CLT and Beyond

Chris Piech
CS109, Stanford University


NOV
29TH

Review

The Insight to Convolution Proofs

What is the probability that X + Y = n?

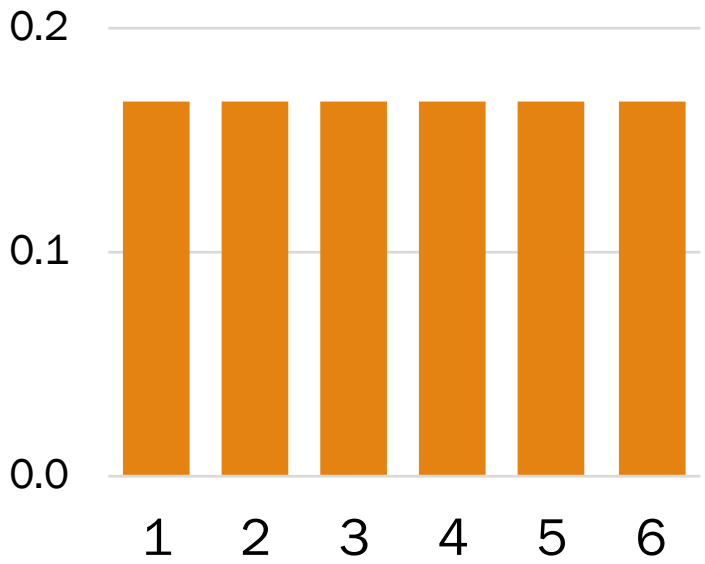
$$P(X+Y=n)?$$

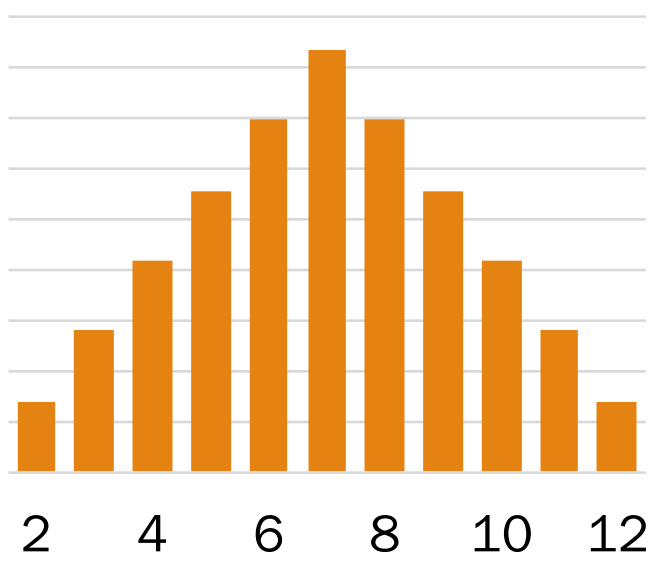
$$P(X+Y=n) = \sum_{i=0}^{n} P(X=i, Y=n-i)$$

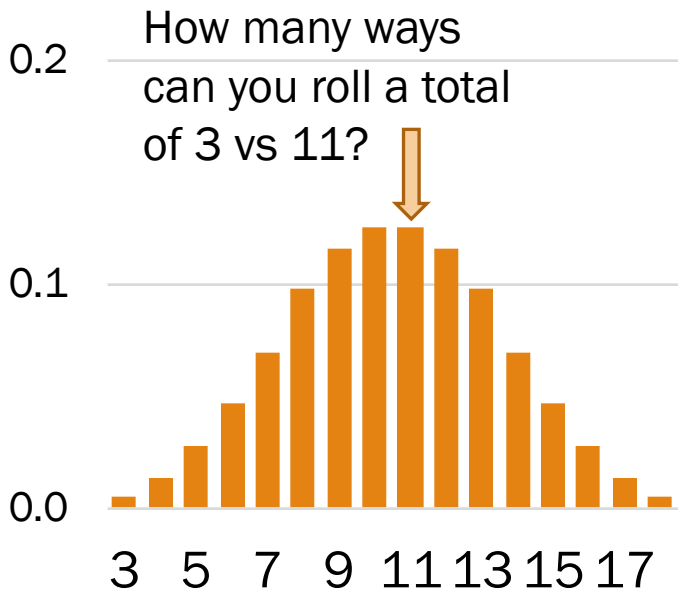
X	Y	i	
0	n	0	$P(X=0, Y=n)$
1	n - 1	1	$P(X=1, Y=n-1)$
2	n - 2	2	$P(X=2, Y=n-2)$
	• • •		
n	0	n	$P(X=n, Y=0)$

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 $n \to \infty$

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

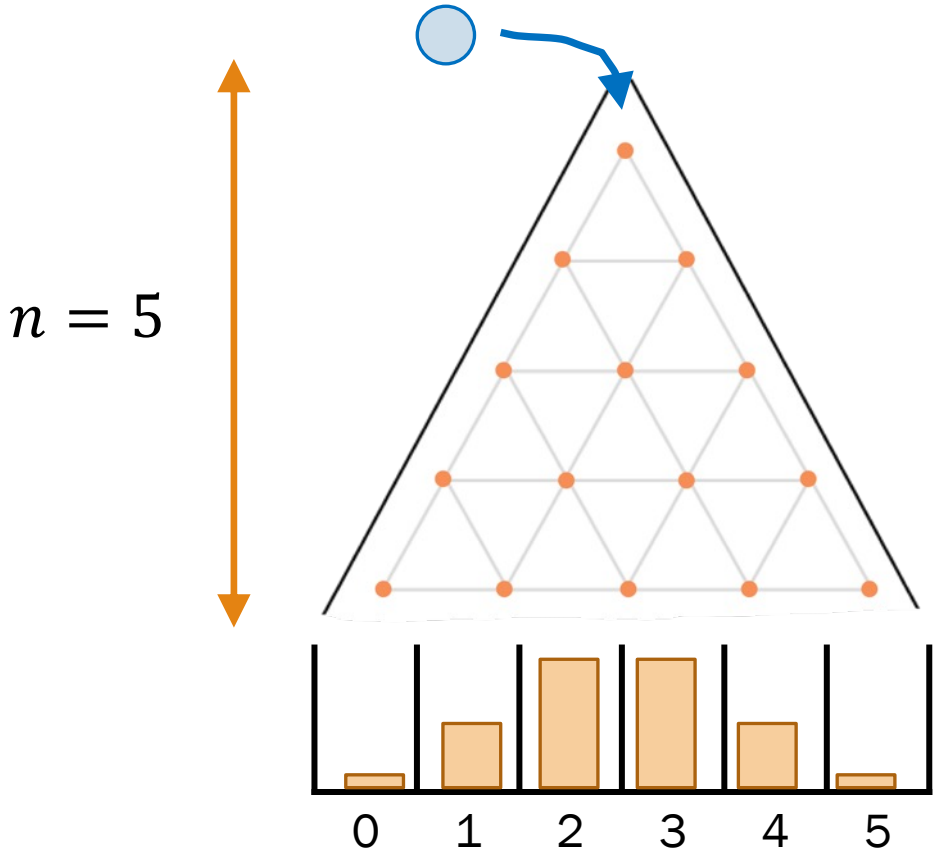
The sum of n i.i.d. random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

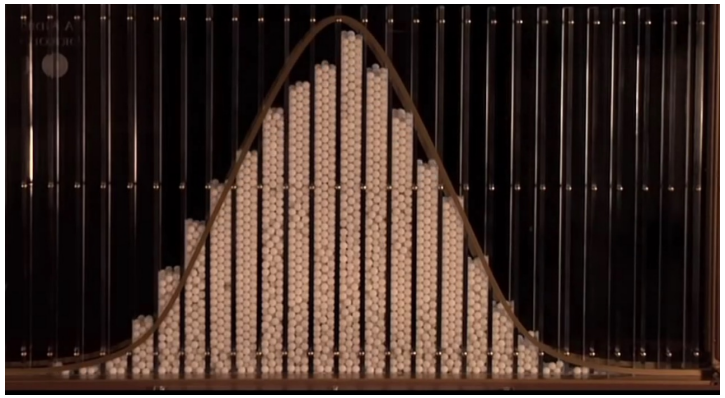
True happiness

CLT explains a lot

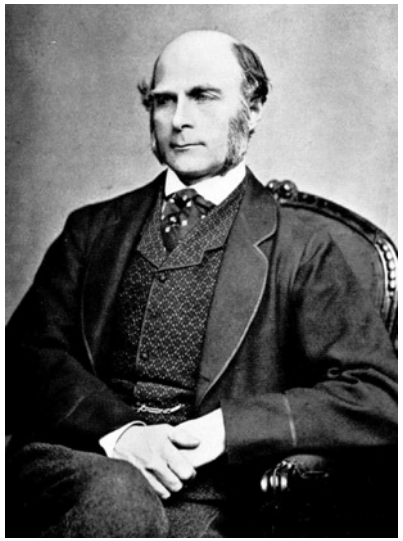
$$\text{As } n \to \infty \qquad \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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





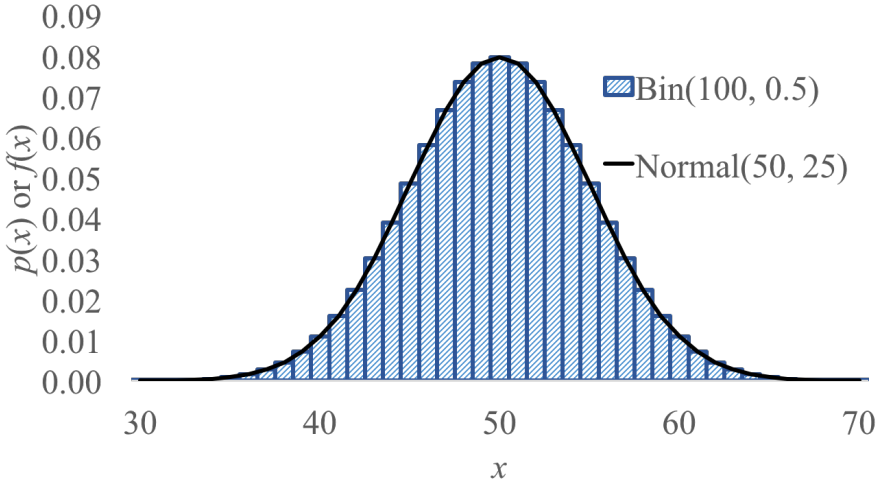
Galton Board, by Sir Francis Galton (1822-1911)



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.

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

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

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

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

It’s play time!

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

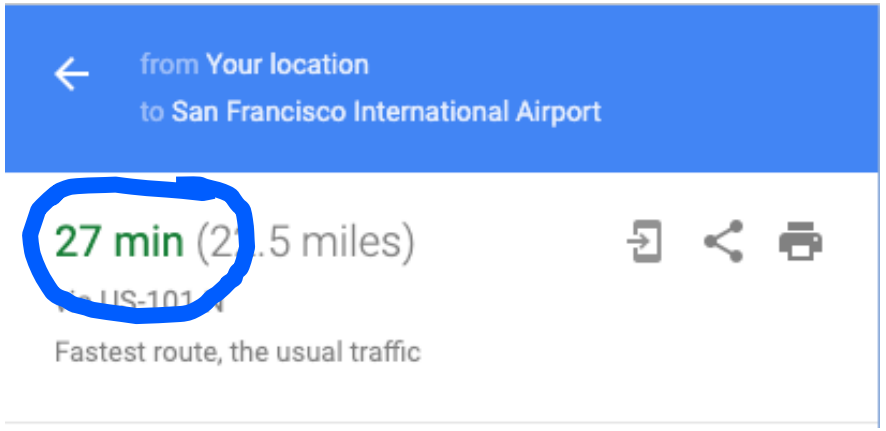
You hit 10 traffic lights on your way to work. You don't know the full distribution of the wait time, but for each you observe the average wait time is 45 seconds and the standard deviation is 5 seconds. You will be on time if your total wait time is less than 8 mins. What is the probability that you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

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

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

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



Proof of CLT

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

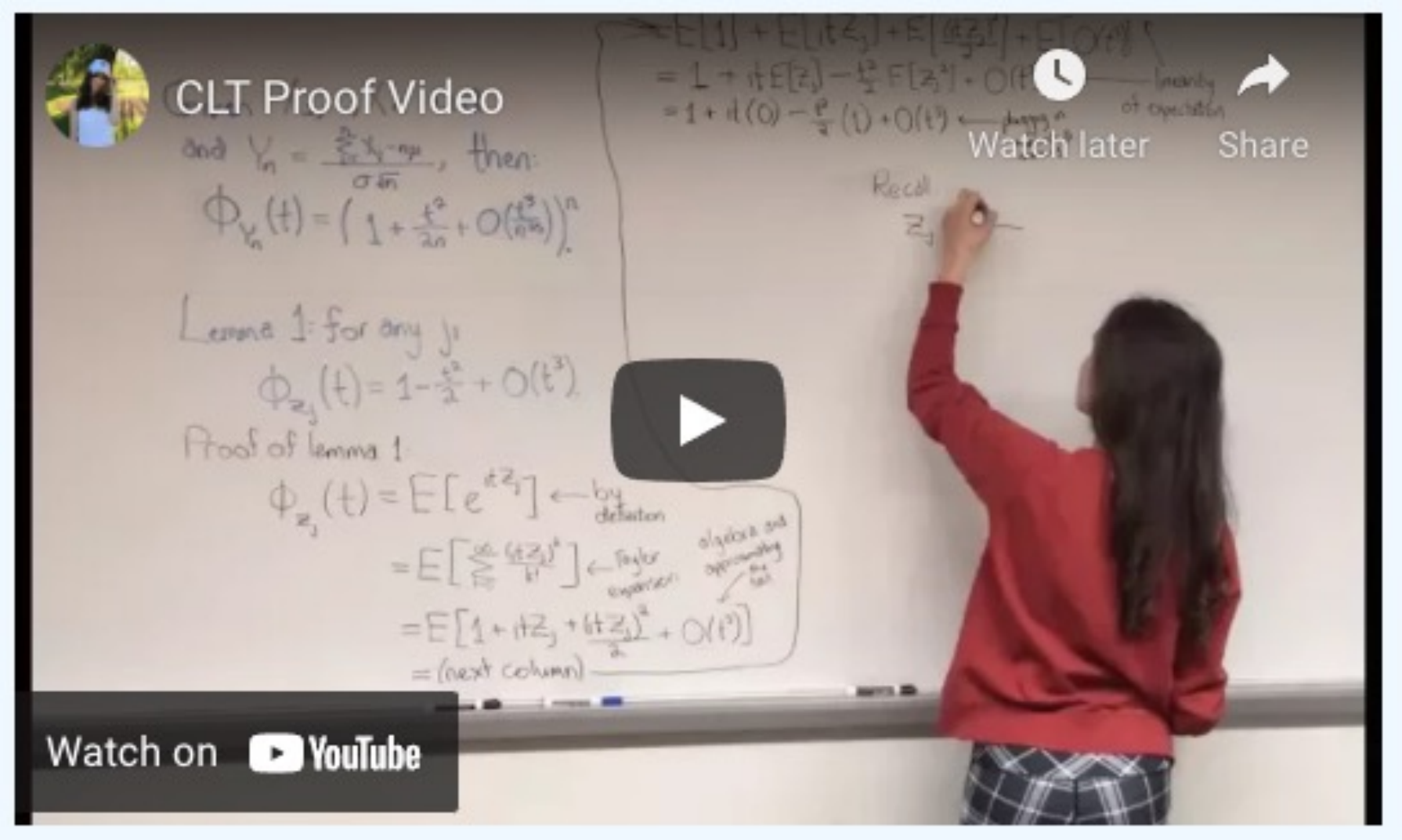
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

Proof:

- The Fourier Transform of a PDF is called a **characteristic function**.
- Take the characteristic function of the probability mass of the sample, normalized
- Show that this approaches an exponential function in the limit as $n \to \infty$: $f(x) = e^{-\frac{x^2}{2}}$
- This function is in turn the characteristic function of a Normal Distribution

(this proof is beyond the scope of CS109)

For Details, 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 $\mu = E[X_i]$

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

What about other functions?

Sum of iid? Normal

Average of iid?

Max of iid?

Average of IID Variables?

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

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

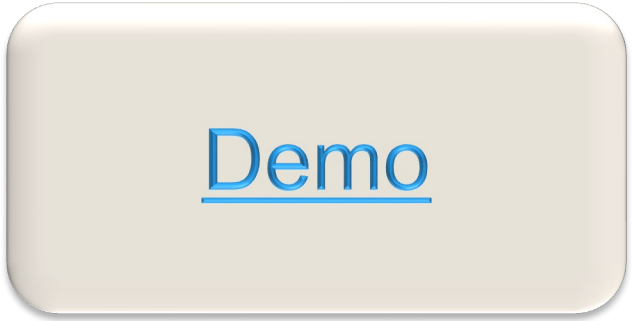
Gaussian by CLT

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

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

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

Average of IID Variables Demo



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

What about other functions?

Sum of iid? Normal

Average of iid? Normal

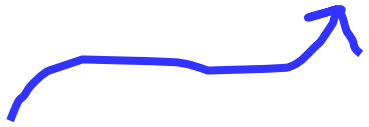
Max of iid?

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: μ = *t* sec.
  - Variance of running time: σ^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}

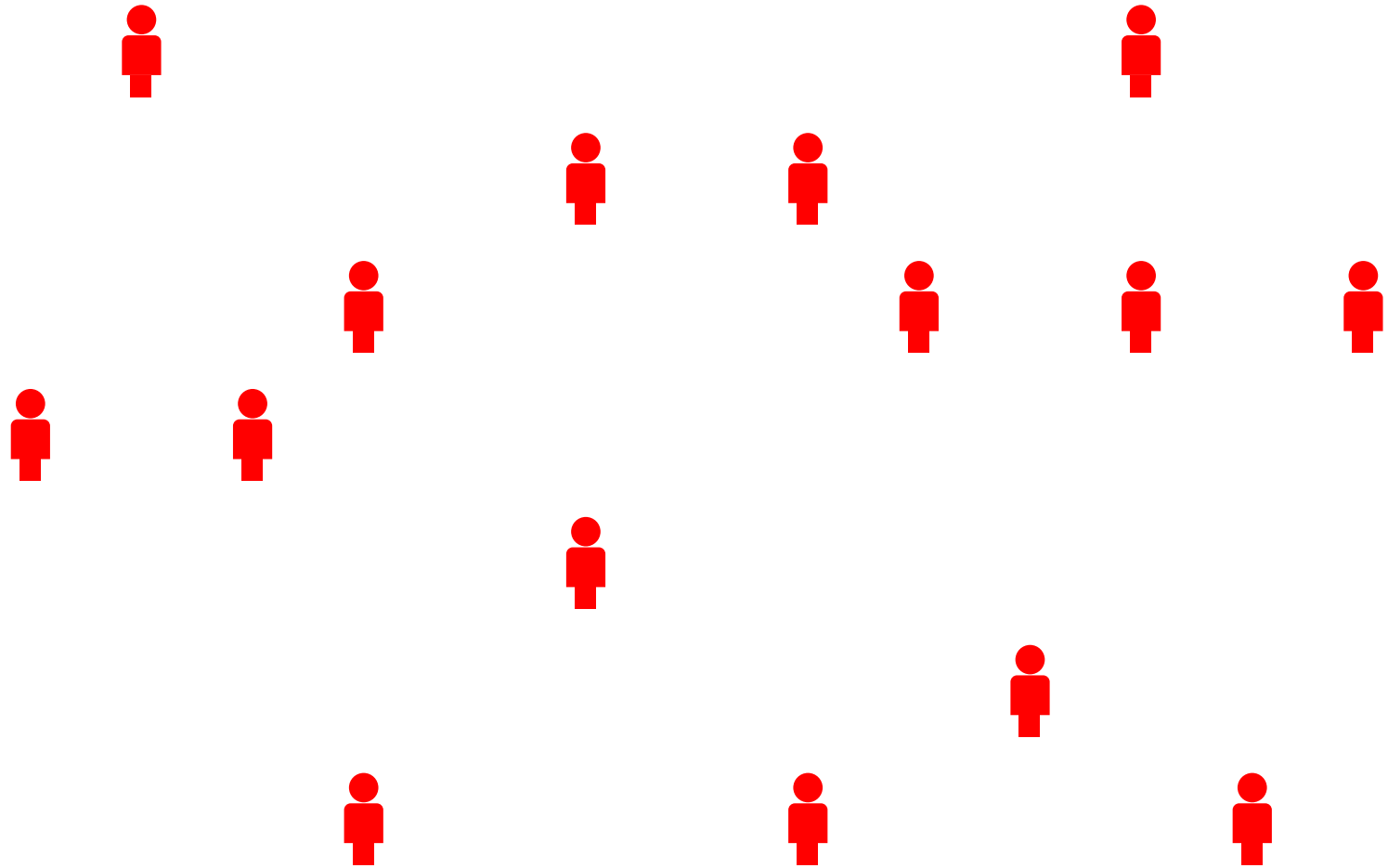
- The mean of all these numbers is 83.

Is this the true mean happiness of Bhutanese people?

Population

Sample

Sample



Collect one (or more) numbers from each person

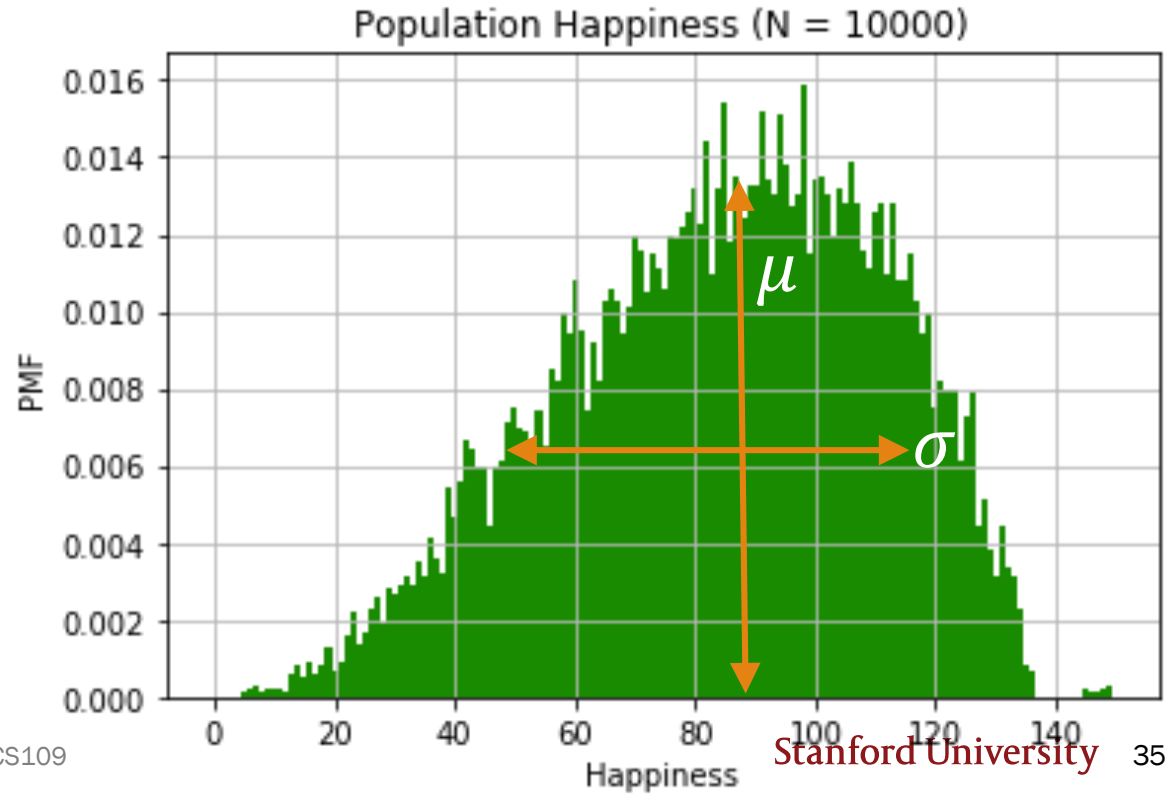
Sample

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



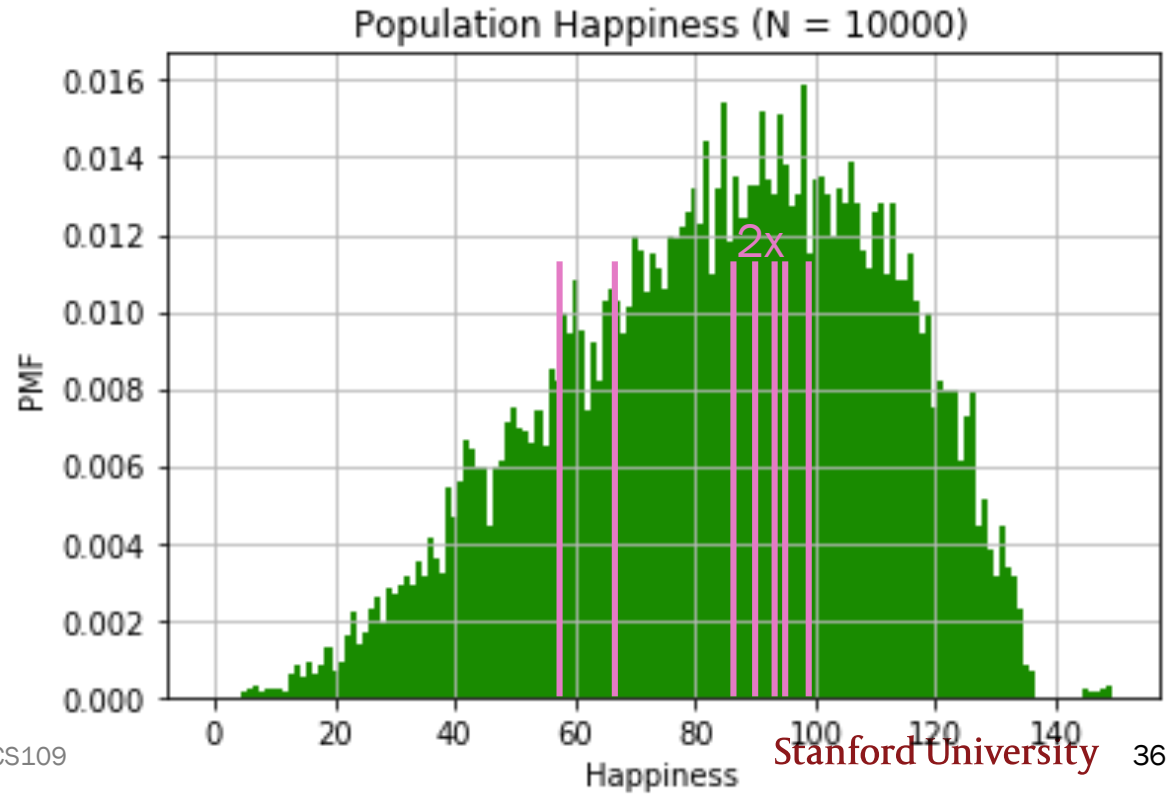
A sample, mathematically

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??*

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

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

Estimating the population mean

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

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

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

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

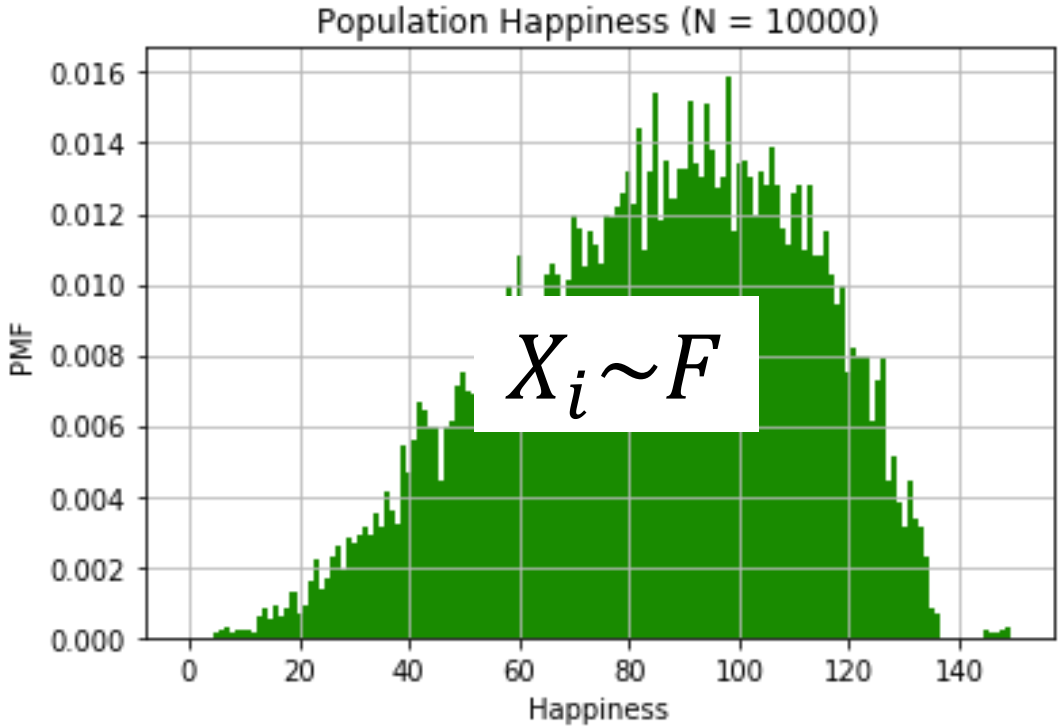
$\bar{X}$ is an **unbiased estimator** of the population mean μ. $E[\bar{X}] = \mu$

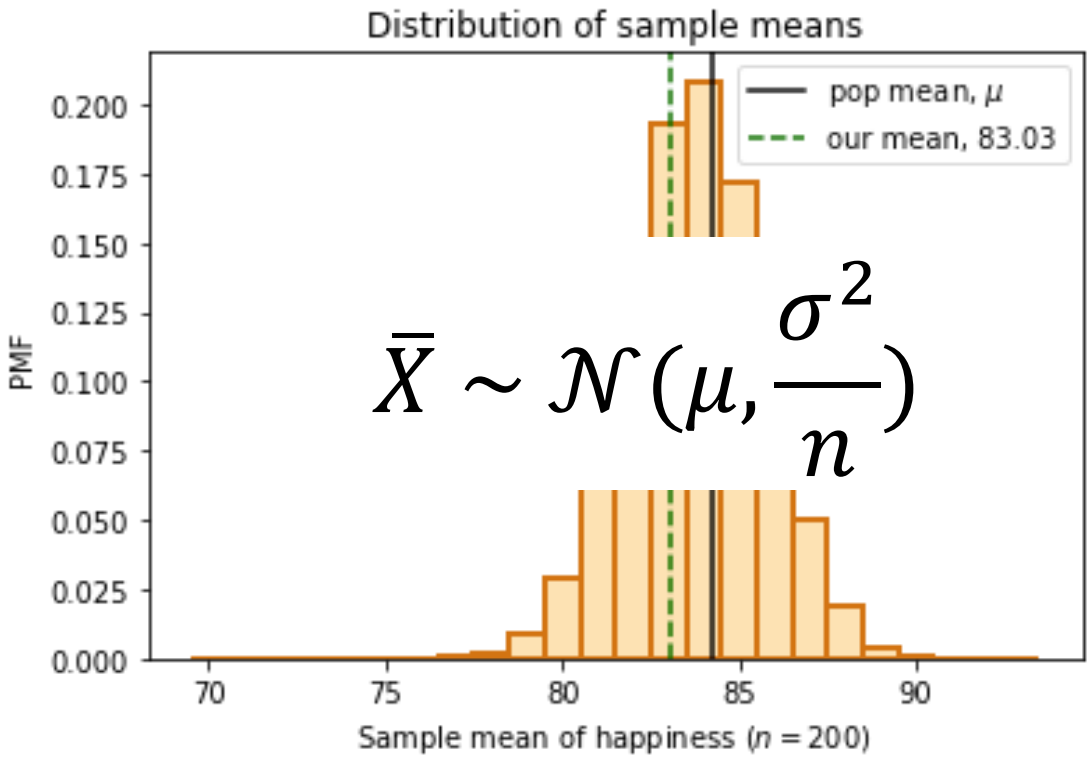
Intuition: By the CLT, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

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

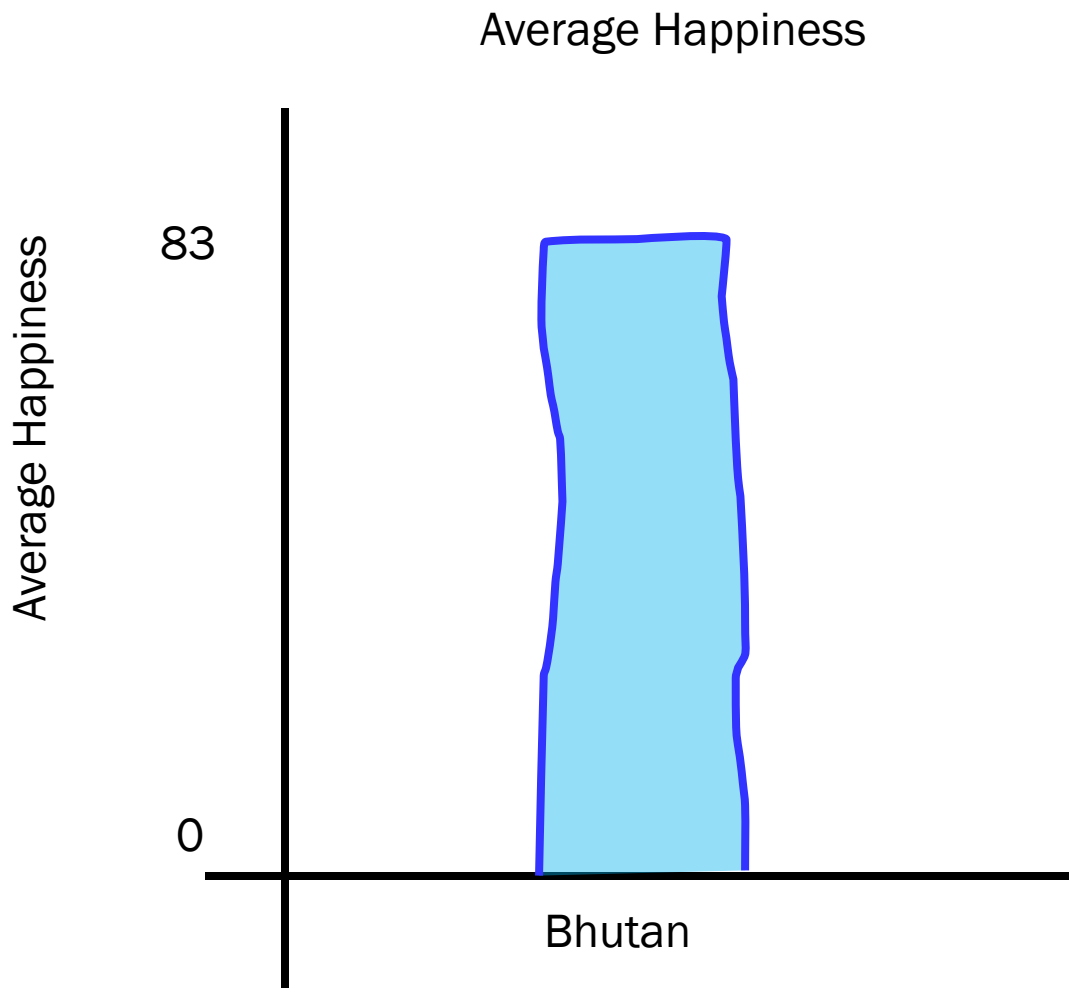
Sample mean





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

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

sample mean

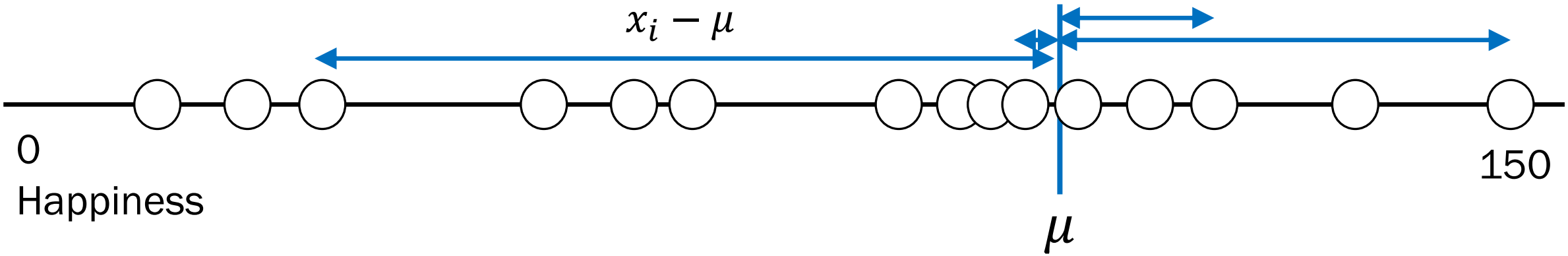
sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

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

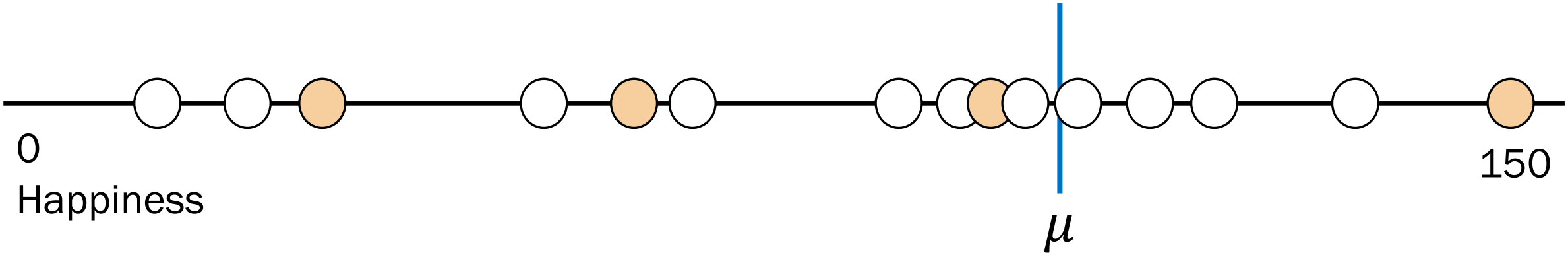
population mean

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

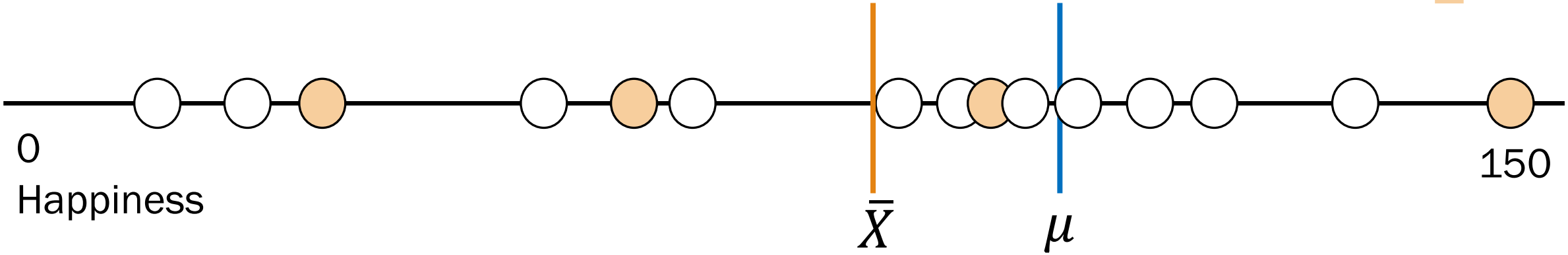
population mean

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

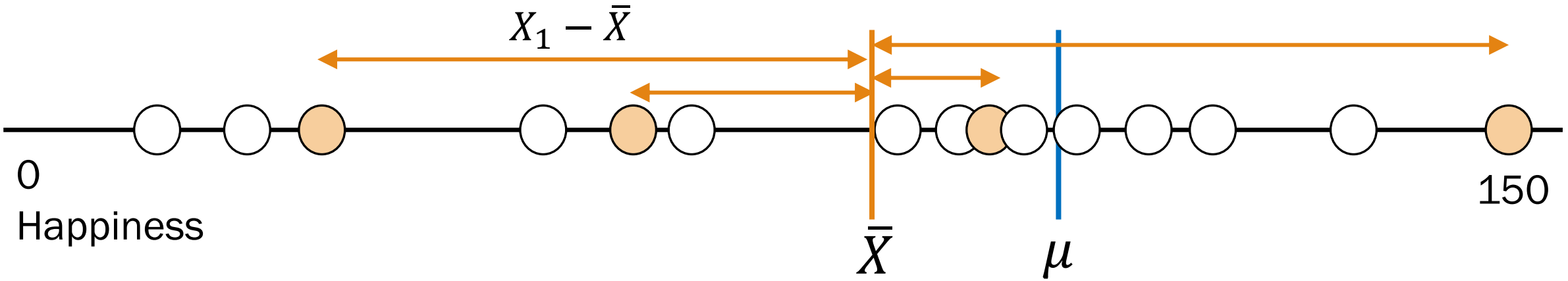
population mean

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?

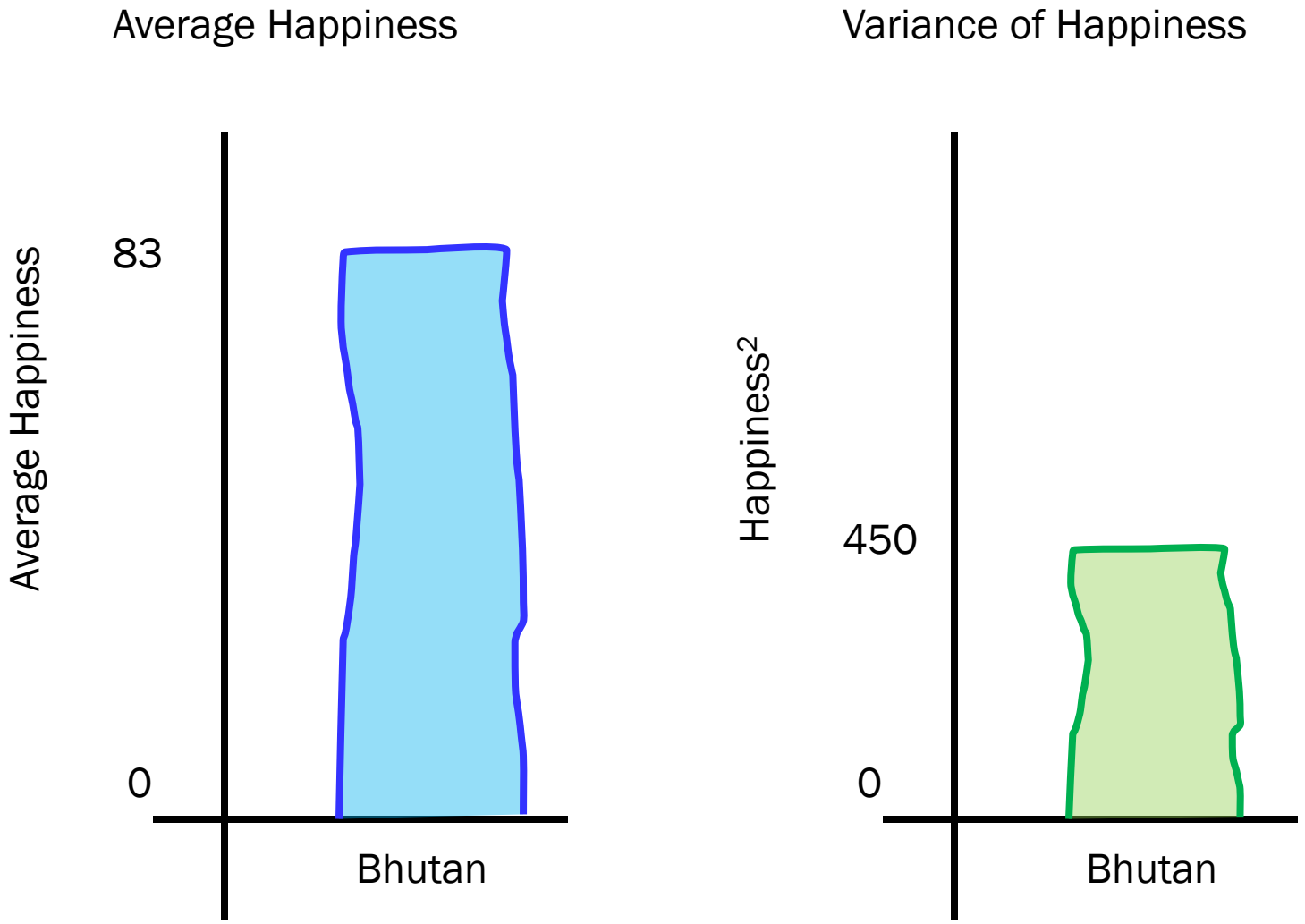
If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

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"

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

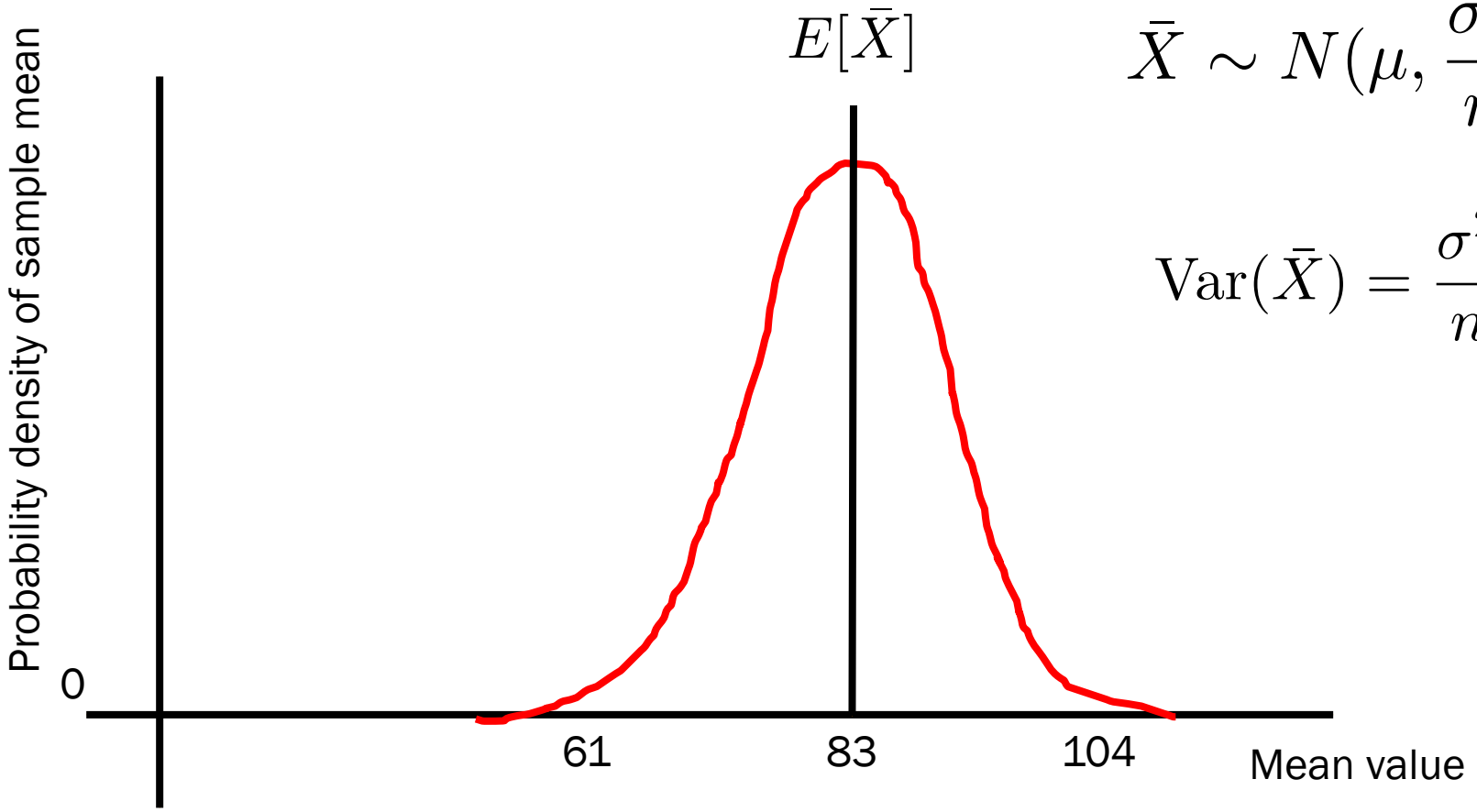
No Error Bars ☹

Insight: Sample Mean is an RV with known Var

By central limit theorem:

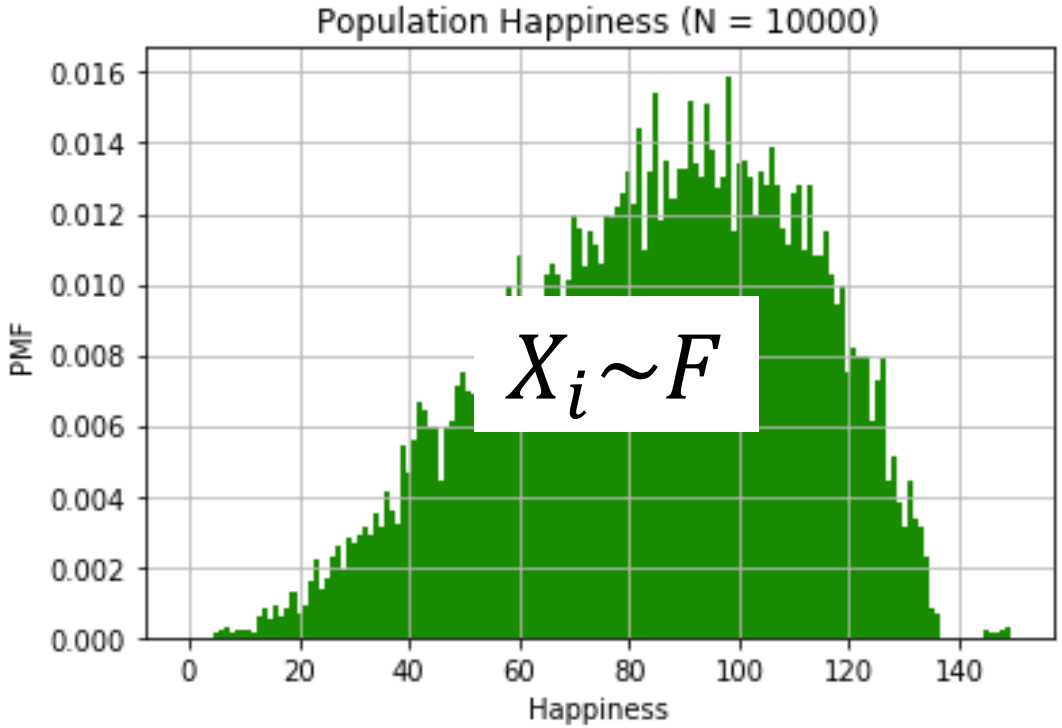
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

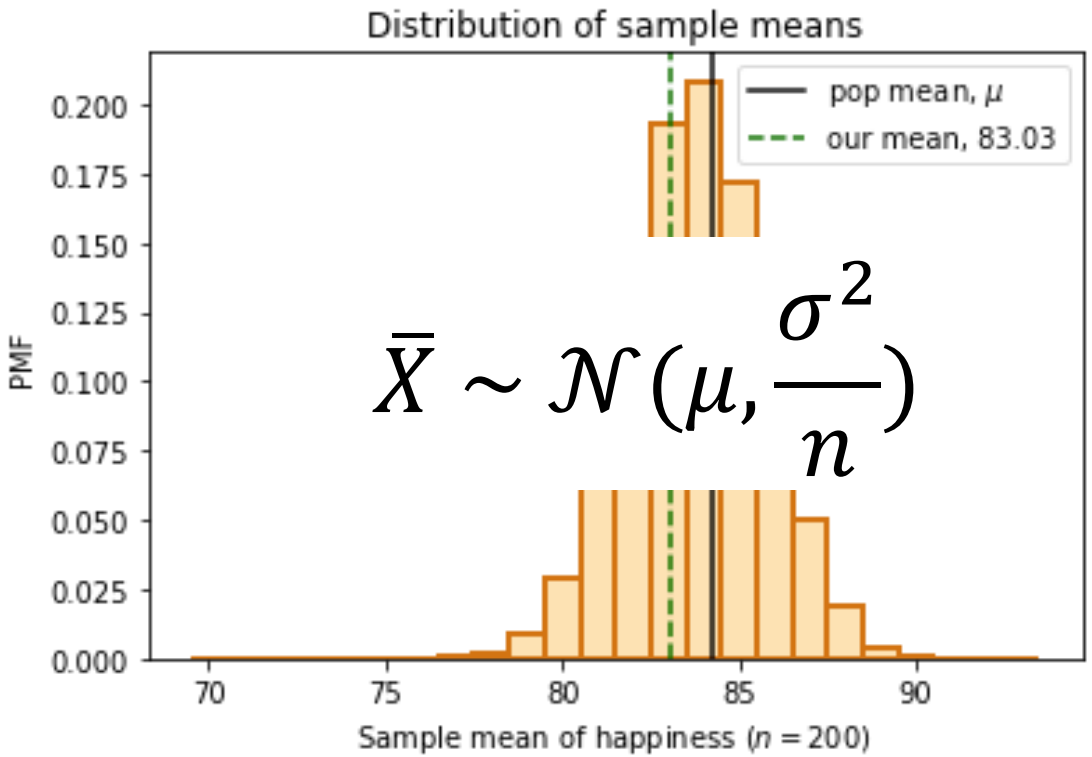
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$



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

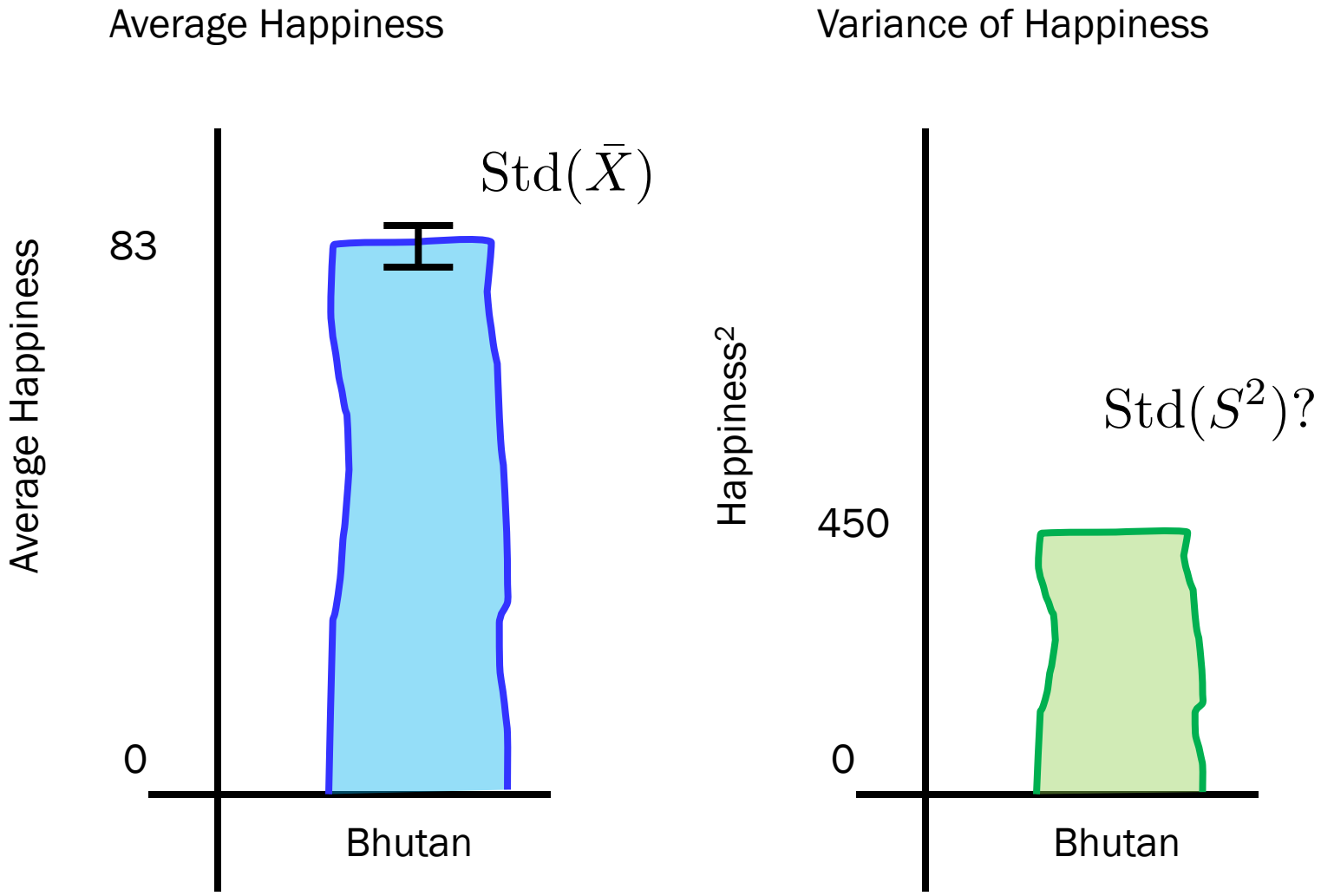
$$= \sqrt{\frac{450}{200}}$$

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

How long does it take CS109 students to complete Psets?

PSet Timings Raw


```
{"pset1-countingcards": [1550.1940480000003, 212.4435679999999, 1722.7378399999996, 531.5406079999998,
2674.1259839999993, 2401.314592, 4181.617599999996, 2420.824192, 276.2297759999999, 1140.
0784479999998, 1206.8157760000001, 560.5036640000001, 931.507056, 1156.8344, 798.4586879999999, 384.
28702399999986, 1050.34776, 1878.2590080000002, 1292.9448640000003, 431.07478400000014, 378.
8209120000001, 818.1577920000002, 2555.404016, 882.420272, 640.076624, 1529.9575360000006, 1232.
9577920000004, 1113.0007039999998, 1846.7954240000001, 1800.8698879999997, 2143.1468569999993, 3184.
684112000002, 1734.1895839999993, 1246.261760000001, 1660.198096000001, 741.8999519999999, 1603.
1233440000017, 710.5813119999996, 5379.367696, 988.8105760000003, 1478.5483519999993, 722.
7267519999994, 578.3234239999999, 518.659968, 1975.309824, 2861.4804320000003, 1358.5260480000002,
1463.927296, 1591.5097599999995, 2572.171295999997, 1580.9689440000004, 2030.1685279999992, 12214.
882255999995, 412.23537600000003, 335.2142080000001, 602.1413119999997, 1871.3165920000006, 1182.
3381600000002, 876.0278399999994, 2358.0574400000014, 627.9507039999999, 1937.6730879999996, 1463.
6703679999996, 478.5180959999999, 1347.6753760000001, 772.8854079999996, 1270.234272, 2764.
7578559999965, 524.3688640000001, 1045.8361919999998, 2220.536159999998, 2045.8206400000013, 2413.
5207999999984, 984.7854719999997, 1006.5262400000004, 869.2664479999995, 745.0748, 0, 385.
2566720000001, 2070.540032, 397.9075039999999, 1666.6341280000004, 751.9160320000001, 2015.
9539999999997, 599.8172480000001, 859.2658560000001, 231.6490880000001, 569.3155360000002, 3276.
267408000002, 3839.6596800000016, 741.7451360000001, 1213.9471680000001, 1824.1588639999995, 586.
0078240000003, 1384.141408, 2077.4697120000005, 1078.9446719999994, 362.7789919999999, 1451.473024,
1229.9791519999997, 388.71222399999994, 261.2879679999998, 870.9573119999999, 1696.6648480000013, 802.
0010240000005, 646.2847200000003, 598.3204320000007, 606.9591680000001, 2392.2958079999985, 508.
41139200000015, 981.5174560000006, 837.928704, 597.6606559999999, 2044.2889280000004, 2387.
9005119999997, 2634.773776000002, 211.579952, 2220.8885120000027, 522.773296, 817.3810079999996, 1136.
7684479999998, 1544.7102240000008, 1365.8311839999997, 883.575952, 448.29513599999984, 3772.
1695199999995, 800.5720319999994, 603.3570560000002, 4195.264224000001, 96.61473600000001, 393.
8498559999999, 533.6176800000002, 325.44255999999984, 706.129216, 2977.463391999999, 3156.
3286400000006, 1732.352831999999, 1258.0552000000002, 214.45435200000009, 862.0855199999994, 2040.
5987840000003, 992.4715680000003, 1434.3783199999996, 1408.3342879999984, 1447.6861600000002, 1884.
4831039999988, 2010.586384000001, 798.7439360000008, 491.58987200000024, 268.505664, 1778.
2884160000015, 1060.6238560000006, 259.94641600000006, 2381.6392799999985, 500.35900800000013, 405.
9288, 1949.8160640000012, 1476.9774559999992, 1552.3747040000007, 504.5615200000002, 1510.
8626239999999, 1034.3021600000004, 3131.245696000002, 1577.3348639999995, 1226.4924480000002, 911.
3914399999999, 234.32780800000006, 1005.4795199999999, 1360.6999999999994, 989.1745439999997, 1390.
004112, 739.4614079999991, 454.34560000000016, 2438.610176000001, 1758.4230880000005, 858.
0467199999998, 462.3268800000001, 227.134608, 569.5653920000002, 859.4128159999997, 663.267392, 778.
0543839999998, 1381.3501920000003, 10107.814288000001, 1320.0240319999998, 1885.6292959999998, 1382.
9131199999993, 430.007232, 457.6967359999998, 1322.0688800000012, 594.5272960000001, 1302.
5745280000006, 427.92484799999994, 762.2496479999998, 280.673488, 2715.640207999996, 1038.
```


For each pset question:
I have a list of how many seconds it took each person

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

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

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

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

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

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

PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

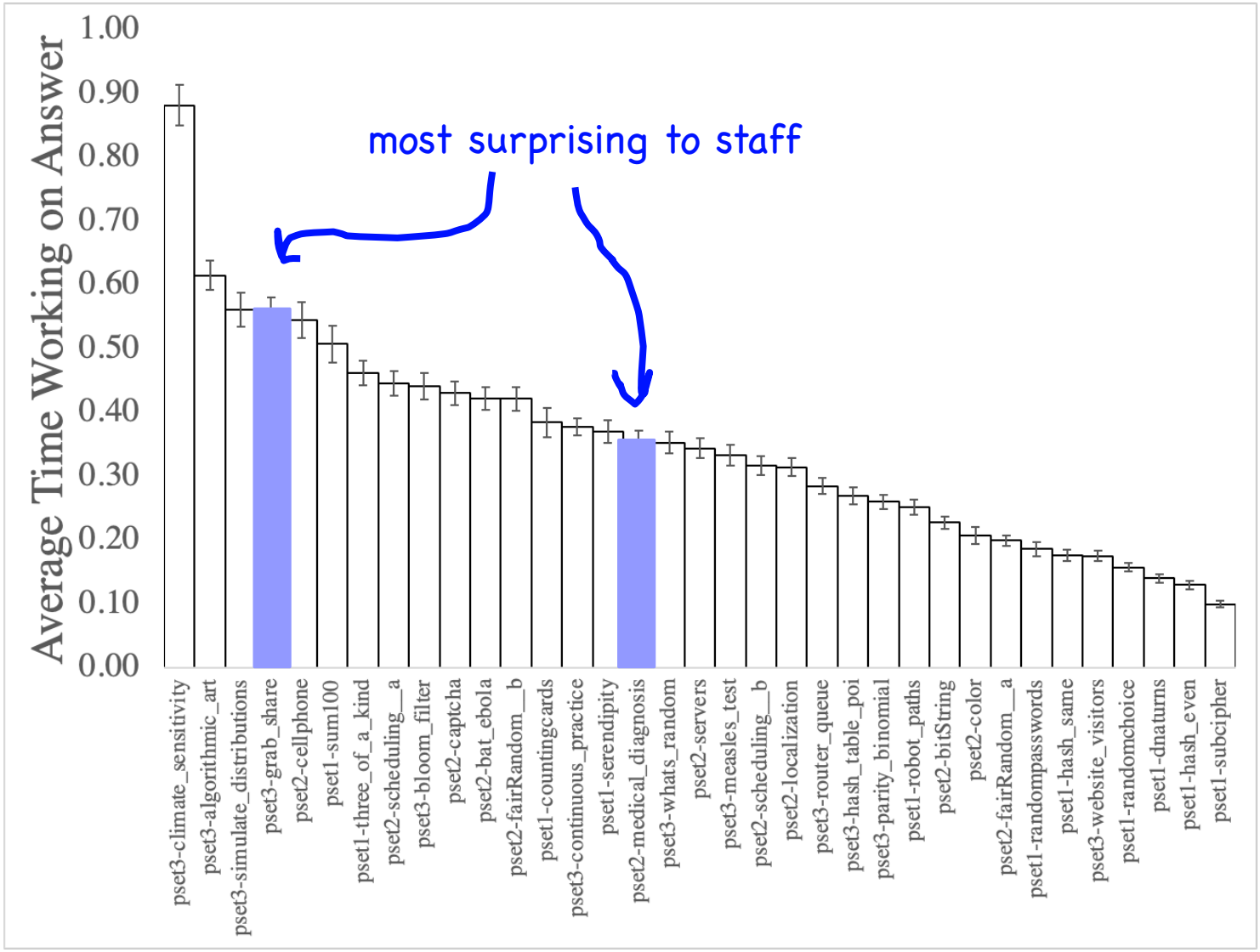
PSet Timings Stats Code

```
def analyse(data):
    for question_key, timings_list in data.items():
        timings_list = remove_zeros(timings_list)
        # calculate n
        n = len(timings_list)
        # estimate the mean
        sample_mean = np.mean(timings_list)
        # estimate the variance (ddof=1 says its a sample... i know...)
        sample_var = np.var(timings_list,ddof=1)
        # calculate the standard error of the mean
        standard_err = math.sqrt(sample_var / n)
        # sample std
        sample_std = math.sqrt(sample_var)
        # print them out
        display_name = question_key[:12]
        print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd:
              {sample_std:.1f}')
```

PSet Timings Stats Output

```
pset1-counti,   mean: 1385.0 ± 83.8,    std: 1288.0
pset1-dnatur,   mean: 504.2 ± 25.4,     std: 390.3
pset1-hash_e,   mean: 467.3 ± 22.0,     std: 336.4
pset1-hash_s,   mean: 634.0 ± 30.7,     std: 471.4
pset1-random,   mean: 567.5 ± 23.3,     std: 357.8
pset1-random,   mean: 671.4 ± 39.6,     std: 606.3
pset1-robot_,   mean: 906.0 ± 42.2,     std: 648.6
pset1-serend,   mean: 1335.3 ± 63.9,    std: 978.0
pset1-subcip,   mean: 359.8 ± 19.9,     std: 304.8
pset1-sum100,   mean: 1827.2 ± 106.2,   std: 1627.3
pset1-three_,   mean: 1663.9 ± 70.2,    std: 1078.7
pset2-bat_eb,   mean: 1520.1 ± 62.3,    std: 942.3
pset2-bitStr,   mean: 819.2 ± 34.0,     std: 523.6
pset2-captch,   mean: 1551.0 ± 67.4,    std: 1030.6
pset2-cellph,   mean: 1962.0 ± 102.7,   std: 1567.3
pset2-color,    mean: 745.1 ± 48.3,     std: 742.9
pset2-fairRa,   mean: 717.5 ± 27.2,     std: 418.8
pset2-fairRa,   mean: 1518.2 ± 67.6,    std: 1035.6
pset2-locali,   mean: 1132.4 ± 49.1,    std: 750.6
pset2-medica,   mean: 1275.3 ± 64.2,    std: 981.5
pset2-schedu,   mean: 1606.8 ± 70.1,    std: 1079.0
```

Sampling statistics on your Psets



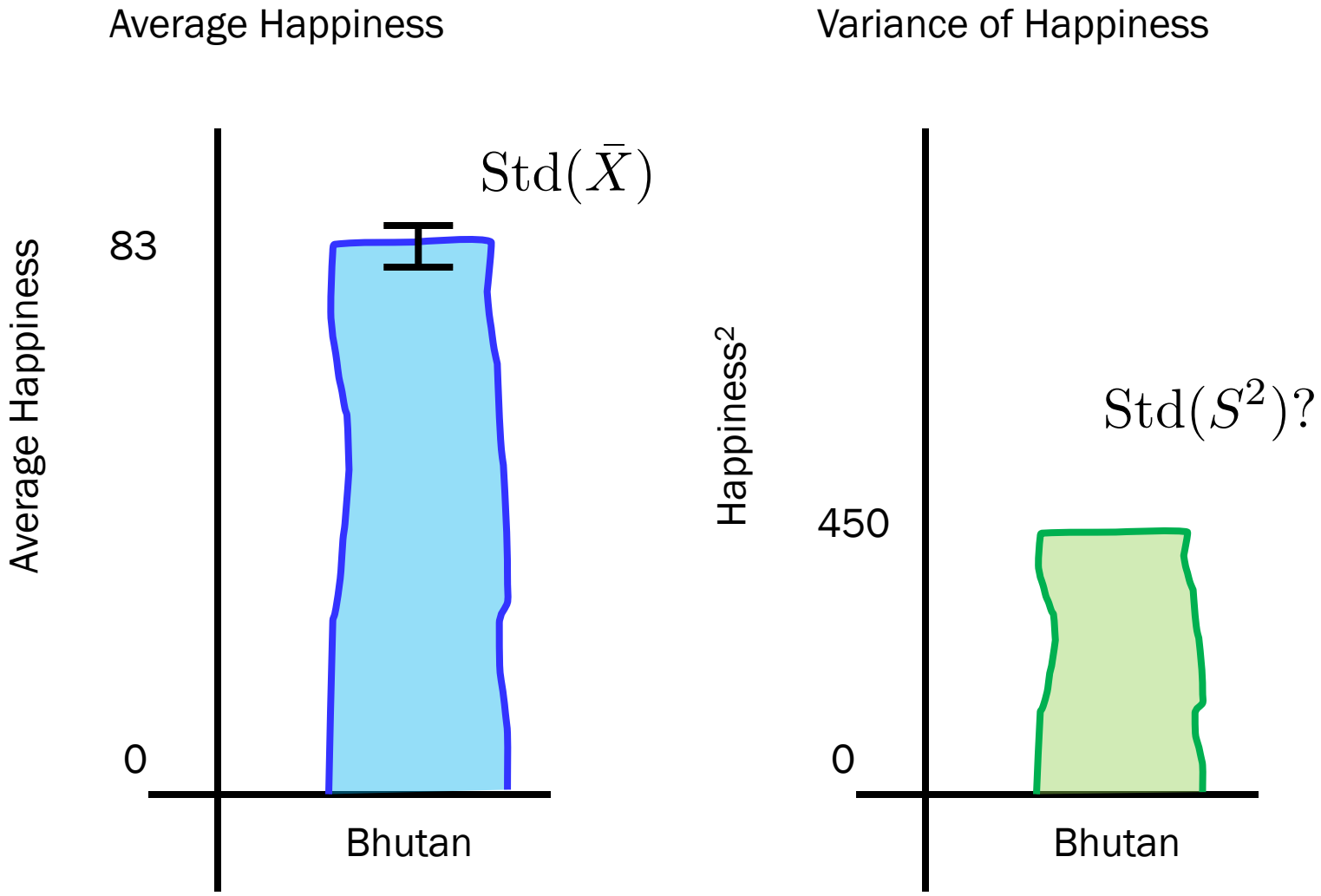
Error bars are standard error of the mean

Expectation of the sum of problems is sum of expectations:

pset1: 2.87 hours on answers
pset2: 4.23 hours on answers
pset3: 5.11 hours on answers

Total: 12.1 hours on answers
Budget: 50 hours for psets

Our Report to Bhutan Government



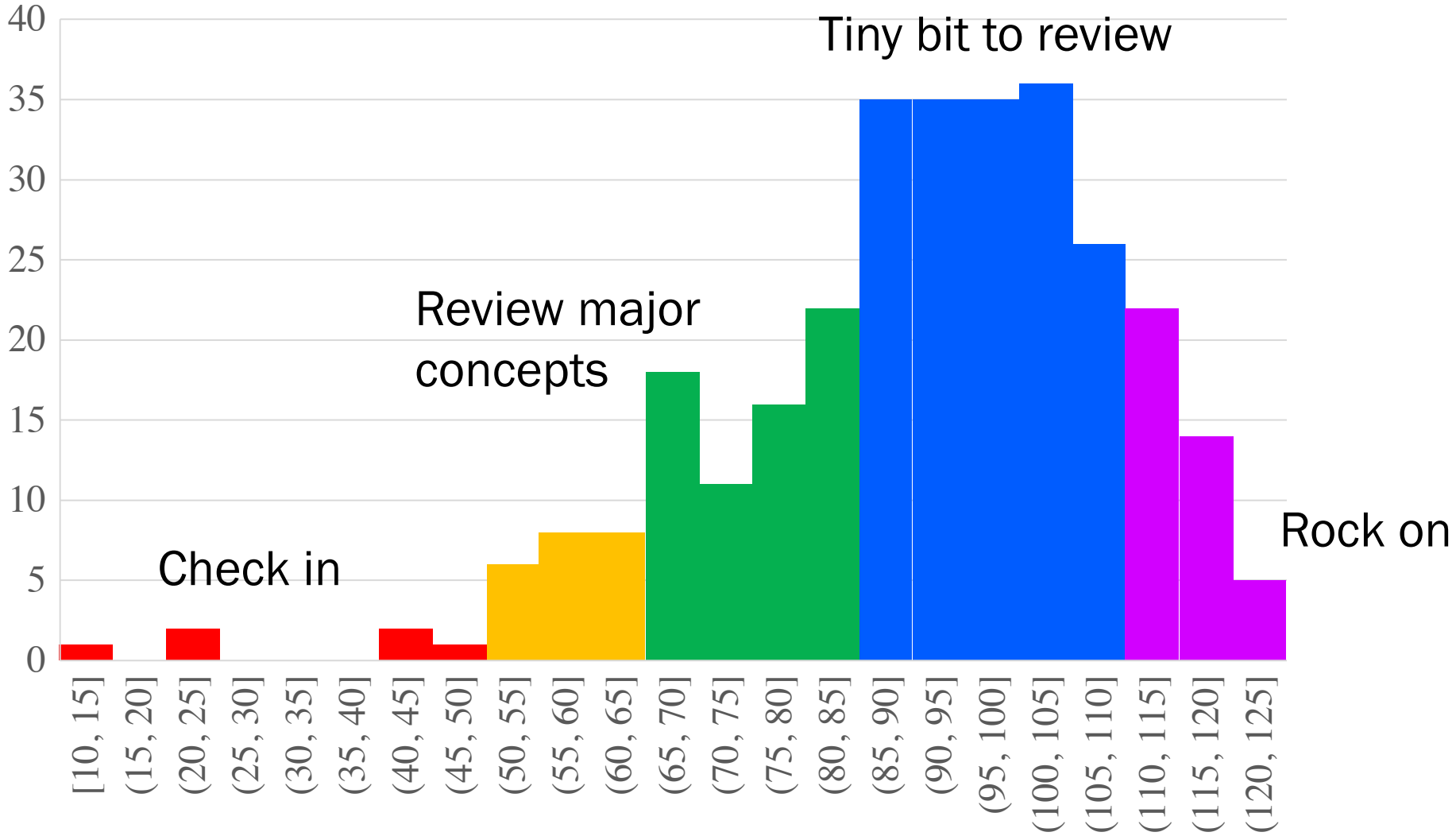
Claim: The average happiness of Bhutan is 83 ± 2

Bootstraping

Come back on Wed!

Midterm

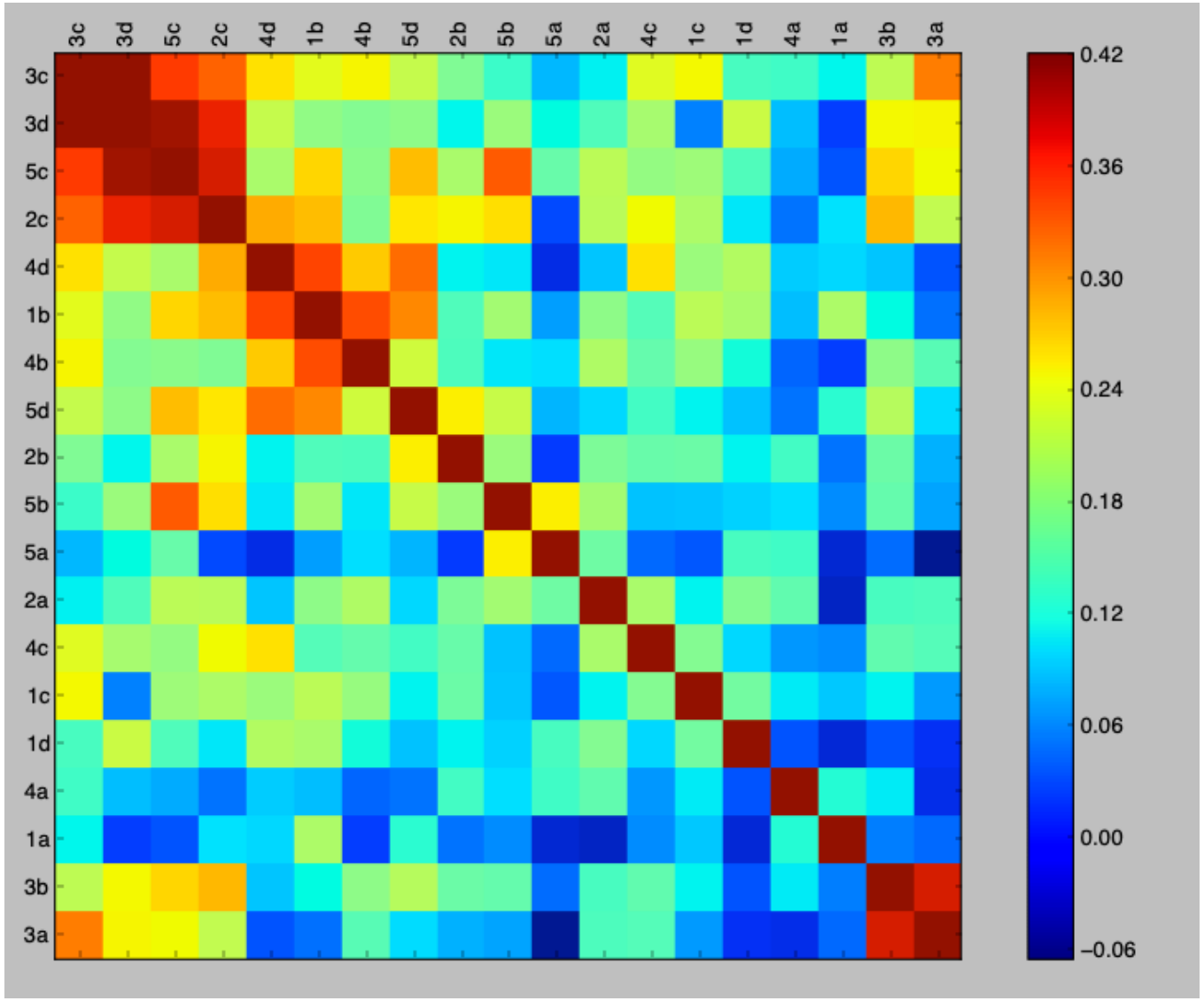
Grade Distribution



$$\mu = \frac{91}{120}$$

$$\sigma = \frac{19}{120}$$

Midterm Correlations



Midterm Logistics (from Chris)

Regrade requests:

- Submit by Friday
- Reserve the right to regrade the whole test

Grade is still under your control:

- Think of the midterm as a diagnostic. Found an area to improve?
- Chris always looks for growth between the midterm and final
- Optional contest

Tuesday / Wednesday exams were slightly similar but no signs of statistical difference, after building a prior based on time to finish psets.


NOV
29TH