CLT and Sampling

Will Song
Slides by Chris Piech
CS109, Stanford University

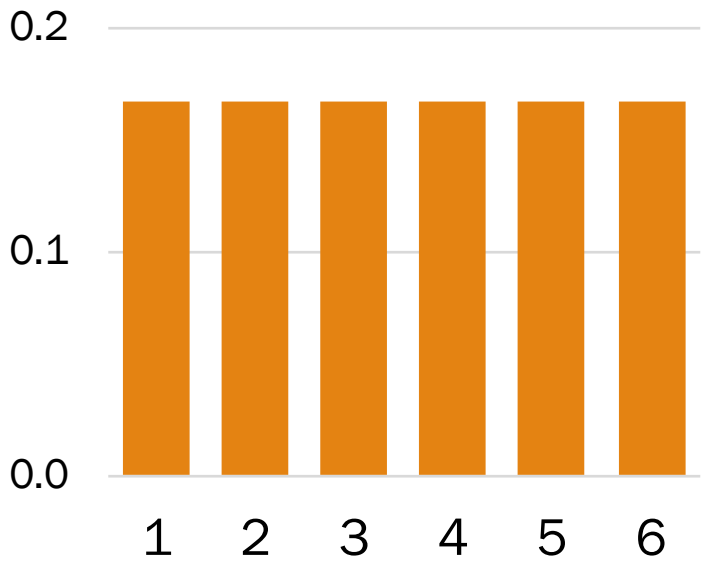
Announcements

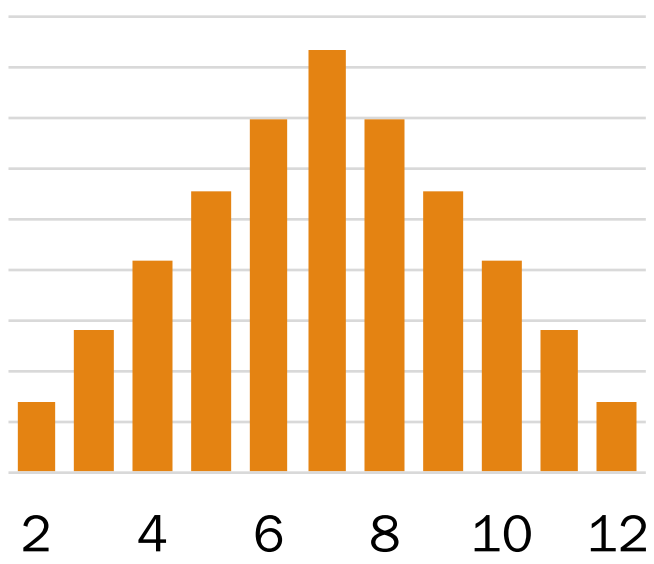
- Change of Grading Basis Deadline!
- Pset 4 Due Monday!
- Pset 5 Released Later Today!

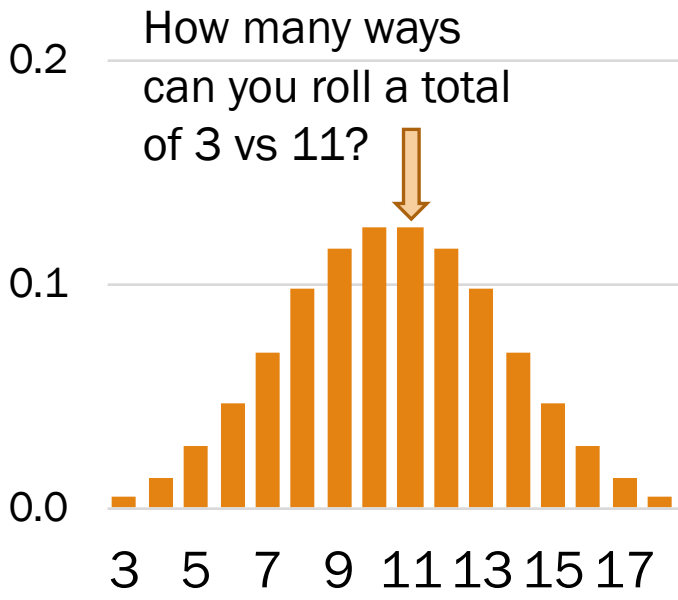
Review

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?

Central Limit Theorem

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 $n \to \infty$

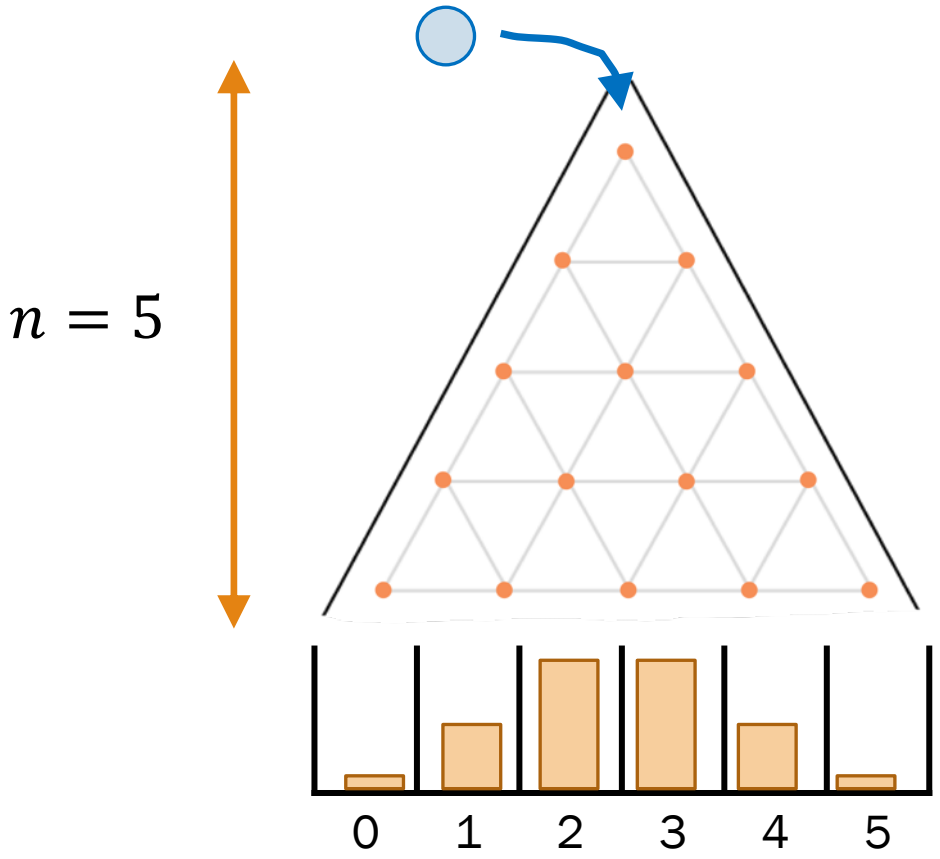
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

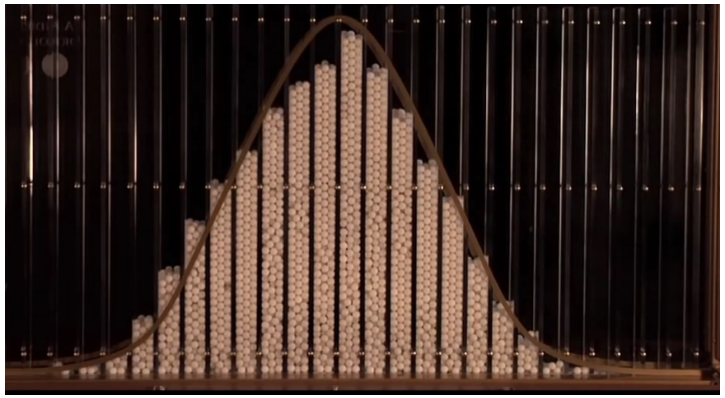
The sum of n i.i.d. random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

CLT explains a lot

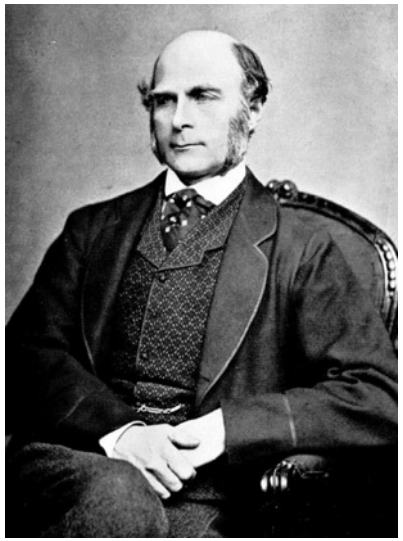
$$\text{As } n \to \infty \qquad \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.





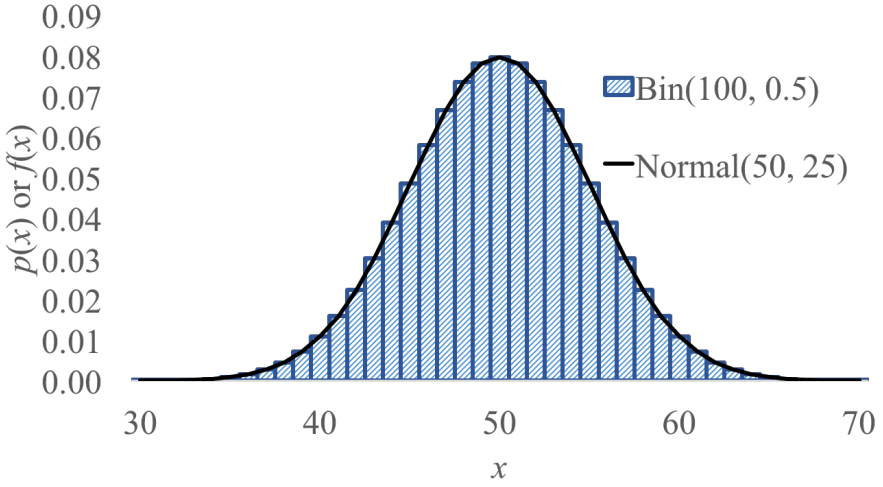
Galton Board, by Sir Francis Galton (1822-1911)



CLT explains a lot

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

The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.



Normal approximation of Binomial

Sum of i.i.d. Bernoulli RVs ≈ Normal

New Explanation:

Let $X_i \sim \text{Ber}(p)$ for $i = 1, \ldots, n$, where X_i are i.i.d.

$E[X_i] = p$, $\text{Var}(X_i) = p(1-p)$

$$X = \sum_{i=1}^{n} X_i \qquad (X \sim \text{Bin}(n, p))$$

$$X \sim \mathcal{N}(n\mu, n\sigma^2) \qquad (\text{CLT, as } n \to \infty)$$

$$X \sim \mathcal{N}(np, np(1-p)) \qquad (\text{substitute mean, variance of Bernoulli})$$

Sum of Dice

- You will roll 10 6-sided dice (X_1, X_2, …, X_{10})
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \leq 25$ or $X \geq 45$
  - Roll!

- And now the truth (according to the CLT)…

Sum of Dice

- You will roll 10 6-sided dice ($X_1, X_2, \ldots, X_{10}$)
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \leq 25$ or $X \geq 45$

- Recall CLT: $X = \sum_{i}^{n} X_i \to N(n\mu, n\sigma^2) \qquad \text{As } n \to \infty$
  - Determine P(X ≤ 25 or X ≥ 45) using CLT:

$$\mu = E[X_i] = 3.5 \qquad \sigma^2 = \text{Var}(X_i) = \frac{35}{12} \qquad X \approx N(35, 29.2)$$

$$1 - P(25.5 < X < 44.5) = 1 - P\left(\frac{25.5 - 35}{\sqrt{29.2}} < Z < \frac{44.5 - 35}{\sqrt{29.2}}\right)$$

$$\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784$$

Example CLT problem

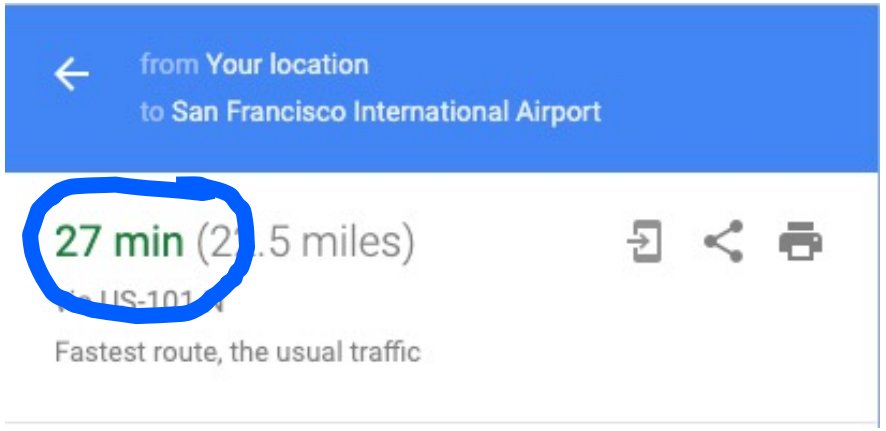
You encounter 10 traffic lights on your way to work. You don't know the full distribution of the wait time, but for each you observe the average wait time is 45 seconds and the standard deviation is 5 seconds. You will be on time if your total wait time is less than 8 mins. What is the probability that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

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

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

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



The sum of independent, identically distributed variables:

$$Y = \sum_{i=0}^{n} X_i$$

Is normally distributed:

$$Y \sim N(n\mu, n\sigma^2)$$

---

where $\mu = E[X_i]$

$$\sigma^2 = \text{Var}(X_i)$$

Average of IID Variables?

Let X_i be i.i.d. variables. There are n. Let $\bar{X}$ be the average

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

Gaussian by CLT

$$N(n\mu, n\sigma^2)$$

By the Central Limit Theorem, the mean of IID variables are distributed normally. As $n \to \infty$

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

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 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

---

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

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

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

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

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

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

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

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

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

$$n = 61.4$$

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}

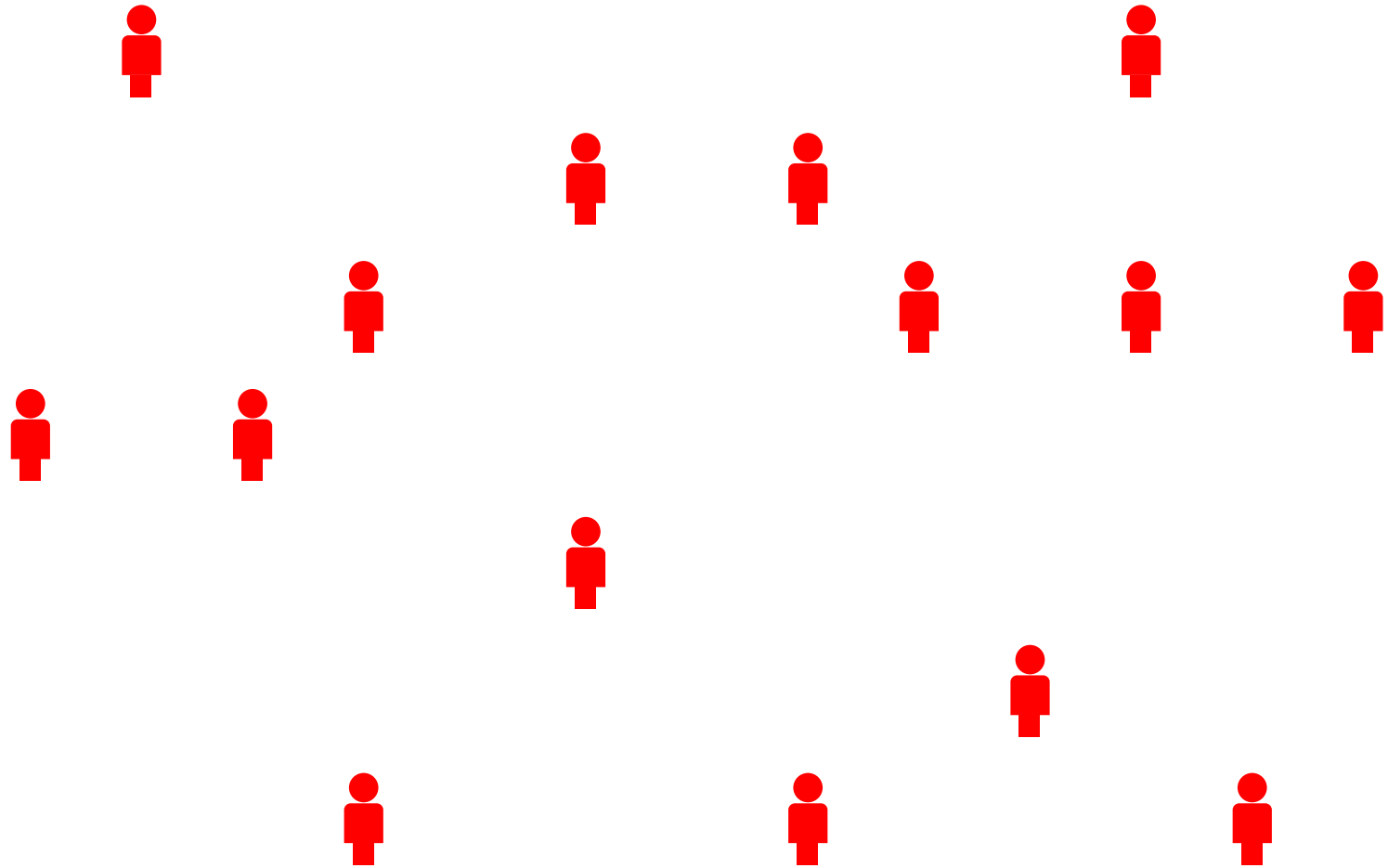
- The mean of all these numbers is 83.

Is this the true mean happiness of Bhutanese people?

Population

Sample

Sample



Collect one (or more) numbers from each person

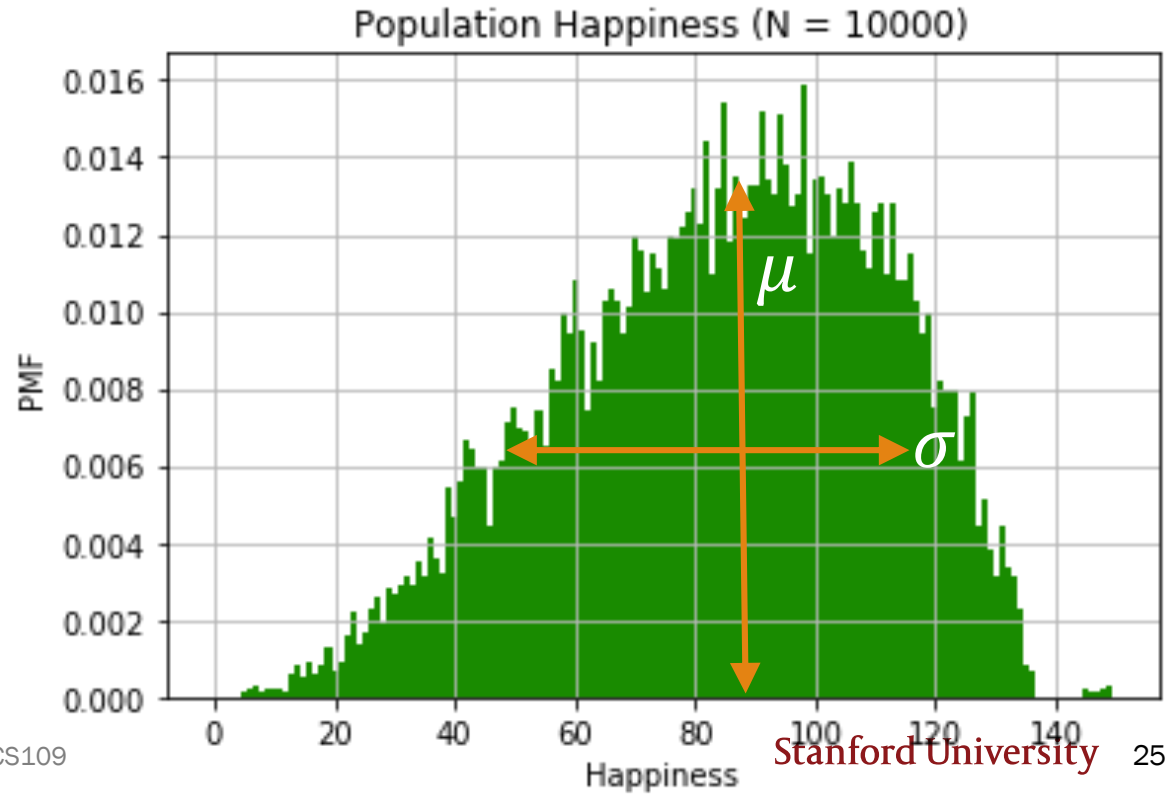
Sample

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



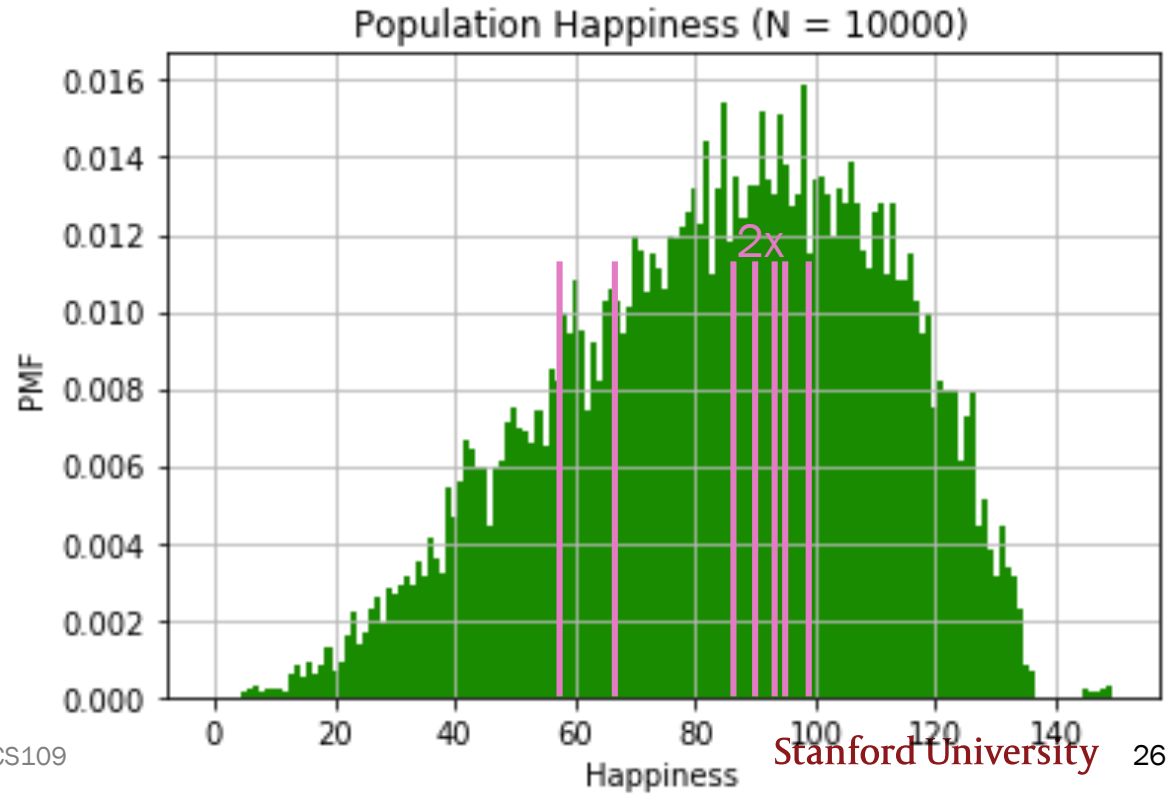
A sample, mathematically

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 single 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)

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).

So, these population statistics are unknown:

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

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 do we define best estimate?

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??*

$$\text{Estimate} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

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

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

Is that estimate any good?

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

Consider *n* random variables X_1, X_2, ... X_n

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

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

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right]$$ Pull out scalar mult

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

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

Estimating the population mean

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

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

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

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

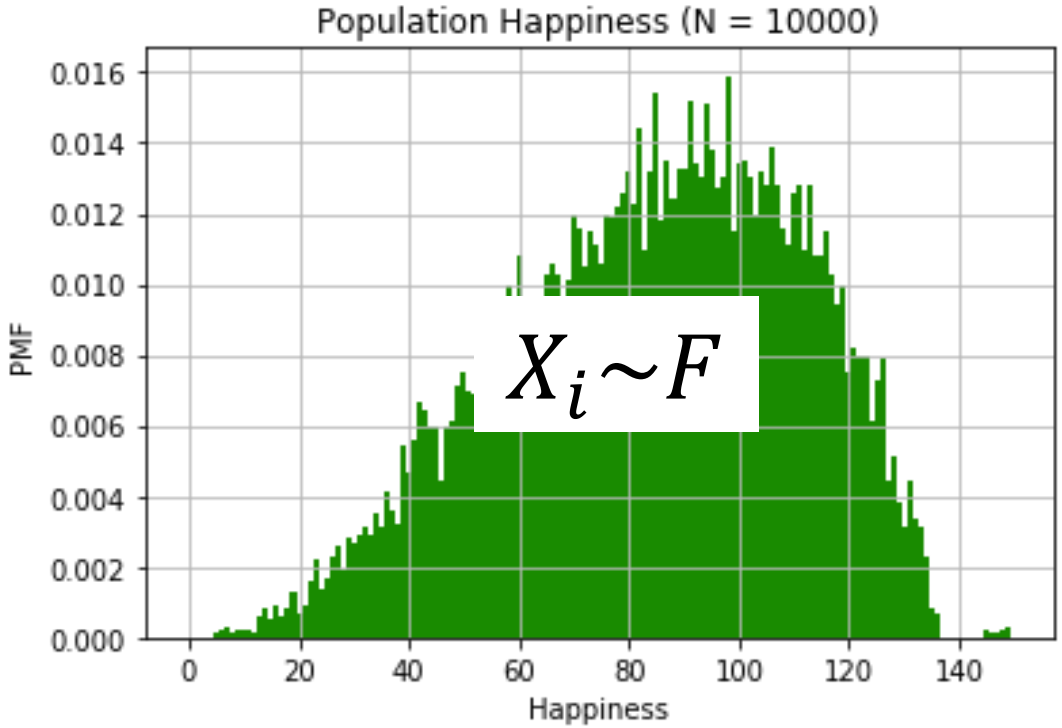
$\bar{X}$ is an **unbiased estimator** of the population mean μ. $E[\bar{X}] = \mu$

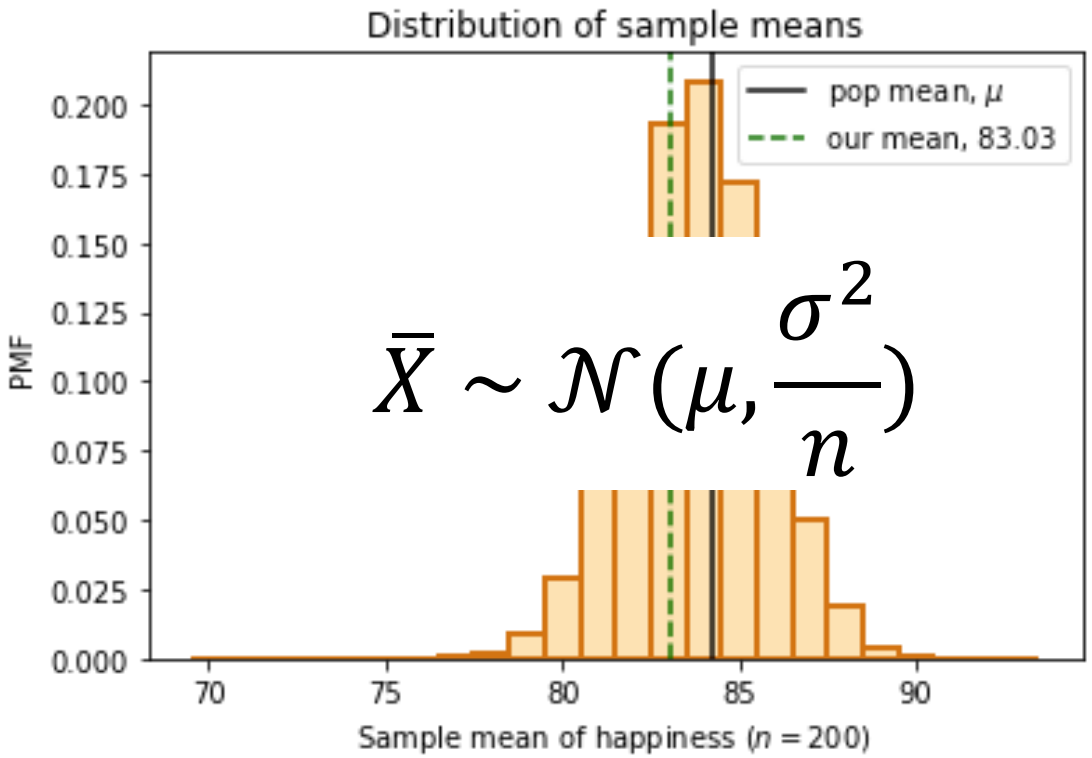
Intuition: By the CLT, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

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

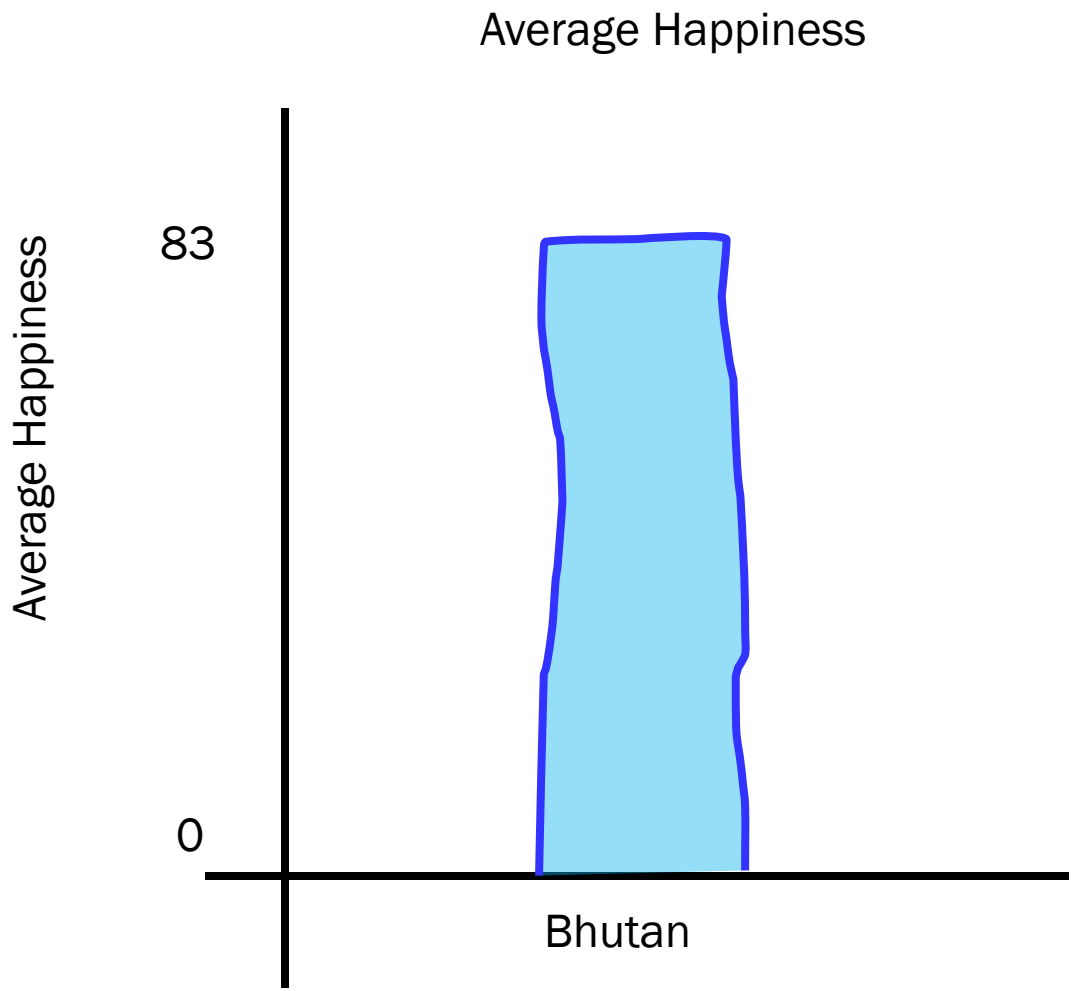
Sample mean





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:

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

ith sample

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

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

sample mean

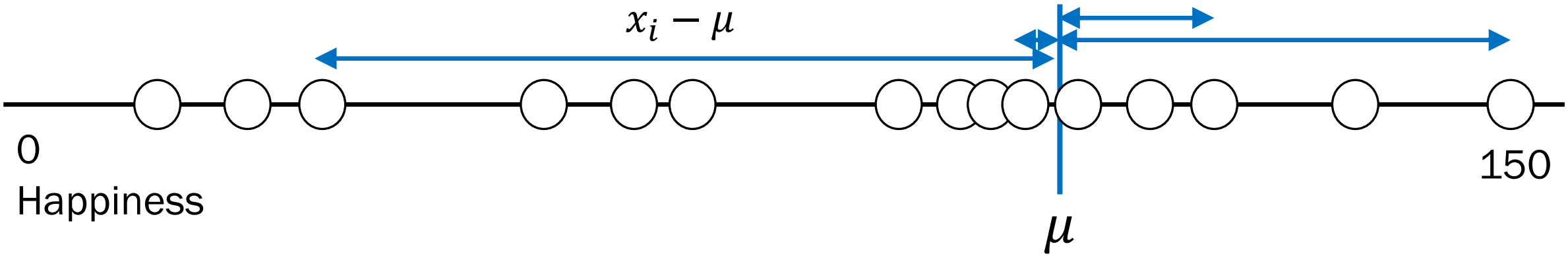
sample variance $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Intuition about the sample variance, S^2

Actual, σ^2

population mean

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



Population size, N

Calculating population statistics **exactly** requires us knowing all N datapoints.

Intuition about the sample variance, S^2

Actual, σ^2

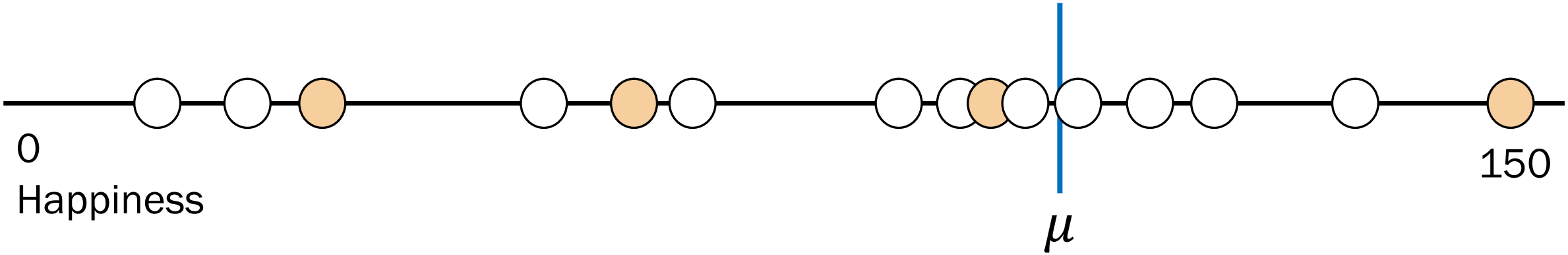
population mean

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

Actual, σ^2

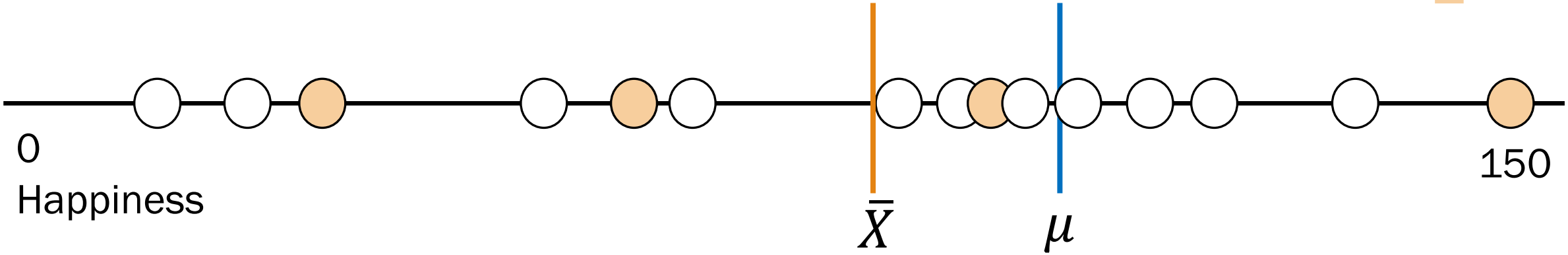
population mean

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

Actual, σ^2

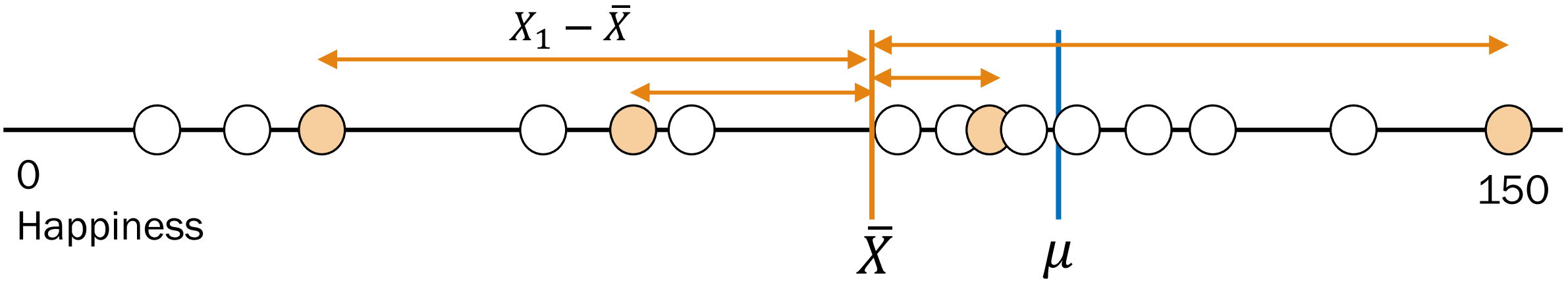
population mean

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Estimating the population variance

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

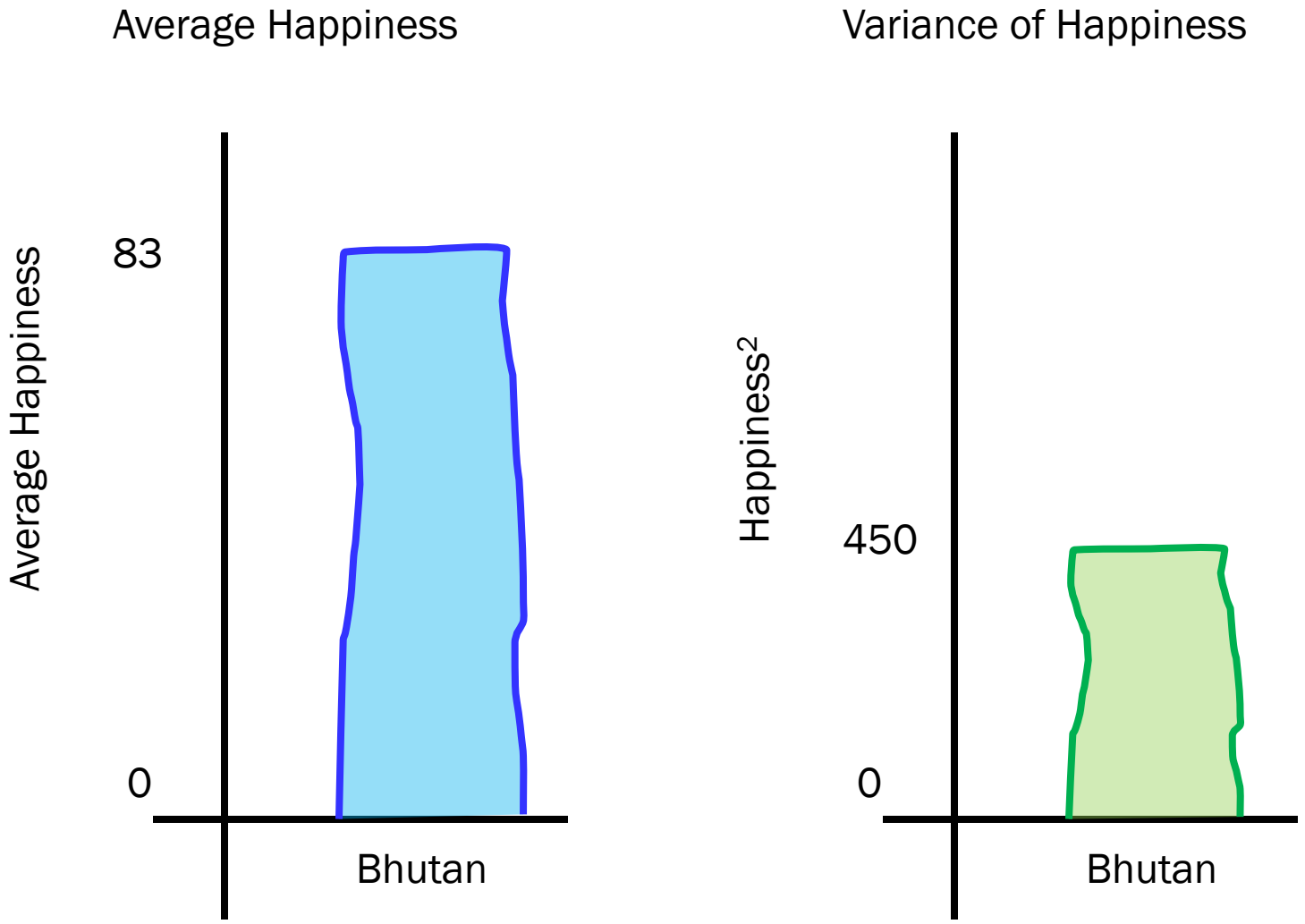
If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

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:

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

Sample mean

Makes it "unbiased"

No Error Bars ☹

Quick check

1. μ, the population mean Value!
2. $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$, a sample RV!
3. σ^2, the population variance Value!
4. $\bar{X}$, the sample mean R.V!
5. $\bar{X} = 83$ Event
6. $(X_1 = 59, X_2 = 87, X_3 = 94, X_4 = 99, X_5 = 87, X_6 = 78, X_7 = 69, X_8 = 91)$ Event

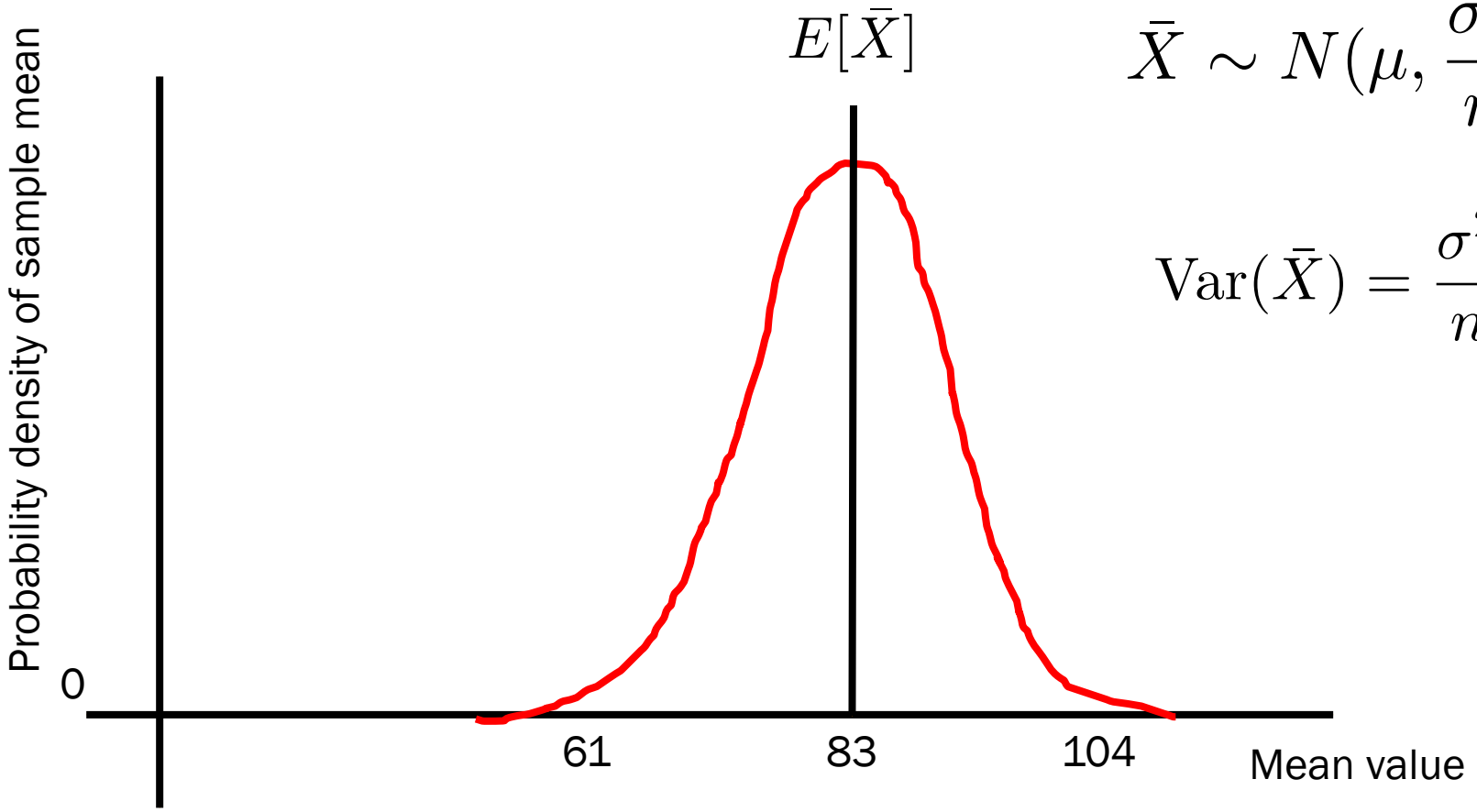
A. Random variable(s)
B. Value
C. Event

Insight: Sample Mean is an RV with known Var

By central limit theorem:

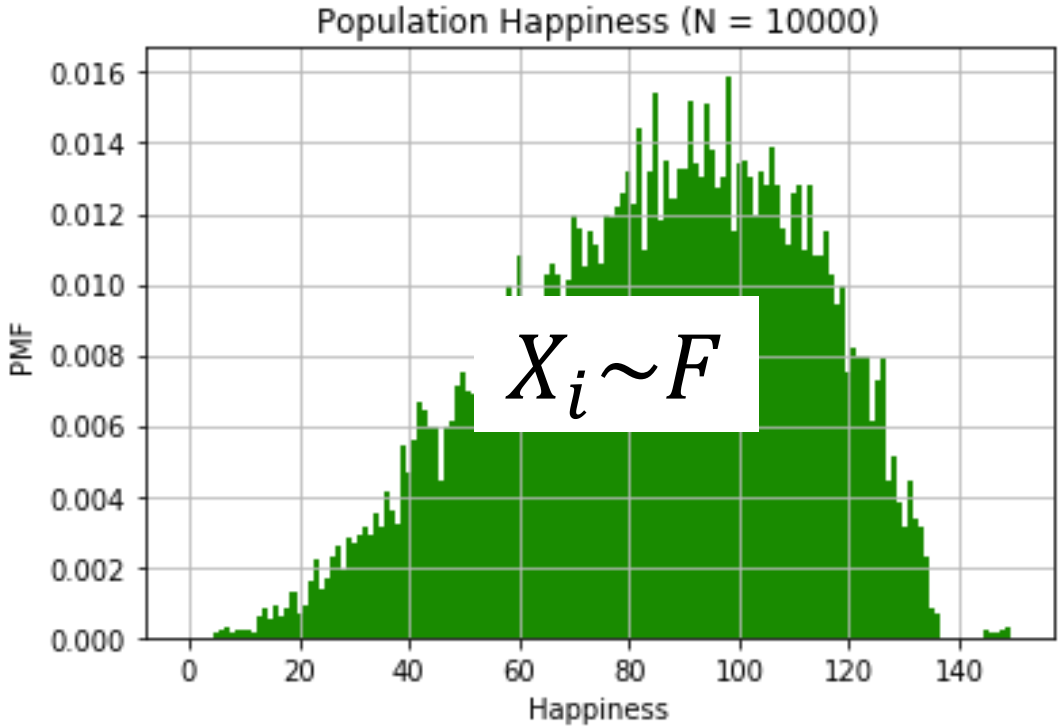
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

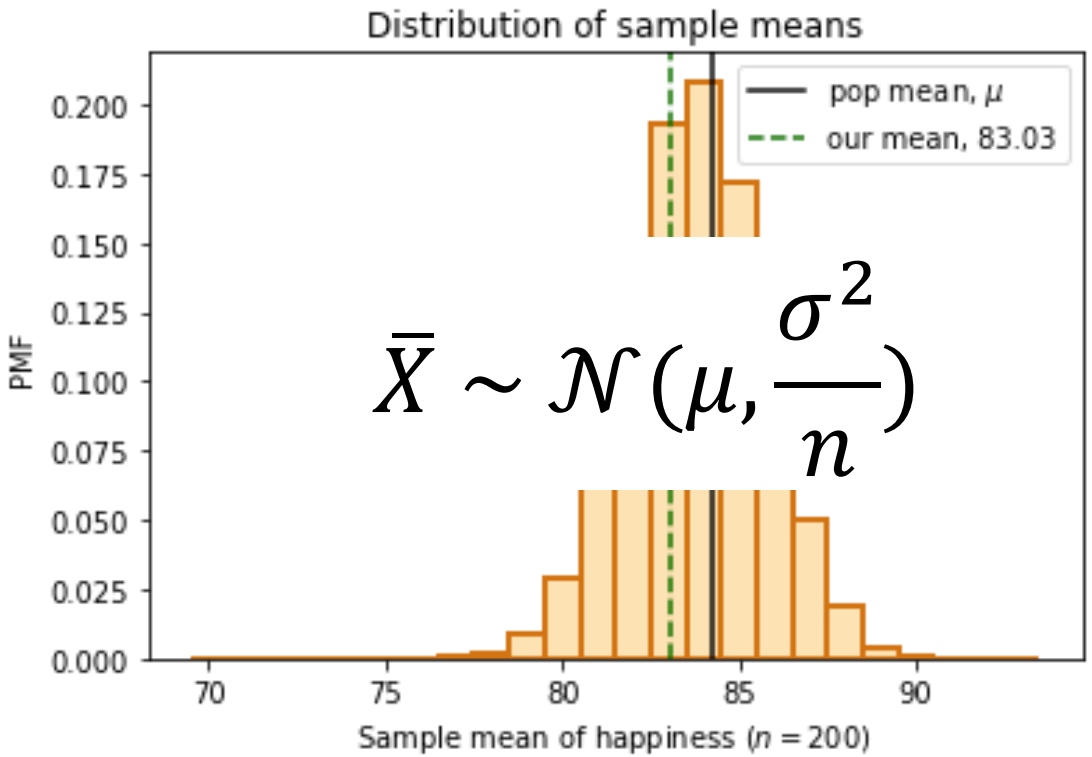
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$



Standard error of the mean

Sample mean





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

Standard Error of the Mean

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

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

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

Standard Error of the Mean

$$\text{Var}(\bar{X}) = \text{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

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

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

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

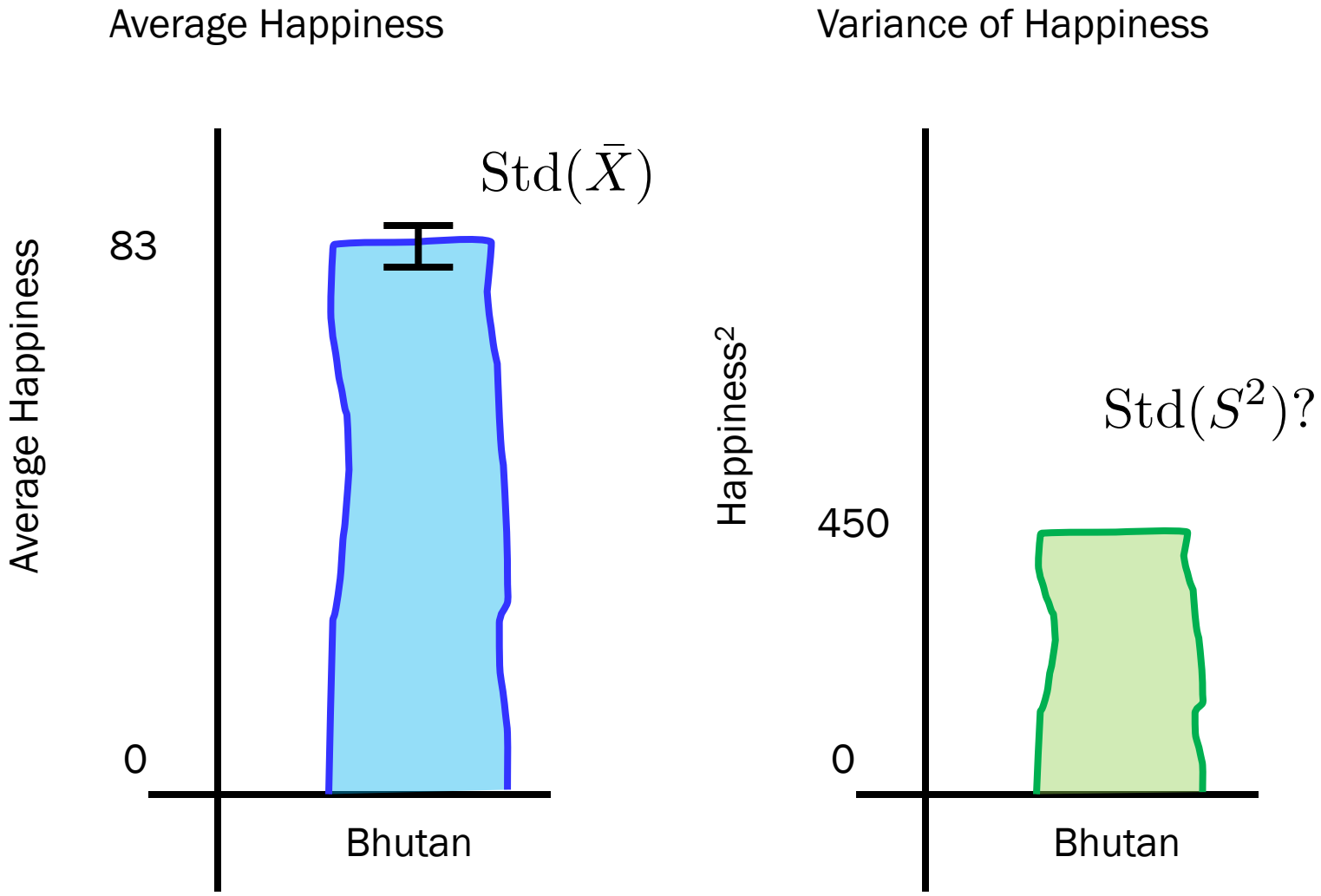
$$= \sqrt{\frac{450}{200}}$$

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

Passwords

How many tries to guess a len 10 password (each letter has 256 options)?

$$256^{10} = HUGE$$

How many tries if you can test 1 char at a time?

$$256 * 10 = Smaller$$

This reduction in difficulty can be a vulnerability.

Passwords

We have this interesting password checker. It's kinda flawed.

```
def check_password(user_input)
    for idx, letter in enumerate(user_input):
        if letter != password[idx]: # Takes time t
            return false
    return true
```

Here's a potential fix (find the flaw) – Hint should take appx **30***10*256 by CLT

```
def rand_time_check(user_input):
  ret = check_password(user_input)
  sleep(rand(0, len(password)))
  return ret
```

Slightly Better (but still can have issues)

```
def check_password(user_input)
    result = true
    for idx, letter in enumerate(user_input):
        if letter != password[idx]: # Takes time t
            result = false
    return result
```

Passwords

Better (not perfect)

```
def time_check(user_input):
  now = time()
  ret = check_password(user_input)
  sleep(len(password) * t - now)
  return ret
```