CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Lab 7 + project proposal** due Wednesday
  - During lab this week: project meetings with all groups
  - Try to come to the same lab section as your partner
  - We can also discuss Lab 7

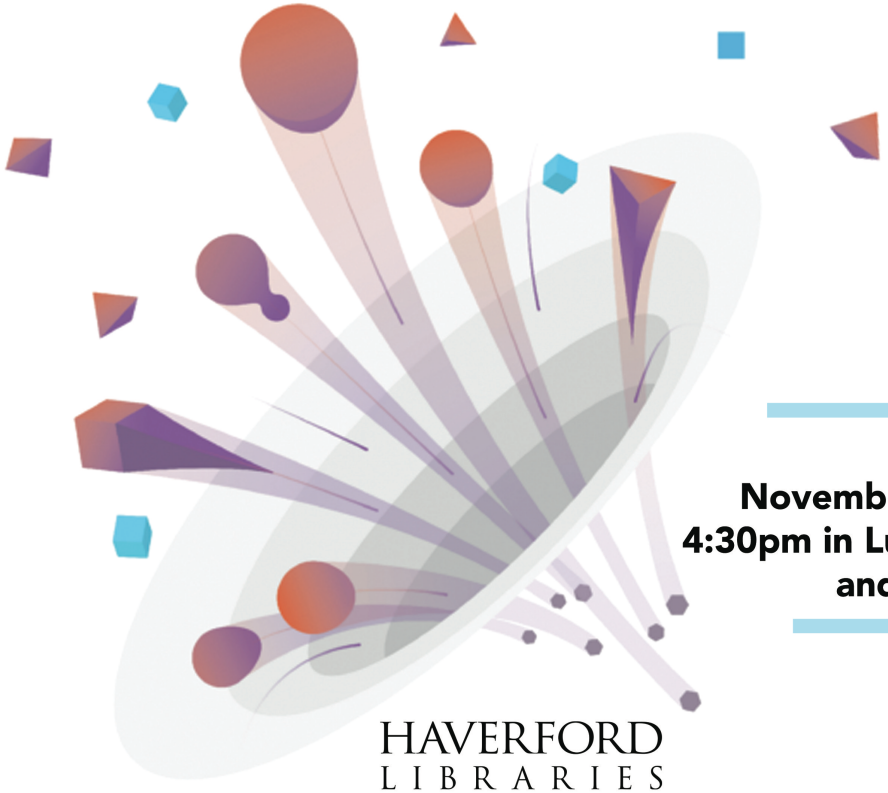
Languages Left Behind:

Why Language Models Fail at Non-English Content Moderation

Aliya Bhatia
Policy Analyst
Free Expression Project

Gabriel Nicholas
Research Fellow

Center for Democracy & Technology



Thursday
November 9, 2023
4:30pm in Lutnick 200
and on Zoom

HAVERFORD
LIBRARIES

Outline for November 7

- Recap PCA and Handout 16
- Begin: statistics and hypothesis testing
- Central limit theorem
- Bootstrapping

Outline for November 7

- Recap PCA and Handout 16
- Begin: statistics and hypothesis testing
- Central limit theorem
- Bootstrapping

Handout 16

$$A = \begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix}$$

$$\vec{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \vec{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

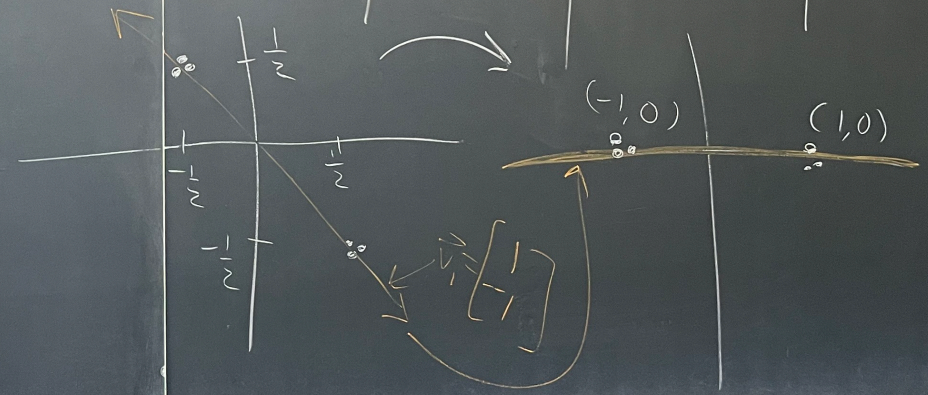
$$\lambda_1 = \frac{3}{5} \qquad \lambda_2 = 0$$

$$\underset{n\times 2}{T} = \underset{n\times p}{X}\ \underset{p\times 2}{W}$$

$$= \begin{array}{cc} f_1 & f_2 \\ \end{array}$$

$$\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{array}{cc} \vec{v}_1 & \vec{v}_2 \\ \end{array} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & ? \end{bmatrix}$$

$n = 6$



Outline for November 7

- Recap PCA and Handout 16
- Begin: statistics and hypothesis testing
- Central limit theorem
- Bootstrapping

Motivation for studying statistics

1) I have a new method that achieves 95% accuracy on a dataset. The previous best method achieved 92% accuracy. Is my method significantly better?

2) I have created a new treatment for high blood pressure. Did it significantly lower the blood pressure of the treatment group over the control group?

3) Which variants in the genome are statistically correlated with a specific disease?

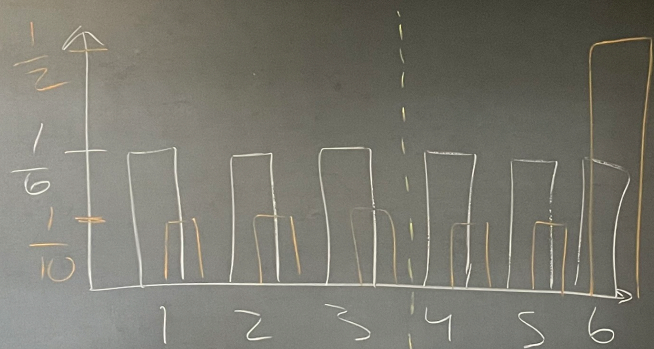
Motivation for studying statistics

- In general there are many questions that can only be answered properly with statistics.
- This one week on statistics is not a substitute for a full stats course, which I recommend everyone take!
- We're going to do a few key examples now to build up intuition, but this is a huge field and not my main area of research.

Distributions, expected value, variance

fair die

weighted die



probability mass functions (pmf)

$$\sum_{x=1}^{6} p(x) = 1 \quad , \quad \sum_{x \in vals(X)} p(x) = 1$$

random variable

"weighted average"

Expected Value

mean

empirical based on data

Theo

$$E[X] = \sum_{x \in vals(X)} x\, p(x)$$

fair $X_f \Rightarrow E[X_f] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots 6 \frac{1}{6}$

$= \boxed{3.5}$

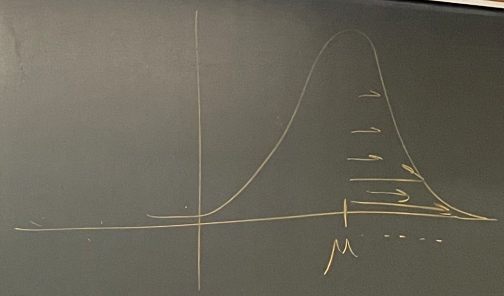
weighted $X_w \Rightarrow E[X_w] = \frac{1}{10}(1+2+\cdots 5) + \frac{1}{2} \cdot 6$

$= \boxed{4.5}$

Variance

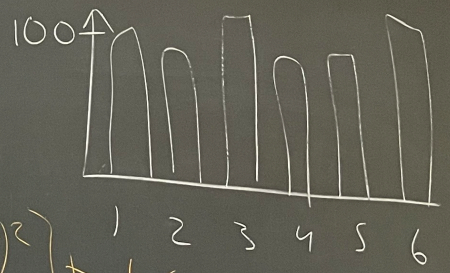
$\mu = E[X]$

$$E[\underbrace{(X-\mu)^2}_{\text{spread}}] = \sum_{x \in vals(X)} (x-\mu)^2 \underbrace{p(x)}$$



fair $\Rightarrow Var(X_f) = \frac{1}{6}\left[(1-3.5)^2 + \cdots + (6-3.5)^2\right]$

$\boxed{\approx 2.92}$



weighted $\Rightarrow Var(X_w) = \frac{1}{10}\left[(1-4.5)^2 \cdots (5-4.5)^2\right] + \frac{1}{2}(6-4.5)^2 = \boxed{3.25}$

(empirical):

Sample variance $\Rightarrow Var(X) = \underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}_{\text{weighting all examples equally.}}$

Outline for November 7

- Recap PCA and Handout 16
- Begin: statistics and hypothesis testing
- Central limit theorem
- Bootstrapping

Central Limit Theorem (CLT)

3 6 1

$X_1, X_2 \ldots X_n$ are samples from a population with mean μ & finite variance σ^2, and $\bar{X}_n$ is the sample mean.

theory

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

normalizing $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

$$\boxed{\text{THEN}} \quad Z = \lim_{n\to\infty} \left(\frac{\bar{X}_n - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right)$$

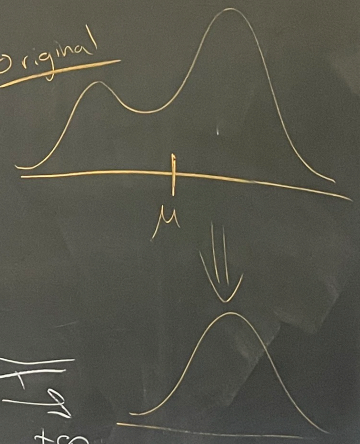
3.

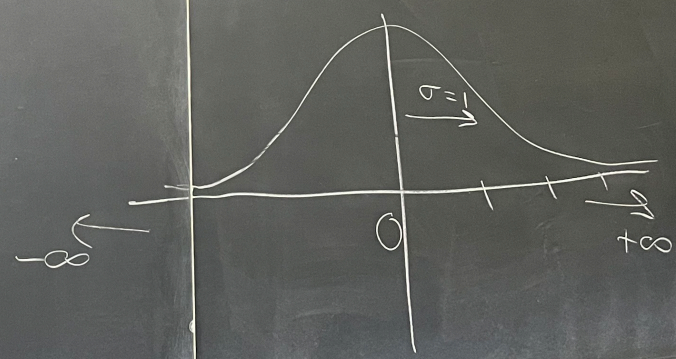
is a standard normal distribution

$N(0, 1)$

mean variance

original





probability density function (pdf)

$$\text{pdf} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \boxed{1}$$

Hypothesis Testing

- H_0 = null hypothesis (ex: die is fair)
- H_1 = alternate hypothesis (ex: die is weighted toward higher values)

apply CLT

$$z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

z-score

$$= \frac{4.2 - 3.5}{\sqrt{\frac{2.92}{20}}}$$

$$= \boxed{1.83}$$

← test statistic

$n = 20$

$\bar{X}_n = 4.2$ (sam m

one observat

variance

$\boxed{\sigma^2 = 2.92}$

stddev

$\boxed{\sigma = \sqrt{2.92}}$

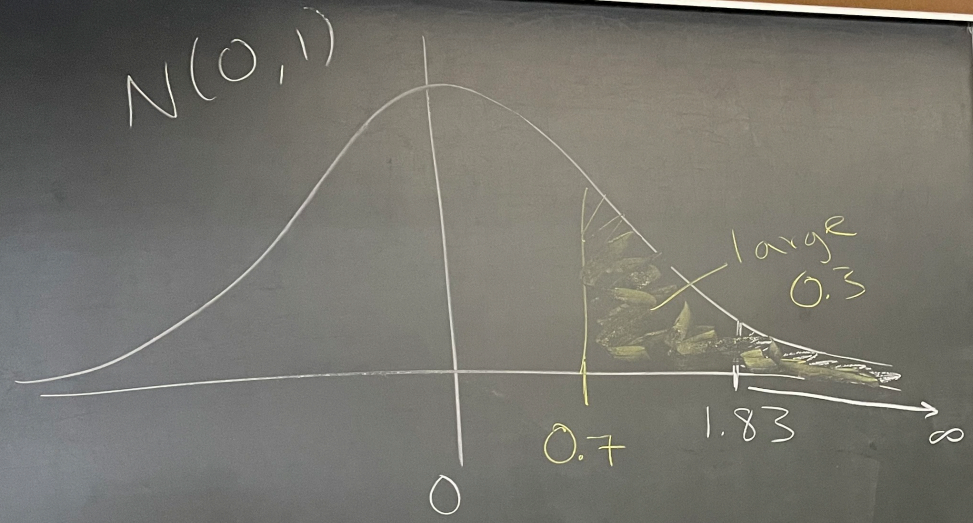
0.7 ← $n = 10$, $\bar{X}_n = 4.2$

mple
mean)
vation

ance

2

2



$$\text{p-value} = \int_{1.83}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

$$\approx \boxed{0.033}$$

usually compare with $\underbrace{0.05}_{\text{Significance level}} = \alpha$

$\Rightarrow$ $\boxed{\text{reject null hypothesis}}$

<u>p-value</u>: probability of observing a result <u>as</u> <u>or</u> <u>more</u> extreme than yours <u>under</u> <u>the</u> <u>null</u> <u>hypothesis</u>

Handout 17

Handout 17

① $E[X] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$

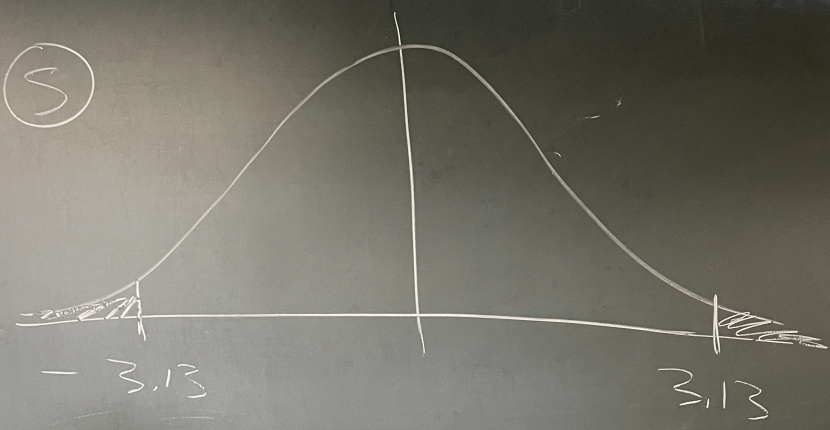
② $Var(X) = \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 = \frac{1}{4}$

③ $\overline{X}_n = \frac{54}{80} = 0.675$

④ $Z = \dfrac{0.675 - 0.5}{\sqrt{\dfrac{0.25}{80}}}$

$\boxed{\approx 3.13}$ z-score test stat

S



p-value = 0.00175 ≤ 0.05

$\sqrt{\frac{\sigma^2}{n}}$

Outline for November 7

- Recap PCA and Handout 16
- Begin: statistics and hypothesis testing
- Central limit theorem
- Bootstrapping

Next time!