CS461 – RECITATION 01

MACHINE LEARNING PRINCIPLES

Daize Dong

2025-09-15

CONTACT

- Teaching Assistant
- Email: *daize.dong@rutgers.edu*
- Office (Hours):
  - Not Yet

BACKGROUND

- First-year PhD
- Deep Learning
- Neural Network Efficiency
- High Performance Computation

TODAY'S CONTENT

Recap what you have learned before

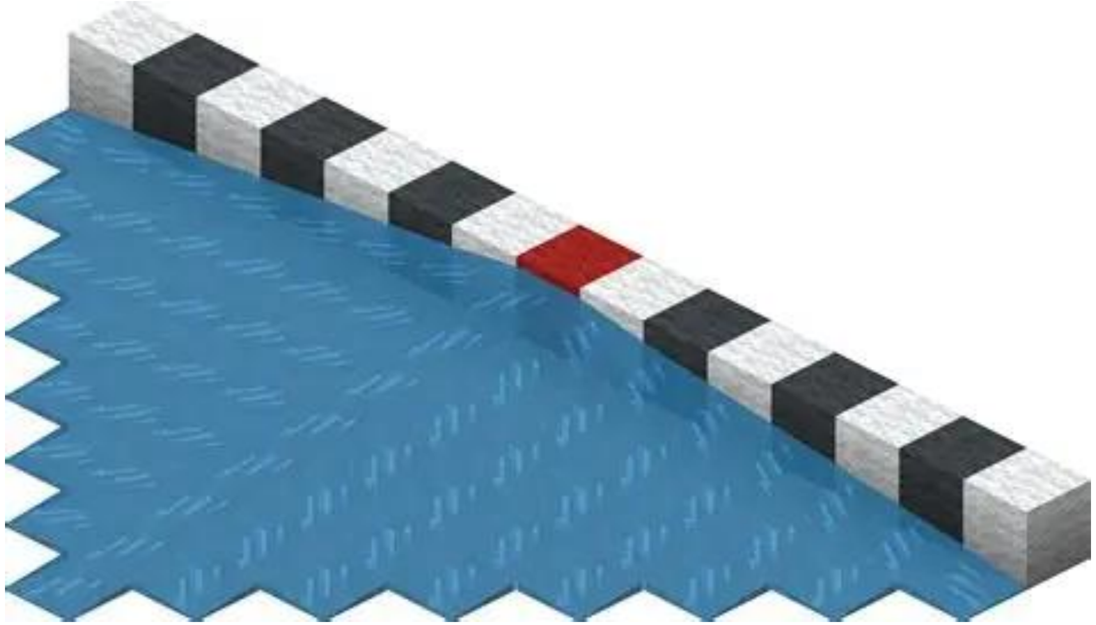
- Common Distributions
  - Continuous
  - Discrete
- Hypothesis Testing
  - P-values
  - χ^2 Test

NOISE

- $Y = \beta_0 + \beta_1 x + \epsilon_i$
- Data points are noised
- How to model it?

DISTRIBUTIONS

Continuous

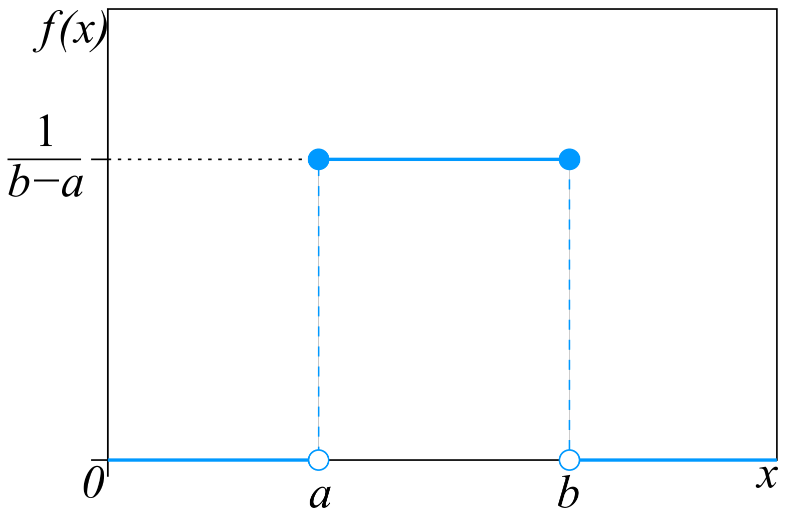


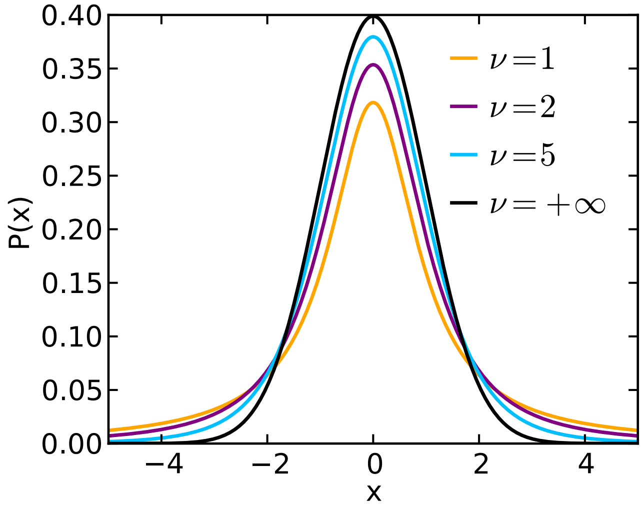
Discrete

Source: https://en.wikipedia.org/wiki/Capillary_wave
Source: https://minecraft.fandom.com/wiki/Water

CONTINUOUS DISTRIBUTIONS

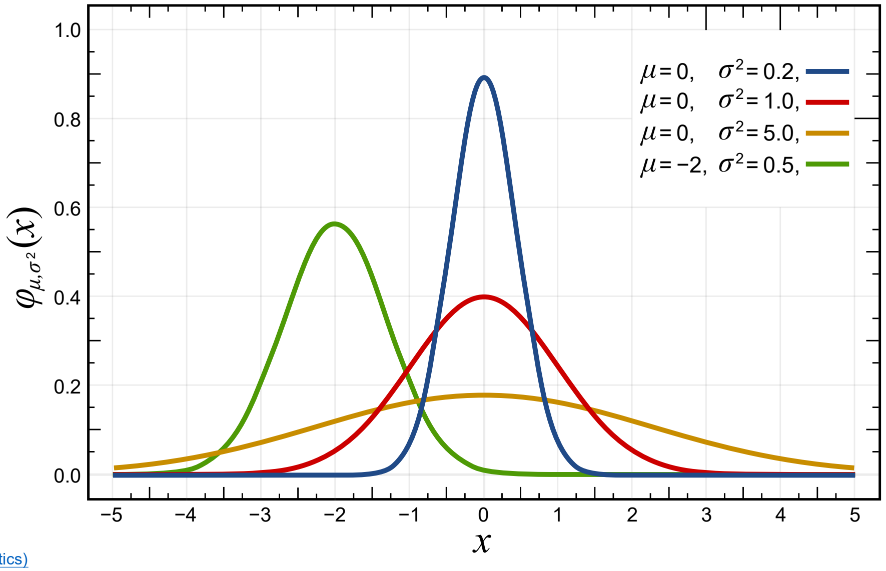
Uniform

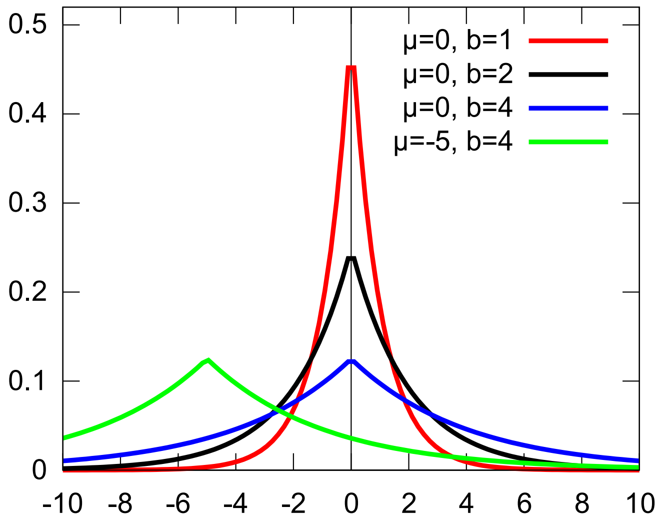




Student's t

Gaussian



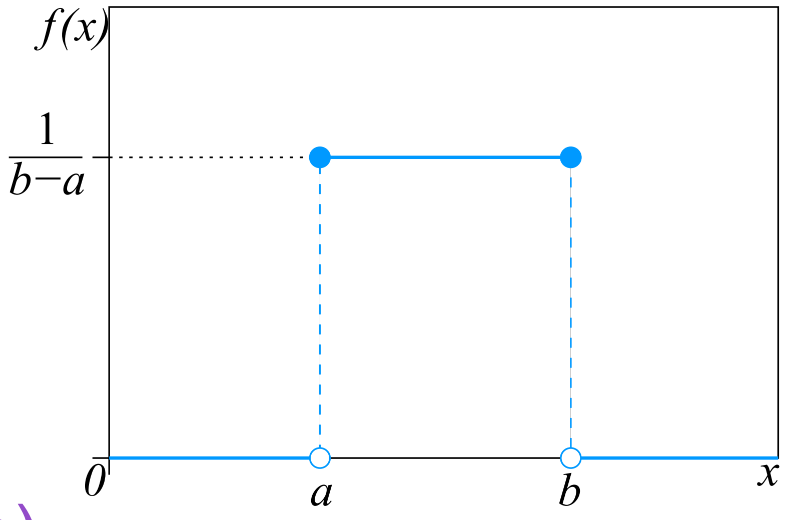


Laplace

Source: https://en.wikipedia.org/wiki/Distribution_(mathematics)

UNIFORM DISTRIBUTION

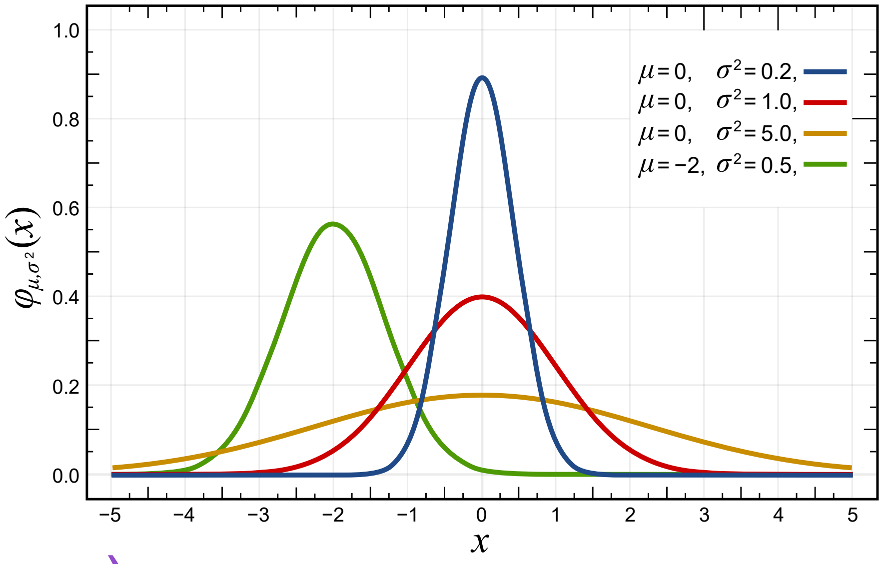
$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \ \text{ or } \ x > b. \end{cases}$$



- Perfectly smooth tabletop
- numpy.random.uniform(low, high)

GAUSSIAN DISTRIBUTION

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

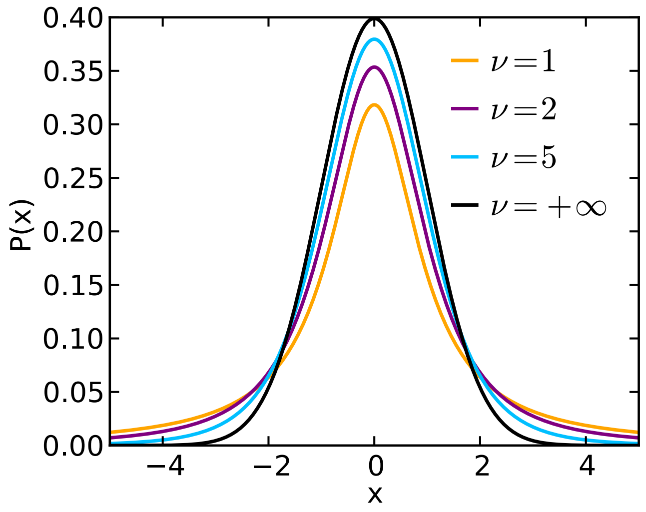


- Exam scores
- Heights and weights
- numpy.random.normal(mu, sigma)

STUDENT'S T DISTRIBUTION

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

- Gaussian distribution w/ fatter tails
- Financial modeling (asset returns)
- numpy.random.student_t(nu)



$f(t) \to N(0,1)$ as $\nu \to \infty$

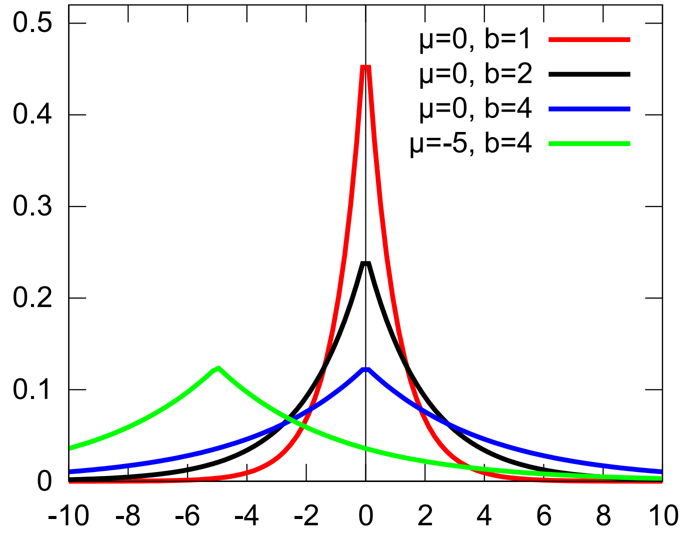
ν: degrees of freedom

Number of data points left after estimating parameters

LAPLACE DISTRIBUTION

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right),$$

- Sharper peak at the center
- Heavier tails
- Signal processing (spiky noise)
- numpy.random.laplace(mu, b)



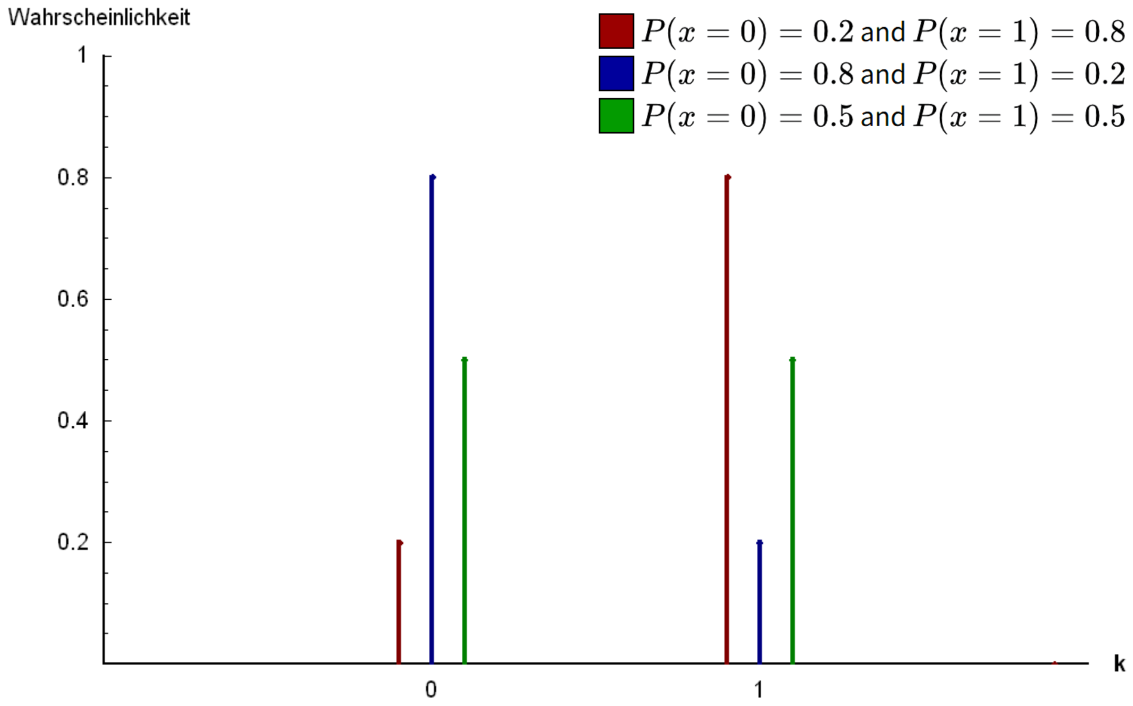
COIN TOSS

- Head=1, Tail=0
- Not all randomness is continuous
- How to measure?

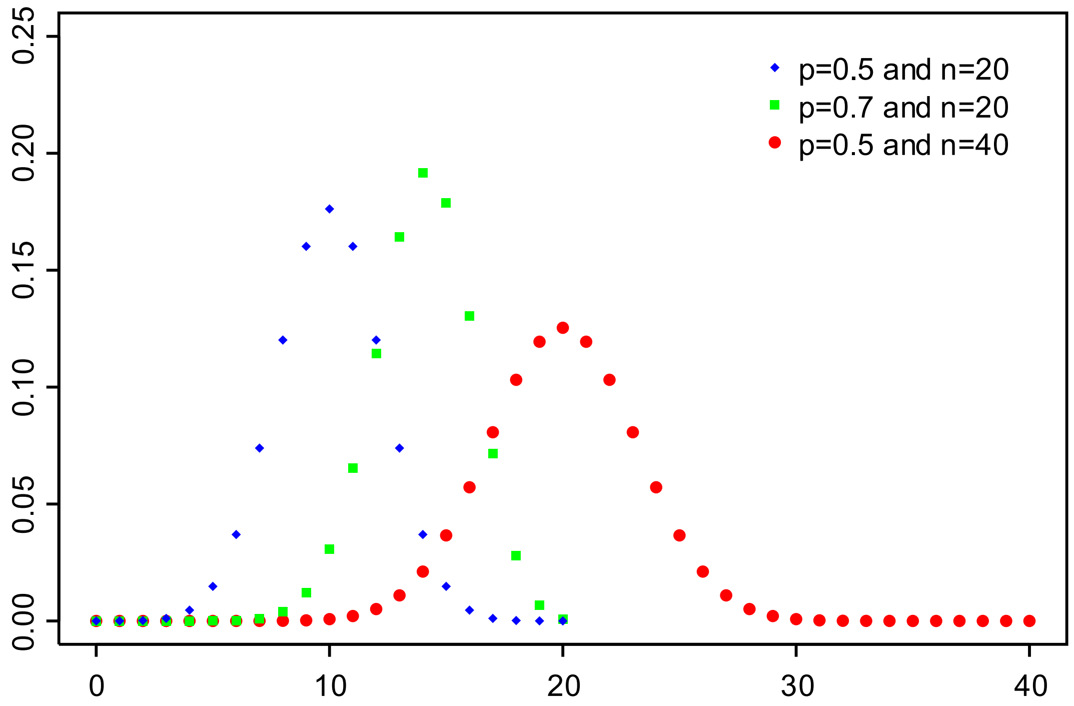
Source: https://en.wikipedia.org/wiki/Coin_flipping

DISCRETE DISTRIBUTIONS

Bernoulli



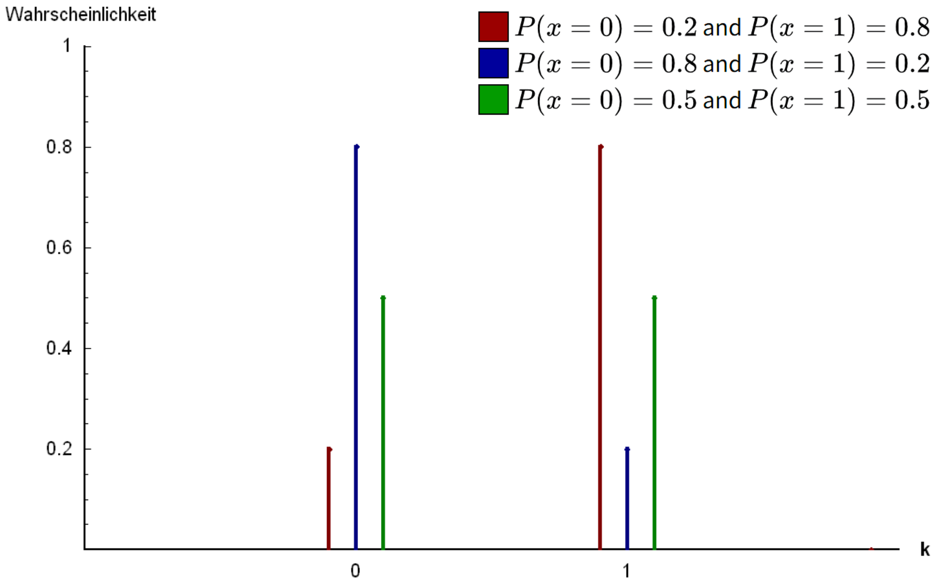
Binomial



Source: https://en.wikipedia.org/wiki/Distribution_(mathematics)

BERNOULLI DISTRIBUTION

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$



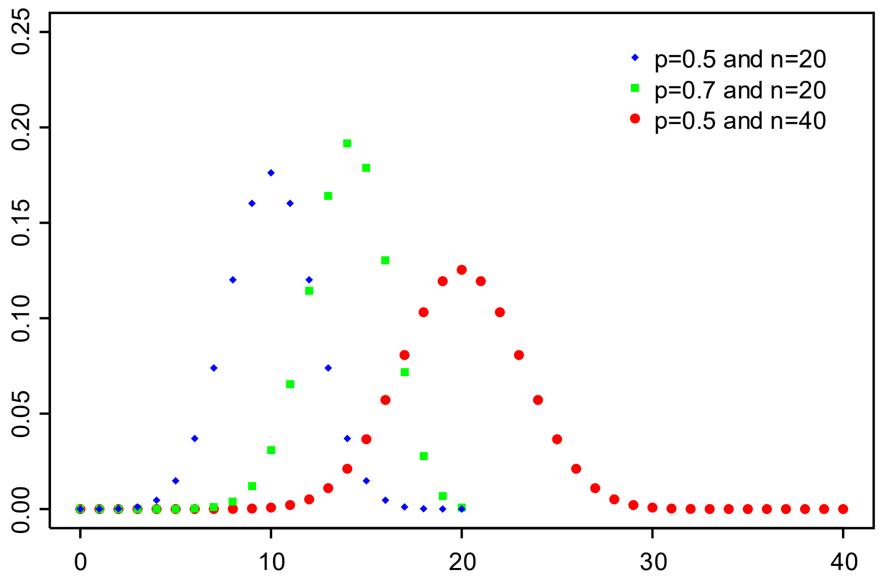
- Tossing a coin → heads/tails
- Light bulb test → work/fail
- numpy.random.binomial(n=1, p)

BINOMIAL DISTRIBUTION

$$f(k,n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Combination of Bernoulli
- Number of heads in 20 tosses
- numpy.random.binomial(n, p)



SO FAR

- Modeling continuous noise —— how much
  - Uniform (all values equally likely)
  - Gaussian (bell curve)
  - Student's t (bell curve with heavy tails)
  - Laplace (sharp peak, heavy tails)
- Modeling discrete noise —— how many
  - Bernouli (single trial)
  - Binomial (number of successes in n trials)

HYPOTHESIS TESTING

How to verify the significance/goodness of our model?

- Question:
  - Is this coin fair?
- Hypothesis:
  - H_0: $p = 0.5$
  - H_1: $p \neq 0.5$
- Data:
  - Head=1, Tail=0
  - 010111010110010101 (eight 0s, ten 1s)

P-VALUES

- The probability of observing data at least as extreme as what we saw, under the **null hypothesis $\boldsymbol{H_0}$**
- $p = P(\text{Test statistic} \geq \text{observed value} \mid H_0)$
  - Small $p \rightarrow$ evidence against H_0
  - Large $p \rightarrow$ evidence consistent with H_0
- Decide whether observed results are "too extreme" under H_0

P-VALUES

- H_0: $p = 0.5$
- X: 010111010110010101 (eight 0s, ten 1s)
  - X~Binomial(n=18, p=0.5)

- $p = P(X \geq 10 \text{ or } X \leq 8 \mid H_0)$

$$= 1 - P(X = 9 \mid H_0)$$
$$= 1 - \binom{18}{9} 0.5^{18}$$
$$\approx 1 - 0.1854$$
$$= 0.8146$$

P-VALUES

- Common conventions
  - $p > 0.1$: Data are very consistent with H_0
  - $0.05 < p < 0.1$: Borderline case
  - $0.01 < p < 0.05$: Moderate evidence against H_0
  - $p < 0.01$: Strong evidence against H_0.

- By convention, people use 0.05 as a threshold
- $p = 0.8146 > 0.05 \rightarrow$ significant

χ^2 TEST

(Chi-Square)

- Whether observed data frequencies match the expected frequencies predicted under the **null hypothesis $\boldsymbol{H_0}$**
- $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
  - Observed counts: $O_1, O_2, \dots, O_k$
  - Expected counts: $E_1, E_2, \dots, E_k$
  - Small $\chi^2 \rightarrow$ good fit under H_0
  - Large $\chi^2 \rightarrow$ bad fit under H_0
- Degree of freedom: $df = k - 1 - r$

Number of estimated parameters e.g. μ, σ

Number of classes

χ^2 TEST

- H_0: $p = 0.5$
- X: 010111010110010101 (eight 0s, ten 1s)
  - Observed: $O_1 = 8,\ O_2 = 10$
  - Expected: $E_1 = 9,\ E_2 = 9$

- $\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(8-9)^2}{9} + \frac{(10-9)^2}{9}$$

$$= \frac{2}{9}$$

$$\approx 0.2222 \qquad df = 2 - 1 - 0 = 1$$

χ^2 TEST

- Reference values for χ^2 with $df = 1$

Significance level (α)	Critical value (χ^2 cutoff)
0.10	2.71
0.05	3.84
0.01	6.63

- By convention, people use 0.05 as a threshold
- $\chi^2 = 0.2222$ (df=1) < 3.84
- Very significant

χ^2 TEST

Chi Square Table & Chi Square Calculator

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588

SO FAR

- P-values
  - Are the results extreme under H_0?
  - Bigger is better
- χ^2 Test
  - Do observed results match expected ones?
  - Smaller is better

Q&A