Official Cheat Sheet, Inference Qual, Gonzalo Mena

Normal Distribution Facts

Normal Distribution: $\Sigma = U\Lambda U^t \to \Sigma^{1/2} = U\Lambda^{1/2}U^t$

$$f(x) = det(2\pi\Sigma)^{-p/2}\exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right)$$

Conditional Normal: For

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^t & \Sigma_{22} \end{pmatrix})$$

then
$X_2|X_1 = x \sim \mathcal{N}\left(\mu_2 + \Sigma_{12}^t\Sigma_{11}^{-1}(x-\mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$
Information Matrix ($\mathcal{N}(\mu, \sigma)$

$$I = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{bmatrix}$$

Other Distributions

T with n degrees of freedom, $N(0,1)/\sqrt{\chi_n^2/n}$. Density

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

F distribution (n,m) $\chi^2 n/n/\chi_m^2/m$, mean $m/(m-2)$ Density

$$\frac{\sqrt{\frac{(nx)^n m^n}{(mx+n)^{m+n}}}}{xB(\frac{n}{2},\frac{m}{2})}$$

Beta Distribution, Mean $\alpha/(\alpha+\beta)$ density

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Gamma distribution: mean α/β, variance α/β^2 //
$X \sim \chi_{k_1}^2, Y \sim \chi_{k_2}^2$ then $X/(X+Y) \sim Beta(k_1, k_2)$

Order Statistics

Joint density $f(u_1, \dots u_n) = n!1_{u_1 \le u_2 \le \dots u_n}\prod_{i=1}^n f(u_i)$
Marginal of r-th order statistic

$$f_r(u) = nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(1-F(u))^{n-r}$$

Joint of r,s-th order statistics $f_{r,s}(u,v)$ (for $u \le v$)
Minimum and maximum:

$$f_{1,n}(u,v) = n(n-1)F(v)-F(u))^{n-2}f(u)f(v)$$

CDF of range: $W = X_{(n)} - X_{(1)}$ then

$$F_W(w) = n\int_{-\infty}^{\infty}(F(x+w)-F(x))^{n-1}f(x)dx$$

$$nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(n-r)f(v)\binom{n-r-1}{n-s}\times$$
$$(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}$$

Example: in a Uniform$(0,\theta)$ context, it is easily seen that

$$f(x_{(1)}, \dots x_{(n-1)}|x_{(n)}) = \frac{(n-1)!}{x_{(n)}}1_{0<x_{(1)}<\dots<x_{(n)}}$$

From that, $X_1, \dots X_n \backslash X_{(n)}$ is uniform on $[0, X_{(n)}]$
Example: $Y_i \sim Exp(1)$ and $S_i \sim_{j=1}^i Y_j$. Then $(S_1/S_{n+1}, \dots S_n/S_{n+1})$ given S_{n+1} is the same as the order statistics of the uniform. Use that to prove the consistency of the order statistics $k = \alpha n$ of the uniform. For this order statistics, if $k_1, k2 \to \infty$, and

$$\sqrt{n}(\frac{k_1}{n} - \alpha_1) \to 0, \sqrt{n}(\frac{k_2}{n} - \alpha_2) \to 0$$

We have

$$\sqrt{n}\begin{pmatrix} U_{(k_1)} \\ U_{(k_2)} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \xrightarrow{d} N\left((0, \begin{bmatrix} \alpha_1(1-\alpha_1) & \alpha_1(1-\alpha_2) \\ \alpha_1(1-\alpha_2) & \alpha_2(1-\alpha_2) \end{bmatrix}\right)$$

To prove the above, define
$(Z_1, Z_2, Z_3) = (S_{k_1}, S_{k_2} - S_{k_1}, S_{n+1} - S_{k_2})/(n+1)$ and show that $\sqrt{n+1}(Z - (\alpha_1, \alpha_2 - \alpha_1, 1-\alpha_2)) \to N(0, \Sigma)$ where $\Sigma = Diag(\alpha_1, \alpha_2 - \alpha_1, 1-\alpha_2)$ and finally use delta method with $g(x_1, x_2, x_3) = \frac{1}{x_1+x_2+x_3}(x_1, x_1+x_2)^t$
Using this we get asymptotics for the sample median!.
$\sqrt{n}(X_{(n/2)} - m) \to N(1/4f^2(m))$

Convex functions

Local minima of convex functions are global minima. If they are strictly convex the minimum is unique.
If g is diff then it is convex if and only if (replace $\le$ by $<$ to obtain strict convexity)

$$g(y) \ge g(x) + \nabla g(x)^t(y-x) \forall y, x$$

Two times diff g it is convex iff $g''(y) \ge 0$ if $g''(y) > 0$ then it is strictly convex Jensen: $E(|X|) < \infty$ g convex then $g(E(X)) \le E(g(X))$. If g is strictly convex then inequality is strict unless $X = a$ a.s. Equality holds, for all X iff g is affine.
Convex functions are continuous
$f : \mathbb{R}^k \to \mathbb{R}$ is convex if and only if the function $g : \mathbb{R} \to \mathbb{R}$ is convex for every $x \in Dom(f)$ and $v \in \mathbb{R}^n$.
Fact: the function $I(f) = \int f'^2(y)/f(y)dy$ is convex. To prove that use

$$\frac{(\alpha u + (1-\alpha)v)^2}{\alpha f + (1-\alpha)g} \le \frac{\alpha u^2}{f} + \frac{(1-\alpha)v^2}{g}$$

Loss functions

Setting: $X \sim P_\theta$, want to estimate $g(\theta), g : \mathcal{E} \to G$ by $\hat{g}(X)$
Def: Loss function $l : \mathcal{G} \times \mathcal{G} \to R$ is a loss function if

$$l(\hat{g}, g(\theta) \ge 0, \forall\theta\forall\hat{g} \quad l(g(\theta), g(\theta) = 0)\forall\theta$$

Examples of convex loss
$l(x,y) = |x-y|, |x-y|^{1/2}, (x-y)^2, 1_{|x-y|>1}$ Def: δ_1 is inadmissible for $g(\theta)$ if there is another estimator of $g(\theta), \delta_2$, s.t. the following two conditions hold
$R(\delta, g(\theta)) = E_\theta(l(\delta, g(\theta)))$

$$R(\delta_2, g(\theta)) \le R(\delta_1, g(\theta), \forall\theta \quad \exists\theta_0 \;\; R(\delta_2, g(\theta_0)) < R(\delta_1, g(\theta_0)$$

Sufficiency & min sufficiency, definitions and properties

The family is: $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ Def: A function T of $X \sim P_\theta$ is sufficient for theta if $X|T = t$ don't depend on θ.
Factorization theorem: If all the P are dominated by the same measure ν and let p_θ denote the pdf. Then T is sufficient iff $p_\theta(X) = g_\theta(T(X))h(X)$ a.e. w/r to ν T is minimal sufficient for θ iff T is sufficient and it is a function of any other sufficient statistic. That is, if U is other suff statistic, there exists g measurable such that $P_\theta(T \ne g(U)) = 0\forall\theta$
Proposition: If $|\Theta| = s$ then $T(X) = \left(\frac{p_{\theta_1}(X)}{p_{\theta_0}(X)}, \dots, \frac{p_{\theta_s}(X)}{p_{\theta_0}(X)}\right)$ is a minimal suff stat for θ. Prove that for any S sufficient $S(x) = S(y)$ implies $T(x) = T(y)$
T sufficient. If $p_\theta(x) \propto_\theta p_\theta(y) \Rightarrow T(x) = T(y)$ (doesn't depend on θ) then T is minimal sufficient

Completeness

T is complete iff $E_\theta(f(T)) = 0\forall\ \theta \ \Rightarrow f(T) = 0 a.e., \forall p \in \mathcal{P}$
Basu: T complete and V ancillary, then $T \perp V$ Theorem: Complete sufficiency $\Rightarrow$ minimal sufficiency
Claim: complete sufficient does not always exists: if the minimal is not complete then no one can be complete
Let P_θ be a dominated family. Let T be MSS and U CSS. Then there exists g such that $U = g(T)$. Take $g_\theta(T) = E_\theta(U|T) = g_{\theta_0}(T)$

Theorems with sufficiency,completeness and Risk

Theorem: for any $\delta(X)$, if T is suff for $g(\theta)$ there is another estimator δ_1 based on T with the same risk: draw X, find $t = T(X)$ and draw a sample $\hat{X} \sim X|T = t$. define $\delta_2(X) = \delta_1(\hat{X}$
Theorem (Rao Blackwell): T sufficient, δ an estimator with $E(|\delta|) < \infty$ l convex function. Define $\nu(t) = E(\delta(X)|T = t)$. Then $R(\delta, g(\theta)) \ge E(\nu, g(\theta)$. If l is strictly convex then inequality is strict unless $P_\theta(\delta(x) = \nu(t)) = 1, \forall\theta$

Sufficiency and completeness in nonparametric families

Def: a family of distributions have common support iff $\mu(A) = 0 \iff \nu(A) = 0$ Let $\mathcal{P}$ be a family of distributions with common support and let $\mathcal{P}_0 \subseteq P$. If T is sufficient for $\mathcal{P}$ and minimal for $\mathcal{P}_0$ then T is minimal sufficient for P
Theorem: if all distributions in a family of distributions P are dominated by another measure, then minimal sufficient statistic exists.
If $\mathcal{P}_0, \mathcal{P}_1$ are two families of distributions, $\mathcal{P}_0 \subseteq \mathcal{P}_1$ and every null set of $\mathcal{P}_0$ is a null set of $\mathcal{P}_1$, then a sufficient statistic T that is complete for $\mathcal{P}_0$ is also complete for $\mathcal{P}_1$. The order statistics are complete for the family of distributions with continuous CDF: they are sufficient and consider $\mathcal{P}_0 = \{C(\theta)\exp(\sum_{i=1}^n \theta_i \sum_{j=1}^n x_j^i - \sum_{i=1}^n x_i^{2n}$. Then show that this statistic is equivalent to the sums of the j products. Then construct polynomial $\Pi(\delta - X_i)$

For $(X_(i), Y_{j_i})$ is complete sufficient for F. To prove consider the exponential family

$$\exp\left(\sum_{j=1}^{n}\alpha_j\sum_{i=1}^{n}x_i^j+\sum_{j=0}^{n-1}\beta_j\sum_{i=1}^{n}y_ix_i^j-\sum_{i=1}^{n}x_i^{2n}-\sum_{i=1}^{n}y_i^{2n}\right)$$

Then express the mixed products in terms of the vandermonde (invertible) matrix and the Y_{ji}.

Examples on sufficiency, completeness and UMVUE

for $X_i \sim U[\theta-1/2, \theta+1/2]$ both $(X_(1), X_(n))$ and $(X_(1), X_(n)-X(1))$ are minimal sufficient for θ. However, there is not CSS $(X_(n)-X(1)$ is ancilliary)
To prove $\sum X_i$ is min sufficient for Poisson(λ) use the $Po(1), Po(2)$

Fisher,Pitman,Koopman,Darmois

Let $X_1, \dots X_n P_\theta$ and there is a k dimensional sufficient statistic, in a family where the support doesn't vary with the density, and continuous in x. For a sample of size $k<n$. Then the family HAS to be an exponential family.
Completeness and sufficiency are preserved by one to one maps!.
UMVUE estimation for σ^r, $X_n \sim N(0,\sigma^2)$. Take $T=\sum X_i^2, Y=T/\sigma^2 \sim \chi^2_n$ and find $E(Y^{r/2})=K(r)\sigma^r, r/2>-n/2+1$. For smaller values there is not UMVUE estimator. Find $E(g(T))=E(g(\sigma^2 Y)$ and make the change $x=\sigma^2 y$

Exponential Families

$$p_\theta(X_1,\dots X_n)=h(X)\exp\left(\sum_{i=1}^{k}\eta(\theta)T_i(X)-B(\theta)\right)$$

Natural parameter space is the set of θ such that

$$B(\theta)=\log\left(\int h(x)\exp\left(\sum_{i=1}^{k}\eta(\theta)T_i(x)\right)d\mu(x)\right)<\infty$$

Theorem: if the convex hull of the parameter space spans a rectangle of dimension k then T is minimal sufficient. (Proof: take likelihood ratios and get $T=AU$ where A is invertible and U is the minimal sufficient. This case corresponds where the parameters satisfies linear constraints. If the interior is nonempty then T is complete. Proof: characteristic function (need to define it in an open set)
Now consider the natural parametrization, η and call $A(\eta)=B(\eta(\theta))$. Prop: $A(\eta)$ is convex.
Prop $H_A(\eta)=cov_\eta(T), \nabla(A)=E_\eta(T)$ (proof just take derivatives of the integral of the density and make it equal zero) Maximum likelihood: $\nabla p_\eta(x)=0 \iff \nabla A(\eta)=T$. If there are no linear constrains between the T there is an unique solution (from above).

$$\log M_T(u)=A(n+u)-A(n)$$

If X comes from an exponential family then T has density $q_\theta(t)=\eta(\theta)\cdot t-B(\theta)$ with respect to $\nu(A)=\int_A \exp(-\eta(\theta_0)t+B(\theta_0)d\nu^*(t), \nu^*$ is the marginal for $\theta=\theta_0$

Unbiased Estimation and UMVU

Def: δ is UMVU if it is unbiased and for all unbiased δ_2

$$Var_\theta(\delta)\le Var_\theta(\delta_2)\forall\theta$$

Theorem:Lehman-Scheffe T is complete sufficient, l is convex in δ. For every unbiased estimable $g(\theta)$ there exists δ that minimizes the risk for every θ. Furthermore, if l is strictly convex the best unbiased estimator is unique (in any case, it is given by $E(\delta_1(x)|T)$ where δ_1 is any unbiased estimator.
Theorem (Bahadur): If every $g(\theta)$ admits UMVU, then a complete sufficient statistic exists
If a complete sufficient statistic does not exist, there is at least one U-estimable of $g(\theta)$ for which there is no UMVU They are not admissible generally!
T, unbiased estimator of $g(\theta)$ is the UMVUE Iff for all unbiased estimator of zero $U, E_\theta(UT)=0, \forall\theta$. Proof: consider estimators $\delta+\Delta U$ and analyze Δ. We can take only functions of the sufficient statistic!.

Unbiased Estimation:Examples

No unbiased estimator: $X\sim Binomial(n,p)$ Then no unbiased estimator exists for $1/p$

$$\sum_{i=1}^{n}\delta(i)\binom{n}{i}p^i(1-p)^{n-i}=\frac{1}{p}$$

Binomial: estimating $p(1-p)$ in the binomial. Consider $\rho=p/(1-p), \rho+1=1/(1-p)$ then from the equality

$$\sum_{i=0}^{n}\delta(x)\binom{n}{i}\rho^i=(\rho+1)^{n-2}\rho$$

we obtain $\delta(T)=\frac{T(n-T)}{n(n-1)}$ Unbiased estimator with zero risk for θ_0. Take an unbiased estimator δ and

$$\delta_\pi(x)=\begin{cases}g(\theta_0) & \text{with probability } 1-\pi\\ \frac{1}{\pi}(\delta(x)-g(\theta_0))+g(\theta_0) & \text{with probability } \pi\end{cases}$$

Has zero risk at θ_0 (as $\pi\Rightarrow 0$) and it is unbiased.
NO UMVUE exists for the common mean μ two normal family problem with common mean. The reason is that there is a UMVUE, function of the CSS, if the ratio of the variances is known. The UMVUE is unique, and a function of the ratio and thus cannot exist when the ratio is unknown.

Fisher Information

For θ_0 being the true parameter, define $G(\theta)=E(\theta_0)(\log(p_\theta(x)))$. By Jensen, $G(\theta)-G(\theta_0)\le 0$, that is, G is minimized at θ_0
Fisher Information: Θ open, all the densities are supported in $A=\{x: p_\theta(x)>0, \forall\theta\}$. p_θ is differentiable w/r to θ and can swap integrals with derivatives. Then

$$I(\theta)=\int\left(\frac{\partial\log(p_\theta(x)}{\partial\theta}\right)^2dx=-\int\frac{\partial^2\log(p_\theta(x)}{\partial\theta^2}dx$$

Multiparameter

$$I(\theta)=-E_\theta(H(\log(p_\theta(x))=E_\theta(\nabla\log(p_\theta(x))\nabla\log(p_\theta(x))^t$$

X, Y independent, the Information of (X,Y) is the sum of the informations. The fisher information of the sufficient statistic $T(x)$ is the same as of X. If $Y=g(X)$ then the information of Y is less or equal than the information of X. To prove it , $\log p_\theta(x)=\log p_\theta(x|g(x))+\log p_\theta(g(x))$, use the variance characterization and prove that $E_\theta(\frac{\partial\log p_\theta(X|Y)}{\partial\theta}\frac{\partial\log p_\theta(Y)}{\partial\theta}$
For location families,

$$I(\theta)=\int frac f'(x)^2 f(x)dx$$

Change of parameter: $x\sim p_\theta(x), \theta=f(\mu)$ Denote $I(\mu), I(\theta)$ the relative mutual informations, then $I(\mu)=JI(\theta)J^t$

Cramer Rao bound

Suppose p_θ is a family of densities for which information is defined. Let δ be an statistic with $E_\theta(\delta^2)<\infty$ and such that $E_\theta(\delta)$ can be differentiated with respect to θ and we can swap integrals. Then

$$Var_\theta(\delta)\ge\frac{\left(\frac{dE_\theta(\delta)}{d\theta}\right)^2}{I(\theta)}$$

The inequality is equality (by cauchy schwarts) only if

$$\delta(x)=a(\theta)\log(p_\theta(x))+b(\theta)$$

, that is, if they come from a exponential family. If δ is an unbiased estimator of $g(x)$ we obtain $b(\theta)=g(\theta)$ and $a(\theta)=\nabla E_\theta(\delta)^t I(\theta)^{-1}$ Multidimensional case

$$Var_\theta(\delta)=\nabla E_\theta(\delta)^t I(\theta)^{-1}\nabla E_\theta(\delta)$$

It comes from

$$Var_\theta(\delta)=\frac{b^t\nabla E_\theta(\delta)^t\nabla E_\theta(\delta)b}{b^T(\theta)b}$$

If there are n unbiased estimators of $g_i(\theta)$ and all of them achieve the CR bound, and $\nabla g_i(\theta)$ are linearly independent, then they come from an exponential family.

Linear Algebra

A symmetric,

$$A=\sum_{i=1}^{n}\lambda_i q_i q_i^t=Q\Lambda Q^t, Q^tQ=QQ^t=I$$

$$\sup_{b^t:b^tb\neq 0}\frac{b^tAb}{b^tb}=\lambda_{\max}(A)$$

$$\inf_{b^t:b^tb\neq 0}\frac{b^tAb}{b^tb}=\lambda_{\min}(A)$$

Rayleigh Theorem: $R_A(x)=\frac{x^tAx}{x^tx}$

$$\lambda_k=\max\{\min\{R_A(x)|x\in U, x\neq 0\}dim(U)=k\}$$

$$\lambda_k=\min\{\max\{R_A(x)|x\in U, x\neq 0\}dim(U)=n-k+1\}$$

Inadmissibility, James Stein

If $X_i \mathcal{N}(\theta, I)$ in p dimensions. Then $\bar{X}$ is UMVUE, but not admissible. In fact, take

$$\sigma_{JS}(X) = \left(1 - \frac{p-2}{n\bar{X}^t\bar{X}}\right)\bar{X}$$

Theorem: σ_{JS} has smaller risk than $barX$ for all θ (whose risk is p/n). Its risk is given by

$$R(\delta_{JS}, \theta) = E_\theta(||\bar{X} - \theta||^2) - \frac{(p-2)^2}{n^2} E_\theta\left(\frac{1}{\bar{X}^t\bar{X}}\right)$$

The estimator

$$\delta_2(X) = \max\left(0, 1 - \frac{p-2}{X^tX}\right)X$$

Dominates the James Stein, use symmetry in the hemispheres!

Bayesian Framework

Minimize the integrated risk with prior $\Lambda(\theta)$

$$R(\delta) = \int R(\delta, \theta)d\Lambda(\theta) = \int\int l(\delta(x), \theta)p_\theta(x)d\Lambda(\theta)d\mu(x)$$

Under squared loss, $\delta_o pt = E(X|\theta)$ under
$l(\delta, \theta) = |X - \theta|, \delta_{opt}$ = Posterior median.
Under $l(\delta, \theta) = 0$ if $|\delta - \theta| < 0$ and 1 otherwise, the bayes estimator is given by δ such that maximizes the probability $P(|\delta - \theta|X)$ Let θ have pdf $\pi(\theta)$. Given $\theta, x \sim f(x|\theta)$ we want to estimate $g(\theta)$ with $l(\delta, g(\theta)$. If

i There exists δ_0 with finite bayesian risk

ii For almost all x with respect to $f(x) = \int f(x|\theta)\pi(\theta)$ there exists a value $\delta_\pi(x)$ minimizing $(E(l(\delta(x), g(\theta)|x)$.

Then δ_π is an optimal Bayes estimator. Theorem: If the loss is stricly convex and define the following distribution

$$Q(A) = \int_\Theta\int_A f(x|\theta)\pi(\theta)d\mu(x)d\theta$$

The if almost everywhere with respect to Q implies almost everywhere with respect to $f(x|\theta), \forall\theta \in \Theta$ then the Bayes estimator is unique (it seems uniqueness is defined almost sure, for $f(x|\theta)$, for all theta.
Def: A family F of probability distributions on Θ is called conjugate for $f(x|\theta)$ if and only if for all $\pi \in F$, it follows that the posterior $\pi(\theta|x) \in F$ Theorem: Any unique Bayes estimator is admissible.
If the risk function is continuous in θ and the prior are continuous, then the bayes is admissible. Also, if $f(x|\theta)$ is continuous and the loss is scrictly convex then the bayes estimator is unique.

Bayes Examples

$$X_i \sim \mathcal{N}(\theta, \sigma^2), \theta \sim \mathcal{N}(0, \tau^2)$$

Then for squared loss

$$\delta_{opt}(X) = \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{X}$$

Let $X \sim Binomial(\theta, n)$ with $P(\theta = 0) = ¶(\theta = 1) = 1/2)$. For squared loss then an optimal estimator assigns $\delta_\pi(0) = 0$ and $\delta_p i(n) = 1$ $X\theta \sim Binomial(n, \theta)$ $F = Beta(\alpha, \beta)$ then $\pi(\theta|x) \sim B(x + \alpha, n - x + \beta)$ $X\theta \sim N(\theta, \sigma^2)$ and $F = N(\mu, A)$ then $\pi(\theta|x) = \mathcal{N}(\mu + B(x - \mu), B\sigma_0^2), B = \frac{A}{A+\sigma_0^2}$

Kullback-Leiber Divergence

Suppose $p(x)$ and $q(x)$ are two pdf's on $\mathbb{R}^p$ with common supports, then the Killback-Liebler Divergence is defined as

$$KL(p, q) = \int p(x)\log\frac{p(x)}{q(x)}d\mu(x)$$

Props: $KL(p, q) \geq 0$ and $KL(p, q) = 0$ iff $p = q$, μ a.s.

$$\lim_{\theta' \to \theta}\frac{KL(p_\theta(x), p'_\theta(x))}{(\theta - \theta')^2} = \frac{1}{2}I(\theta)$$

Paralelogram law. p_0, p_1, Q and p the average of p_0, p_1

$$KL(p_0, Q) + KL(p_1, Q) = KL(p_0, p) + KL(p1, p) + 2KL(p, Q)$$

Jeffrey's Prior

Definition: π is called reference prior if it maximizes the expected KL divergence between itself and the posterior, that is

$$\pi = \arg\max S(\pi^*) = \int KL(\pi^*(\theta), \pi^*(\theta|x)f(x)d\mu(x)$$

Intuitively, it is a weak prior.
Theorem: Let X_1, X_n be iid samples from $f(x|\theta)$ Then for n large

$$S_n(\pi) \approx \frac{p}{2}\log(\frac{n}{2\pi e}) + \int \pi(\theta)\log(\frac{|I(\theta)^{1/2}}{\pi(\theta)}d\theta$$

where p is the dimension of the parameter and $I(\theta)$ the fisher information (of the whole sample) and $\approx$ means convergence of real numbers. Instead of maximzing S one can maximize S_n which leads to minimizing $KL(\pi, cI(\theta)^{1/2}$ That is the Jeffrey's prior. It may not exists (be improper).

Hierarchical Bayes

Framework:

$$x|\theta \sim f(x|theta), \theta|\lambda \sim \pi(\theta|\lambda), \lambda \sim \psi(\lambda)$$

Theorem

$$KL(\psi(\lambda|x), \psi(\lambda)) \leq KL(\pi(\theta|x)\pi(\theta))$$

Hint: define

$$Z = \frac{\psi(\lambda|x}{\psi(\lambda)} = E_{\pi(\theta|\lambda)}\frac{f(x|\theta)}{f(x)}, \quad g(Z) = Z\log(Z)$$

Bayes rule:

$$\frac{f(x|\theta)}{f(x)} = \frac{\pi(\theta|x)}{\pi(\theta)}$$

Empirical Bayes

Estimate the hyper parameters using maximum likelihood, UMVU or method of moments. Example:
$x \sim \mathcal{N}(\theta, \sigma^2 I), \theta\mathcal{N}(0, \tau^2)$. Then for τ fixed

$$\delta_{opt} = E(X|theta) = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Also, $f(x|\tau) = \mathcal{N}(0, (\tau^2 + \sigma^2)I)$

UMVUE for τ^2. is $\hat{\tau} = \frac{x^tx}{p} - \sigma^2$ we obtain $\delta(x) = (1 - \frac{p\sigma^2}{x^tx})x$
UMVUE for $\frac{\sigma^2}{\sigma^2+\tau^2}$ is $\frac{(p-2)\sigma^2}{x^tx}$ since $E(\frac{1}{x^tx}) = \frac{1}{(p-2)(\tau^2+\sigma^2}$.
We obtain the estimator $\delta(x) = \left(1 - \frac{(p-2)\sigma^2}{x^tx}\right)x$ It is the James Stein estimator!
ML for theta: $\hat{\tau} = \max(\frac{x^tx}{p} - \sigma^2, 0)$
MLE for $\frac{\sigma^2}{\sigma^2+\tau^2}$ is $\min(\frac{(p)\sigma^2}{x^tx}, 1)$

Improper priors

Define, in any case

$$\pi(\theta|x) = \frac{p_\theta(x)\pi(x)}{f(x)}$$

Example: Prior for θ lebesgue measure in $\mathbb{R}$, $X|\theta \sim \mathcal{N}(0, \theta)$ then $\delta(X) = X$ is the hayes estimator, even if $f(x)$ is not a pdf (but $\pi(\theta|x)$ is proper).

Numerical Integration, Monte Carlo and MCMC

If $[a, b]$ was partitioned into n subintervals

$$|\int [a, b]^d h(\theta)d\theta - \sum_{i=1}^n h(\theta_i)\delta^d| \leq \frac{c^*}{n^{1/d}}$$

Monte Carlo error $1/\sqrt{n}$ (from CLT) Gibbs Sampler: Draw from $p(\theta_i|\theta_{-i}, X)$

Minimax Estimation

Sup Risk $R(\delta) = \sup_\theta R(\delta, \theta)$. Minimax estimator

$$\delta^* = \arg\min_\delta R(\delta)$$

over a class of estimators (for example linear). $\inf_\delta R(\delta)$ is the minimax risk, denoted by $R(\Theta)$ Lemma:
$\sup_{\pi:supp(\pi)\subseteq\Theta} B(\pi) \leq R(\Theta)$ Definition:Least favorable prior: A least favorable prior π_0 in a class of priors $\mathcal{P}$ is the distribution such that $B(\pi_0) = \sup_{\pi\in\mathcal{P}} B(\pi)$
Definition: A least favorable sequence of priors in a class $\mathcal{P}$ is a sequence that satisfies $\lim B(\pi_n) = \sup_{\pi\in\mathcal{P}} B(\pi)$ Example: $X \sim \mathcal{N}(\theta, I)$ and $\mathcal{P} = \{\mathcal{N}(0, \sigma^2)$ for θ. Then $E(\theta|x) = \frac{\sigma^2}{1+\sigma^2}x$ and $B(\pi_\sigma^2) = \frac{p\sigma^2}{\sigma^2+1}$ Then $\pi_n \sim \mathcal{N}(0, n)$ is a least favorable sequence of priors. Theorem: let π_0 be a prior with Bayes estimator δ_{π_0}. Suppose $R(\delta_{\pi_0}) = B(\pi_0)$ then δ_{π_0} is the minimax estimator and π_0 is the least favorable prior. Furthermore, $\sup_\pi B(\pi) = R(\Theta)$
Extension: let pi_n a sequence and δ_0 such that $R(\delta_0) = \lim B(\pi_n)$ then δ_0 is minimax and pi_n is a l.f.s of priors.
Example: squared lost, $X \sim Binomial(n, \theta)$.
$\mathcal{P} = \{Beta(\alpha, \beta)\}$ then

$$E(\theta|X) = \frac{a + X}{a + b + n} \quad R(\delta_\pi, \theta) = \frac{n\theta(1-\theta) + (a(1-\theta) - b\theta)^2}{(a + b + n)^2}$$

We want $R(\delta_p i) = B(\pi)$. Then, assuming $R(\delta_p i, \theta)$ is continuous in θ $R(\delta_\pi, \theta)$ must be constant in θ (the support is (0,1)). Take $a = \frac{\sqrt{n}}{2}, b = \frac{\sqrt{n}}{2}$ and $R(\delta_\pi, \theta) = \frac{1}{4(\sqrt{n}+1)^2}$ and $\delta_{minimax}(X) = \frac{\sqrt{n}/2 + X}{\sqrt{n}+n}$
Example: $X\mathcal{N}(\theta, 1), |\theta| < \tau$. Suppose $R(\delta_{\pi_0}) = B(\pi_0)$ Since the risk is an analytic function, There are two options: $R(\pi, \theta)$ is constant or $R(\delta_\pi, \theta)$ takes its maximum in a finite number of points.
Theorem: for $|\tau| < 1.05$ $pi_\tau = 1/2\delta_\tau + 1/2\delta_{-\tau}$ is the least favorable prior Minimax is too pessimistic and conservative!
Example: $x \sim \mathcal{N}(\theta, 1)$ with $|\theta| < tau$. assume $\delta(x) = \alpha x$
Minimax $\delta(x) = \frac{\tau^2}{1+\tau^2} x$
Example: $x \sim \mathcal{N}(\theta, 1)$ no constrain on theta. Notice that $\delta(x) = x$ has constant risk equal to 1. Take $\pi_n \sim \mathcal{N}(0, n)$. Then $\delta_n(x) = \frac{1}{1+1/n} x$ and $B(\pi_n) = 1/(1 + 1/n) \to 1$. Thus, $\delta(X) = x$ is minimax. Question: Under which circumstances $\sup_\pi B(\pi) = R(\Theta)$
Theorem: Let $\mathcal{P}$ be a set of prior distributions whose support is a subset of Θ and contains all the δ_{θ_0}. Also, assume that

$$\sup_{\pi \in \mathcal{P}} \inf_{\delta} B(\delta, \pi) = \inf_{\delta} \sup_{\pi \in \mathcal{P}}$$

Then the minimax risk satisfies $R(\Theta) = \sup_{\pi \in \mathcal{P}} B(\pi)$ (Use point masses at the supremum of the risk function for an arbitrary estimator).
Theorem: We can swap infimum and supremum and it is fine. Suppose $f : K \times L \to \setminus$ where K, L are subsets of a metric space. Furthermore, assume f is convex in its first argument and i concave in its second argument. Assume that K is convex and compact, L is convex. Also, assume f is continuous in the first argument. Then

$$\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y)$$

Application: Suppose for each θ, $l(\delta, g(\theta)$ is convex and continuos in δ Assume that the class of estimators is compact and convex. Then We can use the minimax framework!. It seems that for the compactness of the estimator it is enough compactness of the parameter space.
Extension to several dimensions:$\pi(\theta) = \pi_1(\theta_1) \ldots \times \pi_n(\theta_n)$. For squared loss the jth component of the Bayes estimator, is a function only of j. Use jensen!.
For the problem $X \sim \mathcal{N}(\theta, I_n)$ $\theta \in [-\tau, \tau]^n$ there least favorable prior exists and it is a product of priors. For the condition $||\theta||_2 \leq \tau$ then the prior is uniform on the sphere.
Let X be a random variable with distribution F and let $g(F)$ be a function defined over a family $mathcalF_1$. Suppose δ is the minimax of $g(F)$ when defined over the family $\mathcal{F}_0 \subseteq \mathcal{F}_1$. Then if

$$\sup_{F \in \mathcal{F}_0} R(F, \delta) = \sup_{F \in \mathcal{F}_1} R(F, \delta)$$

δ is also minimax when defined on $\mathcal{F}_1$.
If the hayes estimator is unique then the minimax estimator is unique (in the minimax sense).

Analytic functions

Theorem: Let $f : (a, b) \to \mathbb{R}$ be analytic. If there is an accumulation point in the set of zeroes of f in (a, b) then f is zero in (a, b)
Theorem: the set of zeros of an analytic function is countable.

Stein Lemma

$Z_1, Z_2 \sim \mathcal{N}(0, \sigma^2)$ independent and $g : \mathbb{R}^2 \to \mathbb{R}$ differentiable in the first variable. If the expectations are finite then

$$E(Z_1 g(Z_1, Z_2)) = \sigma^2 E(\frac{\partial g(Z_1, Z_2)}{\partial Z_1})$$

If Z_1, Z_2 are not independent (not zero mean either) then

$$Cov(g(Z_1), Y) = E(g'(X)) Cov(X, Y)$$

$E(\frac{1}{x^t x}) = \frac{1}{n-2}$ from stein lemma and using sum properties if x is indep multivariate normal.

Regression

$Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I), X \in \mathbb{R}^{nxp}$ is full rank, $n \geq p$. Then β_{MLE} and $||Y - X\beta_{MLE}||^2$ are complete sufficient for β, σ^2.
Express the exponential family! (the complete sufficient is $X^t Y, Y^T Y$ but the rest is obtained from this. Furthermore, these two are independent by ancilliarity. β_{ML} is the UMVU for β and $||Y - X\beta_{MLE}||^2/(n-p)$ is the UMVU for σ

Convergence in probability, distribution and almost sure

$X_i(t) \to X$ if $F_i(t) \to F(t)$ for all continuity points of F
Let $X_i \sim f_i(x)$ if $f_i(x) \sim f(x)$ then we have convergence in distribution.
$X_i \to X$ in distribution iif $\langle t, X_n \rangle \to \langle t, X \rangle$ for all t
$X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y$ then $X_n Y_n \to XY, X_n/Y_n \to X/Y$
Continuous mapping: under a.s., probability or distribution convergence, if g is continuous in K and $P(X \in K) = 1$ then $g(X_n) \to g(X)$
$X_n \xrightarrow{d} X$ iff $E(g(X_n)) \to E(g(X))$ for all g continuous and bounded. Also (Levy) iff the characteristic function converges.
If $\limsup E(|X_n|^p) < \infty$ for $p > 1$ then the moments converge until $p - 1$
Denote $m_k = E(x^k)$. If the series $\sum_{i=1}^{\infty} m_k r^k/k!$ has a positive radius of convergence then the moments of X characterize the distribution. If that is the case we have $E(X_n^k) \to E(X^k) < \infty$ for all k implies $X_n \to X$ in distribution

Convergence notations

$$a_n = O(b_n) \Leftrightarrow |a_n|/|b_n| < M \ \ \forall n \geq n_0$$

$$a_n = o(b_n) \Leftrightarrow |a_n|/|b_n| \to 0$$

h, g functions Then $h(x) = O(g(X))$ if there are δ, M such that if $|x| < \delta$ then $|h(x)/g(x)| < M$. Also, $h(x) = o(g(x))$ if $h(x)/g(x) \to 0$ when $x \to 0$ If f is differentiable at θ with derivatives continuous then

$$||f(\theta + h) - f(\theta) - f'(\theta)h|| = o(||h||)$$

For the others use taylor!
Stochastic orders $X_n = o_o(Y_n) \Leftrightarrow X_n/Y_n \to 0$ in probability.
$X_n = O(Y_n)$ if $\forall \epsilon > 0 \exists M$ such that $P(|X_n/Y_n| > M) \leq \epsilon$ for all N (it is enough for all $n \geq n_0$. If $X_n \to X$ in distribution then $X_n = O_p(1)$. $X_n = O_p(1/\sqrt{n}) \to X_n = o_p(1)$
If $\lim P(|X_n| > M) = 0$ for some M then $X_n = O_p(1)$
If $R(0) = 0$ then if $X_n \to 0$ in probability we have: $R(h) = o_p(||h||^q) \to R(X_n) = o_p(||X_n||^q)$, and the same happens with O.
If for some $k \geq q, \sup_n E(|X_n|^k) < \infty$ then $X_n = O_p(1)$
$X_n = o_p(1), Y_n = O_p(1) \to X_n Y_n = o_p(1), X_n + Y_n = O_p(1)$.
Moreover, if for all $\delta > 0$, $P(|Y_n| > \delta) \to 1$ then X_n $Y_n \to 1$
If $X_n = O_p(1)$ and f is a continuous function, then $f(X_n) = O_p(1)$. Also, show that if $X_n = o_p(1)$ and f is a continuous function and $f(0) = 0$, then $f(X_n) = o_p(1)$.
Assume $X_n \to X$ in distribution. Then, $X_n = O_p(1)$. Now, if $Z_n = o_p(1)$ and $X_n \xrightarrow{d} X$ then $Z_n = o_p(1)$

Examples on Asymptotics

$X_n \sim F(x - \theta)$ and want to find the asymptotics for the plug in estimator for $|\theta|^p$ when $p \in (0, \infty)$ and $\theta = 0$. Then, $\sqrt{n^p}|\bar{x}|^p \to |N(0, \sigma^2)|^p$.
$X_i \sim U(0, \theta)$ then $n(\theta - X_{(n)}) \to Exp(1/\theta)$

Delta Method

Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be a map differentiable. Suppose T_n is an estimator of θ and $r_n(T_n - \theta) \to T, r_n \to \infty$ Then

$$r_n(\phi(T_n) - \phi(\theta)) \to A(\theta)T$$

Example: $X_i, \sim F$ i.i.d,
$E(X_i) = \theta, Var(X_i) = 1, g(\theta) = cos(\theta)$. Then, if $\theta = 0$ $n(cos(\bar{X}_n - 1) \to -1/2\chi_1^2$ For this, recall that $\cos(h) - 1 + h^2/2 = o(h^2)$ and that we can replace $o(h^2)$ by $o(\bar{X}_n^2)$
Uniform delta method: $\theta_n \to \theta$ and $\sqrt{n}(T_n - \theta_n) \xrightarrow{d} T$, ϕ continuously differentiable at θThen, if for every ϵ and n large enough we have $|\phi(\theta_n + h) - \phi(\theta_n) - \phi'(\theta)h| \leq \epsilon||h||$ we get $\sqrt{n}(\phi(T_n) - \phi(\theta_n)) \xrightarrow{d} \phi'(\theta)T$

Consistency of the MLE and asymptotics

Define

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(X_i)$$

Prop: $G(\theta_0) > G(\theta) \forall \theta \neq \theta_0$
To prove consistency: Show that $P(\theta_{mle} \neq \theta) \leq$ something that goes to zero.
Theorem: Consistency of the MLE Suppose the two following conditions hold

1. $$\sup_{\theta \in \Theta} |\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) - E_{\theta_0}(\log p_\theta(x_i))| \xrightarrow{p} 0$$
2. $$\forall \epsilon > 0 \qquad \sup_{||\theta - \theta_0|| > \epsilon} E_{\theta_0}(\log p_\theta(x)) < E_{\theta_0}(\log p_{\theta_0}(x))$$

Then, $\theta_{MLE} \xrightarrow{p} \theta_0$
Cramer result: if $\log p_\theta(x)$ is two times continuously differentiable and bounded by $K(x) \in L^1(?)$ in some neighborhood of θ_0. Then if the MLE is consistent we have asymptotic normality

For the first condition to hold we use the Uniform law of large numbers.
Theorem: Uniform law of large numbers. Assume the following condition hold

1. Θ is compact
2. $U(x,\theta)$ is continuous of θ for every x
3. There exists $k(x) \in L1$ such that $|U(x,\theta)| < k(x)$ for every x and θ.

Sketch of the proof: define $\tilde{U}(X_i,\theta) = U(X_i,\theta)$ by 3 it is continuous in theta. It is also dominated by a function in L^1, uniformly in θ. Then, define
$\phi(x,\rho,\theta) = \sup_{\theta':||\theta'-\theta\leq\rho} \tilde{U}(x,\theta')$, We have
$\lim_{\rho\to 0}\phi(x,\rho,\theta) = \tilde{U}(x,\theta)$ and $\lim_{\rho\to 0} E(\phi(X_i,\rho,\theta)) = 0$.
This needs the uniform bound of $\tilde{U}$ in θ. From the limits of expectations one conclude there is a covering of balls $B(\theta_i,\rho_{\theta_i})$ such that $U(x,\theta) \leq \phi(x,\rho_{\theta_M},\theta_M)$ where θ belongs to the Mth ball. Thus

$$P(\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\tilde{U}(x,\theta) > 2\epsilon) \leq \sum_{j=1}^{M} P(\frac{1}{n}\sum\phi(x_i,\rho_j,\theta_j) > 2\epsilon)$$

And each of the terms in the right converge to zero, as

$$\frac{1}{n}\sum\phi(x_i,\rho_j,\theta_j) \to E(\phi(X_i,\rho_j,\theta_j)) \leq \epsilon$$

Weakening the compactness condition: Suppose $p_\theta(x) \to 0$ as $||\theta|| \to \infty\ \forall x$. Furthermore, suppose the weak law of large numbers hold for $\sup_{||\theta||>b}\log p_\theta(x)\ \ \forall b$ and $G(\theta_0)$ is finite. Then, there exist b such that

$$P\left(\sup_{||\theta||>b}\frac{1}{n}\sum_{i=1}^{n}\log(p_\theta(x_i)) \geq \frac{1}{n}\sum_{i=1}^{n}\log(p_{\theta_0}(x_i))\right) \to 0$$

Extension: assume not necessarily θ_n maximizes $G_n(\theta)$ but $G_n(\theta_n) \geq G_n(\theta_n) + o_p(1)$. Then $\theta_n \to \theta_0$ Asymptotic distribution of the MLE. Theorem based on third order derivative. Conditions:

1. $\theta_0 \in Int(\Theta)$ and $\int(\Theta) \neq \emptyset$
2. θ_{ML} is consistent
3. Fisher information is well defined (we can differentiate $\log p_\theta(x)$ with respect to θ and pass the derivative under the integral sign
4. Need also second order definition of Fisher Information
5. Third derivative also exists and
$\exists K_3(x) st|\frac{\partial^3}{\partial\theta^3}\log(p_\theta(x))| \leq K_3(x)$ and $E(|K_3(x)|) < \infty$

The key of this proof is to show that $G_n''(\theta) = O(1)$ using the last condition Then, $n^{1/2}(\theta_{MLE}-\theta_0) \to N(0, I(\theta_0)^{-1}$
Cramer theorem on asymptotics for the MLE: Θ is open, second derivatives of $\log p_\theta(x)$ exists and are continuous and information is well defined in both ways. Second derivatives are bounded uniformly by $K(x) \in L^1$, and $I(\theta_0)$ is differentiable. If θ_{mle} is consistent then we have asymptotics. To prove this: Prove $G''(\tilde{\theta_n}) \to -I(\theta_0)$ using the uniform law of large numbers.

Quadratic mean differentiability

Definition: the family $p_\theta, \theta \in \Theta \subseteq \mathbb{R}^k$ is called differentiable in quadratic mean at θ_0 if and only if there exists a function $\eta_1(x,\theta_0)$ such that

$$\int\left[\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \eta_1^T(x,\theta_0)h\right]^2 d\mu(x) = o(||h||^2)$$

η_1 should be, in some sense, the gradient of $\sqrt{p_\theta(x)}$ Definition 2//

$$\int\left[\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}\eta^T(x,\theta_0)\sqrt{p_{\theta_0}(x)}h\right]^2 d\mu(x) = o(||h||^2)$$

Remark: if the support doesn't depend on the parameter then both definitions are essentially equivalent. In definition 2, η should be the derivative of the log likelihood.Some examples, normal, laplace. Uniform is not qmd
Theorem: $\theta \in \Theta$ open set, assume $\sqrt{p_\theta}(x)$ is continuously differentiable for every x. If the elements of
$I_\theta = \int \frac{\nabla p_\theta}{p_\theta}\frac{\nabla p_\theta^t}{p_\theta}p_\theta d\mu$ are well defined and continuous in θ then the family is qmd and $\eta = \nabla\log(p_\theta)$
Theorem. Suppose Θ is open and the model is qmd at θ_0, then

$$E\eta(X,\theta_0) = 0 \quad E(\eta(X,\theta_0),\eta^t(X,\theta_0) < \infty$$

The second expectation is the generalized fisher information.
Theorem: Suppose P_θ is qmd at $\theta \in Int(\Theta)$. Furthermore, suppose that there exists a measurable function K with second moment such that

$$|\log(p\_\theta_1) - \log(p_{\theta_2})| \leq K(x)||\theta_1-\theta_2||$$

If $I(\theta_0)$ is not singular and θ_n is consistent then

$$\sqrt{n}(\theta_n-\theta) \to N(0, I^{-1}(\theta_0))$$

Hilbert spaces identities (useful for qmd)

If $f,g \in L^2(\mu)$ then $f+g, f-g \in L^2(\mu)$
If $g_n \to g, f \in L^2(\mu)$ then $fg_n \to fg$
Examples: $p_\theta(x) = (2-2x/\theta)1_{x\leq\theta}$ is not qmd, but $p_\theta(x) = 3/\theta^3(x-\theta)^2 1_{x\leq\theta}$ is qmd. In both cases, assume in $(\theta,\theta+h)^c$ the integral is $o(||h||^2)$ so we focus on $\theta, \theta+h$

Counterexample for the consistency of the MLE

Consider $h(x) \geq \exp(1/x^2) \geq 1, c < 1$ and a_k defined such that

$$\int_{a_k}^{a_{k-1}}(h(x)-c)dx = 1-c$$

and $f_k(x) = h(x)$ in $[a_k, a_{k-1}]$ and c otherwise. We have $X_{(1)} = o_p(1), nX_{(1)} = O_p(1)$ for all k, and to prove $K_{MLE} \to \infty$ we show instead $P(P_{K^*n}(X) > P_j(X)) \to \infty$ where K^*n is the interval where $X_(1)$ recedes. The above (using decreasingness of h) reduces to show $\frac{1}{n}\log(X_{(1)}) \to \infty$ in probability

Robust Estimation

We used a general M estimator, $\hat{\theta} = \arg\min\sum_{i=1}^{n} m_\theta(X_i)$. Usually $m_\theta(x) = m(x-\theta)$. We want consistency, so $E_{\theta_0}m(X-\theta)$ is minimized at θ_0. Following the same argument as for the asymptotics of the mle, we get the following asymptotics

$$\sqrt{n}(\hat{\theta}-\theta_0) \to N(0, \frac{E_{\theta_0}m'^2(X_i-\theta_0)}{(E_{\theta_0}m''(X_i-\theta_0))^2})$$

For robustness, we consider the escenario
$X_i \sim F(t-\theta) = (1-\epsilon)\Phi(t-\theta) + \epsilon H(t-\theta)$ We want the most robust estimator, the m function such that

$$m* = \arg\min_{m \text{ convex}}\sup_H \frac{E_{\theta_0}m'^2(X_i-\theta_0)}{(E''_{\theta_0}(X_i-\theta_0))^2}$$

Theorem: Huber. The asymptotic variance has the following saddle point $V(m_0,f) \leq V(m_0,f_0) \leq V(m,f_0)$, where f_0 is the least favorable distribution (h_0 don't have any mass on $[-k,k]$

$$m_0(t) = \begin{cases} 1/2t^2 & |t| < k \\ k|t| - 1/2k^2 & |t| \geq k \end{cases}$$

Also, if f denote the pdf of F then we get the following expression to find k as a function of ϵ

$$m'(t) = -\frac{f'}{f}$$

From the above condition one can find the least favorable distribution H (solving a differential equation). It turns out this distribution has density (and making it integrate up to one gives us k as a function of ϵ

$$h_0(y) = \frac{1-\epsilon}{\epsilon\sqrt{2\pi}}\left(\exp(-k|y|-k^2/2) - \exp(-y^2/2)\right)1_{|y|\geq h}$$

Asymptotics for Robust estimation: Consider minimizing
$\hat{\beta} = \arg\min\sum m(y_i - x_i^t\beta)$ and
$E(xx^t) = \Sigma, \epsilon = y_i - x_i^t\beta \perp\!\!\!\perp x_i$. Then

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, \sigma^{-1}E(m'(\epsilon)^2)/E(m''(\epsilon))^2)$$

To prove the above, express $y_i - x_i^t\hat{\beta} = y_i - x_i^t\beta_0 + x_i^t\beta_0 - x_i^t\hat{\beta}$ and then taylor for m' getting $m''(\tilde{z}_n^i)$ with $\tilde{z}_n = y_i^t - \tilde{\beta}_n^i$ and use the usual asymptotics and uniform law of large numbers. (probably there is no dependence on i for the $\tilde{\beta}$

Hypothesis testing

For Kolmogorov Smirnov,
$(F(X_{(1)}), \dots F(X_{(n)}) \sim (U_{(1)}, \dots U_{(n)})$
Behrens-Fisher problem.
$X_1, \dots X_n \sim \mathcal{N}(\mu_x, \sigma_x^2), Y_1, \dots Y_m \sim\sim \mathcal{N}(\mu_y, \sigma_y^2)$ Testing $\mu_x = \mu_y$, the statistic $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_x 2}{n}+\frac{S_y^2}{m}}}$ depends on the unknown σ^y, σ_x
Neyman Pearson Fundamental lemma: Suppose p_0, p_1 are the densities w.r. to μ. For testing $H_0 : p_0$ v. s.$H_1 : p_1$ then

1. There exists a test ϕ and a constant k such that $E_0(\phi(X)) = \alpha$ and

$$\phi(x) = \begin{cases} 1 & p_1(x) > kp_0(x) \\ 0 & p_1(x) < kp_0(x) \end{cases}$$

   $k = k_\alpha$ should be adjusted such that

$$k_\alpha = \inf_k P\left(\frac{p_1(x)}{p_0(x)} \le k\right) \ge 1 - \alpha$$

2. Sufficiency: if a test satisfies the above conditions it is the most powerful test then it has to satisfy the above for some k Also, $E_0(\phi(X)) = \alpha$ unless the $E_1(\phi(X)) = 1$ We use the convention $0 \times \infty = 0$

For composite null (finite) the optimal test is given by

$$\phi(x) = \begin{cases} 1 & p_1(x) > \sum_{i=1}^n c_i p_{\theta_i}(x) \\ 0 & p_1(x) < \sum_{i=1}^n c_i p_{\theta_i}(x) \end{cases}$$

Theorem: if there exists $c_i \ge 0$ that satisfy the above (with a randomized number for the equality) and it also satisfies

$$\int \phi(x) p_{\theta_i}(x) d\mu(x) \le \alpha \ \ \forall i = 1 \dots k$$

$$\int \phi(x) p_{\theta_i}(x) d\mu(x) < \alpha \Rightarrow c_i = 0$$

then ϕ is a most powerful α test
Theorem: if there exists $c(\theta) \ge 0, \int c(\theta) d\lambda(\theta) < \infty$ that satisfies

$$\phi(x) = \begin{cases} 1 & p_1(x) > \int c(\theta) p_\theta(x) d\lambda(\theta) \\ 0 & p_1(x) < \int c(\theta) p_\theta(x) d\lambda(\theta) \end{cases}$$

(with a randomized number for the equality) and also satisfies

$$\int \phi(x) p_\theta(x) d\mu(x) \le \alpha \ \ \forall \theta \in \Theta$$

$$\int_{\{\theta : \int \phi(x) p_\theta(x) d\mu(x) < \alpha\}} c(\theta) d\lambda(\theta) = 0$$

then ϕ is a most powerful α test Def:Least favorable distribution at level α is the one that gives the least bayesian power for the alternative (taking the uniformly most powerful bayesian test).
Theorem: Suppose there exist a distribution Λ such that the bayesian most powerful α test ϕ_Λ is also size α for the original hypothesis. Then

1. ϕ_Λ is most powerful for the original hypothesis
2. if it is unique for the bayesian test it is also unique for the original one
3. Λ is the least favorable distribution

Corollary: Suppose Λ is a probability distribution over ω and ω' is a subset of ω with probability $\Lambda(\omega') = 1$. Let ϕ_Λ be a bayesian most powerful test. Then, if for $\theta' \in \omega', E_{\theta'}(\phi_\lambda(X)) = \sup_{\theta \in \omega} E_\theta(\phi_\Lambda(X)) = \alpha$ ϕ_Λ is most powerful!
Definition: For testing $H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$ a level α test ϕ is unbiased if it is of level α and $E(\phi(X)) \ge \alpha, \forall \theta \in \Theta_1$
For point hypothesis Unbiasedness is obtained via minimization of the power at the null.

Examples on Hypothesis Testing

$p_\theta \sim \mathcal{N}(0,1), H_0 : \theta \in \{-1, 1\} H_1 : \theta = 0$ Reject when $|X| < K$
By the lagrange multipliers lemma, for the test
$p_\theta \sim \mathcal{N}(0,1), H_0 : \theta \in \{-1, 1, 4\} H_1 : \theta = 0$ we still will reject when $|X| < K$
$X \sim \mathcal{N}(\theta, 1)$. $H_0 \theta < 0 \quad vs \quad H_1 : \theta = 1$. The least favorable distribution doesn't exist, but we get a sequence of test converging to the least favorable test with $\theta = 0$
$X \sim \mathcal{N}(\theta, 1)$. $H_0 \theta \le -1 \vee \theta \ge 1 vs H_1 : \theta = 0.5$. Regardless of the value of θ we reject if $|X|$ is small.
$X \sim \mathcal{N}(\theta, 1)$. $H_0 |\theta| \le 1 \quad vs \quad H_1 : |\theta| \ge 1$. No UMP for $\theta > 1$ the least favorable distribution will be a delta at 1.
$X \sim \mathcal{N}(\theta, 1)$. $H_0 |\theta| \ge 1 \quad vs \quad H_1 : |\theta| < 1$. There is a UMP. For each value of the alternative one chooses a least favorable prior such that $\delta\beta(-1) + (1-\delta)\beta(1) = \alpha$. From this, $\beta(1) = \beta(-1) = \alpha$ (otherwise level constraints are not satisfied).
For testing in the case where there are unknowns: find a complete statistic T for the unknown parameters ν , assuming the null. argue that (by completeness) that the problem is equivalent to

$$\max E(\phi(x_1, \dots x_n)|T), \forall (\mu, \nu) \in \Theta_1$$

$$\text{s.t } E(\phi(x_1, \dots x_n)|T) = \alpha, \forall (\mu, \nu) \in \partial\Theta_0$$

Then find an ancillary statistic under the null

Hypothesis Testing from TSH

Theorem: suppose that the distribution of X given by

$$P_{\theta,\mu} = C(\theta, \mu) \exp\left[\theta U(x) + \sum \mu_i T_i(x)\right]$$

and that $V = h(U, T)$ is independent of T when $\theta = \theta_0$. Then ϕ_1 is UMP unbiased for testing H_1 provided the function h is increasing in u for each t, and ϕ_e is UMP unbiased for H_4 provided $h(u,t) = a(t)u + b(t), a(t) > 0$. The tests ϕ_2, ϕ_3 are UMP unbiased for H_2 and H_3 if V is independent of T when $\theta = \theta_1, \theta_2$ and if h is increasing in u for each t. Here H_1 is the simple one sided hypothesis, H_2 is $\theta < \theta_1 \vee \theta > \theta_2$, H_3 is $\theta \in [\theta 1, \theta_2]$ and H_4 is $\theta = \theta_0$ The Rejection regions are $R_1 = u > C_0(t)$, $R_2 = C_1(t) < u < C_2(t)$, $R_3, R_4 = u < C_1(t) \vee u > C_2(t)$. The new rejection regions are expressed in terms of V, without any dependence on T. In this case, R_3 will also have a form like R_2. The constants should be adjusted such that $E(\phi(V)) = \alpha$ or $E(V\phi(V)) = \alpha E(\sigma V)$
Normal examples. Testing $\sigma \le \sigma_0$ with unknown mean. $U = \sum x_i^2, T = \bar{x}$. For all means and σ_0,$h(U,T) = U - T^2$ is independent of T (in fact always). Since h is increasing in U the rejection region is $V \le C_0$ For testing $\mu = 0$ vs $\mu \ne 0$ with unknown variance use $W = \bar{X}/\sum x_i^2$, then one can get the t distribution (need linearity).
Let θ be a real parameter, and let X have density $p_\theta(x) = C(\theta) \exp(Q(\theta)T(x))\, h(x)$ where Q is strictly monotone. Then there exists a UMP test ϕ for testing $\theta < \theta_0$ If Q is increasing it rejects when T is large. If it is decreasing, it rejects for small *delta*

Likelihood Ratio Test

Assume $X_1, \dots X_n$ are i.i.d according to q.m.d family where Ω is an open subset of $\mathbb{R}^k$ and $I(\theta)$ is positive definite.

1. Consider testing the simple null hypothesis $\theta = \theta_0$. Suppose θ_n is an efficient estimator for θ assuming $\theta \in \Omega$ in the sense that is satisfies asymptotic normality conditions for $\theta = \theta_0$ Then the likelihood ratio satisfies, under H_0

$$2\log(R_n) \to X_k^2$$

2. Consider testing the composite null hypothesis $\theta \in \Omega_0$ where

$$\Theta_0 = \{\theta = (\theta_1, \dots, \theta_k) : \ A(\theta - a) = 0\}$$

   And A is a $p \times k$ matrix of rank p and a is a fixed $k \times 1$ vector. Ket $\theta_{n,0}$ be an efficient estimator of θ assuming $\theta \in \Omega_0$, that is, asymptotic normality holds for all $\theta \in \Theta_0$. Then the likelihood ratio converges to a chi square with p degrees of freedom

3. More generally, suppose Θ_0 is represented as $\Theta_0 = \{\theta : g_1(\theta), \dots, g_p(\theta)^T = 0\}$ where $g_i(\theta)$ is a continuously differentiable function from $\mathbb{R}^k$ to $\mathbb{R}$. Let $D = D(\theta)$ be the $p \times k$ matrix with (i,j) entry $\partial g_i(\theta)/\partial theta_j$, assumed to have rank p. Then the limit is as above.
   For testing $H_0 : X_i \sim p(x)$ and $H_1 : X_i \sim q(x)$ where $p(x), q(x)$ are discrete the optimal test is given by

$$KL(\hat{p}, p) - KL(\hat{p}, q) > T$$

Random facts

Law of total (co)variance
$Cov(XY|Z) = E(X, Y|Z) - E(X|Z)E(Y|Z)$. Another formula (for conditional variance): $Var(Y|X) = E((Y - E(Y|X))^2|X)$

$$Cov(X, Y) = E(Cov(X, Y)|Z) + Cov(E(X|Z), E(Y|Z))$$

Suppose $Z \sim \mathcal{N}(0, I_n)$ P projection matrix, $P^2 = P$ with rank r. Then $Z^t P Z \sim \chi_r^2$ Use the diagonalization with eigenvalues 0 and 1.
If the covariance matrix is singular then the elements are linearly dependents, a.e.
$E(X) = \mu, Cov(X) = \Sigma$ If A is symmetric, then $E(X^t A X) = tr(A\Sigma) + \mu^t A \mu$,

$Var(x^tAx) = 2Tr(A\Sigma A\Sigma) + 4\mu^t A\Sigma A\mu$. if $X \sim \mathcal{N}(0, I_N)$ then $Var(x^tAx) = tr(A^2) + tr(AA^t)$

$$Tr(ABC) = Tr(BCA) = Tr(CAB), det(AB) = det(A)det(B)$$

Ratio of $0, 1$ normals is cauchy. $Tan(Unif(-\pi/2, \pi/2)$ is Cauchy. Characteristic function $\phi(t) = \exp(-|t|)$

$X_i \exp(\lambda) \quad X_{(n)} - \log(n)/lambda \quad$ converges to a Gumbell

Stirling Formula:

$$\lim \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1$$

$X_n \xrightarrow{d} X$ and F continuous. Then

$$\sup_x |P(X_n \leq x) - P(X \leq x)| \to 0$$

Levy inversion formula

$$f(x) = \frac{1}{2\pi} \int_R \phi(\theta) \exp(-i\theta x) d\theta$$

The function $F(x) = \log\det(X)$ is concave for X symmetric positive definite

Poems, Prayers and Jokes

Song of Myself (1). Walt Whitman, Leaves of grass

I celebrate myself, and sing myself,
And what I assume you shall assume,
For every atom belonging to me as good belongs to you.
I loafe and invite my soul,
I lean and loafe at my ease observing a spear of summer grass.
My tongue, every atom of my blood, form?d from this soil, this air,
Born here of parents born here from parents the same, and their parents the same,
I, now thirty-seven years old in perfect health begin,
Hoping to cease not till death.
Creeds and schools in abeyance,
Retiring back a while sufficed at what they are, but never forgotten,
I harbor for good or bad, I permit to speak at every hazard,
Nature without check with original energy.

Oh Me, Oh Life, Walt Whitman

Oh me! Oh life! of the questions of these recurring,
Of the endless trains of the faithless, of cities fill?d with the foolish,
Of myself forever reproaching myself, (for who more foolish than I, and who more faithless?)
Of eyes that vainly crave the light, of the objects mean, of the struggle ever renew?d,
Of the poor results of all, of the plodding and sordid crowds I see around me,
Of the empty and useless years of the rest, with the rest me intertwined,
The question, O me! so sad, recurring?What good amid these, O me, O life?
Answer.
That you are here?that life exists and identity,
That the powerful play goes on, and you may contribute a verse.

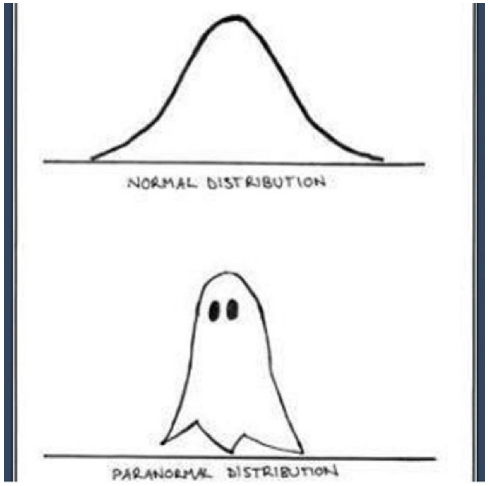
Love is the Water of Life, Rumi

Love is the Water of Life
Everything other than love for the most beautiful God though it be sugar- eating.
What is agony of the spirit?
To advance toward death without seizing hold of the Water of Life.

Serenity Prayer, Reinhold Niebuhr

God, give me grace to accept with serenity
the things that cannot be changed,
Courage to change the things
which should be changed,
and the Wisdom to distinguish
the one from the other.
Living one day at a time,
Enjoying one moment at a time,
Accepting hardship as a pathway to peace,
Taking, as Jesus did,
This sinful world as it is,
Not as I would have it,
Trusting that You will make all things right,
If I surrender to Your will,
So that I may be reasonably happy in this life,
And supremely happy with You forever in the next.
Amen.

(a) Mnemonic rule:The Mr T test is the uniformly most powerful



(b) the paranormal distribution