CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Lab 7** due tonight
- **No lab** tomorrow
- **Final Project proposal** due Friday (Nov 8)
- **Nov 27 class:** work on final project

Outline for today

- Recap PCA

- Begin: statistics and hypothesis testing

- Central limit theorem

Outline for today

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

Principal Component Analysis (PCA)

- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

- PCA is a linear transformation

- Typically, we look at the first few dimensions of the transformed data as a means of dimensionality reduction and visualization

- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

Outline for today

- Recap PCA

- Begin: statistics and hypothesis testing

- Central limit theorem

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?

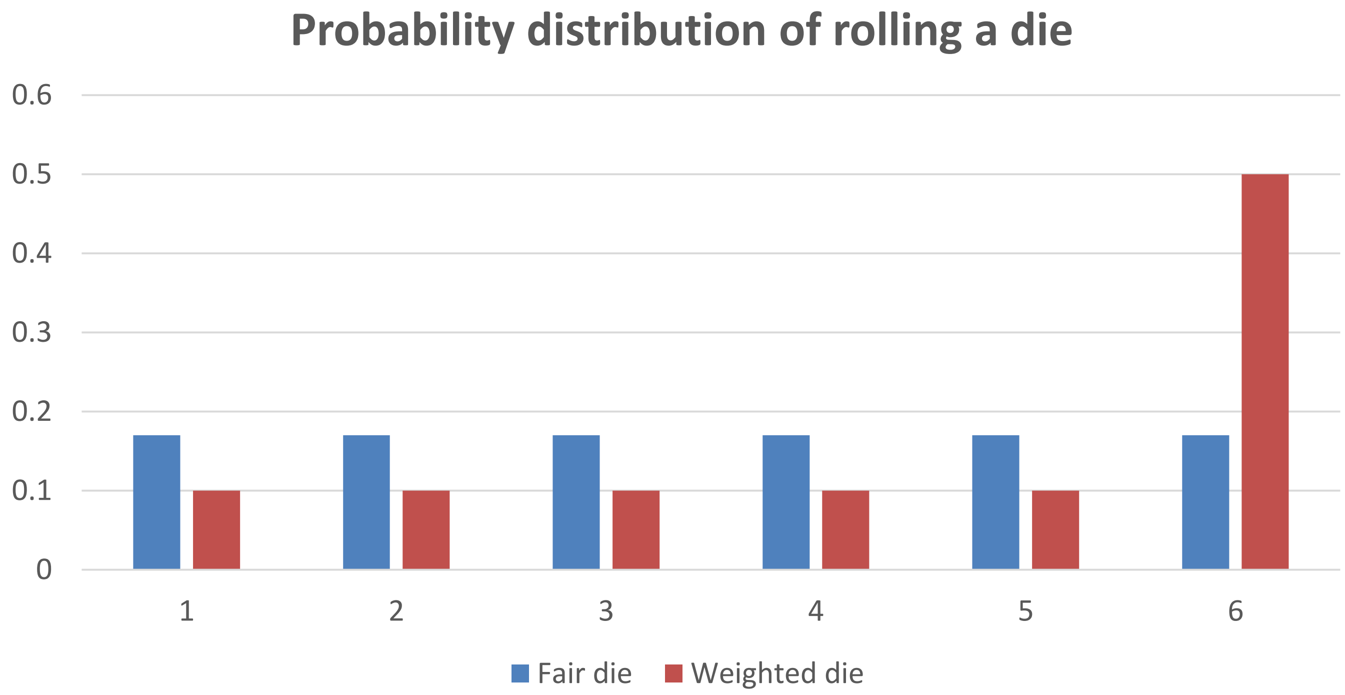
2) I have created a new treatment for high blood pressure. Did it significantly lower the blood pressure of the treatment group over the control group?

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



Probability mass function (pmf): $p(x)$

$$\sum_{x \in vals(X)} p(x) = 1$$

Expected Value

Weighted average:

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

$$E[X_f] = 1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6} = 3.5$$

$$E[X_w] = (1 + 2 + \cdots + 5) * \frac{1}{10} + 6 * \frac{1}{2} = 4.5$$

Sample (empirical) mean:

$$\overline{X_n} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i$$

Variance

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

$$Var(X_f) = \frac{1}{6}[(1 - 3.5)^2 + \cdots + (6 - 3.5)^2] \approx 2.92$$

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

Sample (empirical) variance:

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Outline for today

- Recap PCA and Handout 16

- Begin: statistics and hypothesis testing

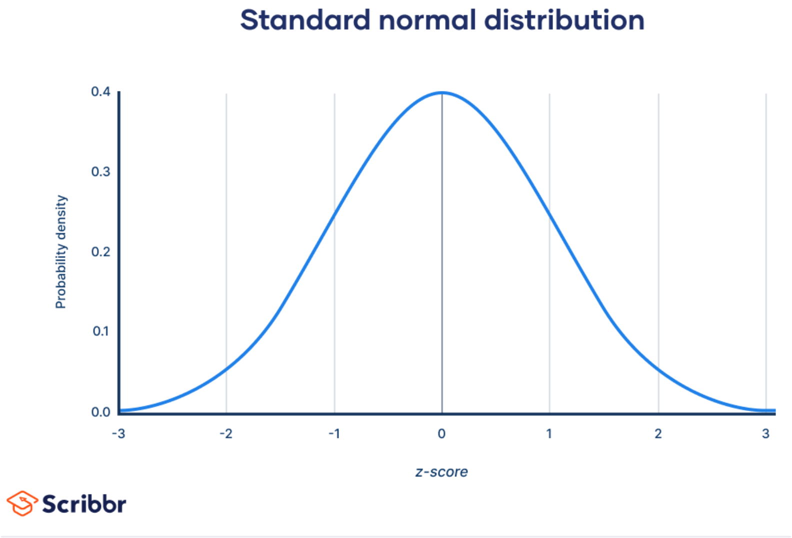
- Central limit theorem

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are samples from a population with expected value μ and finite variance σ^2, and $\overline{X_n}$ is the sample mean, then

$$Z = \lim_{n \to \infty} \left(\frac{\overline{X_n} - \mu}{\sigma / \sqrt{n}} \right)$$

is a standard normal distribution $N(\overset{\text{mean}}{0},\overset{\text{variance}}{1})$.



Hypothesis testing

- H_0: null hypothesis (e.g. die is fair)
- H_1: alternative hypothesis (e.g. die is weighted towards higher values)

- Apply CLT:

$$\text{Z-score} = \frac{\overline{X_n}-\mu}{\sqrt{\sigma^2/n}}$$

n=20

$\overline{X_n} = 4.2$

test statistic

$$\approx \frac{4.2 - 3.5}{\sqrt{2.92/20}} \approx 1.83$$

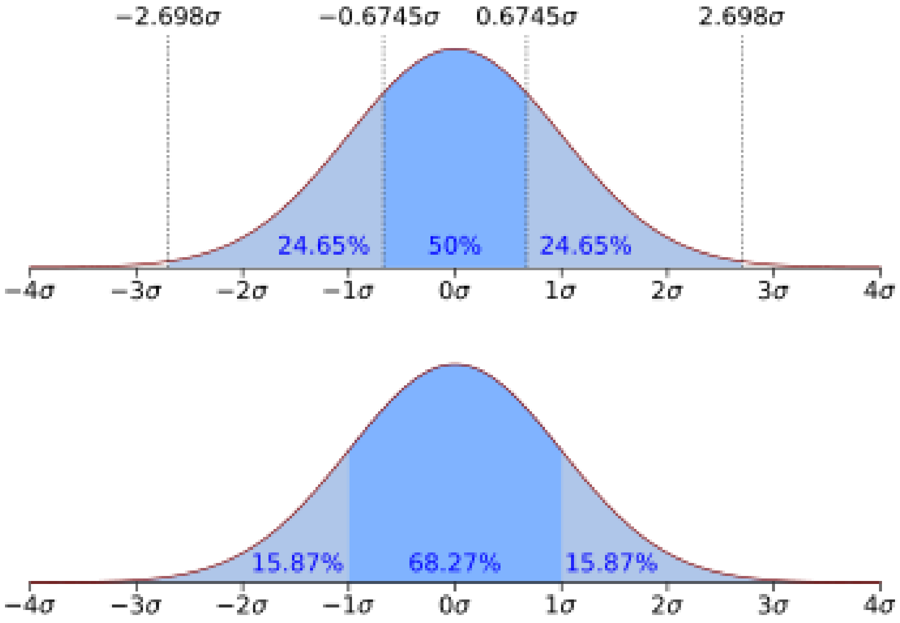
p-value

- Probability of observing a result as or more extreme than ours under the null hypothesis

- Probability density function

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

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



Wikipedia

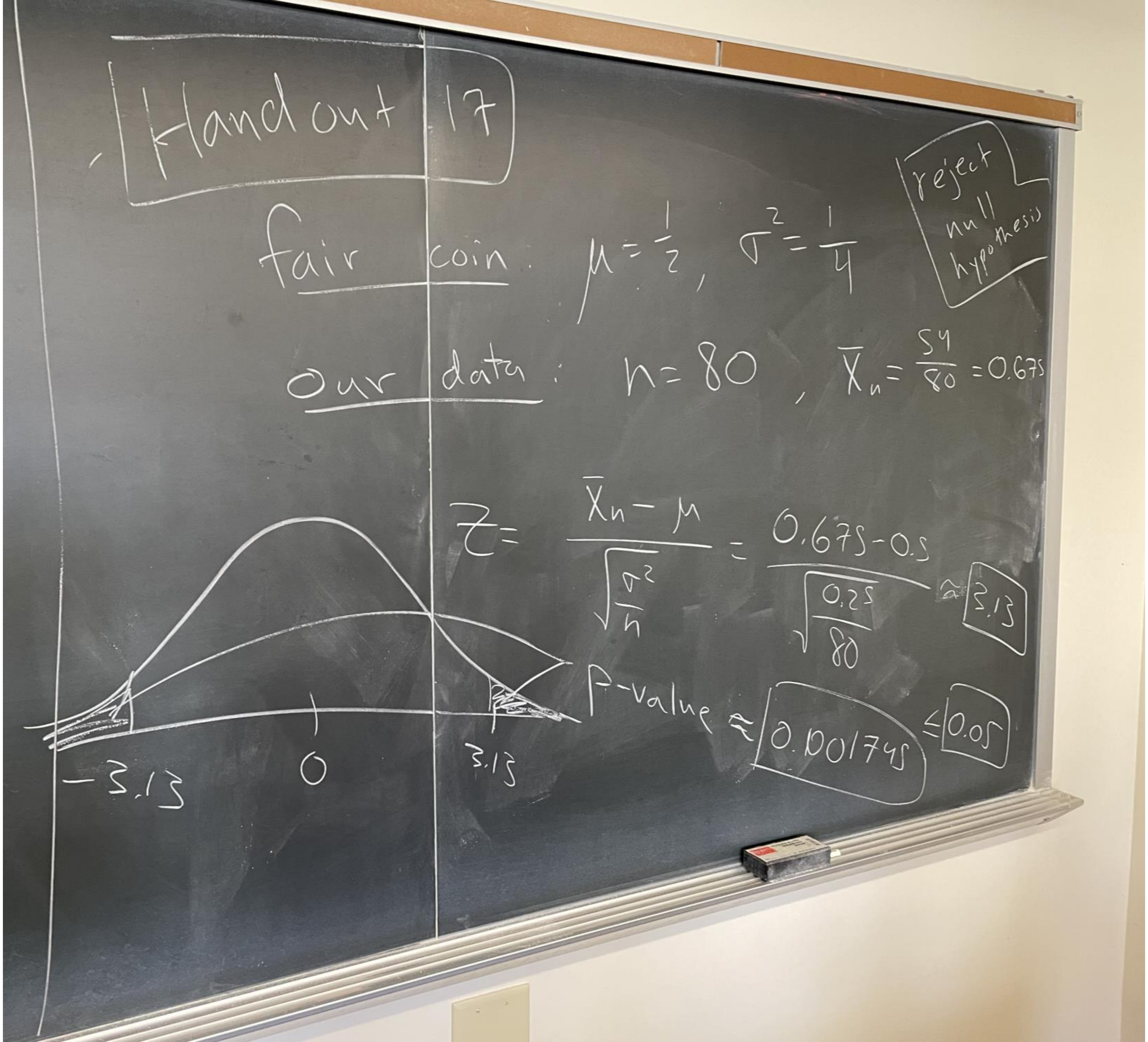
Hypothesis testing

- p-value $= \int_{1.83}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \approx 0.033$

- Usually compare with $\alpha = 0.05$ (significance level)

- $0.033 \leq 0.05 \Rightarrow$ reject the null hypothesis

Handout 17



Handout 17

fair coin: $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{4}$

reject null hypothesis

our data: $n = 80$, $\bar{X}_n = \frac{54}{80} = 0.675$

$$Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.675 - 0.5}{\sqrt{\frac{0.25}{80}}} \approx \boxed{3.13}$$

$$\text{p-value} \approx \boxed{0.001745} \leq \boxed{0.05}$$