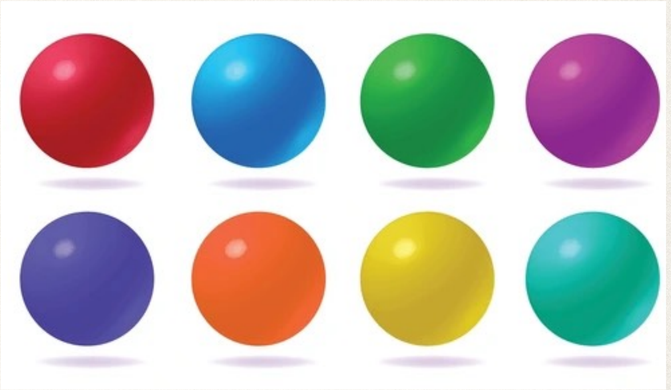
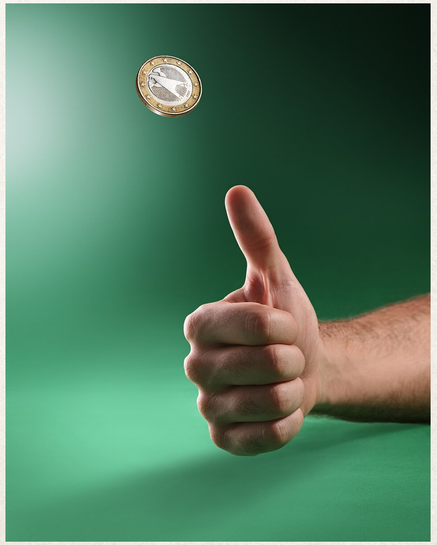
Introduction to Statistics I

Atreyee Sinha

Plan: Recap to probability and stats

- Using basic examples, we will go into
  - Bayesian vs Frequentist
  - Maximum likelihoods and parameter estimations
  - Priors, posteriors
  - Common pdfs
  - Central limit theorem





Performing an experiment

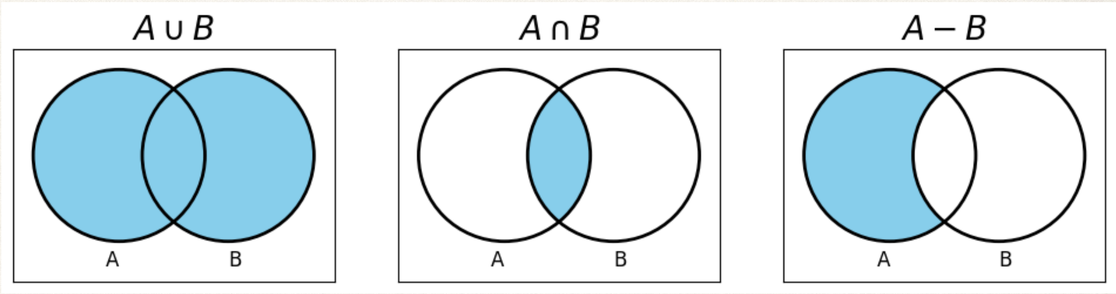
- Experiment design is crucial to drawing statistical inferences
  - Define your experiment and the hypotheses clearly
    - Result of my coin toss is HHHT: Is it a biased coin?
- Determine:
  - Values of certain parameters
  - Check “how well” a parameter matches our hypothesis
- A TED talk on probability in practice: https://www.youtube.com/watch?v=kLmzxmRcUTo

What is a statistic?

- Loosely: “Anything that can be computed from the data”
  - Average of N rolls of a dice
  - Number of time a “6” occurred
  - Probability of getting a 6 : this we have to *estimate*
  - The number of hits on a GM counter with time
- Point statistics: A single value computed from the data
- Statistical intervals: A range of values - how to interpret it?

Recap to probabilities

- If Ω is the sample space (set of all possible outcomes) and A is an event, the probability $P(A)$ must satisfy:
  - $P(A) >= 0$
  - $P(\Omega) = 1$
  - $P(A\ U\ B) = P(A) + P(B) - P(A \cap B)$

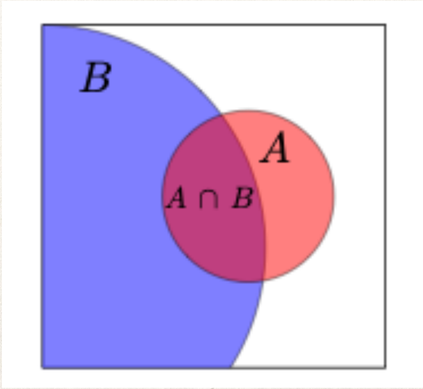


Conditional probability

- Toss a fair coin 3 times.

$$\Omega = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$$

  - What's the probability of getting HHH?
    - Ans: 1/8
  - I know first coin is H. What's is the prob of HHH?
    - Ans: 1/4



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Bayes theorem

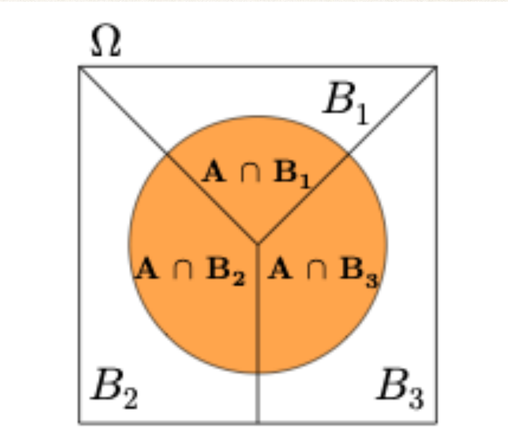
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + P(A \mid B_3)P(B_3)$$

$$P(A) = \Sigma_i P(A \mid B_i)P(B_i)$$

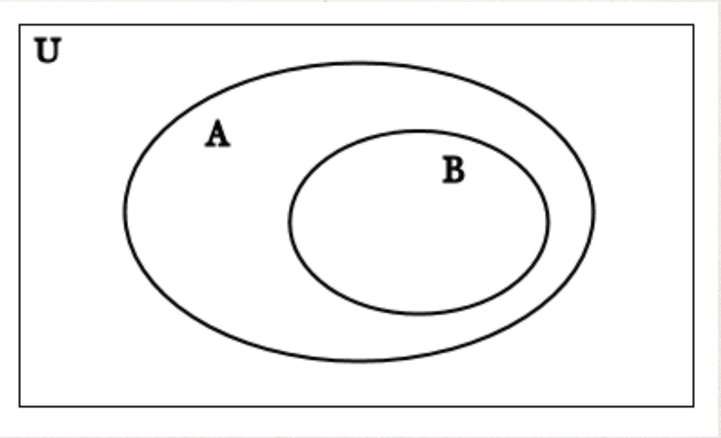
Independent events: $P(A \cap B) = P(A) * P(B)$

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A}$$: Bayes theorem



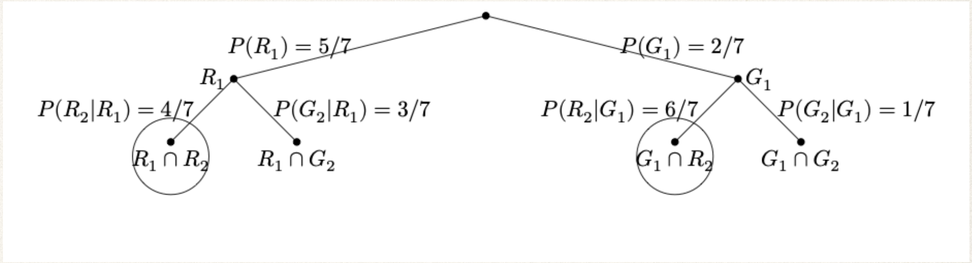
Some examples

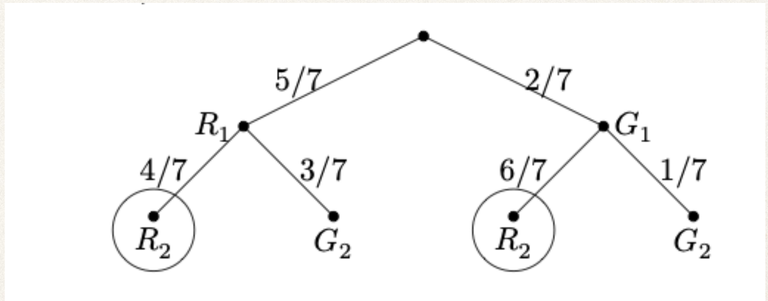
1. Independent events:
   - Two people in an office will have the same genetic disease?
   - Two people in a family will have the same genetic disease?
2. Conditional probability:
   - $P(A \mid B) \neq P(B \mid A)$
   - $\Omega =$ All animals in Kaziranga forest;
   - A: All animals weighing > 500kg in Ω; B: All elephants in Ω
   - P(A | B) ?
   - P (B | A) ?



Probability trees

- An urn contains 5 red balls and 2 green balls. A ball is drawn. If it's green, a red ball is added to the urn and if it's red a green ball is added to the urn. The original ball is not returned. Then a second ball is drawn. What is the probability that the second ball is red?

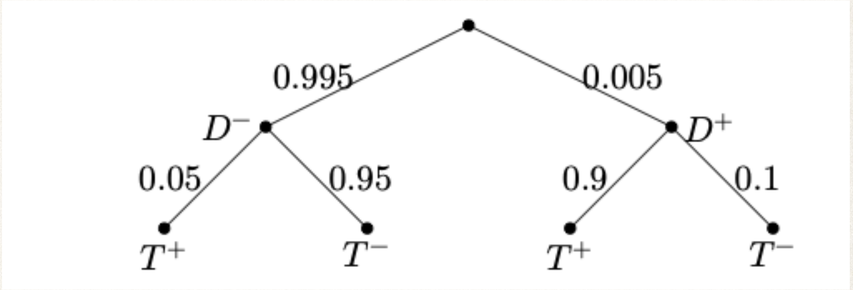




$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = 4/7 * 5/7 + 6/7 * 2/7 = 32/49$$

Base rate fallacy

- Prob of an event: 1 in 200
- You devise a test with 5% false positive, 10% false negative
- A test gives +ve result: What is the probability that the event occurred?
- (Event: disease/ particle id/ bouncing cheques…)



$$P(D^+ \mid T^+) = ?$$

Frequentist vs Bayesian interpretations

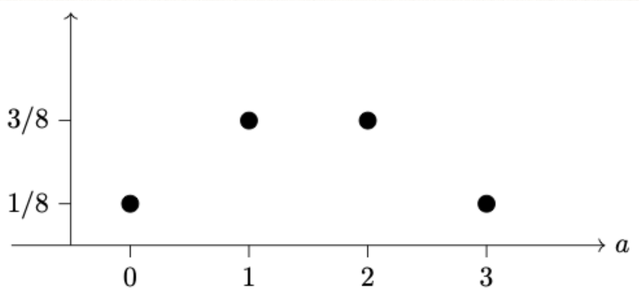
- Frequentist:
  - Parameters: Fixed and unknown
  - Data: Repeatable random sample
  - P(x) ->limit of a frequency of occurrence over many trials.
  - P-values, confidence intervals, null hypothesis, etc
  - No assumption of prior knowledge
- Bayesian:
  - Parameters: Random variable with probability distributions
  - Data: Fixed observed evidence used to update the "prior" belief to a "posterior" belief.
  - P(x) -> Measure of a certainty of a parameter
  - Prior distribution, likelihood, posterior distribution, Bayesian inference

Discrete random variables

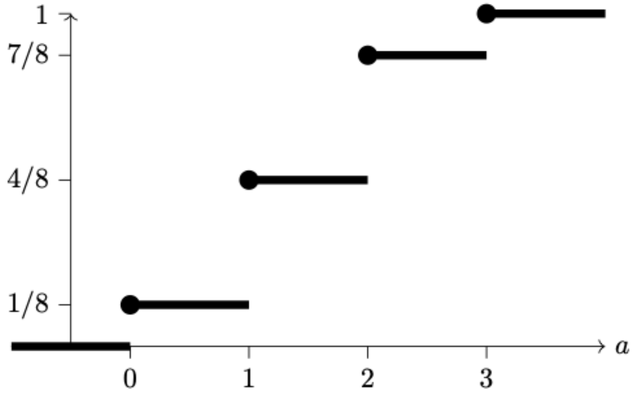
- A random variable assigns a number to each outcome in a sample space. Let Ω be a sample space. A discrete random variable is a function

$$X : \Omega -> \mathscr{R}$$

- Cumulative distribution function: $F(a) = P(X <= a)$
  - F is monotonically increasing
  - $0 <= F(a) <= 1$
  - $F(a) = 1 \ (a -> \infty); \ F(a) = 0 \ (a -> -\infty)$



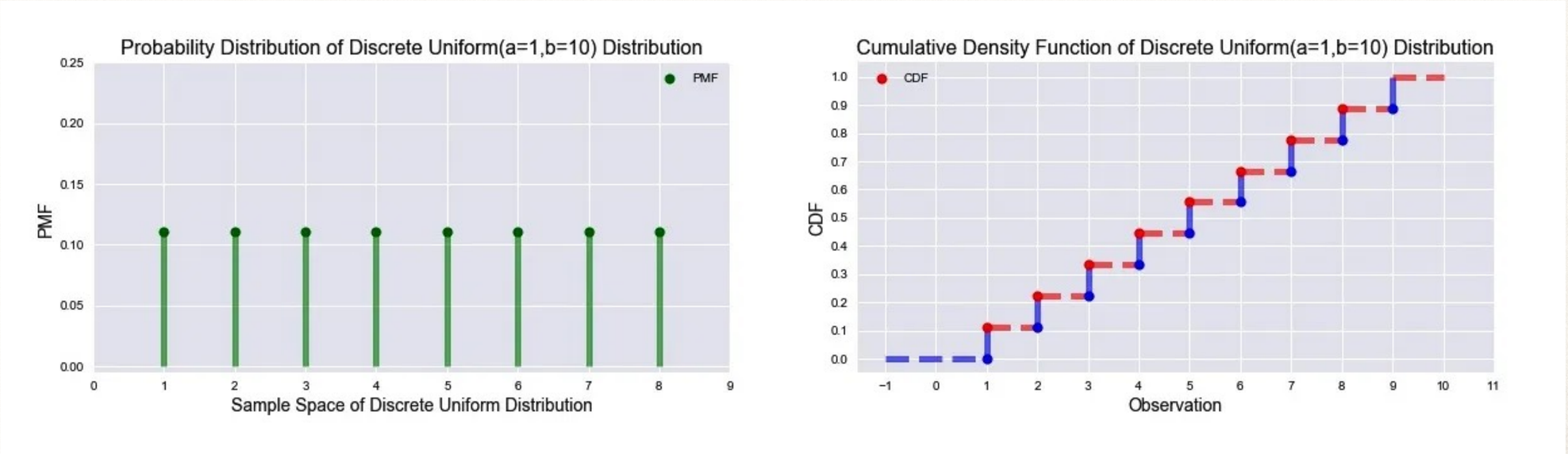
No. Of heads in 3 tosses of fair coin



Some discrete probability functions

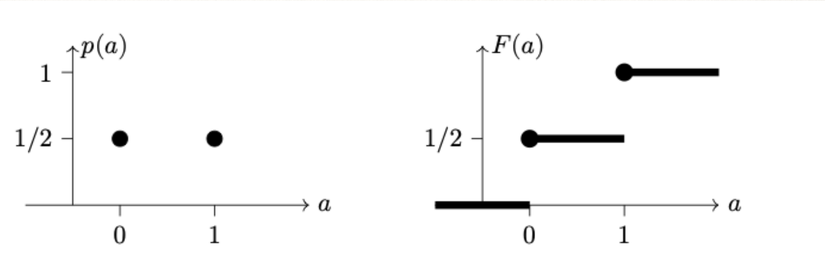
Uniform Distribution

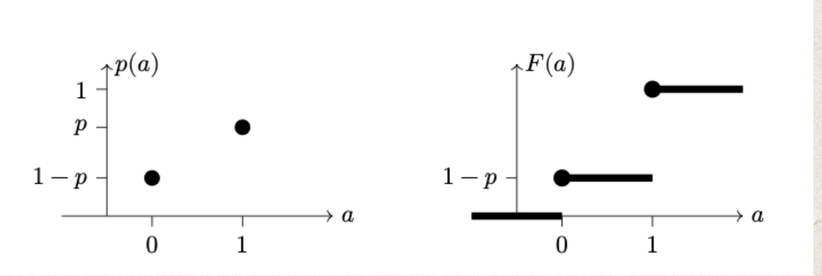
- All outcomes are equally likely within a certain range
- Eg: Roll of unbiased dice, birthday(?)



Bernoulli Distribution

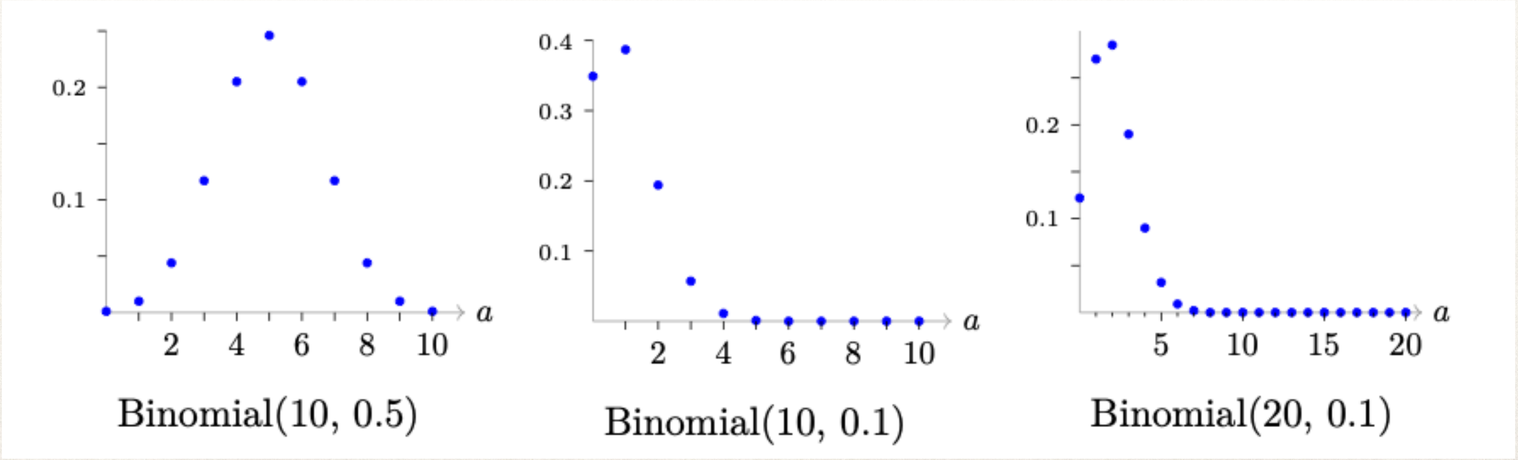
- Bernoulli distribution models **<u>one trial</u>** in an experiment that can result in either *<u>success or failure.</u>* This is the most important distribution and is also the simplest. A random variable X has a Bernoulli distribution with parameter p if:
  - $X \in T, F$
  - $P(X = T) = p; P(X = F) = 1 - p$
- Eg: coin flips, passing a threshold, votes for or against





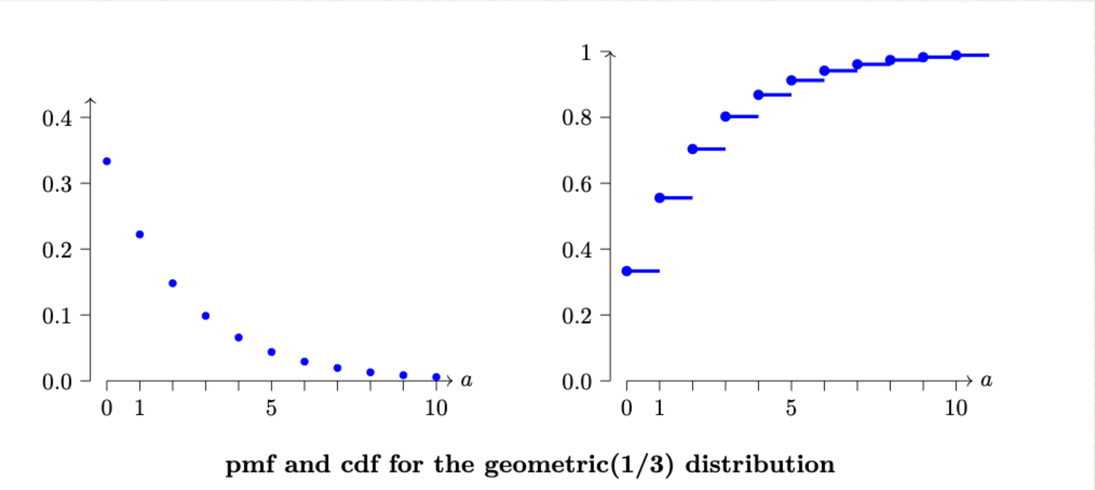
Binomial Distribution

- The binomial distribution Binomial(n,p), or Bin(n,p), models the number of successes in n independent Bernoulli(p) trials.
- Eg: No. Of heads in n flips of a coin with probability p of heads Bin(n,p)
- $P(n;k,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$



Geometric Distribution

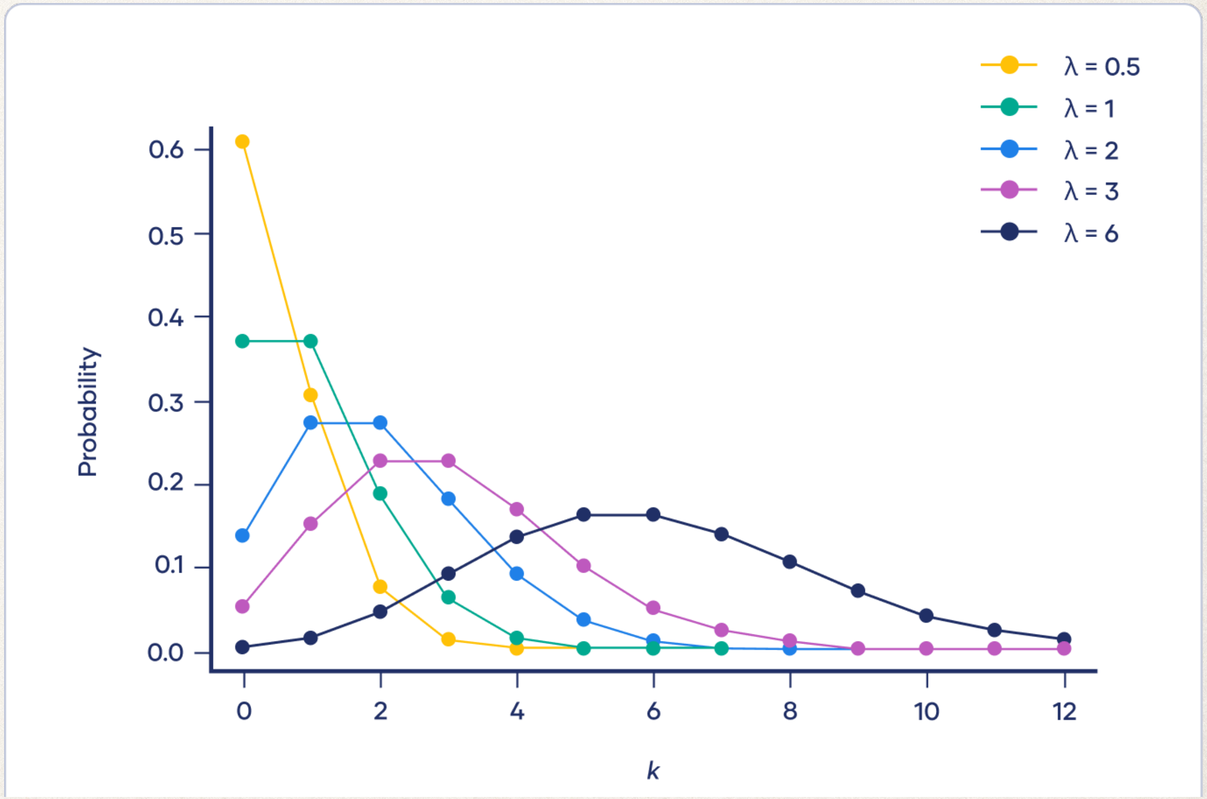
- A geometric distribution models the number of events before the first occurrence of a specific event
  - Eg: No of F before a T (coin toss, running an expt)
  - $X \in 0,1,2,3,\ldots$
  - $p(k) = P(X = k) = (1-p)^k p$
  - Think: difference from prev



pmf and cdf for the geometric(1/3) distribution

Poisson distribution

- For large n and small p, the binomial distribution reduces to
  - $P() = \frac{\lambda^k}{k!} e^{-\lambda}; \lambda = n * p$



Expectation values

- $E[X] = \Sigma_{j=1}^{n} p(x_j) x_j$
- $E[x + Y] = E[X] + E[Y]$
- $E[aX] = aE[X]$
- Compute E[X] for the distributions discussed before

Variance and standard deviation

- $Var(X) = E[(X - \mu)^2] = \Sigma_{j=1}^{n} p(x_j)(x_i - \mu)^2$
- $\sigma = \sqrt{(Var(X)}$
- If X and Y are independent; Var(X + Y) = Var(Y) + Var(Y)
- $Var(aX + b) = a^2Var(X)$
- $Var(X) = E[X^2] - E[X]^2$

Table of expectation and variance

Distribution	range X	pmf $p(x)$	mean $E[X]$	variance $\mathrm{Var}(X)$
Bernoulli(p)	0, 1	$p(0) = 1-p, \quad p(1) = p$	p	$p(1-p)$
Binomial(n, p)	0, 1,..., n	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Uniform(n)	1, 2, ..., n	$p(k) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Geometric(p)	0, 1, 2,...	$p(k) = p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Questions

1. In detector systems, *n-fold coincidences* are often used to suppress random triggers caused by noise or stray light. Consider an N×N array of pixels. During a fixed time window, each pixel independently produces a random trigger with probability 'p'. What is the probability that **three nearest-neighbour pixels** trigger within the same time window purely due to random noise?
   - What are the assumptions here?
   - As a first approach, simply this in 1D as a coin toss problem
2. Simulate N =1000 random points uniformly from a square of side 2 units. What's the number of points satisfying $x^2 + y2 <= 1$. From this, estimate:
   i. the probability of a point lying inside the circle,
   ii. the area of the circle,
   iii. and the value of π
   iv. Try for different values of N and compare the convergence in (iii). What happens when you extend to n-dim?
3. Social evil: Families keep having kids until they have a male child. How many children are expected per family?

1. Suppose X is a random variable with the following pmf:

   - X: 1, 2, 3; pmf: 1/4, 1/2, 1/4

   - Find E[X], E[1/X], E[X*X]

2. A common challenge in gamma ray astronomy is to separate photons from hadrons. (Similar particle identification scenarios occur in other cases of particle physics). Lets say you developed a fancy classifier which correctly reconstructs gammas 95% of the time, and same for protons. 1 in 10,000 triggers corresponds to a gamma, rest are protons. If your classifier reconstructs an event as a gamma-ray, what's the prob it is a gamma-ray? Use Bayes theorem and draw the prob tree