Statistics

Definitions

- [Set-Up]
  - [Model] A <u>model</u> is a mapping from parameter to data distribution i.e. $\theta \longmapsto P_\theta$
    - Commonly just written as a set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$
  - [Parameter] θ
    - [Parameter Set] Θ
  - [Data] X
    - $X \sim P_\theta$
  - [Statistic] A function of data X
  - [Estimand] $g(\theta)$
  - [Estimator] $\delta(X)$
- [Loss Function] The <u>loss function</u> $L(\theta; \delta(X))$ is a function of X. Assuming θ is known, it is a measure of how close $\delta(X)$ and $g(\theta)$ are
  - $L(\theta, g(\theta)) = 0$
  - $L(\theta, d) \geq 0\ \forall \theta, d$
  - [Squared Error Loss] $L(\theta, d) = \|g(\theta) - d\|^2$
  - [Convex Loss] $L(\theta; d)$ is convex in d.
- [Risk Function] The <u>risk</u> of an estimator δ for a loss function $L(\theta, \delta(X))$ is the expected loss i.e. $R(\delta; \theta) = \mathbb{E}_\theta\left[L(\theta, \delta(X))\right]$
  - Judge how good an estimator δ is by its risk function
  - $\mathbb{E}_\theta$ is the expectation when $X \sim P_\theta$ i.e. θ is fixed
  - [Mean Squared Error] $\mathrm{MSE}(\theta, \delta) = \mathbb{E}_\theta\left[\left(g(\theta) - \delta(X)\right)^2\right]$; it is a risk function!
- [Inadmissible] An estimator δ is <u>inadmissible</u> if $\exists$ another estimator δ^* with a uniformly better risk function i.e.
  - $R(\theta, \delta^*) \leq R(\theta, \delta)\ \forall \theta \in \Theta$
  - $\exists \theta' \in \Theta$ s.t. $R(\theta', \delta^*) < R(\theta', \delta)$
  - i.e. the competing estimator δ^* is a strictly better estimator; else δ is <u>admissible</u>
- [Exponential Family] An <u>s-parameter exponential family</u> is a family of distributions $\mathcal{P} = \{P_\eta | \eta \in \Xi \subset \mathbb{R}^s\}$ with densities $P_\eta(x) = e^{\eta^T T(x) - A(\eta)} h(x)$
  - η: natural parameter
  - $s = \dim \eta$
  - $T(x)$: sufficient statistics
  - $h(x)$: carrier density / base density
  - $A(\eta)$: partition function
- [Natural Parameter Space] $\Xi_1 = \{\eta | A(\eta) < \infty\}$
  - Ξ_1 is convex
- [Full Rank] An exponential family with densities $p_\theta(x) = e^{\eta(\theta) \cdot T(x) - A(\theta)} h(x)$ is <u>full rank</u> if interior of $\eta(\Theta)$ is not empty and $\nexists v$ s.t. $v \cdot T = c$ a.e. μ
  - i.e. T does not satisfy a linear constraint
- [Sufficient] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ be a family of distributions. A statistic $T(X)$ is <u>sufficient</u> if $\forall \theta, \forall t$, $P_\theta(X | T = t)$ does not depend on θ. Define $Q_t(B) = \mathbb{P}[X \in B | T = t]$ which is independent of θ.
  - i.e. conditional distribution of X under P_θ given T does not depend on θ
  - i.e. $T(X)$ conveys all of information about θ from data X ($\therefore$ sufficient)
- [Sufficient] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ and $\tilde{\mathcal{P}} = \{\tilde{P}_\theta | \theta \in \Theta\}$ be models. $\tilde{\mathcal{P}}$ is <u>sufficient</u> for $\mathcal{P}$ if $\exists$ a stochastic transition kernel Q s.t. $P_\theta(B) = \int Q_t(B)\, \mathrm{d}\tilde{P}_\theta(t)\ \forall B$ Borel and $\theta \in \Theta$
  - If $\tilde{\mathcal{P}}$ is sufficient for $\mathcal{P}$, then data generation can be done via $T \sim \tilde{P}_\theta$, then $\tilde{X} \sim Q_t$
- [Likelihood] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, then the likelihood function is, given some data X, a function of θ:

- $L(\theta; X) = P_\theta(X)$
- $l(\theta; X) = \log L(\theta; X)$

- [Dominated] A family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated if $\exists$ measure μ s.t. $p_\theta \ll \mu \ \forall \theta \in \Theta$
- [Likelihood Function] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ be a family dominated by μ. Then $p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$. $p_\cdot : \Theta \to (X \to \mathbb{R})$ is the likelihood function
  - i.e. mapping of parameter θ to its density $p_\theta(X)$
- [Likelihood Shape] The likelihood shape is the family of curves spanning the parameter θ space: $S(X) = (0, \infty) \cdot L(\cdot; X) = \{cL(\cdot; X) \mid c \in (0, \infty)\}$.
- [Proportional / Same Shape] Two functions f, g have the same shape if $f \propto g$ i.e. $\exists c$ s.t. $cf(x) = g(x)$
- [a.e. $\mathcal{P}$] A proposition $Q(x)$ a.e. $\mathcal{P}$ means $\forall P \in \mathcal{P}$, $P(\{x \in X : \neg Q(x)\}) = 0$
  - The set on which the proposition fails, i.e. $\{x \in X : \neg Q(x)\}$, is a null set under all distributions
- [Minimal Sufficient] A statistic $T(X)$ is minimal sufficient if:
  - $T(X)$ is sufficient
  - For any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f a.e. $\mathcal{P}$
- [Complete] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ be a family of distributions. A statistic $T(X)$ is complete for $\mathcal{P}$ if $\mathbb{E}_\theta[f(T(X))] = 0 \ \forall \theta \in \Theta$ implies $f(T(X)) = 0$ a.e. $\mathcal{P}$
  - *Typically, prove by directly checking the condition via integration*
- [Completeness] A family of measures $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ on $\mathcal{X}$ is complete if $\int_{\mathcal{X}} f(x) \mathrm{d}P_\theta(x) = 0 \ \forall \theta \Rightarrow P_\theta(\{x : f(x) \neq 0\}) = 0 \ \forall \theta$ i.e. $f(x) = 0$ almost surely for all measure P_θ
  - A family $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ is not complete if there is some nonzero function f that is orthogonal to every P_θ
  - $\mathbb{E}_\theta[\delta_1(T)] = \mathbb{E}_\theta[\delta_2(T)] = g(\theta)$ then $\delta_1 = \delta_2$ a.s.
- [Ancillary] A statistic $V(X)$ is ancillary for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if its distribution is independent of θ
  - V by itself provides no information about θ

Properties

- [Sufficiency]
  - $P_\theta[X \in B] = \mathbb{E}_\theta[P_\theta[X \in B | T]] = \mathbb{E}_\theta[Q_T(B)]$
  - [Fake Data Construction] Given $T = t$, sample $\tilde{X} \sim Q_t$
  - [Factorisation Theorem 3.6] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of distributions dominated by μ. Then, a statistic $T(X)$ is sufficient if and only if $\exists g_\theta \geq 0$, $h \geq 0$ s.t. $p_\theta(X) = g_\theta(T(X)) h(X) \ \forall$ a.e. x under μ
    - i.e. $\mu(\{x : p_\theta(x) \neq g_\theta(T(x)) h(x)\}) = 0$
  - If $T(X)$ is sufficient, then $L(\theta; X) = g_\theta(T(X)) h(X)$
  - If $T(X)$ is sufficient and $T = f(\tilde{T})$, then $\tilde{T}$ is also sufficient
  - Let $T(X)$ be a sufficient statistic. Then $T(X)$ provides enough information to graph out the likelihood shape via $\frac{p_{\theta_1}(X)}{p_{\theta_2}(X)} = \frac{g_{\theta_1}(T)}{g_{\theta_2}(T)}$
- [Minimal Sufficiency]
  - Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family. Then, the shape of the likelihood is minimal sufficient.
    - $T(X)$ is minimally sufficient if it can be recovered from the likelihood shape
  - *Proof technique: Show that $p_\theta(x) \propto_\theta p_\theta(y) \Rightarrow T(x) = T(y)$, then T minimally sufficient*
- [Differential Identities]
  - $\nabla_\eta A(\eta) = \mathbb{E}_\eta[T(x)]$
  - $\nabla_\eta^2 A(\eta) = \mathrm{Var}_\eta[T(x)]$

- [Moment Generating Function] $M_\eta^{T(x)}(u) \coloneqq \mathbb{E}_\eta\left[e^{u^T T(x)}\right] = e^{A(\eta+u)-A(\eta)}$
- [Cumulant Generating Function] $K_\eta^{T(x)}(u) \coloneqq \log M_\eta^{T(x)}(u) = A(\eta+u) - A(\eta)$

- [Exponential Family Properties] $p_\theta(x) = e^{\eta(\theta)\cdot T(x) - A(\theta)}h(x)$
  - $T(X)$ is sufficient (*prove by factorisation theorem*)
  - If $T(x) - T(y) \perp \eta(\theta_0) - \eta(\theta_1)\ \forall \theta_0, \theta_1 \in \Omega$, then $T(X)$ is minimally sufficient
    - i.e. $T(x) - T(y) \in \left(\eta(\Theta) \ominus \eta(\Theta)\right)^\perp$
  - In an exponential family of full rank, T is minimally sufficient
  - [3.19] In an exponential family of full rank, T is complete
  - [12.19] Let $X \sim e^{\eta(\theta)\cdot T(x) - A(\theta)}h(x)$, then $T \sim e^{\eta(\theta)^T t - A(\theta)}$ w.r.t. some measure ν
- [Convex Loss Properties]
  - [3.24] Let f be convex on (a, b) and $t \in (a, b)$. Then $\exists c_t$ s.t. $f(t) + c_t(x - t) \leq f(x)$ $\forall x \in (a, b)$
    - If f strictly convex, this inequality can be upgraded to strict inequality for $x \neq t$
  - [Jensen] Let f be convex on (a, b) and $\mathbb{P}[X \in (a, b)] = 1$ and $\mathbb{E}[X] < \infty$. Then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Theorems

- [Sufficiency Principle] If $T(X)$ sufficient, then any statistical procedure should depend on X only through $T(X)$.
- [Basu] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a model. If $T(X)$ is complete sufficient and $V(X)$ is ancillary for $\mathcal{P}$, then $V \perp T$ under $P_\theta\ \forall \theta \in \Theta$
  - i.e. V and T are independent
  - $P_\theta[T \in B, V \in A] = P_\theta[T \in B]P_\theta[V \in A]$
- [3.3] Let $T = T(X)$ be a sufficient statistic for X with distribution from $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then $\forall \delta(X)$ of $g(\theta)$, $\exists$ randomised estimator with the same risk as $\delta(X)$
  - *Proof: Sample $\tilde{X} \sim Q_T$ and consider $\delta(\tilde{X})$*
- [$\propto_\theta$] $p_\theta(x) \propto_\theta p_\theta(y) \Rightarrow \exists c_{x,y}$ s.t. $p_\theta(x) = c_{x,y}p_\theta(y)\ \forall \theta \in \Theta$
- [3.11] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family and T be a sufficient statistic. If $p_\theta(x) = c(x, y)p_\theta(y)\ \forall \theta \in \Theta$ implies $T(x) = T(y)$, then T is minimal sufficient.
- [3.11] Let T be a sufficient statistic. If $L(\cdot; x) \propto L(\cdot; y) \Rightarrow T(x) = T(y)\ \forall \theta \in \Theta$, then T is minimal sufficient.
- [3.11] Let T be a sufficient statistic. If $l(\cdot; x) = l(\cdot; y) + c(x, y) \Rightarrow T(x) = T(y)\ \forall \theta \in \Theta$, then T is minimal sufficient.
- [3.17] If $T(X)$ is complete and sufficient, then $T(X)$ is minimal sufficient.
- [Rao-Blackwell] Let $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ and $L(\theta, \cdot)$ be a convex loss function, where $\theta \in \Theta$ and $R(\theta, \delta) < \infty$. Let T be a sufficient statistic for $\mathcal{P}$ and δ be an estimator of $g(\theta)$. Define $\tilde{\delta}(T) = \mathbb{E}[\delta(X)|T]$. Then $R(\theta, \tilde{\delta}) \leq R(\theta, \delta)$
  - If $L(\theta, \cdot)$ strictly convex, then inequality will be strict unless $\delta(X) = \tilde{\delta}(T)$ a.e. P_θ
  - For convex loss functions, can upgrade the estimator δ based on T to produce a non-randomised estimator $\tilde{\delta}$ with smaller risk
  - $\therefore$ if L is convex, the only estimators worth considering are functions of T where T is a sufficient statistic
  - $\therefore$ if L is convex, randomised estimators perform no better than non-randomised estimators
  - *Prove via Jensen and law of iterated expectations with the risk*

Exam

- Remember indicator functions in densities (they are functions of X)
- Statistics that might be sufficient: order statistics, max, min, median

Unbiased Estimation

Definitions

• [Unbiased] An estimator $\delta(X)$ is <u>unbiased</u> for $g(\theta)$ if $\mathbb{E}_\theta[\delta(X)] = g(\theta)\ \forall\theta \in \Theta$
  ○ [U-estimable] The function g is <u>U-estimable</u> if ∃ an unbiased estimator for $g(\theta)$
• [UMVU] An estimator $\delta(X)$ is <u>uniform minimum variance unbiased</u> if:
  ○ $\delta(X)$ is unbiased i.e. $\mathbb{E}_\theta[\delta(X)] = g(\theta)\ \forall\theta \in \Theta$
  ○ ∀ unbiased estimator $\tilde{\delta}(X)$, the variance of $\delta(X)$ is uniformly better i.e. $\mathrm{Var}_\theta[\delta(X)] \le \mathrm{Var}_\theta\left[\tilde{\delta}(X)\right]\ \forall\theta \in \Theta$
    ▪ i.e. δ is the best unbiased estimator under squared error loss

Properties

• [Squared Error Loss] Under $L(\theta;\delta) = \left(\delta(X) - g(\theta)\right)^2$:
  ○ Risk of an unbiased estimator δ is $R(\theta;\delta) = \mathrm{Var}_\theta[\delta(X)]$
  ○ Risk of any estimator δ is $R(\theta;\delta) = \mathrm{Var}_\theta[\delta(X)] + \mathrm{Bias}[\delta(X)]^2$
    ▪ $\mathrm{Bias}[\delta(X)] = \mathbb{E}_\theta[\delta(X) - g(\theta)] = \mathbb{E}_\theta[\delta(X)] - g(\theta)$
  ○ [1.10] Let $\delta(X)$ be a Bayes (or UMVU or minimax or admissible) estimator of $g(\theta)$ for squared error loss. Then $a\delta(X) + b$ is Bayes (or UMVU or minimax or admissible) estimator of $ag(\theta) + b$
• [Score]
  ○ Assuming regularity conditions, $\mathbb{E}_{\theta'}[\nabla_\theta l(\theta'; X)] = 0$
    ▪ Expected value of the score, at the true parameter θ', over the sample space $\mathcal{X}$ is 0
    ▪ If one were to resample from some distribution, the mean value of the scores tends to 0 asymptotically
  ○ First order stationary condition for MLE i.e. if $l(\theta; X)$ continuous in θ, then $\nabla_\theta l\left(\hat{\theta}_{\mathrm{MLE}}; X\right) = 0$
• [Exponential Family] $p_\eta(x) = e^{\eta^T T(x) - A(\eta)} h(x)$
  ○ [Score] $\mathcal{S}(\eta) = T(x) - \nabla_\eta A(\eta) = T(x) - \mathbb{E}_\eta[T(x)]$
  ○ [Fisher Information] $\mathcal{I}(\theta) = \nabla_\eta^2 A(\eta)$
  ○ [Cramér-Rao Lower Bound] Unbiased estimator for η has variance $\ge \frac{1}{\mathcal{I}(\theta)} = \left(\nabla_\eta^2 A(\eta)\right)^{-1}$

Theorems

• [Existence of UMVU 4.4] Suppose g is U-estimable and $T(X)$ is complete sufficient. Then ∃! estimator $\delta(T)$ based on T that is UMVU (this implies $\delta(T)$ is unbiased)
  ○ i.e. any other unbiased estimator $\tilde{\delta}(T)$, $\delta \ne \tilde{\delta}$ on a $\mathcal{P}$-null set i.e. $P_\theta\left[\left\{\delta(T) \ne \tilde{\delta}(T)\right\}\right] = 0\ \forall\theta \in \Theta$
  ○ The unbiased estimator $\delta(T)$ could be obtained by transforming any other unbiased estimator $\delta'(X)$ via Rao-Blackwell theorem i.e. $\delta(T) = \mathbb{E}[\delta'(X)\|T]$
  ○ *Proof: show $\mathbb{E}[\delta'(X)\|T]$ is unbiased via law of iterated expectation, then finish with completeness and Rao-Blackwell theorem*
• Let T be complete sufficient. Then if $\delta(T)$ is an unbiased estimator, then $\delta(T)$ is also UMVU.
• Under MSE, a biased estimator can have a better risk function than UMVU estimator if it has a smaller variance than the UMVU estimator as compared to increase in bias.

Exam

• Sometimes, just construct an unbiased estimator $\delta(X)$ via expectation formula
  ○ Taylor expansion and compare coefficients
  ○ May encounter differential equations

Definitions (Variance Bounds)
• [Log-Likelihood] $l(\theta; X) \coloneqq \log p_\theta(X)$

- [Score] The <u>score</u> is: $\mathcal{S}(\theta) \coloneqq \nabla_\theta l(\theta; X)$
  - Locally complete sufficient statistic
  - Given enough regularity, if $\delta(X)$ is unbiased for $g(\theta)$, then $g'(\theta) = \mathbb{E}_\theta[\delta\mathcal{S}]$ i.e. $\delta\mathcal{S}$ is unbiased for $g'(\theta)$
- [Fisher Information] The <u>Fisher information</u> is: $\mathcal{I}(\theta) \coloneqq \mathbb{E}_\theta[\mathcal{S}(\theta)\mathcal{S}(\theta)^T]$
  - Given enough regularity, $\mathcal{I}(\theta) = \mathrm{Var}_\theta[\mathcal{S}(\theta)] = \mathrm{Var}_\theta[\nabla_\theta l(\theta; x)] = -\mathbb{E}_\theta[\nabla_\theta^2 l(\theta; X)]$
  - $\mathcal{I}(\theta) \succcurlyeq 0$
  - $[d = 1]$ $\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\mathcal{S}(\theta)\right)^2\right] = \mathbb{E}_\theta\left[\left(\partial_\theta l(\theta; X)\right)^2\right] = \mathbb{E}_\theta[-\partial_\theta^2 l(\theta; X)]$
  - Intuitively, $\mathcal{I}(\theta)$ is the amount of information that X carries about the parameter θ
  - Expected value of the observed information $\nabla_\theta^2 l(\theta; X)$
  - Curvature of the support curve (the graph of log-likelihood)
  - High Fisher information indicates MLE is sharp
  - Low Fisher information indicates MLE is blunt
  - [Exponential Family] $\mathcal{I}_1(\eta) = \ddot{A}(\eta)$, $\mathcal{I}_n(\eta) = n\ddot{A}(\eta)$
  - $\mathcal{I}_n(\theta)$ is the Fisher information for n observations. Given i.i.d., $\mathcal{I}_n(\theta) = \mathrm{Var}_\theta[\nabla_\theta l_n(\theta; x)] = n\mathcal{I}_1(\theta)$
  - [Transformation] If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $\mathcal{Q} = \{Q_\xi : \xi \in \Xi\}$ are related by bijection $h: \Xi \to \Theta$, then $\mathcal{I}_\mathcal{Q}(\xi) = |h'(\xi)|^2 \mathcal{I}_\mathcal{P}(\theta)$ i.e. Fisher information is dependent on parametrisation
    - [Multivariate] $\mathcal{I}_\mathcal{Q}(\xi) = \left(Dh(\xi)\right)^T \mathcal{I}_\mathcal{P}(\theta)\left(Dh(\xi)\right)$
  - [Independence] If $X \perp Y$, then $\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$
    - If $X_1, \dots, X_n$ are i.i.d., then $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$
- [Efficiency] Let $\delta(X)$ be an unbiased estimator. Then the efficiency of δ is $\mathrm{eff}_\theta(\delta) = \frac{CRLB}{\mathrm{Var}_\theta(\delta)}$
  - $\mathrm{eff}_\theta(\delta) = \mathrm{Corr}_\theta[\delta(X), \nabla_\theta l(\theta; X)]^2$
    - Disguised as the correlation between the estimator and score function
  - "an estimator achieves the Cramér-Rao lower bound to the extent that it is correlated with the score function"
- [Location Family] Let X be an absolutely continuous random variable. The family of distributions $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ where P_θ is the distribution of $\theta + X$ is a location family.
  - i.e. the parameter θ specifies the mean
  - If X has density $f(x)$, then P_θ has density $p_\theta(x) = f(x - \theta)$
  - $\mathcal{I}(\theta) = \int \left(\frac{f'(x)}{f(x)}\right)^2 \mathrm{d}x$ is constant i.e. does not vary with θ

Theorems (Variance Bounds)

- [Hammersley-Chapman-Robbins] Let δ be an unbiased estimator. Then: $\mathrm{Var}_\theta[\delta] \geq \frac{\left(g(\theta+\Delta\theta) - g(\theta)\right)^2}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\Delta\theta}(X)}{p_\theta(X)} - 1\right)^2\right]} \approx \frac{\left(g'(\theta)\right)^2}{\mathbb{E}_\theta[(\partial_\theta \log p_\theta(X))^2]}$
  - *Prove by Cauchy-Schwarz and picking* $\psi = \frac{p_{\theta+\Delta\theta}(X)}{p_\theta(X)} - 1$
- [Cramér-Rao 4.9] Let $\theta \in \mathbb{R}$ and $\delta(X) \in \mathbb{R}$ be an unbiased estimator for $g(\theta) \in \mathbb{R}$. Then $\mathrm{Var}_\theta[\delta(X)] \geq \frac{\left(\nabla_\theta g(\theta)\right)^2}{\mathcal{I}(\theta)}$
  - $\nabla_\theta g(\theta) = \mathrm{Cov}_\theta[\delta(X), \nabla_\theta l(\theta; X)]$
  - Lower bound on the variance of an unbiased estimator $\delta(X)$
- [Cramér-Rao 4.9] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family with $\Theta \subset \mathbb{R}$ open and densities p_θ differentiable w.r.t θ. Provided $\mathbb{E}_\theta[\mathcal{S}] = 0$, $\mathbb{E}_\theta[\delta^2] < \infty$ and $g'(\theta) = \mathbb{E}_\theta[\delta\mathcal{S}]$ $\forall\theta \in \Theta$, then $\mathrm{Var}_\theta[\delta(X)] \geq \frac{\left(\nabla_\theta g(\theta)\right)^2}{\mathcal{I}(\theta)}$
- [Cramér-Rao Multivariate] Let $\theta \in \mathbb{R}^d$ and $\delta(X) \in \mathbb{R}$ be an unbiased estimator for $g(\theta) \in \mathbb{R}$

- $\mathbb{E}_\theta[\nabla_\theta \log p_\theta(X)] = 0$
- $\mathrm{Var}_\theta[\delta] \geq \left(\nabla g(\theta)\right)^T \left(\mathcal{I}(\theta)\right)^{-1} \nabla g(\theta)$

- [Exponential Family] $p_\eta(x) = e^{\eta^T T(x) - A(\eta)} h(x)$
  - $\mathcal{S}(\eta) = T(X) - \nabla_\eta A(\eta)$
  - $\mathcal{I}(\eta) = \mathrm{Var}_\eta[T(X)] = \nabla_\eta^2 A(\eta)$
  - [Cramér-Rao] Let $\mu = \nabla_\eta A(\eta)$, then $\mathrm{Var}_\mu[\delta] \geq \mathrm{Var}_\mu[T]$
    - *Prove by transformation of Fisher information*

Bayes Estimation

Definitions

- [Notation]
  - $\lambda(\theta)$: prior density
  - P_θ: conditional distribution of X given $\Theta = \theta$ i.e. $X|\Theta = \theta \sim P_\theta$
  - $R\big(\theta, \delta(X)\big) = \mathbb{E}\big[L\big(\theta; \delta(X)\big)|\Theta = \theta\big] = \int_{\mathcal{X}} L(\theta; \delta(x))\, \mathrm{d}P_\theta(x)$
  - $\lambda(\theta)p_\theta(x)$: joint density
  - $p_\theta(x) \approx \mathbb{P}[X = x|\Theta = \theta]$
  - $p(x)$: marginal density $\approx \mathbb{P}[X = x]$
    - $p(x) = \int_\Theta \lambda(\theta)p_\theta(x)\, \mathrm{d}\theta$
  - $\lambda(\theta|X)$: posterior density i.e. density of Θ given X
    - $\lambda(\theta|X = x) = \frac{\lambda(\theta)p_\theta(x)}{p(x)}$
  - Λ: prior distribution; probability measure on Θ i.e. $\Theta \sim \Lambda$
  - The expectation is taken over the posterior density $\Theta|X$
- [Bayes Risk] Let δ be an estimator and Λ be a probability distribution on Θ. Then, the <u>Bayes risk</u> is the expected risk over Θ: $r_\Lambda = \mathbb{E}\big[R\big(\Theta, \delta(X)\big)\big] = \int_\Theta R\big(\theta, \delta(X)\big)\, \mathrm{d}\Lambda(\theta) = \int_\Theta \mathbb{E}\big[L\big(\theta; \delta(X)\big)\big]\, \mathrm{d}\Lambda(\theta) = \int_\Theta \int_{\mathcal{X}} L\big(\theta; \delta(x)\big)\mathrm{d}P_\theta(x)\, \mathrm{d}\Lambda(\theta)$
- [Bayes Estimator] A <u>Bayes estimator</u> is an estimator that minimises Bayes risk: $\delta_\Lambda(X) = \arg\min_{\delta(X)} \int_\Theta R\big(\theta, \delta(X)\big)\, \mathrm{d}\Lambda(\theta) = \arg\min_{\delta} \mathbb{E}_{\theta\sim\Theta}\big[R\big(\Theta, \delta(X)\big)\big]$
  - $\delta_\Lambda(x) = \arg\min_{v} \mathbb{E}_{\theta\sim\Lambda(\Theta|X)}\, [L(\theta, v)] = \arg\min_{v} \int\, L(\theta, v)\, \lambda(\theta|x)\, \mathrm{d}\theta$
  - *Prove by Fubini's theorem*
- [Posterior Risk] The <u>posterior risk</u> is the conditional expected loss: $\mathbb{E}[L(\Theta; \delta)|X = x] = \int_\Theta L\big(\theta, \delta(x)\big)\, \lambda(\theta|x)\, \mathrm{d}\theta$
  - i.e. given data $X = x$, returns the expected loss over parameter space using the posterior distribution $\Lambda(\Theta|X)$
- [Conjugate Distribution] The prior distribution $\Lambda(\Theta)$ and posterior distribution $\Lambda(\Theta|X)$ are <u>conjugate distributions</u> if they are in the same probability distribution family.
  - [Conjugate Prior] If the prior distribution $\Lambda(\Theta)$ and posterior distribution $\Lambda(\Theta|X)$ are conjugate distributions, then $\Lambda(\Theta)$ is a <u>conjugate prior</u> for the likelihood function $P_\theta(X)$
- [Empirical Bayes] Data used to estimate parameters of the prior distribution.
  - i.e. as compared to standard Bayesian methods where prior distribution is fixed
- [James-Stein Estimator] Let $X \in \mathbb{R}^d$. Then, the <u>James-Stein estimator</u> is: $\delta_{JS}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right)X$

Theorems

- [7.1] Let $\Theta \sim \Lambda$, $X|\Theta = \theta \sim P_\theta$ and $L(\theta; \delta) \geq 0\ \forall\theta \in \Theta, \delta$. Then δ_Λ is a Bayes estimator if:
  - $\mathbb{E}[L(\Theta; \delta_0)] < \infty$ for some δ_0
  - For a.e. x, $\delta_\Lambda(x) = \arg\min_{d} \mathbb{E}[L(\Theta; d)|X = x]$
- For $L(\theta; \delta) = \big(g(\theta) - \delta(X)\big)^2$, $\delta_\Lambda(X) = \mathbb{E}[g(\Theta)|X]$
  - i.e. the Bayes estimator is just the posterior mean
  - $\delta_\Lambda(x) = \int_\Theta g(\theta)\, \lambda(\theta|x)\, \mathrm{d}\theta$
  - *Prove by dominated convergence theorem*
- [Stein]
  - Let $X \sim N(\mu, \sigma^2)$ and $h: \mathbb{R} \to \mathbb{R}$ differentiable and $\mathbb{E}[|h'(X)|] < \infty$, then $\mathbb{E}[(X - \mu)h(X)] = \sigma^2\mathbb{E}[h'(X)]$
    - *Prove by Fubini with $h(x) = \int_0^x h'(y)\, dy$*

  - Let $X \sim N_d(\mu, \sigma^2 \mathbb{I}_d)$ and $h: \mathbb{R}^d \to \mathbb{R}^d$ differentiable and $\mathbb{E}[\|\mathrm{D}h(X)\|_F] < \infty$, then $\mathbb{E}[(X-\mu)^T h(X)] = \sigma^2 \mathbb{E}\left[\mathrm{tr}\left(\mathrm{D}h(X)\right)\right] = \sigma^2 \sum_{i=1}^d \mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X)\right]$
    - *Prove by law of iterated expectation and 1D Stein*
- [11.3] Let $X_1, \dots, X_d$ independent with $X_i \sim N(\theta_i, d)$. Let $\delta(X)$ be an estimator for θ and $h(X) \coloneqq X - \delta(X)$. Assuming h is differentiable and $\mathbb{E}_\theta[\|Dh(X)\|_F] < \infty$, define $\hat{R} \coloneqq d + \|h(X)\|^2 - 2\mathrm{tr}\left(Dh(X)\right)$. Then $R(\theta, \delta) = \mathbb{E}_\theta[\|\delta(X) - \theta\|^2] = \mathbb{E}_\theta\left[\hat{R}\right]$.
- [Gaussian Sequence Model] $X \sim N_d(\theta, \mathbb{I}_d)$
- [Stein's Unbiased Risk Estimator] Let $\delta(X)$ be an estimator for the Gaussian sequence model. Let $h(X) = X - \delta(X)$. Assuming $\sigma^2 = 1$, $\hat{R}(X) = d + \|h(X)\|^2 - 2\mathrm{tr}\left(\mathrm{D}h(X)\right)$ is an unbiased estimator for the risk $R(\theta; \delta) = \mathbb{E}_\theta[\|\delta(X) - \theta\|^2]$.

Markov Chain Monte Carlo

- [Metropolis-Hasting Algorithm]
  - [Set Up] Goal: construct a Markov chain with stationary distribution π_i proportional to the posterior $\lambda(\theta|X)$
    - Allows sampling from the posterior π
  - θ; $\Theta^{(t)}$: parameters; parameter at time step t
  - $Q\left(\theta^{(j)}\middle|\theta^{(i)}\right)$: transition kernel / proposal distribution; probability of suggesting to go to $\theta^{(j)}$ from $\theta^{(i)}$
  - $a\left(\theta^{(j)}\middle|\theta^{(i)}\right)$: acceptance probability i.e. probability that you adopt the suggestion
    - $a\left(\theta^{(j)}|\theta^{(i)}\right) = \min\left\{1, \frac{\lambda(\theta^{(j)}|X)}{\lambda(\theta^{(i)}|X)} \frac{Q(\theta^{(i)}|\theta^{(j)})}{Q(\theta^{(j)}|\theta^{(i)})}\right\}$
  - [Algorithm]
    - Set $\theta^{(0)}$ to a feasible initial value
    - For t in $\{1,2,3,\dots\}$:
      - Sample $y \sim Q\left(y|\theta^{(t-1)}\right)$ (the proposed value for $\theta^{(t)}$)
      - Compute $A \leftarrow \min\left(1, \frac{\pi(y)Q(\theta^{(t-1)}|y)}{\pi(\theta^{(t-1)})Q(y|\theta^{(t-1)})}\right)$ (the acceptance probability)
      - Set $\theta^{(t)} \leftarrow \begin{cases} y & \text{w.p. } A \\ \theta^{(t-1)} & \text{w.p. } 1-A \end{cases}$
- [Gibbs Sampler] Let $\theta \in \mathbb{R}^d$
  - [Algorithm]
    - Initialise $\theta \leftarrow \theta^{(0)} \in \mathbb{R}^d$
    - For t in $\{1, \dots, T\}$:
      - For j in $\{1, \dots, d\}$:
        - Sample $\theta_j \sim \lambda\left(\theta_j|\theta_1, \dots, \hat{\theta}_j, \dots, \theta_d\right)$ # coordinate-wise update
      - Record $\theta^{(t)} \leftarrow \theta$

Miscellaneous

- [Beta Distribution] $\Theta \sim \mathrm{Beta}(\alpha, \beta)$
  - $\lambda(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$, $\theta \in (0,1)$
  - $\mathbb{E}[\Theta] = \frac{\alpha}{\alpha+\beta}$
  - $\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathrm{d}\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Minimax Estimation

Definitions

• [Minimax Risk] The <u>minimax risk</u> is $r^* = \inf_{\delta} \sup_{\theta} R(\theta; \delta)$
• [Minimax Estimator] δ^* is the minimax estimator if $\delta^* = \arg \inf_{\delta} \sup_{\theta \in \Theta} R(\theta; \delta)$
  ○ i.e. $\sup_{\theta \in \Theta} R(\theta; \delta^*) \leq \sup_{\theta \in \Theta} R(\theta; \delta)\ \forall \delta$
  ○ i.e. δ^* minimises the maximum risk
• [Least Favourable Prior] The <u>least favourable prior</u> is a prior distribution $\Lambda^* = \arg \max_{\Lambda} r_{\Lambda} = \arg \max_{\Lambda} \int_{\Theta} R(\theta; \delta_{\Lambda}) \mathrm{d}\Lambda(\theta)$
  ○ i.e. $r_{\Lambda^*} \geq r_{\Lambda}\ \forall \Lambda$ i.e. Λ^* has the highest Bayes risk out of all prior distributions
  ○ Risk of least favourable prior is the best lower bound for minimax risk
• [Least Favourable Prior Sequence] Let $\{\Lambda_n\}_n$ be a sequence of priors with minimal average risks $\{r_{\Lambda_n}\}_n$ where $r_{\Lambda_n} = \inf_{\delta} \int_{\Theta} R(\theta; \delta) \mathrm{d}\Lambda_n(\theta)$. $\{\Lambda_n\}_n$ is a <u>least favourable prior sequence</u> if $\lim_{n \to \infty} r_{\Lambda_n} = r < \infty$ with $r \geq r_{\Lambda'}$ for any other prior distribution Λ'.
  ○ i.e. the limit of the Bayes risk is highest among all Bayes risk
• [Residual Sum of Squares] $RSS(\hat{\mu}, Y)$
• [Expected Prediction Error] $EPE(\mu, \hat{\mu})$
• [Effective Degrees of Freedom] $DF(\mu, \hat{\mu}) = \frac{1}{2\sigma^2} \mathbb{E}[EPE - RSS]$

Theorems

• Let Λ be a proper prior and δ_{Λ} be the Bayes estimator. The Bayes risk r_{Λ} of any proper prior Λ is less than the minimax risk r^* i.e. $r_{\Lambda} = \int_{\Theta} R(\theta; \delta_{\Lambda}) \mathrm{d}\Lambda(\theta) = \inf_{\delta} \int_{\Theta} R(\theta; \delta) \mathrm{d}\Lambda(\theta) \leq \inf_{\delta} \int_{\Theta} \sup_{\theta} R(\theta; \delta)\, \mathrm{d}\Lambda(\theta) = \inf_{\delta} \sup_{\theta} R(\theta; \delta) = r^*$
  ○ "A minimax estimator is a Bayes estimator for the worst possible prior"
• [1.4] Let Λ be a prior distribution on Θ. If $r_{\Lambda} = \sup_{\theta \in \Theta} R(\theta; \delta_{\Lambda})$, then:
  ○ δ_{Λ} is minimax
  ○ Λ is least favourable
  ○ If δ_{Λ} is the unique Bayes estimator for Λ a.s., then it is also the unique minimax estimator
• [1.5] Let δ_{Λ} be a Bayes estimator. If δ_{Λ} has constant risk, then it is also the minimax estimator, and Λ is the least favourable prior.
• [1.6] Let $\Theta_{\Lambda} = \left\{\theta : R(\theta, \delta_{\Lambda}) = \sup_{\theta' \in \Theta} R(\theta'; \delta_{\Lambda})\right\}$. Then δ_{Λ} is minimax if and only if $\Lambda(\Theta_{\Lambda}) = 1$
  ○ Θ_{Λ} is the set of parameters for which δ_{Λ} attains maximum
• [1.12] Suppose $\{\Lambda_n\}_n$ is a sequence of priors and δ is an estimator that achieves $\sup_{\theta \in \Theta} R(\theta, \delta) = \lim_{n \to \infty} r_{\Lambda_n}$. Then:
  ○ δ is minimax
  ○ $\{\Lambda_n\}_n$ is least favourable

Exam
• Typically, just express Bayes risk in integral form and bound the integrand

Hypothesis Testing

Definitions

- [Set-up]
  - [Model] $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$
  - [Null] $H_0 : \theta \in \Theta_0$
    - Generally represents the status quo
  - [Alternate] $H_1 : \theta \in \Theta_1$
- [Critical Function] Describes behaviour of test on sample X: $\phi(X) = \begin{cases} 0 \\ \text{Bernoulli}(\gamma) \\ 1 \end{cases}$
- [Rejection Region] $\mathcal{R}(\phi) = \{x \in \mathcal{X} : \phi(x) = 1\}$
  - a.k.a critical region
  - $x \in \mathcal{R} \Rightarrow$ "accept" H_1
- [Acceptance Region] $\mathcal{A}(\phi) = \{x \in \mathcal{X} : \phi(x) < 1\}$
  - $x \in \mathcal{A} \Rightarrow$ "accept" H_0
- [Power Function] The power function of a test ϕ is a function $\beta_\phi : \Theta \to [0,1]$ with $\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)] = P_\theta[\phi(X) = 1]$
  - i.e. probability of rejecting H_0 given θ
  - The power function is a measure of performance of test ϕ
- [Level-α Test] A test $\phi(X)$ is a <u>level-α test</u> if $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \le \alpha$ i.e. maximum probability of rejecting null hypothesis, given that null hypothesis is correct
  - α is the significance level a.k.a. worst probability of wrongfully rejecting H_0
  - Ubiquitous choice is $\alpha = 0.05$
  - $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$ also known as Type I error rate
- [Simple] A hypothesis is <u>simple</u> if it is a sub-model that contains a single distribution e.g. $\Theta_0 = \{\theta_0\}$
  - i.e. it completely specifies the distribution of the data
- [Composite] A composite hypothesis is one that is not simple
- [Identifiable] Model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is <u>identifiable</u> if $\theta_1 \ne \theta_2 \Rightarrow P_{\theta_1} \ne P_{\theta_2}$
- [Monotone Likelihood Ratio] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ be an identifiable model with densities p_θ. Let $T(X) \in \mathbb{R}$ be a statistic. Then $\mathcal{P}$ has <u>monotone likelihood ratios</u> in $T(X)$ if $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$ is a non-decreasing function of $T(x)$ $\forall \theta_1 < \theta_2$
  - $T(x_1) \le T(x_2) \Rightarrow \frac{p_{\theta_2}(x_1)}{p_{\theta_1}(x_1)} \le \frac{p_{\theta_2}(x_2)}{p_{\theta_1}(x_2)}$
- [Stochastically Increasing] A real-valued statistic $T(X)$ is <u>stochastically increasing</u> in θ if $\mathbb{P}_\theta[T(X) \le t]$ is non-decreasing in θ $\forall t$
  - $\theta_1 \le \theta_2 \Rightarrow \mathbb{P}_{\theta_1}[T(X) \le t] \le \mathbb{P}_{\theta_2}[T(X) \le t]$
- [Uniformly Most Powerful] A test ϕ^* is <u>uniformly most powerful</u> if $\phi^*(X)$ has level α and any other level α test $\phi(X)$, we have $\mathbb{E}_\theta[\phi^*(X)] \ge \mathbb{E}_\theta[\phi(X)]$ $\forall \theta \in \Theta_1$
  - i.e. ϕ^* has the most power on rejection region across level α tests.
- [Unbiased] A test $\phi(X)$ is <u>unbiased</u> if $\beta_\phi(\theta) \ge \alpha$ $\forall \theta \in \Theta_1$ and $\beta_\phi(\theta) \le \alpha$ $\forall \theta \in \Theta_0$
  - i.e. want $\phi(X)$ to have at least power α in the rejection region and at most power α in acceptance region
  - For $\theta \in \partial\Theta_1 \cup \partial\Theta_2$, typically $\beta_\phi(\theta) = \alpha$
- [Uniformly Most Powerful Unbiased] UMPU
- [Inadmissible] A test $\hat{\phi}$ is <u>inadmissible</u> if $\exists$ a competing test ϕ with better power function i.e. $\beta_{\hat{\phi}}(\theta) \ge \beta_\phi(\theta)$ $\forall \theta \in \Theta_0$ and $\beta_{\hat{\phi}}(\theta) \le \beta_\phi(\theta)$ $\forall \theta \in \Theta_0$ with strict inequality for at least one $\theta \in \Theta_0 \cup \Theta_1$
- [p-value] The <u>p-value</u> of a test ϕ is the α-level at which ϕ barely rejects

- $p(x) = \inf_{\alpha}\{\alpha: \phi_\alpha(x) = 1\}$
- $p(x) \le \alpha \Leftrightarrow \phi_{\alpha'}(x) = 1\ \forall \alpha' > \alpha$
- Note $\left(\phi_\alpha(X)\right)_\alpha$ are tests s.t.:
  - $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\phi_\alpha(X)] = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\phi_\alpha(X) = 1] \le \alpha$ i.e. ϕ_α is test of significance α
  - Increasing w.r.t. α i.e. $\alpha_1 \le \alpha_2 \Rightarrow \phi_{\alpha_1}(x) \le \phi_{\alpha_2}(x)$
    - $R_{\alpha_1} \subset R_{\alpha_2}$

- [Confidence Set] Let $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$. Then $C(X)$ is a <u>$1-\alpha$ confidence set</u> for $g(\theta)$ if $P_\theta[C(X) \ni g(\theta)] \ge 1-\alpha\ \forall \theta \in \Theta$.
  - i.e. no matter which $\theta \in \Theta$, probability that $C(X)$ covers $g(\theta)$ is at least $1-\alpha$
  - Note: $g(\theta)$ is fixed under P_θ; $C(X)$ is a random set
  - $P_\theta[C(X) \ni g(\theta)|X] \in \{0,1\}$ since there are no more randomness
- [Duality] Fix $\alpha \in (0,1)$ i.e. $C(X)$ and $(\phi_a)_a$ are sets and tests created w.r.t. α
  - Let $(\phi_a)_a$ be a family of non-randomised level-α tests indexed by $a \in g(\Theta)$, where $\phi_a(X)$ tests for $H_0: g(\theta) = a$ and $H_1: g(\theta) \ne a$. This gives rise to the confidence set $C(X) = \{a: X \in \mathcal{A}(\phi_a)\} = \{a: \phi_a(X) = 0\}$
  - Let $C(X)$ be a $1-\alpha$ confidence set for $g(\theta)$. Then, construct the family of tests $(\phi_a)_a$ where $\phi_a(X) = \mathbb{1}\{a \notin C(X)\}$ is a level-α test for $H_0: g(\theta) = a$ and $H_1: g(\theta) \ne a$
  - $g(\theta) \in C(X) \Leftrightarrow P_\theta\left[X \in \mathcal{A}\left(\phi_{g(\theta)}\right)\right] \ge 1-\alpha$
- [Confidence Intervals]
  - <u>Lower confidence interval</u>: invert right tailed test of $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$
  - <u>Upper confidence interval</u>: invert left tailed test of $H_0: \theta = \theta_0$ vs $H_1: \theta < \theta_0$
  - <u>Equal-tailed confidence interval</u>: invert the (equal) two-tailed test of $H_0: \theta = \theta_0$ vs $H_1: \theta \ne \theta_0$. Also, can compute via intersection of lower and upper confidence interval for $\frac{\alpha}{2}$ respectively
- [Unbiased] Let $\theta \in \Theta$ be an unknown parameter. Let $\Theta' \subset \Theta$ be a subset that does not contain the true parameter θ and $1-\alpha$ be a given confidence level. A $1-\alpha$ confidence set $C(X)$ is <u>Θ'-unbiased</u> if $\mathbb{P}[\theta' \in C(X)] \le 1-\alpha\ \forall \theta' \in \Theta'$
- [Uniformly Most Accurate Unbiased] Let $C(X)$ be a Θ'-unbiased confidence set with confidence coefficient $1-\alpha$. If $\mathbb{P}[\theta' \in C(X)] \le \mathbb{P}[\theta' \in C_1(X)]\ \forall \theta' \in \Theta'\ \forall C_1(X)$ that is Θ'-unbiased $1-\alpha$ confidence set, then $C(X)$ is <u>Θ'-uniformly most accurate unbiased</u>.

One-sided Test

- [One-sided Test] Given a family of models $\mathcal{P} = \{P_\theta: \theta \in \Theta \subset \mathbb{R}\}$ and $\theta_0 \in \Theta$, want to test:
  - $H_0: \theta \le \theta_0$
  - $H_1: \theta > \theta_0$
- [Likelihood Ratio Test]
  - Define $L(x) = \frac{\mathbb{P}_1[x]}{\mathbb{P}_0[x]}$
  - Critical function of the form $\phi^*(x) = \begin{cases} 1, & L(x) > c \\ \text{Bernoulli}(\gamma), & L(x) = c \\ 0, & L(x) < c \end{cases}$ where (c,γ) chosen s.t. $\beta_{\phi^*}(\theta_0) = \mathbb{E}_0[\phi^*(X)] = \alpha$
- [Type I Error] $P_{\theta_0}[\phi(X) = 1]$
  - Rejecting the null hypothesis when it is indeed true
- [Type II Error] $P_{\theta_1}[\phi(X) = 0]$
  - Failing to reject the null hypothesis
- [Neyman Pearson 12.2] Let $\alpha \in (0,1)$. Then, $\exists$ likelihood ratio test ϕ_α with significance level α. ϕ_α is optimal (maximises $\beta_\cdot(\theta_1)$) among all other tests ϕ with significance level at most α

- [12.3] For another other test ϕ that is optimal at significance level α, $\phi = \phi_\alpha$ i.e. ϕ must be the likelihood ratio test
- [12.9] Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ has monotone likelihood ratios. Then:
  - The likelihood ratio test $\phi^*(X)$ is uniformly most powerful for testing $H_0: \theta \le \theta_0$ vs $H_1: \theta > \theta_0$ and has level $\alpha = \mathbb{E}_{\theta_0}[\phi^*(X)]$
    - $\phi^*(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}$
  - The power function $\beta_{\phi^*}(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is nondecreasing in θ
  - If $\theta' < \theta_0$, the test ϕ^* minimises $\mathbb{E}_{\theta'}[\phi(X)]$ among all tests ϕ with $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha = \mathbb{E}_{\theta_0}[\phi^*(X)]$
    - i.e. the likelihood ratio test not only maximises power for $\theta > \theta_0$, it also minimises power for $\theta < \theta_0$

Two-sided Tests

- [Point Null] Given a family of models $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ and $\theta_0 \in \Theta$, want to test:
  - $H_0^{(P)}: \theta = \theta_0$
  - $H_1^{(P)}: \theta \ne \theta_0$
- [Interval Null] Given a family of models $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ and $\theta_1, \theta_2 \in \Theta$, want to test:
  - $H_0^{(I)}: \theta \in [\theta_1, \theta_2]$
  - $H_1^{(I)}: \theta \notin [\theta_1, \theta_2]$
- [$\mathcal{C}_m$] Let $\mathcal{C}_m$ denote the class of level-α tests ϕ s.t. $\beta'_\phi(\theta_0) = m$
- [Two-Sided Test] A test ϕ is <u>two-sided</u> if $\exists t_1, t_2$ with $t_1 < t_2$ s.t.
$$\phi(X) = \begin{cases} 1, & x \in (-\infty, t_1) \cup (t_2, \infty) \\ 0, & x \in [t_1, t_2] \end{cases}$$
- [Equal-Tailed Test] $\mathbb{P}_{\theta_0}[T(X) < c_1] = \mathbb{P}_{\theta_0}[T(X) > c_2] = \frac{\alpha}{2}$
- [UMP One-Sided] Let ϕ_+ and ϕ_- denote the UMP one-sided tests of level α.
  - Define $m_+ \coloneqq \beta'_{\phi_+}(\theta_0)$ and $m_- \coloneqq \beta'_{\phi_-}(\theta_0)$
- [12.17] Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Suppose $T(X)$ be sufficient for the model. Then, for any test $\phi(X)$, the test $\psi(T) = \mathbb{E}_\theta[\phi(X)|T]$ has the same power function as ϕ i.e. $\beta_\psi(\theta) = \beta_\phi(\theta)$ $\forall \theta \in \Theta$
  - *Prove by law of iterated expectation*
- [12.20] Let η be differentiable at θ and $\theta \in \text{int}(\Theta)$, then $\beta'(\theta) = \eta'(\theta)\mathbb{E}_\theta[T\phi] - B'(\theta)\beta(\theta)$
  - *Prove by differentiating (by applying dominated convergence theorem)*
- [12.22] Assume $X \sim e^{\eta(\theta)^T T(x) - A(\theta)} h(x)$ and $\theta_0 \in \text{int}(\Theta)$ and η differentiable and strictly increasing with $0 < \eta'(\theta_0) < \infty$. Then $\forall m \in (m_-, m_+)$, $\exists$ a two-sided level-α test ϕ^* s.t. $\beta'_\phi(\theta_0) = m$. ϕ^* is uniformly most powerful across all level-α tests with derivative constrained at θ_0
  - i.e. if there is another level-α test ψ s.t. $\beta'_\psi(\theta_0) = m$, $\mathbb{E}_\theta[\psi] \le \mathbb{E}_\theta[\phi^*]$ $\forall \theta \in \Theta$
  - [12.23] If ϕ^* is a two-sided test testing for $H_0: \theta \in [\theta_1, \theta_2]$ and $\mathbb{E}_{\theta_1}[\phi^*] = \alpha_1$ and $\mathbb{E}_{\theta_1}[\phi^*] = \alpha_2$. Then ϕ^* is uniformly most powerful among all tests with $\mathbb{E}_{\theta_1}[\phi] = \alpha_1$ and $\mathbb{E}_{\theta_1}[\phi] = \alpha_2$
- [12.26] Assume $X \sim e^{\eta(\theta)^T T(x) - A(\theta)} h(x)$ and $\theta_0 \in \text{int}(\Theta)$ and η differentiable and strictly increasing with $0 < \eta'(\theta_0) < \infty$. Then $\exists$ two-sided level-α test ϕ^* with $\beta'_{\phi^*}(\theta_0) = 0$. ϕ^* is uniformly most powerful testing $H_0: \theta = \theta_0$ vs $H_1: \theta \ne \theta_0$ among all unbiased tests with level-α.
  - Any two-sided test ϕ^* with level-α that is uncorrelated with T is uniformly most powerful unbiased.

- [Lecture 12.26] Assume $X_i \sim e^{\theta^T T(x) - A(\theta)} h(x)$. Then the unbiased test that rejects extreme values of the sufficient statistic $\sum_{i=1}^n T(x_i)$ with significance level α is UMP among all unbiased test (UMPU)
    - For $H_0^{(P)}$, the UMPU test can be found by solving for c_i, γ_i, $i \in \{1,2\}$, $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$, $\mathbb{E}_{\theta_0}[(\sum_{i=1}^n T(x_i))(\phi(X) - \alpha)] = 0$
    - For $H_0^{(I)}$, the UMPU test can be found by solving for c_i, γ_i, $i \in \{1,2\}$ such that $\mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)] = \alpha$
    - $\phi(X) = \begin{cases} 1, & T(x) \in (-\infty, c_1) \cup (c_2, \infty) \\ \gamma_i, & T(X) = c_i \\ 0, & T(X) \in (c_1, c_2) \end{cases}$

Nuisance Parameters

- [α-Similar] A test ϕ is <u>α-similar</u> if $\beta_\phi(\theta)$ is continuous and $\beta_\phi(\theta) = \alpha\ \forall \theta \in \bar{\Theta}_0 \cap \bar{\Theta}_1$
    - i.e. its power function is α on the common boundary of Θ_0 and Θ_1
    - Warning: α need not be the level of test ϕ
- [Neyman Structure] Let $T(X)$ be sufficient for the subfamily $\mathcal{P}' = \{P_\theta : \theta \in \Omega\} \subset \mathcal{P}$. Then an α-similar test ϕ has <u>Neyman structure</u> if $\mathbb{E}_\theta[\phi | T = t] = \alpha$ for a.e. $t\ \forall \theta \in \Omega$
- [13.3] Let ϕ^* be α-similar and is of level-α and UMP among all α-similar tests. Then ϕ^* is unbiased and uniformly most powerful among <u>all</u> unbiased tests
    - The unbiased test need not be of level-α
- [13.5] Let T be complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$. Then every similar test has Neyman structure.
- [Set Up]
    - [Model] $\mathcal{P} = \{P_{\theta,\lambda} : (\theta, \lambda) \in \Theta\}$, θ: parameter of interest; λ: nuisance parameter
    - [Null] $H_0 : \theta \in \Theta_0$
    - [Alternate] $H_1 : \theta \in \Theta_1$
- [13.6] Let $\theta, \theta_0 \in \mathbb{R}$, $\lambda \in \mathbb{R}^r$, $(\theta, \lambda) \in \Omega$ open. Assume $\mathcal{P}$ is a full-rank exponential family with densities $P_{\theta,\lambda}(x) = e^{\theta^T T(x) + \lambda^T U(x) - A(\theta,\lambda)} h(x)$, where θ is parameter of interest and λ is nuisance parameter.
    - [One-Sided] To test $H_0 : \theta \le \theta_0$, $H_1 : \theta > \theta_0$, $\exists$ a UMPU test $\phi^*(X) = \psi(T(X), U(X))$ where: $\psi(t,u) = \begin{cases} 1, & t > c(u) \\ \text{Bernoulli}(\gamma), & t = c(u) \\ 0, & t < c(u) \end{cases}$ with $\gamma(u), c(u)$ chosen s.t. $\mathbb{E}_{\theta_0}[\phi^*(X) | U(X) = u] = P_{\theta_0}[T(X) > c(u) | U(X) = u] = \alpha$
    - [Point Null] To test $H_0 : \theta = \theta_0$, $H_1 : \theta \ne \theta_0$, $\exists$ a UMPU test $\phi^*(X) = \psi(T(X), U(X))$ where: $\psi(t,u) = \begin{cases} 1, & t \in (-\infty, c_1(u)) \cup (c_2(u), \infty) \\ \text{Bernoulli}(\gamma), & t \in \{c_1(u), c_2(u)\} \\ 0, & t \in (c_1(u), c_2(u)) \end{cases}$ with $\gamma(u), c(u)$ chosen s.t. $\mathbb{E}_{\theta_0}[\phi^*(X) | U(X) = u] = \alpha$, $\mathbb{E}_{\theta_0}[T(X)(\phi^*(X) - \alpha) | U(X) = u] = 0$
- [One Sample t-Test] Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Test $H_0 : \mu = 0$ vs $H_1 : \mu \ne 0$
    - $\bar{X} \perp S_X^2$

t-Test

- [Set Up] $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
    - $H_0 : \mu \le 0$
    - $H_1 : \mu > 0$
- $\frac{X}{||X||} \perp ||X||^2$ (Basu)
- Reject H_0 if $\frac{\sqrt{(n-1)n}\bar{X}}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{1}{n}\bar{X}^2}} > t_{1-\alpha}$

- $\frac{(\bar{X}-\mu)}{\frac{S_X}{\sqrt{n}}} \sim t_{n-1}$

Permutation Test

- [Set Up] Let $X_1, X_2, \ldots, X_{n_a} \sim P$ and $Y_1, Y_2, \ldots, Y_{n_b} \sim Q$
  - $H_0: P = Q$
  - $H_1: P \neq Q$
- [Assumption] Exchangeability i.e. $(X_1, \ldots, X_n)$ is equal in distribution to $\left(X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$ for all permutations π.
- Let $T(X, Y) = \bar{X} - \bar{Y}$. Let $T_0 = T(X, Y)$.
- For $i \in \{1, \ldots, B\}$, obtain $(X_i, Y_i) = \pi(X, Y)$ and $T_i = T(X_i, Y_i)$
- Reject H_0 if T_0 falls in the upper α quantile (i.e. among the top $\alpha(B+1)$ test statistic) of the Monte-Carlo distribution of T.

General Linear Models

Definitions

- [Exponential Family] $Y_i \sim p_{\eta_i}(y) = e^{\eta_i y - A(\eta_i)} h(y)$
  - η is the predictor
  - $\mu(\eta) = \nabla_\eta A(\eta)$
- [Response] Y the random component
- [Covariates / Regressors] The systematic component of GLM i.e. $x_1, x_2, \ldots$
- [Linear Predictor] $\eta_i = \beta^T x_i$
- [Link Function] A <u>link function</u> is a smooth and invertible function g mapping the expectation of the response $\mu_i = \mathbb{E}[Y_i]$ to the predictor η_i
  - $g(\mu_i) = \eta_i$
  - Links the random and the systematic components
- [Mean Function] The <u>mean function</u> g^{-1} is the inverse of the link function
  - g^{-1} is the conditional expectation of the response variable $g^{-1}(\eta_i) = \mu_i$
  - $= \mathbb{E}_{\eta_i}[Y_i]$

Distributions

- [χ^2 Distribution] Let $Z_1, \ldots, Z_d \sim N(0,1)$ and $V = \sum_{i=1}^d Z_i^2 = \|Z\|^2$. Then $V \sim \chi_d^2$.
  - $\mathbb{E}[V] = d$
  - $\operatorname{Var}[V] = 2d$
  - $\mathbb{E}\left[\frac{1}{V}\right] = \frac{1}{d-2}$
  - $\operatorname{Var}\left[\frac{1}{V}\right] = \frac{2}{(d-2)^2(d-4)}$
  - $\frac{n-1}{\sigma^2} S_X^2 \sim \chi_{n-1}^2$
- [t Distribution] Let $Z \sim N(0,1)$, $V \sim \chi_d^2$. Then $\frac{Z}{\sqrt{\frac{V}{d}}} \sim t_d$
  - $\frac{(\bar{X}-\mu)}{\frac{S_X}{\sqrt{n}}} \sim t_{n-1}$
  - $t_d \to N(0,1)$ in distribution as $d \to \infty$
  - Fatter tails
- [F Distribution] Let $V_1 \sim \chi_{d_1}^2$ and $V_2 \sim \chi_{d_2}^2$ be independent. Then $\frac{\frac{V_1}{d_1}}{\frac{V_2}{d_2}} \sim F_{d_1,d_2}$
  - If $d_2 \gg d_1$, then $F_{d_1,d_2} \to \frac{1}{d_1}\chi_{d_1}^2$ in distribution
  - $t_d^2 \sim F_{1,d}$
- [Facts]
  - [Cochran's Theorem] Let $Z_1, \ldots, Z_n \sim N(0,1)$ i.i.d., then $\sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \chi_{n-1}^2$
  - [Sample Variance Properties]
    - $S_X^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$
    - $\bar{X} \perp S_X^2$
    - $(n-1)S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \|X\|^2 - n\bar{X}^2$

Canonical Linear Model

- Let $Z = \begin{bmatrix} Z_0 \\ Z_1 \\ Z_r \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ 0 \end{bmatrix}, \sigma^2 \mathbb{I}_n\right)$ with $Z \in \mathbb{R}^n = \mathbb{R}^{d_0+d_1+d_r}$
- Test
  - $H_0: \mu_1 = 0$
  - $H_1: \mu_1 \neq 0$
- Density: $P_{\mu_0,\mu_1,\sigma}(z) \propto e^{-\frac{1}{2\sigma^2}\|z\|^2 + \frac{\mu_0^T z_0}{\sigma^2} + \frac{\mu_1^T z_1}{\sigma^2}}$
- Case #1 (z-test): σ^2 known, $d_1 = 1$

- Nuisance parameter: $\frac{\mu_0^T Z_0}{\sigma^2}$
  - Condition on Z_0 but Z_1 independent of Z_0 anyways
- Reject extreme values of $\frac{Z_1}{\sigma}$
- $\frac{Z_1}{\sigma} \sim N(0,1)$ under H_0
- $\phi^*(Z) = \begin{cases} 1, & \left|\frac{Z_1}{\sigma}\right| > c \\ 0, & \left|\frac{Z_1}{\sigma}\right| \leq c \end{cases}$

- Case #2 (χ^2-test): σ^2 known, $d_1 \geq 1 \Rightarrow$
  - Reject extreme values of $\frac{\|Z_1\|^2}{\sigma^2}$
  - $\frac{\|Z_1\|^2}{\sigma^2} \sim \chi^2_{d_1}$ under H_0
- Case #3 (t-test): σ^2 unknown, $d_1 = 1$
  - Nuisance parameters: $-\frac{1}{2\sigma^2}\|Z\|^2 + \frac{\mu_0^T Z_0}{\sigma^2}$
  - Reject extreme values of $\frac{Z_1}{\|Z\|}$
    - Equivalently, reject extreme values of $\frac{Z_1}{\sqrt{\frac{\|Z_r\|^2}{d_r}}}$
  - $\frac{Z_1}{\sqrt{\frac{\|Z_r\|^2}{d_r}}} \sim t_{d_r}$ under H_0
- Case #4 (F-test): σ^2 unknown, $d_1 \geq 1$
  - Reject extreme values of $\|Z_1\|^2$
    - Equivalently, reject extreme values of $\frac{\|Z_1\|}{\|Z_r\|}$
  - $\frac{\frac{\|Z_1\|^2}{d_1}}{\frac{\|Z_r\|^2}{d_r}} \sim F_{d_1,d_r}$

General Linear Model

- $Y \sim N_n(\theta, \sigma^2 \mathbb{I}_n)$
- Test the following hypothesises, where $\Theta_0 \subset \Theta_1 \subset \mathbb{R}^n$ are linear subspaces
  - $H_0: \theta \in \Theta_0$
  - $H_1: \theta \in \Theta_1$
- Let $Q = [Q_0 \quad Q_1 \quad Q_r] \in \mathbb{R}^{n \times n}$ orthonormal, where Q_0 is a basis for Θ_0, $[Q_0 \quad Q_1]$ is a basis for Θ_1 and Q is a basis for Θ
  - $Q^T Y \sim N_n(Q^T\theta, \sigma^2 \mathbb{I}_n)$
- Reduces to the following hypothesis:
  - $H_0: Q_1^T \theta = 0$
  - $H_1: Q_1^T \theta \neq 0$

Linear Regression

- $\|Q_1^T Y\|^2 = RSS_0 - RSS$
- $\|Q_1^T Y\|^2 + \|Q_r^T Y\|^2 = RSS_0$
- [F-Statistic] $\frac{\frac{\|Z_1\|^2}{d_1}}{\frac{\|Z_r\|^2}{d_r^2}} = \frac{RSS_0 - RSS}{RSS} \cdot \frac{d_r}{d_1}$

Asymptotic Theory

Definitions

- [Convergence] $X_n \in \mathbb{R}^d$, $c \in \mathbb{R}^d$
    - [Convergence in Probability] $(X_n)_n \xrightarrow{\mathbb{P}} c$ if $\lim_{n\to\infty} \mathbb{P}[\|X_n - c\| > \epsilon] = 0\ \forall \epsilon > 0$
    - [Convergence in Distribution] $(X_n)_n \xrightarrow{d} X$ if $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ $\forall$ bounded, continuous $f: X \to \mathbb{R}$
        - If $d = 1$, then equivalent definition is $\lim_{n\to\infty} \mathbb{P}[X_n \le x] = \mathbb{P}[X \le x]\ \forall x$
- [Consistent] Let $(\mathcal{P}_n)_n$ be a sequence of models (i.e. $\mathcal{P}_n = \{P_{n,\theta}: \theta \in \Theta\}$). Then, a sequence of estimators $\left(\delta_n(X_n)\right)_n$ where $X_n \sim P_{n,\theta}$ is consistent for $g(\theta)$ if $\left(\delta_n(X_n)\right)_n \xrightarrow{\mathbb{P}_\theta} g(\theta)\ \forall \theta \in \Theta$
    - For each $\theta \in \Theta$, $\lim_{n\to\infty} \mathbb{P}_\theta[|\delta_n(X_n) - g(\theta)| > \epsilon] = 0\ \forall \epsilon > 0$
    - As n grows, the upgraded estimator δ_n converges to the actual estimand under the true model
- [Maximum Likelihood Estimator] Let $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$ be a dominated family. Then $\hat{\theta}_{\text{MLE}}(X) = \arg\max_{\theta\in\Theta} p_\theta(X) = \arg\max_{\theta\in\Theta} l(\theta; X)$
    - MLE for $g(\theta)$ is $g(\theta_{\text{MLE}})$
- [Asymptotic Relative Efficiency] Let $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ be asymptotically normal with $\sqrt{n}\left(\hat{\theta}^{(i)} - \theta\right) \xrightarrow{d} N(0, \sigma_i^2)$, then the asymptotic relative efficiency of $\hat{\theta}^{(2)}$ w.r.t. $\hat{\theta}^{(1)}$ is $\frac{\sigma_1^2}{\sigma_2^2}$
    - If $\frac{\sigma_1^2}{\sigma_2^2} = \gamma < 1$, then using $\hat{\theta}^{(2)}$ is asymptotically equivalent to using $\hat{\theta}^{(1)}$ but throwing away $1 - \gamma$ of data.
- [Asymptotically Efficient] An estimator $\hat{\theta}_n$ is asymptotically efficient if $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{P_\theta} N\left(0, \left(\mathcal{J}_1(\theta)\right)^{-1}\right)$
    - i.e. achieves the Cramér-Rao lower bound
    - If $\hat{\theta}_n$ is asymptotically efficient, then $\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{P_\theta} N\left(0, (\nabla g)^T \left(\mathcal{J}_1(\theta)\right)^{-1} (\nabla g)\right)$
- [Kullback-Leibler Divergence] $D_{KL}(\theta_0 || \theta) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right)\right]$
    - $D_{KL}(\theta_0 || \theta) > 0$ unless $p_{\theta_0} = p_\theta$

Tools (Large Sample Theory)

- $(X_n)_n \xrightarrow{\mathbb{P}} c$ if and only if $(X_n)_n \xrightarrow{d} \delta_c$
- [WLLN] If $\mathbb{E}[\|X_n\|] < \infty$, $\mathbb{E}[X_n] = \mu$, then $(\bar{X}_n)_n \xrightarrow{\mathbb{P}} \mu$
- [Central Limit Theorem] If $\mathbb{E}[X_n] = \mu$, $\text{Var}[X_n] = \Sigma$, then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma)$
    - $\bar{X}_n \sim N\left(\mu, \frac{1}{n}\Sigma\right)$
- [Continuous Mapping Theorem] Let f be continuous and $X_1, X_2, \ldots$ be random variables.
    - If $(X_n)_n \xrightarrow{d} X$, then $f(X_n) \xrightarrow{d} f(X)$
    - If $(X_n)_n \xrightarrow{\mathbb{P}} c$, then $f(X_n) \xrightarrow{\mathbb{P}} f(c)$
- [Slutsky] Let $(X_n)_n \xrightarrow{d} X$, $Y_n \xrightarrow{\mathbb{P}} c$. Then:
    - $(X_n + Y_n)_n \xrightarrow{d} X + c$
    - $(X_n Y_n)_n \xrightarrow{d} cX$
    - $\left(\frac{X_n}{Y_n}\right)_n \xrightarrow{d} \frac{X}{c}$ for $c \in \mathbb{R}\backslash\{0\}$

- [Delta Method] Assume $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ and $f(x)$ differentiable at $x = \mu$, then $\sqrt{n}\big(f(X_n) - f(\mu)\big) \xrightarrow{d} N(0, \sigma^2 f'(\mu)^2)$
- [Multivariate Delta Method] Assume $\sqrt{n}(X_n - \mu) \xrightarrow{d} N_d(0, \Sigma)$ and $f: \mathbb{R}^d \to \mathbb{R}^k$ differentiable at $x = \mu$, then $\sqrt{n}\big(f(X_n) - f(\mu)\big) \xrightarrow{d} N_k(0, (Df)\Sigma(Df)^T)$
- [Method of Moments]
- [Good Event Bad Event Lemma 9.15] Suppose $(Y_n) \xrightarrow{d} Y$ and $\lim_{n\to\infty} \mathbb{P}[B_n] = 1$, then for arbitrary $(Z_n)_n$, $Y_n \mathbb{1}_{B_n} + Z_n \mathbb{1}_{B_n^c} \xrightarrow{d} Y$
  - To show convergence in distribution, only care about events with probabilities that converge to 1

Weak Law (Definitions)

- [$C(\Theta)$] Let $\Theta \subset \mathbb{R}^p$ be compact. Then $C(\Theta)$ is the space of continuous functions on Θ.
- [Random Function] Let $\Theta \subset \mathbb{R}^p$ be compact. Define $(X_n)_n$ to be a source of randomness i.i.d. and $W_i(\theta) = h(\theta, X_i)$ where $h(\cdot, x) \in C(\Theta)\ \forall x$. Then $(W_n)_n$ is a sequence of random functions.
- [L^∞ Norm] Let $w \in C(\Theta)$, then $\|w\|_\infty = \sup_{\theta\in\Theta} |w(\theta)|$
- [Convergence in L^∞] Let $(w_n)_n, w \in C(K)$. Then $w_n \to w$ in L^∞ if $\lim_{n\to\infty} \|w_n - w\|_\infty = 0$
- [Banach Space] A <u>Banach space</u> is a complete normed vector space.
- [Dense] Let $B \subset A$. Then B is <u>dense</u> in A if $\forall x \in A$, $\forall \epsilon > 0$, $\exists y \in B$ s.t. $\|x - y\| < \epsilon$
- [Separable] A space is <u>separable</u> if it has a countable dense subset.

Weak Law (Theorems)

- [$(C(\Theta), L^\infty)$] Let Θ be compact.
  - $(C(\Theta), L^\infty)$ is a Banach space (a complete, linear space equipped with a norm)
  - $(C(\Theta), L^\infty)$ is separable
- [Dini] Let $(f_n)_n \to f$ monotonously pointwise on compact space K. If f is also continuous, then the convergence is uniform.
- [9.1] Let Θ be compact and W be a random function in $C(\Theta)$. Let $\mu: \Theta \to \mathbb{R}$ with $\mu(\theta) = \mathbb{E}[W(\theta)]$. Assuming that $\mathbb{E}[\|W\|_\infty] < \infty$, then:
  - μ is continuous (*prove via dominated convergence theorem*)
  - $\lim_{\epsilon\to 0} \sup_{\theta\in\Theta} \mathbb{E}\left[\sup_{\theta':\|\theta'-\theta\|<\epsilon} |W(\theta') - W(\theta)|\right] = 0$ (*prove via Dini*)
    - i.e. uniform convergence of expected difference between close-by points
- [9.2] Let Θ be compact and $(W_n)_n \in C(\Theta)$ i.i.d. with mean μ (i.e. $\mu(\theta) = \mathbb{E}[W(\theta)]$) and $\mathbb{E}[\|W_i\|_\infty] < \infty$. Let $\bar{W}_n = \frac{W_1 + \cdots + W_n}{n}$. Then $\|\bar{W}_n - \mu\|_\infty \xrightarrow{\mathbb{P}} 0$
  - $\forall \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left[\sup_{\theta\in\Theta} \|\bar{W}_n(\theta) - \mu(\theta)\| > \epsilon\right] = 0$
  - This upgrades convergence in probability due to WLLN. Actually, it can be upgraded to convergence almost surely
- [9.4] Let Θ be compact. Let $(G_n)_n \in C(\Theta)$ be random functions and $\|G_n - g\|_\infty \xrightarrow{\mathbb{P}} 0$ with $g \in C(\Theta)$ be a nonrandom function.
  - Let $(X_n)_n \xrightarrow{\mathbb{P}} x^*$ where $x^* \in \Theta$ is a constant, then $\big(G_n(X_n)\big)_n \xrightarrow{\mathbb{P}} g(x^*)$
  - Let g achieve maximum at unique point x^* and $(X_n)_n$ are random variables maximising G_n i.e. $G_n(X_n) = \sup_{X\in\Theta} G_n(X)$, then $(X_n)_n \xrightarrow{\mathbb{P}} x^*$
  - Let $\Theta \subset \mathbb{R}$ and $g(x) = 0$ has a unique solution x^*. If $(X_n)_n$ are random variables s.t. $G_n(X_n) = 0$, then $(X_n)_n \xrightarrow{\mathbb{P}} x^*$

- Upshot: Uniform convergence allows for convergence in probability of sequences, maxima and solutions
- [Consistency of $\hat{\theta}_n$ 9.9] Let Θ be compact and $\mathcal{P}$ be an identifiable model with densities p_θ continuous in θ. Suppose $\mathbb{E}_{\theta_0}\left[\left\|\log p_\theta - \log p_{\theta_0}\right\|_\infty\right] < \infty$ and. Then, under P_{θ_0}, $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
  - $\mathbb{E}_{\theta_0}\left[\left\|\log p_\theta - \log p_{\theta_0}\right\|_\infty\right] < \infty$ is equivalent to $\mathbb{E}_{\theta_0}\left[\sup_{\theta\in\Theta}\left|\log p_\theta - \log p_{\theta_0}\right|\right] < \infty$
  - i.e. MLE $\hat{\theta}_n$ is consistent
  - *Prove via: KL divergence guarantees uniqueness of maximum, then apply (9.2) and (9.4)*
- [Consistency of $\hat{\theta}_n$ 9.11] Let $\Theta \subset \mathbb{R}^n$ and $\mathcal{P}$ be an identifiable model with densities p_θ continuous in θ and $p_\theta(x) \to 0$ as $\|\theta\| \to \infty$. Suppose:
  - $\mathbb{E}_{\theta_0}\left[\left\|\left(\log p_\theta - \log p_{\theta_0}\right)\mathbb{1}_K\right\|_\infty\right] = \mathbb{E}_{\theta_0}\left[\sup_{\theta\in K}\left|\log p_\theta - \log p_{\theta_0}\right|\right] < \infty\ \forall K \subset \Theta$ compact.
  - $\mathbb{E}_{\theta_0}\left[\sup_{\theta:\|\theta\|>M}\left|\log p_\theta - \log p_{\theta_0}\right|\right] < \infty$ for some $M > 0$

  Then, under P_{θ_0}, $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
  - *Prove via considering the ball $\overline{B_r(\theta_0)}$ and showing $\mathbb{P}\left[\hat{\theta}_n \notin \overline{B_r(\theta_0)}\right] \to 0$, then using good-event-bad-event lemma*
- [Asymptotic Efficiency of MLE 9.14] Assume the following conditions:
  - $(X_i)_i$ i.i.d. with common density p_{θ_0} with $\theta_0 \in \Theta \subset \mathbb{R}^d$
  - The MLE estimator $\hat{\theta}_n$ is consistent i.e. $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$
  - $\exists \epsilon > 0$ s.t. $\overline{B_\epsilon(\theta_0)} = \{\theta: \|\theta - \theta_0\| < \epsilon\} \subset \Theta$ and:
    - $\nabla_\theta^2 l(\theta; x)$ exists (i.e. l is twice differentiable in θ $\forall x$)
    - $\mathbb{E}_{\theta_0}\left[\sup_{\theta\in B_\epsilon(\theta_0)}\left\|\nabla_\theta^2 l(\theta; x)\right\|\right] < \infty$
  - Sufficient regularity to interchange derivatives and integrals (e.g. $\frac{\partial^3 l}{\partial\theta^3}$ bounded)

  Then, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \left(\mathcal{J}_1(\theta_0)\right)^{-1}\right)$ under P_{θ_0} (i.e. MLE achieves asymptotic efficiency)
  - *Eventually, $\left\{\hat{\theta}_n \notin \overline{B_\epsilon(\theta_0)}\right\}$ is a measure 0 event. Prove by Taylor expanding $\nabla_\theta l_n\left(\hat{\theta}_n\right)$ around θ_0 and use tools.*

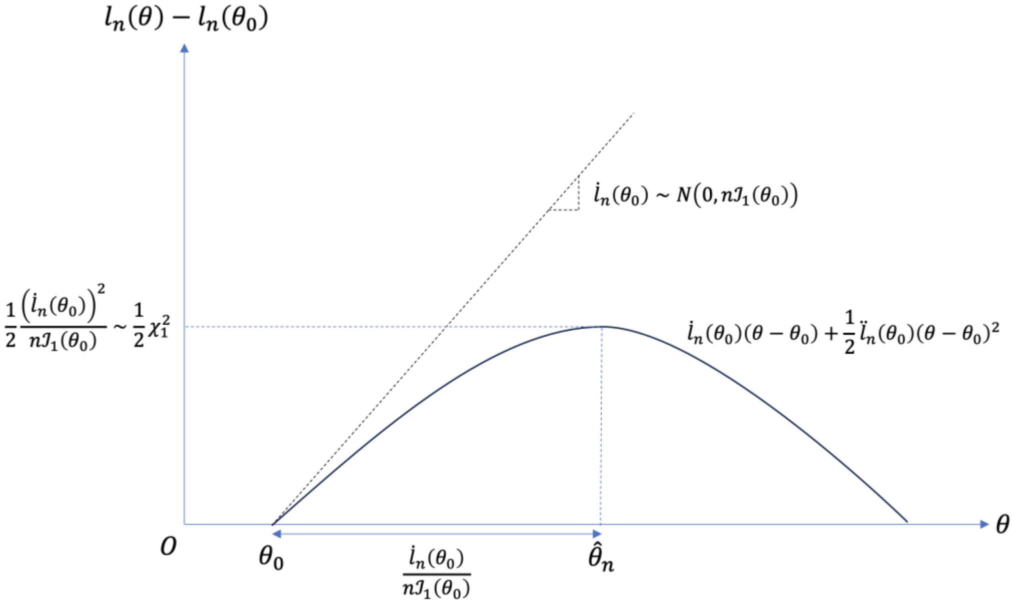
Likelihood Manipulations

- $l_n(\theta) - l_n(\theta_0)$ is a minimal sufficient statistic (log of likelihood ratio)
- $l_n(\theta) - l_n(\theta_0) \approx \dot{l}_n(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ddot{l}_n(\theta_0)(\theta - \theta_0)^2 \approx$
- $\frac{1}{\sqrt{n}}\dot{l}_n(\theta_0) \xrightarrow{d} N\left(0, \mathcal{J}_1(\theta_0)\right)$
- $\frac{1}{n}\ddot{l}_n(\theta_0) \xrightarrow{p_{\theta_0}} -\mathcal{J}_1(\theta_0)$
- $\ddot{l}_n(\theta_0) = -\mathcal{J}_n = -n\mathcal{J}_1$
- $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{J}_1(\theta_0)^{-1}\right)$

Three Musketeers

- [Score Test]
  - [Score Statistic] $\dot{l}_n(\theta_0) \sim N\left(0, n\mathcal{J}_1(\theta_0)\right)$
  - $\left(\mathcal{J}_n(\theta_0)\right)^{-\frac{1}{2}}\nabla_\theta l_n(\theta_0; X) \xrightarrow{d} N_d(0, \mathbb{I}_d)$
  - $H_0: \theta = \theta_0$
    - $(d > 1)$ Reject H_0 if $\left\|\mathcal{J}_n(\theta_0)^{-\frac{1}{2}}\nabla_\theta l_n(\theta_0; X)\right\|_2^2 > \chi_d^2(\alpha)$
    - $(d = 1)$ Reject H_0 if $\left|\frac{\dot{l}_n(\theta_0)}{\sqrt{J_n(\theta_0)}}\right| > Z\left(1 - \frac{\alpha}{2}\right)$ (can do 1-sided or 2-sided test)

- Score Test prioritises alternatives close to Θ_0
- [Wald Test]
  - [Wald Statistic] $\frac{\dot{l}_n(\theta_0)}{n\mathcal{I}_1}$
  - Let $\hat{\mathcal{J}}_n$ be an estimator s.t. $\frac{1}{n}\hat{\mathcal{J}}_n \xrightarrow{p_{\theta_0}} \mathcal{I}_{\theta_0}$
    - $\hat{\mathcal{J}}_n = n\mathcal{I}_1(\hat{\theta}_n)$
    - [Observed Fisher Information] $\hat{\mathcal{J}}_n = -\nabla^2 l_n(\hat{\theta}_n; X)$
  - $\left\|\hat{\mathcal{J}}_n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0)\right\|^2 \xrightarrow{d} \chi_d^2$
  - $H_0: \theta = \theta_0$
    - Reject H_0 if $\left\|\hat{\mathcal{J}}_n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0)\right\|^2 > \chi_d^2(\alpha)$
  - [Confidence Interval] $(\hat{\theta}_n)_j \sim N\left((\theta_0)_j, (\mathcal{I}_n(\theta_0)^{-1})_{jj}\right)$
    - $\left((\hat{\theta}_n)_j - \sqrt{(\hat{\mathcal{J}}_n^{-1})_{jj}} Z_{\frac{\alpha}{2}}, (\hat{\theta}_n)_j + \sqrt{(\hat{\mathcal{J}}_n^{-1})_{jj}} Z_{\frac{\alpha}{2}}\right)$
  - [Confidence Ellipsoid]
- [Generalised Likelihood Ratio Test]
  - $2\left(l_n(\hat{\theta}_n) - l_n(\theta_0)\right) \xrightarrow{d} \chi_d^2$
  - $H_0: \theta = \theta_0,\ H_1: \theta \neq \theta_0$
    - Reject H_0 if $\left\|2\left(l_n(\hat{\theta}_n) - l_n(\theta_0)\right)\right\|^2 > \chi_d^2(\alpha)$
  - $H_0: \theta \in \Theta_0,\ H_1: \theta \in \Theta \backslash \Theta_0$ where Θ_0 is a d-dimensional manifold
    - If $\theta_0 \in \mathrm{relint}(\Theta_0)$, then $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$
    - $2\left(l_n(\hat{\theta}_n) - l_n(\theta_0)\right) \xrightarrow{d} \chi_{d-d_0}^2$
    - Reject H_0 if $\left\|2\left(l_n(\hat{\theta}_n) - l_n(\theta_0)\right)\right\|^2 > \chi_{d-d_0}^2(\alpha)$



Miscellaneous Tests

- [Pearson χ^2 Test] $N = (N_1, \dots, N_d) \sim \mathrm{Multinomial}\left(n, (\pi_1, \dots, \pi_d)\right)$
  - $H_0: \pi = \pi_0,\ H_1: \pi \neq \pi_0$ (Score test in disguise)
  - [Test Statistic] $\sum_{i=1}^d \frac{(N_j - n(\pi_0)_j)^2}{n(\pi_0)_j} \xrightarrow{d} \chi_{d-1}^2$

Bootstrapping

Definitions

- [Set-up]
  - $X_1, \dots, X_n \sim \mathcal{P}$
  - [Functional / Parameter] $\theta(\mathcal{P})$
    - Given a distribution $\mathcal{P}$, can evaluate θ
  - [Empirical Distribution] $\hat{\mathcal{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$
    - Recall that distribution is just push-forward measure
    - $\hat{\mathcal{P}}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \in A\}$
    - Bootstrap is just sampling from $\hat{\mathcal{P}}_n$ with replacement
  - [Plug-in Estimator] The plug-in estimator for $\theta(\mathcal{P})$ is: $\theta(\hat{\mathcal{P}}_n)$
  - [Standard Error] Let $\hat{\theta}_n := \hat{\theta}(X_1, \dots, X_n)$ denote an estimator for $\theta(\mathcal{P})$ after n observations. The standard error is: $se_{\mathcal{P}}(\hat{\theta}_n) := \sqrt{\mathrm{Var}_{\mathcal{P}}[\hat{\theta}_n]}$. *An* estimate for the standard error is $\widehat{\mathrm{se}}(\hat{\theta}_n) := \sqrt{\mathrm{Var}_{\hat{\mathcal{P}}_n}[\hat{\theta}_n]}$
    - $\mathrm{Var}_{\mathcal{P}}[\hat{\theta}_n] = \mathbb{E}_{X\sim\mathcal{P}}\left[\left(\hat{\theta}_n(X) - \mathbb{E}_{X\sim\mathcal{P}}[\hat{\theta}_n(X)]\right)^2\right]$
    - $\mathrm{Var}_{\hat{\mathcal{P}}_n}[\hat{\theta}_n] = \mathbb{E}_{X\sim\hat{\mathcal{P}}_n}\left[\left(\hat{\theta}_n(X) - \mathbb{E}_{X\sim\hat{\mathcal{P}}_n}[\hat{\theta}_n(X)]\right)^2\right]$
    - Typically, use Monte-Carlo to calculate
- [Bias Correction]
  - [True Bias] $\mathrm{Bias}_{\mathcal{P}}[\hat{\theta}_n] = \mathbb{E}_{X\sim\mathcal{P}}[\hat{\theta}_n(X)] - \theta(\mathcal{P})$
    - Cannot compute since do not know $\mathcal{P}$
  - [Estimate] $\mathrm{Bias}_{\hat{\mathcal{P}}_n}[\hat{\theta}_n] = \mathbb{E}_{X\sim\hat{\mathcal{P}}_n}[\hat{\theta}_n(X)] - \theta(\hat{\mathcal{P}}_n)$
    - Can compute via Monte Carlo
- [Estimation Error] Let $\hat{\theta}_n(X_1, \dots, X_n)$ be an estimator after n observations. Then, the <u>estimation error</u> is: $R_n(X, \mathcal{P}) := \hat{\theta}_n(X) - \theta(\mathcal{P})$
  - Remark: not a statistic, since it depends on $\mathcal{P}$
  - $\mathbb{E}_{X\sim\mathcal{P}}[R_n(X,\mathcal{P})] = \mathrm{Bias}_{\mathcal{P}}[\hat{\theta}_n]$
  - $\mathbb{E}_{X\sim\hat{\mathcal{P}}_n}[R_n(X,\hat{\mathcal{P}}_n)] = \mathrm{Bias}_{\hat{\mathcal{P}}_n}[\hat{\theta}_n]$
  - Other possible definitions include:
    - $R_n(X,\mathcal{P}) := \frac{\hat{\theta}_n(X) - \theta(\mathcal{P})}{\theta(\mathcal{P})}$
    - $R_n(X,\mathcal{P}) := \frac{\hat{\theta}_n(X) - \theta(\mathcal{P})}{\hat{\sigma}(X)}$, where $\hat{\sigma}$ is some estimate of standard error of $\hat{\theta}_n$
      - e.g. $\hat{\sigma}(X) := \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$
- [Confidence Interval] Let $\theta(\mathcal{P})$ be a parameter and $\hat{\theta}_n(X_1, \dots, X_n)$ be an estimator for θ after n observations.
  - Define $G_{n,\mathcal{P}}(r) = \mathbb{P}_{X\sim\mathcal{P}}[R_n(X,\mathcal{P}) < r] = \mathbb{P}_{X\sim\mathcal{P}}[\hat{\theta}_n(X) - \theta(\mathcal{P}) < r]$
    - $G_{n,\mathcal{P}}(r)$ is the CDF of $R_n(X,\mathcal{P})$
  - Define $\hat{r}_1 := G^{-1}_{n,\hat{\mathcal{P}}_n}\left(\frac{\alpha}{2}\right)$ and $\hat{r}_2 := G^{-1}_{n,\hat{\mathcal{P}}_n}\left(1 - \frac{\alpha}{2}\right)$
    - $[\hat{r}_1, \hat{r}_2]$ are the $(1-\alpha)$ quantile of the estimation error
  - Then, the $(1-\alpha)$-<u>confidence interval</u> for $\theta(\mathcal{P})$ given n observations is: $C_{n,\alpha} = [\hat{\theta}_n - \hat{r}_2, \hat{\theta}_n - \hat{r}_1]$
- [Coverage Probability] Define the coverage probability of confidence interval $C_{n,\alpha}$ as: $\gamma_{n,\mathcal{P}}(\alpha) := \mathbb{P}_{X\sim\mathcal{P}}[\theta(\mathcal{P}) \in C_{n,\alpha}]$
  - For $C_n = C_{n,\alpha}$, $\gamma_{n,\mathcal{P}}(\alpha) = \mathbb{P}_{X\sim\mathcal{P}}\left[\theta(\mathcal{P}) \in [\hat{\theta}_n - \hat{r}_2, \hat{\theta}_n - \hat{r}_1]\right]$

- Estimate coverage probability via: $\gamma_{n,\hat{\mathcal{P}}_n}(\alpha) = \mathbb{P}_{X\sim\hat{\mathcal{P}}_n}\left[\theta(\hat{\mathcal{P}}_n) \in [\hat{\theta}_n - \hat{r}_2, \hat{\theta}_n - \hat{r}_1]\right]$
- Remark: This could be difference from $1-\alpha$ due to dependency on n by $C_{n,\alpha}$

- [Double Bootstrap]
  - [Idea]
    - First round of bootstrap to get empirical distribution for second round of bootstrap
    - Second round of bootstrap for constructing CDF of estimation error
    - Each iteration of first round of bootstrap gives a collection of $(1-\alpha)$ confidence intervals. Each collection is used to get coverage probability.
  - [Algorithm]
    - For a in $\{1, \dots, A\}$:
      - Sample $X_1^{*a}, \dots, X_n^{*a} \sim \hat{\mathcal{P}}_n$ # first layer of bootstrap
      - $\hat{\mathcal{P}}_n^{*a} \leftarrow \frac{1}{n}\sum_{i=1}^n \delta_{X_i^{*a}}$ # get empirical distribution formula
      - For b in $\{1, \dots, B\}$:
        - Sample $X_1^{**a,b}, \dots, X_n^{**a,b} \sim \hat{\mathcal{P}}_n^{*a}$ # second layer of bootstrap
        - $R_n^{**a,b} \leftarrow \frac{\hat{\theta}_n(X^{**a,b}) - \theta(\hat{\mathcal{P}}^{*a})}{\hat{\sigma}(X^{**a,b})}$
      - $\hat{G}_n^{*a} \leftarrow \mathrm{cdf}(R_n^{**a,1}, \dots, R_n^{**a,B})$
      - For $\alpha \in$ grid:
        - $c_{n,\alpha}^{*a} \leftarrow [\hat{\theta}_n^{*a} - \hat{\sigma}^{*a} r_2(\hat{G}_n^{*a}), \hat{\theta}_n^{*a} - \hat{\sigma}^{*a} r_1(\hat{G}_n^{*a})]$
    - For $\alpha \in$ grid:
      - $\hat{\gamma}(\alpha) \leftarrow \frac{1}{A}\sum_{i=1}^{A} \mathbb{1}\{C_{n,\alpha}^{*a} \ni \theta(\hat{\mathcal{P}}_n)\}$
    - $\hat{\alpha} \leftarrow \hat{\gamma}^{-1}(1-\alpha)$

Diagram