Chapter 3: Facts about Common Statistical Models [*]

3 Facts about common statistical models

3.1 Bayes Models

- <mark>Probability Model on Data</mark> We have distributions $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ for X on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{P} \ll \mu$ ($\sigma-$ finite measure) and R -N derivatives

$$\frac{dP_\theta}{d\mu}(x) = f_\theta(x)$$

- <mark>Prior on Parameter</mark> We now add an assumption of a distribution G on $(\Theta, \mathcal{C})$ with $G \ll \nu(\sigma$ -finite measure) and R-N derivative

$$\frac{dG}{d\nu}(\theta) = g(\theta)$$

- <mark>Joint Distribution</mark> for (X, θ) : Here we consider $f_\theta(x)$ as a function of both x and θ (i.e., measurable in (x, θ)). If $f_\theta(x)$ is $\mathcal{B} \times \mathcal{C}$ -measurable, then there exists a joint probability distribution for (X, θ) on $(\mathcal{X} \times \Theta, \mathcal{B} \times \mathcal{C})$ defined, for $A \in \mathcal{B} \times \mathcal{C}$, by

$$\pi^{X,\theta}(A) \equiv P((X,\theta) \in A) = \int_A f_\theta(x) d(\mu \times G)(x,\theta) = \int f_\theta(x) g(\theta) d(\mu \times \nu)(x,\theta)$$

where

$$\frac{d\pi^{X,\theta}}{d(\mu \times G)} \equiv f_\theta(x), \quad \frac{d\pi^{X,\theta}}{d(\mu \times \nu)} \equiv f_\theta(x) g(\theta)$$

- <mark>Marginal Distributions</mark>

  - for X $(B \in \mathcal{B})$

$$\pi^X(B) \equiv P(X \in B) = \pi^{X,\theta}(B \times \Theta) = \int_{B \times \Theta} f_\theta(x) d(\mu \times G)(x,\theta) \overset{Fubini}{=} \int_B \left[\int_\Theta f_\theta(x) dG(\theta) \right] d\mu(x)$$

$$= \int_B \left[\int_\Theta f_\theta(x) g(\theta) d\nu \right] d\mu(x)$$

$$0 \leq \frac{d\pi^X(x)}{d\mu} = \int_\Theta f_\theta(x) dG(\theta) = \int_\Theta f_\theta(x) g(\theta)$$

  - for θ $(C \in \mathcal{C})$

$$\pi^\theta(C) \equiv P(\theta \in C) = \pi^{X,\theta}(\mathcal{X} \times C) = \int_{\mathcal{X} \times C} f_\theta(x) d(\mu \times G)(x,\theta) = \int_C \left[\int_\mathcal{X} f_\theta(x) d\mu(x) \right] dG(\theta) = G(C)$$

  Marginal distribution of θ is prior distribution G.

- <mark>Conditional distributions</mark>

  - for $X \mid \theta$

$$\pi^{X|\theta}(B \mid \theta) \equiv P_{X|\theta}(X \in B \mid \theta) = \int_B f_\theta(x) d\mu(x) = P_\theta(B), \quad B \in \mathcal{B}$$

$$\frac{d\pi^{X|\theta}(x)}{d\mu} = \frac{dP_\theta(x)}{d\mu} = f_\theta$$

- for $\theta \mid X$

$$\pi^{\theta|X}(C \mid x) \equiv P_{\theta|X}(\theta \in C \mid X = x) = \int_C \left[\frac{f_\theta(x)}{\int_\Theta f_\theta(x)dG(\theta)} \right] dG(\theta) = \int_C \frac{f_\theta(x)g(\theta)}{\int_\Theta f_\theta(x)g(\theta)d\nu(\theta)} d\nu(\theta)$$

$$\frac{d\pi^{\theta|X}(\theta)}{dG} = \frac{f_\theta(x)g(\theta)}{\int_\Theta f_\theta(x)dG(\theta)}, \quad \frac{d\pi^{\theta|X}(\theta)}{d\nu} = \frac{f_\theta(x)g(\theta)}{\int_\Theta f_\theta(x)g(\theta)d\nu(\theta)}, G \ll \nu$$

Note priors does not necessarily have a density, i.e. G is not necessarily dominated by some ν. But you can always write the density of posterior with respect to G.

3.2 Exponential Family of Distributions

- **Exponential family:** Definition 16 : $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma$ -finite measure) is an exponential family if, for some $h(x) \geq 0$, it holds that

$$f_\theta(x) \equiv \frac{dP_\theta}{d\mu}(x) = \exp\left(\alpha(\theta) + \sum_{i=1}^k \eta_i(\theta)T_i(x) \right) h(x), \quad x \in \mathcal{X}$$

for any $\theta \in \Theta$

- **Identifiable:** Definition 17 : A family of distributions, $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ is identifiable if $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$

- **Natural parameter space:** Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) \in \mathbb{R}^k$ and let

$$\Gamma \equiv \left\{ \boldsymbol{\eta} \in \mathbb{R}^k : \int_\mathcal{X} h(x) \exp\left(\sum_{i=1}^k \eta_i T_i(x) \right) d\mu(x) < \infty \right\}$$

note: this kernel with linear combinations of $T_i(X)$ using real numbers $\eta_i, i = 1, \ldots, k$.

Define a new distributional family

$$\mathcal{P}^* \equiv \left\{ P_\eta \text{ has R-N derivative as } f_\eta(x) \equiv \frac{dP_\eta}{d\mu}(x) = K(\boldsymbol{\eta})h(x) \exp\left(\sum_{i=1}^k \eta_i T_i(x) \right) : \boldsymbol{\eta} \in \Gamma \right\}$$

- $\mathcal{P} \subset \mathcal{P}^*$
- Γ is called the natural parameter space for $\mathcal{P}^*$ and Γ is a convex subset of $\mathbb{R}^k$
- If Γ lies in a subspace of dimension less than k, then $f_\eta(x)$ (and $f_\theta(x)$) can be re-written in a form involving fewer than k statistics $T_i(x)$. (We'll assume Γ to be fully k -dimensional.)
- $\mathcal{P}$ may be a proper subset of $\mathcal{P}^*$ or

$$\Gamma_\theta \equiv \left\{ (\eta_1(\theta), \eta_2(\theta), \ldots, \eta_k(\theta)) \in \mathbb{R}^k : \theta \in \Theta \right\}$$

can be a proper subset of Γ.

  * For example, for $f_\theta \propto \exp(\theta, -\theta^2)$,

$$\Gamma_\theta = \{(\theta, -\theta^2) : \theta \in \mathbb{R}\} \subset \Gamma \equiv \{(\eta_1, \eta_2) : \eta \in \mathcal{T}, \eta_2 < 0\}$$

  * The most useful results/theorems about the $\boldsymbol{\eta}$ -parameterization are the ones where Γ contains an open set, i.e. Γ is rich/big enough.
  * If we want to translate results about the $\boldsymbol{\eta}$ -parameterization to θ, then we want Γ_θ to contain an open set in $\mathbb{R}^k$.
  * To use the θ -parameterization, we must want $\boldsymbol{\eta}(\cdot)$ to be 1 -to-1 on Θ.
- Claim 19: The support of P_θ is

$$\{x \in \mathcal{X} : f_\theta(x) > 0\} = \{x \in \mathcal{X} : h(x) > 0\}$$

The distributions in $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ are mutually absolutely continuous.
- Claim 20: The statistic $T = (T_1, \ldots, T_k)$ is sufficient for the exponential family $\mathcal{P}$.
- Claim 21 : $T = (T_1, \ldots, T_k)$ has induced distributions $\{P_\theta^T : \theta \in \Theta\}$, where

$$P_\theta^T(B) = P_\theta(T \in B), \quad B \in \mathcal{B}\left(\mathbb{R}^k\right)$$

which is also an exponential family.

- Claim 22: If Γ_θ contains an *open rectangle* in $\mathbb{R}^k$, then $T = (T_1, \ldots, T_k)$ is complete for the exponential family $\mathcal{P}$.

- Claim 23: If Γ_θ contains an *open rectangle* in $\mathbb{R}^k$ (or under a much weaker assumption by Lehmann (1983)), then $T = (T_1, \ldots, T_k)$ is minimal sufficient for $\mathcal{P}$.

  Lehmann's Geometric Condition: If there exists $k+1$ points $v_0, \ldots, v_k \in \Gamma_\theta \subset \mathbb{R}^k$ *convex hull*
  $$\left\{ \sum_{i=0}^k p_i v_i, v_i \in \mathbb{R}^k, p_i \geq 0, \sum_{i=0}^k p_i = 1 \right\}$$
  contains an open set in $\mathbb{R}^k$ then T is minimally sufficient.

- Claim 24: If $g : \mathcal{X} \to \mathbb{R}$ is a measurable real-valued function with $E_\eta |g(X)| < \infty$ then
  $$E_{\boldsymbol{\eta}} g(X) = \int_{\mathcal{X}} g(x) f_{\boldsymbol{\eta}}(x) d\mu(x)$$
  is continuous on Γ and has continuous partial derivatives of all orders on the interior of Γ. Also,
  $$\frac{\partial^{\alpha_1 + \alpha_2 + \cdots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} E_{\boldsymbol{\eta}} g(X) = \int_{\mathcal{X}} g(x) \frac{\partial^{\alpha_1 + \alpha_2 + \cdots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} f_{\boldsymbol{\eta}}(x) d\mu(x)$$
  holds for $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{Z}_+ \equiv \{0, 1, 2, \ldots\}$
  It is okay to swap partials and expectations.

- Claim 25: Recall the exponential form $f_\eta(x) = K(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^k \eta_i T_i(x) \right) h(x)$ of densities in $\mathcal{P}^*$ where $K(\eta)$ is normalizing constant. If $\boldsymbol{\eta}_0, \boldsymbol{\eta}_0 + \boldsymbol{u} \in \Gamma$ for $\boldsymbol{u} = (u_1, \ldots, u_k)$, then the moment generating function of statistic $T(X)$ is
  $$E_{\boldsymbol{\eta}_0} \exp\left[u_1 T_1(X) + \cdots + u_k T_k(X) \right] = \frac{K(\boldsymbol{\eta}_0)}{K(\boldsymbol{\eta}_0 + \boldsymbol{u})}$$
  and the moments can be calculated by taking derivatives wrt u evaluated at $u = 0$.
  $$E_{\boldsymbol{\eta}_0} \left[T_1^{\alpha_1}(X) T_2^{\alpha_2}(X) \cdots T_k^{\alpha_k}(X) \right] = K(\boldsymbol{\eta}_0) \left. \frac{\partial^{\alpha_1 + \alpha_2 + \cdots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \cdots \partial \eta_k^{\alpha_k}} \frac{1}{K(\boldsymbol{\eta})} \right|_{\boldsymbol{\eta} = \boldsymbol{\eta}_0}$$

- Claim 26: If $X = (X_1, \ldots, X_n)$ with n iid components is such that $X_i \sim P_\theta$ (an exponential family distribution with k-dimensional statistic $T(X_i)$), then X generates a k-dimensional exponential family, say $\mathcal{P}^n \equiv \{P_\theta^n : \theta \in \Theta\}$ on $(\mathcal{X}^n, \mathcal{B}^n)$ with respect to μ^n. The k-dimensional statistic
  $$\sum_{i=1}^n T(X_i), \quad T(X_i) = (T_1(X_i), T_2(X_i), \ldots, T_k(X_i))$$
  is sufficient for this family $\mathcal{P}^n$. And $\sum_{i=1}^n T(X_i)$ is also complete if Γ_θ contains an open rectangle. Here Γ_θ is the parameter space with respect to P_θ.

- Example:
  1. $\mathcal{X} = \mathbb{R}$ and $f_\eta(x) \propto \exp\left(\eta_1 x - \eta_2 x^2 \right)$ for $\boldsymbol{\eta} = (\eta_1, \eta_2) \in \mathbb{R} \times (0, \infty)$
  2. $\mathcal{X} = \mathbb{R}$ and $f_\theta(x) \propto \exp\left(\theta x - \theta x^2 \right) \exp(\theta T_1(x) + \theta T_2(x))$ for $\boldsymbol{\theta} \in (0, \infty)$, where $T_1(x_1) = x, T_2(x) = -x^2$. Remark: $\Gamma_\theta = \{(\theta, \theta) : \theta > 0\}$ contains no open sets and we cannot expect to apply results for $\{P_\eta : \eta \in \Gamma\}$ to $\{f_\theta\}_{\theta > 0}$. This can be fixed by using another parameterization $f_\eta(x) \propto \exp(\eta_1 T_1(x)), T_1(x) = x - x^2, \eta > 0$, then $\Gamma = (0, \infty), \Gamma_\theta = (0, \infty)$ for f_θ as above.
  3. $\mathcal{X} = \mathbb{R}$ and $f_\theta(x) \propto \exp\left(\theta x - \theta^2 x^2 \right) = \exp(\theta T_1(x) + \theta^2 T_2(x))$ for $\boldsymbol{\theta} \in (0, \infty)$, where $T_1(x_1) = x, T_2(x) = -x^2$. Here $\Gamma_\theta \{(\theta, \theta^2) : \theta \neq 0\} \subset \mathbb{R}^2$ does not contain an open set in $\mathbb{R}^2$. In other words, we cannot find $f_\eta(x)$ having the same dimension as $f_\theta(x)$, i.e. $k = 2$ parametric functions (θ, θ^2) larger than $k = 1$ for $\theta \in \mathbb{R} \backslash \{0\}$.

- Curved Exponential Family: when the dimension of the parameterization is less than the dimension of natural parameter space. (need special theory).

3.3 Measures of Statistical Information

- Fisher Information Regularization Conditions: Definition 27 : $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma$-finite $)$ is **FI** regular at $\theta_0 \in \Theta \subset \mathbb{R}^k$ if there is an open neighborhood of θ_0, say O, such that

- (Support) $f_\theta(x) > 0$ for all $x \in \mathcal{X}$ and $\theta \in O$, where $\mathcal{X}$ is all the possible value X can take. We can use support $\{x \in \mathcal{X} : f_\theta(x) > 0\}$ by redefining $\mathcal{X} \equiv$ support.
- (Smoothness) for all x, $f_\theta(x)$ has first-order partial derivatives at θ_0; and
- (Local property: swapping condition) $1 = \int_{\mathcal{X}} f_\theta(x) d\mu(x)$ can be differentiated with respect to each component θ_i at θ_0 so that

$$0 = \int_{\mathcal{X}} \frac{\partial f_\theta(x)}{\partial \theta_i}\bigg|_{\theta_0} d\mu(x), \quad i = 1, \dots, k$$

- **Score function:** The random function of θ given by

$$\left(\frac{\partial \log f_\theta(X)}{\partial \theta_1}, \frac{\partial \log f_\theta(X)}{\partial \theta_2}, \dots, \frac{\partial \log f_\theta(X)}{\partial \theta_k} \right)$$

is called the score function. Note score function always has mean zero.

$$E_{\theta_0}\left(\frac{\partial \log f_\theta(x)}{\partial \theta_i}|_{\theta_0}\right) = E_{\theta_0}\left(\frac{1}{f_{\theta_0}(x)} \frac{\partial f_\theta(x)}{\partial \theta_i}|_{\theta_0}\right) = \int_{\mathcal{X}} \frac{1}{f_{\theta_0}(x)} \frac{\partial f_\theta(x)}{\partial \theta_i} f_{\theta_0}(x) d\mu(x) = 0$$

by the third condition of Fisher Information.

- **Fisher Information about θ contained in X at θ_0:** Definition 28: If $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma-$ finite $)$ is FI regular at $\theta_0 \in \Theta \subset \mathbb{R}^k$ and

$$E_{\theta_0}\left(\frac{\partial \log f_\theta(X)}{\partial \theta_i}\bigg|_{\theta_0} \right)^2 < \infty, \quad i = 1, \dots, k$$

then the $k \times k$ matrix

$$I(\theta_0) = \left[E_{\theta_0}\left(\frac{\partial \log f_\theta(X)}{\partial \theta_i}\bigg|_{\theta_0} \cdot \frac{\partial \log f_\theta(X)}{\partial \theta_j}\bigg|_{\theta_0} \right) \right]_{i,j}, \quad i, j = 1, \dots, k$$

is called the Fisher information about θ contained in X at θ_0.

- Claim 29:
  * Fisher information *does not* depend on the dominating measure μ. Suppose that $\mathcal{P} \ll \mu(\sigma$ -finite $)$ and $\mu \ll \nu$.

  $$\frac{dP_\theta}{d\nu} = \frac{dP_\theta}{d\mu} \frac{d\mu}{d\nu} = f_\theta(x) \frac{d\mu(x)}{d\nu} \implies \frac{\partial \log(\frac{dP_\theta(x)}{d\nu})}{\partial \theta_i} = \frac{\partial \log f_\theta(x)}{\partial \theta_i} + \frac{\partial \log(\frac{d\mu_\theta(x)}{d\nu})}{\partial \theta_i}$$

  where the second part does not depend on θ_0 and thus equals to zero.
  * Fisher Information $I(\theta_0)$ *does* depend on the parameterization. Suppose $\mathcal{P} \ll \mu(\sigma$ -finite $)$ is FI regular at $\theta_0 \in \mathbb{R}$ and let $\eta = h(\theta)$ for 1-to-1 and differentiable function $h : \Theta \to \mathbb{R}$ (so $\theta = h^{-1}(\eta)$). Define distributions $Q_\eta \equiv P_{h^{-1}(\eta)} = P_\theta \ll \mu$ (for η in the range of h) so that we have distributions/densities

  $$\begin{array}{cc} P_\theta & Q_\eta = P_{h^{-1}(\eta)} = P_\theta \\ f_\theta = \frac{dP_\theta}{d\mu} & g_\eta = \frac{dQ_\eta}{d\mu} = \frac{dP_{h^{-1}(\eta)}}{d\mu} = f_{h^{-1}(\eta)} = f_\theta \end{array}$$

  Then, we can compute the Fisher information in X about η at η_0 as

  $$I(\eta_0) = \left(\frac{1}{h'(h^{-1}(\eta_0))} \right)^2 J\left(h^{-1}(\eta_0) \right) = \left(\frac{1}{h'(\theta_0)} \right)^2 J(\theta_0)$$

  using the first order derivative h' and the Fisher information $J\left(h^{-1}(\eta_0) \right) = J(\theta_0)$ in X about θ at $\theta_0 = h^{-1}(\eta_0)$

- Theorem 30: Suppose $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma-$ finite $)$ is FI regular at $\theta_0 \in \Theta \subset \mathbb{R}^k$ using an open neighborhood O around θ_0. In addition, suppose $f_\theta(x)$ has continuous second order partial derivatives with respect to θ in the neighborhood O for all $x \in \mathcal{X}$ with

$$0 = \int_{\mathcal{X}} \frac{\partial f_\theta(x)}{\partial \theta_i}\bigg|_{\theta_0} d\mu(x), \quad i = 1, \dots, k$$

and

$$0 = \int_{\mathcal{X}} \frac{\partial^2 f_\theta(x)}{\partial \theta_i \partial \theta_j}\bigg|_{\theta_0} d\mu(x), \quad i, j = 1, \dots, k$$

4

Then, it holds that

$$I(\theta_0) = -\left[\mathrm{E}_{\theta_0}\left(\left.\frac{\partial^2 \log f_\theta(X)}{\partial \theta_i \partial \theta_i}\right|_{\theta_0}\right)\right]_{i,j}, \quad i,j = 1,\ldots,k$$

Note: For $k = 1$, the above says that

$$I(\theta_0) = \mathrm{E}_{\theta_0}\left(\left.\frac{d \log f_\theta(X)}{d\theta}\right|_{\theta_0}\right)^2 = -\mathrm{E}_{\theta_0}\left(\left.\frac{d^2 \log f_\theta(X)}{d\theta^2}\right|_{\theta_0}\right)$$

- Proposition 31: If $X_1, \ldots, X_n$ are independent with $X_i \sim P_{i,\theta}$, then $X = (X_1, \ldots, X_n)$ carries the Fisher information

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

where $I_i(\theta)$ is the Fisher information carried by $X_i, i = 1, \ldots, n$. Note

$$I(\theta) = Var\left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right) = Var\left(\frac{\sum_i \partial \log f_\theta(x_i)}{\partial \theta}\right) = \sum_i Var\left(\frac{\partial \log f_{\theta_i}(x)}{\partial \theta}\right) = \sum_i I_i(\theta)$$

- Proposition 32: Suppose $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma- \text{finite })$ is FI regular at $\theta_0 \in \Theta \subset \mathbb{R}^k$. If the function $T($ from $(\mathcal{X}, \mathcal{B})$ to $(\mathcal{T}, \mathcal{F}))$ is **1-to-1** then the Fisher information in $T(X)$ is the same as the Fisher information in $X : I_{T(X)}(\theta_0) = I_X(\theta_0)$

- Proposition 33: Suppose $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\} \ll \mu(\sigma \text{ -finite })$ is FI regular at $\theta_0 \in \Theta \subset \mathbb{R}^k$ and that $\mathcal{P}^T \equiv \{P_\theta^T : \theta \in \Theta\}$ (the set of distributions induced by a statistic $T(X)$) is FI regular at θ_0. Then, the $k \times k$ matrix

$$I_X(\theta_0) - I_{T(X)}(\theta_0)$$

is **non-negative definite**. Furthermore, if $T(X)$ **is sufficient**, then

$$I_X(\theta_0) = I_{T(X)}(\theta_0)$$

holds. Also, $I_X(\theta_0) = I_{T(X)}(\theta_0)$ holding for all θ_0 implies that $T(X)$ is sufficient. Note: The proof of this result is really hard and can be found with Theorem 2.86 of Schervish.

- Proposition 34: For an exponential family of distributions as in Claim 28 (i.e., exponential families in natural parameter space), it holds that

$$I(\eta_0) = \mathrm{Var}_{\eta_0}(T(X)) = \left[\left.\frac{\partial^2(-\log K(\eta))}{\partial \eta_i \partial \eta_j}\right|_{\eta_0}\right]_{i,j}, \quad i,j = 1,\ldots,k$$

where $T(X) = (T_1(X), \ldots, T_K(X))$ and

$$f_\eta(x) \equiv \frac{dP_\eta}{d\mu}(x) = K(\boldsymbol{\eta})h(x)\exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\right)$$

Proof: Note that $\log f_\eta(x) = \log K(\eta) + \sum_{i=1}^{k} \eta_i T_i(x) + \log h(x)$. and that

$$\frac{\partial \log f_\eta(x)}{\partial \eta_i} = \frac{\partial \log K(\eta)}{\partial \eta_i} + T_i(x) \overset{\text{Claim } 25}{=} -E_\eta T_i(X) + T_i(X)$$

- <mark>Kullback-Leibler Informaion</mark> Definition 35: If P and Q are probability measures on $(\mathcal{X}, \mathcal{B})$ with R -N derivatives p and q with respect to a dominating σ -finite measure μ, then the Kullback- Leibler information (KL divergence of Q from P) is the P -expected log-likelihood ratio

$$I(P,Q) = \mathrm{E}_P \log\left(\frac{p(X)}{q(X)}\right) = \int_\mathcal{X} \log\left(\frac{p(x)}{q(x)}\right) p(x)d\mu(x)$$

- The choice of μ is immaterial. One could use $\mu = P + Q$
- Non-support sets $N_{0,P} = \{x \in \mathcal{X} : p(x) = 0\} \& N_{0,Q} = \{x \in \mathcal{X} : q(x) = 0\}$ do not impact $I(P,Q)$ above; one can compute $I(P,Q) = \int_{\mathcal{X}^*} \log\left(\frac{p(x)}{q(x)}\right) p(x)d\mu(x)$ using $\mathcal{X}^* \equiv \{x \in \mathcal{X} : p(x) > 0, q(x) > 0\}$ for which $p(x)/q(x) \in (0, \infty)$ holds.
- Claim 36 : In general, $I(P,Q) \neq I(Q,P)$ holds.

- Claim 37 : $I(P,Q) \geq 0$ holds where $I(P,Q) = 0$ if and only if $P = Q$

- Claim 38: Suppose that P and Q are two probability measures on $(\mathcal{X}, \mathcal{B})$ with R-N derivatives p and q with respect to a dominating σ-finite measure μ. Assume $P_P(q(X) = 0) = 0$ and suppose further that $C : (0, \infty) \to \mathbb{R}$ is convex and that $\mathrm{E}_P C^- \left(\frac{q(X)}{p(X)} \right) < \infty$.
  Then $\mathrm{E}_P C \left(\frac{q(X)}{p(X)} \right) \geq C(m), \quad m \equiv \mathrm{E}_P \left(\frac{q(X)}{p(X)} \right) = \int_{\{x \in \mathcal{X}: p(x) > 0, q(x) > 0\}} q(x) d\mu(x) \in (0, 1]$
  If C is strictly convex, then equality holds above if and only if $q(X)/p(X)$ is degenerate under P (or equivalently $q(X)/p(X) = m$ holds a.s. P).
  Prove by Jensen's Inequality.

- Theorem 39: Suppose $X = (X_1, X_2)$ and that, under both distributions P^X and Q^X for X, it holds that X_1 and X_2 are independent. Then,

$$I_X \left(P^X, Q^X \right) = I_{X_1} \left(P^{X_1}, Q^{X_1} \right) + I_{X_2} \left(P^{X_2}, Q^{X_2} \right)$$

where P^{X_i} and Q^{X_i} denote the marginal distributions of $X_i (i = 1, 2)$ under P^X and Q^X, respectively.

- Theorem 40: Suppose that P and Q are probability measures for X on $(\mathcal{X}, \mathcal{B})$ and that $P^{T(X)}$ and $Q^{T(X)}$ are the induced probability measures for $T(X)$. Then,

$$I_X \left(P^X, Q^X \right) \geq I_{T(X)} \left(P^{T(X)}, Q^{T(X)} \right)$$

and there is equality if and only if $T(X)$ is sufficient for $\{P, Q\}$.

- **Theorem 41:** Under the Fisher Information RCs, if

$$\frac{\partial^2}{\partial \theta_i \theta_j} E_{\theta_0} \log f_\theta(X)|_{\theta = \theta_0} = E_{\theta_0} \left(\frac{\partial^2}{\partial \theta_i \theta_j} \log f_\theta(X)|_{\theta = \theta_0} \right)$$

then

$$\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} I(P_{\theta_0}, P_\theta)|_{\theta = \theta_0} \right)_{i,j=1,\dots,k} = I(\theta_0)$$

Note: Fisher's Information is a special form of the curvature of the Kullback-Leibler Information.