Sampling

Chris Piech
CS109, Stanford University

MARCH 10TH

Review

Where are we in CS109?

You are here

Counting Theory

Core Probability

Random Variables

Probabilistic Models

Uncertainty Theory

Machine Learning

Uncertainty Theory

Beta Distributions

Thompson Sampling

Adding Random Vars

Central Limit Theorem

Sampling

Bootstrapping

Algorithmic Analysis

Information Theory + Divergence

As requested by AI faculty

What happens when you Add Two Random Variables?

$$P(A + B = n) = ?$$

Convolution

Discrete

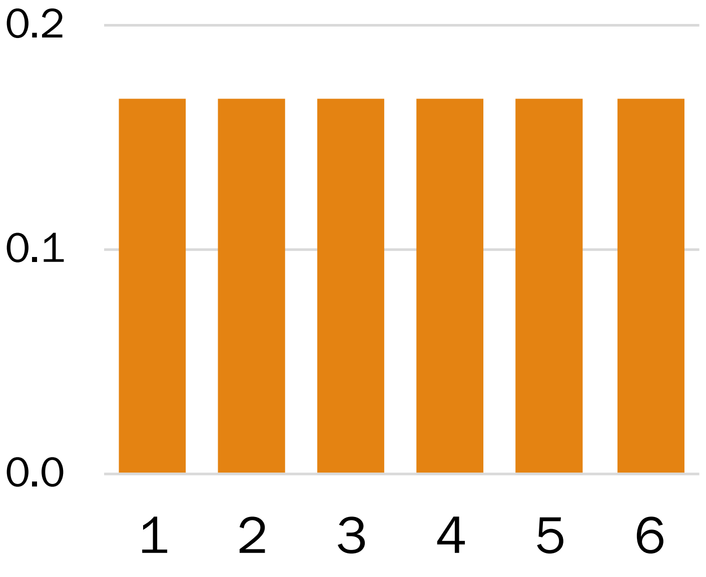
$$P(X+Y=a) = \sum_{y=-\infty}^{\infty} P(X=a-y)P(Y=y)\ dy$$

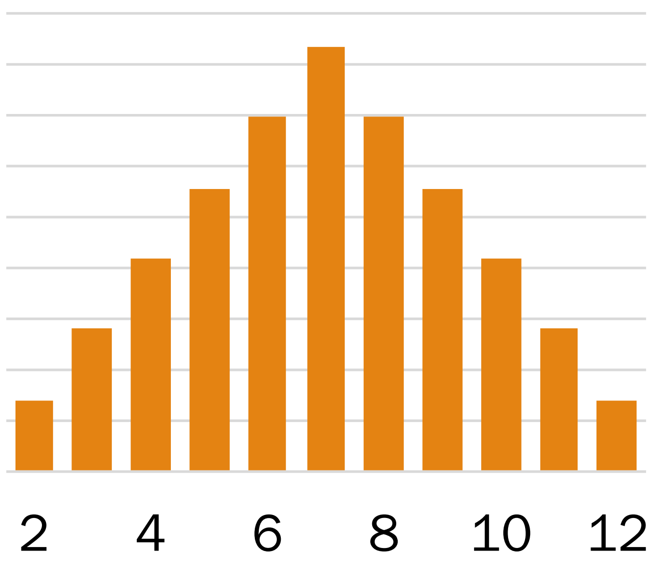
Continuous

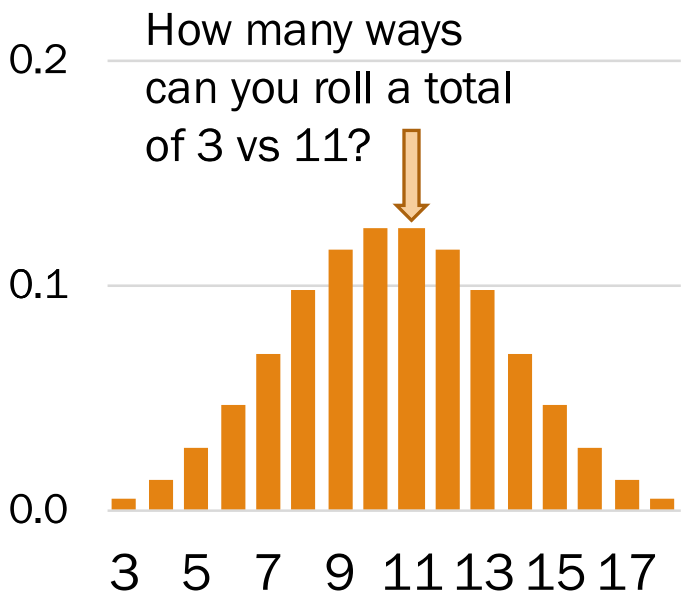
$$f(X+Y=a) = \int_{y=-\infty}^{\infty} f(X=a-y)f(Y=y)\ dy$$

Sum of dice rolls

Roll n independent dice. Let X_i be the outcome of roll i. X_{i} are i.i.d.





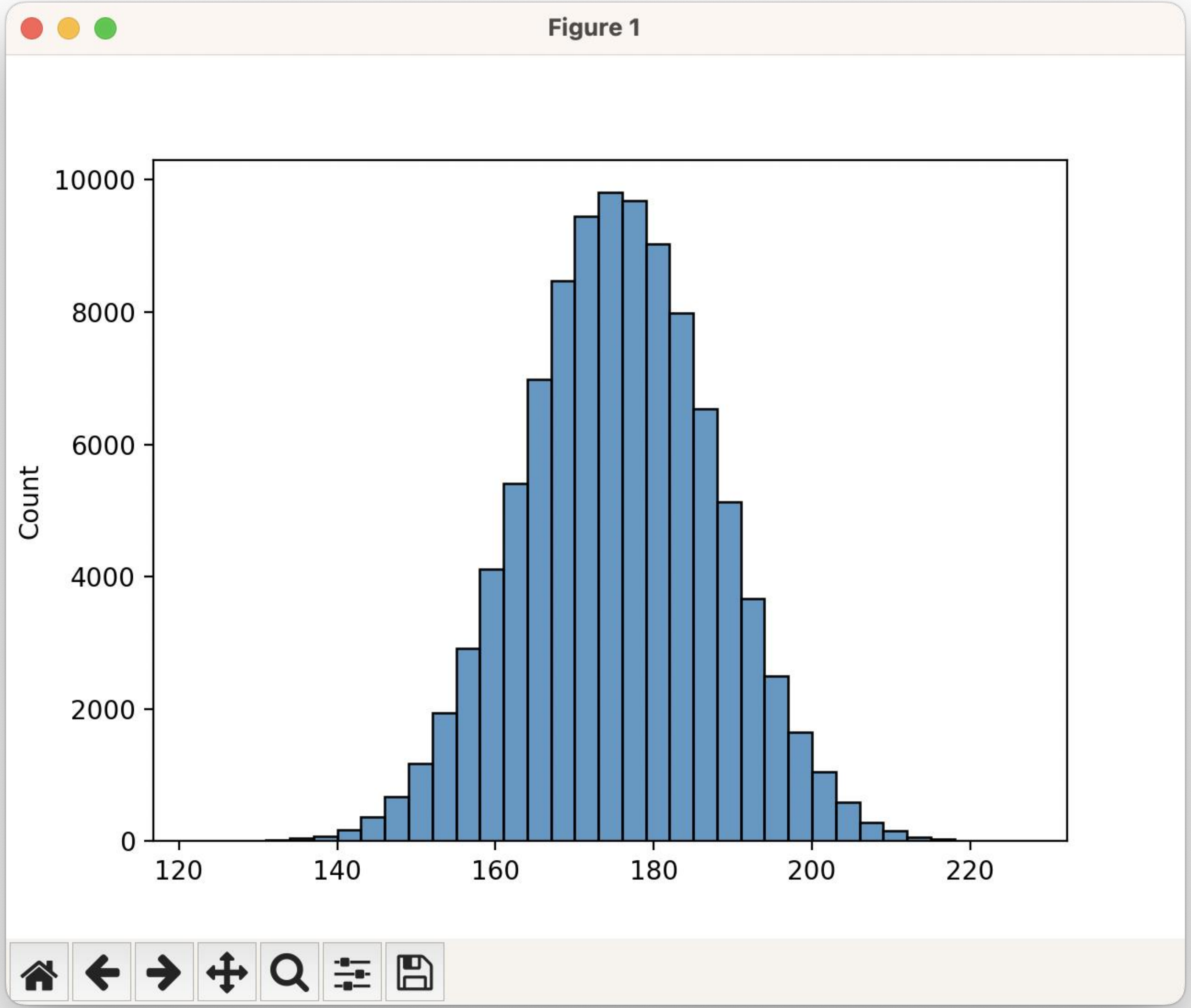


$\sum_{i=1}^{1} X_i$ Sum of 1 die roll

$\sum_{i=1}^{2} X_i$ Sum of 2 dice rolls

$\sum_{i=1}^{3} X_i$ Sum of 3 dice rolls

Sum of 50 dice?




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

Central Limit Theorem (Summation)

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

As $n \to \infty$

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

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

Central Limit Theorem (Average)

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

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

As $n \to \infty$

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

Average of IID Variables Demo

Demo

http://onlinestatbook.com/stat_sim/sampling_dist/

True happiness

Example CLT problem

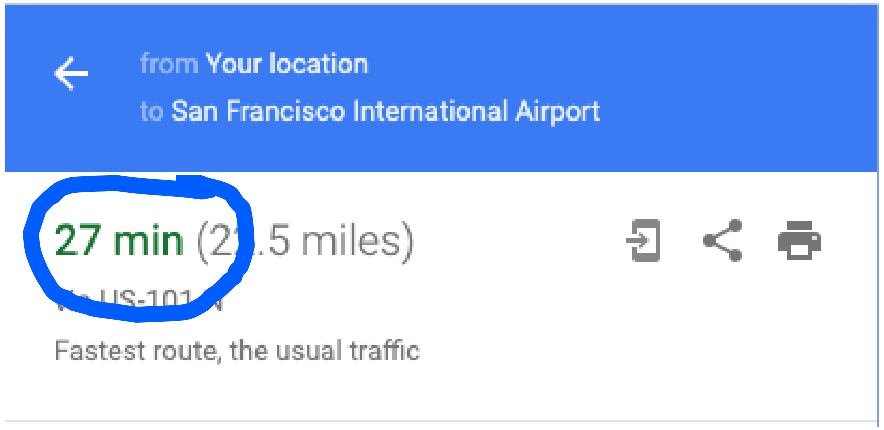
You hit 10 traffic lights on your way to work. You don't know the full distribution of the wait time, but for each you observe the average wait time is 45 seconds and the standard deviation is 5 seconds. You will be on time if your total wait time is less than 8 mins. What is the probability that you are on time? Assume the wait times are IID.

Answer: Let T be the total wait time. It is the sum of the 10 IID wait times. By the CLT

$$T \sim \mathcal{N}(n\mu, n\sigma^2)$$

$$T \sim \mathcal{N}(450, 250)$$

$$P(T \leq 480) \quad = \Phi\left(\frac{480 - 450}{15.8}\right) \quad \approx 0.97$$



What about other functions?

Sum of iid? Normal

Average of iid? Normal

Max of iid? Gumbel

See Fisher Trippett Gnedenko Theorem

Estimating Clock Running Time

- Have new algorithm to test for running time
  - Mean (clock) running time: $\mu = t$ sec.
  - Variance of running time: $\sigma^2 = 4$ sec^2.
  - Run algorithm repeatedly (I.I.D. trials), measure time
    - How many trials do you need s.t. estimated time = $t \pm 0.5$ with 95% certainty?
    - X_i = running time of i-th run (for $1 \leq i \leq n$), $\bar{X}$ is the mean

---

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \sim N(t, \frac{4}{n})$$

$$0.95 = P(-0.5 < \bar{X} - t < 0.5) \qquad \bar{X} - t \sim N(0, \frac{4}{n})$$

$$0.95 = F_{\bar{X}-t}(0.5) - F_{\bar{X}-t}(-0.5)$$

$$= \Phi\left(\frac{0.5 - 0}{\sqrt{4/n}}\right) - \Phi\left(\frac{-0.5 - 0}{\sqrt{4/n}}\right)$$

$$= 2\phi(\frac{\sqrt{n}}{4}) - 1$$

$$0.95 \ = 2\phi(\frac{\sqrt{n}}{4}) - 1$$

$$0.975 = \phi(\frac{\sqrt{n}}{4})$$

$$\phi^{-1}(0.975) = \frac{\sqrt{n}}{4}$$

$$1.96 = \frac{\sqrt{n}}{4}$$

$$n = 61.4$$

William Sealy Gosset
(aka Student)

Uncertainty Theory

- Beta Distributions
- Thompson Sampling
- Adding Random Vars
- Central Limit Theorem
- Sampling
- Bootstrapping
- Algorithmic Analysis
- Information Theory + Divergence

As requested by AI faculty

Sampling definitions

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:

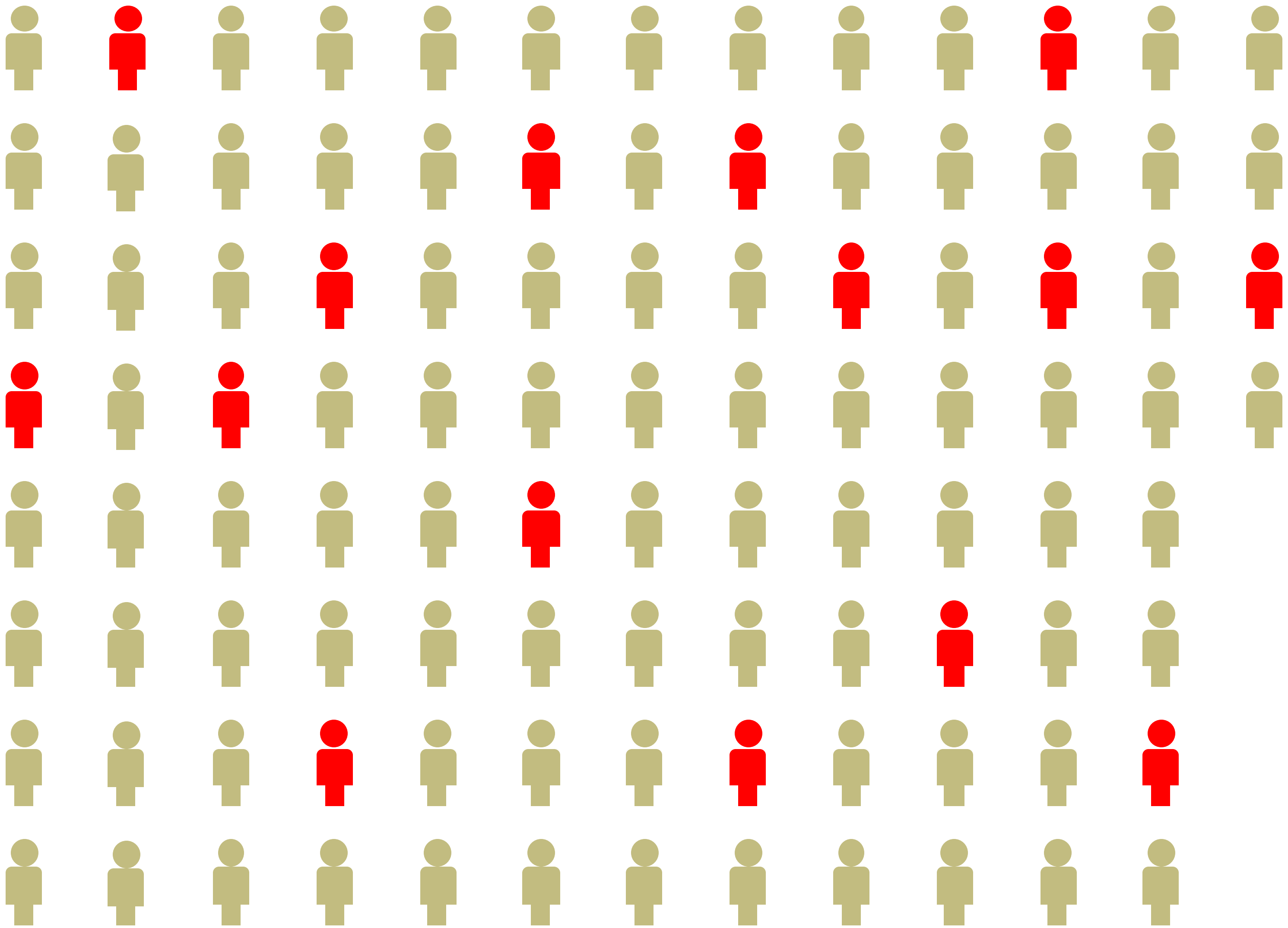
  Happiness = {72, 85, 79, 91, 68, …, 71}

- The mean of all these numbers is 83.

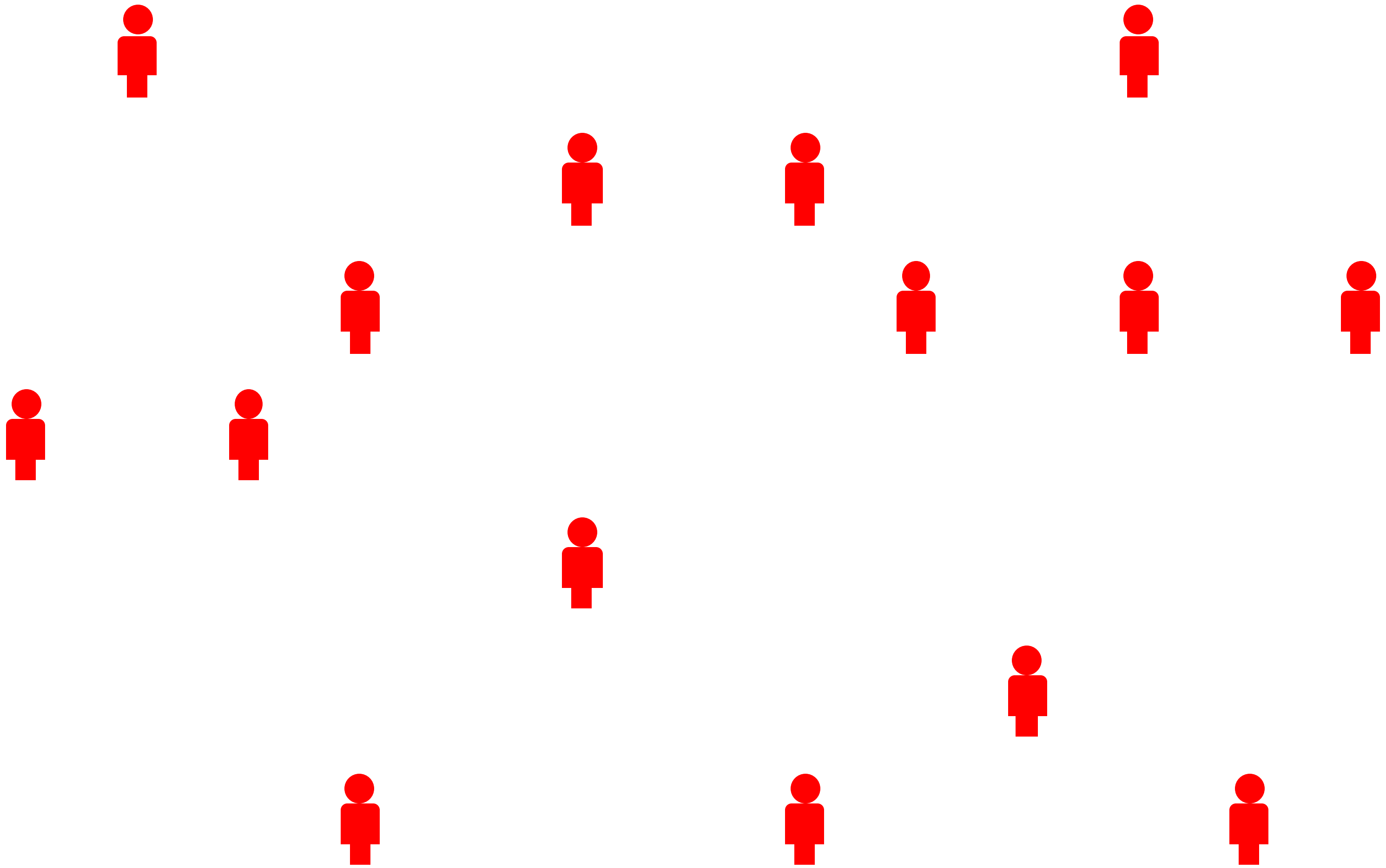
Is this the **true mean happiness** of Bhutanese people?

Population

Sample



Sample



Collect one (or more) numbers from each person

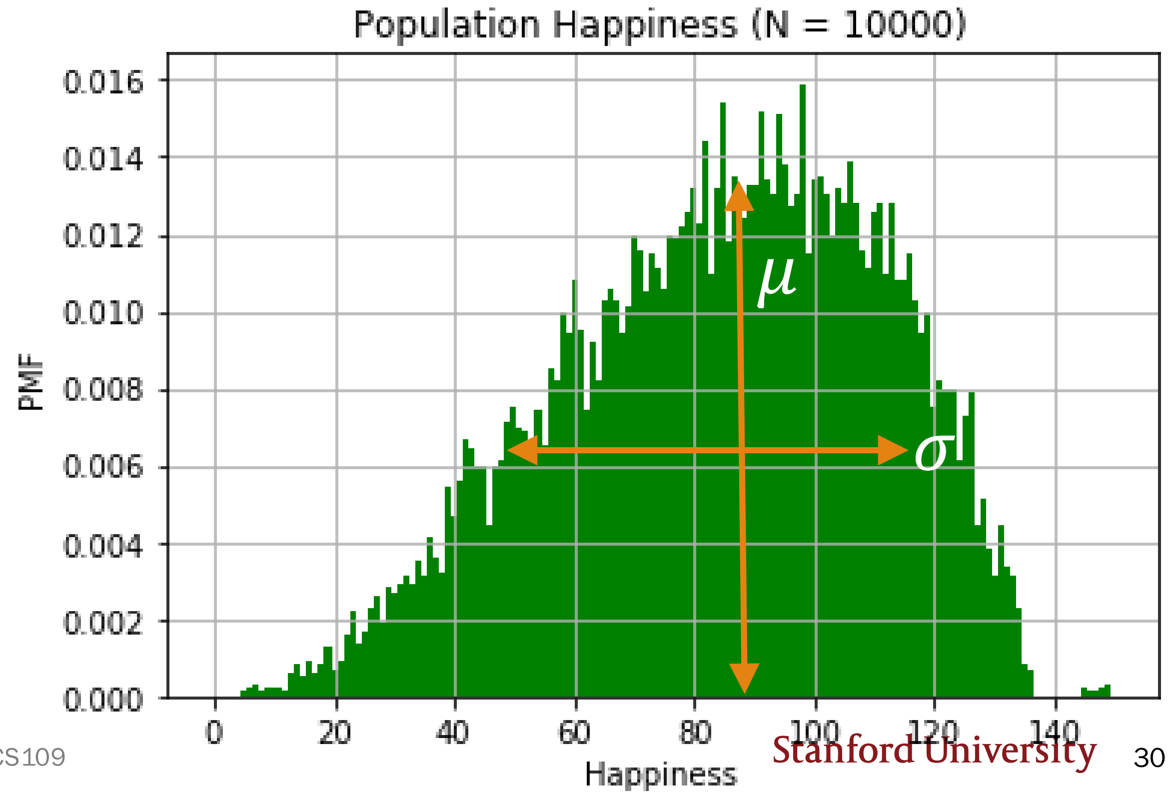
Sample

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 $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$



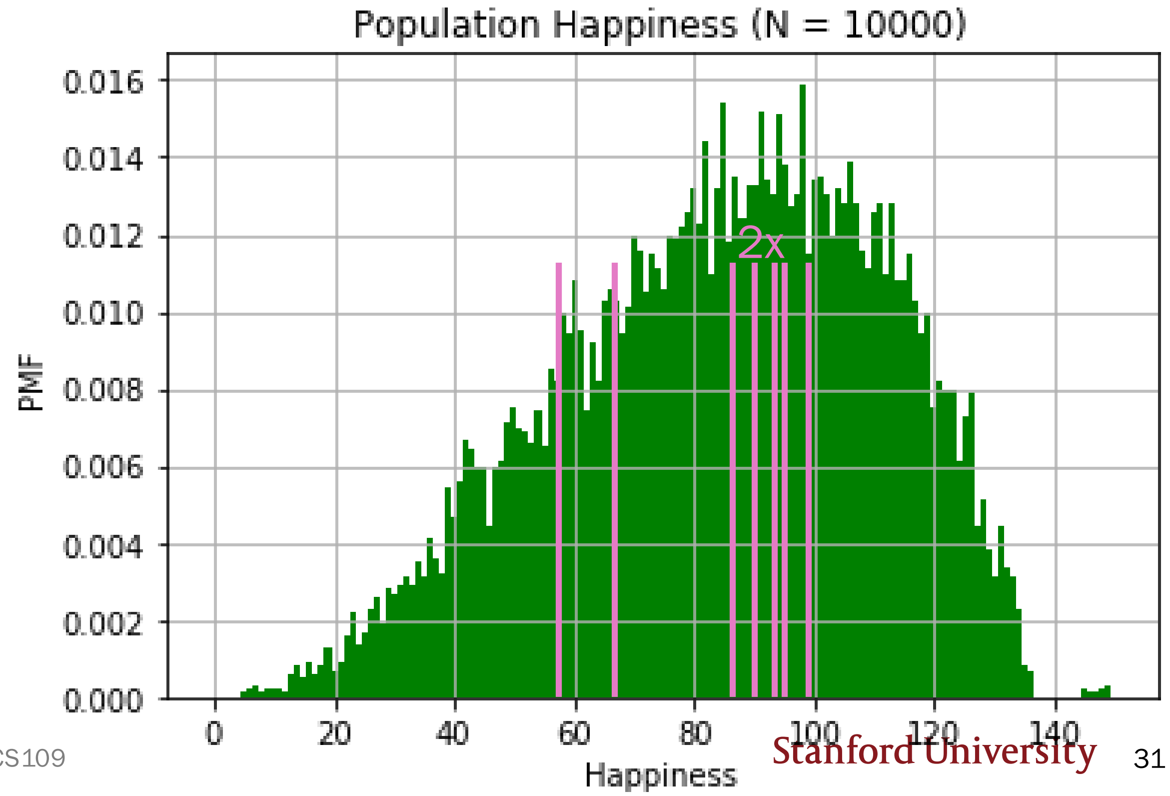
A sample, mathematically

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:

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



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?

Estimating Core Statistics (Mean + Var)

Equations we used to get those values

sample mean estimate

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

Our best guess at the true mean

sample mean

sample variance estimate

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

Our best guess at the true variance

Std error of the mean estimate

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

sample variance

How wrong do we think our mean estimate is?

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).

- From these 200 people, what is our best estimate of population mean and population variance?
- How good are those estimates?

Estimating the Mean

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

- X_i are all independently and identically distributed (I.I.D.)
- Have same distribution function F and $E[X_i] = \mu$
- We call sequence of X_i a **sample** from distribution F
- *How would you estimate the population mean??*

$$\text{Estimated mean} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample Mean: This is a fancy way of writing "your estimate of the mean"

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

Is that estimate any good?

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

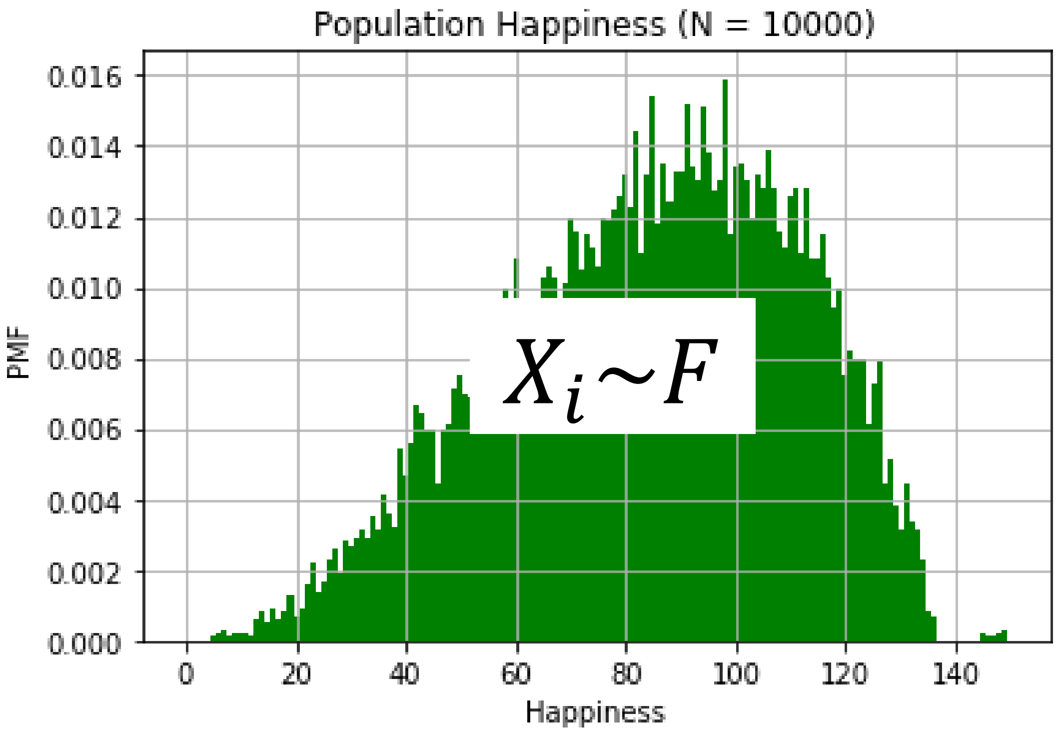
Consider n random variables $X_1, X_2, \dots X_n$

- Have same distribution function F and $E[X_i] = \mu$
- *Is our estimate of mean any good??*

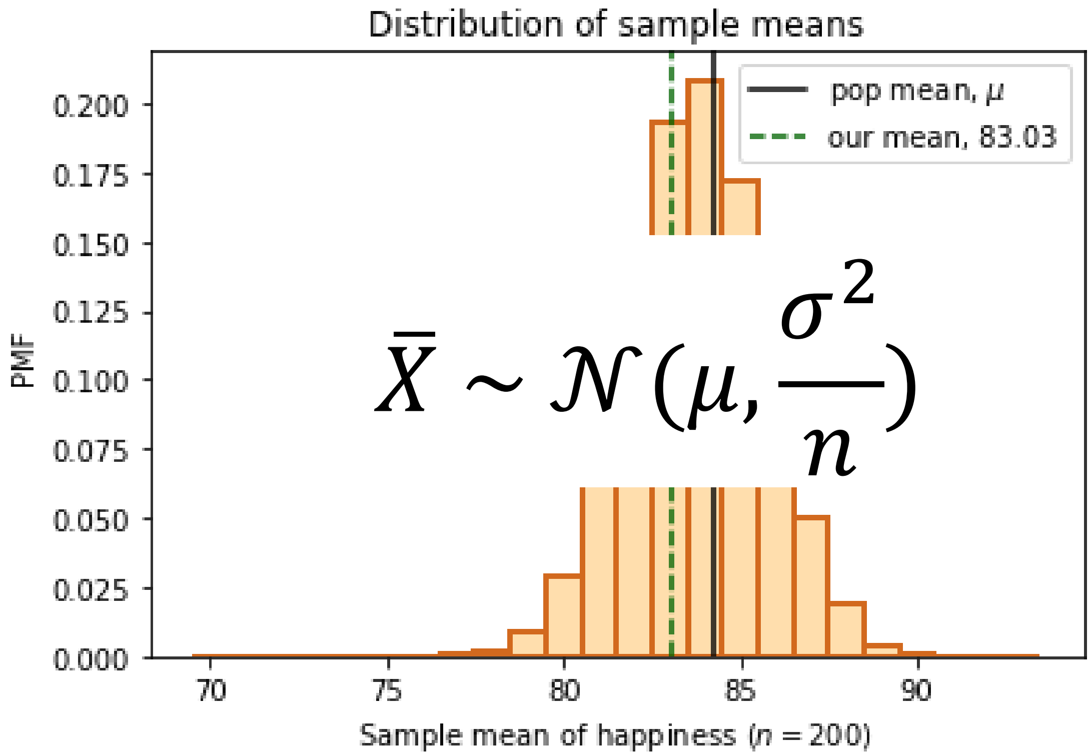
$$E[\bar{X}] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right]$$

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

Sample mean by the CLT



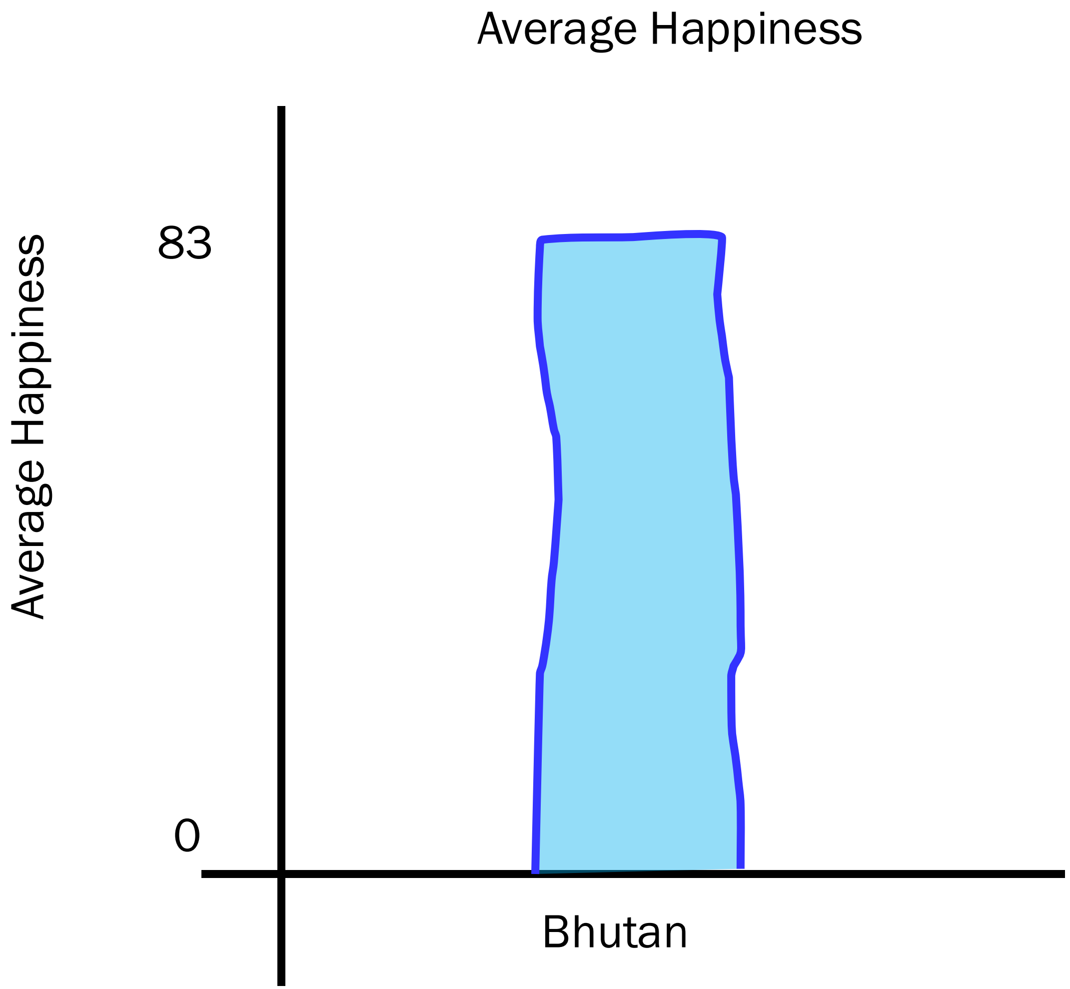
$X_i \sim F$



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

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



Sample Mean:

ith sample

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

Size of the sample

Estimating the population variance

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

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

population mean

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

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

sample mean

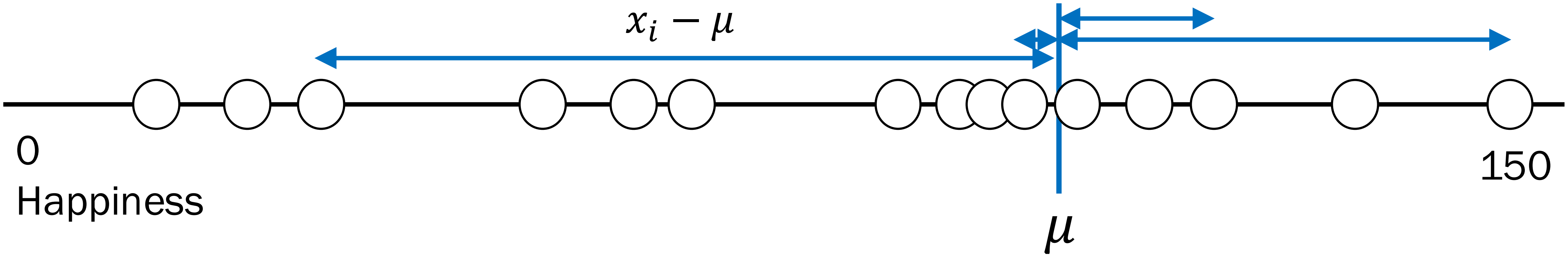
sample variance $$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Intuition about the sample variance, S^2

<u>Actual, σ^2</u>

population mean

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

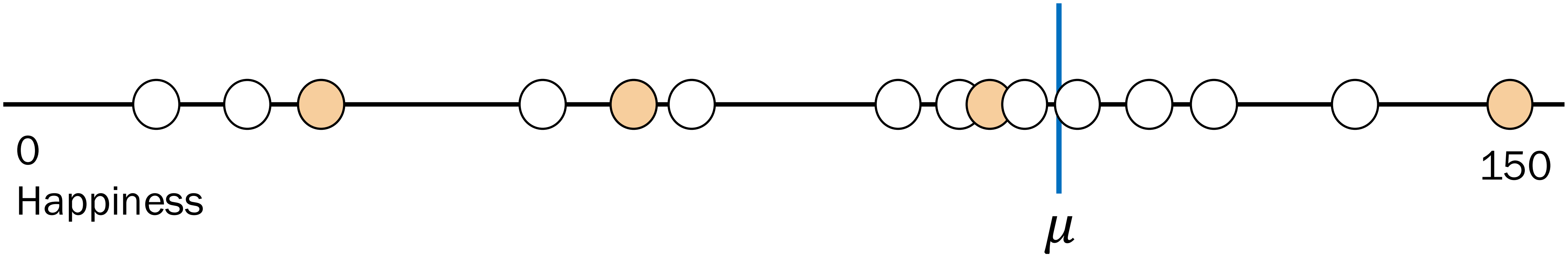


Population size, N

Calculating population statistics <u>**exactly**</u> requires us knowing all N datapoints.

Intuition about the sample variance, S^2

Actual, σ^2	Estimate, S^2
population mean	sample mean
population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	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

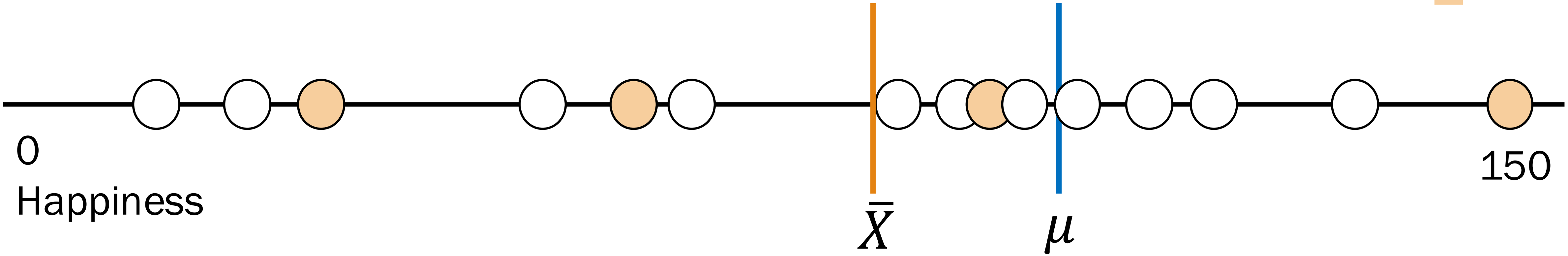
population mean

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

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

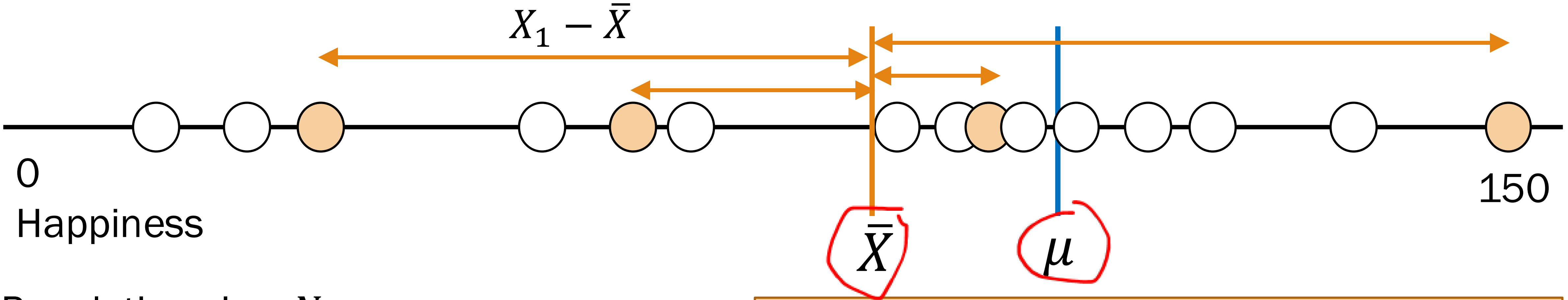
population mean

Estimate, S^2

sample variance

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

sample mean



Population size, N

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 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

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

Bug!

Estimating the population variance

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

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

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, $(X_1, X_2, \ldots, X_n)$:

sample mean

sample variance

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

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

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

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

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

(introduce $\mu - \mu$)

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

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

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

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

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

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

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

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

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

Estimating the population variance

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

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

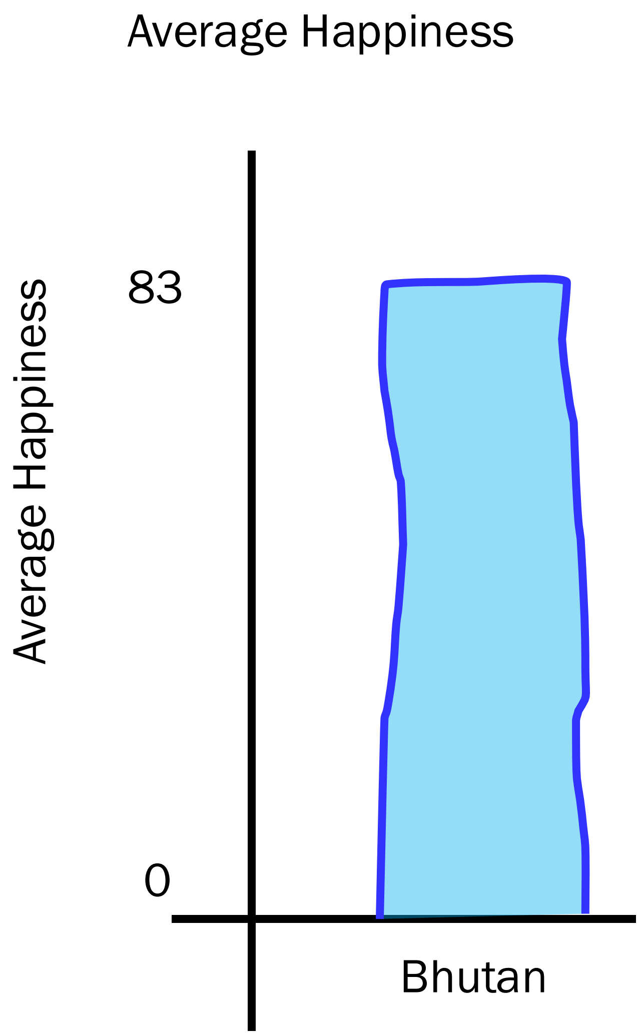
The best estimate of σ^2 is the **sample variance:**

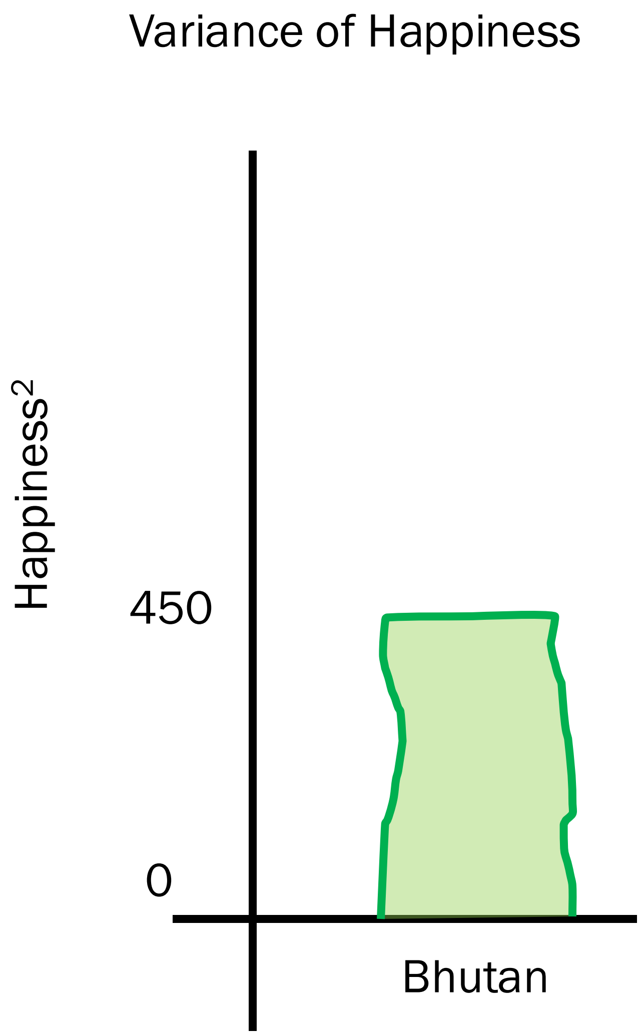
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

S^2 is an **unbiased estimator** of the population variance, σ^2.

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

Our Report to Bhutan Government





Sample Variance:

Sample mean

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

Makes it "unbiased"

Quick check

1. μ, the population mean

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

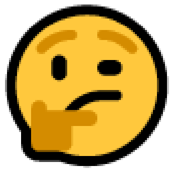
3. σ^2, the population variance

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

5. $\bar{X} = 83$

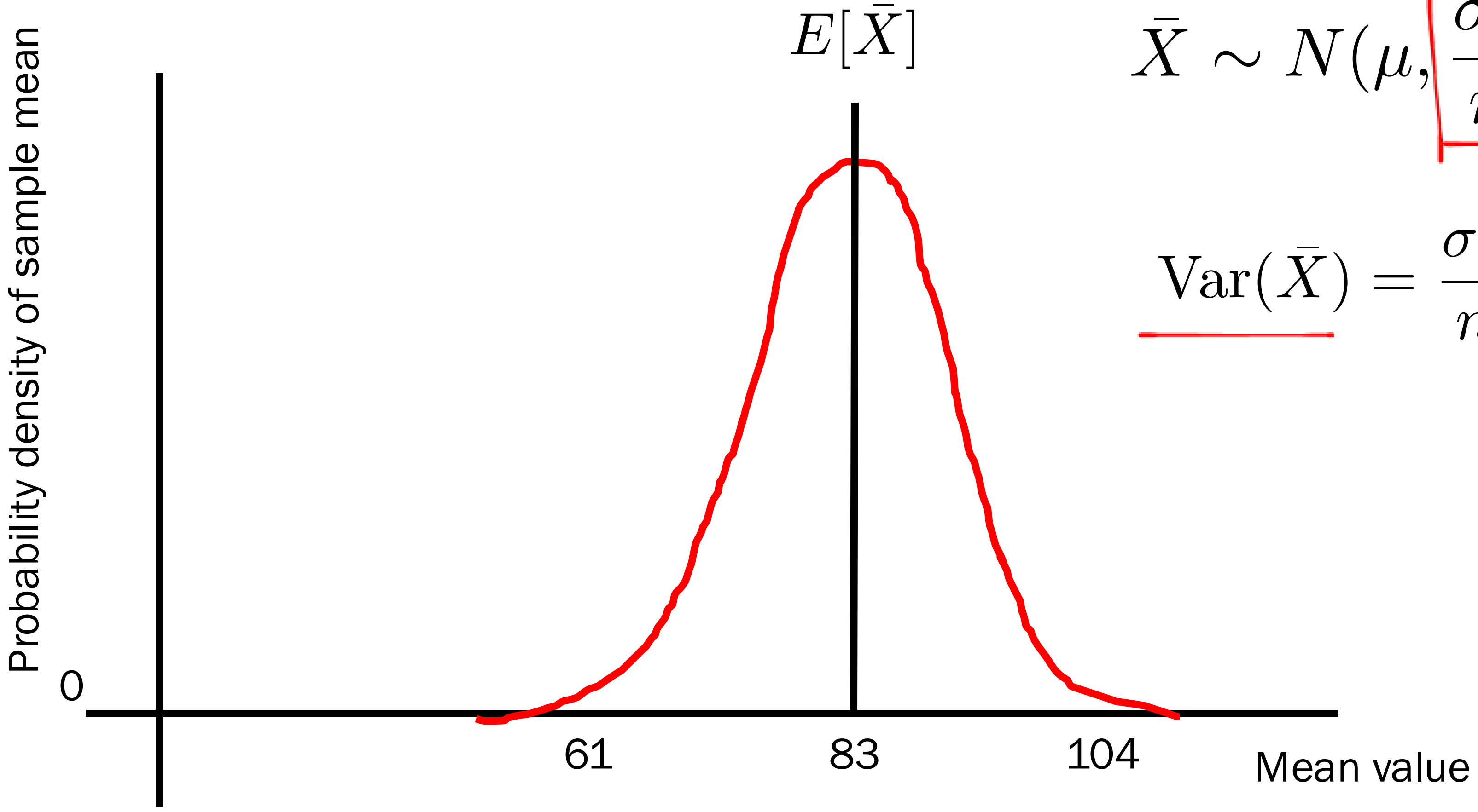
6. $(X_1 = 59, X_2 = 87, X_3 = 94, X_4 = 99,$
   $X_5 = 87, X_6 = 78, X_7 = 69, X_8 = 91)$

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



No Error Bars ☹

Insight: Sample Mean is an RV with known Var



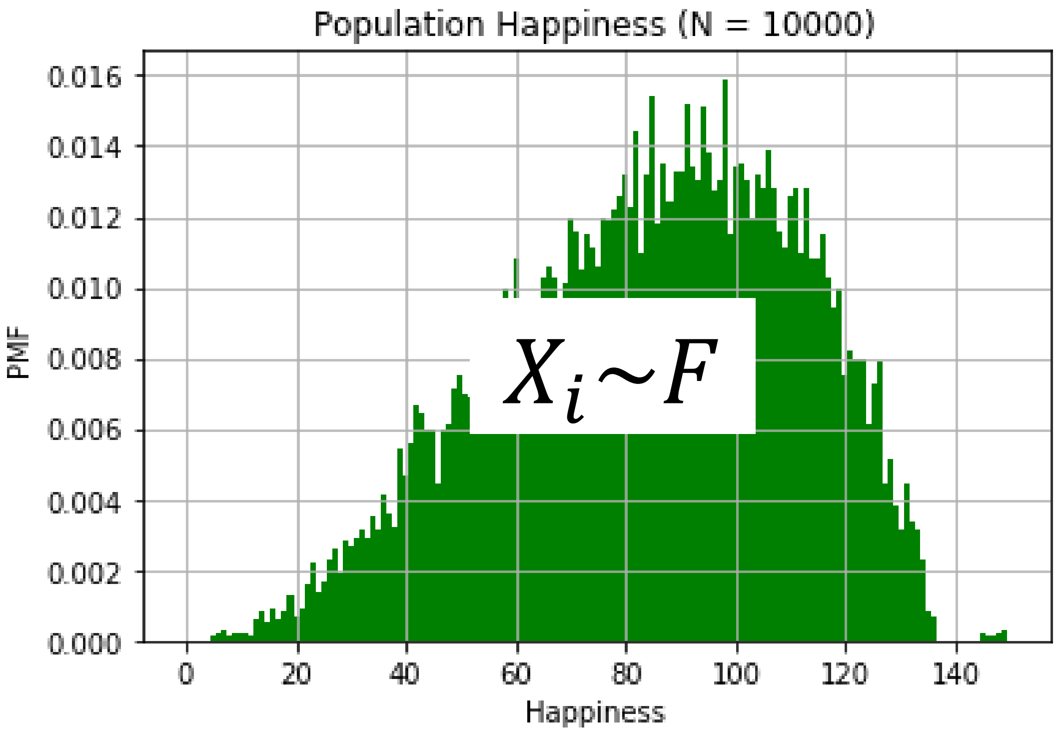
By central limit theorem:

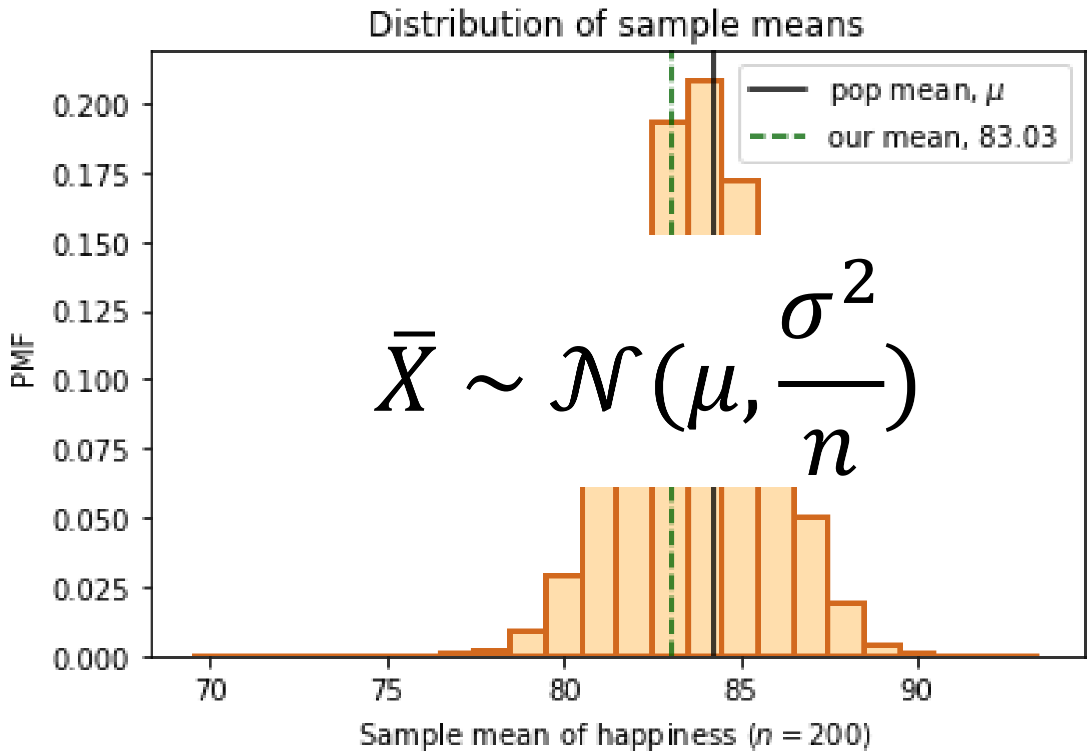
$$\bar{X} \sim N(\mu, \boxed{\frac{\sigma^2}{n}})$$

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

Standard error of the mean

Sample mean





- 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

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

By the Central Limit Theorem

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

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

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

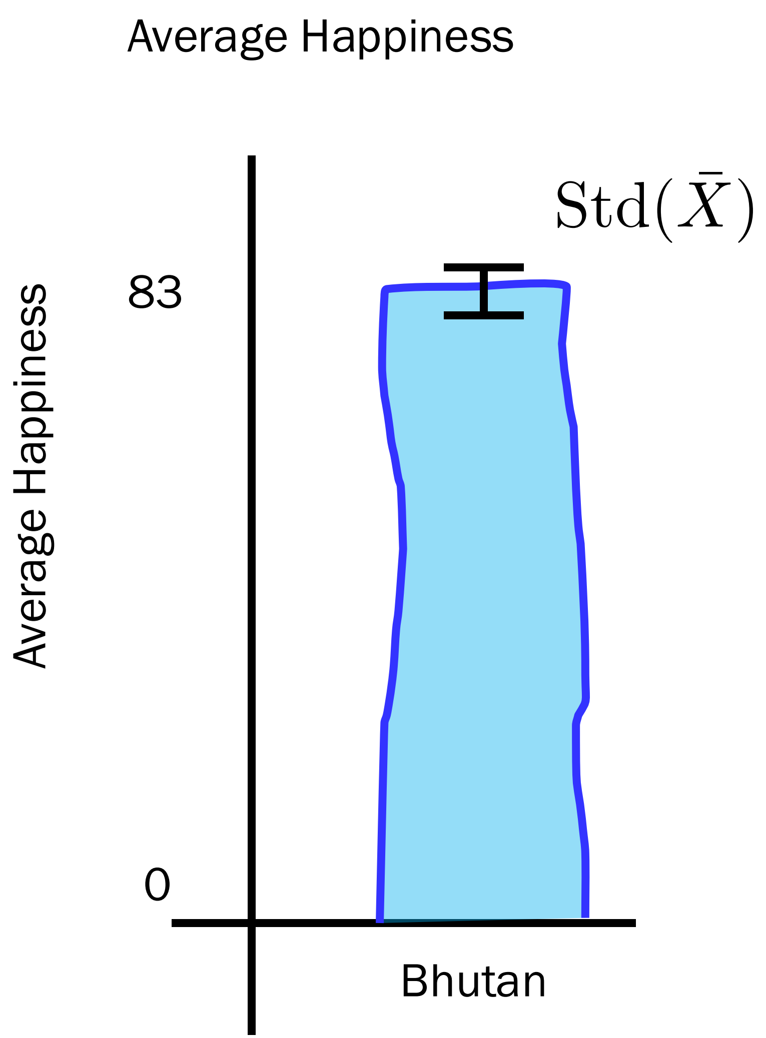
$$= \sqrt{\frac{450}{200}}$$

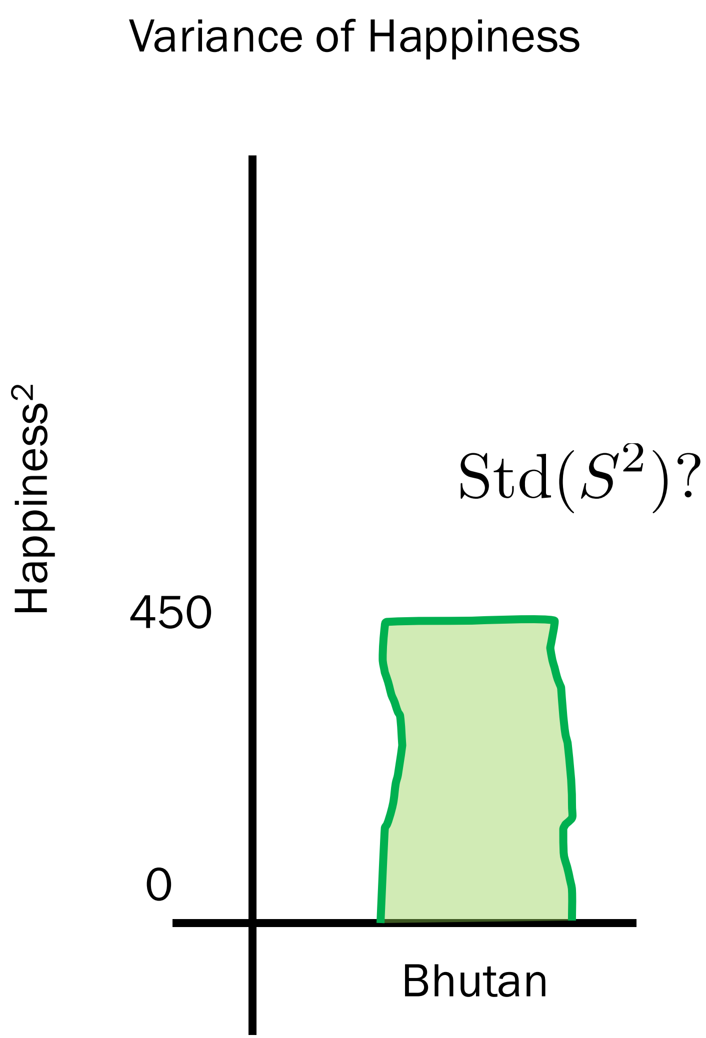
The numbers for our Bhutanese poll

$$= 1.5$$

Bhutanese standard error of the mean

Our Report to Bhutan Government





Claim: The average happiness of Bhutan is 83 ± 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

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

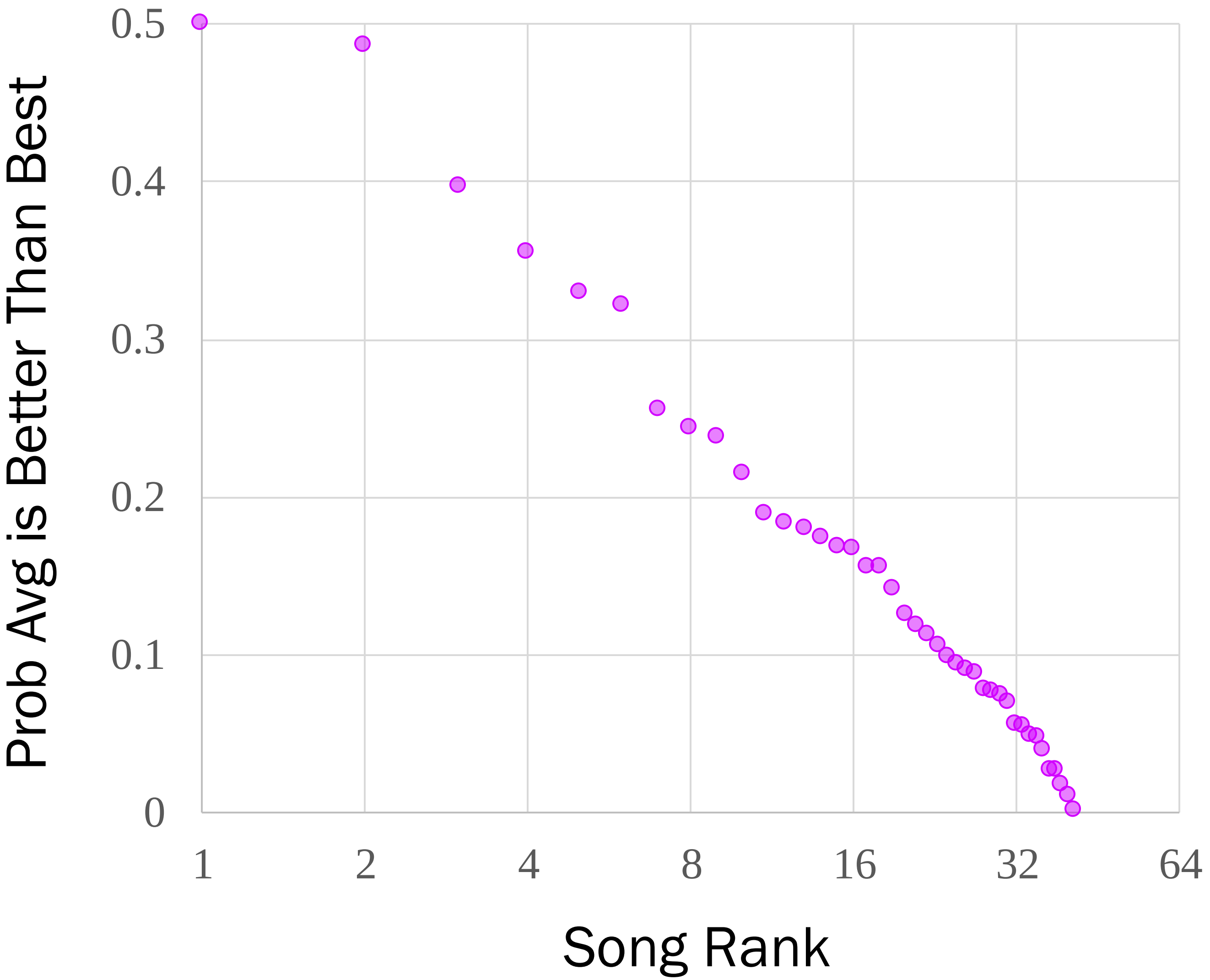
sample variance

How wrong do we think our mean estimate is?

Where Should We Cut Off?

Rank	Sample Mean	Song	NumVotes	AvgScore	SEOM	Pr(Top16)	Pr(Better Than Top)
1	N(μ = 3.81, σ = 0.19)	Daft Punk - Get Lucky	43	3.81	0.19	0.951	0.500
2	N(μ = 3.8, σ = 0.24)	Rascal Flatts - Life is a Highway	35	3.8	0.24	0.904	0.487
3	N(μ = 3.74, σ = 0.19)	The Beatles - Let It Be	39	3.74	0.19	0.905	0.397
4	N(μ = 3.65, σ = 0.31)	Dave Brubeck - Take Five	17	3.65	0.31	0.704	0.330
5	N(μ = 3.7, σ = 0.14)	Jack Johnson - Upside Down	89	3.7	0.14	0.925	0.321
6	N(μ = 3.57, σ = 0.31)	Tame - Let it Happen	21	3.57	0.31	0.610	0.255
7	N(μ = 3.57, σ = 0.29)	Billy Joel - Vienna	23	3.57	0.29	0.615	0.244
8	N(μ = 3.58, σ = 0.26)	Pitbull - Time of Our Lives	24	3.58	0.26	0.640	0.238
9	N(μ = 3.55, σ = 0.27)	ABBA - Voulez-Vous	20	3.55	0.27	0.594	0.215
10	N(μ = 3.6, σ = 0.16)	Earth, Wind & Fire - September	55	3.6	0.16	0.754	0.199
11	N(μ = 3.54, σ = 0.24)	George Michael - Careless Whisper	24	3.54	0.24	0.590	0.189
12	N(μ = 3.54, σ = 0.23)	Grover Washington, Jr. - Just the Two of Us (feat. Bill Withers)	24	3.54	0.23	0.593	0.183
13	N(μ = 3.5, σ = 0.28)	Juna - Clairo	18	3.5	0.28	0.521	0.180
14	N(μ = 3.22, σ = 0.6)	Zach Bryan - Something in the Orange	9	3.22	0.6	0.325	0.174
15	N(μ = 3.41, σ = 0.37)	Smash Mouth - All Star	17	3.41	0.37	0.418	0.168
16	N(μ = 3.4, σ = 0.38)	ILLIT - Magnetic	15	3.4	0.38	0.409	0.167
17	N(μ = 3.5, σ = 0.24)	Queen - We Are The Champions	22	3.5	0.24	0.522	0.156
18	N(μ = 3.38, σ = 0.38)	Otis Redding - The Dock of the Bay	13	3.38	0.38	0.387	0.156
19	N(μ = 3.41, σ = 0.32)	Portugal. The Man - Feel it Still	17	3.41	0.32	0.407	0.141
20	N(μ = 3.27, σ = 0.43)	Bolden - Dawn in LA	11	3.27	0.43	0.306	0.125
21	N(μ = 3.38, σ = 0.31)	Queen - don't stop me now	16	3.38	0.31	0.367	0.118
22	N(μ = 3.36, σ = 0.32)	Kid Cudi - Pursuit of Happiness	14	3.36	0.32	0.343	0.113
23	N(μ = 3.22, σ = 0.43)	Carly Rae Jepsen - Let's Get Lost	9	3.22	0.43	0.266	0.105
24	N(μ = 3.25, σ = 0.39)	Sabrina Carpenter - Please Please Please	12	3.25	0.39	0.272	0.098
25	N(μ = 3.18, σ = 0.44)	Djo - End of the Beginning	11	3.18	0.44	0.239	0.094
26	N(μ = 3.22, σ = 0.4)	Empire Of The Sun - High and Low	9	3.22	0.4	0.250	0.091
27	N(μ = 3.15, σ = 0.45)	Chappell Roan - HOT TO GO!	13	3.15	0.45	0.228	0.088

Where Should We Cut Off?



Have 5 weeks left

Great Practice

4. Song of the Quarter [25 points]

This quarter in CS109 there were 167 songs that were voted on. For each song, we have a list of votes where each vote is an integer in the set $\{1, 2, 3, 4, 5\}$. We assume all votes for a song are IID samples from the "true" distribution of CS109 opinion on the song.

For each song i we have m_i votes stored in a list **`votes[i]`** = $[x_1, x_2, \ldots, x_{m_i}]$. We have already calculated:

$$\mu_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_j \qquad \text{using } \texttt{np.mean(votes[i])}$$

$$\text{var}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} (x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i])}$$

$$\text{svar}_i = \frac{1}{m_i - 1}\sum_{j=1}^{m_i} (x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i], ddof=1)}$$

a. (7 points) Song 1 has $m_1 = 45$ votes. We have calculated:

$\mu_1 = 3.82$ $\quad$ $\text{var}_1 = 1.4$ $\quad$ $\text{svar}_1 = 1.5$

Estimate the probability that the true average rating for song 1 is less than 3.

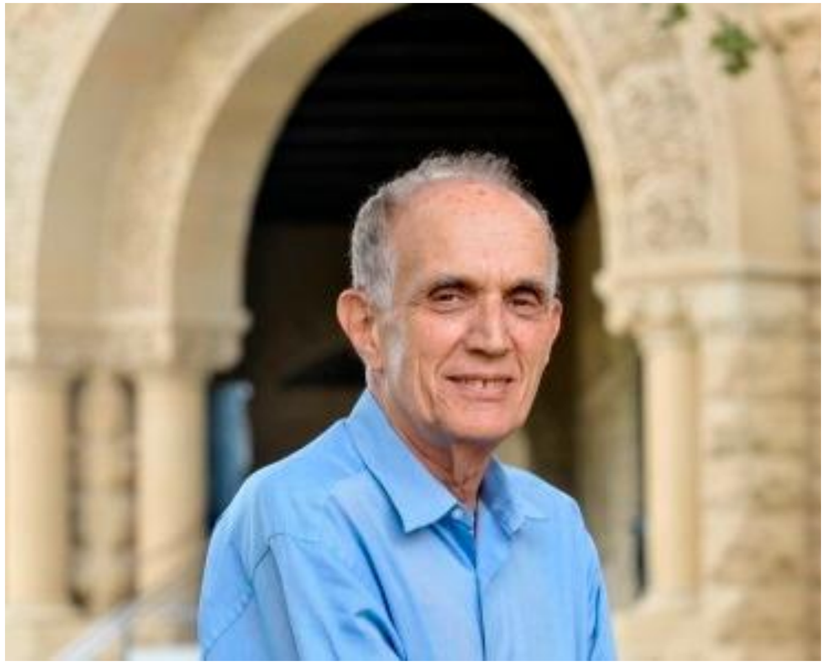
b. (8 points) Song 1 has $m_1 = 45$ votes. Song 2 has $m_2 = 36$ votes. We have calculated:

Song 1:	$\mu_1 = 3.82$	$\text{var}_1 = 1.4$	$\text{svar}_1 = 1.5$
Song 2:	$\mu_2 = 3.79$	$\text{var}_2 = 1.7$	$\text{svar}_2 = 1.8$

What is the probability that the true average of Song 1 is greater than the true average for Song 2?

Bootstraping

One of the Most Important Ideas in Modern Statistics!



Invented bootstrapping in 1979

Still a professor at Stanford

Won a National Science Medal

Bootstraping allows you to:

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

If we have extra time…

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!

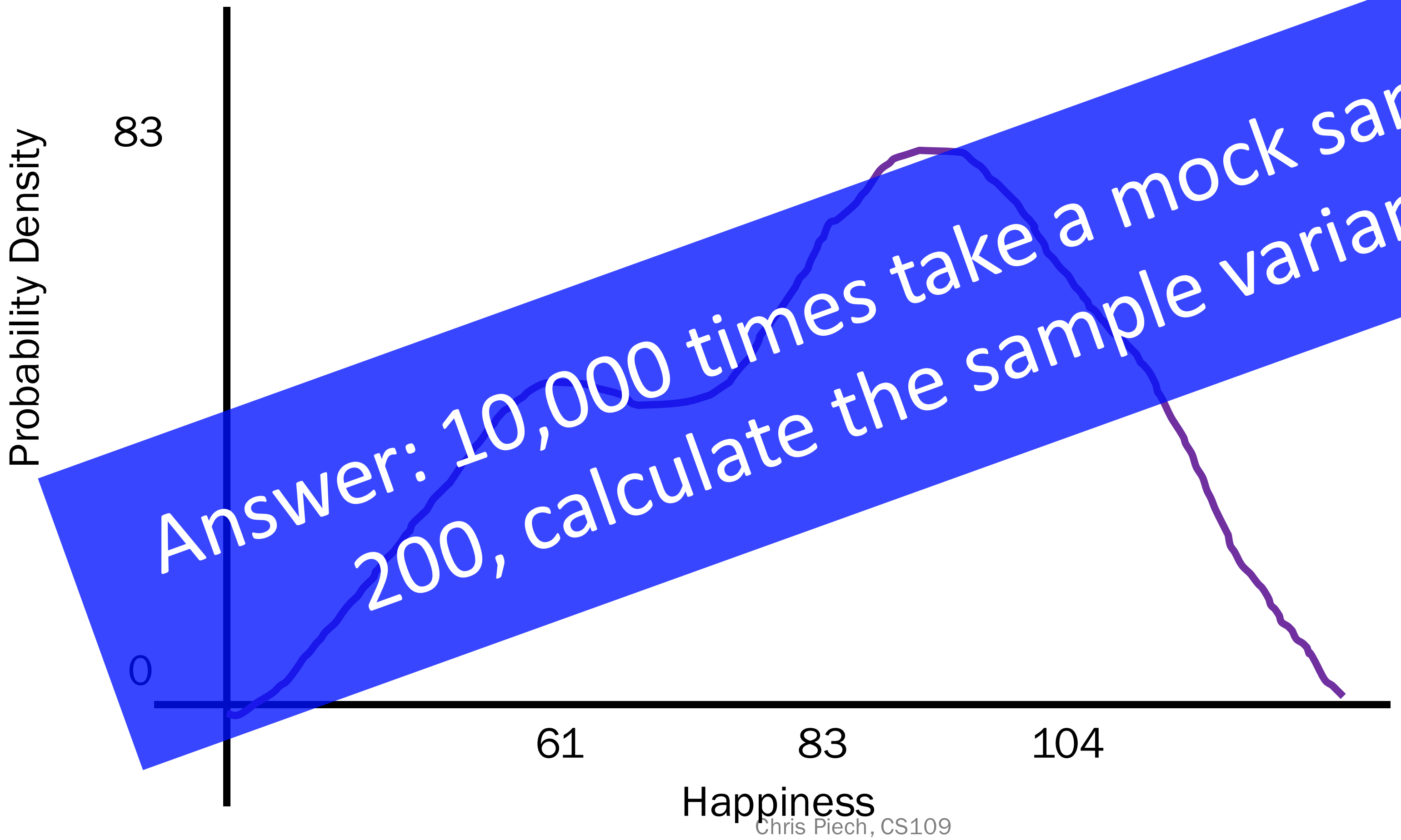
What is the **std** of the **sample variance**, calculated from 200 people?

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

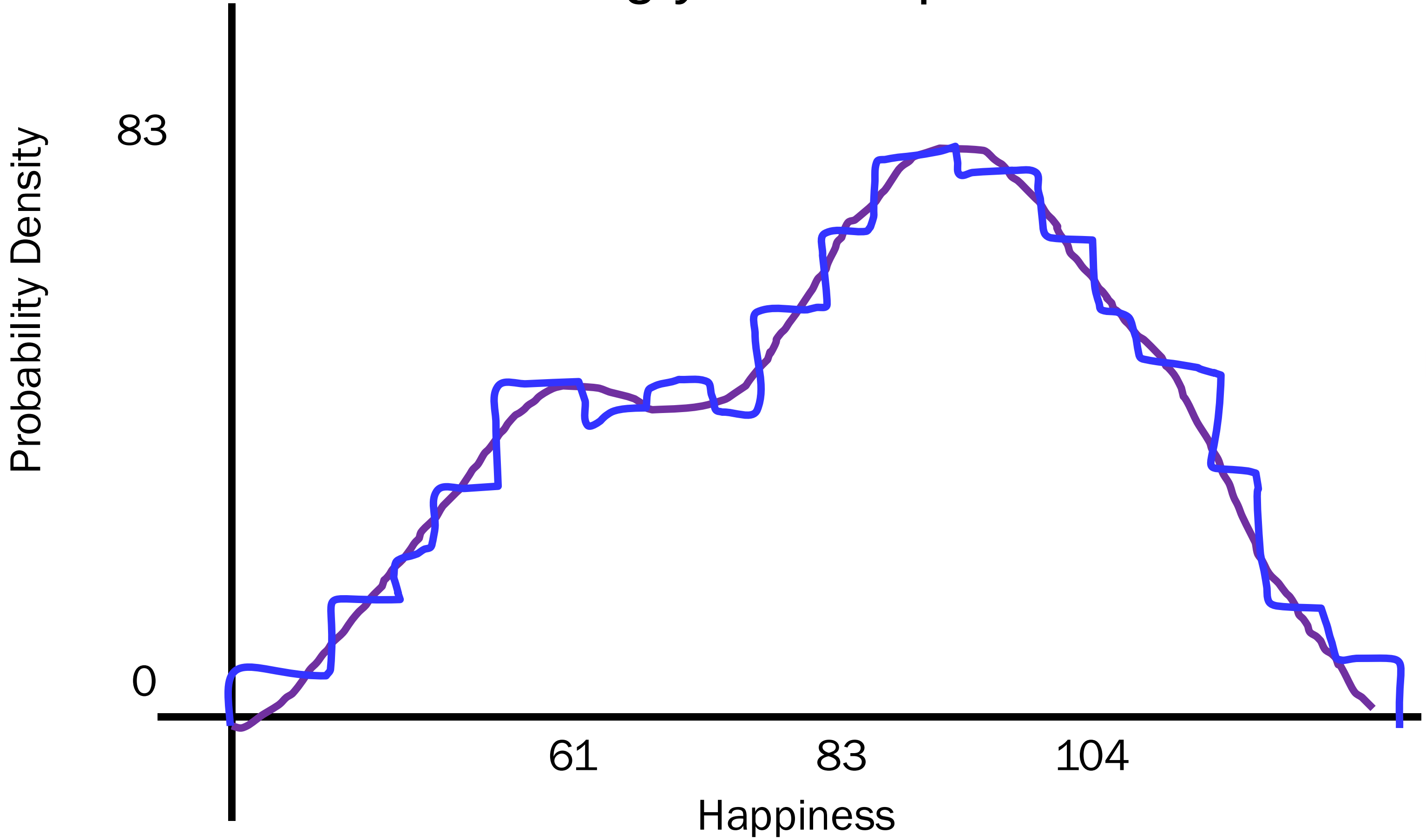


Answer: 10,000 times take a mock sample of 200, calculate the sample variance

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!!

To be continued!