HS2023

Summary - Fundamentals of Mathematical Statistics

Estimator: Given observations $\vec{X} = (X_1, \dots, X_n)$, realizations of a random variable X with distribution $P_\theta \in \{P_\vartheta \mid \vartheta \in \Theta\}$. Then $T(\vec{X})$ is an estimator. $T(\cdot)$ is measurable, and not allowed to depend on unknown parameters. (i.e. θ)
(aka. statistic or decision).

Method of moments: $X \in \mathbb{R}$, Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be the parameter of interest. The first p moments of X are given by

$$\mu_j(\theta) = E_\theta[X^j] = \int x^j \, dF_\theta(x) \qquad j = 1, \dots, p$$

Further assume that $m: \Theta \to \mathbb{R}^p$, $m(\theta) \mapsto [\mu_1(\theta), \dots, \mu_p(\theta)]$ has an inverse $m^{-1}(\mu_1, \dots, \mu_p) \ \forall \mu_1, \dots, \mu_p$.

By estimating the μ_j by their sample counterparts

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j \left(= \int x^j \, d\hat{F}_n(x) \right) \qquad j = 1, \dots, p$$

we obtain the method of moments estimator

$$\hat{\theta} := m^{-1}(\hat{\mu}_1, \dots, \hat{\mu}_p)$$

example: negative Binomial distribution $\theta \in (0,1)$ (thus $p=1$)

$$P_\theta(X = x) = \binom{k+x-1}{x} \theta^k (1-\theta)^x \qquad x \in \{0, 1, 2, \dots\}$$

then $E_\theta[X^1] = k \frac{1-\theta}{\theta} := m(\theta) \Rightarrow m^{-1}(\mu) = \frac{k}{\mu + k}$

$$\Rightarrow \hat{\theta} = \frac{k}{\hat{\mu} + k} = \frac{k}{\bar{X} + k} = \frac{nk}{\sum_{i=1}^n X_i + nk}$$

example: Pareto distribution $p_\theta(x) = \theta (1+x)^{-(1+\theta)}$, $x > 0$

$$E_\theta[X] = \frac{1}{\theta - 1} := m(\theta) \quad (> 0 \text{ by definition; Thus } \theta > 1)$$

$$\Rightarrow m^{-1}(\mu) = \frac{1+\mu}{\mu} \Rightarrow \hat{\theta} = \frac{1 + \bar{X}}{\bar{X}}$$

Maximum Likelihood Estimator (MLE): For data $\vec{X} = (X_1, \dots, X_n)$, the MLE is given by

$$\hat{\theta} := \arg\max_{\vartheta \in \Theta} L_{\vec{X}}(\vartheta)$$

where $L_{\vec{X}}(\vartheta): \Theta \to \mathbb{R}$ is the Likelihood function, given by

$$L_{\vec{X}}(\vartheta) := \prod_{i=1}^n p_\vartheta(X_i) \quad , \ \vartheta \in \Theta .$$

Alternatively, the MLE can be written as the maximizer of the log-likelihood:

$$\hat{\theta} = \arg\max_{\vartheta \in \Theta} \log(L_{\vec{X}}(\vartheta)) = \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \log(p_\vartheta(X_i))$$

(this is easier to differenciate for eg. finding a maximum of the fct.)

↓(MLE)

- The likelihood function may have local maxima, is not always unique, maxima or may not exist (unbounded)

The MLE is a plug-in estimator of θ, because we can write

$$\hat{\theta} = \arg\max_{\vartheta \in \Theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log(P_\vartheta(X_i))}_{\text{this is an empirical expected value}} \Rightarrow \arg\max_{\vartheta \in \Theta} E_\theta[\log(P_\vartheta(X))] = \theta$$

example: Pareto distribution, $\theta \in (0,\infty)$, $P_\theta(x) = \theta(1+x)^{-(1+\theta)}$ $x>0$

For $\underline{X} = (X_1, \dots, X_n)$ we maximize the function

$$\log(L_{\underline{X}}(\vartheta)) = \log\left(\prod_{i=1}^{n} P_\vartheta(X_i)\right) = \sum_{i=1}^{n} \log(P_\vartheta(X_i)) \quad \text{w.r.t. } \vartheta$$

we take the derivative

$$0 = \sum_{i=1}^{n}\left(\frac{1}{\vartheta} - \log(1+X_i)\right)$$

$$\Rightarrow \hat{\theta} = \frac{n}{\left(\sum_{i=1}^{n} \log(1+X_i)\right)}$$

example: standard normal distribution, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0,\infty)$

$$P_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad x \in \mathbb{R}$$

the respective MLE's are given by

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

example: (MLE does not exist). Each observation is distributed either $\mathcal{N}(\mu, 1)$ or $\mathcal{N}(\mu, \sigma^2)$, each with probability $\frac{1}{2}$. $\theta = (\mu, \sigma^2)$

Φ: density of normal distribution

$$P_\theta(x) = \frac{1}{2}\Phi(x-\mu) + \frac{1}{2\sigma}\Phi\left(\frac{x-\mu}{\sigma}\right)$$

then

$$L_{\underline{X}}(\tilde{\mu}, \tilde{\sigma}^2) = \prod_{i=1}^{n}\left(\frac{1}{2}\Phi(X_i - \tilde{\mu}) + \frac{1}{2\tilde{\sigma}}\Phi\left(\frac{X_i - \tilde{\mu}}{\tilde{\sigma}}\right)\right)$$

we can show that $L_{\underline{X}}$ for certain $\tilde{\mu}, \tilde{\sigma}^2$ is unbounded (thus no maximum exists)

take $\tilde{\mu} = X_1$:

$$L_{\underline{X}}(X_1, \tilde{\sigma}^2) = \underbrace{\frac{1}{\sqrt{2\pi}}\left(\frac{1}{2} + \frac{1}{2\tilde{\sigma}}\right)}_{\text{for } \tilde{\sigma} \to 0,\ \to +\infty} \underbrace{\prod_{i=2}^{n}\left(\frac{1}{2}\Phi(X_i - X_1) + \frac{1}{2\tilde{\sigma}}\Phi\left(\frac{X_i - X_1}{\tilde{\sigma}}\right)\right)}_{>0 \quad (*)}$$

(*): further note that $\forall z \neq 0 \quad \lim_{\tilde{\sigma} \searrow 0} \frac{1}{\tilde{\sigma}}\Phi\left(\frac{z}{\tilde{\sigma}}\right) = 0$, thus

$$\lim_{\tilde{\sigma} \searrow 0} (*) = \prod_{i=2}^{n} \frac{1}{2}\Phi(X_i - X_1) > 0$$

$$\Rightarrow \lim_{\tilde{\sigma} \searrow 0} L_{\underline{X}}(X_1, \tilde{\sigma}^2) = \infty.$$

Conditional Distributions (background knowledge)

The conditional probability of set A given set B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{where } P(B) \neq 0.$$

Law of total probability: For B_j a partition of Ω, then

$$P(A) = \sum_j P(A|B_j) \cdot P(B_j)$$

Marginal densities: $f(x,y)$, $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$ be a density

$$f_X(x) = \int_{x \in \mathbb{R}^n} f(x,y)\,dy \quad ; \quad f_Y(y) = \int f(x,y)\,dx, \quad y \in \mathbb{R}^m$$

conditional density: The cond. density of X given $Y=y$ is

$$f_X(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad x \in \mathbb{R}^n$$

thus we get, similarly to the law of total probability

$$f_X(x) = \int f_X(x|y)\, f_Y(y)\, dy$$

conditional expectation: Let $g: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. The conditional expectation of $g(X,Y)$ given $Y=y$ is

$$E[g(X,Y)\,|\,Y=y] := \int f_X(x|y) \cdot g(x,y)\, dx \quad (=h(y))$$

→ this is a function dependent on y.
→ this defines a random variable $h(Y)$

thus note $E[g_1(X)\, g_2(Y)\,|\,Y=y] = g_2(y)\, E[g_1(X)\,|\,Y=y]$

Law of iterated expectation:

$$E[\,E[g(X,Y)\,|\,Y]\,] = E[g(X,Y)]$$

proof: $h(y) = E[g(x,y)|y]$, $E[h(Y)] = \int h(y) f_Y(y)dy = \int E[g(X,y)|y]\, f_Y(y)\, dy$
$= \iint g(x,y)\, f_{X,Y}(x,y)\, dx\, dy = E[g(X,Y)]$

Multinomial distribution: $(N_1, \dots, N_k) \sim \text{Multinomial}(n, p_1, \dots, p_k)$ where $\sum p_i = 1$

$$P(N_1 = n_1, \dots, N_k = n_k) = \underbrace{\binom{n}{n_1 \dots n_k}}_{:= \frac{n!}{n_1! \dots n_k!}} p_1^{n_1} \cdot \dots \cdot p_k^{n_k} \qquad (n = n_1 + \dots + n_k)$$

(multinomial coefficient)

Poisson distribution: $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$, $X \in \{0,1,2,\dots\}$

$$P(X=x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

- For $X \sim \text{Poi}(\lambda_1)$, $Y \sim \text{Poi}(\lambda_2)$ independent: $Z = X+Y \sim \text{Poi}(\lambda_1 + \lambda_2)$
  proof: follows with binomial theorem $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$
- For $X_i \sim \text{Poisson}(\lambda_i)$, $X_1, \dots, X_n$ independent, $Z = \sum_{i=1}^{n} X_i$. Then
  $(X_1, \dots, X_n)\,|\,Z=z \sim \text{Multinomial}(z, p_1, \dots, p_k)$, $p_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$
  proof: compute explicitly $P(X_1, \dots, X_n | z) = \frac{P(X_1, \dots, X_n)}{P(z)}$

max(X_1, X_2)-distribution: Let X_1, X_2 i.i.d, $Z = \max\{X_1, X_2\}$.
The density of Z is

$$f_Z(z) = 2\, F_X(z) \cdot f_X(z)$$

(F_X: cumul. distr., f_X: distr.)

proof: $F_Z(z) = P(Z \le z) = P(\max\{X_1, X_2\} \le z)$
$= (F_X(z))^2$ then take derivative.

Sufficiency

We write X for $(X_1, \dots, X_n)$ copies of a random variable with distribution $P \in \{P_\theta \mid \theta \in \Theta\}$.

A statistic S is <u>sufficient</u> for $\theta \in \Theta$, if for all θ, and all possible s the cond. distribution

$$P_\theta(X = x \mid S = s)$$

does not depend on θ.

example: Bernoulli distribution, $\theta \in (0,1)$, $P_\theta(X_i = 1) = 1 - P(X_i = 0) = \theta$
Take $S = \sum_{i=1}^n X_i$. Then S is sufficient for θ, for all possible s:

$$P_\theta(X_1 = x_1, \dots, X_n = x_n \mid S = s) = \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(S = s)} = \frac{\theta^{x_1}(1-\theta)^{x_1} \cdot \dots \cdot \theta^{x_n}(1-\theta)^{x_n}}{\binom{n}{s} \theta^s (1-\theta)^{n-s}} = \frac{1}{\binom{n}{s}}$$

example: Poisson distribution, $\theta \in (0, \infty)$, $S = \sum_{i=1}^n X_i$ is sufficient

$$P_\theta(X_1 = x_1, \dots, X_n = x_n \mid S = s) = \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(S = s)} = \dots = \binom{s}{x_1 \dots x_n} \left(\frac{1}{n}\right)^s \quad \text{(multinomial)}$$

example: exponential distribution, $X_1, X_2 \sim \exp(\theta)$ independent
then $f_{X_1}(x;\theta) = \theta e^{-\theta x}$, $x > 0$. $S = X_1 + X_2$ is sufficient.
Note $f_S(s;\theta) = s\theta^2 e^{-\theta s}$, $s > 0$ ($\to$ Gamma(2, θ) distribution)
Thus $f_{X_1, X_2}(X_1, X_2 \mid S = s) = \frac{1}{s}$, so independent of θ.

example: Order statistics, $(X_1, \dots, X_n)$ i.i.d. from continuous distribution F.
$S = (X_{(1)}, \dots, X_{(n)})$ is sufficient for F. Because

$$P_\theta((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid (X_{(1)}, \dots, X_{(n)}) = s) = \frac{1}{n!}$$

<u>Factorization Theorem of Neyman</u>: A statistic S is sufficient iff. we can write

$$P_\theta(x) = g_\theta(S(x)) \cdot h(x) \quad \forall x, \theta$$

(density; $\to g_\theta$ is allowed to depend on θ)

with $g_\theta(\cdot), h(\cdot) \geq 0$ are arbitrary functions.

proof (discrete case): Suppose X only takes discrete values $\{a_1, a_2, \dots\}$ $\forall \theta$.
Let Q_θ be the distribution of S, thus $Q_\theta(s) := \sum_{\substack{j \\ \text{where} \\ S(a_j) = s}} P_\theta(X = a_j)$

Then the conditional distribution is

$$P_\theta(X = x \mid S = s) = \frac{P_\theta(X = x)}{Q_\theta(s)}, \quad S(x) = s$$

"$\Rightarrow$": Assume S is sufficient. Then the above term depends only on x, say $h(x)$, by definition of sufficiency. Thus
$P_\theta(X = x) = h(x) \cdot Q_\theta(s)$, where $g_\theta(s) = Q_\theta(s)$, $s = S(x)$.

"$\Leftarrow$": Insert $P_\theta(x) = g_\theta(S(x)) \cdot h(x)$ into the definition of $Q_\theta(s)$.

$$Q_\theta(s) = g_\theta(s) \sum_{S(a_j) = s} h(a_j)$$

Then the conditional distribution is given by

$$P_\theta(X = x \mid S = s) = \frac{h(x)}{\sum_{S(a_j) = s} h(a_j)} \cdot \frac{g_\theta(s)}{g_\theta(s)}$$

which does not depend on θ. □

↓ sufficiency

example: uniform distribution with unknown endpoint, on $[0,\theta]$.

then $p_\theta(x_1,\dots,x_n) = \frac{1}{\theta^n} \mathbb{1}_{\{0 \le \min\{x_1,\dots,x_n\} \le \max\{x_1,\dots,x_n\} \le \theta\}}$

$= g_\theta(S(x_1,\dots,x_n)) \cdot h(x_1,\dots,x_n)$

with $g_\theta(s) = \frac{1}{\theta^n} \mathbb{1}_{\{s \le \theta\}}$ and $h(x_1,\dots,x_n) = \mathbb{1}_{\{0 \le \min\{x_1,\dots,x_n\}\}}$

Thus $S = \max\{x_1,\dots,x_n\}$ is sufficient.

Exponential Families

$\{P_\theta \mid \theta \in \Theta\}$

A distribution P_θ belongs to the k-dimensional exponential family, if all P_θ can be written in the form

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

for any functions T_j, c_j, d and h. Note that $\{T_j\}, \{c_j\}$ are not unique.

Note: The k-dim. statistic $S(X) = (T_1(X), \dots, T_k(X))$ is sufficient for θ.

Note: If each $x_1,\dots,x_n$ is from a k-dim. exp. family, then so is $\vec{X} = (x_1,\dots,x_n)$

$\tilde{p}_\theta(\vec{x}) = \prod_{i=1}^n p_\theta(x_i)$

Canonical Form: $\{P_\theta \mid \theta \in \Theta\}$ is an exponential family in canonical form, if

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

Note: $d(\theta)$ is the normalizing constant, with

$$d(\theta) = \log\left(\int \exp\left(\sum_{j=1}^k \theta_j T_j(x)\right) h(x)\, d\nu(x)\right)$$

$T(x) = \begin{pmatrix} T_1(x) \\ \vdots \\ T_k(x) \end{pmatrix}$

Note: We write $\mathrm{Cov}_\theta(T(x)) := E_\theta[T(x)\,T^T(x)] - E_\theta[T(x)]\, E_\theta[T^T(x)] \in \mathbb{R}^{k\times k}$

Lemma 4.5.1.: It holds that

$\to$ derivative w.r.t. θ

$$E_\theta[T(x)] = \dot{d}(\theta) \quad \text{and} \quad \mathrm{Cov}_\theta(T(x)) = \ddot{d}(\theta)$$

proof:

$$\dot{d}(\theta) = \frac{\partial}{\partial\theta} d(\theta) = \frac{\partial}{\partial\theta} \log\left(\int \exp(\theta^T T(x))\, h(x)\, d\nu(x)\right)$$

$$= \int \exp(\theta^T T(x))\, T(x)\, h(x)\, d\nu(x) \cdot \frac{1}{\int \exp(\theta^T T(x)) h(x) d\nu(x)} \quad \left(= \exp(d(\theta))\right)$$

$\frac{\partial}{\partial\theta}$

$$= \int \exp\left(\theta^T T(x) - d(\theta)\right) T(x)\, h(x)\, d\nu(x)$$

$$= \int p_\theta(x)\, T(x)\, d\nu(x) = E_\theta[T(x)]$$

$$\ddot{d}(\theta) = \frac{\int \exp(\theta^T T(x))\, T(x) T^T(x)\, h(x)\, d\nu(x)}{\int \exp(\theta^T T(x))\, h(x)\, d\nu(x)} - \frac{\left(\int \exp(\theta^T T) T h\, d\nu(x)\right) \cdot \left(\int \exp[\theta^T T] T h\, d\nu\right)^T}{\left(\int \exp(\theta^T T)\, h\, d\nu\right)^2}$$

$$= \int \exp(\theta^T T - d(\theta))\, T T^T h\, d\nu - \left(\int \exp(\theta^T T - d(\theta)) T h\, d\nu\right) \cdot \left(\int \exp(\theta^T T - d(\theta)) T^T h\, d\nu\right)$$

$$= \int T T^T p_\theta\, d\nu - \left(\int p_\theta T\, d\nu\right) \cdot \left(\int p_\theta T^T d\nu\right)$$

$$= E_\theta[T T^T] - E_\theta[T]\, E_\theta[T^T]$$

$$= \mathrm{Cov}_\theta(T(X))$$

□

possibly not relevant for exam

Intermezzo: Score function and Fisher information

The score function is given by $S_\theta(x) := \frac{d}{d\theta} \log(p_\theta(x))$

The Fisher information for estimating θ:

$$I(\theta) := \mathrm{var}_\theta(S_\theta(X))$$

Further, we have that $E_\theta[S_\theta(X)] = 0$ and $I(\theta) = -E_\theta[\dot{S}_\theta(X)]$ (derivative wrt. θ)

For a 1-dim. exponential family we have that

$$S_\theta(x) = \dot{c}(\theta)\,T(x) - \dot{d}(x) \quad \text{and} \quad E_\theta[T(X)] = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}$$

Combining the above equations, we can receive

$$I(\theta) = \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta) \qquad (\text{since } \mathrm{var}_\theta(S_\theta(X)) = -E_\theta[\dot{S}_\theta(X)])$$

Bias: $X = (X_1, \dots, X_n)$ observations from a distribution $p \in \{P_\theta \mid \theta \in \Theta\}$. $(\theta \in \mathbb{R}^p)$

Let $\gamma := g(\theta) \in \mathbb{R}$ be the parameter of interest.

Let $T(X)$ be an estimator of γ.

The bias of T is given by

$$\mathrm{bias}_\theta(T) := E_\theta[T] - g(\theta)$$

we call the estimator T unbiased, if

$$\mathrm{bias}_\theta(T) = 0 \quad \forall \theta \in \Theta.$$

unbiasedness means that there is no systematic error $E_\theta[T] = g(\theta)$.

example: $X \sim \mathrm{Binomial}(n, \theta)$, $\theta \in (0,1)$. Then

$$E_\theta[T(X)] = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} T(k) =: q(\theta)$$

Since $q(\theta)$ is a polynomial of at most degree n, only parameters $g(\theta)$ of such form can be estimated unbiasedly. (NOT: $\sqrt{\theta}$, $\frac{1}{\theta}$, ...)

example: $X \sim \mathrm{Poisson}(\theta)$. $E_\theta[T(X)] = \sum_{k=0}^{\infty} e^{-\theta} \frac{\theta^k}{k!} T(k) =: e^{-\theta} p(\theta)$

$p(\theta)$ is a power series in θ. Unbiased estimators have to be of such form.

example: $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Then

$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

Mean Square Error: The MSE of an estimator T is

$$\mathrm{MSE}_\theta(T) := E_\theta[(T(X) - g(\theta))^2]$$

We can do the following decomposition: "bias-variance decomposition"

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta^2(T) + \mathrm{var}_\theta(T)$$

proof:

$$\mathrm{MSE}_\theta(T) = E_\theta[(T - g(\theta))^2] = E_\theta[(T - g(\theta) + \underbrace{E_\theta[T]}_{\text{const.}} - E_\theta[T])^2]$$

$$= \underbrace{E_\theta[(T - E_\theta[T])^2]}_{\mathrm{var}_\theta(T)} + \underbrace{(E_\theta[T] - g(\theta))^2}_{\mathrm{bias}_\theta^2(T)} + 2(E_\theta[T] - g(\theta)) \underbrace{E_\theta[T - E_\theta[T]]}_{=0}$$

□

Uniform Minimum Variance Unbiased (UMVU):
We call an unbiased estimator T^* UMVU, if for any other unbiased estimator T it holds

$$\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T) \quad \forall \theta$$

not super relevant

For any T unbiased and S sufficient, we have that
$T^* := E[T|S]$ is an unbiased estimator ($E_\theta[T^*] = g(\theta)$)

Complete: We call a sufficient statistic S complete if the following implication holds:

$$E_\theta[h(S)] = 0 \ \forall \theta \Rightarrow h(S) = 0 \quad P_\theta\text{-a.s. } \forall \theta$$

with h any function, not depending on θ.

(Lemma 5.4.1) **Completeness for exponential families:** remember that for $\theta \in \Theta \subseteq \mathbb{R}^k$

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

Consider the set $\mathcal{C} := \{(c_1(\theta), \dots, c_k(\theta)) \mid \theta \in \Theta\} \subseteq \mathbb{R}^k$.
If $\mathcal{C}$ contains an open ball in $\mathbb{R}^k$ (i.e. is truly k-dim.), then the statistik $S := (T_1(X), \dots, T_k(X))$ is complete.

Lehmann-Scheffé Lemma: Let T be an unbiased estimator of $g(\theta)$, with finite variance for all θ.
If further another statistic S is sufficient and complete, then

$$T^* := E(T|S) \quad \text{is UMVU.}$$

proof:
(1) $T^* = T^*(S)$ is unbiased: $E_\theta[T^*] = E_\theta[E[T|S]] \overset{\text{iterated expectation}}{=} E_\theta[T] \overset{T \text{ is unbiased}}{=} g(\theta)$

(2) Show $\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T) \ \forall \theta$: equivalently $\mathrm{var}_\theta(E[T|S]) \leq \mathrm{var}_\theta(T) \ \forall \theta$.
Note that this follows from the following equality for any Y, Z random variables: $\mathrm{var}(Y) = \mathrm{var}(E[Y|Z]) + E[\mathrm{var}(Y|Z)]$

proof of equality:

$$\mathrm{var}(E[Y|Z]) = E[E[Y|Z]^2] - (E[E[Y|Z]])^2$$
$$= \underbrace{E[E[Y|Z]^2]}_{(*)} - E[Y]^2$$

and

$$E[\mathrm{var}(Y|Z)] = E[E[Y^2|Z] - E[Y|Z]^2]$$
$$= E[Y^2] - \underbrace{E[E[Y|Z]^2]}_{(*)}$$

Note that adding both these terms, $(*)$ cancels out, and we are left exactly with the variance

$$\mathrm{var}(Y) = E[Y^2] - E[Y]^2$$

Substituting $T = Y$ and $S = Z$ proves the inequality, together with the fact that $E[\mathrm{var}(Y|Z)] \geq 0$.

(3) UMVU: Let $T'(S)$ be any other unbiased estimator of $g(\theta)$.
Then $E_\theta[T^*(S) - T'(S)] = E_\theta[T^*(S)] - E_\theta[T'(S)] = g(\theta) - g(\theta) = 0 \ \forall \theta$.
But since S is complete, we have that
$h(S) = T^*(S) - T'(S) = 0 \quad P_\theta\text{-a.s. } \forall \theta \Rightarrow T^*(S) = T'(S) \quad P_\theta\text{-a.s. } \forall \theta$. □

$(\theta > 0)$

example: $X_1, \dots, X_n$ i.i.d Poisson(θ), $g(\theta) = e^{-\theta}$ (early failure probability)
$T = 1\!\!1_{\{X_1 = 0\}}$ is an unbiased estimator, and $S = \sum_{i=1}^n X_i$ is a sufficient statistic.

We now check for completeness of S. Thus let h be arbitrary

$$E_\theta[h(S)] = \sum_{k=0}^{\infty} e^{-n\theta} \frac{(n\theta)^k}{k!} h(k) = 0 \quad \forall \theta.$$

It follows that $\sum_{k=0}^{\infty} \frac{(n\theta)^k}{k!} h(k) = 0 \ \forall \theta$. By Taylor expansion at zero:

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} f^{(k)}(0)$$

For this sum to be zero for all x (here of the form $n\theta \ \forall \theta$) all $f^{(k)}(0) \equiv h(k)$ need to be equal to zero $\forall k$. Since S can only take values in $\mathbb{N}$, it follows that $h = 0$ P_θ-a.s. and S is complete.

Now, by Lehmann-Scheffé, $T^* = E[T|S]$ is UMVU. Compute T^*

$$T^* = E[T|S] = 1 \cdot P(X_1 = 0 | S) + 0 \cdot P(X_1 = 1 | S)$$

$$\overset{\text{Bayes}}{=} \frac{P(S = s | X_1 = 0) \cdot P(X_1 = 0)}{P(S = s)} \qquad \text{(note: } S \text{ is poisson distributed)}$$

$$= \left(\frac{n-1}{n}\right)^s$$

example: $X_1, \dots, X_n$ i.i.d. Uniform($[0, \theta]$), $g(\theta) = \theta$. We know (previous example) that $S = \max\{X_1, \dots, X_n\}$ is sufficient. We now show that it is complete.

First find the probability density of S, $F_S(s) = P_\theta(\max\{X_1, \dots, X_n\} \le s) = \left(\frac{s}{\theta}\right)^n$ (cumul. distribution)

take the derivative to receive $f_S(s) = n s^{n-1} / \theta^n \qquad 0 \le s \le \theta$

Thus $E_\theta[h(S)] = \int_0^\theta h(s) \frac{n s^{n-1}}{\theta^n} ds = 0 \quad \forall \theta$

so $\int_0^\theta h(s) s^{n-1} ds = 0 \ \forall \theta$ differentiating wrt. to θ gives

$h(\theta) \theta^{n-1} = 0 \ \forall \theta \Rightarrow h \equiv 0$. Thus S is complet.

Now use the fact from Lehmann-Scheffé that any unbiased estimator $T(S)$ is equal to the unique UMVU. Thus it suffices to find an unbiased estimator.

For this we check if S is unbiased:

$$E_\theta[S] = \int_0^\theta s \frac{n s^{n-1}}{\theta^n} ds = \frac{n}{n+1} \theta$$

S is biased, but we see, that we can easily remove bias with a factor, to receive $T(S) = \frac{n+1}{n} S$.

Score function: If $p_\theta(x)$ is differentiable for all x, we have

$$S_\theta(x) = \frac{d}{d\theta} \log(p_\theta(x)) = \frac{\dot{p}_\theta(x)}{p_\theta(x)} \qquad \text{where } \dot{p}_\theta(x) = \frac{d}{d\theta} p_\theta(x)$$

Fisher Information / score function: $X_1, \dots, X_n$ i.i.d with density p_θ and $s_\theta = \dot{p}_\theta / p_\theta$.

The joint density is $p_\theta^{\text{joint}}(\vec{x}) = \prod_{i=1}^n p_\theta(x_i)$

The score function of n observations is $s_\theta^{\text{joint}}(\vec{x}) = \sum_{i=1}^n s_\theta(x_i)$

The Fisher information of n observations is

$$I^{\text{joint}}(\theta) = \text{var}_\theta\left(s_\theta^{\text{joint}}(\vec{X})\right) = \sum_{i=1}^n \text{var}_\theta(s_\theta(X_i)) = n\, I(\theta)$$

where $I(\theta) = \text{var}_\theta(S_\theta(X)) = E_\theta[s_\theta(X)^2] < \infty$

Cramér-Rao lower bound (CRLB): Let T be an unbiased estimator of $g(\theta)$ with finite variance. Then $g(\theta)$ is differentiable wrt. θ with

$$\dot{g}(\theta) = \frac{d}{d\theta} g(\theta) = \mathrm{cov}\left(T, S_\theta(X)\right)$$

Also

$$\mathrm{var}_\theta(T) \geq \frac{(\dot{g}(\theta))^2}{I(\theta)} \quad \forall \theta.$$

We call this the Cramér-Rao lower bound.

proof: (1) first show differentiability of $g(\theta)$. Since T is unbiased

$$\frac{g(\theta+h)-g(\theta)}{h} = \frac{E_{\theta+h}[T(X)] - E_\theta[T(X)]}{h}$$

$$= \frac{1}{h}\int T\cdot(P_{\theta+h} - P_\theta)\,d\nu = \int T\cdot \frac{P_{\theta+h}-P_\theta}{h}\,\frac{P_\theta}{P_\theta}\,d\nu$$

$$= E_\theta\left[T(X)\cdot \frac{P_{\theta+h}(X)-P_\theta(X)}{h\,P_\theta(X)} + T(X)S_\theta(X) - T(X)S_\theta(X)\right]$$

$$= E_\theta\left[T(X)\cdot\left(\frac{P_{\theta+h}(X)-P_\theta(X)}{h\,P_\theta(X)} - S_\theta(X)\right)\right] + E_\theta[T(X)S_\theta(X)]$$

$$\overset{?}{=} \underbrace{E_\theta\left[(T(X)-g(\theta))\left(\frac{P_{\theta+h}(X)-P_\theta(X)}{h\,P_\theta(X)} - S_\theta(X)\right)\right]}_{(*)} + E_\theta[T(X)S_\theta(X)]$$

as $h\to 0$ $\longrightarrow$ $0 + E_\theta[T(X)S_\theta(X)]$

because by Cauchy-Schwarz we have $|E[XY]|^2 \leq E[X^2]\cdot E[Y^2]$

thus

$$(*) \leq \underbrace{\mathrm{var}_\theta(T(X))}_{\text{const.}} \cdot E\left[\left(\underbrace{\frac{P_{\theta+h}(X)-P_\theta(X)}{h\,P_\theta(X)}}_{\to S_\theta(X)} - S_\theta(X)\right)^2\right]$$

$h\to 0 \longrightarrow 0$

(2) This gives us also the derivative of $g(\theta)$

$$\dot{g}(\theta) = E_\theta[T(X)S_\theta(X)] = \mathrm{cov}(T(X), S_\theta(X)) \qquad (\text{because } E_\theta[S_\theta(X)]=0)$$

(3) The CRLB follows from Cauchy-Schwarz

$$\dot{g}(\theta)^2 = \left(E_\theta[T(X)S_\theta(X)]\right)^2 \overset{?}{\leq} \mathrm{var}_\theta(T)\,\mathrm{var}_\theta(S_\theta(X))$$

$$= \mathrm{Var}_\theta(T)\cdot I(\theta) \qquad \square$$

generally $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$

example: $X_1,\dots,X_n$ i.i.d Exponential(θ), $\theta>0$, $P_\theta(x) = \theta e^{-\theta x}$, $g(\theta) = \frac{1}{\theta}$

$T = \bar{X}$ is unbiased with $\mathrm{var}_\theta(T) = 1/(n\theta^2)$.

We now compute the CRLB $\underbrace{\frac{\dot{g}^2(\theta)}{I^{\text{joint}}(\theta)}}_{\text{single observation}} = \frac{\dot{g}^2(\theta)}{n\,I(\theta)}$

where $\dot{g}(\theta) = -\frac{1}{\theta^2}$, and

single observation: $I(\theta) = \mathrm{var}_\theta(S_\theta(X)) = \mathrm{var}\left(\frac{d}{d\theta}\log(P_\theta(x))\right) = \mathrm{var}\left(\frac{d}{d\theta}(\log(\theta)-\theta x)\right)$

$= \mathrm{var}\left(\frac{1}{\theta} - X\right) = \mathrm{var}(X) = \frac{1}{\theta^2}$

Thus $\mathrm{CRLB} = \frac{1}{n\theta^2}$

↓CRLB

example: $X_1,\dots,X_n$ i.i.d. Poisson(θ), $\theta>0$, then $g(\theta)=e^{-\theta}$
$\log(p_\theta(x)) = -\theta + x\log(\theta) - \log(x!)$ and $S_\theta(x) = -1 + \frac{x}{\theta}$ $\left(= \frac{d}{d\theta}\log(p_\theta(x))\right)$

and $I(\theta) = \mathrm{var}_\theta\left(\frac{X}{\theta}\right) = \frac{\mathrm{var}(X)}{\theta} = \frac{1}{\theta}$

$\rightarrow \mathrm{CRLB} = \frac{\theta e^{-2\theta}}{n}$

Note: The UMVU estimator of $g(\theta)$ is $T := \left(1-\frac{1}{n}\right)^{\sum_{i=1}^n X_i}$.
It does not reach the CRLB, but the gap for large n is small.

CRLB and exponential families: CRLB can only be reached within exponential families.
Let T be unbiased for $g(\theta)$ and reaching the CRLB. Then $\{P_\theta \mid \theta\in\Theta\}$ forms a 1-dim. exp. family and $\exists c,d,h$ s.t. $\forall\theta$

$$p_\theta(x) = \exp\left(c(\theta)T(x) - d(\theta)\right)\cdot h(x) \quad \forall x$$

Also, c,d are diffable with $g(\theta) = \frac{\dot d(\theta)}{\dot c(\theta)}$ $\forall\theta$.

NOT SURE (proof: relevant for exam?)

Tests and Confidence Intervals

Quantile functions: Let F be a cumulative distr. function on $\mathbb{R}$.
Then F is cadlag (right-continuous, and limit from the left exists)
We define the quantile functions

$$q^F_{\sup}(u) := \sup\{x \mid F(x)\le u\}$$

$$q^F_{\inf}(u) := \inf\{x \mid F(x)\ge u\} := F^{-1}(u)$$ (might not necessarily exist)

It further holds that

$$F(q^F_{\inf}(u)) \ge u \quad \text{and} \quad F(q^F_{\sup}(u)-h) \le u \quad \forall h>0.$$

(non-randomized) tests: Consider the model class $\{P_\theta \mid \theta\in\Theta\}$ and $g:\Theta\to\Gamma$, $\theta\mapsto\gamma$ the parameter of interest.
Let $\gamma_0\in\Gamma$ and $\alpha\in[0,1]$ be given a non-randomized test at level α for the hypothesis $H_0: \gamma=\gamma_0$ is a statistic

$$\phi(X,\gamma_0)\in\{0,1\}$$

such that $P_\theta(\phi(X,\gamma_0)=1) \le \alpha \quad \forall\theta\in\{\vartheta \mid g(\vartheta)=\gamma_0\}$
(i.e. prob that hypothesis rejects, $\phi=1$, iff. but γ_0 is actually true, is $\le\alpha$)

Note: A test is typically based on a statistic T, of the form

$$\phi(X) = \begin{cases} 0 & \text{if } T(X)>c \\ 1 & \text{else}\end{cases}$$

c: "critical value".

Pivot: To test $H_{\gamma_0}: \gamma=\gamma_0$, we look for a Pivot. This is a function $Z(X,\gamma)$, such that $\forall\theta\in\Theta$

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

does not depend on θ.
A such pivot does not always exist.

↓

↓Pivot

However if a such Pivot $z(X, \gamma)$, with comulative distribution G, exists, then the test ↙ test statistic

$$\phi(X, \gamma_0) := \begin{cases} 1 & \text{if } z(X, \gamma_0) \notin [q_L, q_R] \\ 0 & \text{else} \end{cases}$$

where $q_L := q^G_{sup}(\frac{\alpha}{2})$ and $q_R := q^G_{inf}(1-\frac{\alpha}{2})$ [Quantiles]

has level α for testing H_{γ_0} with $\gamma_0 = g(\theta_0)$, because

$$P_{\theta_0}(\phi(X, g(\theta_0)) = 1) = P_{\theta_0}(z(X, g(\theta_0)) > q_R) + P_{\theta_0}(z(X, g(\theta_0)) < q_L)$$
$$= 1 - G(q_R) + G(q_L^-) \leq 1 - (1-\tfrac{\alpha}{2}) + \tfrac{\alpha}{2} = \alpha.$$

<u>example:</u> (location model) $X_1, \dots, X_n$ i.i.d. copies of $X = \mu + \varepsilon$, $\varepsilon \sim F_0$.
$H_0: \mu = \mu_0$. Let $\hat{\mu}$ be an equivariant estimator i.e. $\hat{\mu} - \mu$ distr. does not depend on μ)

Case 1: F_0 is known.
Take $z(X, \mu_0) = \hat{\mu} - \mu$ as pivot. By def. of equivariance, its distribution does not depend on θ.

Case 2: F_0 is of the form $\Phi(\frac{\cdot}{\sigma})$ for any $\sigma > 0$.

$$z(X, \mu) := \frac{\sqrt{n}(\bar{X} - \mu)}{S_n} \quad \text{with } S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

Then G is the Student distribution with $n-1$ degrees of freedom.
If σ is known, then G is normally distributed

Case 3: F_0 is symmetric, and continuous at $x = 0$.

$$z(X, \mu) := \sum_{i=1}^n \mathbb{1}_{\{X_i \geq \mu\}}$$

Then G is the Binomial(n, p) distribution with $p = \frac{1}{2}$.

<u>Confidence set/interval:</u> We call $I = I(X) \subseteq \Gamma$ a confidence set for γ at level $1-\alpha$ if

$$P_\theta(\gamma \in I) \geq 1 - \alpha \qquad \forall \theta \in \Theta$$

We call I a confidence interval, if it is of the form $I := [\underline{\gamma}, \bar{\gamma}]$, where the boundaries depend only on X.

Let for any $\gamma_0 \in \mathbb{R}$, $\phi(X, \gamma_0) \in \{0, 1\}$ be a test at level α for the hypothesis $H_{\gamma_0}: \gamma = \gamma_0$.
This means we reject H_{γ_0} iff $\phi(X, \gamma_0) = 1$. It further holds that $P_{\theta: \gamma = \gamma_0}(\phi(X, \gamma_0) = 1) \leq \alpha$.
Then

$$I(X) := \{\gamma \mid \phi(X, \gamma) = 0\}$$

is a $(1-\alpha)$-confidence set for γ.

Conversely, if $I(X)$ is a $(1-\alpha)$-confidence set for γ, then for any γ_0, the test

$$\phi(X, \gamma_0) = \begin{cases} 1 & \text{if } \gamma_0 \notin I(X) \\ 0 & \text{else} \end{cases}$$

is a test at level α of H_{γ_0}.

Level and power of a test:

	H_0 is true	H_0 is false
Test rejects H_0	α (level, Type I error)	$1-\beta$ (power)
Test doesn't reject H_0	$1-\alpha$	β (Type II error)

- p-value: (assume H_0 is true) the probability of obtaining a result equal, or more extreme than what was actually observed.
  (e.g. if p is large, then it was very likely to make this observation, given H_0 is true).

Two-sample problem: We assume our data consists of two independent samples $X=(X_1,\dots,X_n)$ and $Y=(Y_1,\dots,Y_m)$ (each i.i.d) sampled from (possibly unknown) distributions F_X and F_Y.
The testing problem is: $H_0: F_X = F_Y$ against a one-/two-sided alternative. (sometimes only one side of the distribution is relevant to us)

Two-sample student test: We assume that both F_X and F_Y are normally distributed with the same variance (but possibly different mean). $F_X = \mathcal{N}(\mu, \sigma^2)$, $F_Y = \mathcal{N}(\mu+\gamma, \sigma^2)$ where $\gamma \in \Gamma$.
two-sided: $\Gamma = \mathbb{R}$, one-sided: $\Gamma = (-\infty, 0]$

The testing problem is then $H_0: \gamma = 0$.
For constructing a pivot $Z(X, Y, \gamma)$ we consider the sample means $\bar{X}$ (expectation μ and variance σ^2/n) and $\bar{Y}$ (expec. μ, var. σ^2/m).
Then the quantity $\bar{X}-\bar{Y}$ is $\mathcal{N}(\gamma, \sigma^2(n+m)/nm)$ distributed.
But this distribution still depends on unknown parameters.

$$\sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{\sigma}\right) \text{ is } \mathcal{N}(0,1) \text{ distributed.}$$

To arrive at our pivot, we plug in our estimate for σ, given by

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

we receive

$$Z(X, Y, \gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{S}\right)$$

The distribution G is Student $(n+m-2)$.
The test statistic for $H_0: \gamma = 0$ is thus given by

$$T = Z(X, Y, 0)$$

and we receive the one-sided test at level α (against $H_1: \gamma > 0$)

$$\Phi(X, Y) = \begin{cases} 1 & \text{if } T < -t_{n+m-2}(1-\alpha) \\ 0 & \text{if } T \geq -t_{n+m-2}(1-\alpha) \end{cases}$$

→ $(1-\alpha)$-quantile of Student $(n+m-2)$

Two-sample Wilcoxon's test: Assume F_X and F_Y are continuous.
Let $Z = (X_1,\dots,X_n, Y_1,\dots,Y_m)$, $N = n+m$. Define $R_i = \text{rank}(Z_i)$ $i=1,\dots,N$ → index; sorted ascendingly
The Wilcoxon test statistic is

$$Z(X,Y) = T = \sum_{i=1}^n R_i \quad \text{(i.e. sum ranks of all } X_1,\dots,X_n \text{ among } Z\text{)}$$
$$= \#\{Y_j < X_i\} + \frac{n(n+1)}{2}$$

Large values of T indicate that X is typically larger than Y.

Under $H_0: F_X = F_Y$ the vector of ranks $(R_1,\dots,R_n)$ has the same distribution as n random draws without replacement from $\{1,\dots,N\}$.

$$E_{H_0}[T] = \frac{n(N+1)}{2}$$

Neyman Pearson and UMP: Let $\{P_\theta \mid \theta \in \Theta\}$ be a family of probability measures. Let Θ_1, Θ_0 be a partition of Θ.
We consider the general testing problem
$H_0: \theta \in \Theta_0$
against
$H_1: \theta \in \Theta_1$

Neyman Pearson test: Consider the case $\Theta = \{\theta_0, \theta_1\}$, and thus the test $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$.
First, define the **Risk** of a test ϕ

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

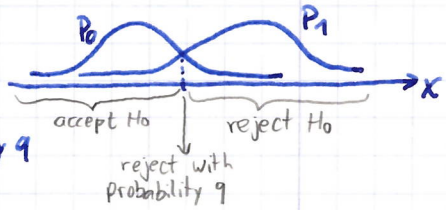
→ probability of rejecting when H_0 is true (Error of first kind)
→ probability of accepting when H_0 is not true (Error of second kind)

Generally we want the risk to be small.
(For a non-randomized test: $R(\theta, \phi) = \begin{cases} P_{\theta_0}(\phi(X)=1) & \theta=\theta_0 \\ 1-P_{\theta_1}(\phi(X)=1) = P_{\theta_1}(\phi(X)=0) & \theta=\theta_1 \end{cases}$)

A Neyman Pearson Test is given by

$$\phi_{NP}(x) := \begin{cases} 1 & \frac{P_1(x)}{P_0(x)} > c \\ q & \frac{P_1(x)}{P_0(x)} = c \\ 0 & \frac{P_1(x)}{P_0(x)} < c \end{cases}$$

→ in this case we reject with probability q



For some constants $q \in [0,1]$, $c \in [0, \infty)$.

Neyman Pearson Lemma: Take same hypotheses as above. For any test ϕ it holds true that

$$R(\theta_1, \phi) - R(\theta_1, \phi_{NP}) \geq c \cdot [\underbrace{R(\theta_0, \phi_{NP})}_{=\text{level of the test}} - R(\theta_0, \phi)]$$

Note that if ϕ_{NP} and ϕ have the same level, the right hand side is zero. Then the Risk (in case of θ_1) of ϕ_{NP} is strictly smaller than the Risk of ϕ.

proof:

$$R(\theta_1, \phi) - R(\theta_1, \phi_{NP}) = E_{\theta_1}[\phi_{NP}(x)] - E_{\theta_1}[\phi(x)]$$

$$= \int (\phi_{NP} - \phi)\, p_1$$

$$= \int_{p_1/p_0 > c} \underbrace{(\phi_{NP} - \phi)}_{=1-\phi \geq 0} \underbrace{p_1}_{> c p_0} + \int_{\frac{p_1}{p_0} = c} (\phi_{NP} - \phi) \underbrace{p_1}_{= c p_0} + \int_{\frac{p_1}{p_0} < c} \underbrace{(\phi_{NP} - \phi)}_{-\phi \leq 0} \underbrace{p_1}_{< c p_0}$$

$$\geq \int_{p_1/p_0 > c} (\phi_{NP} - \phi) \cdot c\, p_0 + \int_{\frac{p_1}{p_0} = c} (\phi_{NP} - \phi)\, c\, p_0 + \int_{\frac{p_1}{p_0} < c} (\phi_{NP} - \phi)\, c\, p_0$$

$$= c \int (\phi_{NP} - \phi)\, p_0 = c \cdot [E_{\theta_0}[\phi_{NP}(X)] - E_{\theta_0}[\phi(X)]]$$

$$= c \cdot [R(\theta_0, \phi_{NP}) - R(\theta_0, \phi)] \qquad \square$$

Level of a (randomized) test: Let $\alpha \in [0,1]$, $\phi: \mathcal{X} \to [0,1]$. We say ϕ is a test at level α if

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \leq \alpha$$

→ probability of rejecting, when $\theta \in \Theta_0$ (corresponds to this)

special case: non-randomized $E_\theta[\phi(x)] = P_\theta(\phi(X)=1)$

Uniformly most powerful: We call a test ϕ UMP, if

- it has level α, and
- for all tests ϕ' with level α we have $E_\theta[\phi'(X)] \le E_\theta[\phi(X)]$ $\forall \theta \in \Theta_1$
  (i.e. ϕ rejects H_0 (correctly) more often than ϕ')

One-sided UMP test and exponential families: Let $\Theta \subseteq \mathbb{R}$ be an interval.
Regard the testing problem $H_0: \theta \le \theta_0$ against $H_1: \theta > \theta_0$.
Let $\{P_\theta \mid \theta \in \Theta\}$ be a 1-dim. exponential family

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

further assume that $c(\theta)$ is a strictly increasing function of θ.
then

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

is UMP.
q and t_0 are chosen s.t. $E_{\theta_0}(\phi(T)) = \alpha$.

($\theta_0 < \theta_1$)

proof: Recall the NP-test for $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, where

$$\phi_{NP}(x) = \begin{cases} 1 & P_1(x)/P_0(x) > c_0 \\ q_0 & P_1(x)/P_0(x) = c_0 \\ 0 & P_1(x)/P_0(x) < c_0 \end{cases}$$

with q_0 and c_0 chosen s.t. $E_{\theta_0}[\phi_{NP}(X)] = \alpha$.
Note that then

$$\frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} = \exp\Big[\underbrace{(c(\theta_1) - c(\theta_0))}_{\ge 0 \text{ because } \theta_1 > \theta_0 \text{, and } c \text{ increasing.}}\,T(x) - (d(\theta_1) - d(\theta_0))\Big]$$

is increasing in $T(x)$. Thus $\left(\frac{P_1(x)}{P_0(x)} \gtrless c \iff T(x) \gtrless t \text{ for some } t \text{ and } c\right)$

Thus ϕ is equivalent to ϕ_{NP} and so most powerful for $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$.
Since above we chose θ_1 ($> \theta_0$) arbitrarily, the above statement holds true for all $\theta_1 > \theta_0$. Thus ϕ is UMP for $H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$.
Hence for the testing problem $H_0: \theta \le \theta_0$ against $H_1: \theta > \theta_0$ ϕ fulfills the second condition for UMP.
It is left to show that ϕ has level α, i.e. $\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \le \alpha$.

For this we define $\beta(\theta) = E_\theta[\phi(T)]$ and use the fact that β is increasing in θ together with $\beta(\theta_0) = \alpha$ (because $\phi = \phi_{NP}$). → see above
Then $\sup_{\theta \le \theta_0} \beta(\theta) = \beta(\theta_0) = E_{\theta_0}[\phi(X)] = \alpha$
and it follows that ϕ also has level α (for the relevant testing problem).
Thus it is UMP.
Left to prove: $\beta(\theta)$ is increasing. Let $\theta, \vartheta \in \Theta$, $\vartheta > \theta$.
We know that

$$\frac{\bar P_\vartheta(t)}{\bar P_\theta(t)} = \exp\Big[\underbrace{(c(\vartheta) - c(\theta))}_{\ge 0 \text{ is increasing in } t.} \cdot t - (d(\vartheta) - d(\theta))\Big]$$

→ new probability measure wrt. T, thus different $\bar P_\theta$.

Since further $1 = \int \bar P_\vartheta(t)\,\bar\nu(t) = \int \bar P_\theta(t)\,\bar\nu(t)$, there must be an s_0, where $\bar P_\vartheta(s_0) = \bar P_\theta(s_0)$ the densities cross. (and $\bar P_\vartheta(t) \le \bar P_\theta(t)$ for $t \le s_0$)
↳ and the other way around

$$\begin{aligned}\beta(\vartheta) - \beta(\theta) &= \int \phi(t)\,(\bar P_\vartheta(t) - \bar P_\theta(t))\,d\bar\nu(t) \\ &= \int_{t \le s_0} \underbrace{\phi(t)}_{\le \phi(s_0)}\,\underbrace{(\bar P_\vartheta(t) - \bar P_\theta(t))}_{\le 0}\,d\bar\nu(t) + \int_{t \ge s_0} \underbrace{\phi(t)}_{\ge \phi(s_0)}\,\underbrace{(\bar P_\vartheta(t) - \bar P_\theta(t))}_{\ge 0}\,d\bar\nu(t) \\ &\ge \phi(s_0) \int \bar P_\vartheta(t) - \bar P_\theta(t)\,d\bar\nu(t) = 0\end{aligned}$$

(note: ϕ is increasing in t)

□

UMP

example: $X_1, \dots, X_n$ i.i.d. $Exp(\theta)$, $\theta > 0$, $H_0: \theta \le \theta_0$, $H_1: \theta > \theta_0$.

$$p_\theta(x) = \theta e^{-\theta x} = \exp[-\theta x - (-\log(\theta))] \cdot 1_{\{x>0\}}(x)$$

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

$$\Rightarrow \phi(T) = \begin{cases} 1 & \text{if } T \le t_0 \\ 0 & \text{if } T > t_0 \end{cases}$$ → no randomization because our distribution is continuous.

Now choose t_0 s.t. $E_{\theta_0}[\phi(T)] = \alpha$ (level of α)

i.e. $\alpha = P_{\theta_0}(T \le t_0) = P(\underbrace{\theta_0 T}_{\text{now has mean 1}} \le \theta_0 t_0) = G(\theta_0 t_0) = \alpha$

where $G = Gamma(n,1)$ (i.e. sum of $Exp(1)$).
Thus $\theta_0 t_0 = G^{-1}(\alpha) \Rightarrow t_0 = G^{-1}(\alpha)/\theta_0$.
Then ϕ is UMP.

Right-/Left-/two-sided alternatives:

Right sided alternative: $H_0: \theta \le \theta_0$, $H_1: \theta > \theta_0$. The UMP test is

$$\phi_R(T) = \begin{cases} 1 & T > t_R \\ q & T = t_R \\ 0 & T < t_R \end{cases} \qquad \beta_R(\theta) = E_\theta[\phi_R(T)] \text{ is strictly increasing in } \theta.$$

Left sided alternative: $H_0: \theta \ge \theta_0$, $H_1: \theta < \theta_0$. The UMP test is

$$\phi_L(T) = \begin{cases} 1 & T < t_L \\ q & T = t_L \\ 0 & T > t_L \end{cases} \qquad \beta_L(\theta) = E_\theta[\phi_L(T)] \text{ is strictly decreasing in } \theta.$$

Two sided alternative: $H_0: \theta = \theta_0$, $H_1: \theta \ne \theta_0$. ϕ_R is most powerful for $\theta > \theta_0$, and ϕ_L is most powerful for $\theta < \theta_0$.
Thus there exists no UMP test for this case.

Unbiased: We call a test ϕ unbiased, if $\forall \theta \in \Theta_0$, $\vartheta \in \Theta_1$

$$E_\theta[\phi(X)] \le E_\vartheta[\phi(X)]$$

Uniformly most powerful unbiased (UMPU): We call a test ϕ UMPU if

- ϕ has level α
- ϕ is unbiased
- For any ϕ' unbiased test with level α we have

$$E_\theta[\phi'(X)] \le E_\theta[\phi(X)] \quad \forall \theta \in \Theta_1$$

UMPU and exponential families: Let $\Theta \subseteq \mathbb{R}$ be an interval. Consider testing $H_0: \theta = \theta_0$ against $H_1: \theta \ne \theta_0$. We have a 1-dim. exp family $p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$, $c(\theta)$ is strictly increasing.

Then

$$\phi(T(x)) := \begin{cases} 1 & \text{if } T(x) < t_L \text{ or } T(x) > t_R \\ q_L & \text{if } T(x) = t_L \\ q_R & \text{if } T(x) = t_R \\ 0 & \text{if } t_L < T(x) < t_R \end{cases}$$

where t_L, t_R, q_L, q_R are chosen such that

$$E_{\theta_0}[\phi(X)] = \alpha \quad \text{and} \quad \frac{d}{d\theta} E_\theta[\phi(X)]\Big|_{\theta=\theta_0} = 0$$

example: mean of $\mathcal{N}(\mu, \sigma_0^2)$, $X_1, \dots, X_n$ i.i.d., $\mu \in \mathbb{R}$ unknown.
we test $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$. We want to find a UMPU test.
Since our distribution is continuous, our test is non-randomized.
Thus we have no q_R and q_L from the previous theorem.
$T = \sum_{i=1}^n X_i$ is a sufficient statistic.
for $t_L < t_R$:

$$E_\mu[\phi(T)] = 1 \cdot P_\mu(T > t_R) + 1 \cdot P_\mu(T < t_L)$$

$$= P_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} > \frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + P_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} < \frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right) \quad \rightarrow \text{std normally distributed.}$$

$$= 1 - \Phi\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + \Phi\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right) = \alpha$$

and $$\frac{d}{d\mu} E_\mu[\phi(T)] = \frac{n}{\sqrt{n}\sigma_0} \dot{\Phi}\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) - \frac{n}{\sqrt{n}\sigma_0} \dot{\Phi}\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right)\Big|_{\mu=\mu_0} = 0$$

gives $$\dot{\Phi}\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right) = \dot{\Phi}\left(\frac{t_L - n\mu_0}{\sqrt{n}\sigma_0}\right)$$

or equivalently: $(t_R - n\mu_0)^2 = (t_L - n\mu_0)^2$. We choose solution $(t_R - n\mu_0) = -(t_L - n\mu_0)$, because otherwise we get a trivial solution.
Plugging this into the first equation gives

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

Risk of an Estimator: Let $X \sim P_\theta$, $\theta \in \Theta$. $T(X)$ an estimator of a parameter of interest $g(\theta) = \mu$.
A Risk function R measures the loss due to the error of the estimator.

$$R(\theta, T) := E_\theta[L(\theta, T(X))], \text{ with } L(\cdot,\cdot) \text{ a loss-function.}$$

Rao-Blackwell Lemma: Let S be a sufficient estimator for θ.
Suppose $\mathcal{A} \subseteq \mathbb{R}^p$ is convex, and $L: \mathcal{A} \to \mathbb{R}$, $a \mapsto L(\theta, a)$ is a convex function for all θ. For any decision $d: \mathcal{X} \to \mathcal{A}$ we define $d'(s) = E[d(X) | S = s]$. Then

$$R(\theta, d') \leq R(\theta, d) \quad \forall \theta$$

Thus in case of convex loss, an estimator based on the original data X can be replaced by another estimator based only on S.

proof:
Recall Jensen inequality: $E[g(X)] \geq g(E[X])$ for g convex.
Thus

$$E[\overset{\text{convex}}{L}(\theta, d(X)) | S = s] \geq L(\theta, E[d(X) | S = s])$$

$$= L(\theta, d'(X))$$

And

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

$$\overset{\text{iterated expectation}}{=} E_\theta[E[L(\theta, d(X)) | S = s]]$$

$$\geq E_\theta[L(\theta, d'(S))] = R(\theta, d')$$

□

Location model: $\theta \in \mathbb{R}$ (location parameter). $X_i = \theta + \varepsilon_i$, $i = 1, \dots, n$
ε_i i.i.d. with distribution $P_0(\cdot)$. We want to estimate θ.

Location-scale model: $\theta = (\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$ (scale parameter).

$X_i = \mu + \sigma \varepsilon_i$, $i = 1, \dots, n$.

Location equivariant: We call a statistic $T = T(X)$ (location) equivariant if $\forall c \in \mathbb{R}$, $\underline{X} = (X_1, \dots, X_n)$ we have

$$T(X_1 + c, \dots, X_n + c) = T(X_1, \dots, X_n) + c$$

(common examples are: $\bar{X}$, $X_{(\frac{n+1}{2})}$ for n odd, ...)

Location invariant: We call a Loss function $L(\theta, a)$ (location) invariant, if $\forall c \in \mathbb{R}$

$$L(\theta + c, a + c) = L(\theta, a) \quad (\theta, a) \in \mathbb{R}^2.$$

Risk of equivariant statistic: If $T(X)$ is equivariant (and $L(\theta, a)$ invariant) then we have that (location model $X = \theta + \varepsilon$)

$$R(\theta, T) = E_\theta[L(\theta, T(X))] = E_\theta[L(0, T(X) - \theta)]$$
$$= E_\theta[L(0, T(X - \theta))] = E_\theta[L(0, T(\varepsilon))] \qquad (\varepsilon = (\varepsilon_1, \dots, \varepsilon_n))$$

Thus, the risk does not depend on θ, and we can write

$$R(\theta, T) = R(0, T).$$

Uniform minimum risk equivariant (UMRE): We call an equivariant statistic T UMRE, if

$$R(\theta, T) = \min_{d \text{ equivariant}} R(\theta, d) \quad \forall \theta \qquad \text{(minimal risk)}$$

or equivalently: $R(0, T) = \min_{d \text{ equivariant}} R(0, d)$

Construction of UMRE estimator: Let $Y_i := X_i - X_n$, $i = 1, \dots, n$.
and $\underline{X} = (X_1, \dots, X_n)$, $\underline{Y} = (Y_1, \dots, Y_n)$. Then

T is equivariant $\Leftrightarrow$ $T(\underline{X}) = T(\underline{Y}) + X_n$

proof:
"$\Rightarrow$": $T(\underline{Y}) = T(X_1 - X_n, \dots, X_n - X_n) \overset{\text{equivariance}}{=} T(\underline{X}) - X_n$.

"$\Leftarrow$": Replace $\underline{X}$ by $\underline{X} + c$ ($\underline{Y}$ is unchanged, because invariant)
Then $T(\underline{X} + c) = T(\underline{Y}) + X_n + c = T(\underline{X}) + c$. □

Moreover define

$$T^*(\underline{Y}) := \arg\min_v E[L(0, V + \varepsilon_n) \mid \underline{Y}]$$

and additionaly

$$T^*(\underline{X}) := T^*(\underline{Y}) + X_n$$

Then T^* is UMRE.

$R(0, T) = E[L_0(T(Y) + \varepsilon_n)] = E[E[L_0(T(Y) + \varepsilon_n) \mid Y]] \geq E[E[L_0(T^*(Y) + \varepsilon_n)] \mid Y] = R(0, T^*)$
(iterated expectation; $\geq$ by def.)

Invariant statistic: We call $Y: \mathbb{R}^n \to \mathbb{R}^n$ maximal invariant, if

$$Y(x) = Y(x') \iff \exists c \in \mathbb{R}: \quad x = x' + c \qquad x, x' \in \mathbb{R}^n$$

(c can depend on x and x')

More general UMRE: Let $d(X)$ be an equivariant estimator.
Let $Y = X - d(X)$, and

$$T^*(Y) := \arg\min_v E[L_0(v + d(\varepsilon)) \mid Y]$$

Then

$$T^*(X) := T^*(Y) + d(X) \quad \text{is UMRE.}$$

Basu's Lemma: Let X have distribution P_θ, $\theta \in \Theta$. Suppose T is sufficient and complete, and that the distribution of $Y = Y(X)$ does not depend on θ. Then $\forall \theta \in \Theta$ T and Y are independent under P_θ.

proof: Let A be a measurable set. $h(T) := P(Y \in A \mid T) - P(Y \in A)$
Since T is sufficient, and $Y = Y(X)$ only depends on X, $P(Y \in A \mid T)$ does not depend on θ.

Further note that since T is complete, we have that

$$E_\theta[h(T)] = E_\theta[P(Y \in A \mid T)] - P(Y \in A) = P(Y \in A) - P(Y \in A) = 0 \quad \forall \theta.$$

completeness $\Rightarrow$ $h(T) = 0$ P_θ-a.s. $\forall \theta$. $\iff P(Y \in A \mid T) = P(Y \in A)$ P_θ-a.s.

Since A was chosen arbitrarily, it follows that $T \perp Y$. □

example: $X_1, \dots, X_n$ i.i.d. $N(\theta, \sigma^2)$, σ^2 known. Then $T = \bar{X}$ is sufficient and complete. Further the distribution of $Y := X - \bar{X}$ does not depend on θ.
Thus, by Basu's Lemma, $\bar{X}$ and $X - \bar{X}$ are independent.
Hence, $\bar{X}$ is UMRE.

Note: For an equivariant (and unbiased) estimator T we have

$$T \text{ is UMRE} \iff E_0[T(X) \mid X - T(X)] = 0$$

Thus UMRE follows directly from independence, since $E_0[T(X)] = 0$

Decisions and Risk: Define $\mathcal{A}$ to be the action space.
$\mathcal{A} = \mathbb{R}$ → real-valued parameter, $\mathcal{A} = \{0, 1\}$ → hypothesis testing,
$\mathcal{A} = [0, 1]$ → randomized test, $\mathcal{A} = \{\text{intervals}\}$ → confidence intervals

A decision $d: \mathcal{X} \to \mathcal{A}$ is a mapping of observations to actions.
Depending on $\mathcal{A}$ we call d a decision, estimator, test, ...

We define the loss function as $L: \Theta \times \mathcal{A} \to \mathbb{R}$, $L(\theta, a)$ the loss.

Lastly the Risk of a decision $d(X)$ is given by

$$R(\theta, d) := E_\theta[L(\theta, d(X))], \quad \theta \in \Theta$$

example: (estimation). $g(\theta) \in \mathbb{R}$ parameter of interest. $\mathcal{A} = \mathbb{R}$.
Important loss functions are

$$L(\theta, a) := w(\theta)\,|g(\theta) - a|^r \qquad \text{where } w(\cdot) \geq 0,\ r \geq 0.$$

the Risk is then

$$R(\theta, d) = w(\theta) \cdot E_\theta[\,|g(\theta) - d(X)|^r\,]$$

For $w \equiv 1$, $r = 2$ we receive the quadratic loss and mean squared error.

example: (Tests) $H_0: \theta \in \Theta_0$, $H_1: \theta \in \Theta_1$, $\mathcal{A} = \{0,1\}$, with loss

$$L(\theta, a) := \begin{cases} 1 & \text{if } \theta \in \Theta_0 \text{ and } a = 1 \\ c & \text{if } \theta \in \Theta_1 \text{ and } a = 0 \\ 0 & \text{else} \end{cases} \qquad \text{for some } c > 0.$$

Then

$$R(\theta, d) = \begin{cases} P_\theta(d(X) = 1) & \text{if } \theta \in \Theta_0 \quad \to \text{type I error probability} \\ c\,P_\theta(d(X) = 0) & \text{if } \theta \in \Theta_1 \quad \to \text{type II error probability} \\ 0 & \text{otherwise} \end{cases}$$

Admissibility: Let d be a decision. We call the decision d' strictly better than d, if

$$R(\theta, d') \leq R(\theta, d) \qquad \forall \theta \in \Theta$$

and additionaly if

$$\exists \theta:\ R(\theta, d') < R(\theta, d).$$

(unzulässig)

We call d inadmissible, if there exists a d' strictly better than d.

(2) example: Consider the NP-test ϕ_{NP} for $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$
Then ϕ_{NP} is admissible iff one of the two cases hold
(1) its power is strictly less than 1
OR (2) it has minimal level among all tests with power 1.

Suppose: $R(\theta_1, \phi_{NP}) > 0$. Then ϕ_{NP} is admissible.
proof: Assume $R(\theta_0, \phi) \leq R(\theta_0, \phi_{NP})$.
then $R(\theta_1, \phi) = \underbrace{R(\theta_1, \phi) - R(\theta_1, \phi_{NP})}_{\geq c \cdot [R(\theta_0, \phi_{NP}) - R(\theta_0, \phi)]} + R(\theta_1, \phi_{NP}) \geq R(\theta_1, \phi_{NP})$

maybe proof in script is better (Lemma 10.22)

□

Minimaxity: We call a decision d minimax, if

$$\sup_{\theta \in \Theta} R(\theta, d) = \inf_{\substack{d' \\ \text{decision}}} \left[\sup_{\theta \in \Theta} R(\theta, d') \right]$$

i.e. the best decision in the worst possible case.

example: A Neyman-Pearson test ϕ_{NP} is minimax if and only if $R(\theta_0, \phi_{NP}) = R(\theta_1, \phi_{NP})$

don't yet understand → proof:
"⇒": Let ϕ be a test. Assume $R(\theta_0, \phi_{NP}) = R(\theta_1, \phi_{NP})$, but ϕ_{NP} is not minimax. Then for some test ϕ

$$\max_{i \in \{0,1\}} R(\theta_i, \phi) < \max_{i \in \{0,1\}} R(\theta_i, \phi_{NP})$$

Then $R(\theta_0, \phi) < R(\theta_0, \phi_{NP})$ and $R(\theta_1, \phi) < R(\theta_0, \phi_{NP})$. But this contradicts the NP-Lemma.
"⇐": $S = \{(R(\theta_0, \phi), R(\theta_1, \phi)) \mid \phi\}$ is convex. Thus for $R(\theta_0, \phi_{NP}) < R(\theta_1, \phi_{NP})$ $\exists \phi$ with $R(\theta_0, \phi_{NP}) < R(\theta_0, \phi) < R(\theta_1, \phi_{NP})$ and $R(\theta_1, \phi) < R(\theta_1, \phi_{NP})$. Then clearly ϕ_{NP} is not minimax. Similarly for $R(\theta_0, \phi_{NP}) > R(\theta_1, \phi_{NP})$ □

Bayes risk: We now think of $\theta \in \Theta$ as a random variable.
π is a probability measure on Θ with density w ("prior density")
We define the bayes risk to be $= w(\theta)\, d\mu(\theta)$

$$r(\pi, d) := \int_\Theta R(\vartheta, d)\, d\pi(\vartheta) = \sum_\Theta R(\vartheta, d) \cdot w(\vartheta)$$

where $R(\vartheta, d) = E_\vartheta[L(\vartheta, d(X))]$. ($\rightarrow$ weighted risks)

We call a decision d bayes (wrt. π), if

$$r_w(d) = r(\pi, d) = \inf_{d'} r(\pi, d') \quad (\rightarrow \text{minimal bayes risk})$$

We write $d = d_{bayes}$.

Bayes test: Consider testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$. $\mathcal{A} = [0,1]$.
Let $L(\theta_0, a) = a$, $L(\theta_1, a) = 1-a$, $w(\theta_0) = w_0$, $w(\theta_1) = w_1 = 1 - w_0$.
Then

$$r_w(\phi) = w_0 R(\theta_0, \phi) + w_1 R(\theta_1, \phi) \quad (\rightarrow \text{want to minimize})$$

The bayes test is $\phi_{Bayes}(x) = \begin{cases} 1 & \text{if } p_1/p_0 > w_0/w_1 \\ q & \text{if } p_1/p_0 = w_0/w_1 \\ 0 & \text{if } p_1/p_0 < w_0/w_1 \end{cases}$.

proof: The bayes risk is given by

$$\begin{aligned} r_w(\phi) &= w_0 \cdot E_{\theta_0}[L(\theta_0, \phi(X))] + w_1 \cdot E_{\theta_1}[L(\theta_1, \phi(X))] \\ &= w_0 E_{\theta_0}[\phi(X)] + w_1 \cdot E_{\theta_1}[1 - \phi(X)] \\ &= w_0 \int \phi\, p_0 + w_1 \cdot (1 - \int \phi\, p_1) \\ &= \int \phi \cdot \underbrace{(w_0 p_0 - w_1 p_1)}_{(*)} + w_1 \end{aligned}$$

We can choose ϕ to minimize the above term as follows

$$\phi = \begin{cases} 1 & w_0 p_0 - w_1 p_1 < 0 \rightarrow (*) \text{ is negative. Thus maximize } \phi \\ q & w_0 p_0 - w_1 p_1 = 0 \rightarrow \text{doesn't really matter (integral is zero anyways)} \\ 0 & w_0 p_0 - w_1 p_1 > 0 \rightarrow (*) \text{ is positive. Thus minimize value of } \phi \end{cases} \quad q \text{ arbitrary } (0,1)$$

□

we can also show that

$$2 \cdot r_w(\phi_{Bayes}) = 1 - \int |w_1 p_1 - w_0 p_0|.$$

Construction of Bayes estimators: We now think of p_θ as the density of X given the value of θ. i.e.

$$p_\theta(x) = p(x|\theta) \quad x \in \mathcal{X}$$

and marginal density $p(\cdot) = \int_\Theta p(\cdot|\vartheta)\, w(\vartheta)\, d\mu(\vartheta)$.

The a posteriori density of θ is

$$w(\vartheta|x) = p(x|\vartheta) \cdot \frac{w(\vartheta)}{p(x)} \quad \vartheta \in \Theta,\ x \in \mathcal{X}$$

Now, given data $X = x$, consider θ as a random variable with density $w(\vartheta|x)$. For

$$L(x, a) := E_\theta[L(\theta, a)|X = x] = \int_\Theta L(\vartheta, a)\, w(\vartheta|x)\, d\mu(\vartheta)$$

$d(x) := \arg\min_a L(x, a)$ is the bayes decision d_{bayes}.

construction of bayes estimator

proof: Let d' be any decision

$$r_\omega(d') = \int_\Theta R(\vartheta, d')\,\omega(\vartheta)\,d\mu(\vartheta)$$

$$= \int_\Theta \left[\int_{\mathcal{X}} L(\vartheta, d'(x))\,p(x|\vartheta)\,d\nu(x)\right]\omega(\vartheta)\,d\mu(\vartheta)$$

switch integrals and us. a posteriori density

$$= \int_{\mathcal{X}} \left[\int_\Theta L(\vartheta, d'(x))\,\omega(\vartheta|x)\,d\mu(\vartheta)\right] p(x)\,d\nu(x)$$

$$= \int_{\mathcal{X}} L(x, d'(x))\,p(x)\,d\nu(x)$$

$$\geq \int_{\mathcal{X}} L(x, d(x))\,p(x)\,d\nu(x) = r_\omega(d).$$

(from above)

□

example: given θ, $X \sim$ Poisson(θ), $\theta \sim$ Gamma(k,λ). Thus
$\omega(\vartheta) = \lambda^k \vartheta^{k-1} e^{-\lambda\vartheta}/\Gamma(k)$, $E[\theta] = k/\lambda$.
The a posteriori density is

$$\omega(\vartheta|x) = p(x|\vartheta)\,\frac{\omega(\vartheta)}{p(x)} = \ldots \propto e^{-\vartheta(1+\lambda)}\,\vartheta^{k+x-1} \sim \text{Gamma}(k+x, 1+\lambda)$$

With quadratic loss, the Bayes estimator is $\frac{k+x}{1+\lambda} = E[\theta|x]$

The maximum a posteriori estimator is $\frac{k+x-1}{1+\lambda}$

Extended Bayes: We call a statistic T extended Bayes, if there exists a sequence of prior densities $\{\omega_m\}_{m=1}^\infty$, such that

$$r_{\omega_m}(T) - \inf_{T'} r_{\omega_m}(T') \to 0, \text{ as } m \to \infty.$$

(Minmax) Lemma 11.1.1: Suppose T is a statistic with risk $R(\theta, T) = R(T)$ independent of θ. Then

(1) T admissible $\Rightarrow$ T minimax
(2) T Bayes $\Rightarrow$ T minimax
(3) T extended Bayes $\Rightarrow$ T minimax

proof:
(1) T is admissible. Thus $\forall T'$ there is either a θ with $R(\theta, T') > R(T)$, or $R(\theta, T') \geq R(T)$ $\forall\theta$. Hence $\sup_\theta(R(\theta, T')) \geq R(T)$.
T takes minimal risk across all decisions $\to$ it is minimax.
(2) This follows from (3), since Bayes implies extended bayes.
(3) Assume that for each prior ω_m there exists a bayes decision: $r_{\omega_m}(T_m) = \inf_{T'} r_{\omega_m}(T')$ $\forall m = 1, \ldots$

By the extended bayes property: $\forall\varepsilon > 0$ $\exists m$ large enough such that

$$R(T) = r_{\omega_m}(T) \leq r_{\omega_m}(T_m) + \varepsilon \leq r_{\omega_m}(T') + \varepsilon \leq \sup_\theta R(\theta, T') + \varepsilon$$

(always true when R independent of θ)

This holds for any arbitrary T'. Further since ε can be chosen arbirarily small it follows that T is minimax. (since $R(T) \leq \sup_\theta R(\theta, T')$.

□

example: $X \sim \text{Binomial}(n,\theta)$, prior $\theta \sim \text{Beta}(r,s)$. The bayes estimator for quadratic loss is

$$T = \frac{X+r}{n+r+s}$$

Its risk is

$$R(\theta,T) = E_\theta[(T-\theta)^2] = \text{var}_\theta(T) + \text{bias}_\theta^2(T)$$
$$= \frac{\dots\theta^2+\dots\theta+\dots}{\dots}$$

To apply the lemma, $R(T)$ must not depend on θ. Thus set the coefficients of θ^2 and θ to be zero.

$\Rightarrow r = s = \sqrt{n}/2$ $\Rightarrow T = \frac{X+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}$ is minimax. (since it is bayes).

Admissibility Lemmas

Now assume Θ is an open set. And $R(\theta,T) < \infty$ $\forall T$.

Lemma 11.2.1: Suppose the statistic T is Bayes with prior density w. Then T is admissible if (1) or (2) hold:

(1) The statistic T is the unique Bayes decision
$r_w(T) = r_w(T') \Rightarrow \forall\theta: T = T'$ P_θ-a.s.

(2) $\forall T'$: $R(\theta,T')$ is continuous in θ, and also, $\forall U \subseteq \Theta$ open the prior probability $\Pi(U) := \int_U w(\vartheta)\,d\mu(\vartheta)$ of U is strictly positive.

proof: (1):
Assume T is inadmissible. Then for some T': $R(\theta,T') \le R(\theta,T)$ $\forall\theta$. Then also $r_w(T') \le r_w(T)$. But because T is bayes, T attains the minimum bayes risk: $r_w(T') = r_w(T)$.
By assumption, then $\forall\theta$: $T = T'$ P_θ-a.s., and thus $\forall\theta$: $R(\theta,T') = R(\theta,T)$. But this is a contradiction to the second part of the admissibility definition.

(2) Suppose T is inadmissible. Then for some T': $R(\theta,T') \le R(\theta,T)$ $\forall\theta$. And for some θ_0: $R(\theta_0,T') < R(\theta_0,T)$. This implies that for some $\varepsilon > 0$ and an open neighborhood $U \subseteq \Theta$ of θ_0 we have

$$R(\vartheta,T') \le R(\vartheta,T) - \varepsilon, \quad \vartheta \in U$$

But then

$$r_w(T') = \int_U R(\vartheta,T')\,w(\vartheta)\,d\nu(\vartheta) + \int_{U^c} \overbrace{R(\vartheta,T')}^{\le R(\vartheta,T)}\,w(\vartheta)\,d\nu(\vartheta)$$
$$\le \int_U R(\vartheta,T)\,w(\vartheta)\,d\nu(\vartheta) - \varepsilon\,\Pi(U) + \int_{U^c} R(\vartheta,T)\,w(\vartheta)\,d\nu(\vartheta)$$
$$= r_w(T) - \varepsilon\,\Pi(U) < r_w(T) \quad \lightning \text{ (contradiction)}$$

□

Lemma 11.2.2: Suppose that T is extended Bayes, and that $\forall T'$: $R(\theta,T')$ is continuous in θ. Assume $\forall U \subseteq \Theta$ open

$$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \to 0 \quad \text{as } m \to \infty$$

Here $\Pi_m(U) := \int_U w_m(\vartheta)\,d\mu_m(\vartheta)$ is the probability of U under the prior Π_m.

Then T is admissible.

(proof)

proof: Suppose T is inadmissible. Then $\exists T': R(\theta,T') \le R(\theta,T)$ and $\exists \theta_0: R(\theta_0,T') < R(\theta_0,T)$. Then $\exists \varepsilon>0\ \exists U \subseteq \Theta$ open s.t. $\theta_0 \in U$

$$R(\vartheta,T') \le R(\vartheta,T) - \varepsilon, \quad \vartheta \in U$$

Then $$r_{w_m}(T') \le r_{w_m}(T) - \varepsilon \Pi_m(U)$$

Further suppose that a Bayes decision T_m for w_m exists $\forall m$, i.e.

$$r_{w_m}(T_m) = \inf_{T'} r_{w_m}(T'), \quad m=1,2,\dots$$

Then $\forall m$:

$$r_{w_m}(T_m) \le r_{w_m}(T') \le r_{w_m}(T) - \varepsilon \Pi_m(U)$$

$$\Rightarrow \frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \ge \varepsilon > 0 \quad \lightning \text{ contradiction} \qquad \square$$

(In-)admissible estimators for the normal mean: $X \sim \mathcal{N}(\theta,1)$, $\theta \in \Theta = \mathbb{R}$.
Let $R(\theta,T) = E_\theta[(T-\theta)^2]$ be the quadratic risk.
Consider estimators of the form $T = aX + b$, $a>0$, $b \in \mathbb{R}$.
Then T is admissible iff either (1) or (2) holds
(1) $a<1$
(2) $a=1$ and $b=0$.

proof:
"$\Leftarrow$" (1): We apply Lemma 11.2.1. First we show T is Bayes for some prior. Then we show cond. (1) of the Lemma, that T is unique. This implies that T is admissible.
First, choose the prior $\theta \sim \mathcal{N}(c,\tau^2)$. From a previous example we know that

$$T_{Bayes} = E[\theta|X] = \frac{\tau^2 X + c}{\tau^2+1}$$

Thus choose $a = \frac{\tau^2}{\tau^2+1}$, $b = \frac{c}{\tau^2+1}$. Then $T = T_{Bayes}$.
Now show uniqueness. From a past example we also know

$$r_w(T') = E[\mathrm{var}[\theta|X]] + E[(T-T')^2]$$

Thus if $r_w(T') = r_w(T)$, then $E[(T-T')^2] = 0$.
We can write $X = \theta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$ indep. of θ.
Then $X \sim \mathcal{N}(c, \tau^2+1)$ corresponds to $p(x) = \int p_\vartheta(x)\, w(\vartheta)\, d\mu(\vartheta)$.
$E[(T-T')^2] = 0 \Rightarrow T = T'$ P-a.s.
Since P dominates all P_θ, we have $T = T'$ P_θ-a.s..
Thus T is unique and admissible.

"$\Leftarrow$" (2): In this case we have $T = X$. We use Lemma 11.2.2.
Because $R(\theta,T) = 1\ \forall\theta$, also $r_w(T) = 1$ for any prior w.
Let w_m be the density of $\mathcal{N}(0,m)$. From previous examples we know

$$T_{Bayes} = T_m = \frac{m}{m+1} X$$

By bias-variance decomposition, it has risk

$$R(\theta,T_m) = \frac{m^2}{(m+1)^2} + \left(\frac{m}{m+1} - 1\right)^2 \theta^2 = \frac{m^2}{(m+1)^2} + \frac{\theta^2}{(m+1)^2}$$

As $E[\theta^2] = 0$: $r_{w_m}(T_m) = m/(m+1)$
Thus

$$r_{w_m}(T) - r_{w_m}(T_m) = \frac{1}{m+1}$$

Then T is extended Bayes. It is left to show the requirement of Lemma 11.2.2

↓(proof)

Instead of an open neighborhood, it suffices to consider open intervals $U=(u,u+h)$ with $u,h>0$ fixed. We have

$$\Pi_m(U) = \Phi\left(\frac{u+h}{\sqrt{m^{-1}}}\right) - \Phi\left(\frac{u}{\sqrt{m^{-1}}}\right) = \frac{1}{\sqrt{m}}\,\phi\left(\frac{u}{\sqrt{m^{-1}}}\right) + o\left(\frac{1}{\sqrt{m^{-1}}}\right)$$

For m large $\phi\left(\frac{u}{\sqrt{m^{-1}}}\right) \approx \phi(0) = \frac{1}{\sqrt{2\pi}} > \frac{1}{4}$ (say)

Thus, for m sufficiently large

$$\frac{r_{\omega_m}(T) - r_{\omega_m}(T_m)}{\Pi_m(U)} \leq \frac{4}{h\sqrt{m^{-1}}} \longrightarrow 0 \quad \text{as } m\to\infty$$

This proves admissibility.

"⇒" (1+2): We show that if (1) or (2) doesn't hold, then T is inadmissible.

case 1: $a>1$. We have $R(\theta, aX+b) \geq \mathrm{var}(aX+b) > 1 = R(\theta, X)$
So $aX+b$ is inadmissible

case 2: $a=1$, $b\neq 0$. The bias term makes $aX+b$ inadmissible:
$R(\theta, X+b) = 1+b^2 > 1 = R(\theta, X)$

□

Least squares estimator (LSE): We now regard our observations now as a matrix $X\in\mathbb{R}^{n\times p}$ (with non-random entries). Given each row, we look for the best linear approximation of y_i (row label)
If X has rank p, the LSE is given by:

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

$$= (X^TX)^{-1}X^Ty$$

→ X has a column of 1's $\begin{pmatrix} 1 & -x_1- \\ 1 & \\ 1 & -x_n- \end{pmatrix} = X$

Then $X\hat\beta$ corresponds exactly to the projection of Y onto the column space of X.

Distribution of LSE: For $f=E[Y]$ define $\beta^* := (X^TX)^{-1}X^Tf$.
We call $X\beta^*$ the best linear approximation of f.

Lemma 12.3.1: Suppose $E[\varepsilon\varepsilon^T] = \sigma^2 I$ $(=\mathrm{var}(\varepsilon))$ where $\varepsilon := Y - f = Y - E[Y]$. Then

(1) $E[\hat\beta] = \beta^*$, $\mathrm{Cov}(\hat\beta) = \sigma^2(X^TX)^{-1}$

(2) $E[\|X(\hat\beta-\beta^*)\|_2^2] = \sigma^2 p$

(3) $E[\|X\hat\beta - f\|_2^2] = \underbrace{\sigma^2 p}_{\text{estimation error}} + \underbrace{\|X\beta^* - f\|_2^2}_{\text{misspecification error}}$

proof:
(1): Compute directly: $E[\hat\beta - \beta^*] = E[\underbrace{(X^TX)^{-1}X^T}_{=:A}\underbrace{(Y-f)}_{=\varepsilon}] = A\underbrace{E[\varepsilon]}_{=0} = 0$ (constant)

and $\mathrm{Cov}(\hat\beta) = \mathrm{Cov}(\hat\beta-\beta^*) = \mathrm{Cov}(A\varepsilon) = A\,\underbrace{\mathrm{Cov}(\varepsilon)}_{\sigma^2 I}\,A^T$

$$= \sigma^2 AA^T = \sigma^2(X^TX)^{-1}$$

↓(2+3)

b(proof)

(2): Define the projection $PP^T := X(X^TX)^{-1}X^T$. Then we can write

$$\|X(\hat{\beta}-\beta)\|_2^2 = \|PP^T\varepsilon\|_2^2 =: \sum_{j=1}^{p} V_j$$

where $V = P^T\varepsilon$ (since: $\|PP^T\varepsilon\|_2^2 = \varepsilon^T P \overbrace{P^TP}^{=I} P^T\varepsilon = (P^T\varepsilon)^T P^T\varepsilon = \|P^T\varepsilon\|_2^2$)
Note that $E[V] = P^T E[\varepsilon] = 0$ and $Cov(V) = P^T Cov(\varepsilon) P = \sigma^2 I$
Then

$$E\left[\sum_{j=1}^{p} V_j^2\right] = \sum_{j=1}^{p} \underbrace{E[V_j^2]}_{\text{diagonal elements of } Cov(V)} = \sigma^2 p.$$

(3): We have for any b

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

because $X\beta^* - f$ is orthogonal to X. Now choose $b = \hat{\beta}$, apply (2), and take the Expectation on both sides.
The statement follows. □

Asymptotic theory

We now regard the i.i.d. $X_1, \dots, X_n \in \mathcal{X}$ as the first n random variables of an infinite sequence. They have distribution P.
Write $\mathbb{P} = P_X P_X \dots$ to be the distribution of the sequence $\{X_i\}_{i=1}^{\infty}$.
We still have that $P = P_\theta$, $\gamma = g(\theta) \in \mathbb{R}^p$ and $\Gamma = \{g(\theta) \mid \theta \in \Theta\}$.

$T_n = T_n(X_1, \dots, X_n)$ is an estimator.

Almost sure convergence: The random variables $Z_n \in \mathbb{R}^p$ converge almost surely to Z, if

$$\mathbb{P}(\{\omega \in \Omega \mid \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1$$

we write $Z_n \xrightarrow{a.s.} Z$.

Convergence in probability: Let $\{Z_n\}_{n=1}^{\infty}$ and Z be $\mathbb{R}^p$-valued r.v.
We say that Z_n converges in probability to Z, if
$\forall \varepsilon > 0$: $\lim_{n\to\infty} \mathbb{P}(\|Z_n - Z\| > \varepsilon) = 0$

we write $Z_n \xrightarrow{\mathbb{P}} Z$.

Note: Chebychev's inequality can be used to prove convergence in probability. For all $\Psi: [0,\infty) \to [0,\infty)$ increasing functions

$$\mathbb{P}(\|Z_n - Z\| \geq \varepsilon) \leq \frac{\mathbb{E}[\Psi(\|Z_n - Z\|)]}{\Psi(\varepsilon)}.$$

Convergence in distribution: Let $\{Z_n\}_{n=1}^{\infty}$ and Z be $\mathbb{R}^p$-valued r.v.
We say that Z_n converges in distribution to Z if
$\forall f$ continuous and bounded $\lim_{n\to\infty} E[f(Z_n)] = E[f(Z)]$

we write $Z_n \xrightarrow{D} Z$.

(→ implies bounded in probability)

Convergence implications:

- conv. in probability ⇒ conv. in distribution
- conv. in distribution to const. $c \in \mathbb{R}$ ⇒ conv. in probability
- almost sure conv. ⇒ conv. in probability
- conv. in probability ⇒ ∃ subsequence with almost sure conv.

Central Limit theorem: (CLT) Let $X_1, \dots, X_n$ be i.i.d. in $\mathbb{R}$. with $E[X_i] = \mu$, $\mathrm{Var}[X_i] = \sigma^2$. Then

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \qquad \left(\bar{X}_n = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i\right)$$

cdf version:

$$\mathbb{P}\left(\sqrt{n}\,\frac{(\bar{X}_n - \mu)}{\sigma} \le z\right) \to \Phi(z) \quad \forall z$$

Cramér-World device: Let $(\{Z_n\}, Z)$ be r.v. then

$$Z_n \xrightarrow{D} Z \iff a^T Z_n \xrightarrow{D} a^T Z \quad \forall a \in \mathbb{R}^p$$

Stochastic order symbols: Let $\{Z_n\}$ be $\mathbb{R}^p$-valued random variables. And $\{r_n\}$ strictly positive random variables. We write

$$Z_n = O_p(1) \quad \text{if} \quad \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(\|Z_n\| > M) = 0$$

We say that Z_n is bounded in probability. (uniform tightness of $\{Z_n\}$).
Further we write $Z_n = O_p(r_n)$ if $\frac{Z_n}{r_n} = O_p(1)$.

Also if $Z_n \xrightarrow{\mathbb{P}} 0$ then write $Z_n = o_p(1)$

Z_n is of small order r_n in probability, if $\frac{Z_n}{r_n} = o_p(1)$.
We write $Z_n = o_p(r_n)$.

Note: $Z_n = o_p(1) \Rightarrow Z_n = O_p(1)$

$Z_n = o_p(r_n) \Rightarrow Z_n = O_p(r_n)$.

Note: Z_n converges in distribution $\Rightarrow Z_n = O_p(1)$

Slutsky's Theorem: Let $(\{Z_n, A_n\}, Z)$ be a collection of $\mathbb{R}^p$-valued random variables. $a \in \mathbb{R}^p$ a vector of constants.
If $Z_n \xrightarrow{D} Z$ and $A_n \xrightarrow{\mathbb{P}} a$, then

$$A_n^T Z_n \xrightarrow{D} a^T Z$$

proof:
Take f lipschitz and bounded.
$|f| \le C_B$, $|f(z) - f(\tilde{z})| \le C_L \|z - \tilde{z}\|$

→ Z_n is bounded, and A_n, a as well. Thus we can define f on a compact set. Thus, since f is continuous, it is automatically Lipschitz

Write $|E[f(A_n^T Z_n) - E[f(a^T Z)]|$

$$\le \underbrace{|E[f(A_n^T Z_n)] - E[f(a^T Z_n)]|}_{(*)} + \underbrace{|E[f(a^T Z_n)] - E[f(a^T Z)]|}_{\to 0}$$

→ also by Cramér world device

Because f is bounded and Lipschitz, the second term goes to 0.
Now show that $(*)$ goes to 0. Let $\varepsilon > 0$, $M > 0$ be arbitrary.
Define $S_n := \{\|Z_n\| \le M, \|A_n - a\| \le \varepsilon\}$. Then

$$|E[f(A_n^T Z_n)] - E[f(a^T Z_n)]| \le E[|f(A_n^T Z_n) - f(a^T Z_n)|]$$
$$= E[|f(A_n^T Z_n) - f(a^T Z_n)|] \cdot 1_{\{S_n\}} + E[|f(A_n^T Z_n) - f(a^T Z_n)|]\, 1_{\{S_n^c\}}$$
$$\le C_L \varepsilon M + 2 C_B \mathbb{P}(S_n^c) \to 0$$

And because $P(S_n^c) \le P(\|Z_n\| > M) + \mathbb{P}(\|A_n - a\| > \varepsilon) \xrightarrow{n\to\infty} 0$ (for M large enough) ← Z_n is bounded

(for ε arbitrarily small). □

Consistency: We call estimators $\{T_n\}$ of $\gamma = g(\theta)$ consistent, if

$$T_n \xrightarrow{P_\theta} \gamma \qquad \text{(for } \theta \text{ fixed)}.$$

Asymptotic normality: We call estimators $\{T_n\}$ of $\gamma = g(\theta)$ asymptotically normal with asymptotic covariance matrix V_θ, if

$$\sqrt{n}\,(T_n - \gamma) \xrightarrow{D_\theta} \mathcal{N}(0, V_\theta) \qquad \text{and if } \{T_n\} \text{ is consistent.}$$

Asymptotic linearity: We call estimators $\{T_n\}$ of $\gamma = g(\theta) \in \mathbb{R}^p$ asymptotically linear, if for a $l_\theta : \mathcal{X} \to \mathbb{R}^p$ with

$$E_\theta[l_\theta(X)] = 0 \quad \text{and} \quad E_\theta[l_\theta(X)\, l_\theta^T(X)] =: V_\theta < \infty$$

it holds that

$$T_n - \gamma = \frac{1}{n} \sum_{i=1}^n l_\theta(X_i) + o_{P_\theta}\left(\frac{1}{\sqrt{n}}\right)$$

→ corresponds to a taylor expansion.

We call l_θ the influence function of $\{T_n\}$. It approximately measures the influence of an additional observation x.

Remark: asymptotically linear $\Rightarrow$ asymptotically normal.
because $\sqrt{n}\,(T_n - \gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^n l_\theta(X) + o_{P_\theta} \xrightarrow{D_\theta} \mathcal{N}(0, V_\theta)$.

example: Assume entries of X have finite variance, $T_n := \bar{X}_n$ is a linear, and thus asymptotically linear, estimator of mean μ with influence function $l_\theta(x) = x - \mu$. (no o_p-term in $T_n - \gamma$)

example: mean μ, $\gamma = \mu^2$, $T_n = \bar{X}_n^2$. Then we have

$$T_n - \gamma = \bar{X}_n^2 - \mu^2 = (\bar{X}_n - \mu)(\bar{X}_n + \mu) = (\bar{X}_n - \mu)\cdot 2\mu - \underbrace{(\bar{X}_n - \mu)^2}_{o_{P_\theta}(\frac{1}{n}) = o_{P_\theta}(\frac{1}{\sqrt{n}})}$$

Thus $l_\theta(x) = 2\mu(x - \mu)$
$\Rightarrow V_\theta = 4\mu^2\sigma^2$.

Same steps for $\gamma = \sigma^2$ and $T_n = \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

δ-technique: Suppose we have estimators $\{T_n\}$ of $\gamma = g(\theta) \in \mathbb{R}^p$, that fulfill for a $l_\theta : \mathcal{X} \to \mathbb{R}^p$

$$E_\theta[l_\theta(X)] = 0 \quad \text{and} \quad \operatorname{Cov}(l_\theta(X)) = V_\theta < \infty$$

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

Let $h : \mathbb{R}^p \to \mathbb{R}$ be differentiable at $\gamma = g(\theta)$. Write $\dot{h}(\gamma) \in \mathbb{R}^p$.
Then ($h(T_n)$ is also asymptotically linear)

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

Note: $\sqrt{n}\,(h(T_n) - h(\gamma)) \xrightarrow{D_\theta} \mathcal{N}(0, \dot{h}(\gamma)^T V_\theta \dot{h}(\gamma))$

example: Bernoulli(θ), $X \in \{0,1\}$, $P_\theta(X=1) = 1 - P_\theta(X=0) = \theta$. $X_1, \dots, X_n \sim X$ i.i.d.
$T = \bar{X}_n$, $\bar{X}_n - \theta = \frac{1}{n}\sum_{i=1}^n (X_i - \theta)$, $\sqrt{n}\,(\bar{X}_n - \theta) \xrightarrow{D_\theta} \mathcal{N}(0, \theta(1-\theta))$.
Let $h(\theta) = \log(\theta/(1-\theta))$, $\dot{h}(\theta) = 1/(\theta(1-\theta))$
Thus by the δ-technique we have

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

M-estimators: We define a new general class of estimators. ($\gamma = g(\theta)$)
But first we need to define some quantities. For each $\gamma \in \Gamma$ (parameter of interest) we define a loss function $\rho_\gamma(x) := L(\theta, d(x))$, for some decision d.

Further the **theoretical risk**

$$R(c) := E_\theta[\rho_c(X)] \quad \text{such that}$$

$$\gamma = \arg\min_{c \in \Gamma} R(c) = \arg\min_{c \in \Gamma} E_\theta[\rho_c(X)]. \quad (\rightarrow \text{given } \rho_c, \text{ this also defines } \gamma)$$

We can write $\dot{R}(\gamma) = 0$, if $c \mapsto \rho_c(x)$ is diff.able. Then write

$$\psi_c(x) := \dot{\rho}_c(x) = \frac{\partial}{\partial c}\rho_c(x) \quad \text{and thus} \quad \dot{R}(c) = E_\theta[\psi_c(X)].$$

The **empirical risk** is defined as

$$\hat{R}_n(c) := \frac{1}{n}\sum_{i=1}^n \rho_c(X_i), \quad c \in \Gamma.$$

Finally an **M-estimator** $\hat{\gamma}_n$ of γ is defined as

$$\hat{\gamma}_n := \arg\min_{c \in \Gamma} \frac{1}{n}\sum_{i=1}^n \rho_c(X_i) = \arg\min_{c \in \Gamma} \hat{R}_n(c).$$

If $\forall x$ $\rho_c(x)$ is differentiable in c, then we define the **Z-estimator** $\hat{\gamma}_n$ of γ to be defined by

$$\dot{\hat{R}}_n(\hat{\gamma}_n) = 0 \quad \text{where} \quad \dot{\hat{R}}_n(c) = \frac{1}{n}\sum_{i=1}^n \psi_c(X_i).$$

example: $X \in \mathbb{R}$, $\Gamma = \mathbb{R}$, $\gamma = \mu = E_\theta[X]$, $\rho_c(x) = (x-c)^2$
then $\mu = \arg\min_c E_\theta[(X-c)^2]$
by bias variance decomposition, we have $E_\theta[(X-c)^2] = \sigma^2 + (\mu - c)^2$
Then clearly the term $\frac{1}{n}\sum_{i=1}^n (X_i - c)^2$ is minimized at $\bar{X}_n$.
This is the M-estimator.

example: $X \in \mathbb{R}$, $\Gamma = \mathbb{R}$, $\gamma = \mu = E_\theta[X]$, $\rho_c(x) = |x-c|$ $\rightarrow$ not diff.able at c.

First show that $R(c) = \int_{x<c} F(x)\,dx + \int_{x>c}(1-F(x))\,dx$
for $F(x) = P(X \le x)$ the density.

$$\rightarrow R(c) = E[\rho_c(X)] = \int_{x \le c}(c-x)\,dF(x) + \underbrace{\int_{x>c}(x-c)\,dF(x)}_{= -\int_{x>c}(x-c)\,d(1-F(x))}$$

$$= \underbrace{[(c-x)F(x)]_{-\infty}^{c}}_{=0} - \underbrace{[(x-c)(1-F(x))]_c^{\infty}}_{=0} + \int_{x<c} F(x)\,dx + \int_{x>c}(1-F(x))\,dx$$

Then $\dot{R}(c) = F(c) - (1-F(c)) = 2F(c) - 1 \doteq 0 \Rightarrow \gamma = F^{-1}(\frac{1}{2})$
Assuming F is strictly increasing and continuous at $F^{-1}(\frac{1}{2})$
We can write the same for the empirical risk

$$\hat{R}_n(c) = \int_{x \le c}\hat{F}_n(x)\,dx + \int_{x>c}(1-\hat{F}_n(x))\,dx \Rightarrow \hat{\gamma}_n = \hat{F}_n^{-1}(\tfrac{1}{2})$$

Thus

$$\hat{\gamma}_n = \begin{cases} X_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}$$

MLE as special case of M-estimation: $X \sim P_\theta$, $\Theta \subseteq \mathbb{R}^p$

suppose $\{x \mid p_\theta(x) > 0\}$ does not depend on θ (to not divide by 0 later)

for the MLE we have:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^n \log(p_\theta(x_i)) = \arg\min_{\theta \in \Theta} \hat{R}_n(\theta)$$

→ then $\rho_\theta(x)$ must be $\rho_\theta(x) = -\log p_\theta(x)$

Define $R(\theta) := E_{\tilde{\theta}}[\rho_\theta(x)] = -E_{\tilde{\theta}}[\log(p_\theta(x))]$ (for $\tilde{\theta}$ the true value of the parameter)

We call

$$K(\theta|\tilde{\theta}) = R(\theta) - R(\tilde{\theta}) = E_{\tilde{\theta}}\left[\frac{p_{\tilde{\theta}}(x)}{p_\theta(x)}\right]$$

the Kullback-Leibler information.

Lemma 14.1.1: The risk $R(\theta) = -E_{\tilde{\theta}}[\log(p_\theta(x))]$ is minimized at the true parameter $\tilde{\theta}$.

i.e. $K(\theta|\tilde{\theta}) = R(\theta) - R(\tilde{\theta}) \geq 0$

proof:

$$R(\theta) - R(\tilde{\theta}) = -E_{\tilde{\theta}}\left[\log\left(\frac{p_\theta(x)}{p_{\tilde{\theta}}(x)}\right)\right] \overset{\text{Jensen}}{\geq} -\log\left(E_{\tilde{\theta}}\left[\frac{p_\theta(x)}{p_{\tilde{\theta}}(x)}\right]\right)$$

$$= -\log\left[\int \frac{p_\theta}{p_{\tilde{\theta}}} p_{\tilde{\theta}}\right] = -\log(1) = 0 \qquad \square$$

Further, if $p_\theta(x)$ is differentiable, then $\psi_\theta(x) = -\frac{\dot{p}_\theta(x)}{p_\theta(x)} = -s_\theta(x)$

and $\dot{R}(\theta) = -E_\theta[s_\theta(x)] = 0$

Theorem 14.2.1: (Consistency of M-estimators)

Suppose the uniform convergence

$$\sup_{c \in \Gamma} |\hat{R}_n(c) - R(c)| \xrightarrow{\mathbb{P}} 0$$

$\hat{R}$: empirical risk

R: theoretical risk

Then

$$R(\hat{\gamma}_n) \xrightarrow{\mathbb{P}} R(\gamma)$$

(γ minimizes R, $\hat{\gamma}_n$ minimizes $\hat{R}$)

proof:

$$0 \leq R(\hat{\gamma}_n) - R(\gamma) = [R_n(\hat{\gamma}_n) - R(\hat{\gamma}_n)] + [\hat{R}_n(\gamma) - R(\gamma)] + \underbrace{[\hat{R}_n(\hat{\gamma}_n) - \hat{R}_n(\gamma)]}_{\leq 0}$$

$$\leq 2 \cdot \sup_c |\hat{R}_n(c) - R(c)| \xrightarrow{\mathbb{P}} 0 \text{ (by assumption)} \qquad \square$$

Theorem 14.2.2: $X \sim P_\theta$, $\Gamma \subseteq \mathbb{R}^p$, $\rho_c : \mathcal{X} \to \mathbb{R}$, $c \in \Gamma$, $\psi_c := \frac{\partial}{\partial c}\rho_c$

remember then

$$R(c) = P\rho_c = E[\rho_c(x)], \quad \hat{R}_n(c) = \hat{P}_n \rho_c = \frac{1}{n}\sum_{i=1}^n \rho_c(x_i)$$

$$\dot{R}(c) = P\psi_c, \quad \dot{\hat{R}}_n(c) = \hat{P}_n \psi_c$$

(Note that here $P_\theta f := E_\theta[f(x)]$ and $\hat{P}_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$)

Assume the Z-estimator, i.e. $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0$ for $\dot{R}(\gamma) = 0$.

theorem: Assume (i) $\Gamma \subseteq \mathbb{R}$, (ii) $c \to \psi_c(x)$ is continuous $\forall x$,

(iii) $E[|\psi_c(x)|] < \infty$ $\forall c$

(iv) $\exists \delta$ s.t. $\dot{R}(c) > 0$ $\forall c \in (\gamma, \gamma+\delta)$ and $\dot{R}(c) < 0$ $\forall c \in (\gamma-\delta, \gamma)$ and $\dot{R}(\gamma) = 0$

Then

$$\mathbb{P}(\exists \hat{\gamma}_n : \dot{\hat{R}}_n(\hat{\gamma}_n) = 0) \longrightarrow 1$$

(→ with high probability there is a solution.)

and $\|\hat{\gamma}_n - \gamma\| = O_{\mathbb{P}}(1)$. (→ the solution $\hat{\gamma}_n$ is in the neighborhood parameter of interest)

↓

proof: Let ε be arbitrary with $0<\varepsilon<\delta$. From (iv) it follows that

$$\dot{R}(\gamma+\varepsilon) \geq 2\tilde{\varepsilon} \quad \text{and} \quad \dot{R}(\gamma-\varepsilon) \leq -2\tilde{\varepsilon} \qquad (*)$$

Let

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

then $P(A^c) \to 0$ i.e. for n large enough $(*)$ also holds for $\dot{\hat{R}}_n$.

So

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

and

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

} on A

Thus by (ii) $\exists\, \hat{\gamma}_n \in (\gamma-\varepsilon, \gamma+\varepsilon)$ such that $\dot{\hat{R}}_n(\hat{\gamma}_n) = 0$ (→ because $\dot{\hat{R}}_n$ is continuous)

And since ε was chosen arbitrarily it follows that

$$\|\hat{\gamma}_n - \gamma\| \xrightarrow{\mathbb{P}} 0$$

□

Influence function of Z-estimator: Suppose the following

(i) $\hat{\gamma}_n \xrightarrow{\mathbb{P}} \gamma$

(ii) $\hat{R}_n(\hat{\gamma}_n) = 0$, $\dot{R}(\gamma) = 0$

(iii) $\nu_n(c) := \sqrt{n}(\hat{P}_n - P)\psi_c = \sqrt{n}(\dot{\hat{R}}_n(c) - \dot{R}(c))$ $c \in \Gamma$ ("empirical process") is asymptotically continuous at γ, i.e.

$\forall$ sequences $\{\gamma_n\} \in \Gamma$ with $\|\gamma_n - \gamma\| = o_{P_\theta}(1)$ it holds $|\nu_n(\gamma_n) - \nu_n(\gamma)| = o_{P_\theta}(1)$.

(iv) $M_\theta = \frac{\partial}{\partial c^T}\dot{R}(\gamma) > 0$ and particularly exists. (→ Hessian)

(v) $J_\theta := P\psi_\gamma\psi_\gamma^T < \infty$ (covariance matrix)

Then $\hat{\gamma}_n$ is asymptotically linear with influence function

$$l_\theta = -M_\theta^{-1}\psi_\gamma$$

Note: $\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{D} \mathcal{N}(0, \underbrace{M_\theta^{-1} J_\theta M_\theta^{-1}}_{\text{"sandwich formula"}})$

Asymptotic relative efficiency: $\gamma \in \Gamma \subseteq \mathbb{R}$. Let $T_{n,1}, T_{n,2}$ be estimators of γ s.t.

$$\sqrt{n}(T_{n,j} - \gamma) \xrightarrow{D_\theta} \mathcal{N}(0, V_{\theta,j}) \quad \text{for } j = 1, 2.$$

Then we call

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

the asymptotic relative efficiency of $T_{n,2}$ with respect to $T_{n,1}$.

If $e_{2:1} > 1$, then estimator $T_{n,2}$ is asymptotically more efficient than $T_{n,1}$. Since an asymptotic $(1-\alpha)$-confidence interval for γ based on $T_{n,2}$ is then narrower than one based on $T_{n,1}$.

↓ asymptotic relative efficiency

example: $X = \mu + \varepsilon$, $\varepsilon \sim F_0$ symmetric around 0, $f_0 := F_0'$, $f_0 > 0$
$\mathrm{var}(\varepsilon) := \sigma^2 < \infty$.
$X_1, \dots, X_n$ i.i.d $\sim X$. $T_{n,1} = \bar{X}_n$, $T_{n,2}$ = sample median.
$V_{\theta,1} = \sigma^2$, $V_{\theta,2} = 1/(4 f_0^2)$. Thus

$$e_{2:1} = 4\sigma^2 f_0^2(\theta)$$

case: for f_0 the std. normal distribution: $e_{2:1} = \frac{2}{\pi} \approx 0.64 < 1$.
→ $T_{n,1}$ is asymptotically more efficient.

Asymptotic pivots (based on MLE): $\Theta \subseteq \mathbb{R}^p$ and remember the MLE

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

Further remember the loss function $\rho_\vartheta = -\log(p_\vartheta)$, $\psi_\vartheta = \dot{\rho}_{\vartheta\theta} = -s_\theta(x)$
where $s_\vartheta := \dot{p}_\vartheta / p_\vartheta$.
And the asymptotic variance of the MLE is $I^{-1}(\theta)$ where
$I(\theta) = P_\theta s_\theta s_\theta^T$ is the Fisher information:

$$\sqrt{n}(\hat{\vartheta}_n - \theta) \xrightarrow{D_\theta} \mathcal{N}(0, I^{-1}(\theta)) \quad \forall \theta.$$

Lemma 14.8.1: Define

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

Under regularity conditions, the term above is an asymptotic pivot for ϑ, i.e. ($\gamma = g(\vartheta)$), its asymptotic distribution does not depend on θ.
Especially, its asymptotic distribution is χ^2-distrib. with p degrees of freedom:

$$2\mathcal{L}_n(\hat{\vartheta}_n) - 2\mathcal{L}_n(\theta) \xrightarrow{D_\theta} \chi^2_p \quad \forall \theta.$$

Likelihood ratio tests and asymptotics: Consider the hypothesis
$H_0: R(\theta) = 0$ where $R(\theta) = \begin{pmatrix} R_1(\theta) \\ \vdots \\ R_q(\theta) \end{pmatrix}$ restrictions on $\theta \in \mathbb{R}^p$
We define the unrestricted MLE

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

and the restricted MLE:

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

Assume $\dot{R}(\theta) = \frac{\partial}{\partial \vartheta^T} R(\vartheta)\big|_{\vartheta=\theta}$ has rank q.

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

is the log-likelihood ratio for testing $H_0: R(\theta) = 0$

Lemma 14.10.1: Under regularity conditions and if $H_0: R(\theta) = 0$ holds, we have

$$2\mathcal{L}_n(\hat{\vartheta}_n) - 2\mathcal{L}_n(\hat{\vartheta}_n^0) \xrightarrow{D_\theta} \chi^2_q.$$