19: Sampling

Lisa Yan

November 4, 2019

Our current trajectory for this course

Start

Counting

Events

Random variables

Joint distributions

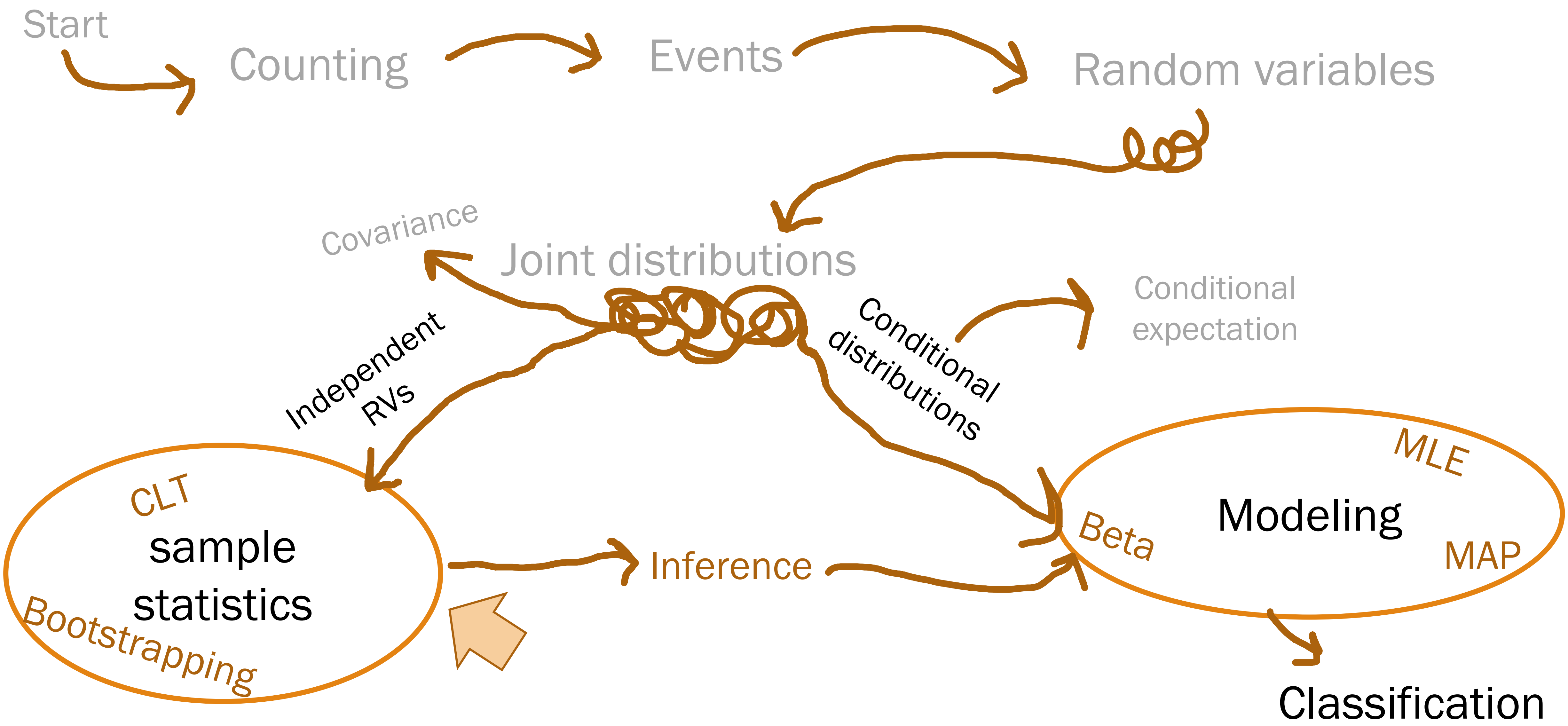
Covariance

Independent RVs

Conditional distributions

Conditional expectation

Our current trajectory for this course



Today's plan

Finishing CLT

Sampling definitions

Unbiased estimates of population statistics

Bootstrapping
- For a statistic
- For a p-value

Working with the CLT

Let $X_1, X_2, \dots, X_n$ i.i.d., where $E[X_i] = \mu$, Var$(X_i) = \sigma^2$. As $n \to \infty$:

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

Sum of i.i.d. RVs

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

Average of i.i.d. RVs
(sample mean)

⚠

If X_i is discrete:

Use the continuity correction on Y!

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

Dice game

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

You will roll 10 6-sided dice$(X_1, X_2, \ldots, X_{10})$.

- Let $X = X_1 + X_2 + \cdots + X_{10}$, the total value of all 10 rolls.
- You win if $X \le 25$ or $X \ge 45$.

And now the truth (according to the CLT)…

1. Define RVs and state goal.
2. Solve.

$E[X_i] = 3.5$,
$\mathrm{Var}(X_i) = 35/12$
$X \approx Y \sim \mathcal{N}(10(3.5), 10(35/12))$

Want:

$P(X \le 25 \text{ or } X \ge 45)$

A. $P(25 \le Y \le 45)$
B. $P(Y \le 25.5) + P(Y \ge 44.5)$
C. $1 - P(25 \le Y \le 45)$
D. $1 - P(25.5 \le Y \le 44.5)$

Dice game

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

You will roll 10 6-sided dice$(X_1, X_2, \ldots, X_{10})$.

- Let $X = X_1 + X_2 + \cdots + X_{10}$, the total value of all 10 rolls.
- You win if $X \le 25$ or $X \ge 45$.

And now the truth (according to the CLT)…

1. Define RVs and state goal.

$E[X_i] = 3.5$,
$\text{Var}(X_i) = 35/12$
$X \approx Y \sim \mathcal{N}(10(3.5), 10(35/12))$

Want:

$P(X \le 25 \text{ or } X \ge 45)$

$\approx P(Y \le 25.5) + P(Y \ge 44.5)$

2. Solve.

$$P(Y \le 25.5) + P(Y \ge 44.5)$$

$$= \Phi\left(\frac{25.5 - 35}{\sqrt{10(35/12)}}\right) + \left(1 - \Phi\left(\frac{44.5 - 35}{\sqrt{10(35/12)}}\right)\right)$$

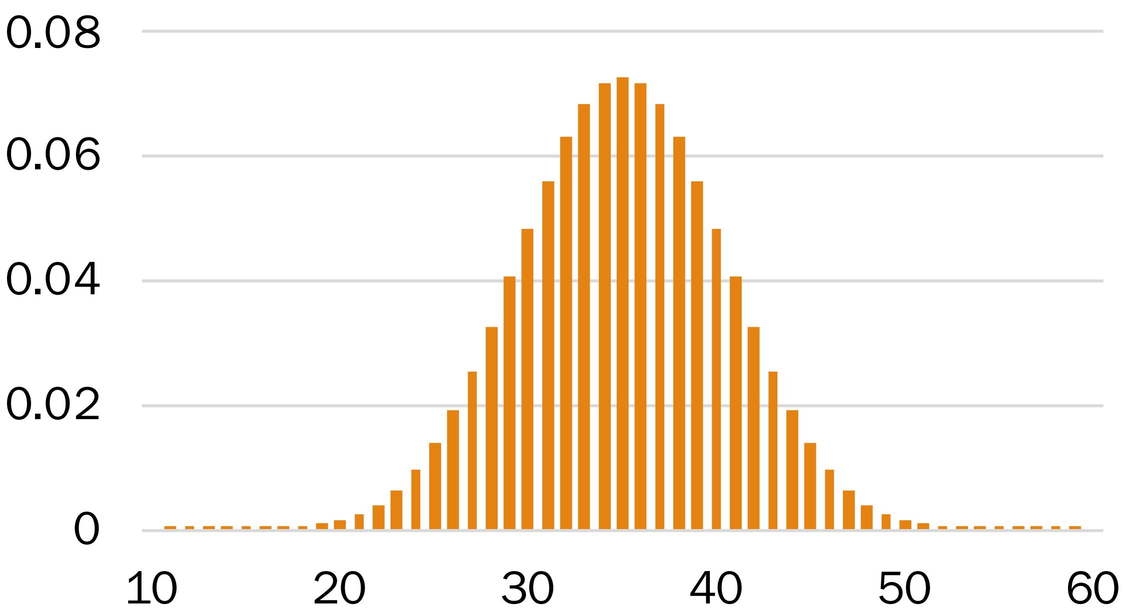
$$\approx \Phi(-1.76) + \left(1 - \Phi(1.76)\right)$$
$$\approx (1 - 0.9608) + (1 - 0.9608)$$
$$= 0.0784$$

Dice game

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

You will roll 10 6-sided dice$(X_1, X_2, \ldots, X_{10})$.

- Let $X = X_1 + X_2 + \cdots + X_{10}$, the total value of all 10 rolls.
- You win if $X \leq 25$ or $X \geq 45$.



(by CLT)

$$P(X \leq 25 \text{ or } X \geq 45) \approx$$
$$P(Y \leq 25.5) + P(Y \geq 44.5)$$
$$\approx 0.0784$$

(by computer)

$$P(X \leq 25 \text{ or } X \geq 45) \approx 0.0780$$

Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (i.i.d. trials):

- $X_i =$ runtime of i-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of n trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with 95% certainty?

1. Define RVs and state goal.
2. Solve.

(CLT) $\bar{X} \sim \mathcal{N}\left(t, \frac{4}{n}\right)$

Want: $P(t - 0.5 \le \bar{X} \le t + 0.5) = 0.95$

(linear transform of a normal) $\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$

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

Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (i.i.d. trials):

- X_i = runtime of i-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of n trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with 95% certainty?

1. Define RVs and state goal.

$\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$

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

2. Solve.

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

$$= \Phi\left(\frac{0.5 - 0}{\sqrt{4/n}}\right) - \Phi\left(\frac{-0.5 - 0}{\sqrt{4/n}}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$

Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (i.i.d. trials):

- X_i = runtime of i-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of n trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with 95% certainty?

1. Define RVs and state goal.

$$\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$$

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

2. Solve.

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

$$= \Phi\left(\frac{0.5 - 0}{\sqrt{4/n}}\right) - \Phi\left(\frac{-0.5 - 0}{\sqrt{4/n}}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$

$$0.975 = \Phi(\sqrt{n}/4)$$

$$\sqrt{n}/4 = \Phi^{-1}(0.975) \approx 1.96 \quad \Rightarrow \quad n \approx 62$$

The Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ i.i.d., where $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

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

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

The Central Limit Theorem allows you to calculate probabilities on sums and means of i.i.d. random variables.

What if we don't know μ or σ^2?

How do we estimate μ and σ^2 from data?

Today's plan

Finishing CLT

Sampling definitions
- Population mean/variance, sample mean/variance
- Standard error

Bootstrapping
- For a statistic
- For a p-value (next time)

Motivating example

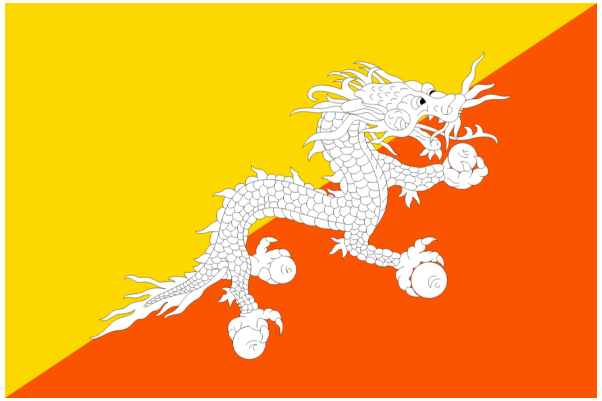
You want to know the true mean and variance of happiness in Bhutan.

- 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?



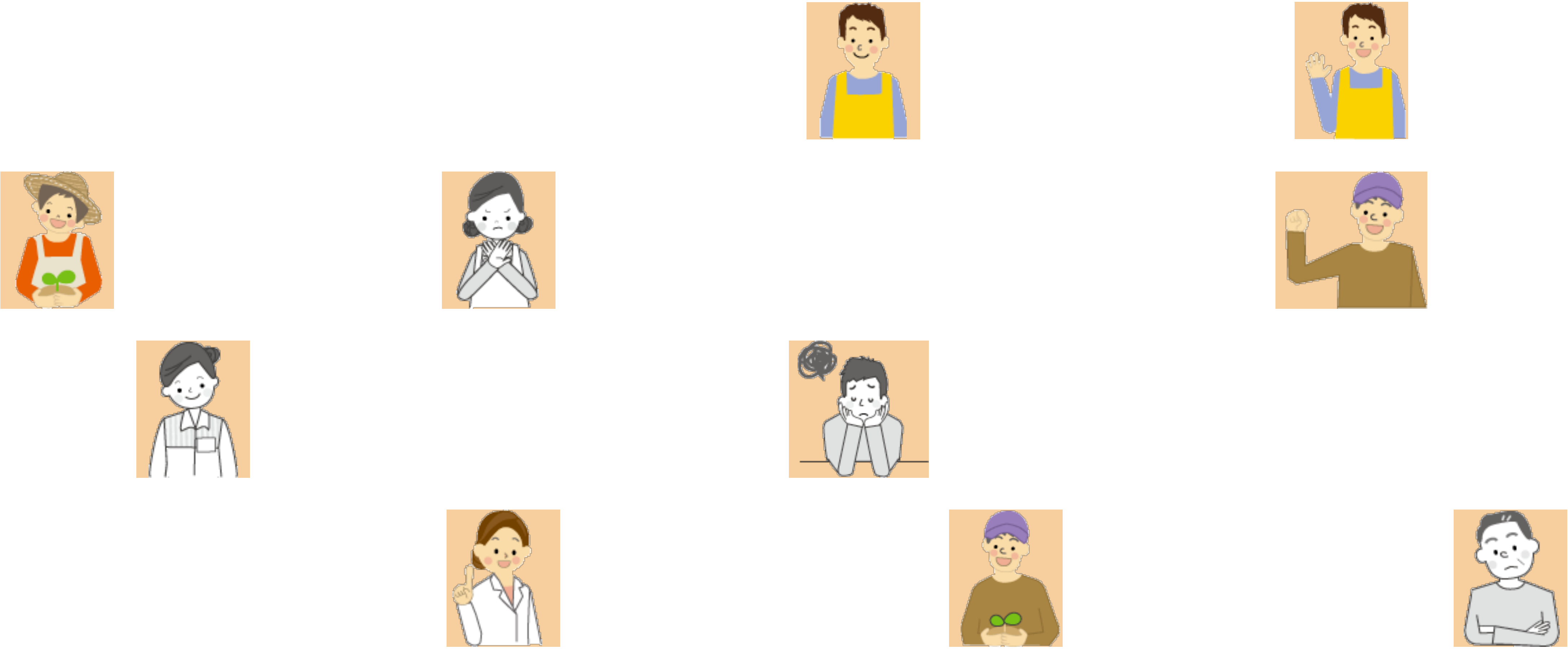
Population

This is a population.

Sample

A **sample** is selected from a population.

Sample



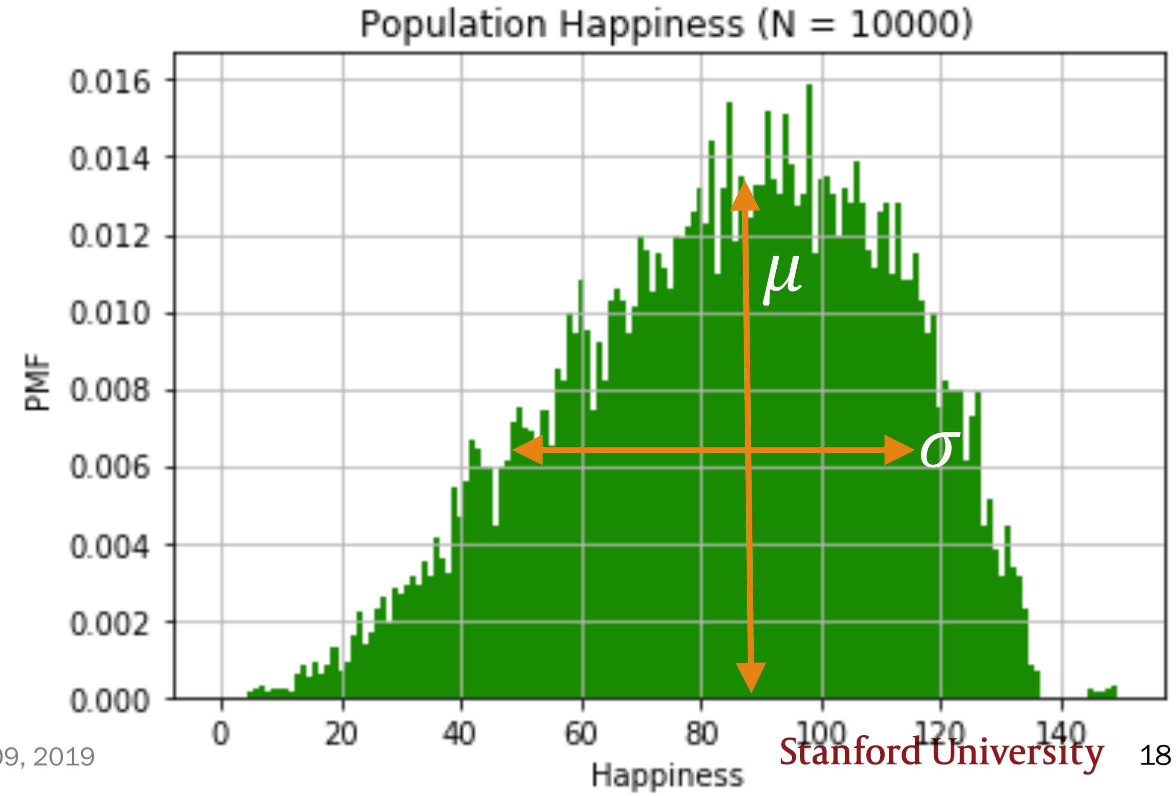
A sample is selected from a population.

A sample, mathematically

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

The sequence $X_1, X_2, \dots, X_n$ 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 $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$



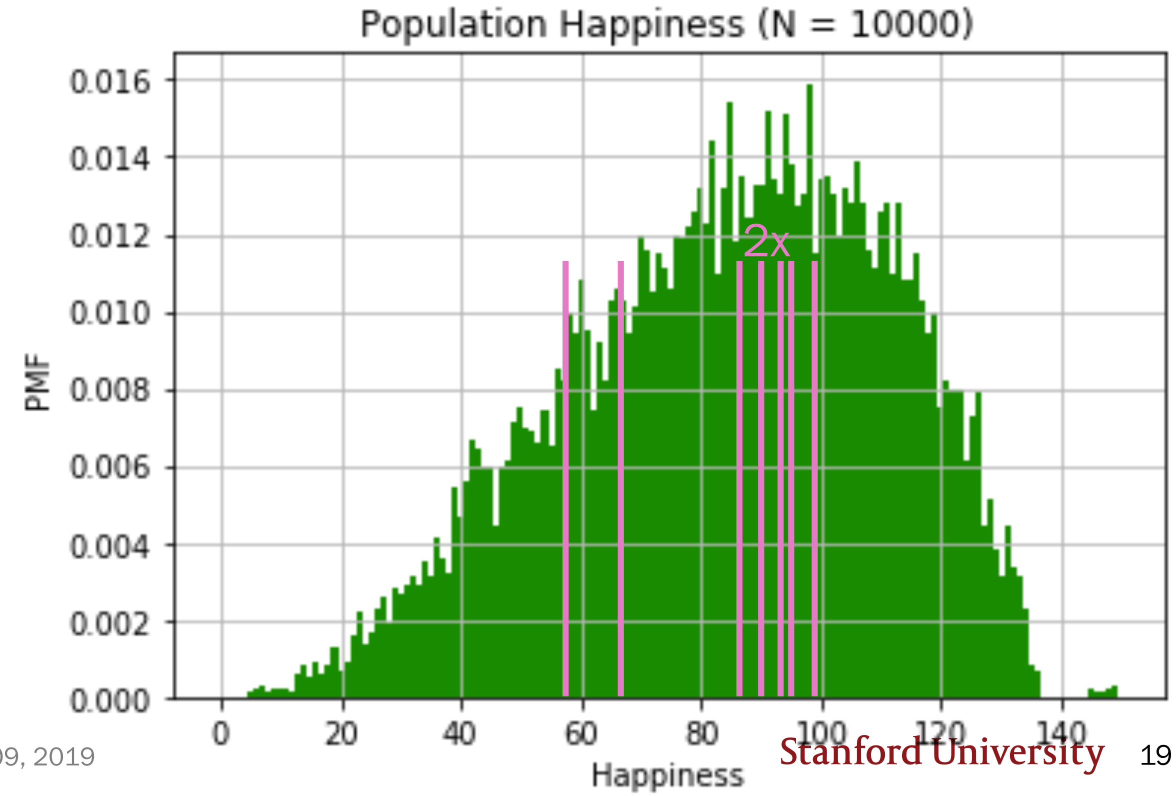
A sample, mathematically

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:

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



Population statistics

A happy
Bhutanese person

The underlying distribution F has unknown statistics:

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

Estimating the population mean

1. What is μ, the **mean happiness** of Bhutanese people?

What if you only have a sample, $(X_1, X_2, \ldots, X_n)$?

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

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

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

1. Take multiple samples of size n
2. For each sample, compute sample means
3. On average, we would get the population mean

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 (frequentist interpretation)
C. Event

Quick check

A. Random variable(s)
B. Value (frequentist interpretation)
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)

These are outcomes from your collected data.

Estimating the population mean

1. What is μ, the **mean happiness** of Bhutanese people?

What if you 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 estimate** of the population mean, μ:

<u>def</u> $E[\text{estimate}] = \text{actual}$

Proof 1: By CLT, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Proof 2: By linearity of expectation (see board)

Break for jokes/ announcements

Announcements

<u>Problem Set 4</u>

Due: Wednesday 11/6
Covers: Up to Law of Total Expectation

Late day reminder: No late days permitted past last day of the quarter, 12/7

Announcements: CS109 contest

Do something cool and creative with probability

Genuinely optional extra credit

Due Monday 12/2, 11:59pm

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

Population size, N

But what if you only have a sample, $(X_1, X_2, \dots, X_n)$?

Estimating the population variance

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

What if you only have a sample, $(X_1, X_2, \dots, X_n)$?

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 estimate** of the population variance, σ^2:

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

If you only have a sample, you can only compute <u>estimates</u> of population statistics.

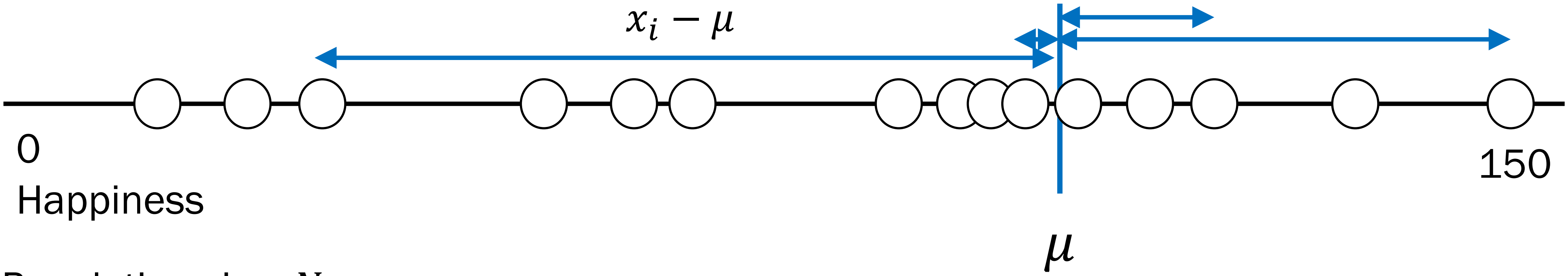
You can only believe what you see.

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

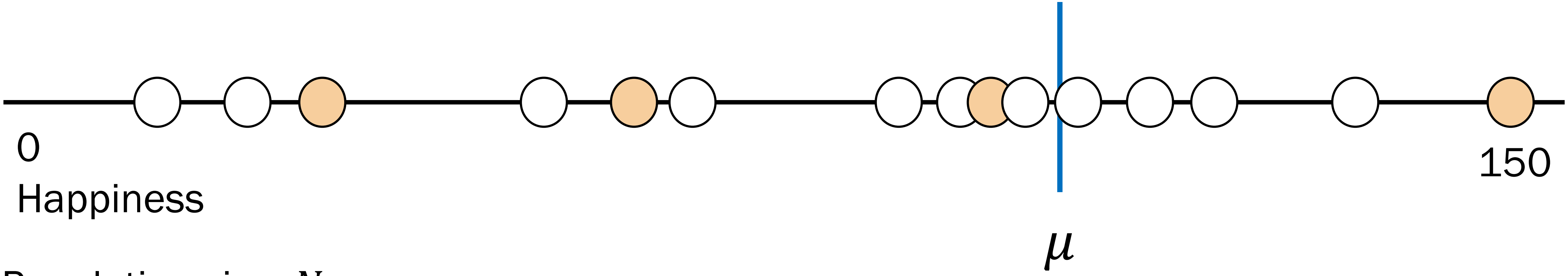
population mean

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

Estimate, S^2

sample mean

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



Population size, N

Intuition about the sample variance, S^2

<u>Actual, σ^2</u>

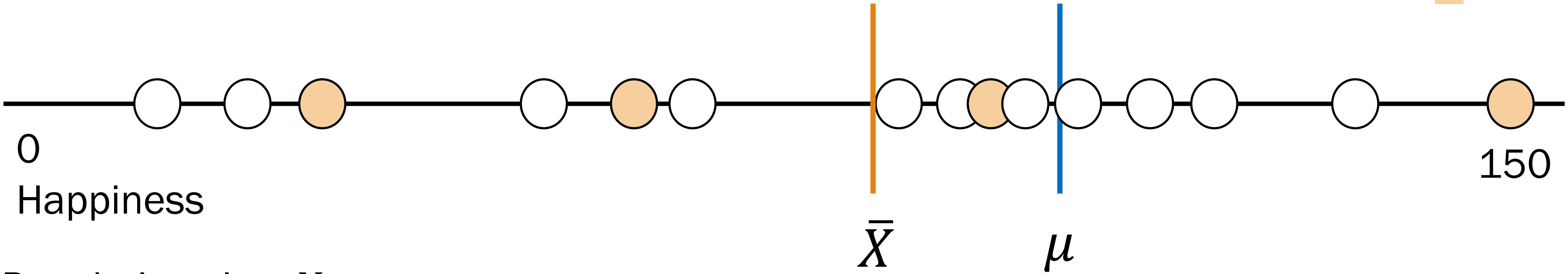
population mean

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

<u>Estimate, S^2</u>

sample mean

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



Population size, N

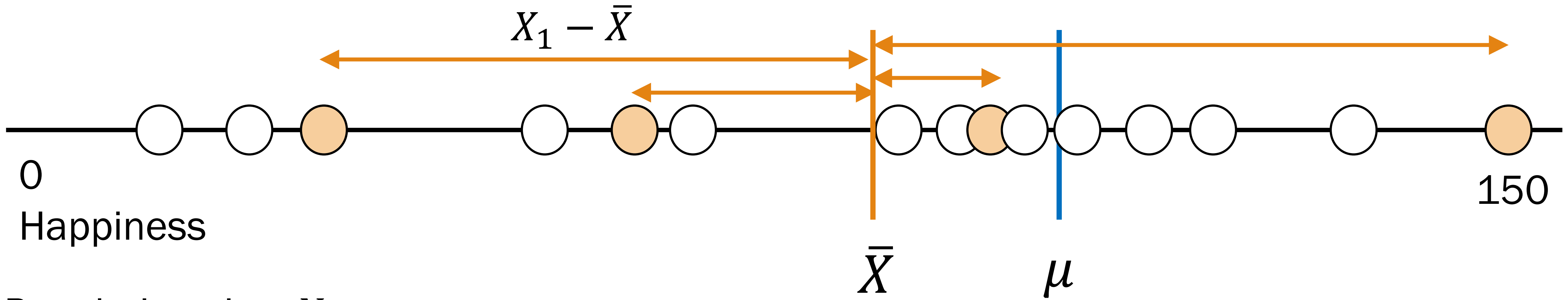
Intuition about the sample variance, S^2

Actual, σ^2

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

Estimate, S^2

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



Population size, N

Sample variance is an estimate using an estimate, so it needs additional scaling.

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

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

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

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

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

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

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

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

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

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

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

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

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

Today's plan

Finishing CLT

Sampling definitions

- Population mean/variance, sample mean/variance
- Standard error

Bootstrapping

- For a statistic
- For a p-value (next time)

Estimating population statistics

1. Collect a sample, $X_1, X_2, \ldots, X_n$. $(72, 85, 79, 79, 91, 68, \ldots, 71)$ $n = 200$

2. Compute **sample mean**, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. $\bar{X} = 83$

3. Compute sample deviation, $X_i - \bar{X}$. $(-11, 2, -4, -4, 8, -15, \ldots, -12)$

4. Compute **sample variance**, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. $S^2 = 793$

How "close" are our estimates $\bar{X}$ and S^2?

How "close" is our estimate $\bar{X}$ to μ?

We know that the sample mean $\bar{X}$ is an unbiased estimate of μ:

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

⚠️ Just knowing the average value of $\bar{X}$ does not inform what the **spread** (e.g., standard deviation) of $\bar{X}$ is.

What is $\text{Var}(\bar{X})$?

A. σ^2, population variance
B. S^2, sample variance
C. σ^2/n, population variance divided by sample size
D. Don't know

🤔

How "close" is our estimate $\bar{X}$ to μ?

We know that the sample mean $\bar{X}$ is an unbiased estimate of μ:

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

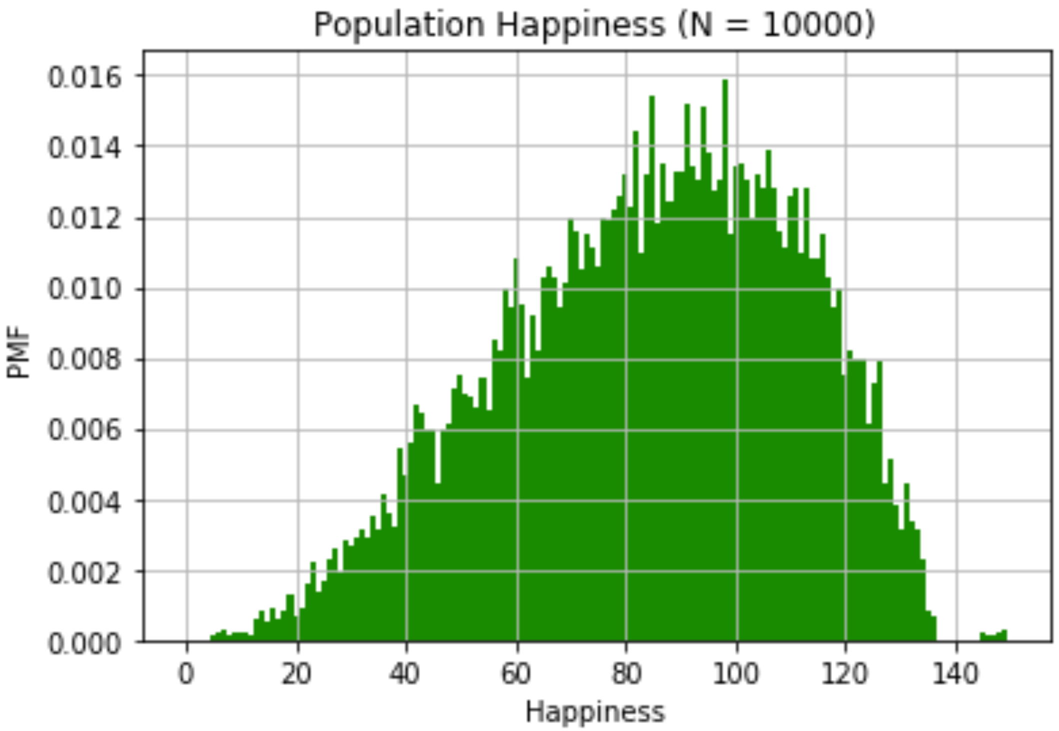
⚠️ Just knowing the average value of $\bar{X}$ does not inform what the **spread** (e.g., standard deviation) of $\bar{X}$ is.

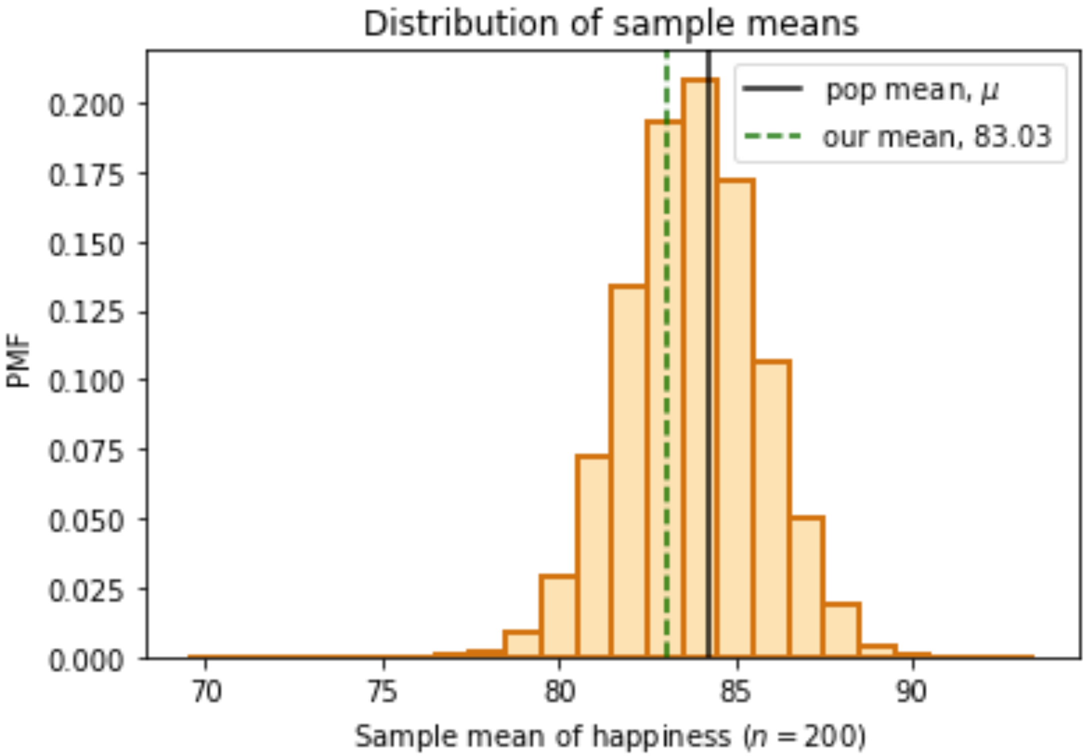
What is Var$\left(\bar{X}\right)$?

A. σ^2, population variance
B. S^2, sample variance
C. σ^2/n, population variance divided by sample size
D. Don't know

🤔

Sample mean





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

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

How "close" is our estimate $\bar{X}$ to μ?

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

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

We want to estimate this

<u>def</u> The **standard error** of the mean is an unbiased 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 $\sigma^2/n = \text{Var}(\bar{X})$
- $\sqrt{S^2/n}$ is an unbiased estimate of $\sqrt{\text{Var}(\bar{X})}$

Standard error

1. Mean happiness:

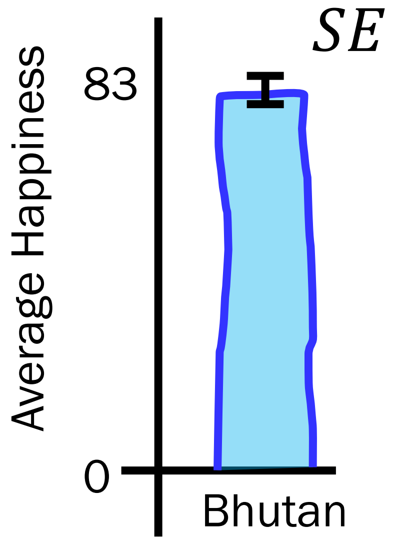
   Claim: The average happiness of Bhutan is 83, with a standard error of 1.99.

this is our best estimate of μ

this is how close we are

Closed form: $SE = \sqrt{\dfrac{S^2}{n}}$ ✅

error bars

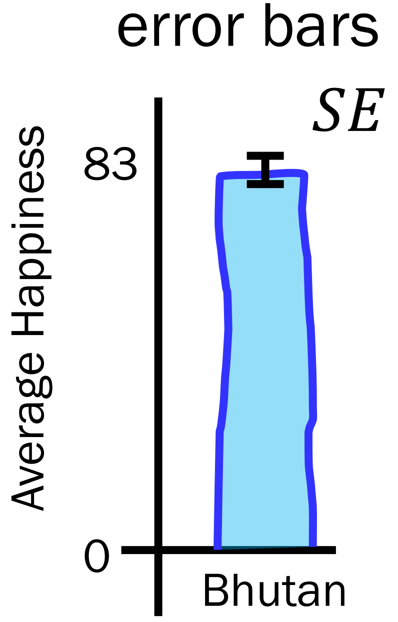


Standard error

1. Mean happiness:

Claim: The average happiness of Bhutan is 83, with a standard error of 1.99.

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



2. Variance of happiness:

this is our best estimate of σ^2

Claim: The variance of happiness of Bhutan is 793.

Closed form: Not covered in CS109

But how close are we?

Up next: Compute Statistics with code!

Today's plan

Finishing CLT

Sampling definitions

- Population mean/variance, sample mean/variance
- Standard error

Bootstrapping

- For a statistic
- For a p-value (next time)

Bootstrap

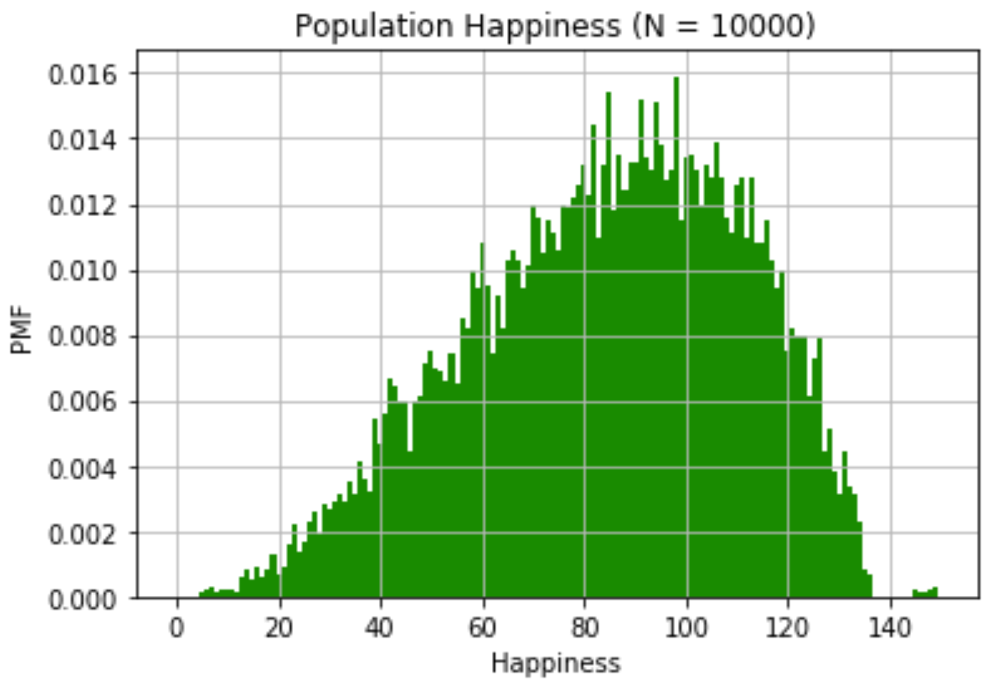
The Bootstrap:

Probability for Computer Scientists

Allows you to do the following:

- Calculate distributions over statistics
- Calculate p values

Bootstrap



Hypothetical questions:

- What is the probability that a Bhutanese peep is just straight up loving life?
- What is the probability that the mean of a subsample of 200 people is within the range 81 to 85?
- What is the variance of the sample variance of subsamples of 200 people?

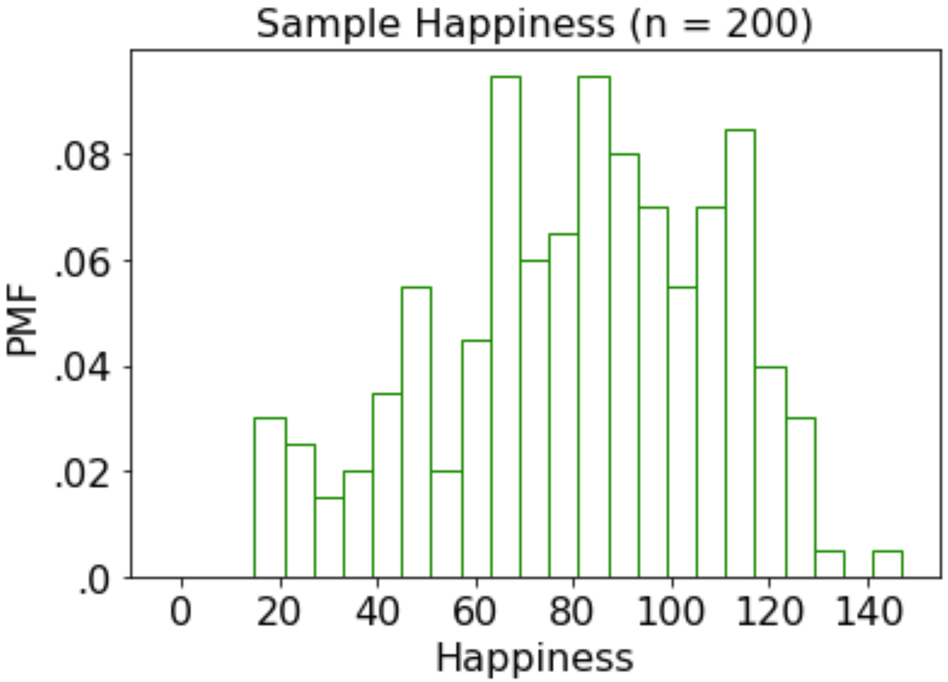
Key insight

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

*This is just a histogram of your data!

72
85
79
79
91
68
...
71

i.i.d. samples



Sample distribution