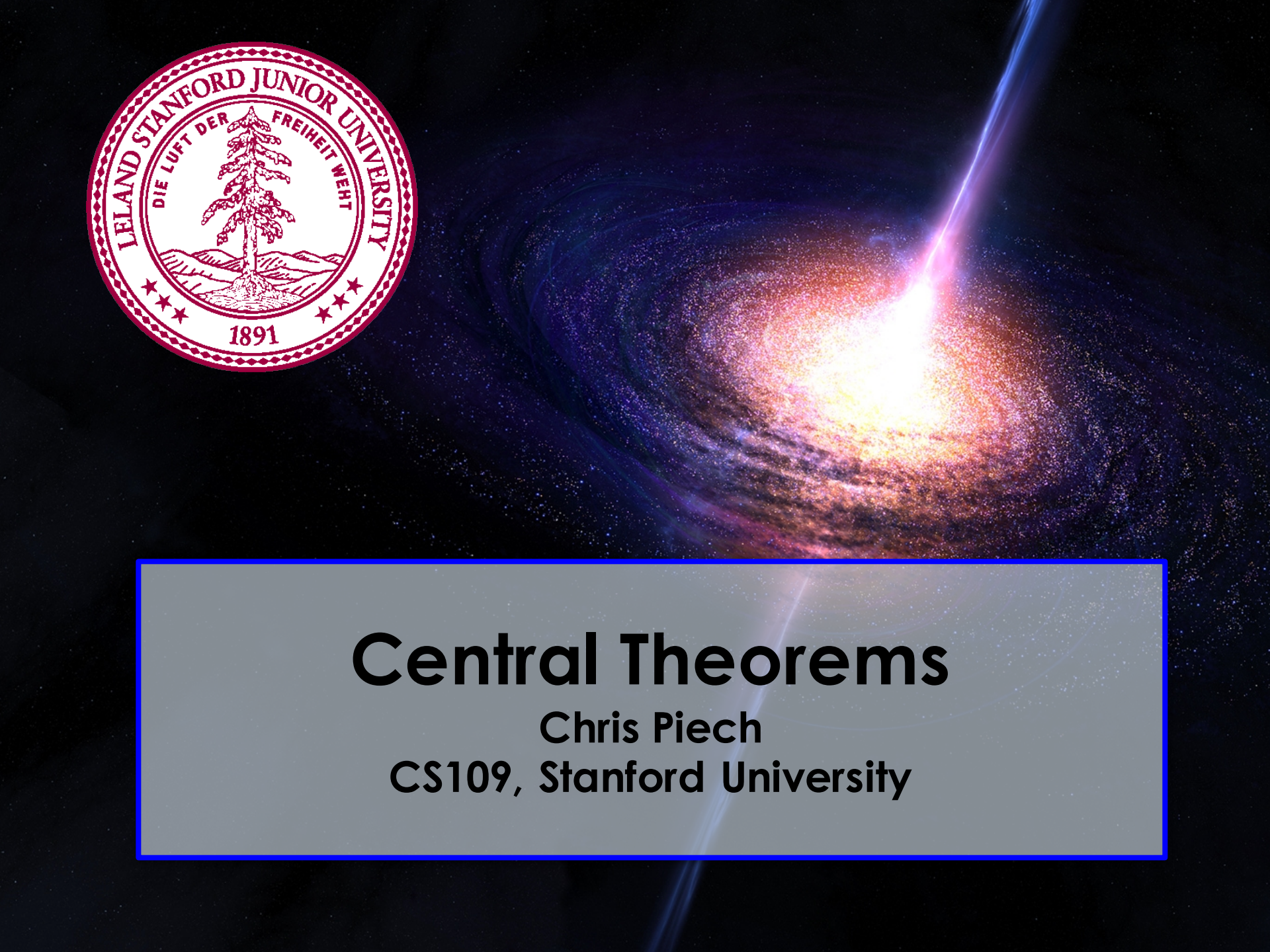
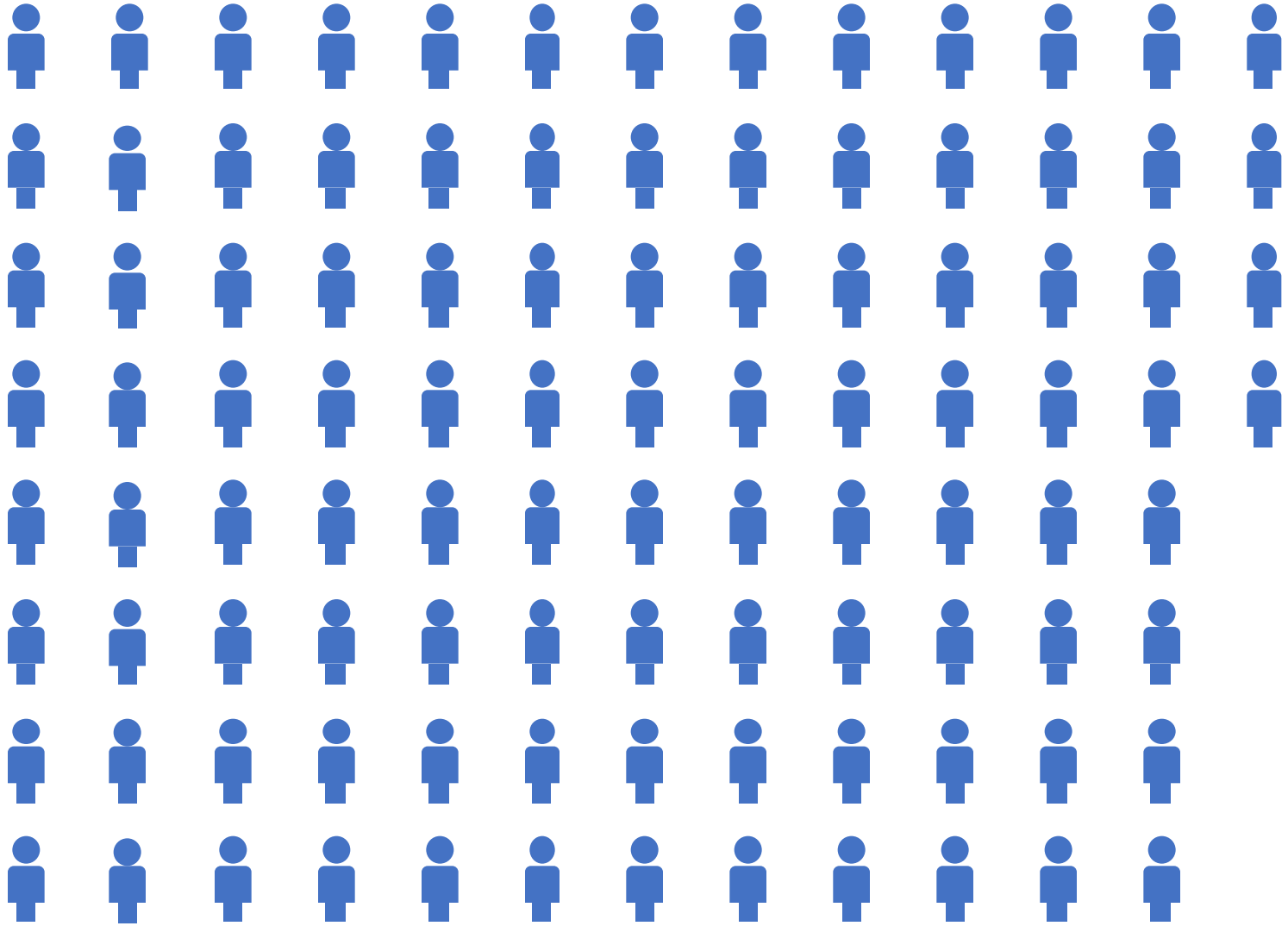


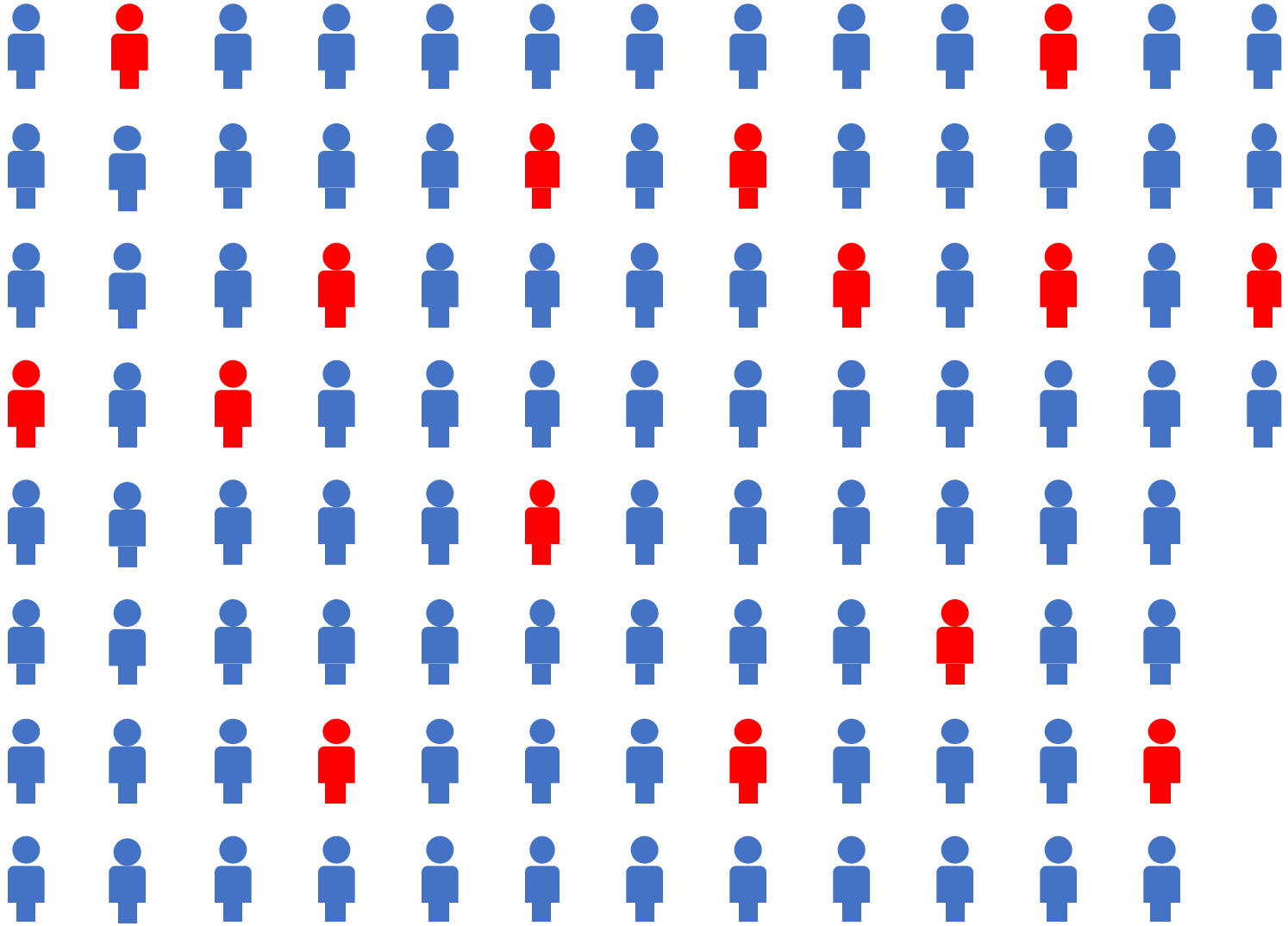
Central Theorems

Chris Piech
CS109, Stanford University

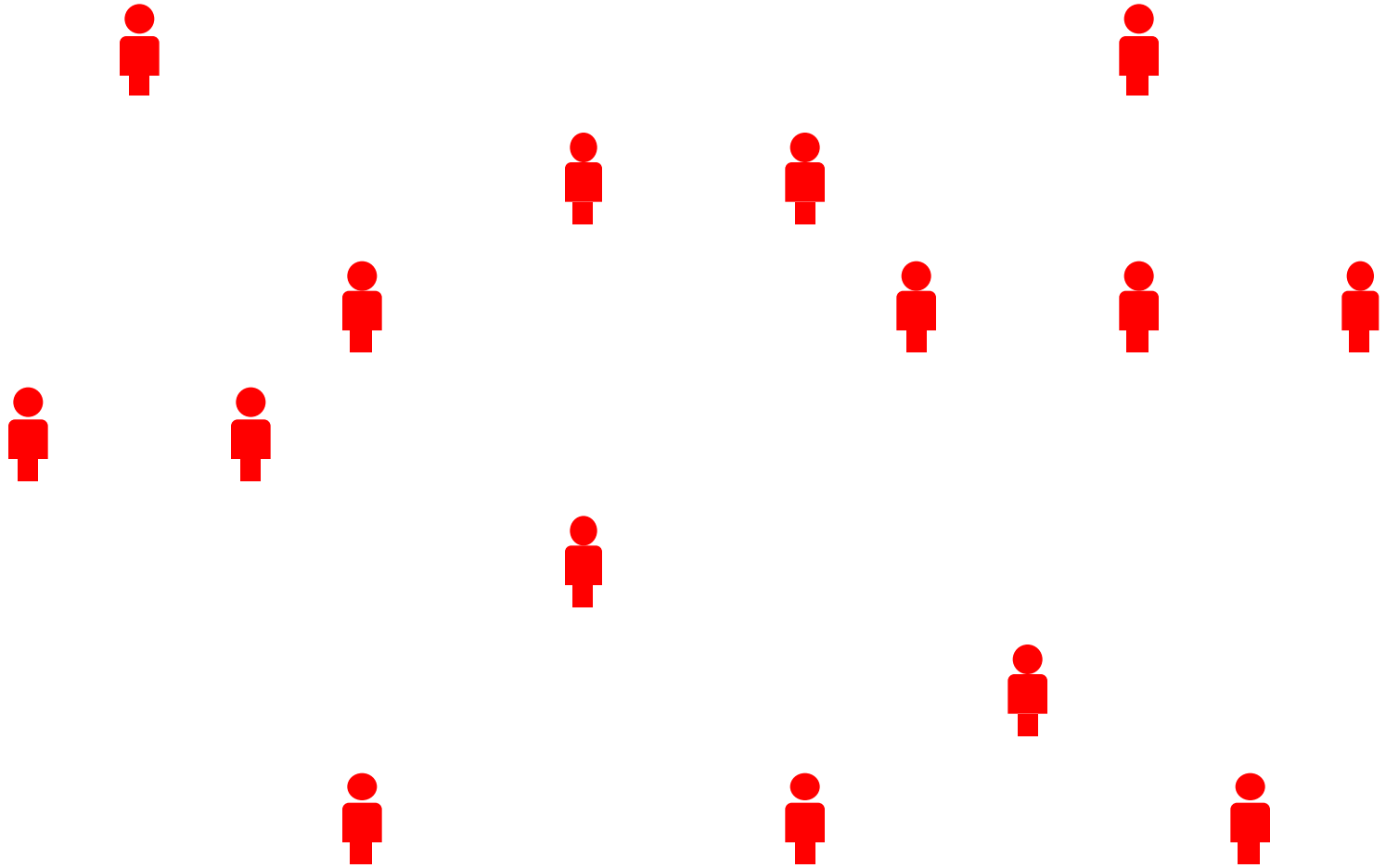
Population



Sample

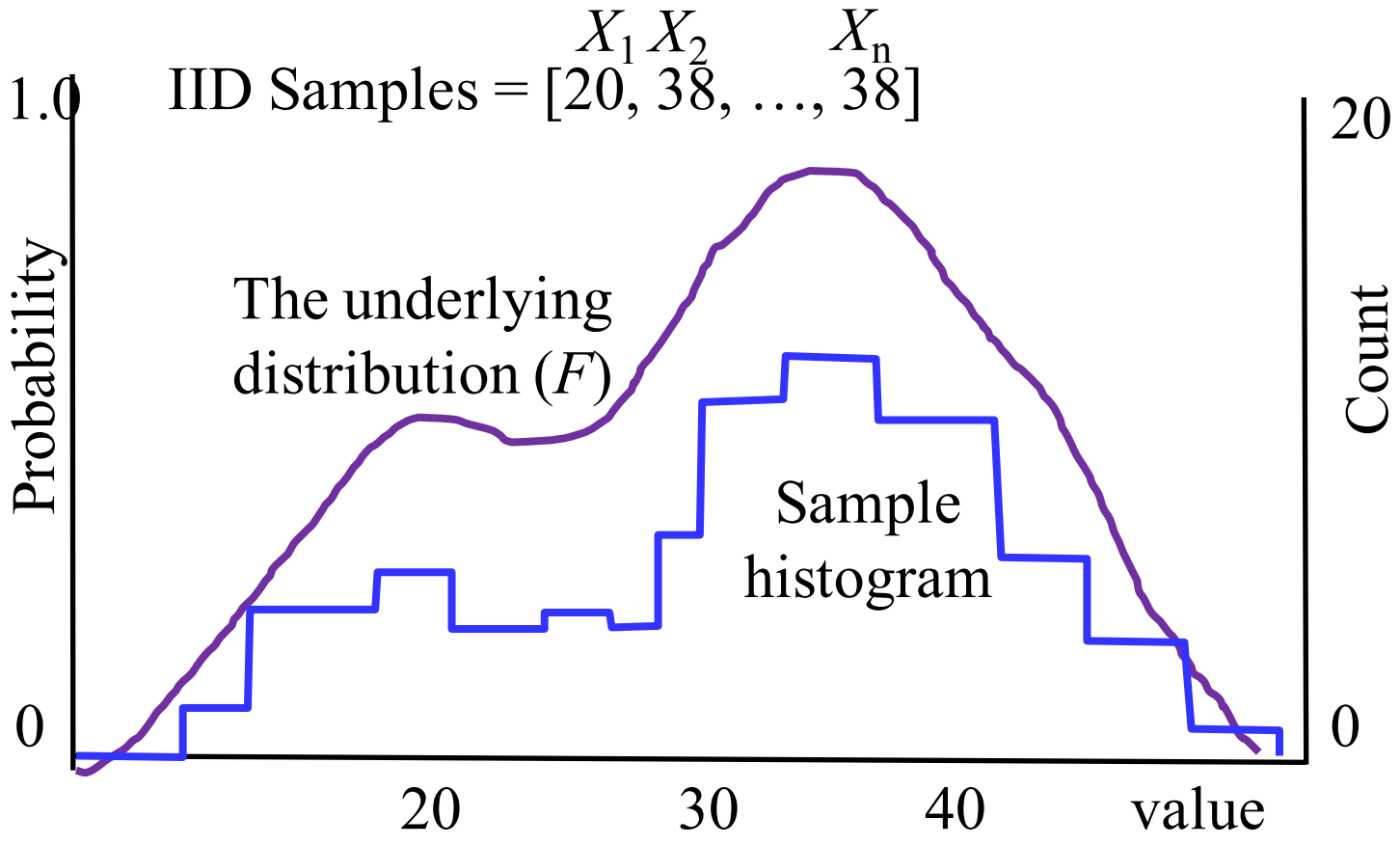


Sample



Collect one (or more) numbers from each person

Samples


X_1 X_2 X_n
IID Samples = [20, 38, …, 38]
1.0
0
Probability
20
Count
0
The underlying
distribution (F)
Sample
histogram
20
30
40
value

Sample Statistics

Sample Mean

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

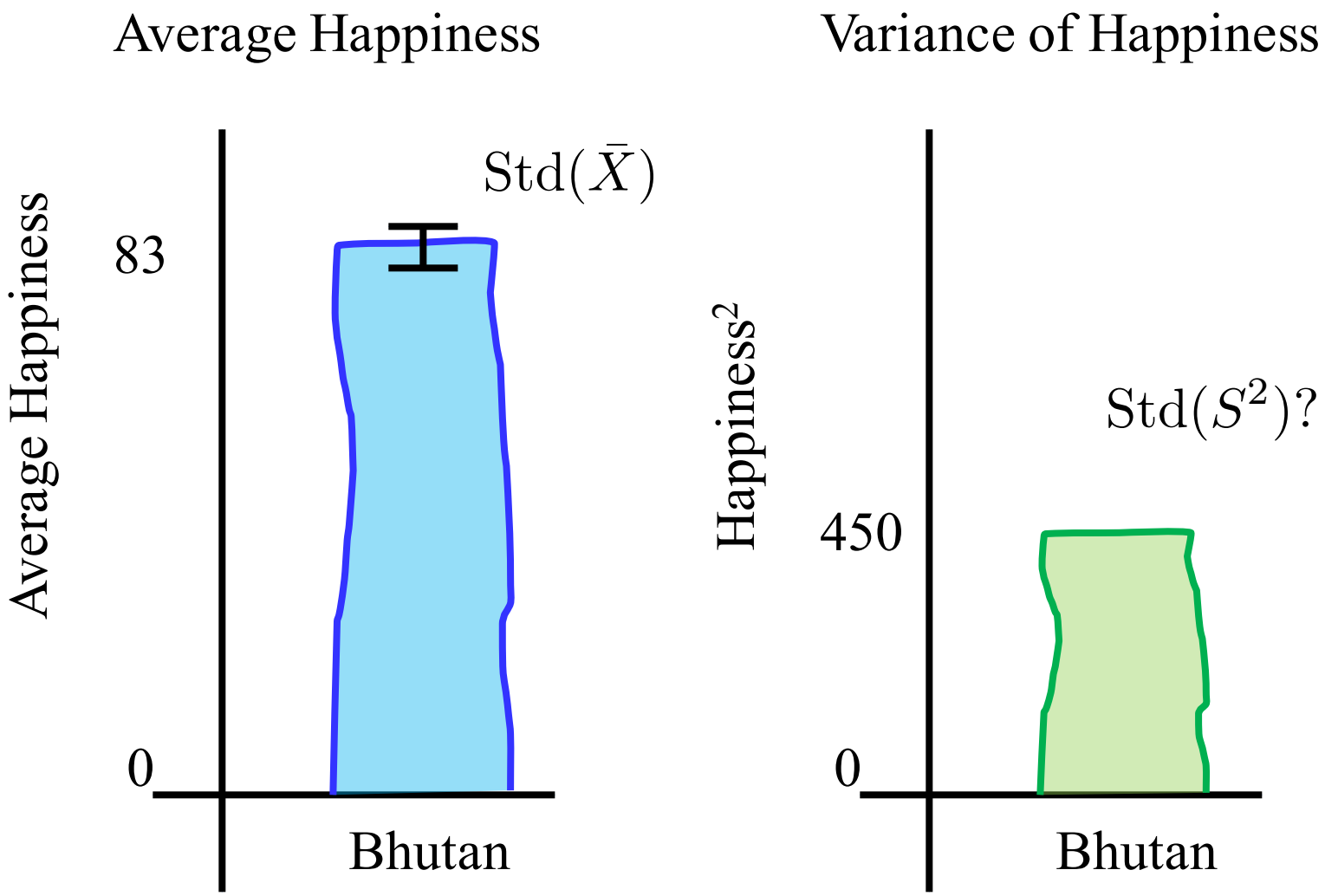
Sample Variance

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

Var of Sample Mean

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

Sample Mean



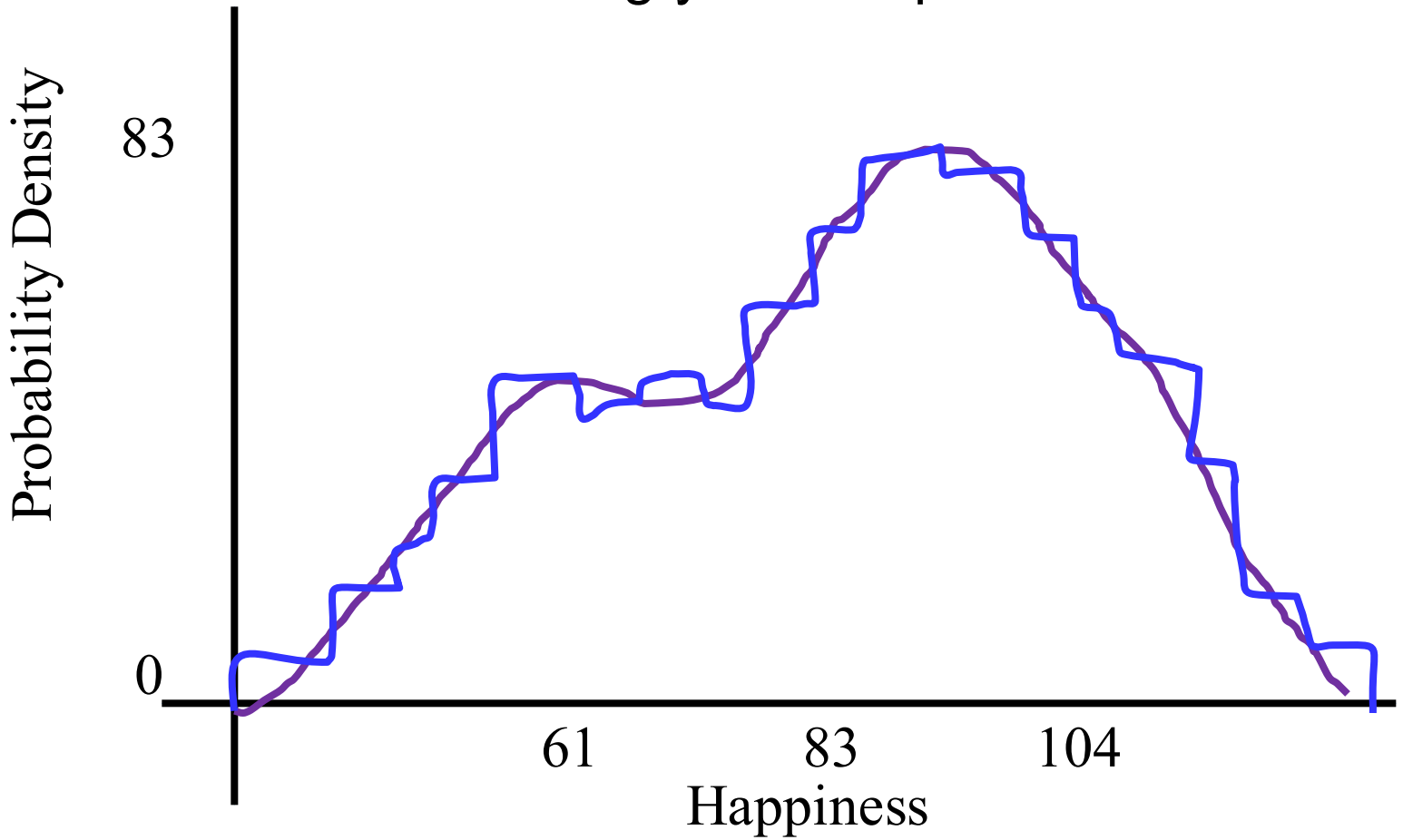
Claim: The average happiness of Bhutan is 83 ± 2

What if you want more?

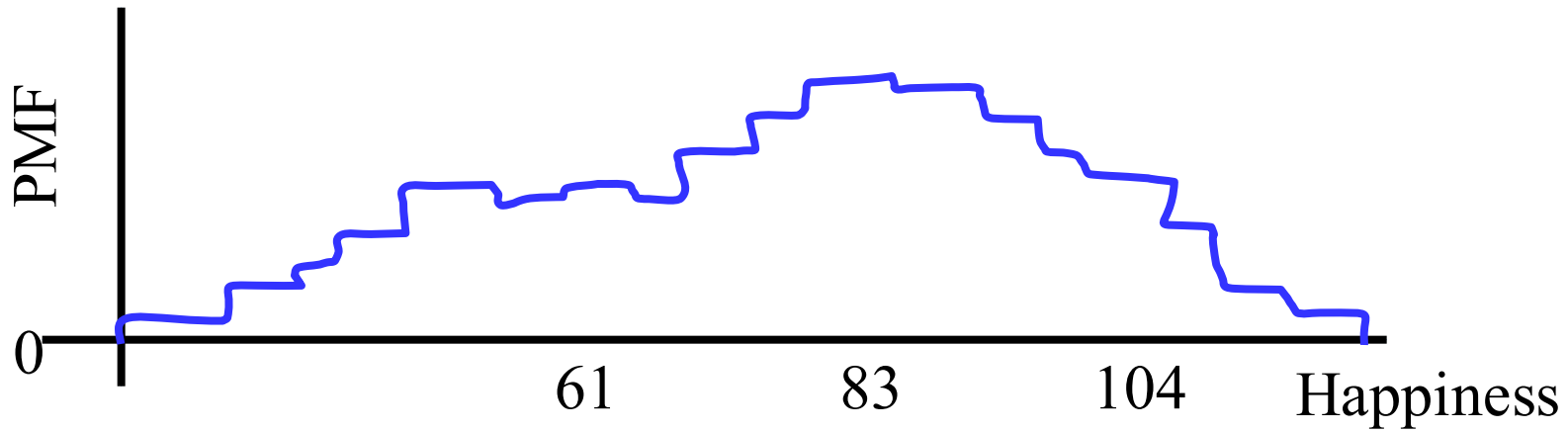
Bootstrap

Bootstrap Insight

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



Bootstrap of Means



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Create a resample with **sample.size()** new samples from PMF
   b. **Recalculate the mean** on the resample
3. You now have a **distribution of your means**

Null Hypothesis Test

Population 1	Population 2
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83
$\mu_1 = 3.1$	$\mu_2 = 2.4$

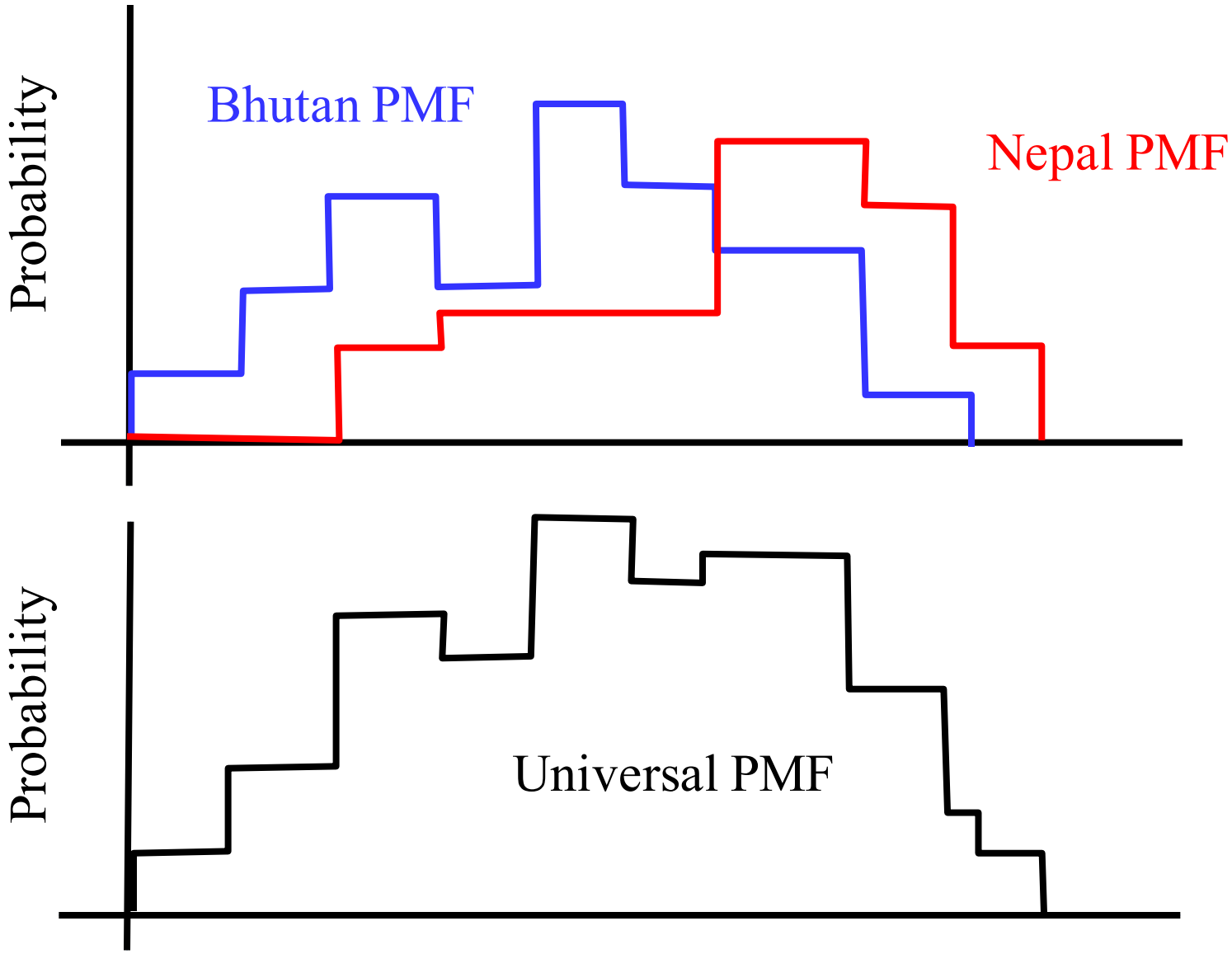
Null Hypothesis Test

Nepal Happiness	Bhutan Happiness
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83
$\mu_1 = 3.1$	$\mu_2 = 2.4$

Claim: The difference in happiness between Nepal and Bhutan is 0.7 happiness points.

Null hypothesis: even if there is no pattern (or two samples are identically distributed) your results might have arisen by chance

Universal Sample



Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal – muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal – muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal - muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal – muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal - muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

Bootstrap for P Values

```
def pvalueBootstrap(bhutanSample,nepalSample):
    N = size of the bhutanSample
    M = size of the nepalSample

    universalSample = combine bhutanSamples and nepalSamples
    universalPmf = estimate the pmf of universalSample

    count = 0

    repeat 10,000 times:
        bhutanResample = draw N resamples from the universalPmf
        nepalResample = draw M resamples from the universalPmf
        muBhutan = sample mean of the bhutanResample
        muNepal = sample mean of the nepalResample
        meanDiff = |muNepal - muBhutan|
        if meanDiff > observedDifference:
            count += 1

    pValue = count / 10,000
```

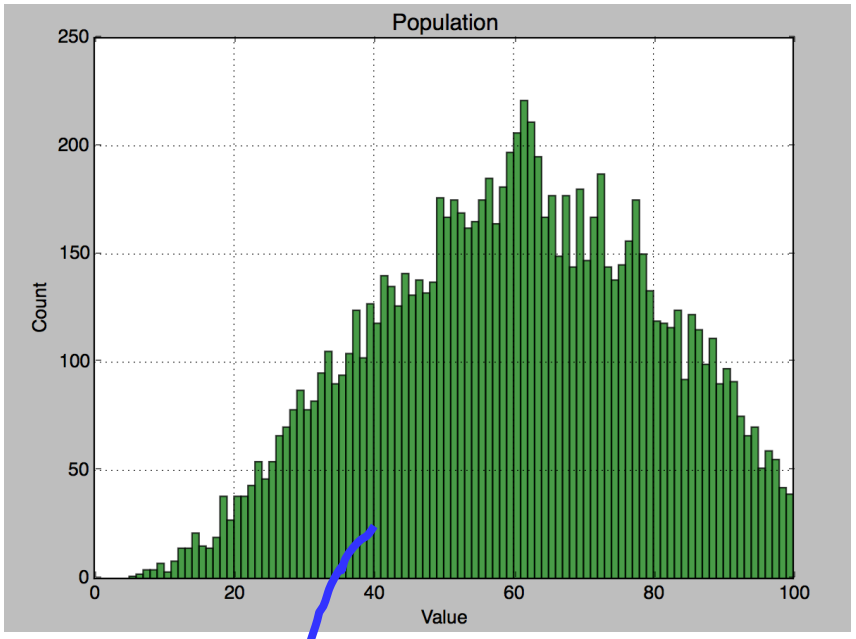
Bootstrap

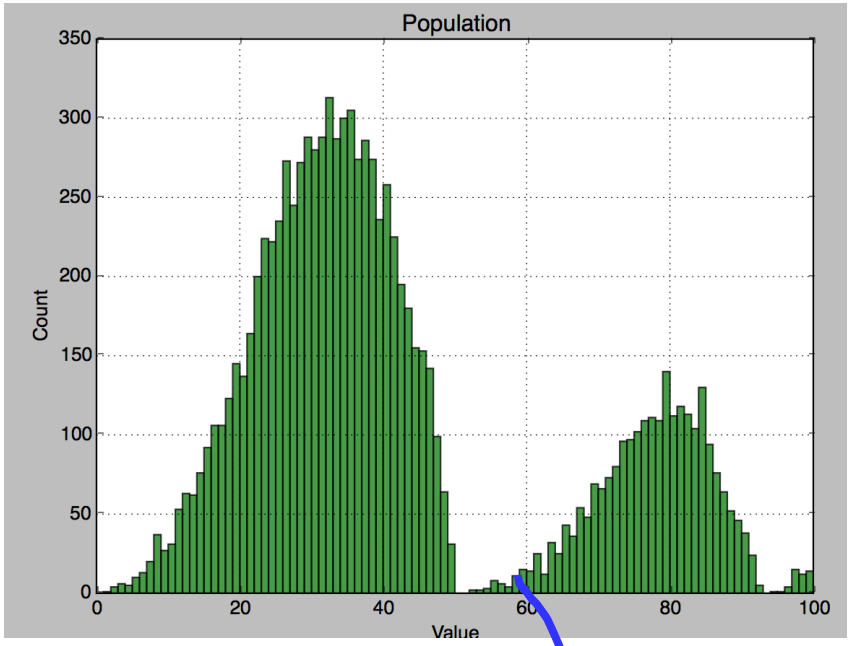
Lets try it!

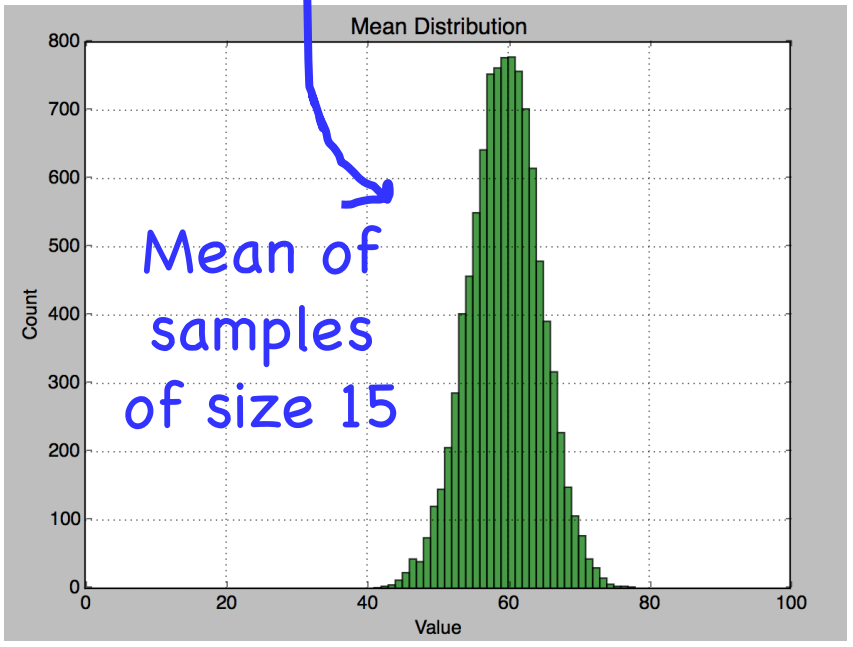
Null Hypothesis Test

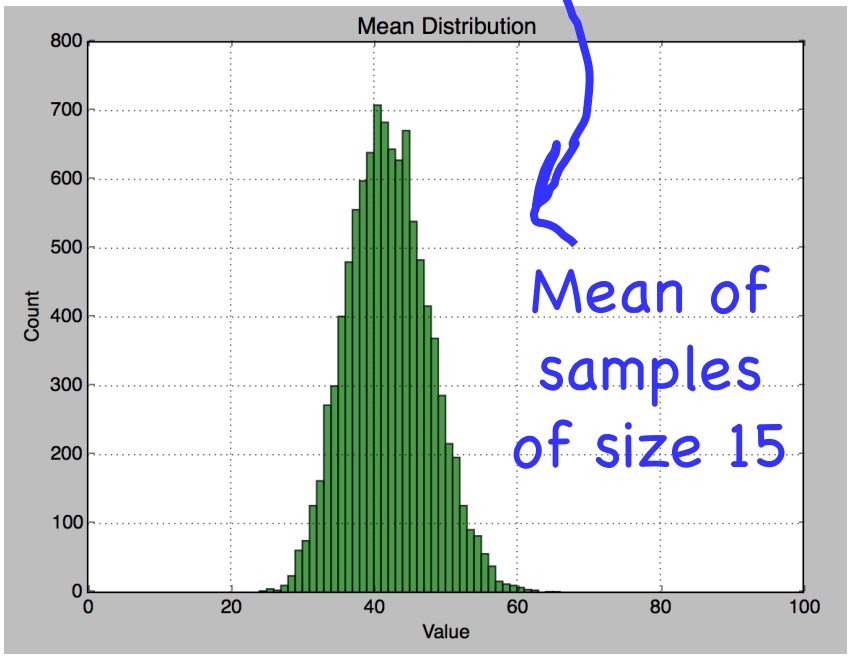
Nepal Happiness	Bhutan Happiness
4.44	2.15
3.36	3.01
5.87	2.02
2.31	1.43
...	...
3.70	1.83
$\mu_1 = 3.1$	$\mu_2 = 2.4$

Claim: The difference in happiness between Nepal and Bhutan is 0.7 happiness points (p = 0.04).


Population
Count
Value
0
50
100
150
200
250
0
20
40
60
80
100


Population
Count
Value
0
50
100
150
200
250
300
350
0
20
40
60
80
100


Mean Distribution
Count
Value
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
Mean of samples of size 15


Mean Distribution
Count
Value
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
Mean of samples of size 15

1. Inequalities

Inequality, Probability and Joviality

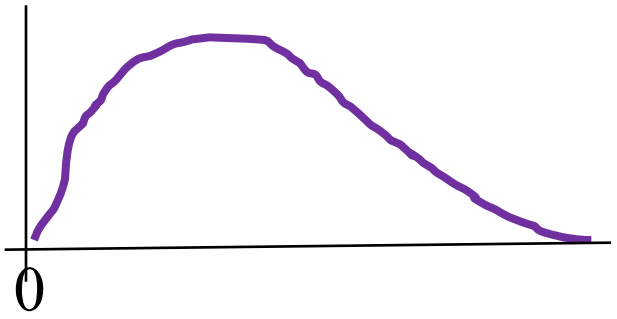
- If we know some statistics of a distribution,
  - E.g., mean, Variance, Non-negativity, Etc.
  - Inequalities and bounds allow us to make analytic claims about the probability distribution
  - May be imprecise compared to knowing true distribution!
  - But a useful tool for proving theorems.

Markov's Inequality

- Say X is a **non-negative** random variable

$$P(X \geq a) \leq \frac{E[X]}{a}, \quad \text{for all } a > 0$$

- Proof:
  - I = 1 if X ≥ *a*, 0 otherwise
  - Since $X \geq 0, \quad I \leq \frac{X}{a}$
  - Taking expectations:



$$E[I] = P(X \geq a) \leq E\left[\frac{X}{a}\right] = \frac{E[X]}{a}$$

Andrey Markov

- Andrey Andreyevich Markov (1856-1922) was a Russian mathematician

  - Markov's Inequality is named after him
  - He also invented Markov Chains…
    - …which are the basis for Google's PageRank algorithm
  - John Snow with a mustache?

Markov and the Midterm

- Statistics from a previous quarter's CS109 midterm
  - X = midterm score
  - Using sample mean $\underline{X}$ = 84.0 ≈ E[X]
  - What is P(X ≥ 100)?

$$P(X \geq 100) \leq \frac{E[X]}{100} = \frac{84}{100} = 0.84$$

  - Markov bound: ≤ 84% of class scored 100 or greater
  - In fact, 20.1% of class scored 100 or greater
    - Markov inequality can be a very loose bound
    - But, it made <u>no</u> assumption at all about form of distribution!

Chebyshev's Inequality

- X is a random variable with E[X] = μ, Var(X) = σ^2

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}, \quad \text{for all } k > 0$$

- Proof:
  - Since $(X - \mu)^2$ is non-negative random variable, apply Markov's Inequality with $a = k^2$

$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

  - Note that: $(X - \mu)^2 \geq k^2 \Leftrightarrow |X - \mu| \geq k$, yielding:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Pafnuty Chebyshev

- Pafnuty Lvovich Chebyshev (1821-1894) was also a Russian mathematician

  - Chebyshev's Inequality is named after him
    - But actually formulated by his colleague Irénée-Jules Bienaymé
  - He was Markov's doctoral advisor
    - And sometimes credited with first deriving Markov's Inequality
  - There is a crater on the moon named in his honor

2. Law of Large Numbers

Weak Law of Large Numbers

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have distribution F with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - For any $\varepsilon > 0$:

$$P(|\overline{X} - \mu| \geq \varepsilon) \xrightarrow{n\to\infty} 0$$

- Proof:

$$E[\overline{X}] = E\left[\frac{X_1+X_2+\ldots+X_n}{n}\right] = \mu \qquad \mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

  - By Chebyshev's inequality:

$$P(|X - \mu| \geq k) \leq \frac{\mathrm{Var}(X)}{k} \qquad P(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon} \xrightarrow{n\to\infty} 0$$

Strong Law of Large Numbers

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have distribution *F* with $E[X_i] = \mu$
  - Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$P\left(\lim_{n\to\infty}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \mu\right) = 1$$

  - Strong Law $\Rightarrow$ Weak Law, but not vice versa
  - Strong Law implies that for any $\varepsilon > 0$, there are only a finite number of values of *n* such that condition of Weak Law: $\left|\bar{X} - \mu\right| \geq \varepsilon$ holds.

Intuitions and Misconceptions of LLN

- Say we have repeated trials of an experiment
  - Let event E = some outcome of experiment
  - Let X_i = 1 if E occurs on trial *i*, 0 otherwise
  - Strong Law of Large Numbers (Strong LLN) yields:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \rightarrow E[X_i] = P(E)$$

  - Recall first week of class: $P(E) = \lim_{n \rightarrow \infty} \frac{n(E)}{n}$
  - Strong LLN justifies "frequency" notion of probability
  - Misconception arising from LLN:
    - Gambler's fallacy: "I'm due for a win"
    - Consider being "due for a win" with repeated coin flips...

La Loi des Grands Nombres

- History of the Law of Large Numbers
  - 1713: Weak LLN described by Jacob Bernoulli

  - 1835: Poisson calls it “La Loi des Grands Nombres”
    - That would be “Law of Large Numbers” in French

  - 1909: Émile Borel develops Strong LLN for Bernoulli random variables

  - 1928: Andrei Nikolaevich Kolmogorov proves Strong LLN in general case

And now a moment of silence...

...before we present...

...a beautiful result of probability theory!

3. Central Limit Theorem

The Central Limit Theorem

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have distribution *F* with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Central Limit Theorem:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \text{as} \quad n \to \infty$$

Demo

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

Once Upon a Time…

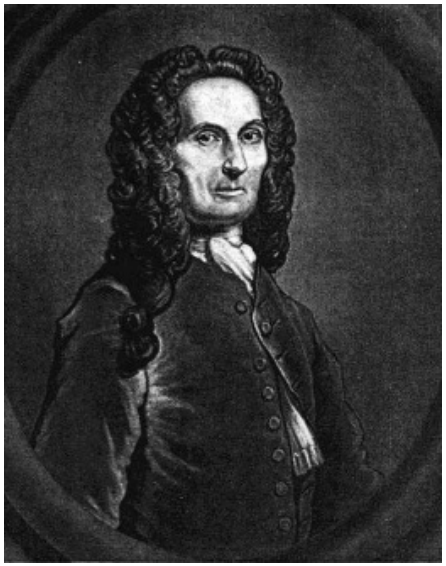
THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.

By *A. De Moivre.* F. R. S.

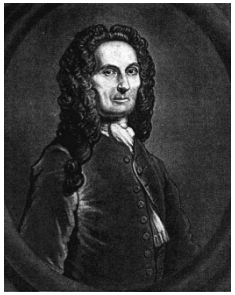
LONDON:
Printed by *W. Pearson,* for the Author. MDCCXVIII.

1733

Abraham De Moivre



Once Upon a Time…



- History of the Central Limit Theorem
  - 1733: CLT for X ~ Ber(1/2) postulated by Abraham de Moivre

  - 1823: Pierre-Simon Laplace extends de Moivre's work to approximating Bin(n, p) with Normal

  - 1901: Aleksandr Lyapunov provides precise definition and rigorous proof of CLT
  - 2017: Kevin Spacey stars in television series "House of Cards"

    - By end of the 4th season, there were 52 episodes
    - Mean quality of subsamples of episodes is Normally distributed (thanks to the Central Limit Theorem)

Central Limit Theorem in the Real World

- CLT is why many things in "real world" appear Normally distributed
  - Many quantities are sum of independent variables
  - Exams scores
    - Sum of individual problems on the SAT
    - Why does the CLT not apply to our midterm?
  - Election polling
    - Ask 100 people if they will vote for candidate X (p_1 = # "yes"/100)
    - Repeat this process with different groups to get $p_1, \ldots, p_n$
    - Will have a normal distribution over p_i
    - Can produce a "confidence interval"
      - How likely is it that estimate for true p is correct

Binomial Approximation

- Consider I.I.D. Bernoulli variables X_1, X_2, ... With probability *p*
  - X_i have distribution *F* with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Let: $Y = n\bar{X}$

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

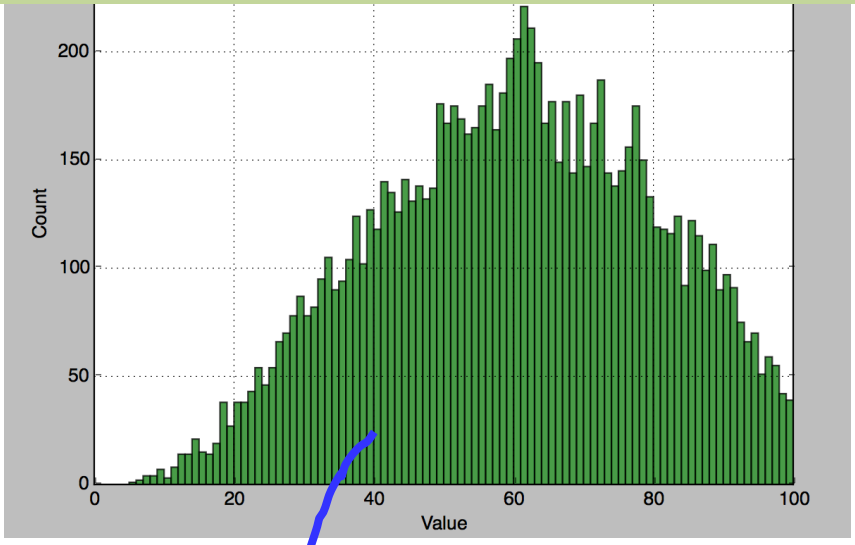
Central Limit Theorem

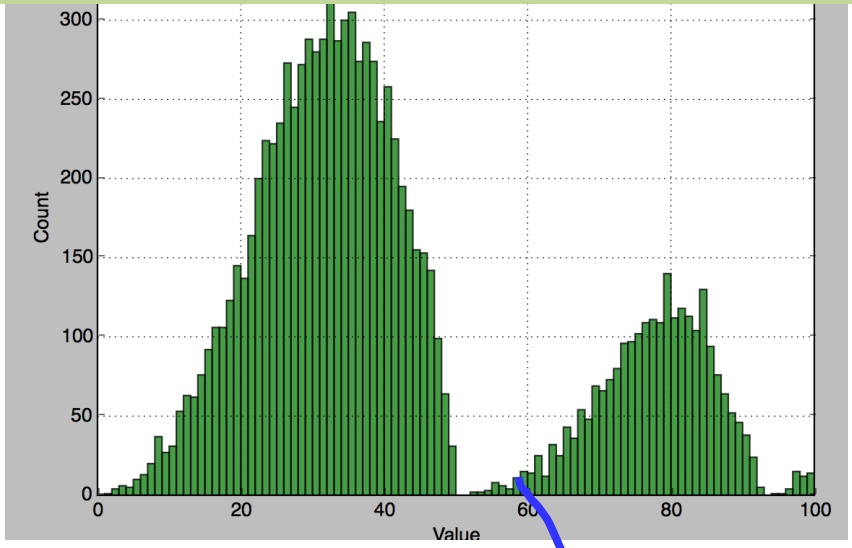
$$Y \sim N(n\mu, n^2 \frac{\sigma^2}{n})$$

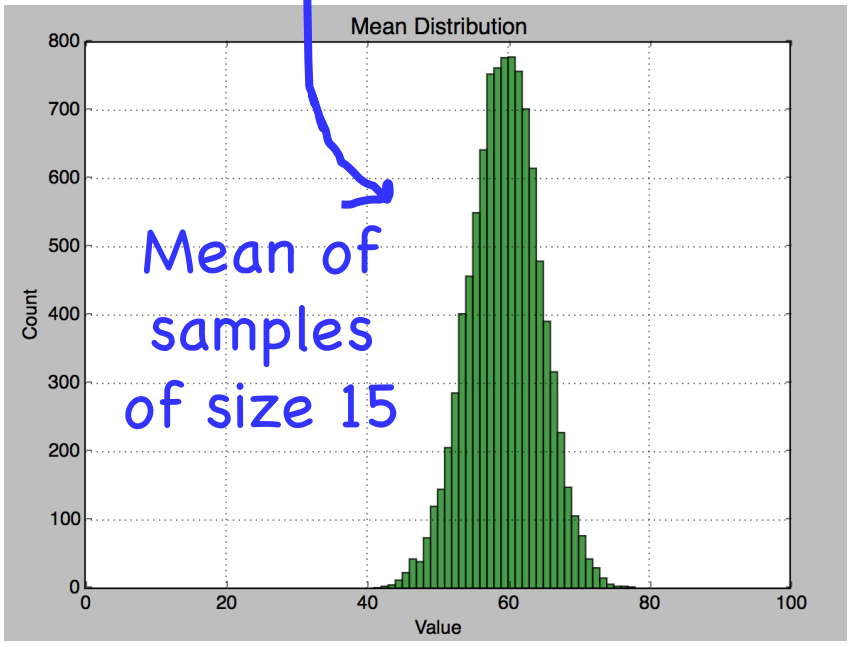
$$Y \sim N(np, np(1-p))$$

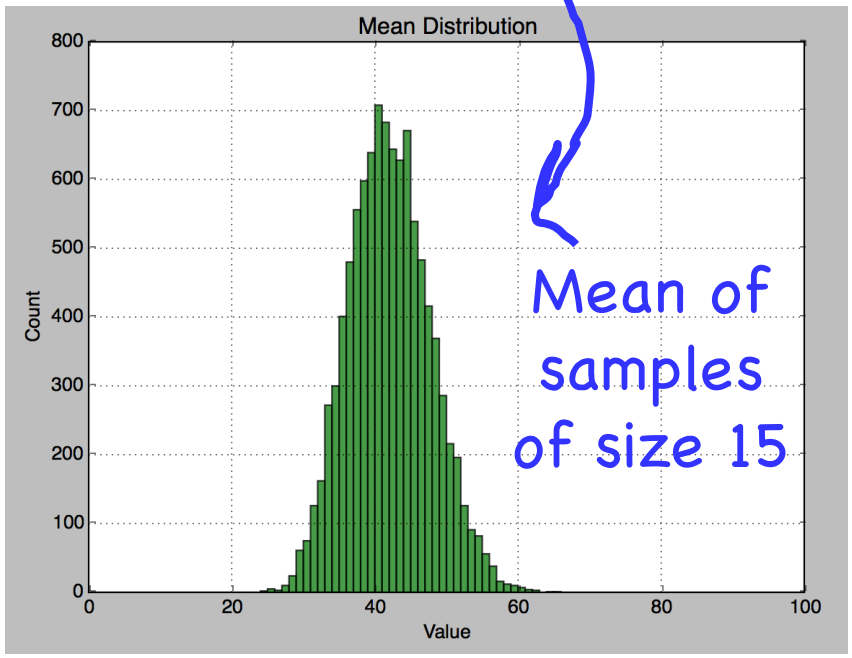
Substituting mean and variance of Bernoilli

C.L.T. Explains This

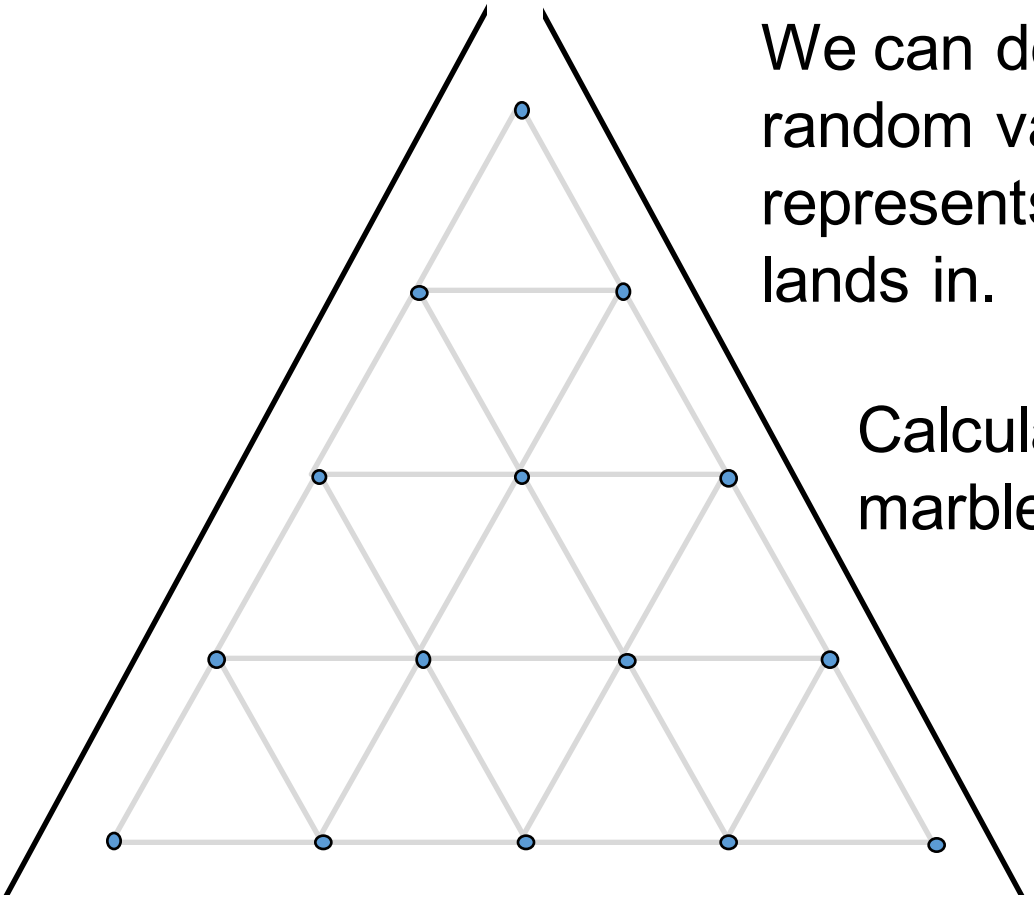






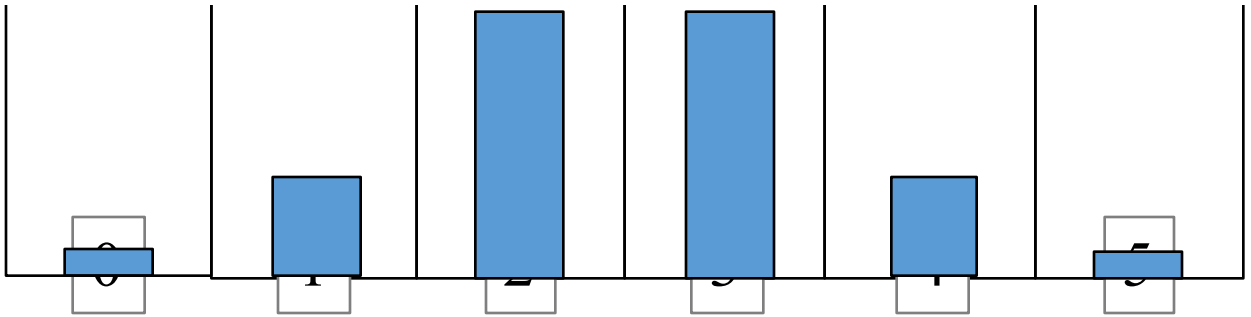


C.L.T. Explains This



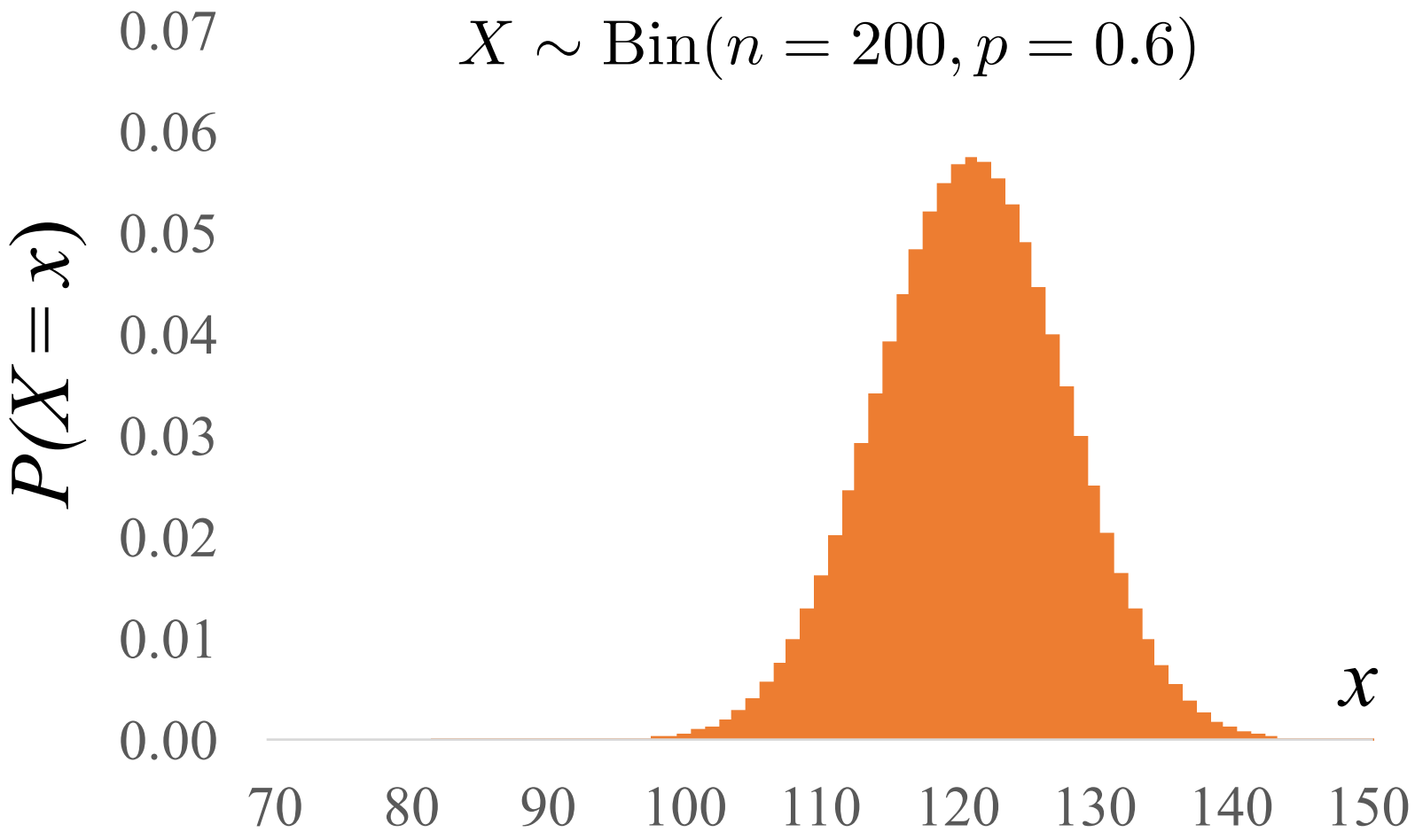
We can define an indicator random variable (***B***) which represents what bucket a marble lands in.

Calculate the probability of a marble landing in a bucket.



PDF

C.L.T. Explains This



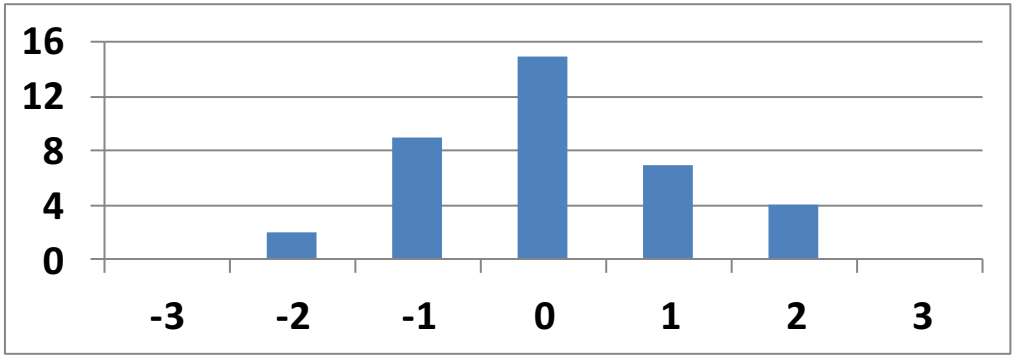
Your Midterm on the Central Limit Theorem

- Start with 236 midterm scores: $X_1, X_2, \ldots, X_{236}$
  - E[X_i] = 84 and Var(X_i) = 370
  - Created 50 samples ($Y_1, Y_2, \ldots, Y_{50}$) of size n = 10
  - Prediction by CLT: $\bar{Y}_i \sim N(84, 37)$

$$Z_i = \frac{\bar{Y}_i - 84}{\sqrt{37}}$$

$$\bar{Z} = \frac{1}{50}\sum_{i=1}^{50} Z_i = 4 \times 10^{-16}$$

$$\mathrm{Var}(\bar{Z}) = 0.997$$



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 $1 \leq i \leq n$)

$$0.95 = P(-0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq 0.5)$$

    - By Central Limit Theorem, $Z \sim N(0, 1)$, where:

$$Z_n = \frac{(\sum_{i=1}^{n} X_i) - n\mu}{\sigma\sqrt{n}}$$

$$= \frac{(\sum_{i=1}^{n} X_i) - nt}{2\sqrt{n}}$$

$$0.95 = P(-0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq 0.5) \qquad Z_n = \frac{(\sum_{i=1}^{n} X_i) - nt}{2\sqrt{n}}$$

$$0.95 = P(-0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq 0.5)$$

$$= P(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sqrt{n}}{2}\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sqrt{n}}{2}t \leq \frac{0.5\sqrt{n}}{2})$$

$$= P(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sum_{i=1}^{n} X_i}{2\sqrt{n}} - \frac{\sqrt{n}}{\sqrt{n}}\frac{\sqrt{n}t}{2} \leq \frac{0.5\sqrt{n}}{2})$$

$$= P(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sum_{i=1}^{n} X_i - nt}{2\sqrt{n}} \leq \frac{0.5\sqrt{n}}{2})$$

$$= P(\frac{-0.5\sqrt{n}}{2} \leq Z \leq \frac{0.5\sqrt{n}}{2})$$

$$0.95 = P(\frac{-0.5\sqrt{n}}{2} \leq Z \leq \frac{0.5\sqrt{n}}{2})$$

$$\begin{aligned}
0.95 &= \phi(\frac{\sqrt{n}}{4}) - \phi(\frac{\sqrt{n}}{4}) \\
&= \phi(\frac{\sqrt{n}}{4}) - (1 - \phi(\frac{\sqrt{n}}{4})) \\
&= 2\phi(\frac{\sqrt{n}}{4}) - 1 \\
0.975 &= \phi(\frac{\sqrt{n}}{4}) \\
\phi^{-1}(0.975) &= \frac{\sqrt{n}}{4} \\
1.96 &= \frac{\sqrt{n}}{4} \\
n &= 61.4
\end{aligned}$$

William Sealy Gosset
(aka Student)

Estimating Time With Chebyshev

- 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 so estimated time = $t \pm 0.5$ with 95% certainty?
    - X_i = running time of i-th run (for $1 \le i \le n$), and $X_S = \sum_{i=1}^{n} \frac{X_i}{n}$
  - What would Chebyshev say? $P(|X_S - \mu_S| \ge k) \le \frac{{\sigma_S}^2}{k^2}$

$$\mu_S = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = t \qquad {\sigma_S}^2 = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \mathrm{Var}\left(\frac{X_i}{n}\right) = n\frac{\sigma^2}{n^2} = \frac{4}{n}$$

$$P\left(\left|\sum_{i=1}^{n} \frac{X_i}{n} - t\right| \ge 0.5\right) \le \frac{4/n}{(0.5)^2} = \frac{16}{n} = 0.05 \quad \Rightarrow \quad n \ge 320$$

Thanks for playing, Pafnuty!

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: $X \le 25$ or $X \ge 45$
- Recall CLT: $\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \to N(0,1)$ as $n \to \infty$
  - Determine P(X ≤ 25 or X ≥ 45) using CLT:

$$\mu = E[X_i] = 3.5 \qquad \sigma^2 = \mathrm{Var}(X_i) = \frac{35}{12}$$

$$1 - P(25.5 \le X \le 44.5) = 1 - P\left(\frac{25.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{X - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{44.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}}\right)$$

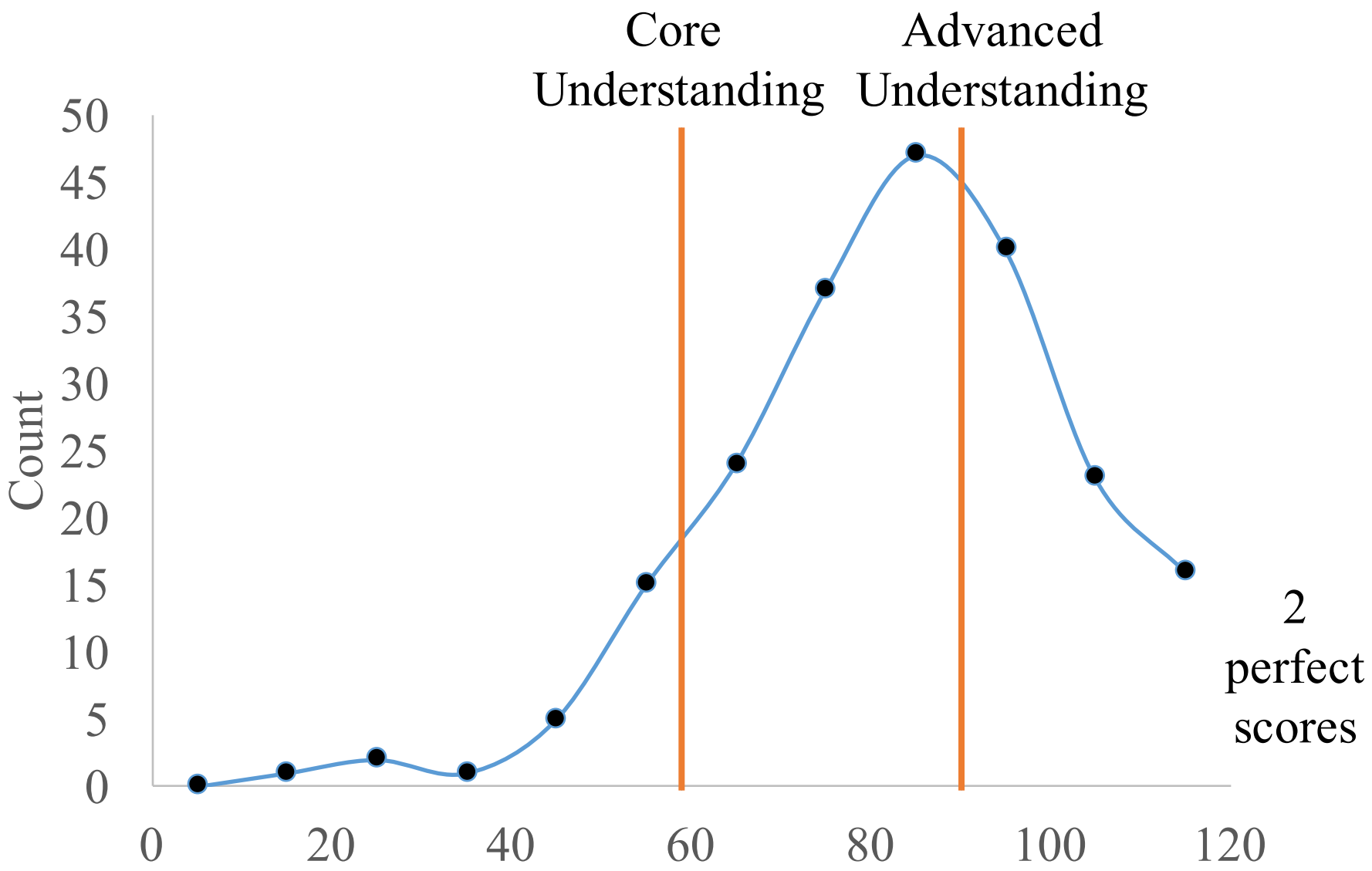
$$\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784$$

Wonderful Form of Cosmic Order

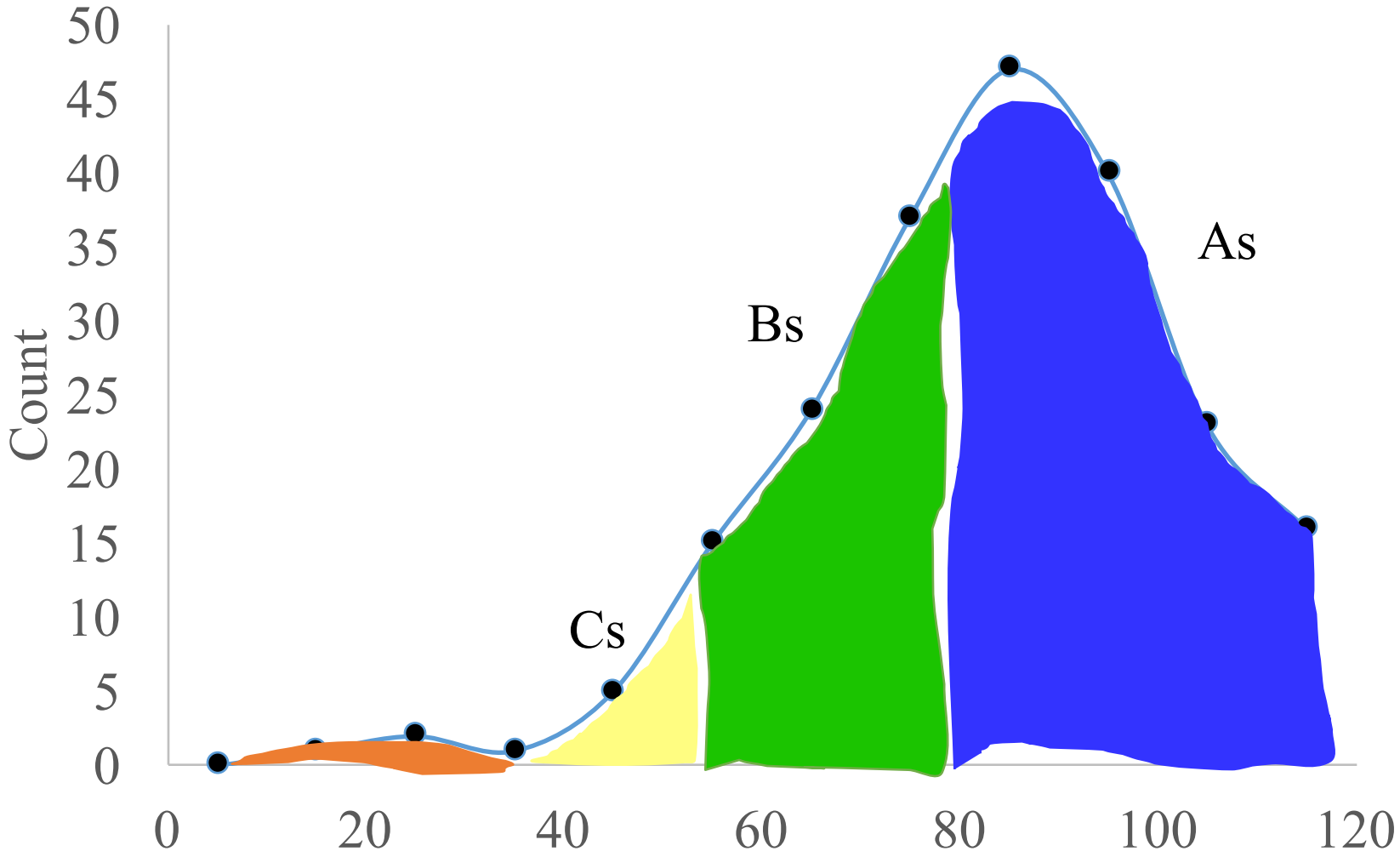
I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ”[Central limit theorem]". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest 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 and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

-Sir Francis Galton

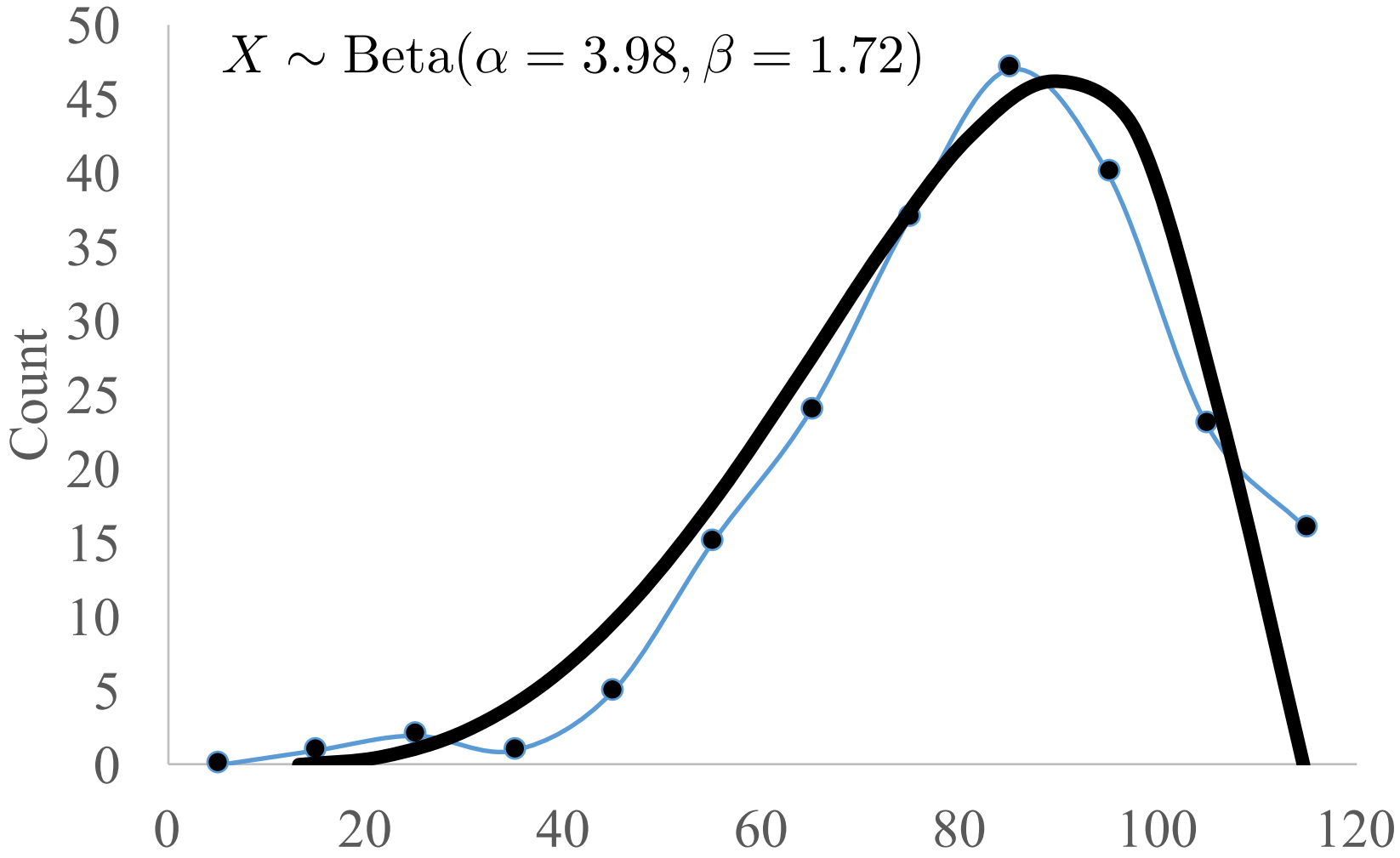
Midterm Distribution



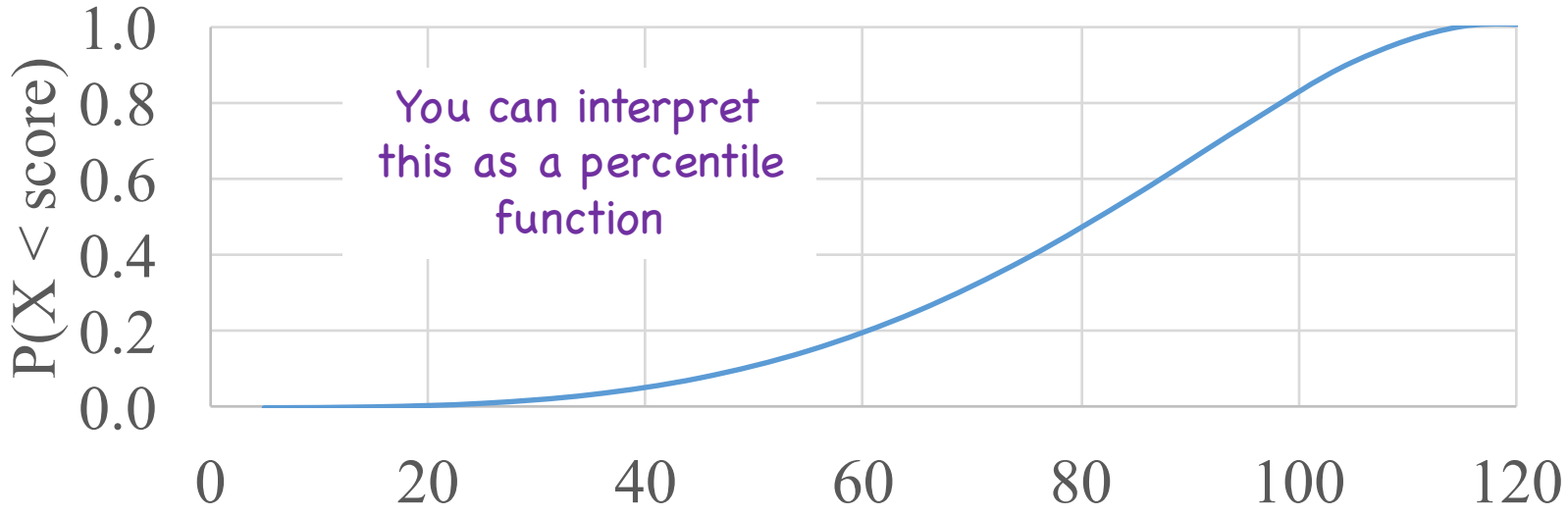
Midterm Distribution

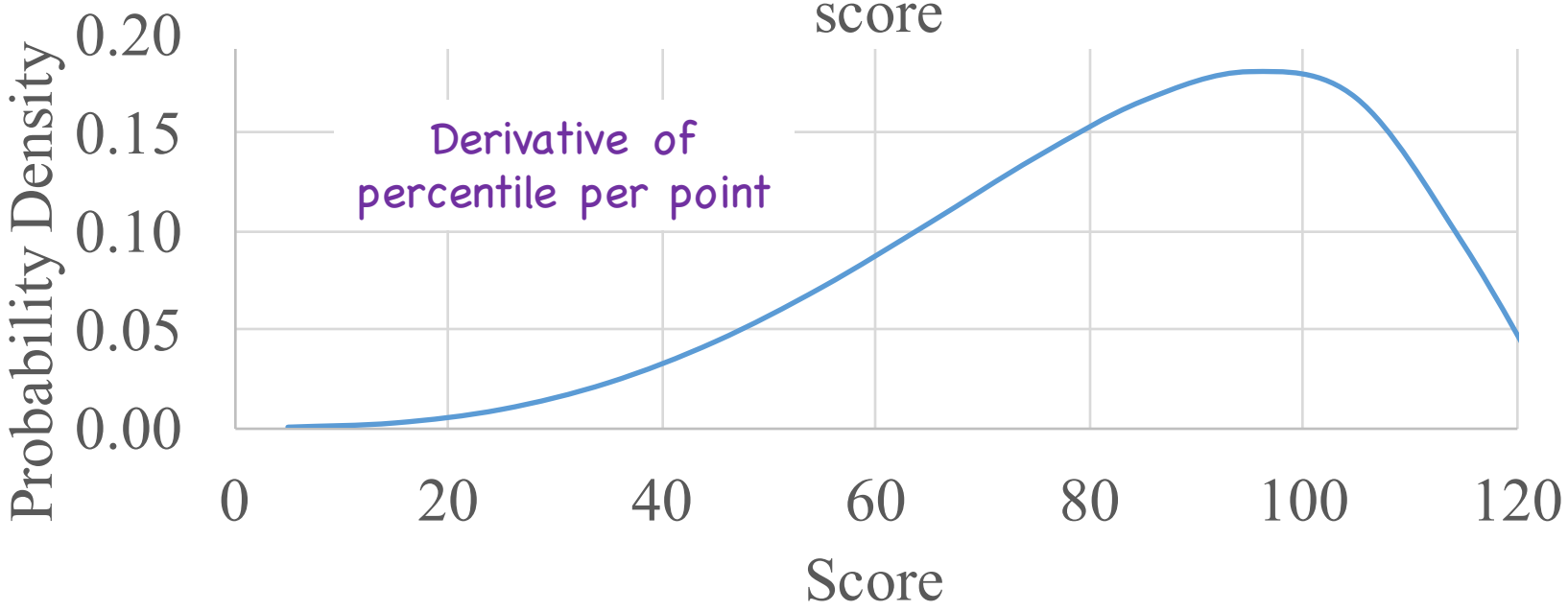


Midterm Distribution

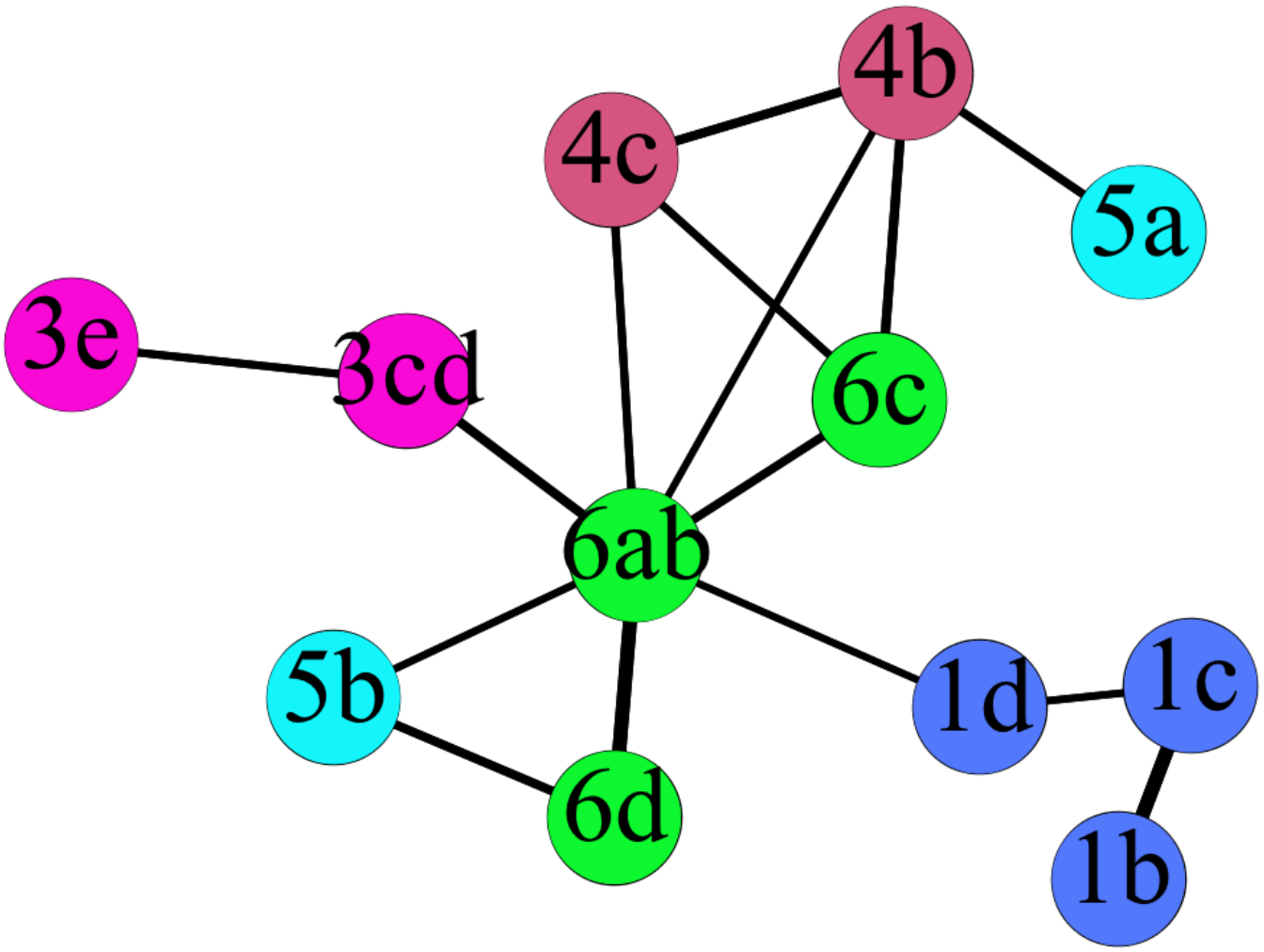


CS109 Midterm CDF





Midterm Question Correlations



Correlation

0.60

0.30

*Correlations bellow 0.34 are not shown

**Almost all correlations were positive

4.b and 6.c?