Convergence Concepts:

Let $\{X_n\}$ be any ~~...~~ sequence of random variables. We say that $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$ if for each $\varepsilon > 0$, It's equivalent to show that $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$ for $\forall \varepsilon > 0$, then $X_n \xrightarrow{P} X$

WLLN: Suppose that $X_1, X_2, \cdots, X_n$ are iid random variable with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ for $i = 1, 2, \cdots$ Define for each $n = 1, 2, \cdots$ $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$: Markov's Inequality

A.S. Convergence: $\{X_n, n = 1, 2, \cdots\}$ of R.V. is said to be converge a.s. to a R.V X if $P(\lim_{n\to\infty} X_n = X) = 1$

SLLN: $X_1, X_2, \cdots$ be iid R.V. with $E(X_i) = \mu$, $Var(X_i) = \sigma^2 < \infty$, then $\bar{X} \xrightarrow{a.s.} \mu$.

Converge in Distribution: $X_1, X_2, \cdots$ converge in distribution to a random variable if $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ for every point x at which $F(x)$ is continuous. It's a pointwise convergence of the distribution function to $F(x)$

a.s > p > d [Convergence Order]

Central Limit Theorem: Let $X_1, X_2 \cdots$ be iid R.V, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$, Define $Z_n = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$, and Let $Z \sim N(0,1)$, we have $Z_n \xrightarrow{D} Z$. i.e. $\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$

Slutsk's Theorem: If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} Y$ is a constant, then: $\begin{cases} X_n Y_n \xrightarrow{D} a \cdot X_n & (a \text{ is } Y) \\ X_n + Y_n \xrightarrow{D} X_n + a & (a \text{ is } Y) \end{cases}$

Student-t distribution: $f(t|k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(k/2)\sqrt{k\pi}} \left(\frac{1}{1 + t^2/k}\right)^{\frac{k+1}{2}}$ standard t-distribution.

Order Statistics: The order statistics for a random sample $X_1, X_2, \cdots, X_n$ arranged in order from smallest to largest. There are denoted: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. And the sample median M is defined in terms of order statistics: $M = \begin{cases} X_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2}, & n \text{ is even} \end{cases}$, And $X_{(n)} = \max\{X_1, X_2 \cdots X_n\}$ $X_{(1)} = \min\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$

pdf of order statistics: $f(x_{(j)}) = \frac{n!}{(j-1)!(n-j)!} f_X(x) F_{(x)}^{j-1} (1 - F(x))^{n-j}$

Delta Method: Let Y_n be a sequence of random variables s.t. $\sqrt{n}(Y_n - \theta) \xrightarrow{D} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta) \neq 0$, Then: $\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2 g'(\theta)^2)$

Properties of Statistic:

Parametric Family: A parametric family of distribution is a collection of distributions indexed by a finite-dim parameter space: eg: $\mathcal{F} = \{F(\cdot|\theta) : \theta \in \Theta\}$

Identifiable: A parametric family is identifiable if $\theta_1 \neq \theta_2 \Rightarrow F(\cdot|\theta_1) \neq F(\cdot|\theta_2)$. each distribution corresponds to a unique value of θ. Since the data depends only on distribution, θ is not estimable if multiple values of θ may corresponds to the same distribution.

Location - Scale Families:

Suppose $Z_i \overset{iid}{\sim} F_Z$, $i = 1, 2, \cdots, n$ and we observe $X_i = \mu + \sigma \cdot Z_i$. Then $F_X(x) = F_Z(\frac{x-\mu}{\sigma})$. Generally, one or both parameters are unknown and must be estimated. Also, $f_X(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$ by variable transformation formula.

Exponential Families:

The exponential family has a pdf or pmf with form: $\begin{cases} f_X(x|\theta) = h(x) \cdot c(\theta) \exp(w_1(\theta)t_1(x) + \cdots + w_k(\theta)t_k(x)) \\ \text{Also, the support of } f_X \text{ does not depend on } \theta. \end{cases}$

Eg: $f_X(x|p) = \binom{n}{x}(1-p)^{n-x} p^x = \exp(\log\binom{n}{x} + x\log p + (n-x)\log(1-p)) = \underbrace{\exp(\log\binom{n}{x})}_{h(x)} \cdot \underbrace{\exp(n\log(1-p))}_{c(\theta)} \cdot \exp(\underbrace{x}_{t(x)} \cdot \underbrace{\log\frac{p}{1-p}}_{w(\theta)})$

$f_X(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) = \exp(\log(\frac{\beta^\alpha}{\Gamma(\alpha)}) + (\alpha-1)\log(x) - \beta x)$

$= \underbrace{\exp(\log\frac{\beta^\alpha}{\Gamma(\alpha)})}_{c(\theta)} \underbrace{\cdot}_{h(x) \equiv 1} \exp(\underbrace{(\alpha-1)}_{w_1(\theta)} \underbrace{\log(x)}_{t_1(x)} \underbrace{-\beta}_{w_2(\theta)} \underbrace{x}_{t_2(x)})$

Sufficient Statistic: $T(x)$ is a sufficient summary of x ($x_i \overset{iid}{\sim} F_X(x|\theta)$) then, any (legitimate) inference about θ depends on the data only through $T(x)$. i.e. If x and y are two samples such that $T(x) = T(y)$, then inference about θ should be the same.

Def of Sufficient statistic: A statistic $T(x)$ is sufficient for θ if the conditional distribution of x given $T(x)$ does not depend on θ.

Corollary: If T is sufficient for θ, then for $\forall A \in \mathbb{R}^n$, $P(x \in A | T)$ does not depend on θ, and for any function $S(x)$, the conditional distribution of $S(x)$ given $T = t$ does not depend on θ.

Sufficiency Ratio:

If $f_X(x|\theta)$ is the sampling distribution, $f_T(t|\theta)$ is the pdf of $T(x)$, then $T(x)$ is a sufficient statistic for θ if for every x in the sample space, the ratio $\frac{f_X(x|\theta)}{f_T(T|\theta)}$ is a constant.

Factorization Theorem:

Let $f_X(x|\theta)$ be a joint pdf (pmf) of the data, then, $T(x) = t(x)$ is sufficient for θ iff

$f_X(x|\theta) = g(T(x)|\theta)h(x)$, $\forall x, \theta$, PS: $f_T(t|\theta) = q(t) g(t|\theta)$ for some function $q(t)$.

Corollary: If T is sufficient for θ and $a(t)$ is a 1-1 function, then $U = a(T)$ is a sufficient for θ.

For exponential families: $x_i \overset{iid}{\sim} f_X(x|\theta)$:

$$f(x) = h(x) c(\theta) \exp(w_1(\theta) t_1(x) + \cdots + w_k(\theta) t_k(x)) \Rightarrow \text{Then } T = \left(\textstyle\sum_{i=1}^n t_1(x_i), \cdots, \sum_{i=1}^n t_k(x_i)\right) \text{ is sufficient for } \theta.$$

Minimal Sufficient Statistics:

A statistic $T(x)$ is minimally sufficient if it's sufficient and it's a function of any other sufficient statistics.

i.e. U is sufficient for θ, $\Rightarrow$ $T = a(U)$ for some function $a(\cdot)$

Theorem: for some sufficient statistic $T(x)$, suppose $\frac{f_X(x|\theta)}{f_X(y|\theta)}$ is constant as a function of θ iff $T(x) = T(y)$

$\Rightarrow$ T is minimally sufficient for θ.

For exponential families: $x_i \overset{iid}{\sim} f_X(x|\theta)$

$$f(x|\theta) = h(x) \cdot c(\theta) \exp(w_1(\theta) t_1(x) + \cdots + w_k(\theta) t_k(x)) \Rightarrow \text{Then } T = \left(\textstyle\sum_{i=1}^n t_1(x_i), \cdots, \sum_{i=1}^n t_k(x_i)\right) \text{ is minimal sufficient for } \theta.$$

Ancillary: An ancillary statistic $R = R(X)$ is one whose distribution does not depend on θ. i.e. If we only observe R, then we cannot say anything about θ.

R does not have information about θ, but it still might have useful information. For example, the residuals, it can be used to check the validity of the model.

Location-Scale families: Let $x_i \overset{iid}{\sim} f_X(x|\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$, then $x_i = \mu + \sigma z_i$, where $z_i \overset{iid}{\sim} g(z)$

Consider the residuals: $\hat{z}_i = \frac{x_i - \bar{x}}{S_X} = \frac{\mu + \sigma z_i - \mu - \sigma \bar{z}_i}{\sigma \cdot S_Z} = \frac{z_i - \bar{z}}{S_Z}$ does not include μ and σ^2.

Completeness: A parametric family is complete if $E(g(x)) = 0$ for all θ $\Rightarrow$ $P(g(x) = 0) = 1$

A statistic T is complete if its family of distribution $\{F_T(\cdot|\theta) : \theta \in \Theta\}$ is complete.

Exponetial Families: Suppose $f_X(x|\theta) = h(x) c(\theta) \exp(w_1(\theta) t_1(\theta) + \cdots + w_k(\theta) t_k(\theta))$ and the range of $w(\theta) = (w_1(\theta), \cdots, w_k(\theta))$ includes an open set in $\mathbb{R}^k$, Then $T = (t_1(x), t_2(x), \cdots, t_k(x))$ is complete and minimal sufficient statistic for θ.

Basu's Theorem: If T is a complete sufficient statistic, then it is independent of any ancillary statistic

Also, let T be a sufficient statistic, and S be an ancillary statistic, if T and S are not independent

Then, T is not complete.

Likelihood: For a fixed sample of X, define the function $L(\theta|x) = f_X(x|\theta)$ for $\theta \in \Theta$. the likelihood function is $L(\theta|x)$. And $L(\theta|x)$ is a function of θ and X is given.

Sufficiency Principle. [Likelihood Principle]: If there are two experiments with likelihood $L_1(\theta)$ and $L_2(\theta)$. while $\frac{L_1(\theta,x)}{L_2(\theta,y)}$ is constant in θ, then the inference about θ should be the same.

Equivariance: Measurement equivariance prescribes that any inference made should not depend on the measurement scale that is used. Formal equivariance states that if the two inference problems have the same formal structure, i.e. mathematical model, then the same inference procedure should be used for both problems.

Equivariance principle: If $Y = g(X)$ is a change of measurement scale such that the model for Y has the same formal structure as the model for X. Then, inference procedure should be both measurement equivariant and formally equivariant.

Group: A set of functions $\{g(\underline{x}) : g \in G\}$, from sample space X to X, is called a group of functions if: For every $g \in G$, there is a $g' \in G$ such that $g(g'(x)) = X$ for $\forall x \in X$ [Inverse]

For every $g \in G$ and $g' \in G$, there exists $g'' \in G$ such that $g'(g(x)) = g(x)$ for $\forall x \in X$.

Invariance: Let G be a group of transformations of the sample space X and let:

$\mathcal{F} = \{f_X(x|\theta) ; \theta \in \Theta\}$ be a parametric family. Suppose for $\forall \theta \in \Theta$, and $g \in G$, there is a unique $\theta^* \in \Theta$ such that if $X \sim f_X(x|\theta)$, then $g(x) \sim f_X(x|\theta^*)$, then we say that $\mathcal{F}$ is invariant under group G.

Invariance Principle: If $f = \{f_X(x|\theta) ; \theta \in \Theta\}$ is invariant under G, then the appropriate inference about θ based on $T(x)$ should be invariant in the sense that $T(g(x)) = g^*(T(x))$

Method of Moments:

$\hat{m}_k = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^k$ is used to estimate $m_k = E(X^k)$, $\hat{m}_1 = \bar{X}$ is a Method of Moment estimator of $\mu = E(X)$, $\sigma^2 = Var(X) = m_2 - \hat{m}_1^2 \Rightarrow \hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}(\sum_{i=1}^{n}(x_i - \bar{x})^2) = \frac{n-1}{n} S^2$

If $\theta = g(m_1, m_2, \cdots, m_k)$ then a method of moment estimator of θ is $\hat{\theta} = g(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)$

Advantage of Method of Moment:

1. Generally easy to define, complete 2. Have a simple justification. 3. Basic properties are easy. 4. Very flexible, can be creative by using various transformations.

Disadvantage of Method of Moments:

1. May result in inappropriate results. 2. Many possibilities, not sure which to use. 3. Do not have to be a function of a minimal sufficient statistic.

Theorem: For exponential family $f_X(x|\theta) = h(x) \cdot c(\theta) \exp(w_1(\theta)t_1(x) + \cdots + w_k(\theta)t_k(x))$, suppose that Θ is an open subset of R^k, $w(\theta) = (w_1(\theta), w_2(\theta), \cdots, w_k(\theta))$ is 1-1 and differentiable, and

$$D = \left[\frac{\partial w}{\partial \theta}\right]_{k \times k} = \begin{pmatrix} \frac{\partial w_1}{\partial \theta_1} & \cdots & \frac{\partial w_k}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_1}{\partial \theta_k} & \cdots & \frac{\partial w_k}{\partial \theta_k} \end{pmatrix}$$

is inversible for all $\theta \in \Theta$. Let $T = (t_1(x), t_2(x), \cdots, t_k(x))$

Then, $E(T) = -[D(\theta)]^{-1}\left(\frac{\partial}{\partial \theta} \log c(\theta)\right)$

Maximum Likelihood Estimation:

Assume we have a parametric family $f(x|\theta)$, $\hat{\theta} = \hat{\theta}(x)$ is called an MLE if it maximizes $L(x|\theta)$ i.e. $L(\hat{\theta}, x) \geq L(\theta|x)$ for $\forall \theta \in \Theta$ and $\hat{\theta} \in \Theta$. $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, L(\theta, x)$. the most plausible value of θ is MLE. And $\hat{\theta}$ need not to be unique and it may be on the boundary of parameter space.

Theorem: An important feature of MLE is their invariance. If $g(\theta)$ is any function of θ, then the MLE of $g(\theta)$ is $g(\hat{\theta}_{MLE})$

Theorem: For exponential family, $f_X(x|\theta) = h(x)c(\theta)\exp(w_1(\theta)t_1(x) + \cdots + w_k(\theta)t_k(x))$, let $D(\theta) = [\frac{\partial w(\theta)}{\partial \theta}]_{k\times k}$ be invertible for all θ. let $G(\theta) = (\frac{\partial}{\partial \theta} \log c(\theta))$. ~~Then~~ The MLE for θ solves the k equations:

$E(t_j(x))\big|_{\theta=\hat{\theta}} = t_j(x)$. in iid case, $\hat{\theta}_{MLE}$ satisfies $E(t_j(x))\big|_{\theta=\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} t_j(x_i)$