Keener, Theoretical Statistics
Chapter 3: Risk, Sufficiency, Completeness, and Ancillarity

Junhyeok Kil

August 13, 2025

The Setup: An Initial Problem

Independent Random Variables

Suppose X and Y are independent and identically distributed (i.i.d.) with the common Lebesgue density:

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

This is the Exponential(θ) distribution.

Let U be independent of X and Y, and uniformly distributed on $(0, 1)$. The distribution of U does not depend on θ.

The Statistic of Interest

We define a statistic T as the sum of X and Y:

$$T = X + Y$$

Constructing "Fake" Data

The Goal

We want to show that the statistic $T = X + Y$ contains all the information about the parameter θ that is available in the original sample (X, Y).

Generating New Data

We can define a new pair of random variables, $(\tilde{X}, \tilde{Y})$, using our statistic T and the ancillary variable U:

$$\tilde{X} = UT$$
$$\tilde{Y} = (1 - U)T$$

Notice that $\tilde{X} + \tilde{Y} = UT + (1 - U)T = T$.

Question

What is the joint distribution of $(\tilde{X}, \tilde{Y})$? If it's the same as the joint distribution of (X, Y), then no information about θ was lost by summarizing the data with T.

Finding the Density of T

To find the joint density of $(\tilde{X}, \tilde{Y})$, we first need the density of $T = X + Y$.

1. **Conditional CDF of T given Y**
   Since X and Y are independent:

$$\begin{aligned} P(T \leq t|Y = y) &= P(X + Y \leq t|Y = y) \\ &= P(X + y \leq t) = P(X \leq t - y) \\ &= F_X(t - y) \end{aligned}$$

   So, $P(T \leq t|Y) = F_X(t - Y)$.

2. **Unconditional CDF of T**
   By the law of total expectation:

$$F_T(t) = P(T \leq t) = E[P(T \leq t|Y)] = E[F_X(t - Y)]$$

Finding the Density of T (cont.)

3. **Applying to the Exponential Case**
   The CDF of $X \sim \text{Exp}(\theta)$ is $F_X(x) = 1 - e^{-\theta x}$ for $x \geq 0$.
   Thus, $F_X(t - Y) = 1 - e^{-\theta(t-Y)}$ for $Y < t$, and 0 otherwise.

$$\begin{aligned} F_T(t) &= E[F_X(t - Y)] = \int_0^t (1 - e^{-\theta(t-y)}) f_Y(y) dy \\ &= \int_0^t (1 - e^{-\theta(t-y)}) \theta e^{-\theta y} dy \\ &= \theta \int_0^t (e^{-\theta y} - e^{-\theta t}) dy \\ &= \theta \left[-\frac{1}{\theta} e^{-\theta y} - y e^{-\theta t} \right]_0^t \\ &= 1 - e^{-\theta t} - t\theta e^{-\theta t} \end{aligned}$$

4. **Finding the PDF of T**
   Differentiating the CDF $F_T(t)$ with respect to t:

$$p_T(t) = F_T'(t) = \theta e^{-\theta t} - (\theta e^{-\theta t} - t\theta^2 e^{-\theta t}) = \theta^2 t e^{-\theta t}$$

Joint Density of $(\tilde{X}, \tilde{Y})$

- Since T and U are independent, their joint density is $p_\theta(t,u) = p_T(t) \cdot f_U(u) = (\theta^2 t e^{-\theta t}) \cdot 1$ for $t \geq 0, u \in (0,1)$.
- We find the distribution of $(\tilde{X}, \tilde{Y})$ via a change of variables from (T, U). Let $x = ut$ and $y = (1-u)t$. This implies $t = x + y$ and $u = x/(x+y)$.
- The joint density of $(\tilde{X}, \tilde{Y})$ is given by $p_{\tilde{X},\tilde{Y}}(x,y) = p_\theta(t(x,y), u(x,y)) \cdot |J|$, where $|J|$ is the determinant of the Jacobian of the transformation.
- A more direct way is to work with the probability integrals, which yields the density of $(\tilde{X}, \tilde{Y})$ as:

$$p(x,y) = p_\theta\left(x+y, \frac{x}{x+y}\right)\frac{1}{x+y}$$

- Substituting the joint density of (T, U):

$$\begin{aligned} p(x,y) &= \theta^2(x+y)e^{-\theta(x+y)} \cdot \frac{1}{x+y} \\ &= \theta^2 e^{-\theta(x+y)} \quad \text{for } x \geq 0, y \geq 0 \end{aligned}$$

Conclusion

Observation

The joint density of the "fake" data $(\tilde{X}, \tilde{Y})$ is

$$p(x, y) = \theta^2 e^{-\theta(x+y)}$$

This is exactly the same as the joint density of the original data (X, Y), since $f_X(x)f_Y(y) = (\theta e^{-\theta x})(\theta e^{-\theta y}) = \theta^2 e^{-\theta(x+y)}$.

Implication

The pair $(\tilde{X}, \tilde{Y})$ is just as informative about θ as (X, Y). But $(\tilde{X}, \tilde{Y})$ was constructed from $T = X + Y$ and a random number U whose distribution is independent of θ. Therefore, T by itself provides as much information about θ as the original data (X, Y).

Formal Definition

Definition: Sufficient Statistic

This works because the conditional distribution of (X, Y) given $T = t$ does not depend on θ. This motivates the formal definition:

A statistic $T = T(X)$ is **sufficient** for a parameter θ if the conditional distribution of the sample X given the value of the statistic $T = t$ does not depend on θ.

Sufficiency and Data Generation

The Conditional Distribution

Suppose $T = T(X)$ is a sufficient statistic. By definition, the conditional distribution of the data X given $T = t$ does not depend on θ. Let's denote it by:

$$Q_t(B) = P_\theta(X \in B | T = t)$$

Reconstructing the Original Distribution

Using the law of total expectation (smoothing), we can recover the original probability:

$$P_\theta(X \in B) = E_\theta[P_\theta(X \in B | T)] = E_\theta[Q_T(B)]$$

Sufficiency and Data Generation

Reconstructing the Original Distribution

Using the law of total expectation (smoothing), we can recover the original probability:

$$P_\theta(X \in B) = E_\theta[P_\theta(X \in B|T)] = E_\theta[Q_T(B)]$$

Generating "Fake" Data $\tilde{X}$

Suppose we use a random number generator to construct "fake" data $\tilde{X}$ from T taking $\tilde{X} \sim Q_t$ when $T = t$. Then $\tilde{X}|T = t \sim Q_t$ and by smoothing:

$$P_\theta(\tilde{X} \in B) = E_\theta[P_\theta(\tilde{X} \in B|T)] = E_\theta[Q_T(B)]$$

Therefore, X and $\tilde{X}$ have the exact same distribution P_θ.

Theorem 3.3

This ability to regenerate statistically equivalent data from a sufficient statistic leads to a powerful theorem.

Theorem (Theorem 3.3)

Suppose X has a distribution from a family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and that $T = T(X)$ is sufficient. Then for any estimator $\delta(X)$ of a function $g(\theta)$, there exists a **randomized estimator based on T** *that has the same risk function as $\delta(X)$.*

What this means

If we have an estimator that uses the full data X, we can find another estimator that depends *only* on the sufficient statistic T (plus some randomness) and performs identically in terms of risk (The randomness is sometimes necessary because not all estimators on X can be exactly written as deterministic functions of T).

Proof of Theorem 3.3

Proof.

1. **Define a Randomized Estimator:** A randomized estimator based on T is one that can be constructed from T and auxiliary random number generation.

Proof of Theorem 3.3

Proof.

1. **Define a Randomized Estimator:** A randomized estimator based on T is one that can be constructed from T and auxiliary random number generation.
2. **Construct the New Estimator:** From the previous slide, we know we can construct "fake" data $\tilde{X}$ from T using the conditional distribution Q_T. Since $\tilde{X}$ is a function of T and random numbers, the estimator $\delta(\tilde{X})$ is a randomized estimator based on T.

Proof of Theorem 3.3

Proof.

1. **Define a Randomized Estimator:** A randomized estimator based on T is one that can be constructed from T and auxiliary random number generation.
2. **Construct the New Estimator:** From the previous slide, we know we can construct "fake" data $\tilde{X}$ from T using the conditional distribution Q_T. Since $\tilde{X}$ is a function of T and random numbers, the estimator $\delta(\tilde{X})$ is a randomized estimator based on T.
3. **Compare Risk Functions:** The risk of an estimator δ is defined by the expectation of a loss function $L(\theta, \delta)$:

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$$

The risk of our new estimator $\delta(\tilde{X})$ is:

$$R(\theta, \delta(\tilde{X})) = E_\theta[L(\theta, \delta(\tilde{X}))]$$

Proof of Theorem 3.3 (cont.)

Conclusion: Since we proved that X and $\tilde{X}$ have the same distribution (P_θ) for any θ, the expectations defining their risks must be identical.

$$R(\theta, \delta) = R(\theta, \delta(\tilde{X}))$$

Thus, $\delta(\tilde{X})$ is a randomized estimator based on T with the same risk as the original estimator $\delta(X)$.

Finding Sufficient Statistics

A Practical Problem

The definition of sufficiency is conceptually powerful:

A statistic $T(X)$ is sufficient if the conditional distribution of X given $T = t$ does not depend on θ.

However, finding a sufficient statistic or trying to determine whether a specific statistic is sufficient is not straightforward.

The Factorization Theorem

Dominated Family of Distributions (Def 3.5)

A family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is **dominated** if there exists a measure μ such that every P_θ is absolutely continuous with respect to μ.

(This means we can write a density $p_\theta(x)$ for each distribution with respect to a common measure μ, like the Lebesgue measure for continuous variables or the counting measure for discrete variables).

The Factorization Theorem (Fisher–Neyman)

Theorem (Factorization Theorem (Thm 3.6))

Let $\mathcal{P}$ be a dominated family of distributions. A statistic $T(X)$ is sufficient if and only if the density $p_\theta(x)$ can be factored into two non-negative functions:

$$p_\theta(x) = g_\theta(T(x)) \cdot h(x)$$

where:

- *g_θ depends on θ and on the data x* **only through** *the value of the statistic $T(x)$.*
- *$h(x)$ depends on the data x but* **does not** *depend on the parameter θ.*

In many simple cases, the probability density function is fully specified by θ and $T(x)$, and $h(x) = 1$

Example 3.7

Setup

Let $X_1, \ldots, X_n$ be i.i.d. with common density:

$$f_\theta(x) = (\theta + 1)x^\theta, \quad x \in (0, 1)$$

for $\theta > -1$. The joint density for the sample $\mathbf{x} = (x_1, \ldots, x_n)$ is:

$$\begin{aligned} p_\theta(\mathbf{x}) &= \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} (\theta + 1)x_i^\theta \cdot \mathbb{1}(0, 1)(x_i) \\ &= (\theta + 1)^n \left(\prod_{i=1}^{n} x_i \right)^\theta \cdot \prod_{i=1}^{n} \mathbb{1}(0, 1)(x_i) \end{aligned}$$

Example 3.7

The Factorization

- Let the statistic be the product of the observations: $T(\mathbf{x}) = \prod_{i=1}^{n} x_i$.
- The part depending on θ and $T(\mathbf{x})$ is:

$$g_\theta(T(\mathbf{x})) = (\theta+1)^n \left(\prod_{i=1}^{n} x_i\right)^\theta = (\theta+1)^n [T(\mathbf{x})]^\theta$$

- The part depending only on $\mathbf{x}$ is:

$$h(\mathbf{x}) = \prod_{i=1}^{n} \mathbb{1}(0,1)(x_i) = \mathbb{1}(0,1)^n(\mathbf{x})$$

Conclusion

Since $p_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$, the Factorization Theorem tells us that $T(\mathbf{x}) = \prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

The Idea of Minimal Sufficiency

Reducing Sufficient Statistics

If T is a sufficient statistic and we can write another statistic $T_{new} = f(T)$, then T_{new} is also sufficient. This suggests we can often "reduce" or "compress" a sufficient statistic further.

The goal is to find the most compressed or "minimal" sufficient statistic that retains all information about θ, and has no other extra information.

Definition (Minimal Sufficient Statistic (Def 3.9))

A statistic T is **minimal sufficient** if:

1. T is sufficient.
2. For every other sufficient statistic $\tilde{T}$, there exists a function f such that $T = f(\tilde{T})$.

In other words, a minimal sufficient statistic is a function of every other sufficient statistic. It represents the maximum possible data reduction.

Minimal Sufficiency and the Likelihood Function

Example (Sufficient but Not Minimal)

If $X_1, \ldots, X_{2n} \sim N(\theta, 1)$, the statistic

$$\tilde{T} = \left(\sum_{i=1}^{n} X_i, \sum_{i=n+1}^{2n} X_i \right)$$

is sufficient. However, the total sum $T = \sum_{i=1}^{2n} X_i$ is also sufficient. Since $T = \tilde{T}_1 + \tilde{T}_2$, T is a function of $\tilde{T}$. In fact, T is minimal sufficient, while $\tilde{T}$ is not.

Minimal Sufficiency and the Likelihood Function

Connection to the Likelihood

The density $p_\theta(x)$, viewed as a function of θ for fixed data x, is called the **likelihood function**.

Two likelihoods $p_\theta(x)$ and $p_\theta(y)$ have the same "shape" if they are proportional as functions of θ:

$$p_\theta(x) \propto_\theta p_\theta(y) \iff p_\theta(x) = c(x, y)p_\theta(y)$$

A minimal sufficient statistic captures exactly the information needed to determine the shape of the likelihood. So a sufficient statistic reduces the data down to something that contains all the information about θ, and a minimal sufficient statistic is the most compressed version that still lets you reconstruct the likelihood shape. **So if x and y give the same likelihood shape, they must have the same value of the minimal sufficient statistic**.

A Test for Minimal Sufficiency

Theorem (Theorem 3.11)

Suppose we have a dominated family with densities $p_\theta(x)$. *Let* $T(x)$ *be a sufficient statistic. If for any two data points* x *and* y,

$$p_\theta(x) \propto_\theta p_\theta(y) \quad \Longrightarrow \quad T(x) = T(y)$$

then T *is* **minimal sufficient**.

Interpretation

This theorem states that if a statistic T takes on the same value for all data points that produce the same likelihood shape (and different values otherwise), then it must be minimal sufficient.
It partitions the sample space into sets where the likelihood shape is constant.

Proof Idea for Theorem 3.11

Let T be a sufficient statistic that satisfies the condition of the theorem. Let $\tilde{T}$ be any other sufficient statistic. We want to show that T is a function of $\tilde{T}$. Assume, for the sake of contradiction, that T is not a function of $\tilde{T}$. This means there exist data points x and y such that $\tilde{T}(x) = \tilde{T}(y)$ but $T(x) \neq T(y)$.

Proof Idea for Theorem 3.11

Let T be a sufficient statistic that satisfies the condition of the theorem. Let $\tilde{T}$ be any other sufficient statistic. We want to show that T is a function of $\tilde{T}$. Assume, for the sake of contradiction, that T is not a function of $\tilde{T}$. This means there exist data points x and y such that $\tilde{T}(x) = \tilde{T}(y)$ but $T(x) \neq T(y)$. Since $\tilde{T}$ is sufficient, we can use the Factorization Theorem:

$$p_\theta(x) = \tilde{g}_\theta(\tilde{T}(x))\tilde{h}(x)$$
$$p_\theta(y) = \tilde{g}_\theta(\tilde{T}(y))\tilde{h}(y)$$

Proof Idea for Theorem 3.11

Because $\tilde{T}(x) = \tilde{T}(y)$, we have $\tilde{g}_\theta(\tilde{T}(x)) = \tilde{g}_\theta(\tilde{T}(y))$. This implies the densities are proportional as functions of θ:

$$p_\theta(x) = \frac{\tilde{h}(x)}{\tilde{h}(y)} p_\theta(y) \implies p_\theta(x) \propto_\theta p_\theta(y)$$

Proof Idea for Theorem 3.11

Because $\tilde{T}(x) = \tilde{T}(y)$, we have $\tilde{g}_\theta(\tilde{T}(x)) = \tilde{g}_\theta(\tilde{T}(y))$. This implies the densities are proportional as functions of θ:

$$p_\theta(x) = \frac{\tilde{h}(x)}{\tilde{h}(y)} p_\theta(y) \implies p_\theta(x) \propto_\theta p_\theta(y)$$

By the theorem's condition on T, if $p_\theta(x) \propto_\theta p_\theta(y)$, then it must be that $T(x) = T(y)$.

Proof Idea for Theorem 3.11

Because $\tilde{T}(x) = \tilde{T}(y)$, we have $\tilde{g}_\theta(\tilde{T}(x)) = \tilde{g}_\theta(\tilde{T}(y))$. This implies the densities are proportional as functions of θ:

$$p_\theta(x) = \frac{\tilde{h}(x)}{\tilde{h}(y)} p_\theta(y) \implies p_\theta(x) \propto_\theta p_\theta(y)$$

By the theorem's condition on T, if $p_\theta(x) \propto_\theta p_\theta(y)$, then it must be that $T(x) = T(y)$. This is a **contradiction** to our starting assumption that $T(x) \neq T(y)$. Therefore, the assumption must be false, and T must be a function of $\tilde{T}$. Since $\tilde{T}$ was an arbitrary sufficient statistic, T is minimal.

The Likelihood Shape

Recap

Theorem 3.11 essentially says that a statistic T is minimal sufficient if it perfectly partitions the sample space according to the **shape of the likelihood function**.

$$p_\theta(x) \propto_\theta p_\theta(y) \iff T(x) = T(y)$$

This means the shape of the likelihood function itself can be considered a minimal sufficient statistic.

Remark

This condition does not need to hold on sets of data that have zero probability under all θ. If the implication fails only on a null set, T is still minimal sufficient.

Example 3.12: Exponential Families

Definition

An s-parameter **exponential family** is a family of distributions with densities of the form:

$$p_\theta(x) = h(x) \exp\left(\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) - B(\theta)\right)$$

where $\mathbf{T}(x) = (T_1(x), \ldots, T_s(x))$ is a vector of statistics and $\boldsymbol{\eta}(\theta) = (\eta_1(\theta), \ldots, \eta_s(\theta))$ is a vector of parameters.

Sufficiency

By the Factorization Theorem, we can immediately see that $\mathbf{T}(x)$ is a sufficient statistic.

- $g_\theta(\mathbf{T}(x)) = \exp\left(\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) - B(\theta)\right)$
- $h(x) = h(x)$

But is $\mathbf{T}(x)$ *minimal* sufficient?

Testing for Minimal Sufficiency

We apply the test from Theorem 3.11. Assume $p_\theta(x) \propto_\theta p_\theta(y)$ for two data points x and y.

$$h(x)e^{\boldsymbol{\eta}(\theta)\cdot\mathbf{T}(x)-B(\theta)} \propto_\theta h(y)e^{\boldsymbol{\eta}(\theta)\cdot\mathbf{T}(y)-B(\theta)}$$
$$e^{\boldsymbol{\eta}(\theta)\cdot\mathbf{T}(x)} \propto_\theta e^{\boldsymbol{\eta}(\theta)\cdot\mathbf{T}(y)}$$

Taking the natural logarithm of the proportionality:

$$\boldsymbol{\eta}(\theta)\cdot\mathbf{T}(x) = \boldsymbol{\eta}(\theta)\cdot\mathbf{T}(y) + c(x,y)$$

where $c(x,y)$ is a constant that does not depend on θ.

The Orthogonality Argument

The equation $\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta) \cdot \mathbf{T}(y) + c$ must hold for all $\theta \in \Omega$. Let's pick any two parameter values, θ_0 and θ_1, from the parameter space Ω.

$$\boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(y) + c$$
$$\boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(y) + c$$

The Orthogonality Argument

The equation $\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta) \cdot \mathbf{T}(y) + c$ must hold for all $\theta \in \Omega$. Let's pick any two parameter values, θ_0 and θ_1, from the parameter space Ω.

$$\boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(y) + c$$
$$\boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(y) + c$$

Subtracting the second equation from the first eliminates the constant c:

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot \mathbf{T}(x) = (\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot \mathbf{T}(y)$$

The Orthogonality Argument

The equation $\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta) \cdot \mathbf{T}(y) + c$ must hold for all $\theta \in \Omega$. Let's pick any two parameter values, θ_0 and θ_1, from the parameter space Ω.

$$\boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_0) \cdot \mathbf{T}(y) + c$$
$$\boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(x) = \boldsymbol{\eta}(\theta_1) \cdot \mathbf{T}(y) + c$$

Subtracting the second equation from the first eliminates the constant c:

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot \mathbf{T}(x) = (\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot \mathbf{T}(y)$$

Rearranging gives:

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot (\mathbf{T}(x) - \mathbf{T}(y)) = 0$$

Conclusion for Exponential Families

The result

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot (\mathbf{T}(x) - \mathbf{T}(y)) = 0$$

means that the vector $\mathbf{T}(x) - \mathbf{T}(y)$ is **orthogonal** to every vector of the form $\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)$.

Conclusion for Exponential Families

The result

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot (\mathbf{T}(x) - \mathbf{T}(y)) = 0$$

means that the vector $\mathbf{T}(x) - \mathbf{T}(y)$ is **orthogonal** to every vector of the form $\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)$.

The Condition for Minimality

Let's define the set of all such difference vectors:

$$\eta(\Omega) \ominus \eta(\Omega) \stackrel{\text{def}}{=} \{\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1) : \theta_0, \theta_1 \in \Omega\}$$

If the linear span of this set is the entire space $\mathbb{R}^s$, then the only vector orthogonal to all of them is the zero vector.

Conclusion for Exponential Families

The result

$$(\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)) \cdot (\mathbf{T}(x) - \mathbf{T}(y)) = 0$$

means that the vector $\mathbf{T}(x) - \mathbf{T}(y)$ is **orthogonal** to every vector of the form $\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1)$.

The Condition for Minimality

Let's define the set of all such difference vectors:

$$\eta(\Omega) \ominus \eta(\Omega) \stackrel{\text{def}}{=} \{\boldsymbol{\eta}(\theta_0) - \boldsymbol{\eta}(\theta_1) : \theta_0, \theta_1 \in \Omega\}$$

If the linear span of this set is the entire space $\mathbb{R}^s$, then the only vector orthogonal to all of them is the zero vector. In this case, we must have $\mathbf{T}(x) - \mathbf{T}(y) = \mathbf{0}$, which means $\mathbf{T}(x) = \mathbf{T}(y)$.

If the parameter space is "rich" enough such that the linear span of $\eta(\Omega) \ominus \eta(\Omega)$ is $\mathbb{R}^s$, then the statistic $\mathbf{T}(x)$ in an exponential family is **minimal sufficient**.

Completeness: A Technical Condition

Motivation

Completeness is a property of a statistic's distribution family. It strengthens the concept of sufficiency and is a key ingredient in theorems about best unbiased estimators.

Definition (Complete Statistic (Def 3.14))

A statistic T is **complete** for a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if for any function f, the condition

$$E_\theta[f(T)] = c, \quad \text{for all } \theta \in \Omega$$

implies that

$$P_\theta(f(T) = c) = 1 \quad \text{for all } \theta \in \Omega$$

(i.e., $f(T) = c$ almost everywhere).

By replacing f with $f - c$, we can simplify the definition: T is complete if $E_\theta[g(T)] = 0$ for all θ implies $g(T) = 0$ almost everywhere.

Completeness: A Technical Condition

Definition (Complete Statistic (Def 3.14))

A statistic T is **complete** for a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if for any function f, the condition

$$E_\theta[f(T)] = c, \quad \text{for all } \theta \in \Omega$$

implies that

$$P_\theta(f(T) = c) = 1 \quad \text{for all } \theta \in \Omega$$

(i.e., $f(T) = c$ almost everywhere).

This use of the word complete is analogous to calling a set of vectors $v1, \ldots, v_n$ complete if they span the whole space, that is, any v can be written as a linear combination $v = \sum a_j v_j$ of these vectors. This is equivalent to the condition that if w is orthogonal to all v_j's, then $w = 0$ (Christian Remling).

Example 3.16: Uniform Distribution

Setup

Let $X_1, \ldots, X_n$ be i.i.d. from a Uniform$(0, \theta)$ distribution, for $\theta > 0$.

Sufficient Statistic

The joint density is:

$$\begin{aligned} p_\theta(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}(0, \theta)(x_i) \\ &= \frac{1}{\theta^n} \mathbb{1}(0, \infty)(\min x_i) \cdot \mathbb{1}(-\infty, \theta)(\max x_i) \end{aligned}$$

Let $T(\mathbf{x}) = \max\{X_1, \ldots, X_n\}$. By the Factorization Theorem:

- $g_\theta(T(\mathbf{x})) = \frac{1}{\theta^n} \mathbb{1}(-\infty, \theta)(T(\mathbf{x}))$
- $h(\mathbf{x}) = \mathbb{1}(0, \infty)(\min x_i)$

Thus, $T = \max\{X_i\}$ is a sufficient statistic for θ.

Example 3.16: Proving Completeness

To prove completeness, we first need the probability density of T.

1. **Find the CDF of T**: For $t \in (0, \theta)$,

$$\begin{aligned} F_T(t) = P_\theta(T \leq t) &= P_\theta(\max\{X_i\} \leq t) \\ &= P_\theta(X_1 \leq t, \ldots, X_n \leq t) \\ &= \prod_{i=1}^{n} P_\theta(X_i \leq t) \quad \text{(by independence)} \\ &= \left(\frac{t}{\theta}\right)^n \end{aligned}$$

2. **Find the PDF of T**: Differentiate the CDF with respect to t:

$$p_T(t) = \frac{d}{dt} F_T(t) = \frac{nt^{n-1}}{\theta^n}, \quad \text{for } t \in (0, \theta)$$

Example 3.16: Proving Completeness (cont.)

3. **Apply the Definition of Completeness**: Suppose $E_\theta[f(T)] = c$ for all $\theta > 0$. We can write this as $E_\theta[f(T) - c] = 0$.

$$\int_0^\theta (f(t) - c) \cdot p_T(t) dt = 0$$

$$\int_0^\theta (f(t) - c) \frac{nt^{n-1}}{\theta^n} dt = 0$$

4. **Simplify the Expression**: Since this must hold for all $\theta > 0$, we can multiply by $\frac{\theta^n}{n}$ (which is non-zero):

$$\int_0^\theta (f(t) - c) t^{n-1} dt = 0 \quad \text{for all } \theta > 0$$

Example 3.16: Proving Completeness (cont.)

5. **Conclusion**: If the integral of a function over the interval $(0, \theta)$ is zero for all possible values of $\theta > 0$, then the function itself must be zero (a.e.). Therefore:

$$(f(t) - c)t^{n-1} = 0 \quad \text{(a.e. for } t > 0)$$

Since $t^{n-1} \neq 0$ for $t > 0$, we must have $f(t) - c = 0$, which implies $f(t) = c$ (a.e.).
Thus, $T = \max\{X_i\}$ is a complete statistic.

Completeness and Minimal Sufficiency

A Powerful Connection

Completeness provides a direct path to establishing minimal sufficiency, which can be simpler than using the likelihood ratio test from Theorem 3.11.

Theorem (Theorem 3.17)

If a statistic T is both **complete** *and* **sufficient**, *then it is also* **minimal sufficient**.

Strategy for finding a minimal sufficient statistic

1. Find a sufficient statistic T (e.g., using the Factorization Theorem).
2. Check if T is complete.
3. If it is, we can immediately conclude that T is minimal sufficient.

Theorem and Proof Strategy

Theorem (Theorem 3.17)

If a statistic T is complete and sufficient, then T is minimal sufficient.

Proof Strategy

To prove that T is minimal sufficient, we must show that it is a function of any other minimal sufficient statistic.

1. Let T be our complete and sufficient statistic.
2. Let $\tilde{T}$ be any other minimal sufficient statistic.
3. We will show that T can be written as a function of $\tilde{T}$.
4. This implies T achieves the same data reduction as $\tilde{T}$, and since $\tilde{T}$ is minimal, T must be minimal as well.

The Proof (Part 1)

Let T be complete and sufficient, and let $\tilde{T}$ be minimal sufficient. For simplicity, we'll first assume T and $\tilde{T}$ are single, bounded random variables.

1. **$\tilde{T}$ is a function of T:** By definition, a minimal sufficient statistic ($\tilde{T}$) must be a function of any other sufficient statistic (T). So, we can write:

$$\tilde{T} = f(T) \quad \text{for some function } f.$$

The Proof (Part 1)

Let T be complete and sufficient, and let $\tilde{T}$ be minimal sufficient. For simplicity, we'll first assume T and $\tilde{T}$ are single, bounded random variables.

1. **$\tilde{T}$ is a function of T:** By definition, a minimal sufficient statistic ($\tilde{T}$) must be a function of any other sufficient statistic (T). So, we can write:

$$\tilde{T} = f(T) \quad \text{for some function } f.$$

2. **Define a new statistic based on $\tilde{T}$:** Let's define a new statistic, $g(\tilde{T})$, as the conditional expectation of T given $\tilde{T}$:

$$g(\tilde{T}) = E[T|\tilde{T}]$$

Since $\tilde{T}$ is sufficient, the function g does not depend on θ.

The Proof (Part 2)

3. **Consider the difference:** Let's look at the statistic $d = T - g(\tilde{T})$.

The Proof (Part 2)

3. **Consider the difference:** Let's look at the statistic $d = T - g(\tilde{T})$.
4. **Express the difference as a function of T:** Since $\tilde{T} = f(T)$, we can substitute this into our expression for d:

$$d = T - g(f(T))$$

We see that d is a function of T alone. Let's call it $h(T)$.

The Proof (Part 2)

3. **Consider the difference:** Let's look at the statistic $d = T - g(\tilde{T})$.
4. **Express the difference as a function of T:** Since $\tilde{T} = f(T)$, we can substitute this into our expression for d:

$$d = T - g(f(T))$$

We see that d is a function of T alone. Let's call it $h(T)$.

5. **Calculate the expectation of $h(T)$:**

$$\begin{aligned} E_\theta[h(T)] &= E_\theta[T - g(\tilde{T})] \\ &= E_\theta[T] - E_\theta[g(\tilde{T})] \\ &= E_\theta[T] - E_\theta[E[T|\tilde{T}]] \quad \text{(by definition of g)} \\ &= E_\theta[T] - E_\theta[T] \quad \text{(by Law of Total Expectation)} \\ &= 0 \end{aligned}$$

This holds for all $\theta \in \Omega$.

The Proof (Conclusion)

6. **Apply Completeness:** We have established that $E_\theta[h(T)] = 0$ for all θ, where $h(T) = T - g(\tilde{T})$.
   Since T is a **complete** statistic, this implies that the function itself must be zero (almost everywhere):

$$h(T) = 0 \quad \text{(a.e. } \mathcal{P})$$

The Proof (Conclusion)

6. **Apply Completeness:** We have established that $E_\theta[h(T)] = 0$ for all θ, where $h(T) = T - g(\tilde{T})$.
   Since T is a **complete** statistic, this implies that the function itself must be zero (almost everywhere):

$$h(T) = 0 \quad \text{(a.e. } \mathcal{P})$$

7. **Final Result:** This means $T - g(\tilde{T}) = 0$, or

$$T = g(\tilde{T}) \quad \text{(a.e. } \mathcal{P})$$

   We have successfully shown that T is a function of $\tilde{T}$.

Conclusion

Since T is a function of any arbitrary minimal sufficient statistic $\tilde{T}$, T must itself be minimal sufficient.

Generalization to Random Vectors

The General Case

The proof assumed T and $\tilde{T}$ were single random variables. The result holds for random vectors as well.

The Argument

- The properties of sufficiency and completeness are preserved by one-to-one transformations.
- Any random vector in $\mathbb{R}^n$ can be mapped to a single random variable in $\mathbb{R}$ via a one-to-one (bimeasurable) function.
- Therefore, we can transform vector statistics T and $\tilde{T}$ into single random variables, apply the proof we just saw, and the result holds.

Full Rank Exponential Families

To apply this strategy to exponential families, we need a condition to ensure completeness.

Definition (Full Rank Exponential Family (Def 3.18))

An exponential family with density

$$p_\theta(x) = h(x) \exp\left(\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(x) - B(\theta)\right)$$

is said to be of **full rank** if two conditions hold:

1. The interior of the natural parameter space, $\eta(\Omega)$, is not empty.
2. The components of the statistic, $T_1, \ldots, T_s$, do not satisfy a linear constraint of the form $\mathbf{v} \cdot \mathbf{T} = c$ (almost everywhere).

Interpretation

1. The parameter space isn't "flat" or confined to a lower-dimensional surface (contains an open region).
2. The statistics $T_1, \ldots, T_s$ are not redundant (are lin. indep).

Completeness of Exponential Families

Theorem (Theorem 3.19)

In an exponential family of **full rank**, *the sufficient statistic* $\mathbf{T}$ *is* **complete**.

Summary of Results for Full Rank Families

If an exponential family is of full rank, the natural statistic $\mathbf{T} = (T_1, \ldots, T_s)$ has all three key properties:

- It is **sufficient** (by factorization).
- It is **complete** (by Theorem 3.19).
- It is **minimal sufficient** (as a consequence of being complete and sufficient, per Thm 3.17).

This is a cornerstone result in statistical theory, simplifying many problems involving exponential families.

Ancillary Statistics

Definition (Ancillary Statistic (Def 3.20))

A statistic V is called **ancillary** if its probability distribution does not depend on the parameter θ.

Intuition

Since its distribution is the same regardless of the true value of θ, an ancillary statistic, by itself, contains no information about θ.
E.g. i.i.d. Gaussian with unknown mean and known variance; Range and sample variance are ancillary.

Remark

Sometimes an ancillary statistic can be a function of a minimal sufficient statistic. In these cases, even though it is uninformative on its own, it can be relevant to inference.

Basu's Theorem

Basu's theorem links the concepts of sufficiency, completeness, and ancillarity. It shows that complete sufficient statistics are independent of any ancillary statistic.

Theorem (Basu's Theorem (Thm 3.21))

If T is a **complete and sufficient** *statistic for a family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, and if V is an* **ancillary** *statistic, then T and V are* **independent** *(under P_θ for any $\theta \in \Omega$).*

Why this is important

This theorem is often used to prove the independence of two statistics without having to derive their joint distribution. If you can show one is complete sufficient and the other is ancillary, independence is guaranteed.

Proof of Basu's Theorem

Let A and B be arbitrary (Borel) sets. We want to show that $P_\theta(T \in B, V \in A) = P_\theta(T \in B)P_\theta(V \in A)$.

1. **Define two functions:**
   - Let $q_A(t) = P_\theta(V \in A | T = t)$.
   - Let $p_A = P_\theta(V \in A)$.

Proof of Basu's Theorem

Let A and B be arbitrary (Borel) sets. We want to show that $P_\theta(T \in B, V \in A) = P_\theta(T \in B)P_\theta(V \in A)$.

1. **Define two functions:**
   - Let $q_A(t) = P_\theta(V \in A | T = t)$.
   - Let $p_A = P_\theta(V \in A)$.
2. **Analyze the functions:**
   - Since T is **sufficient**, the conditional distribution of any other statistic given T cannot depend on θ. Thus, $q_A(t)$ does not depend on θ.
   - Since V is **ancillary**, its marginal distribution cannot depend on θ. Thus, p_A does not depend on θ.

Proof of Basu's Theorem (cont.)

3. **Use the Law of Total Expectation (Smoothing):**

$$p_A = P_\theta(V \in A) = E_\theta[P_\theta(V \in A|T)] = E_\theta[q_A(T)]$$

Since p_A is a constant, we have $E_\theta[q_A(T)] = p_A$ for all θ.

Proof of Basu's Theorem (cont.)

3. **Use the Law of Total Expectation (Smoothing):**

$$p_A = P_\theta(V \in A) = E_\theta[P_\theta(V \in A|T)] = E_\theta[q_A(T)]$$

   Since p_A is a constant, we have $E_\theta[q_A(T)] = p_A$ for all θ.

4. **Use Completeness:** Since T is **complete**, the condition $E_\theta[q_A(T)] = p_A$ implies that $q_A(T) = p_A$ (almost everywhere).

Proof of Basu's Theorem (cont.)

5. **Prove Independence:**

$$\begin{aligned} P_\theta(T \in B, V \in A) &= E_\theta[\mathbb{1}B(T)\mathbb{1}A(V)] \\ &= E_\theta[E_\theta[\mathbb{1}B(T)\mathbb{1}A(V)|T]] \quad \text{(Smoothing)} \\ &= E_\theta[\mathbb{1}B(T)E_\theta[\mathbb{1}A(V)|T]] \text{(since T is fixed)} \\ &= E_\theta[\mathbb{1}B(T)q_A(T)] \quad \text{(Def of } q_A(T)\text{)} \\ &= E_\theta[\mathbb{1}B(T)p_A] \quad \text{(By completeness)} \\ &= p_A \cdot E_\theta[\mathbb{1}B(T)] = P_\theta(V \in A)P_\theta(T \in B) \end{aligned}$$

This shows that T and V are independent.

Example 3.22: The Normal Distribution

The Setup

Let $X_1, \ldots, X_n$ be i.i.d. from a $N(\mu, \sigma^2)$ distribution.
For this example, we consider the standard deviation σ to be a **fixed, known value**. The only unknown parameter is the mean μ.
The family of distributions is indexed only by μ:

$$\mathcal{P}_\sigma = \{N(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$$

The Goal

We will show that the sample mean $\bar{X}$ and the sample variance S^2 are independent.

Step 1: Find a Complete Sufficient Statistic

The Joint Density

The joint density of $X_1, \ldots, X_n$ is:

$$p_\mu(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right) \exp\left(\frac{n\mu}{\sigma^2}\bar{x} - \frac{n\mu^2}{2\sigma^2}\right)$$

Exponential Family Form

This is a full-rank, one-parameter exponential family with respect to the parameter μ:

- Natural Statistic: $T(\mathbf{x}) = \bar{x}$
- Natural Parameter: $\eta(\mu) = \frac{n\mu}{\sigma^2}$

Therefore, the sample mean $\bar{X} = \frac{1}{n}\sum X_i$ is a **complete and sufficient** statistic for the family $\mathcal{P}_\sigma$.

Step 2: Find an Ancillary Statistic

The Sample Variance

Let's consider the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

We want to show that S^2 is an **ancillary** statistic for the family $\mathcal{P}_\sigma$ (i.e., its distribution does not depend on μ).

The Proof Strategy

We will define a set of transformed variables whose distribution is free of μ, and then show that S^2 is a function of only these new variables.

Proof that S^2 is Ancillary

1. **Define new variables:** Let $Y_i = X_i - \mu$ for $i = 1, \ldots, n$.

Proof that S^2 is Ancillary

1. **Define new variables:** Let $Y_i = X_i - \mu$ for $i = 1, \ldots, n$.
2. **Find the distribution of Y_i:** Since $X_i \sim N(\mu, \sigma^2)$, the shifted variable Y_i is distributed as $N(0, \sigma^2)$. The distribution of each Y_i (and thus their joint distribution) **does not depend on** μ.

Proof that S^2 is Ancillary

1. **Define new variables:** Let $Y_i = X_i - \mu$ for $i = 1, \ldots, n$.
2. **Find the distribution of Y_i:** Since $X_i \sim N(\mu, \sigma^2)$, the shifted variable Y_i is distributed as $N(0, \sigma^2)$. The distribution of each Y_i (and thus their joint distribution) **does not depend on** μ.
3. **Rewrite S^2 in terms of Y_i:** First, note that the sample mean of the Y_i is $\bar{Y} = \bar{X} - \mu$. This gives us the key relationship:

$$X_i - \bar{X} = (Y_i + \mu) - (\bar{Y} + \mu) = Y_i - \bar{Y}$$

Proof that S^2 is Ancillary

1. **Define new variables:** Let $Y_i = X_i - \mu$ for $i = 1, \ldots, n$.
2. **Find the distribution of Y_i:** Since $X_i \sim N(\mu, \sigma^2)$, the shifted variable Y_i is distributed as $N(0, \sigma^2)$. The distribution of each Y_i (and thus their joint distribution) **does not depend on** μ.
3. **Rewrite S^2 in terms of Y_i:** First, note that the sample mean of the Y_i is $\bar{Y} = \bar{X} - \mu$. This gives us the key relationship:

$$X_i - \bar{X} = (Y_i + \mu) - (\bar{Y} + \mu) = Y_i - \bar{Y}$$

4. **Substitute into S^2:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Proof that S^2 is Ancillary

1. **Define new variables:** Let $Y_i = X_i - \mu$ for $i = 1, \ldots, n$.
2. **Find the distribution of Y_i:** Since $X_i \sim N(\mu, \sigma^2)$, the shifted variable Y_i is distributed as $N(0, \sigma^2)$. The distribution of each Y_i (and thus their joint distribution) **does not depend on** μ.
3. **Rewrite S^2 in terms of Y_i:** First, note that the sample mean of the Y_i is $\bar{Y} = \bar{X} - \mu$. This gives us the key relationship:

$$X_i - \bar{X} = (Y_i + \mu) - (\bar{Y} + \mu) = Y_i - \bar{Y}$$

4. **Substitute into S^2:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

5. **Conclusion:** Since S^2 is a function of only $Y_1, \ldots, Y_n$, and their joint distribution does not depend on μ, the distribution of S^2 cannot depend on μ. Therefore, S^2 is ancillary for μ.

Step 3: Apply Basu's Theorem

Summary of Findings

- The sample mean $\bar{X}$ is a **complete and sufficient** statistic for the family $\mathcal{P}_\sigma = \{N(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$.
- The sample variance S^2 is an **ancillary** statistic for the same family.

Theorem (Basu's Theorem)

A complete and sufficient statistic is independent of any ancillary statistic.

Conclusion

By Basu's Theorem, we can immediately conclude that for a fixed σ^2, the sample mean $\bar{X}$ and the sample variance S^2 are **independent** random variables. (Refer to https://stat210a.berkeley.edu/fall-2024/reader/completeness.html)

Convex Functions

Definition (Convex Function (Def 3.23))

A real-valued function f on a convex set $C \subseteq \mathbb{R}^p$ is called **convex** if, for any $\mathbf{x} \neq \mathbf{y}$ in C and any $\gamma \in (0, 1)$,

$$f(\gamma\mathbf{x} + (1-\gamma)\mathbf{y}) \leq \gamma f(\mathbf{x}) + (1-\gamma) f(\mathbf{y})$$

The function is **strictly convex** if the inequality is strict ($<$).

Geometric Interpretation

A function is convex if the line segment (chord) connecting any two points on its graph, $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$, lies on or above the graph of the function.
For a strictly convex function, the chord lies strictly above the graph.

The Supporting Line Theorem

Theorem (Theorem 3.24)

If f is a convex function on an open interval C, and if t is any point in C, then there exists a constant c (which depends on t) such that

$$f(t) + c(x - t) \leq f(x), \quad \forall x \in C.$$

If f is strictly convex, the inequality is strict for all $x \neq t$.

Geometric Interpretation

The expression $L(x) = f(t) + c(x - t)$ is the equation of a line passing through the point $(t, f(t))$.
This theorem states that for any convex function, we can always find a "supporting line" that touches the graph at any given point t and lies entirely on or below the graph of the function.

Jensen's Inequality

This is a fundamental result in probability theory that connects convex functions and expected values.

Theorem (Jensen's Inequality (Thm 3.25))

Let X be a random variable such that $P(X \in C) = 1$ for an open interval C, and assume $E[X]$ is finite. If f is a **convex** *function on C, then*

$$f(E[X]) \leq E[f(X)].$$

If f is **strictly convex**, *the inequality is strict unless X is a constant (almost surely).*

Remark (3.26)

Jensen's inequality also holds in higher dimensions, where X is a random vector and C is a convex set in $\mathbb{R}^p$.

Proof of Jensen's Inequality

1. Let $t = E[X]$. Since X takes values in the open interval C, its mean t must also be in C.

Proof of Jensen's Inequality

1. Let $t = E[X]$. Since X takes values in the open interval C, its mean t must also be in C.
2. By the Supporting Line Theorem (Thm 3.24), there exists a constant c such that for any value $x \in C$:

$$f(E[X]) + c(x - E[X]) \leq f(x)$$

Proof of Jensen's Inequality

1. Let $t = E[X]$. Since X takes values in the open interval C, its mean t must also be in C.
2. By the Supporting Line Theorem (Thm 3.24), there exists a constant c such that for any value $x \in C$:

$$f(E[X]) + c(x - E[X]) \leq f(x)$$

3. This inequality must also hold if we replace the fixed value x with the random variable X:

$$f(E[X]) + c(X - E[X]) \leq f(X) \quad \text{(a.s.)}$$

Proof of Jensen's Inequality

4. Take the expectation of both sides. By the linearity of expectation:

$$\begin{aligned} E[f(E[X]) + c(X - E[X])] &\leq E[f(X)] \\ f(E[X]) + c \cdot E[X - E[X]] &\leq E[f(X)] \\ f(E[X]) + c \cdot (E[X] - E[X]) &\leq E[f(X)] \\ f(E[X]) + 0 &\leq E[f(X)] \end{aligned}$$

This proves the main result. For the strict case, the inequality in step 3 is strict whenever $X \neq E[X]$, making the final expectation strict unless $P(X = E[X]) = 1$.

Motivation: Improving Estimators

Recall from Theorem 3.3

If T is a sufficient statistic, then for any estimator $\delta(X)$, we can construct a *randomized* estimator based on T that has the same risk.

The Question

Can we do better? Instead of just matching the risk, can we find a *non-randomized* estimator based on T that has a **smaller risk**?

The Key Ingredient

The answer is yes, provided our loss function is **convex**. The Rao-Blackwell theorem shows us exactly how to construct this improved estimator.

The Rao-Blackwell Theorem

Theorem (Rao-Blackwell (Thm 3.28))

Let T be a **sufficient statistic**, *let $\delta(X)$ be any estimator, and let the loss function $L(\theta, \cdot)$ be* **convex**.
Define a new estimator, $\eta(T)$, by taking the conditional expectation of $\delta(X)$ given T:

$$\eta(T) = E[\delta(X)|T]$$

Then the risk of the new estimator η is less than or equal to the risk of the original estimator δ:

$$R(\theta, \eta) \leq R(\theta, \delta) \quad \text{for all } \theta$$

Furthermore, if the loss is **strictly convex**, *the inequality is strict unless $\delta(X)$ was already a function of T (i.e., $\delta(X) = \eta(T)$ a.s.).*

Proof of the Rao-Blackwell Theorem

The proof is a direct application of Jensen's Inequality to the conditional distribution.

1. Start with the loss of the new estimator $\eta(T)$. By definition, $\eta(T) = E[\delta(X)|T]$.

$$L(\theta, \eta(T)) = L(\theta, E[\delta(X)|T])$$

Proof of the Rao-Blackwell Theorem

The proof is a direct application of Jensen's Inequality to the conditional distribution.

1. Start with the loss of the new estimator $\eta(T)$. By definition, $\eta(T) = E[\delta(X)|T]$.

$$L(\theta, \eta(T)) = L(\theta, E[\delta(X)|T])$$

2. The loss function $L(\theta, \cdot)$ is convex. We can apply Jensen's Inequality to the conditional expectation (the expectation is over the distribution of X given T):

$$L(\theta, E[\delta(X)|T]) \leq E[L(\theta, \delta(X))|T]$$

Proof of the Rao-Blackwell Theorem

3. Now, take the expectation of both sides of this inequality (this time, the expectation is over the distribution of T):

$$E[L(\theta, \eta(T))] \leq E\Big[E[L(\theta, \delta(X))|T]\Big]$$

Proof of the Rao-Blackwell Theorem

3. Now, take the expectation of both sides of this inequality (this time, the expectation is over the distribution of T):

$$E[L(\theta, \eta(T))] \leq E\Big[E[L(\theta, \delta(X))|T]\Big]$$

4. The left side is, by definition, the risk of η, $R(\theta, \eta)$. The right side, by the Law of Total Expectation (smoothing), is simply $E[L(\theta, \delta(X))]$, which is the risk of δ, $R(\theta, \delta)$.

$$R(\theta, \eta) \leq R(\theta, \delta)$$

The statement about strict inequality follows from the strict case of Jensen's.

Implications of the Rao-Blackwell Theorem

In summary, Rao-Blackwell theorem states that if we have an estimator for a metric and a sufficient statistic for the data, we can improve the estimator by conditioning on the sufficient statistic. This improvement is in terms of mean squared error, a measure of the estimator's accuracy.

This means that with convex loss functions, the only estimators worth considering, at least if estimators are judged solely by their risk, are functions of T but not X.

It can also be used to show that any randomized estimator is worse than a corresponding non-randomized estimator.