Probability Notes

1. Def
    a. Sample space- all possible events
    b. Events- any subset of sample space
    c. Random Experiment- All outcomes known, trial of experiment is non reproducible and results in something we don’t previously know about
    d. independent events- If the outcome of one event does not affect the outcome of another
    e. Central tendency
        i. **Mean:** The "balancing point" of a dataset - sum all the values and divide by the number of values. Like the center of gravity of a seesaw.
        ii. **Median:** The middle value of a dataset when ordered from least to greatest. Think of it as the "fair share" value.
        iii. **Mode:** The most frequent value in a dataset. It's the "popular kid" of the data set.
        iv. **Variance:** How "spread out" a dataset is from its mean. Imagine a dance floor - variance tells you how far dancers are typically from the center

(mean).

v. **Standard deviation:** The square root of the variance. It's like the "average distance" dancers are from the center of the dance floor.

f. We define random variable a function which maps from sample space of an experiment to the real numbers. Suppose a dice is thrown (X = outcome of the dice). Here, the sample space S = {1, 2, 3, 4, 5, 6}. The output of the function will be: P(X=1/2/3/4/5/6) = 1/6/ Types

   i. Discrete- Finite possible values like head / tail

   ii. Continuous- infinite values in a range like $0 \leq x \leq 1$

   iii. Mixture. conditional, for example- for values <5, it is discrete and for >5, it is continuous

g. Probability Mass Function- characterizes the distribution of a discrete random variable. This is different from PDF- probability density function because PDF is continuous random variable and this is for discrete. Let X be a discrete random variable of a function, then the probability mass function of a random variable X is given by Px (x) = P(X=x), For all x belongs to the range of X

   i. $P_x(x) \geq 0$ and

   ii. $\sum_{x \varepsilon Range(x)} P_x(x) = 1$ - sum of all probabilities over the range of X is 1

h. Probability Density function- The Probability Density Function(PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution function or just a probability function.

   i. In the case of a continuous random variable, the probability taken by X on some given value x is always 0. In this case, if we find P(X = x), it does not work. Instead of this, we must calculate the probability of X lying in an interval (a, b).

   ii. The probability density function is non-negative for all the possible values, i.e. $f(x) \geq 0$, for all x.

   iii. It is area under curve

i. CDF- cumulative distribution function- The cumulative distribution function (cdf) gives the probability that the random variable X is less than or equal to x and is usually denoted F (x)

    i. Lies between 0 and 1

    ii. The PDF is the derivative of CDF

j. Moment - The moments are the expected values of X, e.g., E(X), $E(X^2)$, $E(X^3)$, … etc.The first moment is **E(X) is called mean** , The second moment is **$E(X^2)$ is called variance**,

    i. the third moment is about the asymmetry of a distribution. - Skewness

    ii. The fourth moment is about how heavy its tails are.- Kurtosis

k. Bernoulli Trial random experiments in probability whose possible outcomes are only of two types, For example Heads/ Tails kind of things. A sequence of this is called a Bernoulli process

l. Popular Discrete distributions

    i. Binomial- Represents the number of successes in a fixed number of independent Bernoulli trials (each with two possible outcomes). For n = 1, i.e. a single experiment, the binomial distribution is a Bernoulli distribution.

        1. Only the number of success is calculated out of n independent trials.

        2. There is ‘n’ number of independent trials or a fixed number of n times repeated trials.

        3. Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

    ii. Negative Binomial Distribution- Extends the geometric distribution to count the number of trials needed for a specified number of successes. For example, we can define rolling a 6 on a die as a success, and rolling any other number as a failure, and ask how many failure rolls will occur before we see the third success (r=3).

    iii. Geometric Distribution- Models the number of Bernoulli trials needed to get the first success. A geometric distribution is a special case of the negative binomial distribution where the number of successes required (r) is equal to 1.

iv. A Poisson distribution is **a discrete probability distribution**. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space.Assumptions are:

1. Individual events happen at random and independently. That is, the probability of one event doesn’t affect the probability of another event.

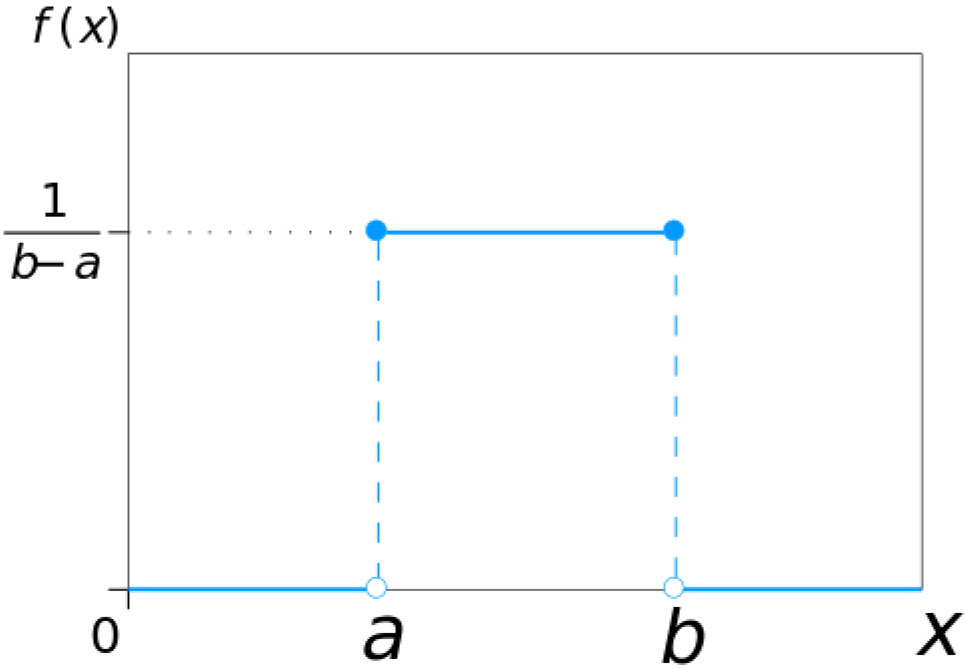
2. You know the mean number of events occurring within a given interval of time or space. This number is called λ (lambda), and it is assumed to be constant. This is calculated over an entire time period. Imagine distribution is calculated for 1 year when sample is for 10 years long so lambda will be per year average.

3. In general, Poisson distributions are often appropriate for count data. Count data is composed of observations that are non-negative integers

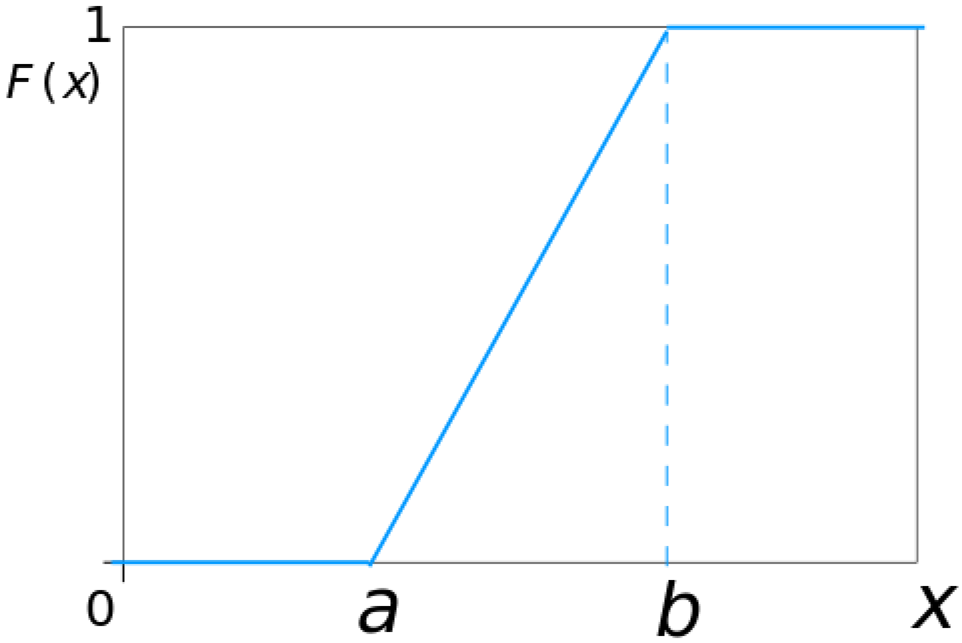
m. Continuous Distributions

i. Uniform Distribution- All outcomes are equally likely over a continuous range

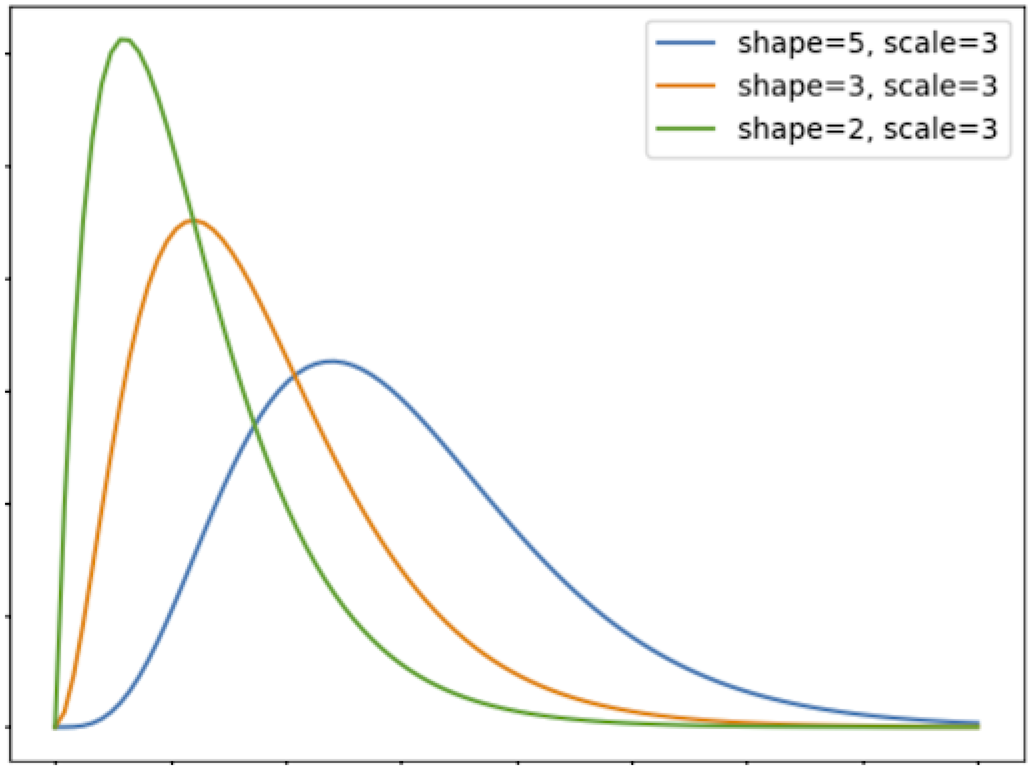
1. Such a distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.The bounds are defined by the parameters, a and , b, which are the minimum and maximum values.



2. The CDF of the distribution looks something like this



ii. Gamma Distribution - Describes the time until a specified number of events occur, with a constant rate of occurrence.



1. In summary, the shape parameter (α) controls the form and skewness of the Gamma distribution, while the scale parameter (β) adjusts the spread and scale of the distribution.
2. widely used in the field of Business, Science and Engineering, in order to model the continuous variable that should have a positive and skewed distribution.
3. The Gamma distribution is a continuous probability distribution that generalizes the exponential distribution. It's often used for the time until an event occurs a certain number of times. In the context of the Gamma distribution,λ is often used as the rate parameter.
4. So in Poisson distribution we get questions like - What is the probability of exactly 3 customers arriving in a given hour given λ=2 (average number of arrivals per hour) but in Gamma we get What is the probability that the third customer will arrive within the first 2 hours?
5. In the context of the Gamma distribution, especially in scenarios like service times or waiting times, the shape parameter α can be intuitively

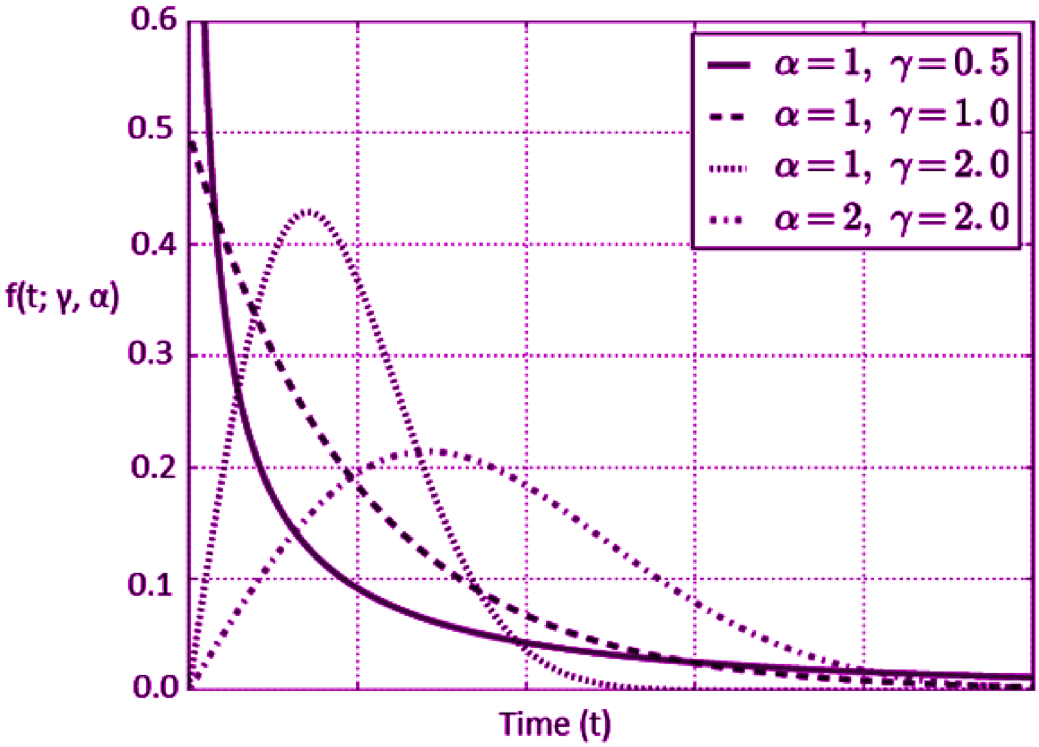
understood as related to the number of events or stages that need to occur for the process to be completed.

   a. For example, if α=3 in a coffee shop, it could mean that there are three key stages in serving a customer: taking the order, preparing the order, and delivering the order.

   b. When α is less than 1, it indicates a process with fewer or even fractional stages.

6. The chi square distribution is a special case of the gamma distribution

iii. weibull distribution- Useful in reliability analysis, focusing on time until a failure of a system.



1. One of the most widely used distributions in reliability engineering

2. it can simulate various distributions like normal and exponential distributions

3. Where to use

a. it can model products that are more likely to fail either early in their life (due to manufacturing defects) or later (due to wear and tear).

b. It's used to analyze failure rates and predict remaining useful life, accommodating a wide range of failure behaviors.

c. In meteorology, the Weibull distribution is often used to model wind speed, which is crucial for wind energy projects.

4. Suppose you're assessing the viability of a wind farm. The wind speed data over several years shows variability. By using the Weibull distribution, you can model the wind speeds effectively, considering the varying rates of occurrence of different wind speeds. The shape parameter will tell you about the distribution of wind speeds (more high-speed winds vs. more moderate winds), and the scale parameter will give an idea of the 'typical' wind speed.

5. The parameter *λ* (lambda) in the Weibull distribution is not the same as *λ* in the Poisson or Gamma distributions, even though the same Greek letter is used. In the Weibull distribution, λ scales the distribution and relates to the distribution of times until an event occurs. Think of λ as a "characteristic life" or "typical scale" of the phenomenon being modeled. It provides a reference point for understanding the typical spread of values within the distribution.λ might represent the average time until the first failure occurs for a typical machine in a population, might correspond to the median survival time for a patient population, could indicate the average wind speed in a particular region. Etc

6. Suppose a machine's failure times follow a Weibull distribution with k = 2 and λ = 1000 hours. The probability of the machine failing within the first 500 hours is F(500). The probability of it lasting beyond 1500 hours is 1 - F(1500)

iv. Normal Distribution- Bell-shaped curve; describes a continuous variable whose probabilities are symmetrically distributed.

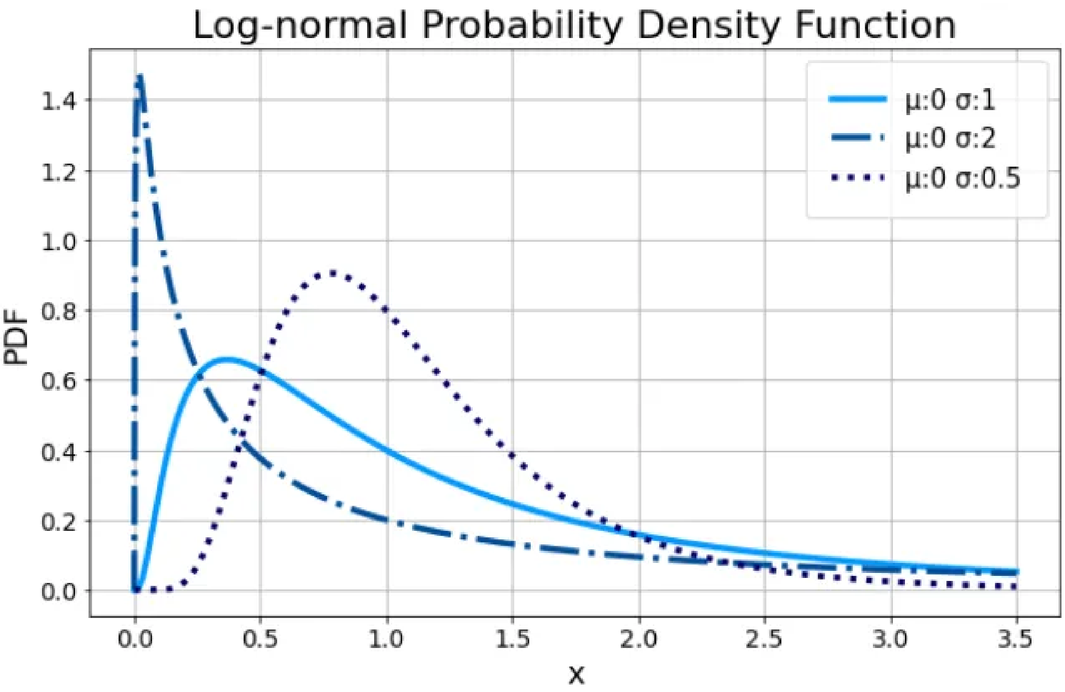
1. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics.

2. Furthermore, it can be used to approximate other probability distributions,

3. The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

4. Whereas, the normal distribution doesn't even bother about the range. The range can also extend to –∞ to + ∞ and still we can find a smooth curve.

v. LogNormal Distribution- Distribution of a variable whose logarithm is normally distributed.



1. The distribution is occasionally referred to as the Galton distribution or Galton's distribution

2. The data points for our log-normal distribution are given by the X variable. When we log-transform that X variable (Y=ln(X)) we get a Y

variable which is normally distributed. Similarly when we take the exponential of a normal distribution we get the log normal distribution

3. The most efficient way to analyse log-normally distributed data consists of applying the well-known methods based on the normal distribution to logarithmically transformed data and then to back-transform results if appropriate.
4. We can estimate our log-normal parameters μ and σ using maximum likelihood estimation (MLE).

n. transformation of random variables

   i. Transformation of 1 dimension random variable which is continious. Given pdf of x we need to find the PDF of y which is a function of x.

      1. If it is a discrete function we have the PMF

         a. when function is 1-1

         b. when function is not 1-1

   ii. Transformation in 2 dimensions

o. Random variables with 2 dimensions - Considers pairs of random variables and their joint behavior.

   i. Given two random variables that are defined on the same probability space, [1] the joint probability distribution is the corresponding probability distribution on all possible pairs of outputs

   ii. The joint distribution can just as well be considered for any given number of random variables.

   iii. It has

      1. Marginal distribution- Distribution of every random variable on it's own
      2. Conditional Probability Distribution on 2 variables

   iv. Discrete joint distribution

	A=Red	A=Blue	P(B)

B=Red	(2/3)(2/3)=4/9	(1/3)(2/3)=2/9	4/9+2/9=2/3
B=Blue	(2/3)(1/3)=2/9	(1/3)(1/3)=1/9	2/9+1/9=1/3
P(A)	4/9+2/9=2/3	2/9+1/9=1/3	

p. Marginal PMF- Probability mass function of a subset of a collection of random variables, disregarding the others.

q. Conditional PMF

   i. Conditional Expectation- Expected value of a random variable given another.

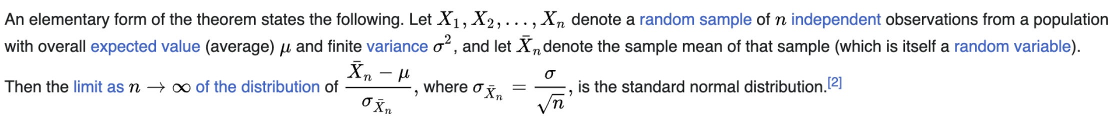
   ii. Conditional Mean and variance- Calculate the average and spread of a variable **given** that we know the value of another variable. They explore how the behavior of one variable changes depending on the specific value of another variable.

   iii. Covariance- Measure of how much two random variables vary together. Covariance is a measure of linear relationship between the random variables. If the relationship between the random variables is nonlinear, the covariance might not be sensitive to the relationship, which means, it does not relate the correlation between two variables.

   iv. Correlation coefficient- Normalized measure of the covariance. The correlation just scales the covariance by the product of the standard deviation of each variable. Consequently, the correlation is a dimensionless quantity that can be used to compare the linear relationships between pairs of variables in different units.

   v. Joint probability- Probability of two events occurring together.

r. Central Limit Theorem- States that, under certain conditions, the sum of a large number of random variables is approximately normal. In probability theory, the central limit theorem (CLT) states that, under appropriate conditions, the distribution of a normalized version of the sample mean converges to a standard normal distribution. This holds even if the original variables themselves are not normally distributed. There are several versions of the CLT, each applying in the context of different conditions.



An elementary form of the theorem states the following. Let $X_1, X_2, \ldots, X_n$ denote a random sample of n independent observations from a population with overall expected value (average) μ and finite variance σ^2, and let $\bar{X}_n$ denote the sample mean of that sample (which is itself a random variable). Then the limit as $n \to \infty$ of the distribution of $\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}}$, where $\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}$, is the standard normal distribution.[2]

s. Sampling from a distribution- Process of selecting a random sample of individuals from a statistical population.

  i. Simple Random Sampling (SRS):

    1. Each individual in the population has an equal chance of being selected.It's analogous to drawing names out of a hat.

  ii. Systematic Sampling:

    1. Selects individuals at regular intervals from a list or ordered population.

    2. Example: Choosing every 10th person on a list.

  iii. Stratified Sampling:

    1. Divides the population into subgroups (strata) with shared characteristics and samples randomly from each stratum.

    2. Ensures representation of different groups within the sample.

  iv. Cluster Sampling:

    1. Divides the population into clusters (groups), randomly selects clusters, and includes all individuals within those clusters in the sample.

  v. Importance Sampling

    1. Involves sampling more frequently from regions of the distribution that are more important for a particular problem. Often used in statistical methods

  vi. Rejection Sampling:

    1. Uses a proposal distribution to generate potential samples and accepts those that meet a specific condition.

    2. Used for sampling from a complex distribution

  vii. Markov Chain Monte Carlo (MCMC):

1. A family of algorithms that generate samples by constructing a Markov chain that has the desired distribution as its stationary distribution.

2. Effective for high-dimensional and complex distributions.

t. Chi-Squared test

i. The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

ii. Use case

1. Goodness of fit test

2. Test of Independence: Examine the relationship between two categorical variables (e.g., gender and preference for a product).

3. Test of Homogeneity: Compare the proportions of different categories across multiple populations (e.g., comparing the proportion of smokers in different age groups).

u. T distribution

i. In probability and statistics, Student's t-distribution (or simply the t-distribution) is a continuous probability distribution that generalizes the standard normal distribution. Like the latter, it is symmetric around zero and bell-shaped.

ii. The heaviness of the tail is controlled by something called the v parameter, for v=0, it is the cauchy distribution but for v$\rightarrow \inf$ it becomes the standard normal distriution

iii. V is called the degrees of freedom

1. In statistics, degrees of freedom represent the number of independent pieces of information that can vary freely when estimating a statistic.

2. When calculating the mean of a sample, we use one degree of freedom to estimate the mean itself. The remaining data points are free to vary around that mean.

3. This is why the degrees of freedom for a t-distribution with a sample size of n is n-1

4. With fewer degrees of freedom (smaller samples), the t-distribution has fatter tails than a normal distribution. This reflects the greater uncertainty in estimating the mean from a smaller sample.

5. As the degrees of freedom increase, the t-distribution becomes more and more similar to the normal distribution. This is because the estimate of the Wmean becomes more precise with larger samples.

iv. Why do we even need t distribution

1. In many real-world scenarios, we don't have access to the entire population's variance. We have to estimate it from the sample we have.

2. When working with small samples (typically less than 30), the sample mean might not perfectly reflect the population mean.

3. The t-distribution is used extensively in hypothesis testing, particularly in t-tests. eg comparing mean of 2 distributions, confidenceinterval calculation

v. Probability computation for T distribution- Used for estimating population parameters for small sample sizes or unknown variances.

v. F Distribution- Arises frequently in ANOVA and in testing whether two variances are equal. It's like a tool to compare and measure variability

i. The F-distribution helps compare these variances l

ii. A high F value means B has more variation than A

iii. You can change degrees of freedom by checking how many samples you are measuring for variability comparison sake

w. Estimation

i. Point estimation. Methods are

1. Unbiassed

2. Biassed

ii. Interval Estimation

iii. Parameter space

x. CR inequality - The Chernoff bound, also known as the Chernoff inequality, is a powerful tool in probability theory that provides upper bounds on the tail probabilities of certain random variables. It's particularly useful in analyzing the probability of rare events, assessing the accuracy of probability approximations, and establishing convergence rates in large-scale systems.

    i. The Chernoff bound estimates the probability that a random variable deviates significantly from its expected value.

    ii. It focuses on the "tails" of the distribution, where extreme events occur.

y. Maximum Likeklihod Estimation - In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

    i. Steps-

        1. First select points that we need to model
        2. Then assume a model that would be able to contain these points for example normal distribution
        3. We know that different parameters of Gaussian distribution will result in different distributions
        4. Maximum likelihood estimation is a method that will find the values of μ and σ that result in the curve that best fits the data.
        5. We have to make an assumption in case we want to calculate the MLE. The assumption is that the data points are independent of each other. Why? Because in case of dependent events we would need to calculate the conditional probabilities. But in case of independent events, we can just use products instead of conditionals.
        6. So now let’s say we have 3 points and we need to calculate the MLE of the three points in a Gaussian plane. We will fir the Gauss formula for all three points and multiply them to get the joint probability. For ex: for numbers 9,9.5 and 11 this is how it would look like

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

7. We just have to figure out the values of *μ* and *σ* that results in giving the maximum value of the above expression.

8. Now we use differentiation to differentiate this function

9. The above expression for the total probability is actually quite a pain to differentiate, so it is almost always simplified by taking the natural logarithm of the expression. This is absolutely fine because the natural logarithm is a monotonically increasing function.

10. In order to find the mean, we differential wrt the mean and in order to find the standard deviation we differentiate wrt the standard deviation

ii. Invariance properties in MLE

z. Methods of moment Estimation

i. Maximum likelihood estimation (MLE) as you saw had a nice intuition but mathematically is a bit tedious to solve. So we use a different technique for estimating parameters called the Method of Moments (MoM)

ii.

aa. Confidence estimation- interval estimation

i. Confidence inetrval - A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

ii. It calculates the margin of error so that we are n% correct

ab. Hypotheiss testing -Hypothesis testing is the process of checking the validity of the claim over a population using evidence found in sample data.

i. Statistical Hypothesis- A hypothesis is nothing but an assumption/claim/proposition made about something which is later on tested statistically to verify its truth.

ii. Null Hypothesis - In a null hypothesis, we claim that there is no relationship or difference between groups with respect to the value of the population parameter. We begin the Hypothesis Test by assuming that the Null Hypothesis to be true and later on we retain/reject the Null Hypothesis based on the evidence found in sample data. If equaltiy appear it will be in nul (=/ ≤ / ≥) Example:

   1. Average life of vegetarians is different than that of meat-eaters
   2. Proportion of married people defaulting on loan repayment is less than the proportion of singles defaulting on loan repayment.

iii. Alternative Hypothesis- In the alternative Hypothesis, we claim that there is some change /relationship between groups with respect to the value of the population parameter. An alternative hypothesis always contradicts the Null Hypothesis and only any one of the above hypotheses could be true. It is denoted using ≠ ≥ ≤ symbols

iv. Steps

   1. Formulate the null and alternative hypothesis
   2. Decide the Significance Level (α) at which we get to reject the null hypothesis. Usually, α is set as 0.05. This means that there is a 5% chance we will reject the Null Hypothesis even when it’s true.
   3. Calculate the test statistic and p value- A test statistic is nothing but the standardized difference between the value of the parameter estimated from the sample (such as sample mean) and the value of the null hypothesis (such as hypothesized population mean). It is a measure of how far the sample mean is from the hypothesized population mean.
   4. Take a decision

v. Critical value- The value of the statistic in the sampling distribution for which the probability is α is called the critical value. The areas beyond the critical values are known as the critical region/rejection region and critical values are the values that indicate the edge of the critical region. Critical regions describe the entire area of values that rejects the null hypothesis. If the test statistic falls in the critical region, the null hypothesis will be rejected.

vi. P-value:A p-value is nothing but the conditional probability of getting the test statistic given the null hypothesis is true.

vii. Types of hypothesis test

   1. One Tailed Test- When the critical/rejection region is on one side of the distribution it is known as a One-Tailed Test. In this case, the null hypothesis will be rejected if the test statistic is on one side of the distribution, either left or right.

      a. **Left Tailed Test:** If the test is left-tailed, the critical/rejection region, with an area equal to α, will be on the left side of the distribution curve. In this case, the null hypothesis will be rejected if the test statistic is very small(as it will fall on the left end of the distribution).

      b. **Right Tailed Test:** If the test is right-tailed, the critical /rejection region, with an area equal to α, will be on the right side of the distribution curve. In this case, the null hypothesis will be rejected if the test statistic is very large (or falls on the right end of the distribution).

   2. Two Tailed Test - If the test is two-tailed, α must be divided by 2 and the critical /rejection regions will be at both ends of the distribution curve. Hence, in this case, the null hypothesis will be rejected when the test value is on either of two rejection regions on either side of the distribution.

viii. T test can be done when

   1. The population distribution is normal or

   2. The sampling distribution is symmetric and the sample size is ≤ 15 or

   3. The sampling distribution is moderately skewed and the sample size is 16 ≤ n ≤ 30 or

   4. The sample size is greater than 30, without outliers.

ix. Z test can be done when

   1. when the population is normally distributed and σ is known.

   2. The sample size n ≥ 30

x. Errors

1. Type 1 Error- FP

   a. Significance level (alpha): The probability of committing a type 1 error, usually set at 0.05 or 0.01.

2. Type 2 error- FN

   a. Power of a test: The probability of correctly rejecting a false null hypothesis (avoiding a type 2 error).

   b. Trade-off: As you decrease the chance of one error, you often increase the chance of the other.

2. Formula

   a. $P(A \cup B) = P(A) + P(B) – P(A \cap B)$

   b. Independent events- $P(A \cap B) = P(A) \cdot P(B)$

   c. Conditional probability of A given that B has already occurred- $P(A | B) = P(A \cap B) / P(B)$

      i. For independent events A and B, $P(B|A) = P(B)$

      ii. By corollary $P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A)$,

   d. Bayes rule- $P(A | B) = P(B | A) \cdot P(A) / P(B)$

   e. Multiplication Theorem for n events - $P(A1 \cap A2 \cap \ldots \cap An) = P(A1) P(A2 | A1) P(A3 | A1 \cap A2) \ldots \times P(An |A1 \cap A2 \cap \ldots \cap An\text{-}1)$

      i. For n independent events, the multiplication theorem reduces to $P(A1 \cap A2 \cap \ldots \cap An) = P(A1) P(A2) \ldots P(An)$.

   f. Total probability theorem- Let events C1, C2 . . . Cn form partitions of the sample space S, where all the events have a non-zero probability of occurrence. For any event, A associated with S, according to the total probability theorem,

$$P(A) = \sum_{k=0}^{n} P(C_k) P(A|C_k)$$

g. Random Variable formula. For any random variable X where P is its respective probability we define its mean as,

  i. Mean(μ) = ∑ X.P

  ii. For any random variable X if it assume the values x1, x2,…xn where the probability corresponding to each random variable is P(x1), P(x2),…P(xn), then the expected value of the variable is- E(x) = ∑ x.P(x)

  iii. Variance - Var(x) = σ2 = E(X2) – {E(X)}2 where

    1. *E(X2)* = ∑*X2P*

    2. *E(X)* = ∑*XP*

h. PDF- P(a ≤ X ≤ b) = P(a < X ≤ b)

  i. $\mathbf{P(a < X < b) = \int_a^b f(x)\,dx}$

  ii. $\mathbf{P(a \le X \le b) = \int_a^b f(x)\,dx}$

i. CDF

  i. $F(x) = \mathrm{P}[X \leq x].$

  ii. Let X be a random variable with cdf F (x) . Then - $P[a < X \leq b] = F(b) - F(a).$

  iii. Continuous value function - $F(x) = \int\limits_{-\infty}^{x} f(t)\mathrm{d}t.$

j. MGF- Moment generating function -

  i. if two random variables have the same MGFs, then their distributions are the same.

  ii. $m(t) = \mathbb{E}(e^{tX})$ ⇒ the nth derrivate wrt t using chain rule $m^{(n)}(t) = \mathbb{E}X^n.e^{tX}$ for example $m^{(n)}(0) = \mathbb{E}X^n$. This substitution fo t must be done only at the end.

$$M_X(t) = E(e^{tX})$$

  iii. There are 2 assumptions calculating MGF

1. Mx(t) exist for X
2. The Mx(t) has a finite value for all t belonging to [-a, a], where a is any positive real number.
3. t is a random. By varying t, we essentially zoom in and out on the distribution along the number line. At t = 0, we focus on the center of the distribution, while larger values of t shift our focus towards the tails. This encodes information about how far individual values are from the average. If two MGFs have the same value for all possible values of t, then the underlying probability distributions are identical. T is the domain of the function.

$$\mu = E(X) = \sum_{-\infty}^{\infty} x \cdot P_X(x) \qquad \text{(First Moment, Discrete)}$$

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) \cdot dx \qquad \text{(First Moment, Continuous)}$$

$$E(X^2) = \sum_{-\infty}^{\infty} x^2 \cdot P_X(x) \qquad \text{(Second Moment, Discrete)}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f_X(x) \cdot dx \qquad \text{(Second Moment, Continuous)}$$

$$E(X^n) = \sum_{-\infty}^{\infty} x^n \cdot P_X(x) \qquad (\text{n}^{\text{th}}\ \text{Moment, Discrete})$$

$$E(X^n) = \int_{-\infty}^{\infty} x^n \cdot f_X(x) \cdot dx \qquad (\text{n}^{\text{th}}\ \text{Moment, Continuous})$$

k. Properties of MGF

i. Moment Gathering Functions when a random variable undergoes a linear transformation for example if y=ax+b. Given b is a constant. This can be extended for an infinitely long series Y = X1 + X2 + … + Xn

$$M_Y(t) = E(e^{tY})$$

$$M_Y(t) = E\left(e^{t(\alpha X+\beta)}\right)$$

$$M_Y(t) = E\left(e^{(t\alpha X+t\beta)}\right)$$

$$M_Y(t) = E\left(e^{t\alpha X} \times e^{t\beta}\right)$$

$$M_Y(t) = e^{t\beta} \times E\left(e^{(t\alpha)X}\right)$$

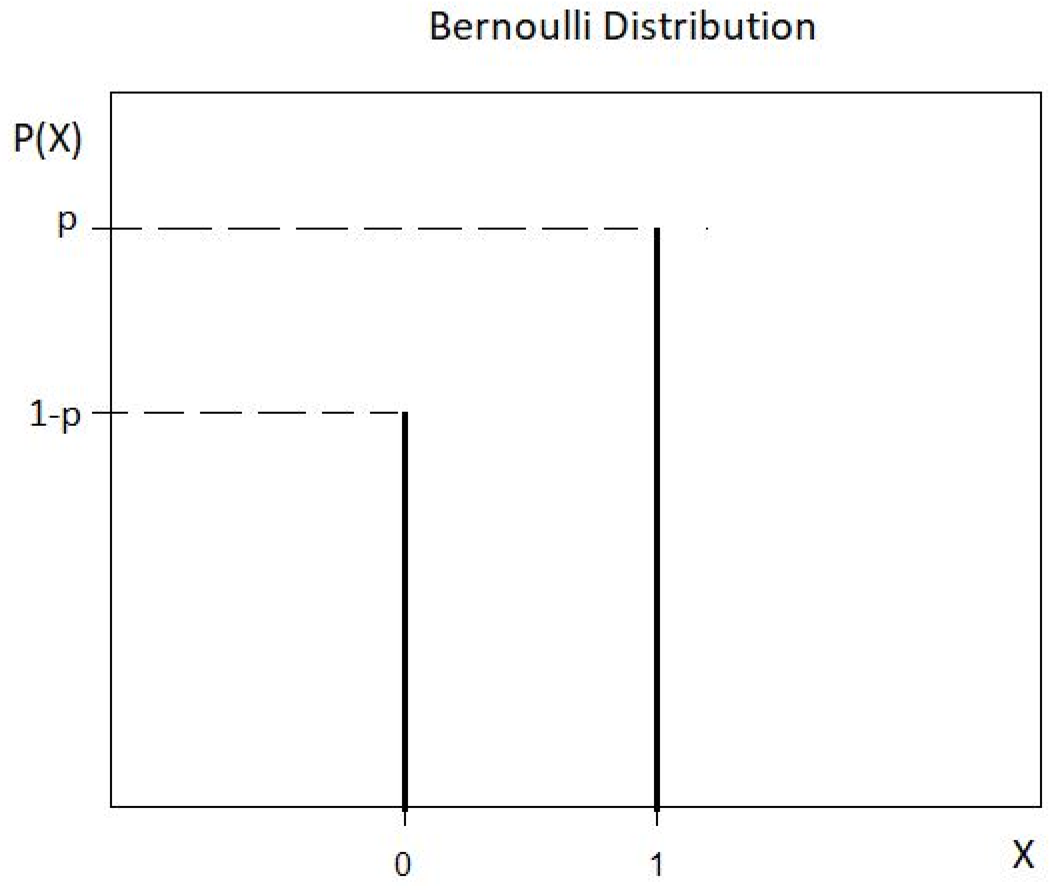
$$M_Y(t) = e^{t\beta} \times M_{\mathrm{X}}(t\alpha)$$

$$M_Y(t) = M_{X_1}(t) \times M_{X_2}(t) \times \cdots \times M_{X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

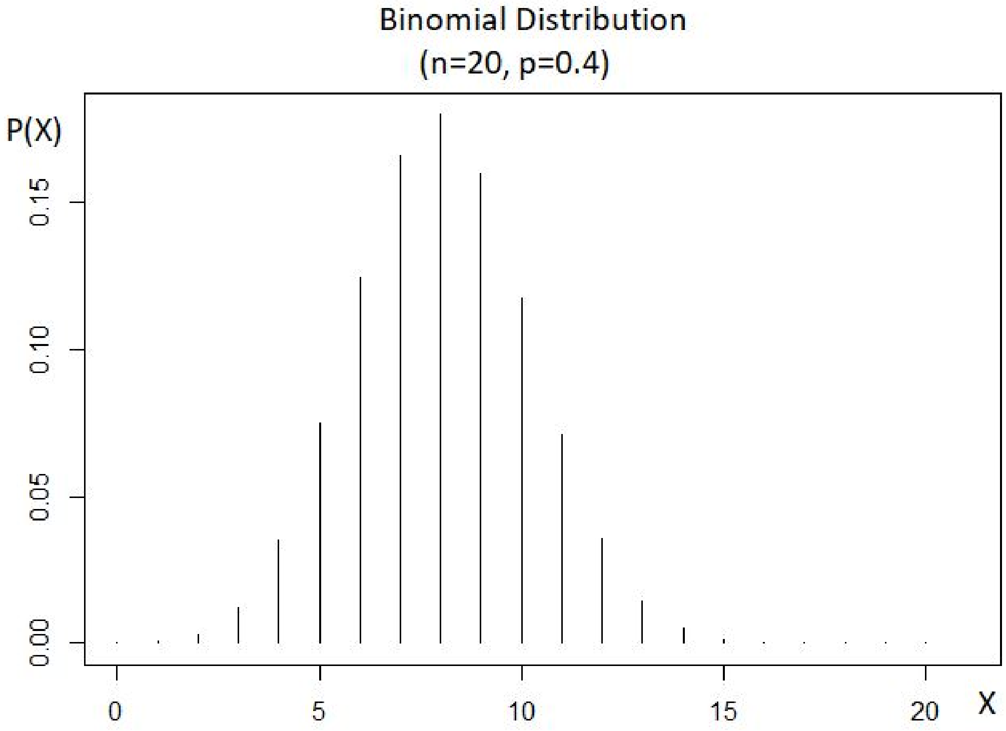
l. Bernouli Distribution- simplest distribution ever

   i. Equation for the distribution is as follows $P(X = k) = p^k(1 - p)^{1-k}, \quad k = 0, 1$

   ii. MGF $M_X(t) = 1 - p + pe^t$



1. Binomial Distribution



1. In order to evaluate the binomial distribution we use combinatorial. Which is what the first term denotes.
2. Equation- $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,1,...,n$
3. MGF- $M_X(t)=(1-p+pe^t)^n$
4. Mean, μ = np and Variance, σ2 = npq where q = 1-p

2. Geometric Distrbution
   a. **k:** Number of successes observed
   b. $P(X=k)=(1-p)^{k-1}p, \quad k=1,2,...$
   c. $M_X(t)=\frac{pe^t}{1-(1-p)e^t}$
   d. **Mean:** μ = 1/p
   e. **Variance:** σ^2 = (1-p)/p^2
3. Negative Binomial Distribution
   a. k is number of successes

   b. $P(X = k) = \binom{k+r-1}{k} p^r (1-p)^k, \quad k = 0, 1, 2, ...$

   c. MGF $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r$

   d. **Mean:** μ = r/p

   e. **Variance:** σ^2 = r(1-p)/p^2

4. Uniform Cost Distribution

   a. As we have seen before a and b are the minimum and maximum values so the range of values for x for the distribution

   b. **PDF:** f(x) = 1/(b-a), for a ≤ x ≤ b; 0, otherwise

   c. **CDF:** F(x) = 0, for x < a; (x-a)/(b-a), for a ≤ x ≤ b; 1, for x > b

   d. **MGF:** M(t) = $(e^{(}tb) - e^{(}ta))/(t(b-a))$

   e. Mean(a+b)/2

   f. Variance - (b-a)^2/12

5. Poisson Distribution

   a. When λ is low, the distribution is much longer on the right side of its peak than its left (i.e., it is strongly. As λ increases, the distribution looks more and more similar to a normal distribution.

   b. As λ increases, the distribution looks more and more similar to a normal distribution. In fact, when λ is 10 or greater, a normal distribution is a good approximation of the Poisson distribution. PMF. K is the number of times an event should occur in the given time period

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

   c. Mean- λ

   d. Variance- λ

   e. When λ is a non-integer, the mode is the closest integer smaller than λ.. When λ is an integer, there are two modes: λ and λ−1.

6. Gamma Distribution

a. Continuous poisson distribution

b. **PDF:** f(x) = $(\lambda^{\alpha}.x^{\alpha-1} * e^{-\lambda x})/\Gamma(\alpha), forx \geq 0; 0,$ otherwise

i. The gamma function can be seen as a solution to the following interpolation problem: "Find a smooth curve that connects the points (x, y) given by y = (x − 1)! at the positive integer values for x.”

ii. The gamma function is represented by **Γ(y)** which is an extended form of factorial function to complex numbers(real). So, if n∈{1,2,3,…}, then Γ(y)=(n-1)! and gamma for 1 is 1

iii. 1/Lambda is called the rate parameter. Lambda is called the scale parameter

iv. wherever the random variable x appears in the probability density, then it is divided by β. (which is inverse of lambda)

v. Discussing the shape parameter- alpha

1. positive real number
2. Smaller α values (e.g., less than 1) lead to right-skewed distributions with heavier tails.
3. Larger α values (e.g., greater than 1) produce more bell-shaped, symmetric distributions.
4. α = 1: The Gamma distribution becomes the Exponential distribution, a special case used for modeling memoryless processes.

c. $CDF:\ F(x) = \int 0^{x} (\lambda^{\alpha} * t^{(}\alpha - 1) * e^{(} - \lambda t))/\Gamma(\alpha) dt$

d. $MGF:\ M(t) = (1 - t/\lambda)^{(} - \alpha),\ fort < \lambda$

e. **Mean:** μ = α/λ

f. **Variance:** = α/λ^{2}

7. Weibull

a. k > 0 is the shape parameter and λ > 0 is the scale parameter of the distribution.

i. A value of k< 1 indicates that the failure rate decreases over time

        ii. A value of = k=1 indicates that the failure rate is constant over time.
        iii. A value of k>1 indicates that the failure rate increases with time.
    b. $PDF: f(x) = (k/\lambda) * (x/\lambda)^{k-1} * e^{-x/\lambda^k}, forx \geq 0; 0, otherwise$
    c. $CDF:\ F(x) = 1 - e^{(} - (x/\lambda)^k)$
    d. $MGF:\ M(t) = \Gamma(1 + k/t)$
    e. $Mean:\ \mu = \lambda * \Gamma(1 + 1/k)$
    f. $Variance:\ \sigma^2 = \lambda^2 * (\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2)$

8. Normal

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

    1. $PDF:\ f(x) = (1/(\sigma\sqrt(2\pi))) * e^{(} - (x - \mu)^2/(2\sigma^2))$
        a. x is the variable
        b. μ is the mean
        c. σ is the standard deviation
    2. $CDF:\ F(x) = \int(-\infty)^x(1/(\sigma\sqrt(2\pi))) * e^{(} - (t - \mu)^2/(2\sigma^2))dt$
    3. $MGF:\ M(t) = e^{(}\mu t + \sigma^2 t^2/2)$
    4. **Mean:** μ
    5. **Variance:** σ^2
    6. Properties
        a. In a normal distribution, the mean, median and mode are equal.(i.e., Mean = Median= Mode).
        b. The total area under the curve should be equal to 1.
        c. The normally distributed curve should be symmetric at the centre.

    d. The normal distribution curve must have only one peak. (i.e., Unimodal)

    e. The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

9. Lognormal Distribution

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

   1. Parameters NOTE- When our log-normal data is transformed using logarithms our μ can then be viewed as the mean (of the transformed data) and σ as the standard deviation (of the transformed data)

      a. μ is the location parameter and σ the scale parameter of the distribution. (And not mean)

      b. The estimations using MLE are listed below and that is how we will calculate the value of these parameters

   2. More stats

      a. The **median** is derived by taking the log-normal cumulative distribution function, setting it to 0.5 and then solving this equation

$$0.5 = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right)\right]$$

$$1 = 1 + \operatorname{erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right)$$

$$0 = \operatorname{erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right)$$

$$\operatorname{erf}^{-1}(0) = \frac{\ln(x) - \mu}{\sigma\sqrt{2}}$$

$$\sigma\sqrt{2} * \operatorname{erf}^{-1}(0) = \ln(x) - \mu$$

$$0 = \ln(x) - \mu$$

$$ln(x) = \mu$$

$$\boxed{\boldsymbol{x = e^{\mu}}}$$

1. The **mode** represents the global maximum of the distribution and can therefore be derived by taking the derivative of the log-normal probability density function and solving it for 0

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

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

$$= -\frac{1}{x^2\sigma\sqrt{2\pi}}\, e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}\left(1+\frac{\ln(x)-\mu}{\sigma^2}\right) = 0$$

$$1+\frac{\ln(x)-\mu}{\sigma^2} = 0$$

$$\frac{\ln(x)-\mu}{\sigma^2} = -1$$

$$\ln(x)-\mu = -\sigma^2$$

$$\ln(x) = \mu - \sigma^2$$

$$\boxed{\boldsymbol{x = e^{(\mu-\sigma^2)}}}$$

1. The **mean** (also known as the expected value) of the log-normal distribution is the probability-weighted average over all possible values
2. The **variance** of the log-normal distribution is the probability-weighted average of the squared deviation from the mean
3. $PDF:\ f(x) = (1/(x\sigma\sqrt{}(2\pi))) * e^(- (ln(x) - \mu)^2/(2\sigma^2)),\ for x > 0; 0,\ otherwise$
4. $CDF:\ F(x) = \Phi((ln(x) - \mu)/\sigma),\ where \Phi is the standard normal CDF$
5. $MGF:\ M(t) = e^(\mu t + \sigma^2 t^2/2)$
6. $Mean:\ e^(\mu + \sigma^2/2)$
7. $Variance:\ (e^(\sigma^2) - 1) * e^(2\mu + \sigma^2)$

10. MLE for Gaussian function

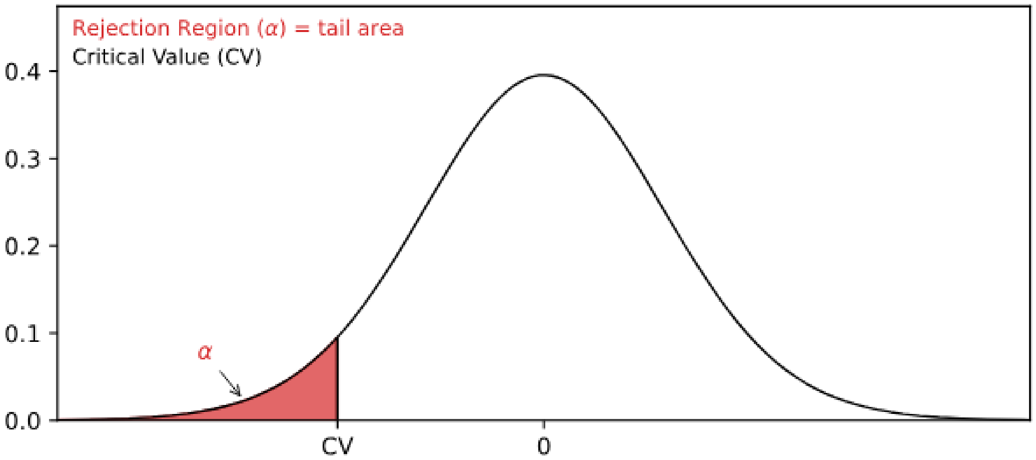
$$\hat{\mu} = \frac{\sum_i x_i}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{n}$$

11. MLE for Log Normal distribution

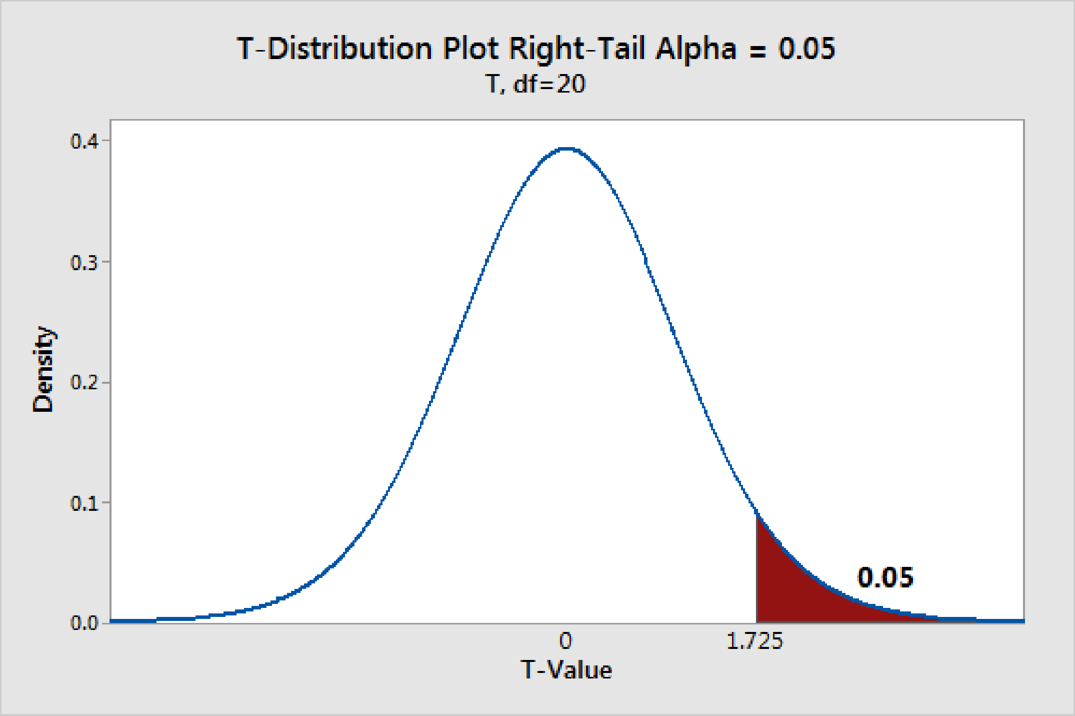
$$\hat{\mu} = \frac{\sum_k ln\, x_k}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_k (\ln x_k - \hat{\mu})^2}{n}$$

    1. These formulas are near identical. We can see that we can use the same approach as with the normal distribution and just transform our data with a logarithm first.

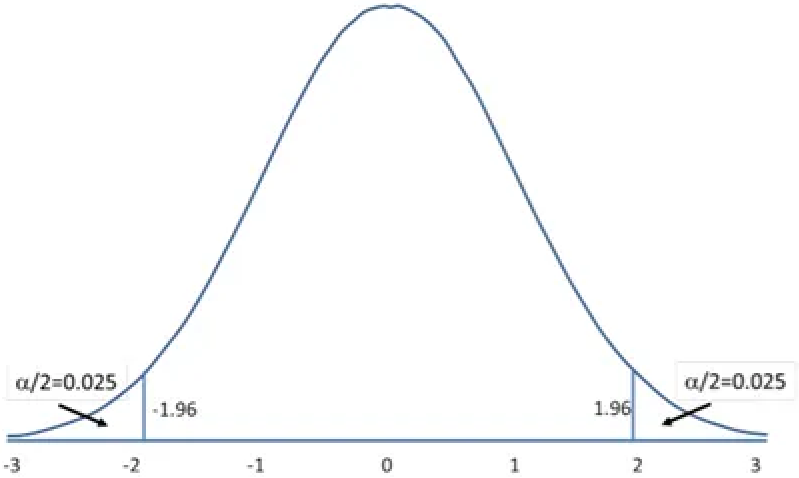
12. Hypotesis testing formula
    a. One tailed test
        i. Left Tailed Test p-value = P[Test statistics <= observed value of the test statistic] Example- H_0: μ ≥ 5 and H_1: μ < 5 . This is a left-tailed test since the rejection region would consist of values less than 5.



1. Right tailed test p-value = P [Test statistics >= observed value of the test statistic] $H_0: \mu \leq 39$ $H_1: \mu > 39$. .This is a right-tailed test since the rejection region would consist of values greater than 39.



b. Two tailed test p-value = 2 * P[Test statistics >= |observed value of the test statistic|].



c. test statistic = (x̄ — μ) / (σ / √n)

  i. x̄ = sample mean

  ii. μ = population mean

  iii. σ = Standard Deviation of Population

  iv. n = Number of Observation

d. P-value=P (Observing the test statistic| Null hypothesis is true)

13. F test

  a. $F = \frac{s_1^2}{s_2^2}$ where s is the variance of the 2 samples

  b. Degree of freedoms is $\frac{n_1 = 1}{n_2 - 1}$ where n1 and n2 are sample sizes

14. T test

  a. $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ where μ_0 is population mean and $\bar{x}$ is sample mean and s is smaple's standard deviation

  b. degree of freedom is n-1

  c. Comparing means t test $t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$. sp^2 is the pooled varaince or covaraince

  d. Degree of freedom is n1 + n2 - 2

  e. Comparing 2 related samples $t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$ $\bar{d}$ is Mean of the differences between paired samples and s_d is standard deviation of differences and n is number of pairs. Degree of freedom is n-1

  f. Pooled varaince

    i. $s_{pooled}^2 = [(n_A - 1) * s_A^2 + (n_B - 1) * s_B^2]/(n_A + n_B - 2)$

    ii. Instead of using just s^2_A or s^2_B as your estimate of the common variance, pooling combines information from both groups to provide a more precise and reliable estimate

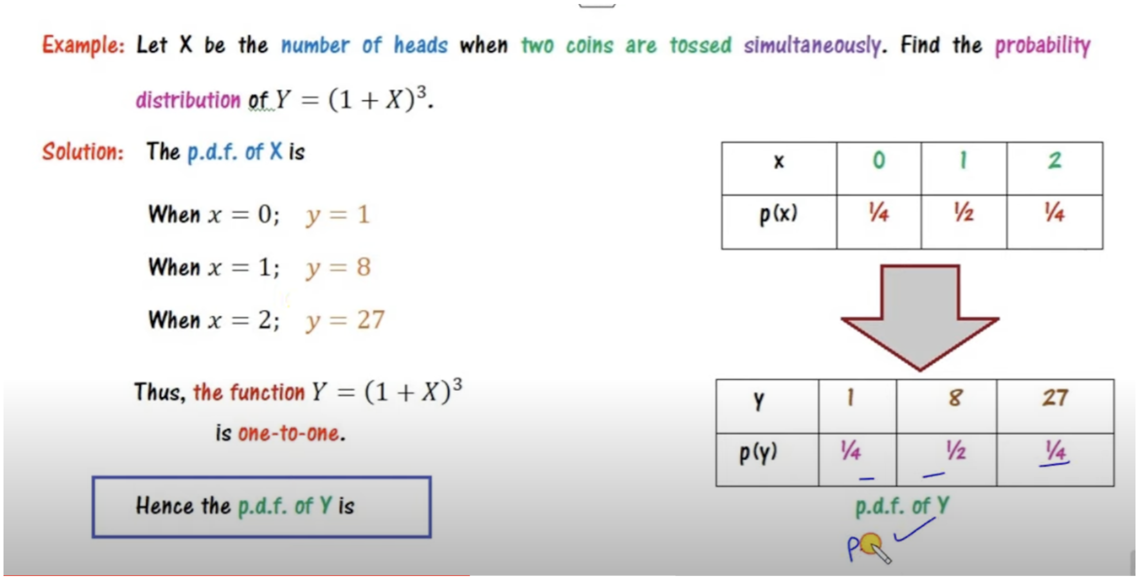
    iii. Improved precision: Pooling combines information from both groups, resulting in a more stable and accurate estimate of the common variance compared to using individual variances.

    iv. **:** Imagine you have data from two groups, group A and group B. You suspect they come from different populations with potentially different means, but you believe they share a common variance. However, you only have estimates of the individual variances (s^2_A and s^2_B) from your samples.

15. Joint discrete distribution
    a. $p(x, y) = P(X = x, Y = y)$
    b. Marginal PMF- $p_X(x) = \sum_y p(x, y)$
    c. Marginal PMF $p_Y(y) = \sum_x p(x, y)$
    d. Joined CDF - $F(x, y) = P(X \leq x, Y \leq y) = \sum_{x' \leq x} \sum_{y' \leq y} p(x', y')$
    e. Conditional mean $E(X|Y = y) = \sum_x x * p(x|y)$
    f. Conditional Variance - $E[(X - E(X|Y = y))^2 | Y = y]$
16. Joint continuous distribution
    a. PDF- $P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x, y)\, dx\, dy$
    b. Marginal PDF $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$
    c. Marginal PDF $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$
    d. Conditional Mean for a given value of y $E(X|Y = y) = \int_{\infty}^{\infty} x * f(x|y) dx$
17. Other formula
    a. Covariance- $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
    b. Correlation - $\rho(X, Y) = Cov(X, Y) / (\sigma_X * \sigma_Y)$ where • σ_X and σ_Y are the standard deviations of X and Y, respectively
        i. **Positive correlation (ρ > 0):** X and Y tend to move in the same direction (both increase or decrease together).
        ii. **Negative correlation (ρ < 0):** X and Y tend to move in opposite directions (one increases while the other decreases).
        iii. **Correlation of 0:** No linear relationship between X and Y.

        iv. **Correlation of 1 or -1:** Perfect linear relationship (all points lie on a straight line).

18. Transformation of hypothesis. Let’s say the transform is y=v(x)
    a. y should be continuous and differentiable
    b. either non decreasing or non increasing for all range of x
    c. $g(x) = f(x(y)).|\frac{dx}{dy}| \Rightarrow g(y) = f(x_1)|\frac{dx_1}{dy}| + f(x_2)|\frac{dx_2}{dy}| + ...$
    d. Proces to do this
        i. Find if the function is continuous and differentiable
        ii. Find if it is strictly increasing / decreasing by evaluating the nature of first derivative
        iii. From y=f(x) find the value of x. Inverse the function
        iv. Differentiate the function wrt y
        v. Find the domain of the differentiated function for variable y, given the domain of variable x from the initial pdf
        vi. Now substitute to the formula we mentioned above and we get the PDF
19. process for the transformation of discrete values
    a. Take the table of PMF which is for every value of x what is the probability
    b. Now transform every x to corresponding y value



Example: Let X be the number of heads when two coins are tossed simultaneously. Find the probability distribution of $Y = (1+X)^3$.

Solution: The p.d.f. of X is

x	0	1	2
p(x)	¼	½	¼

When $x = 0$; $y = 1$

When $x = 1$; $y = 8$

When $x = 2$; $y = 27$

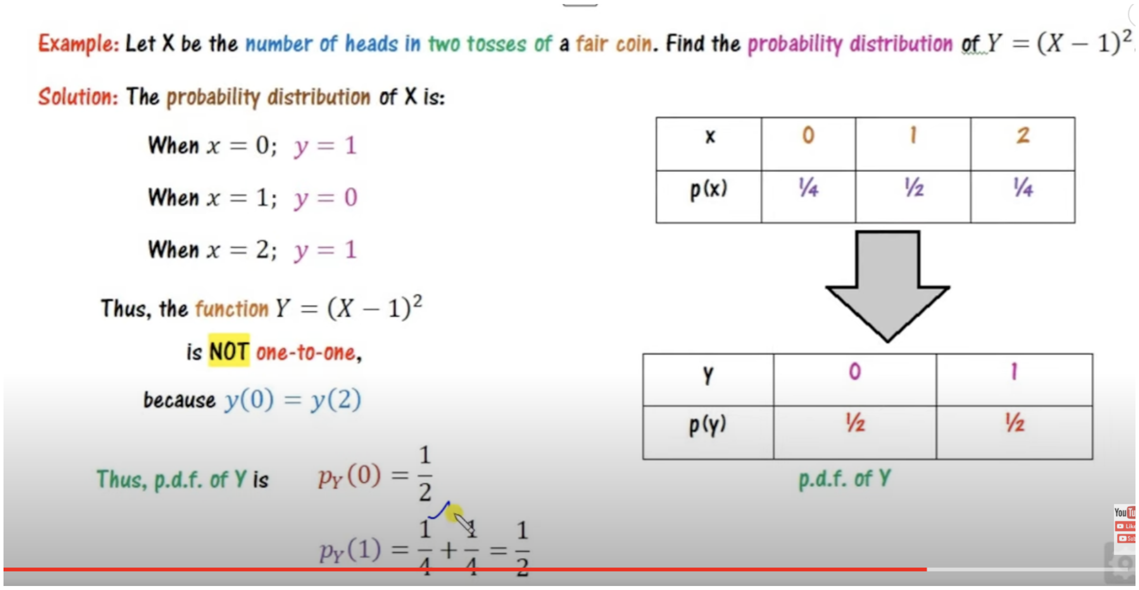
Thus, the function $Y = (1+X)^3$ is one-to-one.

Hence the p.d.f. of Y is

y	1	8	27
p(y)	¼	½	¼

p.d.f. of Y

1. If the value of y is 1-1, we have unique values. If we don’t have 1-1 then it is not unique values of y. In case of 1-1 use the pdf to match with pmf and substitute the value of PMF into the transform depending on whatever value it has. As shown above



Example: Let X be the number of heads in two tosses of a fair coin. Find the probability distribution of $Y = (X-1)^2$.

Solution: The probability distribution of X is:

x	0	1	2
p(x)	¼	½	¼

When $x = 0$; $y = 1$

When $x = 1$; $y = 0$

When $x = 2$; $y = 1$

Thus, the function $Y = (X-1)^2$ is NOT one-to-one, because $y(0) = y(2)$

y	0	1
p(y)	½	½

p.d.f. of Y

Thus, p.d.f. of Y is $p_Y(0) = \frac{1}{2}$

$p_Y(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

1. In case it is not 1-1, you only draw the transformed PMF for unque values. As shown above, in collision values you do the addition of the values of the PMF

20. Transformation of 2 dimensional variales
    a. for a given rv, x and y consider u = u(x,y) and v = v(x,y) are continuous differentiable functions
    b. then JDF of u and v denoted by $g(u,v) = f(x,y)|j|$
    c. J = $\begin{vmatrix} \partial x/\partial u & \partial y/\partial u \\ \partial x/\partial v & \partial y/\partial v \end{vmatrix}$ is called tha jacobian
    d. consider the transform function as u
    e. You can chose v as (if not given) x, y or combination x+y Just so that jacobian ≠0
    f. Calculate the jacobian given these equation.
    g. Substitute x and y as a function of u and v and substitute to the PDF function
    h. Find the domain of u and v given the domain of x and y (and you can use your representation of x and y in terms of x and y) to get domain
    i. Now find marginal density function of u which is integration of substituted function over v and vice versa for v
    j. Now using the domain of u and v find the range of u and v

21. How to calculatye the confidence interval
    a. Know number of samples
    b. Calculate their mean and varaince
    c. $CI = \bar{x} \pm t * (s/\sqrt{n})$
        i. T* is the critical t-value based on the desired confidence level and degrees of freedom (df = n - 1)
        ii. $\bar{x}$ is the sample mean
        iii. s is sample standard deviation
        iv. n is sample size
    d. $CI = \hat{p} \pm z * \sqrt{(\hat{p}(1-\hat{p})/n)}$
        i. p̂ is the sample proportion

   ii. z* is the critical z-value based on the desired confidence level (e.g., 1.96 for a 95% confidence level)

   iii. n is sample size

22. CR inequality

   a. $P(X \geq a) \leq exp(-ta + M(t))$

      i. a and t are real numbers and X is random variable

      ii. P(X ≥ a) ≤ exp(-ta + M(t))

   b. Generalized CR $P(X1 + X2 + ... + Xn \geq a) \leq exp(-ta + n * max(M1(t), M2(t), ..., Mn(t)))$

23. Other probability inequalities

   a. Markov $P(X \geq a) \leq \frac{E(X)}{a}$

   b. Chebyshev's $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$

   c. Booles' $P(\bigcap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} P(A_i^c)$

   d. …

24. Chi SQuare algo

   a. Arrange data into a contingency table with rows and columns representing the different categories of the variables. Each cell in the table contains the observed frequency (count) of data points falling into that combination of categories.

   b. Under the null hypothesis of no association, calculate the expected frequencies for each cell assuming no relationship between the variables.

   c. Compare the observed and expected frequencies using the chi-squared statistic:

      i. $\chi^2 = \Sigma(O - E)^2/E$

      ii. O = observed frequency

      iii. E = expected frequency

      iv. This is summed over all cells

d. Set a significance level (alpha, usually 0.05 or 0.01) to determine the threshold for rejecting the null hypothesis.

e. Calculate the degrees of freedom (df) based on the number of rows and columns in the contingency table: df = (rows - 1) * (columns - 1).

f. Use a chi-squared distribution table or statistical software to find the p-value, the probability of obtaining a chi-squared statistic as extreme or more extreme than the calculated one, assuming the null hypothesis is true.

g. Decision

    i. If the p-value is less than or equal to the significance level, reject the null hypothesis and conclude there's a significant association between the variables.

    ii. If the p-value is greater than the significance level, fail to reject the null hypothesis and conclude there's not enough evidence to support an association.