DEFINITION *How do you compute the MLE estimator for parameter θ of random variable with pdf $p_\theta(x)$?*	DEFINITION *How do you compute the method of moments estimator for parameter of interest γ of a random variable with probability density function $p_\theta(x)$?*
How do you compute the n^{th} moment of a random variable X?	*Let $X_1, \ldots, X_n$ be random variables with pdf* $p_\theta = \begin{cases} \frac{1}{\theta} & \text{, if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$ *What can you say about its likelihood function $L_X(\theta)$? What is the MLE of θ?*
What is the likelihood function $L_X(\theta)$ for i.i.d. random variables $X_1, \ldots, X_n$?	*Suppose $X_1, \ldots, X_n$ are independent random variables with pdf $p_\theta(x)$. What is the likelihood $L_X(\theta)$?*
THEOREM *Law of total probability*	DEFINITION *χ^2_k distribution with k degrees of freedom*
DEFINITION *What is the expectation $\mathbb{E}[\mathbb{E}g(X,Y) \mid Y]$?*	*If $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\mu)$, what can you say about the distribution of $X+Y$?*

Let k be the number of parameters to estimate/dimensions of $\gamma \in \Gamma$. Find a mapping $m : \Gamma \to \mathbb{R}^k$ that maps possible values of γ to the k moments of X. If m is invertible, then plug in the vector of k empirical moments $\hat{\mu}$ into it. Then the method of moments estimate is $\hat{\gamma} = m^{-1}(\hat{\mu})$.	Need to find the $\hat{\theta}$ which solves the optimization problem $\hat{\theta} = \arg\max_\phi p_\phi(x)$.
The likelihood $L_X(\theta)$ is 0 for all $\theta < \max\{X_1, \ldots, X_n\}$, and otherwise it's decreasing. So, MLE of θ is $\max\{X_1, \ldots, X_n\}$	Integrate it as $\mathbb{E}(X^k) = \int X^k dP$
$$L_X(\theta) = \prod_{i=1}^{n} p_\theta(x)$$	It's the probability $X_1, \ldots, X_n$ have a given value if the parameter is θ. Formally $L_X(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$
Let $Y_1, \ldots, Y_k$ be i.i.d. $\mathcal{N}(0,1)$. Then $\sum_{j=1}^{k} Y_j^2 \sim \chi_k^2$.	Let $(B_j)_{j \geq 1}$ be a partition of Ω, i.e. with $$\bigcup_{j \geq 1} B_j = \Omega$$ Then $$\begin{aligned} \mathbb{P}(A) &= \sum_j \mathbb{P}(A \cap B_j) \\ &= \sum_j \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j) \end{aligned}$$
$$X + Y \sim \text{Poisson}(\lambda + \mu)$$	By the iterated expectation lemma, it's $\mathbb{E}[\mathbb{E}g(X, Y) \mid Y] = \mathbb{E}g(X, Y)$.

Let $X_1, \ldots, X_n$ be Poisson (λ_i) distributed random variables, with a sum $Z = \sum_{i=1}^{n} X_i$. What can you say about the distribution $(X_1, \ldots, X_n \mid Z)$?

Given i.i.d. random variables X_1, X_2 with distribution function F_X, what can you say about the distribution of $Z = \max\{X_1, X_2\}$? What is its distribution function and density?

DEFINITION

Sufficient statistics for distribution with random variable X.

THEOREM

Factorization theorem of Neyman

How can you show that the MLE $\hat{\theta}_{\text{MLE}}$ only depends on the sufficient statistic S?

DEFINITION

Exponential family $\{P_\theta : \theta \in \Theta\}$.

How can you show that a random variable $X \sim \text{Poisson}(\theta), \theta > 0$ belongs to an exponential family? The pmf of Poisson is $\mathbb{P}(X = x) = e^{-\theta} \frac{\theta^x}{x!}$.

Show that for independent and identically distributed random variables $X_1, \ldots, X_n$ sampled according to some distribution from the exponential family

$$p_\theta(x) = \exp[\underbrace{c(\theta)^T}_{\in\mathbb{R}^k} \underbrace{T(x)}_{\in\mathbb{R}^k} - d(\theta)] h(x),$$

the sum $\sum_{i-1}^{n} T(X_i)$ is a sufficient statistic for θ.

How can you show that a normal random variable $\mathcal{N}(\mu, \sigma^2)$ with probability density function

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for $x \in \mathbb{R}$ belongs to the exponential family?

DEFINITION

What is the canonical form of a k-dimensional exponential family?

Z has distribution function $F(z) = F_X^2(z)$, and its density is $f(z) = 2F_X(z)f_X(z)$

$(X_1, \ldots, X_n \mid Z) \sim \text{Multinomial}(Z, p_1, \ldots, p_n)$ with

$$p_i = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$$

A statistic S is sufficient if and only if the density function $p_\theta(x)$ can be written as

$$p_\theta(x) = g_\theta(S(x))h(x)$$

or all x, θ and some functions $g_\theta(\cdot) \geq 0$ and $h(\cdot) \geq 0$.

The statistic $S = S(X)$ is called sufficient if the distribution of X giver $S = s$ does not depend on θ.

A k-dimensional exponential family is a family of distributions $\{P_\theta : \theta \in \Theta\}$ if the density functions of distributions in the family can be written in the form

$$p_\theta(x) = \exp\left[\sum_{j=1}^k c_j(\theta)T_j(x) - d(\theta)\right] h(x).$$

You can use the factorization theorem of Neyman:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\max_{\varphi \in \Theta} p_\varphi(x) \\ &= \arg\max_{\varphi \in \Theta} g_\varphi(S)h(x) \\ &= \arg\max_{\varphi \in \Theta} g_\varphi(S)\end{aligned}$$

$$\prod_{i=1}^n p_\theta(x_i) = \underbrace{\exp\left[c(\theta)\sum_{i=1}^n T(x_i) - nd(\theta)\right]}_{g_\theta(S(x))} \underbrace{\prod_{i=1}^n h(x_i)}_{h(x)}$$

The pdf can be written as

$$\begin{aligned}p_\theta(x) &= \mathbb{P}(X = x) = e^{-\theta}\frac{\theta^x}{x!} \\ &= \exp[\underbrace{\log(\theta)}_{c(\theta)} \underbrace{x}_{T(x)} - \underbrace{\theta}_{d(\theta)}] \underbrace{\frac{1}{x!}}_{h(x)}\end{aligned}$$

It's obtained by setting $c(\theta)$ to be an identity function

$$p_\theta(x) = \exp\left[\sum_{j=1}^k \theta_j T_j(x) - d(\theta)\right] h(x)$$

Typically this can be achieved by reparametrization.

Expand the square in the numerator and move σ^2 into the exponential

$$\begin{aligned}p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right] \exp\left[-\log\frac{1}{2}\left(\sigma^2\right)\right] \\ &= \exp\left[-\frac{x^2}{2\sigma^2} + x\frac{\mu}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right] \frac{1}{\sqrt{2\pi}} \\ &= \exp[\underbrace{\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2}\end{bmatrix}}_{c(\theta)} \underbrace{\begin{bmatrix} x^2 \\ x \end{bmatrix}}_{T(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)}_{d(\theta)}] \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)}\end{aligned}$$

DEFINITION

For a collection of probability density functions $\{p_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$, what is a score function?

DEFINITION

What is the Fisher information for a parameter θ?

What can you say about the expected value of the score function $\mathbb{E}[s_\theta(x)]$?

What can you say about the expected value of Fisher information with respect to the score function?

If $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$, how can you write $\mathbb{E}_\theta T(x)$ and $\mathrm{Var}_\theta(T(x))$ in terms of c, d, and/or h ?

If $p_\theta(x)$ is a probability density function from the exponential family in the canonical form

$$p_\theta(x) = \exp\left[\sum_{j=1}^{k} \theta_j T_j(x) - d(\theta)\right] h(x),$$

what can you say about the expectation $\mathbb{E}_\theta[T(x)]$ and $\mathrm{Var}_\theta(T(x))$?

Given $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$, derive $\mathbb{E}_\theta[T(x)]$ in terms of $\dot{c}(\theta)$ and $\dot{d}(\theta)$.

How is the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$ of a parameter θ related to the score function?

Given $X \sim \mathrm{Bernoulli}(\theta)$ for $\theta \in (0,1)$, and knowing that for $x \in \{0,1\}$ we can write

$$\mathbb{P}(X = x) = \exp\left[\left(\log\frac{\theta}{1-\theta}\right)x - (-\log(1-\theta))\right]$$

how can you reparametrize X into canonical exponential family form?

Take a collection of i.i.d. random variables $X_1, \ldots, X_n \sim \mathcal{N}\left(0, \sigma^2\right)$. What s a minimal sufficient statistic for σ^2 ? Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix} \begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln\left(\sigma^2\right)\right)\right] \frac{1}{\sqrt{2\pi}}$$

It's defined as

$$I(\theta) = \mathrm{Var}_\theta\left(s_\theta(x)\right)$$

The score function s_θ is

$$s_\theta(x) = \frac{d}{d\theta} \log p_\theta(x)$$

Under regularity assumptions $I(\theta) = -\mathbb{E}_\theta[\dot{s}_\theta(x)]$

Under regularity conditions, $\mathbb{E}_\theta[s_\theta(x)] = 0$.

$$\begin{aligned} \mathbb{E}_\theta T(x) &= \dot{d}(\theta) \\ \mathrm{Var}_\theta(T(x)) &= \ddot{d}(\theta) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_\theta T(x) &= \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \\ \mathrm{Var}_\theta(T(x)) &= \frac{\left[\ddot{d}(\theta) - \ddot{c}(\theta)\frac{\dot{d}(\theta)}{\dot{c}(\theta)}\right]}{\dot{c}(\theta)^2} \end{aligned}$$

$\hat{\theta}_{\text{MLE}}$ is the solution of

$$\frac{1}{n}\sum_{i=1}^{n} s_\theta\left(x_i\right) = 0.$$

We will use the score function, and $\mathbb{E}_\theta[s_\theta(x)] = 0$

$$\begin{aligned} \log p_\theta(x) &= c(\theta)T(x) - d(\theta) + \log h(x) \\ s_\theta(x) &= \dot{c}(\theta)T(x) - \dot{d}(\theta) \\ \mathbb{E}_\theta s_\theta(x) &= 0 \\ &\implies \mathbb{E}_\theta T(x) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \end{aligned}$$

1. From the formula we see that x_i^2 is minimal sufficient for any particular X_i since $\mu = 0$.
2. We know that for any i.i.d. collection of exponential family random variables, the sum of $T(x)$ is sufficient.

So, the minimal sufficient statistic in this case is the sum (or e.g. mean) of squares $s(x) = \sum_{i=1}^n x_i^2$.

1. Define $\gamma = \left(\log \frac{\theta}{1-\theta}\right)$.
2. Rewrite $(-\log(1-\theta))$ using γ

$$\theta = \frac{e^\gamma}{1+e^\gamma} \implies -\log(1-\theta) = \log\left(1+e^\gamma\right)$$

3. Plug into the equation.

$$\mathbb{P}(X = x) = \exp\left[\gamma x - \log\left(1+e^\gamma\right)\right]$$

Consider $X \sim P_\theta, \theta \in \Theta$, some parameter of interest $\gamma = g(\theta) \in \mathbb{R}$, and an estimator $T(x)$ of γ. What is the bias of T ? What is an unbiased estimator?

Consider $X \sim \text{Binomial}(n, \theta)$. For what class of functions $g(\theta)$ is it possible to construct unbiased estimators? Why? Recall that the pmf of X is

$$\mathbb{P}(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Consider $X \sim \text{Binomial}(n, \theta)$. How could you show that $T(X) = \frac{X(X-1)}{n(n-1)}$ is an unbiased estimator of θ^2 ? Recall that:

1. *Pmf of X is $\mathbb{P}(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$.*
2. *$\mathbb{E}X = n\theta$*
3. *$\mathbb{E}X^2 = n\theta(1-\theta) + (n\theta)^2$.*

Let $X_1, \cdots, X_n$ be i.i.d. copies of X. What is an unbiased estimator of $\text{Var}(X)$?

If $X_1, \cdots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, what

$$\text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right)?$$

Let $X_1, \ldots, X_n$ i.i.d. copies of X. Define $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. What are the steps to show that

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

is an unbiased estimator of σ^2?

Consider $X_1, \ldots, X_n$ i.i.d. copies of X with variance σ^2. Knowing that

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

is an unbiased estimator of σ^2, show that $S = \sqrt{S^2}$ is not an unbiased estimator of σ.

What is the mean square error of an estimator $T(X)$ trying to estimate $g(\theta)$?

What is $\text{MSE}_\theta(T)$ in terms of $\text{Bias}_\theta(T)$ and $\text{Var}_\theta(T)$?

Show that for an estimator $T(X)$ trying to estimate $g(\theta)$

$$\text{MSE}_\theta(T) = \text{Bias}^2_\theta(T) + \text{Var}_\theta(T)$$

Only polynomials. This is because:

1. An estimator is unbiased when $\mathbb{E}_\theta T(X) - g(\theta) = 0$ for all $\theta \in \Theta$.
2. $\mathbb{E}_\theta T(X) = \sum_{x=0}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} T(x) = q(\theta)$ where $q(\theta)$ is a polynomial of degree $\leq n$.
3. $\mathbb{E}_\theta T(X) - g(\theta) = 0$ for all $\theta \in \Theta$ means $g(\theta)$ must be a polynomial.

$$\mathrm{Bias}_\theta(T) := \mathbb{E}_\theta T(X) - g(\theta)$$

An estimator T is unbiased if $\mathrm{Bias}_\theta(T) = 0$ for all $\theta \in \Theta$

An unbiased estimator of $\mathrm{Var}(X)$ is

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

$$\begin{aligned}
\mathbb{E}_\theta T &= \frac{\mathbb{E}_\theta X^2 - \mathbb{E}_\theta X}{n(n-1)} \\
&= \frac{n\theta(1-\theta) + (n\theta)^2 - n\theta}{n(n-1)} \\
&= \frac{-n\theta^2 + n^2\theta^2}{n(n-1)} \\
&= \frac{n(n-1)\theta^2}{n(n-1)} = \theta^2
\end{aligned}$$

so $\forall_{\theta \in (0,1)} \mathbb{E}_\theta T - \theta^2 = 0$

First note the identity

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

$$\begin{aligned}
\mathbb{E}\left(S^2\right) &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^{n} \left([X_i - \mu] + [\mu - \bar{X}]\right)^2\right] && \text{(Insert } \mu - \mu\text{)} \\
&= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^{n} (X_i - \mu)^2 - n\left(\bar{X} - \mu\right)^2\right] && \text{(Above identity)} \\
&= \frac{1}{n-1} \left[n\sigma^2 - \cancel{n}\frac{\sigma^2}{\cancel{n}}\right] = \sigma^2
\end{aligned}$$

Note that $\mathrm{Var}\left(\frac{X_i}{n}\right) = \frac{\sigma^2}{n^2}$. So it's

$$\begin{aligned}
\mathrm{Var}\left(\frac{X_1 + .. + X_n}{n}\right) &= \mathrm{Var}\left(\frac{X_1}{n}\right) + \cdots + \mathrm{Var}\left(\frac{X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}\left(X_1\right) + \cdots + \mathrm{Var}\left(X_n\right)\right) \\
&= \frac{n}{n^2}\sigma^2 \\
&= \frac{1}{n}\sigma^2
\end{aligned}$$

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta(T(X) - g(\theta))^2$$

Observe that

$$\mathbb{E}(S) = \mathbb{E}\left(\sqrt{S^2}\right) \leq \sqrt{\mathbb{E}\left(S^2\right)} = \sigma$$

Since the middle inequality (holds for concave Jensen) is an equality if and only if $\mathrm{Var}(S) = 0, S$ is not an unbiased estimator of σ.

Let $q(\theta) = \mathbb{E}_\theta T(X)$. Then

$$\begin{aligned}
\mathbb{E}_\theta(T - g(\theta))^2 &= \mathbb{E}_\theta(T - q(\theta) + q(\theta) - g(\theta))^2 \\
&= \underbrace{\mathbb{E}_\theta(T - q(\theta))^2}_{\mathrm{Var}_\theta(T)} + \underbrace{(q(\theta) - g(\theta))^2}_{\mathrm{Bias}_\theta(T)^2} \\
&\quad + 2(q(\theta) - g(\theta)) \underbrace{\mathbb{E}_\theta(T - q(\theta))}_{0} \\
&= \mathrm{Bias}_\theta^2(T) + \mathrm{Var}_\theta(T).
\end{aligned}$$

$$\mathrm{MSE}_\theta(T) = \mathrm{Bias}_\theta^2(T) + \mathrm{Var}_\theta(T)$$

When estimating a parameter γ, is it possible for a biased estimator to get a better MSE than an unbiased estimator?	DEFINITION *What is a Uniform Minimum Variance Unbiased (UMVU) estimator?*
LEMMA *State the iterated variance lemma.*	*Show that for any random variables Y and Z,* $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y \mid Z)) + \mathbb{E}(\mathrm{Var}(Y \mid Z))$ *(part-1)*
Show that for any random variables Y and Z, $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y \mid Z)) + \mathbb{E}(\mathrm{Var}(Y \mid Z))$ *(part-2)*	*If T is an unbiased estimator and S is a sufficient statistic, what can you say about the variance of an estimator $T^* = \mathbb{E}(T \mid S)$ in relation to the variance of T ?*
Show that if T is an unbiased estimator for $g(\theta)$ and S is a sufficient statistic, then the estimator defined by $T^* = \mathbb{E}(T \mid S)$ is unbiased and fulfils $\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T)$ for all θ	*What does it mean for a sufficient statistic S to be called complete?*
LEMMA *State the Lehmann-Scheffe lemma.*	*Prove that if T is an unbiased estimator of $g(\theta)$ with finite variance for all θ, and S is a sufficient and complete statistic, then $T^* = \mathbb{E}(T \mid S)$ is UMVU.*

An unbiased estimator T^* is called Uniform Minimum Variance Unbiased (UMVU) estimator if for any other unbiased estimator T,

$$\operatorname{Var}_\theta\left(T^*\right) \leq \operatorname{Var}_\theta(T), \forall_\theta$$

Yes, for instance when estimating σ^2 of a normal distribution. The estimator

$$T_{\text{opt}}=\frac{1}{n+1} \sum_{i=1}^n\left(X_i-\bar{X}\right)^2$$

is biased, but obtains a better MSE than the unbiased sample variance

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

Just repeatedly use $\operatorname{Var}(X)=\mathbb{E}\left(X^2\right)-\mathbb{E}(X)^2$, linearity of $\mathbb{E}$, and iterated $\mathbb{E}$

1.

$$\begin{aligned}
\operatorname{Var}(\mathbb{E}(Y \mid Z)) & =\mathbb{E}(\mathbb{E}(Y \mid Z))^2-(\mathbb{E}(\mathbb{E}(Y \mid Z)))^2 \\
& =\mathbb{E}(\mathbb{E}(Y \mid Z))^2-(\mathbb{E}(Y))^2
\end{aligned}$$

(Iterated $\mathbb{E}$)

Let Y and Z be two random variables. Then

$$\operatorname{Var}(Y)=\operatorname{Var}(\mathbb{E}(Y \mid Z))+\mathbb{E}(\operatorname{Var}(Y \mid Z))$$

For all $\theta, \operatorname{Var}_\theta\left(T^*\right) \leq \operatorname{Var}_\theta(T)$

2. (Linearity, iterated $\mathbb{E}$)

$$\begin{aligned}
\mathbb{E}(\operatorname{Var}(Y \mid Z)) & =\mathbb{E}\left(\mathbb{E}\left(Y^2 \mid Z\right)-\left(\mathbb{E}(Y \mid Z))^2\right)\right. \\
& =\mathbb{E}\left(Y^2\right)-\mathbb{E}(\mathbb{E}(Y \mid Z))^2
\end{aligned}$$

3.

$$\begin{aligned}
& \operatorname{Var}(\mathbb{E}(Y \mid Z))+\mathbb{E}(\operatorname{Var}(Y \mid Z)) \\
& =\mathbb{E}(\mathbb{E}(Y \mid Z))^2-(\mathbb{E}(Y))^2+\mathbb{E}\left(Y^2\right)-\mathbb{E}(\mathbb{E}(Y \mid Z))^2 \\
& =\mathbb{E}\left(Y^2\right)-(\mathbb{E}(Y))^2=\operatorname{Var}(Y)
\end{aligned}$$

A sufficient statistic S is called complete if the following implication holds:

$$\forall_\theta \mathbb{E}_\theta h(S)=0 \Longrightarrow \forall_\theta h(S)=0, \mathbb{P}_\theta\text{-almost surely}$$

where h is a function of S not depending on θ.

T^* is an unbiased estimator for $g(\theta)$ because

$$\mathbb{E}_\theta T^*=\mathbb{E}_\theta \mathbb{E}(T \mid S)=\mathbb{E}_\theta T=g(\theta).$$

Now use iterated variance lemma

$$\begin{aligned}
\operatorname{Var}_\theta(T) & =\operatorname{Var}_\theta \overbrace{(\mathbb{E}(T \mid S))}^{T^*}+\overbrace{\mathbb{E}_\theta \operatorname{Var}(T \mid S)}^{\geq 0} \\
& \geq \operatorname{Var}_\theta\left(T^*\right).
\end{aligned}$$

Let $\tilde{T}$ be unbiased. Define $T^{\prime}=\mathbb{E}(\tilde{T} \mid S)=T^{\prime}(S)$. We know now that $\operatorname{Var}_\theta\left(T^{\prime}\right) \leq \operatorname{Var}_\theta(\tilde{T})$. Now $T^*(S)= \mathbb{E}(T \mid S)$ and $T^{\prime}(S)=\mathbb{E}(\tilde{T} \mid S)$ are unbiased. Since they're unbiased, it's true that for all θ

$$\mathbb{E}_\theta\left[T^*(S)-T^{\prime}(S)\right]=0$$

Since S is complete, $T^*(S)-T^{\prime}(S)=0$
$\mathbb{P}_\theta$ -almost surely for all θ

Let T be an unbiased estimator of $g(\theta)$ with finite variance for all θ. Also let S be a sufficient and complete statistic. Then $T^*=\mathbb{E}(T \mid S)$ is UMVU.

Consider $X_1, \cdots, X_n$ i.i.d. $\text{Poisson}(\theta)$ *random variables with $\theta > 0$. Show that the statistic $S = \sum_{i=1}^n X_i$ is complete. (part-1)*

Consider $X_1, \cdots, X_n$ i.i.d. $\text{Poisson}(\theta)$ *random variables with $\theta > 0$. Show that the statistic $S = \sum_{i=1}^n X_i$ is complete. (part-2)*

If S is a sufficient and complete statistic, and $T'(S) \in \mathbb{R}$ is a function which only depends on S, what is an easy UMVU estimator of $\mathbb{E}_\theta T'$?

LEMMA

State the lemma about sufficiency and completeness of statistics for exponential family random variables.

Consider i.i.d. $X_1, \ldots, X_n \sim \text{Gamma}(k\lambda)$ random variables where $\theta = (k, \lambda) \in \mathbb{R}_+^2$ with density

$$\begin{aligned} p_\theta(x) &= \frac{e^{-\lambda x} x^{k-1} \lambda^k}{\Gamma(k)} \\ &= \exp[-\lambda x + (k-1)\log x + k \log \lambda - \log \Gamma(k)] \end{aligned}$$

Sufficient and complete statistic for θ ? Why?

Consider a collection of distributions on $\mathcal{X}$ parametrized by one-dimensional $\theta \in \Theta$. What is the support condition of the Cramér-Rao lower bound?

Consider a collection of distributions on $\mathcal{X}$ parametrized by one-dimensional $\theta \in \Theta$. What is the differentiability in L^2 condition of the Cramér-Rao lower bound?

Consider a collection of distributions on $\mathcal{X}$ parametrized by one-dimensional $\theta \in \Theta$. What do the following conditions imply about s_θ ?

1. *The support $\{x : p_\theta(x) > 0\}$ does not depend on parameter θ.*
2. *There exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(x) < \infty$ and*

$$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{h p_\theta(x)} - s_\theta(x) \right]^2 = 0$$

for all θ.

Let T be an estimator of $g(\theta)$ where θ is one-dimensional. Assume the support of P_θ does not depend on θ, and the differentiability in L^2 condition holds. If $\text{Var}_\theta(T) < \infty$ for all θ, then what is a simple formula for $\dot{g}(\theta)$?

State the Cauchy-Schwartz inequality for covariances/variances.

Since the monomials $\{1, x, x^2, \cdots\}$ form an independent system, we know that $$\sum_{k=0}^{\infty} \frac{x^k}{k!} h(k) = 0 \text{ for all } x > 0 \Longleftrightarrow h(k) = 0 \text{ for all } k \geq 0$$ which is the desired implication to prove completeness of S.	Note that the sum S is distributed according to Poisson$(n\theta)$. Consider now a function h which has the property that $$\mathbb{E}_\theta h(X) = \sum_{k=0}^{\infty} e^{-n\theta} \frac{(n\theta)^k}{k!} h(k) = 0 \text{ for all } \theta > 0.$$ Rewrite expansion in an obviously independent basis $$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} h(k) = 0 \text{ for all } x > 0.$$
$$p_\theta(x) = \exp[\underbrace{c(\theta)}_{\in \mathbb{R}^k} \underbrace{T(x)}_{\in \mathbb{R}^k} - d(\theta)] h(x)$$ $$\mathcal{C} := \left\{ c(\theta) = \begin{pmatrix} c_1(\theta) \\ \cdots \\ c_k(\theta) \end{pmatrix} \in \mathbb{R}^k ; \theta \in \Theta \right\}$$ contains a k-dimensional open ball. Then $T = \begin{pmatrix} T_1 \\ \cdots \\ T_k \end{pmatrix}$ is sufficient and complete.	T' itself is an UMVU estimator of $\mathbb{E}_\theta T'$
The support $\{x : p_\theta(x) > 0\}$ does not depend on parameter θ	The sufficient and complete statistic is $S = (\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i)$. This is because $c_1(\theta) = -\lambda, c_2(\theta) = k - 1$, so the space of $\mathcal{C} = (-\infty, 0) \times (-1, \infty)$ is two dimensional, and we know from a lemma that this means that T is sufficient and complete.
For all θ it's true that $\mathbb{E}_\theta s_\theta(X) = 0$	There exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(x) < \infty$ and $$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{h p_\theta(x)} - s_\theta(x) \right]^2 = 0$$ for all θ.
$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X) \, \mathrm{Var}(Y)$$	$$\dot{g}(\theta) = \mathrm{Cov}\left(T, s_\theta(X)\right)$$

Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Suppose $s_\theta = \frac{d}{d\theta} \log p_\theta$. How does the Fisher information change as the sample size n increases?

Let $X \sim \text{Geometric}(\theta), 0 < \theta < 1$. What is the score function s_θ in this case? Recall that for $X \sim \text{Geometric}(\theta)$,

1. $p_\theta(x) = \mathbb{P}_\theta(X = x) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots,$
2. $\mathbb{E}_\theta X = \frac{1}{\theta}$
3. $\text{Var}_\theta(X) = \frac{1-\theta}{\theta^2},$
4. $\log p_\theta(x) = \underbrace{(x-1)}_{T(x)} \underbrace{\log(1-\theta)}_{c(\theta)} - \underbrace{(-\log\theta)}_{d(\theta)}.$

Consider a collection of distributions on $\mathcal{X}$ parametrized by one-dimensional $\theta \in \Theta$. What is the Cramér-Rao lower bound for an unbiased estimator of $g(\theta)$?

Let $X \sim \text{Geometric}(\theta), 0 < \theta < 1$. What are the possible ways to find the Fisher information $I(\theta)$ in this case? Recall that for $X \sim \text{Geometric}(\theta)$,

1. $p_\theta(x) = \mathbb{P}_\theta(X = x) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots,$
2. $\mathbb{E}_\theta X = \frac{1}{\theta}$
3. $\text{Var}_\theta(X) = \frac{1-\theta}{\theta^2}$
4. $\log p_\theta(x) = \underbrace{(x-1)}_{T(x)} \underbrace{\log(1-\theta)}_{c(\theta)} - \underbrace{(-\log\theta)}_{d(\theta)}$
5. $s_\theta(x) = \frac{d}{d\theta} \log p_\theta(x) = -\frac{x-1}{1-\theta} + \frac{1}{\theta}.$

Let $X_1, \ldots, X_n \sim \text{Geometric}(\theta)$ be i.i.d with $0 < \theta < 1$. What is the Cramér Rao lower bound for $g(\theta) = \frac{1}{\theta}$? Recall that for $X \sim \text{Geometric}(\theta)$,

1. $p_\theta(x) = \mathbb{P}_\theta(X = x) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots,$
2. $\mathbb{E}_\theta X = \frac{1}{\theta}$
3. $\text{Var}_\theta(X) = \frac{1-\theta}{\theta^2}$
4. $\log p_\theta(x) = \underbrace{(x-1)}_{T(x)} \underbrace{\log(1-\theta)}_{c(\theta)} - \underbrace{(-\log\theta)}_{d(\theta)},$
5. $s_\theta(x) = \frac{d}{d\theta} \log p_\theta(x) = -\frac{x-1}{1-\theta} + \frac{1}{\theta}$
6. $I(\theta) = \frac{1}{(1-\theta)\theta^2}.$

Let $X_1, \ldots, X_n \sim \text{Geometric}(\theta)$ be i.i.d with $0 < \theta < 1$. What is the CramérRao lower bound for $g(\theta) = \theta$? Recall that for $X \sim \text{Geometric}(\theta)$,

1. $p_\theta(x) = \mathbb{P}_\theta(X = x) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots,$
2. $\mathbb{E}_\theta X = \frac{1}{\theta}$
3. $\text{Var}_\theta(X) = \frac{1-\theta}{\theta^2}$
4. $\log p_\theta(x) = \underbrace{(x-1)}_{T(x)} \underbrace{\log(1-\theta)}_{c(\theta)} - \underbrace{(-\log\theta)}_{d(\theta)}$
5. $s_\theta(x) = \frac{d}{d\theta} \log p_\theta(x) = -\frac{x-1}{1-\theta} + \frac{1}{\theta}$
6. $I(\theta) = \frac{1}{(1-\theta)\theta^2}.$

What is the correlation between random variables X and Y ?

When is correlation $\rho(X, Y)$ between X and Y equal to 1?

Which estimators for which class of distributions reach the Cramér-Rao lower bound?

What are the ways to find whether an estimator T reaches the Cramer-Rao lower bound?

$$s_\theta(x) = \frac{d}{d\theta}\log p_\theta(x) = -\frac{x-1}{1-\theta} + \frac{1}{\theta}$$

$$\begin{aligned}
\underbrace{I^{(n)}(\theta)}_{\text{for sample}} &= \operatorname{Var}\left(s_\theta^{(n)}\left(x_1,\ldots,x_n\right)\right) \\
&= \operatorname{Var}\left(\frac{d}{d\theta}\log p_\theta^{(n)}\left(x_1,\ldots,x_n\right)\right) \\
&= \operatorname{Var}\left(\frac{d}{d\theta}\sum_{i=1}^n \log p_\theta\left(x_i\right)\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^n s_\theta\left(x_i\right)\right) = \sum_{i=1}^n \operatorname{Var}\left(s_\theta\left(x_i\right)\right) = n \underbrace{I(\theta)}_{\text{single ob}}
\end{aligned}$$

1. Directly, by $I(\theta) = \operatorname{Var}_\theta\left(s_\theta(X)\right)$:

$$I(\theta) = \frac{\operatorname{Var}_\theta(X)}{(1-\theta)^2} = \frac{1}{(1-\theta)^2}\cdot\frac{1-\theta}{\theta^2} = \frac{1}{(1-\theta)\theta^2}$$

2. By $\mathbb{E}_\theta \dot{s}_\theta(x) = -I(\theta)$:

$$\dot{s}_\theta(x) = \frac{(x-1)}{(1-\theta)^2} - \frac{1}{\theta^2}$$

$$\mathbb{E}_\theta \dot{s}_\theta(X) = \frac{\frac{1}{\theta}-1}{(1-\theta)^2} - \frac{1}{\theta^2} = -\frac{1}{(1-\theta)\theta^2} = -I(\theta)$$

Assume the support of P_θ does not depend on θ, and the differentiability in L^2 condition holds. If $\operatorname{Var}_\theta(T) < \infty$ for all θ, then the Cramér-Rao lower bound is

$$\operatorname{Var}(T) \geq \frac{[\dot{g}(\theta)]^2}{I(\theta)}$$

We have

$$\text{CRLB} = \frac{\dot{g}(\theta)^2}{I(\theta)} = \frac{1}{\frac{n}{(1-\theta)\theta^2}} = \frac{(1-\theta)\theta^2}{n}.$$

Remember that the Fisher information grows linearly in the sample size n.

We have

$$\text{CRLB} = \frac{\dot{g}(\theta)^2}{I(\theta)} = \frac{\left[-\frac{1}{\theta^2}\right]^2}{\frac{n}{(1-\theta)\theta^2}} = \frac{(1-\theta)}{n\theta^2}.$$

Remember that the Fisher information grows linearly in the sample size n.

When there exist constants a, b such that $Y = aX + b$.

Correlation is defined as

$$\rho(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

Two ways:

1. Compute the bound by

$$\text{CRLB} = \frac{(\dot{g}(\theta))^2}{I(\theta)}$$

   and compare it to $\operatorname{Var}(T)$.

2. Check if the distribution is from the exponential family, and if $g(\theta) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}$

Assume the conditions of the Cramér-Rao lower bound hold with $s_\theta = \frac{d}{d\theta}\log p_\theta$. Let T be an unbiased estimator of $g(\theta)$, and suppose $\operatorname{Var}_\theta(T) = \frac{(\dot{g}(\theta))^2}{I(\theta)}$ for all θ. Then for all x and θ,

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$$

and $g(\theta) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}$.

For a random vector taking values in $Z \in \mathbb{R}^k$, what is the covariance Cov (Z) ?

What does it mean for a symmetric matrix $V \in \mathbb{R}^{k \times k}$ to be positive-definite?

PROOF

Show that a covariance matrix $\Sigma := \mathrm{Cov}(Z)$ is positive semi-definite.

Let $V \succ 0$ be a positive definite matrix. How is $\sqrt{V}$ defined? Given $\sqrt{V}$, how can you compute V ? (part-1)

Let $V \succ 0$ be a positive definite matrix. How is $\sqrt{V}$ defined? Given $\sqrt{V}$, how can you compute V ? (part-2)

Let $V \succ 0$ be a positive definite matrix. What is $V^{-\frac{1}{2}}$?

Write $\mathrm{Var}(X)$ *in terms of covariance.*

If X and Y are independent, what is $\mathrm{Cov}(X, Y)$ *?*

Is it true that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ *?*

Expand these statements into simpler forms:

1. $\mathrm{Cov}(aX, Y)$
2. $\mathrm{Cov}(X + c, Y)$
3. $\mathrm{Cov}(X + Y, Z)$

V is positive-definite, written as $V \succ 0$, if $a^T V a > 0$ for all $a \neq 0$.	Covariance is the matrix defined by $$\begin{aligned}\mathrm{Cov}(Z) &= \mathbb{E}\left(ZZ^T\right) - \mathbb{E}(Z)\mathbb{E}(Z)^T \\ &= \mathbb{E}\left((Z-\mu)(Z-\mu)^T\right)\end{aligned}$$
We can write $V = Q\Lambda Q^T$ where • $Q^T Q = I$ • $Q = (q_1, \ldots, q_k)$ is a matrix of eigenvectors q_i, and • $\Lambda = \begin{pmatrix} x_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & x_k \end{pmatrix}$ is a diagonal matrix of eigenvalues x_i.	Note that variance is non-negative, and we have $$\begin{aligned}\mathrm{Var}\left(a^T Z\right) &= \mathbb{E}\left(\left(a^T Z - a^T\mu\right)\left(a^T Z - a^T\mu\right)^T\right) \\ &= \mathbb{E}\left(a^T (Z-\mu)(Z-\mu)^T a\right) \\ &= a^T \mathbb{E}\left((Z-\mu)(Z-\mu)^T\right) a \\ &= a^T \Sigma a\end{aligned}$$ for all $a \in \mathbb{R}^k$.
$$V^{-\frac{1}{2}} := \left(V^{-1}\right)^{\frac{1}{2}} = \left(V^{\frac{1}{2}}\right)^{-1}$$	Then $\sqrt{V} := Q\sqrt{\Lambda}Q^T$ where $\sqrt{\Lambda} = \begin{pmatrix} \sqrt{x_1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sqrt{x_k} \end{pmatrix}$ We can find V by $V^{\frac{1}{2}} V^{\frac{1}{2}} = V$
$$\mathrm{Cov}(X, Y) = 0$$	$$\mathrm{Var}(X) = \mathrm{Cov}(X, X)$$
1. $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$ 2. $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$ 3. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$	Yes

Consider a multidimensional parameter space $\theta \in \Theta \subset \mathbb{R}^k$, and a function $g(\theta)$ for which we want to construct an estimator. How do you define

1. *partial derivative $\dot{g}(\theta)$*
2. *score vector $s_\theta(\cdot)$*
3. *Fisher information matrix $I(\theta)$?*

Consider a multidimensional parameter space $\theta \in \Theta \subset \mathbb{R}^k$, and a function $g(\theta)$. What is the Cramér-Rao lower bound on unbiased estimators of $g(\theta)$?

Consider a parameter space $\Theta = \{\theta_0, \theta_1\}$, an action space $\mathcal{A} = [0, 1]$, a risk

$$R(\theta, \phi) = \begin{cases} E_{\theta_0}\phi(X) & \text{if } \theta = \theta_0 \\ 1 - E_{\theta_1}\phi(X) & \text{if } \theta = \theta_1, \end{cases}$$

(continues next card)

(continued) and a Neyman-Pearson test

$$\phi_{NP}(X) = \begin{cases} 1 & p_1(X)/p_0(X) > c \\ q & p_1(X)/p_0(X) = c \\ 0 & p_1(X)/p_0(X) < c \end{cases}$$

for $c > 0, q \in [0, 1]$. Show that if $R(\theta, \phi_{NP}) \neq 0$, then ϕ_{NP} is admissible.

Proof

Prove for the discrete case that if a statistic $S(X)$ is sufficient then we can factorize the mass function as

$$p_\theta(x) = g_\theta(S(x))h(x)$$

for all x and θ

Proof

Prove for the discrete case that if we can factorize the mass function as

$$p_\theta(x) = g_\theta(S(x))h(x)$$

for all x and θ then $S(X)$ is a sufficient statistic.

Show that for $X_1, \ldots, X_n$ i.i.d. Uniform $[0, \theta]$ random variables with $\theta > 0$ $\max\{X_1, \ldots, X_n\}$ is a sufficient statistic.

If $p_\theta(x) = \exp[\theta T(x) - d(\theta)]h(x)$ is a probability density, then what is an equation for $d(\theta)$?

Show that if $p_\theta(x) = \exp[\theta T(x) - d(\theta)]h(x)$ then $\mathbb{E}_\theta T(X) = \dot{d}(\theta)$

Show that if $p_\theta(x) = \exp[\theta T(x) - d(\theta)]h(x)$ then $\mathrm{Var}_\theta\, T(X) = \ddot{d}(\theta)$. (part-1)

$$\mathrm{Var}_\theta(T) \geq \dot{g}(\theta)^T I(\theta)^{-1} \dot{g}(\theta)$$

1. $\dot{g}(\theta) := \begin{bmatrix} \delta g(\theta)/\delta\theta_1 \\ \vdots \\ \delta g(\theta)/\delta\theta_k \end{bmatrix}$
2. $s_\theta(\cdot) := \begin{bmatrix} \delta \log p_\theta/\delta\theta_1 \\ \vdots \\ \delta \log p_\theta/\delta\theta_k \end{bmatrix}$
3. $I(\theta) = \mathbb{E}_\theta s_\theta(X) s_\theta(X)^T = \mathrm{Cov}\left(s_\theta(x)\right)$

1. If $R(\theta_0, \phi) \leq R(\theta_0, \phi_{NP})$. Then
$$\begin{aligned} R(\theta_1, \phi) &= R(\theta_1, \phi) - R(\theta_1, \phi_{NP}) + R(\theta_1, \phi_{NP}) \\ &\geq c \underbrace{\left[R(\theta_0, \phi_{NP}) - R(\theta_0, \phi)\right]}_{\geq 0} + R(\theta_1, \phi_{NP}) \end{aligned}$$
2. If $R(\theta_1, \phi) \leq R(\theta_1, \phi_{NP})$. Then
$$\begin{aligned} R(\theta_0, \phi) &= R(\theta_0, \phi) - R(\theta_0, \phi_{NP}) + R(\theta_0, \phi_{NP}) \\ &\geq \frac{1}{c} \underbrace{\left[R(\theta_1, \phi_{NP}) - R(\theta_1, \phi)\right]}_{\geq 0} + R(\theta_0, \phi_{NP}) \end{aligned}$$

We simply need to use the Neyman-Pearson lemma in form

- $\left[R(\theta_1, \phi) - R(\theta_1, \phi_{NP})\right] \geq c\left[R(\theta_0, \phi_{NP}) - R(\theta_0, \phi)\right]$, or
- $\left[R(\theta_1, \phi_{NP}) - R(\theta_1, \phi)\right] \leq c\left[R(\theta_0, \phi) - R(\theta_0, \phi_{NP})\right]$

Suppose $S(x) = s$, and $p_\theta(x) = g_\theta(S(x))h(x)$. Then
$$\begin{aligned} P_\theta(X = x \mid S = s) &= \frac{P_\theta(X = x)}{P_\theta(S = s)} \\ &= \frac{g_\theta(S)h(x)}{\sum_{\tilde{x}:S(\tilde{x})=s} g_\theta(s)h(\tilde{x})} \\ &= \frac{h(x)}{\sum_{\tilde{x}:S(\tilde{x})=s} h(\tilde{x})} \end{aligned}$$
does not depend on θ.

Suppose S is sufficient, and that $S(x) = s$.
$$\begin{aligned} p_\theta(x) &= P_\theta(X = x) = P(X = x \mid S = s) P_\theta(S = s) \\ &:= h(x) g_\theta(s) \end{aligned}$$

$d(\theta)$ is a normalizing constant, so we can write
$$d(\theta) = \log\left(\int \exp[\theta T(x)] h(x) dx\right)$$

If we write the density function using indicator functions, we get
$$\begin{aligned} \prod_{i=1}^n p_\theta(x_i) = \prod_{i=1}^n \frac{1_{[0,\theta]}(x)}{\theta} &= \frac{1\left\{0 \leq x_i \leq \theta \forall i\right\}}{x^n} \\ &= \underbrace{1\left\{\min\left\{x_1, \ldots, x_n\right\} \geq 0\right\}}_{h(x_1,\ldots,x_n)} \underbrace{\frac{1\left\{\max\left\{x_1, \ldots, x_n\right\} \leq \theta\right\}}{\theta^n}}_{g_\theta(\max\{x_1,\ldots,x_n\})} \end{aligned}$$
Since we can factorize the density this way, by the Neyman Factorization theorem $\max\left\{X_1, \ldots, X_n\right\}$ is a sufficient statistic.

Since $d(\theta)$ is a normalizing constant, we can write
$$\begin{aligned} d(\theta) &= \log\left(\int \exp[\theta T(x)] h(x) dx\right) \\ &= \log\left(\int \exp[\theta T] h\right) \\ \dot{d}(\theta) &= \frac{1}{\int \exp[\theta T] h} \int \exp[\theta T] T h. \end{aligned}$$

Since $d(\theta)$ is a normalizing constant, we can write
$$\begin{aligned} d(\theta) &= \log\left(\int \exp[\theta T(x)] h(x) dx\right) \\ &= \log\left(\int \exp[\theta T] h\right) \\ \dot{d}(\theta) &= \frac{1}{\int \exp[\theta T] h} \int \exp[\theta T] T h \\ &= \int \exp[\theta T - d(\theta)] T h \\ &= \int p_\theta T = \mathbb{E}_\theta T(X). \end{aligned}$$

Show that if $p_\theta(x) = \exp[\theta T(x) - d(\theta)]h(x)$ then $\mathrm{Var}_\theta\, T(X) = \ddot{d}(\theta)$. *(part-2)*

Proof

Prove that if the support of a density p_θ does not depend on θ, and if there exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(X) < \infty$ and

$$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{hp\theta(x)} - s_\theta(x) \right]^2 = 0$$

for all θ, and T is an unbiased estimator of $g(\theta)$ with finite variance, then $\dot{g}(\theta) = \mathrm{Cov}_\theta\,(T, s_\theta(x))$ (part-1)

Proof

Prove that if the support of a density p_θ does not depend on θ, and if there exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(X) < \infty$ and

$$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{hp\theta(x)} - s_\theta(x) \right]^2 = 0$$

for all θ, and T is an unbiased estimator of $g(\theta)$ with finite variance, then $\dot{g}(\theta) = \mathrm{Cov}_\theta\,(T, s_\theta(x))$ (part-2)

Proof

Prove that if the support of a density p_θ does not depend on θ, and if there exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(X) < \infty$ and

$$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{hp\theta(x)} - s_\theta(x) \right]^2 = 0$$

for all θ, and T is an unbiased estimator of $g(\theta)$ with finite variance, then $\dot{g}(\theta) = \mathrm{Cov}_\theta\,(T, s_\theta(x))$ (part-3)

Proof

Sketch proof that if the support of a density p_θ does not depend on θ, and if there exists a function $s_\theta : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_\theta s_\theta^2(X) < \infty$ and

$$\lim_{h \to 0} \mathbb{E}_\theta \left[\frac{p_{\theta+h}(x) - p_\theta(x)}{hp\theta(x)} - s_\theta(x) \right]^2 = 0$$

for all θ, and T is an unbiased estimator of $g(\theta)$ with finite variance, then

$$\mathrm{Var}_\theta(T) \geq \frac{[\dot{g}(\theta)]^2}{I(\theta)}$$

Suppose the support of P_θ does not depend on θ, and the differentiability in L^2 condition holds. Further assume that T is an unbiased estimator of $g(\theta)$ and it reaches the Cramér-Rao lower bound. What can you say about $p_\theta(x)$? What about $g(\theta)$?

Definition

What is a pivot for a test?

How is a pivot $Z(X, \gamma)$ used to construct a test?

Let X be distributed according to a known symmetric distribution function F_0 shifted by a parameter $\theta = \mu$. If $\hat{\mu}$ is an equivariant estimator of μ, what is a pivot function you could take to test if $\mu = \mu_0$?

Let $X_1, \ldots, X_n$ be distributed i.i.d. according to a distribution function from family $\mathcal{F}_0 = \{F_0(\cdot) = \Phi(\cdot/\sigma) : \sigma > 0\}$ shifted by a parameter $\theta = \mu$. What is a pivot function you could take to test if $\mu = \mu_0$?

$$\begin{aligned}\frac{g(\theta+h)-g(\theta)}{h} &= \frac{\mathbb{E}_{\theta+h}T-\mathbb{E}_\theta T}{h} = \int T\frac{p_{\theta+h}-p_\theta}{h} \\ &= \int T\left[\frac{(p_{\theta+h}-p_\theta)\,p_\theta}{hp_\theta}\right] \\ &= \underbrace{\int T\left[\frac{(p_{\theta+h}-p_\theta)\,p_\theta}{hp_\theta}-s_\theta p_\theta\right]}_{(1)} \\ &+ \underbrace{\int Ts_\theta p_\theta}_{(2)}\left(-\mathbb{E}_\theta Ts_\theta+\mathbb{E}_\theta Ts_\theta\right)\end{aligned}$$

$$\begin{aligned}\ddot{d}(\theta) &= \frac{1}{\underbrace{\int\exp[\theta T]h}_{e^{d(\theta)}}}\int\exp[\theta T]T^2h \\ &- \frac{1}{\underbrace{\left(\int\exp[\theta T]h\right)^2}_{\left(e^{d(\theta)}\right)^2}}\left(\int\exp[\theta T]Th\right)^2 \\ &= \int\exp[\theta T-d(\theta)]hT^2-\left(\int p_\theta(T)\right)^2 \\ &= \mathbb{E}_\theta T^2-(\mathbb{E}_\theta T)^2=\operatorname{Var}_\theta(T)\end{aligned}$$

2. This term is the covariance because

$$\begin{aligned}\operatorname{Cov}_\theta\left(T,s_\theta(X)\right) &= \mathbb{E}_\theta Ts_\theta-(\mathbb{E}_\theta T)\underbrace{(\mathbb{E}_\theta s_\theta)}_{=0} \\ &= \mathbb{E}_\theta Ts_\theta=\int Ts_\theta p_\theta.\end{aligned}$$

$$\begin{aligned}\dot{g}(\theta) &= \lim_{h\to 0}\frac{g(\theta+h)-g(\theta)}{h} \\ &= \mathbb{E}_\theta Ts_\theta=\operatorname{Cov}_\theta\left(T,s_\theta(X)\right).\end{aligned}$$

Now we have

1. By squaring (1) we can use Cauchy-Schwarz bound on it.

$$\left(\int T\left[\frac{(p_{\theta+h}-p_\theta)\,p_\theta}{hp_\theta}-s_\theta p_\theta\right]\right)^2\le$$
$$\underbrace{\left(\int T^2p_\theta\right)}_{<\infty\text{ by finite var.}}\underbrace{\left(\int\left[\frac{p_{\theta+h}-p_\theta}{hp_\theta}-s_\theta\right]^2p_\theta\right)}_{\xrightarrow[h\to 0]{}0\text{ by assum. on }s_\theta}$$

By a lemma we know that

1. $p_\theta(x)=\exp[c(\theta)T(X)-d(\theta)]h(x)$
2. that $c(\theta)$ and $d(\theta)$ are differentiable, and
3. that $g(\theta)=\dot{d}(\theta)/\dot{c}(\theta)$ for all θ

It is possible for us to show that $\dot{g}(\theta)=\operatorname{Cov}_\theta\left(T,s_\theta(X)\right)$. From this point we can just use Cauchy-Schwarz

$$\operatorname{Cov}_\theta\left(T,s_\theta(X)\right)^2\le\operatorname{Var}_\theta(T)\operatorname{Var}_\theta\left(s_\theta(X)\right)$$

Hence

$$\left[\dot{g}(\theta)\right]^2\le\operatorname{Var}_\theta(T)I(\theta)$$

so we get the desired

$$\operatorname{Var}_\theta(T)\ge\frac{[\dot{g}(\theta)]^2}{I(\theta)}$$

Let $G(\cdot)$ be the distribution

$$G(\cdot):=P_\theta(Z(X,g(\theta))\le\cdot).$$

By definition of pivot, it does not depend on X. Now for a test $g(\theta)\neq\gamma_0$ of level α we compute the critical values

$$q_L:=q^G_{\sup}\left(\frac{\alpha}{2}\right),q_R:=q^G_{\inf}\left(1-\frac{\alpha}{2}\right)$$

and define the test

$$\phi\left(X,\gamma_0\right):=\begin{cases}1 & \text{if }Z\left(X,\gamma_0\right)\notin[q_L,q_R]\\ 0 & \text{otherwise.}\end{cases}$$

A pivot is a function $Z(X,\gamma)$ depending on data X and on the parameter γ, such that for all $\theta\in\Theta$, the distribution

$$P_\theta(Z(X,g(\theta))\le\cdot)=:G(\cdot)$$

does not depend on θ.

Recall that $\bar{X}_n$ is equivariant, so the distribution of $\bar{X}_n-\mu$ does not depend on μ. To make it a pivot, we need to make it not depend on σ as well. So, we take

$$Z(X,\mu):=\frac{\sqrt{n}\left(\bar{X}_n-\mu\right)}{S_n}$$

where S_n^2 is the sample variance. Then $Z\left(X,\mu_0\right)$ is distributed according to the Student distribution with $n-1$ degrees of freedom.

We could take $Z(X,\mu):=\hat{\mu}-\mu$ as the pivot. By equivariance, this function has a distribution G depending only on F_0.

Let $X_1, \ldots, X_n$ be distributed i.i.d. according to distribution function from family $\mathcal{F}_0 = \{F_0$ symmetric and continuous at $x = 0\}$ shifted by a parameter $\theta = \mu$. What is a pivot function you could take to test if $\mu = \mu_0$?	*Consider two samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \cdot\mathcal{N}(\mu, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \cdot\mathcal{N}(\mu + \gamma, \sigma^2)$. Find a pivot for testing $\gamma = \gamma_0$. How is such a test called? (part-1)*
Consider two samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \cdot\mathcal{N}(\mu, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \cdot\mathcal{N}(\mu + \gamma, \sigma^2)$. Find a pivot for testing $\gamma = \gamma_0$. How is such a test called? (part-2)	*Consider two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ which are sampled i.i.d. from distributions with distribution functions $F_X(\cdot)$ and $F_Y(\cdot) = F_X(\cdot - \gamma)$ correspondingly. Find a pivot for testing $\gamma = \gamma_0$. How is such a test called?*
Let $V \in \mathbb{R}^{k \times k}$ be positive definite. Let $c \in \mathbb{R}^k$ Show that $$\max_{a \in \mathbb{R}^k} \frac{\left(a^T c\right)^2}{a^T V a} = c^T V^{-1} c$$	*What is the Cramér-Rao lower bound for an unbiased estimator $T \in \mathbb{R}$ if the parameter space $\theta \subset \mathbb{R}^k$ is multidimensional?*
What's a lower bound A such that for the covariance matrix Σ_θ of an unbiased estimator T we have $$\Sigma_\theta - A \succeq 0?$$	*How is the ith diagonal element of the Fisher information matrix $I^{ii}(\theta)$ related to the ith diagonal element of its inverse $I_{ii}(\theta)^{-1}$?*
What is the power of a test ϕ ?	*If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are i.i.d. sequences of random variables with distributions $F_X \sim \mathcal{N}(\mu, \sigma^2)$ and $F_Y \sim \mathcal{N}(\mu + \gamma, \sigma^2)$ respectively. How is $\bar{Y} - \bar{X}$ distributed?*

Note that $\bar{Y}-\bar{X}\sim\mathcal{N}\left(\gamma,\frac{\sigma^2}{m}+\frac{\sigma^2}{n}\right)$ $$\frac{\bar{Y}-\bar{X}-\gamma}{\sigma\sqrt{\frac{m+n}{mn}}}\sim\mathcal{N}(0,1).$$ Now to find the pivot itself, we need to find a distribution not depending on σ. So, we replace σ by the pooled sample variance $$S^2:=\frac{1}{m+n-2}\left[\sum_{i=1}^{n}\left(X_1-\bar{X}\right)^2+\sum_{j=1}^{m}\left(Y_j-\bar{Y}\right)^2\right]$$	We could take $$Z(X,\mu):=\sum_{i=1}^{n}1\left\{X_i\geq\mu\right\}$$ which is equivariant for μ. G is then distributed according to the Binomial(n,p) distribution with parameter $p=1/2$, which does not depend on μ, so it's a pivot.
Let $N=n+m$, and then $Z_1,\ldots,Z_N=(X_1,\ldots,X_n,Y_1,\ldots,Y_m)$. Let R_i be the rank of Z_i if we were to sort all of the Z_i 's. Then the pivot is $$Z(X,Y,\gamma)=\sum_{i=1}^{n}R_i$$ This is the Wilcoxon test.	Now finally our pivot is $$Z(X,Y,\gamma):=\sqrt{\frac{mn}{m+n}}\left[\frac{\bar{Y}-\bar{X}-\gamma}{S}\right].$$ Its distribution G is Student $(n+m-2)$. This is the so called Student's test.
$$\mathrm{Var}(T)\geq\dot{g}(\theta)^T I(\theta)^{-1}\dot{g}(\theta)$$	Let $b:=V^{\frac{1}{2}}a, d:=V^{-\frac{1}{2}}c$. Then $a^TVa=b^Tb=\Vert b\Vert^2$ and $a^Tc=b^Td$. Now we write $$\frac{\left(a^Tc\right)^2}{a^TVa}=\underbrace{\frac{\left\vert b^Td\right\vert^2}{\Vert b\Vert^2}\leq\frac{\Vert b\Vert^2\Vert d\Vert^2}{\Vert b\Vert^2}}_{\text{Cauchy-Schwarz}}=\Vert d\Vert^2=d^Td=c^TV^{-1}c$$
$$I^{ii}(\theta)\geq I_{ii}(\theta)^{-1}$$	It's the inverse of the Fisher information $I(\theta)^{-1}$, so we have $$\Sigma_\theta-I(\theta)^{-1}\succeq 0$$
It's $$\bar{Y}-\bar{X}\sim\mathcal{N}\left(\gamma,\frac{\sigma^2}{m}+\frac{\sigma^2}{n}\right)$$	The power of a test $\phi\left(x,\gamma_0\right)$ is $\mathbb{P}_\theta\left(\phi\left(X,\gamma_0\right)=1\right)$ for $g(\theta)\neq\gamma_0$.

Definition *What is a randomized test?*	*How are non-randomized tests related to randomized tests?*
What is the risk of a randomized test ϕ when $\theta \in \{\theta_0, \theta_1\}$?	Definition *What is a Neyman Pearson test?*
Definition *What is a Neyman Pearson test? (plot)*	Lemma *State the Neyman-Pearson lemma.*
Proof *Sketch a proof of the Neyman-Pearson lemma (part-1)*	Proof *Sketch a proof of the Neyman-Pearson lemma (part-2)*
What is the level of a test?	*What does it mean for a test to be uniformly most powerful?*

A non-randomized test is a special case of randomized tests where $\phi \in \{0,1\}$, and $\mathbb{E}(\phi(X)) = \mathbb{P}(\phi(X) = 1)$

A randomized test at level α is a statistic $\phi : x \to [0,1]$ such that $\mathbb{E}(\phi(X)) \leq \alpha$ for ϕ under the null hypothesis.

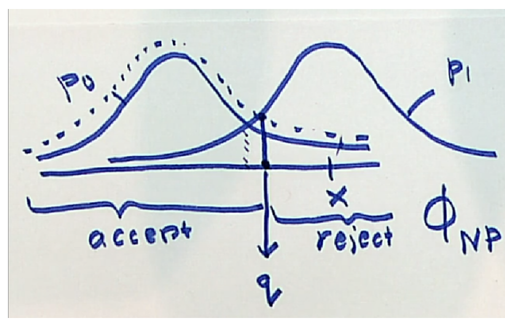
Consider the problem of testing whether $\theta = \theta_0$ or $\theta = \theta_1$. Let $p_0\,(p_1)$ be the density of $P_{\theta_0}\,(P_{\theta_1})$ with respect to some dominating measure ν (for example $\nu = P_{\theta_0} + P_{\theta_1}$). A Neyman Pearson test is then

$$\phi_{\mathrm{NP}}(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > c \\ q & \text{if } \frac{p_1(x)}{p_0(x)} = c \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < c \end{cases}$$

$$R(\theta,\phi) = \begin{cases} \mathbb{E}_{\theta_0}\phi(x) & \text{if } \theta = \theta_0 \\ 1 - \mathbb{E}_{\theta_1}\phi(x) & \text{if } \theta = \theta_1 \end{cases}$$

If $\theta \in \{\theta_0, \theta_1\}$, then for any test ϕ and for a Neyman Perason test ϕ_{NP} with threshold c the risks fulfil the inequality

$$[R(\theta_1,\phi) - R(\theta_1,\phi_{\mathrm{NP}})] \geq c\,[R(\theta_0,\phi_{\mathrm{NP}}) - R(\theta_0,\phi)]$$



$$\begin{aligned} &\geq \int_{\frac{p_1}{p_0}>c} [\phi_{\mathrm{NP}} - \phi]\, c p_{\theta_0} + \int_{\frac{p_1}{p_0}=c} [\phi_{\mathrm{NP}} - \phi]\, c p_{\theta_0} \\ &+ \int_{\frac{p_1}{p_0}<c} [\phi_{\mathrm{NP}} - \phi]\, c p_{\theta_0} \\ &= c\,[\mathbb{E}_{\theta_0}\phi_{\mathrm{NP}}(x) - \mathbb{E}_{\theta_0}\phi(x)] \\ &= c\,[R(\theta_0,\phi_{\mathrm{NP}}) - R(\theta_0,\phi)]\,. \end{aligned}$$

$$\begin{aligned} R(\theta_1,\phi) - R(\theta_1,\phi_{\mathrm{NP}}) &= \mathbb{E}_{\theta_1}\phi_{\mathrm{NP}}(x) - \mathbb{E}_{\theta_1}\phi(x) \\ &= \int [\phi_{\mathrm{NP}} - \phi]\, p_{\theta_1} \\ &= \int_{\frac{p_1}{p_0}>c} \underbrace{[\phi_{\mathrm{NP}} - \phi]}_{1-\phi\geq 0}\, \underbrace{p_{\theta_1}}_{\geq c p_0} \\ &+ \int_{\frac{p_1}{p_0}=c} [\phi_{\mathrm{NP}} - \phi]\, \underbrace{p_{\theta_1}}_{=c p_0} + \int_{\frac{p_1}{p_0}<c} \underbrace{[\phi_{\mathrm{NP}} - \phi]}_{0-\phi\leq 0}\, \underbrace{p_{\theta_1}}_{<c p_0} \end{aligned}$$

Consider $\theta \in \Theta$ in a parameter space with a null hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$ and alternative hypothesis $H_1 : \theta \in \Theta_1 \subset \Theta$ which are disjoint. A test ϕ is uniformly most powerful (UMP) at level α if level$(\phi) \leq \alpha$ and

$$\mathbb{E}_\theta \phi(x) = \sup\{\mathbb{E}_\theta \phi' : \text{level}(\phi') \leq \alpha\}, \forall_{\theta\in\Theta_1}.$$

So, it attains the supremum of power over all tests of this level.

Consider $\theta \in \Theta$ in a parameter space with a null hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$ and alternative hypothesis $H_1 : \theta \in \Theta_1 \subset \Theta$ which are disjoint. Then the level of a test $\phi \in [0,1]$ is

$$\text{level}(\phi) := \sup_{\theta\in\Theta_0} \underbrace{\mathbb{E}_\theta \phi(x)}_{\substack{\text{Probability of error} \\ \text{of first kind}}}\,.$$

What does it mean for a test ϕ to be unbiased?

Definition

What is a uniformly most powerful unbiased (UMPU) test?

Lemma

State the Rao Blackwell lemma.

Proof

Let $d : \mathcal{X} \to \mathcal{A}$ be a decision and $d^* := \mathbb{E}(d(X) \mid S)$ where S is sufficient. Suppose $\mathcal{A}$ is a convex subset of $\mathbb{R}^p$, and $a \mapsto L(\theta, a)$ is convex for all θ. Then for all θ

$$R(\theta, d^*) \leq R(\theta, d)$$

Definition

What does it mean for an estimator T to be location equivariant?

What are some examples of location equivariant estimators?

What does it mean for a loss function L to be called location invariant?

Show that under a location invariant loss, the risk of a location equivariant estimator does not depend on its location parameter.

Definition

What is a uniform minimum risk equivariant (UMRE) estimator?

Lemma

State lemma about construction of UMRE estimators.

An UMPU test is uniformly most powerful among all ϕ' unbiased with level $(\phi') \leq \alpha$	A test ϕ is called unbiased if its power is at least the level, in other words if for all $\theta_1 \in \Theta_1$ $$\mathbb{E}_{\theta_1}\phi(X) \geq \sup_{\theta \in \Theta_0} \mathbb{E}_\theta \phi(X) = \text{level}(\phi)$$
$$\begin{aligned} R(\theta, d) &= \mathbb{E}_\theta L(\theta, d(X)) \\ &= \mathbb{E}_\theta \mathbb{E}(L(\theta, d(X)) \mid S) \text{ (Iterated expectations)} \\ &\geq \mathbb{E}_\theta L(\theta, \mathbb{E}(d(X) \mid S)) \quad \text{(Jensen)} \\ &= \mathbb{E}_\theta L\left(\theta, d^*(S)\right) \\ &= R\left(\theta, d^*\right) \end{aligned}$$	Let $d: \mathcal{X} \rightarrow \mathcal{A}$ be a decision and $d^* := \mathbb{E}(d(X) \mid S)$ where S is sufficient. Suppose $\mathcal{A}$ is a convex subset of $\mathbb{R}^p$, and $a \mapsto L(\theta, a)$ is convex for all θ. Then for all θ $$R\left(\theta, d^*\right) \leq R(\theta, d)$$
Mean, median.	An estimator T is called location equivariant if $T(x+c) = T(x) + c$ for all $c \in \mathbb{R}, x \in \mathbb{R}^n$
$$\begin{aligned} R(\theta, T) &= \mathbb{E}_\theta L(\theta, T(X)) \\ &= \mathbb{E}_\theta L_0(T(X) - \theta) \quad \text{(Loss invariance)} \\ &= \mathbb{E}_\theta L_0(T(X - \theta)) \quad \text{(Estimator equivariance)} \\ &= \mathbb{E}_\theta L_0(T(\varepsilon)) \\ &= \mathbb{E}_0 L_0(T(\varepsilon)) \\ &= R(0, T). \end{aligned}$$	A loss function $L: \underbrace{\mathbb{R}}_{\Theta} \times \underbrace{\mathbb{R}}_{\mathcal{A}}$ is called location invariant if for all $(\theta, a) \in \mathbb{R}^2$, $$L(\theta, a) = L(\theta + c, a + c).$$ We can also write $$L(\theta, a) = L(0, a - \theta) =: L_0(a - \theta)$$
Let T be an equivariant estimator, $Y_i := X_i - X_n, i = 1, \ldots, n, Y := (Y_1, \ldots, Y_n)$, and define $$T^*(Y) = \arg\min_v \mathbb{E}\left(L_0\left(v + \varepsilon_n\right) \mid Y\right).$$ Moreover, let $T^*(X) := T^*(Y) + X_n$. Then T^* is UMRE.	An estimator T is called uniform minimum risk equivariant if it is equivariant and $$R(0, T) = \min\left\{R\left(0, T'\right) : T' \text{ equivariant }\right\}$$

Lemma	Proof
State Basu's lemma.	*Consider a random variable $X \sim P_\theta, \theta \in \Theta$. Let $T = T(X)$ be a sufficient and complete statistic, and let $Y = Y(X)$ have a distribution not depending on θ. Then T and Y are independent.*
If an estimator T is unbiased, sufficient, complete, equivariant, what other property does it have?	*Let $X_1, \ldots, X_n$ be i.i.d. variables distributed according to $\mathcal{N}(\mu, \sigma^2)$. Assume also that $\sigma^2 = \sigma_0^2$ is known. Show that $T(X) = \bar{X}$ is an UMRE estimator of μ*
Let $X_1, \ldots, X_n$ be i.i.d. variables distributed according to $\mathcal{N}(\mu, \sigma^2)$. Assume also that $\sigma^2 = \sigma_0^2$ is known. What is $\mathbb{E}_\theta[\bar{X} \mid X - \bar{X}]$? Why?	*What does it mean for an estimator d' to be strictly better than d ?*
What does it mean for an estimator d to be admissible?	*What does it mean that $\tilde{P} >> P$?*
What does it mean for a decision d to be minimax?	*Show that if for a Neyman-Pearson test ϕ_{NP} it's true that* $$R(\theta_0, \phi_{NP}) = R(\theta_1, \phi_{NP}),$$ *then ϕ_{NP} is minimax.*

1. Label $\mathcal{Y}$ to be the space of values of Y, and take an arbitrary subset. $A \subset \mathcal{Y}$.
2. Define $h(T) := \mathbb{P}(Y \in A \mid T) - \mathbb{P}(Y \in A)$.
3. By iterated expectations, $\mathbb{E}_\theta h(T) = 0$ for all θ.
4. By completeness of $T, h(T) = 0$ almost surely, so

$$\mathbb{P}(Y \in A \mid T) - \mathbb{P}(Y \in A) = 0$$

almost surely.

5. Thus, T and Y are independent.

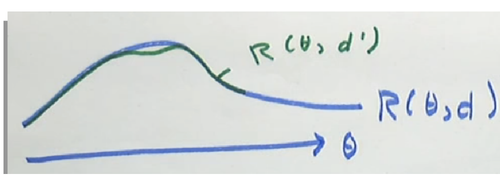
Consider a random variable $X \sim P_\theta, \theta \in \Theta$. Let $T = T(X)$ be a sufficient and complete statistic, and let $Y = Y(X)$ have a distribution not depending on θ. Then T and Y are independent.

Note that T is unbiased and equivariant. We know it is also sufficient and complete from properties of exponential family distributions. From these four properties we can conclude it is UMRE.

It is also uniform minimum risk equivariant (UMRE).

An estimator d' is called strictly better than d if

1. $R(\theta, d') \leq R(\theta, d)$ for all θ, and
2. $R(\theta, d') < R(\theta, d)$ for at least one θ.



Since $\bar{X}$ is sufficient and complete, we can use Basu's lemma to conclude

$$\mathbb{E}_\theta[\bar{X} \mid X - \bar{X}] = \mathbb{E}_\theta \bar{X} = \mu$$

$\tilde{P} >> P$ means "$\tilde{P}$ dominates P", so

$$\tilde{P}(A) = 0 \Longrightarrow P(A) = 0$$

An estimator d is admissible if there is no strictly better estimator d' than it.

Suppose $R(\theta_0, \phi_{NP}) = R(\theta_1, \phi_{NP}) = r$. Consider another test ϕ' with

$$r' = \max\left\{R(\theta_0, \phi'), R(\theta_1, \phi')\right\}.$$

By Neyman-Pearson lemma

$$\begin{aligned}
[r - R(\theta_1, \phi)] &\leq c\,[R(\theta_0, \phi) - r] \\
r(1+c) &\leq cR(\theta_0, \phi) + R(\theta_1, \phi) \leq r'(1+c) \\
r &\leq r'
\end{aligned}$$

so ϕ_{NP} is minimax.

A decision d is minimax if

$$\sup_{\theta \in \Theta} R(\theta, d) = \inf_{d'} \sup_{\theta \in \Theta} R(\theta, d')$$

Suppose that a Neyman-Pearson test ϕ_{NP} is minimax. Show that then

$$R(\theta_0, \phi_{NP}) = R(\theta_1, \phi_{NP})$$

DEFINITION

What is a Bayes risk?

DEFINITION

What is a Bayes estimator?

Consider $\Theta = \{\theta_0, \theta_1\}$ with prior probabilities $w(\theta_0) = w_0, w(\theta_1) = w_1 = 1 - w_0$. The risk of a decision ϕ is then

$$R(\theta, \phi) = \begin{cases} \mathbb{E}_{\theta_0}\phi(X) & \text{if } \theta = \theta_0 \\ 1 - \mathbb{E}_{\theta_1}\phi(X) & \text{if } \theta = \theta_1 \end{cases}$$

Derive a decision $\phi : \mathcal{X} \to [0,1]$ which minimizes the Bayes risk. (part-1)

Consider $\Theta = \{\theta_0, \theta_1\}$ with prior probabilities $w(\theta_0) = w_0, w(\theta_1) = w_1 = 1 - w_0$. The risk of a decision ϕ is then

$$R(\theta, \phi) = \begin{cases} \mathbb{E}_{\theta_0}\phi(X) & \text{if } \theta = \theta_0 \\ 1 - \mathbb{E}_{\theta_1}\phi(X) & \text{if } \theta = \theta_1 \end{cases}$$

Derive a decision $\phi : \mathcal{X} \to [0,1]$ which minimizes the Bayes risk. (part-2)

LEMMA

State the Bayes estimator construction lemma.

PROOF

Given data $X = x$, consider θ as a random variable with density $w(\vartheta \mid x)$. Let

$$\begin{aligned} l(x, a) : &= \mathbb{E}(L(\theta, a) \mid X = x) \\ &= \int_\Theta L(\vartheta, a) w(\vartheta \mid x) d\mu(\vartheta), \end{aligned}$$

$d(x) := \arg\min_a l(x, a)$. Then d is the Bayes decision d_{Bayes}.

What is a maximum a posteriori estimator?

What are the posterior, likelihood, and prior?

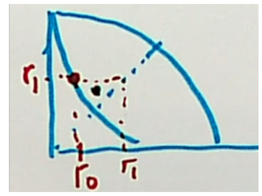
Consider the action space $\mathcal{A} = \mathbb{R}$, parameter space $\Theta \subset \mathbb{R}$, and loss function $L(\theta, a) = (\theta - a)^2$. What is d_{Bayes} for θ ? Why?

Let $\Theta \subset \mathbb{R}^k$. Let $w(\theta)$ be given weights such that $\int_\Theta w(\theta) d\theta = 1$. The Bayes risk of a decision d is then

$$r_w(d) = \int_\Theta R(\theta, d) d\mu(\theta).$$

So, it's the expected risk when θ is a random variable with density w.

Let $S = \{(R(\theta_0, \phi), R(\theta_1, \phi)) : \phi : \mathcal{X} \to [0,1]\}$. Note that S is convex (see image). Thus, if $r_0 < r_1$, we can find a test ϕ with $r_0 < r_0' < r_1$ and $r_1' < r_1$. So then ϕ_{NP} is not minimax. Similarly for $r_0 > r_1$.



$$\begin{aligned} r_w(\phi) &= w_0 \mathbb{E}_{\theta_0} \phi(X) + w_1 \left(1 - \mathbb{E}_{\theta_1 \phi(X)}\right) \\ &= \int (w_0 p_0 - w_1 p_1) \phi + w_1 \\ &= \int_{w_0 p_0 - w_1 p_1 > 0} (w_0 p_0 - w_1 p_1) \phi \\ &+ \int_{w_0 p_0 - w_1 p_1 = 0} (w_0 p_0 - w_1 p_1) \phi \\ &+ \int_{w_0 p_0 - w_1 p_1 < 0} (w_0 p_0 - w_1 p_1) \phi + w_1 \end{aligned}$$

A Bayes estimator is the one which minimizes the Bayes risk for some prion

$$d_{\text{Bayes}} = \arg\min_{d'} r_w(d')$$

Given data $X = x$, consider θ as a random variable with density $w(\vartheta \mid x)$. Let

$$l(x, a) := \mathbb{E}(L(\theta, a) \mid X = x) = \int_\Theta L(\vartheta, a) w(\vartheta \mid x) d\mu(\vartheta)$$

and

$$d(x) := \arg\min_a l(x, a)$$

Then d is the Bayes decision d_{Bayes} .

So the optimal decision is

$$\phi_{\text{Bayes}} = \begin{cases} 1 & \text{if } p_1/p_0 > w_0/w_1 \\ q & \text{if } p_1/p_0 = w_0/w_1 \\ 0 & \text{if } p_1/p_0 < w_0/w_1 \end{cases}$$

The maximum a posteriori estimator is

$$\theta_{\text{MAP}}(x) = \arg\max_{\vartheta \in \Theta} w(\vartheta \mid x)$$

Let d' be some decision.

$$\begin{aligned} r_w(d') &= \textstyle\int_\Theta R(\vartheta, d') w(\vartheta) d\mu(\vartheta) \\ &= \mathbb{E}(R(\vartheta, d')) \\ &= \mathbb{E}(\mathbb{E}(L(\vartheta, d') \mid \vartheta)) \\ &= \mathbb{E}(L(\vartheta, d')) && \text{(Iterated } \mathbb{E}) \\ &= \mathbb{E}(\mathbb{E}(L(\vartheta, d') \mid X)) && \text{(Rev. iterated } \mathbb{E}) \\ &= \mathbb{E}(l(X, d')) \\ &\geq \mathbb{E}(l(X, d)) \\ &= r_w(d) \quad d = \arg\min_{d'} \cdots \end{aligned}$$

$d_{\text{Bayes}}(x) = \mathbb{E}(\theta \mid x)$. This is because we are minimizing

$$l(x, a) = \mathbb{E}\left([\theta - a]^2 \mid X = x\right)$$

and the minimum is simply

$$\min_{a \in \mathbb{R}} l(x, a) = \mathbb{E}(\theta \mid X = x)$$

In the Bayesian context, we have

$$\underbrace{w(\vartheta \mid x)}_{\text{posterior}} \propto \underbrace{p(x \mid \vartheta)}_{\text{likelihood}} \underbrace{w(\vartheta)}_{\text{prior}}$$

Show that for quadratic loss, the Bayes risk of any estimator can be written as

$$r_w(T') = r_w(T_{\text{Bayes}}) + \mathbb{E}\left([T_{\text{Bayes}}(X) - T'(X)]^2\right)$$

Consider $X \mid \theta \sim \text{Poisson}(\theta), \theta \sim \text{Gamma}(k, \lambda)$ with k, λ given. The mass functions are

1. $p(x \mid \theta) = e^{-\theta}\frac{\theta^x}{x!}$ for $x \in \{0, 1, \ldots\}$
2. $w(\vartheta) = e^{-\lambda\vartheta}\vartheta^{k-1}\lambda^k/\Gamma(k)$

Also note that $\mathbb{E}(\theta) = \frac{k}{\lambda}$. What is the Bayes estimator for the mean squared error loss? What is the maximum a posteriori estimator for θ ? (part-1)

Consider $X \mid \theta \sim \text{Poisson}(\theta), \theta \sim \text{Gamma}(k, \lambda)$ with k, λ given. The mass functions are

1. $p(x \mid \theta) = e^{-\theta}\frac{\theta^x}{x!}$ for $x \in \{0, 1, \ldots\}$
2. $w(\vartheta) = e^{-\lambda\vartheta\vartheta}\vartheta^{k-1}\lambda^k/\Gamma(k)$

Also note that $\mathbb{E}(\theta) = \frac{k}{\lambda}$. What is the Bayes estimator for the mean squared error loss? What is the maximum a posteriori estimator for θ ? (part-2)

Consider $X \mid \theta \sim \text{Binomial}(n, \theta), \theta \sim \text{Beta}(r, s)$ with $r, s > 0$ given. The density/mass functions are

1. $p(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$.
2. $w(\theta) = \theta^{r-1}(1-\theta)^{s-1}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}, 0 < \theta < 1$.

Also note that $\mathbb{E}(\theta) = \frac{r}{r+s}$. What is the Bayes estimator under the quadratic loss?

THEOREM

State the theorem about UMP tests for θ of exponential family distributions with increasing $c(\theta)$.

Suppose $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$ where $\theta \in \Theta$ is an interval in $\mathbb{R}$ and $c(\cdot)$ is strictly increasing. Let

$$\phi(T) = \begin{cases} 1 & \text{if } T > t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T < t_0 \end{cases}$$

where t_0 and q are such that $\mathbb{E}_{\theta_0}\phi(T) = \alpha$. Let $\beta(\theta) = \mathbb{E}_\theta\phi(T)$. Show in the discrete case that $\theta \mapsto \beta(\theta)$ is increasing.

(part-1)

Suppose $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$ where $\theta \in \Theta$ is an interval in $\mathbb{R}$ and $c(\cdot)$ is strictly increasing. Let

$$\phi(T) = \begin{cases} 1 & \text{if } T > t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T < t_0 \end{cases}$$

where t_0 and q are such that $\mathbb{E}_{\theta_0}\phi(T) = \alpha$. Let $\beta(\theta) = \mathbb{E}_\theta\phi(T)$. Show in the discrete case that $\theta \mapsto \beta(\theta)$ is increasing.

(part-2)

Suppose $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$ where $\theta \in \Theta$ is an interval in $\mathbb{R}$ and $c(\cdot)$ is strictly increasing. Let

$$\phi(T) = \begin{cases} 1 & \text{if } T > t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T < t_0 \end{cases}$$

where t_0 and q are such that $\mathbb{E}_{\theta_0}\phi(T) = \alpha$. Let $\beta(\theta) = \mathbb{E}_\theta\phi(T)$. Show in the discrete case that $\theta \mapsto \beta(\theta)$ is increasing.

(part-3)

Suppose $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$ where $\theta \in \Theta$ is an interval in $\mathbb{R}$ and $c(\cdot)$ is strictly increasing. Let

$$\phi(T) = \begin{cases} 1 & \text{if } T > t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T < t_0 \end{cases}$$

where t_0 and q are such that $\mathbb{E}_{\theta_0}\phi(T) = \alpha$. Let $\beta(\theta) = \mathbb{E}_\theta\phi(T)$. Knowing that $\theta \mapsto \beta(\theta)$ is increasing, show that ϕ is uniformly most powerful (UMP) for $\tilde{H}_0 : \theta = \theta_0, \tilde{H}_1 : \theta = \theta_1$ for all $\theta_1 > \theta_0$.

Suppose $X_1, \ldots, X_n$ are i.i.d. from Exponential $(\theta), \theta > 0$. The probability density function for the one-dimensional exponential distribution is

$$p_\theta(x) = \exp[-\theta x - (-\log\theta)]1\{x > 0\}.$$

Consider a hypothesis $H_0 : \theta \le \theta_0$, and $H_1 : \theta > \theta_0$. What is a UMP test for H_0? (part-1)

The posterior is

$$\begin{aligned} w(\vartheta \mid x) &\propto p(x \mid \vartheta) w(\vartheta) \\ &\propto e^{-\vartheta} \vartheta^{x} e^{-\lambda \vartheta} \vartheta^{k-1} \\ &= e^{-(1+\lambda)\vartheta} \vartheta^{x+k-1} \end{aligned}$$

so we can see that $\vartheta \mid X = x \sim \text{Gamma}(x+k, 1+\lambda)$ with $\mathbb{E}(\theta \mid X) = \frac{X+k}{1+\lambda}$. This is the Bayes estimator for mean squared error loss.

$$\begin{aligned} r_w(T') &= \mathbb{E} L(\theta, T'(X)) \\ &= \mathbb{E}\mathbb{E}\left[L(\theta, T'(X)) \mid X\right] \quad \text{(Iterated } \mathbb{E}) \\ &= \mathbb{E}\mathbb{E}\left[\left[\theta - T'(X)\right]^2 \mid X\right] \\ &= \underbrace{\mathbb{E}\operatorname{Var}(\theta \mid X)}_{r_w(T_{\text{Bayes}})} + \mathbb{E}\left[(\underbrace{\mathbb{E}(\theta \mid X)}_{T_{\text{Bayes}}(X)} - T'(X))^2 \mid X\right] \\ &= r_w(T_{\text{Bayes}}) + \mathbb{E}\left(\left[T_{\text{Bayes}}(X) - T'(X)\right]^2\right). \end{aligned}$$

4^{th} equality: $MSE = Var + Bias^2$

First compute the posterior

$$\begin{aligned} w(\vartheta \mid x) &\propto \vartheta^{x}(1-\vartheta)^{n-x} \vartheta^{r-1}(1-\vartheta)^{s-1} \\ &= \vartheta^{x+r-1}(1-\vartheta)^{n-x+s-1}. \end{aligned}$$

We can see that it is $\theta \mid X = x \sim \text{Beta}(x+r, n-x+s)$. So, the Bayes lestimator under quadratic loss is $\mathbb{E}(\theta \mid X) = \frac{X+r}{n+r+s}$.

Now we need to find the maximum of $w(\vartheta \mid x)$. We can do it under logarithm, so it's easier

$$\log w(\vartheta \mid x) = (x+k-1)\log\vartheta - (1+\lambda)\vartheta + \underbrace{c(x)}_{\substack{\text{normalizing} \\ \text{constant} \\ \text{to ignore}}}$$

$$\begin{aligned} \frac{\delta}{\delta\vartheta} w(\vartheta \mid x) &= \frac{x+k-1}{\vartheta} - (1+\lambda) \triangleq 0 \\ \theta_{\text{MAP}}(x) &= \frac{x+k-1}{1+\lambda} \end{aligned}$$

We have $P_\theta(T=t) = \exp[c(\theta)t - d(\theta)] \sum_{x:T(x)=t} h(x)$. For $\tilde\theta > \theta$ it's true that the ratio

$$\frac{P_{\tilde\theta}(T=t)}{P_\theta(T=t)} = \exp[\underbrace{\{c(\tilde\theta) - c(\theta)\}}_{>0} t - \{d(\tilde\theta) - d(\theta)\}]$$

is increasing in t. To find the difference $\mathbb{E}_{\tilde\theta}\phi(T) - \mathbb{E}_\theta\phi(T)$ we first need to consider whether $p_{\tilde\theta}(t)$ is greater than, equal, or less than $p_\theta(t)$.

Suppose $\mathbb{P}$ is a one-dimensional exponential family $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$. Assume also that $c(\theta)$ is a strictly increasing function of θ. Then a UMP test is

$$\phi(T) = \begin{cases} 1 & \text{if } T > t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T < t_0, \end{cases}$$

where q and t_0 are chosen in such a way that $\mathbb{E}_\theta\phi(T) = \alpha$.

$$\begin{aligned} &= \sum_{t \le s_0} \underbrace{\phi(t)}_{\le \phi(s_0)} \underbrace{[p_{\tilde\theta}(t) - p_\theta(t)]}_{\le 0} + \sum_{t > s_0} \underbrace{\phi(t)}_{\ge \phi(s_0)} \underbrace{[p_{\tilde\theta}(t) - p_\theta(t)]}_{\ge 0} \\ &\ge \phi(s_0) \sum_{t \le s_0} [p_{\tilde\theta}(t) - p_\theta(t)] + \phi(s_0) \sum_{t > s_0} [p_{\tilde\theta}(t) - p_\theta(t)] \\ &= \phi(s_0) \underbrace{\sum_t [p_{\tilde\theta}(t) - p_\theta(t)]}_{=0} = 0. \end{aligned}$$

Note that if $p_{\tilde\theta}(t) < p_\theta(t)$ for all t, then we would also have

$$1 = \sum_t p_{\tilde\theta}(t) < \sum_t p_\theta(t) = 1$$

which is a contradiction. The same happens if $p_{\tilde\theta}(t) > p_\theta(t)$ for all t. There must be a point s_0 such that for all $t \le s_0, p_{\tilde\theta}(t) \le p_\theta$, but for all $t > s_0$ we instead have $p_{\tilde\theta}(t) > p_\theta$.

$$\mathbb{E}_{\tilde\theta}\phi(T) - \mathbb{E}_\theta\phi(T) = \sum_t \phi(t)\left[p_{\tilde\theta}(t) - p_\theta(t)\right] = (cont)$$

In the case of our sample we have

$$p_\theta(x) = \prod_{i=1}^n p_\theta(x_i) = \exp[\underbrace{-\theta}_{c(\theta)} \underbrace{\sum_{i=1}^n x_i}_{T(x_1,\ldots,x_n)} - (-\log\theta)] 1\{x > 0\}$$

So, using the lemma for exponential family UMP tests with strictly increasing $c(\theta)$ we can construct a UMP test ϕ such that

Since $\beta(\theta) = \mathbb{E}_\theta\phi(T)$ is increasing, we have

$$\sup_{\theta \le \theta_0} \mathbb{E}_\theta\phi(T) = \mathbb{E}_{\theta_0}\phi(T) = \alpha$$

Since t_0 and q are chosen such that $\mathbb{E}_{\theta_0}\phi(T) = \alpha$, we can see that they do not depend on the alternative hypothesis θ_1. Thus, ϕ is uniformly most powerful (UMP).

Suppose $X_1, \ldots, X_n$ are i.i.d. from Exponential $(\theta), \theta > 0$. The probability density function for the one-dimensional exponential distribution is

$$p_\theta(x) = \exp[-\theta x - (-\log \theta)]1\{x > 0\}.$$

Consider a hypothesis $H_0 : \theta \leq \theta_0$, and $H_1 : \theta > \theta_0$. What is a UMP test for H_0? (part-2)

Consider the problem of testing for parameter θ in exponential family distributions with strictly increasing $c(\theta)$. How do you construct UMP tests for the right-sided and left-sided alternative hypotheses? (part-1)

Consider the problem of testing for parameter θ in exponential family distributions with strictly increasing $c(\theta)$. How do you construct UMP tests for, the right-sided and left-sided alternative hypotheses? (part-2)

Theorem

State the theorem about exponential family UMPU tests.

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\sigma^2 = \sigma_0^2$ known. Find a UMPU test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p_\theta(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}\begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right]\frac{1}{\sqrt{2\pi}}$$

(part-1)

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\sigma^2 = \sigma_0^2$ known. Find a UMPU test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p_\theta(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}\begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right]\frac{1}{\sqrt{2\pi}}$$

(part-2)

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\sigma^2 = \sigma_0^2$ known. Find a UMPU test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p_\theta(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}\begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right]\frac{1}{\sqrt{2\pi}}$$

(part-3)

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\sigma^2 = \sigma_0^2$ known. Find a UMPU test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p_\theta(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}\begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right]\frac{1}{\sqrt{2\pi}}$$

(part-4)

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\sigma^2 = \sigma_0^2$ known. Find a UMPU test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. Keep in mind that for any normal random variable $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ you can write the probability density as

$$p_\theta(y) = \exp\left[\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}\begin{bmatrix} y^2 \\ y \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\left(\sigma^2\right)\right)\right]\frac{1}{\sqrt{2\pi}}$$

(part-5)

Let T be an unbiased estimator of $g(\theta)$ where $\theta \in \mathbb{R}^k$. Sketch a proof that then, under regularity conditions,

$$\mathrm{Var}_\theta(T) \geq \dot{g}(\theta)^T I(\theta)^{-1} \dot{g}(\theta)$$

(part-1)

- Right-sided alternative: $c(\cdot)$ strictly increasing,

$$H_0 : \theta \leq \theta_0,\ H_1 : \theta > \theta_0,\ \phi_R(T) = \begin{cases} 1 & T > t_0 \\ q & t = t_0 \\ 0 & T < t_0 \end{cases}$$

- Left-sided alternative: $c(\cdot)$ strictly decreasing,

$$H_0 : \theta \geq \theta_0,\ H_1 : \theta < \theta_0,\ \phi_L(T) = \begin{cases} 1 & T < t_0 \\ q & t = t_0 \\ 0 & T > t_0 \end{cases}$$

$$\phi(T) = \begin{cases} 1 & \text{if } t \leq t_0 \\ 0 & \text{if } t > t_0 \end{cases}$$

with a t_0 for which $\mathbb{E}_{\theta_0}(T) = \alpha$ holds. In particular,

$$P_{\theta_0}(T \leq t_0) = P(\theta_0 T \leq \theta_0 t_0) \text{ (mult. by } \theta_0 \to \text{ mean 1)}$$
$$= G(\theta_0, t_0) = \alpha. \quad \text{(Sum of Exp(1) is Gamma}(n, 1))$$

So $\theta_0 t_0 = G^{-1}(\alpha) \Longrightarrow t_0 = G^{-1}(\alpha)/\theta_0$ and then ϕ is UMP.

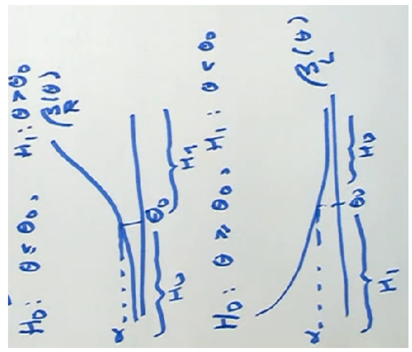
Suppose $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$ where $\theta \in \Theta$ is an interval in $\mathbb{R}$, and $c(\cdot)$ is strictly increasing. Let

$$\phi(T) = \begin{cases} 1 & \text{if } T \notin [t_L, t_R] \\ q_L & \text{if } T = t_l \\ q_R & \text{if } T = t_R \\ 0 & \text{if } T \in (t_L, t_R) \end{cases}$$

where q_L, q_R, t_L, t_R are such that

1. $\mathbb{E}_{\theta_0}\phi(T) = \alpha$
2. $\frac{d}{d\theta}\mathbb{E}_\theta \phi(T)\Big|_{\theta=\theta_0} = 0.$

Then ϕ is UMP unbiased (UMPU) for $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$.



$$\begin{aligned}
\mathbb{E}_\mu \phi(T) &= P_\mu(T \notin (t_L, t_R)) = 1 - P_\mu(T \in (t_L, t_R)) \\
&= 1 - P_\mu(T \leq t_R) - P_\mu(T \leq t_L) \\
&= 1 - P_\mu\left(\frac{T - n\mu}{\sqrt{n\sigma_0^2}} \leq \frac{t_R - n\mu}{\sqrt{n\sigma_0^2}}\right) \\
&\quad - P_\mu\left(\frac{T - n\mu}{\sqrt{n\sigma_0^2}} \leq \frac{t_L - n\mu}{\sqrt{n\sigma_0^2}}\right) \\
&= 1 - \Phi\left(\frac{t_R - n\mu}{\sqrt{n\sigma_0^2}}\right) - \Phi\left(\frac{t_L - n\mu}{\sqrt{n\sigma_0^2}}\right)
\end{aligned}$$

Looking at the density function of a single observation from a normal distribution, we can see that $c(\mu) = \frac{\mu}{\sigma_0^2}$ with $T(x) = x$. So, for the entire sample, our statistic is

$$T(x) = \sum_{i=1}^n x_i \sim \mathcal{N}\left(n\mu, n\sigma_0^2\right).$$

We take $(t_R - n\mu_0) = -(t_L - n\mu_0)$, because the other possibility $(t_R = t_L)$ is a test which always returns 0 . Knowing that $\Phi(-x) = 1 - \Phi(x)$, set

$$P_{\mu_0}(T \notin (t_L, t_R))) = 2\left(1 - \Phi\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right)\right) \triangleq \alpha.$$

$$\Longrightarrow \Phi\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right) = \frac{1-\alpha}{2}$$

Hence we have $t_R = n\mu_0 + \sqrt{n}\sigma_0\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ and $t_L = n\mu_0 - \sqrt{n}\sigma_0\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$

Now we want to find

$$\begin{aligned}
\frac{d}{d\theta}\mathbb{E}_\theta\phi(T)\Big|_{\theta=\theta_0} &= \frac{d}{d\theta}P_\mu(T \notin (t_L, t_R))\Big|_{\theta=\theta_0} \\
&= -\frac{n}{\sqrt{n}\sigma_0}\varphi\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) \\
&\quad + \frac{n}{\sqrt{n}\sigma_0}\varphi\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right)\Big|_{\theta=\theta_0} \triangleq 0 \\
&\Longleftrightarrow (t_R - n\mu_0)^2 = (t_L - n\mu)^2 \\
&\Longleftrightarrow (t_R - n\mu_0) = -(t_L - n\mu_0) \text{ or } t_R = t_L
\end{aligned}$$

As in the one-dimensional Cramer-Rao lower bound proof, we can show that for $j = 1, \ldots, k$,

$$\dot{g}_j(\theta) = \mathrm{Cov}_\theta\left(T, s_{\theta,j}(X)\right).$$

Hence for all $a \in \mathbb{R}^k$,

$$\begin{aligned}
\left|a^T \dot{g}(\theta)\right|^2 &= \left|\mathrm{Cov}_\theta\left(T, a^T s_\theta(X)\right)\right| \\
&\leq \mathrm{Var}_\theta(T)\,\mathrm{Var}_\theta\left(a^T s_\theta(X)\right) \\
&= \mathrm{Var}_\theta(T) a^T I(\theta) a
\end{aligned}$$

Finally, our UMPU test is

$$\phi(T) = \begin{cases} 1 & \text{if } |T - n\mu_0| > \sqrt{n}\sigma_0\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ 0 & \text{otherwise} \end{cases}$$

Let T be an unbiased estimator of $g(\theta)$ where $\theta \in \mathbb{R}^k$. Sketch a proof that then, under regularity conditions, $$\mathrm{Var}_\theta(T) \geq \dot{g}(\theta)^T I(\theta)^{-1} \dot{g}(\theta)$$ *(part-2)*	*Let $X \sim \mathcal{N}(\theta, 1)$ with $\sigma \in \mathbb{R}$, and $\theta \sim \mathcal{N}(c, \tau^2)$ with c, τ^2 known. What is the Bayes estimator for quadratic loss? (part-1)*
Let $X \sim \mathcal{N}(\theta, 1)$ with $\sigma \in \mathbb{R}$, and $\theta \sim \mathcal{N}(c, \tau^2)$ with c, τ^2 known. What is the Bayes estimator for quadratic loss? (part-2)	*What does it mean for an estimator T to be extended Bayes?*
THEOREM *State the theorem about minimax estimators of constant risk*	*Suppose the risk $R(\theta, T) = R(T)$ does not depend on θ. Show that* 1. *T is admissible $\Longrightarrow$ T is minimax,* 2. *T is Bayes $\Longrightarrow$ T is minimax,* 3. *T is extended Bayes $\Longrightarrow$ T is minimax.* *(part-1)*
Suppose the risk $R(\theta, T) = R(T)$ does not depend on θ. Show that 1. *T is admissible $\Longrightarrow$ T is minimax,* 2. *T is Bayes $\Longrightarrow$ T is minimax,* 3. *T is extended Bayes $\Longrightarrow$ T is minimax.* *(part-2)*	*Suppose $X \sim \mathrm{Binomial}(n, \theta)$ with $\theta \in (0,1)$. Suppose also the prior $\theta \sim \mathrm{Beta}(r, s)$ and quadratic loss $L(\vartheta, a) = (a - \vartheta)^2$. Knowing that the Bayes estimator minimizing the loss is* $$T_{\text{Bayes}}(X) = \mathbb{E}(\theta \mid X) = \frac{x+r}{n+r+s}$$ *find a minimax estimator for θ. (part-1)*
Suppose $X \sim \mathrm{Binomial}(n, \theta)$ with $\theta \in (0,1)$. Suppose also the prior $\theta \sim \mathrm{Beta}(r, s)$ and quadratic loss $L(\vartheta, a) = (a - \vartheta)^2$. Knowing that the Bayes estimator minimizing the loss is $$T_{\text{Bayes}}(X) = \mathbb{E}(\theta \mid X) = \frac{x+r}{n+r+s}$$ *find a minimax estimator for θ. (part-2)*	*What does it mean for a Bayes estimator to be unique?*

$$
\begin{aligned}
w(\vartheta \mid x) & \propto p(x \mid \vartheta) w(\vartheta) \\
& =\varphi(x-\vartheta) \frac{1}{\tau} \varphi\left(\frac{\vartheta-c}{\tau}\right) \\
& \propto \exp \left[-\frac{1}{2}(x-\vartheta)^2-\frac{1}{2 \tau^2}(\vartheta-c)^2\right] \\
& \propto \exp \left[x \vartheta-\frac{1}{2} \vartheta^2-\frac{1}{2 \tau^2} \vartheta^2+\frac{\vartheta c}{\tau^2}\right] \\
& =\exp \left[-\frac{1}{2}(-2 \vartheta \underbrace{\left(x+\frac{c}{\tau^2}\right)}_a+\vartheta^2 \underbrace{\left(1+\frac{1}{\tau^2}\right)}_b)\right]
\end{aligned}
$$

Then we can show

$$
\operatorname{Var}_\theta(T) \geq \max_{a \in \mathbb{R}^k} \frac{\left|a^T \dot{g}(\theta)\right|^2}{a^T I(\theta) a}=\dot{g}^T(\theta) I(\theta)^{-1} \dot{g}(\theta)
$$

(Last equality: auxiliary lemma, Cauchy-Schwarz, algebra)

An estimator T is called extended Bayes if there exists a sequence of priors $(w_m)_{m \geq 1}$ such that for the Bayes estimator T_m for each prior w_m we have

$$
r_{w_m}(T)-r_{w_m}(T_m) \underset{m \rightarrow \infty}{\longrightarrow} 0
$$

To come up with an expression which only has one θ, we will complete the square

$$
-2 a \vartheta+b \vartheta^2=b\left(-\frac{2 a}{b} \vartheta+\vartheta^2\right)=b\left(\vartheta-\frac{a}{b}\right)^2-\frac{a^2}{b}
$$

hence from the squared expression $\left(\vartheta-\frac{a}{b}\right)^2$ we can see that the mean is $\frac{a}{b}$ '(the distribution $w(\vartheta \mid x)$ is symmetric). So plugging in the values for a and b, the Bayes estimator for quadratic loss is

$$
T_{\text{Bayes }}=\mathbb{E}(\theta \mid X)=\frac{\tau^2 X+c}{\tau^2+1}
$$

1. Suppose $\sup_\theta R(\theta, T^{\prime}) \leq \sup_\theta R(\theta, T)=R(T)$ for all θ. Since T is admissible, $T^{\prime}$ is never strictly better than T, so $R(\theta, T^{\prime})=R(T)$ for all θ.
2. This is implied by (3).
3. Since T is extended Bayes, for any ε there exists an m such that

$$
r_{w_m}(T) \leq r_{w_m}(T_m)+\varepsilon
$$

Suppose the risk $R(\theta, T)=R(T)$ does not depend on θ. Then

1. T is admissible $\Longrightarrow T$ is minimax,
2. T is Bayes $\Longrightarrow T$ is minimax,
3. T is extended Bayes $\Longrightarrow T$ is minimax.

To find the minimax estimator, we will find parameters r and s for the prior such that risk of $T_{\text{Bayes }}$ is constant in θ.

$$
\begin{aligned}
R\left(\theta, T_{\text{Bayes }}\right) & =\mathbb{E}_\theta\left(T_{\text{Bayes }}-\theta\right)^2 \\
& =\operatorname{Var}_\theta\left(T_{\text{Bayes }}\right)+\operatorname{Bias}_\theta^2\left(T_{\text{Bayes }}\right) \\
& =\frac{n \theta(1-\theta)}{(n+r+s)^2}+\left[\frac{n \theta+r}{n+r+s}-\frac{(n+r+s) \theta}{n+r+s}\right]^2 \\
& =\frac{\left[(r+s)^2-n\right] \theta^2+[n-2 r(r+s)] \theta+r^2}{(n+r+s)^2}
\end{aligned}
$$

3. Now

$$
\begin{aligned}
R(T) & =r_{w_m}(T) \\
& \leq r_{w_m}(T_m)+\varepsilon \\
& \leq r_{w_m}(T^{\prime})+\varepsilon \quad (T_m \text{ is optimal for } r_{w_m}) \\
& \leq \sup_\vartheta R(\vartheta, T^{\prime})+\varepsilon
\end{aligned}
$$

since the ε is arbitrary, we know the risk $R(T)$ is less than the worst case risk $\sup_\vartheta R(\vartheta, T^{\prime})+\varepsilon$ of any other estimator $T^{\prime}$

A w-Bayes estimator T is called unique if $r_w(T^{\prime})=r_w(T)$ implies $P_\theta(T=T^{\prime})=1$ for all θ

Since we don't want the risk to depend on θ, we need to set the coefficients' in front of θ^2 and θ to 0

$$
(r+s)^2-n=0, n-2 r(r+s)=0 .
$$

Solving for r and s gives $r=s=\sqrt{n} / 2$. Plugging these values into the estimator gives

$$
T=\frac{X+\sqrt{n} / 2}{n+\sqrt{n}} .
$$

Since T is Bayes and its risk does not depend on θ, T is minimax.

LEMMA *State the lemma about sufficient conditions for a Bayes estimator T to be admissible.*	*Show that if T is the unique Bayes estimator for the prior density w, then T is also admissible.*
Suppose that for all $T', R(\theta, T')$ is continuous in θ, and for all open $U \subset \Theta$ the prior probability $$\Pi(U) := \int w(\vartheta) d\mu(\vartheta)$$ *of U is strictly positive. Show that if T is a Bayes estimator for the prior w, then T is also admissible. (part-1)*	*Suppose that for all $T', R(\theta, T')$ is continuous in θ, and for all open $U \subset \Theta$ the prior probability* $$\Pi(U) := \int w(\vartheta) d\mu(\vartheta)$$ *of U is strictly positive. Show that if T is a Bayes estimator for the prior w, then T is also admissible. (part-2)*
LEMMA *State the lemma about admissibility of extended Bayes estimators.*	*Suppose that T is extended Bayes, and that for all T', the risk $R(\theta, T')$ is continuous in θ. Furthermore, assume that for all open sets $U \subset \Theta$,* $$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \underset{m\to\infty}{\longrightarrow} 0$$ *where $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ is the probability of U under the prior Π_m. Show that then T is admissible. (part-1)*
Suppose that T is extended Bayes, and that for all T', the risk $R(\theta, T')$ is continuous in θ. Furthermore, assume that for all open sets $U \subset \Theta$, $$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \underset{m\to\infty}{\longrightarrow} 0$$ *where $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ is the probability of U under the prior Π_m. Show that then T is admissible. (part-2)*	*Suppose that T is extended Bayes, and that for all T', the risk $R(\theta, T')$ is continuous in θ. Furthermore, assume that for all open sets $U \subset \Theta$,* $$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \underset{m\to\infty}{\longrightarrow} 0$$ *where $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ is the probability of U under the prior Π_m. Show that then T is admissible. (part-3)*
Suppose $X \sim \mathcal{N}(\theta, 1)$ where $\theta \in \Theta = \mathbb{R}$. Let $R(\theta, T) := \mathbb{E}_\theta(T(X) - \theta)^2$. Consider $$T := aX + b, a > 0, b \in \mathbb{R}.$$ *Show that* $$T \text{ is admissible } \implies \begin{cases} (i) & 0 < a < 1 \\ \text{or} & \\ (\ ii\ ) & a = 1, b = 0 \end{cases}$$	PROOF *Suppose $X \sim \mathcal{N}(\theta, 1)$ where $\theta \in \Theta = \mathbb{R}$. Let $R(\theta, T) := \mathbb{E}_\theta(T(X) - \theta)^2$ Consider* $$T := aX + b, a > 0, b \in \mathbb{R}.$$ *Sketch steps to prove that if $0 < a < 1$ then T is admissible.* *(part-1)*

Suppose some other estimator T' has a lower risk than T for all θ

$$R\left(\theta, T'\right) \leq R(\theta, T).$$

If that's true, then also

$$r_w\left(T'\right) \leq r_w(T).$$

Since T is w-Bayes, this must be an equality $r_w\left(T'\right)=r_w(T)$. Since T is unique, $T=T' P_\theta$-almost surely for all θ. So, $R\left(\theta, T'\right)=R(\theta, T)$ and thus T is admissible

Suppose that T is a Bayes estimator for the prior density w. Then either of these conditions is sufficient for T to be admissible:

1. T is the unique Bayes decision,
2. for all $T', R\left(\theta, T'\right)$ is continuous in θ, and for all open $U \subset \Theta$ the prior probability $\Pi(U):=\int w(\vartheta) d \mu(\vartheta)$ of U is strictly positive.

$$\begin{aligned}
r_w\left(T'\right) & =\int R\left(\vartheta, T'\right) w(\vartheta) d \mu(\vartheta) \\
& =\int_U R\left(\vartheta, T'\right) w(\vartheta) d \nu(\vartheta)+\int_{U^c} R\left(\vartheta, T'\right) w(\vartheta) d \nu(\vartheta) \\
& =\int_U R(\vartheta, T) w(\vartheta) d \nu(\vartheta)-\varepsilon \Pi(U) \\
& +\int_{U^C} R(\vartheta, T) w(\vartheta) d \nu(\vartheta) \\
& =r_w(T)-\varepsilon \Pi(U)<r_w(T) .
\end{aligned}$$

So we came at contradiction, since T is Bayes for prior w.

Suppose T is not admissible, so there exists another estimator T' which fulfils $R\left(\theta, T'\right) \leq R(\theta, T)$ for all θ, and $R\left(\theta_0, T'\right)<R\left(\theta_0, T\right)$ for some specific θ_0. The assumptions imply that there exists an $\varepsilon>0$ for which there is an oper set U with $\theta_0 \in U$ such that

$$R\left(\theta, T'\right) \leq R(\theta, T)-\varepsilon \text{ for all } \theta \in U.$$

Then the risk fulfils

Suppose T is not admissible, so there exists another estimator T' which fulfils $R\left(\theta, T'\right) \leq R(\theta, T)$ for all θ, and $R\left(\theta_0, T'\right)<R\left(\theta_0, T\right)$ for some specific θ_0. So, there exists an $\varepsilon>0$ and U open with $\theta_0 \in U$ such that $R\left(\theta, T'\right) \leq R(\theta, T)$ for all $\theta \in U$.
We can then say (continued)

Suppose that T is extended Bayes, and that for all T', the risk $R\left(\theta, T'\right)$ is continuous in θ. Furthermore, assume that for all open sets $U \subset \Theta$,

$$\frac{r_{w_m}(T)-\inf_{T'} r_{w_m}\left(T'\right)}{\Pi_m(U)} \underset{m \rightarrow \infty}{\longrightarrow} 0$$

where $\Pi_m(U):=\int_U w_m(\vartheta) d \mu_m(\vartheta)$ is the probability of U under the prior Π_m. Then T is admissible.

which means that $\frac{r_{w_m}(T)-\inf_{T'} r_{w_m}\left(T'\right)}{\Pi_m(U)}$ does not converge to 0 , which contradicts our assumptions.

$$\begin{aligned}
r_{w_m}\left(T'\right) & \leq r_{w_m}(T)-\varepsilon \Pi_m(U) \\
\frac{r_{w_m}(T)-r_{w_m}\left(T'\right)}{\Pi_m(U)} & \geq \varepsilon \\
\frac{r_{w_m}(T)-r_{w_m}\left(T_m\right)}{\Pi_m(U)} & =\underbrace{\frac{r_{w_m}(T)-r_{w_m}\left(T'\right)}{\Pi_m(U)}}_{\geq \varepsilon} \\
& +\underbrace{\frac{r_{w_m}\left(T'\right)-r_{w_m}\left(T_m\right)}{\Pi_m(U)}}_{\geq 0}
\end{aligned}$$

1. Show that for any $a \in(0,1), b \in \mathbb{R}, T$ is Bayes for some prior. In particular, $\theta \sim \mathcal{N}\left(c, \tau^2\right)$ works where $a=\frac{\tau^2}{\tau^2+1}, b=\frac{c}{\tau^2+1}$.
2. Show that T is unique Bayes. So, for any other estimator T', show that

$$r_w\left(T'\right)=r_w(T) \Longrightarrow \mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right]=0$$

where the expectation is with θ integrated out, so with respect to some measure P.

Let $T_0:=X$.

1. Suppose $a>1$. Then

$$R(\theta, T) \geq \operatorname{Var}_\theta(T)=a^2>1=\operatorname{Var}_\theta\left(T_0\right)=R\left(\theta, T_0\right)$$

so T is not admissible.

2. Suppose alternatively that $a=1, b \neq 0$. Then

$$R(\theta, T) \geq \operatorname{Bias}_\theta^2(T)+1=b^2+1>\operatorname{Var}_\theta\left(T_0\right)=R\left(\theta, T_0\right)$$

so once again T is inadmissible.

PROOF *Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$ *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Sketch steps to prove that if* $0 < a < 1$ *then* T *is admissible.* *(part-2)*	*Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Knowing that* T *is Bayes for a prior* $\theta \sim \mathcal{N}(c, \tau^2)$ *with* $a = \frac{\tau^2}{\tau^2+1}, b = \frac{c}{\tau^2+1}$, *show that it is unique Bayes.* *(part-1)*
Suppose $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Knowing that* T *is Bayes for a prior* $\theta \sim \mathcal{N}(c, \tau^2)$ *with* $a = \frac{\tau^2}{\tau^2+1}, b = \frac{c}{\tau^2+1}$, *show that it is unique Bayes.* *(part-2)*	*Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Knowing that* T *is Bayes for a prior* $\theta \sim \mathcal{N}(c, \tau^2)$ *with* $a = \frac{\tau^2}{\tau^2+1}, b = \frac{c}{\tau^2+1}$, *show that it is unique Bayes.* *(part-3)*
Suppose $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$ *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Sketch steps to show that if* $a = 1, b = 0$ *then* T *is admissible.* *(part-1)*	*Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$ *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Sketch steps to show that if* $a = 1, b = 0$ *then* T *is admissible.* *(part-2)*
Suppose $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$ *Consider* $T := aX + b, a > 0, b \in \mathbb{R}.$ *Sketch steps to show that if* $a = 1, b = 0$ *then* T *is admissible.* *(part-3)*	*Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Show that* $T = X$ *is extended Bayes. Keep in mind that for the prior* $\theta \sim \mathcal{N}(0, m)$, *the Bayes estimator is* $T_m = \frac{m}{m+1}$
Suppose $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Knowing that* $T = X$ *is extended Bayes for the prior* $\theta \sim \mathcal{N}(0, m)$, *show that it is also admissible. Keep in mind that the risk* $r_{w_m}(T_m) = \frac{m}{m+1}$. *(part-1)*	*Suppose* $X \sim \mathcal{N}(\theta, 1)$ *where* $\theta \in \Theta = \mathbb{R}$. *Let* $R(\theta, T) := \mathbb{E}_\theta (T(X) - \theta)^2$. *Knowing that* $T = X$ *is extended Bayes for the prior* $\theta \sim \mathcal{N}(0, m)$, *show that it is also admissible. Keep in mind that the risk* $r_{w_m}(T_m) = \frac{m}{m+1}$. *(part-2)*

Consider some estimator T' (4eq MSE $=\operatorname{Var}+\operatorname{Bias}^2$)

$$\begin{aligned} r_w\left(T'\right) &= \mathbb{E} R\left(\theta, T'\right) \\ &= \mathbb{E}\left[\left(\theta-T'(X)\right)^2\right] \\ &= \mathbb{E}\mathbb{E}\left[\left(\theta-T'(X)\right)^2 \mid X\right] \\ &= \mathbb{E} \operatorname{Var}(\theta \mid X)+\mathbb{E}\left[\left(\mathbb{E}(\theta \mid X)-T'(X)\right)^2 \mid X\right] \\ &= r_w(T)+\mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right], T=\mathbb{E}(\theta \mid X) \end{aligned}$$

3. Show that

$$\mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right]=\int T(x)-T'(x) dP(x)=0$$

implies $T=T' P_{\theta}$-almost surely for any θ. To do this, show that P is a dominating measure $\left(\mathcal{N}\left(c, \tau^2+1\right)\right)$ for P_θ for any θ.

4. Conclude that T is unique Bayes and hence admissible.

Since all normal distributions dominate each other, P dominates P_θ for all θ, which means that $\mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right]=0$ integrated over $P \Longrightarrow T=T' P_{\theta}$--almost surely for all θ.

So T is unique Bayes.

So, if $r_w\left(T'\right)=r_w(T)$ then $\mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right]=0$. Now we want to show that this implies that indeed $T=T'$ almost surely for all θ. Note that the expectation $\mathbb{E}\left[\left(T(X)-T'(X)\right)^2\right]=0$ is with respect to a measure P with θ integrated out. So, we need to show that P dominates all P_θ. P is the measure of X, which we can write as $X=\theta+\varepsilon$ where $\theta \sim \mathcal{N}\left(c, \tau^2\right)$ and $\varepsilon \sim \mathcal{N}(0,1)$. So, P is the $\mathcal{N}\left(c, \tau^2+1\right)$ distribution.

2. Confirm that the risk converges sufficiently quickly to fulfil the admissibility criterion, i.e. that

$$\frac{r_{w_m}(T)-r_{w_m}\left(T_m\right)}{\Pi_m(U)} \underset{m \rightarrow \infty}{\longrightarrow} 0.$$

To do this, pick any open interval $U=(u, u+h)$ and show by using a Taylor approx that for large m,

$$\Pi_m(U)=\Phi\left(\frac{u+h}{\sqrt{m}}\right)-\Phi\left(\frac{u}{\sqrt{m}}\right) \approx \frac{1}{\sqrt{m}} \phi(0) \geq \frac{1}{\sqrt{m} C}$$

for some constant C.

1. Show that T is extended Bayes for prior $\theta \sim \mathcal{N}(0, m)$. The Bayesian estimator is then $T_m=\frac{m}{m+1} X$, while the risk of T is always 1 , so

$$r_{w_m}(T)-r_{w_m}\left(T_m\right)=1-\frac{m}{m+1} \underset{m \rightarrow \infty}{\longrightarrow} 0.$$

$$\begin{aligned} R\left(\theta, T_m\right) &= \left(\tfrac{m}{m+1} \theta-\theta\right)^2+\tfrac{m^2}{(m+1)^2},(\mathrm{MSE}=\mathrm{B}^2+\operatorname{Var}) \\ &= \tfrac{1}{(m+1)^2} \theta^2+\tfrac{m^2}{(m+1)^2} \\ \mathbb{E} R\left(\theta, T_m\right) &= r_{w_m}\left(T_m\right)=\tfrac{m}{(m+1)^2}+\tfrac{m^2}{(m+1)^2}\left(\mathbb{E} \theta^2=m\right) \\ &= \tfrac{m}{(m+1)} \\ r_w(T) &= 1. \end{aligned}$$

So we can already see that T is extended Bayes, because

$$r_{w_m}(T)-r_{w_m}\left(T_m\right)=1-\frac{m}{m+1} \underset{m \rightarrow \infty}{\longrightarrow} 0$$

3. At this point we get

$$\frac{r_{w_m}(T)-r_{w_m}\left(T_m\right)}{\Pi_m(U)} \leq \frac{1-\frac{m}{m+1}}{\frac{1}{\sqrt{m} C}}=\frac{\frac{1}{m+1}}{\frac{1}{\sqrt{m}}} C \underset{m \rightarrow \infty}{\longrightarrow} 0$$

and so, since the loss function is continuous, T is admissible.

Take an open interval $U=(u, u+h)$. Now

$$\begin{aligned} \Pi_m(U) &= \Phi\left(\frac{u+h}{\sqrt{m}}\right)-\Phi\left(\frac{u}{\sqrt{m}}\right) \\ &= \frac{1}{\sqrt{m}} \phi\left(\frac{u}{\sqrt{m}}\right) h+o(1 / \sqrt{m}) \\ &\approx \frac{1}{\sqrt{m}} \phi(0) \quad \text{(large } m\text{)} \\ &\geq \frac{1}{\sqrt{m} C} \text{ for some constant } C \end{aligned}$$

Since we know T is extended Bayes, we know $r_{w_m}(T)-r_{w_m}\left(T_m\right) \underset{m \rightarrow \infty}{\longrightarrow} 0$. We just need to check if it's sufficiently quick so that for any $U \subset \mathbb{R}$ open it's true that

$$\frac{r_{w_m}(T)-r_{w_m}\left(T_m\right)}{\Pi_m(U)} \underset{m \rightarrow \infty}{\longrightarrow} 0$$

$$\left(\begin{array}{c}\text{Taylor approx.} \\ F(x+h)=F(x)+F'(x) h+o(h)\end{array}\right)$$

Suppose $X \sim \mathcal{N}(\theta, 1)$ where $\theta \in \Theta = \mathbb{R}$. Let $R(\theta, T) := \mathbb{E}_\theta(T(X) - \theta)^2$. Knowing that $T = X$ is extended Bayes for the prior $\theta \sim \mathcal{N}(0, m)$, show that it is also admissible. Keep in mind that the risk $r_{w_m}(T_m) = \frac{m}{m+1}$. (part-3)

DEFINITION

What is the least squares estimator?

LEMMA

State lemma about constructing least squares estimators using linear algebra operations.

Let $f = \mathbb{E}Y$. What is the best linear approximation of f ?

LEMMA

State lemma about estimation and misspecification errors of least squares models.

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also that the noise is uncorrelated, so $\mathbb{E}\epsilon\epsilon^T = \sigma^2 I$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f. Show that then

$$\mathbb{E}\hat{\beta} = \beta^*$$

with $\mathrm{Cov}(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$

Let $\mathbb{E}(Y) = f$, let $\epsilon := Y - f$ be the noise. Suppose also that the noise is uncorrelated, so $\mathbb{E}\epsilon\epsilon^T = \sigma^2 I$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f. Show

$$\mathbb{E}\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = \sigma^2 p$$

(part-1)

Let $\mathbb{E}(Y) = f$, let $\epsilon := Y - f$ be the noise. Suppose also that the noise is uncorrelated, so $\mathbb{E}\epsilon\epsilon^T = \sigma^2 I$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f. Show

$$\mathbb{E}\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = \sigma^2 p$$

(part-2)

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also that the noise is uncorrelated, so $\mathbb{E}\epsilon\epsilon^T = \sigma^2 I$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f. Show that then

$$\mathbb{E}\|X\hat{\beta} - f\|_2^2 = \sigma^2 p + \|X\beta^* - f\|_2^2$$

LEMMA

State the lemma about the distribution of linear approximation errors.

The least squares estimator is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \arg \min_{\begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^{p+1}} \|Y - a - X_b\|_2^2$$

where $\|V\|_2^2 = V^T V = \sum_{i=1}^n V_i^2, V \in \mathbb{R}^n$. Sometimes $\hat{\alpha}$ is replaced instead by a constant term appended to X. Then $\tilde{X} \to X, p+1 \to p, \begin{pmatrix} a \\ b \end{pmatrix} \to b$, and we write

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2.$$

So

$$\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \leq \frac{1 - \frac{m}{m+1}}{\frac{1}{\sqrt{m}C}} = \frac{\frac{1}{m+1}}{\frac{1}{\sqrt{m}}} C \underset{m \to \infty}{\longrightarrow} 0.$$

Since the above criterion is fulfilled and the loss function is continuous, T is admissible.

We set $\beta^* := \left(X^T X\right)^{-1} X^T f$, and then the best linear approximation of f is $X\beta^*$.

Suppose $X \in \mathbb{R}^{n \times p}$ has rank p. Then the least squares estimator is

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[\left(X^T X\right)^{-1} X^T Y\right] \\
&= \left(X^T X\right)^{-1} X^T \mathbb{E}[\epsilon] + \left(X^T X\right)^{-1} X^T f = \beta^* \\
\operatorname{Cov}(\hat{\beta}) &= \operatorname{Cov}\left(\left(X^T X\right)^{-1} X^T \epsilon\right) \\
&= \left(X^T X\right)^{-1} X^T \underbrace{\operatorname{Cov}(\epsilon)}_{\sigma^2 I} X \left(X^T X\right)^{-1} \\
&= \left(X^T X\right)^{-1} \sigma^2.
\end{aligned}$$

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also that the noise is uncorrelated, so $\mathbb{E}\epsilon\epsilon^T = \sigma^2 I$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f.

1. $\mathbb{E}\hat{\beta} = \beta^*, \operatorname{Cov}(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$
2. $\mathbb{E}\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = p\sigma^2$
3. $\mathbb{E}\|X\hat{\beta} - f\|_2^2 = \underbrace{p\sigma^2}_{\text{estimation error}} + \underbrace{\mathbb{E}\|X\beta^* - f\|_2^2}_{\text{misspecification error}}.$

Now just continue the equations we had above

$$\begin{aligned}
\mathbb{E}\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 &= \mathbb{E}\left\|PP^T\epsilon\right\|_2^2 \\
&= \mathbb{E}\left\|P^T\epsilon\right\|_2^2 \quad (\text{ Eqn. above }) \\
&= \mathbb{E}\|V\|_2^2 \\
&= \mathbb{E}V^T V \\
&= \operatorname{Cov}(V) \quad (\mathbb{E}V = 0) \\
&= p\sigma^2
\end{aligned}$$

We have

$$\begin{aligned}
X\left(X^T X\right)^{-1} X^T &= PP^T \\
\implies \left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 &= \left\|PP^T\epsilon\right\|_2^2 \\
&= \epsilon^T PP^T PP^T \epsilon \\
&= \epsilon^T PP^T \epsilon = \left\|P^T\epsilon\right\|_2^2.
\end{aligned}$$

Let $V := P^T\epsilon$. Knowing that $\mathbb{E}P^T\epsilon = 0$, we get

$$\operatorname{Cov}(V) = P^T \operatorname{Cov}(\epsilon) P = I\sigma^2$$

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n \times p}$ be the best linear approximation of f. Then

1. $\hat{\beta} \sim \mathcal{N}\left(\beta^*, \sigma^2 \left(X^T X\right)^{-1}\right)$,
2. $\frac{\|x(\hat{\beta} - \beta^*)\|_2^2}{\sigma^2} \sim \chi_p^2$.

$$X\hat{\beta} - f = \underbrace{X\left(\hat{\beta} - \beta^*\right)}_{\in \text{ column space of } X} + \underbrace{(X\beta^* - f)}_{\text{orthogonal to } X}$$

So by Pythagoras for all b it's true that

$$\|X\hat{\beta} - f\|_2^2 = \|X(b - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2$$

and, taking the expectation,

$$\mathbb{E}\|X\hat{\beta} - f\|_2^2 = \underbrace{\mathbb{E}\|X(b - \beta^*)\|_2^2}_{= \sigma^2 p \text{ from lemma}} + \mathbb{E}\|X\beta^* - f\|_2^2$$

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n\times p}$ be the best linear approximation of f. Show that then

$$\hat{\beta} \sim \mathcal{N}\left(\beta^*, \sigma^2 \left(X^T X\right)^{-1}\right).$$

Let $\mathbb{E}(Y) = f$, and let $\epsilon := Y - f$ be the noise. Suppose also $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Also let $\hat{\beta}$ be the least squares estimator, and $X\beta^*$ with $X \in \mathbb{R}^{n\times p}$ be the best linear approximation of f. Show that then

$$\frac{\left\| X\left(\hat{\beta} - \beta^*\right)\right\|_2^2}{\sigma^2} \sim \chi_p^2$$

Suppose $Y = X\beta + \epsilon$ for $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with a known σ^2. What is a pivot you could use to test for the hypothesis $H_0 : \beta = \beta_0$?

LEMMA

State lemma about testing a linear hypothesis with a restriction.

What does it mean for Z_n to converge to Z in probability?

What does it mean for Z_n to converge to Z almost surely?

What does it mean for Z_n to converge to Z in distribution?

How are convergence in probability and in distribution related?

THEOREM

State the central limit theorem.

THEOREM

State the Cramér-Wold device.

$$\left\|X\left(\hat{\beta}-\beta^*\right)\right\|_2^2 = \left\|PP^T\epsilon\right\|_2^2$$
$$= \|PV\|_2^2 \quad \left(\text{ Let } P^T\epsilon = V\right)$$

Now $\mathbb{E}V = 0, \mathrm{Cov}(V) = \sigma^2 I$

$$\|PV\|_2^2 = V^T P^T P V = \|V\|_2^2 = \sum_{j=1}^p V_j^2$$

where $V_1, \ldots, V_p$ are i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. Hence, $\frac{\sum_{j=1}^p V_j}{\sigma^2} \sim \chi_p^2$

$$\hat{\beta} = \underbrace{\left(X^T X\right)^{-1} X^T f}_{\beta^*} + \underbrace{\left(X^T X\right)^{-1} X^T \epsilon}_{\mathcal{N}\left(0,(X^T X)^{-1}\sigma^2\right)}$$

Let the model be $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Our hypothesis is $H_0 : B\beta = 0$ where $B \in \mathbb{R}^{q\times p}$ is a matrix of restrictions. Estimator $\hat{\beta}_0$ defined as

$$\hat{\beta}_0 = \arg \min_{b\in\mathbb{R}^p, Bb=0} \|Y - Xb\|_2^2.$$

Now under the null hypothesis

$$\frac{\left\|Y - X\hat{\beta}_0\right\|_2^2 - \|Y - X\hat{\beta}\|_2^2}{\sigma^2} \sim \chi_q^2.$$

We could use the distribution function of the χ_p^2 distribution. This is because we know

$$\frac{\|X(\hat{\beta}-\beta)\|_2^2}{\sigma^2} \sim \chi_p^2.$$

Z_n converges to Z almost surely if

$$\mathbb{P}\left(\lim_{n\to\infty} Z_n = z\right) = 1$$

We then write $Z_n \xrightarrow{\text{a.s.}} Z$.

Z_n converges to Z in probability if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\|Z_n - Z\| > \varepsilon\right) = 0$$

We then write $Z_n \xrightarrow{\mathbb{P}} Z$

Convergence in probability implies convergence in distribution, but not the other way around.

Z_n converges in distribution to Z if

$$\lim_{n\to\infty} \mathbb{E}f\left(Z_n\right) = \mathbb{E}f(Z)$$

for all $f : \mathbb{R}^p \to \mathbb{R}$ bounded and continuous. We then write $Z_n \xrightarrow{\mathcal{D}} Z$.

Let $(\{Z_n\}, Z)$ be a collection of $\mathbb{R}^p$-valued random variables. Then

$$Z_n \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} Z \Longleftrightarrow a^T Z_n \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} a^T Z \forall_{a\in\mathbb{R}^p}.$$

If $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \in \mathbb{R}$ with $\mathbb{E}X = \mu, \mathrm{Var}(X) = \sigma^2$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then

$$\frac{\left(\bar{X}_n - \mu\right)}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

THEOREM *State the Portmanteau theorem.*	*Suppose $(z_n)_{n\geq 1}$ is a sequence of $\mathbb{R}^p$ vectors, $(r_n)_{n\geq 1}$ is a sequence of positive constants. What do these order symbols mean?* 1. $z_n = \mathcal{O}(1)$ 2. $z_n = o(1)$ 3. $z_n = \mathcal{O}\left(r_n\right)$ 4. $z_n = o\left(r_n\right)$
What does it mean for a sequence of random variables $(Z_n)_{n\geq 1}$ to be bounded in probability?	*Suppose $(Z_n)_{n\geq 1}$ is a sequence of $\mathbb{R}^p$ random vectors, $(r_n)_{n\geq 1}$ is a sequence of positive constants. What do these symbols mean?* 1. $Z_n = \mathcal{O}_{\mathbb{P}}(1)$ 2. $Z_n = o_{\mathbb{P}}(1)$ 3. $Z_n = \mathcal{O}_{\mathbb{P}}\left(r_n\right)$ 4. $Z_n = o_{\mathbb{P}}\left(r_n\right)$
What does $Z_n \xrightarrow[n\to\infty]{\mathcal{D}} Z$ imply about the order of Z_n ?	THEOREM *State Slutsky's theorem.*
PROOF *Prove the Slutsky's theorem*	*Let $X_1, \ldots, X_n$ be i.i.d. $X \in \mathbb{R}$ with mean μ and variance σ^2. Does $\sqrt{n}\left(\bar{X}_n^2 - \mu^2\right)$ converge to something? If so, to what, and with what type of convergence?*
What does it mean for an estimator T_n to be consistent?	*What does it mean for an estimator T_n to be called asymptotically normal?*

	Order symbol	Meaning
(1)	$z_n = \mathcal{O}(1)$	$\lim_{n\to\infty} \|z_n\| < \infty$
(2)	$z_n = o(1)$	$\lim_{n\to\infty} \|z_n\| = 0$
(3)	$z_n = \mathcal{O}(r_n)$	$\lim_{n\to\infty} \|z_n/r_n\| < \infty$
(4)	$z_n = o(r_n)$	$\lim_{n\to\infty} \|z_n/r_n\| = 0$

The following are equivalent:

- $\mathbb{E}f(Z_n) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbb{E}f(Z)$ for all $f : \mathbb{R}^p \to \mathbb{R}$ bounded and continuous,
- $\mathbb{E}f(Z_n) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbb{E}f(Z)$ for all $f : \mathbb{R}^p \to \mathbb{R}$ bounded and Lipschitz,
- $\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \mathbb{P}(Z \le z) =: G(z)$ for all G-continuity points z.

	Order symbol	Meaning
(1)	$Z_n = \mathcal{O}_{\mathbb{P}}(1)$	$(\star)$
(2)	$Z_n = o_{\mathbb{P}}(1)$	$Z_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$
(3)	$Z_n = \mathcal{O}_{\mathbb{P}}(r_n)$	$Z_n/r_n = \mathcal{O}_{\mathbb{P}}(1)$
(4)	$Z_n = o_{\mathbb{P}}(r_n)$	$Z_n/r_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$

With $(\star) = \lim_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(\|Z_n\| > M) = 0$

A sequence of random variables $(Z_n)_{n\ge 1}$ is bounded in probability if

$$\lim_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(\|Z_n\| > M) = 0$$

Let $(Z_n)_{n\ge 1}$, Z, and $(A_n)_{n\ge 1}$ be (sequences of) random variables in $\mathbb{R}^p$. Furthermore, let $a \in \mathbb{R}^p$ be a constant vector. Then

$$\left\{\begin{array}{l} Z_n \xrightarrow[n\to\infty]{\mathcal{D}} Z \\ \text{and} \\ A_n \xrightarrow[n\to\infty]{\mathbb{P}} a \end{array}\right. \implies A_n^T Z_n \xrightarrow[n\to\infty]{\mathcal{D}} a^T Z.$$

$$Z_n \xrightarrow[n\to\infty]{\mathcal{D}} Z \text{ implies } Z_n = \mathcal{O}_{\mathbb{P}}(1)$$

We know that

1. by the CLT, $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$,
2. by the LLN, $\left(\bar{X}_n + \mu\right) \xrightarrow[n\to\infty]{\mathbb{P}} 2\mu$.

So $\sqrt{n}\left(\bar{X}_n^2 - \mu^2\right) = \sqrt{n}\left(\bar{X}_n - \mu\right)\left(\bar{X}_n + \mu\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, 4\mu^2\sigma^2\right)$. (Slutsky)

See pg 132 skript.

T_n is called asymptotically normal if

$$\sqrt{n}\left(T_n - g(\theta)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, V_\theta)$$

where V_θ is the asymptotic covariance matrix.

T_n is called consistent if

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}_\theta} g(\theta)$$

What does it mean for T_n to be called asymptotically linear?

How are asymptotic linearity and normality related?

Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \in \mathbb{R}$ with $\mathbb{E}_\theta X = \mu, \mathrm{Var}_\theta(X) = \sigma^2 < \infty$. Let $g(\theta) = \mu$, and $T_n = \bar{X}_n$. What is the influence function in this case?

Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \in \mathbb{R}$ with $\mathbb{E}_\theta X = \mu$, $\mathrm{Var}_\theta(X) = \sigma^2 < \infty$. Let $g(\theta) = \mu^2$, and $T_n = \bar{X}_n^2$. What is the influence function and asymptotic variance in this case?

Theorem

State the δ-method theorem. (part-1)

Theorem

State the δ-method theorem. (part-2)

Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \sim \mathrm{Bernoulli}(\theta)$. What is the asymptotic distribution of

$$\sqrt{n}\left(\log \frac{\bar{X}_n}{1-\bar{X}_n} - \log \frac{\theta}{1-\theta}\right)?$$

(part-1)

Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \sim \mathrm{Bernoulli}(\theta)$. What is the asymptotic distribution of

$$\sqrt{n}\left(\log \frac{\bar{X}_n}{1-\bar{X}_n} - \log \frac{\theta}{1-\theta}\right)?$$

(part-2)

Definition

What is the empirical risk?

Definition

What is an M-estimator?

Asymptotic linearity implies asymptotic normality

T_n is called asymptotically linear if there exists a function $l_\theta : \mathcal{X} \to \mathbb{R}^p$ with $\mathbb{E}_\theta l_\theta(X) = 0, \operatorname{Cov}(l_\theta(X)) := V_\theta < \infty$ such that

$$T_n - g(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i) + o_{\mathbb{P}_\theta}\left(\frac{1}{\sqrt{n}}\right)$$

We have

$$T_n - g(\theta) = \bar{X}_n^2 - \mu^2 = 2\mu\left(\bar{X}_n - \mu\right) + \underbrace{\left(\bar{X}_n - \mu\right)^2}_{=\mathcal{O}_{\mathbb{P}_\theta}\left(\frac{1}{n}\right) = o_{\mathbb{P}_\theta}\left(\frac{1}{\sqrt{n}}\right)}$$

hence the influence function is

$$l_\theta(x) = 2\mu(x-\mu)$$

and the asymptotic variance is $V_\theta = \operatorname{Cov}(l_\theta(X)) = 4\mu^2\sigma^2$

We have

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

so the influence function is

$$l_\theta(x) = x - \mu$$

If moreover T_n is an asymptotically linear estimator of γ with influence function l_θ, then $h(T_n)$ is an asymptotically linear estimator of $h(\gamma)$ with influence function

$$\dot{h}(\gamma)^T l_\theta,$$

so we can write

$$h(T_n) - h(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\dot{h}(\gamma)^T l_\theta(X_i) + o_{\mathbb{P}_\theta}\left(\frac{1}{\sqrt{n}}\right).$$

Suppose T_n is an asymptotically normal estimator of $\gamma = g(\theta) \in \mathbb{R}^p$ with asymptotic covariance matrix V_θ. Suppose that a function $h : \mathbb{R}^p \to \mathbb{R}$ is differentiable at γ. Then $h(T_n)$ is an asymptotically normal estimator of $h(\gamma)$ with asymptotic variance

$$\dot{h}(\gamma)^T V_\theta \dot{h}(\gamma),$$

so we can write

$$\sqrt{n}\left(h(T_n) - h(\gamma)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \dot{h}(\gamma)^T V_\theta \dot{h}(\gamma)\right).$$

$$\begin{aligned}\dot{h}(\theta) &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)},\\ \dot{h}(\theta)^T V_\theta \dot{h}(\theta) &= (\dot{h}(\theta))^2 V_\theta \\ &= \frac{\theta(1-\theta)}{(\theta(1-\theta))^2} = \frac{1}{\theta(1-\theta)}\end{aligned}$$

and hence

$$\sqrt{n}\left(\log\frac{\bar{X}_n}{1-\bar{X}_n} - \log\frac{\theta}{1-\theta}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{\theta(1-\theta)}\right)$$

We know $T_n = X_n$ is an asymptotically linear estimator of $\gamma = g(\theta) = \theta$ with

$$l_\theta(x) = x - \theta, V_\theta = \mathbb{E}l_\theta(X)^2 = \theta(1-\theta).$$

Furthermore, we know the function $h(\theta) = \log\frac{\theta}{1-\theta}$ is differentiable for $\theta \in (0,1)$. So, we can use the δ-technique to show that $h(T_n)$ is an asymptotically linear estimator of $h(\gamma)$ with asymptotic variance $\dot{h}(\gamma)^T V_\theta \dot{h}(\gamma)$. So

$$h(\theta) = \log\frac{\theta}{1-\theta},$$

An M-estimator is the estimator which minimizes the empirical risk, so

$$\hat{\gamma}_n = \arg\min_{c\in\Gamma}\hat{R}_n(c).$$

It's also called an empirical risk minimizer (ERM).

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of X. Then the empirical risk is

$$\hat{R}_n(c) := \frac{1}{n}\sum_{i=1}^{n}\rho_c(x_i)$$

Suppose $\gamma = \mathbb{E}X$. What is an M-estimator for γ ?	DEFINITION *What is a Z-estimator?*
Find a Z-estimator γ for the situation $X \in \mathbb{R}, c \in \mathbb{R}, \rho_c(x) = (x-c)^2$.	*Find a Z-estimator γ for the situation $X \in \mathbb{R}, c \in \mathbb{R}, \rho_c(x) = \lvert x-c \rvert$. Describe the ideal, i.e. non-empirical version. (part-1)*
Find a Z-estimator γ for the situation $X \in \mathbb{R}, c \in \mathbb{R}, \rho_c(x) = \lvert x-c \rvert$. Describe the ideal, i.e. non-empirical version. (part-2)	*Find a Z-estimator γ for the situation $X \in \mathbb{R}, c \in \mathbb{R}, \rho_c(x) = \lvert x-c \rvert$. Describe the empirical version. (part-1)*
Find a Z-estimator γ for the situation $X \in \mathbb{R}, c \in \mathbb{R}, \rho_c(x) = \lvert x-c \rvert$. Describe the empirical version. (part-2)	*Consider the problem of finding an MLE estimator where the density of X is p_θ. What is the loss function $\rho_\theta(x)$ in this case?*
What is the Kullback-Leibler information?	*Show that if the loss function is $\rho_\theta(x) = -\log p_\theta(x)$ and the true parameter is $\theta \in \Theta$, then for any other $\vartheta \in \Theta$ the Kullback-Leibler information is non-negative* $K(\vartheta \mid \theta) = R(\vartheta) - R(\theta) \geq 0$ *holds.*

Suppose the parameter space is an open set $\Gamma \subset \mathbb{R}^p$. Suppose the partial derivative $$\psi_c(x) = \frac{\partial}{\partial c}\rho_c(x) = \dot{\rho}_c(x)$$ exists for all x. Then the Z-estimator $\hat{\gamma}_n$ is a solution for which the partial derivative of the empirical risk is 0 , so $\dot{\hat{R}}_n(c)\Big\|_{c=\hat{\gamma}_n} = 0$ where the derivative $\dot{\hat{R}}_n$ is $$\dot{\hat{R}}_n(c) = \frac{\partial}{\partial c}\hat{R}_n(c) = \frac{1}{n}\sum_{i=1}^{n}\psi_c(x_i)$$	Note that $$\gamma = \mathbb{E}X = \arg\min_{c\in\Gamma}\mathbb{E}\left[(X-c)^2\right]$$ so the M-estimator is simply the sample average, i.e. $$\hat{\gamma}_n = \arg\min_{c\in\Gamma}\frac{1}{n}\sum_{i=1}^{n}(X_i - c)^2 = \bar{X}_n$$
Let F be the distribution function $F(\cdot) = P(X \leq \cdot)$. Now the risk is (Partial integration) $$\begin{aligned} \mathbb{E}\|X-c\| &= \int_{x\leq c}(c-x)dF(x) + \int_{x>c}(x-c)dF(x) \\ &= \int_{x\leq c}(c-x)dF(x) - \int_{x>c}(x-c)d(1-F(x)) \\ &= \overbrace{\left[(c-x)F(x)\right\|_{-\infty}^{c}}^{=0} - \overbrace{\left[(x-c)(1-F(x))\right\|_{c}^{\infty}}^{=0} \\ &+ \int_{x\leq c}F(x)dx + \int_{x>c}(1-F(x))dx. \end{aligned}$$	We have $$\begin{aligned} \psi_c(x) &= \frac{\partial}{\partial c}\rho_c(x) = -2(x-c), \\ \dot{\hat{R}}_n(c)\Big\|_{c=\gamma} &= -2\left(\bar{X}_n - c\right)\Big\|_{c=\gamma} \triangleq 0 \\ \gamma &= \bar{X}_n. \end{aligned}$$
The empirical distribution function is $$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}1\{x_i \leq x\}$$ Then we have the empirical risk as $$\hat{R}_n(c) = \int_{x\leq c}\hat{F}_n(x)dx + \int_{x>c}\left(1-\hat{F}_n(x)\right)dx$$ with the derivative $\dot{\hat{R}}_n(c) = 2\hat{F}_n(c) - 1$	So, the derivative of risk is $$\dot{R}(c) = F(c) - (1-F(c)) = 2F(c) - 1.$$ Putting that to 0 , we get $$\dot{R}(\gamma) = 2F(\gamma) - 1 \triangleq 0 \Longrightarrow \gamma = F^{-1}\left(\frac{1}{2}\right).$$
It's $\rho_\theta(x) = -\log p_\theta(x)$	So, we need to set $$\hat{F}_n(\hat{\gamma}_n) \triangleq \frac{1}{2}$$ which makes $\hat{\gamma}_n$ the median, so by convention $$\hat{\gamma}_n := \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{for } n \text{ odd} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{for } n \text{ even} \end{cases}$$
Take any $\tilde{\vartheta} \in \Theta$ $$\begin{aligned} R(\tilde{\vartheta}) - R(\theta) &= -\mathbb{E}_\theta \log\frac{p_{\tilde{\vartheta}(x)}}{p_{\tilde{\theta}(x)}} \\ &\geq -\log\mathbb{E}_\theta\frac{p_{\tilde{\vartheta}(X)}}{p_\theta(X)} \quad (Jensen) \\ &= -\log\left[\int\frac{p_{\tilde{\vartheta}}}{p_\theta}p\theta\right] = -\log 1 = 0. \end{aligned}$$	Suppose $\Theta \subset \mathbb{R}^p$ and that the densities $p_\theta = dP_\theta/d\nu$ exist with respect to some σ-finite measure ν. The Kullback-Leibler information is then $$K(\tilde{\theta} \mid \theta) = \mathbb{E}_\theta \log\left(\frac{p_\theta(X)}{p_{\tilde{\theta}}(X)}\right)$$

LEMMA

State the lemma about uniform convergence of empirical M-estimators.

Show that if we have the uniform convergence

$$\sup_{c\in\Gamma}\left|\hat{R}_n(c)-R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, \mathbb{P}_{\theta^-\text{a.s.}}\ .$$

Then for the empirical risk minimizer $\hat{\gamma}_n$

$$R\left(\hat{\gamma}_n\right) \underset{n\to\infty}{\longrightarrow} R(\gamma), \mathbb{P}_{\theta^-}.\text{a.s.}$$

What does it mean for γ *to be well-separated?*

If γ *is well-separated, what does*

$$R\left(\hat{\gamma}_n\right) \xrightarrow[n\to\infty]{\mathbb{P}} R(\gamma)$$

imply regarding $\hat{\gamma}_n$ *and* γ *?*

What does the notation below concerning the (empirical) measure mean? Let $f:\mathcal{X}\to\mathbb{R}^p$.

1. $\hat{P}_n f$,
2. *Pf.*

Write $R(c)$ *and* $\hat{R}_n(c)$ *using the* P *and* $\hat{P}_n$ *notation.*

THEOREM

State the uniform law of large numbers

Show that if

(i) Γ *is compact,*

(ii) $c\mapsto\rho_c(x)$ *is continuous for all* x,

(iii) $\mathbb{E}\sup_{c\in\Gamma}|\rho_c(X)|<\infty$.

then for all ε *there exists a finite "" ε-bracketing set""* $\left\{\left[f_j^L, f_j^U\right]\right\}_{j=1}^N$ *such that (continued)*

(continued) for $f_j^L:\mathcal{X}\to\mathbb{R}, f_j^U:\mathcal{X}\to\mathbb{R}$

1. $f_j^U\geq f_j^L$
2. $P\left(f_j^U-f_j^L\right)\leq\varepsilon$, *and*
3. *for all* $c\in\Gamma$ *there exists a* $j\in\{1,\ldots,N\}$ *such that*

$$f_j^L\leq\rho_c\leq f_j^U$$

THEOREM

State the uniform law of large numbers

$$
\begin{aligned}
0 &\le R(\hat{\gamma}_n) - R(\gamma) \\
&= -\left[\hat{R}_n(\hat{\gamma}_n) - R(\hat{\gamma}_n)\right] + \left[\hat{R}_n(\gamma) - R(\gamma)\right] \\
&+ \underbrace{\hat{R}_n(\hat{\gamma}_n) - \hat{R}(\gamma)}_{\le 0 (\hat{\gamma}_n \text{ is ERM })} \\
&\le 2 \sup_c \left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, \mathbb{P}_{\theta^-}\text{a.s.}
\end{aligned}
$$

Suppose the uniform convergence

$$\sup_{c\in\Gamma} \left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, \mathbb{P}_\theta\text{-a.s.}$$

Then for the empirical risk minimizer $\hat{\gamma}_n$

$$R(\hat{\gamma}_n) \underset{n\to\infty}{\longrightarrow} R(\gamma), \mathbb{P}_{\theta^-}\text{.a.s.}$$

If γ is well-separated, then

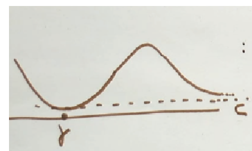
$$R(\hat{\gamma}_n) \xrightarrow[n\to\infty]{\mathbb{P}} R(\gamma)$$

implies

$$\|\hat{\gamma}_n - \gamma\| \xrightarrow[n\to\infty]{P} 0$$

γ is well-separated if for all $\varepsilon > 0$ we have

$$\inf\{R(c) : \|c - \gamma\| > \varepsilon\} > R(\gamma)$$



(Figure: γ not well separated)

1. $R(c) = P\rho_c$
2. $\hat{R}_n(c) = \hat{P}\rho_c$

1. $\hat{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$
2. $Pf = \mathbb{E}f(X)$

Let $\delta > 0$ and define

$$w(x, \delta, c) = \sup_{\tilde{c}\in\Gamma : \|\tilde{c} - c\| < \delta} \left|\rho_c(x) - \rho_{\tilde{c}}(x)\right|.$$

By continuity (ii) $\lim_{\delta\downarrow 0} w(x, \delta, c) = 0$ for all x.
By assumption (iii), the functions are dominated by their supremum which is finite, so we can use dominated convergence to show

$$\lim_{\delta\downarrow 0} Pw(\cdot, \delta, c) = 0$$

for all c.

Suppose

1. Γ is compact,
2. $c \mapsto \rho_c(x)$ is continuous for all x,
3. $\mathbb{E}\sup_{c\in\Gamma} |\rho_c(X)| < \infty$.

Then we have the uniform convergence

$$\sup_{c\in\Gamma} \left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, P_{\theta^-}\text{a.s.}$$

Suppose

1. Γ is compact,
2. $c \mapsto \rho_c(x)$ is continuous for all x,
3. $\mathbb{E}\sup_{c\in\Gamma} |\rho_c(X)| < \infty$.

Then we have the uniform convergence

$$\sup_{c\in\Gamma} \left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, P_{\theta^-}\text{a.s.}$$

So, $\forall\varepsilon > 0$ there exists a δ_c such that $Pw(\cdot, \delta_c, c) \le \frac{\varepsilon}{2}$. Define the open cover $B_c = \{\tilde{c} : \|\tilde{c} - c\| < \delta_c\}$ of Γ. By compactness (i) of Γ, there exists a finite subcover $\left\{B_{c_j}\right\}_{j=1}^N$. Let $c \in \Gamma$ be arbitrary, then there exists a $j : c \in B_{c_j}$ such that

$$
\begin{aligned}
\rho_c &\le \rho_{c_j} + w(\cdot, d_{c_j}, c_j) =: f_j^U \\
\rho_c &\ge \rho_{c_j} - w(\cdot, d_{c_j}, c_j) =: f_j^L
\end{aligned}
$$

Then $P\left(f_j^u - f_j^L\right) = 2Pw(\cdot, \delta_{c_j}, c_j) \le \varepsilon$ as we wished.

Suppose we wish to show that if

(i) Γ *is compact,*

(ii) $c \mapsto \rho_c(x)$ *is continuous for all* x,

(iii) $\mathbb{E}\sup_{c\in\Gamma}|\rho_c(X)| < \infty$.

then we have the uniform convergence

$$\sup_{c\in\Gamma}\left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, P_{\theta^-}\text{a.s.}$$

(continued)

(continued) Show that to show the above it is sufficient that there exists a finite ”” ε-bracketing set”” $\{[f_j^L, f_j^U]\}_{j=1}^N$ *such that for* $f_j^L : \mathcal{X} \to \mathbb{R}, f_j^U : \mathcal{X} \to \mathbb{R}$

1. $f_j^U \geq f_j^L$
2. $P\left(f_j^U - f_j^L\right) \leq \varepsilon$, *and*
3. *for all* $c \in \Gamma$ *there exists a* $j \in \{1, \ldots, N\}$ *such that*

$$f_j^L \leq \rho_c \leq f_j^U$$

Suppose we wish to show that if

(i) Γ *is compact,*

(ii) $c \mapsto \rho_c(x)$ *is continuous for all* x,

(iii) $\mathbb{E}\sup_{c\in\Gamma}|\rho_c(X)| < \infty$.

then we have the uniform convergence

$$\sup_{c\in\Gamma}\left|\hat{R}_n(c) - R(c)\right| \underset{n\to\infty}{\longrightarrow} 0, P_{\theta^-}\text{a.s.}$$

What are the key steps of the proof? (part-1)

LEMMA

State lemma about existence of Z-estimators.

PROOF

Suppose

1. $\Gamma \subset \mathbb{R}$,
2. $c \mapsto \psi_c(x)$ *is continuous for all* x,
3. $\mathbb{E}_\theta|\psi_c(x)| < \infty$ *for all* c
4. *there exists a* $\delta > 0$ *such that* $\dot{R}(c) > 0$ *for* $c \in (\gamma, \gamma+\delta)$ *and* $\dot{R}(c) < 0$ *for* $c \in (\gamma-\delta, \gamma)$.

Then $\mathbb{P}_\theta$*-a.s. there is a solution* $\hat{\gamma}_n$ *of* $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0$ *and this solution is consistent. (part-1)*

PROOF

Suppose

1. $\Gamma \subset \mathbb{R}$,
2. $c \mapsto \psi_c(x)$ *is continuous for all* x,
3. $\mathbb{E}_\theta|\psi_c(x)| < \infty$ *for all* c
4. *there exists a* $\delta > 0$ *such that* $\dot{R}(c) > 0$ *for* $c \in (\gamma, \gamma+\delta)$ *and* $\dot{R}(c) < 0$ *for* $c \in (\gamma-\delta, \gamma)$.

Then $\mathbb{P}_\theta$*-a.s. there is a solution* $\hat{\gamma}_n$ *of* $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0$ *and this solution is consistent. (part-2)*

PROOF

Suppose

1. $\Gamma \subset \mathbb{R}$,
2. $c \mapsto \psi_c(x)$ *is continuous for all* x,
3. $\mathbb{E}_\theta|\psi_c(x)| < \infty$ *for all* c
4. *there exists a* $\delta > 0$ *such that* $\dot{R}(c) > 0$ *for* $c \in (\gamma, \gamma+\delta)$ *and* $\dot{R}(c) < 0$ *for* $c \in (\gamma-\delta, \gamma)$.

Then $\mathbb{P}_\theta$*-a.s. there is a solution* $\hat{\gamma}_n$ *of* $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0$ *and this solution is consistent. (part-3)*

Let $X \in \mathbb{R}, \theta \in \mathbb{R}$. *Take the density*

$$p_\theta(x) = \frac{\exp(x-\theta)}{(1+\exp(x-\theta))^2}, \text{ for } x \in \mathbb{R}.$$

Let the loss ρ_θ *be*

$$\rho_\theta(x) = -(x-\theta) + 2\log(1+\exp(x-\theta))$$

with

$$\frac{\partial}{\partial c}\rho_\theta(x) = \frac{1-\exp(x-\theta)}{1+\exp(x-\theta)}.$$

Does there exist a consistent Z-estimator for this problem?

Consider a function $f : \mathcal{X} \to \mathbb{R}^p$. *What does* Pff^T *mean?*

State the multidimensional central limit theorem using (empirical) measure notation, i.e. using $\hat{P}_n f$ *and* Pf.

Hence

$$\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P\right)\rho_c\right| \le \underbrace{\max_{1\le j\le N}\max\left\{\left|\left(\hat{P}_n - P\right)f_j^L\right|, \left|\left(\hat{P}_n - P\right)f_j^U\right|\right\}}_{\underset{n\to\infty}{\longrightarrow} 0, P_\theta\text{-a.s.}} + \varepsilon$$

If we have this, then

$$\left(\hat{P}_n - P\right)\rho_c \le \hat{P}_n f_j^U - P f_j^U + \underbrace{P\left(f_j^U - \rho_c\right)}_{\le\varepsilon},$$

$$\left(\hat{P}_n - P\right)\rho_c \ge \hat{P}_n f_j^L - P f_j^L + \underbrace{P\left(f_j^L - \rho_c\right)}_{\ge-\varepsilon}.$$

Suppose

1. $\Gamma \subset \mathbb{R}$,
2. $c \mapsto \psi_c(x)$ is continuous for all x,
3. $\mathbb{E}_\theta\left|\psi_c(x)\right| < \infty$ for all c
4. there exists a $\delta > 0$ such that $\dot{R}(c) > 0$ for $c \in (\gamma, \gamma+\delta)$ and $\dot{R}(c) < 0$ for $c \in (\gamma-\delta, \gamma)$.

Then $\mathbb{P}_\theta$-a.s. there is a solution $\hat{\gamma}_n$ of $\dot{\hat{R}}_n\left(\hat{\gamma}_n\right) = 0$ and this solution is consistent.

1. Let $\delta > 0$ and define

$$w(x,\delta,c) = \sup_{\tilde{c}\in\Gamma:\|\tilde{c}-c\|<\delta}\left|\rho_c(x) - \rho_{\tilde{c}}(x)\right|$$

2. Use continuity and dominated convergence to show that for all $\varepsilon > 0$ there exists a δ_c such that $Pw\left(\cdot,\delta_c,c\right) \le \frac{\varepsilon}{2}$
3. Use the above to construct an open cover $B_c = \left\{\tilde{c} : \|\tilde{c} - c\| < \delta_c\right\}$ of Γ such that ρ_c is bounded from below and above by $\pm w\left(\cdot,\delta_c,c\right)$.
4. Use compactness to show that there is a finite number of such bounds. 5. Use previous card

On the set A we have

$$\dot{\hat{R}}_n(\gamma+\varepsilon) = \underbrace{\dot{\hat{R}}_n(\gamma+\varepsilon) - \dot{R}(\gamma+\varepsilon)}_{\ge-\tilde{\varepsilon}} + \underbrace{\dot{R}(\gamma+\varepsilon)}_{\ge 2\tilde{\varepsilon}} \ge \tilde{\varepsilon},$$

$$\dot{\hat{R}}_n(\gamma-\varepsilon) = \underbrace{\dot{\hat{R}}_n(\gamma-\varepsilon) - \dot{R}(\gamma-\varepsilon)}_{\le\tilde{\varepsilon}} + \underbrace{\dot{R}(\gamma-\varepsilon)}_{\le -2\tilde{\varepsilon}} \le -\tilde{\varepsilon}.$$

Now by continuity (ii) on the set A there is a $\hat{\gamma}_n \in (\gamma-\varepsilon, \gamma+\varepsilon)$ such that $\dot{\hat{R}}_n\left(\hat{\gamma}_n\right) = 0$ and $\left\|\hat{\gamma}_n - \gamma\right\| \underset{n\to\infty}{\overset{P}{\longrightarrow}} 0$

Let $0 < \varepsilon < \delta$. By (iv) there is an $\tilde{\varepsilon}$ such that

$$\dot{R}(\gamma+\varepsilon) \ge 2\tilde{\varepsilon} \text{ and } \dot{R}(\gamma-\varepsilon) \le -2\tilde{\varepsilon}.$$

Let

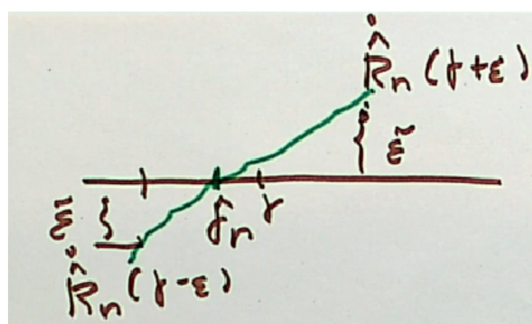
$$A = \left\{\dot{\hat{R}}_n(\gamma+\varepsilon) - \dot{R}(\gamma+\varepsilon) > -\tilde{\varepsilon}, \dot{\hat{R}}_n(\gamma-\varepsilon) - \dot{R}(\gamma-\varepsilon) < \tilde{\varepsilon}\right\}$$

Then by the uniform law of large numbers $\mathbb{P}\left(A^C\right) \underset{n\to\infty}{\longrightarrow} 0$.

The conditions for existence of consistent Z-estimators hold, so yes:

1. $\vartheta \in \Gamma \subset \mathbb{R}$
2. $\vartheta \mapsto \psi_\vartheta(x) = \frac{\partial}{\partial\vartheta}\rho_\vartheta(x)$ is continuous for all x,
3. $\left|\psi_\vartheta(x)\right| \le 1$, so $\mathbb{E}_\theta\left|\psi_\vartheta(X)\right| < \infty$ for all ϑ,
4. since $\psi_\vartheta(x)$ passes 0 and is continuous, for $\dot{R}(\vartheta) = \mathbb{E}_\theta\psi_\vartheta(X)$ there exists a $\delta > 0$ such that

$$\dot{R}(c) > 0 \text{ for } c \in (\gamma, \gamma+\delta) \text{ and } \dot{R}(c) < 0, c \in (\gamma-\delta, \gamma)$$



Suppose $X_1, \ldots, X_n, \ldots$ are i.i.d. copies of $X \in \mathcal{X}$, and $f : \mathcal{X} \to \mathbb{R}^p$ is a function with

$$Pf < \infty \text{ and } \Sigma_f = Pff^T - (Pf)\left(Pf^T\right) < \infty,$$

then

$$\sqrt{n}\left(\hat{P}_n - P\right)f \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}\left(0, \Sigma_f\right).$$

$$Pff^T = \begin{pmatrix} \mathbb{E}f_1^2(X) & \cdots & \mathbb{E}f_1(X)f_p(X) \\ \vdots & \ddots & \vdots \\ \mathbb{E}f_1(X)f_p(X) & \cdots & \mathbb{E}f_p^2(X) \end{pmatrix}$$

What is an empirical process?

Suppose

$$\nu_n(c) = \sqrt{n}\left(\hat{P}_n - P\right)\psi_c$$

is an empirical process. What and how does $\nu_n(c)$ converge if $P\psi_c\psi_c^T = \Sigma_c < \infty$?

What does it mean for an empirical process $\nu_n(\cdot)$ to be asymptotically continuous?

Theorem

State the theorem about asymptotic linearity of Z-estimators.

Sketch proof: suppose

1. $\hat{\gamma}_n \xrightarrow[n\to\infty]{\mathbb{P}} \gamma$
2. $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0, \dot{R}(\gamma) = 0$
3. ν_n is asymptotically continuous at γ,
4. $M_\theta = \frac{\partial}{\partial\gamma^T}\dot{R}(\gamma)$ exists and is invertible,
5. $J_\theta := P\psi_\gamma\psi_\gamma^T < \infty$

Then $\hat{\gamma}_n$ has influence function $l_\theta = -M_\theta^{-1}\psi_\gamma$. (part-1)

Sketch proof: suppose

1. $\hat{\gamma}_n \xrightarrow[n\to\infty]{\mathbb{P}} \gamma$
2. $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0, \dot{R}(\gamma) = 0$
3. ν_n is asymptotically continuous at γ,
4. $M_\theta = \frac{\partial}{\partial\gamma^T}\dot{R}(\gamma)$ exists and is invertible,
5. $J_\theta := P\psi_\gamma\psi_\gamma^T < \infty$

Then $\hat{\gamma}_n$ has influence function $l_\theta = -M_\theta^{-1}\psi_\gamma$. (part-2)

Sketch proof: suppose

1. $\hat{\gamma}_n \xrightarrow[n\to\infty]{\mathbb{P}} \gamma$
2. $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0, \dot{R}(\gamma) = 0$
3. ν_n is asymptotically continuous at γ,
4. $M_\theta = \frac{\partial}{\partial\gamma^T}\dot{R}(\gamma)$ exists and is invertible,
5. $J_\theta := P\psi_\gamma\psi_\gamma^T < \infty$

Then $\hat{\gamma}_n$ has influence function $l_\theta = -M_\theta^{-1}\psi_\gamma$. (part-3)

What condition must be true to be able to write

$$\dot{\hat{R}}_n(\hat{\gamma}_n) - \dot{R}(\hat{\gamma}_n) = \dot{\hat{R}}_n(\gamma) - \dot{R}(\gamma) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)?$$

Suppose $M_\theta = \frac{\partial}{\partial c^T}\dot{R}(\gamma)$. Give an asymptotic Taylor approximation to $\dot{R}(\hat{\gamma}_n) - \dot{R}(\gamma)$?

Suppose we know $J_\theta = P\psi_\gamma\psi_\gamma^T < \infty$. What does that tell us about the order symbol of $\frac{1}{\sqrt{n}}\nu_n(\gamma)$?

If $P\psi_c\psi_c^T = \Sigma_c < \infty$, then $\nu_n(c) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \Sigma_c\right)$

We call

$$\nu_n(c) = \sqrt{n}\left(\hat{P}_n - P\right)\psi_c$$

an empirical process indexed by $c \in \Gamma$.

1. $\hat{\gamma}_n \xrightarrow[n\to\infty]{\mathbb{P}} \gamma$
2. $\dot{\hat{R}}_n\left(\hat{\gamma}_n\right) = 0, \dot{R}(\gamma) = 0$
3. ν_n is asymptotically continuous at γ,
4. $M_\theta = \frac{\partial}{\partial\gamma^T}\dot{R}(\gamma)$ exists and is invertible,
5. $J_\theta := P\psi_\gamma\psi_\gamma^T < \infty$

Then $\hat{\gamma}_n$ has influence function $l_\theta = -M_\theta^{-1}\psi_\gamma$. Furthermore

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \underbrace{M_\theta^{-1}J_\theta M_\theta^{-1}}_{\text{" sandwich formula""}})$$

An empirical process $\nu_n(\cdot)$ is called asymptotically continuous at γ if

$$\left|\nu_n\left(\gamma_n\right) - \nu_n(\gamma)\right| \xrightarrow{\mathbb{P}} 0 \text{ as } \gamma_n \underset{n\to\infty}{\mathbb{P}} \gamma$$

More formally, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\mathbb{P}\left(\sup_{\|c-\gamma\|\leq\delta}\left|\nu_n(c) - \nu_n(\gamma)\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

$$(*) = \frac{1}{\sqrt{n}}\nu_n(\gamma) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + M_\theta\left(\left[\hat{\gamma}_n - \gamma\right]\left[1 + o_{\mathbb{P}}(1)\right]\right)$$

$\left(\text{Assum. (v)} \implies \nu_n(\gamma) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, J_\theta\right) \implies \nu_n(\gamma) = \mathcal{O}_{\mathbb{P}}(1)\right)$

$$= \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + M_\theta\left(\left[\hat{\gamma}_n - \gamma\right]\left[1 + o_{\mathbb{P}}(1)\right]\right)$$

$\implies M_\theta\left(\left[\hat{\gamma}_n - \gamma\right]\left[1 + o_{\mathbb{P}}(1)\right]\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$ by (iv) M_θ is invertible $\implies \|\hat{\gamma}_n - \gamma\| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$

$$\begin{aligned}
0 &= \dot{\hat{R}}_n\left(\hat{\gamma}_n\right) \quad (2.)\\
&= \underbrace{\dot{\hat{R}}_n\left(\hat{\gamma}_n\right) - \dot{R}\left(\hat{\gamma}_n\right)}_{=\frac{1}{\sqrt{n}}\nu_n\left(\hat{\gamma}_n\right)} + \dot{R}\left(\hat{\gamma}_n\right)\\
&= \underbrace{\dot{\hat{R}}_n(\gamma) - \dot{R}(\gamma)}_{=\frac{1}{\sqrt{n}}\nu_n(\gamma)} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + \dot{R}\left(\hat{\gamma}_n\right) \quad \text{(3.asymp cont)}\\
&= \frac{1}{\sqrt{n}}\nu_n(\gamma) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + \dot{R}\left(\hat{\gamma}_n\right) - \underbrace{\dot{R}(\gamma)}_{=0}
\end{aligned}$$

$$\nu_n(\gamma) = \left(\hat{P}_n - P\right)\psi_\gamma = \dot{\hat{R}}_n(\gamma) - \dot{R}(\gamma)$$

must be asymptotically continuous at γ.

$$(*) = \tfrac{1}{\sqrt{n}}\nu_n(\gamma) + o_{\mathbb{P}}\left(\tfrac{1}{\sqrt{n}}\right) + M_\theta\left(\hat{\gamma}_n - \gamma\right) + o_{\mathbb{P}}\left(\tfrac{1}{\sqrt{n}}\right)$$

$$\begin{aligned}
\implies M_\theta\left(\hat{\gamma}_n - \gamma\right) &= \frac{1}{\sqrt{n}}\nu_n(\gamma) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\\
\left(\hat{\gamma}_n - \gamma\right) &= \frac{1}{\sqrt{n}}M_\theta^{-1}\nu_n(\gamma) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\\
&= \hat{P}_n M_\theta^{-1}\psi_\gamma + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\\
&= \hat{P}_n l_\theta + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

$J_\theta = P\psi_\gamma\psi_\gamma^T < \infty$ implies $\nu_n(\gamma) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, J_\theta\right)$, so the order of $\nu_n(\gamma) = \mathcal{O}(1)$, and so finally

$$\frac{1}{\sqrt{n}}\nu_n(\gamma) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

$$\dot{R}\left(\hat{\gamma}_n\right) - \dot{R}(\gamma) = M_\theta\left(\left[\hat{\gamma}_n - \gamma\right]\left[1 + o_{\mathbb{P}}(1)\right]\right)$$

Consider the maximum likelihood estimator $\hat{\gamma}_n$, for $X \sim P_\theta$ with density p_θ, $\theta \in \mathbb{R}^p$, the loss function

$$\rho_\vartheta(x) = -\log p_\vartheta(x)$$

and $\gamma = \theta$. Assuming regularity conditions hold, what can you say about the convergence of

$$\sqrt{n}\,(\hat{\gamma}_n - \gamma)?$$

What are the steps to find the influence function and asymptotic variance of a Z-estimator?

Suppose

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ 2(|x| - k) & \text{if } |x| > k \end{cases}$$

and $\rho_c(x) = \rho(x - c)$ for $c \in \mathbb{R}$, so ρ_c is the Huber loss. What is the influence function and asymptotic variance of an estimator $\hat{\gamma}_n$ which minimizes the Huber loss? (part-1)

Suppose

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ 2(|x| - k) & \text{if } |x| > k \end{cases}$$

and $\rho_c(x) = \rho(x - c)$ for $c \in \mathbb{R}$, so ρ_c is the Huber loss. What is the influence function and asymptotic variance of an estimator $\hat{\gamma}_n$ which minimizes the Huber loss? (part-2)

Suppose

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ 2(|x| - k) & \text{if } |x| > k \end{cases}$$

and $\rho_c(x) = \rho(x - c)$ for $c \in \mathbb{R}$, so ρ_c is the Huber loss. What is the influence function and asymptotic variance of an estimator $\hat{\gamma}_n$ which minimizes the Huber loss? (part-3)

DEFINITION

What is asymptotic relative efficiency?

Consider two estimators $T_{n,1}, T_{n_2}$. What does it mean if the asymptotic relative efficiency $e_{2,1} > 1$?

What is the $(1-\alpha)$ asymptotic confidence interval based on an asymptotically normal estimator T_n of γ with asymptotic variance V_θ ?

Consider two asymptotically normal estimators $T_{\theta,1}, T_{\theta,2}$ with asymptotic relative efficiency $e_{2,1}$. What must be the ratio of sample size n_1/n_2 for them to produce asymptotic confidence intervals of equal length?

Suppose $X = \mu + \varepsilon$, where $\varepsilon \sim F_0$ symmetric around zero with density f_0, and $\mathrm{Var}(\varepsilon) := \sigma^2 < \infty$. Let $X_1, \dots, X_n, \dots$ i. $\overset{\text{i.i.d.}}{\sim}$ X. Consider the estimators

$T_{n,1} := \bar{X}_n$	w/ asymp. var. $V_{\theta,1} = \sigma^2$
$T_{n,2} :=$ sample median	w/ asymp. var. $V_{\theta,2} = \frac{1}{4f_0^2(0)}$

If $F_0 = \Phi$ is the standard normal distribution, what is the asymptotic relative efficiency and which estimator is more efficient? Recall that $\Phi'(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right]$

1. Find $\psi_c = \frac{\partial}{\partial c}\rho_c(x)$.
2. Compute $\dot{R}(c) = \mathbb{E}\psi_c(x)$, and find γ for which $\dot{R}(\gamma) = 0$.
3. Determine $M_\theta = \frac{\partial}{\partial c^T T}\dot{R}(c)\Big|_{c=\gamma}$.
4. The influence function is $l_\theta = -M_\theta^{-1}\psi_\gamma$.
5. Compute $J_\theta = P\psi_\gamma\psi_\gamma^T$.
6. The asymptotic variance is $V_\theta = M_\theta^{-1}J_\theta M_\theta^{-1}$.
7. If the asymptotic variance depends on the unknown distribution and γ, by Slutsky's theorem we can plug in empirical estimates and retain the asymptotic properties.

Under regularity conditions

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, I^{-1}(\gamma)\right)$$

3. The matrix M_θ is then

$$\frac{\partial}{\partial c}\mathbb{E}\psi_c(x)\Big|_{c=\gamma} = 2F(c+k) - 2F(c-k) = M_\theta$$

4. The influence function is then

$$\begin{aligned} l_\theta &= -M_\theta^{-1}\psi_\gamma \\ &= \frac{-1}{2F(\gamma+k) - 2F(\gamma-k)}\begin{cases} -2(x-\gamma) & |x-\gamma| \le k \\ 2k & x-\gamma < -k \\ -2k & x-\gamma > k \end{cases} \end{aligned}$$

1. First we need to find $\psi_c(x) = \frac{\partial}{\partial c}\rho_c(x)$ as

$$\psi(x) = \begin{cases} 2x & |x| \le k \\ -2k & x < -k \\ 2k & x > k, \end{cases}, \psi_c(x) = \begin{cases} -2(x-c) & |x-c| \le k \\ 2k & x-c < -k \\ -2k & x-c > k \end{cases}$$

2. Then we need to find $M_\theta = \frac{\partial}{\partial c^T}\dot{R}(c)\Big|_{c=\gamma}$, so compute $\dot{R}(c)$ as

$$\dot{R}(c) = \mathbb{E}\psi_c(x) = (\text{ long computations }) = 2\left\{\int_{c-k}^{c+k} F(x)dx - k\right\}.$$

so we get γ by setting the above to 0.

Suppose $\gamma \in \Gamma \subset \mathbb{R}$, and for two estimators $T_{n,1}, T_{n,2}$ we have

$$\sqrt{n}\left(T_{n,1} - \gamma\right) \xrightarrow[n\to\infty]{\mathcal{D}_\theta} \mathcal{N}\left(0, V_{\theta,1}\right),$$

$$\sqrt{n}\left(T_{n,2} - \gamma\right) \xrightarrow[n\to\infty]{\mathcal{D}_\theta} \mathcal{N}\left(0, V_{\theta,2}\right)$$

Then the asymptotic relative efficiency (a.r.e) is

$$e_{2,1} := \frac{V_{\theta,1}}{V_{\theta,2}}$$

5. The asymptotic variance is $V_\theta = M_\theta^{-1}J_\theta M_\theta^{-1}$, where $J_\theta = P\psi_\gamma\psi_\gamma^T$. So

$$V_\theta = \frac{1}{(F(\gamma+k) - F(\gamma-k))^2} \times \left\{\int_{|x-\gamma|\le k}(x-\gamma)^2 dF(x) + k^2F(\gamma-k) + k^2(1-F(\gamma+k))\right\}.$$

6. The asymptotic variance depends on the unknown distribution F. However, by Slutsky's theorem the asymptotic variance holds even if we plug in the empirical estimates for F and γ.

The $(1-\alpha)$ asymptotic confidence interval is then

$$\hat{\gamma}_n \pm \sqrt{\frac{V_\theta}{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

If $e_{2,1} > 1$, then $T_{n,2}$ is asymptotically more efficient than $T_{n,1}$.

The a.r.e is $e_{2,1} = 4\sigma^2 f_0^2(0)$. If $F_0 = \Phi$, then $f_0 = \phi$. Hence, $\sigma^2 = 1$, $f_0(0) = \frac{1}{\sqrt{2\pi}}$, so a.r.e. is $e_{2,1} = \frac{4}{2\pi} = \frac{2}{\pi} < 1$ so the estimator $T_{n,1} = \bar{X}_n$ is more efficient.

$$\begin{aligned} \frac{\text{length}_1}{\text{length}_2} &= \frac{2\sqrt{\frac{V_{\theta,1}}{n_1}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{\frac{V_{\theta,2}}{n_2}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)} = \sqrt{\frac{n_2}{n_1}}\sqrt{\frac{V_{\theta,1}}{V_{\theta,2}}} \\ &= \sqrt{\frac{n_2}{n_1}}\sqrt{e_{2,1}} \triangleq 1, \\ \frac{n_1}{n_2} &= e_{2,1}. \end{aligned}$$

Suppose $X = \mu + \varepsilon$, where $\varepsilon \sim F_0$ symmetric around zero with density f_0, and $\operatorname{Var}(\varepsilon) := \sigma^2 < \infty$. Let $X_1, \ldots, X_n, \ldots$ i. $\overset{\text{i.i.d.}}{\sim}$ X. Consider the estimators

$T_{n,1} := \bar{X}_n$	*w/ asymp. var.* $V_{\theta,1} = \sigma^2$
$T_{n,2} :=$ *sample median*	*w/ asymp. var.* $V_{\theta,2} = \frac{1}{4f_0^2(0)}$

If F_0 is the Laplace distribution, what is the asymptotic relative efficiency and which estimator is more efficient?

$$f_0(x) = \frac{1}{\sqrt{2}} \exp[-\sqrt{2}|x|]$$

Take a random variable $Z \in \mathbb{R}^p$ with $Z \sim \mathcal{N}(0, \Sigma), \Sigma > 0$. What is the distribution of $Z^T \Sigma^{-1} Z$?

Definition

What is an asymptotic pivot?

What are the two main ways to construct asymptotic pivots based on asymptotic covariance?

Suppose $\hat{\theta}_n$ is a consistent estimator of θ, and $\theta \mapsto V_\theta$ is continuous. Give a consistent estimator of the asymptotic covariance V_θ.

Consider an M-estimator with asymptotic variance $V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1}$. Give a consistent estimator of the asymptotic covariance V_θ.

Suppose θ_n is the maximum likelihood estimator for θ. What's an asymptotic pivot for θ that does not rely on the covariance V_θ ?

Suppose $\hat{\theta}_n$ is the maximum likelihood estimator for θ, with density $p_\theta(x)$ for $\theta \in \Theta \subset \mathbb{R}^p$. Show three possible asymptotic pivots for $H_0 : \theta = \theta_0$, and what you need to compute to derive them. What are their distributions?
(part-1)

Suppose $\hat{\theta}_n$ is the maximum likelihood estimator for θ, with density $p_\theta(x)$ for $\theta \in \Theta \subset \mathbb{R}^p$. Show three possible asymptotic pivots for $H_0 : \theta = \theta_0$, and what you need to compute to derive them. What are their distributions?
(part-2)

Lemma

State the lemma about asymptotic convergence of restricted MLE.
(part-1)

See that $\Sigma^{-\frac{1}{2}}Z \sim \mathcal{N}(0, I)$. Now we get $Z^T\Sigma^{-1}Z = \left\|\Sigma^{-\frac{1}{2}}Z\right\|^2 \sim \chi_p^2$

The a.r.e is $e_{2,1} = 4\sigma^2 f_0^2(0)$. Variance is

$$\sigma^2 = \mathbb{E}\varepsilon^2 = \sqrt{2}\int_0^\infty x^2 \exp[-\sqrt{2}x]dx = 1.$$

$f_0(0) = \frac{1}{\sqrt{2}}$, so a.r.e is $e_{2,1} = \frac{4}{2} = 2 > 1$, so the median is more efficient than the mean.

1. Suppose the asymptotic variance $V_\theta = V(\gamma)$ depends only on γ. Then an asymptotic pivot is

$$Z_{n,1}(\gamma) := n\left(T_n - \gamma\right)V(\gamma)^{-1}\left(T_n - \gamma\right)$$

and we have $Z_{n,1}(\gamma) \xrightarrow[n\to\infty]{\mathcal{D}} \chi_p^2$.
2. If instead we have a consistent estimator of the variance $\hat{V}_n$, i.e. $\forall\theta$ it's true that $\hat{V}_n \xrightarrow[n\to\infty]{\mathbb{P}_\theta} V_\theta$, then an asymptotic pivot is

$$Z_{n,2}(\gamma) = n\left(T_n - \gamma\right)\hat{V}_n^{-1}\left(T_n - \gamma\right)$$

and using Slutsky's theorem we have again $Z_{n,2}(\gamma) \xrightarrow[n\to\infty]{\mathcal{D}_\theta} \chi_p^2$.

Suppose $X \sim P_\theta, X_1, \ldots, X_n, \ldots$ are i.i.d. copies of X, and we are testing. a hypothesis $H_0 : \gamma = \gamma_0$. We call a function $Z_n(\gamma) = Z_n\left(\gamma, X_1, \ldots, X_n\right)$ an asymptotic pivot if

$$Z_n(\gamma) \xrightarrow[n\to\infty]{\mathcal{D}_\theta} Z$$

for all θ, that is the distribution of Z does not depend on parameters θ.

Just replace the true measure with the empirical measure. So instead of

$$M_\theta = \frac{\partial}{\partial c^T}\mathbb{E}\psi_c(x)\bigg|_{c=\gamma} = P\dot{\psi}_\gamma,$$
$$J_\theta = \mathbb{E}\psi_\gamma(x)\psi_\gamma(x)^T = P\psi_\gamma\psi_\gamma^T,$$

use

$$\hat{M}_n := \hat{P}_n\dot{\psi}_{\hat{\gamma}_n} = \frac{\partial^2}{\partial c\partial c^T}\frac{1}{n}\sum_{i=1}^n \rho_c\left(x_i\right)\bigg|_{c=\hat{\gamma}_n},$$
$$\hat{J} := \hat{P}_n\psi_{\hat{\gamma}_n}\psi_{\hat{\gamma}_n}^T$$
$$\hat{V}_n := \hat{M}_n^{-1}\hat{J}_n\hat{M}_n^{-1}$$

$$\hat{V}_n = V_{\hat{\theta}_n}$$

We need

$$\rho_\vartheta(x) = -\log p_\vartheta(x),$$
$$\psi_\vartheta(x) = -s_\vartheta(x) = -\frac{\dot{p}_\vartheta(x)}{p_\vartheta(x)},$$
$$I(\theta) = -\mathbb{E}\dot{s}_\theta(x),$$
$$M_\theta = -P\dot{s}_\theta = I(\theta),$$
$$J_\theta = Ps_\theta s_\theta^T = I(\theta)$$
$$V_\theta = M_\theta^{-1}J_\theta M_\theta^{-1} = I(\theta)^{-1},$$
$$l_\theta = I(\theta)^{-1}s_\theta.$$

We can construct an asymptotic pivot using the twice log-likelihood ratio

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) := 2\sum_{i=1}^n\left[\log p_{\hat{\theta}_n}\left(X_i\right) - \log p_\theta\left(X_i\right)\right].$$

Under regularity conditions, we have the convergence

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) \xrightarrow[n\to\infty]{\mathcal{D}_\theta} \chi_p^2$$

for all θ.

Suppose we are testing the hypothesis $H_0 : R(\theta) = 0$ where R is a vector of restrictions

$$R(\theta) = \begin{pmatrix} R_1(\theta)\cdots R_q(\theta) \end{pmatrix}^\top.$$

Let $\hat{\theta}_n$ and $\hat{\theta}_n^0$ be the unrestricted and restricted MLE, i.e.

$$\hat{\theta}_n = \arg\max_{\vartheta\in\Theta}\sum_{i=1}^n \log p_\vartheta\left(X_i\right)$$

$$\hat{\theta}_n^0 = \arg\max_{\vartheta\in\Theta:R(\vartheta)=0}\sum_{i=1}^n \log p_\vartheta\left(X_i\right).$$

Then we have the three asymptotic pivots

1. $Z_{n,1}(\theta) = n\left(\hat{\theta}_n - \theta\right)^T I(\theta)\left(\hat{\theta}_n - \theta\right)$,
2. $Z_{n,2}(\theta) = n\left(\hat{\theta}_n - \theta\right)I\left(\hat{\theta}_n\right)\left(\hat{\theta}_n - \theta\right)$,
3. $Z_{n,3}(\theta) = 2\left[\mathcal{L}_n\left(\hat{\theta}_n\right) - \mathcal{L}_n(\theta)\right]$ where $\mathcal{L}_n(\vartheta) = \sum_{i=1}^n \log p_\vartheta\left(x_i\right)$.

They all asymptotically have the χ_p^2 distribution.

Lemma

State the lemma about asymptotic convergence of restricted MLE. (part-2)

Furthermore, let

$$\mathcal{L}_n\left(\hat{\theta}_n\right) - \mathcal{L}_n\left(\hat{\theta}_n^0\right) = \sum_{i=1}^{n}\left[\log p_{\hat{\theta}_n}\left(X_i\right) - \log p_{\hat{\theta}_n^0}\right]$$

be the log-likelihood ratio for testing $H_0 : R(\theta) = 0$. Then under regularity conditions we have

$$2\left[\mathcal{L}_n\left(\hat{\theta}_n\right) - \mathcal{L}_n\left(\hat{\theta}_n^0\right)\right] \xrightarrow[n\to\infty]{\mathcal{D}_\theta} \chi_q^2$$