Fundamentals of Mathematical Statistics
Definitions/Lemmas

by Roman Böhringer

ETH Zürich

January 2, 2021

Probability Theory

Conditional Probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Bayes Theorem:

$$P(B \mid A) = P(A \mid B)\frac{P(B)}{P(A)}$$

Law of Total Probability: For a partition $\{B_j\}$ ($B_j \cap B_k = \emptyset$ for all $j \neq k$ and $P(\cup_j B_j) = 1$):

$$P(A) = \sum_j P(A \mid B_j) P(B_j)$$

Marginal Density:

$$f_X(\cdot) = \int f_{X,Y}(\cdot, y) dy$$

$$f_X(x) = \int f_X(x \mid y) f_Y(y) dy$$

Conditional Density:

$$f_X(x \mid y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$f_Y(y \mid x) = f_X(x \mid y)\frac{f_Y(y)}{f_X(x)}$$

Conditional Expectation:

$$E[g(X,Y) \mid Y = y] := \int f_X(x \mid y) g(x,y) dx$$

Iterated Expectations Lemma:

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

Law of Total Variance:

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

Distributions

Multinomial Distribution:

$$P(N_1 = n_1, \ldots, N_k = n_k) = \left(\begin{array}{c} n \\ n_1 \cdots n_k \end{array}\right) p_1^{n_1} \cdots p_k^{n_k}$$

$$\left(\begin{array}{c} n \\ n_1 \cdots n_k \end{array}\right) := \frac{n!}{n_1! \cdots n_k!}$$

Poisson Distribution:

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

Normal Distribution:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Sum of Independent Normal / Poisson Variables: For X and Y independent: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ with $Z = X + Y$, then $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.
$X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu) \Rightarrow Z \sim \mathcal{P}(\lambda + \mu)$.

Chi-Square Distribution: Let $Z_1, \ldots, Z_p$ be i.i.d. $\mathcal{N}(0,1)$-distributed and define the p-vector:

$$Z := \left(\begin{array}{c} Z_1 \\ \vdots \\ Z_p \end{array}\right)$$

Z is $\mathcal{N}(0, I)$ distributed and the χ^2-distribution with p degrees of freedom is defined as ($\|Z\|_2^2 \sim \chi_p^2$):

$$\|Z\|_2^2 := \sum_{j=1}^p Z_j^2$$

Distribution of Maximum: $Z := \max\{X_1, X_2\}$ with X_1, X_2 independent having distribution F and density f.

$$f_Z(z) = 2F(z)f(z)$$

Exponential Families: A k-dimensional exponential family is a family of distributions with densities of the form:

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

The family is in canonical form if:

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

Where:

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

Estimation

Estimator: An estimator $T(\mathbf{X})$ is a function $T(\cdot)$ evaluated at the observations $\mathbf{X}$. The function $T(\cdot)$ is not allowed to depend on unknown parameters.
Empirical Distribution Function:

$$\hat{F}_n(\cdot) := \frac{1}{n}\#\{X_i \leq \cdot, 1 \leq i \leq n\}$$

Method of Moments: Given the first p moments of X:

$$\mu_j(\theta) = E_\theta X^j = \int x^j dF_\theta(x), j = 1, \ldots, p$$

And the map m with inverse m^{-1}:

$$m(\theta) = [\mu_1(\theta), \ldots, \mu_p(\theta)] \qquad m^{-1}(\mu_1, \ldots, \mu_p)$$

We calculate:

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

And plug in:

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

Maximum Likelihood Estimator: Given the likelihood function:

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

We calculate:

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

Sufficiency

Sufficiency: Some given map $S : \mathcal{X} \to \mathcal{Y}$ is called sufficient for $\theta \in \Theta$ if for all θ, and all possible s, the following conditional distribution does not depend on θ:

$$P_\theta(X \in \cdot \mid S(X) = x)$$

Factorization Theorem of Neyman$^{\mathcal{PR}}$: Given densities p_θ, S is sufficient if and only if there are some functions $g_\theta(\cdot) \geq 0$ and $h(\cdot) \geq 0$ such that we can write:

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

Sufficiency for Exponential Families: For a k-dimensional exponential family, the k-dimensional statistic $S(X) = (T_1(X), \ldots, T_k(X))$ is sufficient for θ. For n i.i.d. samples, the following statistic is sufficient:

$$S(\mathbf{X}) = (\frac{1}{n}\sum_{i=1}^{n} T_1(x_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} T_k(x_i))$$

Expectation/Covariance of Sufficient Statistic for Exponential Families: Given an exponential family in canonical form and:

$$\dot{d}(\theta) := \frac{\partial}{\partial\theta} d(\theta) \quad \ddot{d}(\theta) := \frac{\partial^2}{\partial\theta\partial\theta'} d(\theta) = \left(\frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}} d(\theta)\right)$$

$$T(X) := \begin{pmatrix} T_1(X) \\ \vdots \\ T_k(X) \end{pmatrix}, \quad E_\theta T(X) := \begin{pmatrix} E_\theta T_1(X) \\ \vdots \\ E_\theta T_k(X) \end{pmatrix}$$

$$\mathrm{Cov}_\theta(T(X)) := E_\theta T(X)T'(X) - E_\theta T(X) E_\theta T'(X)$$

We have$^{\mathcal{PR}}$:

$$E_\theta T(X) = \dot{d}(\theta), \mathrm{Cov}_\theta(T(X)) = \ddot{d}(\theta)$$

If the family is not in canonical form:

$$E_\theta T(X) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}$$

$$\mathrm{var}_\theta(T(X)) = \frac{1}{[\dot{c}(\theta)]^2}\left(\ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)\right)$$

Minimal Sufficiency: Two likelihoods $L_x(\theta)$ and $L_{\tilde{x}}(\theta)$ are proportional at $(x, \tilde{x})$ if

$$L_x(\theta) = L_{\tilde{x}}(\theta) c(x, \tilde{x}) \,\forall\theta$$

for some constant $c(x, \tilde{x})$. A sufficient statistic S is called minimal sufficient if $S(x) = S(\tilde{x})$ for all x and $\tilde{x}$ where the likelihoods are proportional.

Completeness: Sufficient statistic S is called complete if (where h is a function not depending on θ):

$$E_\theta h(S) = 0 \forall\theta \Rightarrow h(S) = 0, P_\theta - a.s. \quad \forall\theta$$

Completeness for Exponential Families: Given a k-dimensional exponential family and

$$\mathcal{C} := \{(c_1(\theta), \ldots, c_k(\theta)) : \theta \in \Theta\} \subset \mathbb{R}^k$$

If $\mathcal{C}$ is truly k-dimensional, $S := (T_1, \ldots, T_k)$ is complete.

Fisher Information

Score Function:

$$s_\theta(x) := \frac{d}{d\theta} \log p_\theta(x) = \frac{\dot{p}_\theta(x)}{p_\theta(x)}$$

$$E_\theta s_\theta(X) = 0$$

For n i.i.d. observations:

$$\mathbf{s}_\theta(\mathbf{x}) = \sum_{i=1}^{n} s_\theta(x_i)$$

Fisher Information:

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

$$I(\theta) = -E_\theta \dot{s}_\theta(X)$$

For n i.i.d. observations:

$$\mathbf{I}(\theta) = nI(\theta)$$

Fisher Information for Exponential Families:

$$I(\theta) = \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)$$

And for $\gamma = c(\theta)$:

$$I_0(\gamma) = \ddot{d}_0(\gamma) = \frac{I(\theta)}{[\dot{c}(\theta)]^2}$$

Higher-Dimensional Extensions (Score Vector & Fisher Information Matrix):

$$s_\theta(\cdot) := \begin{pmatrix} \partial \log p_\theta / \partial\theta_1 \\ \vdots \\ \partial \log p_\theta / \partial\theta_k \end{pmatrix}$$

$$I(\theta) = E_\theta s_\theta(X) s'_\theta(X) = \mathrm{Cov}_\theta(s_\theta(X))$$

Bias, Variance

Bias:

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

T is unbiased if $\mathrm{bias}_\theta(T) = 0 \quad \forall\theta$.
Mean Square Error$^{\mathcal{PR}}$:

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

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

Uniform Minimum Variance Unbiased: Unbiased estimator T^* is UMVU if for any other unbiased estimator:

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

Conditioning on Sufficient Statistic: If T is unbiased, S sufficient, and $T^* := E(T \mid S)$:

$$E_\theta(T^*) = g(\theta) \qquad \mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T) \,\forall\theta$$

Lehmann-Scheffé Lemma: If T is an unbiased estimator of $g(\theta)$ with finite variance (for all θ) and S is sufficient and complete, $T^* := E(T \mid S)$ is UMVU.

Cramér Rao Lower Bound: If the support of p_θ does not depend on θ and p_θ is differentiable in L_2, for an unbiased estimator T of $g(\theta)$ (with derivative $\dot{g}(\theta)$), we have:

$$\dot{g}_\theta(x) = \operatorname{cov}(T, s_\theta(X))$$

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

CRLB for Exponential Families: If T is unbiased and reaches the CRLB, then there exist functions $c(\theta)$, $d(\theta)$, and $h(x)$ such that for all θ:

$$p_\theta(x) = \exp[c(\theta)T(X) - d(\theta)]h(x) \quad x \in \mathcal{X}$$

$$g(\theta) = \dot{d}(\theta)/\dot{c}(\theta)$$

Higher-Dimensional CRLB: For an unbiased estimator T of $g(\theta)$:

$$\operatorname{var}_\theta(T) \geq \dot{g}(\theta)' I(\theta)^{-1} \dot{g}(\theta)$$

Comparison

Risk: Given loss function $L(\cdot,\cdot)$:

$$\boxed{R(\theta, T) := \mathbb{E}_\theta(L(\theta, T(X))}$$

Risk and sufficiency: S sufficient for θ and $d: \mathcal{X} \to \mathcal{A}$ some decision. Then there is a randomized decision $\delta(S)$ such that:

$$R(\theta, \delta(S)) = R(\theta, d) \quad \forall\theta$$

Rao-Blackwell$^{\mathcal{PR}}$: S sufficient for θ, $\mathcal{A} \subset \mathbb{R}^p$ convex and $a \mapsto L(\theta, a)$ convex for all θ. For decision $d: \mathcal{X} \to \mathcal{A}$ and $d'(s) := E(d(X) \mid S = s)$:

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

Sensitivity/Robustness: Influence function

$$l(x) := (n+1)\left(T_{n+1}(X_1, \ldots, X_n, x) - T_n(X_1, \ldots, X_n)\right), x \in \mathbb{R}$$

For $m \leq n$:

$$\epsilon(m) := \sup_{x_1^*, \ldots, x_m^*} |T(x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)|$$

Break down point:

$$\epsilon^* := \min\{m : \epsilon(m) = \infty\}/n$$

Equivariant Statistics

Location equivariant statistic: For all constants $c \in \mathbb{R}$ and $\mathbf{x} = (x_1, \ldots, x_n)$:

$$T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c$$

Location invariant loss function: For all constants $c \in \mathbb{R}$:

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

Risk for equivariant statistics/invariant loss functions:

$$R(\theta, T) = E_\theta L(0, T(\mathbf{X} - \theta)) = E L_0[T(\varepsilon)]$$

Uniform Minimum Risk Equivariant:

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

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

Maximal Invariant: Map $\mathbf{Y}: \mathbb{R}^n \to \mathbb{R}^n$ is maximal invariant if:

$$\mathbf{Y}(\mathbf{x}) = \mathbf{Y}(\mathbf{x}') \Leftrightarrow \exists c : \mathbf{x} = \mathbf{x}' + c$$

UMRE estimator construction: $d(\mathbf{X})$ equivariant, $\mathbf{Y} := \mathbf{X} - d(\mathbf{X})$.

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

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

UMRE estimator for quadratic loss:

$$T \text{ is UMRE } \Leftrightarrow E_0(T(\mathbf{X}) \mid \mathbf{X} - T(\mathbf{X})) = 0$$

Pitman estimator:

$$T^*(\mathbf{X}) = X_n - E(\epsilon_n \mid \mathbf{Y})$$

Basu's lemma$^{\mathcal{PR}}$: Let X have distribution P_θ, suppose T is sufficient/complete, and $Y = Y(X)$ has a distribution that does not depend on θ. Then, T and Y are independent under P_θ for all θ.

Tests and Confidence Intervals

Quantile Functions:

$$q_{\sup}^F(u) := \sup\{x : F(x) \leq u\}$$

$$q_{\inf}^F(u) := \inf\{x : F(x) \geq u\} := F^{-1}(u)$$

Test: For $\gamma_0 \in \Gamma$, $\alpha \in [0,1]$ a test for $H_0: \gamma = \gamma_0$ is a statistic $\phi(X, \gamma_0) \in \{0, 1\}$ such that $P_\theta(\phi(X, \gamma_0) = 1) \leq \alpha$ for all $\theta \in \{\vartheta : g(\vartheta) = \gamma_0\}$

Pivot: Function $Z(\mathbf{X}, \gamma)$ such that for all $\theta \in \Theta$, this distribution does not depend on θ:

$$\mathbb{P}_\theta(Z(\mathbf{X}, g(\theta)) \leq \cdot) =: G(\cdot)$$

We can construct test for H_{γ_0}:

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

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

Student's test: Assume data is normal distributed with same variance. Then:

$$T := Z(\mathbf{X}, \mathbf{Y}, 0) = Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X}}{S}\right)$$

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

And one-sided test at level α for $H_0: \gamma = 0$ against $H_1: \gamma < 0$ is:

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

Wilcoxon's test: Let $R_i := \operatorname{rank}(Z_i)$ among the pooled sample. Then:

$$T := \sum_{i=1}^n R_i = \#\{Y_j < X_i\} + \frac{n(n+1)}{2}$$

And (the distribution is often tabulated):

$$\mathbb{P}_{H_0}(T=t)=\frac{\#\left\{\mathbf{r}:\sum_{i=1}^n r_i=t\right\}}{N!}$$

Uniformly Most Powerful Tests

Level: ϕ is a test at level α if:

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

A test ϕ is UMP if it has level α and for all tests ϕ' with level α:

$$E_\theta\phi'(X)\leq E_\theta\phi(X)\quad\forall\theta\in\Theta_1$$

Neyman Pearson Lemma$^{\mathcal{PR}}$: $H_0:\theta=\theta_0$ and $H_1:\theta=\theta_1$.

$$R(\theta,\phi):=\left\{\begin{array}{ll} E_\theta\phi(X), & \theta=\theta_0\\ 1-E_\theta\phi(X), & \theta=\theta_1\end{array}\right.$$

$$\phi_{\mathrm{NP}}:=\left\{\begin{array}{ll} 1 & \text{if } p_1/p_0>c\\ q & \text{if } p_1/p_0=c\\ 0 & \text{if } p_1/p_0<c\end{array}\right.$$

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

One Sided UMP Test$^{\mathcal{PR}}$: Given n i.i.d. copies of a Bernoulli random variable with success parameter θ and with $T:=\sum_{i=1}^n X_i$ as the number of successes, the following test is UMP for $H_0:\theta\geq c$, $H_1:\theta<c$ (and also the weaker hypothesis $H_0:\theta=c$, $H_1:\theta=c_-$ or $H_0:\theta=c$, $H_1:\theta<c$):

$$\phi(T):=\left\{\begin{array}{ll} 1 & \text{if } T<t_0\\ q & \text{if } T=t_0\\ 0 & \text{if } T>t_0\end{array}\right.$$

Where t_0 is chosen such that $P_{\theta_0}\left(T\leq t_0-1\right)\leq\alpha, P_{\theta_0}\left(T\leq t_0\right)>\alpha$ and q such that $P_{\theta_0}\left(H_0\text{ rejected }\right)=P_{\theta_0}\left(T\leq t_0-1\right)+qP_{\theta_0}\left(T=t_0\right):=\alpha$, i.e.:

$$q=\frac{\alpha-P_{\theta_0}\left(T\leq t_0-1\right)}{P_{\theta_0}\left(T=t_0\right)}$$

UMP tests for exponential families: $H_0:\theta\leq\theta_0$, $H_1:\theta>\theta_0$, and $c(\theta)$ strictly increasing. Then a UMP test is:

$$\phi(T(x)):=\left\{\begin{array}{ll} 1 & \text{if } T(x)>t_0\\ q & \text{if } T(x)=t_0\\ 0 & \text{if } T(x)<t_0\end{array}\right.$$

Unbiased tests: Test ϕ is unbiased if for all $\theta\in\Theta_0$, $\vartheta\in\Theta_1$:

$$E_\theta\phi(X)\leq E_\vartheta\phi(X)$$

Uniformly Most Powerful Unbiased: Unbiased test ϕ is UMPU if it has level α and for all unbiased tests ϕ' with level α, $E_\theta\phi'(X)\leq E_\theta\phi(X)\quad\forall\theta\in\Theta_1$

UMPU for a one-dimensional exponential family: $\mathcal{P}$ one-dimensional exponential family with $c(\theta)$ strictly increasing in θ. A UMPU test is then:

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

With constants t_R,t_L,q_R, and q_L such that:

$$E_{\theta_0}\phi(X)=\alpha,\ \frac{d}{d\theta}E_\theta\phi(X)\bigg|_{\theta=\theta_0}=0$$

Confidence Intervals

Confidence Set: Subset $I=I(\mathbf{X})\subset\Gamma$, depending only on the data, is a confidence set for γ at level $1-\alpha$ if:

$$\mathbb{P}_\theta(\gamma\in I)\geq 1-\alpha,\forall\theta\in\Theta$$

Confidence Interval:

$$I:=[\underline{\gamma},\bar{\gamma}]$$

with $\underline{\gamma}=\underline{\gamma}(\mathbf{X})$, $\bar{\gamma}=\bar{\gamma}(\mathbf{X})$.

Confidence Sets / Tests: Given for each $\gamma_0\in\mathbb{R}$ a test at level α for H_{γ_0}, the following is a $(1-\alpha)$-confidence set for γ:

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

Given a $(1-\alpha)$-confidence set for γ, the following test is a test at level α of $H_{\gamma_0}:\gamma=\gamma_0$ for all γ_0:

$$\phi\left(\mathbf{X},\gamma_0\right)=\left\{\begin{array}{ll} 1 & \text{if } \gamma_0\notin I(\mathbf{X})\\ 0 & \text{else}\end{array}\right.$$

Decision Theory

Admissible Decision: A decision d' is strictly better than d if:

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

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

d is called inadmissible when there exists a d' that is strictly better than d.

Admissibility for the Neyman Pearson Test: A Neyman Pearson test is admissible if and only if its power is strictly less than 1 or it has minimal level among all tests with power 1.

Admissible Estimators for the Normal Mean$^{\mathcal{PR}}$**:** $X\sim\mathcal{N}(\theta,1),\Theta:=\mathbb{R}$ and $R(\theta,T):=E_\theta(T-\theta)^2$. If we consider estimators of the form $T=aX+b, a>0, b\in\mathbb{R}$, T is admissible if and only if one of the following cases hold:

1. $a<1$
2. $a=1$ and $b=0$

Minimax Decisions: d minimax if

$$\sup_\theta R(\theta,d)=\inf_{d'}\sup_\theta R\left(\theta,d'\right)$$

Minimax for the Neyman Pearson Test: A Neyman Pearson test is minimax if and only if $R(\theta_0,\phi_{NP})=R(\theta_1,\phi_{NP})$

Bayes Decisions

Bayes Risk: Given probability measure Π (prior distribution) of Θ, and density $w := d\Pi/d\mu$:

$$r(\Pi, d) := \int_\Theta R(\vartheta, d) d\Pi(\vartheta)$$

$$r(\Pi, d) = \int_\Theta R(\vartheta, d) w(\vartheta) d\mu(\vartheta) := r_w(d)$$

Bayes Decision: A decision d is called Bayes if:

$$r(\Pi, d) = \inf_{d'} r(\Pi, d')$$

A posteriori density: Given $p_\theta(x) = p(x \mid \theta)$, and the marginal density:

$$p(\cdot) := \int_\Theta p(\cdot \mid \vartheta) w(\vartheta) d\mu(\vartheta)$$

The a posterior density of θ is:

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

Bayes Decision Construction: Let

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

Then Bayes decision is:

$$d_{\text{Bayes}}(X) = \arg\min_{a \in \mathcal{A}} l(X, a)$$

$$d_{\text{Bayes}}(X) = \arg\min_{a \in \mathcal{A}} \int L(\vartheta, a) g_\vartheta(S) w(\vartheta) d\mu(\vartheta)$$

Bayes Test: Assume $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, $L(\theta_0, a) := a$, $L(\theta_1, a) := 1 - a$, $w(\theta_0) =: w_0$, and $w(\theta_1) =: w_1 = 1 - w_0$. Bayes test is then (for an arbitrary q):

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

Extended Bayes Decision: T is called extended Bayes if there exists a sequence of prior densities $\{w_m\}_{m=1}^\infty$ such that $r_{w_m}(T) - \inf_{T'} r_{w_m}(T') \to 0$ as $m \to \infty$.

Bayes Estimator for Quadratic Loss: $L(\theta, a) := (\theta - a)^2$, then:

$$d_{\text{Bayes}}(X) = E(\theta \mid X)$$

For $T = E(\theta \mid X)$, the Bayes risk of an estimator T' is:

$$r_w(T') = E \operatorname{var}(\theta \mid X) + E(T - T')^2$$

Bayes Estimator/MAP/MLE: For $L(\theta, a) := 1\{|\theta - a| > c\}$ and c small, Bayes rule is approximately the maximum a posteriori estimator, which is equivalent to the MLE for a uniform prior. With quadratic loss, Bayes estimator is the expectation value of the posterior, whereas the MAP is the maximum.

Credibility Interval: A $(1 - \alpha)$-credibility interval is:

$$I := \left[\hat{\theta}_L(X), \hat{\theta}_R(X)\right]$$

$$\int_{\hat{\theta}_L(X)}^{\hat{\theta}_R(X)} w(\vartheta \mid X) d\vartheta = (1 - \alpha)$$

Constructing Estimators

Minimaxity$^{\mathcal{PR}}$: Suppose T is a statistic with risk $R(\theta, T) = R(T)$ not depending on θ. Then:

1. T admissible $\Rightarrow T$ minimax
2. T Bayes $\Rightarrow T$ minimax
3. T extended Bayes $\Rightarrow T$ minimax

Admissibility$^{\mathcal{PR}}$: Suppose T is Bayes for prior density w. Then 1. or 2. are sufficient for the admissibility:

1. T is unique Bayes ($r_w(T) = r_w(T')$ implies $\forall \theta, T = T'$, P_θ-almost surely)
2. For all T', $R(\theta, T')$ is continuous in θ and for all open $U \subset \Theta$, the prior probability $\int_U w(\vartheta) d\mu(\vartheta)$ of U is strictly positive.

Admissibility, Extended Bayes$^{\mathcal{PR}}$: Suppose T is extended Bayes and for all T', $R(\theta, T')$ is continuous in θ. Furthermore, with $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ being the probabilty of U under the prior Π_m:

$$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \to 0$$

Then, T is admissible.

The Linear Model

Least Squares Estimator: Given (augmented) design matrix $X \in \mathbb{R}^{n \times p}$, the least squares estimator is the projection of Y on $\{Xb : b \in \mathbb{R}^p\}$:

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

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

Distribution of the Least Square Estimator$^{\mathcal{PR}}$: For $f = EY$, let $\beta^* := \left(X^T X\right)^{-1} X^T f$ and $X\beta^*$ the best linear approximation of f. For $E\epsilon\epsilon^T = \sigma^2 I$, $\epsilon := Y - f$:

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

Least Squares Estimator Expectation$^{\mathcal{PR}}$: When $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, we have $\hat{\beta} - \beta^* \sim \mathcal{N}\left(0, \sigma^2 \left(X^T X\right)^{-1}\right)$ and $\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{\sigma^2} \sim \chi_p^2$ A test for $H_0 : \beta = \beta_0$ is to reject H_0 when $\|X\left(\hat{\beta} - \beta^0\right)\|_2^2/\sigma_0^2 > G_p^{-1}(1 - \alpha)$ where G_p is the distribution function of a χ_p^2-distributed random variable.

Testing a Linear Hypothesis: $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and we want to test $H_0 : B\beta = 0$. Under H_0, the following fraction is χ_q^2-distributed:

$$\frac{\|Y - X\hat{\beta}_0\|_2^2 - \|Y - X\hat{\beta}\|_2^2}{\sigma^2}$$

Asymptotic Theory

We assume an estimator $T_n(X_1, \ldots, X_n)$ of γ is defined for all n, i.e. we consider a sequence of estimators.

Markov's/Chebyshev's Inequality: For all increasing functions $\psi : [0, \infty) \to [0, \infty)$:

$$\mathbb{P}(\|Z\| \geq \epsilon) \leq \frac{\mathbb{E}\psi(\|Z\|)}{\psi(\epsilon)}$$

Almost Sure Convergence: Z_n converges almost surely to Z if

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

Convergence in Probability: Z_n converges in probability to Z ($Z_n \xrightarrow{\mathbb{P}} Z$) if for all $\epsilon > 0$:

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

Almost sure convergence implies convergence in probability, but not the other way around.

Convergence in Distribution: Z_n converges in distribution to Z ($Z_n \xrightarrow{\mathcal{D}} Z$) if for all continuous and bounded functions f:

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

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

Portmanteau Theorem: The following statements are equivalent:

1. $Z_n \xrightarrow{\mathcal{D}} Z$ (i.e., $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z) \forall f$ bounded and continuous)
2. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z) \forall f$ bounded and Lipschitz (f Lipschitz if for a constant C_L, $|f(z) - f(\tilde{z})| \leq C_L \|z - \tilde{z}\|$)
3. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z) \forall f$ bounded an Q-a.s. continuous (where Q is the distribution of Z).
4. $\mathbb{P}(Z_n \leq z) \to G(z)$ for all G-continuity points z (where $G = Q(Z \leq \cdot)$ is the distribution function of Z)

Cramér-Wold Device:

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

Slutsky's Theorem$^{\mathcal{PR}}$: Assume that $Z_n \xrightarrow{\mathcal{D}} Z, A_n \xrightarrow{\mathbb{P}} a$. Then:

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

Central Limit Theorem: Let $X_1, X_2, \ldots$ be i.i.d. with mean μ, variance σ^2. Then:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

Stochastic Order Symbols: Let r_n be strictly positive random variables. $Z_n = \mathcal{O}_{\mathbf{P}}(1)$ (Z_n bounded in probability) if:

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

$Z_n = \mathcal{O}_{\mathbf{P}}(r_n)$ if $Z_n/r_n = \mathcal{O}_{\mathbf{P}}(1)$. When Z_n converges in distribution, $Z_n = \mathcal{O}_{\mathbf{P}}(1)$. If Z_n converges in probability to zero, $Z_n = o_{\mathbf{P}}(1)$ and $Z_n = o_{\mathbf{P}}(r_n)$ if $Z_n/r_n = o_{\mathbf{P}}(1)$.

Consistent Estimators: Sequence of estimators T_n is consistent if:

$$T_n \xrightarrow{\mathbb{P}_\theta} \gamma$$

Asymptotically Normal Estimators: Sequence of estimators T_n is asymptotically normal with covariance matrix V_θ:

$$\sqrt{n}(T_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta)$$

Asymptotically Linear Estimators: Sequence of estimators T_n is asymptotically linear if for a (influence) function $l_\theta : \mathcal{X} \to \mathbb{R}^p$ with $E_\theta l_\theta(X) = 0$ and $E_\theta l_\theta(X) l_\theta^T(X) =: V_\theta < \infty$:

$$T_n - \gamma = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_i) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

The δ-Technique: Let h be differentiable at c and suppose:

$$(T_n - c)/r_n \xrightarrow{\mathcal{D}} Z$$

Then:

$$(h(T_n) - h(c))/r_n \xrightarrow{\mathcal{D}} \dot{h}(c)^T Z$$

$$h(T_n) - h(c) = \dot{h}(c)^T (T_n - c) + o_{\mathbf{P}}(r_n)$$

For an asymptotically normal estimator of $\gamma = g(\theta)$ with asymptotic covariance matrix V_θ, $h(T_n)$ (h differentiable at γ) is an asymptotically normal estimator of $h(\gamma)$ with asymptotic variance:

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

If T_n is an asymptotically linear estimator of γ with influence function l_θ, $h(T_n)$ is an asymptotically linear estimator of $h(\gamma)$ with influence function $\dot{h}(\gamma)^T l_\theta$.

M-Estimators

For each $\gamma \in \Gamma$, ρ_γ is some loss function. The theoretical risk $\mathcal{R}(c) := E_\theta \rho_c(X)$ is minimized at the value $c = \gamma$ and if $c \mapsto \rho_c(x)$ is differentiable for all x, we write:

$$\psi_c(x) := \dot{\rho}_c(x) := \frac{\partial}{\partial c}\rho_c(x)$$

We then have $\dot{\mathcal{R}}(c) = E_\theta \psi_c(X)$ and $\dot{\mathcal{R}}(\gamma) = 0$. The empirical risk is defined as:

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

And the M-estimator $\hat{\gamma}_n$:

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

Assuming $\rho_c(x)$ is differentiable, the Z-estimator is $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$ with

$$\dot{\hat{\mathcal{R}}}_n(c) = \frac{\partial}{\partial c}\frac{1}{n}\sum_{i=1}^{n} \rho_c(X_i) = \frac{1}{n}\sum_{i=1}^{n} \psi_c(X_i)$$

MLE as M-Estimator: With $\rho_{\tilde{\theta}}(x) = -\log p_{\tilde{\theta}}(x)$, the M-estimator is the Maximum Likelihood Estimator.

Conditions for Uniform Convergence$^{\mathcal{PR}}$: If Γ is compact, $c \mapsto \rho_c(x)$ is continuous for all x, and

$$E_\theta\left(\sup_{c\in\Gamma}|\rho_c|\right) < \infty$$

, we have uniform convergence.

Consistency of M-Estimators$^{\mathcal{PR}}$: Suppose uniform convergence:

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

Then:

$$\mathcal{R}(\hat{\gamma}_n) \to \mathcal{R}(\gamma), \mathbb{P}_\theta - \text{a.s.}.$$

If the minimizer γ of $\mathcal{R}(c)$ is well-separated, $\hat{\gamma}_n \to \gamma, \mathbb{P}_\theta$−a.s... Well-separated means that for all $\epsilon > 0$:

$$\inf\{\mathcal{R}(c) : c \in \Gamma, \|c-\gamma\| > \epsilon\} > \mathcal{R}(\gamma)$$

Consistency of a one-dimensional Z-Estimator$^{\mathcal{PR}}$: Suppose $\Gamma \subset \mathbb{R}$, $\psi_c(x)$ continuous in c for all x, $P_\theta|\psi_c| < \infty, \forall c$, and that there is a δ s.t. $\dot{\mathcal{R}}(c) > 0, \gamma < c < \gamma+\delta$ and $\dot{\mathcal{R}}(c) < 0, \gamma-\delta < c < \gamma$. Then there is a consistent solution of $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$.

Asymptotic Linearity of Z-Estimators$^{\mathcal{PR}}$: Suppose $\hat{\gamma}_n$ is a consistent Z-estimator of γ and $|\nu_n(\gamma_n) - \nu_n(\gamma)| = o_{\mathbf{P}_\theta}(1)$ for all sequences $\gamma_n \to \gamma$ (asymptotically continuous at γ), where $\nu_n(c) = \sqrt{n}\left(\dot{\hat{\mathcal{R}}}_n(c) - \dot{\mathcal{R}}(c)\right)$. We assume

$$M_\theta := \frac{\partial}{\partial c^T}\dot{\mathcal{R}}(c)\Big|_{c=\gamma}$$

exists and is invertible, and $J_\theta := P_\theta \psi_\gamma \psi_\gamma^T$ exists. Then $\hat{\gamma}_n$ is asymptotically linear with influence function:

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

and $\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta)$ with $V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1}$.

If the map $c \mapsto \psi_c(x)$ is differentiable for all x and $\left\|\dot{\psi}_c(x) - \dot{\psi}_{\tilde{c}}(x)\right\| \le H(x)\|c-\tilde{c}\|$ for all $c, \tilde{c}$ in a neighborhood of γ with $P_\theta H < \infty$, the same result holds.

Asymptotic Normality of the MLE: Under regularity:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, I^{-1}(\theta)\right)$$

Asymptotic Relative Efficiency: Let $T_{n,1}$ and $T_{n,2}$ be two estimators of γ with:

$$\sqrt{n}(T_{n,j} - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_{\theta,j}), j = 1, 2$$

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

$$\mathrm{e}_{2:1} := \frac{V_{\theta,1}}{V_{\theta,2}}$$

Asymptotic Pivots: A function $Z_n(\gamma) := Z_n(X_1, \ldots, X_n, \gamma)$ such that its asymptotic distribution does not depend on the unknown parameter θ, i.e.:

$$Z_n(\gamma) \xrightarrow{\mathcal{D}_\theta} Z, \quad \forall\theta$$

Given

$$\sqrt{n}(T_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta), \quad \forall\theta$$

, can be constructed:

1. If V_θ only depends on γ, i.e. $V_\theta = V(\gamma)$:

$$Z_{n,1}(\gamma) := n(T_n - \gamma)^T V(\gamma)^{-1}(T_n - \gamma) \sim \chi_p^2$$

2. If we have for all θ a consistent estimator $\hat{V}_n$ of V_θ (e.g. $V_{\hat{\theta}_n}$ or $\hat{M}_n^{-1}\hat{J}_n\hat{M}_n^{-1}$):

$$Z_{n,2}(\gamma) := n(T_n - \gamma)^T \hat{V}_n^{-1}(T_n - \gamma) \sim \chi_p^2$$

3. For the MLE:

MLE Asymptotic Pivot:

$$Z_{n,3}(\theta) := 2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) :=$$

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

MLE for the multinomial distribution: $P_\theta(X = j) := \pi_j, j = 1, \ldots, k$.

$$\sum_{j=1}^{k}\frac{(N_j - n\pi_j)^2}{n\pi_j} \sim \chi_{k-1}^2$$

$$\sum_{j=1}^{k}\frac{(N_j - n\pi_j)^2}{N_j} \sim \chi_{k-1}^2$$

Likelihood Ratio Tests: Given $H_0 : R(\theta) = 0$ where the R are q restrictions on θ, $\hat{\theta}_n$ is the unrestricted MLE and $\hat{\theta}_n^0$ the restricted one, we have under H_0:

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

Appendix

Derivatives

$$\frac{d}{dx}\frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{d}{dx}\frac{x}{x+1} = \frac{1}{(x+1)^2}$$

Standardization Normal Distribution

When $X \sim \mathcal{N}(\mu, \sigma^2)$, $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, i.e.

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Showing Sufficiency

$P_\theta(X = x \cap S = s)$ often simplifies to $P_\theta(X = x)$, $S(x) = s$ because $\{X = x\} \subseteq \{S = s\}$.

Factorization Theorem of Neyman. We assume the discrete case where X only takes the values $a_1, a_2, \ldots \forall\theta$. $Q_\theta(s)$ is the distribution of S:

$$Q_\theta(s) := \sum_{j:S(a_j)=s} P_\theta(X = a_j)$$

The conditional distribution of X, given S is then:

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

Where $P_\theta(X = x \ \cap S = s) = P_\theta(X = x)$, $S(x) = s$ since the event $\{X = x\}$, $S(x) = s$ is a subset of $\{S = s\}$.
$\Rightarrow$: If S is sufficient for θ, $P_\theta(X = x \mid S = s)$ does not depend on θ per definition, but is only a function of x, say $h(x)$. Therefore (with $g_\theta(s) = Q_\theta(S = s)$):

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

$\Leftarrow$: We can write $p_\theta(x) = g_\theta(S(x))h(x)$, inserting this into $Q_\theta(s)$ gives:

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

Replacing both $p_\theta(x)$ and $Q_\theta(s)$ in the formula for $P_\theta(X = x \mid S = s)$, we get:

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

Which does not depend on θ. □

Expectation/Covariance of Sufficient Statistic for Exponential Families: By the definition of $d(\theta)$, we have:

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

Evaluating this derivative and taking the derivative inside the integral in the numerator:

$$\frac{\int \exp[\theta^\mathsf{T} T(x)] T(x) h(x) d\nu(x)}{\int \exp[\theta^\mathsf{T} T(x)] h(x) d\nu(x)}$$

The denominator is equal to $e^{d(\theta)}$, we can therefore take it inside the exp:

$$\int \exp[\theta^\mathsf{T} T(x) - d(\theta)] T(x) h(x) d\nu(x)$$

Where we can use the definition of $p_\theta(x)$:

$$\int p_\theta(x) T(x) d\nu(x) = E_\theta T(X)$$

For the second derivative, we get by applying the quotient rule (to the first derivative):

$$\ddot{d}(\theta) = \frac{\int \exp[\theta^\mathsf{T} T] T T^\mathsf{T} h d\nu}{\int \exp[\theta^\mathsf{T} T] h d\nu} - \frac{(\int \exp[\theta^\mathsf{T} T] T h d\nu)(\int \exp[\theta^\mathsf{T} T] T h d\nu)^\mathsf{T}}{(\int \exp[\theta^\mathsf{T} T] h d\nu)^2}$$

Again using that the denominator is equal to $e^{d(\theta)}$ and $e^{d(\theta)} e^{d(\theta)}$:

$$\int \exp\left[\theta^\mathsf{T} T - d(\theta)\right] T T^\mathsf{T} h d\nu - \left(\int \exp\left[\theta^\mathsf{T} T - d(\theta)\right] T h d\nu\right) \times \left(\int \exp\left[\theta^\mathsf{T} T - d(\theta)\right] T^\mathsf{T} h d\nu\right)$$

Using the definition of $p_\theta(x)$:

$$\int T T^\mathsf{T} p_\theta d\nu - \left(\int T p_\theta d\nu\right)\left(\int T^\mathsf{T} p_\theta d\nu\right) = E_\theta T(X) T^\mathsf{T}(X) - (E_\theta T(X))\left(E_\theta T^\mathsf{T}(X)\right)$$

Which is the definition of $\mathrm{Cov}_\theta(T(X))$. □

Bias/Variance Decomposition: With $E_\theta := q(\theta)$

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

□

Neyman Pearson Lemma:

$$R(\theta_1, \phi_{\mathrm{NP}}) - R(\theta_1, \phi) = 1 - E_{\theta_1}\phi_{\mathrm{NP}}(X) - (1 - E_{\theta_1}\phi(X)) = E_{\theta_1}\phi(X) - E_{\theta_1}\phi_{\mathrm{NP}}(X)$$
$$= \int (\phi - \phi_{\mathrm{NP}}) p_1 = \int_{p_1/p_0 > c} (\phi - \phi_{\mathrm{NP}})\, p_1 + \int_{p_1/p_0 = c} (\phi - \phi_{\mathrm{NP}})\, p_1 + \int_{p_1/p_0 < c} (\phi - \phi_{\mathrm{NP}})\, p_1$$

When $p_1/p_0 > c$, $p_1 > c \times p_0$ and $\phi_{\mathrm{NP}} = 1$ per definition, therefore $(\phi - \phi_{\mathrm{NP}}) \leq 0$. On the other hand, when $p_1/p_0 < c$, $p_1 < c \times p_0$ and $\phi_{\mathrm{NP}} = 0$ per definition, therefore $(\phi - \phi_{\mathrm{NP}}) \geq 0$. Putting this together, we get:

$$\leq c \int_{p_1/p_0 > c} (\phi - \phi_{\mathrm{NP}})\, p_0 + c \int_{p_1/p_0 = c} (\phi - \phi_{\mathrm{NP}})\, p_0 + c \int_{p_1/p_0 < c} (\phi - \phi_{\mathrm{NP}})\, p_0$$
$$= c\left[R\left(\theta_0, \phi\right) - R\left(\theta_0, \phi_{\mathrm{NP}}\right)\right]$$

□

One Sided UMP Test: When we only consider testing $H_0 : \theta = c$, $H_1 : \theta = c_-$, it follows from the Neyman Pearson Lemma, that $\phi = \phi_{\mathrm{NP}}$ is the most powerful test at level α. Its power is $\beta(\theta_1)$ with $\beta(\theta) := E_\theta \phi(T)$, i.e. $E_{\theta_1}\phi(T)$.
The construction of the test ϕ is independent of the value c_- (as long as it is smaller), the test is therefore also uniformly most powerful for the alternative $H_1 : \theta < c$.
When $\theta_1 < \theta_0$, small values of T are more likely under P_{θ_1} than under P_{θ_0}. $\beta(\theta)$ is therefore a decreasing fucntion of θ. The level is per definition the sup for all $\theta \in \Theta_0$, therefore $\sup_{\theta \geq \frac{1}{2}} \beta(\theta) = \beta(\frac{1}{2}) = \alpha$. This implies that every other test ϕ' with level α has to have $\beta(\frac{1}{2}) = \alpha$, which makes ϕ UMP by the Neyman Pearson Lemma. □

Rao-Blackwell Lemma: By Jensen's inequality:

$$E(L(\theta, d(X)) \mid S = s) \geq L(\theta, E(d(X) \mid S = s)) = L\left(\theta, d'(s)\right)$$

Applying the iterated expectation lemma gives:

$$R(\theta, d) = E_\theta L(\theta, d(X)) = E_\theta E(L(\theta, d(X)) \mid S) \geq E_\theta L\left(\theta, d'(S)\right)$$

□

Basu's Lemma: Let A be some measurable set and define:

$$h(T) := P(Y \in A \mid T) - P(Y \in A)$$

Y does not depend on θ by assumption and $Y \mid T$ by sufficiency (if one of the variables would depend on θ, we could not apply the iterated expectation lemma for all θ). By the iterated expectation lemma:

$$E_\theta H(T) = E_\theta[E(\mathbb{1}_{Y \in A} \mid T) - P(Y \in A)] = P(Y \in A) - P(Y \in A) = 0,\ \forall \theta$$

The completeness of T now implies that $h(T) = 0$, P_θ-a.s., $\forall \theta$. Therefore for an arbitrary A:

$$P(Y \in A \mid T) = P(Y \in A),\ P_\theta - \text{a.s.},\ \forall \theta$$

□

Minimaxity: (i): When T is admissible, there is for all T' either a θ with $R(\theta, T') > R(T)$ or $R(\theta, T') \geq R(T)$ for all θ. Therefore, $sup_\theta R(\theta, T') \geq R(T)$.
(ii): For any T', Bayes risk is bounded by the supremum risk:

$$r_w\left(T'\right) = \int R\left(\vartheta, T'\right) w(\vartheta) d\mu(\theta) \leq \int \sup_\vartheta R\left(\vartheta, T'\right) w(\vartheta) d\mu(\theta) = \sup_\vartheta R\left(\vartheta, T'\right)$$

Assume that T' is a statistic with $sup_\theta R(\theta, T') < R(T)$. We then have

$$r_w\left(T'\right) \leq \sup_\vartheta R\left(\vartheta, T'\right) < R(T) = r_w(T)$$

which is a contradiction, as T is Bayes.
(iii): We assume that a Bayes decision T_m for the prior w_m exists for all m, i.e. $r_{w_m}\left(T_m\right) = \inf_{T'} r_{w_m}\left(T'\right), m = 1, 2, \ldots$. Because T is extended Bayes, for all $\epsilon > 0$, there exists an m such that:

$$R(T) = r_{w_m}(T) \leq r_{w_m}\left(T_m\right) + \epsilon \leq r_{w_m}\left(T'\right) + \epsilon \leq \sup_\theta R\left(\theta, T'\right) + \epsilon$$

Where we used again that Bayes risk is bounded by supremum risk. □

Admissibility: (i): Assume that for some T', $R(\theta, T') \leq R(\theta, T)$ for all θ. Then also $r_w(T') \leq r_w(T)$ and because T is Bayes, $r_w(T') = r_w(T)$. Because T is the unique Bayes decision per assumption, T' and T are equal P_θ-a.s. and therefore $R(\theta, T') = R(\theta, T)$, i.e. T' is not strictly better than T.
(ii): Suppose T is inadmissible, i.e. there is some T' such that $R(\theta, T') \leq R(\theta, T)$ for all θ and $R(\theta_0, T') < R(\theta_0, T)$ for some θ_0. This implies that for some $\epsilon > 0$ and some open neighborhood $U \subset \Theta$ of θ_0, $R\left(\vartheta, T'\right) \leq R(\vartheta, T) - \epsilon, \vartheta \in U$. Then:

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

Which is a contradiction as T is Bayes. □

Admissibility, Extended Bayes: As in the previous proof, we can arrive at $r_{w_m}\left(T'\right) \leq r_{w_m}(T) - \epsilon \Pi_m(U)$ which is no contradiction on its own because T is extended Bayes. If we assume that a Bayes decision T_m for the prior w_m exists for all m, i.e. $r_{w_m}\left(T_m\right) = \inf_{T'} r_{w_m}\left(T'\right), m = 1, 2, \ldots$, we have for all m:

$$r_{w_m}\left(T_m\right) \leq r_{w_m}\left(T'\right) \leq r_{w_m}(T) - \epsilon \Pi_m(U)$$

This can be rewritten to arrive again at a contradiction:

$$\frac{r_{w_m}(T) - r_{w_m}\left(T_m\right)}{\Pi_m(U)} \geq \epsilon > 0$$

□

Admissible Estimators of the Normal Mean: ($\Leftarrow$) (i): For quadratic loss, the Bayes estimator is $E(\theta \mid X)$. If we take $\theta \sim \mathcal{N}(c, \tau^2)$ as a prior, the Bayes estimator is therefore $\frac{\tau^2 X + c}{\tau^2+1}$. If we set $\frac{\tau^2}{\tau^2+1} = a$, $\frac{c}{\tau^2+1} = b$, T is Bayes for the normal prior. We need to show that T is unique such that the admissibility follows from (i) of the Admissibility Lemma: For quadratic loss and $T = E(\theta \mid X)$, we have $r_w(T') = E\operatorname{var}(\theta \mid X) + E(T - T')^2$. Therefore, if $r_w(T') = r_w(T) = E\operatorname{var}(\theta \mid X)$, we have $E(T-T')^2 = 0$, which implies $T = T'$ (P-a.s. which implies P_θ-a.s., where P is the measure with θ integrated out). Therefore, T is unique and admissible.
($\Leftarrow$) (ii): $T = X$, $R(\theta, T) = 1$ by the bias-variance decomposition and therefore $r_w(T) = 1$ for any prior. For $w_m \sim \mathcal{N}(0, m)$, the Bayes estimator is $T_m = \frac{m}{m+1}X$ and applying the bias-variance decomposition gives:

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

With $E\theta^2 = m$:

$$r_{w_m}(T_m) = ER(\theta, T_m) = \frac{m^2}{(m+1)^2} + \frac{m}{(m+1)^2} = \frac{m}{m+1}$$

Therefore $r_{w_m}(T) - r_{w_m}(T_m) = 1 - \frac{m}{m+1} = \frac{1}{m+1} \to 0$, i.e. T is extended Bayes. Furthermore, for m sufficiently large (considering open intervals $U = (u, u+h)$), $\Pi_m(U) \geq \frac{1}{4\sqrt{m}}h$ and therefore $\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \leq \frac{4}{h\sqrt{m}}$. This allows application of the Admissibility Lemma for extended Bayes estimators.
($\Rightarrow$): We have to show that if (i) or (ii) do not hold, T is not admissible. If (i) does not hold, $a > 1$, $R(\theta, aX+b) \geq \operatorname{var}(aX+b) > 1 = R(\theta, X)$ (using the bias-variance decomposition). If (ii) does not hold, $a = 1$ and $b \neq 0$, and we have $R(\theta, X+b) = 1 + b^2 > 1 = R(\theta, X)$. $\square$

Distribution of the Least Squares Estimator: i)

$$\hat{\beta} - \beta^* = \left(X^T X\right)^{-1} X^T (Y - f) = \underbrace{\left(X^T X\right)^{-1} X^T}_{:=A} \epsilon$$

Therefore, $E\left(\hat{\beta} - \beta^*\right) = AE\epsilon = 0$ and $\operatorname{Cov}(\hat{\beta}) = \operatorname{Cov}(A\epsilon) = A\operatorname{Cov}(\epsilon)A^T = \sigma^2 AA^T = \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$
ii) With the projection $PP^T := X(X^TX)^{-1}X^T$

$$\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = \left\|X(X^TX)^{-1}X^T\epsilon\right\|_2^2 = \left\|PP^T\epsilon\right\|_2^2 = (\epsilon^T PP^T)(PP^T\epsilon)$$

As projections are idempotent, this is equal to $\epsilon^T PP^T \epsilon = V^T V =: \sum_{j=1}^p V_j^2$ with $V := P^T\epsilon$, $EV = P^T E\epsilon = 0$ and $\operatorname{Cov}(V) = P^T \operatorname{Cov}(\epsilon) P = \sigma^2 I$ (where we used $P^TP = I$, which can be derived by the idempotence of the projection). Now we have: $E\sum_{j=1}^p V_j^2 = \sum_{j=1}^p EV_j^2 = \sigma^2 p$
iii) By Pythagoras' rule:

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

Since $X(\hat{b} - \beta^*) = PP^T\epsilon$ is in the column space of X and $X\beta^* - f = PP^Tf - f$ is orthogonal to the column space. $\square$

Least Square Estimator Expectation: i) As shown in the previous proof, $\hat{\beta} - \beta^*$ is a linear combination of ϵ and therefore also normally distributed if ϵ is.
ii) As in the previous proof,

$$\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = \sum_{j=1}^p V_j^2$$

with $V_j \sim \mathcal{N}(0, \sigma^2)$. $\square$

Slutsky's Theorem: Take a bounded Lipschitz function f:

$$|f| \leq C_B, |f(z) - f(\tilde{z})| \leq C_L\|z - \tilde{z}\|$$

By Cauchy Schwarz:

$$\begin{aligned}\left|\mathbb{E}f\left(A_n^T Z_n\right) - \mathbb{E}f\left(a^T Z\right)\right| &= \left|\mathbb{E}f\left(A_n^T Z_n\right) - \mathbb{E}f\left(a^T Z_n\right) + \mathbb{E}f\left(a^T Z_n\right) - \mathbb{E}f\left(a^T Z\right)\right| \\ &\leq \left|\mathbb{E}f\left(A_n^T Z_n\right) - \mathbb{E}f\left(a^T Z_n\right)\right| + \left|\mathbb{E}f\left(a^T Z_n\right) - \mathbb{E}f\left(a^T Z\right)\right|\end{aligned}$$

The function $z \mapsto f(a^t z)$ is bounded and Lipschitz (with constant $||a||C_L$), it therefore follows from the Portmanteau Theorem that the second term goes to zero. For the first term, we define $S_n := \{\|Z_n\| \leq M, \|A_n - a\| \leq \epsilon\}$ and apply Jensen's inequality:

$$\begin{aligned}\left|\mathbb{E}f\left(A_n^T Z_n\right) - \mathbb{E}f\left(a^T Z_n\right)\right| &\leq \mathbb{E}\left|f\left(A_n^T Z_n\right) - f\left(a^T Z_n\right)\right| \\ = \mathbb{E}\left|f\left(A_n^T Z_n\right) - f\left(a^T Z_n\right)\right| \mathbb{1}\{S_n\} &+ \mathbb{E}\left|f\left(A_n^T Z_n\right) - f\left(a^T Z_n\right)\right| \mathbb{1}\{S_n^c\} \\ &\leq C_L\epsilon M + 2C_B\mathbb{P}\left(S_n^c\right)\end{aligned}$$

We have $\mathbb{P}\left(S_n^c\right) \leq \mathbb{P}\left(\|Z_n\| > M\right) + \mathbb{P}\left(\|A_n - a\| > \epsilon\right)$ and can therefore make both terms arbitrary small by choosing ϵ small and n and M large. $\square$

Consistency of M-Estimators: As the theoretical risk is minimized at γ, we have $P_\theta(\rho_{\hat{\gamma}_n} - \rho_\gamma) \geq 0$. Subtracting and adding $\hat{P}_n(\rho_{\hat{\gamma}_n} - \rho_\gamma)$, we get:

$$0 \leq P_\theta(\rho_{\hat{\gamma}_n} - \rho_\gamma) = -\left(\hat{P}_n - P_\theta\right)(\rho_{\hat{\gamma}_n} - \rho_\gamma) + \hat{P}_n(\rho_{\hat{\gamma}_n} - \rho_\gamma)$$

As $\hat{\gamma}_n$ minimizes the empirical risk, $\hat{P}_n(\rho_{\hat{\gamma}_n} - \rho_\gamma)$ is negative, therefore:

$$\leq -\left(\hat{P}_n - P_\theta\right)(\rho_{\hat{\gamma}_n} - \rho_\gamma) \leq \left|\left(\hat{P}_n - P_\theta\right)\rho_{\hat{\gamma}_n}\right| + \left|\left(\hat{P}_n - P_\theta\right)\rho_\gamma\right|$$

Both terms are smaller than the supremum, so we get:

$$\leq \sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right| + \left|\left(\hat{P}_n - P_\theta\right)\rho_\gamma\right| \leq 2\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right|$$

Which goes to 0 by assumption, i.e. $0 \leq P_\theta(\rho_{\hat{\gamma}_n} - \rho_\gamma) \leq 0$. □

Consistency of a one-dimensional Z-Estimator: Let $0 < \epsilon < \delta$ be arbitrary. By the law of large numbers, $\mathbb{P}_\theta$-a.s. for n sufficiently large:

$$\dot{\hat{\mathcal{R}}}_n(\gamma + \epsilon) > 0, \quad \dot{\hat{\mathcal{R}}}_n(\gamma - \epsilon) < 0$$

Because of the continuity of $c \mapsto \psi_c$, $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$ for some $|\hat{\gamma}_n - \gamma| < \epsilon$, i.e. $\hat{\gamma}_n$ gets arbitrarily close to γ. □

Asymptotic Linearity of Z-Estimators: We have by definition $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0, \dot{\mathcal{R}}(\gamma) = 0$. Using the definition of ν_n, we arrive at:

$$0 = \dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = \nu_n(\hat{\gamma}_n)/\sqrt{n} + \dot{\mathcal{R}}(\hat{\gamma}_n) = \nu_n(\hat{\gamma}_n)/\sqrt{n} + \dot{\mathcal{R}}(\hat{\gamma}_n) - \dot{\mathcal{R}}(\gamma)$$

Because $\hat{\gamma}_n \to \gamma$, we can use the asymptotic continuity and definition of ν_n:

$$\nu_n(\hat{\gamma}_n)/\sqrt{n} = \nu_n(\gamma)/\sqrt{n} + o_{\mathbf{P}_\theta}(1/\sqrt{n}) = \dot{\hat{\mathcal{R}}}_n(\gamma) + \dot{R}(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) = \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

We assume $\hat{R}(c)$ is differentiable at $c = \gamma$ and can therefore perform a Taylor expansion:

$$\dot{\mathcal{R}}(\hat{\gamma}_n) - \dot{\mathcal{R}}(\gamma) = M_\theta(\hat{\gamma}_n - \gamma) + o(\|\hat{\gamma}_n - \gamma\|)$$

Putting those two results together, we get:

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) + M_\theta(\hat{\gamma}_n - \gamma) + o(\|\hat{\gamma}_n - \gamma\|)$$

By the CLT, $\dot{\hat{\mathcal{R}}}_n(\gamma) = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$ and we have $-\dot{\hat{\mathcal{R}}}_n(\gamma) = o_{\mathbf{P}_\theta}(1/\sqrt{n}) + M_\theta(\hat{\gamma}_n - \gamma) + o(\|\hat{\gamma}_n - \gamma\|)$. $o_{\mathbf{P}_\theta}(1/\sqrt{n})$ and $o(\|\hat{\gamma}_n - \gamma\|)$ is negligible compared to $M_\theta(\hat{\gamma}_n - \gamma)$ (which is $\mathcal{O}_{\mathbf{P}_\theta}(\|\hat{\gamma}_n - \gamma\|)$ as division by $\|\hat{\gamma}_n - \gamma\|$ results in a constant). Therefore, we can conclude that $\|\hat{\gamma}_n - \gamma\| = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$ (otherwise the equality would not hold).
Now that we know the rate of $\|\hat{\gamma}_n - \gamma\|$, we can conclude:

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

Which we can rewrite (using $\dot{\hat{\mathcal{R}}}_n(\gamma) = \hat{P}_n\psi_\gamma$ as:

$$(\hat{\gamma}_n - \gamma) = -\hat{P}_n M^{-1}\psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

□

Conditions for Uniform Convergence: We define for $\delta > 0$ and $c \in \Gamma$:

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

Because of the continuity of ρ_c, for all x as $\delta \to 0$, $w(x, \delta, c) \to 0$. The dominated convergence theorem implies $P_\theta w(\cdot, \delta, c) \to 0$, therefore there is for all $\epsilon > 0$ a δ_c such that $P_\theta w(\cdot, \delta_c, c) \leq \epsilon$. We define balls around c as follows:

$$B_c := \{\tilde{c} \in \Gamma : \|\tilde{c} - c\| < \delta_c\}$$

Because of the compactness of Γ, there exists finite sub-covering $B_{c_1}, \ldots, B_{c_N}$. By definition of w, δ_{c_j}, and B_{c_j}, we have:

$$\left|\rho_c - \rho_{c_j}\right| \leq w\left(\cdot, \delta_{c_j}, c_j\right)$$

We can now bound the supremum of the empirical/theoretical difference by the maximal difference of a ball centroid and the maximal difference of points to their centroid (i.e. $\left|\rho_c - \rho_{c_j}\right|$, which is bounded by w):

$$\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right| \leq \max_{1\leq j\leq N}\left|\left(\hat{P}_n - P_\theta\right)\rho_{c_j}\right| + \max_{1\leq j\leq N}\hat{P}_n w\left(\cdot, \delta_{c_j}, c_j\right) + \max_{1\leq j\leq N} P_\theta w\left(\cdot, \delta_{c_j}, c_j\right)$$

Because we are considering a finite number of balls, the first term converges to 0 by the law of large numbers and the second to $P_\theta w\left(\cdot, \delta_{c_j}, c_j\right)$. So the whole term converges to

$$2\max_{1\leq j\leq N} P_\theta w\left(\cdot, \delta_{c_j}, c_j\right) \leq 2\epsilon, \mathbb{P}_\theta - \text{a.s.}.$$

□