Notes for online course 18.6501x by Siddharth Begwani (sbegwani@hotmail.com)

Contents

General Concepts

- Raw moments of a normal distribution:
  - Let $X \sim N(\mu, \sigma^2)$. We need to compute $E[X], E[X^2], E[X^3], E[X^4]$
    - The third and fourth moments of a distribution are called skewness and kurtosis. The odd-numbered moments of a normal distribution around the mean are always equal to 0
  - To do this, we can make use of the raw moments of a standard normal distribution $Z \sim N(0,1)$
    - $E[Z] = 0, E[Z^2] = 1, E[Z^3] = 0, E[Z^4] = 3$
  - Since $X = \sigma Z + \mu$
    - $E[X^2] = E[(\sigma Z + \mu)^2] = \int_{-\infty}^{+\infty} (\sigma z + \mu)^2 f_Z(z) dz$
      - Or $E[X^2] = \sigma^2 E[Z^2] + \mu^2 + 2\mu\sigma E[Z] = \sigma^2 + \mu^2$
    - $E[X^3] = E[(\sigma Z + \mu)^3] = \int_{-\infty}^{+\infty} (\sigma z + \mu)^3 f_Z(z) dz$
      - Or $E[X^3] = \sigma^3 E[Z^3] + \mu^3 + 3\sigma^2\mu E[Z^2] + 3\sigma\mu^2 E[Z] = \mu^3 + 3\sigma^2\mu$
    - Similarly, it can be shown that $E[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
- **Hoeffding's inequality**: Hoeffding's inequality can be used to get bounds on the deviation of a random variable (generally, the mean estimator) from its mean even when the number of observations n is not very large. Hoeffding's inequality states that:
  - If $X_1, X_2, \ldots X_n$ are independent and identically distributed random variables drawn from a distribution which is bounded between $[a, b]$ $almost\ surely$ and $\mu = E[X]$, then for any value of n,
    - $P\left[\left|\hat{X}_n - \mu\right| \geq \varepsilon\right] \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \forall\ \varepsilon > 0$ where $\hat{X}_n = \frac{X_1 + X_2 + \cdots X_n}{n}$
  - Hoeffding's inequality gives much more conservative (looser) estimates than the Central Limit Theorem, but it can be used for any value of n
- **Useful Properties of a Normal Distribution**: For a normal distribution, the following properties are applicable:
  - It is invariant under an affine transformation – this simply means that the linear function of a normal random variable also yields a normal random variable
    - If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$
    - This property is not applicable for all distributions. For instance, a linear transformation of a Poisson distribution is not a Poisson distribution.
  - It is possible to standardize a normal random variable. A standardized normal random variable is a linear function of the normal random variable.
    - $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$
    - The standard normal distribution is often represented as Z, i.e. $Z \sim N(0,1)$
  - The normal distribution is symmetrical about its mean. This means that if $X \sim N(0, \sigma^2)$, then $-X \sim N(0, \sigma^2)$
    - This allows us to compute $P(|X| > x)$ as:
      - $P(|X| > x) = P(X > x) + P(-X > x) = 2P(X > x)$
- **Convergence**: If $(T_n)_{n \geq 1}$ is a sequence of random variables and T is a random variable which may be deterministic, there can be three broad types of convergence:

- Almost surely (a.s.) convergence (also known as convergence with probability 1 or strong convergence) – Strongest type of convergence
  - $T_n \underset{n \to \infty}{\overset{a.s.}{\to}} T \; iff \; P[\{\omega: \; T_n(\omega) \to T(\omega)\}] = 1$
- Convergence in probability – This is weaker than almost surely convergence, but stronger than convergence in distribution
  - $T_n \underset{n \to \infty}{\overset{P}{\to}} T \; iff \; P[|T_n - T| \geq \; \varepsilon] \to 0 \; as \; n \to \; \infty \; \forall \varepsilon > 0$
- Convergence in distribution (also known as convergence in law or weak convergence): This is the weakest form of convergence
  - $T_n \underset{n \to \infty}{\overset{d}{\to}} T \; iff \; E[f(T_n)] \to E[f(T)] \; as \; n \to \; \infty$ for all continuous and bounded functions $f(x)$
  - Convergence in distribution also means that:
    - $P(a \leq \; T_n \leq b) \to \; P(a \leq \; T \leq b) \; as \; n \to \; \infty$
    - Converge in distribution actually implies convergence of CDFs
- If a strong form of convergence is satisfied, all weaker forms of convergence will be satisfied
- In the case of almost sure convergence and convergence in probability, we can add, multiply and divide limits of two random variables just as we would work with numbers
  - If $T_n \underset{n \to \infty}{\overset{a.s./P}{\to}} T$ and $U_n \underset{n \to \infty}{\overset{a.s./P}{\to}} U$ then:
    - $T_n + U_n \underset{n \to \infty}{\overset{a.s./P}{\to}} T + U$
    - $T_n U_n \underset{n \to \infty}{\overset{a.s./P}{\to}} TU$
    - $\frac{T_n}{U_n} \underset{n \to \infty}{\overset{a.s./P}{\to}} \frac{T}{U}$ if $U \neq 0$
  - However, the above identities do not hold for two random variables that converge in distribution
  - Slutsky's theorem allows us to perform mathematical operations on two random variables, one of which converges in distribution and the other converges almost surely or in probability. If $T_n \underset{n \to \infty}{\overset{d}{\to}} T$ and $U_n \underset{n \to \infty}{\overset{P}{\to}} u$ where u is a given real number, then:
    - $T_n + U_n \underset{n \to \infty}{\overset{a.s./P}{\to}} T + u$
    - $T_n U_n \underset{n \to \infty}{\overset{a.s./P}{\to}} Tu$

- $\frac{T_n}{U_n} \underset{n \to \infty}{\overset{a.s./P}{\to}} \frac{T}{u}$ if $u \neq 0$

- **Continuous mapping theorem**: The continuous mapping theorem says that for all three types of convergence, if f is a continuous function:
  - If $T_n \to T \ as \ n \to \infty$, then $f(T_n) \to f(T) \ as \ n \to \infty$
  - The function need not be continuous throughout the real-number line; it needs to be continuous around the vicinity of values for which we use the above convergence equality
- **Statistical models**: Let the observed outcome of a statistical experiment be a sample $X_1, X_2, .. X_n$ which are independent and identically distributed random variables in some measurable space E (usually E is a subset of the set of all real numbers, R). Let P denote the common distribution of the identically distributed random variables.
  - A statistical model associated with the statistical experiment is a pair $(E, (P_\theta)_{\theta \in \Theta})$
    - E is the sample space
    - $(P_\theta)_{\theta \in \Theta}$ is a family of probability distributions defined on E
    - Θ is the parameter set
  - The sample space of a random variable may not be unique. For instance, {0,1} is the sample space of a Bernoulli random variable, but the whole of the real number line can also be the sample space (since both 0 and 1 are part of it). Generally, when referring to the sample space of a random variable, we refer to the smallest sample space.
    - Also, a sample space cannot depend on a parameter. For instance, if we have a uniform distribution between $[0, a]$ where a is unknown, the sample space cannot be $[0, a]$. It would have to be R_+.
  - Some examples of the notation for statistical parametric models are:
    - Bernoulli random variable: $(\{0,1\}, Ber(p)_{p \in (0,1)})$ – the parameter set excluded the values 0 and 1 for which the distribution becomes degenerate
    - Random normal variable: $(R, N(\mu, \sigma^2)_{(\mu, \sigma^2) \in R x (0,\infty)})$
    - Poisson random variable: $(N, Poisson(\lambda)_{\lambda \in (0,\infty)})$
  - For a non-parametric statistical model, the set of probability distributions is infinitely large (all probability distributions defined on E) and is not indexed by any parameters.
    - For instance, if we have a unimodal distribution, it can take on any shape. All we know is that the pdf of the distribution will be a rising curve in $(-\infty, a)$ and a falling curve in (a, ∞) for some $a \in R$. For this distribution, the parameter set is the set of all pdfs which give a unimodal distribution
  - A linear regression model has in its sample space pairs of variables $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n) \in R^d \ x \ R$. Here d is the number of explanatory variables.
    - The regression model is $Y_i = \beta^T X_i + \varepsilon_i$ where $\beta \in R^d$. β is the parameter matrix
      - If we assume that the explanatory variables follow a multivariate Gaussian distribution, and all the explanatory variables are independent and each follows a standard normal distribution, then $X_i \sim N_d(0, I_d)$ where I_d is the covariance matrix which in this case will simply be the identity matrix in R^d
      - ε_i is noise which can be considered to be Gaussian. It is independent of X_i and $\varepsilon_i \sim N(0,1)$

- Identifiability of parameters. The parameter θ is called identifiable if and only if the map $\theta \in \Theta \mapsto P_\theta$ is injective (one-to-one mapping) This means that if $\theta \neq \theta'$, then $P_\theta \neq P_{\theta'}$ or alternatively, if $P_\theta = P_{\theta'}$, then $\theta = \theta'$
  - Consider a case where we have a set of independent and identically distributed random variables $Y_1, Y_2, .. Y_n \sim N(\mu, \sigma^2)$. However, what we observe is a set of random variables $X_1, X_2, .. X_n$ where $X_i = I(Y_1 > 0)$
  - The random variable X_i is a Bernoulli random variable, and the parameter p is $P(Y_i > 0) = 1 - P(Y_i \leq 0) = 1 - \Phi(\frac{-\mu}{\sigma})$
    - Or, $\frac{\mu}{\sigma} = -\Phi^{-1}(1-p)$
  - With the given set of observations, we can determine the value of $\frac{\mu}{\sigma}$ uniquely but not the value of both μ and σ
  - The model $(\{0,1\}, Bernoulli\left(1 - \Phi\left(\frac{-\mu}{\sigma}\right)\right)_{\mu,\sigma^2 \in Rx(0,\infty)})$ does not have identifiable parameters. The model $(\{0,1\}, Bernoulli\left(1 - \Phi\left(\frac{-\mu}{\sigma}\right)\right)_{\frac{\mu}{\sigma} \in R})$ has identifiable parameters

- A statistic is any measurable function of the sample space. For instance, $\bar{X}_n$, $\max_i X_i$, $Var(X_i)$ are all measurable functions and hence valid statistics
- An estimator of θ is a statistic that does not depend on θ. The estimator should not depend on any unknown parameters.
- An estimator $\hat{\theta}_n$ of θ is asymptotically normal if $\sqrt{n}(\hat{\theta}_n - \theta) \underset{n \to \infty}{\overset{d}{\to}} N(0, \sigma^2)$ and σ^2 is called the asymptomatic variance of $\hat{\theta}_n$. When defined in this way, the asymptotic variance is not the limit of $Var(\hat{\theta}_n)$ as $\to \infty$; this limit will be equal to 0.
- **Cox-Proportional Hazard Model**: This is a semi-parametric model used for survival analysis. A semi-parametric model is a product of two variables, one of which is parametric and the other is non-parametric. Generally, the non-parametric variable is considered to be a nuisance function, and we are not interested in computing this exactly
  - Let $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n) \in R^d \, x \, R$. Here d is the number of explanatory variables.
  - The conditional distribution of Y given $X = x$ has the form:
    - $F(t) = 1 - \exp(-\int_0^t h(u) e^{\beta^T x} du)$
      - Where h is an unknown non-negative nuisance function and $\beta \in R^d$ is the parameter of interest
- **Jensen's inequality**: Jensen's inequality for probability distributions says that if X is a random variable and f is a function, then:
  - $f(E[X]) \leq E(f(X))$ if f is a convex function
  - $f(E[X]) \geq E(f(X))$ if f is a concave function
  - Thus, for example:
    - $\sqrt{E[X]} \geq E[\sqrt{X}]$ since the square root is a concave function
    - $E[X]^2 \leq E[X^2]$

- The bias of an estimator $\hat{\theta}_n$ of θ is defined as $E[\hat{\theta}_n] - \theta$, and the quadratic risk of an estimator can be computed as $E[(\hat{\theta}_n - \theta)^2]$ which is equal to Variance of the estimator + bias squared

Confidence Intervals and Hypothesis Testing

- **Computing confidence intervals for a Bernoulli distribution:** Let us consider independent and identically distributed variables $X_1, X_2, .. X_n$ where $X_i = Ber(p)$
  - Let the mean estimator be $\bar{X}_n$
  - If we want to estimate p with a confidence level of $1-\alpha$, we can use the central limit theorem given $\frac{X_n - p}{(\frac{\sqrt{p(1-p)}}{\sqrt{n}})}$ is a standard normal variable
  - For a standard normal variable Z,
    - $P(|Z| \leq k) = P(-k \leq Z \leq k) = 1 - 2P(Z \geq k)$
    - Or, $P(|Z| \leq k) = 2P(Z \leq k) - 1 = 2\Phi(\mathrm{k}) - 1$
  - Therefore, in the equation $P(|\frac{X_n - p}{(\frac{\sqrt{p(1-p)}}{\sqrt{n}})}| \leq z) \geq 1-\alpha$, given a value of α, $z \geq \Phi^{-1}(1-\frac{\alpha}{2})$
    - $\Phi^{-1}(1-\frac{\alpha}{2})$ is by definition, $q_{\alpha/2}$ – this is actually the $(1-\frac{\alpha}{2})$ quantile of the standard normal equation
  - Hence, $P(|\frac{X_n - p}{(\frac{\sqrt{p(1-p)}}{\sqrt{n}})}| \leq q_{\alpha/2}) \geq 1-\alpha$ or,
    - $P\left(\bar{X}_n - \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}} \leq p \leq \bar{X}_n + \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}}\right) \geq 1-\alpha$
  - Here the confidence interval bounds depend on the unknown parameter p
  - A conservative bound would involve taking the maximum value of $\sqrt{p(1-p)}$ which is ½ would give a conservative confidence interval
  - A more accurate confidence interval can be constructed if we consider that:
    - $\bar{X}_n - \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}} \leq p \leq \bar{X}_n + \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}}$ implies
      - $\bar{X}_n - \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}} \leq p$ and $p \leq \bar{X}_n + \frac{\sqrt{p(1-p)}\, q_{\alpha/2}}{\sqrt{n}}$
      - Hence, the solutions of the quadratic equation $(p - \bar{X}_n)^2 \leq \frac{p(1-p)\, q^2_{\alpha/2}}{n}$ will give us the bounds of the confidence interval
    - The confidence interval bounds can be found by solving the quadratic equation:
      - $\left(1 + \frac{q^2_{\alpha/2}}{n}\right)p^2 - \left(2\bar{X}_n + \frac{q^2_{\alpha/2}}{n}\right)p + \bar{X}_n{}^2 \leq 0$ (Since the coefficient of p^2 is positive, this is a convex equation)
      - If the solutions to the above equation are p_1 and p_2, then $[p_1, p_2]$
- **One-sided confidence interval**: Sometimes, we are only interested that the estimator is greater than or lesser than the true value of the parameter with a certain probability. If $\hat{\theta}_n$ is the estimator and θ is the estimand, we might be interested in computing:
  - The value of z such that $P(\hat{\theta}_n - \theta \leq z) \geq 1-\alpha$ (The interval is bounded on the upper end, and is of the form $[-\infty, z]$
  - The value of z such that $P(\theta - \hat{\theta}_n \geq z) \geq 1-\alpha$ which can be rewritten as $P(\hat{\theta}_n - \theta \leq -z) \geq 1-\alpha$

- In case the estimator is the sample average and the unknown parameter is the mean of the underlying distribution of the sample, the values of z can be computed using the central limit theorem

- **Confidence intervals centered around the estimator**: There can be multiple confidence intervals constructed for a given set of data
  - As an example, if we have the mean estimator $\bar{X}_n$ based on independent and identically distributed Bernoulli samples with the true value of the parameter being p, and the realized value of $\bar{X}_n$ is 0.5, then for a 50% confidence interval:
    - $\frac{\bar{X}_n - p}{\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)}$ is a standard normal variable $Z \sim N(0,1)$
    - $P(Z \geq 0) = 0.5 \rightarrow P\left(\frac{\bar{X}_n - p}{\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)} \geq 0\right) = 0.5$ or $P(\bar{X}_n - p \geq 0) = 0.5$
      - Or, $P(\bar{X}_n \geq p) = 0.5$ and hence $[\bar{X}_n, 1]$ is a 50% confidence interval. Given that $\bar{X}_n = 0.5$ and p must lie between 0 and 1, $[0.5,1]$ is a 50% confidence interval for p
    - Also, $P(Z \leq 0) = 0.5$ and thus $[0, \bar{X}_n]$ is also a 50% confidence interval for p. Given that $\bar{X}_n = 0.5$, $[0,0.5]$ is a 50% confidence interval for p
  - If the estimator follows a Gaussian distribution, the Gaussian confidence interval centered at the estimator can be shown to be the confidence interval with the least width for a given confidence level

- **The Delta Method**: In some cases, the mean estimator or any linear function of the mean estimator does not enable us to obtain confidence intervals for the parameter required
  - As an example, consider an exponential distribution. If we draw n independent and identically distributed samples from it, the mean estimator $\bar{X}_n$ converges asymptotically (by the central limit theorem) to $1/\lambda$ where λ is the parameter of the exponential distribution
    - Also, by Jensen's inequality, since $f(x) = 1/x$ is a convex function,
      - $E\left[\frac{1}{\bar{X}_n}\right] \geq \lambda$
  - In such cases, we can use the Delta method
    - Let $(Z_n)_{n \geq 1}$ is a sequence of random variables that satisfies
      - $\sqrt{n}(Z_n - \theta) \underset{n \to \infty}{\overset{d}{\rightarrow}} Z(0, \sigma^2)$ for some $\theta \in R$ (the sequence $(Z_n)_{n \geq 1}$ is said to be asymptotically normal around θ)
    - Let $g: R \rightarrow R$ be continuously differentiable at the point θ. Then $g((Z_n)_{n \geq 1})$ is also continuously differentiable at the point θ
    - The first order expansion of Taylor's theorem gives us that:
      - $g(Z_n) - g(\theta) \cong (Z_n - \theta) g'(\theta)$
    - Thus,
      - $\sqrt{n}\big(g(Z_n) - g(\theta)\big) \cong \sqrt{n}\,(Z_n - \theta) g'(\theta) \underset{n \to \infty}{\overset{d}{\rightarrow}} Z(0, (g'(\theta))^2 \sigma^2)$
  - In the case of the mean estimator for the parameter of the exponential distribution, we want the function g such that $g(x) = 1/x$ and $g'(x) = -1/x^2$

- Then, $\sqrt{n}\left(g(\bar{X}_n) - g\left(1/\lambda\right)\right) = \sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right)$
- By the delta method, $\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \cong \sqrt{n}\left(\bar{X}_n - 1/\lambda\right) g'\left(1/\lambda\right)$
- Since $g'\left(1/\lambda\right) = -\lambda^2$, this means that $\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \cong -\lambda^2\sqrt{n}\left(\bar{X}_n - 1/\lambda\right)$ –(A)
- We know that $\sqrt{n}\left(\bar{X}_n - 1/\lambda\right) \underset{n \to \infty}{\overset{d}{\to}} Z(0, \frac{1}{\lambda^2})$
  - Hence $-\lambda^2\sqrt{n}\left(\bar{X}_n - 1/\lambda\right) \underset{n \to \infty}{\overset{d}{\to}} Z(0, \lambda^2)$ – (B)
- From (A) and (B), we have:
  - $\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \underset{n \to \infty}{\overset{d}{\to}} Z(0, \lambda^2)$
- We can now find the confidence interval at a significance level α
  - $\frac{\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right)}{\lambda} = Z(0,1)$ and hence
  - $P\left(\frac{1}{\bar{X}_n} - \frac{q\alpha/2\lambda}{\sqrt{n}} \leq \lambda \leq \frac{1}{\bar{X}_n} + \frac{q\alpha/2\lambda}{\sqrt{n}}\right) \geq 1 - \alpha$
- For the confidence interval, the bounds include the parameter to be estimated. We cannot use a conservative bound in this case since the upper bound of λ is infinity. We can replace λ with its estimator $\frac{1}{\bar{X}_n}$ in the bounds. Alternatively, we can solve exactly for λ. To solve exactly for λ, we observe that:
  - $\lambda\left(1 + \frac{q\alpha/2}{\sqrt{n}}\right) \geq \frac{1}{\bar{X}_n}$, or $\lambda \geq \frac{1}{\bar{X}_n}\left(1 + \frac{q\alpha/2}{\sqrt{n}}\right)^{-1}$
  - Also, $\lambda\left(1 - \frac{q\alpha/2}{\sqrt{n}}\right) \leq \frac{1}{\bar{X}_n}$, or $\lambda \leq \frac{1}{\bar{X}_n}\left(1 - \frac{q\alpha/2}{\sqrt{n}}\right)^{-1}$
  - Thus, the $1 - \alpha$ confidence interval for λ is:
    - $[\frac{1}{\bar{X}_n}\left(1 + \frac{q\alpha/2}{\sqrt{n}}\right)^{-1}, \frac{1}{\bar{X}_n}\left(1 - \frac{q\alpha/2}{\sqrt{n}}\right)^{-1}]$

- **Shifted exponential distribution**: A shifted exponential distribution is an exponential distribution that has been shifted right by an amount $a > 0$.
  - If the exponentially shifted random variable is X, the pdf of such a distribution $f_X(x)$ is given by:
    - $f_X(x) = 0 \; \forall \; x < a$
    - $f_X(x) = \lambda e^{-\lambda(x-a)} \; \forall \; x \geq a$
  - One way of calculating the confidence interval for the parameter a is to use the sample mean as the estimator $\bar{X}_n$. By the weak law of large numbers, the sample mean estimator will converge in probability to the mean of the distribution underlying the independent and identically distributed samples
    - $E[X] = \int_a^\infty x\lambda\, e^{-\lambda(x-a)}\, dx$
      - Using integration by parts:
        - $\int_a^\infty x\lambda\, e^{-\lambda(x-a)} dx = [-e^{-\lambda(x-a)}\, x]_a^\infty - \int_a^\infty -e^{-\lambda(x-a)} dx$, or
        - $\int_a^\infty x\lambda\, e^{-\lambda(x-a)}\, dx = a - [\frac{1}{\lambda} e^{-\lambda(x-a)}]_a^\infty = a + \frac{1}{\lambda}$

- It may be noted that for the term $-e^{-\lambda(x-a)}x$, the decay of the exponential term is much faster than the growth of the polynomial term, and hence as $x \to \infty$, the whole expression goes to 0.

- Similarly, $E[X^2] = \int_a^\infty x^2 \lambda e^{-\lambda(x-a)}$
  - Using integration by parts, we have:
    - $\int_a^\infty x^2 \lambda e^{-\lambda(x-a)} = [-e^{-\lambda(x-a)}x^2]_a^\infty - \int_a^\infty -2xe^{-\lambda(x-a)}\, dx$, or
    - $\int_a^\infty x^2 \lambda e^{-\lambda(x-a)} = a^2 + \frac{2}{\lambda}\int_a^\infty x\lambda\, e^{-\lambda(x-a)} dx$, or
    - $\int_a^\infty x^2 \lambda e^{-\lambda(x-a)} = a^2 + \frac{2}{\lambda}E[X] = a^2 + \frac{2}{\lambda}(a + \frac{1}{\lambda})$
- Thus, $Var[X] = E[X^2] - E[X]^2 = a^2 + \frac{2}{\lambda}\left(a + \frac{1}{\lambda}\right) - (a + \frac{1}{\lambda})^2$, or
  - $Var[X] = \frac{1}{\lambda^2}$
- Using the central limit theorem:
  - $\frac{\bar{X}_n - (a + \frac{1}{\lambda})}{(1/\sqrt{n}\lambda)} \to N(0,1)$
  - The confidence interval $1 - \alpha$ for a is given by:
    - $[\bar{X}_n - \frac{1}{\lambda} - \frac{q\alpha/2}{\lambda\sqrt{n}}, \bar{X}_n - \frac{1}{\lambda} + \frac{q\alpha/2}{\lambda\sqrt{n}}]$

- Another approach to construct a confidence interval is to use the estimator as the minimum of the sample observations. This allows construction of a one-sided confidence interval whose width is less than that of a two-sided confidence interval. A one-sided confidence interval is useful when a parameter can take on only positive or only negative values
  - Let the estimator be $\hat{X} = Min(X_1, X_2, \ldots X_n)$
  - If our confidence interval is of the form $[\hat{X} - s, \hat{X}]$ and the significance level is α, then $P(\hat{X} - s \le a \le \hat{X}) = 1 - \alpha$ which can be rewritten as:
    - $P(\hat{X} - s \le) = 1 - \alpha$ (we need to find the value of s) – (A)
  - $P(\hat{X} > t) = 0 \; \forall \; t < a$
    - Hence, $f_{\hat{X}}(t) = 0 \; \forall \; t < a$
  - $P(\hat{X} > t) = \prod_{i=1}^n \int_t^\infty \lambda e^{-\lambda(x-a)} dx = e^{-n\lambda(t-a)} \; \forall \; t \ge a$
    - Hence, $P(\hat{X} \le t) = 1 - e^{-n\lambda(t-a)}$
    - Or, $f_{\hat{X}}(t) = n\lambda e^{-n\lambda(t-a)}$ – (B)
  - Also, from the expression $P(\hat{X} > t) = e^{-n\lambda(t-a)} \; \forall \; t \ge a$,
    - If we write $\tilde{t} = n\lambda(t - a)$, then $\tilde{t} \ge 0$
    - $P(\hat{X} > t) = P\left(\hat{X} > \frac{\tilde{t}}{n\lambda} + a\right) = e^{-\tilde{t}}$, or $P(n\lambda(\hat{X} - a) > \tilde{t}) = e^{-\tilde{t}}$
    - It can be easily seen that the variable $n\lambda(\hat{X} - a)$ follows an exponential distribution with parameter 1
  - The distribution of $n\lambda(\hat{X} - a)$ for $\hat{X} \ge a$ can also be derived by considering
    - $Y = n\lambda(\hat{X} - a)$ where Y is a monotonically increasing function of $\hat{X}$
    - Hence $P(Y \le y) = P(\hat{X} \le x)$ *where* $x = {}^{y}/_{n\lambda} + a$
    - Hence, by differentiating the CDFs using (B), we get:

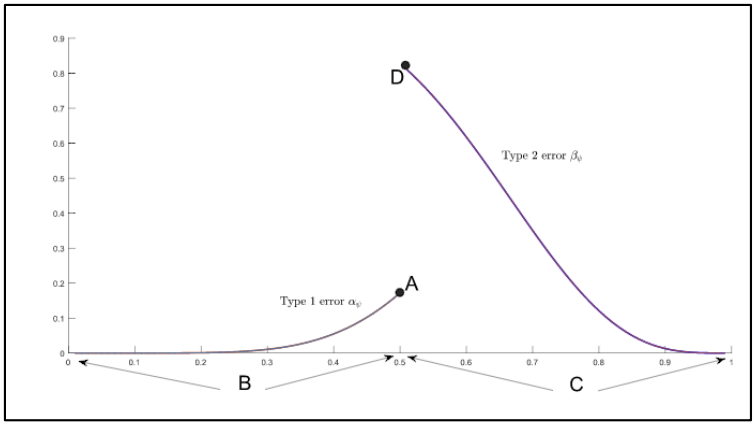
$$\circ \quad f_Y(y) = f_{\hat{X}}\left({}^{y}/_{n\lambda} + a\right) * {}^{1}/_{n\lambda} = e^{-y}$$

- Hence the distribution of $n\lambda(\hat{X} - a)$ follows an exponential distribution with parameter 1 for $\hat{X} \geq a$ which is an inequality which will always be true given the choice of the estimator
- From (A), we need to find the value of s such that:
  - $P(\hat{X} - s \leq a) = 1 - \alpha$ which is equivalent to :
  - $P(\hat{X} - a \leq s) = 1 - \alpha$, or $P(n\lambda(\hat{X} - a) \leq n\lambda s) = 1 - \alpha$
  - Since $n\lambda(\hat{X} - a)$ is an exponential distribution with parameter 1:
    - $P(n\lambda(\hat{X} - a) \leq n\lambda s) = 1 - e^{-n\lambda s}$
  - Thus, $1 - e^{-n\lambda s} = 1 - \alpha$, or $n\lambda s = \ln(\frac{1}{\alpha})$, or $s = \ln\left(\frac{1}{\alpha}\right)/(n\lambda)$
- Thus, the confidence interval desired is $[\hat{X} - \frac{\ln\left(\frac{1}{\alpha}\right)}{n\lambda}, \hat{X}]$

- If we compare the two confidence intervals:
  - The mean estimator uses asymptotic properties as it is based on the Central Limit Theorem. The size of the confidence interval is of the order $\sqrt{n}$
  - The minimum estimator is an approach specifically tailored to the problem, and is non-asymptotic. The size of the confidence interval is smaller than when using the mean estimator. The size of the confidence interval is of the order n and the interval shrinks more rapidly with increasing n than the confidence interval based on the mean estimator
  - The mean estimator is a more robust distribution (at least to bounded outliers) than the minimum estimator. The minimum estimator should be used if we are clear that the estimand is guaranteed to be positive.

- **Hypothesis Testing**: In hypothesis testing, the null hypothesis H_0 and the alternative hypothesis H_1 are not symmetrical in their treatment. Specifically, the data is used to try and disprove H_0; it cannot be used to prove H_0
  - A hypothesis test can be described by an indicator function ψ (which does not depend on the unknown parameter) and whose output is either 0 or 1 such that:
    - $\psi(X_1, X_2, \ldots X_n) \in \{0,1\}$
      - If $\psi = 0$, H_0 is not rejected
      - If $\psi = 1$, H_0 is rejected
    - The parameter space for the null hypothesis is denoted as Θ_0 and for the alternative hypothesis, it is defined as Θ_1
    - The type I error can be defined as:
      - $\sup_{\theta} \alpha_{\psi}(\theta)$ where $\alpha_{\psi}(\theta) = P(\psi(X_1, X_2, \ldots X_n) = 1)\ for\ \theta \in \Theta_0$
    - The type II error can be defined as:
      - $\sup_{\theta} \beta_{\psi}(\theta)$ where $\beta_{\psi}(\theta) = P(\psi(X_1, X_2, \ldots X_n) = 0)\ for\ \theta \in \Theta_1$
  - A one-sided test has both the hypothesis as composite hypothesis: For instance, $H_0: \mu \leq 0.5, H_1: \mu > 0.5$ is a one-sided test. A two-sided test has one hypothesis as a simple hypothesis and the other as a composite hypothesis. For instance, $H_0: \mu = 0.5, H_1: \mu \neq 0.5$ is a two-sided test

- In the case of a one-sided test, the significance level or the Type 1 error is the upper bound on the worst-case probability of making an error under the null hypothesis. As an example, consider a hypothesis test where the sample comprises Bernoulli random variables.
  - We define $H_0{:}p \geq 0.33$ and $H_1{:}p < 0.33$ where p is the actual unknown parameter of the Bernoulli distribution
  - The sample average $\hat{p}$ has a Gaussian distribution for a large sample size (from the Central Limit theorem)
  - We would reject the null hypothesis if $\hat{p}$ was significantly less than 0.33. We can use as a test statistic $\frac{\sqrt{n}\,(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$ which would be a Gaussian distribution. Though we do not know the exact value of p, we know that under the null hypothesis, the value of p is in the interval [0.33,1). p_0 is a value in that interval. The objective is to determine the value of p_0 to choose. The correct value in this case would be the boundary value 0.33
    - This is because if we have $P\left(\frac{\sqrt{n}\,(\hat{p}-0.33)}{\sqrt{0.33(1-0.33)}} > C\right) \geq 1-\alpha$ where C is a negative value given the nature of the one-sided hypothesis test, then
      - The distribution of $\frac{\sqrt{n}\,(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$ for any p_0 in the interval $(0.33,1)$ is to the right of the distribution $\frac{\sqrt{n}\,(\hat{p}-0.33)}{\sqrt{0.33(1-0.33)}}$
      - Hence $P\left(\frac{\sqrt{n}\,(\hat{p}-0.33)}{\sqrt{0.33(1-0.33)}} > C\right) \geq 1-\alpha \;\Rightarrow\; P\left(\frac{\sqrt{n}\,(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}} > C\right) \geq 1-\alpha$
      - Clearly, $C = q_{1-\alpha}$
      - Hence, our criterion of rejection of the null hypothesis is:
        - $\frac{\sqrt{n}\,(\hat{p}-0.33)}{\sqrt{0.33(1-0.33)}} < q_{1-\alpha}$
  - As another example, consider that for a sample involving Bernoulli random variables, the null hypothesis and alternative hypothesis are $H_0{:}p \leq 0.5$ and $H_1{:}p > 0.5$
    - The test statistic will be of the form:
      - $\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{0.5(1-0.5)}}$
    - The rejection criteria with a significance level α will be:
      - $\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{0.5(1-0.5)}} > q_{\alpha}$
    - The graphs of the two types of errors is as shown below for $\alpha = 20\%$:



-
- B represents the region of the null hypothesis, and C represents the region of the alternative hypothesis.

Distance Measures between Distributions

- **Total Variation Distance**: Let a statistical model be defined by $(E, (P_\theta)_{\theta \in \Theta})$ for a set of independent and identically distributed random variables $X_1, X_2, .. X_n$. If $\theta^* \in \Theta$ is the true parameter, then the goal of the estimator is to ensure $P_{\hat{\theta}}$ is close to P_{θ^*} for each point in the sample space E.
  - The total variation distance between two probability measures P_θ and $P_{\theta'}$ is defined as:
    - $TV\ (P_\theta, P_{\theta'}) = \max_{A \in E} |P_\theta(A) - P_{\theta'}(A)|$
  - For discrete distributions, the total variation distance is given by:
    - $TV\ (P_\theta, P_{\theta'}) = \frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|$
    - It can be easily shown that the maximum value of $|p_\theta(x) - p_{\theta'}(x)|$ must be at least as large as $\frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|$
      - Let us consider the set $A = \{x \in E: p_\theta(x) \geq p_{\theta'}(x)\}$
      - $P_\theta(A) - P_{\theta'}(A) = \sum_{x:\, p_\theta(x) \geq p_{\theta'}(x)} (p_\theta(x) - p_{\theta'}(x))$ which can also be written as:
        - $|P_\theta(A) - P_{\theta'}(A)| = \sum_{x:\, p_\theta(x) \geq p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$ - (A)
      - Considering the complement of the set A, we have:
        - $P_\theta(A^C) - P_{\theta'}(A^C) = \sum_{x:\, p_\theta(x) < p_{\theta'}(x)} (p_\theta(x) - p_{\theta'}(x))$, or
        - $P_\theta(A^C) - P_{\theta'}(A^C) = -\sum_{x:\, p_\theta(x) < p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$ since in the summation, each of the elements is negative
        - Since the right hand side in the above equation is negative, $P_\theta(A^C) - P_{\theta'}(A^C)$ is negative
        - $-|P_\theta(A^C) - P_{\theta'}(A^C)| = -\sum_{x:\, p_\theta(x) < p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$, or
        - $-|P_{\theta'}(A) - P_\theta(A)| = -\sum_{x:\, p_\theta(x) < p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$, or
        - $-|P_\theta(A) - P_{\theta'}(A)| = -\sum_{x:\, p_\theta(x) < p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$, or
        - $|P_\theta(A) - P_{\theta'}(A)| = \sum_{x:\, p_\theta(x) < p_{\theta'}(x)} |p_\theta(x) - p_{\theta'}(x)|$ - (B)
      - Adding corresponding sides of (A) and (B), we get:
        - $|P_\theta(A) - P_{\theta'}(A)| = \frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|$
  - For continuous distributions, the total variation distance is given by:
    - $TV\ (P_\theta, P_{\theta'}) = \frac{1}{2} \int_{x \in E} |f_\theta(x) - f_{\theta'}(x)|\, dx$
  - The total variation distance (TV) satisfies the four criteria mathematically required for a distance function:
    - Symmetric: $TV\ (P_\theta, P_{\theta'}) = TV\ (P_{\theta'}, P_\theta)$
    - Non-negative: $TV\ (P_\theta, P_{\theta'}) \geq 0$
      - The upper bound on $TV\ (P_\theta, P_{\theta'})$ is equal to 1. This is because, by the triangle inequality:
        - $\frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)| \leq \frac{1}{2} (\sum_{x \in E} |p_\theta(x)| + \sum_{x \in E} |p_{\theta'}(x)|) \leq 1$ (the same logic holds for continuous variables)
    - Definite: $TV\ (P_\theta, P_{\theta'}) = 0 \Rightarrow P_\theta(A) = P_{\theta'}(A)\ \forall\ A \in E$ where E is the common sample space for P_θ and $P_{\theta'}$
    - Triangle inequality: $TV\ (P_\theta, P_{\theta'}) \leq TV\ (P_\theta, P_{\theta''}) + TV\ (P_{\theta''}, P_{\theta'})$

- This is because:
  - $\frac{1}{2}\sum_{x\in E}|p_\theta(x)-p_{\theta'}(x)| = \frac{1}{2}\sum_{x\in E}|p_\theta(x)-p_{\theta''}(x)+p_{\theta''}(x)-p_{\theta'}(x)|$, or by the triangle inequality:
  - $\frac{1}{2}\sum_{x\in E}|p_\theta(x)-p_{\theta'}(x)| \leq \frac{1}{2}\sum_{x\in E}|p_\theta(x)-p_{\theta''}(x)| + \frac{1}{2}\sum_{x\in E}|p_{\theta''}(x)-p_{\theta'}(x)|$, or
  - $TV(P_\theta, P_{\theta'}) \leq TV(P_\theta, P_{\theta''}) + TV(P_{\theta''}, P_{\theta'})$

- In many cases, we might have to compute the total variation distance for two probability functions which may not have the same sample space.
  - As an example, consider $P_\theta \sim Exp(1)$ and $P_{\theta'} \sim Unif(0,1)$
  - $TV(P_\theta, P_{\theta'}) = \frac{1}{2}\int_{x\in R}|f_\theta(x)-f_{\theta'}(x)|\,dx$
    - $f_\theta(x) = e^{-1x}\ \forall\ x \in [0,\infty]$, and $f_{\theta'}(x) = 1\ x \in [0,1]$
  - Therefore $TV(P_\theta, P_{\theta'}) = \frac{1}{2}\int_0^1|e^{-x}-1|\,dx + \frac{1}{2}\int_1^\infty|e^{-x}|\,dx$
    - $e^{-x}-1 \leq 0\ \forall\ x \in [0,\infty]$ and $e^{-x} \geq 0\ \forall\ x \in [0,\infty]$
  - Hence, $TV(P_\theta, P_{\theta'}) = \frac{1}{2}\int_0^1(1-e^{-x})\,dx + \frac{1}{2}\int_1^\infty e^{-x}\,dx$, or
  - $TV(P_\theta, P_{\theta'}) = \frac{1}{2}(1+e^{-1}-1) + \frac{1}{2}e^{-1} = e^{-1}$
- For two probability distributions which have disjoint support (i.e. the random variables in the two distributions can never take the same value), the total variation distance will be equal to 1.
  - Let us consider the case of two distributions: P_θ which is the distribution of the mean estimator $\bar{X}_n$ where X_i is Bernoulli with parameter 0.5, and $P_{\theta'} = N(0,1)$
  - Asymptotically, as $n \to \infty$, the distribution of $2\sqrt{n}(\bar{X}_n - \frac{1}{2})$ will converge to that of a standard normal distribution
  - However, for any finite n, $\bar{X}_n$ has a discrete distribution which can take the values in the set $\{\frac{0}{n}, \frac{1}{n}, \ldots \frac{n}{n}\}$. If we call this set S, then from the definition of total variation distance:
    - $TV(P_\theta, P_{\theta'}) = \max_A |P_\theta(A) - P_{\theta'}(A)| = 1$ for $A = S$
    - This is because the probability of a point value for the continuous normal distribution is always equal to 0.

- **Kullback-Leibler divergence**: The total variation distance does not suggest a method for producing an estimator of θ which can be minimized in order to minimize the distance from the true value.
  - The Kullback-Leibler (KL) divergence (also called relative entropy) between two probability measures P_θ and $P_{\theta'}$ is given by:
    - $$KL(P_\theta, P_{\theta'}) = \begin{cases} \sum_{x\in E} p_\theta(x)\log(\frac{p_\theta(x)}{p_{\theta'}(x)}) & if\ E\ is\ discrete \\ \int_{x\in E} f_\theta(x)\log(\frac{f_\theta(x)}{f_{\theta'}(x)})\,dx & if\ E\ is\ continuous \end{cases}$$
    - The integration or summation has to be done only over the support of P_θ if the two distributions have different support
    - The Kullback-Leibler divergence is considered to be infinite if there is a value of $x \in E$ for which $p_{\theta'}(x)$ $(or\ f_{\theta'}(x))$ is zero but $p_\theta(x)$ $(or\ f_\theta(x))$ is non-zero. We

ignore the cases where both the numerator and the denominator of the logarithmic expression are zero.

- The Kullback-Leibler divergence is:
  - Non-symmetric
  - Positive: This is because the logarithmic function is concave
    - In the continuous case, If we take $g(x) = \frac{f_{\theta'}(x)}{f_\theta(x)}$ and $f(x) = \log(x)$, then:
      - $f(g(X)) = \log(\frac{f_{\theta'}(X)}{f_\theta(X)})$
      - $E_{f_\theta}[f(g(X))] = -\int_{x\in E} f_\theta(x) \log(\frac{f_\theta(x)}{f_{\theta'}(x)})\, dx$
      - $f[E_{f_\theta}(g(X))] = \log\left(\int_{x\in E} f_\theta(x) \left(\frac{f_{\theta'}(x)}{f_\theta(x)}\right) dx\right) = \log(\int_{x\in E} f_{\theta'}(x)dx)$
      - Clearly, $f[E_{f_\theta}(g(X))] = 0$
    - For a concave function, $f[E(g(X))] \geq E[f(g(X))]$, or in this case:
      - $-\int_{x\in E} f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx \leq 0$, or
      - $\int_{x\in E} f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx \geq 0$
  - Definite: If $KL(P_\theta, P_{\theta'}) = 0$, then $P_\theta = P_{\theta'}$
  - Does not follow the triangle inequality in general
- The Kullback-Leibler divergence is not mathematically a distance.
- The Kullback-Leibler divergence can be considered to be an expectation:
  - $KL(P_\theta, P_{\theta'}) = E_\theta[\log(\frac{p_\theta(x)}{p_{\theta'}(x)})]$ in the discrete case
  - $KL(P_\theta, P_{\theta'}) = E_\theta[\log(\frac{f_\theta(x)}{f_{\theta'}(x)})]$ in the continuous case
- If θ^* is the true parameter, and θ is the estimator, then we can write in the continuous case (the discrete case will have an identical formulation):
  - $KL(P_{\theta^*}, P_\theta) = \int_{x\in E} f_{\theta^*}(x) \log\left(\frac{f_{\theta^*}(x)}{f_\theta(x)}\right) dx = E_{\theta^*}[\log(\frac{f_{\theta^*}(x)}{f_\theta(x)})]$
  - By linearity of expectations, $KL(P_{\theta^*}, P_\theta) = E_{\theta^*}[\log f_{\theta^*}(x)] - E_{\theta^*}[\log f_\theta(x)]$
  - $E_{\theta^*}[\log f_{\theta^*}(x)]$ is a constant because the true distribution is known
  - Hence, $KL(P_{\theta^*}, P_\theta) = constant - E_{\theta^*}[\log f_\theta(x)]$
  - If we have a very large number of observations, then based on the law of large numbers, the mean estimator converges in probability to the expected value. Hence, for large values of n,
    - $KL(P_{\theta^*}, P_\theta) = constant - {}^1\!/_n \sum_{i=1}^n \log f_\theta(X_i)]$
- We want to find the parameter θ that minimizes the Kullback-Leibler divergence between the true distribution and the estimated distribution. The following formulations are equivalent for finding the parameter value that minimizes the Kullback-Leibler divergence:
  - $\underset{\theta\in\Theta}{\operatorname{argmin}} - {}^1\!/_n \sum_{i=1}^n \log f_\theta(X_i)]$

- $\underset{\theta\in\Theta}{\operatorname{argmax}}\ {}^{1}/_{n}\sum_{i=1}^{n}\log f_\theta(X_i)]$ (since minimizing the negative of an expression is equivalent to maximizing the expression)
- $\underset{\theta\in\Theta}{\operatorname{argmax}}\ \sum_{i=1}^{n}\log f_\theta(X_i)]$ ($1/n$ is a constant and does not affect where the maximum is reached)
- $\underset{\theta\in\Theta}{\operatorname{argmax}}\ \prod_{i=1}^{n} f_\theta(X_i)$ (for a positive valued function, the value of the variable at which the function reaches its maximum or minimum is the same as the value of the variable at which the logarithm of the function reaches its maximum or minimum)

- Estimating the parameter via the approach of finding $\underset{\theta\in\Theta}{\operatorname{argmax}}\ \prod_{i=1}^{n} f_\theta(X_i)$ is called maximum likelihood estimation
- As an example of Kullback-Leibler divergence for discrete distributions, let us consider:
  - $P_\theta = Binomial(n,p)$ and $P_{\theta'} = Binomial(n,q)$ where n is fixed
  - Effectively, we can denote the two distributions as:
    - $P_p(k) = C_k^n p^k(1-p)^{n-k}$ and $P_q(k) = C_k^n q^k(1-q)^{n-k}$ for k=0,1,2,..
  - $KL(P_p,P_q) = KL(P_p||P_q) = \sum_{k=0}^{\infty} P_p(k)\log(\frac{P_p(k)}{P_q(k)})$
  - Using the probability mass functions, we have:
    - $KL(P_p,P_q) = KL(P_p||P_q) = \sum_{k=0}^{n} P_p(k)\log(\frac{C_k^n p^k(1-p)^{n-k}}{C_k^n q^k(1-q)^{n-k}})$, or
    - $KL(P_p,P_q) = \sum_{k=0}^{n} kP_p(k)\log(\frac{p}{q}) + (n-k)\sum_{k=0}^{n} P_p(k)\log(\frac{1-p}{1-q})$, or
    - $KL(P_p,P_q) = \log(\frac{p}{q})\sum_{k=0}^{n} kP_p(k) + \log(\frac{1-p}{1-q})(\sum_{k=0}^{n}(nP_p(k) - kP_p(k))$
    - We know that $\sum_{k=0}^{n} kP_p(k)$ is the expectation of the binomial distribution with parameters n and p, and this is equal to np. Therefore,
    - $KL(P_p,P_q) = np\log(\frac{p}{q}) + (n-np)\log(\frac{1-p}{1-q})$
  - If $q = 0$, then the Kullback-Leibler divergence will become infinite
- As an example of Kullback-Leibler divergence for continuous distributions, let us consider
  - $P_\theta = N(a,1)$ and $P_{\theta'} = N(b,1)$ (The two distributions have the same variance)
  - $P_{a,1}(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2}}$ and $P_{b,1}(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-b)^2}{2}}$
  - Using the probability density functions, we have:
    - $KL(P_p,P_q) = KL(P_p||P_q) = \int_{-\infty}^{+\infty} P_{a,1}(x)\log(\frac{\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-b)^2}{2}}})\,dx$, or
    - $KL(P_p,P_q) = \int_{-\infty}^{+\infty} P_{a,1}(x)\log(\frac{e^{\frac{-(x-a)^2}{2}}}{e^{\frac{-(x-b)^2}{2}}})dx$, or
    - $KL(P_p,P_q) = \int_{-\infty}^{+\infty} P_{a,1}(x)\left(ax - bx - \frac{a^2}{2} + \frac{b^2}{2}\right)dx$, or
    - $KL(P_p,P_q) = (a-b)\int_{-\infty}^{+\infty} xP_{a,1}(x)dx + (-\frac{a^2}{2} + \frac{b^2}{2})\int_{-\infty}^{+\infty} P_{a,1}(x)dx$
    - Since $\int_{-\infty}^{+\infty} xP_{a,1}(x)$ is the expectation of a normal distribution with mean a, $\int_{-\infty}^{+\infty} xP_{a,1}(x) = a$. Also $\int_{-\infty}^{+\infty} P_{a,1}(x)dx = 1$. Therefore,

- $KL(P_p, P_q) = (a-b)a + \left(-\frac{a^2}{2} + \frac{b^2}{2}\right) = \frac{1}{2}(a-b)^2$

Maximum Likelihood Estimation

- For a discrete distribution, if $(E, (P_\theta)_{\theta\in\Theta})$ is the statistical model associated with a sample of identically distributed random variables $X_1, X_2, .. X_n$, the likelihood of the model is the map L_n (or just L) defined as:
  - $L_n: E^n x\, \Theta \to$ R, or $L_n(x_1, x_2, .. x_n; \theta) \mapsto P_\theta[X_1 = x_1, X_2 = x_2, .., X_n = x_n]$
    - If the samples are independent, then the joint pmf is the product of the marginal pmfs
  - In the case of the Bernoulli distribution, with independent and identically distributed random variables
    - $L_n(x_1, x_2, .. x_n; p) = \prod_{i=1}^n P_p(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$, or
    - $L_n(x_1, x_2, .. x_n; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$
  - In the case of the Poisson distribution, with independent and identically distributed random variables
    - $L_n(x_1, x_2, .. x_n; \lambda) = \prod_{i=1}^n P_p(X_i = x_i) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$, or
    - $L_n(x_1, x_2, .. x_n; \lambda) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1!\, x_2! \ldots x_n!}$
- For a continuous distribution, , if $(E, (P_\theta)_{\theta\in\Theta})$ is the statistical model associated with a sample of identically distributed random variables $X_1, X_2, .. X_n$, the likelihood of the model is the map L_n (or just L) defined as:
  - $L_n: E^n x\, \Theta \to$ R, or $L_n: (x_1, x_2, .. x_n; \theta) \mapsto f_\theta[X_1 = x_1, X_2 = x_2, .., X_n = x_n]$
    - If the samples are independent, then the joint pdf is the product of the marginal pdfs
  - In the case of the Gaussian distribution, with independent and identically distributed random variables:
    - $\Theta = \mathrm{R}\ \mathrm{x}\ (0, \infty)$
    - $L_n(x_1, x_2, .. x_n; \mu, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}}$
  - For the exponential distribution, with independent and identically distributed random variables:
    - $L_n(x_1, x_2, .. x_n; \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$
- **Concave and convex functions**: Let us consider an interval $I = (x_1, x_2)\ where\ x_2 > x_1 \quad x_1, x_2 \in R$
  - Any point x between x_1 and x_2 can be written as:
    - $x = x_2 - t(x_2 - x_1)\ where\ 0 < t < 1$ or $x = tx_1 + (1-t)x_2$
  - For a concave function $g: I \to R$, we have:
    - $g(tx_1 + (1-t)x_2) \geq tg(x_1) + (1-t)g(x_2) \quad \forall \quad 0 < t < 1$
    - It is strictly concave if $g(tx_1 + (1-t)x_2) > tg(x_1) + (1-t)g(x_2)$
    - In graphical terms, it means for $x_1 < x < x_2$, the secant line connecting the two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ is below the graph of g

- For a convex function $g: I \rightarrow R$, we have:
  - $g(tx_1 + (1-t)x_2) \leq tg(x_1) + (1-t)g(x_2) \quad \forall \quad 0 < t < 1$
  - It is strictly convex if $g(tx_1 + (1-t)x_2) < tg(x_1) + (1-t)g(x_2)$
  - In graphical terms, it means for $x_1 < x < x_2$, the secant line connecting the two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ is above the graph of g
- From a differentiability standpoint, if a function g is twice differentiable in the interval I, i.e. $g''(x)\ exists\ \forall\ x \in I$, then g is :
  - Concave if $g''(x) \leq 0\ \forall\ x \in I$, strict concavity implies $g''(x) < 0\ \forall\ x \in I$
  - Convex if $g''(x) \geq 0\ \forall\ x \in I$, strict convexity implies $g''(x) > 0\ \forall\ x \in I$
- The sum of two concave functions is also concave, and the sum of two convex functions is also convex. However, the product of concave functions may not be concave, and the product of convex functions may not be convex.
- When we evaluate functions with more than one variable, convexity and concavity are defined by evaluating gradients and hessians. Gradient of a function h: $\Theta \in \mathrm{R}^d \rightarrow R$ is a vector of partial derivatives:
  - Gradient vector: $\nabla h(\theta) = \begin{pmatrix} \frac{\partial h(\theta)}{\partial \theta_1} \\ \frac{\partial h(\theta)}{\partial \theta_2} \\ \cdots \\ \frac{\partial h(\theta)}{\partial \theta_d} \end{pmatrix}$
  - The Hessian is the second derivative of the function. Each of the entries in the gradient vector have to be partially differentiated with respect to all the parameters. The Hessian is denoted as $\boldsymbol{H}h(\theta)$ and is a matrix in R^{dxd}
    - $\boldsymbol{H}h(\theta) = \begin{pmatrix} \frac{\partial h(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial h(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial h(\theta)}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial h(\theta)}{\partial \theta_2 \partial \theta_1} & \cdots & \cdots & \frac{\partial h(\theta)}{\partial \theta_2 \partial \theta_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h(\theta)}{\partial \theta_d \partial \theta_1} & \cdots & \cdots & \frac{\partial h(\theta)}{\partial \theta_d \partial \theta_d} \end{pmatrix} \in \mathrm{R}^{dxd}$
  - The function h is concave if and only if $x^T \boldsymbol{H}h(\theta)x \leq 0\ \forall\ x \in \mathrm{R}^d, \theta \in \Theta$. In this case, $\boldsymbol{H}h(\theta)$ is negative semi-definite. Strict concavity exists if for all non-zero x, $x^T \boldsymbol{H}h(\theta)x < 0$, and thin this case $\boldsymbol{H}h(\theta)$ is negative definite. Both negative semi-definite and negative definite matrices are symmetrical.
  - The function h is convex if and only if $x^T \boldsymbol{H}h(\theta)x \geq 0\ \forall\ x \in \mathrm{R}^d, \theta \in \Theta$. In this case, $\boldsymbol{H}h(\theta)$ is positive semi-definite. Strict convexity exists if for all non-zero x, $x^T \boldsymbol{H}h(\theta)x > 0$. In this case, $\boldsymbol{H}h(\theta)$ is positive definite. Both positive semi-definite and positive definite matrices are symmetrical.
  - As an example, if $h(\theta) = -\theta_1^2 - 2\theta_2^2$, then
    - $\nabla h(\theta) = \begin{pmatrix} -2\theta_1 \\ -4\theta_2 \end{pmatrix}$, and $\boldsymbol{H}h(\theta) = \begin{pmatrix} -2 & 0 \\ 0 & -4 \end{pmatrix}$. The matrix is symmetrical, has negative eigenvalues (-2 and -4), and is a negative definite matrix. If we take $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, then $x^T \boldsymbol{H}h(\theta)x = -2x_1^2 - 4x_2^2 > 0\ \forall\ x \neq zero\ vector$
  - As another example, if we have $h(\theta) = \log(\theta_1 + \theta_2)$, then:

- $\nabla h(\theta) = \begin{pmatrix} \frac{1}{\theta_1+\theta_2} \\ \frac{1}{\theta_1+\theta_2} \end{pmatrix}$, and $\boldsymbol{H}h(\theta) = \begin{pmatrix} \frac{-1}{(\theta_1+\theta_2)^2} & \frac{-1}{(\theta_1+\theta_2)^2} \\ \frac{-1}{(\theta_1+\theta_2)^2} & \frac{-1}{(\theta_1+\theta_2)^2} \end{pmatrix}$. This is not an invertible matrix; it is singular. $\boldsymbol{H}h(\theta) = \frac{-1}{(\theta_1+\theta_2)^2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Hence, if take $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, then $x^T\boldsymbol{H}h(\theta)x = \frac{-1}{(\theta_1+\theta_2)^2}\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ which gives:
    - $x^T\boldsymbol{H}h(\theta)x = \frac{-1}{(\theta_1+\theta_2)^2}(x_1+x_2)^2 < 0 \; \forall \; x \neq zero\ vector$
    - Hence the function $h(\theta)$ is strictly concave
  - If a function is strictly concave over an interval, it has a unique maximum in that interval. Similarly, if a function is strictly convex over an interval, it has a unique minimum in that interval.
- The maximum likelihood estimator or MLE is the value of the parameter at which the likelihood is maximized.
  - In other words, if the likelihood function is $L_n(x_1,x_2,..x_n;\theta)$, then $\hat{\theta}_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}} L_n(x_1,x_2,..x_n;\theta)$ provided it exists
  - The likelihood is always positive-valued, and instead of using the likelihood function, we use the log of the likelihood function, which will have its maximum at the same value of θ for which the likelihood function achieves its maximum
  - For the Bernoulli distribution, $\hat{p}_{MLE} = \underset{p\in(0,1)}{\operatorname{argmax}} L_n(x_1,x_2,..x_n;p)$
    - $L_n(x_1,x_2,..x_n;p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$
    - Taking the log of the likelihood, we have:
      - $\log L_n(x_1,x_2,..x_n;p) = S_n \log p + (n-S_n)\log(1-p)$ where $S_n = \sum_{i=1}^n x_i$
      - The first derivative of the log-likelihood is $\frac{S_n}{p} - \frac{(n-S_n)}{1-p}$ and the second derivative is $-\frac{S_n}{p^2} - \frac{(n-S_n)}{(1-p)^2} \leq 0$ since $S_n \leq n, S_n \geq 0$. Thus, the likelihood function is concave and reaches its maximum at the point where the first derivative is 0.
      - Setting $\frac{S_n}{p} - \frac{(n-S_n)}{1-p} = 0$ gives $p = \frac{S_n}{n}$
      - Thus, $\hat{p}_{MLE} = \underset{p\in(0,1)}{\operatorname{argmax}} L_n(x_1,x_2,..x_n;p) = \frac{S_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$
  - For the Poisson distribution, we have:
    - $L_n(x_1,x_2,..x_n;\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{x_1!\,x_2!\ldots x_n!}$
    - Taking the log of the likelihood, we have:
      - $\log L_n(x_1,x_2,..x_n,\lambda) = -n\lambda + S_n \log\lambda - \log(x_1!x_2!..x_n!)$ where $S_n = \sum_{i=1}^n x_i$
      - The first derivative of the log likelihood is $-n + \frac{S_n}{\lambda}$, and the second derivative is $-\frac{S_n}{\lambda^2} \leq 0 \; as \; S_n \geq 0, and \; \lambda > 0$. Thus, the likelihood

function is concave and reaches its maximum at the point where the first derivative is 0.

- Setting $-n + \frac{S_n}{\lambda} = 0$ gives $\lambda = \frac{S_n}{n}$
- Thus, $\hat{\lambda}_{MLE} = \underset{\lambda \in (0,\infty)}{\operatorname{argmax}} L_n(x_1, x_2, .. x_n; \lambda) = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

- For the Gaussian distribution, we have:
  - $L_n(x_1, x_2, .. x_n; \mu, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}}$
  - Taking the log of the likelihood, we get:
    - $\log L_n(x_1, x_2, .. x_n; \mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$
  - The first derivative of the log likelihood is (for the variance, we take the derivative with respect to σ^2):
    - $\begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}$
  - The second derivative of the log likelihood is:
    - $\begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu) \\ -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu) & \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}$
    - The second derivative can be used to show that the log likelihood function is strictly concave
  - To get the maximum likelihood estimate values, we set the first derivative to zero. This gives:
    - $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$
    - $\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2}{n}$

- In some cases, the maximum likelihood estimator cannot be found by taking derivatives and setting the expression to zero.
  - As an example, consider that the observations are drawn from a uniform distribution $U(0, \theta)$. The likelihood function can be written as:
    - $L_n(x_1, x_2, .. x_n; \theta) = \frac{1}{\theta^n}$. Setting the first derivative equal to 0 yields $\hat{\theta}_{MLE} \to \infty$ which is not correct
    - The error arises because there are additional constraints on θ. In this case, θ cannot be lesser than the maximum of the realized observations. In fact, the graph of $L_n(x_1, x_2, .. x_n; \theta)$ versus θ is zero-valued for $\theta \in (-\infty, \max_i X_i)$ and a downward sloping convex function from the point corresponding to $\theta = \max_i X_i$. There is a discontinuity in the likelihood function graph at $\theta = \max_i X_i$. In effect, the correct expression of the likelihood function is:
      - $L_n(x_1, x_2, .. x_n; \theta) = \frac{1}{\theta^n} I(\max_i X_i \leq \theta)$
      - The maximum likelihood estimator is $\hat{\theta}_{MLE} = \max_i X_i$

- The maximum likelihood estimator has the following properties under certain regularity conditions:
    - **Convergence**: From the Kullback-Leibler divergence, we know that if the true parameter is θ^*, then for a large number of observations:
        - $KL(P_{\theta^*}, P_\theta) \underset{n \to \infty}{\overset{P}{\to}} constant - {}^1\!/_n \sum_{i=1}^{n} \log f_\theta(X_i)]$, or
        - $KL(P_{\theta^*}, P_\theta) - constant \underset{n \to \infty}{\overset{P}{\to}} -\frac{1}{n} \log L_n(x_1, x_2, .. x_n; \theta)$
    - If we minimize $KL(P_{\theta^*}, P_\theta)$, we are maximizing the log likelihood. Conversely, maximizing the log likelihood leads to minimizing the KL divergence, and it can be shown that as $n \to \infty$, P_{θ^*} converges in distribution to P_θ, and for an identifiable model, this means that θ converges to θ^*.

Multivariate random variables

- **Multivariate random variable**: A multivariate random variable is a vector of random variables. It is also called a random vector. Specifically, a random vector $X = (X^{(1)}, X^{(2)}, \dots X^{(d)})^T$ of dimension $dx1$ is a vector-valued function from a probability space Ω to R^d.
    - The probability distribution of a random vector X is the joint distribution of its components $X^{(1)}, X^{(2)}, \dots X^{(d)}$
    - The cumulative distribution function of a random vector X is defined as:
        - $P(X^{(1)} \leq x^{(1)}, X^{(2)} \leq x^{(2)}, \dots X^{(d)} \leq x^{(d)})$
    - For a single random variable $\hat{\theta}_n^{MLE} \underset{n \to \infty}{\overset{P}{\to}} \theta^*$. For a random vector, each of the components must converge in probability. Thus,
        - $\hat{X}_n \underset{n \to \infty}{\overset{P}{\to}} X \iff \hat{X}_n^{(k)} \underset{n \to \infty}{\overset{P}{\to}} X^{(k)} \; \forall \, 1 \leq k \leq d$
    - In the case of two random variables X and Y, if the variables are independent, then $Cov(X,Y) = 0$. However, it is not always true that if $Cov(X,Y) = 0$, X and Y are independent.
        - $Cov(X,Y) = 0 \Rightarrow X \; and \; Y \; are \; independent$ only if $(X,Y)^T$ is a Gaussian vector, i.e. $\alpha X + \beta Y$ is Gaussian for all $(\alpha, \beta) \in R^2, (\alpha, \beta) \neq (0,0)$.
        - As an example, consider $X \sim N(0,1)$ and $Y = XR$ where R is a Rademacher random variable. R takes only two values, -1 and +1 both with equal probability ½. R is independent of X.
            - If B is a Bernoulli random variable with $p = 0.05$, then $R = 2B - 1$.
        - $Y = \begin{matrix} X \; with \; probability \; 1/2 \\ -X \; with \; probability \; 1/2 \end{matrix}$. However, since $X \sim N(0,1)$, $-X \sim N(0,1)$ due to the symmetry of the random normal variable. This means that $Y \sim N(0,1)$
        - $Cov(X,Y) = Cov(X, XR) = E[X^2 R] - E[X]E[XR]$, or
            - $Cov(X,Y) = E[R]E[X^2] - E[X]E[XR] = 0$, as $E[X] = 0, E[R] = 0$
        - However, the variables X and Y are not independent. $|X| = |Y|$. Hence, in this case, $Cov(X,Y) = 0 \not\Rightarrow X \; and \; Y \; are \; independent$
            - This is because despite X and Y being both Gaussian variables, $(X,Y)^T$ is a not Gaussian vector, i.e. $\alpha X + \beta Y$ is not Gaussian for all $(\alpha, \beta) \in R^2, (\alpha, \beta) \neq (0,0)$. Specifically, if we take $\alpha = \beta = 1$, then:
                - $X + Y = \begin{matrix} 0 \; with \; probability \frac{1}{2} \\ 2X \; with \; probability \frac{1}{2} \end{matrix}$
                - $2X$ is a Gaussian random variable, but 0 is not (0 is a degenerate normal random variable)
    - Let $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ be independent and identically distributed observations of the random vector $(X,Y)^T$. This means that each random variable pair (e.g. (X_1, Y_1)) has the same distribution as the random variable pair (X,Y) and the pairs are independent of one another.
        - Let $E[X] = \mu_X, E[Y] = \mu_Y, and \;\; E[XY] = \mu_{XY}$
        - An unbiased estimator $\hat{S}_{XY}$ of the covariance between X and Y is:

- $\hat{S}_{XY} = \frac{1}{(n-1)} (\sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n))$
- $E[\hat{S}_{XY}] = \frac{1}{(n-1)} (E[\sum_{i=1}^{n} (X_i Y_i)] - E[\sum_{i=1}^{n} X_i \bar{Y}_n] - E[\sum_{i=1}^{n} Y_i \bar{X}_n] + nE[\bar{X}_n \bar{Y}_n])$, or
- $E[\hat{S}_{XY}] = \frac{1}{(n-1)} (E[\sum_{i=1}^{n} (X_i Y_i)] - nE[\bar{X}_n \bar{Y}_n] - nE[\bar{X}_n \bar{Y}_n] + nE[\bar{X}_n \bar{Y}_n])$, or
- $E[\hat{S}_{XY}] = \frac{1}{(n-1)} (E[\sum_{i=1}^{n} (X_i Y_i)] - nE[\bar{X}_n \bar{Y}_n])$
- $E[X_i Y_j] = \begin{matrix} \mu_{XY} \; if \; i = j \\ \mu_X \mu_Y \; if \; i \neq j \end{matrix}$
- $E[\bar{X}_n \bar{Y}_n]$ has n terms of the form $E[X_i Y_i]/n^2$ and $n^2 - n$ terms of the form $E[X_i Y_j]/n^2 \; where \; i \neq j$.
    - Hence $E[\bar{X}_n \bar{Y}_n] = \mu_{XY}/n + (\frac{n-1}{n}) \mu_X \mu_Y$
- Thus, $E[\hat{S}_{XY}] = \frac{1}{(n-1)} (n\mu_{XY} - \mu_{XY} - (n-1)\mu_X \mu_Y)$
- Or, $E[\hat{S}_{XY}] = \frac{1}{(n-1)} ((n-1)\mu_{XY} - (n-1)\mu_X \mu_Y) = \mu_{XY} - \mu_X \mu_Y = Cov(X, Y)$

- **Covariance Matrices**: A multivariate random variable, or a random vector can be written as $X = (X^{(1)}, X^{(2)}, \ldots X^{(d)})^T$.
    - The covariance matrix or the variance-covariance matrix is defined as:
        - $\mathbf{\Sigma} = Cov(X) = E[(X - E[X])(X - E[X])^T]$ which is a matrix of size dxd.
        - The term i, j of this matrix is $E[(X^{(i)} - E[X^{(i)}])(X^{(j)} - E[X^{(j)}])]$
        - This is a symmetrical matrix and its diagonal terms represent the variance of the individual variables and the off-diagonal terms represent the covariance.
    - If we have a matrix A of dimension nxd, and another matrix B with dimension $nx1$, then:
        - $Cov(AX + B) = Cov(AX) = ACov(X)A^T = A\mathbf{\Sigma}A^T$
- **Multivariate Gaussian distribution**: A Gaussian vector $X \in R^d$ is completely determined by its expected value $E[X] = \mu \in R^d$ and its covariance matrix $\Sigma \in R^{dxd}$.
    - Notationally, we write $X \sim N_d(\mu, \Sigma)$
    - It may be noted that for a random vector $X \in R^d$ to be a Gaussian vector, $\alpha^T X$ is Gaussian for any column vector $\alpha \in R^d, \alpha \neq 0$.
    - The pdf of the multivariate Gaussian distribution is:
        - $\frac{1}{(2\pi)^{d/2} \det(\Sigma)} exp^{(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu))}$
    - A Gaussian vector is considered to be a standard normal random vector if $\mu = 0$, and Σ is the identity matrix. If Σ is the identity vector, it means that the components of X are independent, and the joint pdf is the product of the marginal pdfs.
    - In the case where the covariance matrix is singular, the Gaussian vector is said to be degenerate.
        - As an example, if $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, then clearly Σ is singular – the second column is twice the first column. The null space vector of $\Sigma = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$. If we denote this vector by M, then $M^T X$ is a random variable with zero variance, i.e. it is a constant

- The covariance matrix is at least a positive semi-definite matrix. In most cases, it is a positive definite matrix. Also, any positive definite matrix can be a potential covariance matrix.

- **Multivariate Central Limit Theorem**: Let $X_1, X_2, .. X_n$ be independent copies of a random vector X such that $E[X] = \mu,\ Cov(X) = \Sigma$. Then:
  - $\sqrt{n}(\bar{X}_n - \mu) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, \Sigma)$, and equivalently:
  - $\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{X}_n - \mu) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, \mathrm{I}_d)$

- **Multivariate delta method**: The multivariate delta method is an extension of the delta method to multivariate random variables, or random vectors.
  - If $(T_n)_{n \geq 1}$ is a sequence of random vectors in R^d such that $\sqrt{n}(T_n - \theta) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, \Sigma)$ for some $\theta \in R^d$ and $\Sigma \in R^{dxd}$, and if $g: R^d \to R^k$ $(k \geq 1)$ is a function continuously differentiable at θ, then:
    - $\sqrt{n}\big(g(T_n) - g(\theta)\big) \underset{n \to \infty}{\overset{d}{\to}} N_k(0, \nabla g(\theta)^T \Sigma \nabla g(\theta))$
    - g can be considered a set of k different functions, each of which has to be partially differentiated with respect to each of the d variables to get the first derivative of the function:
      - $\nabla g(\theta) = \begin{pmatrix} | & | & | & | \\ \nabla g_1 & \nabla g_2 & \dots & \nabla g_k \\ | & | & | & | \end{pmatrix} \in R^{dxk}$, and $\nabla g(\theta)_{i,j} = (\frac{\partial g_j}{\partial \theta_i})_{\substack{1 \leq j \leq k \\ 1 \leq i \leq d}}$
      - $\nabla g(\theta)$ is the transpose of the Jacobian matrix J_g of g
  - As an example, if $g(x, y, z) = \begin{pmatrix} x^2 + y^2 + z^2 \\ 2xy \\ y^3 + z^3 \\ z^4 \end{pmatrix}$, then:
    - $\nabla g(x, y, z) = \begin{pmatrix} 2x & 2y & 0 & 0 \\ 2y & 2x & 3y^2 & 0 \\ 2z & 0 & 3z^2 & 4z^3 \end{pmatrix}$

- **Inference for variance of a Gaussian distribution**: Let us consider a sample of n independent and identically distributed random variables $X_1, X_2, \dots X_n$ with unknown parameters $\mu \in R$, and $\sigma^2 > 0$.
  - One approach to constructing an estimator for the variance is to express the variance in terms of expectations and replace expectations with sample averages. The estimator for the variance can be expressed as $\hat{\sigma}^2$ where:
    - $Var(X) = E[X^2] - E[X]^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2$
    - $\hat{\sigma}^2$ is clearly a consistent estimator, as the averages converge to the expectations as $n \to \infty$.

- If we consider X and X^2 as two separate variables, then we can consider $\begin{pmatrix} X_i^2 \\ X_i \end{pmatrix}$ $\forall i \in [1,n]$ as independent copies of the random vector $\begin{pmatrix} X^2 \\ X \end{pmatrix}$. Using the multivariate central limit theorem, we can state that:
  - $\sqrt{n}\left(\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \frac{1}{n}\sum_{i=1}^n X_i \end{pmatrix} - \begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}\right) \underset{n \to \infty}{\overset{d}{\to}} N_2(0,\Sigma)$ where Σ is the covariance matrix
  - $\Sigma = \begin{pmatrix} Var(X^2) & Cov(X^2,X) \\ Cov(X^2,X) & Var(X) \end{pmatrix}$, or
  - $\Sigma = \begin{pmatrix} E[X^4] - E[X^2]^2 & E[X^3] - E[X]E[X^2] \\ E[X^3] - E[X]E[X^2] & E[X^2] - E[X]^2 \end{pmatrix}$
    - $E[X^2] = \sigma^2 + \mu^2$
    - $E[X^3] = \mu^3 + 3\sigma^2\mu$
    - $E[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
  - Thus, $\Sigma = \begin{pmatrix} 4\mu^2\sigma^2 + 2\sigma^4 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & \sigma^2 \end{pmatrix}$
- Next we can apply the delta method. If we take a function g such that $g(a,b) = (a - b^2)$, then $g: R^2 \to R$. Also, $\nabla g(a,b) = \begin{pmatrix} 1 \\ -2b \end{pmatrix}$.
  - $g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}$ is continuously differentiable if X is a Gaussian random variable. Therefore:
  - $\sqrt{n}\left(g\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \frac{1}{n}\sum_{i=1}^n X_i \end{pmatrix} - g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}\right) \underset{n \to \infty}{\overset{d}{\to}} N_1(0, \nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}^T \Sigma \nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix})$
  - Since $\nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix} = \begin{pmatrix} 1 \\ -2E[X] \end{pmatrix} = \begin{pmatrix} 1 \\ -2\mu \end{pmatrix}$,
    - $\nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}^T \Sigma \nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix} =$
      $(1 \quad -2\mu)\begin{pmatrix} 4\mu^2\sigma^2 + 2\sigma^4 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & \sigma^2 \end{pmatrix}\begin{pmatrix} 1 \\ -2\mu \end{pmatrix}$
    - Or, $\nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}^T \Sigma \nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix} = (1 \quad -2\mu)\begin{pmatrix} 2\sigma^4 \\ 0 \end{pmatrix}$
    - Or, $\nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}^T \Sigma \nabla g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix} = 2\sigma^4$
  - Thus, $\sqrt{n}\left(g\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \frac{1}{n}\sum_{i=1}^n X_i \end{pmatrix} - g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix}\right) \underset{n \to \infty}{\overset{d}{\to}} N_1(0,2\sigma^4)$
  - Since $g\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \frac{1}{n}\sum_{i=1}^n X_i \end{pmatrix} = \hat{\sigma}^2$, and $g\begin{pmatrix} E[X^2] \\ E[X] \end{pmatrix} = \sigma^2$,

- $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \underset{n \to \infty}{\overset{d}{\to}} N(0, 2\sigma^4)$

- To obtain a $(1-\alpha)$ confidence interval for σ^2, we use the fact that $\frac{\sqrt{n}(\hat{\sigma}^2 - \sigma^2)}{\sqrt{2}\sigma^2}$ is a standard random variable, and hence:
  - $P\left(\hat{\sigma}^2 - \frac{\sqrt{n} q_{\alpha/2}}{\sqrt{2}\sigma^2} \leq \sigma^2 \leq \hat{\sigma}^2 + \frac{\sqrt{n} q_{\alpha/2}}{\sqrt{2}\sigma^2}\right) = 1 - \alpha$
  - Since the bounds of the confidence interval in the above expression depend on the unknown parameter, we replace it with the estimator which is consistent and asymptotically converges to the parameter. Thus:
  - $P\left(\hat{\sigma}^2 - \frac{\sqrt{n} q_{\alpha/2}}{\sqrt{2}\hat{\sigma}^2} \leq \sigma^2 \leq \hat{\sigma}^2 + \frac{\sqrt{n} q_{\alpha/2}}{\sqrt{2}\hat{\sigma}^2}\right) = 1 - \alpha$

Fisher Information

- If we define the log-likelihood for one observation as $l(\theta) = \log L_1(X, \theta),\ \theta \in \Theta\ in\ R^d$, and if we assume that $l(\theta)$ is twice differentiable, then under some regularity conditions, the Fisher information for the statistical model is defined as:
  - $I(\theta) = Cov\big(\nabla l(\theta)\big) = E[\nabla l(\theta)\nabla l(\theta)^T] - E[\nabla l(\theta)]E[\nabla l(\theta)]^T = -E[\boldsymbol{H} l(\theta)]$ where $\boldsymbol{H} l(\theta)$ is the second derivative or the Hessian
    - If $\Theta \in R$, i.e. we have only one unknown parameter, then:
      - $I(\theta) = E[(l'(\theta))^2] - E[l'(\theta)]^2 = Var\big(l'(\theta)\big) = -E[l''(\theta)]$
    - In the one-dimensional case, i.e. $\Theta \in R$, we can write:
      - $f_\theta(x) = L_1(x, \theta)$ where $f_\theta(x) = P(X = x; \theta)$
        - $\int_{-\infty}^{+\infty} f_\theta(x)dx = 1$
        - $\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty} f_\theta(x)dx = 0$, and $\frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{+\infty} f_\theta(x)dx = 0$, since in both cases we are taking the derivative of a constant
        - $\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty} f_\theta(x)dx = 0 \Rightarrow \int_{-\infty}^{+\infty}\frac{\partial}{\partial\theta} f_\theta(x)dx = 0 \Rightarrow \int_{-\infty}^{+\infty}\frac{\partial}{\partial\theta} L_1(x, \theta)dx = 0$ - **(A)**
        - $\frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{+\infty} f_\theta(x)dx = 0 \Rightarrow \int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2} f_\theta(x)dx = 0 \Rightarrow \int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2} L_1(x, \theta)dx = 0$ – **(B)**
      - $l'(\theta) = \frac{\partial}{\partial\theta}\log L_1(x, \theta) = \frac{\frac{\partial}{\partial\theta}(L_1(x,\theta))}{L_1(x,\theta)}$,
        - Thus $E[l'(\theta)] = \int_{-\infty}^{+\infty}\frac{\frac{\partial}{\partial\theta}(L_1(x,\theta))}{L_1(x,\theta)} L_1(x, \theta)dx = 0$ **from (A)**
        - Thus, $Var[l'(\theta)] = E[(l'(\theta))^2] - E[l'(\theta)]^2 = E[(l'(\theta))^2]$ as $E[l'(\theta)] = 0$, or
        - $Var[l'(\theta)] = E[(l'(\theta))^2] = \int_{-\infty}^{+\infty}(\frac{\frac{\partial}{\partial\theta}(L_1(x,\theta))}{L_1(x,\theta)})^2 L_1(x, \theta)dx$, or
        - $Var[l'(\theta)] = \int_{-\infty}^{+\infty}\frac{(\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{L_1(x,\theta)}\ dx$ – **(C)**
      - $l''(\theta) = \frac{\partial}{\partial\theta} l'(\theta) = \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}(L_1(x,\theta))}{L_1(x,\theta)}\right) = \frac{L_1(x,\theta)\frac{\partial^2}{\partial\theta^2}(L_1(x,\theta)) - (\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{(L_1(x,\theta))^2}$
      - Or, $E[l''(\theta)] = \int_{-\infty}^{+\infty}\frac{L_1(x,\theta)\frac{\partial^2}{\partial\theta^2}(L_1(x,\theta)) - (\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{(L_1(x,\theta))^2} L_1(x, \theta)dx$,
      - Or, $E[l''(\theta)] = \int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2}\big(L_1(x, \theta)\big)dx - \int_{-\infty}^{+\infty}\frac{-(\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{L_1(x,\theta)}dx$,
      - Or, $E[l''(\theta)] = -\int_{-\infty}^{+\infty}\frac{(\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{L_1(x,\theta)}dx$ as $\int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2}\big(L_1(x, \theta)\big)dx = 0$ **from (B)**
      - Or, $-E[l''(\theta)] = \int_{-\infty}^{+\infty}\frac{(\frac{\partial}{\partial\theta}(L_1(x,\theta)))^2}{L_1(x,\theta)}dx$ - **(D)**
      - Comparing **(C)** and **(D)**, we have $Var[l'(\theta)] = -E[l''(\theta)]$
- As an example, for a Bernoulli random variable parameterized by p,
  - $L_1(X, p) = p^X(1-p)^{1-X}$

- $l(p) = \log L_1(X,p) = X \log p + (1-X)\log(1-p)$
- $\frac{\partial}{\partial p}(l(p)) = l'(p) = \frac{X}{p} - \frac{(1-X)}{(1-p)} = \frac{X}{p(1-p)} - \frac{p}{p(1-p)}$
- $\frac{\partial^2}{\partial p^2}\left(L_1(x,p)\right) = l''(p) = -\frac{X}{p^2} - \frac{(1-X)}{(1-p)^2}$
- $Var[l'(p)] = Var\left(\frac{X}{p(1-p)}\right) = \frac{1}{p^2(1-p)^2} Var(X) = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}$
- $-E[l''(p)] = \frac{E[X]}{p^2} + \frac{E[1-X]}{(1-p)^2} = \frac{1}{p} + \frac{1}{(1-p)} = \frac{1}{p(1-p)}$
- Clearly, Fisher information = $I(p) = Var[l'(p)] = -E[l''(p)] = \frac{1}{p(1-p)}$

- Since the Fisher information $I(\theta) = -E[l''(\theta)]$, it specifies, on average, how curved the graph of $\log L_1(X,\theta)$ vs. θ is. This actually holds not just for one sample, but for several samples of the observed variable as well. In particular, $I(\theta^*)$ specifies how curved, on average, the log-likelihood is near the true parameter θ^*. As a rule of thumb, if the Fisher information $I(\theta^*)$ is large, then we can expect the maximum likelihood estimator to give a good estimate for θ^*.
- Under certain conditions, the Fisher information (or the Fisher information matrix, in the case of multivariate distributions) controls the covariance matrix of the maximum likelihood estimator.
  - The following conditions have to be satisfied for the Fisher information to provide the asymptotic variance of the maximum likelihood estimator
    - The true parameter $\theta^* \in \Theta$ is identifiable
    - For all $\theta \in \Theta$, the support of P_θ does not depend on θ
      - This means, for example that if we have a uniform distribution between $[0,\theta]$, the Fisher information cannot be utilized. The support of the distribution depends on the unknown parameter
    - θ^* is not on the boundary of Θ – the true parameter situated on the boundary will create issues in differentiability
    - $I(\theta)$ is invertible in a neighborhood of θ^* or in the interval containing θ^*- for a one-dimensional case, $I(\theta)$ is a number and as long as it is non-zero, the condition is satisfied. For a multivariate case, $I(\theta)$ is a matrix, and we have to check the matrix is non-singular
    - Some other regularity conditions
  - With the above conditions satisfied, we have:
    - $\hat{\theta}_{MLE} \underset{n \to \infty}{\overset{P}{\to}} \theta^*$
    - $\sqrt{n}\left(\hat{\theta}_{MLE} - \theta^*\right) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, I(\theta^*)^{-1})$ (this shows that the bigger the Fisher information, the lesser is the variance and the better is the maximum likelihood estimator)
- **Multinomial models and MLE estimation**: Consider a random variable X that taken on values in a finite set $E = \{a_1, a_2, \ldots . a_K\}$ with $P(X = a_j) = p_j \forall\, j \in [1,K]$. The parameters $p = [p_1, p_2, \ldots p_K]$ have to be estimated using a set of independent and identically distributed samples $X_1, X_2, \ldots X_n$. The total number of outcomes possible, i.e. size of the sample space, is sometimes called the modality of the sample space.

- We can write the family of probability models as $(P_{\boldsymbol{p}})_{\boldsymbol{p}\in\Delta_K}$ where Δ_K is called the probability simplex and can be written as:
  - $\Delta_K = \{\, \boldsymbol{p} = (p_1, p_2, \dots p_K) \in (0,1)^K ; \sum_{j=1}^{K} p_j = 1\}$
  - Δ_K may be considered as the set of all possible probability mass functions that can be assigned to the sample space E
  - $P_{\boldsymbol{p}}$ is a vector of probabilities in R^K (more specifically, in $(0,1)^K$) which must sum up to 1. Hence $(P_{\boldsymbol{p}})^T(\mathbf{1})_K = 1$ where $(\mathbf{1})_K$ is a vector of all 1s in R^K
- We can write the likelihood function as:
  - $L_n(X_1, X_2, \dots X_n; p) = \prod_{i=1}^{n}\prod_{j=1}^{K}(p_j)^{I(X_i = a_j)}$
  - Since this is a product of products, this can be written as:
    - $L_n(X_1, X_2, \dots X_n; p) = \prod_{j=1}^{K}\prod_{i=1}^{n}(p_j)^{I(X_i = a_j)}$
    - $\prod_{i=1}^{n}(p_j)^{I(X_i = a_j)}$ is simply p_j raised to the power N_j where N_j – number of times a_j occurs in the n sample realizations of X
    - Hence, $L_n(X_1, X_2, \dots X_n; p) = \prod_{j=1}^{K}(p_j)^{N_j}$ with $\sum_{j=1}^{K} N_j = n$
- $\log L_n(X_1, X_2, \dots X_n; p) = \sum_{j=1}^{K} N_j \log p_j$
- To find the optimal estimator of p_j, we can take the partial derivatives of the log likelihood with respect to p_j and set it to 0. However, this does not take into account the fact that there is a constraint on p_j which is $\sum_{j=1}^{K} p_j = 1$.
  - To incorporate this constraint, we can write the log likelihood as:
    - $\log L_n(X_1, X_2, \dots X_n; p) = N_1 \log p_1 + N_1 \log p_2 + \cdots N_K \log\ (1 - \sum_{j=1}^{K-1} p_j)$
  - Partially differentiating the log likelihood equation with respect to $p_1, p_2, \dots, p_{K-1}$ and setting the derivatives equal to 0 gives $K-1$ equations of the form:
    - $\frac{N_j}{p_j} - \frac{N_K}{\left(1 - \sum_{j=1}^{K-1} p_j\right)} = 0$, or $p_j = \frac{N_j}{Y}\ for\ j \in [1, K-1]$ -**(A)**
      - where $Y = \frac{N_K}{\left(1 - \sum_{j=1}^{K-1} p_j\right)}$
    - Since $\sum_{j=1}^{K} p_j = 1$
      - $\frac{N_1}{Y} + \frac{N_2}{Y} + \cdots \frac{N_{K-1}}{Y} + p_K = 1$
      - Also, $p_K = 1 - \sum_{j=1}^{K-1} p_j = \frac{N_K}{Y}$
      - Or, $\frac{N_1}{Y} + \frac{N_2}{Y} + \cdots \frac{N_{K-1}}{Y} + \frac{N_K}{Y} = 1$
      - Since $N_1 + N_2 + \cdots . N_K = n$ where n is the total number of samples, we have $Y = n$ –**(B)**
    - Using (A) and (B), we have:
      - $\hat{p}_j = \frac{N_j}{n}\ \forall\ j \in [1, K]$ where $\hat{p}_j$ is the MLE estimator of p_j
    - The MLE estimators are also intuitively correct as they state that the estimated probability of a value is the frequency with which the value is observed in the sample
    - It can be shown that the equation $\hat{p}_j = \frac{N_j}{n}\ \forall\ j \in [1, K]$ holds even when $N_j = 0$ for some j and we cannot differentiate the log likelihood with respect to p_j – in this case, the corresponding estimator is 0. This is also

intuitively logical because if a particular value is never observed in a sample, it cannot be accorded a non-zero probability.

- The log-likelihood function can be shown to be concave, and therefore the MLE estimators represent the global maximum of the likelihood function.
- The asymptotic variance of the MLE estimator can be obtained by observing that the estimators are in the form of sample averages and hence we can apply the central limit theorem
  - $\hat{p}_j = \frac{N_j}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = a_j)\ \forall\ j \in [1, K]$
    - $\hat{p}_j$ is the sample average of Bernoulli variables $I(X_i = a_j)$
      - $I(X_i = a_j) = \begin{matrix} 1\ with\ probability\ p_j \\ 0\ with\ probability\ (1-p_j) \end{matrix}$
      - $E\left[I(X_i = a_j)\right] = p_j$
  - To find the asymptotic variance, we need to obtain the covariance matrix of
    - $\begin{pmatrix} I(X = a_1) \\ I(X = a_2) \\ \dots \\ I(X = a_K) \end{pmatrix}$
  - The covariance matrix Σ will be of size KxK.
  - Its diagonal elements will be variances of the random variables and non-diagonal elements will be the covariances
    - $\Sigma_{i,i} = Var\left(I(X = a_i)\right) = p_i(1-p_i)\ \forall\ i \in [1, K]$
    - $\Sigma_{i,j\ (i \neq j)} = Cov\left(I(X = a_i), I\left(X = a_j\right)\right) = -p_i p_j\ \forall\ i, j\ \in [1, K]$
      - This is because $E\left[I(X = a_i) I\left(X = a_j\right)\right] = 0$ as $I(X = a_i) I\left(X = a_j\right) = 0$ always
  - Hence, $\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \dots & -p_1 p_K \\ -p_2 p_1 & p_2(1-p_2) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ -p_K p_1 & \dots & \dots & p_K(1-p_K) \end{pmatrix}$
  - This covariance matrix can be shown to be non-invertible. As an example, if $K = 2$, then $\Sigma = p_1(1-p_1)p_2(1-p_2) - p_1^2 p_2^2 = 0$, as $(1-p_1) = p_2$ and $(1-p_2) = p_1$
    - The non-singular nature of the covariance matrix is on account of the fact that the individual probabilities are not independent, and are constrained by the equation $\sum_{j=1}^{K} p_j = 1$
- The Fisher information computed by taking the negative of the expectation of the second derivative of the log likelihood function using one copy of the random variable will not be meaningful in this case. This is principally because the true parameter is situated on the boundary of the parameter space
  - For instance, if we consider two dimensions only, while the parameter space would be a triangle with base 1 and height 1 (each of p_1 and p_2 can take values between 0 and 1), the true parameter must be situated on the hypotenuse, because of the constraint $p_1 + p_2 = 1$

- Also, as shown using the multivariate central limit theorem, the covariance matrix is singular, and non-invertible. The Fisher information matrix will also be a non-invertible matrix

Method of moments

- If $X_1, X_2, .. X_n$ are independent and identically distributed samples of a random variable X with an associated statistical model $(E, (P_\theta)_{\theta \in \Theta})$,
  - Let $E \in R$ and $\Theta \in R^d$, for some $d > 1$ (this means that there are d unknown parameters)
  - There will be d population moments and the k^{th} population moment, where $1 \leq k \leq d$ is given by $m_k(\theta) = E[X^k]$ where $m_k(\theta)$ is the k^{th} moment of the population distribution parameterized by θ
  - The d population moments are estimated using d sample moments:
    - $\widehat{m}_k(\theta) = \bar{X}_n^k = \frac{1}{n}\sum_{i=1}^n X_i^k$ for $1 \leq k \leq d$
  - From the law of large numbers, we can say that:
    - $\widehat{m}_k(\theta) \underset{n \to \infty}{\overset{P/a.s.}{\to}} E[X^k]$ for $1 \leq k \leq d$, or $\widehat{m}_k(\theta) \underset{n \to \infty}{\overset{P/a.s.}{\to}} m_k(\theta)$ for $1 \leq k \leq d$
  - Notationally, we can write that there exists a function M such that:
    - $M(\theta) = [m_1(\theta), m_2(\theta), \ldots . m_d(\theta)]$ where it may be remembered that θ is in R^d.
      - The inverse function M^{-1} is such that $M^{-1}\big(m_1(\theta), m_2(\theta), \ldots . m_d(\theta)\big) = \theta$
    - As an example, for a Gaussian distribution, there are two unknown parameters, μ and σ^2.
      - $M(\theta) = M(\mu, \sigma^2) = [\, \mu, \mu^2 + \sigma^2]$, and $M^{-1}(\mu, \mu^2 + \sigma^2) = (\mu, \sigma^2)$
    - Generally, we use the sample moments to estimate the unknown parameters, and hence:
      - $\hat{\theta} = M^{-1}\big(\widehat{m}_1(\theta), \widehat{m}_2(\theta), \ldots . \widehat{m}_d(\theta)\big)$
  - The central limit theorem can be applied to the method of moments as well; the estimator for X_i^k is $\widehat{m}_k(\theta)$ which is a sample average ($1 \leq k \leq d$), and hence:
    - $\sqrt{n}\big(\widehat{m}_k(\theta) - E\big[X_i^k\big]\big) \underset{n \to \infty}{\overset{d}{\to}} N(0, Var\big[X_i^k\big])$
    - All the moments and their estimators can be combined into one expression using the multivariate central limit theorem:
      - $\sqrt{n}\left(\begin{pmatrix} \widehat{m}_1(\theta) \\ \widehat{m}_2(\theta) \\ \ldots \\ \widehat{m}_d(\theta) \end{pmatrix} - \begin{pmatrix} E[X^1] \\ E[X^2] \\ \ldots \\ E[X^d] \end{pmatrix}\right) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, \Sigma(\theta))$
      - where $\Sigma(\theta)$ is the covariance matrix. If $\mathrm{X} = \begin{pmatrix} X^1 \\ X^2 \\ \ldots \\ X^d \end{pmatrix}$, and $\mu = \begin{pmatrix} E[X^1] \\ E[X^2] \\ \ldots \\ E[X^d] \end{pmatrix}$, then $\Sigma(\theta) = \mathrm{E}[(\mathrm{X} - \mu)(\mathrm{X} - \mu)^T] = \mathrm{E}[\mathrm{X}\mathrm{X}^T] - E[\mu\mu^T]$
  - The method of moments can have a more generalized formulation, where $m_k(\theta) = E[g_k(X)]$ and its estimator is $\widehat{m}_k(\theta) = = \frac{1}{n}\sum_{i=1}^n g_k(X)$. In this case, we can write the multivariate central limit theorem as:

- $\sqrt{n}(\begin{pmatrix}\widehat{m}_1(\theta)\\ \widehat{m}_2(\theta)\\ \dots\\ \widehat{m}_d(\theta)\end{pmatrix}-\begin{pmatrix}E[g_1(X)]\\ E[g_2(X)]\\ \dots\\ E[g_d(X)]\end{pmatrix}) \underset{n\to\infty}{\overset{d}{\to}} N_d(0,\Sigma(\theta))$
- If $\mathrm{X}=\begin{pmatrix}g_1(X)\\ g_2(X)\\ \dots\\ g_d(X)\end{pmatrix}$, and $\mu=\begin{pmatrix}E[g_1(X)]\\ E[g_2(X)]\\ \dots\\ E[g_d(X)]\end{pmatrix}$, then $\Sigma(\theta)=\mathrm{E}[(\mathrm{X}-\mu)(\mathrm{X}-\mu)^T]$

  - The multivariate delta method can be applied on the result of the multivariate central limit theorem.
    - $M(\theta)=[m_1(\theta),m_2(\theta),\dots.m_d(\theta)]$
      - In the most common case, $M(\theta)=\begin{pmatrix}E[X^1]\\ E[X^2]\\ \dots\\ E[X^d]\end{pmatrix}$ but in the more general case, it can be $\begin{pmatrix}E[g_1(X)]\\ E[g_2(X)]\\ \dots\\ E[g_d(X)]\end{pmatrix}$.
    - $M^{-1}\big(M(\theta)\big)=\theta$
    - If we assume M^{-1} is continuously differentiable at $M(\theta)$, then by the delta method:
      - $\sqrt{n}(M^{-1}\begin{pmatrix}\widehat{m}_1\\ \widehat{m}_2\\ \dots\\ \widehat{m}_d\end{pmatrix}-\begin{pmatrix}\theta_1\\ \theta_2\\ \dots\\ \theta_d\end{pmatrix}) \underset{n\to\infty}{\overset{d}{\to}} N_d(0,\Gamma(\theta))$
      - where $\Gamma(\theta)=[\frac{\partial M^{-1}}{\partial\theta}(M(\theta))]^T\Sigma(\theta)[\frac{\partial M^{-1}}{\partial\theta}\big(M(\theta)\big)]$

- In general, the MLE estimate is always preferred to the method of moments estimate, principally because its asymptotic variance is the lowest amongst all estimators, and it is quite robust to small misspecifications in the model. However, the computation of MLE can be intractable in many cases, and in such situations, we have to take recourse to the method of moments estimator.

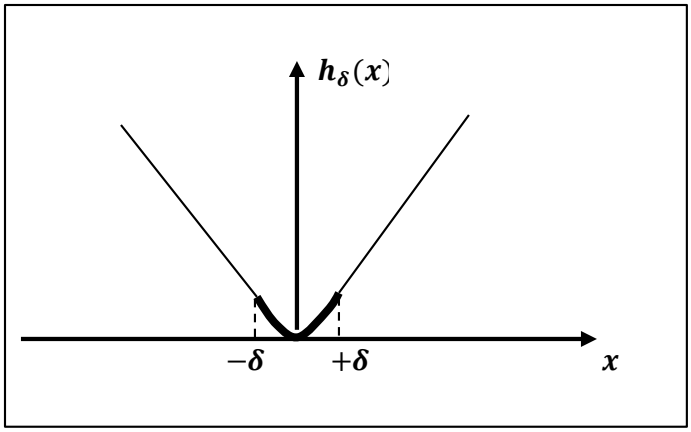
M-Estimation

- M-estimation techniques define a loss function that has to be minimized. No underlying distribution is assumed. If we assume that $X_1, X_2, \dots X_n$ are independent and identically distributed with respect to some unknown distribution P in some sample space E ($E \in R^d, d \geq 1$), then the goal of M-estimation is to estimate some parameter μ^* associated with P.
  - μ^* may be the mean, variance, median, a specific quartile, etc. of the unknown distribution
  - In mathematical terms, the objective of M-estimation is to find a function $\rho: ExM \rightarrow R$ where M is the set of all possible values for the unknown parameter μ^*, such that
    - $Q(u) = E[\rho(X, \mu)]$ achieves its minimum at $\mu = \mu^*$
    - It may be noted that that X is a random variable, with an associated probability distribution P which however is unknown. Using expectations and the law of large numbers allows us to use the empirical data on X (i.e. sample averages) to estimate the expected value without knowing the distribution.
    - The choice of the function ρ is important and will vary depending on the parameter whose estimate we seek. The key behind using M-estimation for finding statistical parameters is to define the function ρ such that it is minimized at the statistical parameter of interest.
  - As an example, if $E = M = R$, and $\rho(X, \mu) = (X - \mu)^2 \ \forall X \in R, \mu \in R$, then:
    - $Q(u) = E[(x - \mu)^2] \Rightarrow \frac{\partial Q(u)}{\partial u} = -2\mu E[X] + 2\mu$
    - Setting $\frac{\partial Q(u)}{\partial u} = 0$, we get $\mu^* = E[X]$
      - The function $\rho(X, \mu) = (X - \mu)^2$ has been chosen because it is minimized when μ is the expected value of X
    - In the case where X is a multi-dimensional vector (multivariate random variable or random vector) in R^d, then $\rho(X, \mu) = ||X - \mu||_2^2$ (which is the L2-norm or the square of the length of the vector). In this case $\mu^* = E[X]$ where now μ^* is a vector in R^d and is a collection of expectations of components of X
  - It can be shown that if $E = M = R$, and $\rho(X, \mu) = |X - \mu| \ \forall X \in R, \mu \in R$, then $\mu^* =$ median of X.
    - To estimate the quantiles, we use "check" functions. The check function can be considered to be a left tilt of the modulus function. The extent of the tilt is controlled by a parameter α. The check function is described as:
      - $C_\alpha(x) = \begin{cases} -(1-\alpha)x & if\ x < 0 \\ \alpha x & if\ x > 0 \end{cases}, \alpha \in (0,1)$
      - If we have $\rho(X, \mu) = C_\alpha(X - \mu)$, then μ^* is the α^{th} quantile of P
- To define the asymptotic normality of M-estimators, let us consider X to be a random vector in R^d which is associated with an unknown distribution P, and a sample space E.
  - $Q(u) = E[\rho(X, \mu)]$ is a function that achieves its minimum at $\mu = \mu^*, \mu \in R^d$.
  - We can define:

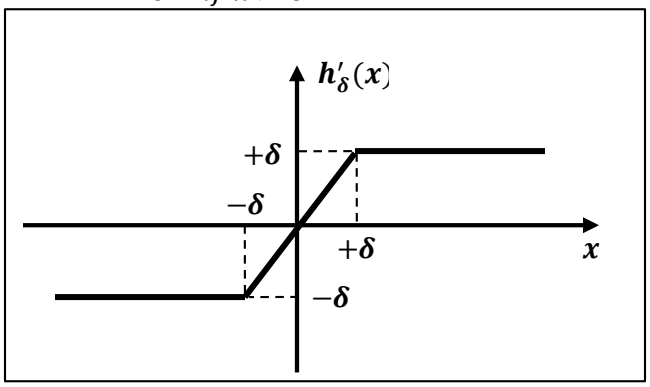
- $K(\mu) = Cov\big(\nabla\rho(X,\mu)\big) = Cov\begin{pmatrix} \frac{\partial\rho(X,\mu)}{\partial\mu_1} \\ \frac{\partial\rho(X,\mu)}{\partial\mu_2} \\ \cdots \\ \frac{\partial\rho(X,\mu)}{\partial\mu_d} \end{pmatrix}$ (This is a dxd matrix)
- $J(\mu) = -E(\boldsymbol{H}\rho) = -E\begin{pmatrix} \frac{\partial^2\rho(X,\mu)}{\partial\mu_1\partial\mu_1} & \frac{\partial^2\rho(X,\mu)}{\partial\mu_1\partial\mu_2} & \cdots & \frac{\partial^2\rho(X,\mu)}{\partial\mu_1\partial\mu_d} \\ \frac{\partial^2\rho(X,\mu)}{\partial\mu_2\partial\mu_1} & \frac{\partial^2\rho(X,\mu)}{\partial\mu_2\partial\mu_2} & \cdots & \frac{\partial^2\rho(X,\mu)}{\partial\mu_2\partial\mu_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2\rho(X,\mu)}{\partial\mu_d\partial\mu_1} & \frac{\partial^2\rho(X,\mu)}{\partial\mu_d\partial\mu_2} & \cdots & \frac{\partial^2\rho(X,\mu)}{\partial\mu_d\partial\mu_d} \end{pmatrix}$ (This is a dxd matrix)
    - J is also called the loss function

- If μ^* is a unique minimizer of $Q(u) = E[\rho(X,\mu)]$, the matrix J is invertible for all $\mu \in M$, and a few additional conditions hold, then the sample average of $\rho(X,\mu)$, $\hat{\mu}_n$ will be such that:
    - $\hat{\mu}_n \underset{n\to\infty}{\overset{P}{\to}} \mu^*$
    - $\sqrt{n}(\hat{\mu}_n - \mu^*) \underset{n\to\infty}{\overset{d}{\to}} N_d(0, J(\mu^*)^{-1}K(\mu^*)J(\mu^*)^{-1})$

- **Cauchy distribution**: The Cauchy distribution is a long-tailed distribution. A random variable X which follows a Cauchy distribution with parameter m has the pdf given by:
    - $f_m(x) = \frac{1}{\pi}\frac{1}{1+(x-m)^2}, x \in R$
    - The distribution is symmetrical about the parameter m, but its expectation tends to infinity, i.e. $E[X]$ is not well-defined.
    - Since the distribution is symmetrical about m, the median of the distribution is equal to m.
    - The method of moments cannot be used with the Cauchy distribution since even the first moment is not well-defined.
- **Huber's loss function**: In M-estimation, computing median and other parameters which involve expectation of a function involving modulus of a number is common. The modulus function is not continuously differentiable.
    - The Huber loss function is a modified modulus function where a small area around the discontinuity is replaced by a convex function. The rest of the function graph is linear.
        - $h_\delta(x) = \begin{cases} \frac{x^2}{2} \ if \ |x| < \delta \\ \delta\left(|x| - \frac{\delta}{2}\right) \ if \ |x| > \delta \end{cases}$



- The first derivative of the Huber loss function is continuous.
  - $\frac{\delta(h_\delta(x))}{\delta x} = \begin{cases} x & if\ |x| < \delta \\ -\delta & if\ x < -\delta \\ \delta & if\ x > \delta \end{cases}$



  - The loss function is clipped at $-\delta$ on the negative side and $+\delta$ on the positive side
- The second derivative of the Huber loss function is 1 for $-\delta \leq x \leq +\delta$ and 0 elsewhere

- **Applying Huber's loss to compute MLE of Laplace distribution**: The Laplace distribution is also known as the double exponential distribution. It has a location parameter m, which is also the point of symmetry of the distribution.
  - The pdf of the Laplace distribution with parameter m is:
    - $f_m(x) = \frac{1}{2}e^{-|x-m|},\ x \in R$
    - m is both the mean and median of the Laplace distribution
    - The variance and higher order moments of the Laplace distribution are well-defined. In general, $x^k e^{-|x|}$ is integrable on R for all $k > 0$
  - The log-likelihood of the Laplace distribution is:
    - $\log L(x_1, x_2, x_3, \ldots x_n; m) = -n \log 2 - \sum_{i=1}^{n} |x_i - m|$
    - Therefore, $\hat{m}_{MLE} = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - m|$
    - $\underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - m|$ is equal to the median of the sampled observations (empirical median)

- This is because, for M-estimation, it can be shown that if $\rho(X,\mu) = |X - \mu|\ \forall\ X \in R, \mu \in R$, then $\mu^* =$ median of X

- It may be noted that the log likelihood is one of the functional forms of the M-estimator. Specifically, for the log likelihood, $\rho(X,\mu) = \log L(X,\mu)$. Here, the unknown parameter is being denoted by m instead of μ.

- To compute asymptotic variance of the estimator within the framework of M-estimation, we need to obtain the first and second order derivatives of $\log L(X;m) = -\log 2 - |X - m|$.
  - Also, we need to establish that there is a unique minimizer of $E[\log L(X;m)]$. This can be easily proved given that we know that $\log L(x_1, x_2, x_3, \ldots x_n; m)$ minimizes when we set m = empirical median of the observations. The expected value is the integral of $\log L(x_1, x_2, x_3, \ldots x_n; m)$ multiplied by $f_m(x)$ – if we take the minimum value of $\log L(x_1, x_2, x_3, \ldots x_n; m)$, we get the minimum value of the expectation.
  - Finally, we also need to establish that the expected value of the second derivative is invertible.
  - The constant logarithmic term will vanish in the first and second order derivatives.
- We can approximate $|X - m|$ as $h_\delta(X - m)$, or
  - $\log L(X;m) = -\log 2 - h_\delta(X - m)$ where
    - $h_\delta(X - m) = \begin{cases} \frac{(X-m)^2}{2} & if\ |X - m| < \delta \\ \delta\left(|X - m| - \frac{\delta}{2}\right) & if\ |X - m| > \delta \end{cases}$
  - Also, $X - m$ follows a Laplace distribution with location parameter 0
  - We rewrite the estimator as $\widehat{m}_{MLE} = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} h_\delta(X - m)$
- $J(m) = \frac{\delta^2 h_\delta(X-m)}{\delta m^2} = \begin{cases} 1 & if\ |X - m| < \delta \\ 0 & if\ |X - m| > \delta \end{cases}$
  - $-E\left[\frac{\delta^2 \log L(X;m)}{\delta m^2}\right] = E\left[\frac{\delta^2 h_\delta(X-m)}{\delta m^2}\right]$, or
    - $-E\left[\frac{\delta^2 \log L(X;m)}{\delta m^2}\right] = \int_{m-\delta}^{m+\delta} \frac{1}{2} e^{-|x-m|}\, dx = 2\int_{m}^{m+\delta} \frac{1}{2} e^{-|x-m|}\, dx$ (by symmetry of the distribution of X around its location parameter which is m here)
    - $-E\left[\frac{\delta^2 \log L(X;m)}{\delta m^2}\right] = \int_{m}^{m+\delta} e^{(m-x)}\, dx = 1 - e^{-\delta} = J(m)$
    - $J(m)$ is invertible for the range of possible values m can take as it is independent of m <Is this correct?>
- $\frac{\delta \log L(X;m)}{\delta m} = \begin{cases} (X - m) & if\ |X - m| < \delta \\ -\delta & if\ X - m > \delta \\ +\delta & if\ m - X > \delta \end{cases}$
  - $E\left[\frac{\delta \log L(X;m)}{\delta m}\right] = \frac{1}{2}\left(\int_{-\infty}^{m-\delta} \delta\, e^{(x-m)} dx + \int_{m+\delta}^{\infty} -\delta\, e^{-(x-m)} dx + \int_{m-\delta}^{m+\delta} (x - m)\, e^{-|x-m|} dx\right)$
    - $\int_{-\infty}^{m-\delta} \delta\, e^{(x-m)}\, dx = \delta e^{-\delta}$
    - $\int_{m+\delta}^{\infty} -\delta\, e^{-(x-m)} dx = -\delta e^{-\delta}$

- $\int_{m-\delta}^{m+\delta}(x-m)\,e^{-|x-m|}dx = \int_{-\delta}^{+\delta} y\,e^{-|y|}dy$ (on replacing $x-m$ with y)
  - $\int_{-\delta}^{+\delta} y\,e^{-|y|}dy = \int_{-\delta}^{0} y\,e^{y}dy + \int_{0}^{\delta} y\,e^{-y}dy$, or
  - $\int_{-\delta}^{+\delta} y\,e^{-|y|}dy = 0$ (using integration by parts)
- Thus, $E\left[\frac{\delta \log L(X;m)}{\delta m}\right] = 0$

- $K(m) = Cov(\frac{\delta \log L(X;m)}{\delta m})$
  - $Cov\left(\frac{\delta \log L(X;m)}{\delta m}\right) = Var(\frac{\delta \log L(X;m)}{\delta m})$
  - We know that:
    - $\frac{\delta \log L(X;m)}{\delta m} = \begin{cases} (X-m) & if\ |X-m| < \delta \\ -\delta & if\ X-m > \delta \\ +\delta & if\ m-X > \delta \end{cases}$
  - $Var\left(\frac{\delta \log L(X;m)}{\delta m}\right) = E\left[\left(\frac{\delta \log L(X;m)}{\delta m}\right)^2\right]$ as $E\left[\frac{\delta \log L(X;m)}{\delta m}\right] = 0$
  - $E\left[\left(\frac{\delta \log L(X;m)}{\delta m}\right)^2\right] = \frac{1}{2}\left(\int_{-\infty}^{m-\delta}\delta^2 e^{(x-m)}dx + \int_{m+\delta}^{\infty}(-\delta)^2 e^{-(x-m)}dx + \int_{m-\delta}^{m+\delta}(x-m)^2 e^{-|x-m|}dx\right)$ – (A)
    - $\int_{-\infty}^{m-\delta}\delta^2 e^{(x-m)}dx = \delta^2 e^{-\delta}$ – (B)
    - $\int_{m+\delta}^{\infty}(-\delta)^2 e^{-(x-m)}dx = \delta^2 e^{-\delta}$ – (C)
    - $\int_{m-\delta}^{m+\delta}(x-m)^2 e^{-|x-m|}dx = \int_{-\delta}^{+\delta} y^2 e^{-|y|}dy$ (on replacing $x-m$ with y)
      - $\int_{-\delta}^{+\delta} y^2 e^{-|y|}dy = \int_{-\delta}^{0} y^2 e^{y}dy + \int_{0}^{\delta} y^2 e^{-y}dy$, or
        - $\int_{-\delta}^{0} y^2 e^{y}dy = -\delta^2 e^{-\delta} - 2\int_{-\delta}^{0} y\,e^{y}dy$, or
        - $\int_{-\delta}^{0} y^2 e^{y}dy = -\delta^2 e^{-\delta} - 2(\delta e^{-\delta} - \int_{-\delta}^{0} e^{y}\,dy)$, or
        - $\int_{-\delta}^{0} y^2 e^{y}dy = -\delta^2 e^{-\delta} - 2\delta e^{-\delta} + 2 - 2e^{-\delta}$
        - Also, $\int_{0}^{\delta} y^2 e^{-y}dy = -\delta^2 e^{-\delta} - 2\delta e^{-\delta} + 2 - 2e^{-\delta}$
      - Thus, $\int_{-\delta}^{+\delta} y^2 e^{-|y|}dy = -2\delta^2 e^{-\delta} - 4\delta e^{-\delta} + 4 - 4e^{-\delta}$- (D)
  - Using (B), (C) and (D) in the expression for (A) gives:
    - $E\left[\left(\frac{\delta \log L(X;m)}{\delta m}\right)^2\right] = \delta^2 e^{-\delta} - \delta^2 e^{-\delta} - 2\delta e^{-\delta} + 2 - 2e^{-\delta}$- , or
    - $E\left[\left(\frac{\delta \log L(X;m)}{\delta m}\right)^2\right] = -2\delta e^{-\delta} + 2 - 2e^{-\delta}$
  - Thus, $Var\left(\frac{\delta \log L(X;m)}{\delta m}\right) = -2\delta e^{-\delta} + 2 - 2e^{-\delta} = K(m)$
- The asymptotic variance of the estimator $\hat{m}_{MLE}$ is given by $J(m^*)^{-1}K(m^*)J(m^*)^{-1}$.
  - $J(m^*)^{-1}K(m^*)J(m^*)^{-1} = \frac{-2\delta e^{-\delta}+2-2e^{-\delta}}{(1-e^{-\delta})^2}$
  - The asymptotic variance as computed using Huber's loss can be shown to be the minimum when $\delta = 0$, and at $\delta = 0$, the asymptotic variance is 1.
    - Also, when $\delta = 0$, the Huber's loss function becomes the same as the modulus function, and $\hat{m}_{MLE} = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} h_\delta(X-m)$

- The maximum value of the asymptotic variance is 2 and this happens when $\delta \to \infty$. It may be noted that by construction of the Huber loss function, δ cannot take negative values.
    - When $\to \infty$, the Huber's loss function becomes equal to the squared loss function, i.e. $h_\delta(X-m) = \frac{(X-m)^2}{2}$
    - $\widehat{m}_{MLE} = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} (\frac{(X-m)^2}{2}) = \bar{X}_n$

Advanced Hypothesis Testing

- **Chi-squared distribution**: For a positive integer d, the χ^2 distribution with d degrees of freedom is the sum of d independent standard normal variables.
  - $\chi_d^2 = Z_1^2 + Z_2^2 + \cdots Z_d^2$ where $Z_i \sim N(0,1)$
  - d is known as the degrees of freedom of the chi-squared distribution
  - The chi-squared distribution can also be considered to be the square of the L2 norm of a Gaussian vector in R^d with mean 0 and covariance matrix equal to the identity matrix – such a Gaussian vector has components each of which is a standard normal random variable and the components are independent of each other
    - $\chi_d^2 = \|Z\|_2^2$ where $Z \sim N_d(0, I_d)$
  - The support of the chi-squared distribution comprises non-negative real numbers
  - The probability density of the chi-squared distribution is:
    - $f_{\chi_k^2}(x) = \frac{x^{(k/2-1)} e^{-\frac{x}{2}}}{2^{k/2}\,\Gamma(\frac{k}{2})}$ for $x > 0$ where $\Gamma(\frac{k}{2})$ denotes the gamma function
      - For a positive integer n, $\Gamma(n) = (n-1)!$
    - $E[\chi_k^2] = E[Z_1^2 + Z_2^2 + \cdots Z_k^2] = k$ (for the standard normal variable Z, $E[Z^2] = 1$
    - $Var[\chi_k^2] = Var[Z_1^2 + Z_2^2 + \cdots Z_k^2] = 2k$
      - For the standard normal variable Z, $Var[Z^2] = E[Z^4] - E[Z^2]^2 = 3 - 1 = 2$
      - The standard deviation is of the order of the square root of the expectation
    - For a large enough value of k, by the Central Limit Theorem, $\chi_k^2 \cong N(k, 2k)$
    - For $k = 2$, $f_{\chi_k^2}(x) = \frac{e^{-\frac{x}{2}}}{2} = Exp(\frac{1}{2})$
- **Cochran's theorem**: If $X_1, X_2, \ldots X_n$ are independent and identically distributed normal variables such that $X_i \sim N(\mu, \sigma^2)\ \forall\ i \in [1, n]$ and $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ (where S_n is the sample variance and $\bar{X}_n$ is the sample mean), then:
  - S_n and $\bar{X}_n$ are independent; in vector terms, they are perpendicular vectors
  - $\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$
  - Generally, we take the sample variance as $\tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ – this is an unbiased estimator of the population variance
    - From the equation $\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$, we have $E[\frac{nS_n}{\sigma^2}] \sim E[\chi_{n-1}^2]$, or $E[S_n] = \sigma^2 \frac{n-1}{n}$
    - $\tilde{S}_n = \frac{n}{n-1} S_n$, and hence $E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \sigma^2$
    - Also, since $nS_n = (n-1)\tilde{S}_n$, $\frac{(n-1)\tilde{S}_n}{\sigma^2} \sim \chi_{n-1}^2$
- **Student's T-distribution**: The Student's T-distribution is used in the case of small sample sizes.
  - If $X_1, X_2, \ldots X_n$ are independent and identically distributed normal variables such that $X_i \sim N(\mu, \sigma^2)\ \forall\ i \in [1, n]$, then:
    - $Z = \frac{\bar{X}_n - \mu}{(\frac{\sigma}{\sqrt{n}})}$

- $\tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, and $\frac{(n-1)S_n}{\sigma^2} \sim \chi^2_{n-1}$
- Since $\bar{X}_n$ and $\tilde{S}_n$ are independent by the Cochran theorem, the ratio $\frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}}$ is the ratio of two independent random variables

  - $\frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{S_n}}$ follows the t-distribution with degrees of freedom (n-1)

- **T-Tests**: When doing hypothesis testing using the T-distribution, we are not operating under asymptotic conditions. Hence, we need that the independent and identically distributed samples come from a normal distribution.
  - If $X_1, X_2, \ldots X_n$ are independent and identically distributed normal variables such that $X_i \sim N(\mu, \sigma^2)\ \forall\, i \in [1, n]$ and we have a two-sided, one-sample test with the hypothesis of the form $H_0: \mu = 0$ and $H_1: \mu \neq 0$, then:
    - The t-statistic is of the form $T_n = \frac{\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma})}{\sqrt{(\frac{S_n}{\sigma^2})}} = \frac{\sqrt{n}\bar{X}_n}{\sqrt{S_n}}$ where $\tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is the unbiased estimator of the population variance
    - The t-distribution has $(n-1)$ degrees of freedom, and hence the null hypothesis would have to be rejected at a significance level α if:
      - $|T_n| > q_{\alpha/2}$ where $q_{\alpha/2}$ is the $(1 - \frac{\alpha}{2})$ quantile of the t_{n-1} distribution
  - If we have a one-sided test where the hypotheses are of the form $H_0: \mu \leq \mu_0$ and $H_1: \mu > \mu_0$, then the rejection criteria would be:
    - $\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{S_n}} > q_\alpha$ where q_α is the $(1-\alpha)$ quantile of the t_{n-1} distribution
  - For a one-sided test where the hypotheses are of the form $H_0: \mu \geq \mu_0$ and $H_1: \mu < \mu_0$, the rejection criteria would be:
    - $\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{S_n}} < q_{(1-\alpha)}$ where $q_{(1-\alpha)}$ is the α quantile of the t_{n-1} distribution
  - For two-sample tests, the test statistic is of the form:
    - $\frac{(\bar{X}_n - \bar{Y}_m) - (Value\ of\ difference\ under\ null\ hypothesis)}{\sqrt{\frac{\hat{\sigma}^2_X}{n} + \frac{\hat{\sigma}^2_Y}{m}}}$
    - For the degrees of freedom, we can take a conservative approach and choose the degrees of freedom as $min\ (m, n)$.
    - Alternatively, we can use the Welch-Satterthwaite formula which gives the minimum degrees of freedom as:
      - $\left\lfloor \frac{(\frac{\hat{\sigma}^2_X}{n} + \frac{\hat{\sigma}^2_Y}{m})^2}{\frac{\hat{\sigma}^4_X}{n^2(n-1)} + \frac{\hat{\sigma}^4_Y}{m^2(m-1)}} \right\rfloor$ which is greater than $min\ (m-1, n-1)$ but less than $(m+n)$
  - If the number of samples if large, the t-distribution will converge to the normal distribution since the central limit theorem kicks in. With significantly large sample sizes, the sample need not be independent and identically distributed normal variables – when we can apply the Central Limit theorem, this is not a limitation. Often, in statistical tests,

we compute t-statistics and use t-tests even though the sample size is large, because the results will be more or less identical to the case when we used normal distributions. This also allows for more conservative testing, rather than relying on asymptotics completely.

- **Comparing two proportions**: If we have two samples $(X_1, X_2, \dots X_n)$ and $(Y_1, Y_2, \dots Y_n)$ such that:
  - $X_1, X_2, \dots X_n$ are independent and identically distributed according to a Bernoulli distribution with parameter $p_X \; where \; p_X \in (0,1)$
  - $Y_1, Y_2, \dots Y_n$ are independent and identically distributed according to a Bernoulli distribution with parameter $p_Y \; where \; p_Y \in (0,1)$
  - X and Y are independent of each other with an equal number of sample observations of both variables
  - If we want to test that the two sets of observations come from the same underlying population, our hypotheses are of the form:
    - $H_0: p_X = p_Y$ and $H_1: p_X \neq p_Y$
  - The estimator for p_X is $\bar{p}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the estimator for p_Y is $\bar{p}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$
    - By the law of large numbers, the estimators converge in probability to the estimand as the sample size increases
    - By the multivariate central limit theorem, we have:
      - $\sqrt{n}\left(\begin{pmatrix}\bar{p}_X \\ \bar{p}_Y\end{pmatrix} - \begin{pmatrix}p_X \\ p_Y\end{pmatrix}\right) \underset{n \to \infty}{\overset{d}{\to}} N_2\left(0, \begin{pmatrix}p_X(1-p_X) & 0 \\ 0 & p_Y(1-p_Y)\end{pmatrix}\right)$
      - Since we want to find the distribution of the estimator for $p_X - p_Y$, we use the delta method with $g(x,y) = x - y$ and $\nabla g(x,y) = \begin{pmatrix}1 \\ -1\end{pmatrix}$
      - $\sqrt{n}((\bar{p}_X - \bar{p}_Y) - (p_X - p_Y)) \underset{n \to \infty}{\overset{d}{\to}} N(0, p_X(1-p_X) + p_Y(1-p_Y))$
  - Under the null hypothesis, $p_X = p_Y$. Let us consider that $p_X = p_Y = p$
    - $p_X(1-p_X) + p_Y(1-p_Y) = 2p(1-p)$
    - Since we do not know the values of p_X or p_Y, we use an estimator for p.
      - $\hat{p} = \frac{1}{2}(\bar{p}_X + \bar{p}_Y)$ which will asymptotically converge in probability to p
    - Hence the test statistic becomes:
      - $T_n = \frac{\sqrt{n}(\bar{p}_X - \bar{p}_Y)}{\sqrt{2\hat{p}(1-\hat{p})}}$ which converges to $Z \sim N(0,1)$ under the null hypothesis
      - At a significance level of α, the criterion for rejecting the null hypothesis is $\left|\frac{\sqrt{n}(\bar{p}_X - \bar{p}_Y)}{\sqrt{2\hat{p}(1-\hat{p})}}\right| > q_{\alpha/2}$
- **Positive semi-definite matrices**: A matrix A of size dxd is positive semi-definite if $x^T A x \geq 0 \; \forall \; x \in R^d$.
  - Diagonal matrices with non-negative entries are always positive semi-definite, i.e. if D is such a diagonal matrix, then $x^T D x \geq 0$
  - If P is an invertible matrix, and D is a diagonal matrix whose entries are non-negative then $P^T D P$ is a positive semi-definite matrix. This is because:
    - $x^T P^T D P x = (Px)^T D (Px)$. If we write $Px = y$, this is equal to $y^T D y \geq 0$
  - The square root of a positive semi-definite matrix B is a matrix A such that $AA = B$. In other words, $A = B^{\frac{1}{2}}$.

        - For a diagonal matrix whose entries are non-negative, the square root of the matrix is simply the matrix formed by taking the square root of the elements on the principal diagonal.
            - As an example, if $D = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $D^{1/2} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{pmatrix}$

- **Wald's test**: Under the Wald statistical test, the maximum likelihood estimate of the parameter of interest is compared with the proposed value, with the assumption that the difference between the two will be approximately normally distributed. Typically the square of the difference is compared to a chi-squared distribution.
    - In the case of maximum likelihood estimation of a multivariate random variable, we have:
        - $\sqrt{n}\left(\hat{\theta}_{MLE} - \theta^*\right) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, I(\theta^*)^{-1})$ where:
            - θ^* is the true parameter value (this would also be the value under the null hypothesis, and $I(\theta^*)$ is the Fisher information- this will be a matrix for a multivariate distribution)
        - We can rewrite the equation as: $\sqrt{n} I(\theta^*)^{1/2}\left(\hat{\theta}_{MLE} - \theta^*\right) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, I_d)$
        - If we take the Euclidean norm of both sides of the above equation (Euclidean norm of a vector X, $\|X\|_2^2 = X^T X$), we have:
            - $n\left(\left(\hat{\theta}_{MLE} - \theta^*\right)^T I(\theta^*)^{\frac{1}{2}^T} \right) I(\theta^*)^{\frac{1}{2}} \left(\hat{\theta}_{MLE} - \theta^*\right) \underset{n \to \infty}{\overset{d}{\to}} \chi_d^2$
            - Or, $n\left(\hat{\theta}_{MLE} - \theta^*\right)^T I(\theta^*)\left(\hat{\theta}_{MLE} - \theta^*\right) \underset{n \to \infty}{\overset{d}{\to}} \chi_d^2$ since $I(\theta^*)^{\frac{1}{2}}$ is a symmetrical matrix
    - In the Wald test, the test statistic T_n is $n\left(\hat{\theta}_{MLE} - \theta^*\right)^T I(\theta^*)\left(\hat{\theta}_{MLE} - \theta^*\right)$ and it follows a Chi-square distribution.
        - The rejection criteria is $T_n > q_\alpha$ where q_α is the $(1-\alpha)$ quantile of the χ_d^2 distribution
    - Wald's test in dimensions higher than 2 is testing the degree of divergence of the estimator vector and the true vector of the parameters. As such, there is no notion of the estimated vector being greater than or smaller than the true vector
        - The Wald's test in 1 dimension can be shown to be equivalent to a two sided test
            - In one dimension, the rejection criteria of the Wald's test is:
                - $n\frac{\left(\hat{\theta}_{MLE} - \theta^*\right)^2}{\sigma^2} > q_\alpha(\chi_1^2)$
                - Or, $\frac{|\hat{\theta}_{MLE} - \theta^*|}{\frac{\sigma}{\sqrt{n}}} > \sqrt{q_\alpha(\chi_1^2)}$ – (A)
            - By definition, $P\left(Z^2 > q_\alpha(\chi_1^2)\right) = \alpha$
                - Or, $P\left(|Z| > \sqrt{q_\alpha(\chi_1^2)}\right) = \alpha$
                - Or, $2(1 - P\left(Z < \sqrt{q_\alpha(\chi_1^2)}\right) = \alpha$
                - Or, $P\left(Z < \sqrt{q_\alpha(\chi_1^2)}\right) = 1 - \frac{\alpha}{2}$

- This implies that $\sqrt{q_\alpha(\chi_1^2)} = q_{\alpha/2}(N(0,1)) - (B)$
- Using (B) in (A), we have:
  - $n\frac{\left(\hat{\theta}_{MLE}-\theta^*\right)^2}{\sigma^2} > q_\alpha(\chi_1^2) \Leftrightarrow \frac{|\hat{\theta}_{MLE}-\theta^*|}{\frac{\sigma}{\sqrt{n}}} > \sqrt{q_\alpha(\chi_1^2)} \Leftrightarrow$ $\frac{|\hat{\theta}_{MLE}-\theta^*|}{\frac{\sigma}{\sqrt{n}}} > q_{\alpha/2}(N(0,1))$

- **Likelihood ratio test**: The likelihood ratio test is based on the premise that if a certain hypothesis is more likely than an alternative hypothesis, the maximum of the likelihood function under that hypothesis should be greater than the maximum of the likelihood function under the alternative hypothesis. It compares goodness of fit of the observed data under the null and the alternative hypothesis to make a judgement.
  - If $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_0^C$ where $\Theta_0 \cup \Theta_0^C = \Theta$ where Θ is the entire parameter space, then the test-statistic for the basic likelihood ratio test can take on one of the following two forms:
    - $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0^C)}$ or $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta)}$
    - The criteria for rejection is when $T_n < C$ where C is a defined threshold. The principle here is that if the maximum likelihood under the null hypothesis is much lesser than the maximum likelihood under the alternative, then the null hypothesis can be rejected.
    - We can also formulate $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0^C)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0)}$ or $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta \in \Theta_0)}$ in which case the criteria for rejection would become $T_n > C$
    - If both the null and the alternative hypothesis are simple hypothesis such as $H_0: \theta = \theta_0$ and $H_1: \theta = \theta_1$, then the test has to be formulated as:
      - $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta = \theta_0)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta = \theta_1)}$ with rejection criteria of the form $T_n < C$
      - Or, $T_n = \frac{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta = \theta_1)}{\sup_\theta L(x_1,x_2 \dots x_2;\theta \,|\theta = \theta_0)}$ with rejection criteria $T_n > C$
  - Often, when dealing with multi-dimensional parameters, we want only some of the parameters to be specified under the null and the alternative hypothesis, and allow the rest of the parameters to vary.
    - We can use the Wilk's test for this purpose.
      - We have independent and identically distributed random variables $X_1, X_2, \dots X_n$ with the statistical model $(E, (P_\theta)_{\theta \in \Theta})$ where $\theta \in R^d$
      - The null hypothesis and alternative hypothesis have the form
        - $H_0: (\theta_{r+1}, \theta_{r+2}, \dots \theta_d) = \left(\theta_{r+1}^{(0)}, \theta_{r+2}^{(0)}, \dots \theta_d^{(0)}\right)$ for $r \geq 0$ where $\left(\theta_{r+1}^{(0)}, \theta_{r+2}^{(0)}, \dots \theta_d^{(0)}\right)$ is specified
        - $H_1: (\theta_{r+1}, \theta_{r+2}, \dots \theta_d) \neq \left(\theta_{r+1}^{(0)}, \theta_{r+2}^{(0)}, \dots \theta_d^{(0)}\right)$

        - Basically, this means that the first r components of the parameter vector are variable for the purpose of the hypothesis test
    - Let $\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \left(\log(X_1, X_2, \dots X_n; \theta)\right)$ (maximum likelihood estimator of θ, and this is not based on the hypothesis)
    - Let $\hat{\theta}_n^c = \underset{\theta \in \Theta_0}{\operatorname{argmax}} \left(\log(X_1, X_2, \dots X_n; \theta\right)$ where $(\theta_{r+1}, \theta_{r+2}, \dots \theta_d) = \left(\theta_{r+1}^{(0)}, \theta_{r+2}^{(0)}, \dots \theta_d^{(0)}\right)$ (this is a constrained maximum likelihood estimator)
  - The test statistic under the Wilk's test is:
    - $T_n = 2(\log(X_1, X_2, \dots X_n; \hat{\theta}_n) - \log(X_1, X_2, \dots X_n; \hat{\theta}_n^c))$
    - $T_n \underset{n \to \infty}{\overset{(d)}{\to}} \chi^2_{d-r}$
    - The rejection criteria is $T_n > q_\alpha$ where q_α is the $(1-\alpha)$ quantile of χ^2_{d-r}
      - This essentially means that if the maximum likelihood without any constraints imposed is much greater than the maximum likelihood under the null hypothesis, then the null hypothesis can be rejected
- **Implicit Hypothesis Testing**: Implicit hypothesis testing is a general framework for carrying out multiple types of hypothesis tests when the parameter vector is multi-dimensional. For instance, for $\theta \in R^d$, we may have as the null hypothesis the proposition that the first and second components of θ are equal, i.e. $\theta_1 = \theta_2$. In this case, we do not know the value of the parameter under the null hypothesis fully, and hence the name implicit hypothesis test
  - To carry out implicit hypothesis testing, we need an asymptotic estimator – this could be based on the multivariate central limit theorem, or on the Fisher information for MLE estimators, or on a more general approach involving first and second order derivatives for method of moments and M-estimation. We use the delta method and Wald's test structure to obtain a test statistic. In implicit hypothesis testing, we do not know the value of the parameter under the null hypothesis fully.
    - We assume that we have an asymptotic estimator $\hat{\theta}_n$ of the true parameter θ such that:
      - $\sqrt{n}(\hat{\theta}_n - \theta) \underset{n \to \infty}{\overset{d}{\to}} N_d(0, \Sigma(\theta))$ and $\Sigma(\theta)$ is invertible
    - g is a function such that g: $R^d \to R^k$ is continuously differentiable $(k < d)$.
      - The function g is chosen based on our partial hypothesis. For instance, if we want to test $\theta_1 = \theta_2$, we want $g(\theta_1, \theta_2) = \theta_1 - \theta_2$
      - If we want to test $(\theta_{r+1}, \theta_{r+2}, \dots \theta_d) = \left(\theta_{r+1}^{(0)}, \theta_{r+2}^{(0)}, \dots \theta_d^{(0)}\right)$, we would have $g(\theta_{r+1}, \theta_{r+2}, \dots \theta_d) = \begin{pmatrix} \theta_{r+1} - \theta_{r+1}^{(0)} \\ \dots \\ \theta_d - \theta_d^{(0)} \end{pmatrix}$
    - In the implicit hypothesis test, the null and the alternative hypothesis are:
      - $H_0: g(\theta) = 0$ and $H_1: g(\theta) \neq 0$

- Using the multivariate delta method, we have:
  - $\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \underset{n \to \infty}{\overset{d}{\to}} N_k(0, \Gamma(\theta))$
    - where $\Gamma(\theta) = \nabla g(\theta)^T \Sigma(\theta) \nabla g(\theta) \in R^{kxk}$ is invertible. This requires $\nabla g(\theta)$ to have rank k (i.e. it must be a full column rank matrix) such that $\nabla g(\theta)^T \Sigma(\theta) \nabla g(\theta)$ is a product of three invertible matrices and is itself invertible
  - We can write $\sqrt{n}\Gamma(\theta)^{-\frac{1}{2}}(g(\hat{\theta}_n) - g(\theta)) \underset{n \to \infty}{\overset{d}{\to}} N_k(0, \mathrm{I}_k)$
- Taking the squared Euclidean norm of both sides of the above vector and considering that under the null hypothesis, we have $g(\theta) = 0$, we have:
  - Under the null hypothesis, $ng(\hat{\theta}_n)^T \Gamma(\theta)^{-1} g(\hat{\theta}_n) \underset{n \to \infty}{\overset{d}{\to}} \chi_k^2$
- Since we do not know θ fully under the null hypothesis, in the expression for $\Gamma(\theta)$, we use the MLE estimator for θ, $\hat{\theta}_n$ knowing that if the null hypothesis is true, the estimator will asymptotically converge to θ as defined by the null hypothesis.
  - Thus, we have $ng(\hat{\theta}_n)^T (\nabla g(\hat{\theta}_n)^T \Sigma(\hat{\theta}_n) \nabla g(\hat{\theta}_n))^{-1} g(\hat{\theta}_n) \underset{n \to \infty}{\overset{d}{\to}} \chi_k^2$
- The test statistic is $ng(\hat{\theta}_n)^T (\nabla g(\hat{\theta}_n)^T \Sigma(\hat{\theta}_n) \nabla g(\hat{\theta}_n))^{-1} g(\hat{\theta}_n)$

- **Goodness of fit testing for the multinomial distribution**: In the multinomial distribution, the null hypothesis is generally of the form that the PMF of the variable is a uniform distribution, i.e. all values are equally likely. The alternative hypothesis is that the PMF is not a uniform distribution
  - If $((a_1, a_2, .. a_K), (P_p)_{p \in \Delta_K})$ is the statistical model for the multinomial distribution with K discrete values in the sample space and Δ_K is in R^K, then the null and alternative hypothesis generally take the form:
    - $H_0: p = p^0$, $H_1: p \neq p^0$ where $p^0 = \left[\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right] \in R^K$
    - In case the population proportions are known, and we want to test whether a sample comes from the population, then p^0 is the vector of population proportions
  - If $\hat{p}$ is the MLE estimator of p, then $\sqrt{n}(\hat{p} - p)$ is asymptotically normal, or $\sqrt{n}(\hat{p} - p)$ is a Gaussian vector.
    - However, $\sum_{j=1}^K \hat{p}_j = 1$ (by derivation) and of course, $\sum_{j=1}^K p_j = 1$
    - This means that $(\hat{p}_j - p_j)^T (\mathbf{1})_K = 0$ where $(\mathbf{1})_K$ is the all 1s vector in dimension K. 0 is a degenerate Gaussian random variable, all its mass is at 0 and it has no variance.
    - Since a linear combination of the vectors in $(\hat{p}_j - p)$ gives zero, $\sqrt{n}(\hat{p} - p)$ is not asymptotically normal in N_K. This conclusion is further reinforced by examining the covariance matrix of $\sqrt{n}(\hat{p} - p)$ which non-invertible.

- $\sqrt{n}(\hat{p}-p)$ is asymptotically normal in N_{K-1} – one degree of freedom is lost because of the necessity that the individual probabilities must add up to 1. If we know $(K-1)$ of the probabilities, we know with certainty the remaining one.

- If H_0 is true, then $n\sum_{j=1}^{K}\frac{(\hat{p}_j-p_j^0)^2}{p_j^0} \underset{n\to\infty}{\overset{(d)}{\to}} \chi^2_{K-1}$
- The Chi-squared test can also be used in the case when the null distribution parameters have to be estimated from the data itself. The estimators are MLE estimators.
  - Let a random variable X be described by a family of discrete distributions $\{P_\theta\}_{\theta\in\Theta}$ with $\Theta \in R^d$
    - The support space of X is $\{0,1,2,\dots K\}$
    - The probability mass function is $f_{\hat{\theta}}$ (this is the MLE estimator computed based on the data given)
  - The null and alternative hypothesis are of the form:
    - H_0: $Actual\ distribution \in \{P_\theta\}_{\theta\in\Theta}$ and H_1: $Actual\ distribution \notin \{P_\theta\}_{\theta\in\Theta}$
  - Then, if H_0 is true:
    - $n\sum_{j=0}^{K}\frac{(\frac{N_j}{n}-f_{\hat{\theta}}(j))^2}{f_{\hat{\theta}}(j)} \underset{n\to\infty}{\overset{(d)}{\to}} \chi^2_{(K+1)-(d+1)}$
    - where N_j is the number of observations where the random variable takes the value j
    - $(K+1)$ is the total number of elements in the support space of P

- **Goodness of fit testing for continuous distributions:** If we divide the support space of a continuous random variable into B bins, and associate with each bin the probability that the continuous random variable lies in that bin, then we get a discrete random variable which follows a multinomial distribution with B modalities. This approximate discrete random variable can be used for goodness of fit testing. However, this process may not give very accurate results – it also requires judgement to select the number of bins and the size of each bin
  - A common approach in goodness of fit tests for continuous distributions is to use cumulative distribution functions. A cumulative distribution function can be written as an expectation using indicator variables:
    - For any continuous random variable X, we can write:
      - $F(t) = P(X \le t) \quad \forall\, t \in R$
      - Or, $F(t) = E[I(X \le t)] \quad \forall\, t \in R$
      - The cumulative distribution function completely characterizes the distribution of X
    - If we have a set of independent and identically distributed random variables $X_1, X_2, \dots X_n$, we can estimate the cumulative distribution function. This is known as the empirical cumulative distribution function.
      - $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t) = \frac{1}{n} Count\{i = 1,2,..n;\ X_i \le t\}\ \forall\, t \in R$
    - When we consider convergence of a function to another function, we can consider point-wise convergence or uniform convergence

- A sequence of functions $g_n(x)$ converges point-wise to a function $g(x)$ if $\lim_{n\to\infty} g_n(x) = g(x)\ \forall\ x \in R$ (Here, we are looking at the values of the two functions only in the case when $n \to \infty$)
- A sequence of functions $g_n(x)$ converges uniformly to a function $g(x)$ if $\lim_{n\to\infty} \sup_{x\in R} |g_n(x) - g(x)| = 0$ (Here, we first look at the value of x for which the supremum function gives the largest value and then evaluate it in the limit of $n \to \infty$)
- Consider two functions related by the equation $g_n(x) = g(x) + \frac{x}{n}$. There is point-wise convergence, but not uniform convergence.
    - This is because for any finite value of x, $\lim_{n\to\infty} g_n(x) = g(x)$.
    - However, $\sup_{x\in R} |g_n(x) - g(x)|$ is achieved when $x \to \infty$, and $\lim_{n\to\infty} \sup_{x\in R} |g_n(x) - g(x)| = 1$
- The Glivenko-Cantelli theorem (also called the fundamental theorem of statistics) says that for CDF functions $F_n(t)$ and $F(t)$, point-wise convergence implies uniform convergence:
    - $F_n(t) \underset{n\to\infty}{\overset{a.s}{\to}} F(t)\ \forall\ t \in R \implies \sup_{t\in R} |F_n(t) - F(t)| = 0$

- Since $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\ (X_i \le t)$ is the average of the random variable $I(X_i \le t)$, using the central limit theorem, it will asymptotically converge in distribution to a normal distribution with asymptotic variance equal to the variance of $I(X_i \le t)$.
    - $I(X_i \le t)$ is a Bernoulli random variable with $E[I(X_i \le t)] = F(t)$ and $Var[I(X_i \le t)] = F(t)(1 - F(t))$
    - Hence, $\sqrt{n}\big(F_n(t) - F(t)\big) \underset{n\to\infty}{\overset{(d)}{\to}} N\Big(0, F(t)\big(1 - F(t)\big)\Big)\ \forall\ t \in R$
    - In this case, the variance is not constant. It depends on t. Specifically, for $t \to -\infty$ or $t \to +\infty$, the variance is 0. However, for values of t between $-\infty$ and $+\infty$, it will fluctuate based on the sample we have, and this fluctuation is of a random nature. This fluctuation can be considered similar to Brownian motion.
    - In the case of empirical CDFs, we make use of a theorem called the Donsker theorem. Donsker's theorem states that:
        - $\sqrt{n} \sup_{t\in R} |F_n(t) - F(t)| \underset{n\to\infty}{\overset{(d)}{\to}} \sup_{0\le x\le 1} |B(x)|$ where B is a random curve called the Brownian bridge
        - $\sup_{0\le x\le 1} |B(x)|$ is a pivotal distribution, i.e. it does not depend on the unknown distribution of the data. Its distribution is known and is independent of the underlying distribution that generated the data.
        - The Donsker theorem gives a distribution for the maximum difference between two CDFs

- If the empirical CDF is $F_n(t)$ and the distribution under the null hypothesis is $F^0(t)$, then the **Kolmogorov-Smirnov test** has as the test statistic $T_n = \sqrt{n}\ \sup_{t\in R}|F_n(t) - F^0(t)|$
  - The rejection criteria is $T_n > q_\alpha$
  - To compute $\sup_{t\in R}|F_n(t) - F^0(t)|$, we consider the fact that the empirical CDF function is a step function, while the CDF of the null hypothesis is a continuous function.
    - We organize the sample observations in increasing order and create an ordered set $(X_{(1)}, X_{(2)}, \ldots X_{(n)})$ such that $X_{(1)} \leq X_{(2)} \leq .. X_{(n)}$
      - Clearly $F_n(X_{(i)}) = \frac{i}{n}\ \forall\ i\ \in [1,n]$, and $F_n(X_{(1)}^-) = 0$ where $X_{(1)}^-$ is any value less than $X_{(1)}$. Also, $F_n(X_{(n)}^+) = 1$ where $X_{(n)}^+$ is any value greater than $X_{(n)}$
    - At every value of $X_{(i)}\ \forall\ i\ \in [1,n]$, the empirical CDF function has a step jump and we need to evaluate the difference between $F^0(X_{(i)})$ and the base of the jump, and the difference between $F^0(X_{(i)})$ and the top of the jump. The greater of these two values is the difference between the empirical CDF and the CDF of the null hypothesis at $X_{(i)}$. This computation has to be done for every $X_{(i)}\ \forall\ i\ \in [1,n]$.
      - Difference between $F^0(X_{(i)})$ and the base of the jump = $|F^0(X_{(i)}) - F_n(X_{(i-1)})|$ = $|F^0(X_{(i)}) - \frac{(i-1)}{n}|$
      - Difference between $F^0(X_{(i)})$ and the top of the jump = $|F^0(X_{(i)}) - F_n(X_{(i)})|$ = $|F^0(X_{(i)}) - \frac{i}{n}|$
    - $\sup_{t\in R}|F_n(t) - F^0(t)| = \max_{i=1,2,..n}\{\max(\left|F^0(X_{(i)}) - \frac{(i-1)}{n}\right|, |F^0(X_{(i)}) - \frac{i}{n}|)$
  - In evaluating the function $\sup_{t\in R}|F_n(t) - F^0(t)|$, we consider the fact that if F_X is the CDF of a random variable X, then the CDF function mapping when applied to the random variable X itself gives an expression involving the random variable X. This expression is the uniform random variable with the support $[0,1]$.
    - In the expression $T_n = \sqrt{n}\ \sup_{t\in R}|F_n(t) - F^0(t)|$, if we set $\bar{t} = F^0(t)$, then $t = F^{0^{-1}}(\bar{t})$
    - $T_n = \sqrt{n}\ \sup_{t\in R}|F_n(t) - F^0(t)| = \sqrt{n}\ \sup_{t\in R}|(\frac{1}{n}\sum_{i=1}^n I\,(X_i \leq t)) - F^0(t)|$
    - Or, $T_n = \sqrt{n}\ \sup_{t\in R}|\left(\frac{1}{n}\sum_{i=1}^n I\,(F^0(X_i) \leq F^0(t))\right) - F^0(t)|$
      - This is because under the CDF function is an increasing function, and $X_i \leq t \Longrightarrow F(X_i) \leq F(t)$ and under the null hypothesis, this is equivalent to $F^0(X_i) \leq F^0(t)$
    - Or, $T_n = \sqrt{n}\ \sup_{\bar{t}\in[0,1]}|\left(\frac{1}{n}\sum_{i=1}^n I\,(Y_i \leq \bar{t})\right) - \bar{t}|$ where $Y_i \sim Unif(0,1)$
    - T_n is a pivotal statistic under the null hypothesis for any value of n. Its distribution does not depend on the underlying data

- The same conclusion can also be arrived at by considering the expression $\sup_{t \in R} |F_n(t) - F^0(t)| = \max_{i=1,2,\ldots n} \{\max(\left|F^0\left(X_{(i)}\right) - \frac{(i-1)}{n}\right|, |F^0\left(X_{(i)}\right) - \frac{i}{n}|)$
    - $F^0\left(X_{(i)}\right)$ is a draw from a uniform distribution, and does not depend on knowledge of any distribution

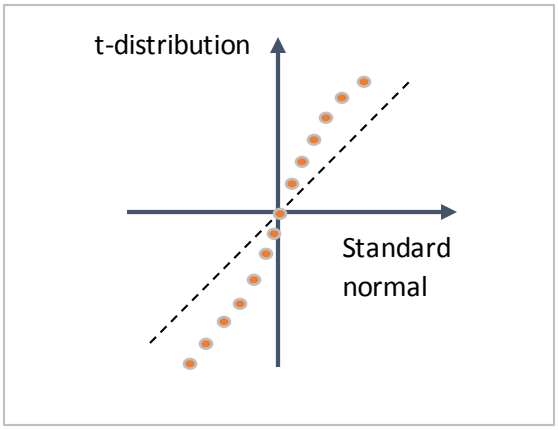
- q_α depends on n, and is obtained through simulations with a very large number of runs. The quantiles are available from tables.
- The Kolmogorov-Smirnov test can also be used to compare between two samples.
    - If we have two samples $(X_1, X_2, \ldots X_n)$ and $(Y_1, Y_2, \ldots Y_m)$ where each of the samples comprises independent and identically distributed random variables, then we can undertake a goodness of fit test to see whether the two samples come from the same distribution. While we do not know the underlying distribution of the two samples, we assume that the underlying distributions are continuous
    - Let $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \leq t)\ \forall\ i \in [1, n]$ be the empirical CDF of the first distribution
    - Let $G_m(t) = \frac{1}{m}\sum_{j=1}^{m} I\left(Y_j \leq t\right) \forall\ j \in [1, m]$ be the empirical CDF of the second distribution
    - The Kolmogorov-Smirnov test statistic here is:
        - $T_n = \sqrt{\frac{nm}{(n+m)}} \sup_{t \in R} |F_n(t) - G_m(t)|$

- The **Kolmogorov-Lilliefors test** is used when we want to test whether a sample has been drawn from a Gaussian distribution, but we do not know the mean and variance of the Gaussian distribution. The test is to find whether the sample fits any Gaussian distribution.
    - When using t-test, one of the assumptions is that the sample is drawn from a Gaussian distribution. The Kolmogorov-Lilliefors test is used to test that assumption.
    - We cannot use the Kolmogorov Smirnov test if we do not know the mean and variance of the normal distribution – in this case, the null hypothesis is not precisely defined. Computing the mean and variance of the normal distribution from the sample is a violation of the hypothesis testing principle whereby the hypothesis must involve a deterministic quantity or distribution (this is what is proposed and must be tested). Furthermore, such a computation would make the resulting normal distribution fit to the data and rejection of the null hypothesis would be very unlikely.
    - The Kolmogorov-Lilliefors test uses the sample mean and the sample variance for estimating the mean and variance of the hypothetical normal distribution, but adjusts for the fact that the hypothesis depends on the data by having more aggressive rejection criteria. The test statistic for the Kolmogorov-Lilliefors test is:

- $T_n = \sqrt{n} \sup_{t \in R} |F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)|$ where $\Phi_{\hat{\mu},\hat{\sigma}^2}$ is the CDF of a normal distribution where the mean is the sample mean $\hat{\mu}$ and the variance is the sample variance $\hat{\sigma}^2$
- In this case also, T_n is a pivotal statistic, and it does not depend on the true distribution. In fact, in this case, the true distribution is not even known.

- Both the Kolmogorov-Smirnov and the Kolmogorov-Lilliefors tests are non-asymptotic in the sense that, for any fixed n, the distribution of the test statistic under the null can be consulted via tables. Hence, it is possible to specify the non-asymptotic level of the test and not just the asymptotic level.

- **Quantile-Quantile (QQ) plots**: This is a visual approach to testing the goodness of fit of a sample to a hypothesized distribution. This approach is utilized because looking at the CDFs of the empirical and the hypothesized distribution may not give a clear indication whether the distributions are similar or different
  - Given a set of independent and identically distributed random variables $X_1, X_2, \ldots X_n$, we can reorganize this set into an ordered set $(X_{(1)}, X_{(2)}, \ldots X_{(n)})$ such that $X_{(1)} \leq X_{(2)} \leq .. X_{(n)}$
  - For each value in the set $(X_{(1)}, X_{(2)}, \ldots X_{(n-1)})$, we can compute an x-y pair comprising of quantiles $(F_n^{-1}(i/n), F^{0^{-1}}(i/n)) = (X_{(1)}, F^{0^{-1}}(i/n))$.
    - We do not use $X_{(n)}$ because $F^{0^{-1}}\left(\frac{n}{n}\right) \to \infty$
  - Putting these pairs of points on a graph where the y-axis is $F_n^{-1}(t)$ and the x-axis is $F^{0^{-1}}(t)$ give a scatter plot, and the degree of scatter from the x=y line is a measure of how close the empirical and hypothesized distributions are. This is a plot of the actual versus the theoretical distribution.
  - Since we are plotting inverses of CDFs, or quantiles, this plot is called a quantile-quantile plot
  - When the theoretical distribution is the standard normal distribution, and the actual distribution is the t-distribution:
    - Both distributions share the same support and both are symmetrical
    - The t-distribution has fatter tails on either side of the point of symmetry
    - The 0.5 quantile is the same for both the distributions and is equal to 0
    - For quantiles <0.5, the t-distribution on account of the fatter tails will have a higher absolute value of the quantile than the standard normal. However, the quantiles being negative, this means that for quantiles less than 0.5, quantile of the t-distribution < quantile of the standard normal
    - For quantiles > 0.5, the t-distribution on account of the fatter tails will have a higher value of the quantile than the standard normal.
    - The Q-Q plot will be as below (this is the Q-Q plot shape for heavy tailed distributions:



- Three other patterns are possible for the Q-Q plot when the theoretical distribution is the standard normal distribution
    - Right-skewed: The Q-Q plot cuts the x=y line from the left, crosses it, and then retraces its path back, cutting the x-y line from the right and crossing it. Other variations are possible, the key feature being that most of the Q-Q plot will be to the left of the x-y line. The exponential distribution will exhibit a right-skewed curve. For the exponential distribution, all quantiles are positive, and the whole of the Q-Q plot will lie above the line y=0 or in other words, lie above the x-axis,
    - Left-skewed: The Q-Q plot cuts the x=y line from the right, crosses it, and then retraces its path back, cutting the x-y from the left and crossing it. Other variations are possible, the key feature being that most of the Q-Q plot will be to the right of the x-y line. The random variable $-X$ where X is an exponential distribution will exhibit a left-skewed curve – in this case, all the quantiles of $-X$ are negative, and the whole of the Q-Q plot will lie to the left of the line x=0 (y-axis).
    - Light tails: This will be the mirror image (reflection) of the heavy tailed distribution on the line x=y. The uniform distribution will exhibit a light-tailed curve.
- A distribution P is said to have a heavier right tail than a distribution Q if:
    - $P(X \geq t) \geq P(Y \geq t)$ for $t > 0$ with t being sufficiently large
- A distribution P is said to have a heavier left tail than a distribution Q if:
    - $P(X \leq -t) \geq P(Y \leq -t)$ for $t > 0$ with t being sufficiently large

Bayesian Statistics

- In Bayesian statistics, the true parameter is accorded a probability distribution, even though the true parameter is actually a constant. The basic premise of the Bayesian approach is that any quantity which is unknown can be given a probability distribution, which represents our belief about the range of values that the parameter can take.
    - Based on observed data, we can update our beliefs and arrive at a posterior distribution of the parameter. The posterior distribution takes into account the fact that our prior belief must be modified by the actual observed data
    - The choice of the prior distribution is critical, as the support of the posterior distribution will be a subset of the support of the prior distribution
- In settings where the observed variable follows a Binomial distribution or a Bernoulli distribution with a certain parameter, the parameter's prior is taken as a beta distribution. The beta distribution has a support [0,1] and has two parameters, α and β.
    - The PDF of the beta distribution is $f_X(x) = Cx^{\alpha-1}(1-x)^{\beta-1}\ for\ x \in (0,1)\ and\ \alpha, \beta > 0$ where C is the normalizing constant
        - Since $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$, $C = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$
        - The expected value of the beta distribution is $\frac{\alpha}{\alpha+\beta}$
    - The mode of the beta distribution can take different values depending on the values of α and β.
        - Setting $\frac{d(x^{\alpha-1}(1-x)^{\beta-1})}{dx} = 0$ gives $x = \frac{\alpha-1}{\alpha+\beta-2}$
            - However, this is the mode only when $\alpha > 1$ and $\beta > 1$
        - When $\alpha = \beta$, we have $f_X(x) = 1\ \forall\ x$ (any value in (0,1) can be the mode, the mode is not uniquely defined)
        - For $\alpha > 1$ and $\beta \leq 1$, the equation $x = \frac{\alpha-1}{\alpha+\beta-2}$ cannot be used as it may give negative results and the support of the distribution is strictly non-negative
            - For $\alpha > 1$ and $\beta \leq 1$, $x^{\alpha-1}$ increases with increase in x, and $(1-x)^{\beta-1}$ also increases with increase in x. Hence the mode is $x = 1$
        - For $\alpha \leq 1$ and $\beta > 1$, again the equation $x = \frac{\alpha-1}{\alpha+\beta-2}$ cannot be used as may give negative results and the support of the distribution is strictly non-negative
            - For $\alpha \leq 1$ and $\beta > 1$, $x^{\alpha-1}$ decreases with increase in x, and $(1-x)^{\beta-1}$ also decreases with increase in x. Hence the mode is $x = 0$
        - If both $\alpha < 1$ and $\beta < 1$, the distribution is bimodal with both $x = 0$ and $x = 1$ being modes of the distribution
- The Gamma distribution and the inverse Gamma distribution are two other distributions that are commonly used in Bayesian statistics
    - The Gamma distribution has a pdf of the form $Constant * x^{\alpha-1}e^{-\beta x}$ and its support is $(0,\infty)$
        - α and β are parameters of the distribution with $\alpha > 0$ and $\beta > 0$
        - The mean of the distribution is α/β and the variance is α/β^2
        - There is not simple closed form for the median of the distribution

        - The mode of the distribution is $(\alpha - 1)/\beta$ if $\alpha \geq 1$
    - The inverse Gamma distribution has a pdf of the form $Constant * x^{-\alpha-1} e^{-\frac{\beta}{x}}$ and its support is $(0, \infty)$
        - α and β are parameters of the distribution with $\alpha > 0$ and $\beta > 0$
        - The mean of the distribution is $\beta/(\alpha - 1)$ if $\alpha > 1$
        - The variance of the distribution is $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$
        - The mode of the distribution is $\beta/(\alpha + 1)$
- In the case of estimation using Bayesian statistics, we have two basic formulas:
    - $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$ (when the observation follows a continuous distribution)
        - $f_{X|\Theta}(x|\theta)$ can be written as $L(X_1, X_2, \dots X_n \,|\theta)$ – with multiple observations, X is a vector
            - $L(X_1, X_2, \dots X_n \,|\theta) = L(X_1|\theta) * L(X_2|\theta) * \dots L(X_n|\theta)$ if the observations are independent and identically distributed
        - $f_X(x) = \int_\theta f_\Theta(\theta) f_{X|\Theta}(x|\theta) d\theta = \int_\theta f_\Theta(\theta) L(X_1, X_2, \dots X_n \,|\theta) d\theta$ (this will not depend on θ)
    - $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) p_{X|\Theta}(x|\theta)}{p_X(x)}$ (when the observation follows a discrete distribution)
        - Again, $p_{X|\Theta}(x|\theta)$ can be written as $L(X_1, X_2, \dots X_n \,|\theta)$ – with multiple observations, X is a vector
            - $L(X_1, X_2, \dots X_n \,|\theta) = L(X_1|\theta) * L(X_2|\theta) * \dots L(X_n|\theta)$ if the observations are independent and identically distributed
        - $p_X(x) = \int_\theta f_\Theta(\theta) p_{X|\Theta}(x|\theta) d\theta = \int_\theta f_\Theta(\theta) L(X_1, X_2, \dots X_n \,|\theta) d\theta$ (this will not depend on θ)
- **Improper Priors**: In many cases, we have no prior information on the distribution of the parameter, and we want to pick a distribution that is uniform over the support of the parameter
    - For instance, if the observations follow a Bernoulli distribution, then we can choose the uniform distribution $Unif([0,1])$ as the prior distribution of p, the probability of success.
        - The $Unif([0,1])$ distribution is a special case of the beta distribution where both α and β are set to 1, i.e. $Unif([0,1]) = Beta(1,1)$
    - However, if the observations are Gaussian with distribution $N(\theta, 1)$ and we want to pick a uniform prior for θ, we have to select a uniform distribution whose support is $(-\infty, +\infty)$. Such a distribution would have a density 0 at all points.
        - In such a case, we choose an improper prior. We assume $f_\Theta(\theta) = 1 \; \forall \; \theta \in R$. It is an improper prior because it does not represent a valid probability density function. However, it is a measurable, non-negative function.
        - In general, we can use an improper prior and still get a valid posterior distribution
    - If we have independent and identically distributed observations $X_1, X_2, \dots X_n$ which follow a Gaussian distribution $N(\theta, 1)$ and the parameter θ is drawn from an improper prior, i.e. $f_\Theta(\theta) = 1 \; \forall \; \theta \in R$, then:
        - $f_{\Theta|X}(\theta|x) = Constant * \; e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2}$

- Let us consider the expression $-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2$
- $-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2 = -\frac{1}{2}(\sum_{i=1}^{n}(X_i)^2 + n\theta^2 - 2\theta\sum_{i=1}^{n}X_i)$
- Or, $-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2 = -\frac{n}{2}(\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 + \theta^2 - 2\theta\bar{X}_n)$
- Or, $-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2 = -\frac{n}{2}(\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 + \theta^2 - 2\theta\bar{X}_n + \bar{X}_n^2 - \bar{X}_n^2)$
- Or, $-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2 = -\frac{n}{2}(\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 + (\theta-\bar{X}_n)^2 - \bar{X}_n^2)$

- Therefore, $f_{\Theta|X}(\theta|x) = Constant * e^{-\frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - \bar{X}_n^2\right)} * e^{-\frac{n}{2}(\theta-\bar{X}_n)^2}$
- Since $e^{-\frac{1}{2n}(\sum_{i=1}^{n}(X_i)^2 - X_n^2)}$ does not depend on θ and serves as a normalizing constant, we can write:
  - $f_{\Theta|X}(\theta|x) = Constant * e^{-\frac{n}{2}(\theta-\bar{X}_n)^2}$
  - The posterior distribution of θ is a normal distribution with mean $\bar{X}_n$ and variance $\frac{1}{n}$

- **Jeffrey's Prior**: Jeffrey's prior attempts to incorporate frequentist ideas of likelihood in Bayesian statistics. It is a non-informative prior that also looks at the statistical model used for the observations, specifically the likelihood function.
  - Jeffrey's prior for a parameter θ is:
    - $\pi_J(\theta) \propto \sqrt{\det(I(\theta))}$ where $I(\theta)$ is the Fisher information matrix. The prior can only be defined when the Fisher information matrix exists.
      - $\pi_J(\theta)$ represents the prior distribution of θ
      - If the parameter θ is one-dimensional, then $\pi_J(\theta) \propto \sqrt{I(\theta)}$
    - The variance of the MLE estimator is $I(\theta)^{-1}$, or in the one-dimensional case, $1/I(\theta)$
      - Thus, Jeffrey's prior gives more weight to values of θ where the MLE estimate has lesser variance, and the observed data provides more information in deciding the parameter (the observed data can allow us to estimate the parameter more precisely)
    - For another interpretation of Jeffrey's prior, we observe that $I(\theta) = Var(\frac{\partial}{\partial\theta}(L(X,\theta))$
      - Clearly the values of $I(\theta)$ will be higher where the variance of the likelihood estimator is higher to changes in θ.
      - Jeffrey's prior gives more weight to parameter values where a small change in the parameter value influences the likelihood function significantly
  - In the case of a Bernoulli distribution:
    - $\pi_J(\theta) \propto \frac{1}{\sqrt{p(1-p)}}$, $p \in (0,1)$
    - Here, $\pi_J(\theta) \propto Beta(\frac{1}{2},\frac{1}{2})$
    - In this case, Jeffrey's prior is not an improper prior
  - In the case of a normal distribution, if we assume the variance is known, then the prior specifies the distribution of the mean.

- Assuming the variance as 1, we have $L(X,\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X-\theta)^2}$
  - $\log L(X,\theta) = \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(X-\theta)^2$
  - $\frac{\partial \log L(X,\theta)}{\partial \theta} = (X-\theta)$
  - $I(\theta) = Var\left(\frac{\partial \log L(X,\theta)}{\partial \theta}\right) = Var(X-\theta) = Var(X) = 1$
- Thus, $\pi_J(\theta) \propto \sqrt{I(\theta)} \propto 1$
- This is an improper prior. Qualitatively, it signifies that no value of the mean θ is more likely than any other value

o For the Poisson distribution, the prior distribution specifies the distribution of λ. In this case,
- $L(X,\theta) = \frac{\lambda^X e^{-\lambda}}{X!}$
- $\log L(X,\theta) = X \log \lambda - \lambda + \log \frac{1}{X!}$
- $\frac{\partial \log L(X,\theta)}{\partial \theta} = \frac{X}{\lambda} - 1$
- $I(\theta) = Var\left(\frac{\partial \log L(X,\theta)}{\partial \theta}\right) = Var\left(\frac{X}{\lambda} - 1\right) = Var\left(\frac{X}{\lambda}\right) = \frac{1}{\lambda^2} Var(X) = 1/\lambda$
- Thus, $\pi_J(\theta) \propto \sqrt{I(\theta)} \propto \sqrt{1/\lambda}$
- This is an improper prior as $\lambda \in (0,\infty)$ and $\int_0^\infty \sqrt{1/\lambda}\, d\lambda = \infty$

o For the normal distribution where both the mean and variance are unknown, the prior distribution of θ is a joint distribution of the mean μ and variance σ^2.
- $L(X,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X-\mu)^2}$
- $\log L(X,\mu,\sigma^2) = \log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sqrt{\sigma^2}} - \frac{1}{2\sigma^2}(X-\mu)^2$
- The first order partial derivatives with respect to μ and σ^2 are:
  - $\frac{\partial \log L(X,\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}(X-\mu)$
  - $\frac{\partial \log L(X,\mu,\sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(X-\mu)^2$
- The second order derivative is a 2x2 matrix
  - $\begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{1}{(\sigma^2)^2}(X-\mu) \\ -\frac{1}{(\sigma^2)^2}(X-\mu) & \frac{1}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}(X-\mu)^2 \end{pmatrix}$
  - The expectation of this matrix is $\begin{pmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2(\sigma^2)^2} \end{pmatrix}$
  - Hence, $I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2(\sigma^2)^2} \end{pmatrix}$
  - $\pi_J(\theta) \propto \sqrt{\det(I(\theta))} \propto \sqrt{\frac{1}{2(\sigma^2)^3}} \propto \frac{1}{\sigma^3}$
- In this case also, Jeffrey's prior is an improper prior, as $\sigma^3 \in R$ and the integral of $\frac{1}{\sigma^3}$ between $-\infty$ and ∞ is not equal to 1

o Jeffrey's prior has the property of having reparametrization invariance.

- If $\pi_J(\theta) \propto \sqrt{\det(I(\theta))}$ and we have $\eta = \phi(\theta)$, then:
  - $\pi_J(\theta) \propto \sqrt{\det(I(\theta))} \implies \bar{\pi}_J(\eta) \propto \sqrt{\det(I(\eta))}$ (the functional form of Jeffrey's prior does not change)
  - When we say that the functional form of Jeffrey's prior does not change, it implies that the prior can be derived in an identical manner from the likelihood function. The specific mathematical expression of the prior will change and is not invariant, but its dependence on the likelihood function is invariant.
- Also, $\pi_J(\theta) = \bar{\pi}_J(\eta)|\frac{d\eta}{d\theta}|$
  - The above expression can be easily derived if we consider $\phi(\theta)$ to be a monotonically increasing or decreasing function of θ, and assume that both $\pi_J(\theta)$ and $\bar{\pi}_J(\eta)$ are proper probability density functions
  - If we know $\pi_J(\theta)$, we can compute $\bar{\pi}_J(\eta)$
  - $\bar{\pi}_J(\eta) = \pi_J(\theta)|\frac{d\theta}{d\eta}|$
  - $\theta = \phi^{-1}(\eta)$ and $\frac{d\theta}{d\eta} = \frac{1}{\phi'(\theta)}$
  - Therefore, $\bar{\pi}_J(\eta) = \pi_J(\phi^{-1}(\eta))|\frac{1}{\phi'(\phi^{-1}(\eta))}|$
  - As an example, consider a prior $\pi_J(q)$ when the observations follow a Bernoulli distribution with parameter q
    - $\pi_J(q) \propto \frac{1}{\sqrt{q(1-q)}}$
    - We change the parameter to p such that $p = q^{1/10}$. Here, ϕ is a function that maps $x \rightarrow \ , x \in (0,1)$. ϕ^{-1} is a function that maps $x^{1/10} \rightarrow x \ , x \in (0,1)$
    - $\phi^{-1}(p) = p^{10}$
    - $\pi_J\left(\phi^{-1}(p)\right) = \frac{1}{\sqrt{p^{10}(1-p^{10})}}$
    - $\frac{dp}{dq} = \frac{1}{10} q^{-\frac{9}{10}} = \frac{1}{10}\, p^{-9}$
    - $\phi'\left(\phi^{-1}(p)\right) = \frac{dp}{dq} = \frac{1}{10}\, p^{-9}$
    - $\left|\frac{1}{\phi'(\phi^{-1}(p))}\right| = 10p^9$
    - Thus, $\bar{\pi}_J(p) \propto \frac{1}{\sqrt{p^{1/10}(1-p^{1/10})}} 10p^{9/10} \propto \frac{1}{\sqrt{p^{10}(1-p^{10})}} p^9$
- We know that $I(\theta) = Var(\frac{\partial}{\partial\theta}(\log L(X,\theta))$ and it measures how sensitive the likelihood is to marginal movements of θ. High values of $I(\theta)$ also imply more certainty about our MLE estimate. The main motivation for using Jeffrey's prior is because certain parametrizations may compress meaningful differences into a small interval, whilst yielding large room for less impactful differences. In this case, a naive approach of using the uniform distribution would give an undue large weight to areas where modifying the parameter will not change the outcome much.

- **Bayesian confidence regions**: A Bayesian confidence region (also called credible interval) is an interval in which an unobserved parameter falls with a particular subjective probability.
    - For $\alpha \in (0,1)$, the Bayesian confidence region with level α is a random subset R of the parameter space Θ, which depends on the sample $X_1, X_2, \dots X_n$ such that:
        - $P(\theta \in R | X_1, X_2, \dots X_n) = 1 - \alpha$
        - Once we have the observations, the randomness in the posterior distribution comes from the randomness in the prior.
- **Bayesian estimation**: When using Bayesian framework for analysis, we generally look at either the maximum a posteriori estimate or the mean of the posterior distribution as an estimator for the unknown parameter
    - $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} f_{\Theta|X}(\theta|x) = \underset{\theta}{\operatorname{argmax}} f_{\Theta}(\theta) L(X_1, X_2, \dots X_n | \theta)$
        - Here the likelihood function is weighted by the prior. If we have a lot of observations, the importance of the prior diminishes. If the prior is uninformative (e.g. a uniform prior), then the MAP and MLE estimator will coincide.
    - $\hat{\theta}_{LMS} = \int_{\theta} \theta f_{\Theta|X}(\theta|x) d\theta$
        - The Least Mean Squares (LMS) estimator is the preferred estimator because the mean squared error (MSE) is the least with the LMS estimator.
- **Dirichlet distribution**: The Dirichlet distribution, often denoted by $Dir(\alpha)$, is a family of continuous multivariate probability distributions parameterized by a vector α of positive real numbers.
    - The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \alpha_2, \dots \alpha_K > 0$ has a probability density function given by:
        - $f_{X_1, X_2, \dots X_K}(x_1, x_2, \dots . x_k; \alpha_1, \alpha_2, \dots \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} {x_i}^{(\alpha_i - 1)}$
            - Where $\prod_{i=1}^{K} x_i = 1$ and $x_i \geq 0 \; \forall \; i \in [1, K]$
            - $\frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}$ is a normalizing constant
        - Since the Dirichlet distribution is a multivariate distribution, the mean of the Dirichlet distribution is a vector of means in $R^K = [\frac{\alpha_1}{\sum_{i=1}^{K} \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{K} \alpha_i}, \dots \frac{\alpha_K}{\sum_{i=1}^{K} \alpha_i}]$
    - Consider the case when we have sample observations $X_1, X_2, \dots X_n$ which follow a multinomial distribution with the probability simplex $\Delta_K = \{ \boldsymbol{p} = (p_1, p_2, \dots p_K) \in (0,1)^K ; \sum_{j=1}^{K} p_j = 1 \}$
        - This means that the random variable X takes on values in a finite set $E = \{a_1, a_2, \dots . a_K\}$ with $P(X = a_i) = p_i \; \forall \; i \in [1, K]$
        - Also, $f(X_1, X_2, .. X_n; \boldsymbol{p}) \propto \prod_{i=1}^{K} (p_i)^{N_i}$ with $\sum_{i=1}^{K} N_i = n$ and N_i is the number of times a_i was observed in the sample
        - If we want a prior distribution on p, the Dirichlet distribution can serve as a prior distribution since in the Dirichlet distribution, the variables are in the nature of probability measures since the variable values must add up to 1
    - In a Bayesian setup, we can write:
        - The prior for $\boldsymbol{p}$ as $f(\boldsymbol{p}) = f(p_1, p_2, \dots . p_k; \alpha_1, \alpha_2, \dots \alpha_K) \propto \prod_{i=1}^{K} {p_i}^{(\alpha_i - 1)}$
        - The likelihood function as $f(X_1, X_2, .. X_n | \boldsymbol{p}) \propto \prod_{j=1}^{K} (p_j)^{N_j}$
        - This gives us a posterior distribution for $\boldsymbol{p}$ as:

- $f(\boldsymbol{p}|X_1,X_2,..X_n) \propto f(\boldsymbol{p})f(X_1,X_2,..X_n|\boldsymbol{p})$
- Or, $f(\boldsymbol{p}|X_1,X_2,..X_n) \propto \prod_{i=1}^{K} p_i{}^{(\alpha_i-1)} * \prod_{j=1}^{K}(p_j)^{N_j}$

▪ Or, $f(\boldsymbol{p}|X_1,X_2,..X_n) \propto \prod_{i=1}^{K} p_i{}^{(N_i+\alpha_i-1)}$ which is again a Dirichlet distribution with parameters $N_1+\alpha_1, N_2+\alpha_2, \ldots. N_K+\alpha_K$

Linear Regression

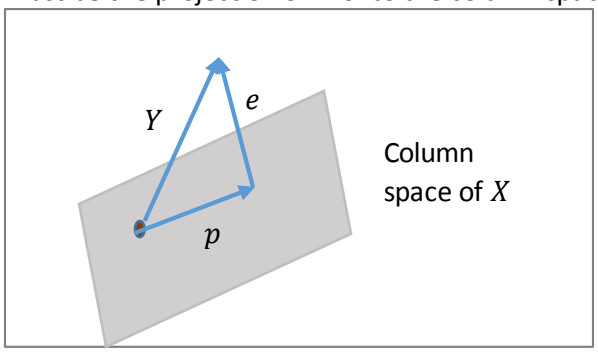
- In linear regression, we do not know the exact relationship between the response variable (Y) and the explanatory variables $X^{(1)}, X^{(2)}, .. X^{(n)}$. Ideally, we would want to find the joint pdf of these variables but this may not be possible.
    - We settle for the conditional expectation of Y. The regression function is:
        - $E[Y|X = x]$ where X could be random variable or a random vector
    - Theoretically, $E[Y|X = x] = \int_y y f_{Y|X}(y|x)\, dy$
        - Instead of using the mean as the summary statistic, it is also possible to use other conditional summary statistics like median or quantiles. For instance, if we use the conditional median, for a given value of $X = x$, we would compute Y as m where:
            - $\int_{-\infty}^{m} f_{Y|X}(y|x)dy = 1/2$
- In practice, we are not able to compute $f_{Y|X}(y|x)$ from the data. We make the assumption that $E[Y|X = x] = f(x)$ and have to choose the function f which is not a probability density function.
    - In linear regression, we choose the function f as a linear function in X and introduce an error term. In this case, we are making the strong assumption that the conditional expectation of Y given X is a linear function of X.
        - As an example, if $Y = 3X + 5 + \varepsilon$ where $\varepsilon \sim N(0,1)$ and X is independent of ε, then
            - $E[Y|X = x] = E[3X + 5 + \varepsilon|X = x]$ = 3x+5
        - Independence of the explanatory variables from the error term, i.e. $Cov(X, \varepsilon) = 0$ is an important assumption in linear regression.
- Given the equation $Y = a + bX + \varepsilon$, and a set of data points $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \ldots (X_{(n)}, Y_{(n)})$, we want to minimize the sum of squared errors. In other words, we want to find the **least squares estimator (LSE)** for a and b.
    - The objective is to minimize $SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_{(i)} - a - bX_{(i)})^2$
    - Taking the first order partial derivatives with respect to the parameters a and b, we have:
        - $\frac{\partial SSE}{\partial a} = -2\sum_{i=1}^{n}(Y_{(i)} - a - bX_{(i)})$
            - Setting $\frac{\partial SSE}{\partial a} = 0$ gives $-2\sum_{i=1}^{n}(Y_{(i)} - a - bX_{(i)}) = 0$, or dividing throughout by $2n$,
                - $\sum_{i=1}^{n}\left(\frac{Y_{(i)}}{n} - \frac{a}{n} - \frac{bX_{(i)}}{n}\right) = 0$, or
                - $\bar{Y} = a + b\bar{X}$ - **(A)** where $\bar{X}$ and $\bar{Y}$ are **empirical means**
        - $\frac{\partial SSE}{\partial b} = -2\sum_{i=1}^{n}(Y_{(i)} - a - bX_{(i)})X_{(i)}$
            - Setting $\frac{\partial SSE}{\partial b} = 0$ gives $-2\sum_{i=1}^{n}(Y_{(i)} - a - bX_{(i)})X_{(i)} = 0$, or
                - $\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - bX_{(i)}^2 - aX_{(i)}) = 0$
                - $\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - bX_{(i)}^2 - (\bar{Y} - b\bar{X})X_{(i)}) = 0$ – using (A)
                - $\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - bX_{(i)}^2 - X_{(i)}\bar{Y} + bX_{(i)}\bar{X}) = 0$, or
                - $b = \frac{\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - X_{(i)}\bar{Y})}{\sum_{i=1}^{n}(X_{(i)}^2 - X_{(i)}\bar{X})} = \frac{1/n\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - X_{(i)}\bar{Y})}{1/n\sum_{i=1}^{n}(X_{(i)}^2 - X_{(i)}\bar{X})}$
            - Considering the expression $b = \frac{1/n\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - X_{(i)}\bar{Y})}{1/n\sum_{i=1}^{n}(X_{(i)}^2 - X_{(i)}\bar{X})}$, we observe that:

- $\frac{1}{n}\sum_{i=1}^{n}(X_{(i)}Y_{(i)} - X_{(i)}\bar{Y}) = \frac{\sum_{i=1}^{n} X_{(i)}Y_{(i)}}{n} - \bar{X}\bar{Y} = Cov(X,Y)$
- $\frac{1}{n}\sum_{i=1}^{n}(X_{(i)}^2 - X_{(i)}\bar{X}) = \frac{\sum_{i=1}^{n} X_{(i)}^2}{n} - \bar{X}^2 = Var(X)$
- It must be noted that we are talking about the **empirical variance and covariance** here since we do not know the underlying distribution of X or Y

- Thus, $b = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} * \frac{\sigma_Y}{\sigma_X} = \rho\frac{\sigma_Y}{\sigma_X}$ – **(B)**
  - Here ρ is the correlation coefficient
- Using the value of b in (A), we can get the value of a as:
  - $a = \bar{Y} - \frac{Cov(X,Y)}{Var(X)}\bar{X}$

- The second-order derivative is:
  - $\begin{pmatrix} \frac{\partial^2 SSE}{\partial a \partial a} & \frac{\partial SSE}{\partial a \partial b} \\ \frac{\partial^2 SSE}{\partial b \partial a} & \frac{\partial^2 SSE}{\partial b \partial b} \end{pmatrix} = \begin{pmatrix} 2n & 2\sum_{i=1}^{n} X_{(i)} \\ 2\sum_{i=1}^{n} X_{(i)} & 2\sum_{i=1}^{n} X_{(i)}^2 \end{pmatrix} = 2n\begin{pmatrix} 1 & \frac{\sum_{i=1}^{n} X_{(i)}}{n} \\ \frac{\sum_{i=1}^{n} X_{(i)}}{n} & \frac{\sum_{i=1}^{n} X_{(i)}^2}{n} \end{pmatrix}$, or
  - $det\begin{pmatrix} \frac{\partial^2 SSE}{\partial a \partial a} & \frac{\partial SSE}{\partial a \partial b} \\ \frac{\partial^2 SSE}{\partial b \partial a} & \frac{\partial^2 SSE}{\partial b \partial b} \end{pmatrix} = 2nVar(X)$, and also $\frac{\partial^2 SSE}{\partial a \partial a} = 2n$
    - Since the determinant and sub-determinants are always positive, the solutions for a and b arrived at by setting the partial derivatives to 0 is a global minimum
  - An important assumption in this derivation is that $Var(X) \neq 0$
- Given the equation $Y = a + bX + \varepsilon$, since the optimal values of a and b satisfy the equation $\bar{Y} = a + b\bar{X}$,
  - $E[Y] = E[a + bX + \varepsilon] \implies \bar{Y} = a + b\bar{X} + E[\varepsilon] \implies E[\varepsilon] = \bar{Y} - a - b\bar{X}$, or $E[\varepsilon] = 0$
- Also, $Cov(X, \varepsilon) = Cov(X, Y - a - bX) = Cov(X,Y) - bCov(X,X) = 0$
  - The explanatory variables and the error are independent
- If $Var(X) = 0$, then any line passing through $(\bar{X}, \bar{Y})$ will have the same least sum of square errors. In this case, an infinite number of least squares estimators are present.
- **Multivariate regression**: In multivariate regression, there are more than one explanatory variables.
  - In the case of multivariate regression, the regression equation is written as:
    - $Y_i = X_i^T\beta + \varepsilon_i$
    - Here X_i^T is a vector of explanatory variables or covariates. If $X_i \in R^{p+1}$, X_i is of the form $(X_i^{(0)}, X_i^{(1)}, X_i^{(2)}, \ldots X_i^{(p)})^T$ – generally $X_i^{(0)}$ is taken as 1 for all i to allow for an intercept term. p is the number of explanatory variables
    - $X_i \in R^{p+1}$, then $\beta \in R^{p+1}$ (to allow for the intercept term)
    - Given n observations, we have n equations of the form:
      - $Y_i = \beta_0 + X_i^{(1)}\beta_1 + X_i^{(2)}\beta_2 + \cdots X_i^{(p)}\beta_p + \varepsilon_i \; \forall \, i \in [1,n]$
    - $\{\varepsilon_i\}_{i=1,2,3,..n}$ are noise terms satisfying the equation $Cov\left(X_i^{(j)}, \varepsilon_i\right) = 0 \; \forall \, j \in [1,p]$
  - For multivariate regression, the least squares estimator (LSE) of β is the value of β that minimizes $SSE = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - X_i^T\beta)^2$

- In the equation $SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$, X_i is of the form $(X_i^{(0)}, X_i^{(1)}, X_i^{(2)}, \dots X_i^{(p)})^T$, i.e. each observation is a column vector
- If we instead choose to represent the vector X such that each observation is in a row, then X is called the design matrix; each row corresponds to one observation, and for the intercept, we take $X_i^{(0)} = 1 \forall i \in [1, n]$. So, for instance, the first row of the design matrix would be $(1, X_1^{(1)}, X_1^{(2)}, \dots X_1^{(p)})$. Clearly, X is a $nx(p+1)$ matrix.
- With this change of representation,
  - $SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - X_i \beta)^2 = ||Y - X\beta)||_2^2$
  - β is a $(p+1)x1$ matrix of the form $(\beta_0, \beta_1, \beta_2, \dots \beta_p)$
  - Y is a $nx1$ matrix of the form $(Y_1, Y_2, .. Y_n)$

- To minimize $||Y - X\beta||_2^2$
  - Let us consider the solution to the equation $X\beta = Y$
  - The equation can only be solved if Y is in the column space of X. If Y is not in the column space of X, we must find a vector $\hat{\beta}$ such that the difference between $X\hat{\beta}$ and Y is as small as possible.
    - $X\hat{\beta}$ lies in the column space of X, and let $X\hat{\beta} = p$
    - It can be easily shown geometrically that for $Y - p$ (error) to be the least, p must be the projection of Y onto the column space of X
  - 

  - $X\hat{\beta} = Y - e$, or
  - $X^T X\hat{\beta} = X^T (Y - e)$ – however X^T has as its rows column vectors of X which are perpendicular to e. Hence $X^T e = 0$.
  - $X^T X\hat{\beta} = X^T Y$, or
  - $\hat{\beta} = (X^T X)^{-1} X^T Y$ -for the inverse to be possible, X must be a full column rank matrix. The number of observations must be greater than the number of explanatory variables, and there must be no linear relationship between the explanatory variables.
    - It need not be a full rank matrix (both full row rank and full column rank) and indeed will not be a full rank matrix in most practical cases where the number of observations is much higher than the number of explanatory variables
  - Since the error $(Y - X\beta)$ is minimized when $\beta = \hat{\beta} = (X^T X)^{-1} X^T Y$, $\hat{\beta}$ is the least squares estimator of β
  - The same conclusion can also be arrived at using matrix calculus

        - The projection $p = X\hat{\beta} = X(X^TX)^{-1}X^TY = PY$ where $P = X(X^TX)^{-1}X^T$ is the projection matrix. The projection matrix when left-multiplied with any vector gives us the projection of the vector onto the column space of X.
            - Clearly, $P(PY) = p \Rightarrow P^2 = P$
- **Deterministic setting for linear regression**: In the setting of deterministic design for linear regression, we assume that the relationship between y and x is deterministic, i.e. is completely known. However, due to noise, what we can observe is the relationship:
    - $Y = X\beta + \varepsilon$
        - In this equation, X and β are not random variables – they are deterministic quantities
        - ε is a random variable with mean 0, and therefore Y is a random variable
        - $E[Y] = E[X\beta + \varepsilon] = X\beta$
        - The deterministic assumption is necessary to ensure that we can compute an estimator for β and the distribution of that estimator. If we treat X as a random variable, this becomes very difficult as the underlying distribution of X is not known. The assumption is also practical because in a regression setting, the values of X are observed values known with certainty.
    - Now, $\hat{\beta} = (X^TX)^{-1}X^TY$
        - $E[\hat{\beta}] = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TE[Y] = (X^TX)^{-1}X^TX\beta = \beta$
    - The errors ε_i have a mean of 0, and are considered independent and identically distributed.
        - The zero mean is characteristic of any phenomena that is described as noise
        - The independence assumption implies there is no serial correlation, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall \; i \neq j$
        - The identical distribution assumption implies that the errors are homoscedastic, that is the variance of ε_i is constant for all i. If the variance changes across the range of the explanatory variables, we will get fan-shaped plots of the errors (residuals) versus the explanatory variables. This means that the errors are not independent of the explanatory variables.
    - The assumptions of zero mean, independence and identical distribution (homoscedasticity) are often compressed into one assumption:
        - $\varepsilon \sim N_n(0, \sigma^2 I_n)$ were ε is the random vector of errors in dimension n, and $\sigma^2 I_n$ is its covariance matrix
        - Also, since $Y = X\beta + \varepsilon$
            - $Y \sim N_n(X\beta, \sigma^2 I_n)$
        - If we are assuming that the errors or the residuals are normally distributed, we must check this assumption by using Kolmogorov-Smirnov test or a basic Q-Q plot. Hypothesis testing for linear regression coefficients using t-tests involves the assumption that the errors are normally distributed.
    - If we assume that $Y \sim N_n(X\beta, \sigma^2 I_n)$, then the least squares estimator is also the maximum likelihood estimator of β
        - Y_i has the pdf $P_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - x_i\beta)^2}$
        - Therefore $L(Y_1, Y_2, \ldots Y_n; \beta) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2}$

- The log likelihood is:
  - $\log L(Y_1, Y_2, \ldots Y_n; \beta) = \frac{n}{2}\log\frac{1}{2\pi} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$
- To maximize the log likelihood, we need to minimize $\sum_{i=1}^{n}(y_i - x_i\beta)^2$ since all other quantities in the right hand side of the above equation are known
- Minimizing $\sum_{i=1}^{n}(y_i - x_i\beta)^2$ gives us the least squares estimator of β, $\hat{\beta}$. Hence the maximum likelihood estimator and the least squares estimator of β are the same.

- **Distribution of $\widehat{\boldsymbol{\beta}}$ in a deterministic linear regression setting**: In a deterministic linear regression setting where we assume $\varepsilon \sim N_n(0, \sigma^2 I_n)$:
  - $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $Y = X\beta + \varepsilon$
  - Or, $\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \varepsilon)$,
  - Or, $\hat{\beta} = ((X^T X)^{-1} X^T X)\beta + (X^T X)^{-1} X^T \varepsilon$
  - Or, $\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$
  - Since ε follows a normal distribution, $\hat{\beta}$ also follows a normal distribution
    - $\hat{\beta} \sim N_{p+1}(\beta, \Sigma_{p+1})$ where
      - $\Sigma_{p+1} = ((X^T X)^{-1} X^T)\sigma^2 I_n((X^T X)^{-1} X^T)^T$
      - The size of the matrix $((X^T X)^{-1} X^T)$ is $(p+1)xn$. When this matrix is multiplied by an identity matrix of size nxn, the original matrix. i.e. $((X^T X)^{-1} X^T)$ is returned
      - Therefore, $\Sigma_{p+1} = ((X^T X)^{-1} X^T)\sigma^2((X^T X)^{-1} X^T)^T$
      - σ^2 is a scalar, and hence $\Sigma_{p+1} = \sigma^2((X^T X)^{-1} X^T)((X^T X)^{-1} X^T)^T$
      - Or, $\Sigma_{p+1} = \sigma^2((X^T X)^{-1} X^T X(X^T X)^{-1}) = \sigma^2(X^T X)^{-1}$ (since $X^T X$ is a symmetric matrix, and the inverse of a symmetric matrix is also symmetric with its transpose being equal to the matrix itself)
    - The variance of $\hat{\beta}$ increases with the variance of the error term (residuals). It also increases as $(X^T X)^{-1}$ increases. $(X^T X)^{-1}$ is high when the range of values of X is small – in this case, the estimator becomes unreliable (in the extreme case, when X has no variance, an infinite number of estimators are possible).
  - The quadratic risk for the estimator $\hat{\beta}$ is computed as:
    - $E[||\hat{\beta} - \beta||_2^2] = (\hat{\beta}_0^2 - \beta_0^2) + (\hat{\beta}_1^2 - \beta_1^2) + \cdots + (\hat{\beta}_p^2 - \beta_p^2)$
    - $(\hat{\beta}_0^2 - \beta_0^2) + (\hat{\beta}_1^2 - \beta_1^2) + \cdots + (\hat{\beta}_p^2 - \beta_p^2)$ is the sum of the elements on the principal diagonal of the matrix $(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$. The sum of the elements on the principal diagonal of a matrix is the trace of the matrix
    - Hence, $E[||\hat{\beta} - \beta||_2^2] = E[trace((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T)]$
    - The trace of a matrix is a linear operation, and it is possible to swap the expectation and the trace, and:
      - $E[trace((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T)] = trace(E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T])$
    - Since $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^T X)^{-1})$,
      - $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = Cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$
    - Therefore, $E[||\hat{\beta} - \beta||_2^2] = trace(\sigma^2(X^T X)^{-1}) = \sigma^2 trace((X^T X)^{-1})$
- **Prediction Error**: The prediction error is the sum of squared residuals.
  - Prediction error = $E[||Y - X\hat{\beta}||_2^2]$

- However, $X\hat{\beta}$ is the projection of Y onto the column space of X, and can be written as PY where P is the projection matrix $X(X^TX)^{-1}X^T$
- Hence, prediction error = $E[||Y - PY||_2^2] = E[||(I_n - P)Y||_2^2]$
- If P is the projection matrix onto a subspace, then $(I - P)$ is the projection matrix for the orthogonal subspace. In this case, $(I_n - P)$ is a matrix which projects a vector onto a subspace that is perpendicular to the subspace of the column vectors of X
- We can write the prediction error as $E[||Y - PY||_2^2] = E[||(I_n - P)(X\hat{\beta} + \varepsilon)||_2^2]$
  - Or, $E[||Y - PY||_2^2] = E[||(I_n - P)\varepsilon||_2^2]$ (since $(I_n - P)(X\hat{\beta})$ is the zero matrix)
- It can be shown that using the property of translational invariance of ε, we can say that effectively:
  - $(I_n - P)\varepsilon \sim N_{n-(p+1)}(0, \sigma^2 I_{n-(p+1)})$
  - Prediction error = $E[||Y - X\hat{\beta}||_2^2] = (n - p - 1)\sigma^2$ where p is the number of explanatory variables
- The prediction error also provides a way to compute an unbiased estimator for σ^2
  - $\hat{\sigma}^2 = \frac{E[||Y - X\hat{\beta}||_2^2]}{(n-p-1)}$
- Using Cochran's theorem for random vectors, we have:
  - $\hat{\beta}$ is orthogonal to $\hat{\sigma}^2$
  - $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$

- **Significance tests**: The most typical test in the case of linear regression is whether the j^{th} explanatory variable is significant where $0 \leq j \leq p$
  - Here the hypothesis are: H_0: $\beta_j = 0$ and H_1: $\beta_j \neq 0$
  - We know that the estimator $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^TX)^{-1})$
    - Hence $\widehat{\beta_j} \sim N(\beta_j, (\sigma^2(X^TX)^{-1})_{j,j})$ for $0 \leq j \leq p$
      - Here $(\sigma^2(X^TX)^{-1})_{j,j}$ is the $(j,j)th$ element of the covariance matrix for $\hat{\beta}$ and gives the variance of $\widehat{\beta_j}$
    - If we denote $((X^TX)^{-1})_{j,j}$ by γ_j, then:
      - $\widehat{\beta_j} \sim N(\beta_j, \sigma^2\gamma_j)$ for $0 \leq j \leq p$
      - $\frac{(\widehat{\beta_j} - \beta_j)}{\sigma\sqrt{\gamma_j}} \sim N(0,1)$ – (A)
    - Also, $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$
      - Or, $\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} = \sqrt{\frac{\chi^2_{n-(p+1)}}{n-(p+1)}}$ – (B)
    - Dividing corresponding sides of (A) by (B), we get:
      - $\frac{(\widehat{\beta_j} - \beta_j)}{\hat{\sigma}\sqrt{\gamma_j}} = t_{n-(p+1)}$
      - Here, $\widehat{\beta_j}$ is an unbiased estimator for the mean of β_j, while $\hat{\sigma}$ is an unbiased estimator for the variance of β_j. Also, $\widehat{\beta_j}$ is orthogonal to $\hat{\sigma}$
  - The null hypothesis is that $\beta_j = 0$, and for this null hypothesis, the test statistic is:
    - $T_n = \frac{\widehat{\beta_j}}{\hat{\sigma}\sqrt{\gamma_j}}$ and the test statistic follows a t-distribution with $n - (p + 1)$ degrees of freedom, where p is the number of explanatory variables

- Other types of hypothesis tests can also be undertaken. For instance, if we want to test whether $\beta_2 > \beta_3$, we have the following hypothesis:
  - $H_0: \beta_2 - \beta_3 \leq 0$ and $H_1: \beta_2 - \beta_3 > 0$
  - We can choose a unit vector u such that:
    - $u\beta \leq 0 \iff \beta_2 - \beta_3 \leq 0$
    - For example, if we have $p = 5$, the unit vector u would be $(0,0,\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},0,0)^T$
  - We know that $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2 (X^T X)^{-1})$
    - Therefore, $u^T\hat{\beta} = N_{p+1}(u^T\beta, \sigma^2 u^T (X^T X)^{-1} u)$
  - The test statistic in this case is:
    - $T_n = \frac{u^T\hat{\beta}}{\hat{\sigma}\sqrt{u^T(X^T X)^{-1}u}}$
    - In this case, we will have to conduct a one-sided test

- **Bonferroni Correction in Hypothesis Testing**: If we are testing a large number of hypothesis, each with a significance level of α, the composite probability of Type 1 error is going to be larger than α
  - For instance, if we have 10 tests each of which has a probability of type I error of 5%, then the probability that all the tests are accurate is:
    - $(0.95)^{10} \cong 0.5987 = 59.87\%$ and the probability that at least of the tests has a type I error is close to 40%
    - In such cases, we have to reduce the significance level for each individual test such that the composite significance level is 5%. If we choose a significance level of 5/10 = 0.5% for each of the tests, then the probability that all the tests are accurate is:
      - $(0.995)^{10} \cong 0.9511 = 95.11\%$ and the probability that at least of the tests has a type I error is close to 5%
  - Formally, Bonferroni's correction is used to test whether a group of explanatory variables is significant in linear regression. if we have the null and alternative hypothesis of the form:
    - $H_0: \beta_j = 0 \; \forall \, j \in S$, and $H_1: \exists \, j \in S, \; \beta_j \neq 0$ where $S = \{0,1,2 \ldots p\}$
    - At a significance level of α, let $R_{j,\propto}$ be the rejection region for coefficient j where $0 \leq j \leq p$
    - The overall rejection region for all the coefficients has an upper bound of:
      - $R_{S,\alpha} \leq \bigcup_{j=0}^{p} R_{j,\alpha}$ where $S = \{0,1,2 \ldots p\}$ and $0 \leq j \leq p$
      - If all the rejection regions are disjoint then:
        - $R_{S,\propto} = (p+1)\alpha$
    - Bonferroni's correction states that for each individual test, the significance level must be taken as:
      - $\alpha/(p+1)$
      - In this case, $R_{S,\alpha} \leq \bigcup_{j=0}^{p} R_{j,\frac{\alpha}{p+1}}$ and the maximum value of $R_{S,\alpha}$ is α. In other words, the non-asymptotic level of the composite test has a significance level which is at most α

- Bonferroni's correction is very conservative, as the rejection regions for each of the hypothesis are unlikely to be disjoint. The significance level for each of the tests when applying Bonferroni's correction may be very aggressive and it may become very difficult to reject the null hypothesis.

Generalized Linear Models

- In a linear regression setup, we have:
  - The conditional expectation of the response variable is a linear function of the covariates, or the explanatory variables
  - The errors are independent and identically distributed random variables with mean 0 and variance σ^2 (strictly speaking, this is not a requirement for linear regression, but is used in almost all practical setups)
  - We assume that the covariates are known and are not random variables
    - This deterministic assumption allows us to compute the distribution of $Y|X = x$ where Y is the response variable in R, and X are the explanatory variables in R^{p+1} where p is the number of explanatory variables and we include an additional variable for the intercept
    - The expectation of $Y|X = x$ is simply $x^T\beta$ (or $x\beta$ if we represent x in the design matrix form) since the errors have a zero mean. Its distribution is a normal distribution since the errors follow a normal distribution and $x^T\beta$ is a constant. The covariance matrix is the covariance matrix of the errors.
- In generalized linear models, we make two departures from the linear regression setting.
  - The expectation of $Y|X = x$ is not a linear function of the covariates. In other words, the response variable Y is not a linear function of the explanatory variables
  - The distribution of $Y|X = x$ is not a Gaussian distribution. In other words, the response variable is not a Gaussian random variable.
- As an example, if the response variable takes on only binary values, then $E[Y|X = x]$ must lie in the interval $(0,1)$
  - In such a case, we can still use a linear combination of the covariates but must find a function f such that $f(x^T\beta)$ lies between 0 and 1
  - Also, the distribution of $Y|X = x$ is a Bernoulli distribution in this case
- **Link function**: In generalized linear models, we use a link function g such that:
  - $g\big(\mu(x)\big) = x^T\beta$ where $\mu(x)$ is $E[Y|X = x]$
  - Essentially, we want to find a function that takes as input only the range of possible values of the conditional expectation but can output any real value
  - The above equation can also be written as:
    - $\mu(x) = g^{-1}(x^T\beta)$ where g^{-1} is the inverse of the link function. It takes any real value as input but restricts the output to the range of values the conditional expectation can take.
  - The choice of the link function depends on the range of values of the response variable, and there can be multiple link functions for a specific distribution of Y. The choice of the link function is made with a view to be able to compute $\hat{\beta}$, generally using MLE estimation techniques
  - As an example of choosing a mean function and a link function, consider a case where a fox is in a farm which has grape plants. The maximum number of grapes the fox can consume is determined by the number of grapes in the farm. The response variable, Y, is the number of grapes consumed. Y can only be non-negative integers, and the Poisson distribution would be a good candidate distribution for Y

- The explanatory variable, X is taken to be the number of grapes in the farm
- The expected value of a Poisson distribution, or $E[Y|X=x]$ would be λ or the parameter of the distribution. However, since we are performing regression, we want to make $E[Y|X=x]$ a function of x
- One possible choice for $E[Y|X=x]=\frac{mx}{h+x}$ where m represents the maximum consumption rate (the fox will not be able to consume all the grapes in the farm) and h represents the grape density at which the consumption rate is half of the maximum rate (h controls how fast x goes to m)
- Here, $\mu(x)=\frac{mx}{h+x}$ and we choose the link function g such that $g: x \to 1/x$
  - $g\big(\mu(x)\big)=\frac{h+x}{mx}=\frac{1}{m}+\frac{h}{m}\frac{1}{x}$ which is a linear function in $\frac{1}{x}$
  - In this case, g is a reciprocal link function. Also, the range of g, i.e. $\frac{1}{m}+\frac{h}{m}\frac{1}{x}$ does not span the entire real line

- **Exponential family of distributions**: A family of distributions $\{P_\theta : \theta \in \Theta\}, \Theta \subset R^k$ is said to be a k-parameter exponential family on R^q (we are taking the general case of a q-dimensional response variable, generally q=1), if there exist real-valued functions:
  - $\eta_1, \eta_2, .. \eta_k$ and B of $\theta \in R^k$ (these functions do not depend on the response variable Y)
    - Each of these functions is a map from $R^k \to R$
  - $T_1, T_2, .. T_k$ and h of the explanatory variable $Y \in R^q$ (these functions do not depend on θ)
    - Each of these functions is a map from $R^q \to R$

    such that the probability density or probability mass function can be written as:
    - $P_\theta(y) = h(y)e^{[\sum_{i=1}^{k}(\eta_i(\theta)T_i(y)]-B(\theta)}$
    - If we take $\eta(\theta)=\begin{pmatrix}\eta_1(\theta)\\ \eta_2(\theta)\\ \dots\\ \eta_k(\theta)\end{pmatrix}$ and $T(y)=\begin{pmatrix}T_1(y)\\ T_1(y)\\ \dots\\ T_k(y)\end{pmatrix}$, then the above equation can also be written as:
      - $P_\theta(y) = h(y)e^{(\eta(\theta)^T T(y)-B(\theta))}$
  - In the case of a univariate normal distribution, we have $Y \in R^1$, while $\Theta \in R^2$ (there are two unknown parameters, μ and σ^2). We can write the pdf of Y as:
    - $P_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{-\frac{y^2}{2\sigma^2}+\frac{\mu y}{\sigma^2}}e^{\log\frac{1}{\sqrt{2\pi\sigma^2}}}e^{-\frac{\mu^2}{2\sigma^2}}$, or
    - $P_\theta(y) = 1e^{-\frac{y^2}{2\sigma^2}+\frac{\mu y}{\sigma^2}}e^{-\frac{\mu^2}{2\sigma^2}-\log\sqrt{2\pi\sigma^2}}$
      - Here, $\eta(\theta)=\begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix}$, $T(y)=\begin{pmatrix}y\\ y^2\end{pmatrix}$, $B(\theta)=\frac{\mu^2}{2\sigma^2}+\log\sqrt{2\pi\sigma^2}$ and $h(y)=1$
      - This is not a unique choice of $\eta(\theta), T(y), B(\theta)$ and $h(y)$
    - If σ^2 is given, then $\Theta \subset R$ (only the mean μ is unknown) and we would have:
      - $\eta(\theta)=\frac{\mu}{\sigma^2}$, $T(y)=y$, $B(\theta)=\frac{\mu^2}{2\sigma^2}+\log\sqrt{2\pi\sigma^2}$ and $h(y)=e^{-\frac{y^2}{2\sigma^2}}$
  - For a Bernoulli distribution with parameter p, we have:

- $P_\theta(y) = p^y(1-p)^{1-y} = e^{(y\log p+(1-y)\log(1-p))} = 1e^{(y\log\frac{p}{(1-p)}-\log\frac{1}{(1-p)})}$
- Here, $\eta(\theta) = \log\frac{p}{(1-p)}$, $T(y) = y$, $B(\theta) = \log\frac{1}{(1-p)}$, and $h(y) = 1$

- For the Poisson distribution, we have:
  - $P_\theta(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!}e^{(y\log\lambda-\lambda)}$
  - Here, $\eta(\theta) = \log\lambda$, $T(y) = y$, $B(\theta) = \lambda$, and $h(y) = \frac{1}{y!}$

- **One-parameter canonical exponential families**: In the case when there is only one parameter, i.e. $\{P_\theta : \theta \in \Theta\}, \Theta \subset R^1$ and the response variable Y is also in R^1, then for most of the common distributions, we can express the pdf in the form:
  - $P_\theta(y) = e^{(\frac{y\theta-b(\theta)}{\phi}+c(y,\phi))}$
    - where the parameter θ may be a function of the original parameter - θ is called the canonical parameter
    - b and c are known functions - $b(\theta)$ does not depend on y and $c(y,\phi)$ does not depend on θ. $b(\theta)$ is also known as the log-partition function.
    - ϕ is known as the dispersion parameter
  - As an example, for the normal distribution where the variance σ^2 is given, we can express the pdf as:
    - $P_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{\frac{\mu y}{\sigma^2}-\frac{\mu^2}{2\sigma^2}}e^{\log\frac{1}{\sqrt{2\pi\sigma^2}}}e^{-\frac{y^2}{2\sigma^2}}$
    - Or, $P_\theta(y) = e^{(\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}+\log(\frac{1}{\sqrt{2\pi\sigma^2}})-\frac{y^2}{2\sigma^2})}$
      - Here, b is a function which maps $\theta \to \theta^2/2$
      - $\phi = \sigma^2$
      - $c(y,\phi) = \log(\frac{1}{\sqrt{2\pi\phi}}) - \frac{y^2}{2\phi^2}$
  - In the case of the Bernoulli distribution, we have:
    - $P_\theta(y) = p^y(1-p)^{1-y} = e^{(y\log p+(1-y)\log(1-p))} = 1e^{(y\log\frac{p}{(1-p)}-\log\frac{1}{(1-p)})}$
    - Here, θ is equal to $\log\frac{p}{(1-p)}$
      - Or, $\frac{p}{(1-p)} = e^\theta$
      - Or, $(1-p)e^\theta = p$, or $p = \frac{e^\theta}{1+e^\theta}$
      - Therefore, $\log\frac{1}{(1-p)} = \log\ (1+e^\theta)$
    - Here, $\theta = \log\frac{p}{(1-p)}$, $\phi = 1$, b is a function which maps $\theta \to \log\ (1+e^\theta)$, and $c(y,\phi) = 0$
  - In the case of the Poisson distribution, we have:
    - $P_\theta(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!}e^{(y\log\lambda-\lambda)} = e^{(y\log\lambda-\lambda+\log\frac{1}{y!})}$
    - Here, $\theta = \log\lambda$, or $\lambda = e^\theta$
    - We have $\theta = \log\lambda$, $\phi = 1$, b is a function which maps $\theta \to e^\theta$ and $c(y,\phi) = \log\frac{1}{y!}$

- It can also be shown that when we have the one-parameter canonical exponential family, the canonical parameter θ which is a function of the original parameter of the distribution is such that it can take any real value
  - For instance, in the normal distribution, $\theta = \mu$ and since μ can take any real value, θ can also take on any real value
    - In the Bernoulli case, $\theta = \log\frac{p}{(1-p)}$ – for $p \in (0,1)$, $\log\frac{p}{(1-p)}$ can take any value on the real number line
    - Similarly, for the Poisson distribution, $\theta = \log\lambda$, for positive values of λ, θ can take any value on the real number line

- **Mean and Variance of One-parameter canonical exponential family**: In the case of one-parameter canonical exponential family, we have:
  - $E\left[\frac{d\log P_\theta(y)}{d\theta}\right] = 0$ where $\log P_\theta(y)$ is the log likelihood (this is from the derivation of Fisher information for a one parameter distribution)
    - When $P_\theta(y) = e^{(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi))}$, this implies:
      - $E\left[\frac{Y - b'(\theta)}{\phi}\right] = 0$, or $E[Y] = b'(\theta)$
  - Also, $Var\left[\frac{d\log P_\theta(y)}{d\theta}\right] = E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right] - E\left[\frac{d\log P_\theta(y)}{d\theta}\right]^2 = E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right]$
    - $Var\left[\frac{d\log P_\theta(y)}{d\theta}\right] = E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right]$
    - From the derivation of the Fisher information for a one-parameter distribution, we know that:
      - $E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right] = \frac{(\frac{d(P_\theta(y))}{d\theta})^2}{P_\theta(y)}$
  - Also, from the derivation of the Fisher information for a one-parameter distribution, we know that:
    - $Var\left[\frac{d\log P_\theta(y)}{d\theta}\right] = -E[\frac{d^2\log P_\theta(y)}{d\theta^2}]$
    - Since $Var\left[\frac{d\log P_\theta(y)}{d\theta}\right] = E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right]$, this implies
    - $E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right] + E\left[\frac{d^2\log P_\theta(y)}{d\theta^2}\right] = 0$
  - If we denote $\log P_\theta(y)$ as l, we can write the above equation as:
    - $E\left[\left(\frac{dl}{d\theta}\right)^2\right] + E\left[\frac{d^2 l}{d\theta^2}\right] = 0$
  - Using the identity $E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right] + E\left[\frac{d^2\log P_\theta(y)}{d\theta^2}\right] = 0$ along with the result that $E[Y] = b'(\theta)$, we can derive an expression for $Var[Y]$
    - $E\left[\left(\frac{d\log P_\theta(y)}{d\theta}\right)^2\right] = E\left[\left(\frac{Y - b'(\theta)}{\phi}\right)^2\right] = E\left[\left(\frac{Y - E[Y]^2}{\phi}\right)^2\right] = \frac{Var(Y)}{\phi^2}$
    - $E\left[\frac{d^2\log P_\theta(y)}{d\theta^2}\right] = -\frac{b''(\theta)}{\phi}$
    - Therefore, we have $\frac{b''(\theta)}{\phi} = \frac{Var(Y)}{\phi^2}$, or $Var(Y) = \phi b''(\theta)$
  - It may be noted that while we have derived expressions for $E[Y]$ and $Var(Y)$, these are actually conditional on X in a regression setting. In a regression setup, all computed

parameters must depend on observed values, and the expectation of Y or its variance will be dependent on the observed values of X.

- For various distributions, we can compute $\phi b''(\theta)$ and check that the results are correct
  - For the normal distribution where σ^2 is given,
    - $b(\theta) = \theta^2/2$ where $\theta = \mu$ and hence $b''(\theta) = 1$
    - $\phi = \sigma^2$
    - Hence $Var(Y) = \phi b''(\theta) = \sigma^2$
  - For the Bernoulli distribution, we have:
    - $b(\theta) = \log\ (1 + e^{\theta})$ where $\theta = \log\frac{p}{(1-p)}$ or $p = \frac{e^{\theta}}{1+e^{\theta}}$
      - $b''(\theta) = \frac{e^{\theta}}{(1+e^{\theta})^2} = p\frac{1}{1+e^{\theta}} = p\frac{1}{1+\frac{p}{1-p}} = p(1-p)$
    - $\phi = 1$
    - Hence $Var(Y) = \phi b''(\theta) = p(1-p)$
  - For the Poisson distribution, we have:
    - $b(\theta) = e^{\theta}$ where $\theta = \log\lambda$ or $\lambda = e^{\theta}$
    - $b''(\theta) = e^{\theta} = \lambda$
    - $\phi = 1$
    - Hence $Var(Y) = \phi b''(\theta) = \lambda$

- In generalized linear models, we take $Y|X = x$ to be a distribution in the exponential family.
  - The mean $E[Y|X = x]$ is a function of x, $\mu(x)$
  - The link function g is such that $g\big(\mu(x)\big) = x^T\beta$, i.e. its range is a linear function of x
  - g needs to be monotonic and differentiable
- For Poisson distribution, we have:
  - $\mu(x) > 0$
  - We can choose the logarithmic function as the link function. The logarithmic function maps $(0, \infty) \to R$ and is monotonically increasing and differentiable
  - $g\big(\mu(x)\big) = \log\mu(x)$ is a natural choice for the link function
- For the Bernoulli distribution,
  - $0 < \mu(x) < 1$
  - The link function should map $(0,1) \to R$
  - One of the choices for the link function is a function that maps $x \to \log\left(\frac{x}{1-x}\right)$. The domain of this function (i.e. values of x) when restricted to $(0,1)$ gives the range (i.e. $\log\left(\frac{x}{1-x}\right)$) as the entire real line
    - This is the logit link function $g\big(\mu(x)\big) = \log\frac{\mu(x)}{(1-\mu(x))}$. This is monotonically increasing and differentiable.
    - The inverse function for the logit function is a function that maps $\log\left(\frac{x}{1-x}\right) \to x$.
      - $g^{-1}(x) = \frac{e^x}{1+e^x}$ and $g^{-1}\left(\log\left(\frac{x}{1-x}\right)\right) = x$
  - Another choice for the link function is the probit function, which relies on the fact that the cdf of a distribution is always a value between 0 and 1. If choose a normal distribution, then the inverse cdf spans the entire real number line.

        - For the probit link function $g(\mu(x)) = \Phi^{-1}(\mu(x))$. The inverse cdf is monotonically increasing and differentiable.
        - The inverse function for the probit link function is the cdf function Φ
- **Canonical Link Function**: The canonical link function is one which takes as input the conditional expectation and outputs the canonical parameter θ.
    - In other words, we want $g(\mu(x)) = \theta$
    - The canonical exponential family has the pdf/pmf of the form:
        - $P_\theta(y) = e^{(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi))}$
    - We know that $\mu(x) = E[Y|X = x] = b'(\theta)$
        - Therefore, $g(b'(\theta)) = \theta$
        - Since $(b')^{-1}(b'(\theta)) = \theta$, hence $g = (b')^{-1}$ and $g(\mu(x)) = (b')^{-1}(\mu(x))$
    - Since b' is differentiable, its inverse $(b')^{-1}$ is also differentiable and hence g is differentiable
    - If the dispersion parameter ϕ is positive then given that $Var[Y|X = x] = b''(\theta)\phi > 0$,
        - $b''(\theta) > 0$
        - Since the derivative of $b'(\theta)$, $b''(\theta)$ is always positive, this implies that b' is a monotonically increasing function
        - Since b' is a monotonically increasing function, its inverse is also a monotonically increasing function and hence $g = (b')^{-1}$ is a monotonically increasing function
        - In the case where the dispersion parameter is negative, we can reparameterize the equation $P_\theta(y) = e^{(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi))}$ by multiplying both the numerator and denominator in $\frac{y\theta - b(\theta)}{\phi}$ by -1 and making appropriate changes, and this will give us a new function for $b(\theta)$ whose first derivative is monotonically increasing.
    - For the Bernoulli distribution, $b'(\theta) = \frac{e^\theta}{(1+e^\theta)} = p$
        - $(b')^{-1}$ must be a function that maps $\frac{e^\theta}{(1+e^\theta)} \to \theta$. This is satisfied by a function that maps $\log\frac{x}{(1-x)} \to x$
        - $(b')^{-1}(\mu(x)) = (b')^{-1}(p) = \log\frac{p}{(1-p)}$
        - Also, the dispersion parameter $\phi = 1$ or the dispersion parameter is positive
        - Thus, the canonical link function for the Bernoulli distribution is the logit function
    - Similarly, for the Poisson distribution, the log function is the canonical link function as $\log\lambda = \theta$. For the normal distribution with known variance, the link function is the identity function as $\mu = \theta$. For both the Poisson and the normal distributions, the dispersion parameter is positive ($\phi = 1$ and $\phi = \sigma^2$respectively)
- **Beta parameter for regression using generalized regression models**: Once we have expressed a distribution as a one-parameter canonical exponential family, and chosen a link function, we can start applying regression
    - For the one-parameter exponential family, we have:
        - $P_\theta(y) = e^{(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi))}$
            - where θ is the canonical parameter. θ can take any value on the real number line – this is an important consideration for regression

- $E[Y] = b'(\theta)$ – this can actually be expressed as $E[Y|X = x] = b'(\theta)$ because the expected value of Y will change depending on the observed data. Since this is a regression setting, we are actually using a strong assumption that the observed values of Y follow a particular exponential distribution. The observed values of Y are themselves dependent on the observed values of the covariates or the independent/explanatory variables, and hence $E[Y] = E[Y|X = x]$
- $Var[Y|X = x] = b''(\theta)\phi$

- The link function g is such that $g(E[Y|X = x]) = g(\mu(x))$ has as its range the entire real number line.
  - We can express $g(\mu(x)) = X\beta$ where X is the design matrix of size $nx(p + 1)$, and β has dimensions $(p + 1)x1$
    - In this equation, we assume we have n pairs of observations $(X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n)$
    - There are p explanatory variables and an additional variable for the intercept term. This means that $X \in R^{p+1}$
  - If the link function g is the canonical link function, then:
    - $g(\mu(x)) = (b')^{-1}(b'(\theta)) = \theta$
    - Or, since $g(\mu(x)) = X\beta$, this means $\theta = X\beta$
    - It may be noted that θ is a $nx1$ matrix. For every observation, there is a separate computed value of θ based on the data. As an example, if we assume the distribution of the response variable to be Bernoulli, then $\theta = \log\frac{p}{(1-p)}$ and the value of p is different for different observations, i.e. p is equal to either 0 or 1 for any particular observation
  - If the link function is not the canonical link function, then:
    - $g(\mu(x)) = X\beta$
    - Or, $g(b'(\theta)) = X\beta$
    - Or, $b'(\theta) = g^{-1}(X\beta)$
    - Or, $\theta = (b')^{-1}(g^{-1}(X\beta))$
    - In mathematical terms, this can be written as $\theta = h(X\beta)$ where:
      - $h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}$

- **Log-likelihood function for the exponential family**: If we have n observations $(X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n)$, we can write the likelihood function as:
  - $L((X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n);\ \theta_1,\theta_2,..\theta_n) = \prod_{i=1}^{n} e^{(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi))}$
  - Or, $\log L((X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n);\ \theta_1,\theta_2,..\theta_n) = \sum_{i=1}^{n}(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi))$
  - To simplify notation, we can represent $\log L((X_1,Y_1),(X_2,Y_2),\dots(X_n,Y_n);\ \theta_1,\theta_2,..\theta_n)$ as l
  - $l = \sum_{i=1}^{n}(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi))$
  - Or, $l = \sum_{i=1}^{n}(\frac{y_ih(x_i\beta) - b(h(x_i\beta))}{\phi} + c(y_i,\phi))$ since $\theta = h(X\beta)$ where $h = (g \circ b')^{-1}$
  - Or, $l = \sum_{i=1}^{n}(\frac{y_ih(x_i\beta) - b(h(x_i\beta))}{\phi}) + constant$ since $c(y_i,\phi)$ does not depend on the unknown parameter β

- If we are using the canonical link function, h is the identity function as $\theta = X\beta$ and the above expression simplifies to:
  - $l = \sum_{i=1}^{n}(\frac{y_i x_i\beta - b(x_i\beta)}{\phi}) + constant$
- There are n observations, and there are $(p+1)$ explanatory variables. This means that:
  - $\frac{dl}{d\beta} = \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \dots \\ \frac{\partial l}{\partial \beta_{p+1}} \end{pmatrix}$ – **(A)** (this is a $(p+1)x1$ matrix)
  - Let us consider the expression for $\frac{\partial l}{\partial \beta_k}$ for $0 <= k <= p$
    - $\frac{\partial l}{\partial \beta_k} = \frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}((\frac{y_i x_i\beta - b(x_i\beta)}{\phi}) + constant)$ – **(B)**
    - $\frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}(y_i x_i\beta) = \sum_{i=1}^{n} y_i x_i^k$ - **(C)** where x_i^k is the value of the k^{th} explanatory variable in the i^{th} observation with $x_i^0 = 1$
      - This is because $y_i x_i\beta = y_i(x_i^0\beta_0 + x_i^1\beta_1 + \cdots x_i^k\beta_k + \cdots x_i^p\beta_p)$
      - Thus, $\frac{\partial}{\partial \beta_k}(y_i x_i\beta) = y_i x_i^k$
    - $\frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}\big(b(x_i\beta)\big) = \sum_{i=1}^{n}\frac{\partial}{\partial (x_i\beta)}(b(x_i\beta))x\frac{\partial (x_i\beta)}{\partial \beta_k}$ (Using chain rule for derivatives)
      - For the canonical link function $x_i\beta = \theta_i$ and hence $\frac{\partial}{\partial (x_i\beta)}\big(b(x_i\beta)\big) = b'(\theta_i) = \mu_i$
      - Also, $\frac{\partial}{\partial \beta_k}(x_i\beta) = x_i^k$
      - Hence, $\frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}\big(b(x_i\beta)\big) = \sum_{i=1}^{n}(\mu_i x_i^k)$ – **(D)**
    - Using **(C)** and **(D)** in **(B)** we get:
      - $\frac{\partial l}{\partial \beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}(y_i x_i^k - \mu_i x_i^k) = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)x_i^k$ – **(E)**
  - Consider the matrix product:
    - $\frac{1}{\phi}\begin{pmatrix} x_1^0 & x_2^0 & \dots & x_n^0 \\ x_1^1 & x_2^1 & \dots & x_n^1 \\ \dots & \dots & \dots & \dots \\ x_1^{p+1} & x_2^{p+1} & \dots & x_n^{p+1} \end{pmatrix}\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \dots \\ y_n - \mu_n \end{pmatrix}$ – **(F)** (The first matrix is a $(p+1)xn$ matrix, and is the transpose of the design matrix)
    - It is easily seen that:
      - $\frac{dl}{d\beta} = \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \dots \\ \frac{\partial l}{\partial \beta_{p+1}} \end{pmatrix} = \frac{1}{\phi}\begin{pmatrix} x_1^0 & x_2^0 & \dots & x_n^0 \\ x_1^1 & x_2^1 & \dots & x_n^1 \\ \dots & \dots & \dots & \dots \\ x_1^{p+1} & x_2^{p+1} & \dots & x_n^{p+1} \end{pmatrix}\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \dots \\ y_n - \mu_n \end{pmatrix}$
        - where $\frac{\partial l}{\partial \beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)x_i^k$
- Hence, $\frac{dl}{d\beta} = \frac{1}{\phi}x^T(y - \mu)$ [in matrix form]

- where x is the design matrix with dimensions $nx(p+1)$
- y is the matrix of response variables with dimensions $nx1$
- μ is the matrix of conditional means with dimensions $nx1$
- In this equation, μ is the only term that depends on β

- In the equation, $\frac{\partial l}{\partial \beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)x_i^k$
  - $\mu_i = b'(x_i\beta)$
  - And so, $\frac{\partial l}{\partial \beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - b'(x_i\beta))x_i^k$
  - These are $(p+1)$ equations in $(p+1)$ variables and has to be solved using optimization techniques
- The second derivative of l with respect to β is:
  - $\frac{d^2 l}{d\beta^2} = \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0 \beta_0} & \frac{\partial^2 l}{\partial \beta_0 \beta_1} & \dots & \frac{\partial^2 l}{\partial \beta_0 \beta_{p+1}} \\ \frac{\partial^2 l}{\partial \beta_1 \beta_0} & \frac{\partial^2 l}{\partial \beta_1 \beta_1} & \dots & \frac{\partial^2 l}{\partial \beta_1 \beta_{p+1}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 l}{\partial \beta_{p+1} \beta_0} & \frac{\partial^2 l}{\partial \beta_{p+1} \beta_1} & \dots & \frac{\partial^2 l}{\partial \beta_{p+1} \beta_{p+1}} \end{pmatrix}$ which is a $(p+1)x(p+1)$ matrix
    - The (j,k) element of this matrix is $\frac{\partial^2 l}{\partial \beta_j \beta_k}$
  - Let us consider the value of $\frac{\partial^2 l}{\partial \beta_j \beta_k}$
    - $\frac{\partial^2 l}{\partial \beta_j \beta_k} = \frac{\partial}{\partial \beta_j}\left(\frac{\partial l}{\partial \beta_k}\right) = \frac{\partial}{\partial \beta_j}(-\frac{1}{\phi}\sum_{i=1}^{n} b'(x_i\beta)x_i^k)$
    - Or, $\frac{\partial^2 l}{\partial \beta_j \beta_k} = -\frac{1}{\phi}\sum_{i=1}^{n}\frac{\partial}{\partial (x_i\beta)}(b'(x_i\beta))x\frac{\partial}{\partial \beta_j}(x_i\beta)\, x_i^k$
    - Or, $\frac{\partial^2 l}{\partial \beta_j \beta_k} = -\frac{1}{\phi}\sum_{i=1}^{n} b''(x_i\beta))x\frac{\partial}{\partial \beta_j}(x_i\beta)\, x_i^k$
    - Or, $\frac{\partial^2 l}{\partial \beta_j \beta_k} = -\frac{1}{\phi^2}\sum_{i=1}^{n} Var(y_i)x_i^j\, x_i^k$ since $Var(Y) = \phi b''(\theta)$
  - Consider the matrix:
    - $-\frac{1}{\phi^2}\begin{pmatrix} x_1^0 & x_2^0 & \dots & x_n^0 \\ x_1^1 & x_2^1 & \dots & x_n^1 \\ \dots & \dots & \dots & \dots \\ x_1^{p+1} & x_2^{p+1} & \dots & x_n^{p+1} \end{pmatrix}\begin{pmatrix} Var(y_1) & 0 & \dots & 0 \\ 0 & Var(y_2) & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & Var(y_n) \end{pmatrix}\begin{pmatrix} x_1^0 & x_1^1 & \dots & x_1^{p+1} \\ x_2^0 & x_2^1 & \dots & x_2^{p+1} \\ \dots & \dots & \dots & \dots \\ x_n^0 & x_n^1 & \dots & x_n^{p+1} \end{pmatrix}$
    - Or, $-\frac{1}{\phi^2}\begin{pmatrix} x_1^0 & x_2^0 & \dots & x_n^0 \\ x_1^1 & x_2^1 & \dots & x_n^1 \\ \dots & \dots & \dots & \dots \\ x_1^{p+1} & x_2^{p+1} & \dots & x_n^{p+1} \end{pmatrix}\begin{pmatrix} Var(y_1)x_1^0 & Var(y_1)x_1^1 & \dots & Var(y_1)x_1^{p+1} \\ Var(y_2)x_2^0 & Var(y_2)x_2^1 & \dots & Var(y_2)x_2^{p+1} \\ \dots & \dots & \dots & \dots \\ Var(y_2)x_n^0 & Var(y_2)x_n^1 & 0 & Var(y_2)x_n^{p+1} \end{pmatrix}$
    - The (j,k) element of this matrix is clearly $-\frac{1}{\phi^2}\sum_{i=1}^{n} Var(y_i)x_i^j\, x_i^k$
  - Therefore, $\frac{d^2 l}{d\beta^2} = -\frac{1}{\phi^2}x^T\Sigma x$
    - where x is the design matrix of dimensions $nx(p+1)$
    - Σ is a diagonal matrix of dimensions nxn with positive elements on its diagonal

        - Any matrix of the form $x^T \Sigma x$ where Σ is a diagonal matrix with positive values on its principal diagonal can be shown to be a positive definite matrix. Hence $-\frac{1}{\phi^2} x^T \Sigma x$ is negative definite
        - This means that the solution to $\frac{dl}{d\beta} = 0$ gives a unique minimum which is the global minimum
    - The log-likelihood is strictly concave using the canonical link function when $\phi > 0$ and the design matrix x is full column rank. As a consequence, the maximum likelihood estimator is unique.
        - If a function other than the canonical link function is used, then the maximum likelihood function may not be strictly concave and several local maxima may appear
    - The maximum likelihood estimator for β is asymptotically normal if the statistical model and the parameter space satisfy the conditions for the asymptotic normality of the maximum likelihood estimator
- **Optimization approach to solving the maximum likelihood equation**: To obtain the maximum likelihood estimate for β, we need to solve for $(p+1)$ equations in $(p+1)$ variables.
    - There are $(p+1)$ equations of the form $\frac{\partial l}{\partial \beta_k} = \frac{1}{\phi} \sum_{i=1}^{n} (y_i - b'(x_i \beta)) x_i^k$
    - The gradient descent method is an iterative approach to finding the local minima or maxima given an initial solution. In the case of maximum likelihood estimation using the canonical link function, the local minima is also the global minima
    - The steps in the gradient descent method are:
        1. Choose a starting solution for $\beta = \beta_{initial}$
        2. Compute the likelihood function for the n observations given the current value of β and compute the derivative of the likelihood function with respect to β. Let the value of the derivative of the likelihood function be denoted by $\Delta\beta$
        3. Set the new value of β, β_{new} for the next iteration to be:
            i. $\beta_{new} = \beta_{old} - t\Delta\beta$
            ii. where t is a step size (learning rate) that has to be chosen
        4. If the absolute difference between the likelihood function value using the old and new values of β is less than some chosen value ε, then the algorithm is stopped and the value of β_{new} is taken as the final estimate of β. Otherwise, the algorithm goes back to step 2
    -