CLT and Beyond

CS109, Stanford University


AUG
14TH

Review

Where are we in CS109?

You are here → Uncertainty Theory

- 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

What happens when you Add Two Random Variables?

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

The Insight to Convolution Proofs

What is the probability that X + Y = n?

$$P(X + Y = n)?$$

$$P(X + Y = n) = \sum_{i=0}^{n} P(X = i, Y = n - i)$$

X	Y	i	
0	n	0	$P(X = 0, Y = n)$
1	n - 1	1	$P(X = 1, Y = n - 1)$
2	n - 2	2	$P(X = 2, Y = n - 2)$
	. . .		
n	0	n	$P(X = n, Y = 0)$

Convolution

Discrete

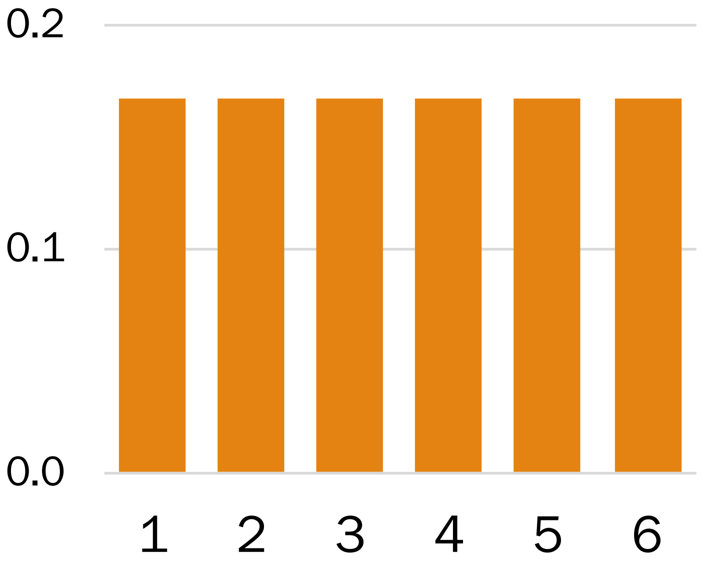
$$P(X+Y=a) = \sum_{y=-\infty}^{\infty} P(X=a-y)P(Y=y)\ dy$$

Continuous

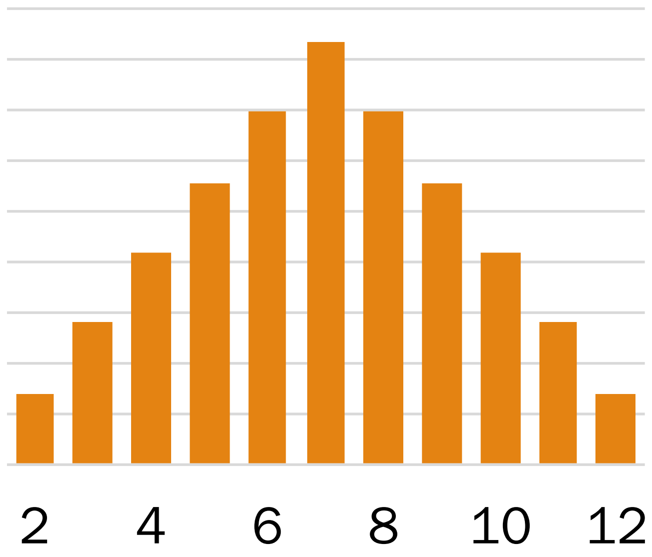
$$f(X+Y=a) = \int_{y=-\infty}^{\infty} f(X=a-y)f(Y=y)\ dy$$

Sum of dice rolls

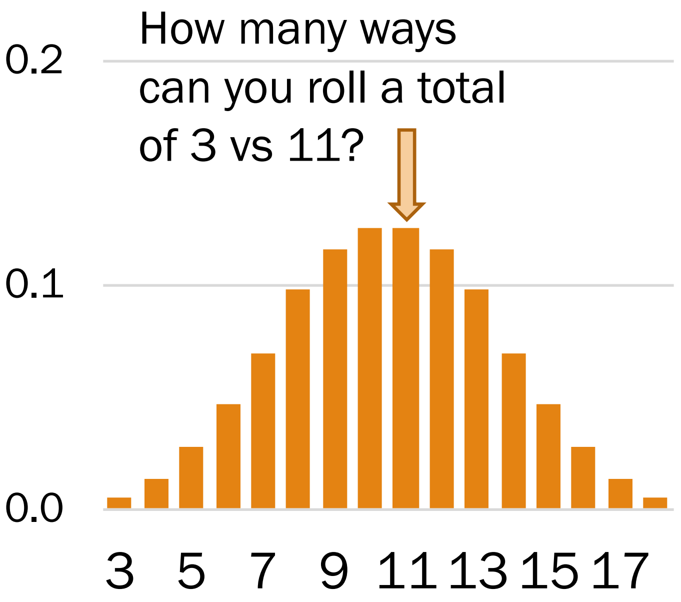
Roll n independent dice. Let X_i be the outcome of roll i. X_i are i.i.d.



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

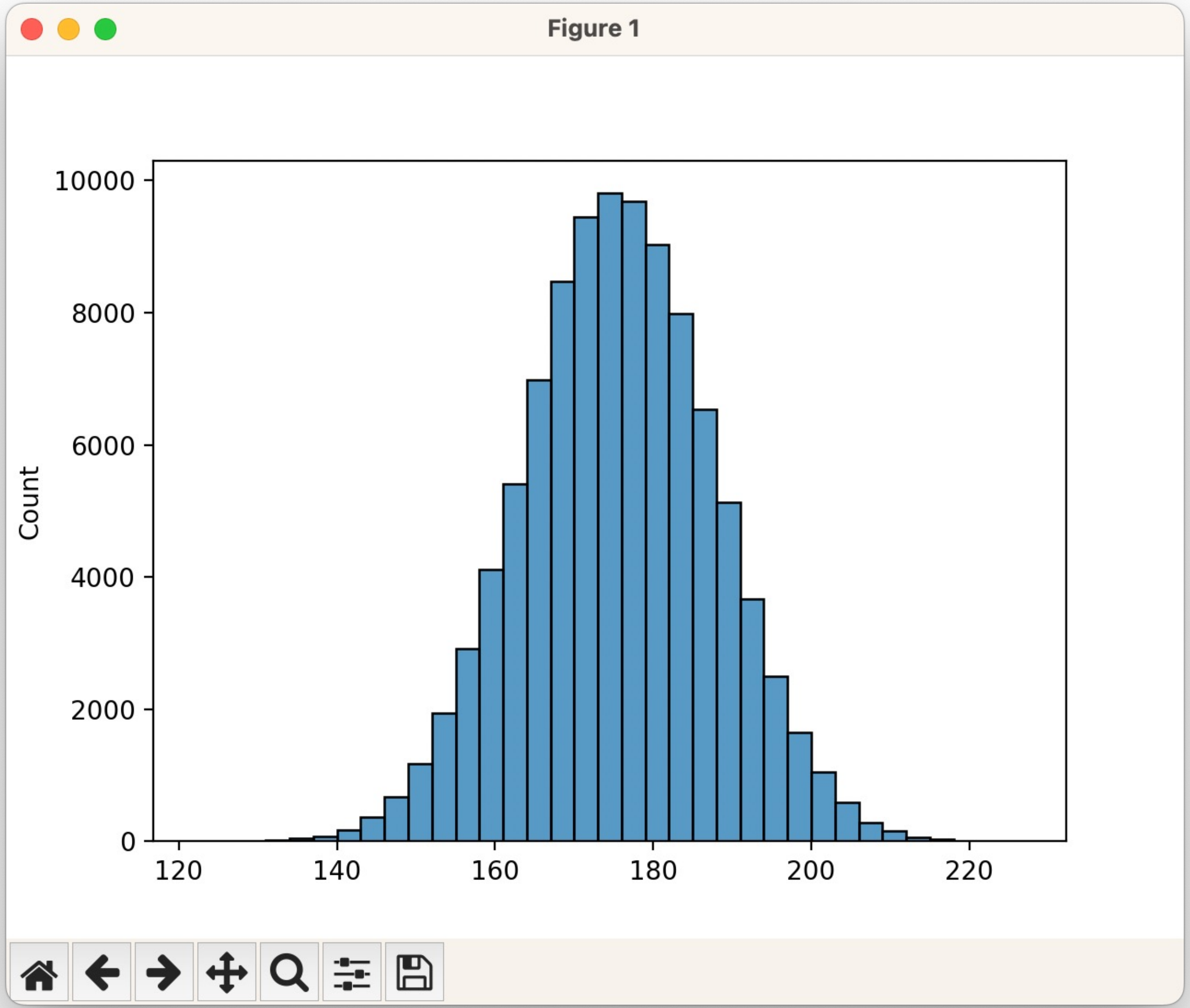


$$\sum_{i=1}^{2} X_i$$ Sum of 2 dice rolls

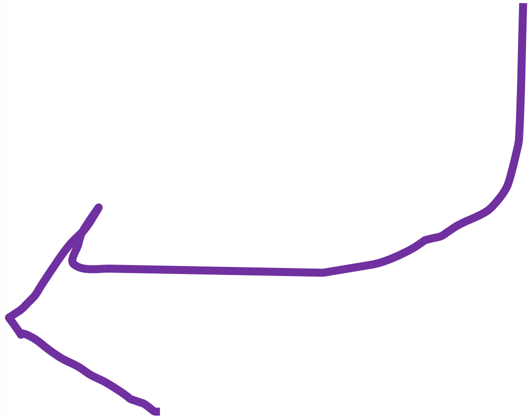


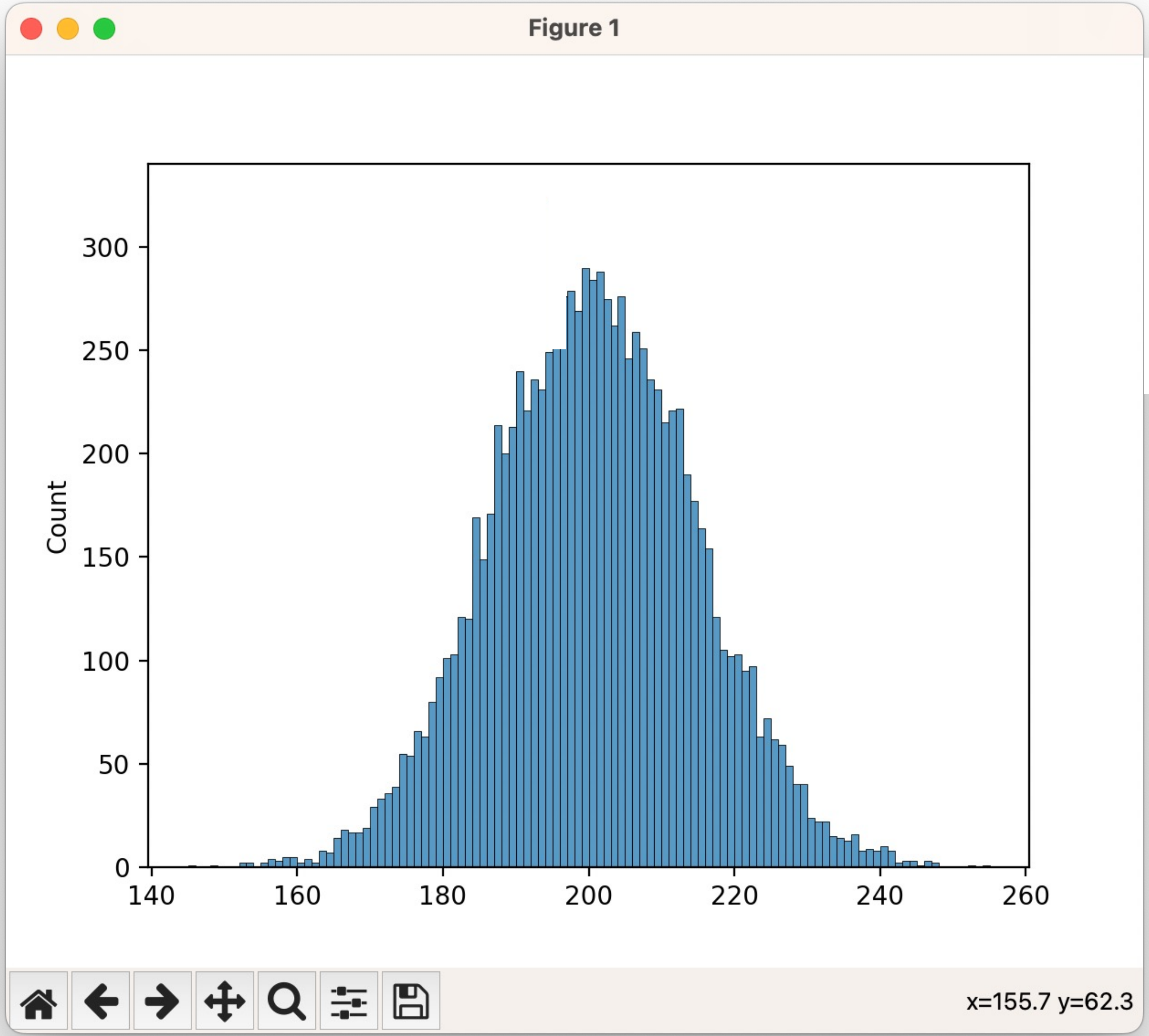
$$\sum_{i=1}^{3} X_i$$ Sum of 3 dice rolls

Sum of 50 dice?




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






```python
def run_experiment():
    total = 0
    for i in range(100):
        total += stats.poisson.rvs(2)
    return total
```

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 $n\sigma^2$.

True happiness

Wonderful Form of Cosmic Order

I know of scarcely anything so apt to impress the imagination as the **wonderful form of cosmic order** expressed by the ”[Central limit theorem]". The law would have been personified by the Greeks and **deified**, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. **Whenever a large sample of chaotic elements are taken in hand (summed) an unsuspected and most beautiful form of regularity proves to have been latent all along**.

-

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 $X \geq 45$
  - 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: $X = \sum_{i}^{n} X_i \rightarrow N(n\mu, n\sigma^2)$ As $n \rightarrow \infty$
  - Determine P($X \leq 25$ or $X \geq 45$) using CLT:

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

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

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

Quick Concept Check CLT problem

You hit 10 boba shops on your way to work for your 10 work besties. You don't know the full distribution of the wait time, but for each you observe the average wait time is 45 sec. and the Std. of 5 seconds. You will be on time if your total wait time is less than 8 mins across all boba shops. What is the probability that you are on time? *Assume the wait times are IID.*

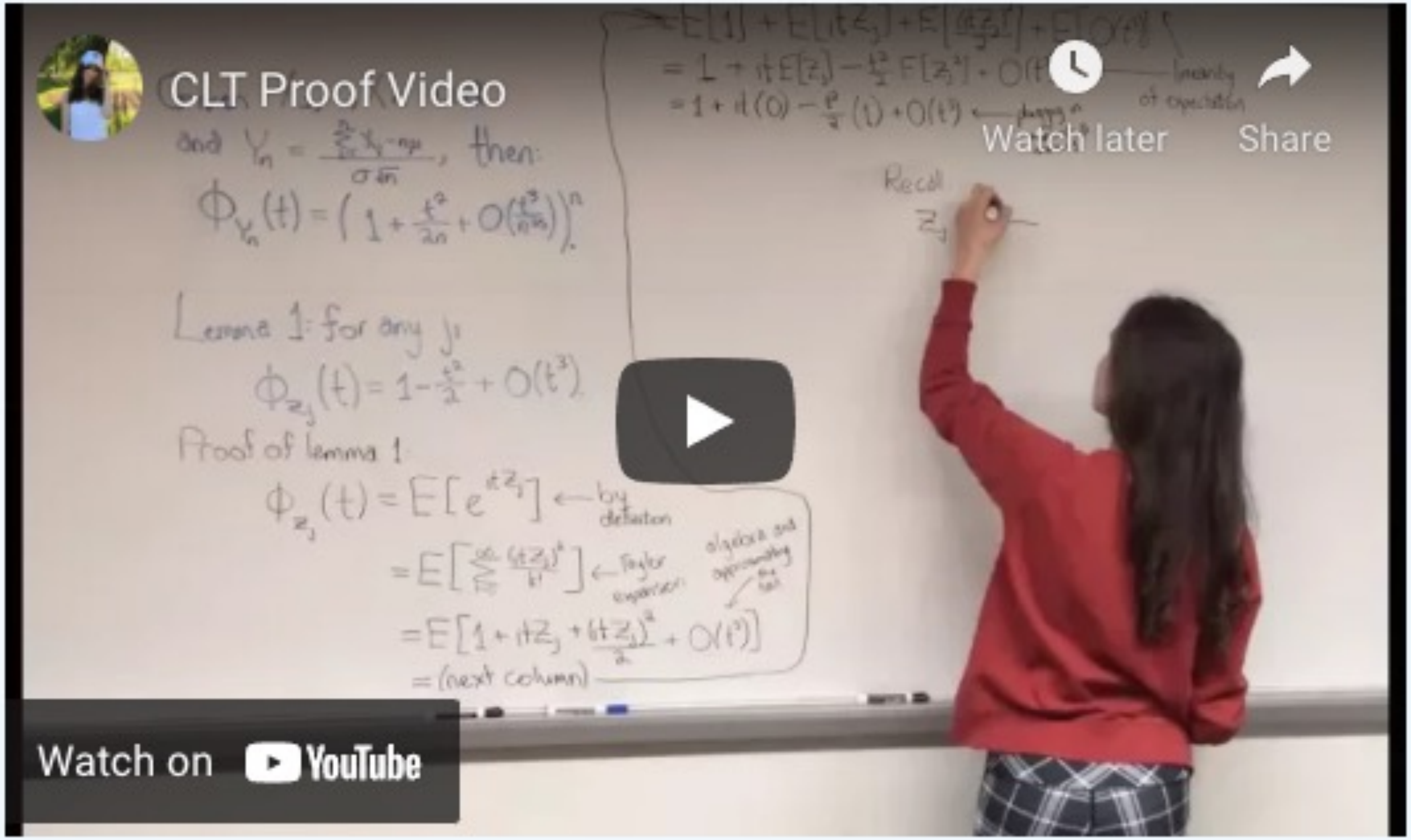
Answer: Let T be the total wait time. It is the sum of the 10 IID wait times. By the CLT

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

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

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

For Proof, See Video



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

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

What about other functions?

Sum of iid? Normal

Average of iid?

Max of iid?

By the Central Limit Theorem, the mean of IID variables are distributed normally. As *n* -> ∞

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

What about other functions?

Sum of iid? Normal

Average of iid? Normal

Max of iid?

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)

William Sealy Gosset (13 June 1876 – 16 October 1937) was an English statistician, chemist and brewer who served as Head Brewer of Guinness and Head Experimental Brewer of Guinness and was a pioneer of modern statistics. He pioneered small sample experimental design and analysis with an economic approach to the logic of uncertainty. Gosset published under the pen name **Student** and developed most famously Student's t-distribution – originally called Student's "z" – and "Student's test of statistical significance".[1]

Sampling definitions

Motivating example

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

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

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

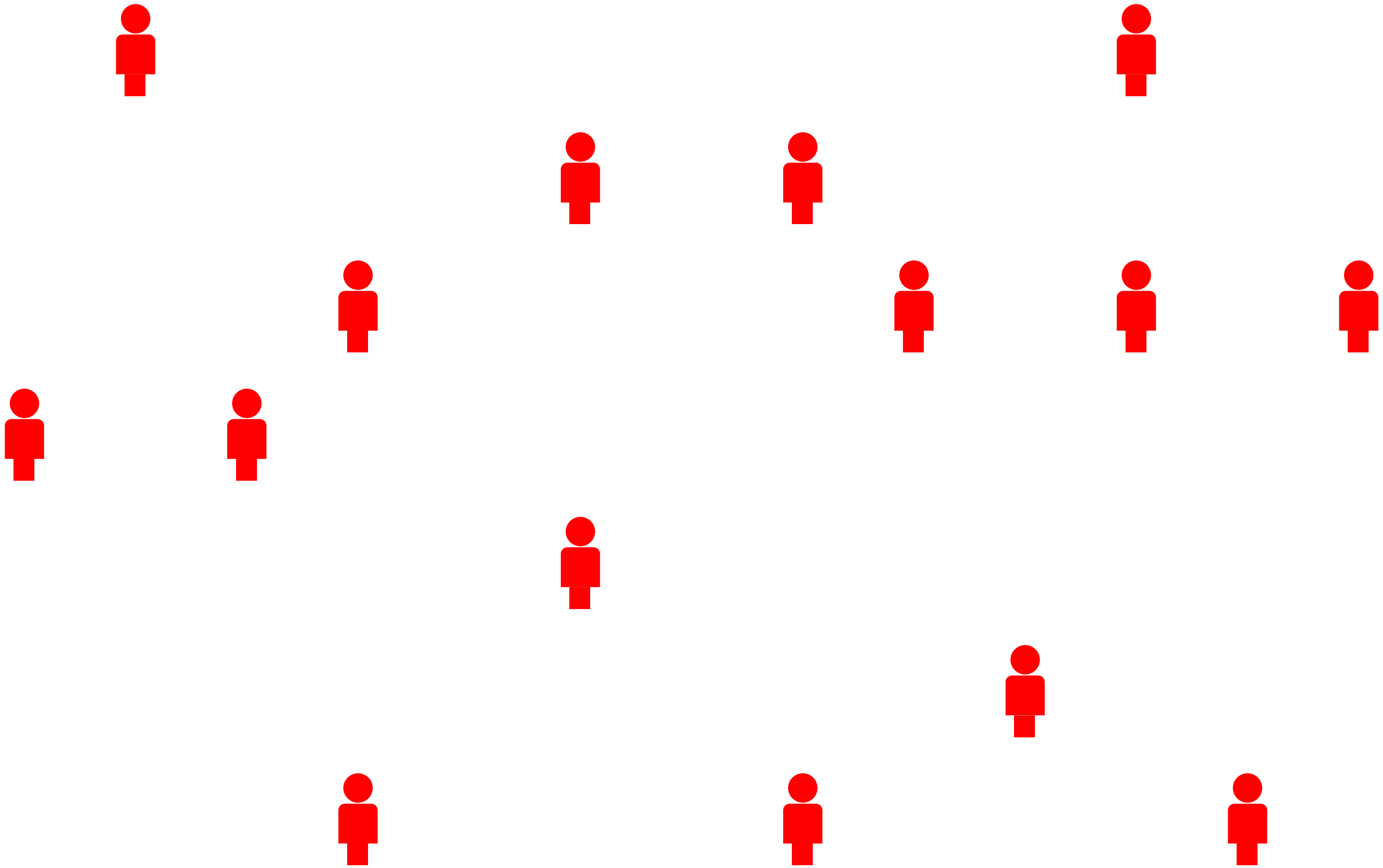
- The mean of all these numbers is 83.

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

Population

Sample

Sample



Collect one (or more) numbers from each person

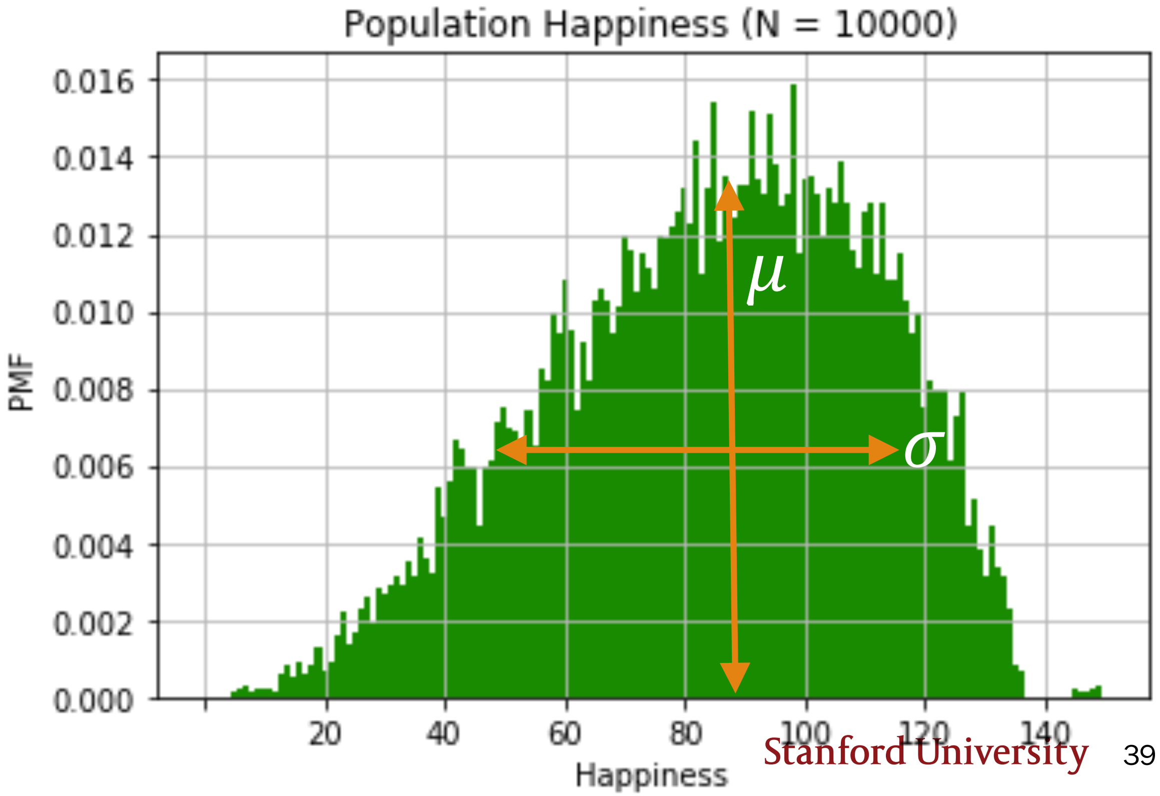
Sample

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



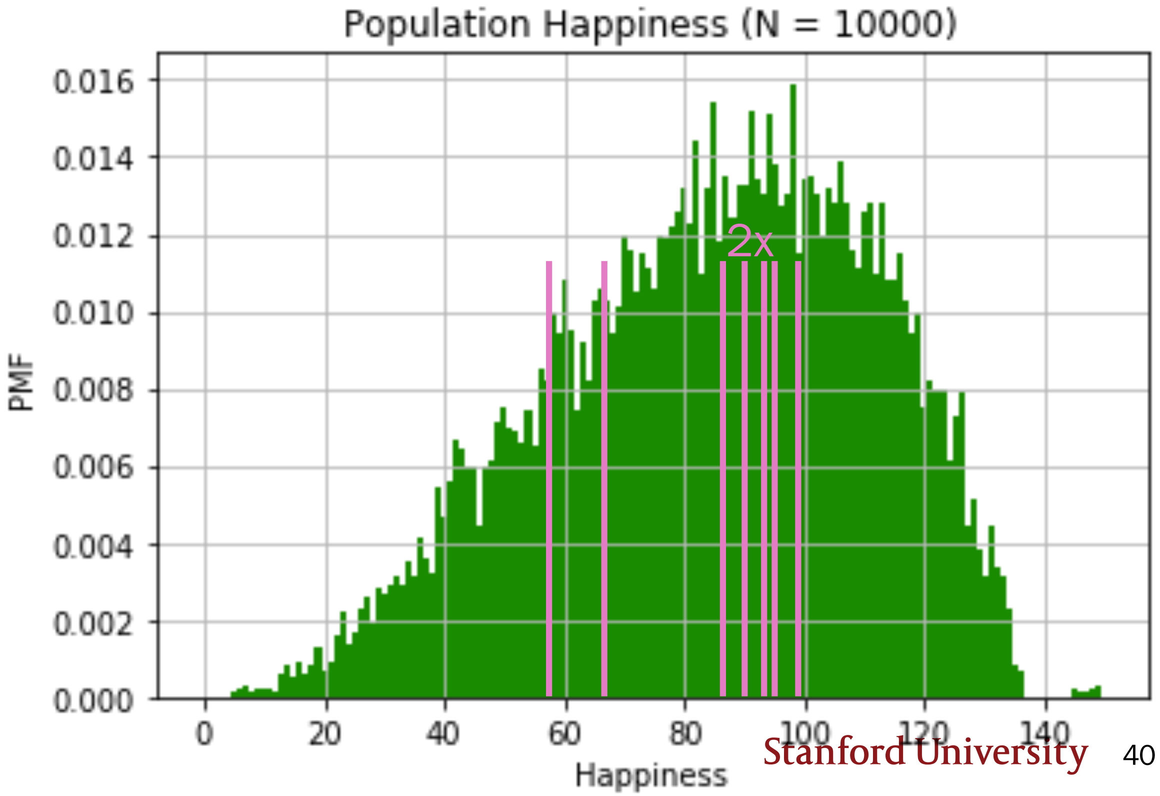
A sample, mathematically

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:

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



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)

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??*

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

Sample Mean: This is a fancy way of saying "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 $X_1, X_2, \ldots X_n$

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

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

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

Estimating the population mean

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

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

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

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

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

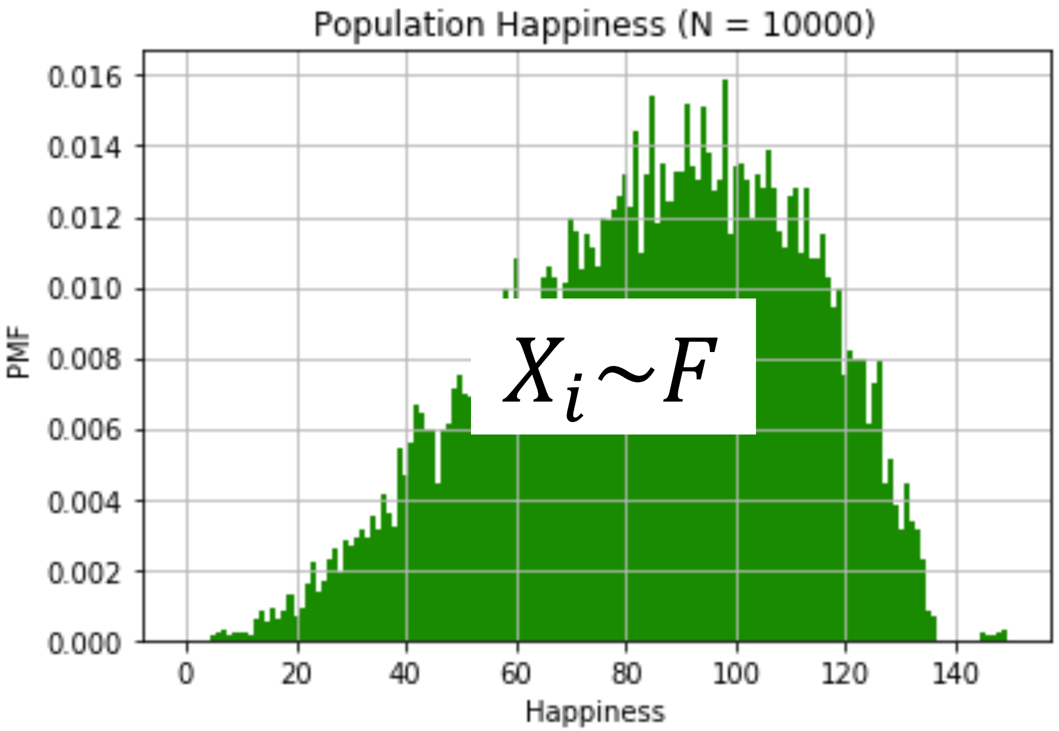
$$E[\bar{X}] = \mu$$

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

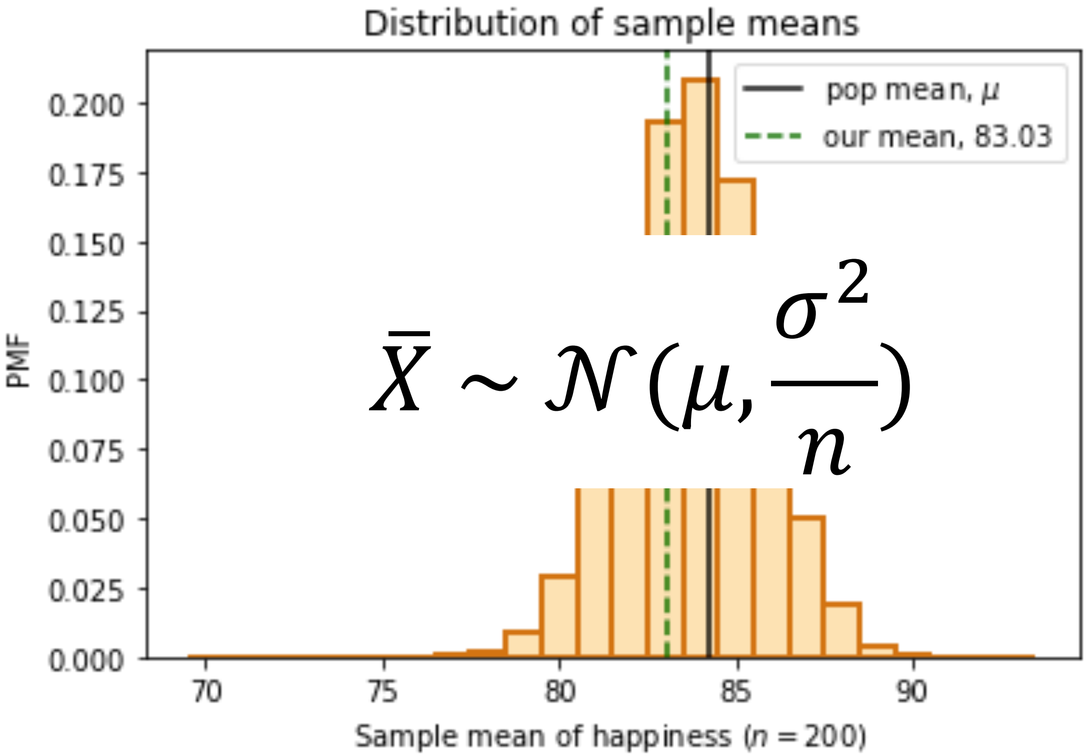
If we could take *multiple* samples of size n:

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

Sample mean



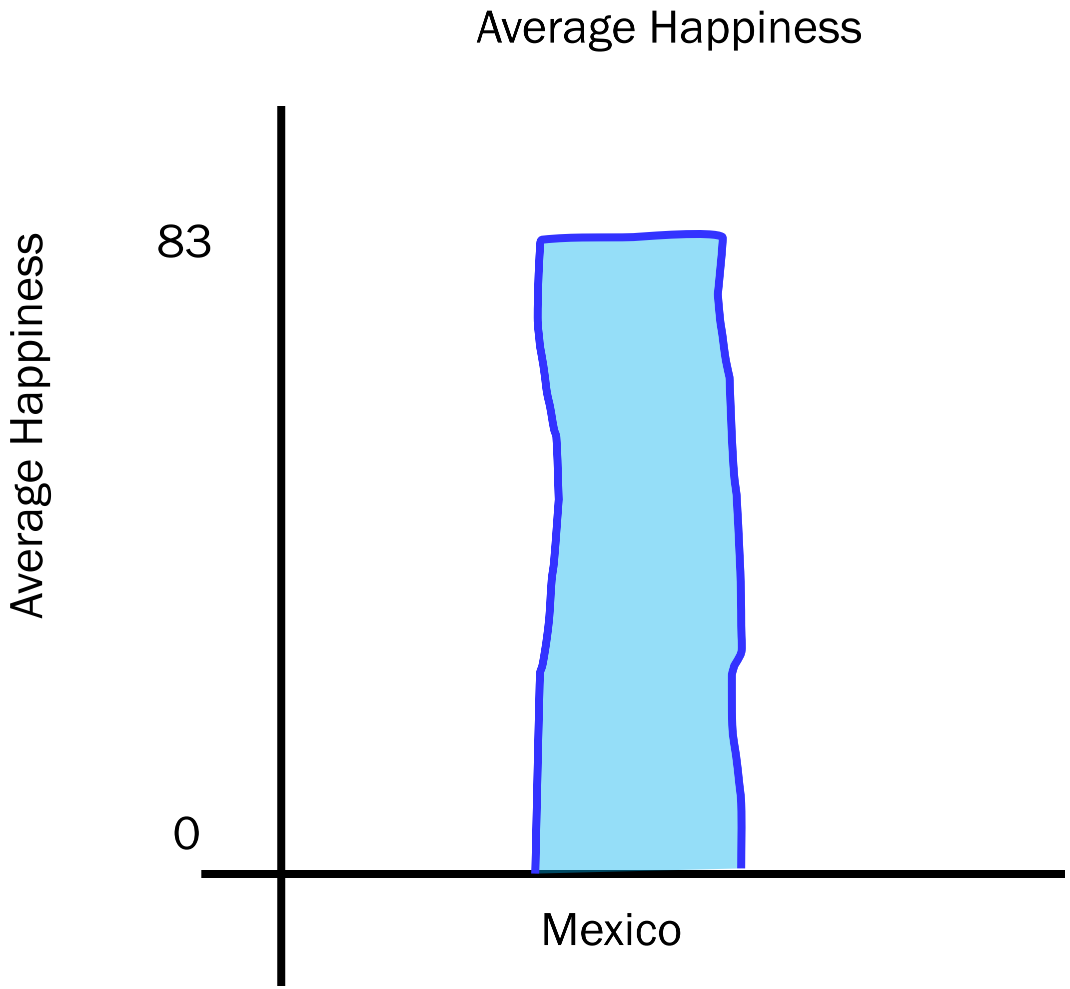
$X_i \sim F$



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

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 the Mexican 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 Mexican people?

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

population variance

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

population mean

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

sample variance

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

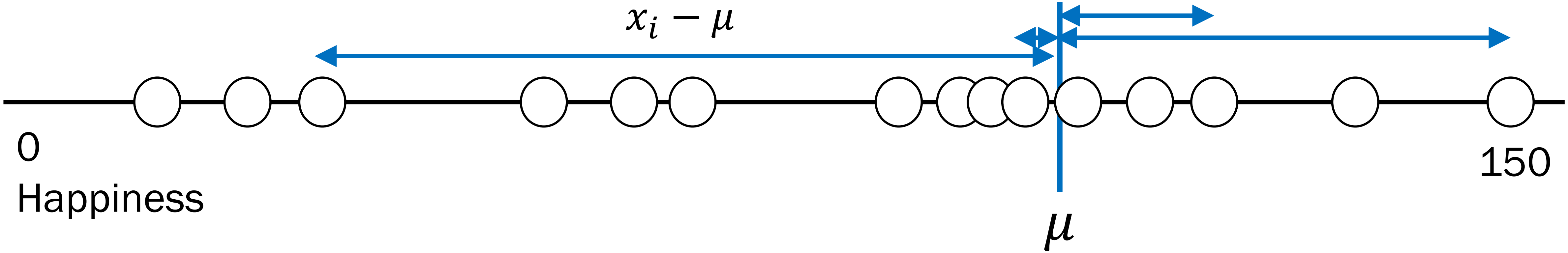
sample mean

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

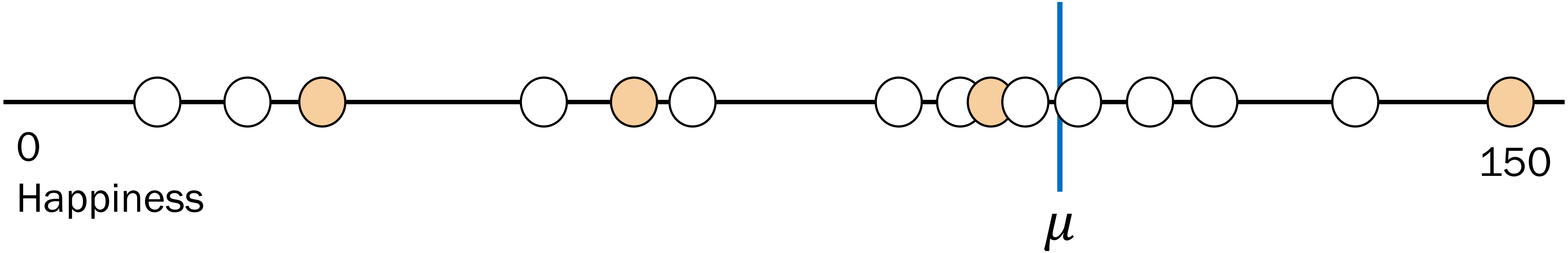
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Estimate, S^2

sample mean ↓

sample variance

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



Population size, N

Intuition about the sample variance, S^2

<u>Actual, σ^2</u>

population mean

population variance

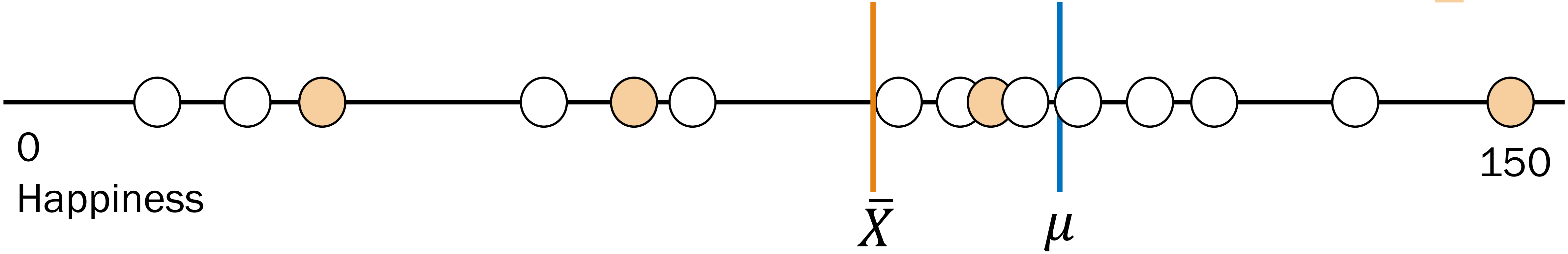
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

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

sample mean

sample variance

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



Population size, N

Intuition about the sample variance, S^2

Actual, σ^2

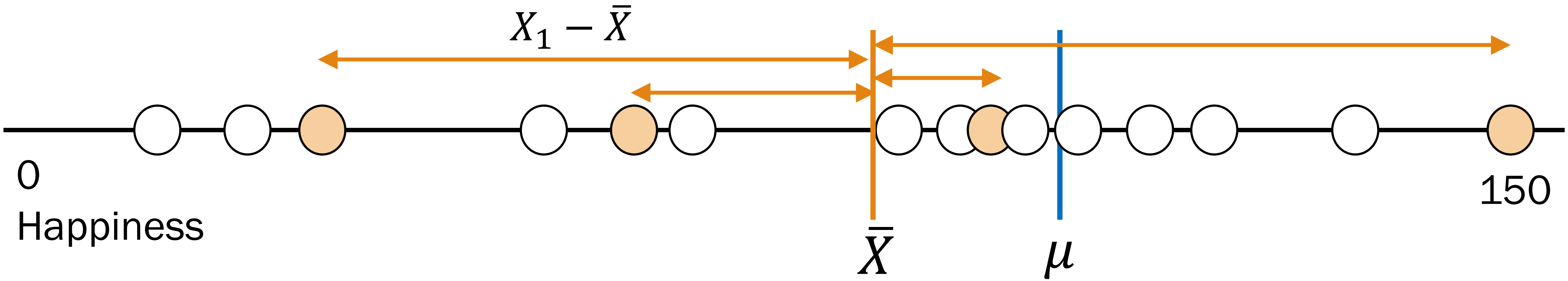
population mean

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

Estimate, S^2

sample mean

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



Population size, N

This formula will always underestimate the variance…

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

Estimating the population variance

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

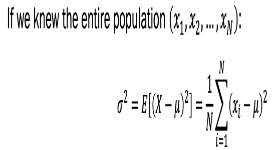
If we knew the entire population $(x_1, x_2, \dots, x_N)$:

population mean

population variance

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

Zoinks!



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

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

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

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

sample mean

sample variance

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

Estimating the population variance ….unbiasedly…

2. What is σ^2, the variance of happiness of Mexican 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 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Estimating the population variance ....unbiasedly...

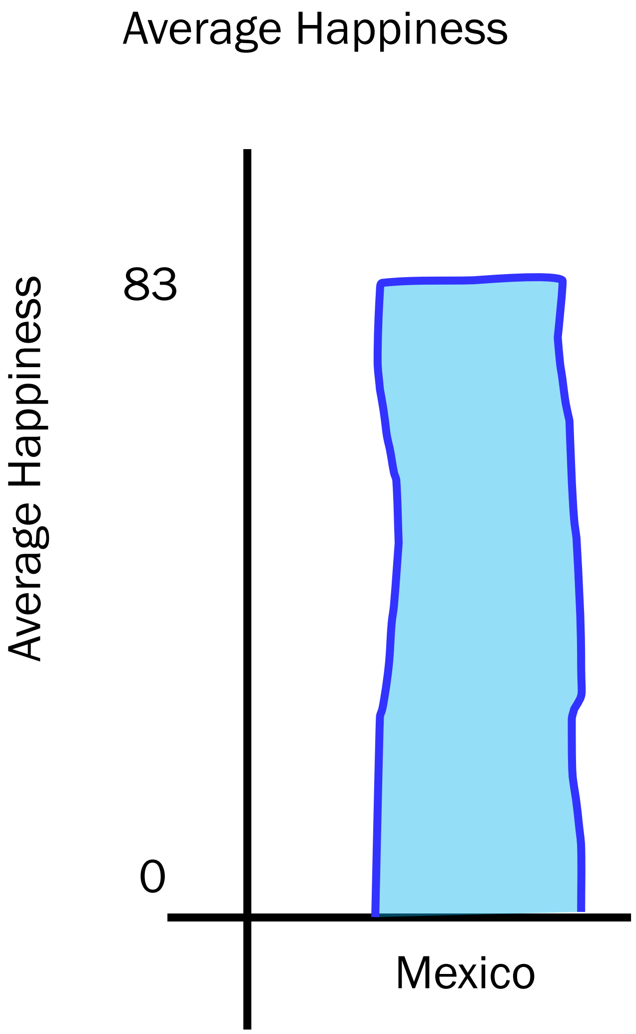
2. What is σ^2, the variance of happiness of Mexican people?

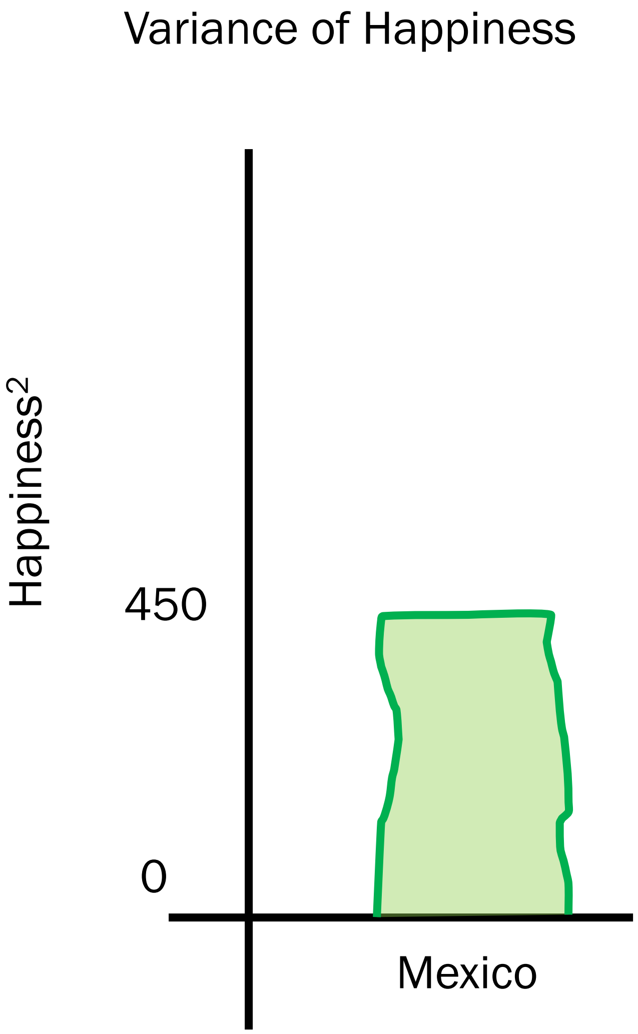
If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

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 Mexico Government





Sample Variance:

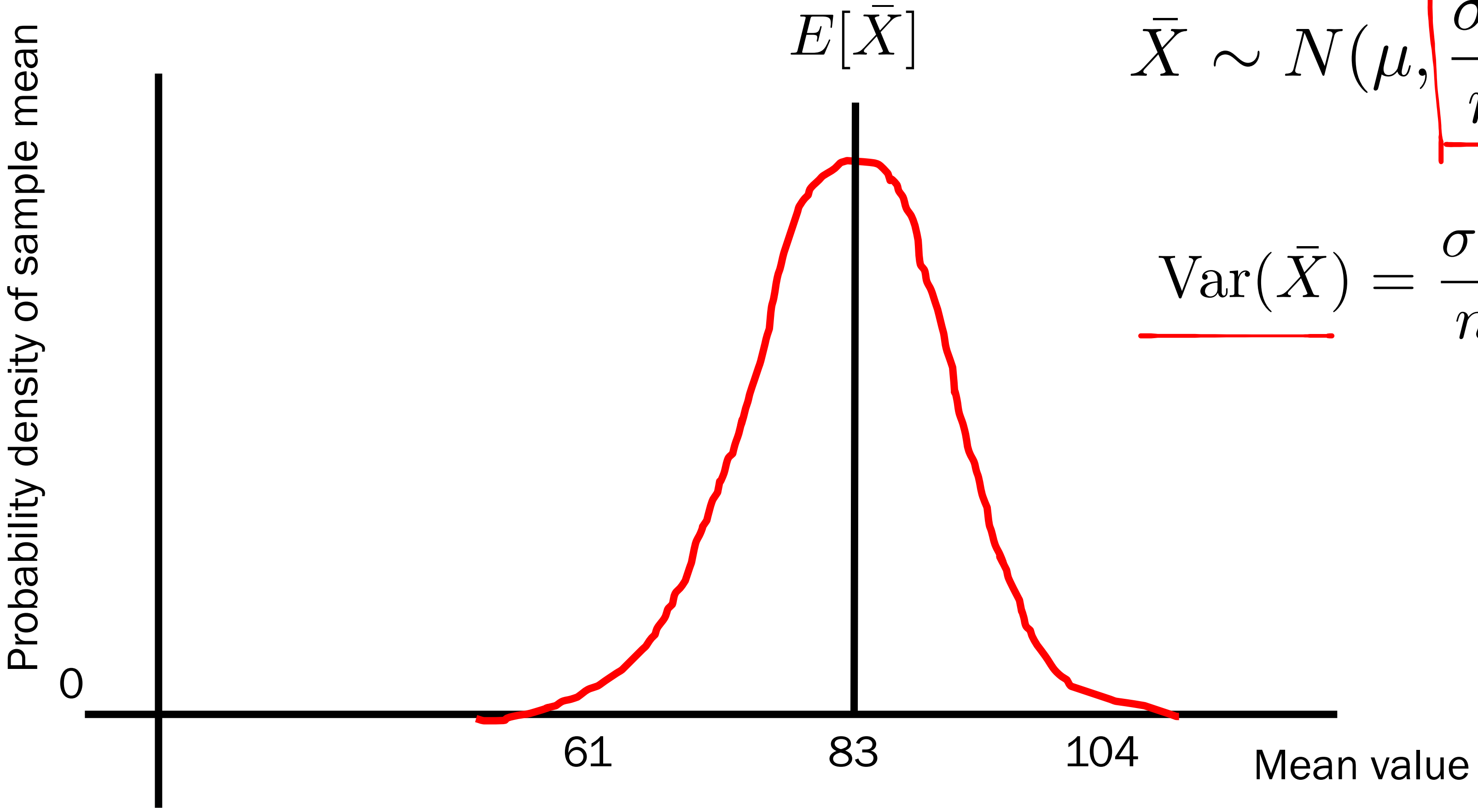
Sample mean

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

Makes it "unbiased"

No Error Bars ☹

Insight: Sample Mean is an RV with known Var



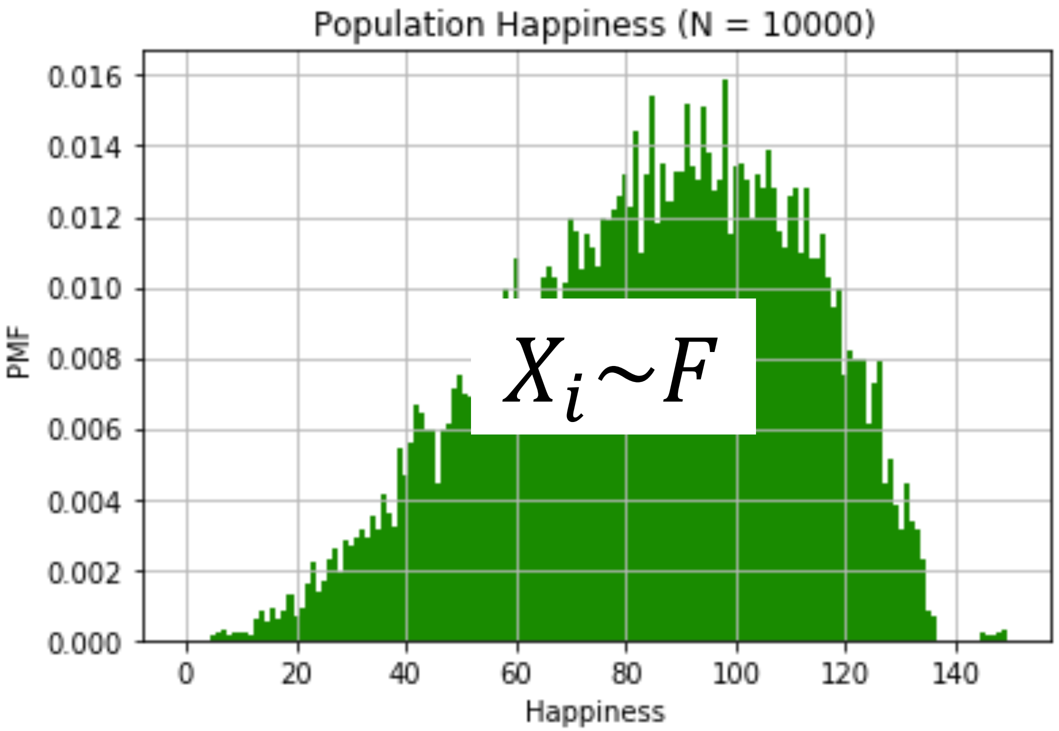
By central limit theorem:

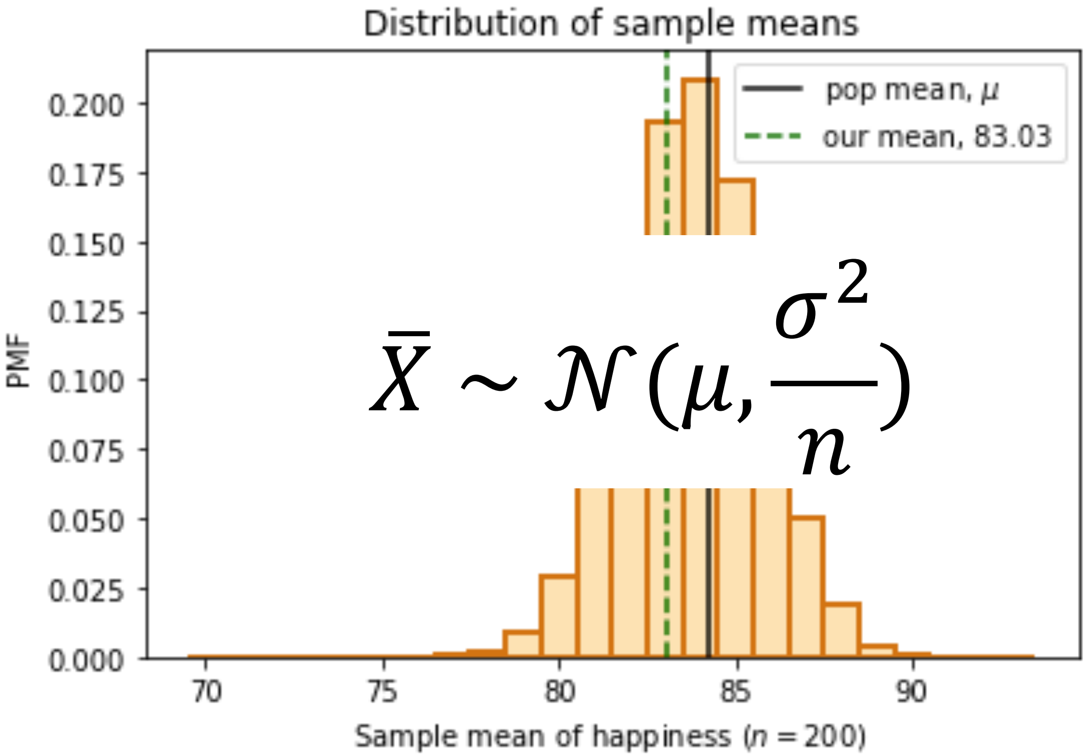
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

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

Standard error of the mean

Sample mean





- 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

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

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

We want to estimate this

<u>def</u> 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

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

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

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

Since S_2 is an unbiased estimate

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

Change variance to standard deviation

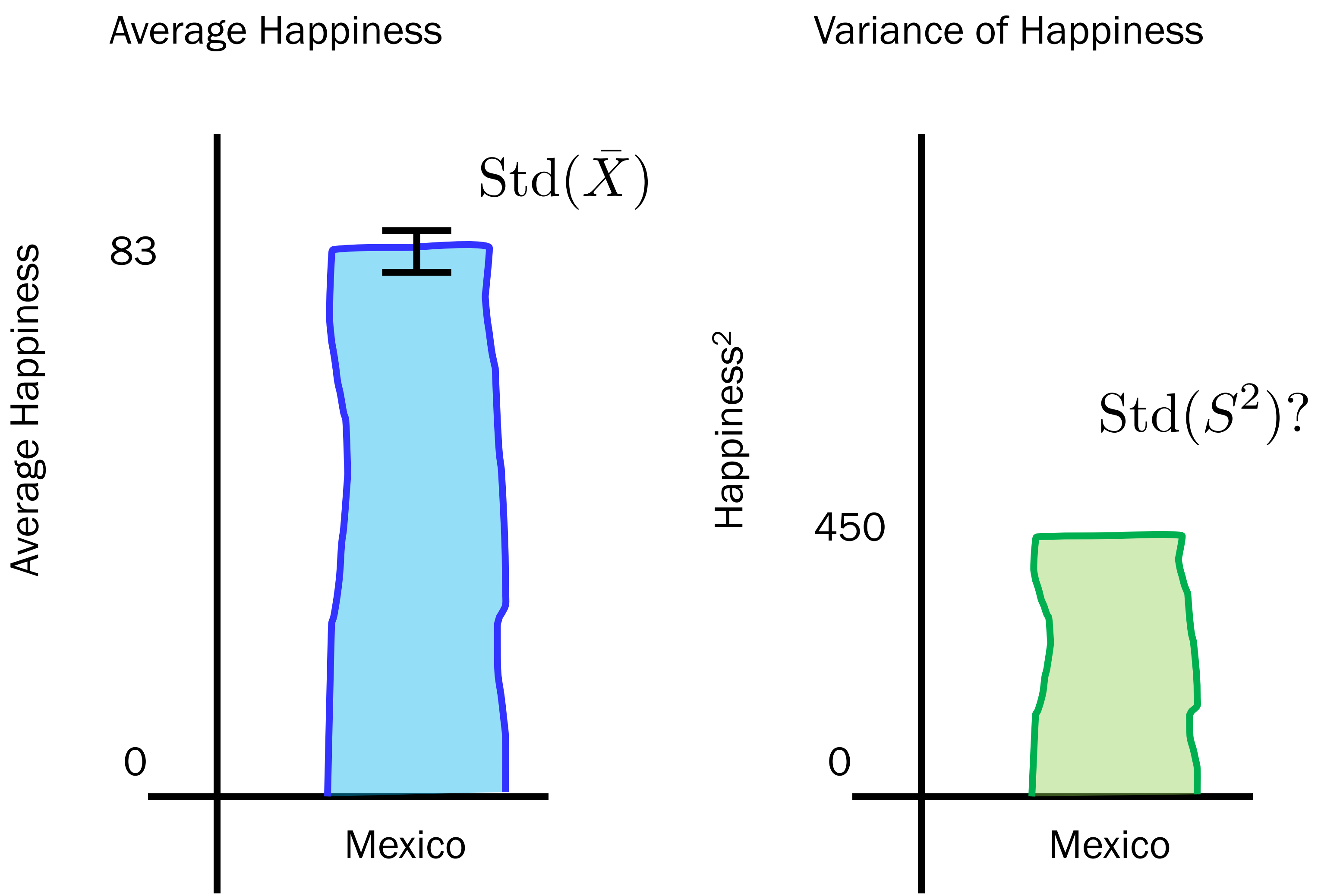
$$= \sqrt{\frac{450}{200}}$$

The numbers for our Mexican poll

$$= 1.5$$

Mexican standard error of the mean

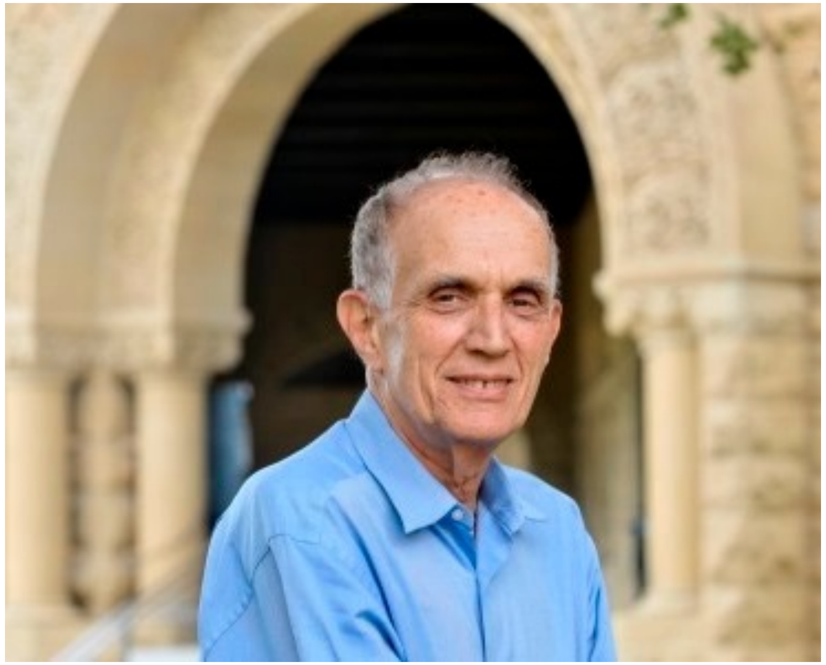
Our Report to Mexico Government

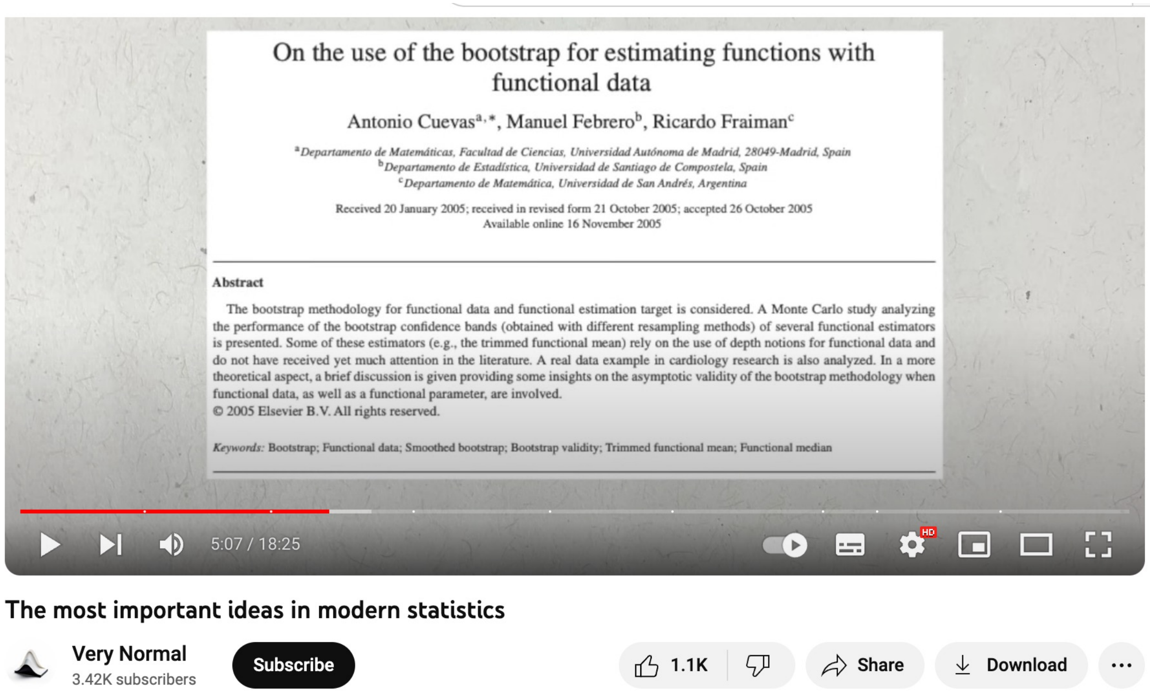


Claim: The average happiness of Mexico is 83 ± 2

Bootstraping

One of the Most Important Ideas in Modern Statistics!





Invented bootstrapping in 1979

Still a professor at Stanford

Won a National Science Medal

Come back on Friday!!

AUG 14TH