Contents

1 Measure Theory (The Basics)

Definition 1 (Basic measure theoretic definitions)**.**

- A *σ-algebra* on a set Ω is the collection of all possible outcomes or all subsets of a set Ω.
- Given a *probability space*, $(\Omega \text{ (set)}, \mathcal{F} \text{ (σ-alg.)}, p \text{ (prob measure)})$, a **random variable** is a measurable map from the set to the real numbers: $\Omega \to \mathbb{R}$.
- The **distribution** of a random variable X is defined as $p_X = p \cdot X^{-1}$, meaning the probability measure applied to the inverse map of the random variable (the set). The distribution of a random variable is a *measure* on $\mathbb{R}$.
- The **Radon-Nikodym theorem** says that under certain conditions, any measure can be described using another measure defined on the same space by assigning a density to each point in space and integrating over the measurable subset of interest. The strategy is like so:

$$\nu(A) = \int_A f d(\mu)$$

  The function f, the **Radon-Nikodym derivative**, is defined as $\frac{d\nu}{d\mu}$, the derivative of one measure with respect to another.
- The **probability density function** is the Radon-Nikodym derivative of the distribution with respect to the Lesbegue measure on $\mathbb{R}$ or $\mathbb{R}^k$: $f_X = \frac{dP_x}{d\lambda}$.

The dominated convergence theorem (DCT) is a very useful device that allows us to link pointwise convergence of a sequence of functions to convergence of the *integral* of the sequence of functions. Essentially provides conditions under which we can push the integral inside the limit.

Theorem 1 (DCT)**.**
Suppose $f_n \to f$ pointwise within a measurable space $(\Omega, \mathcal{F}, \mu)$ (of which a probability space is a special example). Suppose f_n is dominated by some integrable function, i.e.,

$$|f_n(x)| \leq g(x) \quad \text{s.t.} \int_\Omega |g(x)|\, d\mu < \infty$$

for all n and $x \in \Omega$. Then:

$$\begin{aligned} \lim_{n\to\infty} \int_S |f_n - f| d\mu = 0 \\ \equiv \lim_{n\to\infty} \int_S f_n d\mu = \int_S f d\mu \end{aligned}$$

Note: this result is very useful in probability theory, because it allows us to swap limits and expectations. Suppose $X_n \overset{p}{\to} X$ are random variables and $Pr(|X_n| < Y)$ for some other random variable Y with $\mathbb{E}(Y) < \infty$. Then:

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X) = \mathbb{E}\left(\lim_{n\to\infty} X_n\right)$$

It also allows us to swap integration (expectation) and differentiation.

2 Decision Theory

Decision theory is a general framework that unites hypothesis testing and estimation. Based on a data realization x, we can take an action a in the action space $\mathcal{A}$.

1. Point estimation example: suppose $\theta = (\mu, \sigma)^2$, and objective is to estimate $\Psi(\theta) = \mu$ of a $N(\mu, \sigma^2)$ random variable. The action space may be $\mathcal{A} = \mathbb{R}$. A typical loss function is $L(\theta, a) = |\theta - a|^2$
2. Hypothesis testing example: suppose we want to test whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. The action space, $\mathcal{A} = \{0, 1\}$, denotes the choice of $\theta \in \Theta_a$. A typical loss function could be the modified 0-1 loss: $L(0, \theta) = \ell_0 \mathbb{I}_{\Theta_1}(\theta)$ and $L(1, \theta) = \mathbb{I}_{\Theta_0}(\theta)$

2.1 Basic terminology

Definition 2 (Decision rule, loss, risk)**.**

- A **decision function**, $D : \mathcal{X} \times \mathcal{A} \to [0, 1]$, $D(a, x) = d(a|X = x)$, is a probability of action A conditional on $X = x$. Decision functions/rules can be either stochastic or deterministic (with probabilities 0/1)– we denote the class of all stochastic decision rules $\mathcal{D}$, and deterministic decision rules $\mathcal{T}$.
    1. Note: a particular decision is a random quantity that depends on variability in the data X and variability in the decision a.
- The **loss** for a given action a, $L(a, \theta)$, describes the quality of a decision at θ.
- The **risk** of a decision rule D at θ is the expected loss, $\mathcal{R}(D, \theta) \equiv \int_{\mathcal{X}} \int_{\mathcal{A}} L(a, \theta) D(a|x) dP_\theta(x)$. Notice that the risk is the average loss, marginal over the two layers of randomness: the randomness of the data and randomness of the decision. Smaller risk indicates better performance of a decision rule.

Example 1 (Neyman-Pearson hypothesis Testing: constrained minimax)**.** We can consider hypothesis testing under the **Neyman-Pearson** paradigm as a constrained minimax approach. Consider testing whether θ belongs in Θ_0 or Θ_1, with loss function $L(0, \theta) = \ell_0 \mathbb{I}_{\Theta_1}(\theta)$ and $L(1, \theta) = \mathbb{I}_{\Theta_0}(\theta)$. $\ell_0 > 1$ implies making T2 errors more costly than T1 errors. The risk:

$$\begin{aligned}
R(\theta, D) &= \int \sum_{a=0}^{1} L(a, \theta) D(a|X) dP_\theta(x) \\
&= \int \left[\ell_0 \mathbb{I}_{\theta_1}(\theta) D(0|X) + \mathbb{I}_{\theta_0}(\theta) D(1|X) \right] dP_\theta(x) \\
&= \begin{cases} P_\theta(\text{declaring } \theta \in \Omega_1) = \text{type 1 error} & \text{if } \theta \in \Theta_0 \\ \ell_0 P_\theta(\text{declaring } \theta \in \Omega_0) = \ell_0 \times \text{type 2 error} & \text{if } \theta \in \Theta_1 \end{cases}
\end{aligned}$$

The Neyman-Pearson paradigm advocates choosing decision rule D^* s.t.

$$\sup_{\theta \in \Theta_1} \mathcal{R}(D^*, \theta) = \inf_{D \in \mathcal{D}} \sup_{\theta \in \Theta_1} \mathcal{R}(D, \theta) \text{ subject to constraint } \sup_{\theta \in \Theta_0} \mathcal{R}(\theta, D) \leq \alpha$$

2.2 Bayesian Inference

Definition 3 (Bayes risk, Bayes rule)**.**
In the Bayesian paradigm, we define a *prior distribution* Π on the parameter space Θ. The **Bayes risk** of a decision rule D is the expected risk of D over the the prior on θ:

$$r(D, \Pi) = \int \mathcal{R}(D, \theta) d\Pi(\theta)$$

A **Bayes rule**, D_Π is optimal with regard to the Bayes risk

$$r(D_\Pi, \Pi) = \inf_{D \in \mathcal{D}} r(D, \Pi) = \inf_{D \in \mathcal{D}} \mathbb{E}_\Pi \left[\int_{\mathcal{A}} L(a, \theta) D(da|x) \middle| X = x \right]$$

Definition 4 (Prior, Posterior, Kernel, Conjugate Prior)**.** Let Π be a **prior distribution** on Θ. Let $p(X|\theta)$ and $\pi(\theta)$ be associated densities. The **posterior distribution** of $\theta|X = x$ is defined as:

$$\begin{aligned} p(\theta|x) &= \frac{p(x, \theta)}{p(x)} \\ &= \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta')\Pi(d\theta')} \end{aligned}$$

The **kernel** of the posterior is a function f such that the posterior distribution factorizes into a component that depends on X only, $c(X)$, and f, a component that depends on both X and θ:

$$p(\theta|x) = c(x)f(x, \theta) \propto f(x, \theta)$$

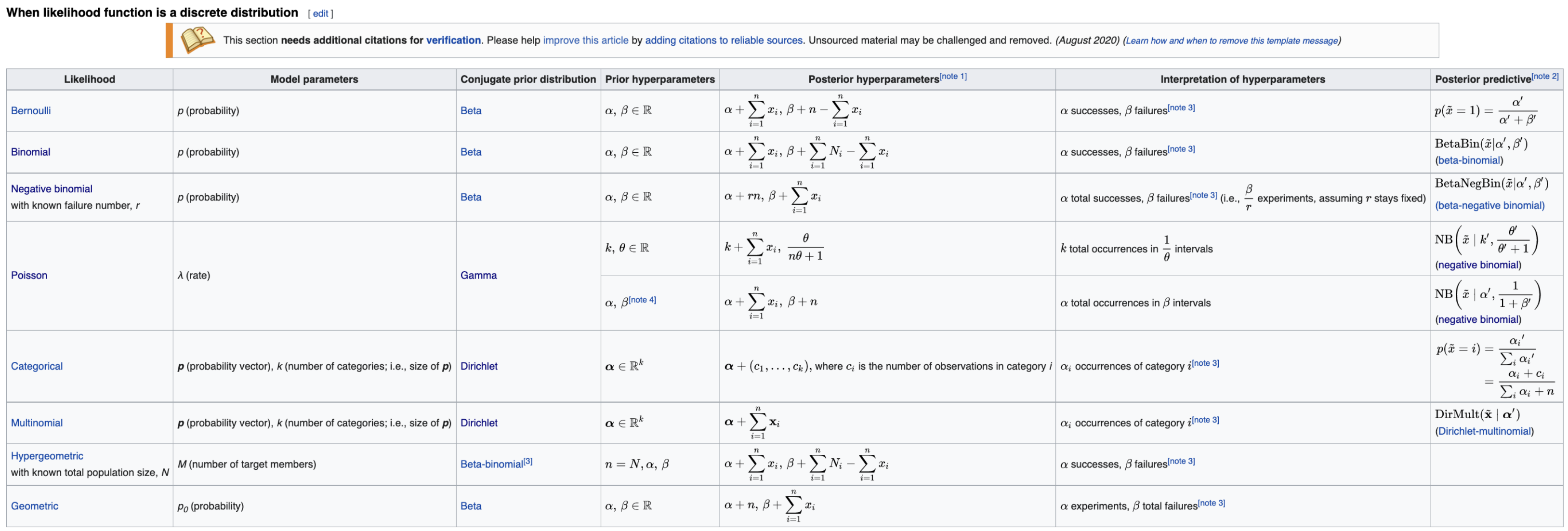
Importantly, the kernel *uniquely determines the distribution.*

A **conjugate prior** is a prior that belongs to a family $\mathcal{P}_\Pi$, and ensures for almost all x, the posterior distribution $P(\theta|x)$ also falls in $\mathcal{P}_\Pi$: the posterior belongs to the same family as the prior.

Strategy 1 (Finding the posterior)**.** One can find the posterior by setting it proportional to the conditional likelihood times the prior, and factoring to identify the kernel. For example, suppose $X|\theta \sim \text{Pois}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta) \times \pi(\theta) \\ &= e^{-\theta}\frac{\theta^x}{x!} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &= \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)x!}}_{c(x)} \underbrace{\theta^{x+\alpha-1}e^{-(\beta+1)\theta}}_{f(x,\theta)} \\ &\propto \theta^{x+\alpha-1}e^{-(\beta+1)\theta} \end{aligned}$$

Which is the kernel of a $\text{Gamma}(\alpha + x, \beta + 1)$, so $\theta|x \sim \text{Gamma}(\alpha + x, \beta + 1)$.

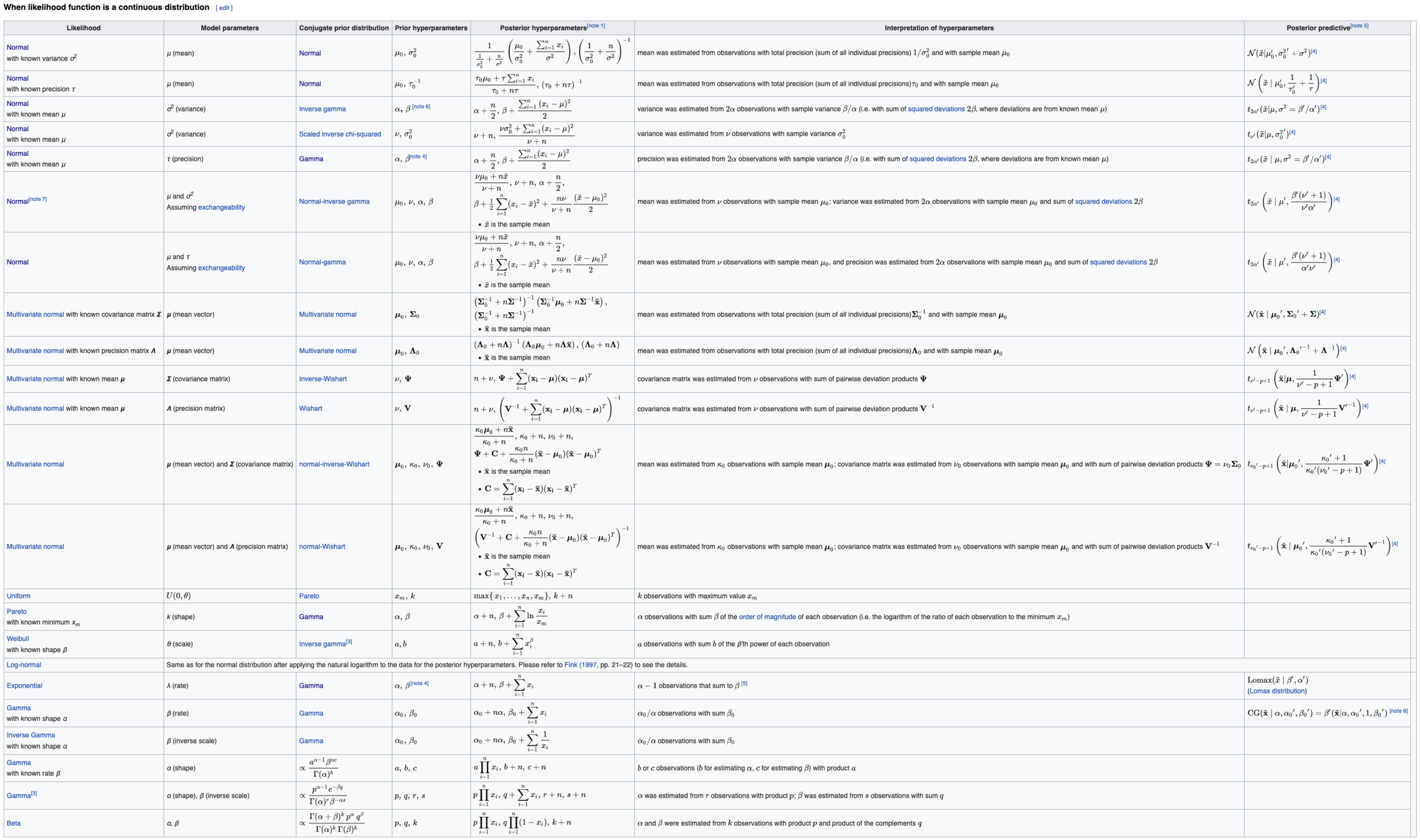


When likelihood function is a discrete distribution [edit]

This section **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(August 2020)* (*Learn how and when to remove this template message*)

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters	Posterior predictive[note 2]
Bernoulli	*p* (probability)	Beta	$\alpha,\, \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^n x_i,\, \beta + n - \sum_{i=1}^n x_i$	α successes, β failures[note 3]	$p(\tilde{x}=1) = \frac{\alpha'}{\alpha'+\beta'}$
Binomial	*p* (probability)	Beta	$\alpha,\, \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^n x_i,\, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	α successes, β failures[note 3]	$\operatorname{BetaBin}(\tilde{x}\mid\alpha',\beta')$ (beta-binomial)
Negative binomial with known failure number, *r*	*p* (probability)	Beta	$\alpha,\, \beta \in \mathbb{R}$	$\alpha + rn,\, \beta + \sum_{i=1}^n x_i$	α total successes, β failures[note 3] (i.e., $\frac{\beta}{r}$ experiments, assuming r stays fixed)	$\operatorname{BetaNegBin}(\tilde{x}\mid\alpha',\beta')$ (beta-negative binomial)
Poisson	λ (rate)	Gamma	$k,\, \theta \in \mathbb{R}$	$k + \sum_{i=1}^n x_i,\, \frac{\theta}{n\theta+1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}\left(\tilde{x} \mid k', \frac{\theta'}{\theta'+1}\right)$ (negative binomial)
			$\alpha,\, \beta$[note 4]	$\alpha + \sum_{i=1}^n x_i,\, \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1+\beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), *k* (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category *i*	α_i occurrences of category i[note 3]	$p(\tilde{x}=i) = \frac{\alpha_i{}'}{\sum_i \alpha_i{}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), *k* (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$	α_i occurrences of category i[note 3]	$\operatorname{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, *N*	*M* (number of target members)	Beta-binomial[3]	$n = N, \alpha,\, \beta$	$\alpha + \sum_{i=1}^n x_i,\, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	α successes, β failures[note 3]	
Geometric	p_0 (probability)	Beta	$\alpha,\, \beta \in \mathbb{R}$	$\alpha + n,\, \beta + \sum_{i=1}^n x_i$	α experiments, β total failures[note 3]	

Figure 1: Conjugate priors for discrete distributions.



When likelihood function is a continuous distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters	Posterior predictive[note 5]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\, \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\vert\mu_0', {\sigma_0^2}' + \sigma^2)$[4]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\, \tau_0^{-1}$	$\frac{\tau_0 \mu_0 + \tau \sum_{i=1}^n x_i}{\tau_0 + n\tau}, (\tau_0 + n\tau)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions)τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x} \mid \mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[4]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\, \beta$ [note 6]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\vert\mu, \sigma^2 = \beta'/\alpha')$[4]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\, \sigma_0^2$	$\nu+n, \frac{\nu\sigma_0^2 + \sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\vert\mu, {\sigma_0^2}')$[4]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\, \beta$[note 4]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x} \mid \mu, \sigma^2 = \beta'/\alpha')$[4]
Normal[note 7]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\, \nu,\, \alpha,\, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n}, \nu+n, \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2 + \frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x} \mid \mu', \frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[4]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0,\, \nu,\, \alpha,\, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n}, \nu+n, \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2 + \frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x} \mid \mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[4]
Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0,\, \boldsymbol{\Sigma}_0$	$\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right),$ $\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions)$\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\tilde{\mathbf{x}} \mid {\boldsymbol{\mu}_0}', {\boldsymbol{\Sigma}_0}' + \boldsymbol{\Sigma})$[4]
Multivariate normal with known precision matrix $\boldsymbol{\Lambda}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0,\, \boldsymbol{\Lambda}_0$	$\left(\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Lambda}\bar{\mathbf{x}}\right), \left(\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda}\right)$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions)$\boldsymbol{\Lambda}_0$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}\left(\tilde{\mathbf{x}} \mid {\boldsymbol{\mu}_0}', {{\boldsymbol{\Lambda}_0}'}^{-1} + \boldsymbol{\Lambda}^{-1}\right)$[4]
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$ (covariance matrix)	Inverse-Wishart	$\nu,\, \boldsymbol{\Psi}$	$n+\nu, \boldsymbol{\Psi} + \sum_{i=1}^n (\mathbf{x_i}-\boldsymbol{\mu})(\mathbf{x_i}-\boldsymbol{\mu})^T$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\boldsymbol{\Psi}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\vert\boldsymbol{\mu}, \frac{1}{\nu'-p+1}\boldsymbol{\Psi}'\right)$[4]
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Lambda}$ (precision matrix)	Wishart	$\nu,\, \mathbf{V}$	$n+\nu, \left(\mathbf{V}^{-1} + \sum_{i=1}^n (\mathbf{x_i}-\boldsymbol{\mu})(\mathbf{x_i}-\boldsymbol{\mu})^T\right)^{-1}$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol{\mu}, \frac{1}{\nu'-p+1}\mathbf{V}'^{-1}\right)$[4]
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Sigma}$ (covariance matrix)	normal-inverse-Wishart	$\boldsymbol{\mu}_0,\, \kappa_0,\, \nu_0,\, \boldsymbol{\Psi}$	$\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0+n}, \kappa_0+n, \nu_0+n,$ $\boldsymbol{\Psi} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^n (\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\boldsymbol{\Psi} = \nu_0\boldsymbol{\Sigma}_0$	$t_{{\nu_0}'-p+1}\left(\tilde{\mathbf{x}}\vert{\boldsymbol{\mu}_0}', \frac{{\kappa_0}'+1}{{\kappa_0}'({\nu_0}'-p+1)}\boldsymbol{\Psi}'\right)$[4]
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Lambda}$ (precision matrix)	normal-Wishart	$\boldsymbol{\mu}_0,\, \kappa_0,\, \nu_0,\, \mathbf{V}$	$\frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{\kappa_0+n}, \kappa_0+n, \nu_0+n,$ $\left(\mathbf{V}^{-1} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C} = \sum_{i=1}^n (\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{{\nu_0}'-p+1}\left(\tilde{\mathbf{x}} \mid {\boldsymbol{\mu}_0}', \frac{{\kappa_0}'+1}{{\kappa_0}'({\nu_0}'-p+1)}\mathbf{V}'^{-1}\right)$[4]
Uniform	$U(0,\theta)$	Pareto	$x_m,\, k$	$\max\{x_1,\ldots,x_n,x_\mathrm{m}\}, k+n$	k observations with maximum value x_m	
Pareto with known minimum x_m	k (shape)	Gamma	$\alpha,\, \beta$	$\alpha+n, \beta+\sum_{i=1}^n \ln\frac{x_i}{x_\mathrm{m}}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)	
Weibull with known shape β	θ (scale)	Inverse gamma[3]	a, b	$a+n, b+\sum_{i=1}^n x_i^\beta$	a observations with sum b of the βth power of each observation	
Log-normal	Same as for the normal distribution after applying the natural logarithm to the data for the posterior hyperparameters. Please refer to Fink (1997, pp. 21–22) to see the details.					
Exponential	λ (rate)	Gamma	$\alpha,\, \beta$[note 4]	$\alpha+n, \beta+\sum_{i=1}^n x_i$	$\alpha-1$ observations that sum to β [5]	$\operatorname{Lomax}(\tilde{x} \mid \beta', \alpha')$ (Lomax distribution)
Gamma with known shape α	β (rate)	Gamma	$\alpha_0,\, \beta_0$	$\alpha_0+n\alpha, \beta_0+\sum_{i=1}^n x_i$	α_0/α observations with sum β_0	$\operatorname{CG}(\tilde{\mathbf{x}} \mid \alpha, {\alpha_0}', {\beta_0}') = \beta'(\tilde{\mathbf{x}}\vert\alpha, {\alpha_0}', 1, {\beta_0}')$ [note 8]
Inverse Gamma with known shape α	β (inverse scale)	Gamma	$\alpha_0,\, \beta_0$	$\alpha_0+n\alpha, \beta_0+\sum_{i=1}^n \frac{1}{x_i}$	α_0/α observations with sum β_0	
Gamma with known rate β	α (shape)	$\propto \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$	$a,\, b,\, c$	$a\prod_{i=1}^n x_i, b+n, c+n$	b or c observations (b for estimating α, c for estimating β) with product a	
Gamma[3]	α (shape), β (inverse scale)	$\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}}$	$p,\, q,\, r,\, s$	$p\prod_{i=1}^n x_i, q+\sum_{i=1}^n x_i, r+n, s+n$	α was estimated from r observations with product p; β was estimated from s observations with sum q	
Beta	α, β	$\propto \frac{\Gamma(\alpha+\beta)^k\, p^\alpha\, q^\beta}{\Gamma(\alpha)^k\,\Gamma(\beta)^k}$	$p,\, q,\, k$	$p\prod_{i=1}^n x_i, q\prod_{i=1}^n (1-x_i), k+n$	α and β were estimated from k observations with product p and product of the complements q	

Figure 2: Conjugate priors for continuous distributions.

Theorem 2 (Existence of deterministic Bayes Rules (Thm 1.3.2))**.**
If $L(a, \theta)$ is convex for all $\theta \in \Theta$, $\mathcal{D}$ is unrestricted, $\mathcal{A}$ is a convex set, and there exists a Bayes rule $D_\Pi \in \mathcal{D}$, then there exists a deterministic Bayes rule – i.e., a Bayes rule $D(\cdot, x)$ that places point mass at $a_x \in \mathcal{A}$.

Proof: by assumption, $D_\Pi \in \mathcal{D}$ is a Bayes rule. Choose $D_1(\cdot|x)$ to be a distribution that places a point mass at $\int aD_\Pi(a|x)$, the expected action under $D_\Pi(\cdot|x)$. Clearly, D_1 is deterministic because it is a point mass. D_1 is a Bayes rule because, via Jensen's inequality:

$$L(a,\theta)D_1(da|x) = \underbrace{L\left(\int aD_\Pi(da|x), \theta\right) \leq \int L(a,\theta)D_\Pi(da|x)}_{\text{Jensen}} \leq \int L(a,\theta)D(da|x)$$

Example 2 (Point estimation with squared error loss)**.**
Suppose our objective is to estimate $\Psi(\theta)$ with squared error loss $L : (a, \theta) \to \{a - \psi(\theta)\}^2$.

Let f_x be the Bayes risk function:

$$f_x : a \to \mathbb{E}[(a - \Psi(\theta))^2 | X = x]$$

Then the Bayes rule elects the action that minimizes the Bayes risk: $D_\Pi : x \to \underset{a \in \mathcal{A}}{\text{argmin}} f_x(a)$.

If we differentiate the Bayes risk function:

$$\frac{d}{da} f_x(a) = 2(a - \mathbb{E}(\Psi(\theta)|X = x))$$
$$\frac{d}{da} f_x(a) = 0 \implies \underset{a \in \mathcal{A}}{\text{argmin}} f_x(a) = \mathbb{E}(\Psi(\theta)|X = x)$$

Thus, the **posterior mean** is the Bayes rule under a squared loss.

For example, suppose:

$$X|\theta \sim \text{Poisson}(\theta) \qquad \theta \sim \text{Gamma}(\alpha, \beta)$$
$$\implies \theta | X = x \sim \text{Gamma}(\alpha + x, \beta + 1)$$

To estimate θ using mean squared error, the Bayes rule is the posterior mean, which is just a convex combination of the MLE and prior mean:

$$T_\Pi : x \to \frac{\alpha + x}{\beta + 1} = \frac{\beta}{\beta + 1}(\alpha/\beta) + \frac{1}{\beta + 1}x$$

Example 3 (Point estimation with absolute deviation loss)**.**
Suppose our objective is to estimate $\Psi(\theta)$ with squared error loss $L : (a, \theta) \to |\psi(\theta) - a|$.

Notice that $\frac{d|a|}{da} = \text{sign}(a)$. We want to find $T_\Pi : x \to \underset{a \in \mathcal{A}}{\text{argmin}}\ \mathbb{E}\left[\left\{|\psi(\theta) - a|\middle| X = x\right\}\right]$.

When we differentiate we obtain:

$$\frac{d}{da} f_x(a) = \mathbb{E}\left[\text{sign}(\psi(\theta) - a)\right] = 0$$
$$\implies a = \text{median}(\psi(\theta)|X = x)$$

2.3 Minimax framework

The minimax framework concerns itself with trying to find the decision rule with the smallest maximal risk. There are two key ways to construct minimax estimators, 1) via information theoretic approaches and 2) by building connections to Bayes rules. We focus on the second strategy.

Definition 5 (Minimax)**.**
The minimax framework posits that we should prefer decision rules with lower maximal risk. A **minimax rule** is optimal with respect to the maximal risk criterion, meaning it achieves the smallest maximal risk over all decision rules:

$$\sup_{\theta\in\Theta}\mathcal{R}(D^*,\theta) = \inf_{D\in\mathcal{D}}\sup_{\theta\in\Theta}\mathcal{R}(D,\theta)$$

Definition 6 (Least favorable prior)**.**
A prior is a least favorable prior if it (and its associated Bayes rule) yields the maximum Bayes risk:

$$r(D_{\Pi^*},\Pi^*) = \sup_{\Pi} r(D_\Pi,\Pi)$$

The following theorem establishes a connection between Bayes rules and Minimax rules.

Theorem 3 (Theorem 1.4.2)**.**
If Π and the Bayes rule D_Π have a bayes risk (optimal wrt Π) equal to the maximum risk of D_Π over all $\theta\in\Theta$ (maximmal), i.e.,

$$r(D_\Pi,\Pi) = \sup_{\theta} \mathcal{R}(D_\Pi,\theta)$$

Then,

1. D_Π is minimax.
2. If D_Π is a unique Bayes rule, then D_Π is the unique minimax.
3. Π is least favorable.

Proof:

1. Consider a general $D\in\mathcal{D}$, then:

$$\begin{aligned}\sup_{\theta\in\Theta}\mathcal{R}(D,\theta) &\geq \int_\Theta \mathcal{R}(D,\theta)\Pi(d\theta) \quad \text{(Max > average)}\\ &\geq \int_\Theta \mathcal{R}(D_\Pi,\theta)\Pi(d\theta) \quad (D_\pi \text{ optimal})\\ &= \sup_{\theta\in\Theta} R(D_\pi,\theta) \quad \text{(Assumed condition)}\end{aligned}$$

2. If D_π is unique, then $\int \mathcal{R}(D,\theta)\Pi(d\theta) > \int \mathcal{R}(D_\pi,\theta)\Pi(d\theta) \implies \sup_{\theta\in\Theta} R(D,\theta) > \sup_{\theta\in\Theta} R(D_\pi,\theta)$, showing D_π is unique minimax.

3. To prove Π is least favorable, consider another prior Π':

$$\begin{aligned} r(D_{\Pi'},\Pi') &\le r(D_\Pi,\Pi') \quad \text{(Bayes rule optimal)} \\ &\le \sup_{\theta\in\Theta} \mathcal{R}(D_\pi,\theta) \quad \text{(Max > average)} \\ &= r(D_\Pi,\Pi) \quad \text{(Theorem condition)} \end{aligned}$$

Theorem 4 (Corollary 1.4.3)**.**
If Π is a prior s.t. $\mathcal{R}(D_\Pi,\theta)$ is constant, i.e., $\mathcal{R}(D_\Pi,\theta)$ does not depend on θ, then D_Π is minimax.

Proof: trivial. If D_Π has constant risk than $r(D_\Pi,\Pi) = \sup_{\theta\in\Theta} \mathcal{R}(D_\Pi,\theta)$, so we can apply Theorem 1.4.2 to obtain minimaxity of D_Π.

By defining (least favorable) sequences of priors and taking the limit, we can begin to explore behavior of Bayes estimators under improper priors.

Definition 7 (Least favorable sequence)**.** Let $\{\Pi_k; k = 1, 2, \ldots\}$ be a sequence of priors on Θ and let:

$$r_0 := \liminf_{k\to\infty} r(D_{\Pi_k},\Pi_k)$$

A sequence is a least favorable prior sequence if $\forall\Theta$:

$$r(D_\pi,\pi) \le r_0$$

We can generalize Theorem 1.4.2 to the setting of prior sequences:

Theorem 5 (Theorem 1.4.7)**.**
Suppose $\{\Pi_k\}$ is a prior sequence and let r_0 be as defined in Definition 7. If $D \in \mathcal{D}$ satisfies:

$$\sup_{\theta\in\Theta} \mathcal{R}(D,\theta) = r_0$$

Then D is minimax, and $\{\Pi_k\}$ is a LFP sequence.

Proof: Consider a general decision $D' \in \mathcal{D}$ then for all $k = 1, 2, \ldots$:

$$\begin{aligned} \sup_\theta \mathcal{R}(D',\theta) &\ge \int_\Theta \mathcal{R}(D',\theta)\Pi_k(d\theta) \\ &\ge r(D_{\Pi_k},\Pi_k) \end{aligned}$$

Since this was true $\forall k$:

$$\begin{aligned}\sup_{\theta} \mathcal{R}(D', \theta) &\geq \liminf_{k\to\infty} r(D_{\Pi_k}, \Pi_k) \\ &= r_0 \\ &= \sup_{\theta\in\Theta} \mathcal{R}(D, \theta)\end{aligned}$$

Thus, D is minimax. To show that $\{\Pi_k\}$ is a LFP sequence, consider any Π:

$$\begin{aligned}r(D_\Pi, \Pi) &\leq r(D, \Pi) \\ &\leq \sup_{\theta} \mathcal{R}(D, \theta) = r_0\end{aligned}$$

Example 4 (Sample mean is Minimax under normal mean model)**.**
The general proof idea is to show that the risk of the sample mean is constant wrt θ, then define a prior sequence that achieves the bayes risk equal to the constant risk asymptotically with k. Then we can apply Theorem 1.4.7 to show minimax.

$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ^2 known (with squared error loss). We claim $\bar{X}_n$ is minimax. For $T : x \to \bar{x}_n$, $\mathcal{R}(T, \theta) = \sigma^2/n$ which is constant wrt θ, implying:

$$\sup_{\theta} \mathcal{R}(T, \theta) = \frac{\sigma^2}{n}$$

To show that $\bar{X}_n$ is minimax via Theorem 1.4.7, we need to find a prior sequence such that $r_0 := \lim_{k\to\infty} r(D_{\Pi_k}, \Pi_k) = \frac{\sigma^2}{n}$.

Let $\Pi_k := N(0, k)$. Under this model, the posterior distribution is:

$$\begin{aligned}\theta | X = x &\sim N\left(\frac{\bar{x}_n n/\sigma^2}{1/k + n/\sigma^2}, \frac{1}{1/k + n/\sigma^2}\right) \\ \implies T_{\Pi_k} &: x \to \frac{\bar{x}_n n/\sigma^2}{1/k + n/\sigma^2} \text{ is Bayes}\end{aligned}$$

Let $\mathbb{E}_k$ is the expectation with respect to $\theta \sim \Pi_k$ and $X|\theta \sim N(\theta, \sigma^2)$. Then asymptotically:

$$r(T_{\Pi_k}, \Pi_k) - \mathbb{E}_k\left[(\bar{x}_n - \theta)^2\right] = \mathbb{E}_k\left[\left(\frac{\bar{x}_n n/\sigma^2}{1/k + n/\sigma^2} - \theta\right)^2\right] - \mathbb{E}_k\left[(\bar{x}_n - \theta)^2\right] \overset{k\to\infty}{\longrightarrow} 0$$

Thus $\lim_{k\to\infty} r(T_{\Pi_k}, \Pi_k) = \lim_{k\to\infty} \mathbb{E}_k\left[(\bar{x}_n - \theta)^2\right] = \frac{\sigma^2}{n} = r_0$.

But we had just showed that $\sup_{\theta} \mathcal{R}(T, \theta) = \frac{\sigma^2}{n} = r_0$, implying that $T : x \to \bar{x}_n$ is minimax by Theorem 1.4.7.

Theorem 6 (Lemma 1.4.9)**.**
Let $\mathcal{P}_1 \subset \mathcal{P}_2$ denote two models. If D_1 is minimax over $\mathcal{P}_1$ and:

$$\sup_{P\in\mathcal{P}_1} \mathcal{R}(D_1, P) = \sup_{P\in\mathcal{P}_2} \mathcal{R}(D_1, P)$$

Then D_1 is also minimax over $\mathcal{P}_2$.

Proof: STAC that D_1 is not minimax over $\mathcal{P}_2$, then there exists $D_2 \in \mathcal{D}$ s.t. D_2 achieves a smaller worst risk in $\mathcal{P}_2$. Then:

$$\begin{aligned}\sup_{P \in P_1} \mathcal{R}(D_2, P) &\leq \sup_{P \in P_2} \mathcal{R}(D_2, P) && \text{(b/c } P_1 \subset P_2\text{)} \\ &< \sup_{P \in P_2} \mathcal{R}(D_1, P) && \text{(by } D_1 \text{ not minimax)} \\ &= \sup_{P \in P_1} \mathcal{R}(D_1, P) && \text{(by condition (ii) in theorem)}\end{aligned}$$

But this shows that D_1 is not minimax over $\mathcal{P}_1$, which is a contradiction. Thus, D_1 must be minimax over $\mathcal{P}_2$.

Example 5 (Sample mean minimax under bdd variance)**.**
If we consider $\mathcal{P}_2 = \{P = Q^n, \text{support}(Q) \subset \mathbb{R}, \text{Var}_Q(X) \leq \sigma^2\}$, then $\bar{X}_n$ is minimax wrt $\mathcal{P}_2$. This is because $\mathcal{R}(T, P) = \frac{\sigma^2}{n}$ which is independent of $\mathcal{P}_2$, therefore, the max risks are equal between $\mathcal{P}_1$ and $\mathcal{P}_2$.By Lemma 1.4.9, $\bar{X}_n$ is also minimax over $\mathcal{P}_2$.

2.4 Admissibility

Admissibility is the ”lowest-bar” criterion for an estimator or decision – essentially, there does not exist another rule that is uniformly as good or better based on the risk criterion.

Definition 8 (Admissibility)**.**
A minimal requirement for a good decision rule is that there does not exist a uniformly better rule. A rule D is called **inadmissible** if there exists another rule $\tilde{D}$ s.t.

$$\begin{aligned}&\mathcal{R}(\tilde{D}, \theta) \leq \mathcal{R}(D, \theta) \text{ for all } \theta \in \Theta, \text{ and} \\ &\mathcal{R}(\tilde{D}, \tilde{\theta}) < \mathcal{R}(D, \tilde{\theta}) \text{ for some } \tilde{\theta} \in \Theta\end{aligned}$$

The rule is called **admissible** otherwise.

Definition 9 (Uniqueness of Bayes and Minimax Rules)**.**
For a prior Π, a rule D_Π is **unique Bayes** if a rule is Bayes iff it is equal to D_Π a.e. P_θ.
A rule $D*$ is **unique minimax** if a rule is minimax iff it is equal to $D*$ a.e. P_θ.

Theorem 7 (Admissibility of unique Bayes/minimax rules: Theorems 1.5.2-1.5.4)**.**
Any unique Bayes/minimax rule is admissible.

Proof: 1. STAC that D_Π is unique Bayes but not admissible. Then $\exists D \in \mathcal{D}$ s.t.

$$\mathcal{R}(D,\theta) \leq \mathcal{R}(D_\Pi,\theta) \; \forall \; \theta \in \Theta$$
$$\mathcal{R}(D,\theta*) < \mathcal{R}(D_\Pi,\theta*) \text{ for some } \theta* \in \Theta$$

However, this implies $r(D,\Pi) \leq r(D_\Pi,\Pi) \implies r(D,\Pi) = r(D_\Pi,\Pi)$ since the Bayes rule is optimal. However, this implies that the Bayes rule is not unique, yielding a contradiction, and showing that a unique Bayes rule must be admissible.
2. STAC that D^* is unique minimax but not admissible. Then $\exists D \in \mathcal{D}$ s.t.

$$\mathcal{R}(D,\theta) \leq \mathcal{R}(D_\Pi,\theta) \; \forall \; \theta \in \Theta$$
$$\mathcal{R}(D,\theta*) < \mathcal{R}(D_\Pi,\theta*) \text{ for some } \theta* \in \Theta$$

But since D^* is minimax ($\sup\limits_\theta \mathcal{R}(D^*,\theta) = \inf\limits_D \sup\limits_\theta \mathcal{R}(D,\theta)$), $\sup\limits_\theta \mathcal{R}(D,\theta) = \sup\limits_\theta \mathcal{R}(D^*,\theta)$, because D^* achieves optimal max risk and D is uniformly as good or better by the risk criterion. However, this yields a contradiction, because we showed two distinct rules yield the same minimax risk despite assuming D^* was unique minimax. Thus, we conclude D^* is unique minimax.

Unique Bayes/minimax rules guarantee admissibility! How do we find these rules? Some helpful theorems will come in handy!

Theorem 8 (Unique bayes rule: Theorem 1.5.5)**.**
Let Π be a prior and D_Π be the associated Bayes rule. If the following hold:

(i) The loss function is squared error loss

(ii) $r(D_\Pi,\Pi) < \infty$

(iii) $P_\theta << Q$ (probability measure is absolutely continuous wrt some marginal measure): for any subset A of the σ-algebra $\mathcal{A}$, $Q(A) \equiv \int P_\theta(X \in A) d\Pi(\theta) = 0 \implies P_\theta(X \in A) = 0$ for all $\theta \in \Theta$

Then D_Π is unique Bayes.
Proof left available in paper by Larry Brown.
Addendum: a sufficient condition for item (iii) is that as long as we can find a new measure η (not necessarily a prob measure) on the measure space, s.t.

$$P_\theta << \eta \quad \text{AND} \quad \eta << P_\theta$$

Condition (iii) holds. A useful example of this is that the normal distribution is absolutely continuous wrt the Lesbegue measure and vice versa.

Proof: Fix $\theta_0 \in \Theta$. The goal is to show $P_{\theta_0}(A) > 0 \implies Q(A) > 0$. Suppose $P_{\theta_0}(A) > 0$, because lemma conditions say:

$$P_{\theta_0} << \eta \implies \text{if } \eta(A) > 0 \implies P_\theta(A) > 0 \; \forall \; \theta \in \Theta$$
$$\implies Q(A) = \int_\Theta \underbrace{P_\theta(A)}_{>0} \Pi(d\theta) > 0$$

Example 6 (Bayesian normal mean/sample mean is admissible)**.**
$X_1, \ldots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$

We can show that $T_\Pi : x \to (1-P_n)\cdot \bar{X}_n + P_n \cdot \mu$ with $P_n := \frac{1/\tau^2}{1/\tau_n^2/\sigma^2} \in (0,1)$ is admissible using Thoerem 1.5.5. (i) is trivial, (ii) follows because optimal, and the Addendum and knowing that $P_\theta << \lambda$ and $\lambda << P_\theta$ where λ is the Lesbesgue measure and P_θ is $N(\theta, \sigma^2)$ shows that it is unique Bayes and therefore admissible.

However, this begs the question of whether the Bayesian normal mean is admissible when $P_n = 0, 1$. When $P_n = 0$, $T : x \to \mu$ is admissible because it is a constant estimator that achieves Risk of 0 when the θ is μ. When $P_n = 1$, $T : x \to \bar{x}_n$ is minimax and admissible, but this requires a rigorous proof.

Proof: Claim: $\bar{X}_n$ is admissible in normal mean model. $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where σ^2 known and $\theta \in \mathbb{R}$.

For all rules, we will establish admissibility by proving:

(a) $\mathcal{R}(T_1, \theta) \geq \mathcal{R}(T, \theta)\ \forall\ \theta$

(b) there exists $\theta \in \Theta$ for which $\mathcal{R}(T_1, \theta) > \mathcal{R}(T, \theta)$.

WLOG consider $\sigma^2 = 1$. Then there exists $\theta_1 \in \Theta$ at which:

$$\mathcal{R}(T_1, \theta_1) < \mathcal{R}(T, \theta_1)$$

Since the risk function is continuous, we can build a δ-bubble around θ_1 where the risk difference is greater than some ϵ:

$$\mathcal{R}(T_1, \theta) < \mathcal{R}(T, \theta) - \epsilon = \frac{1}{n} - \epsilon \qquad \text{for } \theta \in (\theta_1 - \delta, \theta_1 + \delta)$$

Let's specify a prior $\Pi_\tau \equiv N(0, \tau^2)$ and let $T_\tau := T_{\Pi_\tau}$ be the Bayes estimator wrt this prior. Via some algebra, we obtain:

$$r(T_\tau, \Pi_\tau) - \mathcal{R}(T, \theta) = \frac{\tau^2}{1+n\tau^2} - \frac{1}{n} = -\frac{1}{n(1+n\tau^2)}$$

Thus,

$$\begin{aligned}
-\frac{1}{n(1+n\tau^2)} &= r(T_\tau, \Pi_\tau) - \frac{1}{n} \\
&\leq r(T_1, \Pi_\tau) - \frac{1}{n} \qquad \text{(B/c optimality bayes rule)} \\
&= \int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right] \Pi_\tau(d\theta) \\
&= \int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^+ \Pi_\tau(d\theta) - \int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- \Pi_\tau(d\theta) \qquad \text{(Splits pos and neg regions)}
\end{aligned}$$

Recall that for $\theta_1 \in (\theta - \delta, \theta + \delta)$, $R(T_1, \theta_1) < \frac{1}{n} - \epsilon \implies R(T_1, \theta) - \frac{1}{n} < -\epsilon \implies \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- > \epsilon$. Therefore:

$$\begin{aligned}
\int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- d\Pi_\tau(\theta) &\geq \int_{\theta_1 - \delta}^{\theta_1 + \delta} \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- d\Pi_\tau(\theta) \\
&\geq \epsilon \int_{\theta_1 - \delta}^{\theta_1 + \delta} d\Pi_\tau(\theta) = \epsilon \Pi_\tau(\theta_1 - \delta \leq \Theta \leq \theta_1 - \delta)
\end{aligned}$$

Implying:

$$\int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^+ \Pi_\tau(d\theta) \geq -\frac{1}{n(1+n\tau^2)} + \epsilon \Pi_\tau(\theta_1 - \delta \leq \Theta \leq \theta_1 - \delta) := L(\tau)$$

Note that:

$$\sqrt{2\pi}\tau L(\tau) \overset{\tau\to\infty}{\longrightarrow} 2\epsilon\delta$$

Because $\frac{-\sqrt{2\pi}\tau}{n(1+n\tau^2)} \overset{\tau\to\infty}{\longrightarrow} 0$ by LH rule and $\sqrt{2\pi}\tau\Pi_\tau(\theta_1 - \delta \leq \Theta \leq \theta_1 - \delta) = \int_{\theta_1-\delta}^{\theta_1+\delta} \exp\left(-\frac{1}{2\tau^2}\theta^2\right) = 2\delta e^{-\theta^2/2\tau^2} \overset{\tau\to\infty}{\longrightarrow} 2\delta$. Now choose $\tau_0 > 0$ s.t. $\sqrt{2\pi}\tau_0 L(\tau) > \delta\epsilon$ and plug-in:

$$\int \left[\mathcal{R}(T_1,\theta) - \frac{1}{n}\right]^+ \Pi_\tau(d\theta) \geq L(\tau_0) > \frac{\delta\epsilon}{\sqrt{2\pi}\tau_0} > 0$$

We just showed that there exists a θ for which $R(T_1,\theta) > R(T,\theta) = 1/n$ implying that condition (b) holds. So either T has uniformly lower/equal risk (as in case (a)) or T has a lower risk for some θ (as in case (b)) for a general estimator T_1. Thus, T is admissible!

2.5 Inadmissibility of the sample mean in dim ≥ 3

Turns out that the sample mean is inadmissible in higher dimensions under MSE loss, because as the dimension increases, the distribution's mass gets pulled further into the tails and the sample mean vector increasingly deviates from the true mean.

Under $X_1,\ldots,X_n \overset{iid}{\sim} N(\theta,\sigma^2 I_d)$, Charles Stein proved that **T is inadmissible when** $d \geq 3$ by introducing an estimator that dominated T under $d \geq 3$, the **James-Stein estimator**.

Definition 10 (James-Stein Estimator)**.** The James-Stein estimator is as follows

$$T^{JS} : x \to \begin{cases} \left(1 - \frac{(d-2)}{n||\bar{x}_n||^2}\right)\bar{x}_n & \text{if } \bar{x}_n \neq (0,\ldots,0) \\ 0, & \text{otherwise} \end{cases}$$

Note that it is just a shrinkage of the sample mean estimator towards 0, where shrinkage is controlled by the dimension, sample size, and $||\bar{x}_n||$.
Important: despite the observations being independent, the shrinkage property ensures that the estimate of θ_j depends on X_k despite X_j being independent of X_k and θ_j and θ_k being variationally independent.

Theorem 9 (Stein's Lemma (Lemma 1.6.2))**.**
Let $Y \sim N(\mu,\sigma^2 I_d)$ and let $g_1,\ldots,g_d$ be $\mathbb{R}^d \to \mathbb{R}$ functions such that $\mathbb{E}\left|\frac{\partial}{\partial y_j} g_j(y)|_y = Y\right| < \infty$. Defining $g : y \to (g_1(y),\ldots,g_d(y))$ we have:

$$\mathbb{E}\left[\langle g(Y), Y-\mu\rangle\right] = \sigma^2\mathbb{E}[\nabla \cdot g(Y)] \quad \text{where } \nabla \cdot g(Y) \equiv \sum_{j=1}^{d} \frac{\partial}{\partial y_j} g_j(y)$$

Proof: Step 1 is to show for $Y \sim N(\mu,\sigma^2)$ and $g : \mathbb{R} \to \mathbb{R}$ s.t. $\mathbb{E}(g'(Y)) < \infty$:

$$\mathbb{E}[g(Y)(Y-\mu)] = \sigma^2\mathbb{E}[g'(Y)]$$

Note:

$$\begin{aligned}
\mathbb{E}[g(Y)(Y-\mu)] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} g(x)(y-\theta)e^{-(y-\theta)^2/(2\sigma^2)}dy \\
&= \frac{1}{\sqrt{2\pi}\sigma}\left[\underbrace{-\sigma^2 g(y)e^{-(y-\theta)^2/(2\sigma^2)}|_{-\infty}^{\infty}}_{=0} + \sigma^2\int_{-\infty}^{\infty} g'(y)e^{-(y-\theta)^2/(2\sigma^2)}dy\right] \\
&= \sigma^2\mathbb{E}[g'(y)]
\end{aligned}$$

By Fubini's theorem, multiple integrals over each dimension can be reduced to iterated integrals because the expected absolute value partial derivatives are all finite (condition required for Fubini's theorem). This yields the multivariate result!

Theorem 10 (James-Stein estimator dominates the sample mean)**.** Consider the simplifed setting where $X_1, \ldots, X_n \sim N(\theta, I_d)$ and the J-S estimator takes the form:

$$T^{JS} : x \to \begin{cases} \left(1 - \frac{d-2}{||x||^2}\right)x & \text{if } x \neq (0, \ldots, 0) \\ 0 & \text{else} \end{cases}$$

Then we write the risk as:

$$\begin{aligned}
\mathcal{R}(T^{JS}, \theta) &= \mathbb{E}[||\tau^{JS}(||X||)X - \theta||^2] \\
&= \mathbb{E}[||[\tau^{JS}(||X||) - 1]X + [X - \theta]||^2] \\
&= \mathbb{E}[||[\tau^{JS}(||X||) - 1]X||^2] + \mathbb{E}[||X - \theta||^2] - 2\mathbb{E}[\langle [1 - \tau^{JS}(||X||)]X, X - \theta\rangle] \\
&= \mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right] + \mathcal{R}(T, \theta) - 2(d-2)\mathbb{E}_\theta\left[\left\langle \frac{X}{||X||^2}, X - \theta\right\rangle\right]
\end{aligned}$$

The goal will be to show that term 3 is equal to $-2\mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right]$ which shows that $\mathcal{R}(T^{JS}, \theta) < \mathcal{R}(T, \theta)$. We so via Stein's Lemma, letting $g_j : x \to \frac{z}{||z||^2}$, we see that:

$$\begin{aligned}
\nabla \cdot g(z) &= \sum_{j=1}^{d}\left[\frac{1}{||z||^2} - \frac{2z_j^2}{||z||^4}\right] \qquad \text{(Quotient rule)} \\
&= \frac{d}{||z||^2} - \frac{2||z||^2}{||z||^4} = \frac{d-2}{||z||^2}
\end{aligned}$$

Thus, the third term is equal to $-2(d-2)\mathbb{E}(\nabla \cdot g(y)) = -2(d-2)^2\mathbb{E}\left[\frac{1}{||X||^2}\right]$ as desired, proving $\mathcal{R}(T^{JS}, \theta) < \mathcal{R}(T, \theta)$ for all θ.

3 Elementary Asymptotics

Asymptotic statistics allows us to evaluate statistical procedures on the basis of increasing sample size and repeated sampling from the superpopulation.

3.1 Modes of convergence

Definition 11 (Convergence almost surely, in probability, and in distribution)**.** The following definitions concern a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ and a random variable X defined on a common probability space $(\Omega, \mathcal{F}, P)$.

Almost surely convergence: a sequence of random variables converges almost surely to X if:

$$P(\lim_{n\to\infty} ||X_n - X|| = 0) = 1 \qquad \equiv \qquad \lim_{n\to\infty} ||X_n(\omega) - X(\omega)|| = 0$$

Convergence in probability: a sequence of random variables converges in probability to X if:

$$P(||X_n - X|| > \epsilon) \overset{n\to\infty}{\longrightarrow} 0$$

Convergence in distribution: a sequence of random variables converges in distribution/weakly converges to X iff or all bounded continuous functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \text{ as } n \to \infty$$

Note that boundedness and continuity are essential to the definition of weak convergence.

- Suppose we didn't require f be continuous. Let $X_n = 1/n$ and let $f : a \to I(a > 0)$. $\mathbb{E}(f(X_n)) \to 1$ which does not equal $\mathbb{E}[f(X)] = 0$, so X_n would not converge in distribution to 0 despite converging in every other sense (a.s., in prob).
- Suppose we didn't require f. to be bounded. Let $X_n = n$ w.p, $1/n$ and $X_n = 0$ otherwise. Note that X_n converges in probability to 0 (b/c the probability mass increasingly gets concentrated at 0 as $n \to \infty$. Let $f : a \to \min(|a|, 1)$ and $g : a \to a$. Then $\mathbb{E}[f(X_n)] = 1/n \to 0$ where the limit equals $\mathbb{E}[f(X)]$ iff $X \overset{a.s.}{=} 0$. However if $X \overset{a.s.}{=} 0$, $\mathbb{E}[g(X_n)] \to 1$ and $\mathbb{E}[g(X)] \neq 1$. Thus, convergence in distribution can't hold despite convergence in prob.

Note in addition that convergence a.s. $\implies$ convergence in prob $\implies$ weak convergence.

The Portmanteau theorem provides linkage between the many definitions of convergence in distribution!

Theorem 11 (Portmanteau)**.**
Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X be a random variable. TFAE:

1. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$ for all bounded continuous functions f.
2. For all continuity points $t \in \mathbb{R}^d$, $P(X_n \leq t) \to P(X \leq t)$ as $n \to \infty$.
3. . . . MANY others
4. **Levy's continuity theorem**: for all $t \in \mathbb{R}^d$, $\mathbb{E}[\exp(it^T X_n)] \to \mathbb{E}[\exp(it^T X)]$, convergence in characteristic functions
5. **Cramer-Wold device**: for all $t \in \mathbb{R}^d$, $t^T X_n \implies t^T X$

3.2 Continuous Mapping Theorem and Slutsky's Theorem

CMT and Slutsky's theorem allows us to describe the behavior of functions of convergent sequences!

Theorem 12 (Continuous Mapping Theorem)**.** Let X_n be a $\mathbb{R}^d$-valued sequence of random variables and $g : \mathbb{R}^d \to \mathbb{R}^m$ be continuous at every point of a set C s.t. $P(X \in C) = 1$, the following are valid:

(i) if $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$

(ii) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$

(iii) if $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$

(iv) if $X_n \Rightarrow X$ and $||X_n - Y_n|| \xrightarrow{p} 0$, $Y_n \Rightarrow X$

(v) of $X_n \Rightarrow X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \Rightarrow (X, c)$

Theorem 13 (Slutsky's Lemma)**.** Let X_n be a $\mathbb{R}^d$-valued sequence of random variables and $X_n \Rightarrow X$. If the $\mathbb{R}^d$-valued random variable Y_n satisfies $Y_n \overset{p}{c}$ for a constant c, then the following are valid

(i) $X_n + Y_n \Rightarrow X + c$

(ii) $X_n \cdot Y_n \Rightarrow c \cdot X$

(iii) $X_n / Y_n \Rightarrow X/c$ if $c \neq 0$

3.3 Law of Large Numbers and Central Limit Theorem

Law of large numbers allows us to describe the consistency of the sample mean, while the central limit theorem yields the asymptotic normal distribution of the sample mean.

Theorem 14 (Law of large numbers)**.** For $X_1, \ldots, X_n \overset{iid}{\sim} P$ and letting $\bar{X}_n = \frac{1}{n}\sum X_i$,

$$\textbf{WLLN: } \text{if } \mathbb{E}_P|X| < \infty, \bar{X}_n \xrightarrow{p} \mathbb{E}_P[X]$$
$$\textbf{SLLN: } \text{if } \mathbb{E}_P|X| < \infty, \bar{X}_n \xrightarrow{a.s.} \mathbb{E}_P[X]$$

Theorem 15 (Univariate CLT)**.**
For the moment, we assume that the CLT under a univariate parameter drawn iid from a fixed distribution P. If $\mathbb{E}_P(X^2) < \infty$, then where $\sigma_P^2 := \mathrm{Var}_P(X)$

$$\sqrt{n}(\bar{X}_n - \mathbb{E}_p[X]) \Rightarrow N(0, \sigma_P^2)$$

Example 7 (t-statistics)**.**
The goal is to show that the t-statistic: $\sqrt{n}\bar{X}_n/S_n \Rightarrow N(0,1)$ when $\mathbb{E}(X) = 0$. To do this, we write:

$$t - \frac{\sqrt{n}\mathbb{E}(X)}{S_n} = \sqrt{n}\frac{\bar{X}_n - \mathbb{E}(X)}{S_n} \Rightarrow \frac{N(0,\sigma^2)}{S_n}$$

Now it suffices to show that $S_n \overset{p}{\to} \sigma \equiv S_n^2 \overset{p}{\to} \sigma^2$. We do so via WLLN: $\bar{X}_n \overset{p}{\to} \mathbb{E}_p(X)$ and $\frac{1}{n}\sum X_i^2 \overset{p}{\to} \mathbb{E}(X^2)$, and $\frac{n}{n-1} \overset{p}{\to} 1$. Hence:

$$\left(\bar{X}_n, Y_n, \frac{n}{n-1}\right) \overset{p}{\to} (\mathbb{E}(X), \mathbb{E}(X^2), 1)$$
$$\implies S_n^2 = \frac{n}{n-1}[\frac{1}{n}\sum X_i^2 - \bar{X}_n^2] \overset{p}{\to} \sigma^2 \qquad \text{(by CMT)}$$

Thus, by CMT $S_n \to \sigma$, yielding that $t \to N(0,1)$ when $E(X) = 0$.

3.4 Stochastic Order Notation and Prokhorov's Theorem

Suppose we have two real valued sequences of random variables X_n and R_n and we which to compare the magnitude of the two sequences as $n \to \infty$.

Definition 12 (Big-O and little-o notation)**.**

(i) $X_n = O_p(R_n)$ means that X_n *is within a multiplicative constant of* R_n, i.e., x_n variable is stochastically bounded. In other words

$$P\left(\left|\frac{X_n}{R_n}\right| > \delta\right) < \epsilon, \forall n > N$$

Meaning we can find a tail in the sequence ($n > N$) such that the probability of the ratio being larger than some constant number (δ) is essentially 0. In other words, x_n is asymptotically within a finite constant of r_n. Equivalently, for all $\epsilon > 0$, there exists an $M > 0$ s.t. $\liminf_{n\to\infty} P(|X_n| \leq M|R_n|) \overset{n\to\infty}{\longrightarrow} 1-$

(ii) $X_n = o(r_n)$ means that x_n *grows more slowly than* r_n and refers to convergence in probability towards 0. $X_n = o_p(1)$ means:

$$\lim_{n\to\infty} P(|X_n| \geq \epsilon) = 0 \forall, \epsilon > 0 \implies X_n \overset{p}{\to} 0$$

While $X_n = o_p(r_n)$ means:

$$\frac{x_n}{r_n} = o_p(1) \implies \lim_{n\to\infty} P(\left|\frac{X_n}{r_n}\right| \geq \epsilon) = 0 \forall, \epsilon > 0 \implies \frac{X_n}{r_n} \xrightarrow{p} 0$$

Equivalently, for all $M > 0$ s.t. $P(|X_n| \leq M|R_n|) \overset{n\to\infty}{\longrightarrow} 1$

The Prokhorov theorem shows that if a sequence converges in distribution then it is bounded in probability (uniformly tight), and if it is bounded in probability (uniformly tight), it has a convergent subsequence (ala Bolzano-Weierstrass).

Theorem 16 (Prokhorov)**.**
Let X_n be a random vector in $\mathbb{R}^p$.

(i) If $X_n \Rightarrow X$ for some X, then $X_n = O_p(1)$

(ii) If $X_n = O_p(1)$, there exists a subsequence $\{X_{n_j}\} \subset \{X_n\}$ such that $X_{nj} \Rightarrow X$ for some X.

Note that $X_n = O_p(1)$ means is referred to as the sequence being uniformly tight.

Theorem 17 (Operations using big/little-o notation)**.**

1. $X_n = o_P(R_n)$ iff $X_n = R_n Y_n$ for some $Y_n = o_P(1)$
2. $X_n = O_P(R_n)$ iff $X_n = R_n Y_n$ for some $Y_n = O_P(1)$
3. $o_P(1) + o_P(1) = o_P(1)$ (sum of two things that conv in prob to 0)
4. $o_P(1) + O_P(1) = O_P(1)$ (sum of thing that conv in prob to 0 and bdd in prob)
5. $O_P(1) \cdot O_P(1) = O_P(1)$ (product of two things bdd in prob)
6. $o_P(1) \cdot O_P(1) = o_P(1)$ (product of conv in prob to 0 and bdd in prob)
7. $[1 + o_P(1)]^{-1} = O_P(1)$
8. $X_n = o_P(1) \implies X_n = O_P(1)$ (conv in prob to 0 implies bdd in prob)

We can use big/little-o notation to explore the convergence rates of common estimators!

Example 8 (Convergence rates of sample mean and variance)**.**
Sample mean: Suppose $X_1, \dots, X_n \overset{iid}{\sim} P$ and $\mathbb{E}_p[X^2] < \infty$. By CLT, $\sqrt{n}(\bar{X}_n - \mathbb{E}_P[X]) \Rightarrow N(0, \text{Var}_P(X))$. By Prokhorov's theorem, this implies $\sqrt{n}(\bar{X}_n - \mathbb{E}_P[X]) = O_P(1)$ yielding $\bar{X}_n - \mathbb{E}_P(X) = O_p(n^{-1/2})$, the convergence rate of the sample mean to the population mean based on CLT.

Sample variance: Suppose $X_1, \dots, X_n \overset{iid}{\sim} P$ and $\mathbb{E}_p[X^4] < \infty$. We also know that:

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

Where by CLT: $\frac{1}{n}\sum_{i=1}^{n}(X_i^2) = \mathbb{E}_P(X^2) + O_p(n^{-1/2})$ and $\bar{X}_n = \mathbb{E}_P(X) + O_p(n^{-1/2})$ and $(\bar{X}_n)^2 = \left(\mathbb{E}_P(X) + O_p(n^{-1/2})\right)^2 = \mathbb{E}_P(X)^2 + O_p(n^{-1/2}) + O_p(n^{-1})$ and since $O_p(n^{-1/2}) + O_p(n^{-1}) = O_p(n^{-1/2})$, then:

$$\begin{aligned} S_n^2 &= \frac{n}{n-1}(\mathbb{E}_p[X^2] + O_p(n^{-1/2}) - (\mathbb{E}_P(X)^2 + O_p(n^{-1/2}))) \\ &= \frac{n}{n-1}\left(\text{Var}_P(X) + O_P(n^{-1/2})\right) \end{aligned}$$

Thus, $S_n^2 - \text{Var}_p(X) = O_p(n^{-1/2})$

3.5 Multivariate and Lindeberg Feller CLT

We presented the Central Limit Theorem only in the case of univariate observations iid from some fixed distribution. Here, we generalize the result to multivariate iid observations and general independent observations (not required to be identically distributed).

Theorem 18 (Multivariate CLT)**.** Suppose $X_1, \dots, X_n \overset{iid}{\sim} P$ where P is fixed distribution with support in $\mathbb{R}^d$ and $\mathbb{E}_P[||X||^2]$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0_d, \Sigma)$$

Where $\mu := \mathbb{E}[X]$ and $\Sigma := \mathbb{E}_P[(X-\mu)(X-\mu)^T]$

Proof: use the Cramer-Wold device. Fix $t \in \mathbb{R}^d$ and $\bar{Y}_n = t^T \bar{X}_n$ meaning $\bar{Y}_n$ is the sample mean of the iid observations $Y_i := t^T X_i$ for all observations X_i.

We can confirm the bounded second moment of Y_1 via the Cauchy Schwartz inequality ($E[XY] \leq \sqrt{E(X^2)E(Y^2)}$):

$$\mathbb{E}[Y_1] = \sum_j^d \sum_k^d t_j t_k \mathbb{E}[X_{1j}, X_{1k}] \overset{C-S}{\leq} \sum_j^d \sum_k^d t_j t_k \mathbb{E}[X_{1j}^2]^{1/2} \cdot \mathbb{E}[X_{1k}^2]^{-1/2}$$
$$\leq \mathbb{E}[||X_1||^2] \sum_{j=1}^d \sum_{k=1}^d |t_j t_k| < \infty$$

Thus, we can apply the univariate CLT, where $\mathbb{E}(Y_1) = t^T\mu$, and $\text{Var}(Y_1) = \text{Var}(t^T Y_1) = t^T \text{Var}(X_1) t = t^T \Sigma t$

$$\sqrt{n}(\bar{X}_n - t^T\mu) \Rightarrow N(0, t^T\Sigma t)$$

Thus,

$$t^T[\sqrt{n}(\bar{Y}_n - \mu)] \Rightarrow N(0, t^T\Sigma t)$$
$$\implies \sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0, \Sigma)$$

Can we generalize to independent (but not identical) observations? The LF CLT is the way to do it!

Theorem 19 (Lindeberg-Feller CLT)**.** For each n, let $\{X_{ni}\}_{i=1}^n$ be an independent collection of $\mathbb{R}$-valued random variables. Suppose that the means $\mu_{ni} := \mathbb{E}(X_{ni})$ and variances $\sigma_{ni}^2 := \text{Var}(X_{ni})$ exist and are finite. Suppose $\sigma_n^2 := \sum_{i=1}^n \sigma_{ni}^2 > 0$ for all n. Finally let $Y_{ni} = (X_{ni} - \mu_{ni})/\sigma_n$. If the Lindeberg condition holds:

$$\text{for all } \epsilon > 0, \quad \sum_{i=1}^n \mathbb{E}[Y_{ni}^2 I(|Y_{ni} \geq \epsilon)] \overset{n\to\infty}{\longrightarrow} 0$$

then

$$\sum_{i=1}^n Y_{ni} \Rightarrow N(0,1)$$

Note: we can also replace the Lindeberg condition with the Lyapunov condition:

$$\sum_{i=1}^n \mathbb{E}[|Y_{ni}^{2+\delta}|] \overset{n\to\infty}{\longrightarrow} 0 \text{ for some } \delta > 0$$

Example 9 (OLS is ASN)**.**
We can use the LF-CLT to show that the OLS estimator under a fixed design, $\hat{\beta}$ is ASN. Using the orthogonal decomposition of the OLS estimator we obtain:

$$\begin{aligned}\hat{\beta} &= (X_n^TX_n)^{-1}X_n^TY = (X_n^TX_n)^{-1}X_n^T(X\beta+\epsilon_n)\\ &= \beta + (X_n^TX_n)^{-1}X_n^T\epsilon_n\\ &\implies (X_n^TX_n)^{1/2}(\hat{\beta}-\beta) = (X_n^TX_n)^{-1/2}x_n^T\epsilon_n\end{aligned}$$

Goal is to show RHS converges weakly to $N(0,\sigma^2 I_{d+1})$ random variable. We execute the proof in steps:

1. Cramer-Wold device: for a_{ni} being the ith column of $(X_n^TX_n)^{-1/2}x_n^T$:

   $$t^T(X_n^TX_n)^{-1/2}x_n^T\epsilon_n = \sum_{i=1}^n (t^Ta_{ni})\epsilon_i$$

   and observe that:

   $$\sigma_{ni}^2 := \text{Var}([t^Ta_{ni}]\epsilon_i) = [t^Ta_{ni}]^2\text{Var}(\epsilon_i) = [t^Ta_{ni}]^2\sigma^2$$

   Hence:

   $$\begin{aligned}\sigma_n^2 &= \sum_{i=1}^n \sigma_{ni}^2 = \sigma^2\sum_{i=1}^n [t^Ta_{ni}]^2\\ &= \sigma^2 t^T(X_n^TX_n)^{-1/2}X_n^TX_n(X_n^TX_n)^{-1/2}t = \sigma^2||t||^2\end{aligned}$$

2. Defined scaled RV and write out Lindeberg condition: Define $Z_{ni} = \frac{(t^Ta_{ni})\epsilon_i}{\sigma_n}$. For fixed $\epsilon > 0$:

   $$\sum_{i=1}^n \mathbb{E}[Z_{ni}^2 I(|Z_{ni}| \geq \epsilon)]$$

3. Simplify and eliminate randomness in i by taking the max:

   $$\begin{aligned}\sum_{i=1}^n \mathbb{E}[Z_{ni}^2 I(|Z_{ni}| \geq \epsilon)] &= \sigma_n^{-2}\sum_{i=1}^n (t^Ta_{ni})^2\mathbb{E}[\epsilon_i^2 I(|t^Ta_{ni}||\epsilon_i| \geq \sigma_n\epsilon)]\\ &= \sigma_n^{-2}\left[\sum_{i=1}^n (t^Ta_{ni})^2\right]\max_i \mathbb{E}[\epsilon_i^2 I(|t^Ta_{ni}||\epsilon_i| \geq \sigma_n\epsilon)]\\ &= \sigma^{-2}\max_i \mathbb{E}[\epsilon_i^2 I(|t^Ta_{ni}||\epsilon_i| \geq \sigma_n\epsilon)] \quad (\text{B/c } \sigma_n^2 = \sigma^2||t||^2)\end{aligned}$$

4. Cauchy-Schwartz to the expected value statement

   $$\begin{aligned}\max_i \mathbb{E}[\epsilon_i^2 I(|t^Ta_{ni}||\epsilon_i| \geq \sigma_n\epsilon)] &\overset{C-S}{\leq} \max_i \mathbb{E}[\epsilon_i^2 I(||t||||a_{ni}||\epsilon_i| \geq \sigma_n\epsilon)]\\ &= \max_i \mathbb{E}[\epsilon_i^2 I(||t||||a_{ni}||\epsilon_i| \geq \sigma\epsilon)] \quad (\text{b/c } \sigma_n = \sigma||t||)\end{aligned}$$

5. Dominated convergence theorem (DCT) to establish convergence criterion

   $$\max_i \mathbb{E}[\epsilon_i^2 I(||t||\cdot||a_{ni}|\cdot|\epsilon_i| \geq \sigma\epsilon)] << \mathbb{E}[\epsilon_i^2 I(\max_i ||a_{ni}|| > \epsilon)]$$

Thus, if $\max_i ||a_{ni}||$ i.e., the **maximum leverage goes to 0**, then by DCT, the Lindeberg condition holds. This implies

$$t^T(X_n^TX_n)^{-1/2}x_n^T\epsilon_n \Rightarrow N(0,\sigma^2||t||^2) \equiv N(0,t^T[\sigma^2 I_{d+1}]t)$$

Thus,

$$\begin{aligned}&(X_n^TX_n)^{-1/2}x_n^T\epsilon_n \Rightarrow N(0,\sigma^2 I_{d+1})\\ \implies &(X_n^TX_n)^{1/2}(\hat{\beta}-\beta) \Rightarrow N(0,\sigma^2 I_{d+1})\end{aligned}$$

3.6 Multivariate delta method

Suppose that $X_1,\dots,X_n \overset{iid}{\sim} P_{\theta_0}$ from a collection of distributions $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$ with support on $\mathbb{R}^d$. Suppose $\psi \equiv \Psi(\theta_0) \in \mathbb{R}^d$ is an arbitrary function of the input parameter and ψ_n (an estimate of ψ_0) satsifies:

$$r_n(\psi_n - \psi_0) \Rightarrow Z$$

for some weak limit Z and sequence of reals $r_n \to \infty$.

Suppose we are interested in estimating $f(\psi_0)$ where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at ψ_θ. The following theorem allows us to calculate the limiting distribution of $f(\psi_n)$:

Theorem 20 (MV Delta method $\mathbb{R}^d \to \mathbb{R}$)**.**
If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at ψ_0 and $r_n(\psi_n - \psi_0) \Rightarrow Z$ holds, then:

$$\begin{aligned}&f(\psi_n) - f(\psi_0) - \langle \phi_n - \phi_0, \nabla f(\psi_0)\rangle = o_P(r_n^{-1})\\ \iff &r_n(f(\psi_n) - f(\psi_0)) \Rightarrow \langle Z, \nabla f(\phi_0)\rangle\end{aligned}$$

Proof: f is differentiable at ϕ_0 iff f is uniformly converging to 0, i.e., g is continuous at 0 for:

$$g : \epsilon \to \begin{cases} \sup\limits_{h\in\mathbb{R}^d:||h||=1} \frac{|f(\psi_0+\epsilon h) - f(\psi_0) - \epsilon\langle h, \nabla f(\psi_0)\rangle|}{\epsilon} & \text{if } \epsilon \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note: that we can replace the uniform convergence condition with the condition that f is partially differentiable in a neighborhood about ψ_0 AND the partial derivatives are continuous.

Let $\epsilon_n = ||\psi_n - \psi_0||$ and let:

$$h_n = \begin{cases} (\psi_n - \psi_0)/\epsilon_n & \text{if } \epsilon_n \neq 0 \\ 0 & \text{else} \end{cases}$$

Then:

$$\begin{aligned}|f(\psi_n) - f(\psi_0) - \langle \psi_n - \psi_0, \nabla f(\psi_0)\rangle| &= |f(\psi_0 + \epsilon_n h_n) - f(\psi_0) - \epsilon_n\langle h_n, \nabla f(\psi_0)\rangle| \\ &\leq \sup_{h:||h||=1} |f(\psi_0 + \epsilon_n h) - f(\psi_0) -_n, \nabla f(\psi_0)\rangle| \\ &= \epsilon_n g(_n)\end{aligned}$$

We know $r_n(\psi_n - \psi_0) \equiv r_n(\epsilon_n h_n) \Rightarrow Z$ and $h_n = o_p(1)$ therefore $\epsilon_n = O_p(r_n^{-1})$. As $n \to \infty$, $\epsilon_n = o_p(1)$. By CMT $g(_n) = o_p(1)$ so $\epsilon_n g(_n) = o_p(r_n^{-1})$.

Thus $|f(\psi_n) - f(\psi_0) - \langle \psi_n - \psi_0, \nabla f(\psi_0)\rangle| = o_p(r_n^{-1})$.

Since the first and third terms would cancel we obtain:

$$\begin{aligned} r_n(f(\psi_n) - f(\psi_0)) &= \langle r_n(\psi_n - \psi_0), \nabla f(\psi_0)\rangle + \underbrace{r_n(f(\psi_n) - f(\psi_0) - \langle \psi_n - \psi_0, \nabla f(\psi_0)\rangle)}_{o_p(1)} \\ &\Rightarrow \langle Z, \nabla f(\psi_0)\rangle \qquad \text{(by CMT and Slutsky)} \end{aligned}$$

Can we generalize to a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^p$? Yes we can!

Theorem 21 (MV Delta Method $\mathbb{R}^d \to \mathbb{R}^p$)**.**
Suppose $r_n(\psi_n - \psi_0) \Rightarrow Z$ for some weak limit Z and real numbers $r_n \to \infty$ and $f : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable at ψ_0 (meaning we replace the gradient with the Jacobian and dot product by matrix multiplication). It holds that:

$$f(\psi_n) - f(\psi_0) - J_f(\psi_n - \psi_0) = o_p(r_n^{-1})$$

Where $J_f = \begin{pmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_p \end{pmatrix}$. And:

$$r_n[f(\psi_n) - f(\psi_0)] \Rightarrow J_f Z$$

Example 10 (Estimating relative risk)**.**
Suppose we observe n iid copies of $X = (T, Y)$ where T and Y are binary and $P_{\theta_0}(T = 1) = 1/2$. Let:

$$\psi_0 = \begin{pmatrix} \mathbb{E}_{\theta_0}[YT] \\ \mathbb{E}_{\theta_0}[Y(1-T)] \end{pmatrix}$$

The objective is to estimate $f(\psi_0)$ where $f(z) = \frac{z_1}{z_2}$. Our estimator is as follows:

$$\psi_n = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} Y_i T_i \\ Y_i(1-T_i) \end{pmatrix}$$

By the CLT:

$$\sqrt{n}(\psi_n - \psi_0) \Rightarrow N\left(0, \Sigma = \begin{pmatrix} \psi_{0,1}(1-\psi_{0,1}) & -\psi_{0,1}\psi_{0,2} \\ -\psi_{0,1}\psi_{0,2} & \psi_{0,2}(1-\psi_{0,2}) \end{pmatrix}\right)$$

By noting that $\nabla f(\psi_0) = (1/\psi_{0,2}, -\psi_{0,1}/\psi_{0,2}^2) = (1/\psi_{0,2}, -f(\psi_0)/\psi_{0,2})$, gthe multivariate delta method yields that:

$$\begin{aligned} \sqrt{n}[f(\psi_n) - f(\psi_0)] &\Rightarrow \langle N(0,\Sigma), \nabla f(\psi_0)\rangle \\ &\equiv N(0, \nabla f(\psi_0)^T \Sigma \nabla f(\psi_0)) \\ &\equiv N\left(0, f(\psi_0)\frac{1-\psi_{0,1}}{\psi_{0,2}} + f(\psi_0)^2\left[2 + \frac{1-\psi_{0,2}}{\psi_{0,2}}\right]\right) \end{aligned}$$

4 M-estimation, Z-estimation, and Maximum likelihood estimation

M-estimation and Z-estimation are closely related estimation procedures for $\phi_0 = \Phi(\theta_0)$ (some functional of the true DGP P_{θ_0} where $\theta_0 \in \Theta$ is unrestricted) that involve maximizing and finding the root of an population-based estimating equation, and replacing the population quantity by its empirical estimator.

We introduce empirical process notation for expectations, which can be read as a probability measure applied to a function:

$$Pf \equiv \int f(x)dP(x)$$

We also introduce *empirical process at f* which is just the centered sums:

$$\mathbb{G}_n f = n^{-1/2}\sum_{i=1}^{n}(f(X_i) - Pf)$$

4.1 M-estimation

Definition 13 (M-estimation framework)**.**
For a collection of functions $\{m_\phi : \phi \in S\}$, m_ϕ identifies ϕ_0 if:

$$\phi_0 \in \operatorname*{argmax}_{\phi} \mathbb{E}_\phi[m_\phi(X)] \text{ is a singleton} \equiv \operatorname*{argmax}_{\phi} P_{\theta_0} m_\phi$$

In the m-estimation framework, we replace the expectation over P_{θ_0} (the true DGP) with the expectation over the empirical distribution P_n:

$$\phi_n \in \operatorname*{argmax}_{\phi} \frac{1}{n}\sum_{i=1}^{n} m_\phi(X_i) \equiv \operatorname*{argmax}_{\phi} P_n m_\phi$$

More generally, suppose $\{M_\theta : \theta \in \Theta\}$ are a collection of real-valued functions satisfying:

$$\phi_0 \in \operatorname*{argmax}_{\phi} M_\theta(\phi) \text{ of all } \theta \in \Theta$$

Then the M-estimator is given by

$$\phi_n \in \operatorname*{argmax}_{\phi} M_n(\phi)$$

Where M_n is an estimator of M_{θ_0}.

Example 11 (MLE as M-estimator)**.**
Suppose any two distribution functions P, Q are absolutely continuous with respect to the Lesbegue/counting measure – then the pdfs/pmfs can be defined via the Radon-Nikodyn derivative: $P := \frac{dP}{d\mu}$.

Define the **KL-divergence** as:

$$D_{KL}(P||Q) := -P\left[\log\left(\frac{q}{p}\right)\right]$$

The KL-divergence satisfies positivity and identification criteria:

1. $D_{KL}(P||Q) \geq 0$.
2. $D_{KL}(P||Q) = 0$ iff $P = Q$.

Let:

$$\begin{aligned}
\theta_0 \in \underset{\theta\in\Theta}{\operatorname{argmax}}\,[-D_{KL}(P_{\theta_0}||P_\theta)] \\
= \underset{\theta\in\Theta}{\operatorname{argmax}}\; P_{\theta_0}\left[\log\frac{p_\theta}{p_{\theta_0}}\right] \\
= \underset{\theta\in\Theta}{\operatorname{argmax}}\; P_{\theta_0}\underbrace{[\log p_\theta]}_{m_\phi} \\
= \underset{\theta\in\Theta}{\operatorname{argmax}} P_{\theta_0} m_\phi
\end{aligned}$$

Then:

$$\begin{aligned}
\theta_n \in \underset{\theta\in\Theta}{\operatorname{argmax}}\; P_n m_\theta \equiv \underset{\theta\in\Theta}{\operatorname{argmax}}\; P_n(\log p_\theta) \\
= \underset{\theta\in\Theta}{\operatorname{argmax}}\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \\
= \underset{\theta\in\Theta}{\operatorname{argmax}}\prod_{i=1}^n p_\theta(X_i) = \hat{\theta}_{MLE}
\end{aligned}$$

4.2 Z-estimation

Z-estimation is closely related to M-estimation, but instead of choosing a ϕ_n that maximizes an empirical expectation, we choose one that finds the root of an empirical estimation equation.

Definition 14 (Z-estimation framework)**.**
A second approach to estimating ϕ_0 is to set it as the root of the equation:

$$\mathbb{E}_\theta[z_\phi(X)] = (0, \ldots, 0)^T$$

Z-estimation estimates ϕ_0 with a solution ϕ_n to:

$$\frac{1}{n}\sum_{i=1}^n z_\phi(X_i) = (0, \ldots, 0)^T$$

In a more general setting, suppose $\{Z_\theta : \theta \in \Theta\}$ is a bunch of $\mathbb{R}^b$-valued functions satisfying for which ϕ_0 is a solution in ϕ to:

$$Z_\theta(\phi) = (0, \ldots, 0)^T$$

And letting Z_n denote an estimator of Z_{θ_0}, the Z-estimator ϕ_n is given as a solution in ϕ to:

$$Z_n(\phi) = (0, \ldots, 0)^T$$

Example 12 (Sample median as a Z-estimator)**.**
Let $\phi_0 := \text{median}(X)$. We know

$$\mathbb{E}_\theta[\text{sign}(X - \phi_0)] = 0$$

Implying that a the sample median ϕ_n will be a root of the following estimating equation with probability 1:

$$\frac{1}{n}\sum_{i=1}^{n} \text{sign}(X_i - \phi) = 0$$

Theorem 22 (Z-estimation as M-estimation and vice versa)**.**
Let

$$M_\theta : \phi \to -||Z_\theta(\phi)||$$

implying that we can cast any Z-estimator as an M-estimator that obtains its maximum iff $Z_\theta(\phi) = 0$. Many (but not all) M-estimators can be cast as Z-estimator. We require that derivatives exist:

$$Z_\theta : \phi \to \nabla M_\theta(\phi)$$

An example of an M-estimator that is NOT writable as a Z-estimator is Manski's estimator of binary choice:

$$\hat{\beta}_n := \underset{\beta}{\text{argmax}} \sum \left(Y_i \, \mathbb{I}(X_i^T \beta > 0)\right)$$

Which does not have a natural Z-estimator.

4.3 Consistency of M and Z-estimators

There are several approaches to establishing the consistency of M and Z estimators including the uniform consistency criterion on M_n and the weak law of large numbers.

Suppose an M-estimator $\hat{\theta}_n$ maximizes the random criterion function $M_n(\theta)$. Under suitable regularity conditions, there exists an asymptotic criterion function such that $M_n(\theta) \xrightarrow{p} M(\theta)\ \forall\theta$ but this pointwise convergence is too weak to ensure that $\hat{\theta}_n \xrightarrow{p} \theta_0$. We need a stronger version of functional convergence: one such version is uniform convergence. See chapter 5 in VdV for more details.

Theorem 23 (Consistency of an M-estimator (VdV 5.8))**.**
In order for an M-estimator to be consistent, $\phi_n \xrightarrow{p} \phi_0$, the following three criteria must hold:

(i) Near maximizer is available: $M_n(\phi_n) \geq \sup_\phi M_n(\phi) - o_P(1)$ (maximizes up to a small mistake)

(ii) Identification: $\forall\epsilon > 0$, $M_0(\phi_0) > \sup_{\phi:||\phi-\phi_0||>\epsilon} M_0(\phi)$, i.e., ϕ_0 is a well-separated maximum

(iii) **Uniform consistency**: $\sup_\phi |M_n(\phi) - M_0(\phi)| \xrightarrow{p} 0$

Proof:

$$\begin{aligned}
M_0(\phi_0) - M_0(\phi_n) &\geq 0 \text{ by (ii)} \\
M_0(\phi_0) - M_0(\phi_n) &\leq (M_0(\phi_0) - M_0(\phi_n)) - \underbrace{(M_n(\phi_0) - \sup \ M_n(\phi))}_{\leq 0} \\
&\leq (M_0(\phi_0) - M_0(\phi_n)) - (M_n(\phi_0) - M_n(\phi_n)) + o_P(1) \\
&= \underbrace{[M_0(\phi_0) - M_n(\phi_0)]}_{\leq \sup\limits_{\phi} |(M_n - M_0)(\phi)} + \underbrace{[M_n(\phi_n)) - M_0(\phi_n)]}_{\leq \sup\limits_{\phi} |(M_n - M_0)(\phi)} + o_P(1) \\
&\leq 2 \cdot \sup |(M_n - M_0)(\phi)| + o_P(1) \\
&= o_P(1) \text{ by part (iii), uniform consistency}
\end{aligned}$$

Thus, we've shown that $M_0(\phi_0) - M_0(\phi_n) = o_P(1)$. Remains to show consistency. For a fixed $\epsilon > 0$, let $\delta = M_0(\phi_0) - \sup\limits_{||\phi - \phi_0|| > \epsilon} M_0(\phi) > 0$ by (ii). Notice that $\{||\phi_n - \phi_0|| > \epsilon\} \subset \{M_0(\phi_0) - M_0(\phi_n) \geq \delta\}$ because the former event implies tha latter. Therefore:

$$\begin{aligned}
P_0(||\phi_n - \phi_0|| > \epsilon) &\leq P(M_0(\phi_0) - M_0(\phi_n) \geq \delta) \\
&\overset{n \to \infty}{\longrightarrow} 0 \text{ b/c we showed } M_0(\phi_0) \xrightarrow{p} M_0(\phi_n) \\
&\implies \phi_n \xrightarrow{p} \phi_0
\end{aligned}$$

Theorem 24 (Consistency of a Z-estimator (General, VdV 5.9))**.**
We obtain the consistency of general Z-estimators, largely for free, based on the previous proof. If we notice that a zero of $Z_n(\phi)$ *maximizes* the function $-||Z_n(\phi)||$.

Let $Z_n(\phi)$ be a random values estimating equation and $Z_0(\phi)$ be the population-based estimating equation such that $\forall \ \epsilon > 0$

$$\sup_{\phi} ||Z_n(\phi) - Z_0(\phi)|| \xrightarrow{p} 0$$

$$\inf_{\phi: ||\phi - \phi_0|| \geq 0} ||Z_0(\phi)|| > 0 = ||Z_0(\phi_0)||$$

Then any sequence of estimators $\hat{\phi}_n$ such that $Z_n(\hat{\phi}_n) = o_P(1)$ yields $\hat{\phi}_n \xrightarrow{p} \phi_0$.

Proof: follows from the preceding theorem on applying the function $M_n(\phi) = -||Z_n(\phi)||$ and $M_0(\phi) = -||Z_n(\phi)||$.

We can derive the consistency of a 1-dimensional Z-estimator under slightly weaker conditions.

Theorem 25 (Consistency of a Z-estimator (1-dim, VdV Lemma 5.10))**.**
To show that ϕ_n (the root of the empirical estimating equation $Z_n(\phi)$) converges in probability to ϕ_0 (the root of the population estimating equation $Z_0(\phi)$), we rely on the following conditions:

1. Pointwise consistency: $\forall \phi$, $Z_n(\phi) \xrightarrow{p} Z_0(\phi)$ by WLLN (weaker than uniform consistency)

2. Either (a) each $\phi \to Z_n(\theta)$ is continuous and has exactly one root OR (b) $\phi \to Z_n(\theta)$ is non-decreasing.

3. $\forall \epsilon > 0$, $Z_0(\phi_0 - \epsilon) < 0 < Z_0(\phi_0 + \epsilon)$

Note this theorem only applies to the 1-dimensional case.

Proof:

$$\begin{aligned}&P[Z_n(\phi_0 - \epsilon) < 0, Z_n(\phi_0 + \epsilon) > 0]\\ &\leq P(\exists \text{ a root of } Z_n \text{ between } \phi_0 - \epsilon \text{ and } \phi_0 + \epsilon)\\ &= P(\phi_n \in (\phi_0 - \epsilon, \phi_0 + \epsilon))\end{aligned}$$

(1) and (3) implies that:

$$\begin{aligned}&Z_n(\phi_0 - \epsilon) \xrightarrow{p} Z_0(\phi_0 - \epsilon) \quad \& \quad Z_n(\phi_0 + \epsilon) \xrightarrow{p} Z_0(\phi_0 + \epsilon)\\ &\implies P[Z_n(\phi_0 - \epsilon) < 0, Z_n(\phi_0 + \epsilon) > 0] \xrightarrow{p} P[Z_0(\phi_0 - \epsilon) < 0, Z_0(\phi_0 + \epsilon) > 0] \xrightarrow{p} 1 \quad \text{(part (3))}\end{aligned}$$

Implying that $P(\phi_n \in (\phi_0 - \epsilon, \phi_0 + \epsilon)) \xrightarrow{p} 1$ and showing that $\phi_n \xrightarrow{p} \phi_0$

Wald offers an alternative set of conditions that permit the consistency of M-estimators. It works best if the parameter set is compact, and if not, we must show that estimators are eventually in a compact set of lie in a suitable compactification. We also require that $m_\theta(x)$ is upper-semicontinuous for almost all x, i.e.,

$$\limsup_{\theta_n \to \theta} m_{\theta_n}(x) \leq m_\theta(x)$$

Let:

$$M_n(\theta) = P_n m_\theta \qquad M(\theta) = P m_\theta$$

Theorem 26 (Wald's consistency for M-estimators (VdV 5.14))**.**
Let $m_\theta(x)$ be upper-semicontinuous for almost all x and let the criterion over a locally maximum choice of θ have finite measure – i.e., for every small ball $U \subset \Theta$:

$$P \sup_{\theta \in U} m_\theta < \infty$$

Typically there exists a unique maximum, but we allow multiple maxima with θ_0 describing this set. Then for any estimators with $\hat{\theta}_n$ s.t. $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$, then for every $\epsilon > 0$ and every compact set $K \subset \Theta$, the joint probability:

$$P(||\hat{\theta}_n, \Theta_0|| \geq \epsilon \ \wedge \ \hat{\theta}_n \in K) \to 0$$

4.4 Proving uniform consistency

The real meat of proving consistency of M/Z-estimators is showing the uniform consistency condition holds, i.e., $\sup_{\phi \in S} |(P_n - P_0) m_\phi| = o_P(1)$. In other words, a set of functions $\{m_\phi : \phi \in S\}$ that satisfies uniform consistency is said to be P_0-Glivenko-Cantalli. We can demonstrate the function space is P_0-Glivenko-Cantalli a variety of ways, using symmetrization and VC-bounds, Martingale theory, or via bracketing entropy. We pursue the final strategy. See chapter 19 in VdV for more details.

Definition 15 (P_0-Glivenko-Cantalli and Bracketing Number)**.**
A class of functions $\{m_\phi : \phi \in S\}$ is said to be P_0**-Glivenko-Cantalli** if it satisfies $\sup_{\phi \in S} |(P_n - P_0)m_\phi| = o_P(1)$.
Given two functions ℓ, u in $L^1(P_0)$ where $L^1(P_0)$ is the space of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying:

$$||f||_{L^1(P_0)} \equiv \int |f(x)| dP_0(x) < \infty$$

A *bracket* $[\ell, u]$ contains the set of all functions f with $\ell \leq f \leq u$. An $L^1(P_0)$ ϵ-bracket is a bracket $[\ell, u]$ for which $||u - \ell||_{L^1(P_0)} \leq \epsilon$.

The **bracketing number** denoted $N_{[]}(\epsilon, \mathcal{F}, L^1(P_0))$ of $\mathcal{F}$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.

And turns out, finite bracketing numbers of sets of functions imply that the set is P_0-Glivenko-Cantelli!

Theorem 27 (P_0-G-C via bracketing)**.**
If $\mathcal{F}$ is a class of functions for which $N_{[]}(\epsilon, \mathcal{F}, L^1(P_0)) < \infty$ for every $\epsilon > 0$, $\mathcal{F}$ iS P_0-G-C, i.e,

$$||P_n - P_0||_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |(P_n - P_0)f| = o_P(1)$$

Proof: the proof relies on the union bound result. For $A_i \in \mathcal{A}$ for $i \in \{1, 2, \ldots\}$, then:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i)$$

The general proof idea involves (a) within each ϵ-bracket, recognizing the f's are nearly identical (b) there exist finitely many ϵ-brackets and (c) using the union bound result.

Fix $\epsilon > 0$. By condition $\exists$ $[\ell_j, u_j]$ for $j \in \{1, 2, \ldots, N\}$ where N is the bracketing number s.t. the collection of brackets covers $\mathcal{F}$. Define the following quantities:

$$A_{n,\epsilon} := \{\sup_{f \in \mathcal{F}} (P_n - P)f > 2\epsilon\}$$
$$B_{n,\epsilon} := \{\sup_{f \in \mathcal{F}} (P_n - P)f < -2\epsilon\}$$

By symmetry it suffices to study $A_{n,\epsilon}$ and show that it is vanishing, i.e., the LHS is converging in probability to 0. Fix $f \in \mathcal{F}$, and by condition there exists one j s.t., $\ell_j \leq f \leq u_j$ pointwise, implying:

$$\begin{aligned}
&P_n f \leq P_n u_j \qquad P_0 f \geq P_0 u_j - P_0(u_j - \ell_j) \\
&(P_n - P_0)f \leq P_n u_j - P_0 u_j + P_0(u_j - \ell_j) \\
&= (P_n - P_0)u_j + P_0(u_j - \ell_j) \\
&\leq (P_n - P_0)u_j + \underbrace{P_0|u_j - \ell_j|}_{=||u_j - \ell_j||_{L_1(P_0)} \leq \epsilon} \quad \text{(by abs value)} \\
&\leq (P_n - P_0)u_j + \epsilon
\end{aligned}$$

Putting all the pieces together:

$$\begin{aligned}
P_0(A_{n,\epsilon}) &= P_0(\sup_{f\in\mathcal{F}}(P_n - P)f > 2\epsilon) \\
&\leq P_0\left(\sup_{j\in\{1,\ldots,N\}}(P_n - P_0)u_j + \epsilon + 2\epsilon\right) \\
&\leq \sum_{i=1}^{N} P_0((P_n - P_0)u_j > \epsilon) \\
&= o(1) \equiv\xrightarrow{n\to\infty} 0
\end{aligned}$$

Thus, we've shown that the probability of $\sup_{f\in\mathcal{F}}(P_n - P)f > 2\epsilon$ is asymptotically shrinking to 0, meaning $\sup_{f\in\mathcal{F}}|(P_n - P)f| = o_P(1)$

This begs the question, how do we show that a given class of functions has a finite bracketing number. To do so, we rely on Example 19.8 from Van der Vaart.

Theorem 28 (Finite bracketing number)**.** Suppose

(i) $\mathcal{F} \equiv \{f_\phi : \phi \in K\}$ is a collection of functions where $K \subset \mathbb{R}^d$ is compact.

(ii) $\forall x, \phi \to f_\phi(x)$ is continuous.

(iii) There exists an envelope function F s.t. both of the following as satisfied:

  (a) $\sup_{\phi\in K}|f_\phi(x)| \leq F(x)$ for all x.

  (b) $P_0|F| = P_0F < \infty$

Then $\forall\epsilon > 0$, the bracketing number is finite: $N_{[]}(\epsilon, L^1(P_0), \mathcal{F}) < \infty$.

Proof: Let $[f_B, f^B]$ be the bracket formed by the infimum and supremum of f for an open ball B about θ.

Construct a sequence of balls centered at ϕ called B_m with radii decreasing to 0. This implies $f^{B_m} - f_{B_m} \xrightarrow{m\to\infty} 0$ by the continuity of f.

Thus for any given $\epsilon > 0$ and for all ϕ, we can find an open ball B about ϕ such that the bracket $[f_B, f^B]$ is at most size ϵ.

Sine K is compact, the open cover of collection of brackets (the union of the open balls) has a finite subcover. The brackets in the finite subcover cover $\mathcal{F}$, are finite in number, and have size at most ϵ, implying that the bracketing number for $\mathcal{F}$ is finite.

4.5 Asymptotic normality of M and Z-estimators

We like the consistency of M and Z-estimators, but the next question is how quickly the M and Z-estimators converge to their targets. Turns out that for estimators based on N replications of an experiment, that the order is often $n^{-1/2}$ and multiplication with the inverse rate creates a balance, allowing $\sqrt{n}(\hat{\theta}_n - \theta)$ to converge to an (often) normal distribution! This is a powerful tool for inference!

Example 13 (Heuristic for ASN of 1-d Z-estimator)**.**
Suppose ϕ_n is the root of the equation $Z_n(\phi) = P_n z_\phi = 0$, and ϕ_0 is the root of $Z_0(\phi) = Pz_{\phi_0} = 0$. Let

$\phi_n \xrightarrow{p} \phi_0$.

$$\begin{aligned}
0 &= Z_0(\phi_0) = -Z_0(\phi_0)\\
&= [Z_n(\phi_0) - Z_0(\phi_0)] - Z_n(\phi_0) \quad \text{(add subtract)}\\
&= [Z_n(\phi_0) - Z_0(\phi_0)] + Z_n(\phi_n) - Z_n(\phi_0) \quad \text{(bc } Z_n(\phi_n) = 0)\\
&= \underbrace{[Z_n(\phi_0) - Z_0(\phi_0)]}_{\text{Term 1}} + \underbrace{Z_0(\phi_n) - Z_0(\phi_0)}_{\text{Term 2}} + \underbrace{Z_n(\phi_n) - Z_n(\phi_0) - Z_0(\phi_n) + Z_0(\phi_0)}_{\text{Term 3}}
\end{aligned}$$

Let's analyze the three terms separately!

$$\begin{aligned}
\text{Term 1} &= (P_n - P_0)z_{\phi_0}\\
\text{Term 2} &= (\phi_n - \phi_0)\dot{Z}_0(\phi_0) + \underbrace{\frac{1}{2}(\phi_n - \phi_0)^2\ddot{Z}_0(\tilde{\phi}_n)}_{=(\phi_n-\phi_0)o_p(1)O_p(1)} \quad (\text{Since } \ddot{Z}_0(\tilde{\phi}_n) = O_p(1))\\
&= (\phi_n - \phi_0)\dot{Z}_0(\phi_0) + o_P(\phi_n - \phi_0)\\
\text{Term 3} &= (P_n - P_0)(z_{\phi_n} - z_{\phi_0})
\end{aligned}$$

Let's pretend that ϕ_n is deterministic. Now we invoke Chebychev's inequality:

$$\begin{aligned}
P_0\{|(P_n - P_0)(z_{\phi_n} - z_{\phi_0})| > t/\sqrt{n}\} &\leq \frac{n\text{Var}_0[(P_n - P_0)(z_{\phi_n} - z_{\phi_0})]}{t^2}\\
&= \frac{\text{Var}_0(z_{\phi_n} - z_{\phi_0})}{t^2}
\end{aligned}$$

Suppose $\phi_n \to \phi_0$, the variance of the RHS will typically go to 0. This holds for example when there exists a function G with $P_0G^2 < \infty$ such that for every ϕ in some neighborhood of ϕ_0:

$$|z_\phi(x) - z_{\phi_0}(x)| \leq ||\phi - \phi_0||G(x)$$

If $\phi_n \to \phi_0$, then Term 3 $= o_p(n^{-1/2})$. To show this for a random sequence ϕ_n is outside the scope of this course (will be covered in 582-583).

Plugging in Terms from above, we obtain:

$$\begin{aligned}
0 &= (P_n - P_0)z_{\phi_0} + (\phi_n - \phi_0)\Big(\dot{Z}_0(\phi_0) + o_P(1)\Big) + o_P(n^{-1/2})\\
\implies \phi_n - \phi_0 &= -\frac{(P_n - P_0)z_{\phi_0}}{\dot{Z}_0(\phi_0) + o_P(1)} + o_P(n^{-1/2})
\end{aligned}$$

And under a finite second moment $P_0 z_{\phi_0}^2 < \infty$, then $\sqrt{n}(P_n - P_0)z_{\phi_0}/\dot{Z}_0(\phi_0) \Rightarrow N(0, P_0z_{\phi_0}^2/\dot{Z}_0(\phi_0)^2)$, implying:

$$\sqrt{n}(\phi_n - \phi_0) \Rightarrow N\left(0, \frac{P_0 z_{\phi_0}^2}{\dot{Z}_0(\phi_0)^2}\right)$$

Note that the preceding derivation requires that the criterion function $z_\phi(x)$ possesses two continuous derivatives with respect to the parameter θ for all x. This fails when the criterion function is $z_\theta(x) = \text{sign}(x - \theta)$ for which the median is a root, yet the sample median is still ASN! This motivates the need for additional conditions to achieve ASN.

The following two theorems describe the asymptotic normality of M and Z estimators. We omit their proofs:

Theorem 29 (ASN of General Z-estimators (VdV 5.21))**.**
Let $\phi_0 = P_0 z_\phi$ and $\phi_n = P_n z_\phi$. Suppose the following conditions hold:

1. Interior: suppose ϕ is in a open subset of $\mathbb{R}^d$ and that z_ϕ (the EE) is a map from $\mathcal{X} \to \mathbb{R}^d$.
2. EE has finite second moment: $\mathbb{E}_0||z_{\phi_0}(X)||^2 < \infty$
3. Smoothness and strong convexity: $\phi \to P_0 z_\phi$ is differentiable at a zero ϕ_0 with nonsingular Jacobian (derivative matrix) V_{ϕ_0} (strongly convex).
4. Envelope function: there exists $G : \mathcal{X} \to \mathbb{R}$ s.t.
   (i) Finite second moment: $P_0 G^2 < \infty$
   (ii) Lipschitz condition: $\forall x \in \mathcal{X}$ and every $\phi, \tilde{\phi} \in U(\phi_0)$ (neighborhood of ϕ_0):
   $$||z_\phi(z) - z_{\tilde{\phi}}(x)|| \leq ||\phi - \tilde{\phi}||G(x)$$
5. Root-n consistency: $\{\phi_n\}$ is a sequence of estimators of ϕ_0 s.t. $P_n z_{\phi_n} = o_p(n^{-1/2})$ and $\phi_n \xrightarrow{p} \phi_0$.

Under these conditions, the Z-estimator ϕ_n is ASN:

$$\sqrt{n}(\phi_n - \phi_0) \Rightarrow N\left(0, V_{\phi_0}^{-1} P[z_{\phi_0} z_{\phi_0}^T](V_{\phi_0}^{-1})^T\right)$$

Note: when $z_\phi(x)$ is continuously differentiable, a natural candidate for G in the above equation is $\sup\limits_{\phi \in U_{\phi_0}} ||\dot{z}_\phi||$.

Then the main condition reduces to partial derivatives are locally dominated by a square integrable function, i.e., there should exists a square-integrable function G s.t. $||\dot{z}_\phi(x)|| \leq G(x)$ for all ϕ close to ϕ_0. If $\dot{z}_\phi$ is also continuous at ϕ_0, DCT allows us to move the derivative inside the expectation, yielding $V_{\phi_0} = P\dot{z}_{\phi_0}$

Proof: Let $\mathbb{G}_n f := \sqrt{n}(P_n - P)f$ denote the empirical process evaluated at f. Note that Jensen's inequality, the Lipschitz condition on z_ϕ, and and consistency of $\phi_n \xrightarrow{p} \phi_0$ implies:

$$\begin{aligned}
||\mathbb{G}_n z_{\phi_n} - \mathbb{G}_n z_{\phi_0}|| &\underbrace{\leq}_{\text{Jensen}} \mathbb{G}_n ||z_{\phi_n} - z_{\phi_0}|| \\
&\underbrace{\leq}_{\text{4(ii)}} \mathbb{G}_n G(x) ||\phi - \tilde{\phi}|| \\
&= O_p(1) o_p(1) = o_p(1) \\
&\implies \mathbb{G}_n z_{\phi_n} - \mathbb{G}_n z_{\phi_0} \xrightarrow{p} 0
\end{aligned}$$

Note we can rewrite:

$$\begin{aligned}
\mathbb{G}_n z_{\phi_n} &= \sqrt{n}(P_n - P) z_{\phi_n} \\
&\equiv \sqrt{n}(\underbrace{P_n z_{\phi_n}}_{=0} - P z_{\phi_n}) \\
&\equiv \sqrt{n}\, \underbrace{P(z_{\phi_0}}_{=0} - z_{\phi_n}) + o_P(1)
\end{aligned}$$

Since Pz_ϕ is differentiable, we can apply the delta method to $\sqrt{n}P(z_{\phi_0} - z_{\phi_n}) + o_P(1)$:

$$\begin{aligned}
\mathbb{G}_n z_{\phi_n} &= \mathbb{G}_n z_{\phi_0} + o_P(1) \\
\implies \sqrt{n} P(z_{\phi_0} - z_{\phi_n}) &= \mathbb{G}_n z_{\phi_0} + o_P(1) \\
\implies \sqrt{n}\left(V_{\phi_0}(\phi_0 - \phi_n) + o_P(||\phi_n - \phi_0||)\right) &= \mathbb{G}_n z_{\phi_0} + o_P(1) \qquad \text{(Taylor expansion)}
\end{aligned}$$

Now we can write the following, knowing that nonsingularity of the variance matrix yields:

$$\begin{aligned}\sqrt{n}||\phi_n-\phi_0|| &\underbrace{\leq}_{\text{C-S}} \sqrt{n}||V_{\phi_0}^{-1}||\cdot||V_{\phi_0}(\phi_n-\phi_0)||\\ &= o_P(\sqrt{n}||\phi_n-\phi_0||)+O_p(1)\end{aligned}$$

This shows that $\phi_n \xrightarrow{p} \phi_0$ at rate at least $n^{-1/2}$. Thus, we have that $o_P(\sqrt{n}||\phi_n-\phi_0||)o_P(1)$ and then"

$$\begin{aligned}\sqrt{n}V_{\phi_0}(\phi_n-\phi_0) &= -\mathbb{G}_n z_{\phi_0}+o_P(1)\\ \implies \sqrt{n}(\phi_n-\phi_0) &= -[V_{\phi_0}]^{-1}\mathbb{G}_n z_{\phi_0}+o_P(1)\\ &\Rightarrow -[V_{\phi_0}]^{-1}N\left(0, Pz_{\phi_0}{}^{T}_{\phi_0}\right)\\ &\equiv N\left(0, [V_{\phi_0}]^{-1}]Pz_{\phi_0}{}^{T}_{\phi_0}[V_{\phi_0}]^{-1}]^T\right)\end{aligned}$$

Turns out, the Lipschitz condition is even stronger than is necessary and does not work for the sample median! We can still obtain convergence of the empirical processes under the weaker conditions where $z_\phi(x)$ are a *Donsker class* and is continuous in probability. For example, the $z_\phi(x) = \text{sign}(x-\phi)$ (which generates the median) do satisfy these criteria.

Theorem 30 (ASN of General M-estimators (VdV 5.23))**.**
Let $\phi_0 = \underset{\phi}{\text{argmax}} P_0 m_\phi$ and $\phi_n = \underset{\phi}{\text{argmax}} P_n m_\phi$. Suppose the following conditions hold:

1. m_ϕ differentiable: suppose ϕ is in a open subset of $\mathbb{R}^d$ and that $m_\phi(x)$ (the max criterion) is differentiable at ϕ_0 for P_0-almost everywhere x with derivative $\dot m_{\phi_0}$.
2. Envelope function: $\forall \phi, \tilde\phi \in U(\phi_0)$ (neighborhood of ϕ_0), assume there exists a function $G : \mathcal{X} \to \mathbb{R}$ satisfying:
   (i) Finite second moment: $P_0 G^2 < \infty$
   (ii) Lipschitz condition: $\forall x \in \mathcal{X}$ and every $\phi, \tilde\phi \in U(\phi_0)$ (neighborhood of ϕ_0):
   $$||m_\phi(x) - m_{\tilde\phi}(x)|| \leq ||\phi - \tilde\phi||G(x)$$
   Note: we can identify this G by defining $\dot m_\theta(x) = \nabla_\theta m_\theta(X)$ for a continuously differentiable h neighborhood U around θ, pick $G(x) = \theta \underset{\sup}{\in} U||\dot m_\theta(x)||_2$ because we've taken the largest magnitude derivative in the neighborhood. Second step is showing $P_\theta G(x)^2 < \infty$ (pg 53 VdV).
3. Uniform convergence: assume there exists a non-singular symmetric matrix V_{ϕ_0} s.t.
   $$\lim_{\epsilon\to 0}\sup_{||h||=1}\frac{\left|P_0 m_{\phi_0+\epsilon h} - P_0 m_{\phi_0} - \frac{1}{2}\epsilon^2 h^T V_{\phi_0} h\right|}{\epsilon^2} \xrightarrow{\epsilon\to 0} 0$$
   Note: we will verify this condition using QMD for MLE. **Note:** we can replace this condition by the supposition that $P_0 m_\phi$ is twice continuously differentiable at θ_0, affording a two-term Taylor expansion:
   $$P_0 m_\phi = P_0 m_{\phi_0} + \frac{1}{2}(\phi-\phi_0)^T V_{\phi_0}(\phi-\phi_0) + o(||\phi-\phi_0||^2)$$
4. Near maximizer and consistent: $P_n m_{\phi_n} \geq \underset{\phi}{\sup} P_n m_\phi - o_P(n^{-1})$ and $\phi_n \xrightarrow{p} \phi_0$

Under these conditions, the M-estimator is ASN:

$$\sqrt{n}(\phi_n-\phi_0) \Rightarrow N\left(0, V_{\phi_0}^{-1}P_0\dot m_{\phi_0}\dot m_{\phi_0}^T V_{\phi_0}^{-1}\right)$$

4.6 Maximum Likelihood Estimation

Maximum likelihood estimators can be viewed as maximizers of the log-likelihood criterion function. For mathematical convenience, we can subtract a constant p_{θ_0} too:

$$\begin{aligned}
&M_{\theta_0} : \theta \to \mathbb{E}_{\theta_0}\left[\log \frac{dP_\theta}{d\mu}(X)\right] = \mathbb{E}_{\theta_0}[\log(p_\theta(X))] \equiv \\
&\text{equivalent to maximizing: } \mathbb{E}_{\theta_0} \log \left[\frac{p_\theta}{p_{\theta_0}}\right] = P_0 \log \left[\frac{p_\theta}{p_{\theta_0}}\right] \\
&M_{\theta_n} : \theta \to \frac{1}{n}\sum \log p_\theta(X_i) \\
&\text{equivalent to maximizing: } P_n \log \left[\frac{p_\theta}{p_{\theta_0}}\right]
\end{aligned}$$

$-M_{\theta_0}$ is the Kullback-Leibler divergence of p_θ and p_{θ_0}. Thus, The MLE by definition minimizes the (empirically estimated) KL divergence and by consistency, converges in probability to the θ that minimizes the true KL divergence. This corresponds to the true value θ_0 when the model is *identifiable*, i.e.:

$$P_\theta \neq P_{\theta_0} \qquad \forall \theta \neq \theta_0$$

The MLE can also be viewed as a Z-estimator:

$$Z_{\theta_0} : \theta \to \nabla_\theta M_{\theta_0} \equiv \nabla_\theta \mathbb{E}_{\theta_0}[\log p_\theta(X)]$$

In STAT513, we assumed that we could exchange integration and differentiation, yielding that the score function has mean 0 and the asymptotic variance of the MLE. However, exchanging integration and differentiation is a strong condition, which we can replace by QMD in the next subsection.

Example 14 (Properties of MLE under strong condition)**.**
Suppose we can exchange integration and differentiation (a strong and non-necessary condition).

Claim 1: the score has function has mean 0, i.e., $Z_{\theta_0}(\theta_0) = 0$.

$$\begin{aligned}
Z_{\theta_0}(\theta_0) &= \mathbb{E}_0[\dot{\ell}_{\theta_0}] \\
&= \int \dot{\ell}_{\theta_0}(x) p_{\theta_0}(x) d\mu(x) \\
&= \int \frac{\dot{p}_{\theta_0}(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) d\mu(x) \\
&= \int \dot{p}_{\theta_0}(x) d\mu(x) \\
&= \int \nabla_\theta p_{\theta 0} d\mu(x) \\
&= \nabla_\theta \underbrace{\int p_{\theta 0} d\mu(x)}_{=1} \\
&= 0
\end{aligned}$$

Claim 2: the MLE is asymptotically normal. Suppose $\theta \in \mathbb{R}$. Recall we showed for Z-estimators that under some conditions:

$$\sqrt{n}(\theta_n - \theta_0) \Rightarrow N\left(0, \frac{P_0 z_{\theta_0}^2}{\dot{Z}_0(\theta_0)^2}\right)$$

We know $P_0 z_{\theta_0}^2 = P_0 \dot{\ell}_{\theta_0}^2$ where $\dot{\ell}_{\theta_0}$ is the score. We know the following:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\dot{\ell}_\theta(x) &= \frac{\ddot{p}_{\theta_0}(x)p_{\theta_0}(x) - \dot{p}_{\theta_0}(x)^2}{p_{\theta_0}(x)^2} \\
\dot{Z}_0(\theta_0) &= \mathbb{E}_0\left[\frac{\partial}{\partial\theta}\dot{\ell}_\theta(x)\right] \\
&= \int \frac{\ddot{p}_{\theta_0}(x)p_{\theta_0}(x) - \dot{p}_{\theta_0}(x)^2}{p_{\theta_0}(x)^2} dP_{\theta_0}(x) \\
&= \int \ddot{p}_{\theta_0}(x)d\mu(x) - \int \dot{\ell}_{\theta_0}(x)^2 dP_{\theta_0}(x) \\
&= \nabla_\theta^2 \underbrace{\int p_{\theta_0}(x)d\mu(x)}_{=1} - \int \dot{\ell}_{\theta_0}(x)^2 dP_{\theta_0}(x) \\
&= -\mathbb{E}_0[\dot{\ell}_{\theta_0}(x)^2] \equiv -P_0 z_{\theta_0}^2
\end{aligned}$$

Therefore the asymptotic variance of the MLE is $[P_0 z_{\theta_0}^2]^{-1} = [P_0 \dot{\ell}_\theta \dot{\ell}_\theta^T]^{-1} = I_\theta^{-1}$.

Claim 3: MLE is ASN multivariate. Let $\theta \in \mathbb{R}^d$. By similar arguments, we note that the FIM is defined as:

$$\begin{aligned}
I_\theta &= \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^T] \\
\sqrt{n}(\theta_n - \theta_0) &\Rightarrow N(0, I_{\theta_0}^{-1} P_0[\dot{\ell}_\theta \dot{\ell}_\theta^T] I_{\theta_0}^{-1}) \equiv N(0, I_{\theta_0}^{-1})
\end{aligned}$$

Note: these derivations implicitly require that the density p_θ has at least two derivatives with respect to the parameter. This is not the case with uniform distributions for example!

4.7 Quadratic Mean Differentiablility (QMD)

Note that we can show the score has mean 0 without requiring the exchange of integration and differentiation! Also note that the asymptotic variance of the MLE depends on the score but not its derivative, motivating the need for QMD.

QMD is also critical for allowing the asymptotic expansion of the local log-likelihood ratio, which allows us to conclude that likelihood ratio processes tend to a Gaussian process after reparametrization.

Definition 16 (QMD)**.**
The root density $\sqrt{p_\theta}$ is called QMD (or differentiable in quadratic mean) at θ if there exists a function $\dot{\ell}_\theta$ s.t.:

$$\sup_{h\in\mathbb{R}^d:||h||=1} \int \left[\frac{\sqrt{p_{\theta+\epsilon h}(x)} - \sqrt{p_\theta(x)}}{\epsilon} - \frac{1}{2}h^T\dot{\ell}_\theta(x)\sqrt{p_\theta(x)}\right]^2 d\mu(x) \xrightarrow{\epsilon\to 0} 0$$

Or equivalently, for any h (VdV pg 93)

$$\begin{aligned}
&\int \left[\sqrt{p_{\theta+\epsilon h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta}\right]^2 \mu(dx) = o(||h||^2) \text{ as } h \to 0 \\
&\equiv \frac{1}{||h||^2}\int \left[\sqrt{p_{\theta+\epsilon h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta}\right]^2 \mu(dx) \xrightarrow{p} 0
\end{aligned}$$

A model $\{P_\theta : \theta \in \Theta\}$ is called QMD at θ if the root density is QMD at θ. A model is called QMD if the root density is QMD at all $\theta \in \Theta$.

Here is a theorem that we can use to verify QMD-ness.

Theorem 31 (Thm 7.6 VdV)**.**
Suppose that $\sqrt{p_\theta(x)}$ is continuously differentiable for every x. If the elements of the matrix:

$$I_\theta = \int \frac{\dot{p}_\theta(x)\dot{p}_\theta(x)^T}{p_\theta(x)^2} dP_\theta(x)$$

are well-defined and continuous in θ, then $\sqrt{p_\theta}$ is QMD and $\dot{\ell}_\theta$ is given by $\frac{\dot{p}_\theta}{p_\theta}$.

See Theorem 15 for a result showing QMD-ness for exponential families (Vdv Ex. 7.7)!
One theorem will help us link the results of applying M-estimation ASN to known properties of the MLE

Theorem 32 (Thm 7.2 VdV)**.**
Suppose the following:

- Θ is an open subset of $\mathbb{R}^d$
- $\{P_\theta : \theta \in \Theta\}$ is QMD at θ.

Then $P_\theta \dot{\ell}_\theta = 0$ and the FIM $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ exists.

With QMD in hand, we can establish the asymptotic normality and expected properties of maximum likelihood estimators rigorously: without requiring exchange of integration and differentiation and without assuming that second derivatives exist.

Theorem 33 (Properties of MLE under QMD (VdV Theorem 5.39))**.** Suppose the model $\{P_\theta : \theta \in \Theta\}$ is QMD at an inner point $\theta_0 \in \Theta \subset \mathbb{R}^k$. Also suppose there exists a measurable function G (natural choice would be $\dot{\ell}$) with $P_0 G^2 < \infty$ s.t. for every θ_1, θ_2 in a neighborhood of θ_0:

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq G(x)||\theta_1 - \theta_2||$$

If the Fisher information matrix I_{θ_0} is nonsingular and $\hat{\theta}_n$ is consistent, then:

$$\begin{aligned}
\sqrt{n}[\hat{\theta}_n - \theta_0] &= I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_P(1) \\
&= I_{\theta_0}^{-1} \sqrt{n}(P_n - P_0)\dot{\ell}_{\theta_0} + o_P(1) \\
&\underset{\text{CLT}}{\Rightarrow} I_{\theta_0}^{-1} N(0, \mathbb{E}[(\dot{\ell}_{\theta_0} - \mathbb{E}[\dot{\ell}_{\theta_0}])^2]) \\
&\equiv I_{\theta_0}^{-1} N(0, \text{Var}(\dot{\ell}_{\theta_0})) \\
&\equiv N(0, I_{\theta_0}^{-1}) \quad \text{(b/c FIM is variance of score)}
\end{aligned}$$

Proof: This is a corollary of Theorem 5.23 (M-estimator ASN). We will show that condition (iii) (uniform convergence) holds, i.e., we WTS:

$$\sup_{||h||=1} \left| P_0 \ell_{\theta_0 + \epsilon h} - P_0 \ell_{\theta_0} + \frac{1}{2}\epsilon^2 h^T V_0 h \right| = o(\epsilon^2) \quad \text{ as } \epsilon \to 0$$

Let $p_\epsilon := p_{\theta_0 + \epsilon h}$ and let $p_0 := p_{\theta_0}$. Note that $P_0 \ell_{\theta_0 + \epsilon h} - P_0 \ell_{\theta_0}$ can be written as:

$$\begin{aligned}
P_0[\log p_\epsilon - \log p_0] &= P_0[2\log\sqrt{p_\epsilon} - 2\log\sqrt{p_0}] \\
&= 2P_0\left[\log\sqrt{\frac{p_\epsilon}{p_0}} - \underbrace{\log(1)}_{=0}\right] \\
&= 2\int f(w_\epsilon) - f(0) P_0(dx) \left(\text{Letting } w_\epsilon := \sqrt{\frac{p_\epsilon}{p_0}} - 1 \text{ and } f(x) = \log(1+w)\right)
\end{aligned}$$

We did the following bc f has a nice Taylor expansion:

$$\begin{aligned}
f(w) - f(0) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(w-a)^n \\
&= 0 + \frac{1}{1!}\frac{1}{1+w}\Big|_{w=0}(w-0)^1 - \frac{1}{2!}\frac{1}{(1+w)^2}\Big|_{w=0}(w-0)^2 + w^2 r(w) \\
&= w - \frac{w^2}{2} + w^2 r(w)
\end{aligned}$$

With $r(w) \to 0$ as $w \to 0$.

$$\begin{aligned}
P_0[\log p_\epsilon - \log p_0] &= 2\left(\int w_\epsilon P_0(dx)\right) - 2\left(\int \frac{w_\epsilon^2}{2} P_0(dx)\right) + 2\left(\int w_\epsilon^2 r(w_\epsilon) P_0(dx)\right) \\
\text{Term 1} &= 2\int \left(\sqrt{\frac{p_\epsilon}{p_0}} - 1\right) p_0 \ \mu(dx) \\
&= 2\int \left(\sqrt{p_\epsilon} - \sqrt{p_0}\right) \sqrt{p_0} \ \mu(dx) \\
&= -2\int \left(p_0 - \sqrt{p_\epsilon p_0}\right) \ \mu(dx) \\
&= -2\int \frac{p_\epsilon + p_0 - 2\sqrt{p_\epsilon p_0}}{2} \mu(dx) \qquad (\text{b/c} \int p_\epsilon = \int p_0 = 1) \\
&= -2\frac{1}{2}\int \left(\sqrt{p_\epsilon} - \sqrt{p_0}\right)^2 \mu(dx) \\
&= -H^2(P_\epsilon, P_0) \text{ the Hellinger Distance} \\
\text{Term 2} &= \int w_\epsilon^2 P_0(dx) \\
&= \int \left(\sqrt{\frac{p_\epsilon}{p_0}} - 1\right)^2 p_0 \ \mu(dx) \\
&= \int \left(\frac{p_\epsilon}{p_0} + 1 - 2\sqrt{\frac{p_\epsilon}{p_0}}\right) p_0 \mu(dx) \\
&= \int \left(p_\epsilon + p_0 - 2\sqrt{p_\epsilon p_0}\right) \mu(dx) \\
&= \int (\sqrt{p_\epsilon} - \sqrt{p_0})^2 \mu(dx) \\
&= H^2(P_\epsilon, P_0) \\
\text{Term 3} &= 2\int (\sqrt{p_\epsilon} - \sqrt{p_0})^2 r(W_\epsilon) \mu(dx) \\
&= o(\epsilon^2) \text{ tricky to show}
\end{aligned}$$

Thus,

$$P_0[\log p_\epsilon - \log p_0] = -2H^2(P_\epsilon, P_0)$$

Now it remains to show that $-2H^2(P_\epsilon, P_0)$ and $\frac{1}{2}\epsilon^2 h^T V_0 h$ and $o(\epsilon^2)$ close. To do this, we use the following:

- Reverse triangle inequality: $||| a|| - ||b||| \leq ||a - b||$.
- Let's introduce the $L^2(\mu)$ metric space which is a collection of functions f s.t. $\{f : ||f||_{L_2(\mu)} < \infty\}$ equipped with norm $||f||_{L_2(\mu)} = \left[\int f^2(x)\mu(dx)\right]^{1/2}$.

Thus we obtain:

$$
\begin{aligned}
\left| ||\sqrt{p_\epsilon} - \sqrt{p_0}||_{L_2(\mu)} - ||\frac{1}{2}\epsilon h^T \dot{\ell}_{\theta_0}\sqrt{p_0}||_{L_2(\mu)} \right| &\le \underbrace{||\sqrt{p_\epsilon} - \sqrt{p_0} - \frac{1}{2}\epsilon h^T \dot{\ell}_{\theta_0}\sqrt{p_0}||_{L_2(\mu)}}_{\text{QMD}} \text{ (by rev-tri inequal)} \\
&= o(\epsilon^2) \\
\implies H^2(P_\epsilon, P_0) = ||\sqrt{p_\epsilon} - \sqrt{p_0}||_{L_2(\mu)} &= \frac{1}{4}\epsilon^2 ||h^T \dot{\ell}_{\theta_0}\sqrt{p_0}||^2_{L_2(\mu)} + o(\epsilon^2) \\
\implies -2H^2(P_\epsilon, P_0) &= -\frac{1}{2}\epsilon^2 ||h^T \dot{\ell}_{\theta_0}\sqrt{p_0}||^2_{L_2(\mu)} + o(\epsilon^2) \\
&= -\frac{1}{2}\epsilon^2 h^T \left[\int \dot{\ell}_{\theta_0}\dot{\ell}^T_{\theta_0} p_0 \mu(dx)\right] h \\
&= \frac{1}{2}\epsilon^2 h^T P_0[\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}]h
\end{aligned}
$$

Implying

$$P_0[m_{\theta_0+\epsilon h} - m_{\theta_0}] = P_0[\log p_\epsilon - \log p_0] = \frac{1}{2}\epsilon^2 h^T V_0 h + o(\epsilon^2)$$

Where $V_0 = P_0[\dot{\ell}_{\theta_0}\dot{\ell}^T_{\theta_0}]$.

4.8 Local Asymptotic Normality (Ch 7 VdV)

A sequence of statistical models is LAN if their likelihood ratio processes are similar to those of a normal location parameter (asymptotically). This holds if the likelihood ratio processes admit a quadratic expansion. An important example involves sampling from a smooth parametric model. **The power of LAN** implies convergence of the models to a Normal model after rescaling the parameter; i.e., statistical experiments can be approximated by Gaussian experiments after suitable reparameterization.

First we introduce some background.

- Suppose we observe a sample $X_1, \ldots, X_n$ from a distribution P_θ on a measurable space indexed by $\theta \in \Theta \subset \mathbb{R}^k$ an open subset. The distribution of the sample is equivalent to sampling over $\{P^n_\theta := \prod_{i=1}^n P_\theta : \theta \in \Theta\}$
- **Statistical experiment**: procedure that can be infinitely repeated, has well-defined set of possible outcomes, produces only one outcome at conclusion of each trial.
- **Local parameter**: $h := \sqrt{n}(\theta - \theta_0)$ for a fixed, known parameter θ_0. We can rewrite $P^n_\theta = P^n_{\theta_0+h/\sqrt{n}}$, meaning that the experiment is with respect to unknown parameter h.

Remarkably, we can show that for large n, the following experiments are similar in statistical properties whenever the original experiments P_θ are smooth in the parameter:

$$(P^n_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^k) \equiv (N(h, I^{-1}_{\theta_0}) : h \in \mathbb{R}^k)$$

meaning a single observation from a normal distribution with mean h and covariance matrix equal to inverse FIM.

Theorem 34 (Expanding the likelihood ratio (VdV 7.2))**.**
To show that the likelihood ratio process approximates a Gaussian process (in a local neighborhood of θ),

we must first expand the likelihood. Suppose for simplicity that $\ell_\theta(x)$ is twice differentiable wrt θ for all x with derivatives $\dot\ell_\theta(x)$ and $\ddot\ell_\theta(x)$. A Taylor series expansion of $p_{\theta+h}$ at θ yields the log likelihood ratio is:

$$\begin{aligned}\log\frac{p_{\theta+h}}{p_\theta} &= \log p_{\theta+h} - \log p_\theta \\ &= \log p_\theta + h\dot\ell_\theta(x) + \frac{1}{2}h^2\ddot\ell_\theta(x) + o_x(h^2) - \log p_\theta \\ &= h\dot\ell_\theta(x) + \frac{1}{2}h^2\ddot\ell_\theta(x) + o_x(h^2)\end{aligned}$$

It follows that:

$$\begin{aligned}\log\prod_{i=1}^n \frac{p_{\theta_0}}{p_\theta}(X_i) &:= \log\prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) \\ &= \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot\ell_\theta(X_i) + \frac{1}{2}\frac{h^2}{n}\sum_{i=1}^n \ddot\ell_\theta(X_i) + \text{Remainder}_n \\ &= h\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \dot\ell_\theta(X_i) - 0\right) + \frac{1}{2}\frac{h^2}{n}\sum_{i=1}^n \ddot\ell_\theta(X_i) + \text{Remainder}_n \\ &= h\underbrace{\sqrt{n}(P_n\dot\ell_\theta - P_\theta\dot\ell_\theta)}_{\mathbb{G}_n\dot\ell_\theta} + \frac{h^2}{2}\underbrace{P_n\ddot\ell_\theta}_{\xrightarrow{p}-I_\theta} + \text{Remainder}_n \qquad \text{(Score has mean 0)} \\ \text{asymptotically } &\equiv hN(0,I_\theta) - \frac{h^2}{2}I_\theta + o_{P_\theta}(1)\end{aligned}$$

In the next step, we will see that this is similar to the likelihood ratio process for a normal experiment! Note that we refer to this as *"local"* ASN because the expansion was in the neighborhood of θ.

Note: the preceding derivation can be made rigorous under continuity conditions on the log likelihood OR under the weaker condition that the model is QMD.

Formally: Suppose $\Theta \subset \mathbb{R}^k$ is an open subset and the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ. As stated earlier, $P_\theta\dot\ell_\theta = 0$ and the FIM $I_\theta = P_\theta\dot\ell_\theta\dot\ell_\theta^T$ exists. Then for every converging sequence $h_n \to h$, as $n \to \infty$:

$$\log\prod_{i=1}^n \frac{p_{\theta+h_n/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\dot\ell_\theta(X_i) - \frac{1}{2}h^T I_\theta h + o_{P_\theta}(1)$$

The asymptotic expansion of the local log likelihood hinges on the model being QMD. We can establish QMD-ness via showing $p_\theta(x)$ is differentiable and is dominated by an integrable function. Alternatively we can use Lemma 7.6 VdV to establish QMD-ness. Or even better, we can establish QMD-ness for useful classes of models!

Example 15 (QMD-ness and Local Asymptotic Expansion of Exponential Families (VdV Ex. 7.7))**.** Suppose we have an exponential family of the form:

$$p_\theta(x) = d(\theta)h(x)\exp(Q(\theta)^T t(x))$$

If $Q(\theta)$ is continuously differentiable and map the parameter set Θ into the interior of the natural parameter space, then the three conditions of VdV Lemma 7.6 are satisfied, making the exponential family model QMD.

Additionally, the score function and information matrix are:

$$\dot{\ell}_\theta(x) = Q'_\theta(t(x) - \mathbb{E}_\theta(t(X))) \qquad I_\theta = Q'_\theta \text{cov}_\theta(t(X))(Q'_\theta)^T$$

Thus, the asymptotic expansion of the local log likelihood is valid for most exponential family members.

Example 16 (QMD-ness and local asymptotic expansion of location models (VdV Ex. 7.8))**.**
Consider all location models $\{f(x-\theta) : \theta \in \mathbb{R}\}$ for a positive, continuously differentiable density f with finite Fisher information:

$$I_f = \int \left(\frac{f'}{f}\right)(x) f(x) dx$$

The score function $\dot{\ell}_\theta(x)$ can be equal to $-\left(\frac{f'}{f}\right)(x-\theta)$, and the Fisher information is equal to I_f for all θ and hence is continuous in θ. Then the location family is QMD and the asymptotic expansion of the local

5 Hypothesis Testing

The intuition: suppose we observe some iid data from a distribution P_θ belonging to $\{P_{\theta'} : \theta' \in \Theta \subset \mathbb{R}^d\}$. The objective is to test:

$$H_0 : \theta \in \Theta_0 \text{ against } H_1 : \theta \notin \Theta_0$$

Let ϕ_n be a **test function**, which outputs values on $[0, 1]$. Deterministic tests take on values 0 or 1, while randomized functions can output a probability that we reject H_0.

The **power function** denotes the probability of rejecting H_0 based on the test ϕ_n:

$$\pi_n(\theta) \equiv \mathbb{E}_\theta[\phi_n(X_1, \ldots, X_n)]$$

The **Neyman-Pearson** testing paradigm encourages us to choose a ϕ_n such that

- Type 1 error rate control: $\sup\limits_{\theta_0 \in \Theta_0} \pi_n(\theta_0) \leq \alpha$
- Achieve high power at alternatives: make $\pi_n(\theta)$ large for $\theta \notin \Theta_0$

How do we motivate these tests asymptotically? By defining an asymptotically α-level test.

Definition 17 (Asymptotically α-level tests)**.**
A sequence of tests $\{\phi_n\}_{n=1}^\infty$ is a *asymptotically level-α* tests if:

$$\limsup_n \pi_n(\theta_0) \leq \alpha \text{ for all } \theta \in \Theta_0$$

5.1 Testing framework, Wald, Likelihood Ratio, and Score tests

Definition 18 (Testing framework and three famous parametric tests)**.**
In our testing paradigm, we can split the data generating parameter $\theta := (\psi, \eta)$ where ψ is the POI and η is a nuisance. So $\theta \in \mathbb{R}^d, \psi \in \mathbb{R}^m$ and $\eta \in R^{d-m}$.

Then $\Theta := T \times N$ where $T \subset \mathbb{R}^m$ and $N \subset \mathbb{R}^{d-m}$ where $\times$ is the tensor product. We consider θ on the interior of Θ.

WLOG we can define the null parameter set $\Theta_0 := \{\theta = (\psi, \eta) : \psi = 0\}$.

- If $m = d$, i.e., $\Theta_0 = \{\theta = \psi = 0\}$ which is denoted as a *simple null hypothesis.*
- If $m < d$, then the nuisance can take on values and Θ_0 may have multiple elements, creating a *composite null hypothesis.*

The following are three classic statistical tests:

1. **Wald Test**: motivated by the fact that we should reject H_0 when an estimate $\hat{\psi}$ is far from 0. We know that for $\hat{\theta} = (\hat{\psi}, \hat{\eta})$:

$$\begin{aligned} & n^{1/2}[\hat{\theta} - \theta] \Rightarrow N(0, I_\theta^{-1}) \\ \Longrightarrow\ & n^{1/2}[\hat{\psi} - \psi] \Rightarrow N(0, A_\theta^{-1}) \quad \text{by Woodbury } A_\theta = I_{\theta,11} - I_{\theta,12} I_{\theta,22}^{-1} I_{\theta,12}^T \\ \Longrightarrow\ & n^{1/2} A_\theta^{1/2}[\hat{\psi} - \psi] \Rightarrow N(0, Id_m) \\ \overset{\text{Slutsky}}{\Longrightarrow} \Longrightarrow\ & n^{1/2} A_{\hat{\theta}}^{1/2}[\hat{\psi} - \psi] \Rightarrow N(0, Id_m) \\ \Longrightarrow\ & n[\hat{\psi} - \psi]^T A_{\hat{\theta}}[\hat{\psi} - \psi]^T \Rightarrow \chi^2(m) \end{aligned}$$

Which when $\psi = 0$, suggests rejecting H_0 when $W_n = n[\hat{\psi}]^T A_{\hat{\theta}}[\hat{\psi}]^T$ is larger than the $(1-\alpha)$ quantile of $\chi^2(m)$. Hence, we expect $W_n \xrightarrow{p} \infty$ so the test will attain asymptotic power of 1.

2. **Likelihood ratio test**: heuristically, the LRT compares $D_{KL}(P_\theta, P_{\theta_0})$ and will reject if the following is too large:

$$\begin{aligned}\inf_{\theta_0 \in \Theta_0} D_{KL}(P_\theta, P_{\theta_0}) &= P_\theta[\ell_\theta - \ell_{\theta_0}] \\ &= \int \frac{\log p_\theta}{\log p_{\theta_0}} P_\theta(dx)\end{aligned}$$

We allow $\theta_0 \in \Theta_0$ nonempty with potentially multiple elements. In practice, we don't know θ so we us a consistent estimator, so we replace P_θ by its empirical plugin estimator P_n and replace θ by an *unrestricted MLE* $\hat{\theta}$. We give the likelihood ratio test statistic:

$$\begin{aligned}L_n &:= 2nP_n[\ell_{\hat{\theta}}] - \underbrace{\sup_{\theta_0 \in \Theta_0} P_n[\ell_{\theta_0}]}_{\text{Restricted MLE}} \\ &= 2nP_n[\ell_{\hat{\theta}} - \ell_{\hat{\theta}_0}] \\ &= -2\sum_{i=1}^n [\ell_{\hat{\theta}_0} - \ell_{\hat{\theta}}] \text{ (2nd order Taylor expansion next)} \\ &= -2\sum_{i=1}^n [\ell_{\hat{\theta}}(X_i) - \ell_{\hat{\theta}}(X_i) + (\hat{\theta}_0 - \hat{\theta})^T \dot{\ell}_{\hat{\theta}}(X_i) + \frac{1}{2}(\hat{\theta}_0 - \hat{\theta})^T \ddot{\ell}_{\tilde{\theta}}(X_i)(\hat{\theta}_0 - \hat{\theta})] \\ &= -2(\hat{\theta}_0 - \hat{\theta})^T \sum_{i=1}^n \dot{\ell}_{\hat{\theta}}(X_i) - (\hat{\theta}_0 - \hat{\theta})^T \sum_{i=1}^n \ddot{\ell}_{\tilde{\theta}}(X_i)(\hat{\theta}_0 - \hat{\theta}) \\ &= -(\hat{\theta}_0 - \hat{\theta})^T \sum_{i=1}^n \ddot{\ell}_{\tilde{\theta}}(X_i)(\hat{\theta}_0 - \hat{\theta}) \qquad \text{(b/c score has mean 0)} \\ &= -\sqrt{n}(\hat{\theta}_0 - \hat{\theta})^T [P_n \ddot{\ell}_{\tilde{\theta}}] \sqrt{n}(\hat{\theta}_0 - \hat{\theta})\end{aligned}$$

Under H_0, $\hat{\theta}$ AND $\hat{\theta}_0 \xrightarrow{p} \theta$ and $P_n \ddot{\ell}_{\tilde{\theta}} \xrightarrow{p} P_\theta \ddot{\ell}_\theta = -I_\theta$. This implies:

$$L_n \overset{H_0}{=} [\sqrt{n} I_\theta (\hat{\theta}_0 - \hat{\theta})^T] I_\theta^{-1} [\sqrt{n} I_\theta (\hat{\theta}_0 - \hat{\theta})^T] + o_P(1)$$

We also know that:

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta) &= I_\theta^{-1} \underbrace{\left(\sqrt{n}(P_n - P_\theta)\dot{\ell}_\theta\right)}_{=\mathbb{G}_n \dot{\ell}_\theta} + o_p(1) \\ \implies (\hat{\theta} - \theta) &= I_\theta^{-1}(P_n - P_\theta)\dot{\ell}_\theta + o_p(n^{-1/2}) \\ \text{And} \quad \hat{\theta}_0 - \theta &\overset{H_0}{=} \begin{pmatrix} 0 \\ \hat{\eta}_0 - \eta \end{pmatrix} = \begin{pmatrix} 0 \\ I_{\theta,22}^{-1}(P_n - P_\theta)\dot{\ell}_{\theta,2} + o_P(n^{-1/2}) \end{pmatrix}\end{aligned}$$

Where $\dot{\ell}_{\theta,2} := \nabla_\eta \log p_\theta$ and $I_{\theta,22} \equiv P_\theta \dot{\ell}_{\theta,2} \dot{\ell}_{\theta,2}^T$

Combining these results we obtain:

$$\begin{aligned}
\sqrt{n}I_\theta(\hat{\theta}_0-\hat{\theta}) &= \sqrt{n}I_\theta((\hat{\theta}_0-\theta)-(\hat{\theta}-\theta))+o_P(1)\\
&= \sqrt{n}I_\theta(P_n-P_\theta)\left(\begin{pmatrix}0\\ I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\end{pmatrix}-I_\theta\dot{\ell}_\theta\right)+o_P(1)\\
&= \sqrt{n}(P_n-P_\theta)\left(\begin{pmatrix}I_{\theta,11} & I_{\theta,12}\\ I_{\theta,21} & I_{\theta,22}\end{pmatrix}\begin{pmatrix}0\\ I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\end{pmatrix}-\begin{pmatrix}\dot{\ell}_{\theta,1}\\ \dot{\ell}_{\theta,2}\end{pmatrix}\right)+o_P(1)\\
&= \sqrt{n}(P_n-P_\theta)\left(\begin{pmatrix}-[\dot{\ell}_{\theta,1}-I_{\theta,12}I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}]\\ 0\end{pmatrix}\right)+o_P(1)\\
&\overset{\text{CMT and Slutsky}}{\Rightarrow}\begin{pmatrix}N(0,A_\theta)\\ 0\end{pmatrix}
\end{aligned}$$

Thus, under H_0

$$\begin{aligned}
L_n = [\sqrt{n}I_\theta(\hat{\theta}_0-\hat{\theta})^T]I_\theta^{-1}[\sqrt{n}I_\theta(\hat{\theta}_0-\hat{\theta})^T]+o_P(1) &= \begin{pmatrix}V^T & 0\end{pmatrix}\begin{pmatrix}(I_{\theta,11}-I_{\theta,12}I_{\theta,22}^{-1}I_{\theta,21})^{-1} & \dots\\ \dots & \dots\end{pmatrix}\begin{pmatrix}V^T\\ 0\end{pmatrix}\\
&\equiv V^TA_\theta^{-1}V\Rightarrow\chi^2(m)
\end{aligned}$$

Thus, we should reject H_0 when L_n is larger than the $1-\alpha$ quantile of the $\chi^2(m)$ dist.

3. **Score test**: Heuristically, scores have mean 0: $P_\theta\dot{\ell}_0=0$. Under when WLOG $\psi=0$, $H_0: P_\theta\dot{\ell}_{(0,\eta)}=0$. Thus, the score tests rejects when the estimate of the latter expectation is far from 0. Let $\hat{\theta}_0$ be the restricted MLE over Θ_0 and define Z_n is our estimate of the latter:

$$\begin{aligned}
Z_n(\theta_0) := \frac{1}{\sqrt{n}}\sum_{i=1}^n\dot{\ell}_{\theta_0}(X_i) &\overset{H_0}{=} 0+o_P(1)\\
Z_n(\hat{\theta}_0) &\overset{H_0}{=} 0+o_P(1)
\end{aligned}$$

So we showed our estimator should be unbiased and consistent for 0 under the null, but now we want it's limiting distribution under the null. Note we can break our estimator up into terms and show it has asymptotic distribution under very similar arguments as for LRT:

$$Z_n(\hat{\theta}_0) = \underbrace{n^{1/2}(P_n-P_0)\dot{\ell}_\theta}_{\text{CLT}}+\underbrace{n^{1/2}\left(P_\theta\dot{\ell}_{\hat{\theta}_0}-P_\theta\dot{\ell}_\theta\right)}_{\text{Delta method}}+\underbrace{n^{1/2}(P_n-P_\theta)(\dot{\ell}_{\hat{\theta}_0}-\dot{\ell})}_{\text{Donsker } o_P(1)}$$

$$\overset{H_0}{\Rightarrow}\begin{pmatrix}N(0,A_\theta)\\ 0\end{pmatrix}$$

Therefore, the estimator:

$$S_n := [Z_n(\hat{\theta}_0)]^T I_{\hat{\theta}_0 -1[Z_n(\hat{\theta}_0)]\Rightarrow\chi^2(m)}$$

In practice: the tests are all asymptotically equivalent under the null! The score test has the advantage of not requiring calculation of the unrestricted MLE. The Wald test has the advantage of not having to calculate the restricted MLE. The likelihood ratio test requires calculating both, but has nice theoretical guarantees in some cases (most powerful test under two point hypothesis). In small samples, the LRT and score test offer better T1 error rate control than the Wald test.

5.2 Contiguity

Contiguity generalizes the concept of absolute continuity to sequences of measures in asymptopia. In other words, contiguity is ”asymptotic absolute continuity”.

Note that we care about absolute continuity because it allows us to change the integration measure simply by reweighting the integration by the likelihood ratio! This allows us to take what we've learned about distributions of test statistics under the null and define their distributions under the alternative (a new measure)!

For instance:

$$\int f dQ \geq \int f \frac{dQ}{dP} dP$$

$$\int f dQ = \int f \frac{dQ}{dP} dP \iff Q << P$$

Definition 19 (Absolute continuity, orthogonal measures, Lebesgue decomposition)**.**
Let P and Q be measures on a measurable space $(\Omega, \mathcal{A})$, then Q is *absolutely continuous* wrt P if $P(A) = 0 \implies Q(A) = 0$.

Furthermore, P and Q are *orthogonal* if Ω can be partitioned as $\Omega = \Omega_P \cup \Omega_Q$ with $\Omega_P \cap \Omega_Q = \emptyset$ and $P(\Omega_Q) = Q(\Omega_P) = 0$. An example of orthogonal measures on $\mathbb{R}$ is the counting and Lebesgue measures.

Lebesgue decomposition: for any two probability distribution P and Q, there exist unique measures $Q^a(A) := Q(A \cap \Omega_P)$ (where $\Omega_P := \{p > 0\}$ i.e., where density of P has positive support) and $Q^{\perp} := Q(A \cap \Omega_P^c)$ (where $\Omega_P^c := \{p = 0\}$ i.e., where density of P has no support), such that $Q = Q^a + Q^{\perp}$ and $Q^a << P$ and $Q^{\perp} \perp P$. Essentially, any measure Q can be decomposed into absolutely continuous and orthogonal components wrt P.

Theorem 35 (Lemma 6.2 VdV)**.**
Let P and Q be probability measures with densities p and q with respect to μ. Let

$$Q^a(A) = Q(A \cap \{p > 0\}) \quad Q^{\perp}(A) = Q(A \cap \{p = 0\})$$

For these measures

(i) Lebesgue decomposition: $Q = Q^a + Q^{\perp}$, $Q^a << P$, $Q^{\perp} \perp P$

(ii) $Q^a(A) = \int_A \frac{q}{p} dP$ for every measurable A and likelihood ratio $\frac{q}{p}$.

(iii) $Q << P \iff \int \frac{q}{p} dP = 1 \iff Q = Q^a$ and $Q^{\perp} = 0$

Proof: in VdV page 86

Contiguity is simply absolute continuity for sequences of measures.

Definition 20 (Contiguity)**.**
A sequence $\{Q_n\}_{n=1}^{\infty}$ is contiguous with respect to $\{P_n\}_{n=1}^{\infty}$ if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for every sequence of measurable sets $\{A_n\}_{n=1}^{\infty}$. It is denoted by $Q_n \lhd P_n$. Mutual contiguity is denoted $Q_n \lhd \rhd P_n$

5.3 Le Cam's Lemmas

Le Cam's First Lemma generalizes part (iii) of Theorem 35 to asymptotic sequences of measures. I provide a heuristic understanding.

The likelihood ratios are nonnegative and satisfy:

$$\mathbb{E}_{P_n}\left[\frac{dQ_n}{dP_n}\right] \leq 1 \text{ and } \mathbb{E}_{Q_n}\left[\frac{dP_n}{dQ_n}\right] \leq 1$$

This means that the likelihood ratios $\frac{dQ_n}{dP_n}$ and $\frac{dP_n}{dQ_n}$ are uniformly tight under P_n and Q_n respectively, i.e., $\frac{dQ_n}{dP_n} = O_{P_n}(1)$ and $\frac{dP_n}{dQ_n} = O_{Q_n}(1)$. By Prokhorov's theorem, this implies that each sequence of measures has a weakly convergent subsequence. Le Cam's Lemma states that contiguity is determined by the limit points of this sequence. In other words, we recast part (iii) of Theorem 35:

$$Q << P \iff \int E_P \frac{dQ}{dP} = 1 \iff Q\left(\frac{dP}{dQ} = 0\right) = 0$$

By replacing the measures with measure sequences and their likelihood ratios with the limit points of their likelihood ratios. For $\frac{dP_n}{dQ_n} \overset{Q_n}{\Rightarrow} U$ for a subsequence of Q_n and $\frac{dQ_n}{dP_n} \overset{P_n}{\Rightarrow} V$ for a subseqeunce of P_n:

$$Q_n \lhd P_n \iff E_{P_n}[V] = 1 \iff Q_n\left(U = 0\right) = 0$$

Theorem 36 (Le Cam's First Lemma (VdV 6.4))**.**
Suppose P_n and Q_n are sequences of measures on a measurable space $(\Omega_n, \mathcal{A}_n)$. Then TFAE:

(i) Contiguity: $Q_n \lhd P_n$

(ii) Weak limit points of $\frac{dP_n}{dQ_n}$ give mass 0 to 0 under Q_n: if $\frac{dP_n}{dQ_n} \overset{Q_n}{\Rightarrow} U$ along a subsequence of Q_n, then $P(U > 0) = 1$.

(iii) Weak limit points of $\frac{dQ_n}{dP_n}$ have mean 1 under P_n: if $\frac{dQ_n}{dP_n} \overset{P_n}{\Rightarrow} V$ along a subsequence of Q_n, then $\mathbb{E}[V] = 1$.

(iv) For any statistics $T_n : \Omega_n \to \mathbb{R}^k$, If $T_n \overset{P_n}{\to} 0$ then $T_n \overset{Q_n}{\to} 0$

Proof: in VdV pg 88.

The following special case plays an important role in the asymptotic theory of smooth parametric models.

Theorem 37 (Le Cam's First Lemma – Smooth Parametric Models (VdV Ex. 6.5))**.**
Let P_n and Q_n be probability measures on arbitrary measure spaces such that the likelihood ratio converges weakly to a lognormal and the log likelihood ratio converges weakly to a normal:

$$L_n := \frac{dP_n}{dQ_n} \overset{Q_n}{\Rightarrow} \exp(N(\mu, \sigma^2))$$
$$\log L_n \overset{Q_n}{\Rightarrow} N(\mu, \sigma^2)$$

If this condition holds, $Q_n \lhd P_n$. And $Q_n \lhd \rhd P_n$ iff $\mu = -\frac{\sigma^2}{2}$.

Proof: To show $Q_n \lhd P_n$, let $U = \exp(N(\mu, \sigma^2))$ by definition as in Theorem 36 part (ii). Note that $U > 0$ because the exponential guarantees that U is always positive. Thus $Q_n(U > 0) = 1$ and by Le Cam's First Lemma, $Q_n \lhd P_n$.

To show $Q_n \lhd \rhd P_n$, we take the path of Theorem 36 part (iii) to show contiguity the other direction, switching the roles of Q_n and P_n. Let $V := \exp(N(\mu, \sigma^2))$ as defined in part (iii), which is true because $\frac{dP_n}{dQ_n} \overset{Q_n}{\Rightarrow} \exp(N(\mu, \sigma^2))$. Hence, V is obtained for any subsequence of Q_n. Then $\mathbb{E}[V] = \mathbb{E}(\exp(N(\mu, \sigma^2)))$ equals 1 iff $\mu = -\frac{\sigma^2}{2}$ (since mean of a lognormal RV is $\exp(\mu + \sigma^2/2)$.

Now we're ready for the big one: Le Cam's third Lemma, which allows us to obtain the limiting distribution of a sequence of random vectors under laws Q_n (an alternative distribution) based on laws P_n (a null distribution).

Theorem 38 (Le Cam's Third Lemma (VdV 6.6))**.**
Let P_n and Q_n be sequences of probability measures on $(\Omega_n, \mathcal{A}_n)$ (set and sigma alg., often borel alg.) and let $T_n : \Omega_n \to \mathbb{R}^k$ be a sequence of random vectors (your test statistics). Suppose that $Q_n \lhd P_n$ and

$$\left(T_n, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\Rightarrow} (X, V)$$

Define a new probability measure s.t. $\forall A \in \mathbb{R}^k$, $R(A) \equiv \mathbb{E}_{(T,V)}[\mathbb{I}(T \in A)V]$. Then:

$$T_n \overset{Q_n}{\Rightarrow} R$$

Proof: VdV pg. 90

Many people use the User-friendly version of Le Cam's third lemma!

Theorem 39 (User Friendly Le Cam's Third Lemma (Ex 6.7 VdV))**.**
If the following is true:

$$\left(T_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\Rightarrow} N_{k+1}\left(\begin{pmatrix} \mu \\ -\frac{\sigma^2}{2} \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$$

Then:

$$T_n \overset{Q_n}{\Rightarrow} N_k(\mu + \tau, \Sigma)$$

Proof: VdV pg. 90-91. Uses characteristics functions!

How do we find τ, the shift under the local alternative?
Note that local asymptotic normality gives us the fact that:

$$\log \frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta} \overset{P_\theta}{\Rightarrow} N\left(-\frac{1}{2} h^T I_\theta h, h^T I_\theta h\right)$$

If we can write the estimator as:

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(X_i) + o_{P_\theta}(1) \overset{P_n}{\Rightarrow} N(0, P_\theta\ \psi_\theta\ \psi_\theta^T)$$

If we know the asymptotic variance, we can work backwards to the influence function.

Once we have the influence function, we can work back to the τ via the following fact:

$$\left(\sqrt{n}(T_n-\theta), \log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\right) \overset{P_n}{\Rightarrow} N\left(\begin{pmatrix}0\\ -\frac{1}{2}h^T I_\theta h\end{pmatrix}, \begin{pmatrix}P_\theta\psi_\theta\psi_\theta^T & P_\theta\psi_\theta h^T\dot{\ell}_\theta\\ P_\theta\psi_\theta^T h^T\dot{\ell}_\theta & h^T I_\theta h\end{pmatrix}\right)$$

One fruitful example is the distribution of the MLE:

Example 17 (MLE under local alternative)**.**
Under a QMD model at inner point θ, with Lipschitz condition on the density, with nonsingular fisher infromation, we showed that in Theorem 33 the following is true for the MLE:

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underbrace{I_\theta^{-1}\dot{\ell}_\theta(X_i)}_{\text{Infl. func MLE}} + o_{P_n}(1)$$

If the model is QMD, we obtain a second order Taylor expansion of the log-likelihood enabling us to describe the distribution of the log likelihood via LAN:

$$\begin{aligned}\log L_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\dot{\ell}_\theta(X_i) - \frac{1}{2}h^T I_\theta h + o_{P_n}(1)\\ &\equiv N\left(-\frac{1}{2}h^T I_\theta h, h^T I_\theta h\right)\end{aligned}$$

And thus, the log likelihood ratio and MLE converge to a bivariate normal distribution:

$$\begin{aligned}\begin{pmatrix}\sqrt{n}(\hat{\theta}_n-\theta)\\ \log L_n\end{pmatrix} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\begin{pmatrix}I_\theta^{-1}\dot{\ell}_\theta(X_i)\\ h^T\dot{\ell}_\theta(X_i)\end{pmatrix} + \begin{pmatrix}0\\ -\frac{1}{2}h^T I_\theta h\end{pmatrix} + o_{P_n}(1)\\ &\Rightarrow N\left(\begin{pmatrix}0\\ -\frac{1}{2}h^T I_\theta h\end{pmatrix}, \begin{pmatrix}I_\theta^{-1} & h\\ h & h^T I_\theta h\end{pmatrix}\right)\end{aligned}$$

Implying that $\sqrt{n}(\hat{\theta}_n-\theta) \overset{Q_n}{\Rightarrow} N(h, I_\theta^{-1})$.

Thus, the asymptotic distribution of the MLE is invariant to local pertubations of θ!