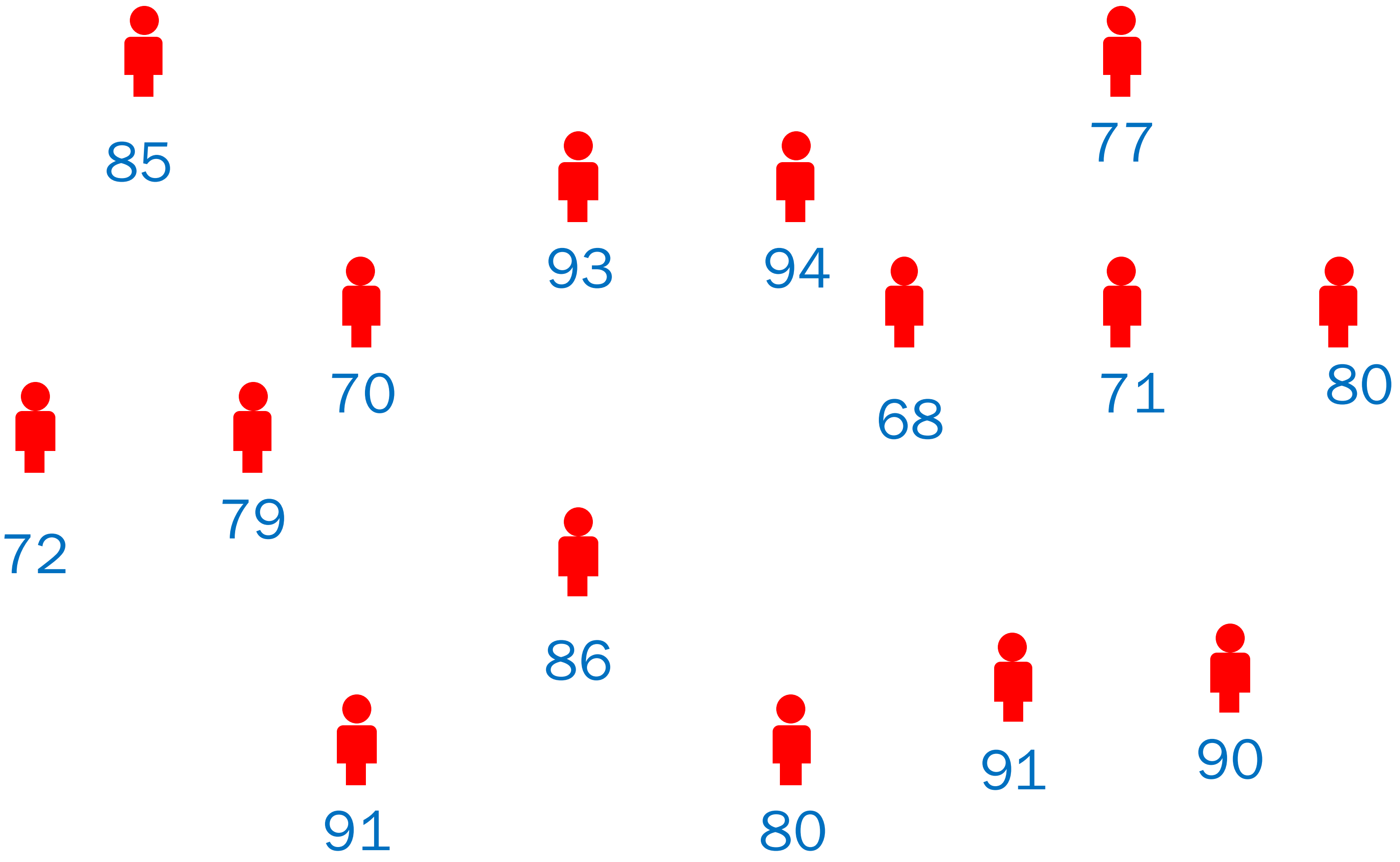
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Size of the sample

Sample Variance

Sample

sample = [72, 85, 79, 91, 68, ..., 71]



Sample Mean

Unbiased estimator of population mean, μ

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

= 83

Sample Variance

Unbiased estimator of population variance, σ²

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

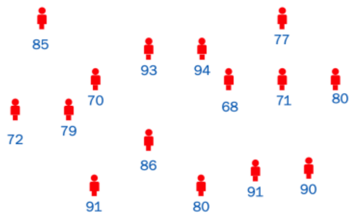
= 40

Is that estimate any good? Biased vs Unbiased Estimators

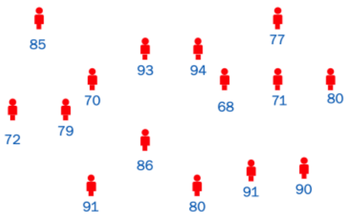
Suppose that we knew the population variance. How can we evaluate whether the sample variance estimates the population variance well enough?

$\sigma^2 = 45$

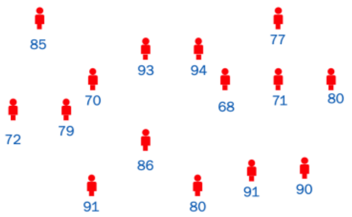
Test the estimator and keep track of the sample variances:



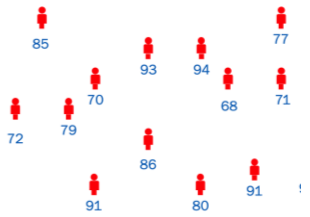
$S^2 = 40$



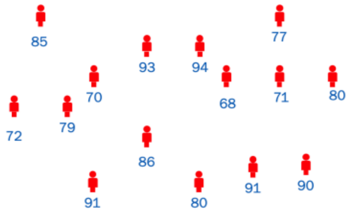
$S^2 = 55$



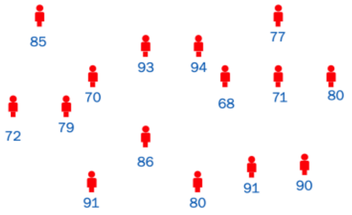
$S^2 = 67$



$S^2 = 43$



$S^2 = 35$



$S^2 = 30$

Take many, many more samples and compute the sample variances..

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

The average of the sample variances wil be the population variance

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

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

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

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

(introduce $\mu - \mu$)

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

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

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

$$2(\mu - \bar{X})n(\bar{X} - \mu)$$

$$-2n(\mu - \bar{X})^2$$

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

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

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

Therefore $\boxed{E[S^2] = \sigma^2}$

Biased sample variance

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

sample mean

sample variance

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

Intuition about the sample variance, S^2

Actual, σ^2

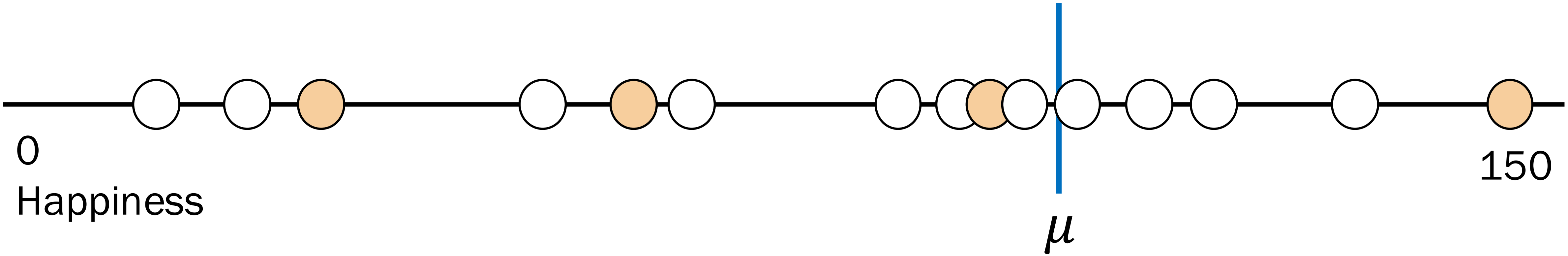
population mean

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

Estimate, S^2

sample mean

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



Population size, N

Intuition about the sample variance, S^2

Actual, σ^2

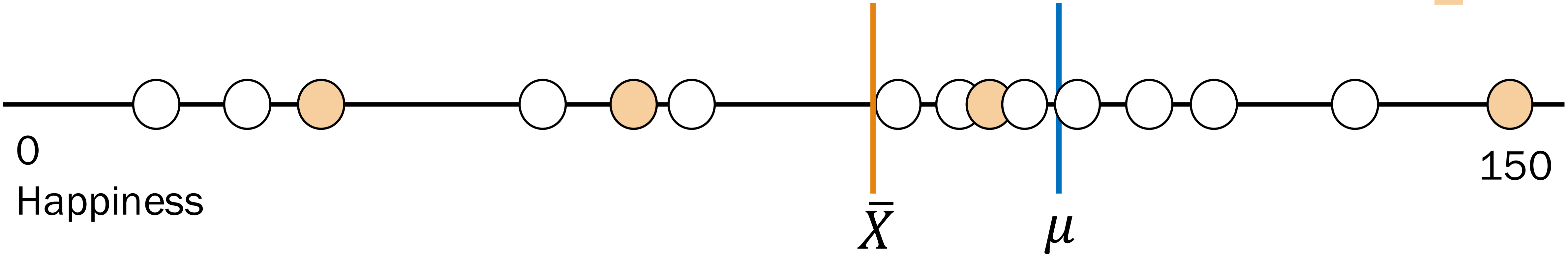
population mean

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

Estimate, S^2

sample mean

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



Population size, N

Intuition about the sample variance, S^2

Actual, σ^2

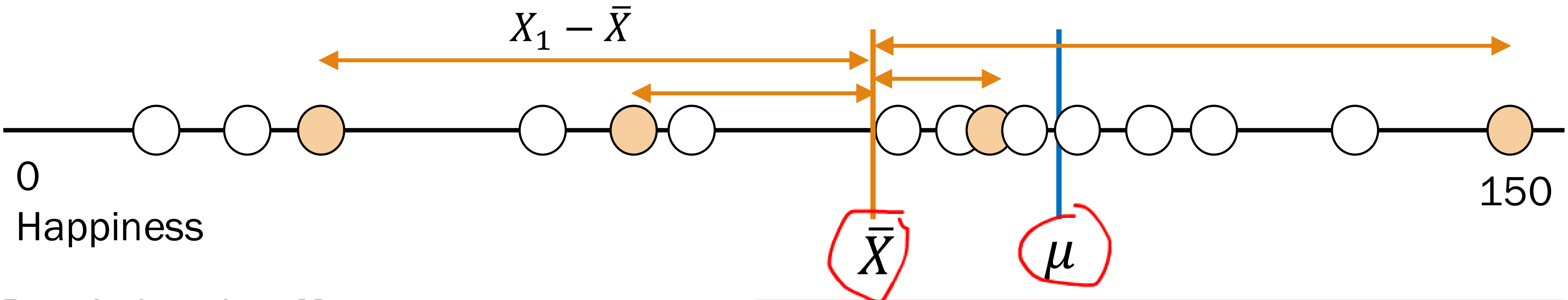
population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overset{\text{population mean}}{\overset{\downarrow}{\mu}})^2$$

Estimate, S^2

sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(\underbrace{X_i - \overset{\text{sample mean}}{\overset{\downarrow}{\bar{X}}}})^2$$



Population size, N

This formula will always underestimate the variance…

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

Estimating the population variance

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

population mean

population variance

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

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

sample mean

sample variance

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

Sample Variance:

Sample mean

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

Makes it “unbiased”

Quick check

A. Random variable(s)
B. Value
C. Event

1. μ, the population mean B

2. $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$, a sample A.

3. σ^2, the population variance B

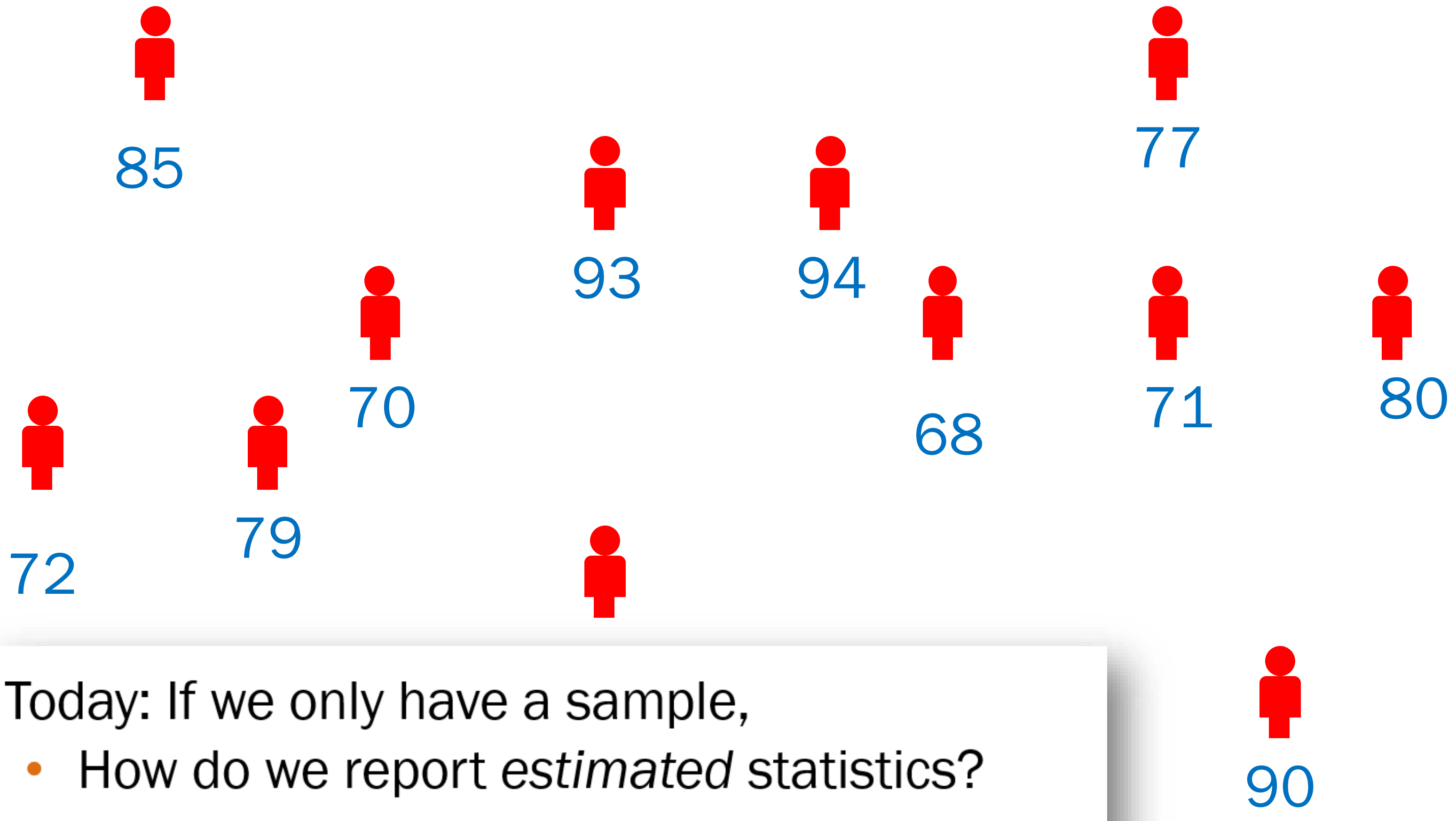
4. $\bar{X}$, the sample mean A.

5. $\bar{X} = 83$ C

6. $(X_1 = 59, X_2 = 87, X_3 = 94, X_4 = 99,$
   $X_5 = 87, X_6 = 78, X_7 = 69, X_8 = 91)$ C

Sample

sample = [72, 85, 79, 91, 68, ..., 71]



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?

Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

= 83

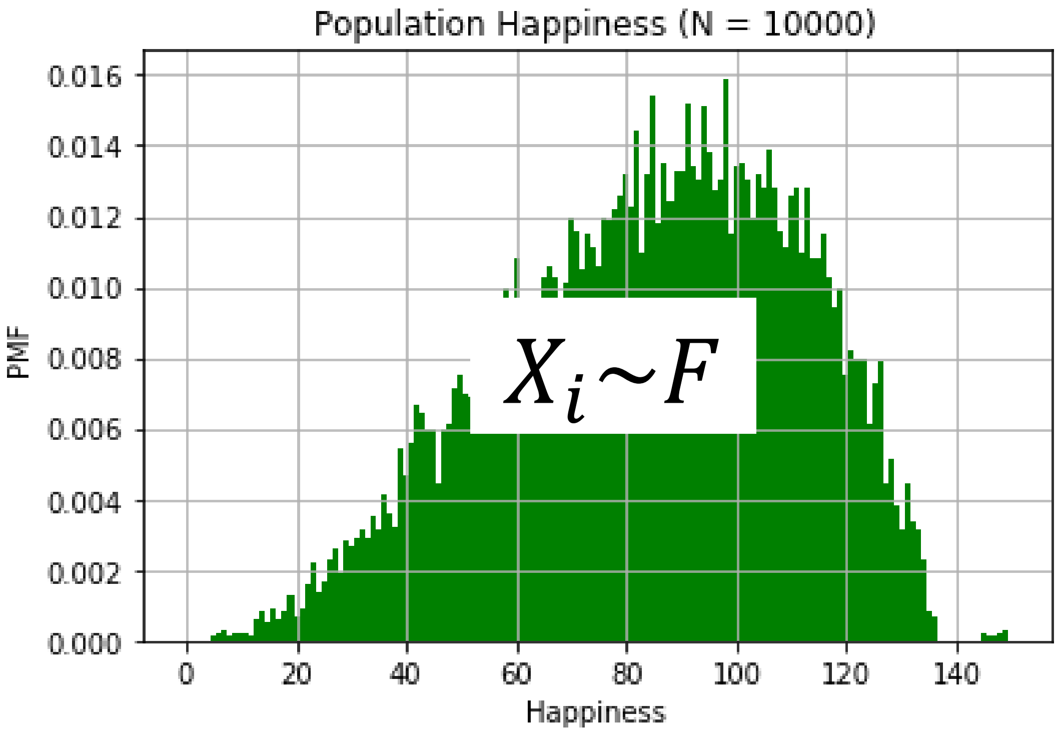
Sample Variance

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

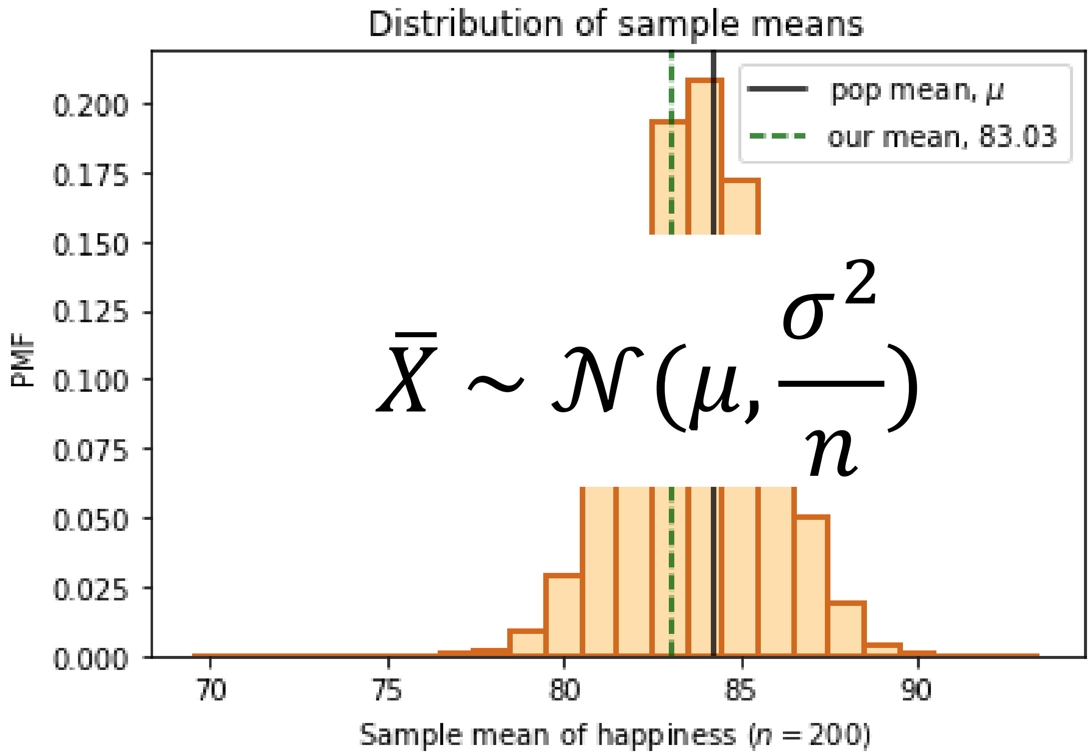
= 40

Reporting estimates

Sample mean



$X_i \sim F$

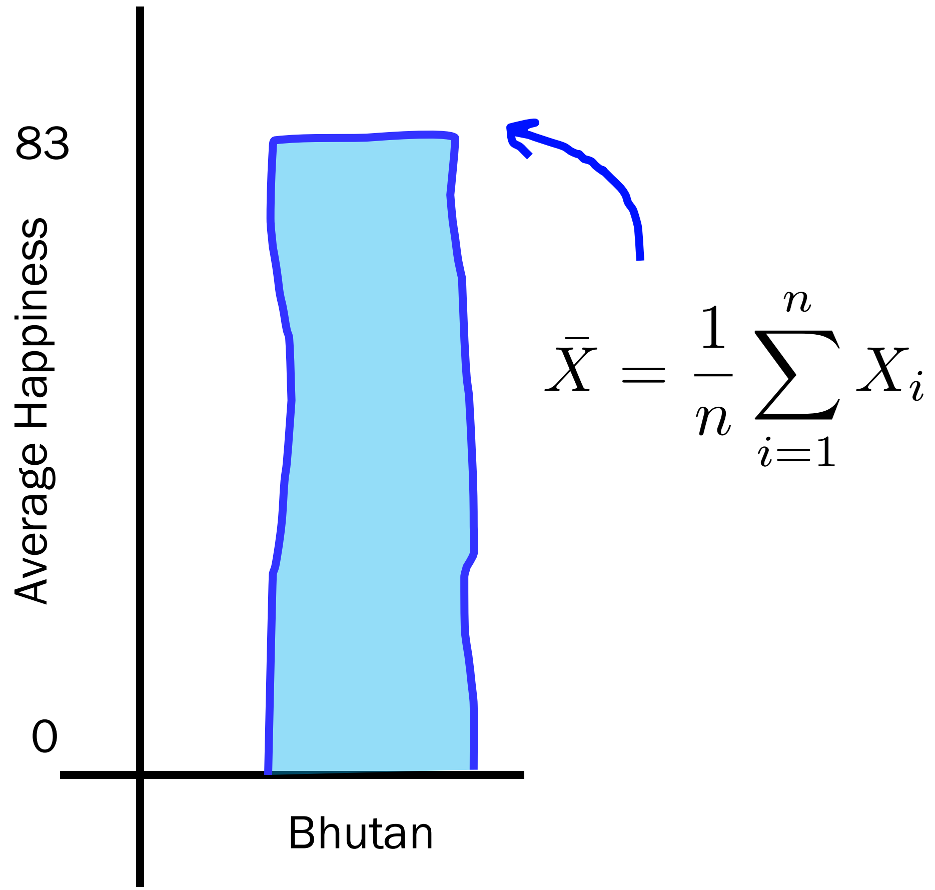


$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

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

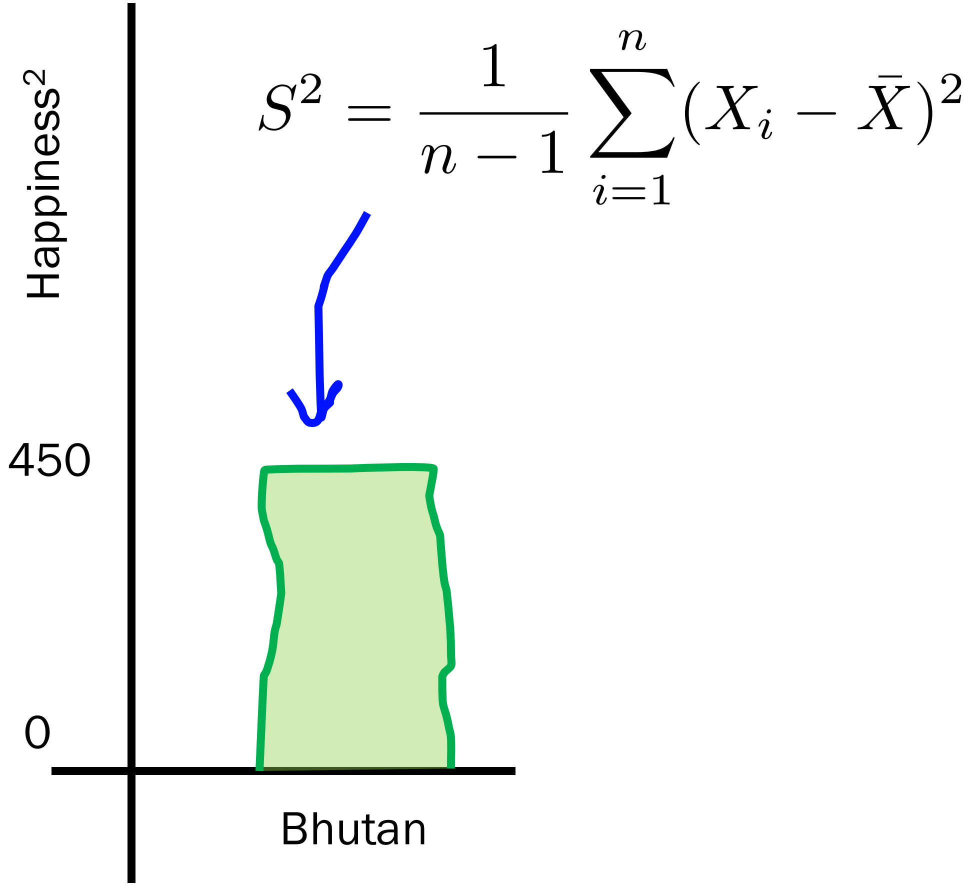
Our Report to Bhutan Government (after talking to 200 ppl)

Average Happiness



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

Variance of Happiness



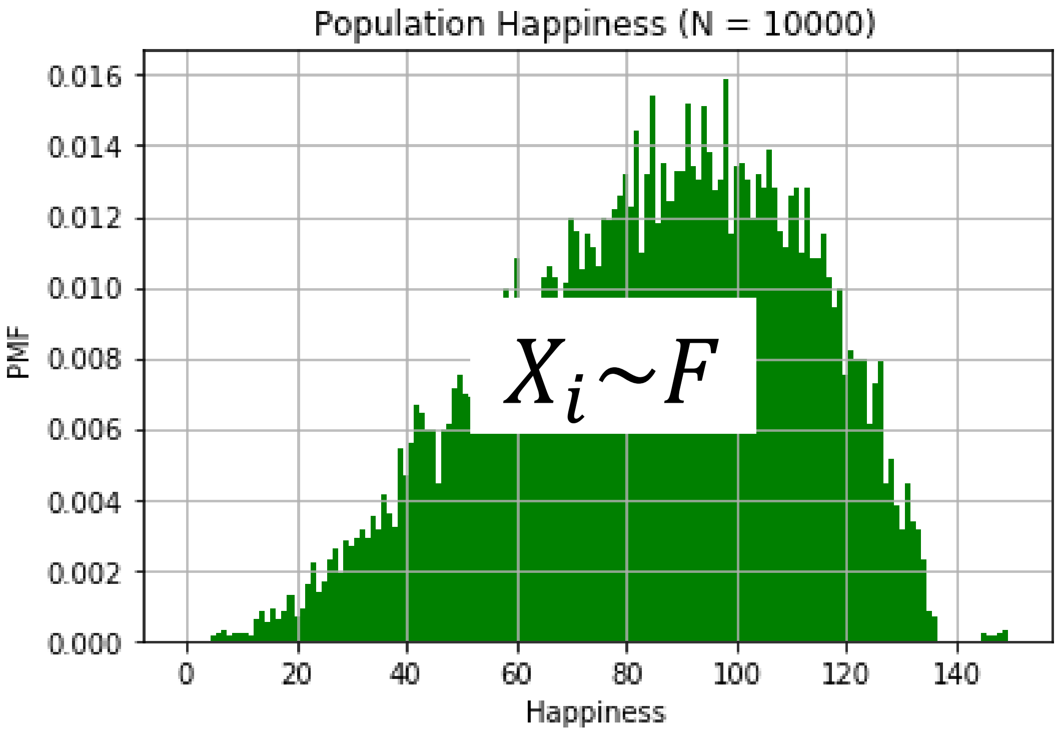
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

No Error Bars ☹

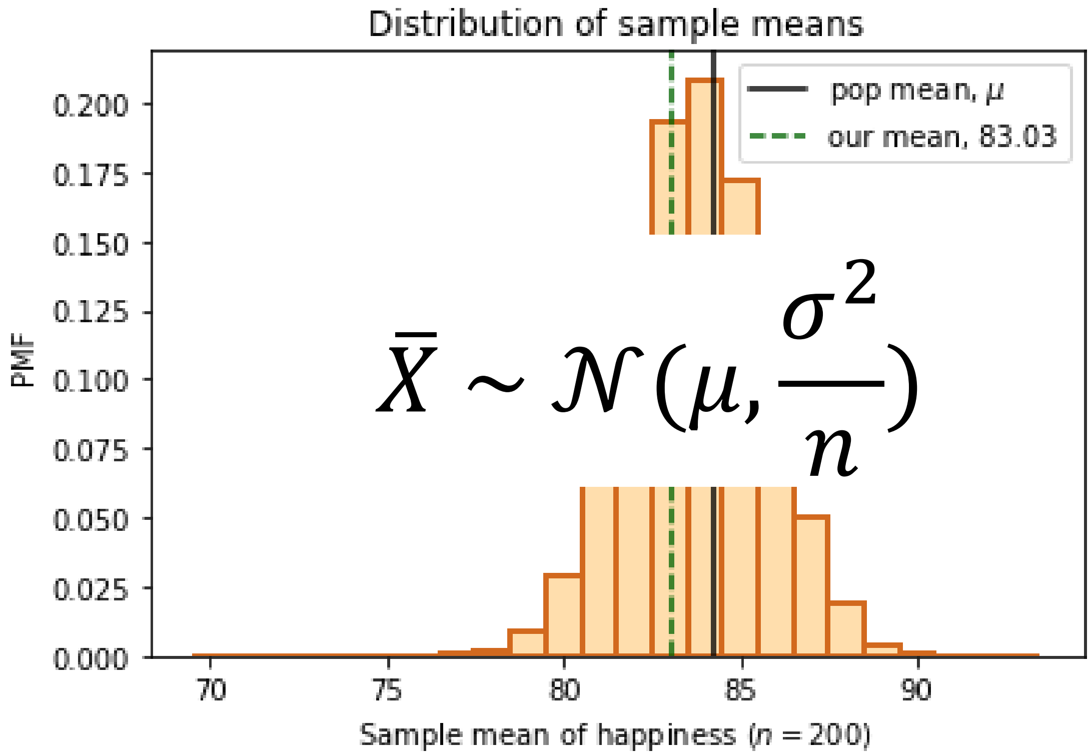
19c_standard_error

Standard error of the mean

Sample mean



$X_i \sim F$



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

- Var$(\bar{X})$ is a measure of how “close” $\bar{X}$ is to μ.
- **How do we estimate Var$(\bar{X})$?**

Standard Error of the Mean

$$E[\bar{X}] = \mu$$

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

We want to estimate this

def The **standard error** of the mean is an estimate of the standard deviation of $\bar{X}$.

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

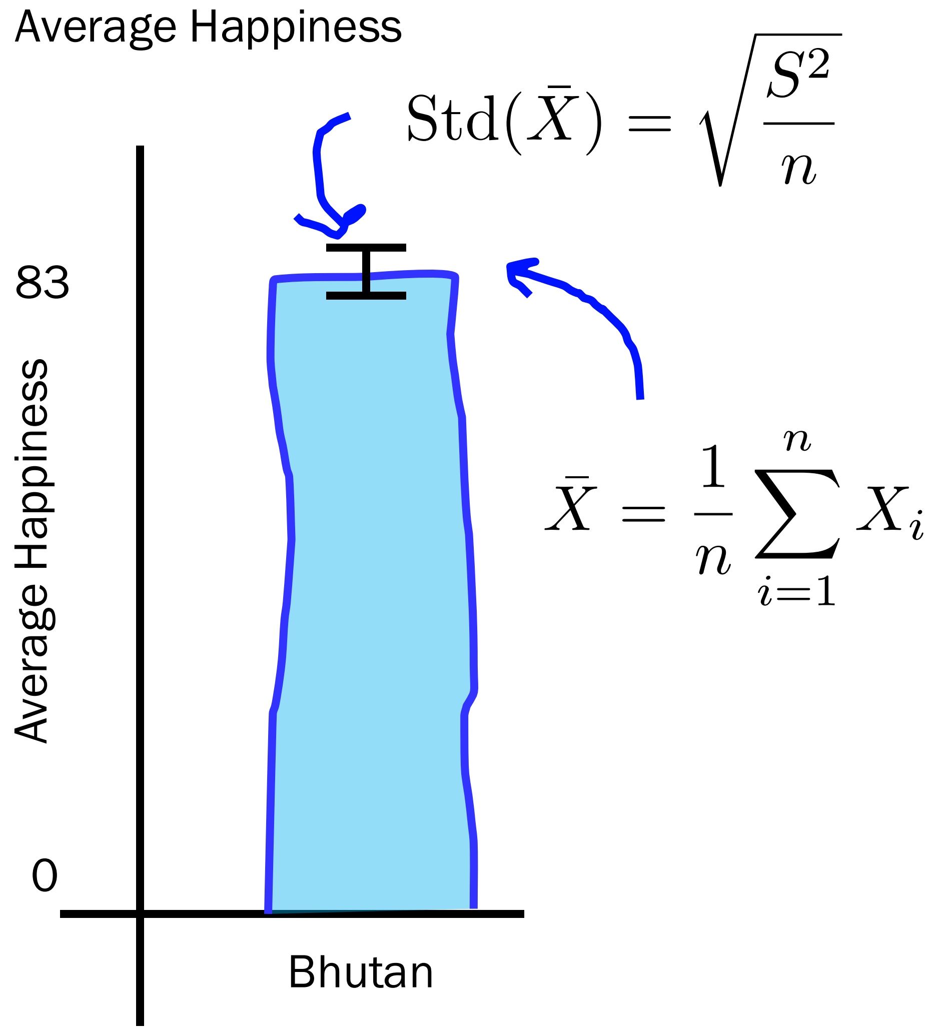
Intuition:

- S^2 is an unbiased estimate of σ^2
- S^2/n is an unbiased estimate of $\sigma^2/n = \text{Var}(\bar{X})$
- $\sqrt{S^2/n}$ can estimate $\sqrt{\text{Var}(\bar{X})}$

More info on bias of standard error: wikipedia

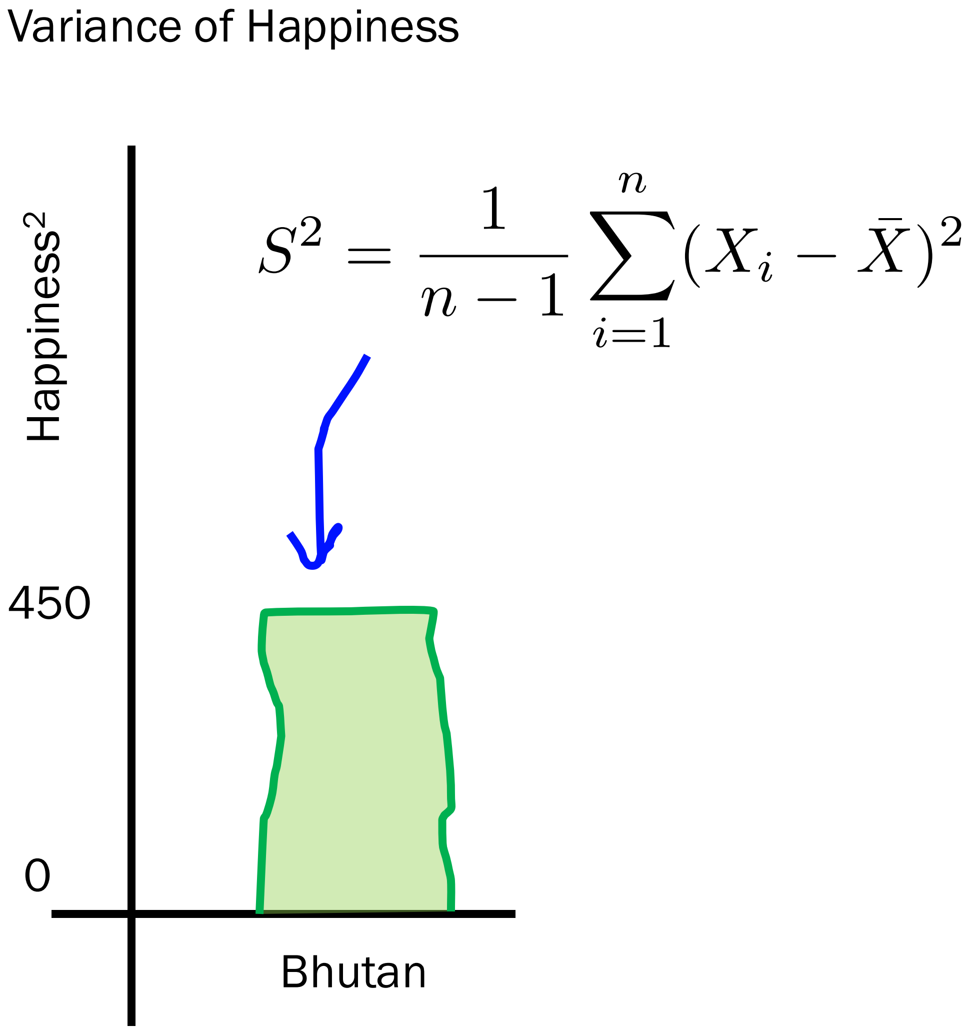
But what about **error bars**???

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



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

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



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

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

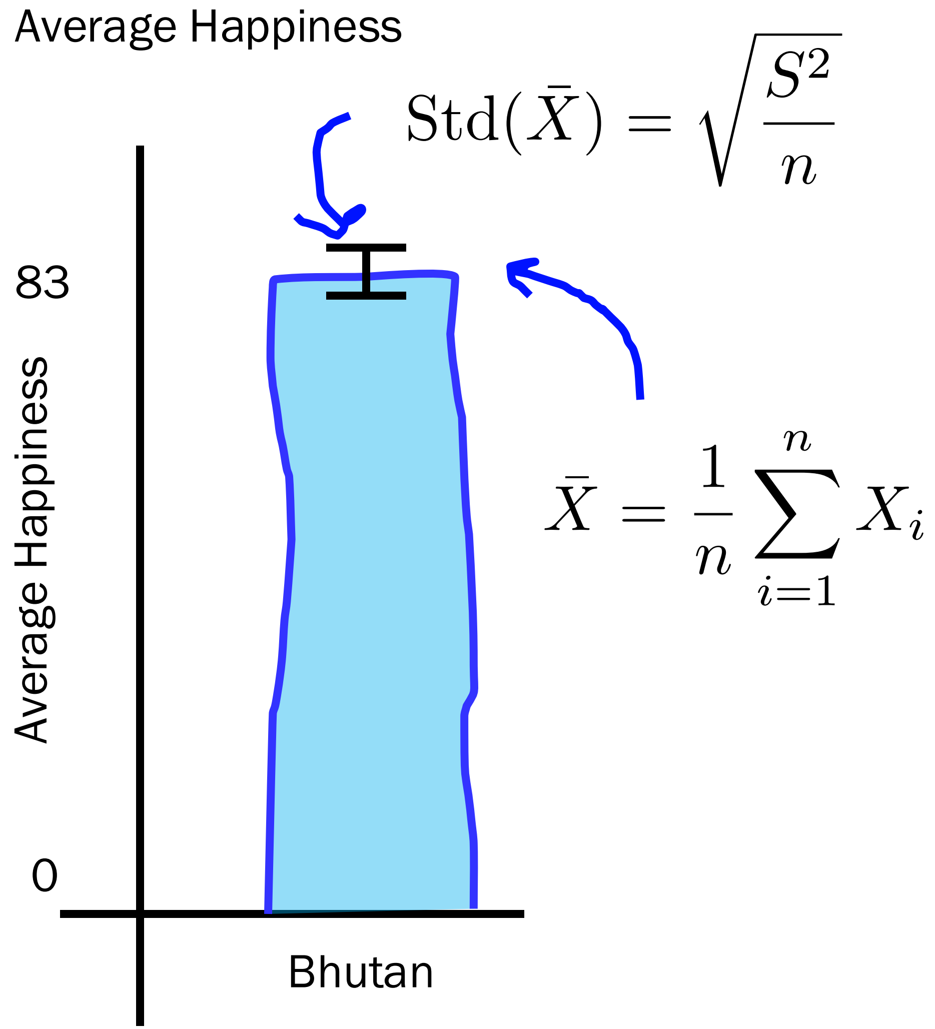
$$\mathrm{Std}(\bar{X}) = \sqrt{\frac{S^2}{n}}$$

sample variance

How wrong do we think our mean estimate is?

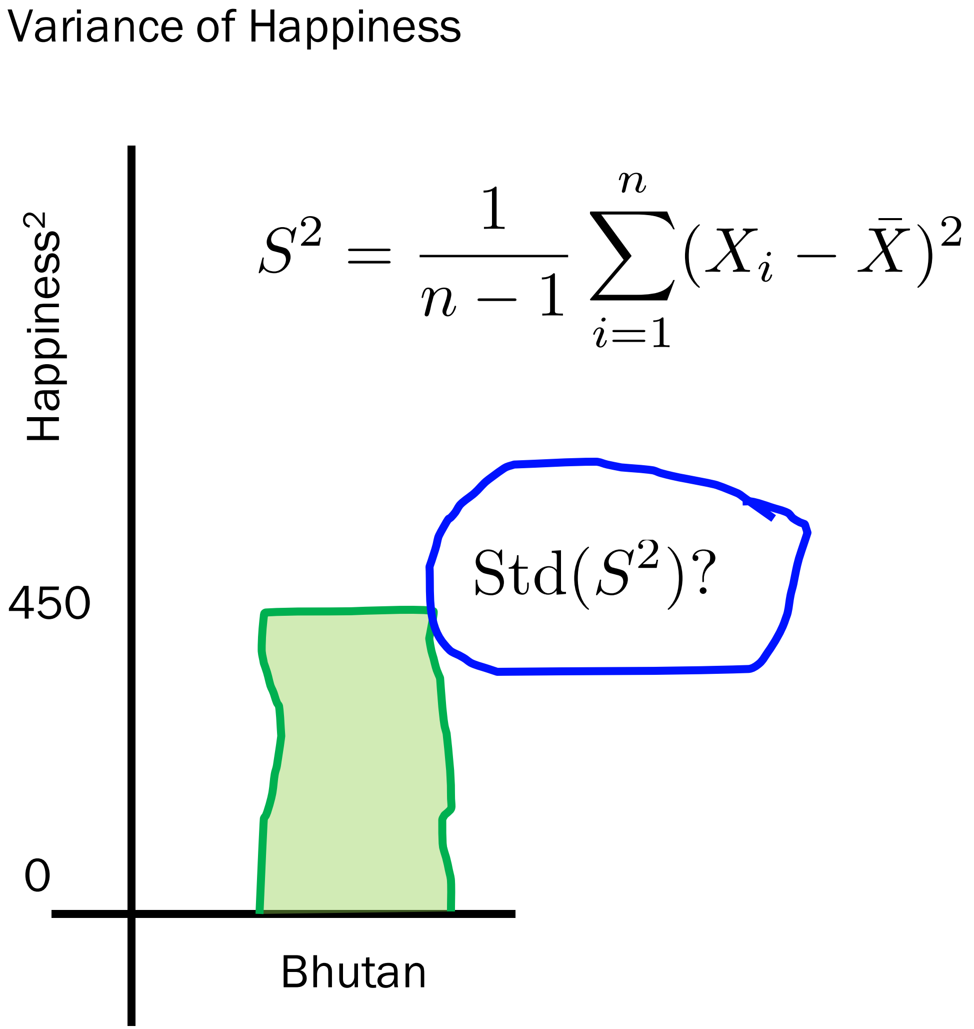
But what about **error bars**???

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



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

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



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

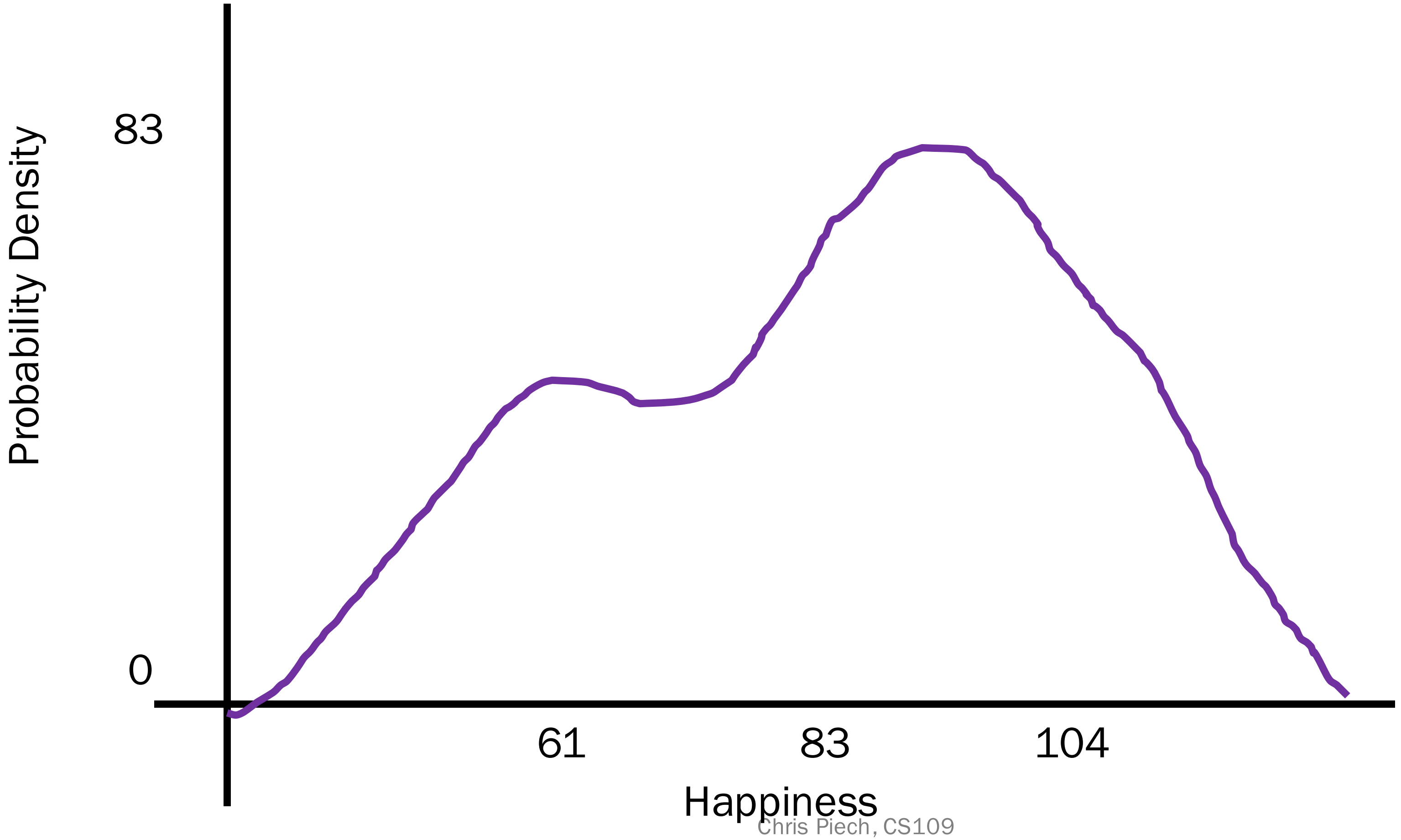
$\text{Std}(S^2)$?

Claim: The average happiness of Bhutan is 83 ± 1.5

Hypothetical

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

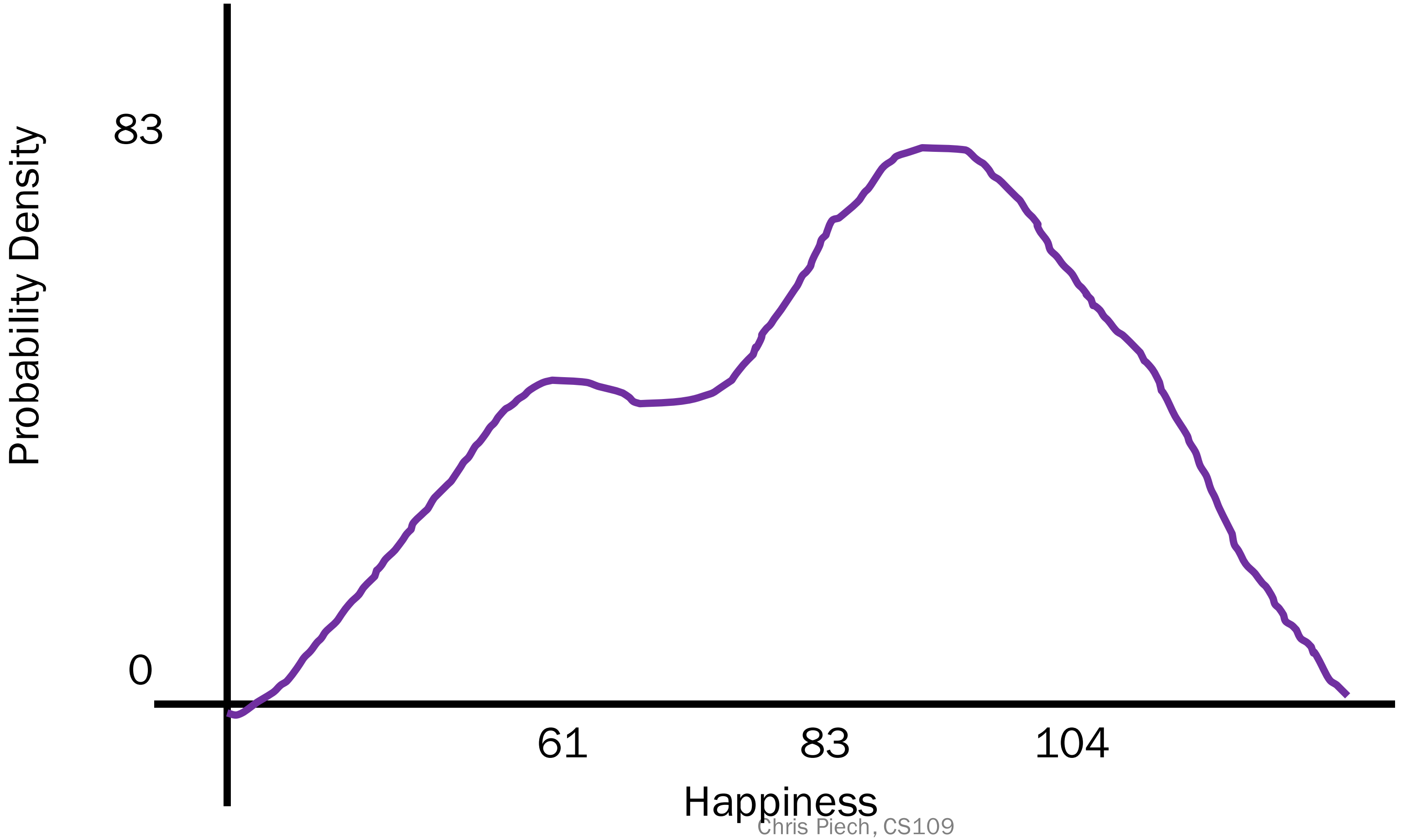
Plot twist: I give you the ***entire*** underlying distribution



Hypothetical

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

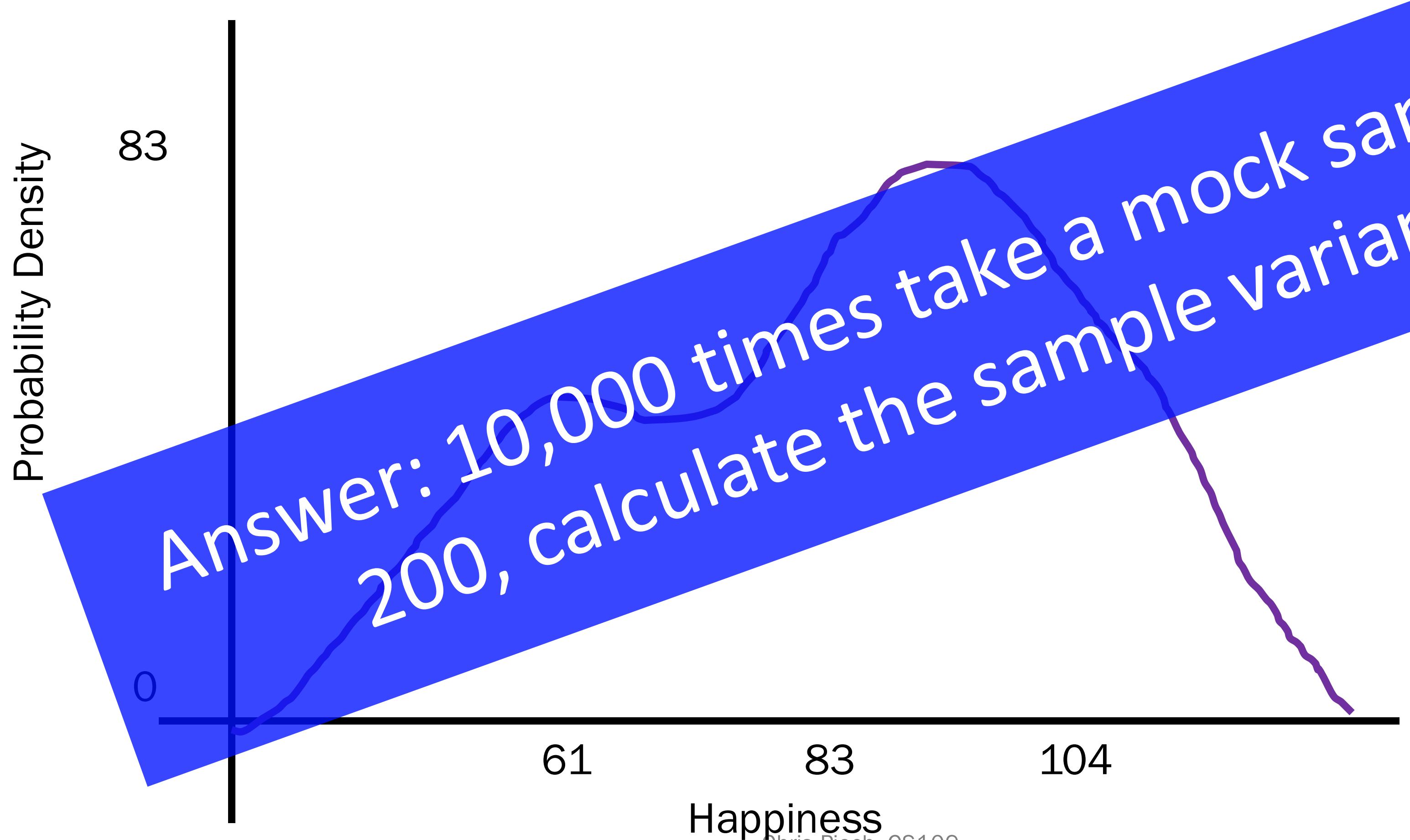
Plot twist: I give you the ***entire*** underlying distribution



Hypothetical

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

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



Answer: 10,000 times take a mock sample of 200, calculate the sample variance

Hypothetical

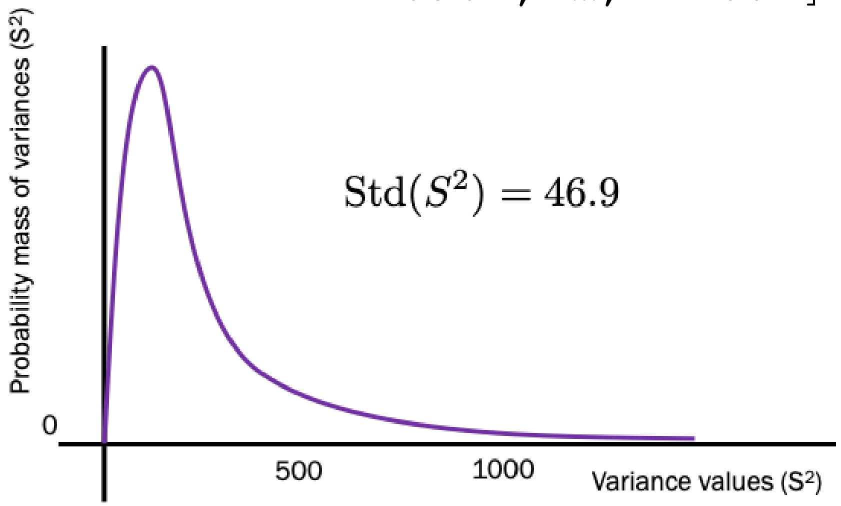
What is the **std** of the **sample variance**, calculated from 200 people?

```
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

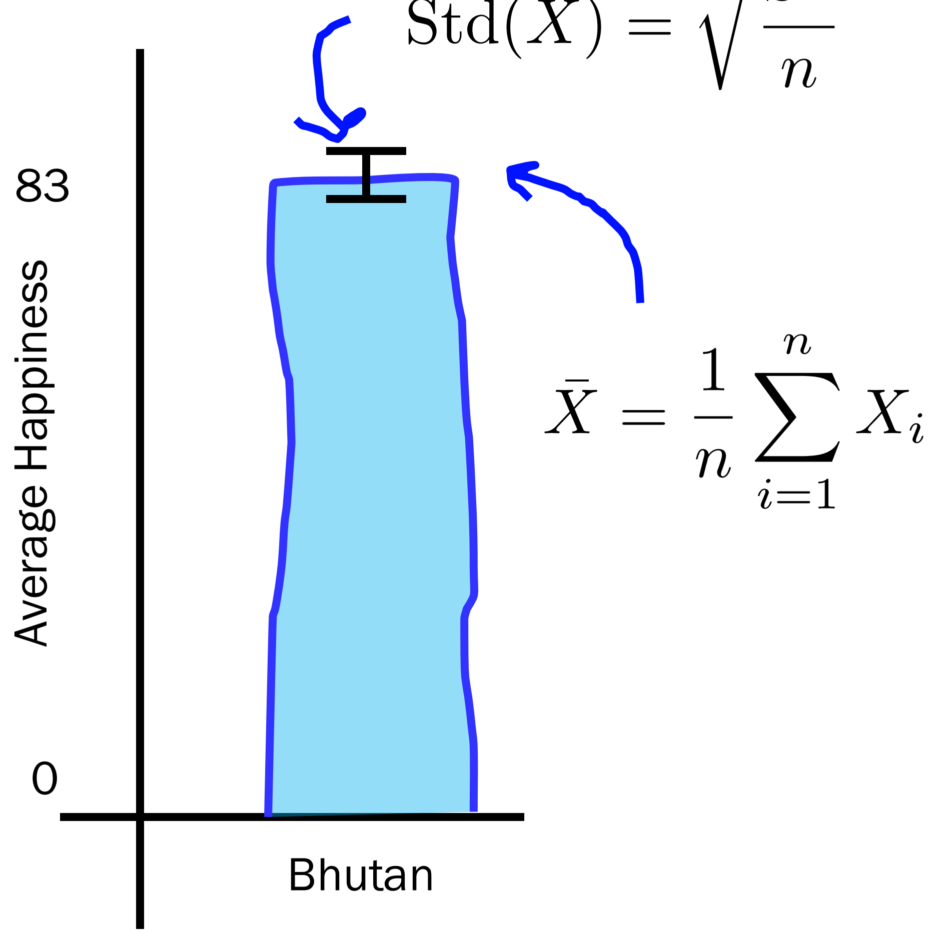
Bootstrapping allows you to:

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

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

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

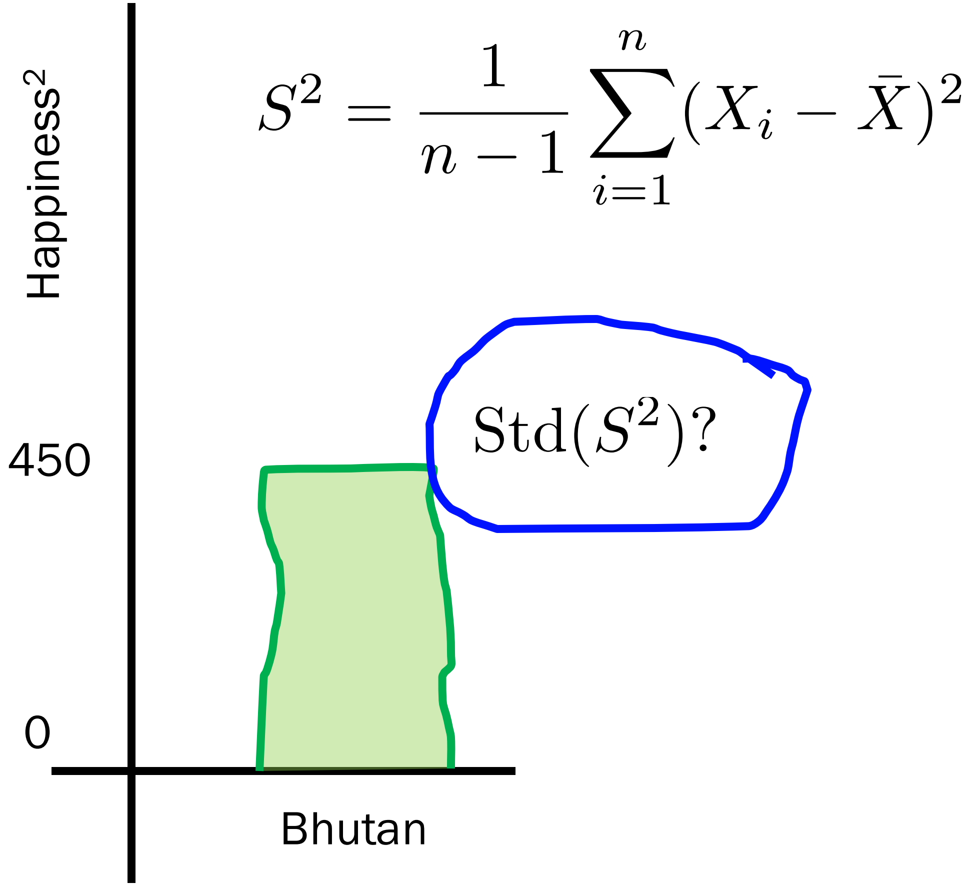
Average Happiness



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

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

Variance of Happiness



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

$\text{Std}(S^2)$?

Hypothetical – You have the underlying distribution!

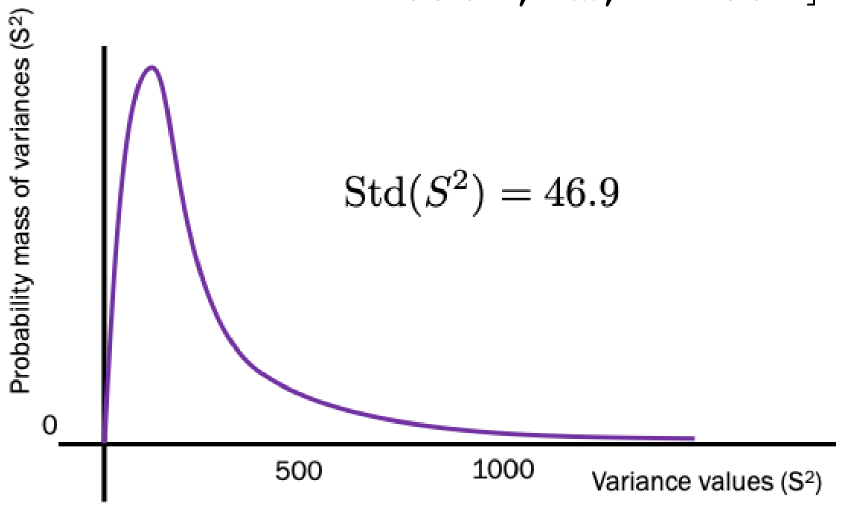
What is the **std** of the **sample variance**, calculated from 200 people?

```
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
```

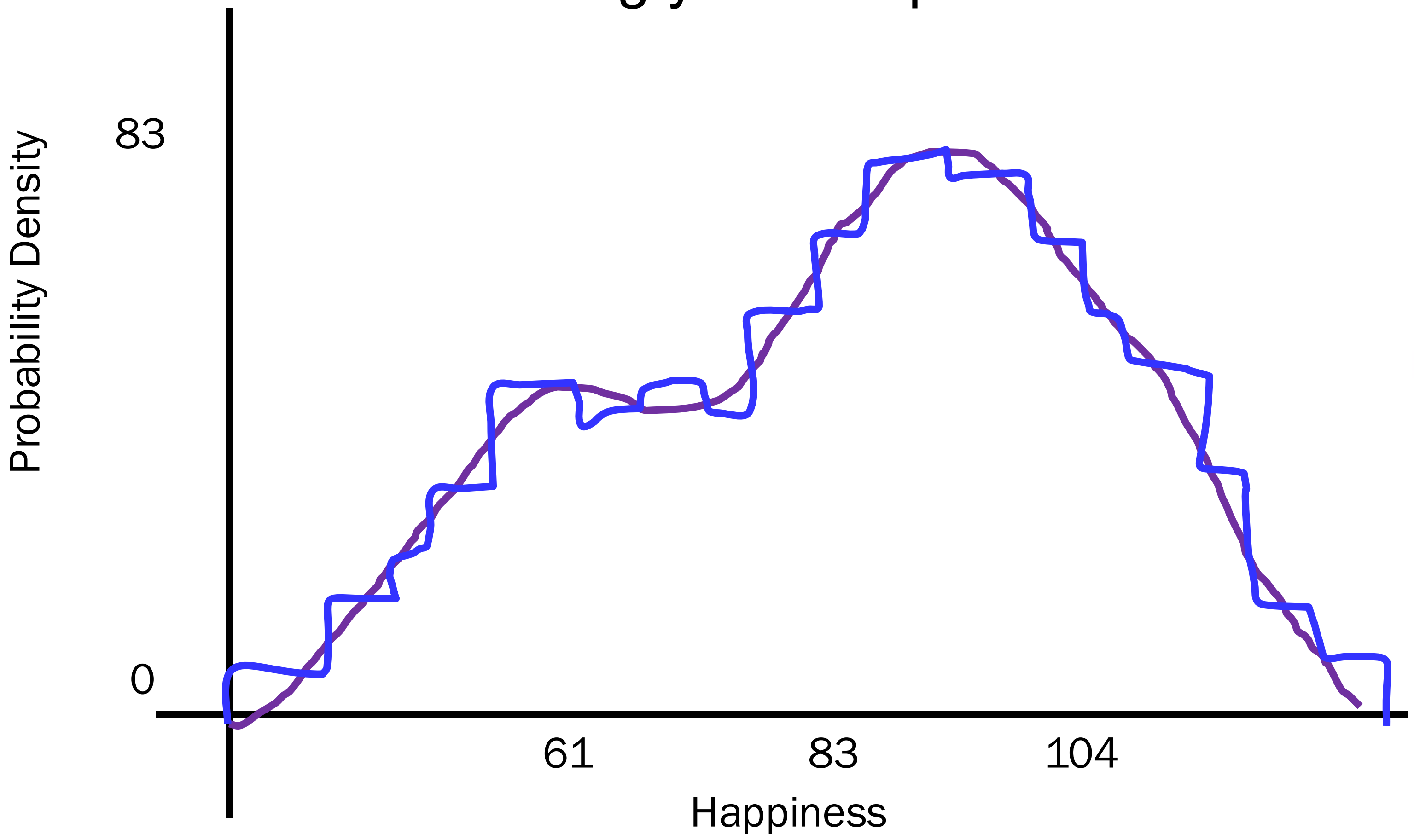
```
sample_vars = [472.7, 478.4,
               469.2, …, 476.2]
```



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.*



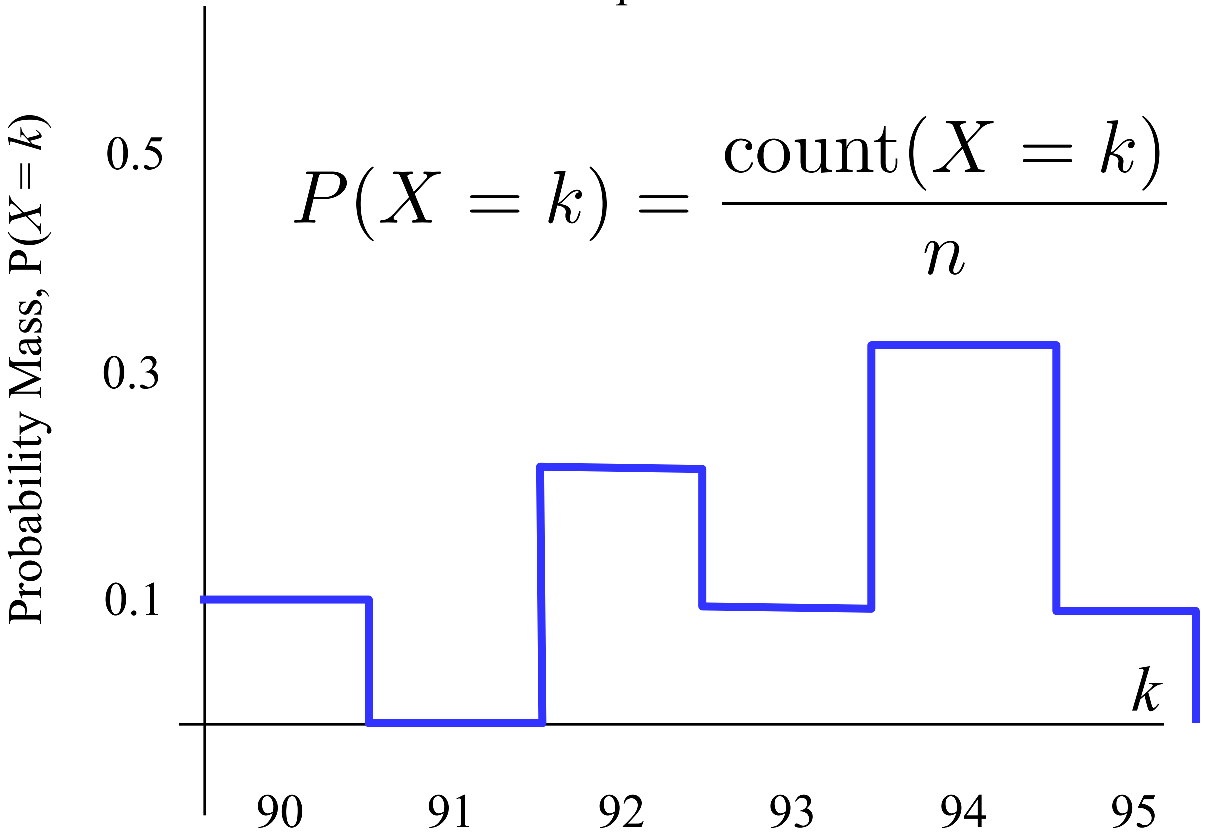
* This is just a histogram of your data!!

Key Insight

IID Samples

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

Sample Distribution



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

Bootstrapping Assumption

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

The underlying distribution

The sample distribution

(aka the histogram of your data)

Algorithm

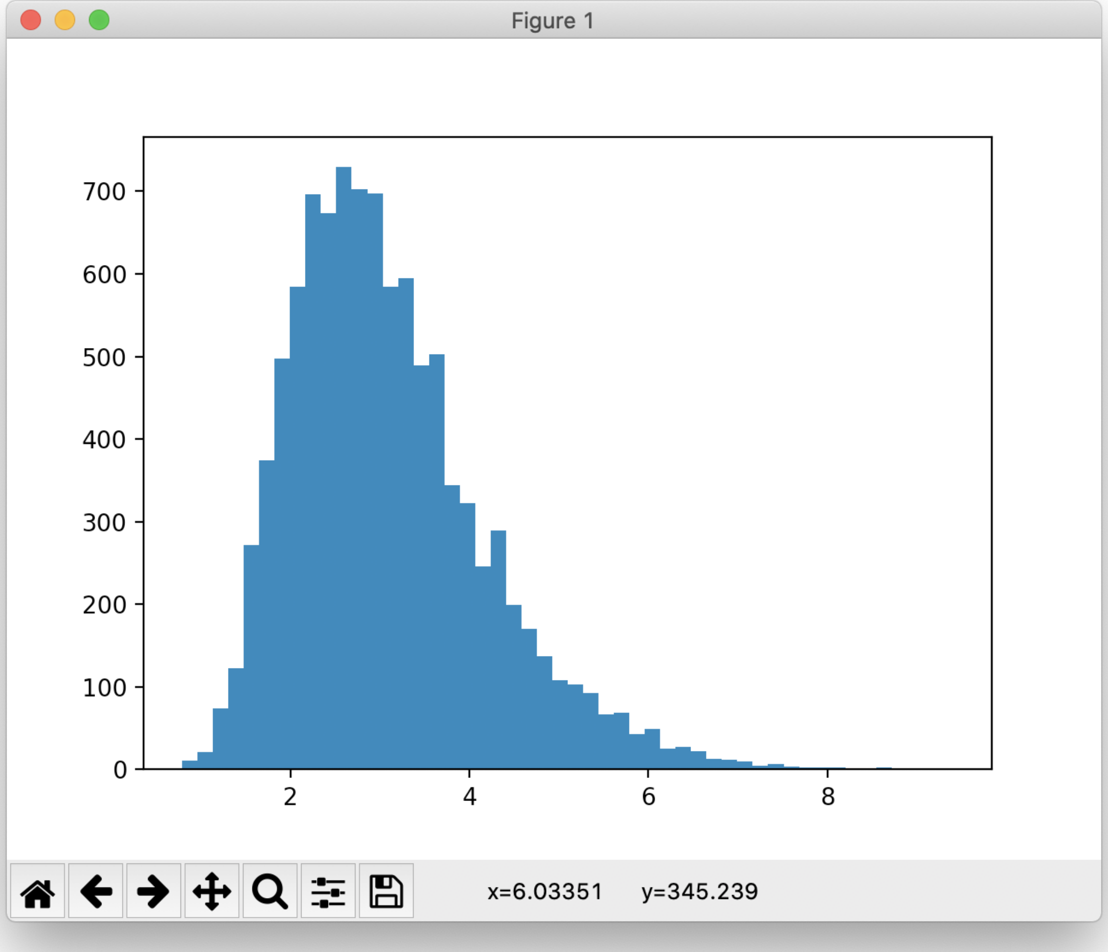
```
Bootstrap Algorithm (sample):
  Estimate the PMF using the sample
  Repeat 10,000 times:
    a. Draw len(sample) new samples from PMF
    b. Recalculate the stat on the resample
  You now have a distribution of your stat
```

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
```

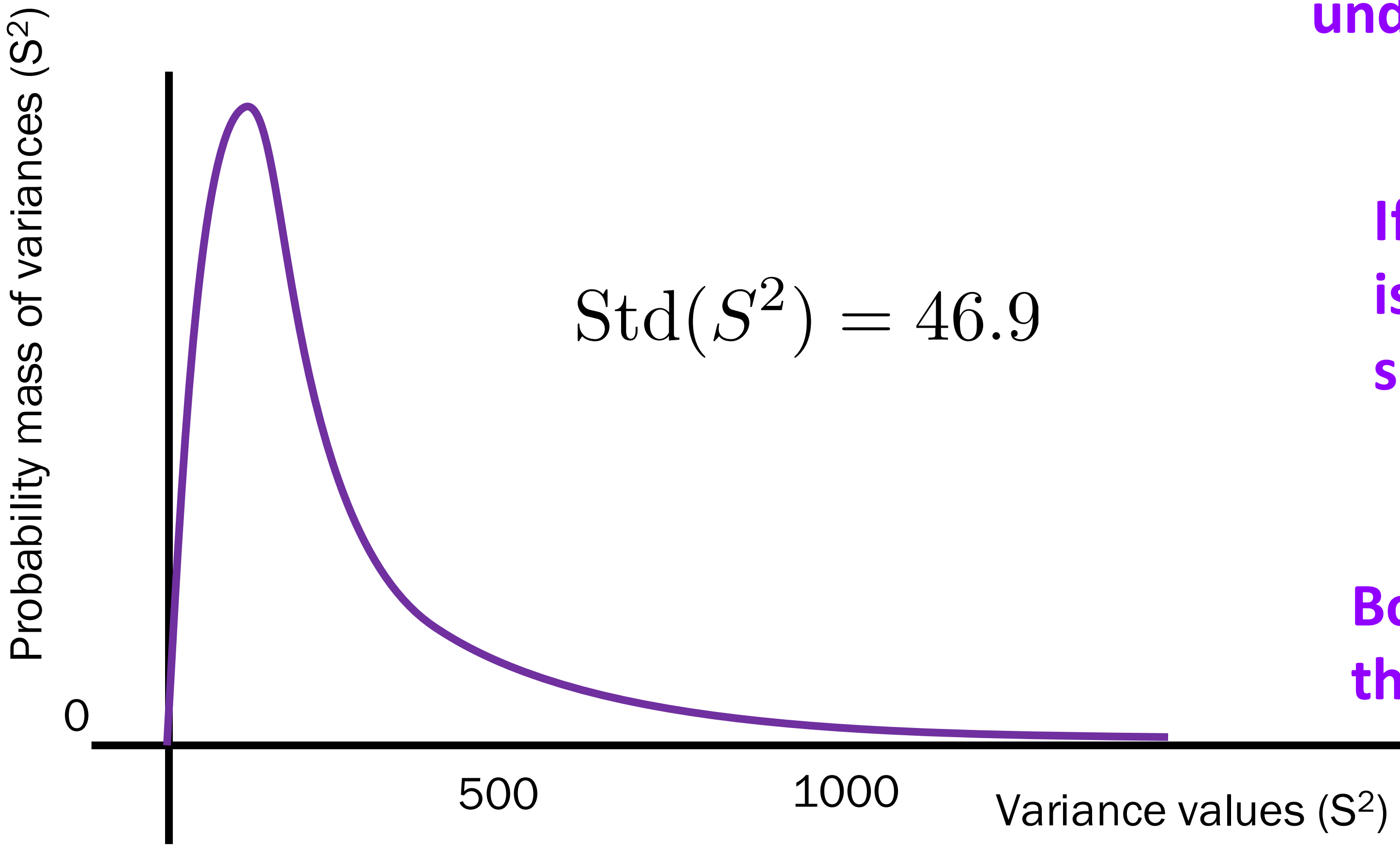
To the code!

The Distribution of the Sampling Variance



Bootstrapping of Variance

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



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

Aside: the distribution of variance depends on the underlying distribution

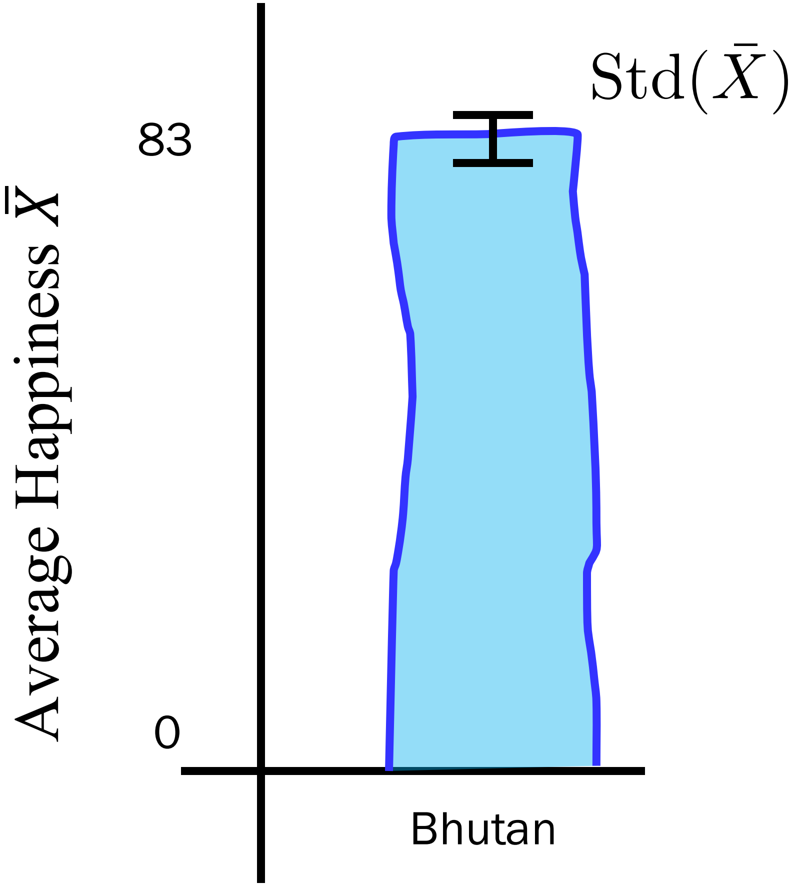
↓

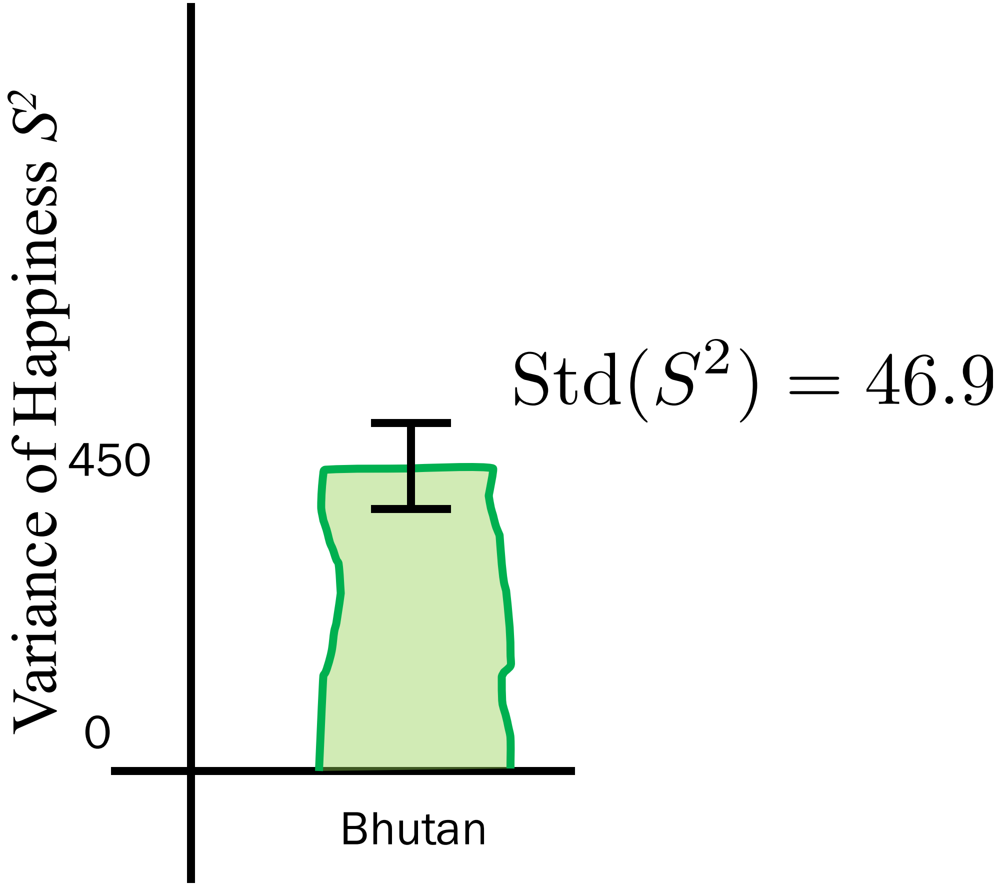
If the underlying distribution is Gaussian, variance is "chi-squared"

↓

Bootstrapping doesn't need to kno that…

Our Report to Bhutan Government





Claim: The average happiness of Bhutan is 83 ± 2

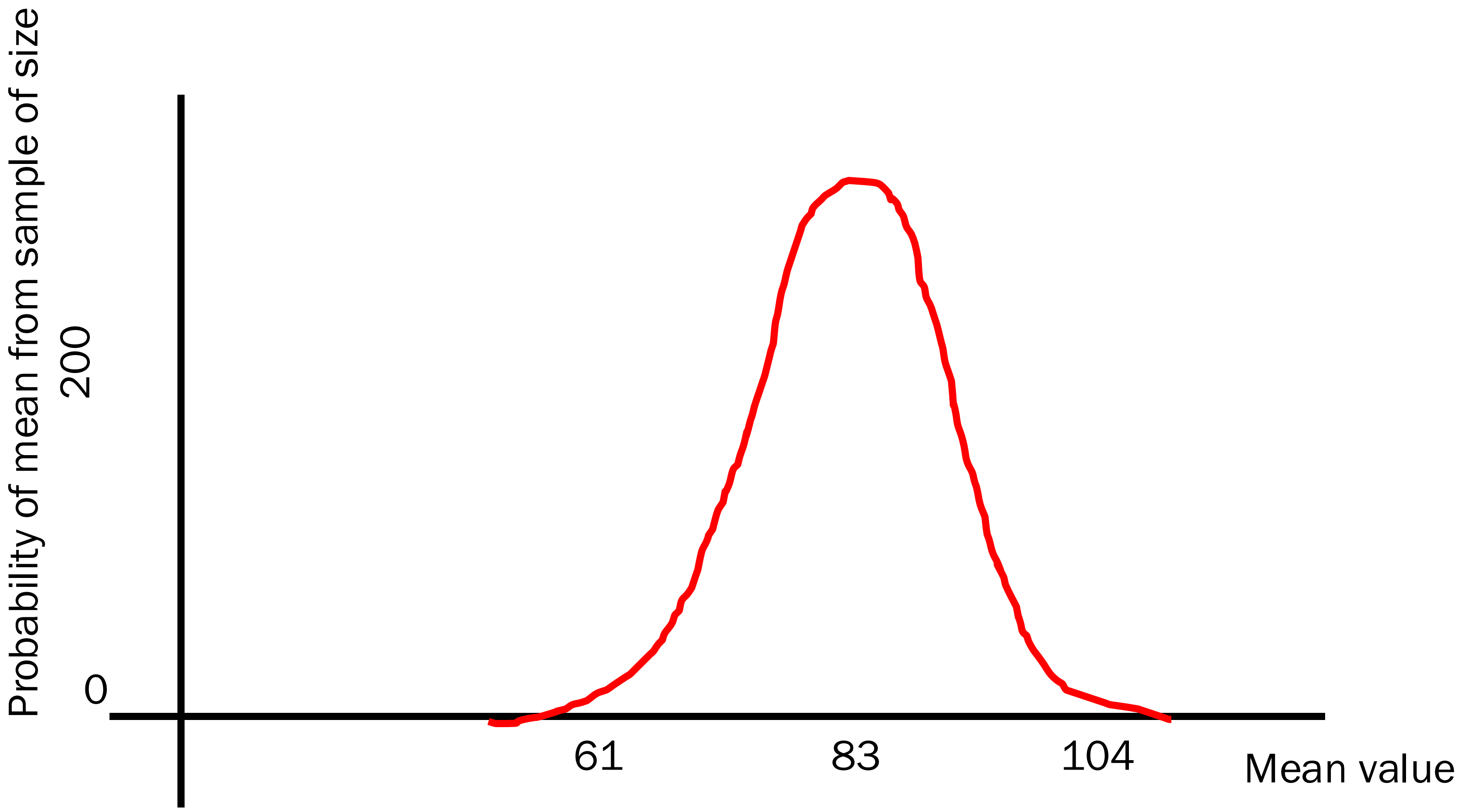
Validation with Sample Mean

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
```

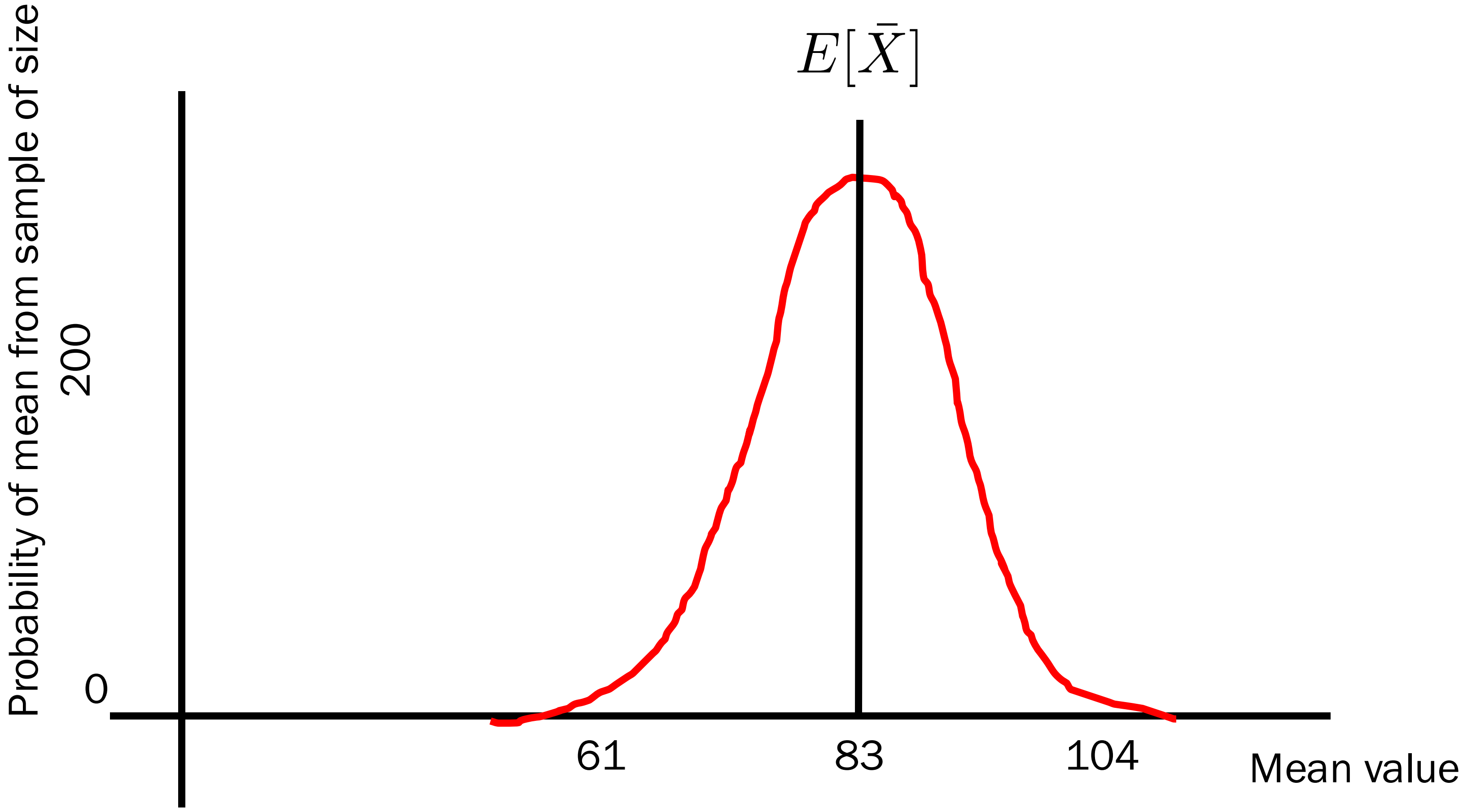
Bootstrapping of Means

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



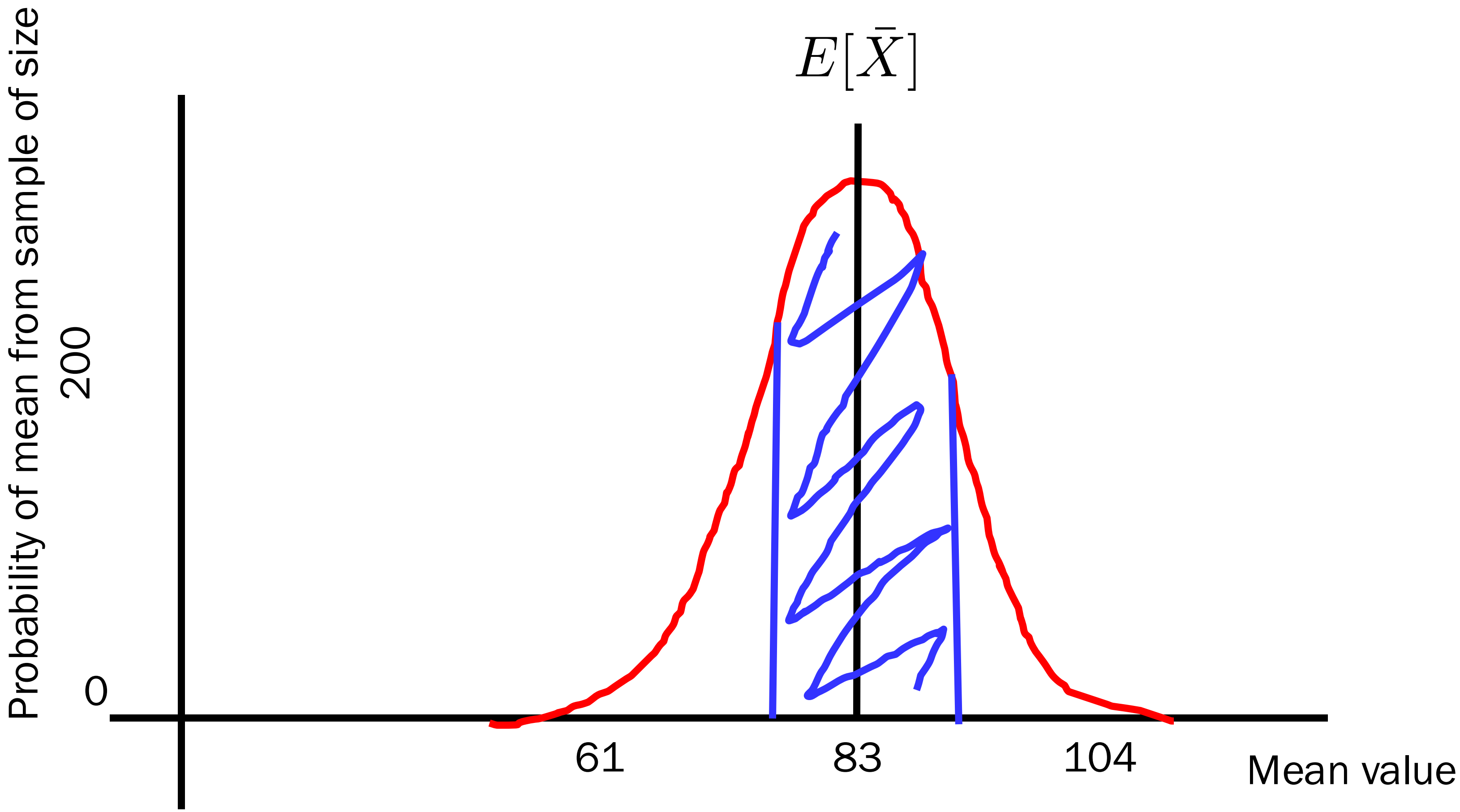
Bootstrapping of Means

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?

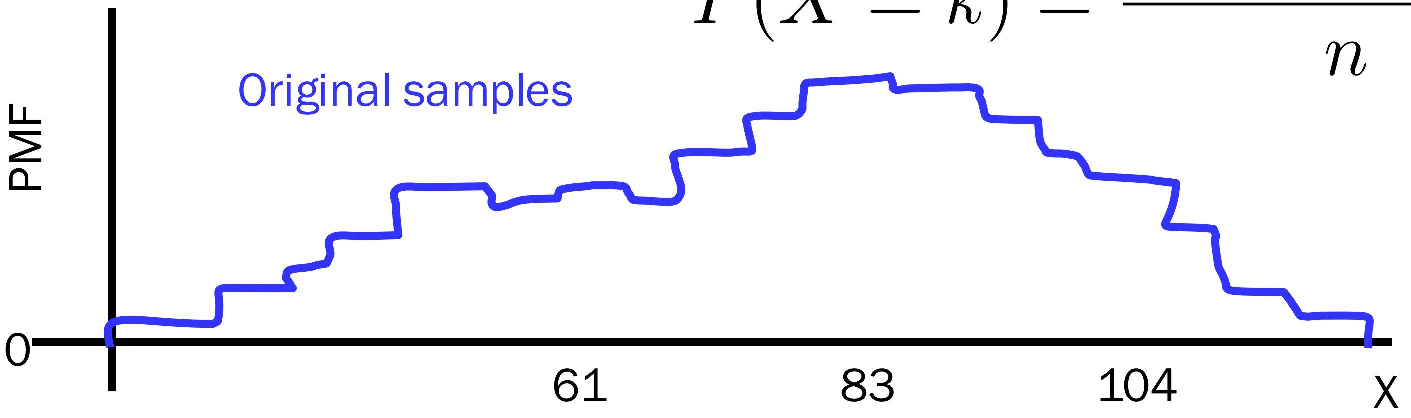


Ok Good!

Resampling in Bootstrapping

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

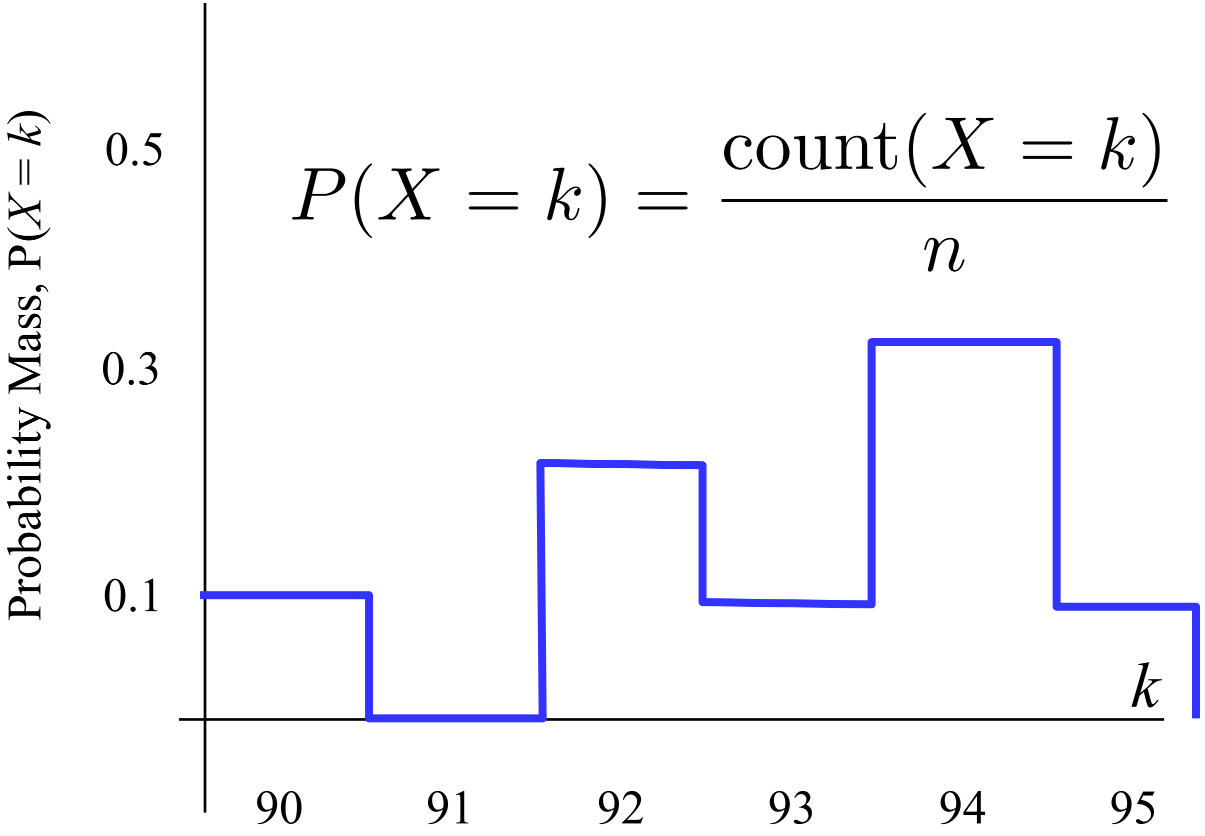
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$



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

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

Resample:

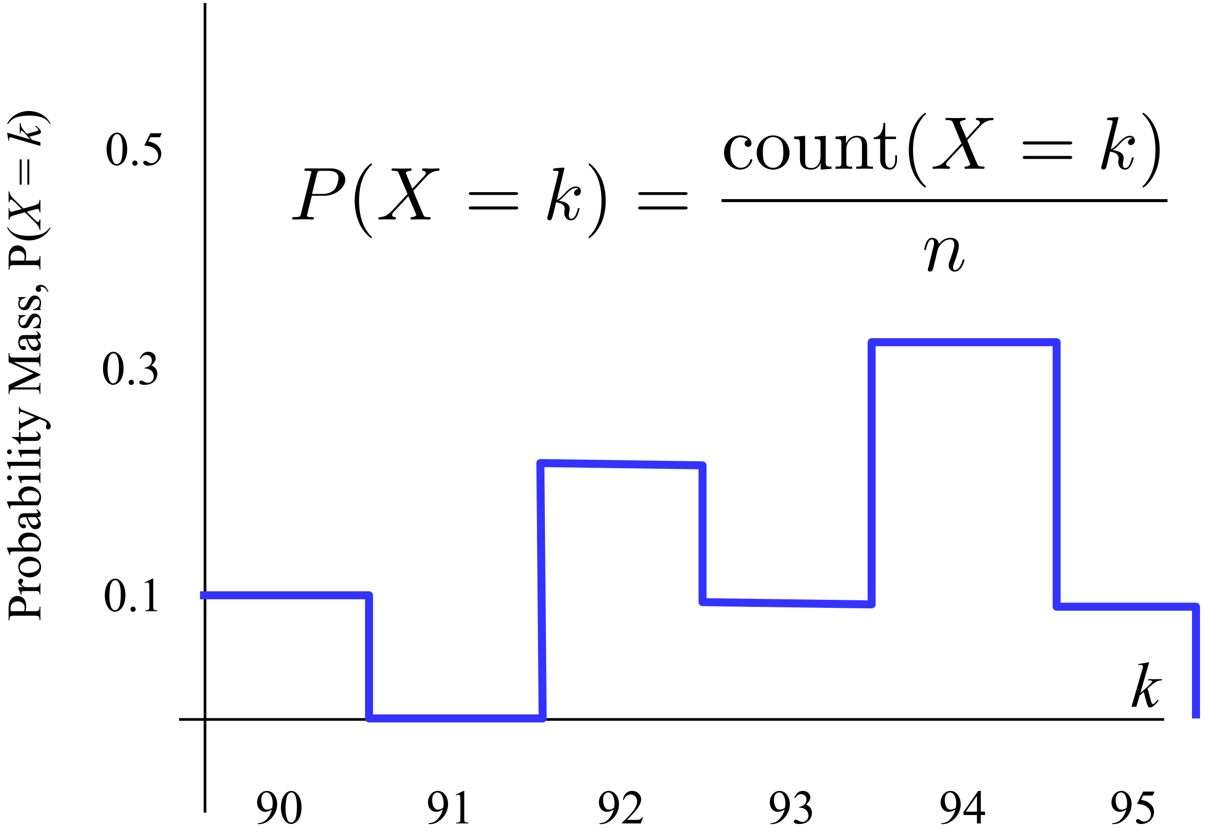


$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:



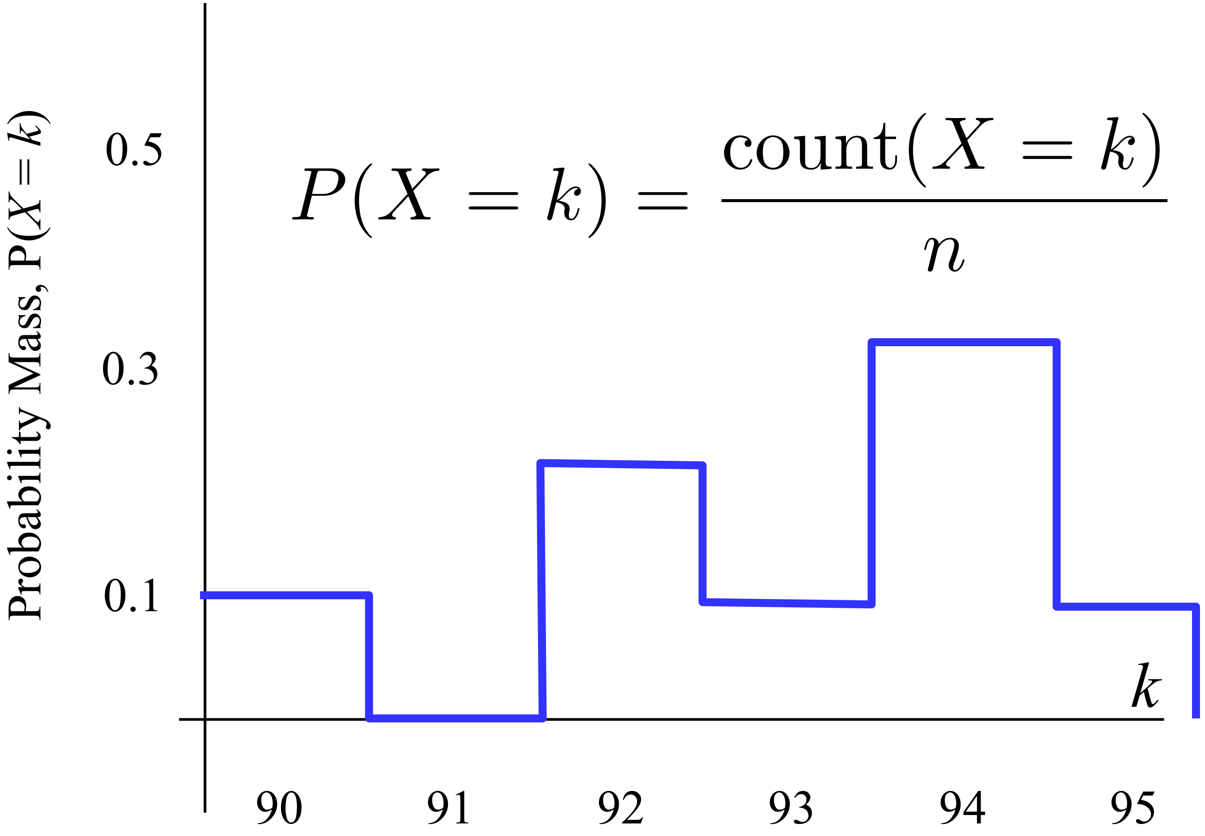
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[**94**]

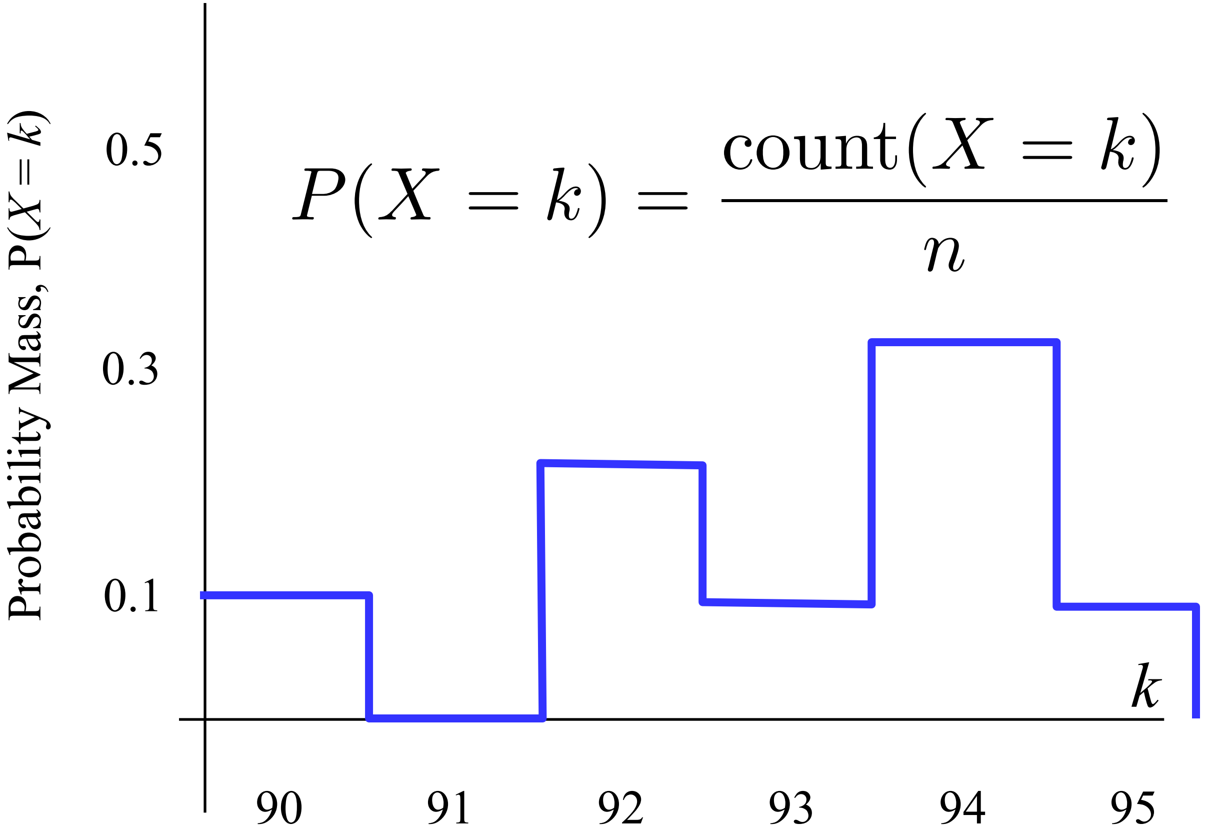


np.random.choice(samples, K, replace = True)

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

Resample:

[94]



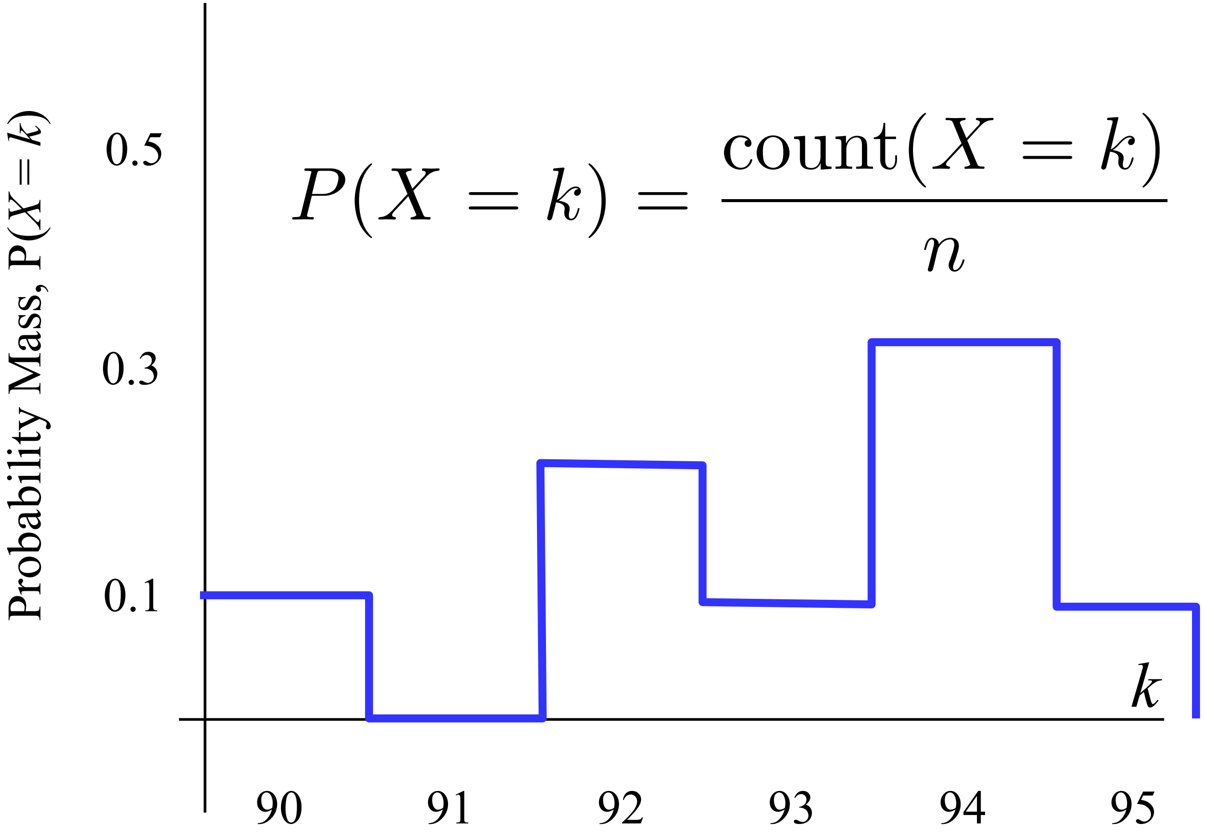
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[94]



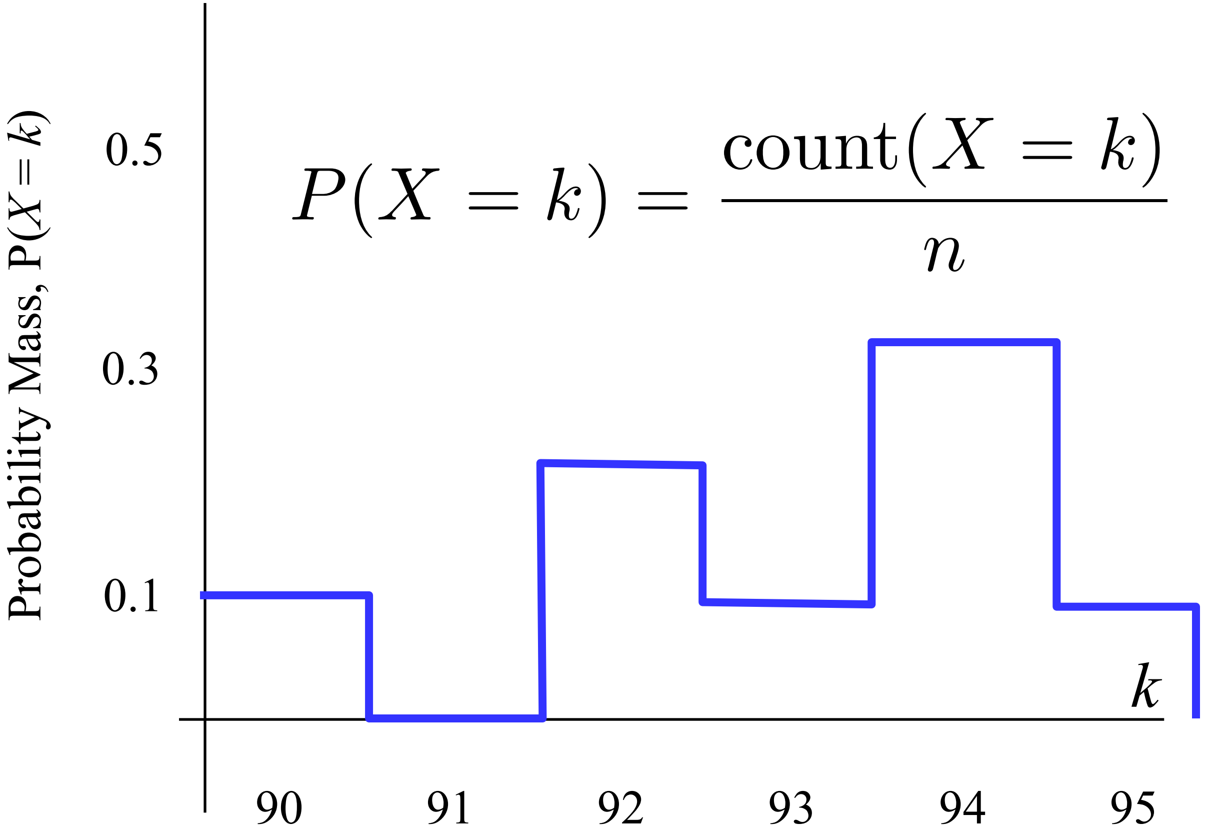
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[94, **90**]



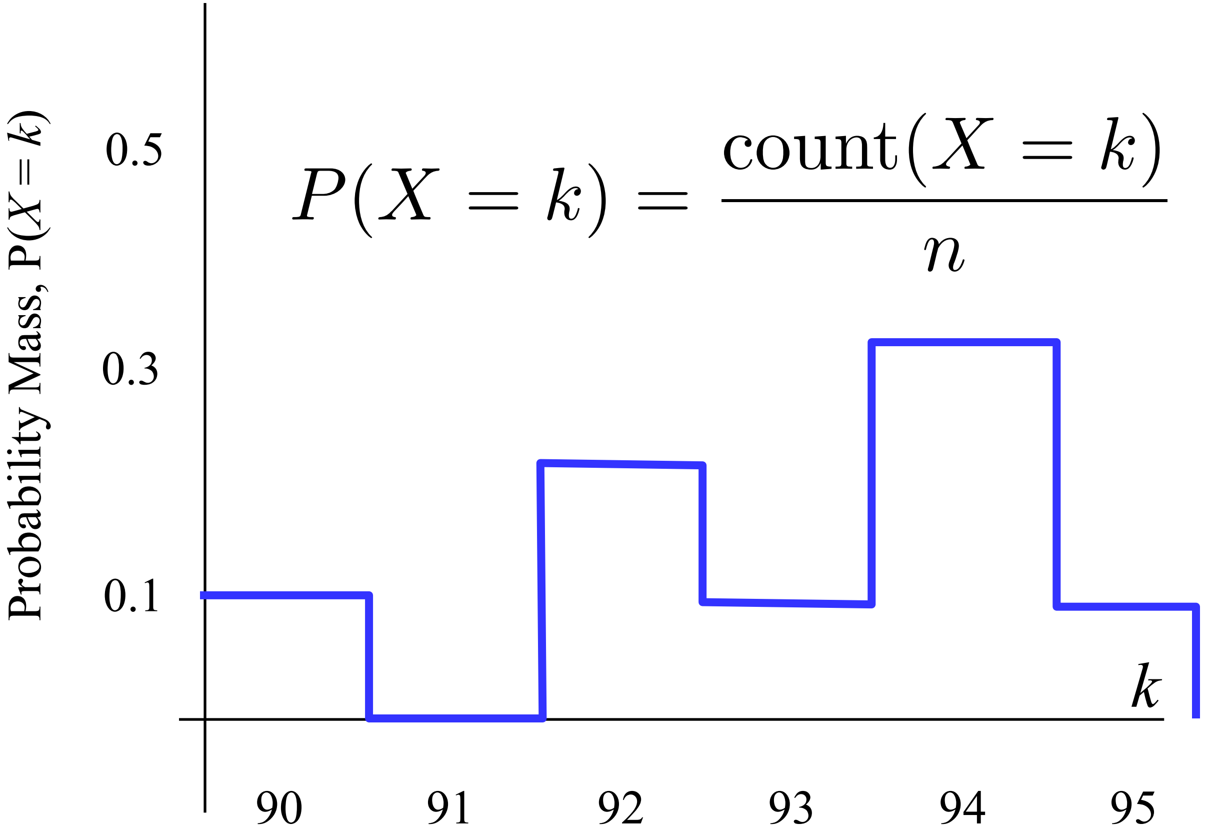
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[94, 90]



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

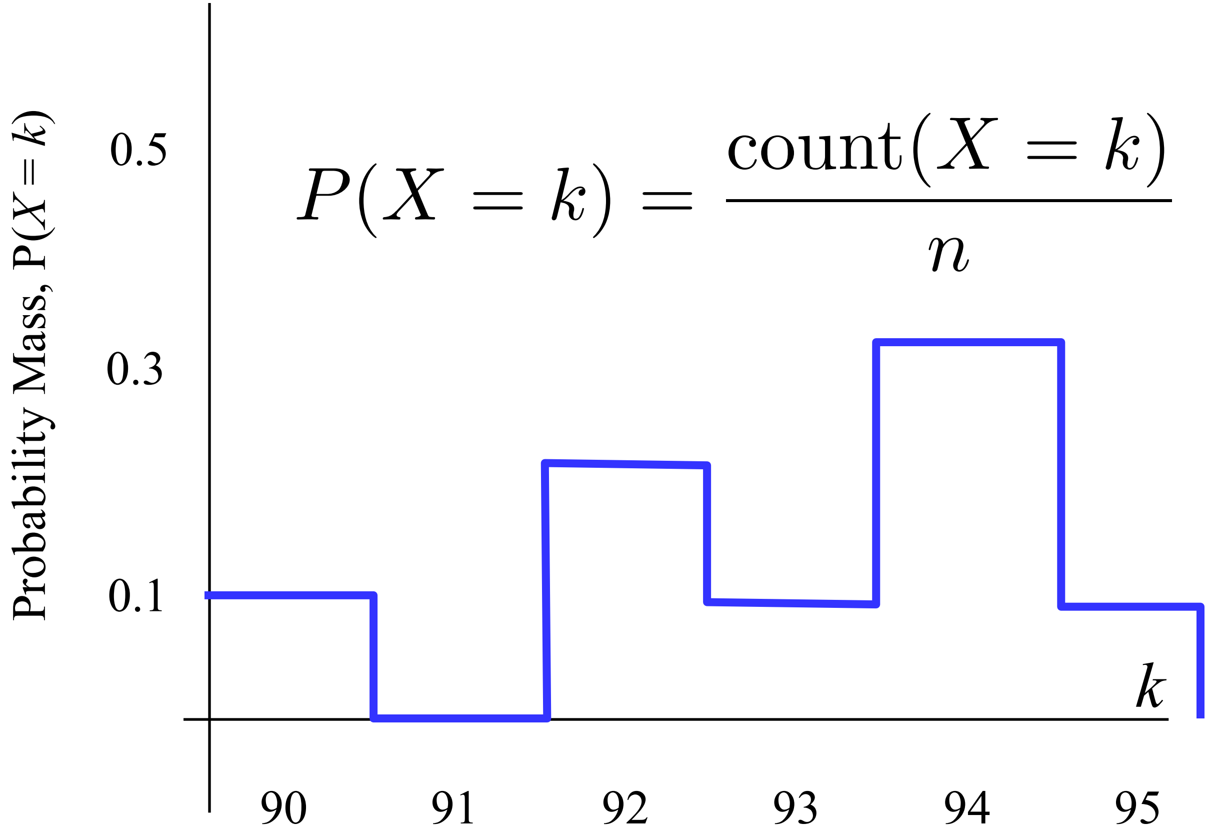
`np.random.choice(samples, K, replace = True)`

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

Resample:

[94, 90]

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

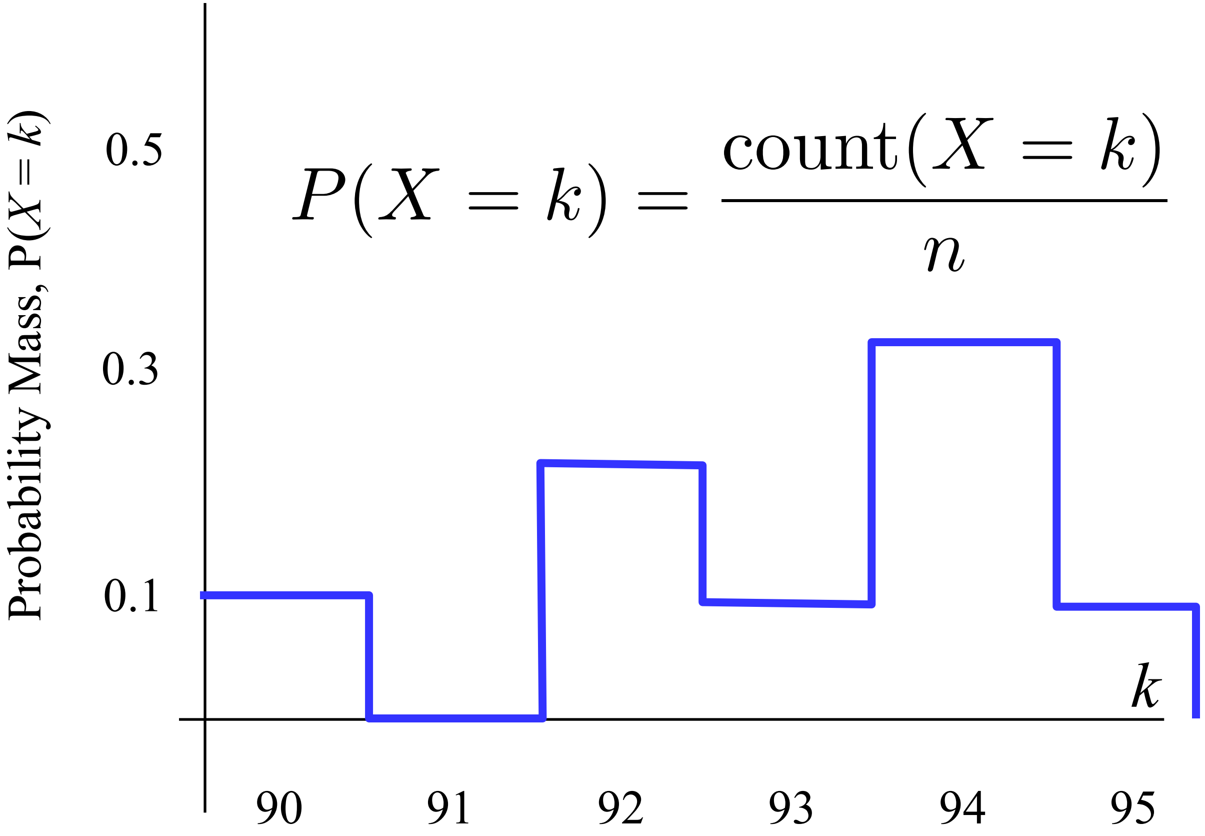


np.random.choice(samples, K, replace = True)

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

Resample:

[94, 90, **90**]

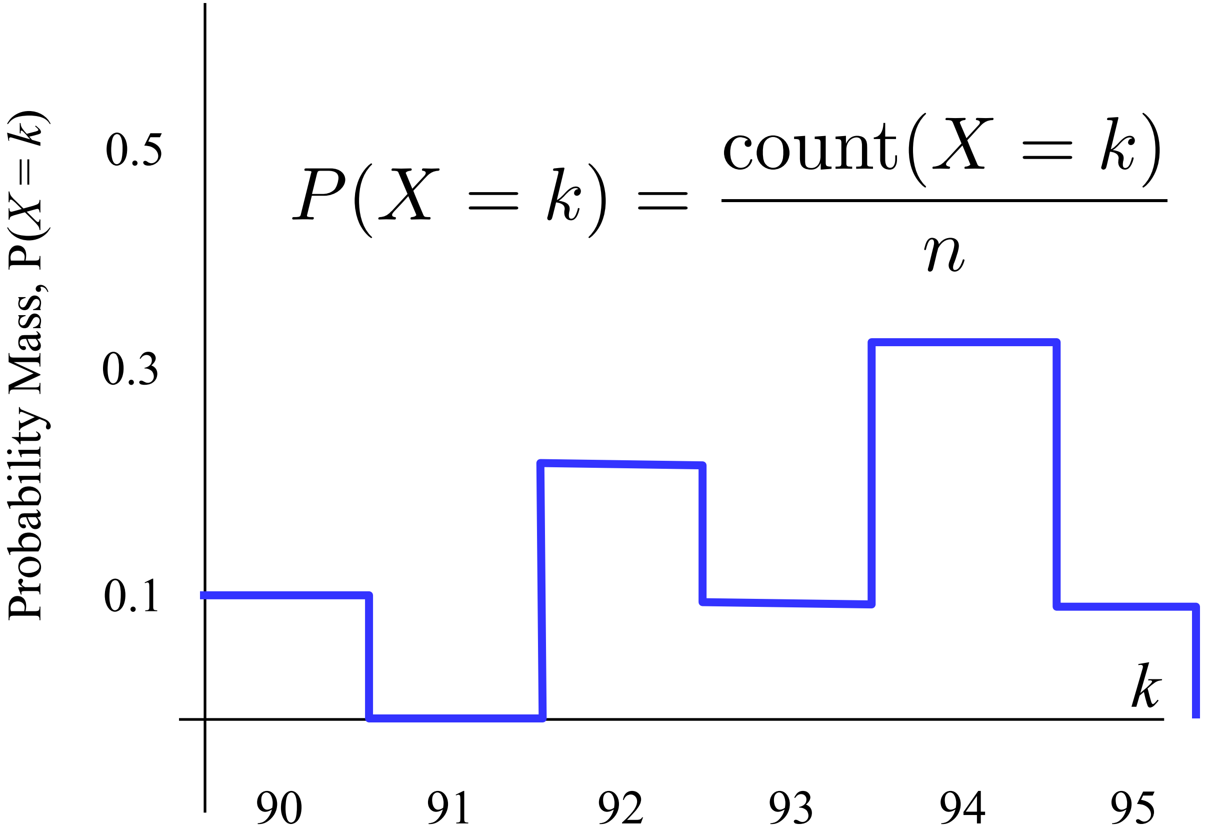


`np.random.choice(samples, K, replace = True)`

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

Resample:

[94, 90, 90]



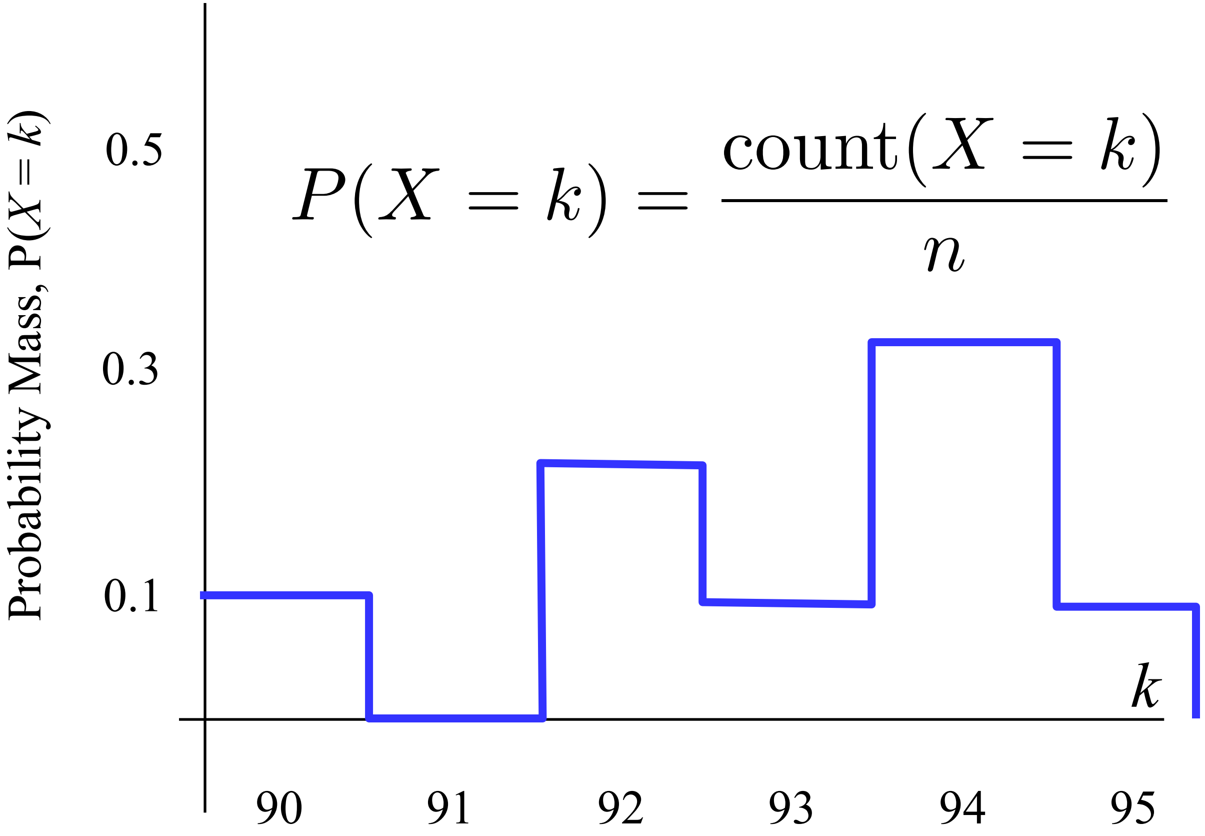
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

Now with replace = False

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

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

Resample:

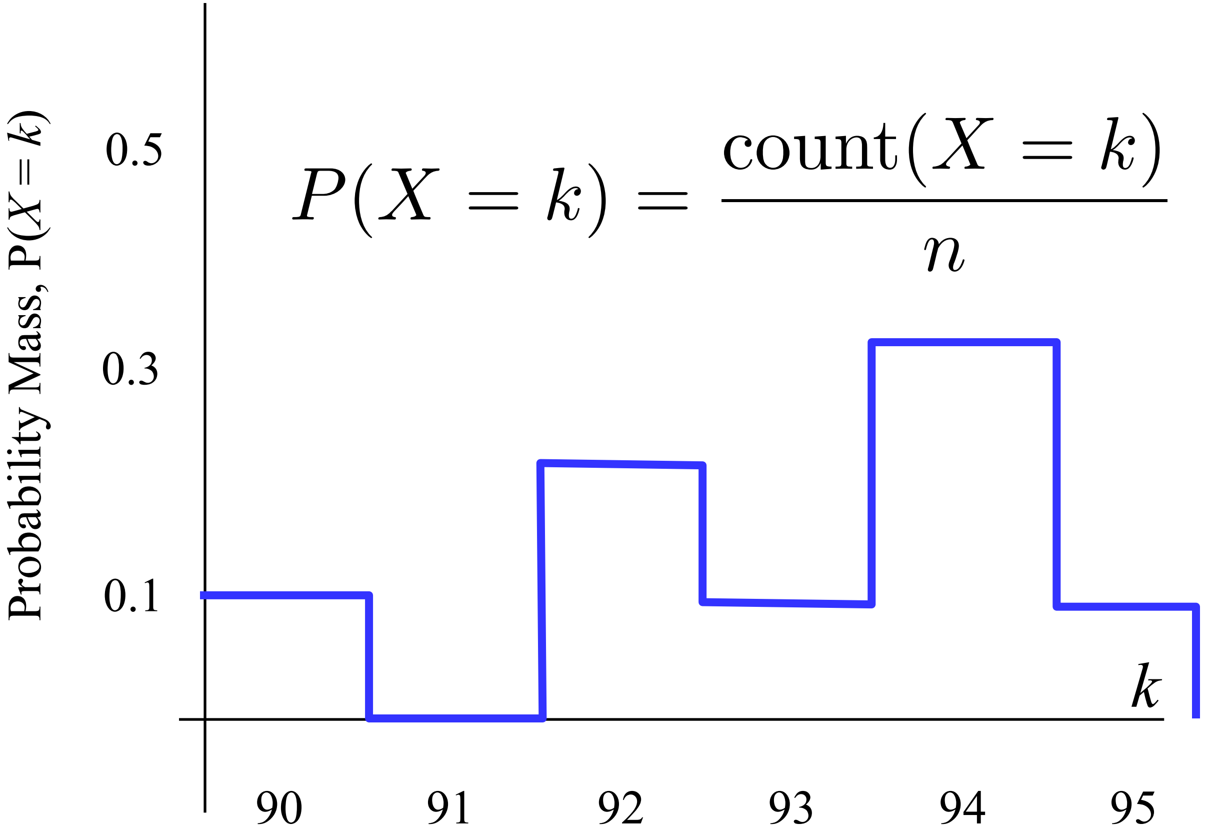


$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

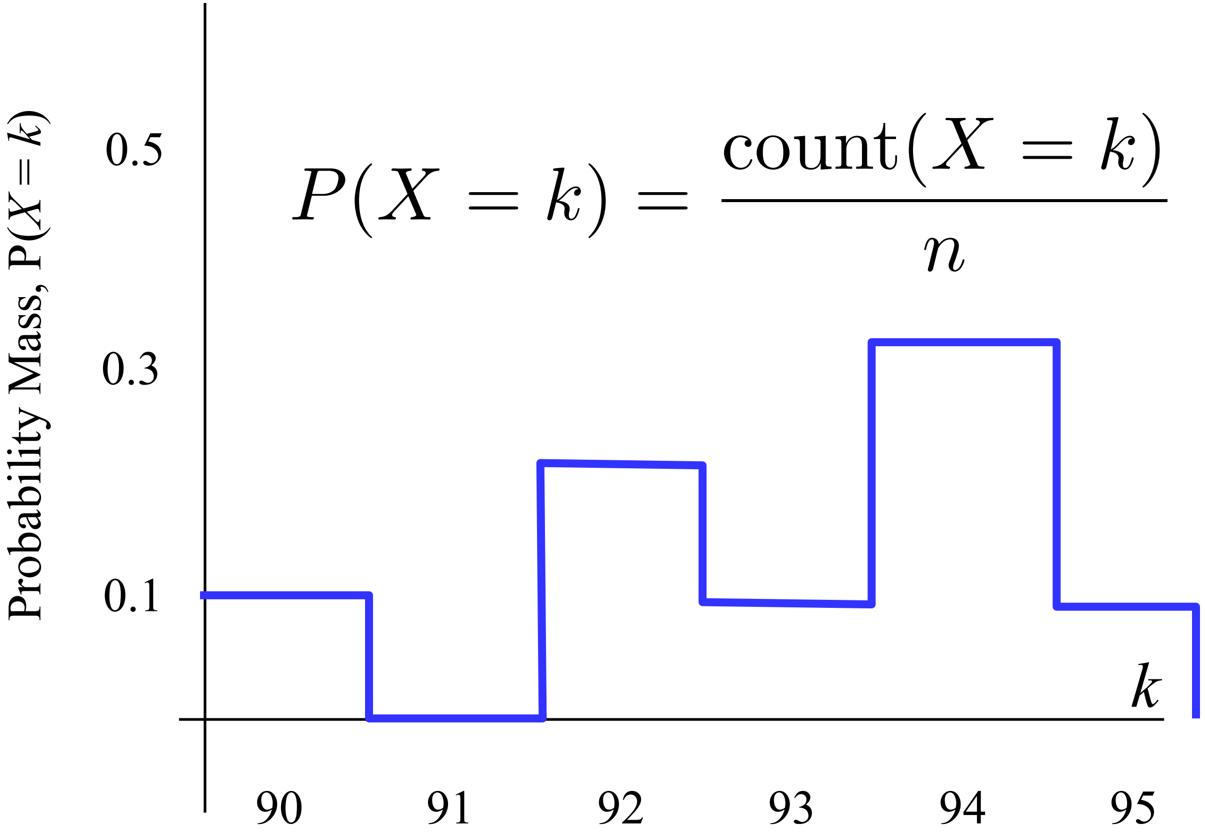


np.random.choice(samples, K, replace = False)

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

Resample:

[**94**]



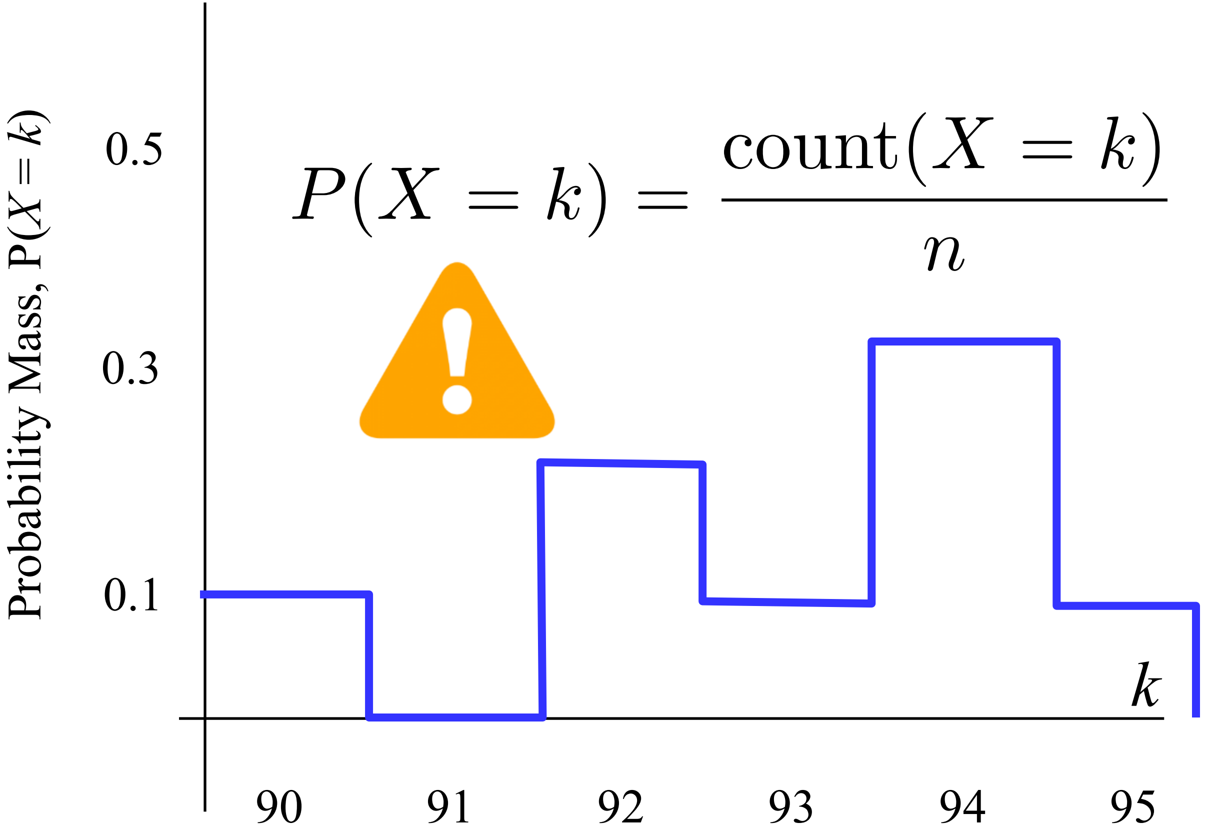
np.random.choice(samples, K, replace = False)

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

Resample:

Removed 94

[94]



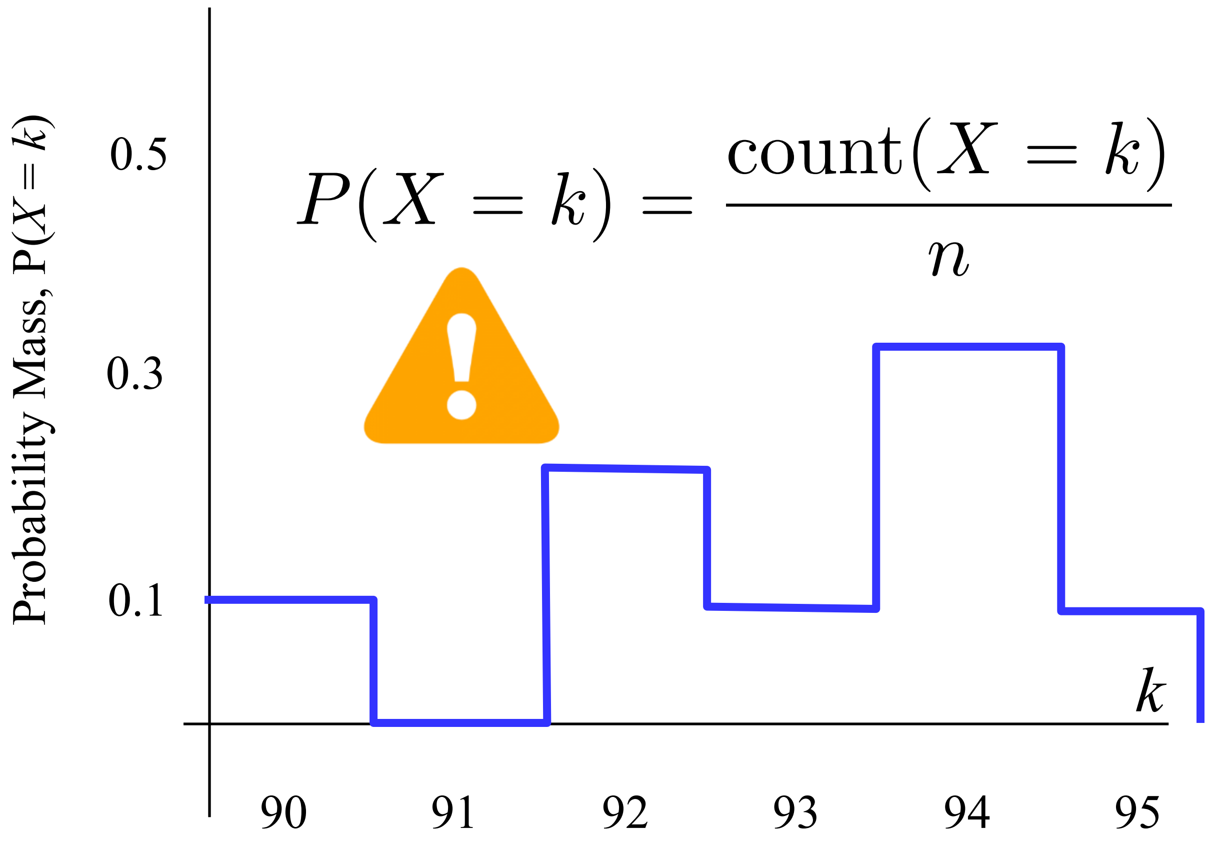
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[94]



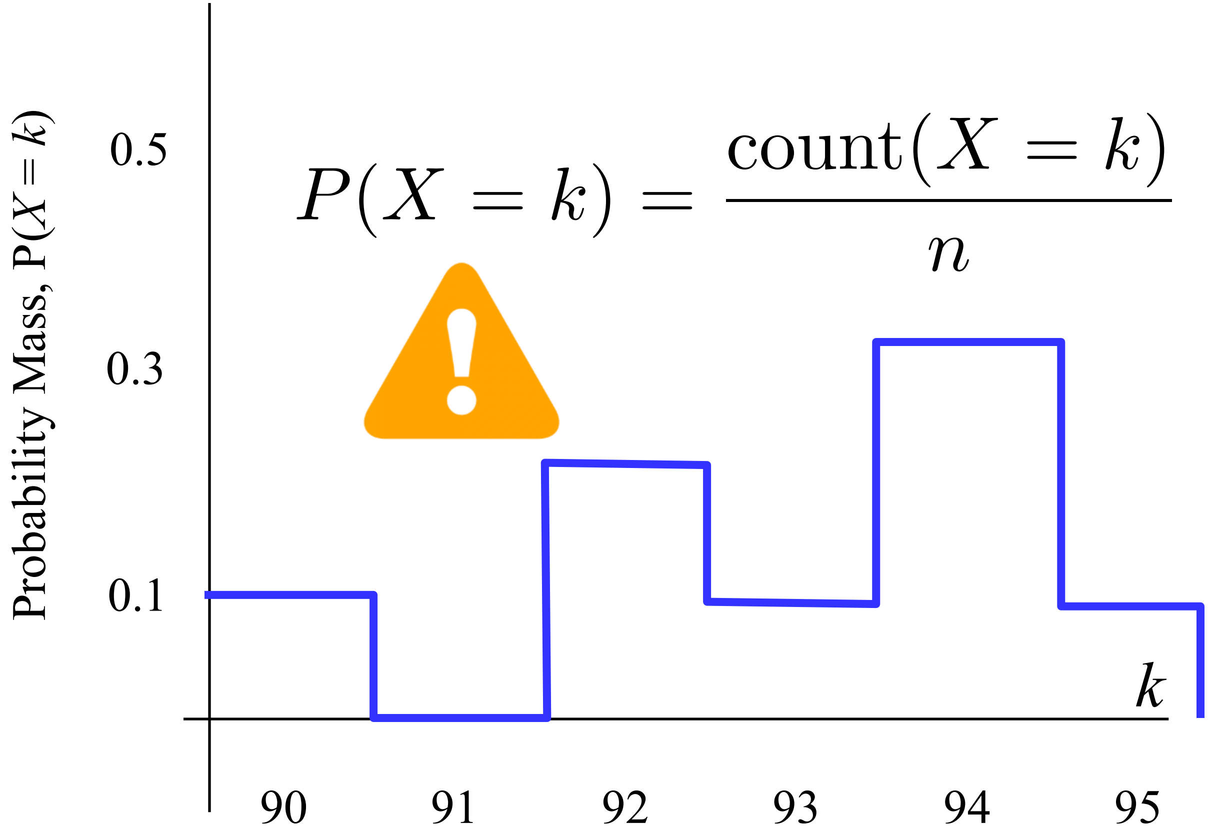
$$P(X = k) = \frac{\text{count}(X = k)}{n}$$

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

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

Resample:

[94]

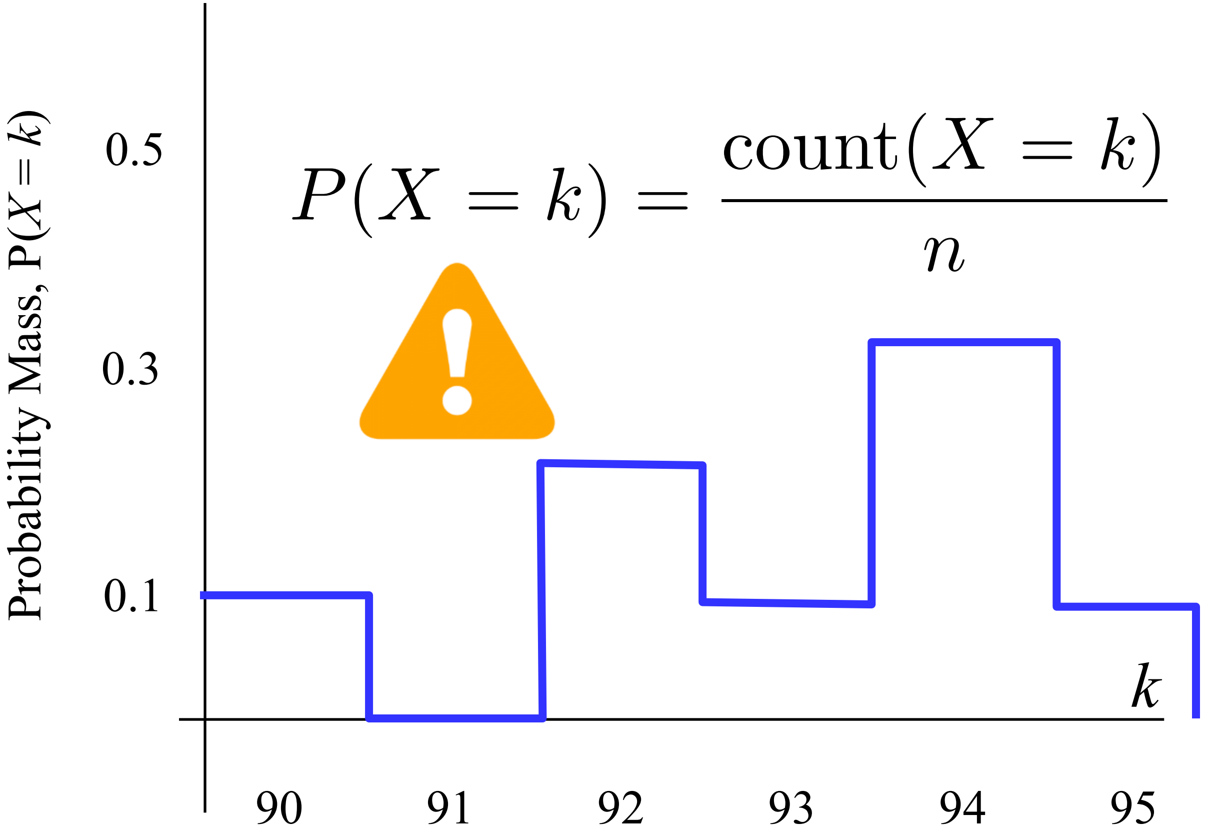


`np.random.choice(samples, K, replace = False)`

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

Resample:

[94, **90**]



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

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

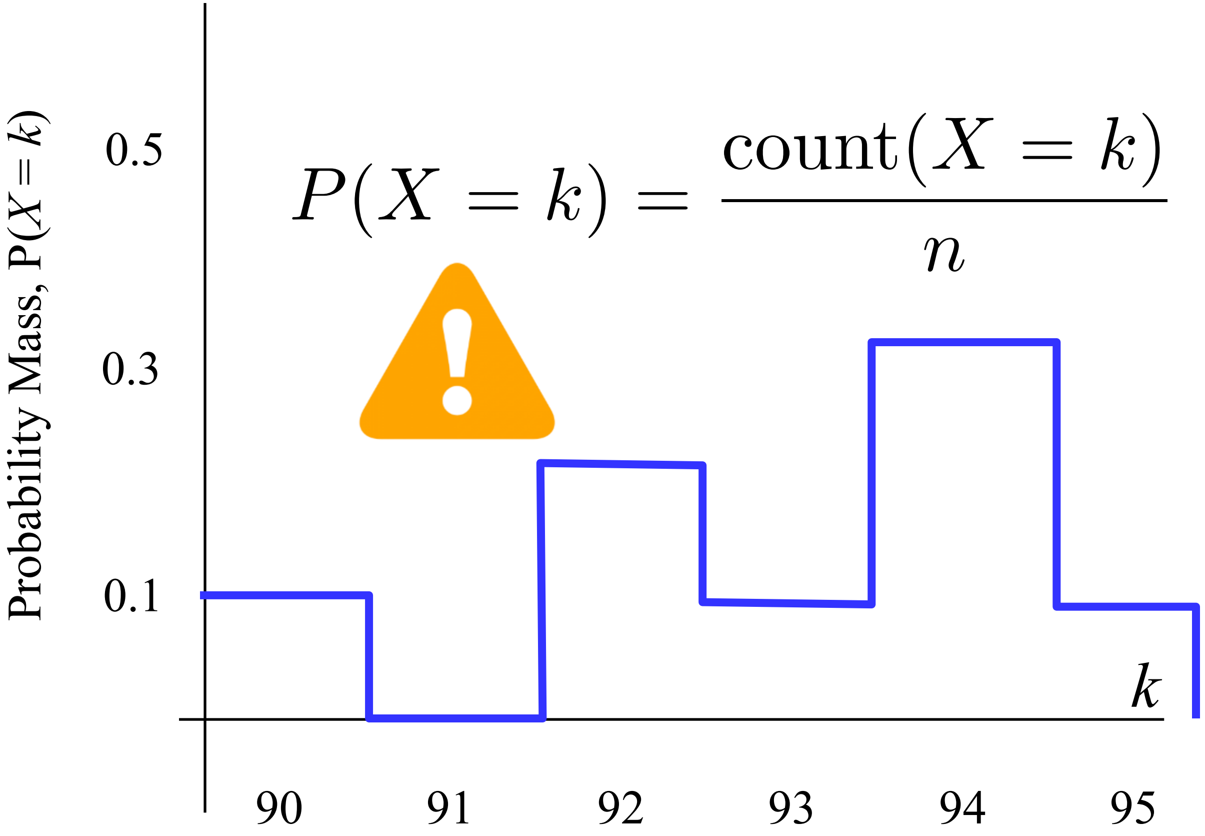
Original Samples: [92, 92, 93, 94, 94, 95]

Resample:

Removed 90

[94, 90]

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



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

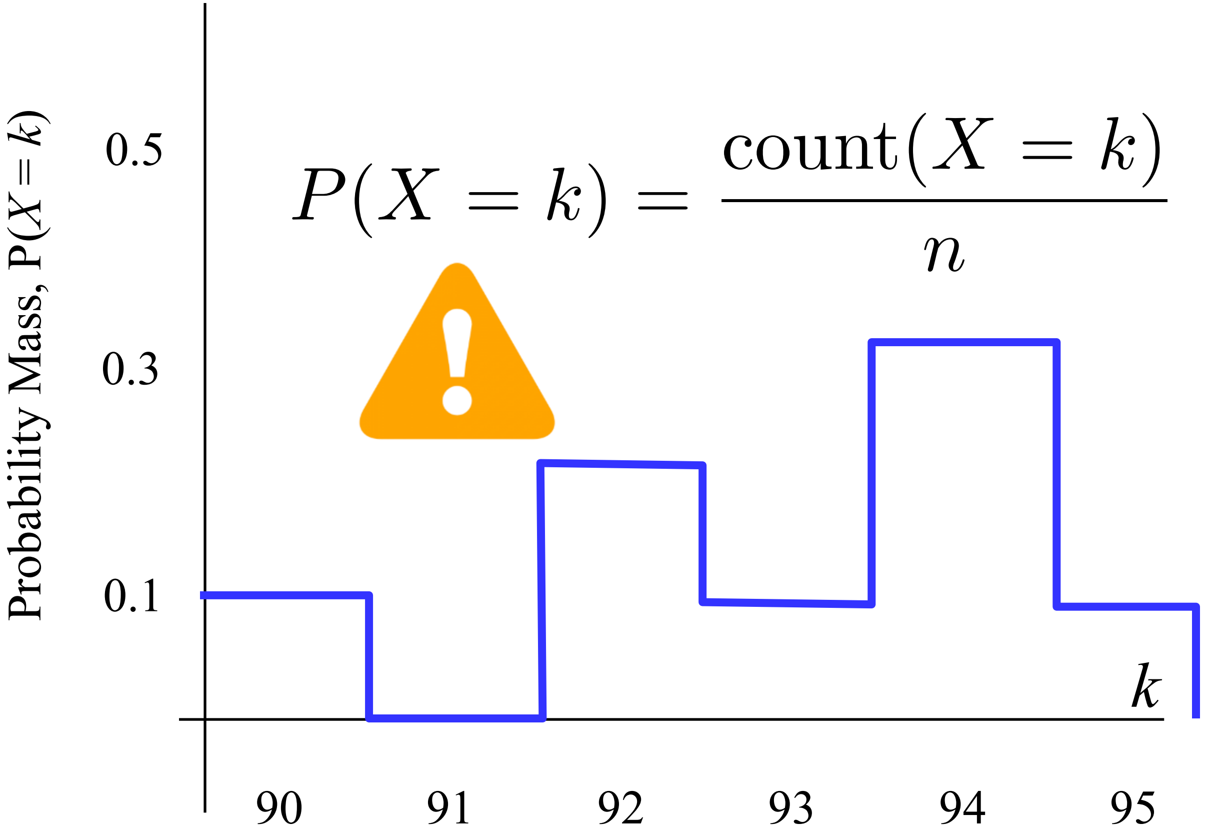
Original Samples: [92, 92, 93, 94, 94, 95]

Resample:

Removed 90

[94, 90]

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

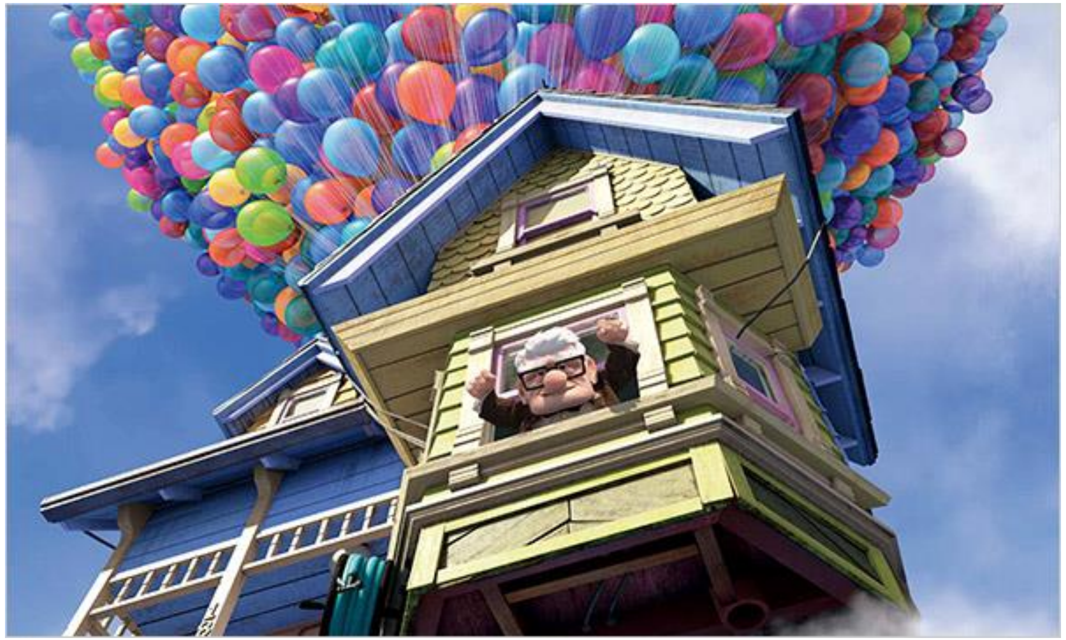


The probability of sampling a 90 is no longer 0.1

The probability of sampling 94 is no longer 0.3

Bootstrapping in Practice

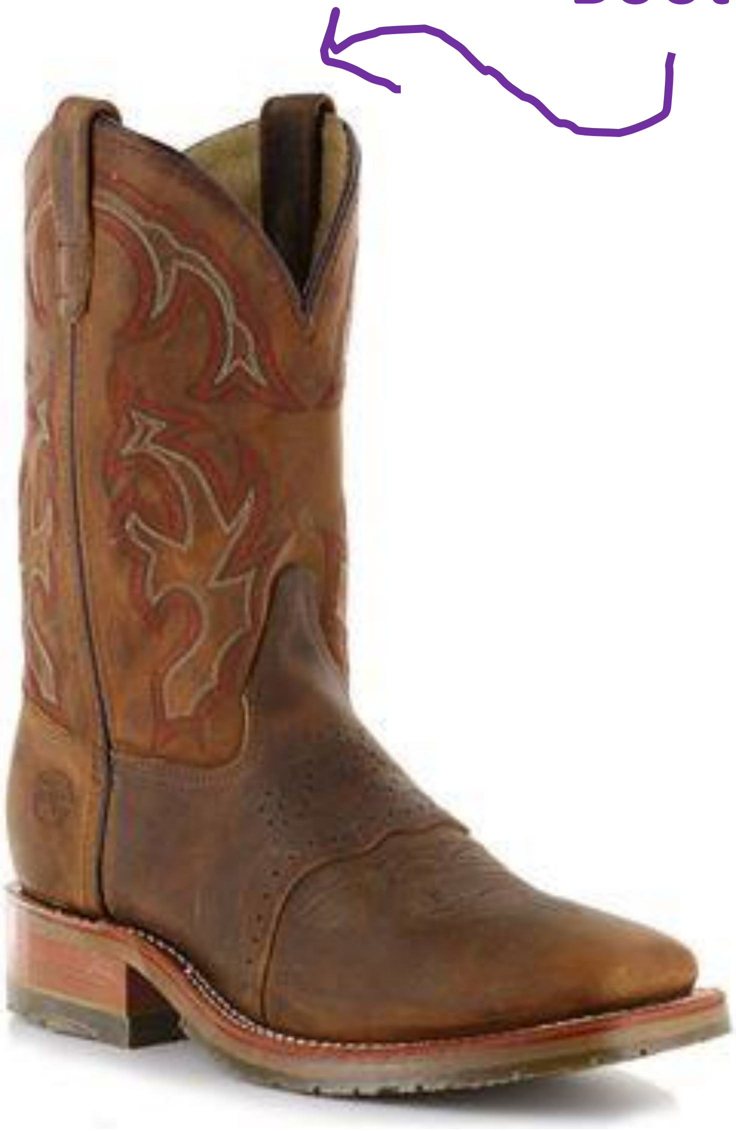
```
Bootstrap Algorithm (sample):
 1.  Repeat 10,000 times:
   a. Choose len(sample) elems from sample,
      with replacement
   b. Recalculate the stat on the resample
 2.  You now have a distribution of your stat
```



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

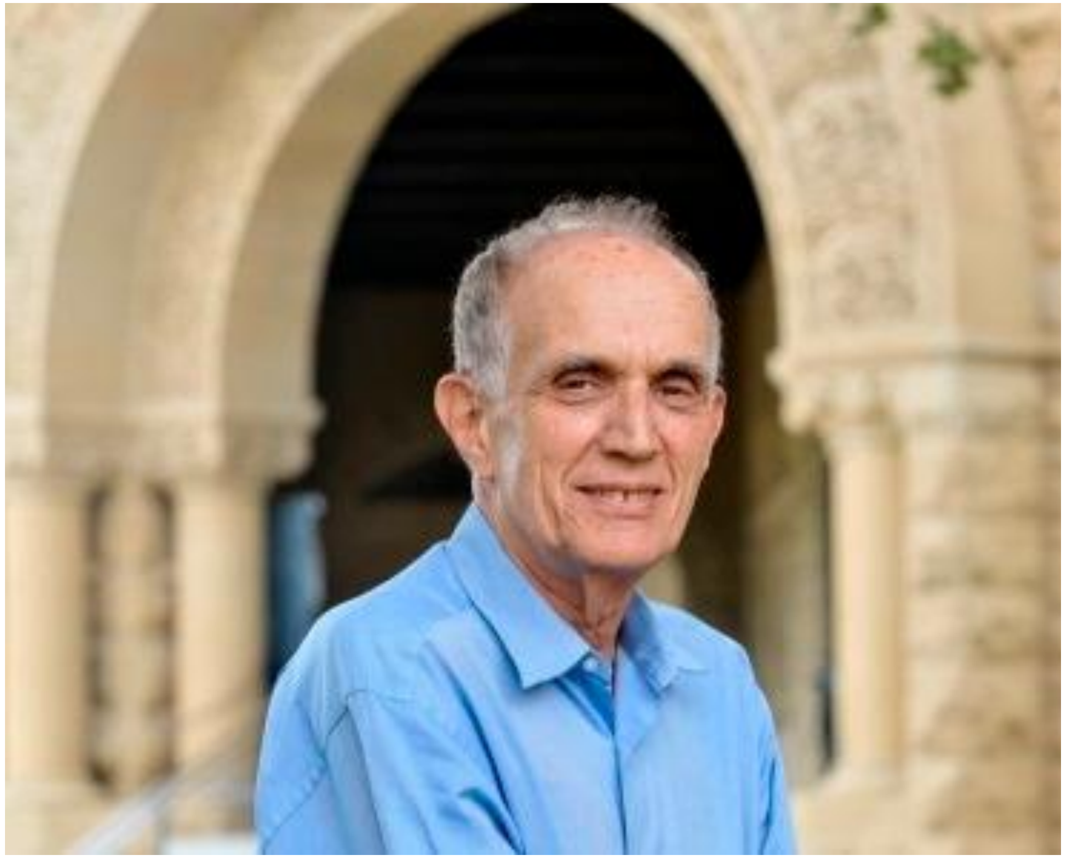
Works for any statistic*

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



Bootstrap

Bradley Efron



Invented bootstrapping in 1979

Still a professor at Stanford

Won a National Science Medal

According to starbyface.com: Dolph Lundgren

Hypothesis Testing

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?

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

23 students

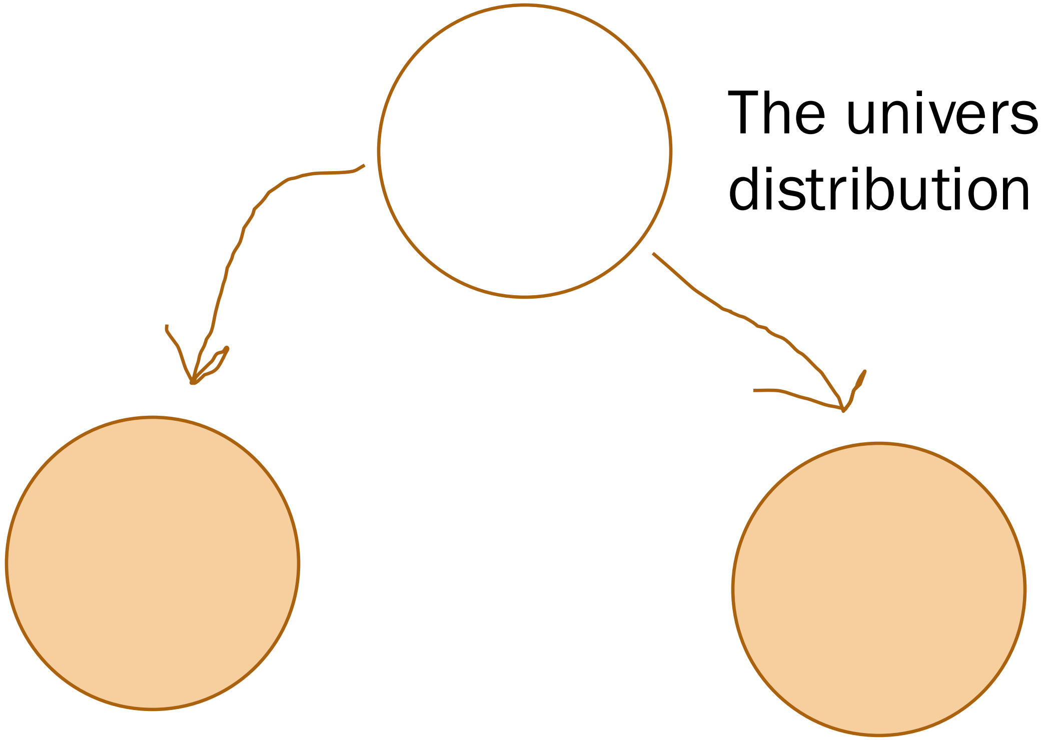
$\mu_2 = 2.4$

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.



The universal distribution

Group A Samples

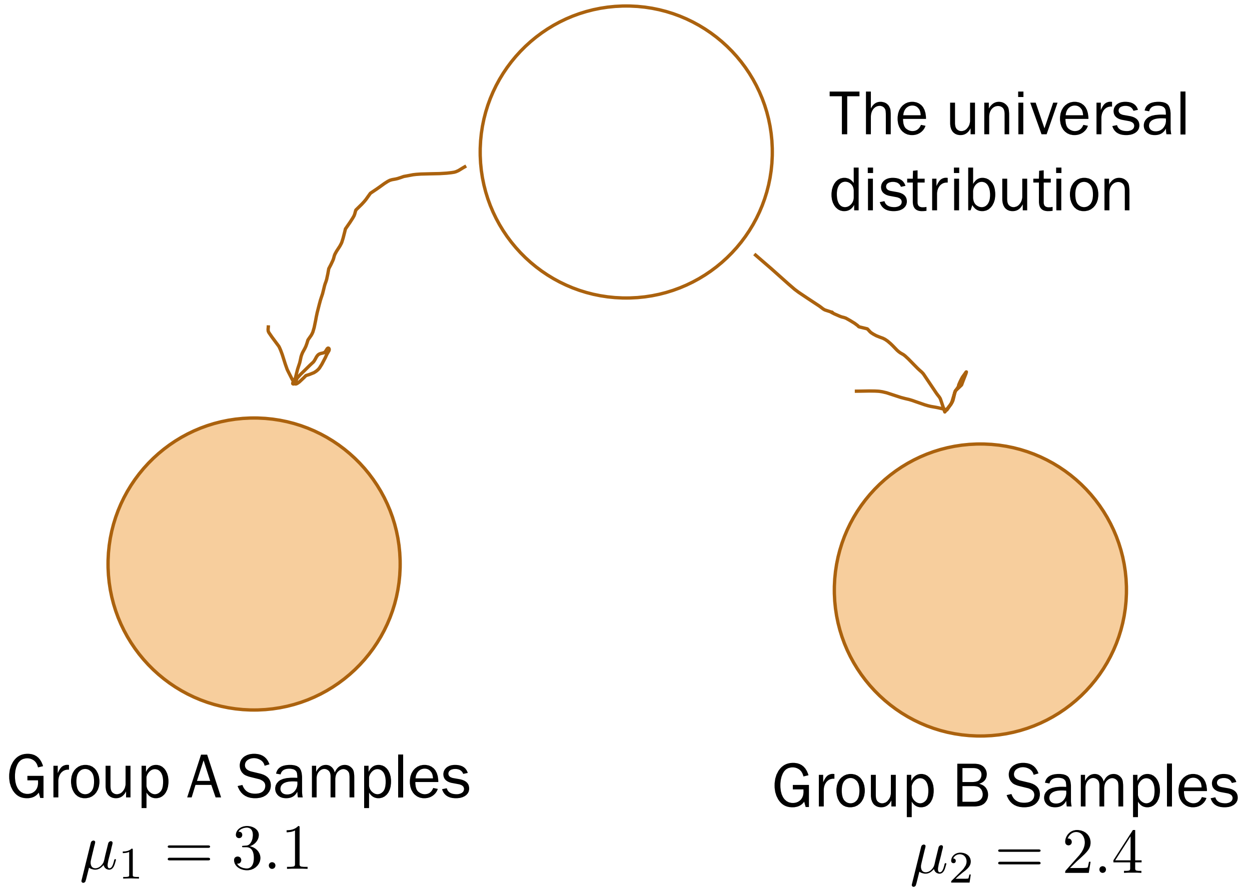
$\mu_1 = 3.1$

Group B Samples

$\mu_2 = 2.4$

P-Value

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



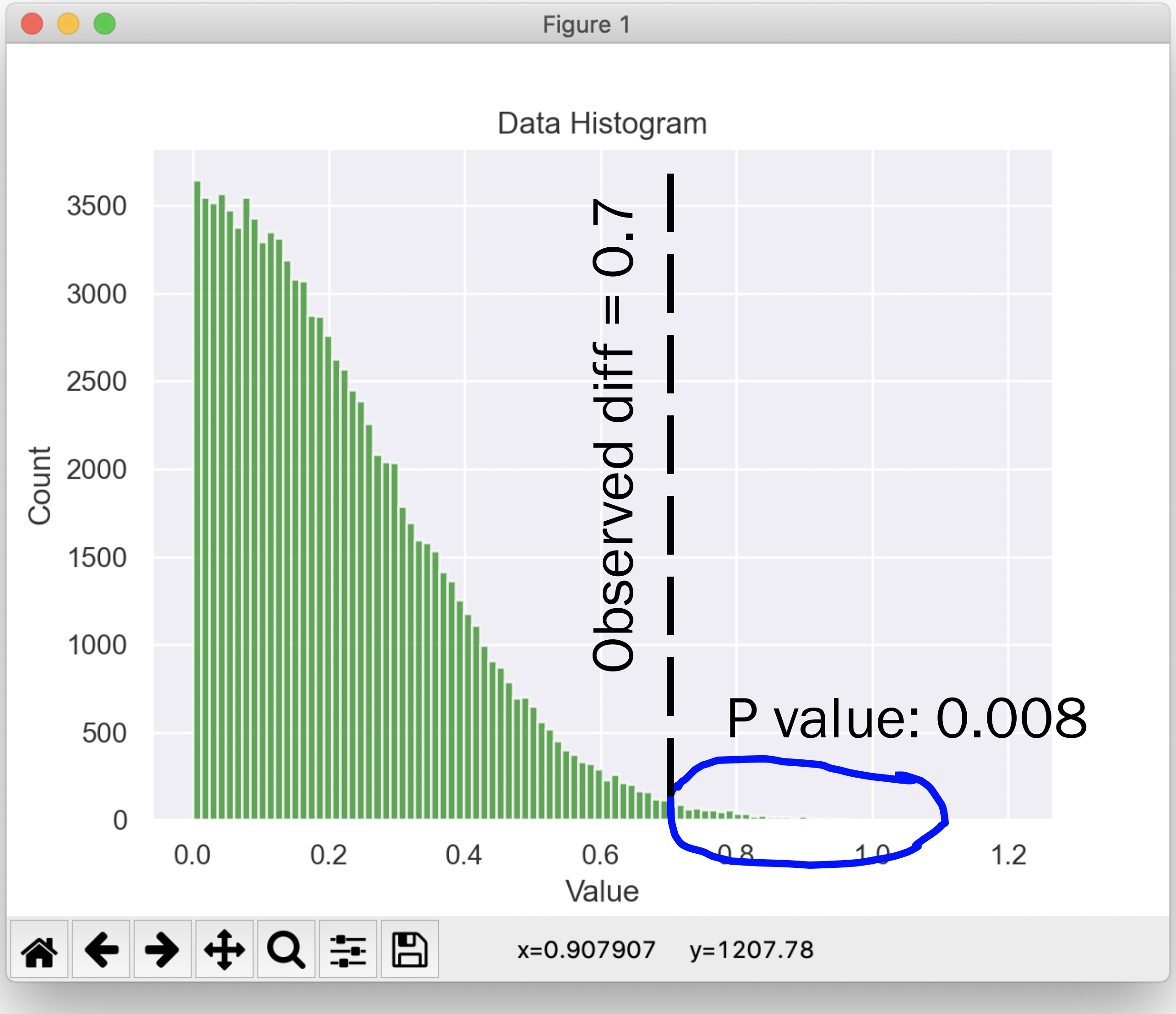
The universal distribution

Group A Samples
$\mu_1 = 3.1$

Group B Samples
$\mu_2 = 2.4$

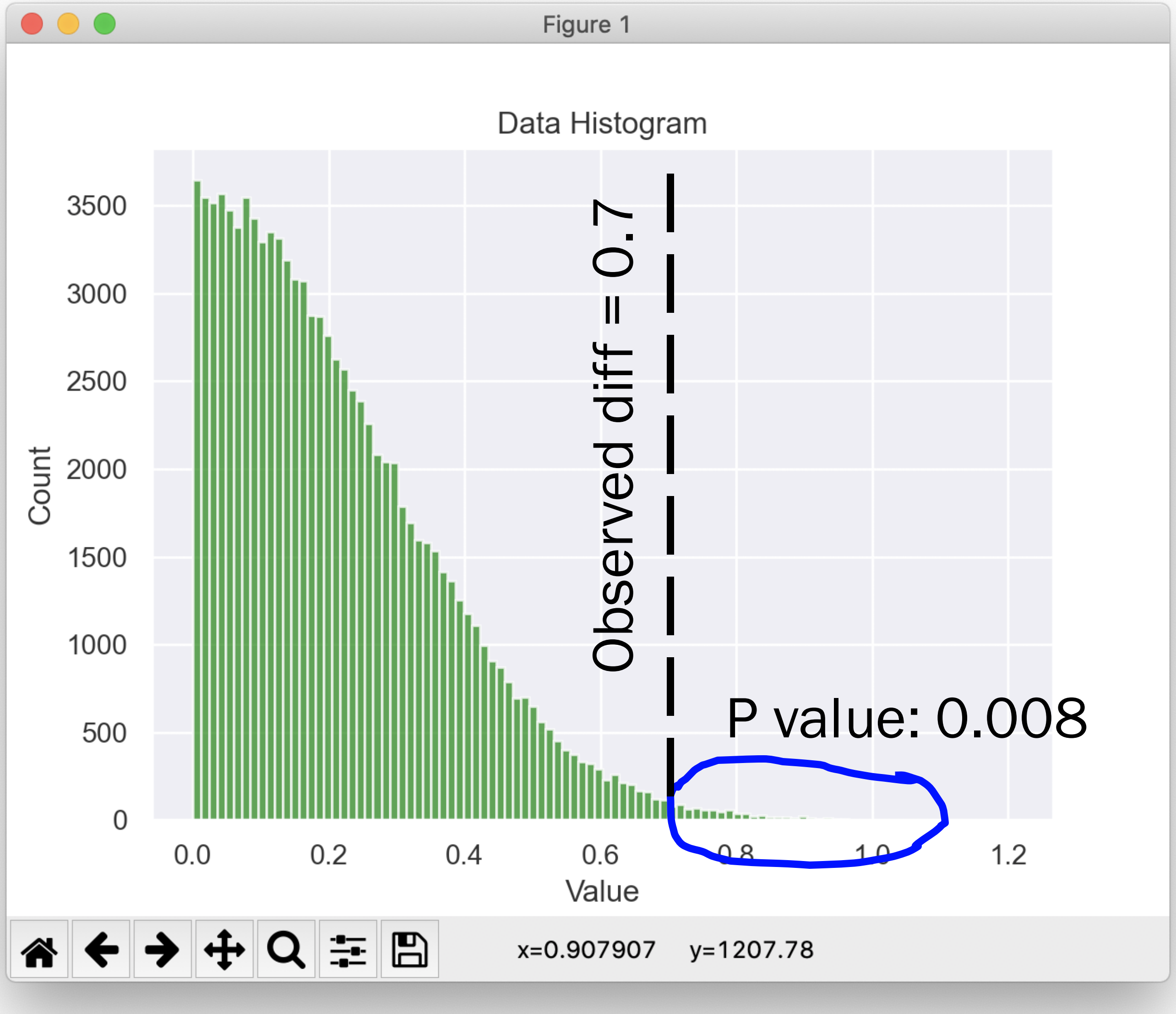
To the code!

Distribution of Mean Diffs under Null Hypothesis



Every* Science Result needs a p-value!

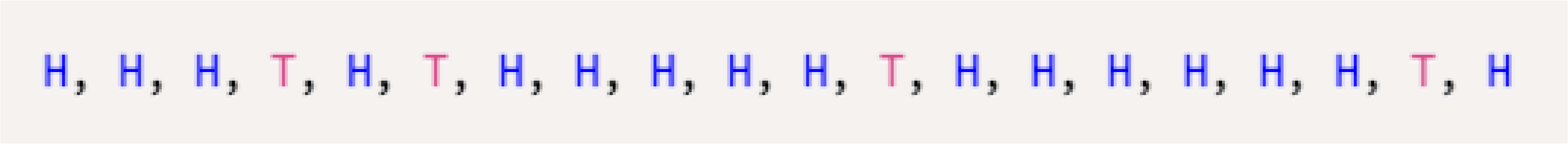
* almost



Food For Thought
(if extra time)

Puzzle

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



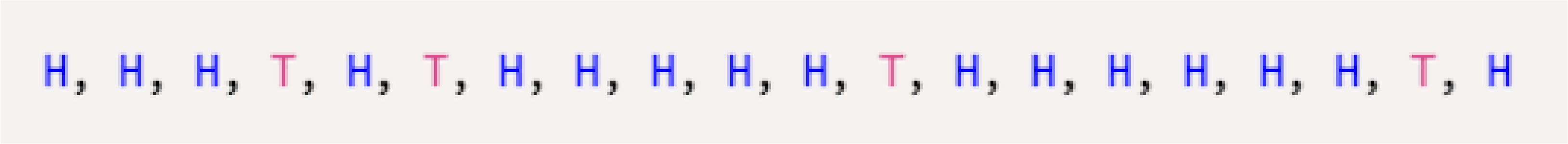
```
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H
```

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:



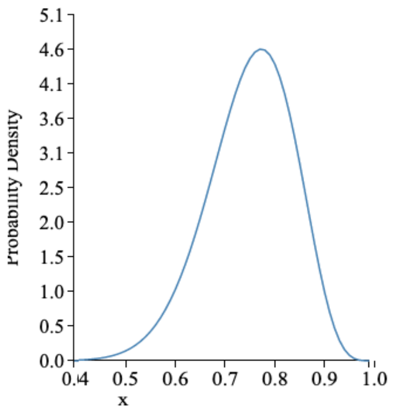
H, H, H, T, H, T, H, H, H, H, H, T, H, H, H, H, H, H, T, H

4 tails, 16 heads

Bayesian:

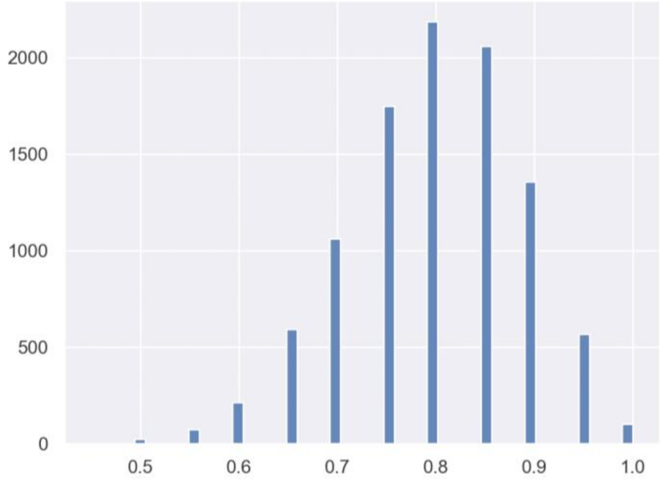
Let's use Laplace prior X ~ Beta(2, 2)

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



Frequentist:

Let's bootstrap


2000
1500
1000
500
0
0.5
0.6
0.7
0.8
0.9
1.0