CS365
Foundations of Data Science

Lectures 2, 3, 4
(1/23, 25, 30)

Charalampos E. Tsourakakis
ctsourak@bu.edu

What is a fair coin?

$\frac{1}{2}$

Are all coin flips random?

- Can a coin flip be “rigged”?
  - Yes!

- Dynamical Bias in the Coin Toss
  by
  P Diaconis, S Holmes, R Montgomery

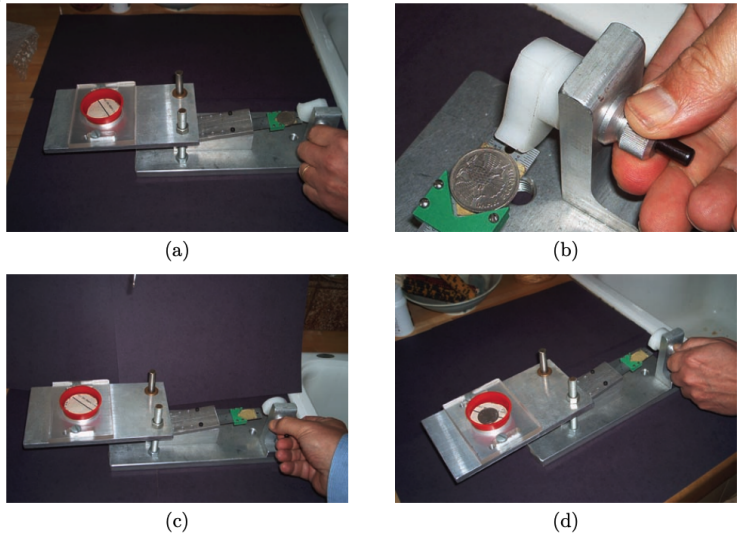
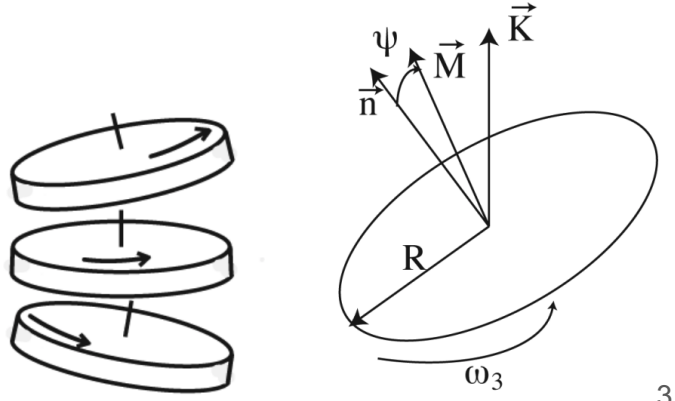
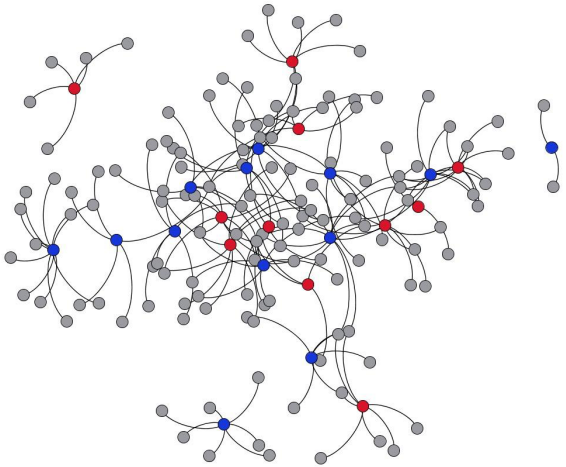
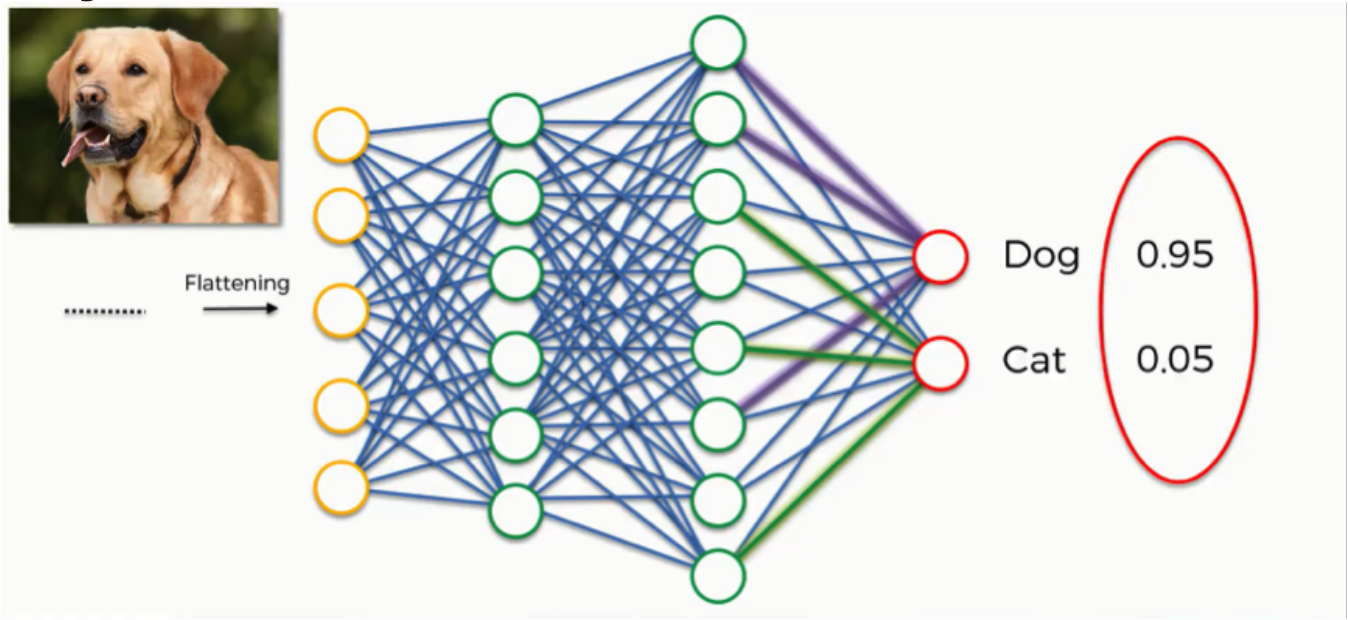


Fig. I



Modeling uncertainty



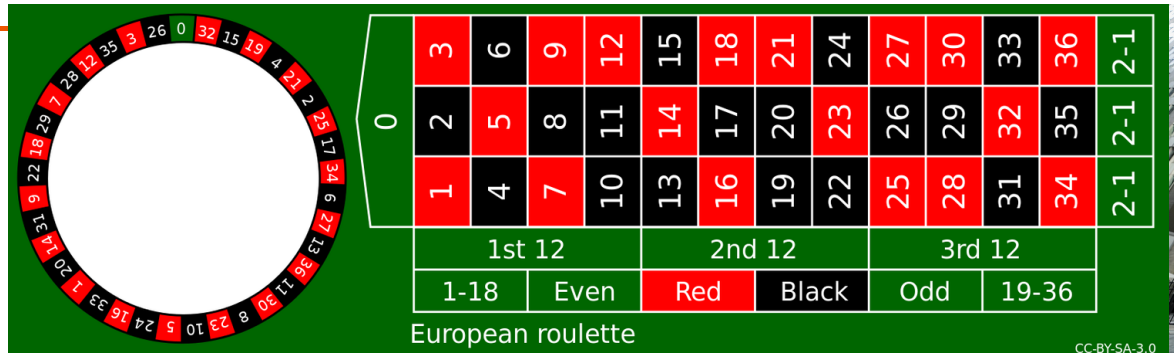


Disease spreading

Source: Blog

covariance?

Modeling uncertainty



- Information theory, modeling the reliability of numerous complex systems, insurance companies, investments etc.
- **Today's agenda**: reminders of prereq probability material through problem solving.

Roulette

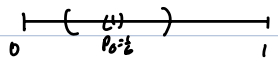
Odds & Payouts at European & American Roulette			
Roulette Bet	**Payout**	**European Roulette Odds**	**American Roulette Odds**
Single Number	35 to 1	2.70%	2.60%
2 Number Combination	17 to 1	5.4%	5.3%
3 Number Combination	11 to 1	8.1%	7.9%
4 Number Combination	8 to 1	10.8%	10.5%
5 Number Combination	6 to 1	13.5%	13.2%
6 Number Combination	5 to 1	16.2%	15.8%
Column	2 to 1	32.40%	31.6%
Dozen	2 to 1	32.40%	31.6%
Even/Odd	1 to 1	48.60%	47.4%
Red/Black	1 to 1	48.60%	47.4%
Low/High	1 to 1	48.60%	47.4%

- Even American: {2,4,6,8,...,34,36}, hence Pr(even)=18/38=0.47368
- Even European: same favorable outcomes {2,4,6,8,...,34,36}, but Pr(even)=18/37=0.4864

전체개수가 달라서 확률이 다르다.

$\hat{p}_6 = \frac{\#\text{ times I observe 6}}{n}$

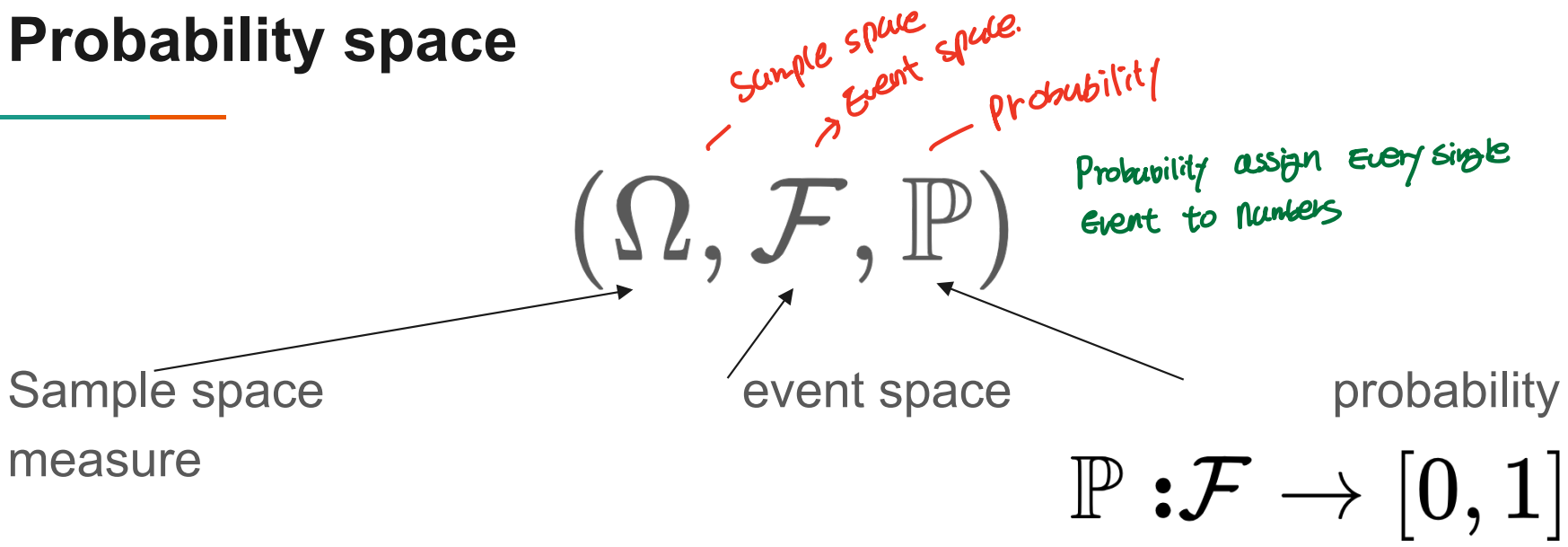
$\hat{p}_6 \xrightarrow{n\to\infty} \frac{1}{6}$



$Pr(|\hat{p}_6 - p_6| \geq 0.01) \leq 0.05$

when $n \to \infty$ $\downarrow$ is 0 since $\hat{p}_6 = p_6$

Probability space



Questions: what is a random variable? What is the difference between continuous and discrete random variables?

Random variable : Toss coin twice

Ω {HH, TT, HT, TH}

$P(\Omega) = 1$

Central limit

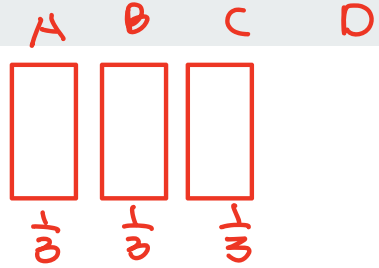
Discrete $X: \Omega \to D$ (Domain D) D $X(\omega)$: # heads in ω
$X(\omega_1)=2$ $X(\omega_2)=1$ $X(\omega_3)=1$ $X(\omega_4)=0$

theorem

Bayes

Monty-hall problem

conditional



Suppose you're on a game show, and you're given the choice of three doors:

- Behind one door is a car; behind the others, goats.
- You pick a door, say No. A, and the host, who knows what's behind the doors, opens another door, say No. C, which has a goat.
- He then says to you, "Do you want to pick door No. B?" Is it to your advantage to switch your choice?

$\frac{P(A \cap D)}{P(D)} = \frac{P(D \cap A)}{P(A)} \times \frac{P(A)}{P(D)}$ $\frac{1}{6} + \frac{1}{3} = \frac{2}{6}$

$\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0$

open C

A: car $P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{1}{2}$

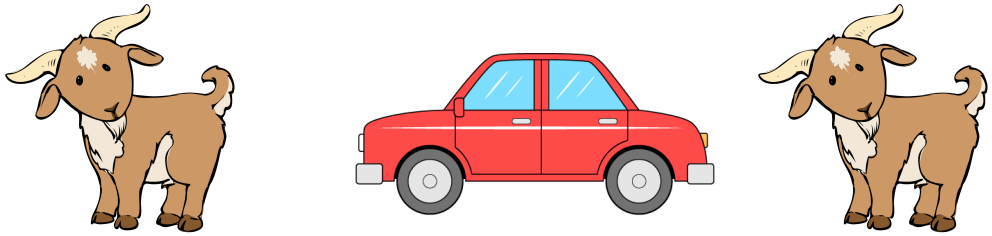
B: car $P(D|B) = 1$ C로열릴수밖에

C: car $P(D|C) = 0$

$P(A|D) = \frac{P(D|A)P(A)}{P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C)}$

$= \frac{P(D|A)P(A)}{P(D)}$

$= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$



P(A|D)

$P(B|D) = \frac{P(D|B)P(B)}{P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C)}$

$= \frac{1 \times \frac{1}{3}}{\frac{3}{6}} = \frac{1}{3} \times \frac{6}{3} = \frac{2}{3}$

Assumptions

Let's make the problem concrete by specifying certain assumptions.

- Let's say the car is placed uniformly at random (**uar**) behind a door.
- Our initial guess is also **uar** uniformly at random $\frac{1}{3}$ $\frac{1}{3}$
- The host opens a door with a goat. When there exist two such doors, i.e., our guess is the car, he chooses **uar**.

$P(A|C) = \frac{P(C|A)P(A)}{P(C)} = \frac{P(C|A)P(A)}{P(C|A)P(A)+P(A|A^c)P(A^c)}$

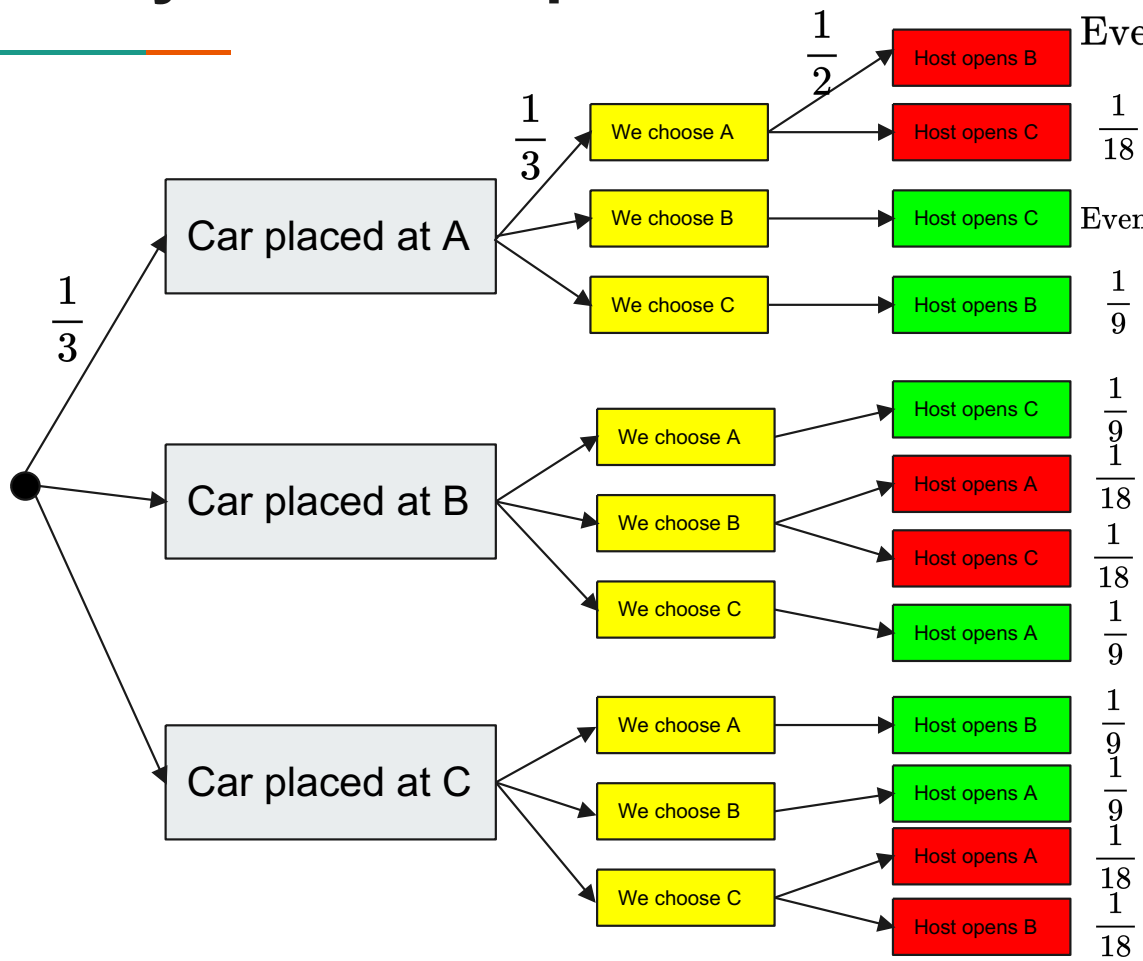
$P(C) =$ 

$\frac{P(A \cap C)}{P(A)} \times P(A) + \frac{P(C \cap A^c)}{P(A^c)} \times P(A^c)$

CS131 Reminder: Four step-method

1. Find the sample space
2. Define events of interest
3. Determine outcome probabilities
4. Compute event probabilities

Monty hall - 4 steps in one slide

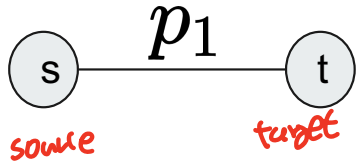


$$\text{Event (AAB)} : \frac{1}{3} \times \frac{1}{3} \times \frac{1}{2} = \frac{1}{18}$$

$$\text{Event (ABC)}: \frac{1}{3} \times \frac{1}{3} \times 1 = \frac{1}{9}$$

Therefore, switching yields
Probability of winning 6/9=⅔
Vs not switching 6/18=⅓

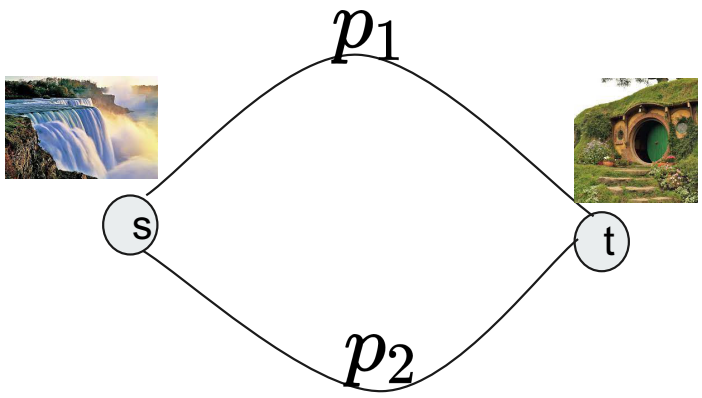
Transfer water



- Consider a water source **s** and a destination village **t**.
- Each pipe *i* has probability of failure p_i. Pipes fail ***independently***.
- **Question**: What is the probability we cannot get water from s to t? In other words: p_1
  - when is the village **t** not reachable from the water source **s**?

$1 - p_1$

Exercise 1



- Clearly, there is no path if both pipes fail
- Since they are independent, the probability of this event is the product of the probabilities of the individual events

Thus, failure probability is p_1p_2

안되는것 끼리곱한것

$P_1 \times P_2$

independent

Reminders: Independent events, conditional probability

Intuitive two events A,B are dependent if A's occurrence or non-occurrence provides us with some information about event B.

Formally, A,B are independent if and only iff $\Pr(A \cap B) = \Pr(A)\Pr(B)$

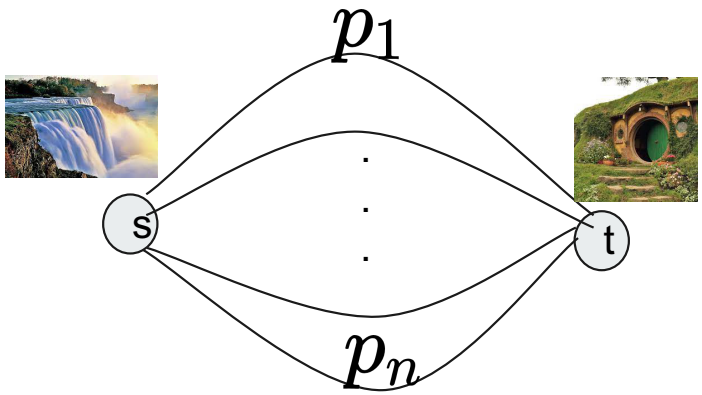
By rearranging we get $\Pr(A) = \frac{\Pr(A \cap B)}{\Pr(B)}$

$\Pr(A) = \frac{\Pr(A \cap B)}{\Pr(B)}$ independent

Recall that by the law of conditional probability $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$

Therefore, when A,B are independent Pr(A)=Pr(A|B) and of course Pr(B)=Pr(B|A).

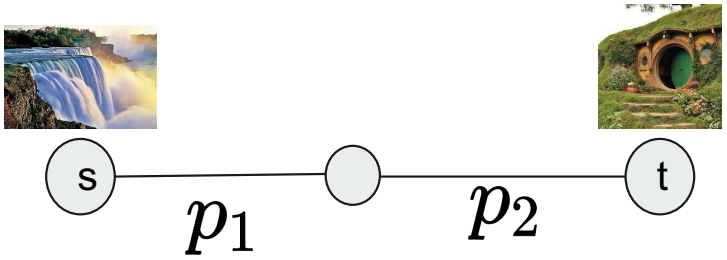
Exercise 2



- Clearly, there is no path if **all** pipes fail
- Since they are independent, the probability of this event is the product of the probabilities of the individual events

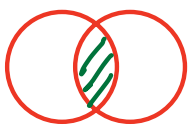
Thus, failure probability is $p_1 p_2 \ldots p_n$

Exercise 3

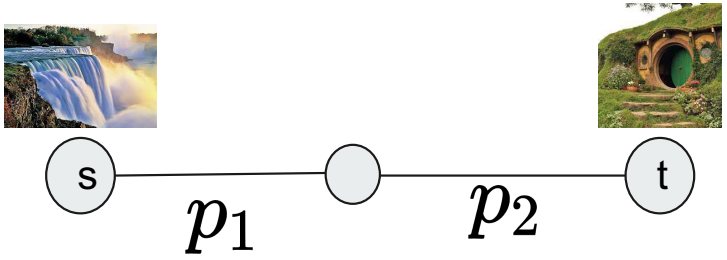


- Clearly, there is no path if **at least one of** the pipes fail
- Let A_i be the event that pipe i fails.
- Then,

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\ &= p_1 + p_2 - \Pr(A_1)\Pr(A_2) \\ &= p_1 + p_2 - p_1 p_2\end{aligned}$$



Exercise 3



- **Using conditional probability**

- We condition on whether the one of the two pipes (say the first) is broken or not.

- Let A_i be the event that pipe i fails.

A_1 fail 하면 이어 못가짐.

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + (1 - \Pr(A_1))\Pr(A_2) \\ &= p_1 + (1 - p_1)p_2 \\ &= p_1 + p_2 - p_1 p_2\end{aligned}$$

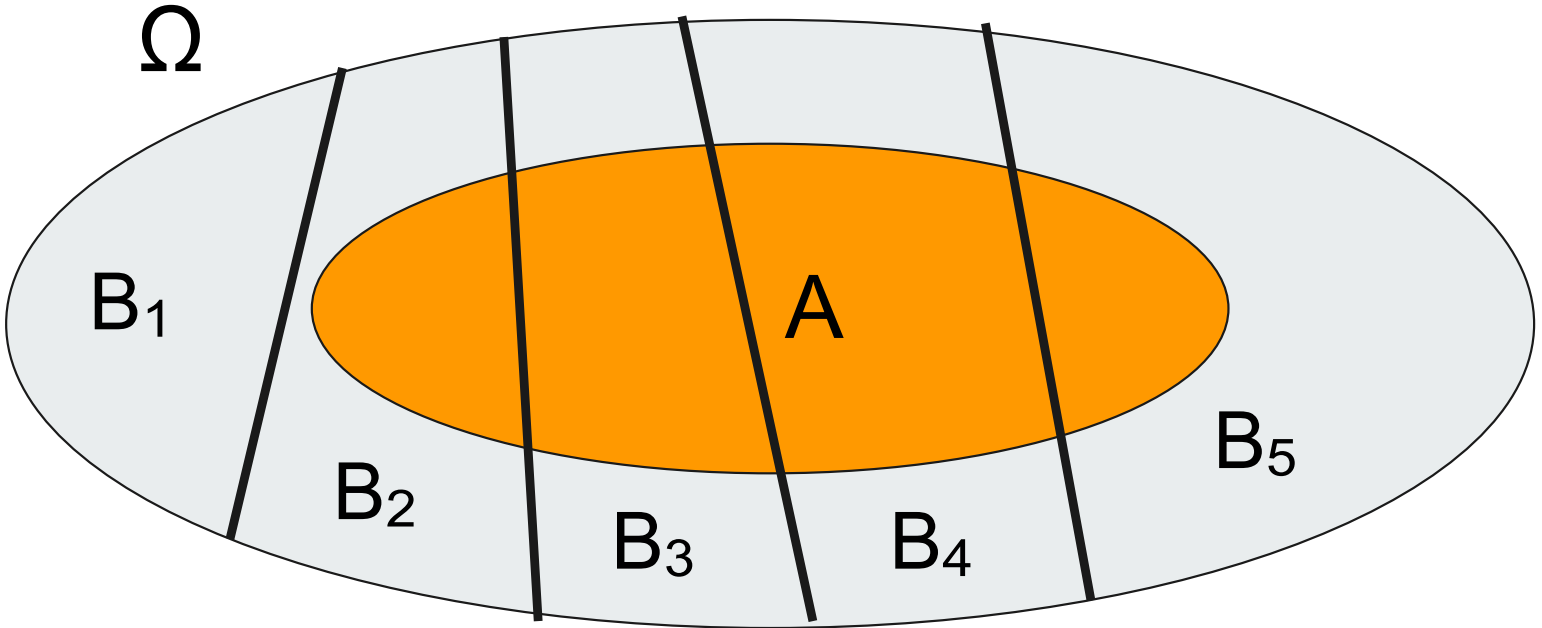
success and fail

$\Pr(A_1) + (1 - \Pr(A_1) \times \Pr(A_2)$

We used the law of total probability

Let Ω be a probability space. Let $B_1,\ldots,B_m$ be a partition of Ω. Then,

$$\Pr(A) = \sum_{i=1}^{m} \Pr(A \cap B_i) = \sum_{i=1}^{m} \Pr(B_i)\Pr(A|B_i)$$



Exercise 4

- Instead of thinking the probability that t will not be reachable from s, we think of the probability that it is. **Reminder**: $\Pr\left(\bar{A}\right) = 1 - \Pr(A)$
  - Let $\bar{A}_i$ be the event that pipe i does not fail.
- The probability of not failing is $\Pr\left(\cap_{i=1}^{n} \bar{A}_i\right) = \prod_{i=1}^{n} \Pr\left(\bar{A}_i\right) = \prod_{i=1}^{n} (1 - p_i)$

Therefore, the right answer is $1 - \prod_{i=1}^{n} (1 - p_i)$.

Reminder: chain rule

Chain rule:

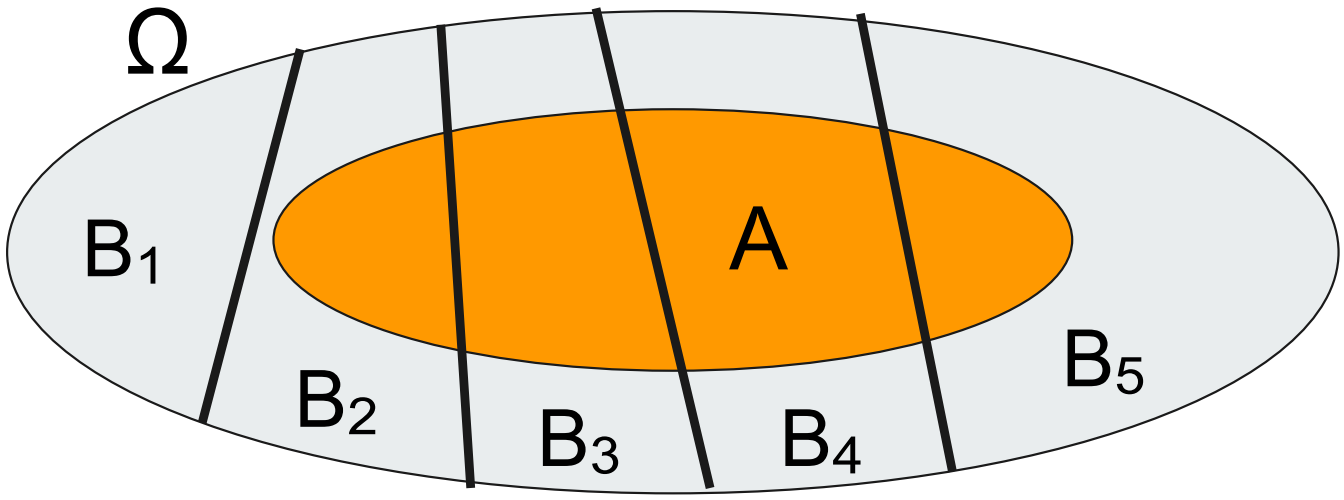
$$\Pr(A_1 \cap \ldots \cap A_n) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_2A_1)\ldots\Pr(A_n|A_{n-1}..A_1)$$

In our case the events are mutually independent, so this simplifies to the product of the individual probabilities of the events A_i.

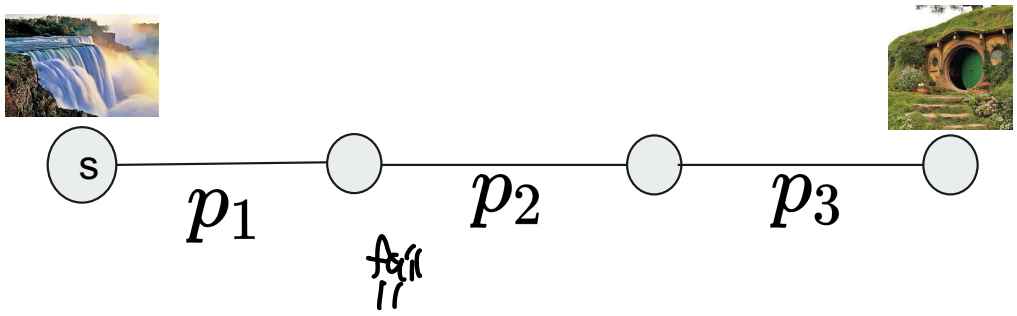
Question: what is the difference between pairwise and mutually independent events?

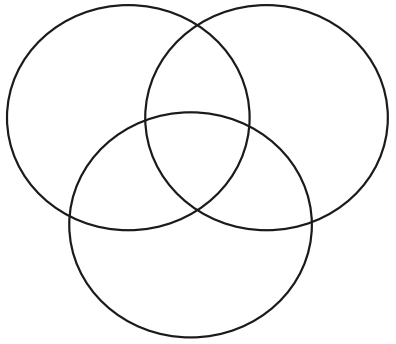
Conditional probability + Law of total probability → Bayes rule

$$\Pr(B_i|A) = \frac{\Pr(B_i \cap A)}{\Pr(A)} = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{n} \Pr(B_j)\Pr(B_j|A)}$$



Exercise n=3





$$1-(1-p_1)(1-p_2)(1-p_3) = 1-(1-p_1)(1-p_2-p_3+p_2p_3)$$

$$= 1-(1-p_2-p_3+p_2p_3-p_1+p_1p_2+p_1p_3-p_1p_2p_3)$$

$$= p_1+p_2+p_3-p_1p_2-p_1p_3-p_2p_3+p_1p_2p_3$$

Does this remind of something from CS131?

Exercise 4

- We condition on whether the one of the two pipes (say the first) is broken or not.
- Let A_i be the event that pipe i fails.
- We are interested in $\Pr(A_1 \cup \ldots \cup A_n)$.

Inclusion-exclusion

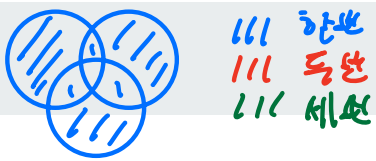
$$\Pr\left(\cup_{i=1}^{n} A_i\right) = \sum_{\mathcal{J}\subseteq\{1,\ldots,n\};|\mathcal{J}|=k} (-1)^{k+1} P\left(\bigcap_{i\in\mathcal{J}} A_i\right)$$

Proof sketch (inductive proof)

When n=1 the statement is obvious. Use the IH and the fact that

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) + P\left(A_{n+1} \setminus \bigcup_{i=1}^{n} A_i\right)$$

$$= P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) - P\left(\bigcup_{i=1}^{n}(A_i \cap A_{n+1})\right).$$



Inclusion exclusion

Another convenient way to write the IE formula is the following

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) = S_1 - S_2 + S_3 - \ldots + (-1)^{n-1} S_n$$

where $$S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}).$$

In our setting, due to the independence of the events A_i we can write the following expression

$$\Pr(\cup A_i) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq [n], |I|=k} \prod_{i \in I} \Pr(A_i)$$

let's write down some terms

$$\begin{aligned} \Pr(\cup A_i) &= p_1 + \ldots + p_n \\ &\quad -(p_1 p_2 + \ldots + p_{n-1} p_n) \\ &\quad +(p_1 p_2 p_3 + \ldots + p_{n-2} p_{n-1} p_n) \\ &\quad - \ldots \end{aligned}$$

Union bound

upper bound

Let $A_1,..,A_n$ be events in a probability space. Then, we get the following upper bound on the probability of their union.

$$\Pr(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} \Pr(A_i)$$

Hats off



.....


n men with hats enter a room, and take off their hats. On their way out, hats are mixed up according to a random permutation, so each person receives a random hat. The probability each man gets his own hat is 1/n.

uniformly at Random

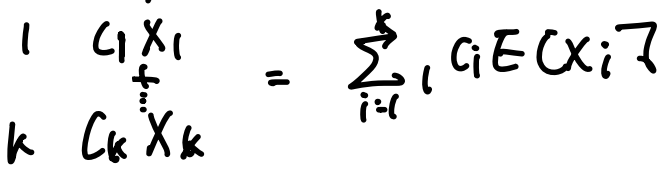
- Since the permutation is chosen uar, the probability each man gets his own hat is 1/n.
- Question: How many men in expectation will get their own hat?

Linearity of Expectation

- Expected value of the sum of random variables is equal to the sum of their individual expectations
  - Important point: this holds regardless of whether they are independent!

- Let $X_1,..,X_k$ be random variables, $c_1,..,c_k$ constants.

$$\mathbb{E}\left[\sum_{i=1}^{k} c_i X_i\right] = \sum_{i=1}^{k} c_i \mathbb{E}[X_i]$$



Hats off - Variance computation

- We define for each man an indicator variable that indicates whether he got his own hat.
  - Let X_i=1 iff the i-th man receives his hat. We know that $Pr(X_i=1)=1/n$
  - Are the Xi's independent?
- We define the variable $S_n = \sum_{i=1} X_i$
  - What is the meaning of this variable?

probability very low 1/n

Nope → well as n changes $n \to \infty$ independent

$n \to 1$ Not independent

- We compute $E[S_n]$ using

$$E[S_n] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \left(1 \times \frac{1}{n} + 0 \times \left(1 - \frac{1}{n}\right)\right) = \frac{n}{n} = 1$$

Expectation of single $E[X_i]$ and they are all same

Hats off - Variance computation

- Let's compute the variance of S_n
  - Question: Why would we like to do this? What insights can this computation provide us with?
- Attempt 1: The variance of the bernoulli variable X_i is $\frac{1}{n}(1-\frac{1}{n})$
  - Therefore $\mathbb{V}\mathrm{ar}[S_n] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] = 1 - \frac{1}{n}$. Is this correct?

Error! Why?

$$\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] + 2\sum_{i<j} \mathbb{C}\mathrm{ov}[X_i, X_j]$$

Reminder: $\mathbb{C}\mathrm{ov}[X,Y] = [(X-EX)(Y-EY)]$
$= E[XY] - (EX)(EY).$

Hats 0ff - Variance computation

- It is clear that the variables X_i are not even pairwise independent!
  - It is straightforward to argue that for distinct i,j, $E[X_i X_j] = \frac{1}{n(n-1)}$
  - Conditioned on Frank Sinatra picking first hat, and getting the right one, the probability Clint Eastwood getting his is 1/(n-1)>1/n
- Furthermore, $E[X_i^2] = \frac{1}{n}$

- In order to compute the variance of Sn, we will compute first the expectation of its square as follows:

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = (\text{expand+linearity of expectation})$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E}[X_i X_j] = n\frac{1}{n} + n(n-1)\frac{1}{n(n-1)} = 2$$

Hats 0ff - Variance computation

- Therefore, we obtain that the variance is equal to 1:

$$\mathbb{V}\mathrm{ar}[S_n] = \mathbb{E}[S_n^2] - \mathbb{E}[\mathbb{S}_n]^2 = 2 - 1 = 1$$

- While it is greater by 1-1/n, it is only greater by the additive term 1/n.
  - **Note:** As n grows, this term goes to 0, so even the wrong answer is not far from the correct one.
  - This is not always the case, sometimes variance can be huge (example?)!

- What is the actual distribution of S_n?
  - Intuitively our Jupyter notebook suggests that it is a Poisson distribution in the limit of n->inf.
  - But what does this mean?

Markov's inequality

Let X be a non-negative random variable and suppose that E[X] exists.

For any t>0,

$$\boxed{\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}}$$

Proof:

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx = \int_0^t x f(x) dx + \int_t^\infty x f(x) dx \geq \int_t^\infty x f(x) dx \geq t \int_t^\infty f(x) dx = t \Pr(X > t)$$

Examples

- Let X be a non-negative RV. Then $\Pr(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$

  $t = k \cdot E[X]$

  - For example X~Bin(1000, 0.1). Then, E[X]=100. The tail probability $\Pr(X \geq 400) \leq \frac{1}{4}$

    Probability $\Pr(X=400) = \sum_{k=400}^{1000} \binom{1000}{k} (0.1)^k (0.9)^{1000-k} \leq \frac{100}{400}$

  - The bound is not tight, we can get much tighter bounds (to be seen)
    - Exponential method
    - Nifty manipulation of the binomial coefficients, i.e.,

$$\sum_{k=400}^{1000} \binom{n}{k} p^k (1-p)^{1000-k} = \sum_{k=400}^{1000} \binom{n}{k}^k \left(\frac{1}{10}\right)^k \left(\frac{9}{10}\right)^{1000-k} \leq \dots$$

Let's see another way to get something tighter using the **2nd moment** method

Chebyshev's inequality

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

- Let $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}\text{ar}[X]$

Then $\Pr(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$

- **Example 1**

Let's consider the same example as before, namely bound the tail $\Pr(X \geq 400)$

where X~Bin(1000, 0.1). Let's apply Chebyshev.

$$\Pr(X \geq 400) = \Pr(X - 100 \geq 300) \leq \Pr(|X - 100| \geq t) \leq \frac{\mathbb{V}\text{ar}[X]}{t^2} = \frac{90}{300^2} = 0.001$$

Chebyshev's inequality

- **Example 2**
  Suppose that we toss a fair coin 100 times. What is the probability we see tails more than 60 times or less than 40?

Solution

Let X be the number of tails. Then X~Bin(100,½). By applying Chebyshev's inequality we obtain:

$$\Pr(X \leq 40 \cup X \geq 60) = \Pr(|X - 50| \geq 10) \leq \frac{25}{100} = \frac{1}{4}$$

Next week, we will see a method that gives a much tighter upper bound, close to the true probability ~0.05

Weak law of large numbers

- Let $X_1, X_2, \ldots$ be a sequence of iid RVs with mean μ, and standard deviation σ. Consider the sum

$$S_n = X_1 + \ldots + X_n$$

Then as $n \to \infty$, for all ε>0 the empirical average converges to μ in probability, i.e.,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0.$$

Weak law of large numbers

- Proof of WLLN using Chebyshev's inequality

(more details on blackboard)

$$\Pr\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathbb{V}\mathrm{ar}(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

Examples

1. Consider rolling a regular die n times. Then by WLLN we obtain that the sum of the first n rolls as n grows, for any ε>0 satisfies

$$\Pr\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| > \epsilon\right) \to 0$$

Weak law of large numbers - Confidence intervals

The proof of WLLN gives us a way to construct confidence intervals

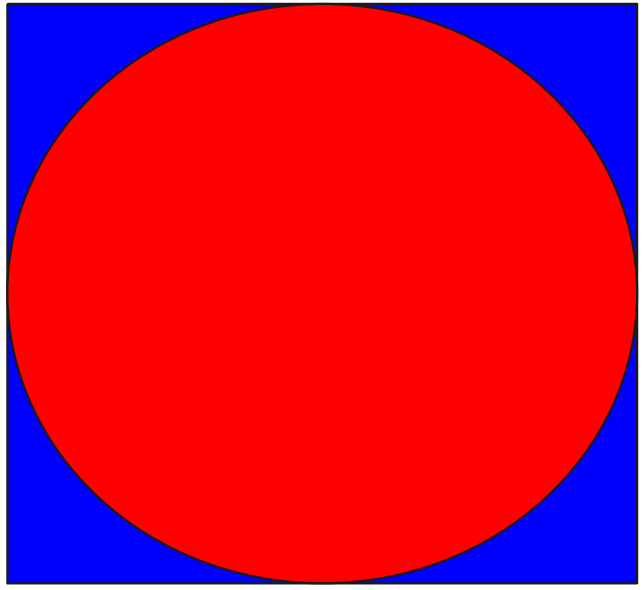
$$\Pr\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathbb{V}\mathrm{ar}(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

Specifically, if 0<α<1 we get the following confidence interval

$$\mu - \frac{\sigma}{\sqrt{\alpha n}} \leq \frac{S_n}{n} \leq \mu + \frac{\sigma}{\sqrt{\alpha n}} \text{ wp at least } 1 - \alpha$$

How to estimate π?

- We generate a point uniformly at random within the unit square [0,1]x[0,1]
- Let E be the event that the random point will fall within the circle.
  - What is the probability of this event?



$$\Pr(\mathcal{E}) = \frac{\text{circle area}}{\text{area of square}} = \frac{\pi r^2}{1 \times 1} = \frac{\pi}{4}$$

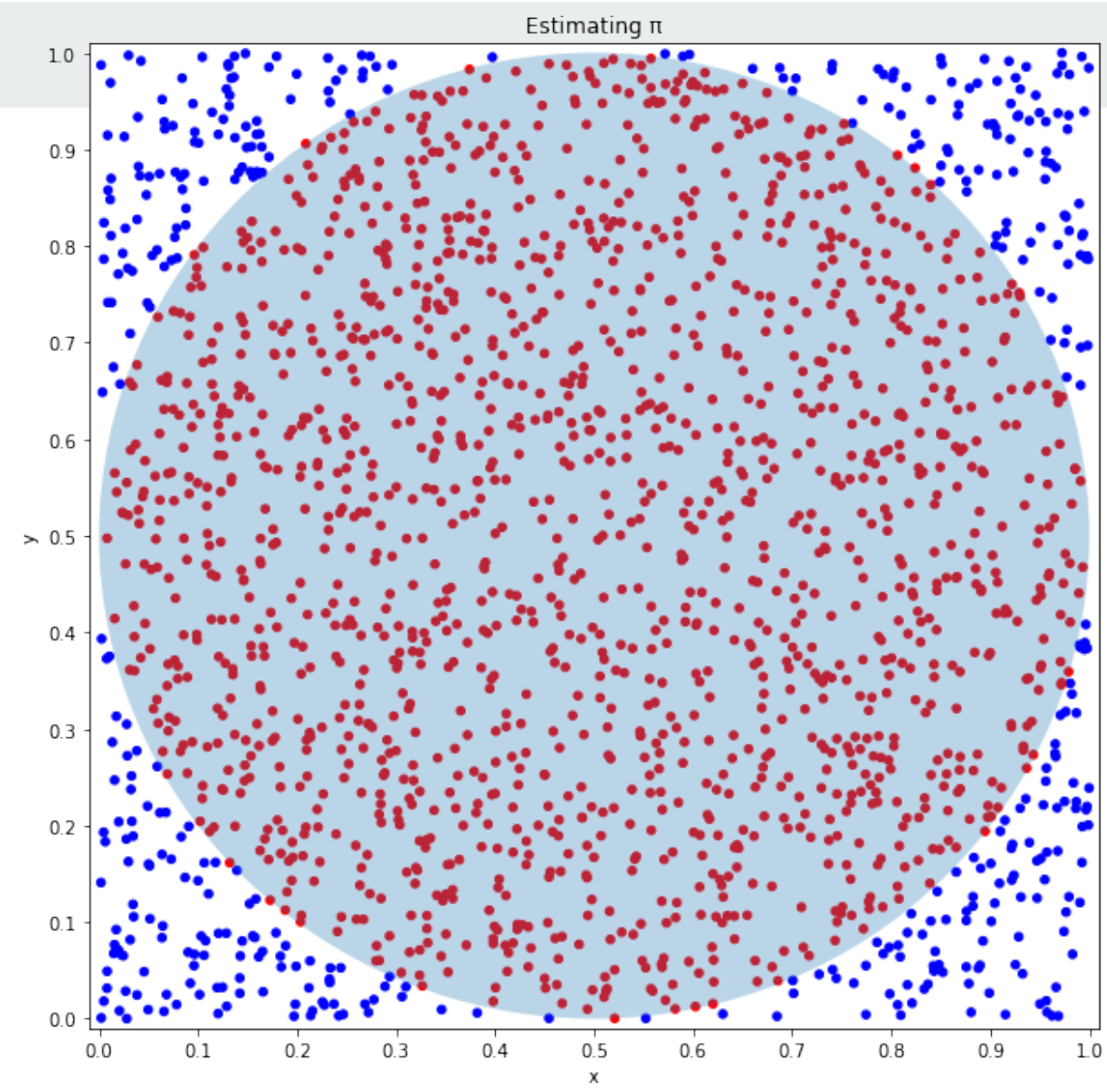
Algorithm for estimating π

- Consider the following algorithm for estimating π:
  - We generate n random points (x_i, y_i) in the unit square [0,1]x[0,1]
  - We test for each point if it falls with the circle
  - Define $S_n = \#\text{points inside circle}$
  - Let our π estimate be $\tilde{\pi} = 4\frac{S_n}{n}$

Since $S_n \approx n \cdot \frac{\pi}{4}$

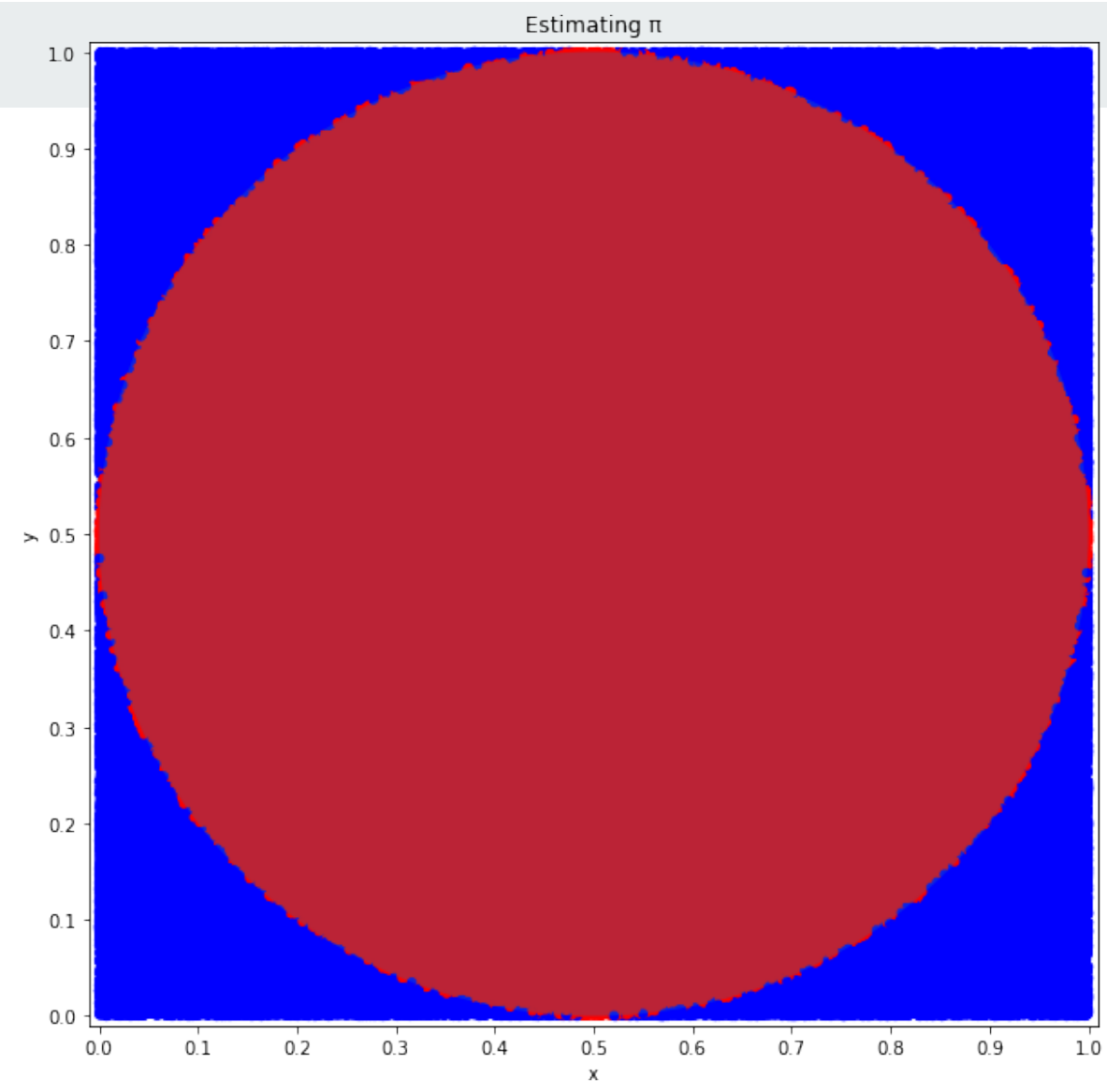
$\frac{S_n}{n} \approx \frac{\pi}{4}$, $4\frac{S_n}{n} \approx \pi$

Why does this algorithm make sense?



The red points are the ones that “fall” within the circle, whereas the blue outside the circle.

- Jupyter notebook (Python)
- Jupyter notebook (Julia)



- Clearly, sampling more points cannot hurt the approximation.
- On the contrary, we know by the WLLN that our approximation of π should improve
- How large does n have to be to make sure that we have a good approximation with high enough probability?

Algorithm for estimating π

We can write the RV S_n as the sum of n independent Bernoulli variables:

$$S_n = \sum_{i=1}^{n} X_i,\ X_i = \begin{cases} 1 & \text{if i-th point inside circle} \\ 0 & \text{otherwise (o/w)} \end{cases}$$

- QQ (quick question): What is the distribution of S_n? → Binomial
- $\mathbb{E}[S_n] = n\frac{\pi}{4}, \mathbb{V}\text{ar}[S_n] = n\frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)$

Expectation : np

Variance : npq

Thus in expectation $f = \dfrac{\#\,\text{points inside circle}}{n}$ in expectation is $\dfrac{\pi}{4}$

Confidence intervals

- Using the WLLN that we obtained using Chebyshev's inequality, we get that:

$$Pr(|\frac{X_1 + \ldots + X_n}{n} - \frac{\pi}{4}| \geq \epsilon) \leq \frac{\pi/4(1-\pi/4)}{n\epsilon^2} \leq \delta$$

- We can use this inequality in various ways.
  - E.g., if ε, n are given we can compute the confidence 1-δ.

Chernoff bound

- Let X_i=1 with probability p_i, 0 with prob. 1-p_i Bernoulli
- Define $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$ → ×1 ⇒ 베르누이 확률 변수 합.

  Then, the following probability inequality holds:

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\frac{\delta^2}{3}\mu} \text{ for all } 0 < \delta < 1$$

평균으로 갈수록 0~1사이의 값들

cheby shev's $\Pr(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$: σ^2 알아야함. 시험에 물어볼 수 있다고함.

Office Hour 1. why $\frac{X}{n}$ not X

X is ↙ in this case.

Application: flipping a fair coin

Suppose we flip a fair coin n times. Let X be the number of Heads.

- Chebyshev's inequality

$$\Pr\left(\left|\frac{X}{n}-\frac{1}{2}\right|\geq\epsilon\right)=\Pr\left(\left|X-\frac{n}{2}\right|\geq\epsilon n\right)\leq\frac{\mathbb{V}\mathrm{ar}[X]}{n^2\epsilon^2}=\frac{1}{4n\epsilon^2}$$

why $\frac{X}{n}$?

- Chernoff bound (for which ε is this valid? Always check the assumptions before applying a theorem)

Chernoff

Ratio (probability)?

$$\Pr\left(\left|\frac{X}{n}-\frac{1}{2}\right|\geq\epsilon\right)=\Pr\left(\left|X-\frac{n}{2}\right|\geq\epsilon n\right)\leq 2e^{-(2\epsilon)^2\frac{n}{6}}=2e^{-\frac{2n\epsilon^2}{3}}$$

IXP

IXH to X ?

↓ fair coin

Sampling theorem

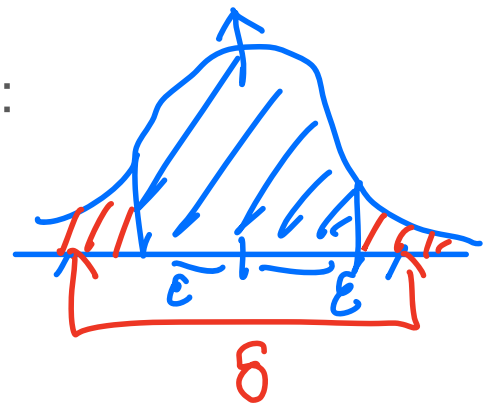
- Chernoff bound comes in many variations, see here

A beautiful corollary of Chernoff bound is the following.

Sampling theorem: Given n independent 0-1 RVs X_i such that $\Pr(X_i=1)=p$ (i=1…n),

where $n \geq \frac{3}{\epsilon^2} \ln\left(\frac{2}{\delta}\right)$ then the following holds:

$$\Pr\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \leq \epsilon\right) \geq 1 - \delta$$



Sampling theorem

- Let’s assume we want to be ε-accurate with a certain confidence 1-δ about the fraction of population that will vote for Biden

$$n \geq \frac{3}{\epsilon^2} \ln\left(\frac{2}{\delta}\right)$$

  - The poll size does not depend on the size of the total population

Caveat: Doing polls with independent samples who truthfully report their vote is another business, and a challenging problem

Lindeberg-Levy Central Limit theorem

Let $X_1, X_2, \ldots$ be a sequence of iid RV with mean μ, variance σ^2. Consider the sum $S_n = X_1 + \ldots + X_n$ and normalize it to obtain the RV Z_n with zero mean, and unit variance as follows:

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}\text{ar}[S_n]}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

Then as $n \to \infty$, $Z_n \rightsquigarrow \mathcal{N}(0,1)$

office hour

In other words $\Pr(Z_n \geq t) \to \Pr(g \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-\frac{x^2}{2}} dx$ as $n \to \infty$.

Applications: Confidence intervals

Exercise

Suppose we have collected 100 points from an unknown distribution, for which we know that the true population mean is 500, and the standard deviation 80.

(a) Find the probability that the sample mean will be inside the interval (490, 510)

(b) Find an interval such that 95% of the sample average is covered.

Applications: Confidence intervals

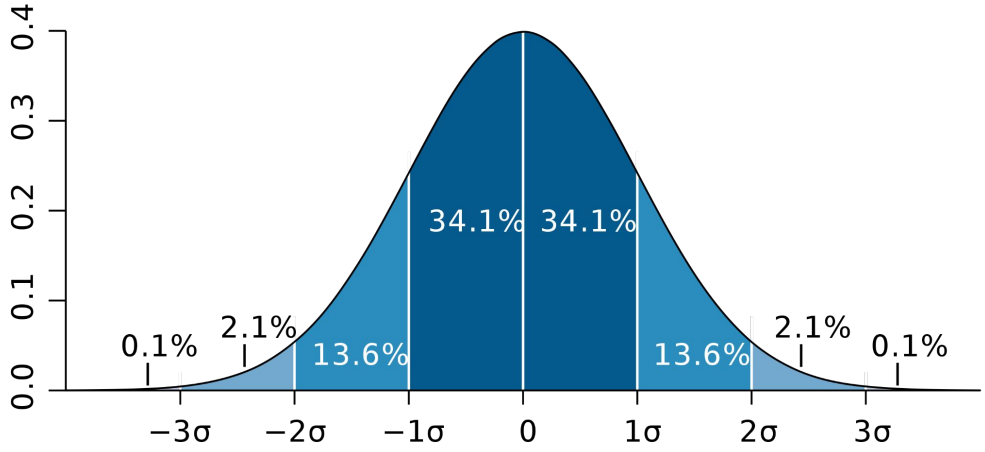
a) By the CLT we know that the sample mean converges to

$$\bar{X}_n \to N\left(500, \left(\frac{80}{\sqrt{100}}\right)^2\right) \quad \text{in distribution. Therefore,}$$

$$\Pr\left(490 \leq \bar{X}_n \leq 510\right) = \Pr\left(\frac{490-500}{\frac{80}{\sqrt{n}}} \leq \bar{Z}_n \leq \frac{510-500}{\frac{80}{\sqrt{n}}}\right) = \Phi(1.25) - \Phi(-1.25) = 0.789$$

Here $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$ is the usual cdf.

Applications: Confidence intervals



- CLT conveniently provides us with confidence intervals tighter than WLLN
- If X is a gaussian RV with mean μ and std σ, then we know

$$\Pr(-\sigma < X - \mu < \sigma) \approx 0.68$$
$$\Pr(-2\sigma < X - \mu < 2\sigma) \approx 0.95$$
$$\Pr(-3\sigma < X - \mu < 3\sigma) \approx 0.99$$

- Thus, when the CLT applies, we can use it in combination with the knowledge of properties of the Normal distribution to get confidence intervals as the following example shows.

Applications: Confidence intervals

b) The mass on each side of the tails is at most (1-0.95)/2=0.025

Let's find the value z for which Φ(z)=1-0.025=0.975.

- Calculator online yields x_{up}=1.96, and by symmetry x_{low}=-1.96.

$$\sqrt{100}\frac{z-500}{80} = \pm 1.96 \Rightarrow z_{low} = 484.32,\ z_{up} = 515.68$$

In other words, the interval of 95% confidence is

$$\Pr\left(484.32 \leq \bar{X}_n \leq 515.68\right) = 0.95$$

The Thumbtack problem

- A billionaire from the suburbs of Boston asks you a question:

  – He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?

  – You say: Please flip it a few times:

Heads

- **You say**: Pr(Heads)=4/7.
- **He says**: Why?

The Thumbtack problem

Assumptions

1. Pr(Heads)=p, Pr(Tails)=1-p

2. Flips are iid

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

- Consider the sequence HTHHTHT?

Pr(HTHHTHT)= $p(1-p)pp(1-p)p(1-p) = p^4(1-p)^3$

office Hour 3

Comprehension question: Where is the “n choose k” factor in binomial?

The Thumbtack problem

- We shall denote Pr(HTHHTHT) as Pr(HTHHTHT;p)
  - p is not a random variable (at least for now..)

- D data → HTHHTHT
  θ parameter(s)
  We refer to Pr(D|θ) as the likelihood of the data under the model.

- In our problem

$$\Pr(\mathcal{D};p) = p^4(1-p)^3$$

why?

$\binom{7}{4} p^4(1-p)^3$

independent

Maximum Likelihood Estimate (MLE)

- **Data**: thumbtack tosses

- **Hypotheses**: A flip is a Bernoulli distributed variable. Independence of flips.

derivative find 0 ?

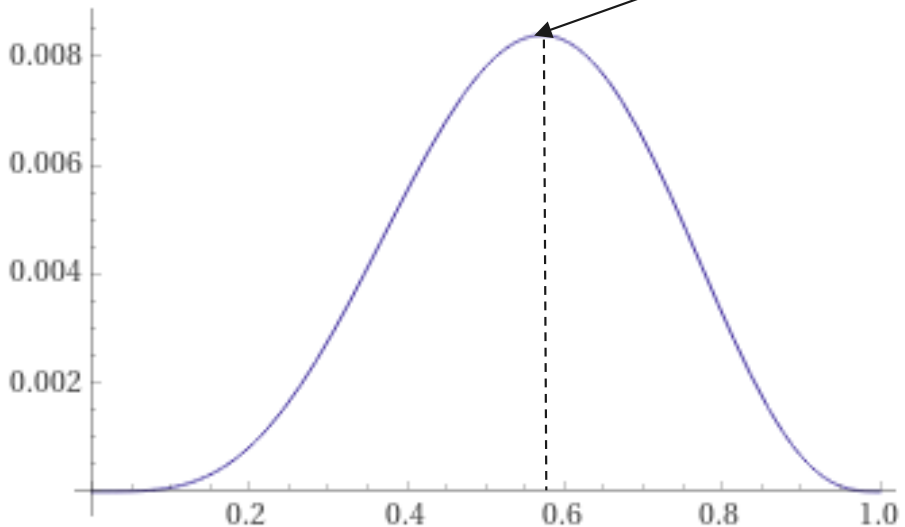
- **Learning**: Find p* that maximizes the data likelihood

$$\Pr(\mathcal{D}; p) = p^4(1-p)^3$$

Maximum Likelihood Estimate (MLE)

plot $p^4 (1-p)^3$ $p = 0$ to 1

Plot



Maximum Likelihood Estimate (MLE)

- Choose p that maximizes the likelihood

$$p^\star = \arg\max_p \Pr(\mathcal{D}; p)$$

$$= \arg\max_p \log \Pr(\mathcal{D}; p)$$

multiple of likelihood

same thing but less cost

Optimize to find p*

$$p^* = \arg\max_p \log\left(p^{a_H}(1-p)^{a_T}\right)$$

$$\frac{d}{dp}(*) = \frac{d}{dp}\left(a_H \log p + a_T \log(1-p)\right) =$$

$$= a_H \frac{1}{p} + a_T\left(-\frac{1}{1-p}\right) = 0 \Rightarrow$$

$$\Rightarrow a_H/p = a_T/1-p \Rightarrow p a_T = a_H(1-p) \Rightarrow$$

$$\Rightarrow p(\overbrace{a_H + a_T}^{n}) = a_H \Rightarrow \boxed{p_{MLE} = \frac{a_H}{n}}$$

Almost done:

- We need to verify that we get a maximum for p_{MLE}

(whiteboard)

$$p^* = \arg\max_p \log\left(p^H (1-p)^T\right)$$

$$\frac{d}{dp}\left(\log p^H + \log(1-p)^T\right)$$

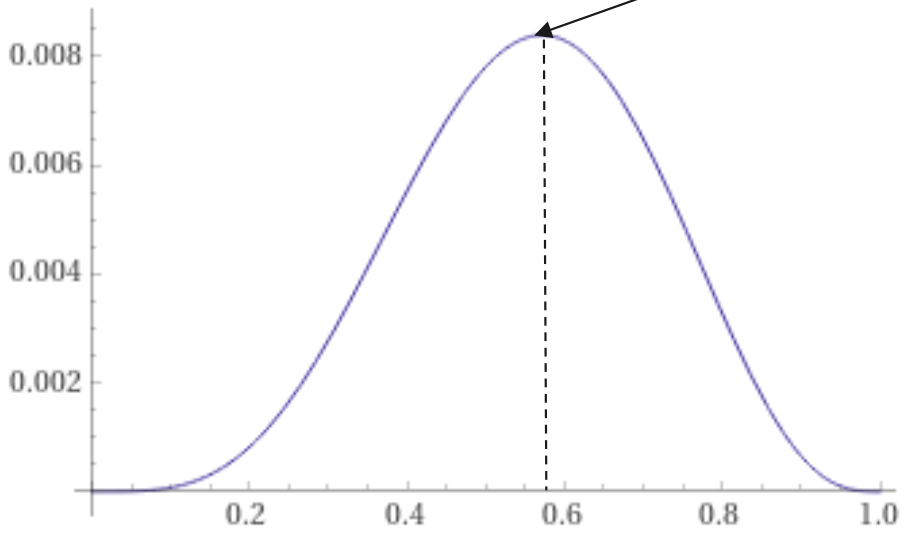
$$= \frac{d}{dp}\left(H \cdot \log p + T \cdot \log(1-p)\right)$$

$$= \frac{H}{p} + \frac{T}{p}\cdot(-1) = \frac{H(1-p) - Tp}{p(1-p)} = \frac{H - Hp - Tp}{p - p^2}$$

Maximum Likelihood Estimate (MLE)

plot	$p^4 (1-p)^3$	$p = 0$ to 1

Plot



$$p^\star = \frac{\#Heads}{n} = \frac{4}{7} = 0.5714$$

Confidence intervals (reminder)

- Boston billionaire says: I want to know the true p within 0.01 accuracy, with confidence at least 95%.

Sampling theorem: Given n independent 0-1 RVs X_i such that $\Pr(X_i=1)=p$ (i=1…n),

$$n \geq \frac{3}{\epsilon^2} \ln\left(\frac{2}{\delta}\right)$$

where then the following holds:

$$\Pr\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \leq \epsilon\right) \geq 1 - \delta$$

Two important properties of the MLE

- Consistent

$$\theta_{\text{MLE}} \rightarrow \theta_{\text{true}} \text{ in probability}$$

- Equivariant

$$\text{If } \theta_{\text{MLE}} \text{ is the MLE of } \theta_{\text{true}} \Rightarrow g(\theta_{\text{MLE}}) \text{is the MLE of } g(\theta_{\text{true}}$$

Billionaire with prior beliefs

- He says: Wait! I know that the thumbtack should be close to 50-50.

- You say: let's be Bayesian!

- Rather than learn a single value for p, we learn ***a probability distribution***

  p now becomes a random variable

Inference using Bayes' rule

Notice the notation Pr(D|p) instead of Pr(D;p)

$$Pr(p|D) = \frac{Pr(D|p)Pr(p)}{Pr(D)}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Does not depend on p.

Bayesian inference - Summary

1. We choose the prior distribution $f(\theta)$. This distribution expresses our prior beliefs on the parameter θ.

2. We choose the statistical model for the likelihood function $f(D|\theta)$.

3. After observing the data $D=X_1,..,X_n$, we update our beliefs and calculate the posterior distribution $f(\theta|D)$.

Maximum a posteriori (MAP) estimate is the mode of the posterior distribution (as we did in this lecture).

Important observation

- If we impose a uniform prior on p, then

$$Pr(p|D) \propto Pr(D|p)$$

- Image denoising lecture
  - Had we imposed a uniform prior on images x, then our MAP inference would be the same as the MLE
  - Choosing a good prior is important in applications of Bayesian inference

Conjugate priors

Definition

“ If the posterior distribution $p(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x \mid \theta)$.”

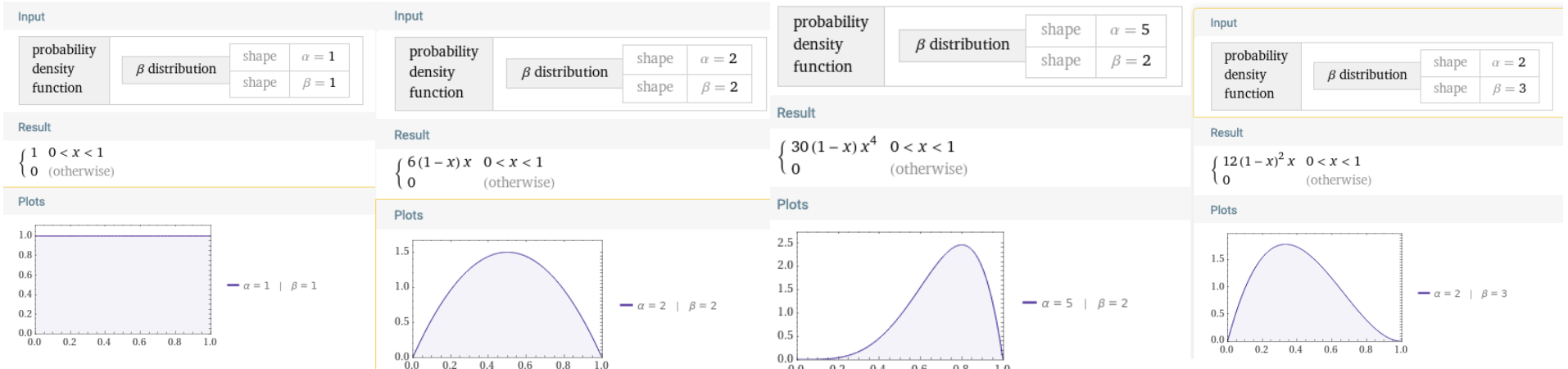
- For Binomial, conjugate prior is Beta distribution.

Bayesian inference for the thumbtack problem

- The probability density for beta distribution is

$$f(x;a,b) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)}, \quad 0 \le x \le 1, a, b > 0$$

Where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function



☆ understand this

Bayesian inference for the thumbtack problem

one hour

$$P(p) \propto p^{\beta_H - 1} (1-p)^{\beta_T - 1}$$ PRIOR

$$P(D|p) = p^{a_H} (1-p)^{a_T}$$ LIKELIHOOD.

How you get this from Beta

POSTERIOR

$$P(p|D) \propto P(D|p) P(p) \propto p^{a_H + \beta_H - 1} (1-p)^{a_T + \beta_T - 1}$$

$$\text{Beta}(a_H + \beta_H, a_T + \beta_T)$$

fictitious coin tosses reflecting our prior belief.

Posterior has same form as prior!

Method of moments

2nd moment is max

Suppose our model has parameters $\boldsymbol{\theta}=(\theta_1,..,\theta_k)$.

- Recall that the j-th moment of a RV X is $E[X^j]$. To denote that this is a function of the model, we write $E_{\boldsymbol{\theta}}[X^j]$.
  - Compute analytically $\alpha_j = \mathbb{E}_\theta\left[X^j\right], j = 1, \ldots, k$
- Consider the j-th *sample* moment for data $x_1,...,x_n$ is $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j, j = 1, \ldots, k$
- Equate the analytical moment expressions with the sample moments, and solve a system of k equations with k unknowns to learn $\boldsymbol{\theta}=(\theta_1,..,\theta_k)$.

Method of moments: example 1

Let $X_1,\ldots,X_n \sim \text{Bernoulli}(p)$.

- We have one parameter, so k=1.
- The first moment is the mean $\alpha_1 = E_p[X] = p$.
- The first sample moment is the sample mean.
- Thus we directly get $p_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Remark: Here, MoM is same as MLE, but this is not always the case!

Method of moments: example 2

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

$k=2$, so we need the first two moments!

$a_1 = \mu_1$ $\quad a_2 = \mu^2 + \sigma^2$

Our MoM estimators: $\hat{\mu}_{MoM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$ (1)

$E(X) = \mu$

$Var(X) = \sigma^2$

$Var(X) = E(X^2) - (E(X))^2$

$E(X^2) = \sigma^2 + \mu^2$

$\hat{\mu}^2_{MoM} + \hat{\sigma}^2_{MoM} = \frac{1}{n}\sum X_i^2$ (2)

Solving for $\hat{\sigma}^2_{MoM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

MLE for Gaussian

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$. The likelihood function

(ignoring the constants $\sqrt{2\pi}$)

$$\mathcal{L}_n(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \to = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma}\right\} \exp.\left\{-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}\right\}.$$

→ we used the FACT $\sum_{i=1}^{n}(X_i-\mu)^2 = \sum(X_i-\bar{X}+\bar{X}-\mu)^2 = nS^2 + n(\bar{X}-\mu)^2$

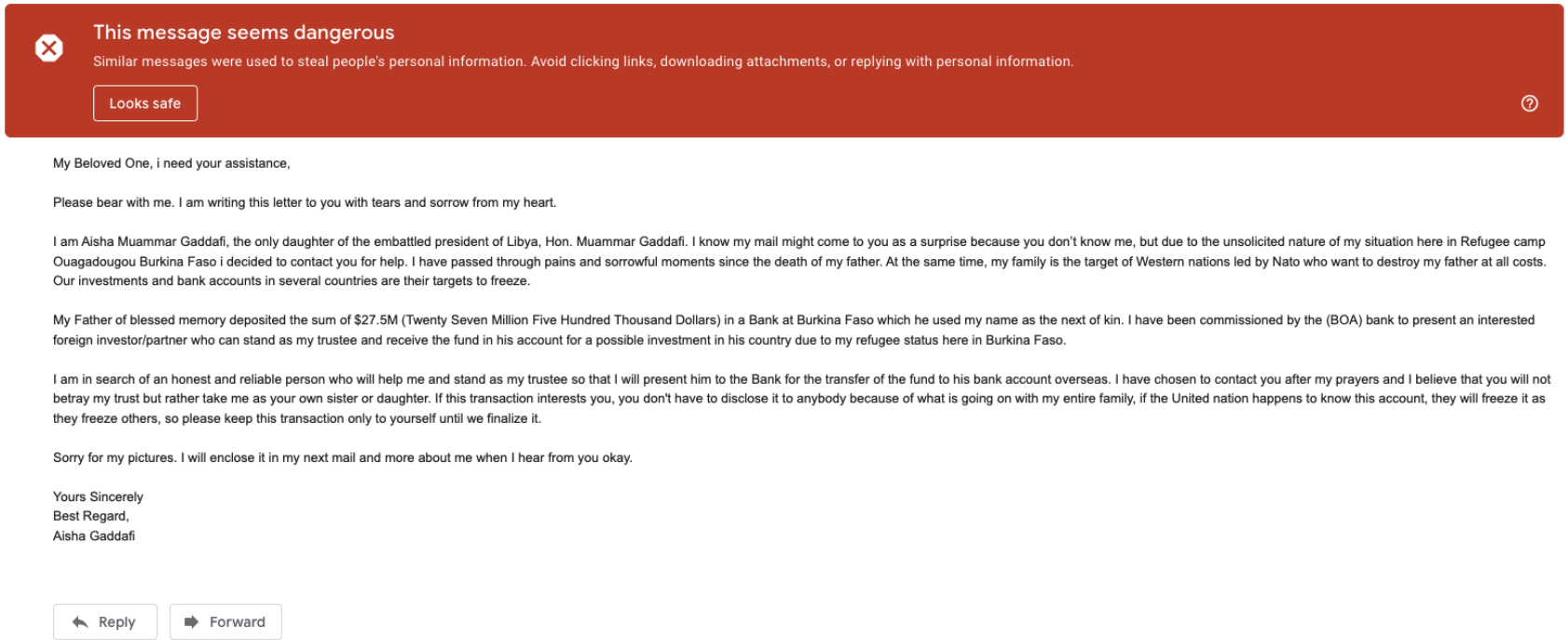
where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

The log-likelihood is $\ell(\mu,\sigma) = -n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}$.

Solving $\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = 0$, $\frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = 0 \Rightarrow \mu_{MLE} = \bar{X}$, $\sigma_{MLE} = S$

Spam or Ham?

- Real-world problem: how can we classify an email as spam?



Spam classifier

- Consider the email as a collection of words $w_1, w_2, .., w_n$
  - Certain words that appear often in all emails, e.g., “the”, can be removed from the email.
- Formulate mathematically our problem:
  We are interested in posterior probability $\Pr(\text{spam} | w_1, w_2, .., w_n)$
- Classes ={spam, not spam}
  - Suppose an email is equally likely to be spam or non-spam.
- We apply Bayes’ rule

$$\Pr(\text{spam} \mid w_1, \ldots, w_n) = \frac{\Pr(w_1, \ldots, w_n \mid \text{spam}) \Pr(\text{spam})}{\Pr(w_1, \ldots, w_n \mid \text{spam}) \Pr(\text{spam}) + \Pr(w_1, \ldots, w_n \mid \text{not spam}) \Pr(\text{not spam})}$$

Bayes Classifier

- More generally suppose we have k classes, $\{c_1,...c_k\}$ with a given prior $Pr(C=c_j)=p_j$
  - In our example k=2
- We have some data D
  - In our example, the set of words in the e-mail
- Suppose we somehow know exactly $Pr(C=c_i \mid D)$ for each class c_i.
- What class would you assign to D?

$$c^\star = h_{\text{Bayes}}(\mathcal{D}) = \arg\max_i \Pr(C = i \mid \mathcal{D})$$

Naive Bayes Classifier

- A popular classifier is known as Naive Bayes and makes the following *conditional* independence assumption:

$$\begin{aligned}\Pr(w_1, \ldots, w_n \mid \text{spam}) &= \Pr(w_1 \mid \text{spam}) \cdot \ldots \cdot \Pr(w_n \mid \text{spam}) \\ \Pr(w_1, \ldots, w_n \mid \text{not spam}) &= \Pr(w_1 \mid \text{not spam}) \cdot \ldots \cdot \Pr(w_n \mid \text{not spam})\end{aligned}$$

  namely words (attributes/features) are conditionally independent given the class.

- Decision rule: Output the class c that maximizes the posterior probability

$$c^{\star}_{\text{Naive-Bayes}} = \arg\max_{c} \; \Pr(c) \prod_{i=1}^{n} \Pr(w_i \mid c)$$

Naive Bayes Classifier

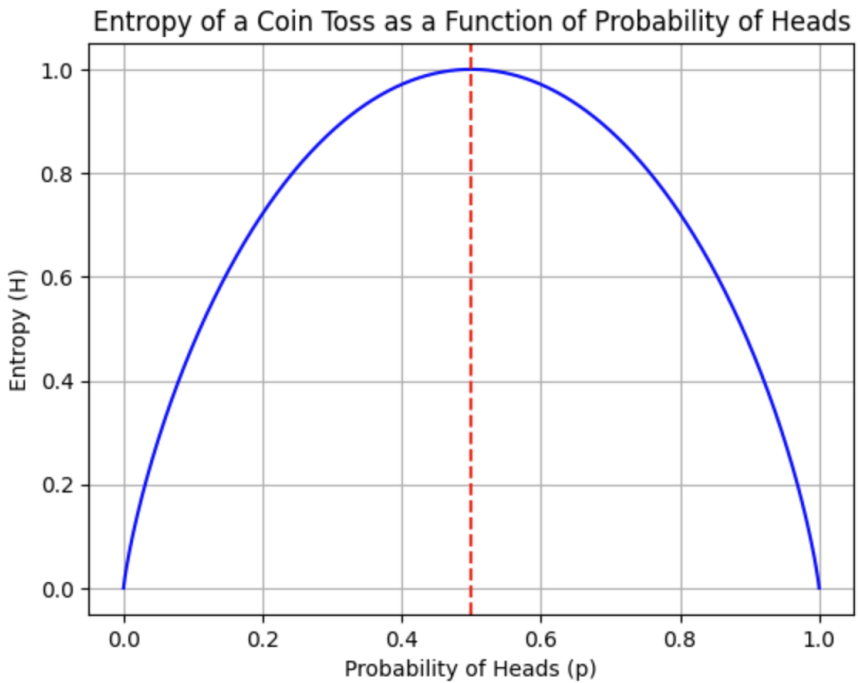
- Suppose we have access to a training set, namely a set of emails that are associated with a label {spam, non-spam}.
  - How can we use this information to classify an unseen email?
- We learn the probabilities of Naive Bayes classifier from the training data.

Prior: $\Pr(\text{spam}) = \frac{\#\text{spam emails}}{\#\text{emails}}, \ \Pr(\text{not spam}) = \frac{\#\text{not spam emails}}{\#\text{emails}} = 1 - \Pr(\text{spam})$

Likelihood: $\Pr(Word = w \mid C = c) = \frac{\#(Word = w, Class = c)}{\#\text{emails of Class c}}$

Coin toss entropy (plotting code)

- H(p)=-p log(p)- (1-p) log(1-p) for 0<p<1
- H(0)=H(1)=0



How Shannon Entropy Imposes Fundamental Limits on Communication

- The value of the random variable X is presented to A who has to describe it to B
- A and B decide on encoding of the range of (X)
- H[X] is the average number of bits A needs to send to B to reveal the value of X under the best encoding.

Joint Entropy

- For two random variables X,Y the joint entropy is defined

$$H[X,Y] = -\sum_{x \in range(X),\, y \in range(Y)} \Pr(X=x,\, Y=y) \log_2 \Pr(X=x, Y=y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$$

Exercise: compute the joint entropy for the following X,Y

	Y=1	**Y=2**
X=0	¼ (i.e. Pr(X=0,Y=1)=¼)	1/2
X=10	0	1/4

Facts about Entropy

- Conditional entropy of X given Y: $H[X \mid Y] = \mathbb{E}[H[X_Y]]$

  where X_y is a random variable such that

  $$\Pr(X_y = x) = \Pr(X = x \mid Y = y)$$

- H[X,Y]=H[X] + H[Y|X]
- H[X,Y|Z]=H[X|Z] + H[Y|X,Z]

How do we generalize this to n variables?

Facts about Entropy

Consider a random vector (X_1,...,X_n). Then,

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$$

Entropy subadditivity: $H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$

Facts about Entropy

- $H[X] \leq \log_2 |range(X)|$ With equality when X is uniform
- $H[X] \geq H[X|Y]$ (information never hurts)

lower the better.

- Let X be a random variable and let g(X) be some deterministic function of X.

$$H[X] \geq H[g(X)]$$

Equality holds iff g is invertible.

Readings

Reading the material is essential, while the slides are only supplementary. Please refer to the website for the assigned readings and GitHub for the Jupyter notebook, which should be reviewed thoroughly!