2 Fundamental of statistics

2.1 Populations, samples and models

2.1.1 Populations and samples

- Suppose that our data are realizations of a random vector $(X_1, \ldots, X_n)$ on a probability space $(\Omega, \mathcal{F}, P)$. Then $(X_1, \ldots, X_n)$ is called a sample.
- Suppose that we have a sample $(X_1, \ldots, X_n)$ and X_i's are IID. Then we say $(X_1, \ldots, X_n)$ is a random sample.

2.1.2 Parametric models

- Suppose that for every $\theta \in \Theta$, P_θ is a probability measure and all P_θ's are on the same measurable space. Then $\{P_\theta : \theta \in \Theta\}$ is called a family (of probability measures).
- Example 1. $\{N(\mu, \sigma^2) : \mu \in R, \sigma > 0\}$ is a parametric family.
- Example 2. Let λ denote the Lebesgue measure on $(R, \mathcal{B})$ and let $S = \{f : f \geq 0,\ f \text{ is even and continuous, and } \int f d\lambda = 1.\}$. For $f \in S$, let P_f denote the probability measure defined by $P_f(A) = \int_A f d\lambda$ for $A \in \mathcal{B}$. Then $\{P_f : f \in S\}$ is a nonparametric family.
- A family $\{P_\theta : \theta \in \Theta\}$ is identifiable if $P_{\theta_1} = P_{\theta_2}$ implies that $\theta_1 = \theta_2$.
- Example 3. The family $\{N(\mu + c, 1) : (\mu, c) \in R^2\}$ is not identifiable.

2.1.3 Exponential and location scale families

- Definition. Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measure and there exists a measure ν such that for every $\theta \in \Theta$, P_θ is absolute continuous with respect to ν. Then ν is called a dominating measure of $\mathcal{P}$.
- Definition of an exponential family (Definition 2.2). Suppose that $\Theta \subset R^d$. A parametric family $\{P_\theta : \theta \in \Theta\}$ with a σ-finite dominating measure ν is called an exponential family if for every $\theta \in \Theta$,

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\Big(\eta(\theta)^t T(\omega) - \xi(\theta)\Big)\, h(\omega), \tag{1}$$

where T is a $p \times 1$ random vector and $\exp(\xi(\theta)) = \int \exp(\eta(\theta)^t T) h d\nu$.

- Natural exponential family. Consider the exponential family with p.d.f. given in (1). By the reparameterization $\eta = \eta(\theta)$, we have an exponential family $\{P_\eta : \eta \in \Xi\}$, where $\Xi = \{\eta(\theta) : \theta \in \Theta\}$,

$$\frac{dP_\eta}{d\nu}(\omega) = \exp\left(\eta^t T(\omega) - \zeta(\eta)\right) h(\omega), \tag{2}$$

and $\exp(\zeta(\eta)) = \int \exp(\eta^t T) h d\nu$. The family $\{P_\eta : \eta \in \Xi\}$ is called the natural exponential family and η is called the natural parameter.

- Properties of a natural exponential family (modified version of Theorem 2.1). Consider the natural exponential family with p.d.f. given in (2). Let $T^t = (Y^t, U^t)$ and $\eta^t = (\vartheta^t, \varphi^t)$, where Y and ϑ are of the same dimension. Suppose that $P_T = P_\eta \circ T^{-1}$. Then we have the following results.

  (i) Y has the p.d.f.
  $$f_\eta(y) = \exp(\vartheta^t y - \zeta(\eta))$$
  with respect to a σ-finite measure depending on φ.

  (ii) The conditional distribution of Y given $U = u$ has the p.d.f.
  $$f_{Y|U}(y|u) = \exp(\vartheta^t y - \zeta_u(\vartheta))$$
  with respect to a σ-finite measure depending on u.

  (iii) Suppose that η_0 is an interior point of the natural parameter space Ξ. Then the m.g.f. of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0. Denote the m.g.f. by ψ_{η_0}. Then
  $$\psi_{\eta_0}(t) = \exp(\zeta(\eta_0 + t) - \zeta(\eta_0)).$$

  (iv) Suppose that there exists I: an open set in R^p such that $I \subset \Xi$ and $\int |f| dP_\eta < \infty$ for every η in I. Then for every $\eta_0 \in I$, $k \in \{1, \ldots, p\}$,
  $$\frac{\partial}{\partial \eta_k}\left(\int f e^{\eta^t T} h d\nu\right)\Big|_{\eta=\eta_0} = \int \frac{\partial}{\partial \eta_k}\left(f e^{\eta^t T} h\right)\Big|_{\eta=\eta_0} d\nu = \int f T_k e^{\eta_0^t T} h d\nu,$$
  where T_k is the k-th component of T. Also, $\int |f T_k| dP_\eta < \infty$ for every η in I.

– The proofs of (i) and (ii) are based on the fact that $P_\eta \circ (Y, U)^{-1}$ has the p.d.f.

$$f_\eta^*(y, u) = \exp((\vartheta - \vartheta_0)^t y + (\varphi - \varphi_0)^t u - \zeta(\eta) + \zeta(\eta_0))$$

with respect to $P_{\eta_0} \circ (Y, U)^{-1}$ for every $\eta_0 \in \Xi$.

– The proof of (iv) is based on the dominated convergence theorem (Example 1.8 in Section 1.2).

- Location-scale families. Definition 2.3 in Section 2.1.3.

  – Suppose that X is a $k \times 1$ random vector with distribution P, then the distribution of the random vector $\mu + \Sigma^{1/2} X$ is $P_{(\mu,\Sigma)}$.

2.2 Statistics, sufficiency and completeness

2.2.1 Statistics and their distributions

- Suppose that X is a sample that is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. A statistic is a function defined on $\mathcal{X}$ and is measurable from $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ to some measurable space $(\Lambda, \mathcal{G})$, where $\mathcal{G}$ contains all singletons in Λ.

- Suppose that $X = (X_1, \ldots, X_n)$ is a sample. The following are some common statistics.

  – $\overline{X} = (X_1 + \cdots + X_n)/n$

  – $S = \sqrt{(n-1)^{-1} \sum_{k=1}^n (X_k - \overline{X})^2}$.

  – $X_{(k)}$: the k-th order statistic, $1 \le k \le n$.

- Suppose that X is a random sample. Then the asymptotic distribution of $(\overline{X}, S^2)$ can be obtained from the multivariate CLT and delta method.

- (Example 2.9 in Section 2.2.1) Suppose that X is a random sample from a distribution with Lebesgue p.d.f. f. Then the joint distribution of $(X_{(1)}, \ldots, X_{(n)})$ has a p.d.f. f_n with respect to the Lebesgue measure on $(R^n, \mathcal{B}^n)$, where

$$f_n(x_1, \ldots, x_n) = \begin{cases} n! f(x_1) \cdots f(x_n) & \text{if } x_1 < \cdots < x_n; \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2 Suffciency and minimal suffciency

- Suppose that X is a sample and P_X is in a family $\mathcal{P}$. Suppose that $T = T(X)$ is a statistic. For every $P \in \mathcal{P}$, let $P_{X|T=t,P}$ denote the conditional distribution of X given $T = t$ when $P_X = P$. Then T is sufficient for $P \in \mathcal{P}$ means that $P_{X|T=t,P}$ does not depend on P.

  - Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then T is sufficient for $\theta \in \Theta$ if and only if the conditional distribution of X given $T = t$ when $P_X = P_\theta$ does not depend on θ.

- The factorization theorem (Theorem 2.2). Suppose that X is a sample and P_X is in a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with a σ-finite dominating measure ν. Suppose that T is a statistic. Then $T = T(X)$ is sufficient for θ if and only if there exist nonnegative measurable functions g_θ and h such that
$$\frac{dP_\theta}{d\nu}(x) = g_\theta(T(x))h(x). \tag{3}$$

- The proof of the factorization theorem uses the following fact (Lemma 2.1).

  Fact 1. Suppose that $\mathcal{P}$ is a family with a σ-finite dominating measure ν. Then there exists nonnegative constants c_i's and measures P_i's in $\mathcal{P}$ such that $\sum_{i=1}^\infty c_i = 1$ and $\mathcal{P}$ is dominated by the probability measure $Q = \sum_{i=1}^\infty c_i P_i$.

  Proof of Fact 1. Suppose that ν is a finite measure. Let
$$\mathcal{P}_0 = \left\{ \sum_{i=1}^\infty c_i P_i : \text{ for each } i,\ c_i \geq 0 \text{ and } P_i \in \mathcal{P}, \text{ and } \sum_{i=1}^\infty c_i = 1. \right\}$$
  and
$$\mathcal{C} = \left\{ C : \frac{dP}{d\nu} > 0 \ \nu\text{-a.e. on } C \text{ for some } P \in \mathcal{P}_0 \text{ and } \nu(C) > 0 \right\}.$$
  Let $\{C_k\}$ be a sequence in $\mathcal{C}$ such that $\nu(C_k) \to \sup_{C \in \mathcal{C}} \nu(C)$ as $k \to \infty$. Let P_k be an element in $\mathcal{P}_0$ such that $dP_k/d\nu > 0$ ν-a.e. on C_k. Let $\{a_k\}$ be a sequence of positive numbers such that $\sum_k a_k = 1$ and let $Q = \sum_k a_k P_k$. Then $Q \in \mathcal{P}_0$ and the following claims are true.

Claim 1. $\mathcal{C}$ is closed under the finite union operation.

Claim 2. Q is a dominating measure for $\mathcal{P}$.

Proof of Claim 2. Suppose that $Q(A) = 0$. For $P \in \mathcal{P}$, let $D = \{dP/d\nu > 0\} \cap A$. Then we have the following.

(i) For every k, $P_k(A) = 0 \Rightarrow P_k(C_k \cap D) = 0 \Rightarrow \nu(C_k \cap D) = 0$.

(ii) Suppose that $\nu(D) > 0$. Then $D \in \mathcal{C}$ and for every k, $C_k \cup D \in \mathcal{C}$. From (i), $\nu(C_k \cup D) = \nu(C_k) + \nu(D)$. Let $k \to \infty$, then we have $\lim_{k\to\infty} \nu(C_k \cup D) = \sup_{C\in\mathcal{C}} \nu(C) + \nu(D)$. Since $\lim_{k\to\infty} \nu(C_k \cup D) \leq \sup_{C\in\mathcal{C}} \nu(C)$, we have $\nu(D) = 0$, which contradicts the assumption that $\nu(D) > 0$.

From (ii), we can conclude that $\nu(D) = 0$, which implies that $P(A) = 0$. Thus Claim 2 holds true.

In the above proof, it is assumed that ν is a finite measure. If ν is σ-finite but not finite, some modification is needed by considering A_i's such that $\Omega = \cup_{i=1}^{\infty} A_i$ and $\nu(A_i) < \infty$ for each i.

- The proof of factorization theorem. Let Q be the $\sum_{i=1}^{\infty} c_i P_{\theta_i}$ guaranteed by Fact 1. Let X_0 be a random vector such that $P_{X_0} = Q$ and let $T_0 = T(X_0)$. Then the joint p.d.f. of (X, T) with respect to P_{X_0,T_0} is

$$\frac{dP_\theta}{dQ}(x) = \frac{dP_\theta}{dQ}(x) I_{\{T(x)\}}(t) \quad P_{X_0,T_0}\text{-a.e..}$$

To prove the "if" direction, suppose that (3) holds. Apply the chain rule, and we have

$$\frac{dP_\theta}{dQ}(x) = \frac{\frac{dP_\theta}{d\nu}(x)}{\frac{dQ}{d\nu}(x)} = \frac{g_\theta(T(x))}{\sum_{i=1}^{\infty} c_i g_{\theta_i}(T(x))} \; Q\text{-a.e.,}$$

so the p.d.f. of (X, T) with respect to P_{X_0,T_0} is

$$\frac{g_\theta(T(x))}{\sum_{i=1}^{\infty} c_i g_{\theta_i}(T(x))} = \frac{g_\theta(t)}{\sum_{i=1}^{\infty} c_i g_{\theta_i}(t)} \; P_{X_0,T_0}\text{-a.e.}$$

and the conditional p.d.f. of X given $T = t$ with respect to $P_{X_0|T_0=t}$ is 1. Therefore, T is sufficient for θ.

To prove the "only if" direction, suppose that T is sufficient for θ. Then, there exists μ_t such that μ_t does not depend on θ and $\mu_t = P_{X|T=t}$. It can be shown that

$$\int \mu_t(A) I_B(t) dP_{T_0}(t) = P_{X_0,T_0}(A \times B), \tag{4}$$

so $\mu_t = P_{X_0|T_0=t}$ and the conditional p.d.f. of X given $T = t$ with respect to $P_{X_0|T_0=t}$ is 1. Let $f_{X,T}(x,t) = \dfrac{dP_\theta \circ T^{-1}}{dQ \circ T^{-1}}(t)$, then when $P_X = P_\theta$, it can be shown that

$$\int_{A \times B} f_{X,T}(x,t) dP_{X_0,T_0}(x,t) = P_X(A \cap T^{-1}(B))$$

and $f_{X,T}$ is the p.d.f. for (X,T) w.r.t. P_{X_0,T_0}. The p.d.f of X with respect to Q is then

$$\int f_{X,T}(x,t) dP_{T_0|X_0=x}(t) = \int \frac{dP_\theta \circ T^{-1}}{dQ \circ T^{-1}}(t) dP_{T_0|X_0=x}(t) = \frac{dP_\theta \circ T^{-1}}{dQ \circ T^{-1}}(T(x))$$

and the p.d.f. of X with respect to ν is

$$\frac{dP_\theta}{dQ}(x) \times \frac{dQ}{d\nu}(x) = \frac{dP_\theta \circ T^{-1}}{dQ \circ T^{-1}}(T(x)) \frac{dQ}{d\nu}(x).$$

Thus (3) holds with $g_\theta(T(x)) = \dfrac{dP_\theta \circ T^{-1}}{dQ \circ T^{-1}}(T(x))$ and $h = dQ/d\nu$.

It remains to prove (4). Since $P_{T_0} = Q \circ T^{-1} = \sum_i c_i P_{\theta_i} \circ T^{-1}$, we have

$$\begin{aligned} \int \mu_t(A) I_B(t) dP_{T_0}(t) &= \int \mu_t(A) I_B(t) dQ \circ T^{-1}(t) \\ &= \sum_i c_i \int \mu_t(A) I_B(t) dP_{\theta_i} \circ T^{-1}(t) \\ &= \sum_i c_i P_{\theta_i}(A \cap T^{-1}(B)) = P_{X_0,T_0}(A \times B) \end{aligned}$$

and (4) holds.

- Minimum sufficiency. Suppose that X is a sample measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and P_X is in a family $\mathcal{P}$. Suppose that T is a sufficient statistic for $P \in \mathcal{P}$. If for every S that is a sufficient statistic for $P \in \mathcal{P}$, there exist a function ψ and and a set $A \in \mathcal{B}_\mathcal{X}$ such that $T(x) = \psi(S(x))$ for $x \in A$ and $P(A) = 1$ for every $P \in \mathcal{P}$. Then T is called a minimum sufficient statistic for $P \in \mathcal{P}$.

- Finding minimum sufficient statistics. Theorem 2.3
  - Theorem 2.3 (iv). Suppose that $\mathcal{P}$ has a σ-finite dominating measure ν and let $f_P = dP/d\nu$ for $P \in \mathcal{P}$. Suppose that $\mathcal{P}_0$ is a countable sub-collection of $\mathcal{P}$ such that $\mathcal{P}_0$-a.s. implies that $\mathcal{P}$-a.s. and A is a Borel set such that $P(A) = 1$ for every $P \in \mathcal{P}_0$. For $x \in A$, let $D(x)$ be the collection of points $y \in A$ such that there exists some measurable function ϕ such that

    $$f_P(x) = f_P(y)\phi(x, y) \text{ for every } P \in \mathcal{P}_0.$$

    Suppose that there exists a statistic $T = T(X)$ that is sufficient for $\mathcal{P}$ such that for x, $y \in A$,

    $$y \in D(x) \Rightarrow T(x) = T(y),$$

    then T is minimal sufficient for $\mathcal{P}$.
  - Proof of Theorem 2.3 (iv). Suppose that T is measurable from $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ to $(\Lambda_T, \mathcal{G}_T)$, where $\mathcal{G}_T$ contains all singletons in Λ_T. Suppose that S is a sufficient statistic for $\mathcal{P}_0$. Then by the factorization theorem, there exist measurable functions g_P and h such that

    $$f_P(x) = g_P(S(x))h(x) \ \nu\text{-a.e. for every } P \in \mathcal{P}_0.$$

    Let

    $$A_1 = \{x \in A : h(x) > 0 \text{ and } f_P(x) = g_P(S(x))h(x) \text{ for every } P \in \mathcal{P}_0 \ \}.$$

    Then $P(A_1^c) = 0$ for every $P \in \mathcal{P}_0$. For x, $y \in A_1$, $S(x) = S(y)$ implies that $f_P(x) = f_P(y)h(x)/h(y)$ and $y \in D(x)$, so $T(x) = T(y)$. Therefore, $T(x) = \psi(S(x))$ for $x \in A_1$ for some function ψ and T is minimal sufficient for $\mathcal{P}_0$. By Theorem 2.3 (i), T is minimal sufficient for $\mathcal{P}$.

- Example 4. Suppose that $X = (X_1, X_2)$ and X_1 and X_2 are independent normal random variables with $E(X_i) = \mu_i$ and $Var(X_i) = 1$ for $i = 1, 2$. Suppose that the range for (μ_1, μ_2) is R^2. Then (X_1, X_2) is minimum sufficient for (μ_1, μ_2).

2.2.3 Complete statistics

- Ancillary statistic: a statistic S is ancillary if the distribution of S does not depend on P (or θ).
- Suppose that $T = (U, V)$ is a sufficient statistic and V is ancillary. It is not necessary that U is sufficient.
  - Example. Consider a random sample from $U(\theta, \theta + 1)$ (Example 2.13)
- Completeness.
  - A statistic T is complete if $E(g(T)) = 0$ for all P implies that $g(T) = 0$ $\mathcal{P}$-a.e..
  - A statistic T is boundedly complete if for every bounded function g, $E(g(T)) = 0$ for all P implies that $g(T) = 0$ $\mathcal{P}$-a.e..
- Suppose that T is complete, then $g(T)$ cannot be an ancillary statistic unless $g(T)$ is constant.
- Suppose that T is sufficient and boundedly complete, and takes values in R^k. Then T is minimum sufficient.
  - Proof. Suppose that $T = (T_1, \ldots, T_k)$ and let $T^* = (T_1^*, \ldots, T_k^*)$, where $T_j^* = e^{T_j}/(1 + e^{T_j})$ for $1 \le j \le k$. Then T^* is a one-to-one function of T and T^* is measurable. For a sufficient statistic S and $1 \le j \le k$, consider

    $$h_j(T) = E(E(T_j^*|S)|T) - T_j^*,$$

    then $h_j(T)$ is a bounded function of T and $Eh_j(T) = 0$, which implies that $h_j(T) = 0$ $\mathcal{P}$-a.e. since T is boundedly complete. Note that $T_j^* = E(E(T_j^*|S)|T)$ $\mathcal{P}$-a.e. implies that $T_j^* = E(T_j^*|S)$ $\mathcal{P}$-a.e., so $T^* = (E(T_1^*|S), \ldots, E(T_k^*|S))$ $\mathcal{P}$-a.e. and $T = \psi(S)$ $\mathcal{P}$-a.e. for some measurable function $\psi(S)$. Let $A = \{x : T(x) = \psi(S(x))\}$, then $T(x) = \psi(S(x))$ for $x \in A$ and $P(A) = 1$ for every $P \in \mathcal{P}$, so T is minimal sufficient for $\mathcal{P}$.
  - Example (Proposition 2.1). Consider a random sample from an exponential family.

  - Example (Example 2.16). Consider a random sample from $U(0, \theta)$.

- A minimum sufficient statistic is not necessarily complete (Example: $U(\theta, \theta + 1)$).

- Basu's Theorem (Theorem 2.4). Suppose that T is a complete and sufficient statistic and V is ancillary. Then T and V are independent.

  - Proof. Consider $g(T) = E(I_A(V)|T) - E(I_A(V))$. Then $E(g(T)) = 0$ for all $P \in \mathcal{P}$ and $g(T) = 0$, which implies the independence of T and V.
  - Example (Example 2.18). Consider a random sample from $N(\mu, \sigma^2)$. Then $\bar{X}$ and S^2 are independent.

2.3 Statisitical decision theory

2.3.1 Decision rules, loss functions, and risks

- A decision problem is characterized by an action space Ω_2, a family $\mathcal{P}$ for the distribution of the sample, and a loss function L from $\mathcal{P} \times \Omega_2$ to $[0, \infty)$.

- A decision rule is a function from the range of the sample to the action space.

- Suppose that δ is a decision rule and L is the loss function. Then $E(L(P, \delta(X)))$ is called the risk function of δ under loss L. $E(L(P, \delta(X)))$ is often denoted by $R_\delta(P)$ (or $R_\delta(\theta)$).

- Example 5. Consider the problem of estimating θ based on X: a random sample from $N(\theta, 1)$. The action space is R. Consider the loss function $L(P, a) = (a - \theta)^2$. Then for a decision rule δ, its risk is $E(\delta(X) - \theta)^2$.

- A randomized decision rule is characterized by a random probability measure $\delta(X)$, where $\delta(x)$ repesents the conditional distribution of a random action Y given $X = x$.

- For a randomized decision rule $\delta(X)$ that represents a random action Y ($P_{Y|X=x} = \delta(x)$), its risk function $R_\delta(P)$ is $E(L(P, Y)) = E(E(L(P, Y)|X)) = \int \int L(P, a) d\delta(x)(a) dP_X(x)$.

- Example 6. Consider the problem of testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ based on X: a random sample from $N(\theta, 1)$. The action space is $\{0, 1\}$, where 1 respents the action of rejecting H_0. Consider the loss function

$$L(P, a) = \begin{cases} 0 & \text{if } a = 1 \text{ and } \theta > 0 \text{ or } a = 0 \text{ and } \theta \leq 0; \\ 1 & \text{if } a = 0 \text{ and } \theta > 0 \text{ or } a = 1 \text{ and } \theta \leq 0. \end{cases}$$

Consider the randomized rule that rejects H_0 with probability 0.9 when $\bar{X} > 1$ and rejects H_0 with probability 0.5 when $\bar{X} \leq 1$. Then the rule is characterized by $\delta(X)$, which is defined by

$$\delta(X)(A) = \begin{cases} 0.9d_{\{1\}}(A) + 0.1d_{\{0\}}(A) & \text{if } \bar{X} > 1; \\ 0.5d_{\{0\}}(A) + 0.5d_{\{1\}}(A) & \text{if } \bar{X} \leq 1, \end{cases}$$

where $d_{\{a\}}$ is the probability measure that puts all its mess on the point a. Note that $\delta(x) = P_{Y|X=x}$, where $P_{Y|X=x}$ is $Ber(0.9)$ if $\bar{x} > 1$ and $P_{Y|X=x}$ is $Ber(0.5)$ if $\bar{x} \leq 1$.

The risk of δ is $E(L(P, Y)) = \int \int L(P, a)d\delta(x)(a)dP_X(x)$, where

$$\int L(P, a)d\delta(x)(a) = E(L(P, Y)|X = x) = \begin{cases} 0.9L(P, 1) + 0.1L(P, 0) & \text{if } \bar{x} > 1; \\ 0.5L(P, 0) + 0.5L(P, 1) & \text{if } \bar{x} \leq 1. \end{cases}$$

Therefore,

$$\begin{aligned} E(L(P, Y)) &= E(E(L(P, Y)|X)) \\ &= E(0.9L(P, 1) + 0.1L(P, 0))I_{(1,\infty)}(\bar{X})) \\ &\quad +E((0.5L(P, 0) + 0.5L(P, 1))I_{(-\infty,1]}(\bar{X})) \\ &= \begin{cases} E(0.9I_{(1,\infty)}(\bar{X}) + 0.5I_{(-\infty,1]}(\bar{X})) & \text{if } \theta \leq 0; \\ E(0.1I_{(1,\infty)}(\bar{X}) + 0.5I_{(-\infty,1]}(\bar{X})) & \text{if } \theta > 0, \end{cases} \\ &= \begin{cases} 0.9 - 0.4\Phi(\sqrt{n}(1 - \theta)) & \text{if } \theta \leq 0; \\ 0.1 + 0.4\Phi(\sqrt{n}(1 - \theta)) & \text{if } \theta > 0, \end{cases} \end{aligned}$$

where Φ is the cdf for $N(0, 1)$.

2.3.2 Admissibility and optimality

- We say a decision rule δ_1 is as good as a decision rule δ_2 if $R_{\delta_1}(P) \leq R_{\delta_2}(P)$ for all $P \in \mathcal{P}$.

- Suppose that a decision rule δ_1 is as good as a decision rule δ_2 and $R_{\delta_1}(P) < R_{\delta_2}(P)$ for some $P \in \mathcal{P}$, then we say that δ_1 is better than δ_2.

- A decision rule δ is admissible if no decision rule is better than δ.

- A decision rule is optimal if it is as good as any other decision rule. However, sometimes it is impossible to find an optimal rule.

  Example 7. Consider the problem of estimating θ based on X: a random sample from $N(\theta, 1)$ under the square error loss $L(P, a) = (a - \theta)^2$. There is no optimal decision rule.

- To obtain optimal rules, there are two types of approaches:

  (i) Changing the criteria. Examples: minimizing Bayes risk $\int R_\delta(P) d\pi(P)$ or the max risk $\sup_P R_\delta(P)$.

  (ii) Considering a sub-collection of decision rules. Examples: unbiased estimators or invariant decision rules.

- Suppose that $\mathcal{T}$ is a collection of some decision rules. A decision rule δ in $\mathcal{T}$ is $\mathcal{T}$-admissible if no decision rule in $\mathcal{T}$ is better than δ. A decision rule δ in $\mathcal{T}$ is $\mathcal{T}$-optimal if δ is as good as any other decision rule in $\mathcal{T}$.

- Invariance. Suppose that X is a sample and $P_X \in \mathcal{P}$.

  – Group of transformations. Let $\mathcal{G}$ be a collection of transformations of X. $\mathcal{G}$ is called a group if it is closed under composition and for each transformation g in $\mathcal{G}$, g^{-1} is also in $\mathcal{G}$.

    Example 8. Let $g_c(x) = x + c\mathbf{1}$ for $c \in R$ and $x \in R^n$, where $\mathbf{1}$ is the vector of 1's in R^n. Then $\mathcal{G} = \{g_c : c \in R\}$ is a group of transformations.

  Supppose that $\mathcal{G}$ is a group of transformations.

  – The family $\mathcal{P}$ is invariant under $\mathcal{G}$ means that for every $g \in \mathcal{G}$ and $P \in \mathcal{P}$, $P \circ g^{-1} \in \mathcal{P}$.

- A loss function is invariant under $\mathcal{G}$ means that for every $g \in \mathcal{G}$, $P \in \mathcal{P}$ and $a \in \Omega_2$ (action space), there exists an action $h(a, g)$ such that $L(P, a) = L(P \circ g^{-1}, h(a, g))$.
- A decision problem is invariant under a group of transformations $\mathcal{G}$ means that the family $\mathcal{P}$ is invariant under $\mathcal{G}$ and the loss function is invariant under $\mathcal{G}$.
- For a decision problem that is invariant under $\mathcal{G}$, a decision rule $\delta(X)$ is invariant under $\mathcal{G}$ means that $\delta(g(X)) = h(\delta(X), g)$, where h satisfies $L(P, a) = L(P \circ g^{-1}, h(a, g))$.

Example 9. Suppose that X is a random sample from $N(\theta, 1)$. Consider the problem of estimating θ under square error loss. Then $\bar{X}$ is an invariant decision rule under the location transformation group, and an invariant decision rule for this decision problem is of the form $\bar{X} + D_0(X_1 - X_n, \ldots, X_{n-1} - X_n)$ for some function D_0. In addition, the optimal invariant decision rule for this decision problem is $\bar{X}$.

- Two decision rules δ_1 and δ_2 are equivalent if they have the same risk function.
- Proposition 2.2. Suppose that the action space is a subset of R^k, δ is a decision rule with finite risk, and $T = T(X)$ is a sufficient statistic. Then there exists a randomized decision rule δ_1 that is based on T such that δ_1 is equivalent to δ.

  Proof. Suppose that Y is a random action such that $P_{Y|X=x} = \delta(x)$. Define

  $$\delta^*(X)(A) = E[\delta(X)(A)|T] = E(E[I_A(Y)|X]|T) = E[I_A(Y)|T]$$

  for every A. Suppose that Y^* is a random action such that $P_{Y^*|X=x} = \delta^*(x)$. Then $P_{Y^*|X=x} = P_{Y|T=T(x)}$, so $E[L(P, Y)|T] = E[L(P, Y^*)|X]$ and $R_\delta(P) = EL(P, Y) = EL(P, Y^*) = R_{\delta^*}(P)$.
- Theorem 2.5. Suppose that the action space Ω_2 is a convex subset in R^k and the loss function $L(P, a)$ is convex in a when P is fixed. Then (i) and (ii) hold.

(i) Suppose that δ is a randomized rule and $\int_{\Omega_2} \|a\| d\delta(x)(a) < \infty$ for each x in the range of X. Let $T_1(x) = \int a d\delta(x)(a)$. Then $L(P, T_1(x)) \leq \int_{\Omega_2} L(P, a) d\delta(x)(a)$.

(ii) Rao-Blackwell Theorem. Suppose that T_0 is a non-randomized rule such that $E(\|T_0\|) < \infty$. Let $T_1 = E(T_0|T)$, then T_1 is as good as T_0.

- Jensen's inequality. Suppose that $\mathcal{X}$ is a convex subset of R^k, X is a random vector that takes values in $\mathcal{X}$, and g is a convex function defined on $\mathcal{X}$. If $E(X)$ is finite, then $E(X) \in \mathcal{X}$ and $g(E(X)) \leq E(g(X))$. The equality holds only if $g(X) = a + b^T X$ a.e. for some constants a and b.