Introduction to Probability

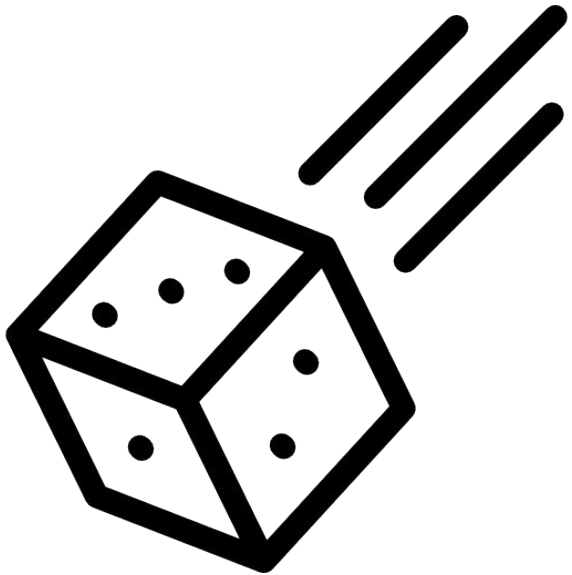
Group assignment 1

Questions

- What are Random variables and Random functions?
- What are probability axioms?
- What are basic probability rules?
- What is the advantage of representing data using probability distributions?

Just imagine ...

- Consider a regular dice.
- Now throw the dice ‘n’ times.
  (let $n \in \mathbb{N}$)
- Let the outcome of the i^{th} throw be represented as X_i .
  (let $i \in \mathbb{N}$ and $i \leq n$)
- So now the outcomes of ‘n’ throws can be listed down as :
  $\{ X_1, X_2, X_3, ... X_n \}$



Dice icon made by Freepik from Flaticon.

Let's formulate 'random variables'...

- Now each and every time we threw the dice, we basically performed an experiment. Hence we can say that, i^{th} throw of dice = i^{th} experiment.
- Now the outcome of i^{th} throw was expressed as X_i. X_i is what we mathematically call as a random variable.
- Simply put, "*A random variable is simply an expression whose value is the outcome of a particular experiment*".
- Mathematically put, "*A random variable X on a sample space S is a function X : S → $\mathfrak{R}$ that assigns a real number X(s) to each sample point s ∈ S*".

Let's formulate 'random functions' ...

- Now, our set of 'n' throws basically form a collection of random variables.
- And if we index these random variables by a parameter (such as time), then this collection of 'n' throws basically becomes a stochastic process.
- A stochastic process is also known as a random process or a random function.
- Mathematically put, "*A stochastic process is a family of random variables* $\{ X(t) \mid t \in T \}$*, defined on a given probability space, indexed by the parameter t, where t varies over an index set T*".

Axioms of Probability

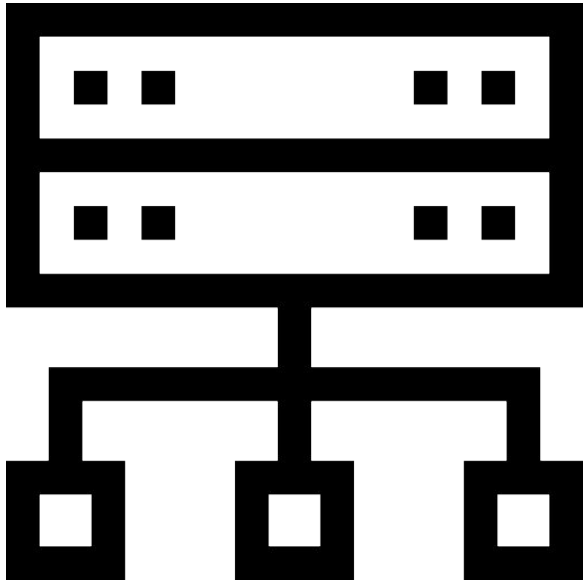
- Also known as **Kolmogorov's axioms**.
- For a sample space S and events A & B (such that $A, B \subseteq S$), a probability function P(.) should satisfy the following axioms:
  a. $P(E) \geq 0$
  b. $P(S) = 1$
  c. $P(A \cup B) = P(A) + P(B)$, provided A and B are mutually exclusive events (i.e., when $A \cap B = \varnothing$)

Basic Rules of Probability

- Commutative laws:
  - $A \cup B = B \cup A \quad A \cap B = B \cap A.$
- Associative laws:
  - $A \cup B = B \cup A \quad A \cap B = B \cap A.$
- Distributive laws:
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$
- Identity laws:
  - $A \cup \varnothing = A \quad A \cap S = A.$
- Complementation laws:
  - $A \cup A = S \quad A \cap A = \varnothing.$

Imagine a web server ..

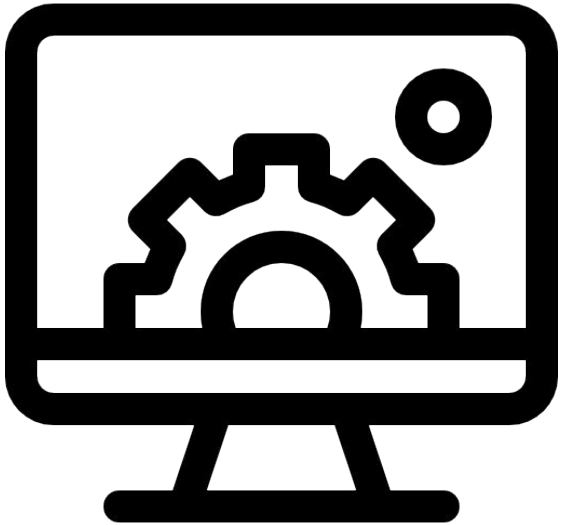
- Assume that you have data for the incoming requests for a particular time duration from your past deployments.
- Now how do you re-organise the resources of your web server to fulfill the maximum no. of requests possible?



Server icon made by Freepik from Flaticon.

Imagine a software ...

- Now let's say that it has been deployed in the production environment and you have accumulated data regarding the issues it might face.
- Now you need to determine the fault-tolerance capability of your software in order to decide whether or not it needs improvement. So how will you do it?



CMS icon made by Freepik from Flaticon.

Data and Probability Distributions

- What are the common factors amongst both the problems?
  - Probability and data
- Now if you can make use of both of them, then you can, to some degree predict the behaviour of your environment. So, how do you make use of both of them?
  - The answer lies in creation of Probability Model.
  - Simply put, it's association of data with probability.
  - And to exaggerate its benefits, you can simply assume that *you can predict the future of your system!*

Questions

- How you will find out distributions of X1 + X2, X1 - X2, X1/X2, if X1 and X2 are random variables
- How you will find out distributions of f(X) where X is random variable
- What is expected value and variance? How will you find out for any distribution?

Expected value

- Dealing with large quantity of numbers (Sample space) is difficult.
- We often like to find out its mean or average for easy understanding.

For example: A game of dice where we get profit for even numbers and loss at occurence of odd numbers. I.e +2,+4,+6 and -1,-3,-5.

Now in this problem we are much more interested in the net gain at the end of say 100 or 1000 plays or trials. This is where expected value comes in.

- The game appears to be favourable towards player.
- We can also see that relative frequency is approximately equal to the probability of each number occurring = 0.16.
- Therefore E = -1 * ⅙ +2 * ⅙ -3 * ⅙ +4 * ⅙ - 5 * ⅙ + 6 * ⅙ = + 0.50
- For large values of n our E (Expected value) and our calculated value appears to be same

Hence Expected value can be given as summation of each event occurring multiplied by its probability of occuring.

$E(X) = \sum x.p(x)$ or $\int x.p(x)\,dx$

Where p(x) is distribution function

	n=100		n=1000	
Winnings	Frequency	Relative Frequency	Frequency	Relative Frequency
-1	17	0.17	1678	0.1678
+2	16	0.16	1681	0.1681
-3	17	0.17	1626	0.1626
+4	18	0.18	1686	0.1686
-5	16	0.16	1696	0.1696
+6	16	0.16	1633	0.1633
Total	+48	0.48	+4868	0.49

Variance

- In ideal case we would like our expected value to not deviate from our absolute value.
- The measure to this deviation is nothing but variance.
- And just like we find out the error :

$$V(X) = E\left((x-\mu)^2\right) \quad \text{where } \mu = E(X)$$

$$= \sum (x-\mu)^2 . p(x) \text{ or } \int (x-\mu)^2 . p(x)\, dx$$

Joint Cumulative Distribution

- In our problem we have X_1 and X_2 as random variables.
- Let X_1 and X_2 have a joint distribution function $f_{X1,X2}(x1, x2)$.
- We introduce a new random variable Z.
  1. $Z = X_1 + X_2$
     Now we can find the distribution of
     $F(Z) = P(X_1 + X_2) = \iint_{x1+x2} f_{X1,X2}(x_1,x_2)\, dx_1 dx_2$

  2. $Z = X_1 - X_2$
     Now we can find the distribution of
     $F(Z) = P(X_1 - X_2) = \iint_{x1-x2} f_{X1,X2}(x_1,x_2)\, dx_1 dx_2$

Conditional Distribution

Let X_1 and X_2 be continuous random variables with joint pdf f(x, y). The conditional density $f(X_1 \mid X_2)$ is defined by

$f_{X1|X2}(x_1|x_2) = f(x_1, x_2) / f_X(x)$, if $0 < f_X(x) < \infty$.

Derivation :

$f_{x1|x2}(x_1|x_2)$

$= f_{X1,X2}(x_1,x_2)f_{X2}(x_2)$

$= f_{X1,X2}(x_1,x_2)\int_{\infty-\infty} f_{X1,X2}(x_1,x_2)dx_1 f_{X2|X1}(x_2|x_1)$

$= f_{X1,X2}(x_1,x_2)f_{X1}(x_1)=f_{X1,X2}(x_1,x_2)\int_{\infty-\infty} f_{X1,X2}(x_1,x_2)dx_2.$

By rewriting the above as $\boxed{f_{x1}|_{x2}(x_1|x_2) = f_{X2}|_{X1}(x_2|x_1)f_{X1}(x1)\int_{\infty-\infty} f_{X2}|_{X1}(x_2|x_1)f_{X1}(x_1)dx}$

Probability Distribution Function

- Till now we have learned about random variables and how to find the probability at occurence of specific event.
- Now we will learn about computing probability of set $\{ s \mid X(s) \in A\}$ where A is subset of R with endpoints (a,b), $-\infty < a < b < \infty$.
- Let f(x) be the probability mass function of random variable X then $P(X \in A) = \sum_{xi \in A} f_x(x_i)$
- Let us take an example

- A dice is thrown 2 times. Probability of sum of number on two attempts is < or = 5 is?
- Total possibilities = 36
- We know that minimum sum = 2
- Therefore P(2<=X<=5) = P(X=2) + P(X=3) + P(X=4) + P(X=5) = 10/36

Question

What is the utility of normal distribution?

Explain central limit theorem.

What is joint distribution and marginal distribution?

Utility of normal distribution

Whenever we measure things like :

- people’s height, weight, salary, opinions or votes.
- IQ of particular population
- Technical stock market
- Distribution of income among rich and poor community.
- Average shoe size based on country, gender etc
- Birth weight
- Student’s Average report

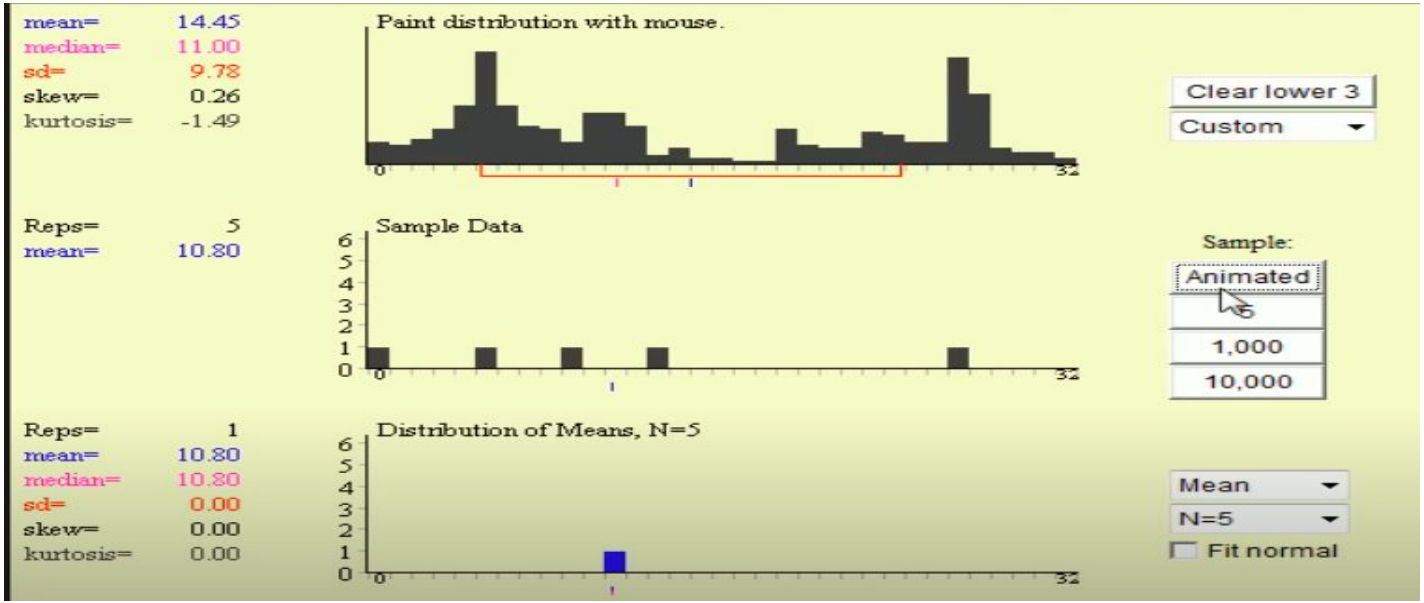
Central Limit Theorem

- It states that under certain (fairly common) conditions, the sum of many random variables will have an approximately normal distribution.

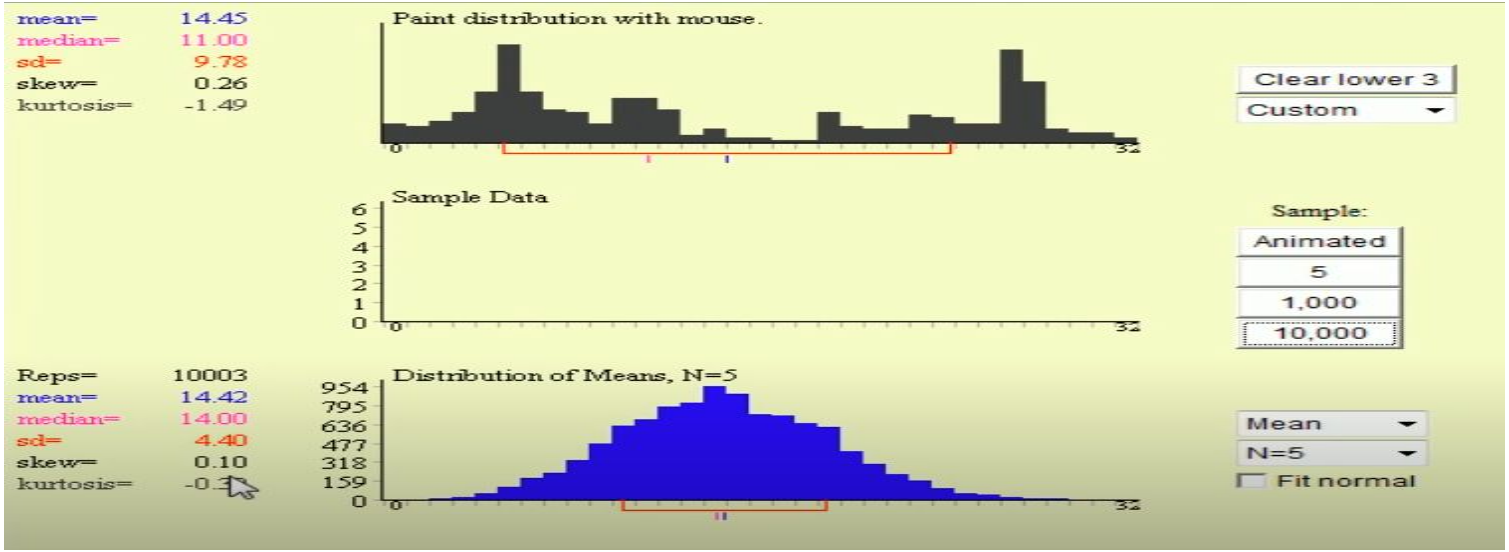
$$Z = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

- Even if the original variables themselves are not normally distributed, sum of independent random variables will tends toward a normal distribution.

- Taking sample space N=5



- Taking sample space N=10000



- As N increases, the probability distribution of Z will tend to normal distribution (a bell curve).

Joint Distribution

- Often, we need to consider the relationship between two or more events.
- Joint distributions allow us to reason about the relationship between multiple events.
- If X and y are discrete random variables, the function is given by

$$f(x,y) = P(X = x, Y = y)$$

for each pair of values(x,y) within range of X.

- Example:


Male Female TOTAL
Game of thrones 80 120 200
West World 100 25 125
Other 50 125 175
TOTAL 230 270 500
Male Female TOTAL
Game of thrones 0.16 0.24 0.4
West World 0.2 0.05 0.25
Other 0.1 0.25 0.35
TOTAL 0.46 0.54 1
Joint probability distribution
Sums to 1

Marginal Distribution

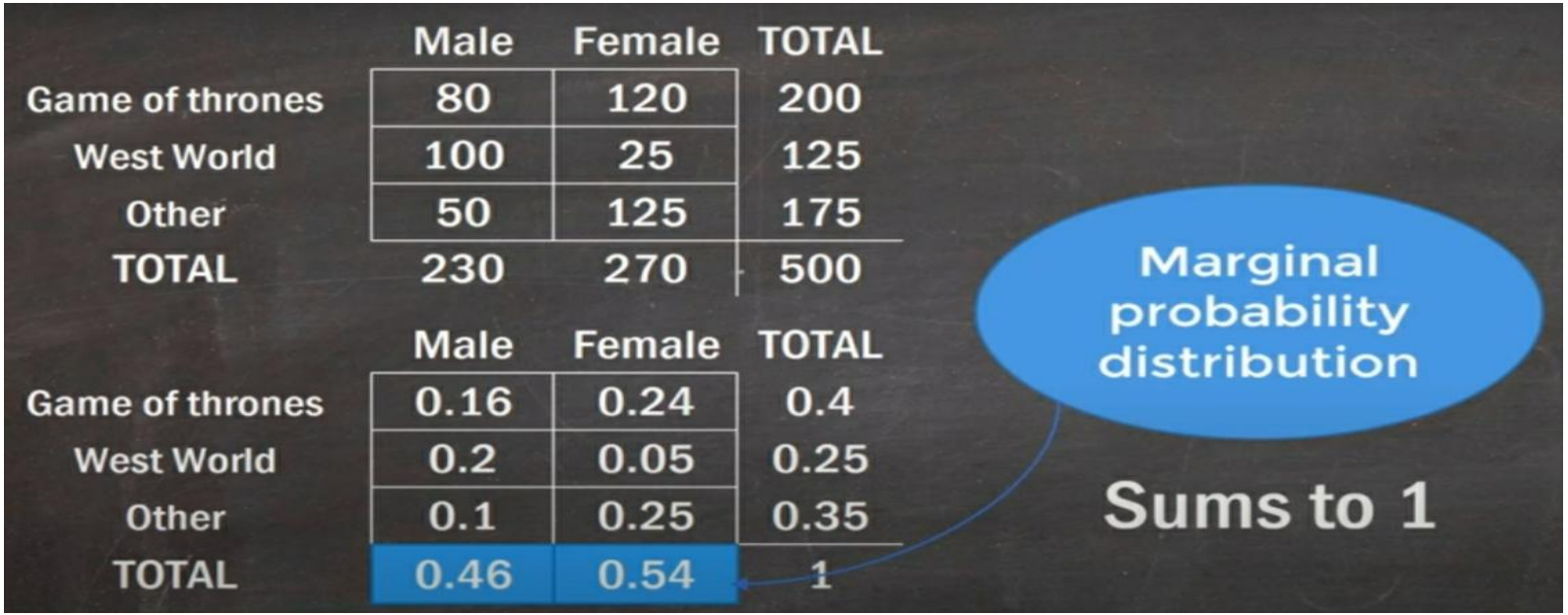
Marginalisation refers to the process of ‘removing’ the influence of one or more events from a probability.

Definition:

If X and Y are discrete random variable and f(x,y) is the value of the joint probability distribution at (x,y), the functions are given by:

$$g(x) = \sum_{y} f(x,y) \quad \text{and} \quad h(y) = \sum_{x} f(x,y)$$

Example:


Male Female TOTAL
Game of thrones 80 120 200
West World 100 25 125
Other 50 125 175
TOTAL 230 270 500
Male Female TOTAL
Game of thrones 0.16 0.24 0.4
West World 0.2 0.05 0.25
Other 0.1 0.25 0.35
TOTAL 0.46 0.54 1
Marginal probability distribution
Sums to 1

Questions

- What is the concept of PGF? What is the practical utility?
- Which are common distributions , both discrete and continuous?
- How you will estimate parameters of distributions from the data?

Probability Generating Functions:

- Let X be a random variable defined over the non-negative integers. The probability generating function (PGF) is given by the polynomial.

  $G_x(t) = p_0 + p_1 t + p_2 t^2 + \ldots = \sum_{x=0}^{\infty} P(X=x) t^x$

- Example:

  Different outcomes of throwing a fair die would be {1, 2, 3, 4, 5, 6} and probability associated with each outcome is 1/6. We can define PGF for the experiment as:

  $G(t) = 0t^0 + 1/6\ t^1 + 1/6\ t^2 + 1/6\ t^3 + 1/6\ t^4 + 1/6\ t^5 + 1/6\ t^6$

Properties of PGF:

- $G(0) = P(X = 0)$

- $G_x(1) = P(X=0) + P(X=1) + P(X=2) + \ldots = 1$
  In the PGF of the fair dice we can see if t = 1,
  $G_x(t) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$

- $G(t) = E(t^X)$
  we know that, $G_X(t) \sum_r t^r.P(X = r)$
  The general formula for the expectation of an arbitrary function of a random variable.
  $E(f(X)) = \sum_r f(r).P(X = r)$

taking f(X) as t^X

$E(t^X) = \sum_r t^r .P(X = r) = G_X(t)$

(The crucial point to notice, in the power series expansion of G(t), is that the coefficient of t^r is the probability P(X = r).)

Applications of PGF:

- Using the probability generating function to calculate probabilities.

Given only the PGF $G_x(t) = E(t^x)$, we can recover all probabilities $P(X = x)$.

Let, $G_x(t) = p_0 + p_1t + p_2t^2 + p_3t^3 + p_4t^4 + \ldots$

$G'_X(t) = p^1 + 2p_2t + 3p_3t^2 + 4p_4t^3 + \ldots$ then $p_1 = P(X = 1) = G'_X(0)$.

$G''_X(t) = 2p_2 + (3 \times 2)p_3t + (4 \times 3)p_4t^2 + \ldots$ then $p_2 = P(X = 2) = (1/2)\, G''_X(0)$

$G'''_X(t) = (3 \times 2 \times 1)p_3 + (4 \times 3 \times 2)p_4t + \ldots$ then $p_3 = P(X = 3) = (1/3!)\, G'''_X(0)$.

In general,

$p_n = P(X = n) = (1/n!)\ G^{(n)}_x(0)$

- Expectation and variance from the PGF

Consider first and second derivatives G' (t) and G''(t) (the differentiation is with respect to t of course):

$$G'(t) = 1\,P(X = 1)t^{0} + 2\,P(X = 2)t^{1} + 3\,P(X = 3)t^{2} + 4\,P(X = 4)t^{3} + \cdots$$

$$G''(t) = 2.1\,P(X = 2)t^{0} + 3.2\,P(X = 3)t^{1} + 4.3\,P(X = 4)t^{2} + \cdots$$

Now, consider G(1), G' (1) and G''(1):

$$G'(1) = 1\,P(X = 1) + 2\,P(X = 2) + 3\,P(X = 3) + 4\,P(X = 4) + \cdots$$

$$G''(1) = 2.1\,P(X = 2) + 3.2\,P(X = 3) + 4.3\,P(X = 4) + \cdots$$

express G(1), G' (1) and G''(1) in sigma notation

$G'(1) = \sum_r r.P(X = r) = E(X)$
$G''(1) = \sum_r r(r-1).P(X = r) = E(X(X-1))$

By differentiating we have obtain the expectation of E(X). By differentiating again we get E(X(X-1)) from where we can get variance V(X).

$$G''(1) + G'(1) - (G'(1))^2 = E(X(X-1)) + E(X) - E(X)$$
$$= E(X^2) - E(X) + E(X) - E(X)^2$$
$$= E(X^2) - E(X)^2$$
$$= V(X)$$

- Probability generating function for a sum of independent random variables:

Let $G_x(t)$ be the generating function associated with X and $G_y(t)$ be the generating function associated with Y .

$G_x(t) = E(t^x)$ and $G_y(t) = E(t^y)$

Consider the generating function associated with X + Y and call this G_{x+y} .

$G_{x+y}(t) = E(t^{(x+y)})$

$$= \sum_r \sum_s t^{r+s} P(X = r, Y = s)$$

$$= \sum_r \sum_s t^r t^s P(X = r).P(Y = s)$$

Note that P(X = r, Y = s) may be split only because X and Y are independent.

Hence: $G_{x+y}(t) = \sum_r t^r .P(X = r) \sum_s t^s .P(Y = s)$

$= E(t^x) .E(t^y)$

$= G_x(t).G_y(t)$

The generating function of the sum of two independent random variables is the product of the separate generating functions of the random variables. This idea can be extended to sum of any number of independent random variables.

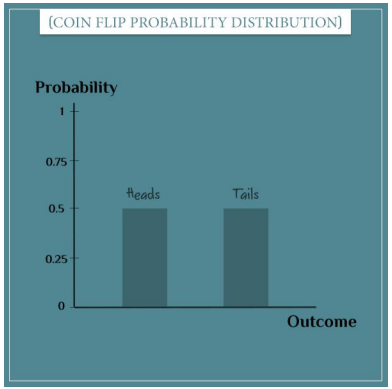
Discrete Probability Distributions:

The probability distribution of a discrete random variable X is a listing of all possible values of X and their probabilities of occuring.

E.g For the toss of an unbiased coin

X can take 0(head) or 1(tail) with ½ probability each. Then, the probability distribution will be

Value of X	0	1
Probability	1/2	1/2



- All discrete probability distributions must satisfy:

  1. 0 <= p(x) <= 1 for all x
  2. $\sum_{x}$ p(x) = 1.

Bernoulli Trial and Bernoulli Distribution:

A Bernoulli Trial is

- A single trial
- The trial can result in one of two possible outcomes, labelled success and failure.
- P(Success) = p
- P(failure) = 1 - p

E.g. The randomly chosen student in VNIT is mtech.

In a Bernoulli trial, let X = 1 if a success occurs, and X = 0 if a failure occurs. Then X has a Bernoulli distribution.

$$P(X = x) = p^x(1-p)^{1-x}$$

With mean = p
And variance = p(1-p).

E.g. Approximately 1 in 200 Indian adults are lawyers. One Indian is randomly selected.
The distribution of the above problem is Bernoulli with p = 1/200

Binomial Distribution:

The binomial distribution gives the discrete probability distribution of obtaining exactly x successes out of n Bernoulli trials.

$$P(X = x) = {}^{n}C_{x}\, p^{x}(1-p)^{n-x}$$

For x = 0, 1, 2, .., n

Mean = np

Variance = np(1-p)

- E.g. A balanced, six-sided die is rolled 3 times. What is the probability a 5 comes up exactly twice.

  Let X represent the number of fives in 3 rolls
  X has a binomial distribution with n = 3 and p = ⅙

$$P(X = 2) = {}^{3}C_{2}(1/6)^{2}(1 - 1/6)^{3-2}$$
$$= 0.0694$$

Poisson Distribution:

- Suppose Events are occurring independently.
- The probability that an event occurs in a given length of time does not change through time. (rate of event occurring)

Then X, the number of events in a fixed unit of time, has a Poisson distribution. The Poisson probability mass function is:

$$P(X = x) = (\lambda^{x} . e^{-\lambda}) / x!$$

Where λ is the mean number of occurrences or rate of event

Mean = λ, Variance = λ

- E.g. One nanogram of Plutonium-239 will have an average of 2.3 radioactive decays per second, and the number of decays will follow a Poisson distribution. What is the probability that in a 2 sec period there are exactly 3 radioactive decays?

  Here, λ will be = 2 * 2.3
  I.e. mean no. of decays in 2 second
  We can apply the equation
  $P(X = 3) = (\lambda^{3} . e^{-\lambda}) / 3!$
  $= 0.163$

Continuous Probability Distributions:

- In Continuous Probability Distributions the random variable X can take on any value. Because there are infinite values that X could assume, the probability of X taking on any one specific value is zero.

- We model a continuous random variable with a curve f(x), called probability density function (pdf).
  - f(x) represents the height of the curve at point x.
  - For continuous random variables probabilities are area under the curve.

- For any continuous probability distribution:

  f(x) >= 0 for all x.

  The are under the entire curve is equal to one.

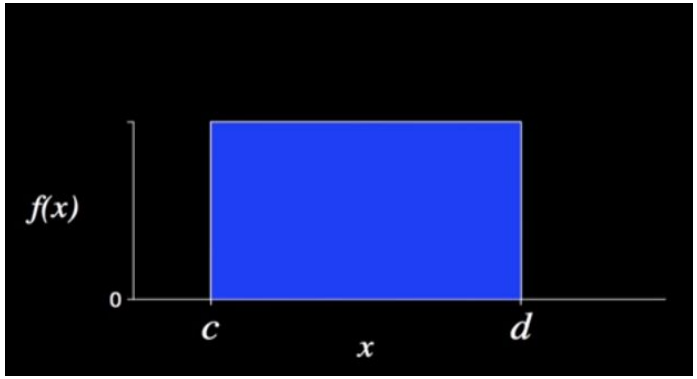
Continuous Uniform Distribution:

- For the uniform distribution, the probability density function f(x) is constant over the possible values of x.
- Any intervals of equal length are equally likely to occur.

f(x) = 1/(d-c) for c <= x <= d

0 elsewhere.

Mean = (c+d)/2, variance = $(1/12).(d - c)^2$



- Suppose X is a random variable that has uniform distribution with c = 200 and d = 250. What is the probability of X > 230.

  f(x) = 1/(d - c) = 1/50

  P(X > 230) will be the area under the rectangle formed by the region b/w 230 and 250

  P(X > 230) = length * width
  = (250 - 230) * 1/50
  = 0.4

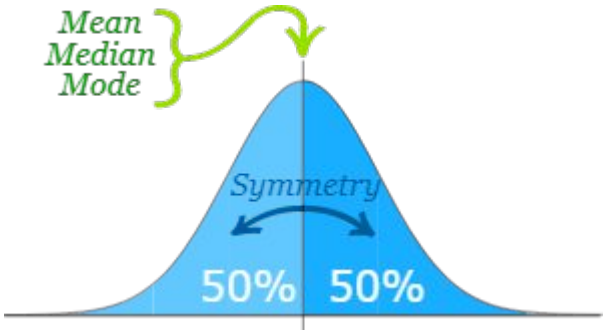
Normal Distribution (Gaussian Distribution):

- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- Probability Density Function is:

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

For -∞ <= x <= ∞

μ is mean, σ is standard deviation.



- For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

Parameter Estimation From Data:

- Given a model, the parameters are the numbers that yield the actual distribution. In the case of a Bernoulli random variable, the single parameter is the value p.
- Most of the time, we are in the situation of processing data where we don’t know the underlying distribution and the parameters of the distribution. Therefore we need the estimation.

  For e.g The sample mean x, which helps statisticians to estimate the population mean, μ.

- A good estimate should possess: consistency, unbiasedness and efficiency (in terms of low variance).

- Methods for parameter estimation:
  - Maximum Likelihood
  - Maximum A Posteriori

Maximum Likelihood:

- The central idea behind MLE is to select that parameters (θ) that make the observed data most likely.

  Suppose that the parameters are independent and identically distributed (IID) samples: X1,X2,...Xn.

  We define a likelihood function as:

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

  where θ is the parameter.

- In maximum likelihood estimation (MLE) our goal is to choose values of our parameters (θ) that maximizes the likelihood function.

  Let θ’ be the best approximate of the parameter θ. Then,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

  The argmax of a function is the same as the argmax of the log of the function. So instead of writing the likelihood we write the log likelihood.

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

- E.g. Suppose we have n data points which we will refer to as IID random variables X1,X2,...Xn.
  Every one of these random variables is assumed to be a sample from the same Bernoulli, with the same p, The likelihood of Bernoulli is the pmf.

$$L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$LL(\theta) = \sum_{i=1}^{n} \log p^{X_i}(1-p)^{1-X_i}$$

$$= \sum_{i=1}^{n} X_i(\log p) + (1-X_i)log(1-p)$$

$$= Y \log p + (n-Y)log(1-p)$$

Now for the value of p that maximizes the log likelihood,

$$\frac{\delta LL(p)}{\delta p} = Y\frac{1}{p} + (n - Y)\frac{-1}{1-p} = 0$$

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

which is just the sample mean.

Maximum a Posteriori

- The paradigm of MAP is that we should chose the value for our parameters that is the most likely given the data. MAP works on

$$\theta_{MAP} = = \text{argmax}_{\theta} P(\theta \mid X)$$

- With the Bayes rule,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\propto P(X|\theta)P(\theta)$$

Using the Bayes rule we get,

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log P(X|\theta) + \log P(\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta) + \log P(\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta) + \log P(\theta)$$

where first term is same as the likelihood, while the second term is the log of the prior of θ.

References

- Probability and Statistics with Reliability, Queuing and Computer Science Applications by Kishor S. Trivedi
- Introduction to Probability by Charles M. Grinstead and J. Laurie Snell