Quals Review

Background Info / Things that Showed up in Multiple Classes

Distributions

- (Univariate) $Normal(\mu, \sigma^2):\ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)};\ -\infty < x < \infty;\ -\infty < \mu < \infty$
- $Gamma(\alpha, \beta) : f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta};\ 0 \le x < \infty;\ \alpha, \beta > 0$
- $Beta(\alpha, \beta) : f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1};\ 0 \le x \le 1;\ \alpha, \beta > 0$
- $Chi - Square(n) : f(x) = \frac{1}{\Gamma(n/2)} (\frac{1}{2})^{n/2} x^{n/2-1} e^{-x/2};\ 0 < x < \infty$
- $Poisson(\lambda) : p(X = x) = \frac{e^{-\lambda}\lambda^x}{x!};\ x = 0, 1, ...;\ 0 \le \lambda < \infty$
- $Binomial(n, p) : p(X = x) = \binom{n}{x} p^x (1-p)^{n-x};\ x = 0, 1, ..., n;\ 0 \le p \le 1$
- $Negative\ Binomial(n, p) : p(X = x) = \binom{n+x-1}{x} p^n (1-p)^x;\ x = 0, 1, ..., n;\ 0 \le p \le 1$
- $Multinomial(n, \pi_1, ..., \pi_k)$: $p(X = x) = \binom{n}{x_1, ..., x_k} \pi_1^{x_1} \cdots \pi_k^{x_k}$ where $\sum_i \pi_i = 1;\ x = 0, 1, ..., n$
- $(Multivariate)\ Normal_n(\mu, \Sigma) : f(\mathbf{x}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} e^{\frac{-1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$; If we partition X into X_1 and X_2 such that $X = \binom{X_1}{X_2}$ where X_1 has dimension k and X_2 has dimension n-k
  - Marginal: $X_1 \sim N_k(\mu_1,\ \Sigma_{11})$
  - Conditional: $X_1 | X_2 \sim N_k(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
  - Linear Transformation: Let A_{rxn} and b_{rx1}. Then $AX + b \sim N_r(A\mu + b, A\Sigma A')$
  - Any linear combination, linear transformation, marginal, or conditional of a MVN RV is also MVN
  - The density of the MVN distribution only exists when Σ is positive definite (and thus non-singular) but the MGF of MVN always exists even if Σ is singular
- Location-Scale Families: If X is in the location-scale family and $Y = aX + b$, then $f_Y(y) = \frac{1}{a} f_X(\frac{x-b}{a})$

Distributional Relationships

- If $X_1, ..., X_n$ i.i.d. $Bernoulli(p)$ RVs, then $\sum_{i=1}^n X_i \sim Binomial(n, p)$
- If $X_1, ..., X_n$ i.i.d. $Exponential(\lambda) \equiv Gamma(1, \lambda)$ RVs, then $\sum_{i=1}^n X_i \sim Gamma(n, \lambda)$ and $Z_i = \frac{X_1 + ... + X_i}{X_1 + ... + X_n} \sim Beta(i, n - i + 1)$
- If $X_i \sim Gamma(k_i, \lambda)\ for\ i = 1, .., n$ then $\sum_{i=1}^n X_i \sim Gamma\left(\sum_{i=1}^n k_i, \lambda\right)$
- If $X_1, ..., X_n$ i.i.d. $Normal(0, 1)$ RVs, then $\sum_{i=1}^n X_i^2 \sim \chi_n^2 \equiv Gamma(n/2, 2)$
- If $X_1, ..., X_n$ independent $Pois(\lambda_i)$ RVs, then $\sum_{i=1}^n X_i \sim Pois(\sum_{i=1}^n \lambda_i)$
- $NegBin(1, p) \equiv Geometric(p)$ and a sum of n i.i.d. $Geometric(p)$ RVs is $NegBin(n, p)$
- If $Y_1 \sim Pois(\lambda_1)$ and $Y_2 \sim Pois(\lambda_2)$ are independent, then $Y_1 | Y_1 + Y_2 = m \sim Bin(m, \frac{\lambda_1}{\lambda_1 + \lambda_2})$
- Asymptotically speaking...
  - $Bin(n, p) \to_{n\to\infty} Pois(\lambda)$ where $\lambda = np$
  - $NegBin(n, p) \to_{n\to\infty} Pois(\lambda)$ where $\lambda = n(1-p)$
  - $t(\nu) \to_{\nu\to\infty} N(0, 1)$
  - $Gamma(r, \lambda)$ (as $r \to \infty$), $Beta(\alpha, \beta)$ (as $\alpha = \beta \to \infty$), $Pois(\lambda)$ (as $\lambda \to \infty$), and $Bin(n, p)$ (as $n \to \infty$) all go to $N(\mu, \sigma^2)$
- Let $X \sim N(0, 1),\ Y \sim \chi_m^2,\ Z \sim \chi_n^2$, all independent. Then

  - Student's t: $\frac{X}{\sqrt{Y/m}} \sim t_m$
  - F: $\frac{Y/m}{Z/n} \sim F_{m,n}$
  - Beta: $\frac{Y}{Y+Z} \sim Beta(m/2, n/2)$
- Now let Y be as above and $X \sim N(\mu, 1)$, still independent. Then $\frac{X}{\sqrt{Y/m}} \sim t_{m,\mu}$ where μ is the non-centrality parameter of the t distribution
- Now let Y and Z be independent non-central chi-square distributions such that $Y \sim \chi^2_{m,\gamma_Y}$ and $Z \sim \chi^2_{n,\gamma_Z}$. Then $\frac{Y/m}{Z/n} \sim F_{m,n,\gamma_Y,\gamma_Z}$ is the doubly non-central F distribution where γ_Y, γ_Z are non-centrality parameters for the F distribution
  - In linear models, under the null hypothesis the test statistic will have central F distribution and under the alternative hypothesis it will have singly non-central F distribution ($\gamma_Z = 0$). Only time doubly non-central F arises is when both null and alternative are not true
- If $X \sim t_q$ then $X^2 \sim F_{1,q}$
- Quadratic Forms
  - $Y_{nx1} \sim N_n(0, \Sigma) \Rightarrow Y'\Sigma^{-1}Y \sim \chi^2_n$
  - $Y_{nx1} \sim N_n(\mu, \sigma^2 I) \Rightarrow Y'MY \sim \chi^2_r(\delta)$, where $\delta = \frac{\mu' M \mu}{2\sigma^2}$ is the non-centrality parameter and M is o.p.o. rank r
  - If $Y \sim N_n(\mu, \sigma^2 I)$ then $\frac{1}{\sigma^2}(Y'MY) \sim \chi^2(r, \gamma) \iff M$ is an o.p.o. of rank r, where $\gamma = \frac{\mu'\mu}{2\sigma^2}$
  - If A is symmetric of rank $k \leq n$ and $Y \sim N_n(\mu, \Sigma)$ with Σ full rank, then $Y'AY = \sum_{i=1}^k \lambda_i W_i$ where $W_i \sim \chi^2(1, \mu'\Sigma^{-1}\mu)$ and the λ_i are the eigenvalues of $\Sigma^{1/2}A\Sigma^{1/2}$
  - If A is symmetric of rank $k \leq n$ and $Y \sim N_n(\mu, \Sigma)$ with Σ full rank and either 1. $A\Sigma$ or 2. ΣA is a projection operator of rank r or 3. Σ is a g-inverse of A, then $(Y'AY) \sim \chi^2(r, \gamma)$ where $\gamma = \frac{\mu' A \mu}{2}$
  - Note that the non-centrality parameter is always found by replacing Y with $E[Y]$ in the quadratic form and dividing by 2
  - If $Y \sim N_n(\mu, \Sigma)$ with Σ positive definite and A, B positive semi-definite, then $Y'AY \perp Y'BY \iff A\Sigma B = 0$
- Bayes' Rule: $f_{X|Y}(X|Y) = \frac{f_{X,Y}(X,Y)}{f_Y(Y)} = \frac{f_{Y|X}(Y|X)f_X(X)}{f_Y(Y)}$
- Independence: Two random vectors are independent if their joint density $f(x, y)$ factors into $f(x, y) = f_1(x)f_2(y)$ where $f_1(x)$ is the marginal density of X and $f_2(y)$ is the marginal density of Y, and if their supports are not tangled.
  - If X and Y are independent random vectors then $G(X)$ and $H(Y)$ are independent for arbutrary functions $G(.)$ and $H(.)$
  - If X and Y are MVN, then only need to show $Cov(X, Y) = 0$ to prove independence

Exponential Family

- Definition (uni): If the density of a random variable X can be put into the form $p(x|\xi) = exp\left\{\phi[x\theta - b(\theta) - c(x)] - \frac{1}{2}s(x, \phi)\right\}$ where $\xi = (\theta, \phi)$ and $c(.), b(.), s(.,.)$ are known functions, then X belongs to the exponential family. Here θ is the natural parameter and ϕ is the dispersion parameter.
- Definition (multi): Let $Y = (X_1, ..., X_n)$ with $X_1, ..., X_n$ i.i.d.. Then Y is in the multivariate exponential family if its density can be written in the form $p(y|\xi) = exp\{T(y)Q(\xi) - b(\xi) - c(y)\}$, where ξ is a k-dim parameter vector.
- Properties:
  - θ is some expression of the given parameters of the distribution (called the natural/canonical parameter)
  - MGF: $M_X(t) = exp\{\phi[b(\theta + t/\phi) - b(\theta)]\}$ and CGF: $K_X(t) = \phi[b(\theta + t/\phi) - b(\theta)]$
    - $b(\theta)$ is a convex function in θ, and all derivatives (in the interior of the parameter space Θ) exist, thus all moments of X exist
  - $\sum_{i=1}^n E[X_i] = n\delta_\theta b(\theta)$ and $\sum_{i=1}^n Var(X_i) = n\phi^{-1}\delta^2_\theta b(\theta)$

  - If all components of $Q(\xi)$ and $T(y)$ are distinct then the exponential family is said to be full rank
  - Any k-parameter exponential family is full rank if at every value of the ξ parameter vector, the kxk covariance matrix $\left(\left(\frac{\delta^2}{\delta\xi_i\delta\xi_j}b(\xi)\right)\right)$ is nonsingular
  - $T(y)$ are CSS for the natural parameter vector $Q(\xi)$ if the distribution is full rank; in univariate form, x is CSS for the natural parameter θ

- Examples: Normal, Chi-Square, Binomial (fixed n), Poisson, Exponential, Gamma, Beta, Bernoulli, Multinomial (fixed n), Negative Binomial (fixed r)

Expectation, Variance, Covariance Rules

- $E_X[X] = E_Y[E_{X|Y}[X|Y]]$
- $Var_X(X) = E_Y[Var_{X|Y}(X|Y)] + Var_Y(E_{Y|X}[Y|X])$
- $Cov_{X,Y}(X,Y) = E_Z[Cov_{X,Y|Z}(X,Y|Z)] + Cov_Z(E_{X|Z}[X|Z], E_{Y|Z}[Y|Z])$
- $Var_X(X) = E_X[(X - E[X])^2] = E_X[X^2] - (E_X[X])^2$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$
- $Var(XY) = E[(XY)^2] - (E[XY])^2 = (if\ indep)\ E[X^2]E[Y^2] - (E[X])^2(E[Y])^2$
- $Cov(X,Y) = E[XY] - E[X]E[Y]$
- $Cov(aX + b, cY + d) = acCov(X,Y)$
- $Cov(aX + bY, cW + dZ) = acCov(X,W) + adCov(X,Z) + bcCov(Y,W) + bdCov(Y,Z)$
- For some A_{rxn}, b_{nx1}, and Y_{nx1} is a random variable with mean μ_{nx1} and covariance matrix Σ_{nxn}, we have that $E[AX + b] = AE[X] + b = A\mu + b$ and $Cov(AX + b) = ACov(X)A' = A\Sigma A'$

Quality Inequalities

- Chebyshev: For a random variable X with finite mean μ and finite variance σ^2, $P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$
- Markov: $P(|X| > \epsilon) \leq \frac{E[g(|X|)]}{g(\epsilon)}$ (swap $|X - \mu|$ for X to get Chebyshev's, with $g(x) = x^2$)
- Holder: $\int |f(x)g(x)|dx \leq \left(\int |f(x)|^p dx\right)^{1/p} \left(\int |g(x)|^q dx\right)^{1/q}$ (Cauchy-Schwartz is special case where p=q=2)
- Jensen: $E[\phi(X)] \geq \phi(E[X])$ for any convex function ϕ
- Minkowski: For $r > 1$, $||X + Y||_r \leq ||X||_r + ||Y||_r$ where $||X||_r = (\sum |X|^r dx)^{1/r}$

Probability Stuff

- Law of Total Probability: If $\{B_n : n = 1, 2, ...\}$ is a partition of the whole space Ω (i.e. $\cup_n B_n = \Omega$), then the probability of an event A can be written $P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n)$
  - Popular convenient form: $P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$
- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Order Statistics

- Pdf of k^{th} order statistic: $f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}F_X(x)^{k-1}[1 - F_X(x)]^{n-k}f_X(x)$
- Pdf and Cdf of 1^{st} order statistic: $f_{X_{(1)}}(x) = n[1 - F_X(x)]^{n-1}f_X(x)$ and $F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n$
- Pdf and Cdf of last order statistic: $f_{X_{(n)}}(x) = nF_X(x)^{n-1}f_X(x)$ and $F_{X_{(n)}}(x) = F_X(x)^n$

Complete, Sufficient, and Ancillary Statistics

- Complete: A statistic $T(X)$ is complete for θ if $E[g(T(X))] = 0 \Rightarrow g(.) = 0$, i.e. that $g(.)$ is the zero function

- Sufficient: A statistic $T(X)$ is sufficient for θ if the conditional distribution $X|T(X) \perp \theta$. Also, $T(X)$ is sufficient $\iff$ the pdf $p_\theta(x)$ can be factorized into $g(T(X))h(x)$
- Ancillary: A statistic V is ancillary for $\theta \iff$ the distribution of V doesn't depend on θ ; note though that the expression of V may depend on θ

Calculus Factoids

- FTC 2: $F(x) = \int_{-\infty}^{x} f(t)dt$
- Lagrange Multipliers: Useful when we wish to optimize (maximize or minimize) a given function $f(x, y)$ subject to a given constraint $g(x, y) = k$. Solve the following system of equations for x and y:

$$\bigtriangledown f(x, y) = \lambda \bigtriangledown g(x, y) \quad and \quad g(x, y) = k$$

  Then plug in all solutions (x, y) you've found into $f(x, y)$ and identify which solution achieves the optimum.

- Taylor Series Expansion of a function $f(x)$ about a point a: $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$
  - Note that when we expand this out to the number of terms we are interested in, this can be written 2 ways depending on your end goal:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + o(x^2) \quad or \quad f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + O(x^3)$$

  - Generally, any time a proof involves an asymptotic result, a Taylor series expansion will be needed

- Expansions to know:
  - $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \lim_{n\to\infty}[1 + \frac{x}{n}]^n$
  - $sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$
  - $cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$
  - For $|x| < 1$, $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$ (and more generally $\frac{x^a - x^{b+1}}{1-x} = \sum_{n=a}^{b} x^n$)
  - Binomial Formula: $\sum_{x=0}^{n} \binom{n}{x} u^x v^{n-x} = (u+v)^n$
- Convex Function: A function $f(x)$ is convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$

Big O and Little o

- Definition: Let A_n and B_n be sequences of random variables. Then,
  - A_n is of order $\leq B_n$, i.e. $A_n = O_p(B_n)$ if A_n/B_n is bounded in probability $P\left(\left|\frac{A_n}{B_n}\right| > M(\epsilon)\right) \leq \epsilon$
  - A_n is of smaller order than B_n, i.e. $A_n = o_p(B_n)$ if $A_n/B_n \to_p 0$
- Properties:
  - $o_p(A_n) = A_n o_p(1)$, $O_p(A_n) = A_n O_p(1)$
  - $o_p(1) + o_p(1) = o_p(1)$
  - $O_p(1) + O_p(1) = O_p(1)$
  - $O_p(1) + o_p(1) = O_p(1)$
  - $O_p(1)o_p(1) = o_p(1)$
  - $o_p(O_p(1)) = o_p(1)$
  - $o_p(1) = O_p(1) \neq o_p(1)$
  - $[1 + o_p(1)]^{-1} = 1 + o_p(1) = O_p(1) + o_p(1) = O_p(1)$
  - $O_p(n^{-1}) \to o_p(1)$
- If we have an estimator/random variable X such that $\sqrt{n}(\bar{X} - E[X]) \to_d$ *some distribution*, then

$$\Rightarrow \sqrt{n}(\bar{X} - E[X]) = O_p(1) \Rightarrow \bar{X} - E[X] = O_p(n^{-1/2})$$

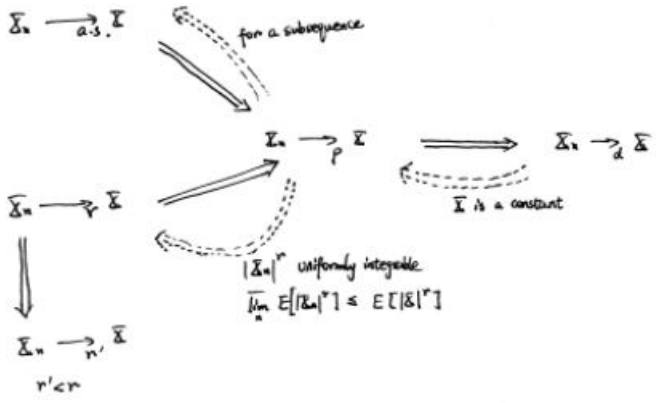
760

Distribution Theory

- Characteristic Function: $\phi_X(t) = E[e^{itX}]$
  - Uniquely determines the distribution; if characteristic functions of two RVs are equal then these two RVs are identically distributed
  - For i.i.d. $X_1, ..., X_n$, $\phi_{a_1X_1+...+a_nX_n} = \phi_{X1}(a_1t) \cdots \phi_{X_n}(a_nt)$
- Moment Generating Functions: $MGF = M_Y(t) = E[e^{tY}] = \int exp(ty)dF_Y(y)$; $M_Y(0) = 1$
- Cumulant Generating Functions: $CGF = log(MGF) = K_Y(t) = log(M_Y(t))$; $K_Y(0) = 0$
- Distributions of Sums, Products, and Quotients of Random Variables. Assume $X \perp Y$ and $Y > 0$:
  - $Z = X + Y \ \rightarrow \ f_Z(z) = \int f_X(z-y)f_Y(y)dy = \int f_Y(z-x)f_X(x)dx$
  - $Z = XY \ \rightarrow \ f_Z(z) = \int \frac{1}{y} f_X(\frac{z}{y})f_Y(y)dy$
  - $Z = X/Y \ \rightarrow \ f_Z(z) = \int y f_X(zy)f_Y(y)dy$

Large Sample Theory

- Modes of Convergence
  - Almost Surely: $X_n \rightarrow_{a.s.} X \iff P(sup_{m \geq n}|X_m - X| > \epsilon) \rightarrow 0$. In words, the sets on which $X_n \nrightarrow X$ have measure zero
  - Probability: $X \rightarrow_p X$ if for every $\epsilon > 0$, $P(|X_n - X| > \epsilon) \rightarrow 0$
  - r^{th} Moment: $X \rightarrow_r X$ if $E[|X_n - X|^r] \rightarrow 0$ as $n \rightarrow \infty$ where $\int |X|^r dP < \infty$ and $\int |X_n|^r dP < \infty$ (i.e. where the r^{th} moment of X and X_n are finite)
  - Distribution: $X_n \rightarrow_d X$ if the distribution functions F_n and F of X_n and X satisfy $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ for each continuity point x of F. (Number of discontinuity points can be at most countable)
  - Relationships among modes of convergence:



- Uniform Integrability: A sequence of RVs X_n is uniformly integrable if

$$\lim_{\lambda \rightarrow \infty} \limsup_{n \rightarrow \infty} E[|X_n| I(|X_n| \geq \lambda|)] = 0$$

  - Liapunov Condition is sufficient to show U.I.: If there exists a positive constant ϵ_0 such that $\limsup_n E[|X_n|^{r+\epsilon_0}] < \infty$, then $E[|X_n|^r I(|X_n|^r \geq \lambda)] \leq \frac{E[|X_n|^{r+\epsilon_0}}{\lambda^{\epsilon_0}}$
- Vitali's Theorem: Suppose that $||X_n||_r < \infty$ where $0 < r < \infty$ and $X_n \rightarrow_p X$. Then TFAE:
  A. $|X_n|^r$ are uniformly integrable
  B. $X_n \rightarrow_r X$
  C. $E[|X_n|^r] \rightarrow E[|X|^r]$
- Cramer-Wold Device: A random vector X_n in R^k satisfies $X_n \rightarrow_d X \iff t'X_n \rightarrow_d t'X$ in $R \ \forall \ t \in R^k$

- Continuous Mapping Theorem: Suppose $X_n \to_{a.s.} X$, or $X_n \to_p X$, or $X_n \to_d X$. Then for any continuous function $g(.)$, $g(X_n)$ converges to $g(X)$ almost surely, in probability, or in distribution respectively.
- Slutsky's Theorem: Suppose $X_n \to_d X$, $Y_n \to_p y$, and $Z_n \to_p z$ for some constants y and z. Then $Z_n X_n + Y_n \to zX + y$.
- Weak Law of Large Numbers: If $X, X_1, ..., X_n$ are iid with mean $E[|X|] = \mu < \infty$, then $\bar{X}_n \to_p \mu$.
- Strong Law of Large Numbers: If $X_1, ..., X_n$ are iid with mean $\mu < \infty$, then $\bar{X}_n \to_{a.s.} \mu$.
- Central Limit Theorems
  - Classic (Univariate): If $X_1, ..., X_n$ are iid with mean μ and variance σ^2, then $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma^2)$.
  - Classic (Multivariate): If $X_1, ..., X_n$ are iid random vectors in R^k with mean μ and covariance $\Sigma = E[(x - \mu)(X - \mu)']$ then $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \Sigma)$.
  - Weighted: If $X_1, ..., X_n$ are iid with mean μ and variance σ^2 and $max_i \frac{w_i^2}{\sum_{i=1}^n w_i^2} \to 0$, then $\frac{\sum_{i=1}^n (x_i w_i - \mu w_i)}{\sqrt{\sum_{i=1}^n w_i^2 \sigma^2}} \to_d N(0, 1)$.
  - Liapunov: Let $X_{1n}, ..., X_{nn}$ be independent random variables with $\mu_{ni} = E[X_{ni}]$ and $\sigma_{ni}^2 = Var(X_{ni})$. Let $\mu_n = \sum_{i=1}^n \mu_{ni}$ and $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. If $\sum_{i=1}^n \frac{E[|X_{ni} - \mu_{ni}|^3]}{\sigma_n^3} \to 0$, then $\sum_{i=1}^n \frac{(X_{ni} - \mu_{ni})}{\sigma_n} \to_d N(0, 1)$.
    - Note that for this to work, the 3rd moment must be finite
  - Lindeberg-Fell: Let $X_{1n}, ..., X_{nn}$ be independent random variables with $\mu_{ni} = E[X_{ni}]$ and $\sigma_{ni}^2 = Var(X_{ni})$. Let $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. Then both

$$\sum_{i=1}^n \frac{(X_{ni} - \mu_{ni})}{\sigma_n} \to_d N(0, 1) \text{ and } max(\sigma_{ni}^2/\sigma_n^2 \ : \ 1 \le i \le n) \to 0 \iff$$

    the Lindeberg condition $\frac{1}{\sigma_n^2} \sum_{i=1}^n E[|X_{ni} - \mu_{ni}|^2 I(|X_{ni} - \mu_{ni}| \ge \epsilon \sigma_n)] \to 0$ holds for all $\epsilon > 0$.
  - In summary:
    1. If see sum of i.i.d. RVs, of course do classic CLT
    2. If $X_1, ..., X_n$ i.i.d. but $w_1 X_1, ..., w_n X_n$ not i.d., do weighted CLT
    3. If condition for weighted CLT doesn't hold, try using Liapunov CLT
    4. If Liapunov condition doesn't hold, i.e. 3^{rd} moment not finite, try using the Lindeberg-Fell CLT
- Delta Method: For random vectors X_n and X in R^k, if there exists two constants a_n and μ such that $a_n(X_n - \mu) \to_d X$ and $a_n \to \infty$, then for any function $g : R^k \mapsto R^l$ such that g has a derivative at μ (denoted $\bigtriangledown g(\mu)$), we have that $a_n(g(X_n) - g(\mu)) \to_d \bigtriangledown g(\mu) X$.

Point Estimation and Efficiency

- Uniformly Minimum Variance Unbiased Estimators (UMVUEs)
  - Method 1 for finding UMVUE of θ:
    1. Find a complete and sufficient statistic $T(X)$ for θ (and should justify how we know it's CSS)
    2. Find a function $g(T(X))$ such that $E[g(T(X))] = \theta$, then $g(T(x))$ is the UMVUE for θ
  - Method 2 for finding UMVUE of θ:
    1. Find a complete and sufficient statistic $T(X)$ for θ (and should justify how we know it's CSS)
    2. Find an unbiased estimator for θ, denoted $\tilde{T}(X)$
    3. Calculate $E[\tilde{T}(X)|T(X)]$ to yield the UMVUE for θ
  - Rao-Blackwell Theorem: Suppose $\hat{\theta}(X)$ is an unbiased estimator for θ. If $T(X)$ is a sufficient statistic of X, then $E[\hat{\theta}(X)|T(X)]$ is unbiased and moreover, $Var(E[\hat{\theta}(X)|T(X)]) \le Var(\hat{\theta}(X))$ with the equality $\iff$ $E[\hat{\theta}(X)|T(X)] = \hat{\theta}(X)$ with probability 1.
  - Lehmann-Sheffe Theorem: For finding the UMVUE of a function $g(\theta)$ of your parameter, first get a CSS $T(X)$ for θ, then find an unbiased estimator for θ and use it to create an unbiased estimator for $g(\theta)$, then computing $E[\hat{g(\theta)}|T(X)]$ gives the unique UMVUE for $g(\theta)$

  - Basu's Theorem: If T is a complete and sufficient statistic for a family of models dependent on parameter θ, then for any ancillary statistic V of θ, $V \perp T$
  - If working with a linear model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$, then the unique UMVUE of $\lambda'\beta$ (for estimable $\lambda'\beta$) is the same as the unique LSE and the unique BLUE, which is $\rho' MY$ where $M = X(X^T X)^{-1} X^T$

- Estimating Equations: $\sum_{i=1}^n f(X_i; \theta) = 0$ is a set of estimating equations for the parameter θ, where $\frac{1}{n}\sum_{i=1}^n f(X_i;\theta) \to_{a.s.} E_X[f(X;\theta)]$. Let θ_0 be the true value of θ and let $\hat{\theta}$ be the estimator for θ that satisfies the estimating equations. Then we know

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Sigma) \; where \; \Sigma = E[-\dot{f}_{\theta_0}(X)]^{-1} Cov(f_{\theta_0}(X)) E[-\dot{f}_{\theta_0}(X)]^{-1}$$

- The Asymptotic Relative Efficiency of some estimators of θ, $\hat{\theta}$ with respect to $\tilde{\theta}$ is the ratio of their asymptotic variances, i.e. $ARE(\hat{\theta}, \tilde{\theta}) = \frac{aVar(\tilde{\theta})}{aVar(\hat{\theta})}$. If $ARE(\hat{\theta}, \tilde{\theta}) < 1$, then $\hat{\theta}$ is the more efficient estimator
  - When comparing the asymptotic variances of 2 random variables (estimators) with different asymptotic distributions (or any 2 random variables with different convergence rates), then the estimator that converges at the higher rate (larger power of n) will always be more efficient. E.g. $\tilde{\theta_2}$ will always be more efficient no matter the values of μ and σ^2 in the following example: $\sqrt{n}(\tilde{\theta_1} - \theta) \to_d N(0, \sigma^2)$ and $n(\tilde{\theta_2} - \theta) \to_d Exp(\mu)$

- Cramer-Rao Lower Bound (CRLB) (i.e. Information Bound)
  - If the likelihood function of a random variable X is differentiable w.r.t. θ and $T(X)$ is an estimator for some function $q(\theta)$, then $Var_\theta(T(X)) \geq \frac{(\dot{q}(\theta) + \dot{b}(\theta))^2}{I(\theta)}$ where $b(\theta)$ is the bias of the estimator $T(X)$, i.e. $b(\theta) = E[T(X)] - q(\theta)$ and $I(\theta)$ is the Fisher Information. Additionally, if the likelihood function of X is twice differentiable w.r.t. θ then $I(\theta) = -E_\theta[\ddot{l}_\theta(X)]$
  - Smallest possible lower bound for the variance of any estimator of $q(\theta)$. Some UMVUEs may attain the CRLB, but not all.
  - When calculating the Fisher Information, first differentiate the log likelihood, then take the negative expectation, and last we plug in the MLE if we are looking for the Fisher information of the MLE. Don't plug in the MLE if we're just looking for the general Fisher information.

- Locally Regular Estimator ?

Maximum Likelihood Estimation

- MLEs (here $\hat{\theta}$) have 2 important properties, provided that certain regularity conditions are met (automatic with exponential family) – Consistency and Asymptotic Efficiency

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

  - Consistency: $\hat{\theta} \to \theta_0$ (asymptotically unbiased)
  - Asymptotic Efficiency: Asymptotic variance achieves CRLB $I^{-1}(\theta_0)$
  - It is possible to have an MLE that doesn't have these properties, e.g. the estimator $\hat{\alpha}$ of α from $X \sim Unif(0, \alpha)$

- Newton-Raphson Algorithm is an iterative algorithm used to compute the MLE when it doesn't have a closed form:

$$\theta^{k+1} = \theta^k + I^{-1}(\theta^k)\left(\frac{1}{n}\sum_{i=1}^n \dot{l}_{\theta^k}(X_i)\right), \; where \; I(\theta^k) = \frac{-1}{n}\sum_{i=1}^n \ddot{l}_{\theta^k}(X_i)$$

- E-M Algorithm
  - Used to compute the MLE(s) when there is missing data present
  - Types of missing data:
    - MCAR: $P(R_i|Y_i^O, Y_i^M, X_i) = P(R_i|X_i)$; Probability that responses are missing is unrelated to the values of both the observed and missing responses
    - MAR: $P(R_i|Y_i^O, Y_i^M, X_i) = P(R_i|Y_i^O, X_i)$; Probability that responses are missing depends on the set of observed responses but is conditionally unrelated to the values of the missing responses

    - MNAR: $P(R_i|Y_i^O, Y_i^M, X_i)$; Probability that responses are missing depends on the values of both the observed and missing responses
  - Framework: $Y = (Y^M, Y^O)$; R is a vector of 0/1 indicating which responses are missing/not missing, so $Y^O = RY$; the density function for the observed data (Y^O, R) is $\int_{Y^M} f(Y;\theta)P(R|Y)dY^M$
  - For both EM methods below, we use the log likelihood of the complete data (X_i, R_i, Y_i) and then take the expectation with respect to *complete|observed*
  - Method 1 – Usual Way
    - E-Step: Evaluate the conditional expectation

$$E\left[log(f(Y;\theta))|Y^O, \theta^{(k)}\right]$$

    - M-Step: Maximize by taking the derivative of the equation above and setting equal to 0, and solving for $\theta^{(k+1)}$
    - For this way, often easier to take the derivative of the log likelihood first (M-step) and then take the expectation (E-step)
  - Method 2 – For when the maximization step is too complicated
    - E-Step: Evaluate the conditional expectations

$$E\left[\frac{\delta}{\delta\theta}log(f(Y;\theta))|Y^O, \theta^{(k)}\right] \quad and \quad E\left[\frac{\delta^2}{\delta\theta^2}log(f(Y;\theta))|Y^O, \theta^{(k)}\right]$$

    - M-Step: Maximize by setting the first equation above to 0 and solving for $\theta^{(k+1)}$ using the one-step Newton-Raphson algorithm

$$\theta^{(k+1)} = \theta^{(k)} - \left(E\left[\frac{\delta^2}{\delta\theta^2}log(f(Y;\theta))|Y^O, \theta^{(k)}\right]\right)^{-1}\left(E\left[\frac{\delta}{\delta\theta}log(f(Y;\theta))|Y^O, \theta^{(k)}\right]\right)\Big|_{\theta=\theta^{(k)}}$$

  - Iterate between E- and M-step until the convergence of θ, i.e. until the difference between $\theta^{(k+1)}$ and $\theta^{(k)}$ is less than a given criteria
  - If we don't know what type of missing data we have, can't use the EM algorithm
  - The following theorem explains why EM algorithm works: At each iteration of the EM algorithm, $log(f(Y^O;\theta^{(k+1)})) \geq log(f(Y^O;\theta^{(k)}))$ is true and the equality holds $\iff \theta^{(k+1)} = \theta^{(k)}$
  - To get the asymptotic distribution of the MLEs, use the joint log likelihood of the observed data to get the CRLB
  - MLE theory doesn't apply to distributions whose support depends on an unknown paramter (e.g. Unif(0,α), where α unknown)

762 - Linear Models

General Definitions/Theorems

- Linearly (In)Dependent: A set of vectors $D = \{x_1, ..., x_n\}$ is linearly dependent if there is a set of scalars $a_1, ..., a_n$, not all zero, such that $\sum_{i=1}^n a_i x_i = 0$. If $\sum_{i=1}^n a_i x_i = 0 \Rightarrow a_i = 0 \ \forall \ i = 1, ..., n$ then $D = \{x_1, ..., x_n\}$ are linearly independent.
  - Theorem: A set of vectors is linearly dependent $\iff$ some vector of the set can be written as a linear combination of the others.
  - Two orthogonal vectors are necessarily linearly independent
- Orthogonal Vectors: Two vectors x and y are orthogonal, $x \perp y$ if $x'y = 0$
- Any basis (set of vectors) can be written as an orthonormal basis using the Gram-Schmidt process
- Orthogonal Compliment: Let N be a subspace of a vector space $M \subset R^n$. The orthogonal compliment $N^\perp$ is $N^\perp = \{y \in M : y \perp N\}$, the set of all vectors perpendicular to the subspace N.

- Column Space: The space spanned by the columns of a matrix A_{nxp} is called the column space of A, $C(A) = \{z : Ax = z,\ x \in R^p\}$.
  - Post-multiplying a matrix A by another matrix B of greater or equal rank will not change the column space, i.e. $C(A) = C(AB)$; there is no such rule for pre-multiplying matrices
- Null Space: For a matrix A, $N(A) = \{x : Ax = 0\}$
  - Theorem: If A_{nxn} is symmetric and of rank $r \leq n$, then $N(A) = C(A)^{\perp}$, $N(A) \cap C(A) = 0$, $N(A) + C(A) = R^n$, and $r(A) = r,\ r(N(A)) = n - r$
  - If A is any nxp matrix, then $N(A) = C(A')^{\perp}$
- Rank: The rank of a matrix A_{nxp}, $r(A)$, is the number of linearly independent columns in the matrix (or the number of nonzero eigenvalues).
  - $r(A) = r(A')$, $r(A) \leq min(n, p)$
- Trace: The trace of a matrix A is $tr(A) = \sum_{i=1}^{n} a_{ii}$, i.e. the sum of all diagonal elements of the matrix.
  - Trace is invariant under cyclic permutations
  - For square A and nonsingular B, $tr(A) = tr(BAB^{-1}) = tr(B^{-1}AB)$
- Singularity: A square matrix is non-singular $\iff$ it is a full rank matrix $\iff$ all columns are linearly independent $\iff$ all eigenvalues of the matrix are nonzero $\iff$ it's invertible
- Eigenvalues and Eigenvectors: If A is a square matrix (nxn), then an eigenvector of A is any nonzero vector x satisfying $Ax = \lambda x$, and λ is the corresponding eigenvalue of A.
  - Eigenvalues of matrix A are found by finding the zeroes of the equation $det(A - \lambda I) = 0$.
  - $det(A) = \Pi_{i=1}^{n} \lambda_i$
  - $tr(A) = \sum_{i=1}^{n} \lambda_i$ and $tr(A^{-1}) = \sum_{i=1}^{n} \lambda_i^{-1}$
  - Eigenvalues of A' are the same as those of A, so $tr(A) = tr(A')$
  - If A also symmetric, then $tr(A^r) = \sum_{i=1}^{n} \lambda_i^r$ for any integer r
- Symmetry: A matrix A is symmetric if $A = A'$
  - Any symmetric matrix A can be written $A = A^{1/2}A^{1/2}$
  - A product of symmetric matrices ABC is not symmetric unless $ABC = (ABC)' = C'B'A'$
- Orthogonal Matrix: A square matrix A is orthogonal if $A' = A^{-1}$, i.e. if $AA' = A'A = I$
  - If A is orthogonal then so is A'
  - The product of two orthogonal matrices is orthogonal
- Positive (Semi)Definite: A symmetric nxn matrix A is positive (semi)definite if $x'Ax > (\geq)\ 0\ \forall x \neq 0, x \in R^n$
  - The eigenvalues of a positive (semi)definite matrix are all positive (nonnegative)
  - All positive definite matrices are invertible
  - A is positive semidefinite $\iff$ it has a Cholesky decomposition, i.e. $A = QQ'$ for Q nonsingular and $Q = A^{1/2} = P\Lambda^{1/2}P'$ (See Spectral Decomposition below)
  - A is positive semidefinite $\iff$ 1. All diagonal elements $a_{ii} \geq 0$ and 2. The determinant of every submatrix is ≥ 0
  - All covariance matrices are at least positive semidefinite
  - For any nxp matrix A, $A'A$ and AA' are positive semi-definite
- Spectral Decomposition: Allows representation of any symmetric matrix in terms of an orthogonal matrix and a diagonal matrix of eigenvalues. If A_{nxn} is a symmetric matrix then there exists an orthogonal matrix P such that $A = Pdiag\{\lambda_1, ..., \lambda_n\}P'$ where $\lambda_1, ..., \lambda_n$ are the eigenvalues of A and P is the orthogonal matrix of eigenvectors.

- Singular Value Decomposition: Let A_{nxp} be a matrix of rank $r \leq min(n,p)$. Then there exists orthogonal matrices U_{pxp} and V_{nxn} such that $V'AU = D = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}$ where $\Delta = diag\{\delta_1, ..., \delta_r\}$ is the diagonal matrix of ordered singular values of A and $\delta_i^2 = \lambda_i$ where λ_i are the eigenvalues of $A'A$. Then we have too that $A = VDU'$.
- Q-R Factorization: Let A_{nxp} have linearly independent columns (be full rank p). Then A can be written uniquely as $A = QR$ where Q_{nxp} has orthonormal columns and R_{pxp} is an upper triangular matrix with positive diagonal elements.
  - Most useful in computational problems where n and p are large
  - Q and R are constructed using the Gram-Schmidt process
- A square matrix A is said to be a projection operator (matrix) onto $C(A)$ along $N(A)$ if for any $v \in C(A)$, $Av = v$. Projection operators are not unique unless they are orthogonal. All eigenvalues of A are either 0 or 1
- Idempotent: If $A^2 = A$, then A is an idempotent matrix
  - A is idempotent $\iff$ it is a projection matrix
- Orthogonal Projection Operators: M is an orthogonal projection operator onto $C(M) \iff M$ is idempotent and symmetric
  - If M is an o.p.o. onto $C(X)$ then $C(M) = C(X)$ (i.e. their column spaces are equivalent)
  - O.p.o.'s are unique
  - Suppose M_{nxn} is an o.p.o. of rank $r \leq n$. Then 1. The eigenvalues of M are 0 or 1; 2. $r(M) = tr(M) = r$; 3. M is positive semidefinite
  - M is the unique o.p.o. onto $C(X)$ and $I - M$ is the unique o.p.o. into $C(X)^\perp$
  - Let X_{nxp} have rank r, and $M = X(X'X)^-X'$. Then M is the unique o.p.o. onto $C(X)$
    - Note that if $r(X) = r < p$, then M can be equivalently written (for ease of computation) as $M = X^*(X^{*\prime}X^*)^-X^{*\prime}$ where X^* is the nxr matrix made up of only the linearly independent columns of X
- Generalized Inverses: If a matrix X_{nxp} has rank $r < min(n,p)$ then $(X'X)^{-1}$ doesn't exist, but it has a generalized inverse. If A_{nxp}, then a generalized inverse of A is A^-_{pxn} such that $AA^-y = y \ \forall y \in C(A)$; equivalently, A^- is a generalized inverse of A if $AA^-A = A$
  - Generalized inverses are not unique
  - A^-A is idempotent, but not symmetric in general
  - Every matrix A has a Moore-Penrose generalized inverse A^+, and it is unique and AA^+ and A^+A are o.p.o.'s
    - Construct MPG-inverse using the SVD of A and inverting the SVD of A to get A^+
  - Note that only square matrices can have inverses, but all matrices have a generalized inverse
- If $Y \sim N_n(\mu, \Sigma)$ then $E[Y'AY] = \mu'A\mu + tr(A\Sigma)$ for all A_{nxn}
- $||MY||^2 = Y'MY$ where M is an o.p.o.
- Nifty trick: If have a matrix A that is dependent on a (univariate) parameter θ (e.g. a covariance matrix), then $\frac{d}{d\theta} log(|A|) = tr(A^{-1}\frac{d}{d\theta}A)$, assuming A invertible (usually A here is a covariance matrix)

Estimability

- If have more parameters than equations, then none of your parameters are individually estimable/identifiable, only functions involving multiple parameters. If have fewer or equal number of parameters as you have equations, each parameter itself will be estimable/identifiable
- Golden Rule of Estimability: If our model is $E[Y] = \eta = X\beta$, the rule is that η is estimable, i.e. any linear combination of elements of η is estimable
- While the estimate of element(s) of β will depend on $Cov(Y)$, the conditions for estimability don't depend on $Cov(Y)$

- Can also think of as: For some vector of constants λ, $\lambda'\beta$ is estimable $\iff$ there exists a vector ρ such that $\rho'X = \lambda'$
- If a random variable $X \sim N_n(\mu, \Sigma)$ then we know all the means, variances, and covariances are identifiable
- More generally, to check identifiability we set the pdf of our random variable $p(\theta)$ equal to another pdf of the same type but with different parameters $p(\theta^*)$; then we either show the parameters can differ or that each parameter must be equal to its analog

Common Estimation Methods

- Least Squares Estimation (LSE): The LSE of β, $\hat{\beta}$ minimizes the squared Euclidean distance between Y and its mean $\mu = X\beta$.
  - $\hat{\beta}$ is a LSE for $\beta \iff X\hat{\beta} = MY$, where $\hat{\beta}$ is not necessarily unique. If $\lambda'\beta$ is estimable then the LSE $\lambda'\hat{\beta}$ is unique.
  - The unique LSE of $E[Y] = \mu = X\beta$ is MY and an unbiased estimate of σ^2 is $\frac{||(I-M)Y||^2}{n-r(X)} = \frac{Y'(I-M)Y}{n-r(X)}$, otherwise known as the Mean Squared Error (MSE)
- Best Linear Unbiased Estimator (BLUE): If we have the linear model $Y = X\beta + \epsilon$ where $E[\epsilon] = 0$ and $Cov(\epsilon) = \sigma^2 I$ and if $\lambda'\beta$ is estimable, then the unique LSE of $\lambda'\beta$, $\rho'MY$, is also the unique BLUE (Gauss-Markov Theorem)
  - "Best" here means minimum variance
- Weighted Least Squares (WLS): Consider the linear model $Y = X\beta + \epsilon$ where $E[\epsilon] = 0$ and $Cov(\epsilon) = \sigma^2 V$, with V a known positive definite matrix. Since $V = QQ'$, can transform this model by pre-multiplying it by Q^{-1} which leads us to the usual linear model. Then all the same properties apply to this model for computing the LSEs, so can get them and then back-transform to the WLS model to get the WLSEs of β and σ^2
- Note that there have been no distributional assumptions made thus far
- Maximum Likelihood Estimation: If we have the usual linear model $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I)$ then we have that $Y \sim N_n(X\beta, \sigma^2 I)$. The MLEs for estimable functions of β, σ^2 are obtained by maximizing the (log) likelihood function of Y.
  - Can only be used if the likelihood function is differentiable and if support doesn't depend on any unknowns
  - $\hat{\beta} = (X'X)^- X'Y$ ($- = -1$ if X full rank) and $\hat{\sigma^2} = \frac{Y'(I-M)Y}{n}$
  - Note that $\hat{\sigma^2}$ is biased, but asymptotically unbiased; $\frac{n}{r(I-M)}\hat{\sigma}^2$ is unbiased
  - If X is full rank p, then $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$
  - The ordinary residual can be written $\hat{\epsilon}_i = y_i - x_i^T\hat{\beta}$
  - Newton-Raphson algorithm (written in terms of β): $\beta^{(k+1)} = \beta^{(k)} + \{-\delta_\beta^2 l_n(\beta^{(k)})\}^{-1}\delta_\beta l_n(\beta^{(k)})$
    - Remember to always take the derivatives first and plug in $\beta^{(k)}$ last

Hypothesis Testing

- We are interested only in testing nested linear models; if models we're testing weren't nested then the F-test wouldn't make sense
- Remember that when testing hypotheses, the expression of β's being tested (i.e. in H_0) must be estimable
- Consider the usual linear model $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I)$ and the reduced model $Y = X_0\gamma_0 + \epsilon$ where $C(X_0) \in C(X)$ We wish to test the hypothesis $H_0 : E(Y) \in C(X_0)$ vs $H_1 : E(Y) \in C(X) \cap C(X_0)^c$. Let $M = X(X'X)^-X'$ be the o.p.o. onto $C(X)$ and $M_0 = X_0(X_0'X_0)^-X_0'$ be the o.p.o. onto $C(X_0)$ and let $r_0 = r(X_0) < r(X) = r$. Then, under H_1:

$$F = \frac{||(M-M_0)Y||^2/(r-r_0)}{||(I-M)Y||^2/(n-r)} \sim F(r-r_0, n-r, \gamma), \text{ where } \gamma = \frac{||(M-M_0)X\beta||^2}{2\sigma^2} = \frac{||(I-M)X\beta||^2}{2\sigma^2}$$

And under H_0:

$$F = \frac{||(M-M_0)Y||^2/(r-r_0)}{||(I-M)Y||^2/(n-r)} \sim F(r-r_0, n-r)$$

So, a level α test of this hypothesis rejects H_0 if $F > F(1-\alpha, r-r_0, n-r)$

- General Procedure: Suppose we wish to test $H_0 : \Lambda'\beta = d$ where $\Lambda' = P'_{sxn}X_{nxp}$ is known and $d_{sx1} = P'_{sxn}X_{nxp}b_{px1}$ is a known vector of constants (could be all 0's). Then the F-test can be written as

$$F = \frac{(\Lambda'\beta - d)'(\Lambda'(X'X)^-\Lambda)^-(\Lambda'\beta - d)/r(\Lambda)}{MSE} \sim F(r(\Lambda), r(I-M), \gamma)$$

  where

$$MSE = \frac{Y'(I-M)Y}{r(I-M)} \text{ and } \gamma = \frac{(\Lambda'\beta - d)'(\Lambda'(X'X)^-\Lambda)^-(\Lambda'\beta - d)}{2\sigma^2}$$

  So, a level α test of this hypothesis rejects H_0 if $F > F(1-\alpha, r(\Lambda), r(I-M))$

  - Note that the vector b is not unique
  - Remember that $\Lambda' = P'X$, so can substitute this into the expression above, and it can also be helpful to re-state the null hypothesis as $H_0 : P'X\beta = P'Xb \iff H_0 : P'(X\beta - Xb) = 0$

- Confidence Regions: A $100(1-\alpha)\%$ CR for the estimable vector $\Lambda'\beta$ is

$$\left\{\beta : \frac{(\Lambda'\hat{\beta} - \Lambda'\beta)'(\Lambda'(X'X)^-\Lambda)^-(\Lambda'\hat{\beta} - \Lambda'\beta)/r(\Lambda)}{MSE} \leq F_{1-\alpha, r(\Lambda), r(I-M)}\right\}$$

  which is the inverted acceptance region of the test above

- Likelihood Ratio Test: The F test for testing nested linear models is equivalent to the LRT. Let Θ denote the parameter space and $\Theta_0 \subset \Theta$. Let θ be a vector in Θ and let y denote the data. Then the LRT for testing $H_0 : \theta \in \Theta_0$ *vs.* $\theta \in \Theta_0^c$ is given by $\lambda(y) = \frac{\sup_{\Theta_0} L(\theta|y)}{\sup_{\Theta} L(\theta|y)}$ and it has a rejection region of the form $\{y : \lambda(y) \leq c\}$ where $0 \leq c \leq 1$

  - If any of the parameters in the vector θ are not being tested, still need to find their MLEs under H_0 and $H_0 \cup H_1$ to plug into $\lambda(y)$

ANOVA

- One way (unbalanced): We have the model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ where ϵ_{ij}'s are i.i.d. $N(0, \sigma^2)$ and $i = 1, ..., t$; $j = n_1, ..., n_t$.

  - Remember that $E[Y_{ij}] = \mu + \alpha_i$ is always estimable, and here so are any contrasts, i.e. functions of the form $\sum_{i=1}^t \lambda_i\alpha_i$ where $\sum_{i=1}^t \lambda_i = 0$
  - We have $\lambda = (0, \lambda_1, ..., \lambda_t)'$, and so if $\lambda'\beta$ is estimable then $\lambda' = \rho'X$ and the BLUE of $\lambda'\beta$ is $\rho'MY$ where ρ' is of the form $\rho' = \left(\frac{\lambda_1}{n_1}J'_{n_1}, ..., \frac{\lambda_t}{n_t}J'_{n_t}\right)$ where J_{n_i} is a column vector of n_i 1's
    - This choice for ρ' is not the only choice, but has many convenient properties for ease of calculations, most notably $\rho \in C(X)$ so $\rho'M = \rho'$ and the o.p.o. for a contrast $\rho_i'X\beta$ is $M_i = \frac{\rho_i\rho_i'}{\rho_i'\rho_i}$

- Two way (balanced, no interaction): We have the model $Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk}$ where ϵ_{ijk}'s are i.i.d. $N(0, \sigma^2)$ and $i = 1, ..., a$ (number levels of trt i); $j = 1, ..., b$ (number levels of trt j); and $k = 1, ..., N$ (number obs per trt combo) for a total of $n = abN$ observations.

  - We break up the estimation space into a sum of orthogonal subspaces such that $C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta)$ where $M_\mu = \frac{JJ'}{J'J} = \frac{J_n^n}{n}$ (same for all ANOVA), $M_\alpha = Z_\alpha(Z_\alpha'Z_\alpha)^-Z_\alpha'$, and $M_\eta = Z_\eta(Z_\eta'Z_\eta)^-Z_\eta'$. Here the Z_*'s represent the columns of the design matrix that correspond to the $*^{th}$ treatment

Source	DF	SS	MS
Mean	1	$Y'\left(\frac{J_n^n}{n}\right)Y$	$Y'\left(\frac{J_n^n}{n}\right)Y$
Treatments (α)	$a-1$	$Y'M_\alpha Y$	$\frac{Y'M_\alpha Y}{a-1}$
Treatments (η)	$b-1$	$Y'M_\eta Y$	$\frac{Y'M_\eta Y}{b-1}$
Error	$n-a-b+1$	$Y'(I-M)Y$	$\frac{Y'(I-M)Y}{n-a-b+1}$
Total	n	$Y'Y$	

  - Remember that $\lambda'\beta$ is estimable; $\lambda'\beta$ is a contrast in the α_i's $\iff M\rho = M_\alpha\rho$
  - Choose ρ similarly as before

762 - Generalized Linear Models

Tools/Definitions in Likelihood Theory

- Let $l_n = l_n(\xi)$ be the log likelihood function of a multivariate (n-dim) random variable Y (note that the following expectations are taken w.r.t. Y not ξ, and derivatives taken w.r.t. ξ) and $\delta_{\xi_j} = \delta/\delta\xi_j$ where $\xi = (\beta, \phi)$
  - Score Equation: $E_\xi[\delta_{\xi_j} l_n] = 0$ (basic equation for defining the MLE $\hat{\xi}$)
  - Fisher Info: $I_n(\xi) = -E_\xi(\delta^2_\xi l_n)$ (for n random variables). When components $Y_1, ..., Y_n$ are i.i.d. then $I(\xi) = \frac{1}{n} I_n(\xi)$
- Within the exponential family, the MLE of $E[Y] = \mu$ is always the sample mean
- If $X_1, ..., X_n$ i.i.d. with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$, then
  - $\frac{1}{n}\sum_{i=1}^n X_i \to_p \mu = O_p(1)$ (LLN)
  - $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \to_d N(0, \sigma^2) = O_p(1)$ (CLT)
  - $\frac{1}{n}\sum_{i=1}^n [f(X_i) - E[f(X_i)]] = o_p(1)$ (LLN; note here that the X_i don't need to be i.d.)
- Once you go outside the classic linear model no hypothesis test is exact, they all involve asymptotics
- General Hypothesis Testing Procedures for GLMs, testing the hypothesis $H_0 : h_0(\xi) = b_0$ vs. $H_1 : h_0(\xi) \neq b_0$, where $h_0(.)$ and b_0 are rx1 vectors
  - Wald Test: $W_n = [h_0(\hat{\xi}) - b_0]^T \{H(\hat{\xi}) E[-\delta^2_\xi l_n(\hat{\xi})] H(\hat{\xi})^T\}^{-1} [h_0(\hat{\xi}) - b_0]$ where $H(\xi) = \delta h_0(\xi)/\delta\xi$ is an rxq matrix
  - Rao's Score Test: $SC_n = \delta_\xi l_n(\tilde{\xi})^T \{-E[\delta^2_\xi l_n(\xi)]\}\Big|_{\xi=\tilde{\xi}} \delta_\xi l_n(\tilde{\xi})$ where "$\sim$" denotes the estimate of the parameter under H_0
  - Likelihood Ratio Test: $LRT_n = 2[l_n(\hat{\xi}) - l_n(\tilde{\xi})]$, which is essentially 2 times the difference between the likelihood function evaluated at the global MLE and at the MLE under H_0
  - In each of the above, note that we always take the derivative first, then expectations, then plug in the MLEs last
  - All 3 of the above testing procedures have the same asymptotic distribution. Namely, the test statistic for each is distributed χ^2 with degrees of freedom equal to the number of parameters being tested (Can be shown, of course, using Taylor expansions for each of the main components of each test statistic)
- Delta Method: If $g(.)$ is a differentiable function at ξ_* and $\sqrt{n}(\hat{\xi} - \xi_*) \to_d N(0,\ I^{-1}(\xi_*))$ then

$$\sqrt{n}(g(\hat{\xi}) - g(\xi_*)) \to_d N(0,\ \delta_\xi g(\xi_*) I^{-1}(\xi_*) \delta_\xi g(\xi_*)^T)$$

  - Allows us to find the asymptotic covariance matrix of any function $g(.)$ (even non-linear) of a random variable
  - Remember to differentiate the $g(.)$ function generally first, then plug in the values specific to the problem
- GLMs Definition: Let $y = (y_1, ..., y_n)^T$ where the components of y are mutually independent and the conditional density of $y_i|x_i$ is $D(\theta_i, \phi/\omega_i)$ which is a member of the exponential family, $\theta_i = \theta_i(x_i)$, and ω_i is a weight. GLMs are characterized by the distribution of the y vector and by the link function $g(\mu_i) = x_i^T\beta = \eta_i$ which "links" the mean μ_i to the covariates x_i
  - Categorical Link Functions:
    - Logit: $g(\pi) = log(\pi/(1-\pi))$
    - Probit: $g(\pi) = \Phi^{-1}(\pi)$
    - Log-Log: $g(\pi) = -log(-log(\pi))$
    - Complimentary Log-Log: $g(\pi) = log(-log(1-\pi))$
  - Continuous Link Functions:
    - Identity: $g(\mu) = \mu = \eta$
    - Canonical: $g(\mu) = -\mu^{-1} = \eta$

    - Log: $g(\mu) = log(\mu) = \eta$
  - Can determine the link function of a model by setting $\theta = x_i^T\beta$ where θ is the canonical parameter obtained from the exponential family form
  - When $g(.) = \dot{b}^{-1}(.)$, we call it the canonical link function ; the canonical link θ is whatever sits next to the CSS in exponential family form!

  Note that in classic linear models, we are always dealing with the situation where the y vector is assumed to have the Normal distribution and the identity link is used (yielding a linear relationship between the mean and the covariates). GLMs extend LMs to other distributions within the exponential family and other relationships between the mean and the covariates

- What can we do when given a distribution of a response y_i that belongs to the exponential family?
  1. Put density in exponential family form to see what sits next to y_i
  2. This is the canonical link $\theta_i = g(\mu_i) = \eta_i = x_i^T\beta$ (aka the μ-link)
  3. To get the θ-link, take $\theta_i = g(\mu_i)$ and solve for $\mu_i \to \mu_i = g^{-1}(x_i^T\beta)$ where $\mu_i = \delta_{\theta_i} b(\theta_i) \Rightarrow \theta_i = \dot{b}^{-1}(g^{-1}(x_i^T\beta))$
  4. Have pdf in exponential family form, so easy to get the likelihood function
  5. Take derivatives of the likelihood function w.r.t. β to get $\dot{l}(\beta)$ and $\ddot{l}(\beta)$
  6. Can us the N-R algorithm to get MLE of β
  7. With the above info, can compute $W_n,\ SC_n,\ LRT_n$ to test hypotheses of interest
- Deviance: The deviance for all n observations is defined as $Dv(y;\hat{\mu}) = \phi^{-1}LRT_n = \sum_{i=1}^n Dv_i$
  - Used to assess GOF of a model
  - Want deviance to be small to indicate good GOF because the closer in value the likelihood functions are under H_0 and H_1, the better the fit of the model being tested (because we assume the full model fits well)

GLMs for Continuous Data

- Gamma Regression: Response $y_i|x_i$ assumed to have a $Gamma(\mu_i, \nu)$ distribution. Mainly used for modeling continuous, positive responses having constant coefficient of variation, or count data where the counts are relatively large.
- Inverse Gaussian Regression: Response $y_i|x_i$ assumed to have a $IG(\mu_i, \lambda)$ distribution. Also used for modeling continuous, positive responses (e.g. time to first passage of a Brownian motion).

GLMs for Categorical Data

- Contingency Tables
  - Sampling Methods - X=exposure, Y=disease:
    - Cross-Sectional sampling - Grand total n is fixed and the Multinomial distribution is used to model the joint probability of X and Y
    - Prospective sampling (cohort studies)- Marginal of X is fixed, Binomial distribution used to model $P(Y|X)$
    - Retrospective sampling (case-control studies)- Marginal of Y is fixed, Binomial distribution used to model $P(X|Y)$
    - The odds and its estimate $\hat{R}$ and the standard error of the estimate $se(\hat{R})$ (formulas below) are invariant to sampling scheme
  - We denote the joint distribution of (X, Y) as $\pi_{i,j} = P(X = i, Y = j)$, the marginal distributions of X and Y respectively as $\pi_{i,.} = \sum_{j=1}^J \pi_{i,j}$ and $\pi_{.,j} = \sum_{i=1}^I \pi_{i,j}$, and the conditional distribution of Y given X is $\pi_{j|X=i} = \pi_{i,j}/\pi_{i,}$.
  - To test whether X and Y are associated with each other, we consider the null hypothesis of independence $H_0 : \pi_{i,j} = \pi_{i,.}\pi_{.,j}$
  - Relative Risk: Let X and Y both be binary variables taking values 1 or 2. The relative risk (or equivalently the odds) of $Y = 2$ over $Y = 1$ for $X = 2$ is $odds(P(Y = 2|X = 2)) = \frac{P(Y=2|X=2)}{1-P(Y=2|X=2)} = \frac{P(Y=2|X=2)}{P(Y=1|X=2)} = \frac{n_{22}/n_{2.}}{n_{21}/n_{2.}} = \frac{n_{22}}{n_{21}}$

  - Odds Ratio: The odds ratio of $X = 2$ and $X = 1$ is $R = \frac{odds(P(Y=2|X=2))}{odds(P(Y=2|X=1))} \approx \hat{R} = \frac{n_{22}n_{11}}{n_{21}n_{12}}$. Additionally, the standard error of the odds ratio estimate is $se(\hat{R}) = \frac{\hat{R}}{\sqrt{n}}\sqrt{\frac{1}{\hat{\pi}_{11}} + \frac{1}{\hat{\pi}_{12}} + \frac{1}{\hat{\pi}_{21}} + \frac{1}{\hat{\pi}_{22}}}$, where $\hat{\pi}_{i,j} = n_{ij}/n$ are the MLEs of the cell probabilities

- Logistic Regression: Assumes the response variable $y_i|x_i$ has a $Bin(n_i, p_i)$ distribution. Commonly used for modeling binary response data as a function of the covariates.
  - The parameters in this model can be interpreted as the log odds ratios comparing the two categories of their corresponding covariate; exponentiating the parameters yields the odds ratios.
  - Model-predicted probabilities of response for a particular combination of covariates can be obtained by taking the exponentiated relevant parameters divided by 1+ the exponentiated relevant parameters
- Models for Polytomous (>2 nominal categories) and Ordinal (inherent ordering) Responses
  - Polytomous Response: Use Multicategorical Logit Model. Assumes $Y_i|x_i \sim Multi(n_i; \pi_1(x_i), ..., \pi_J(x_i))\ \forall i = 1, ..., N$ where $\pi_j(x_i; \beta) = \frac{e^{x_i^T \beta_j}}{\sum_{j=1}^J e^{x_i^T \beta_j}}\ \forall j = 1, ..., J$. Note that $\sum_{j=1}^J \pi_j = 1$, so need to set one $\beta_j = 0$ to make model identifiable
  - Ordinal Response: Use Proportional Odds Model. Assumes $Z_i|x_i \sim Multi(1; \pi_1(x_i), ..., \pi_J(x_i))\ \forall i = 1, ..., N$ where $P(Z_I \leq j|x_i) = \sum_{k=1}^j \pi_k(x_i)\ \forall j = 1, ..., J$ and $g(\frac{P(Z_I \leq j|x_i)}{1-P(Z_I \leq j|x_i)}) = \alpha_j + x_i^T\beta$. Here $g(.)$ is the logit link and α_j are cumulative probabilities (so expect them to increase with j). Note that the odds ratio here is invariant to choice of j, i.e. how we dichotomize the response variable (this is the assumption of the P.O.M., and if it isn't true then P.O.M. not a good fit)
- Poisson Regression: Assumes the components of $y = (y_1, ..., y_n)^T$ are mutually independent and $y_i|x_i \sim Pois(\mu(x_i))\ \forall i = 1, ..., n$. Used to model count data. If the canonical link (log link) is used, then we call this a loglinear model.
  - Loglinear Models for Contingency Tables: Analogous to ANOVA, except that the response variable is a count now instead of a continuous variable. Cell probabilities are $\{\pi_{ij}\}$, cell counts are $\{n_{ij}\}$, and $n_{ij} \sim Pois(\mu_{ij})$ and all n_{ij} are independent. Expected counts are $\mu_{ij} = n\pi_{ij}$. The saturated model can be written $log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$ and

	Y	
X	1	2
1	$e^{\lambda+\lambda_1^X+\lambda_1^Y+\lambda_{11}^{XY}}$	$e^{\lambda+\lambda_1^X-\lambda_1^Y-\lambda_{11}^{XY}}$
2	$e^{\lambda-\lambda_1^X+\lambda_1^Y-\lambda_{11}^{XY}}$	$e^{\lambda-\lambda_1^X-\lambda_1^Y+\lambda_{11}^{XY}}$

Nuisance Parameters

- Statistical models generally involve a set of parameters $\xi = (\psi, \lambda)$ in which ψ is/are the parameter(s) of interest and λ is/are the nuisance parameter(s). In most applications, β is of interest and σ^2 is nuisance. Can do inference on ψ in the presence of λ, but need specific methods that will eliminate λ while still accounting for its influence on ψ
- Conditional Likelihood: Key idea is to identify a CSS for λ, s_λ, such that the conditional distribution of $Y|s_\lambda$ depends only on ψ.
  - First, we must identify s_λ and determine that it satisfies either S- or P-ancillarity. Then we determine the distribution of $s_\lambda(\psi_0)$ for some fixed ψ_0. Finally, with this information we can write out the conditional likelihood function as $l_C(\psi) = log(p_{y|s_\lambda}) = log(p_y(\psi, \lambda)) - log(p_{s_\lambda}(\psi_0))$
  - Once we have the expression for the conditional likelihood, we can calculate the conditional score statistic $U_\psi = \frac{\delta l_C}{\delta \psi}\Big|_{\psi_0=\psi}$
  - 3 most common examples of this in practice are using Hypergeometric Distributions, Binary Matched Pairs, and Conditional Logistic Regression. Used in situations where sample size is very small or where we have many nuisance parameters

Models for Over-Dispersion

- Over-dispersion occurs when the actual variability of discrete data exceeds the variance predicted by the standard GLM. Have two ways of handling this:
- Assume a more general form of the variance function: Define the variance to include a dispersion parameter; this additional parameter may be estimated using moment methods and quasi-likelihood methods.
- Use two-level hierarchical models: Start out with the standard GLM. Then the parameter in the standard GLM is assumed to follow some distribution which contains additional (known) parameters. Use maximum likelihood estimation
  - Popular examples include the Beta-Binomial ($Y_i|p_i \sim Bin(n_i, p_i)$ and $p_i \sim Beta(\alpha, \beta) \Rightarrow Y_i \sim BetaBin(n_1, \alpha, \beta)$) and Poisson ($Y_i|\theta_i \sim Pois(\theta)$ and $\theta_i \sim \Gamma(k, \mu_i) \Rightarrow Y_i \sim NegBin(k, \mu_i)$)
- Purpose of accounting for over-dispersion is to model the data as accurately as possible (not underestimate SE of parameter estimates, etc.).
- These methods allow us to find an expression for the score function, compute the covariance matrix of the score function ($I(\beta) \rightarrow$ Fisher Info of β), and ultimately get the asymptotic covariance matrix of $\hat{\beta}$. This score test tests whether or not there is significant over-dispersion present in the data

Quasi-Likelihood ?
Diagnostics for GLMs ?

761

Decision Theory

- Non-Randomized Decision Rule: A NRDR $d(X)$ is a function such that $d : X \rightarrow A$. If $X = x$ is the observed value of X, then $d(x)$ is the action that will be taken (with probability 1).
  - Note that $d(X)$ is a rule when X is random, and $d(x)$ is the action once $X = x$ is observed
  - Will always have a NRDR in estimation problems
  - The Frequentist Risk Function for a NRDR $d(X)$ is defined by $R(\theta, d) = E_{X|\theta}[L(\theta, d(x)] = \int_X L(\theta, d(x)p(x|\theta)dx$ where $L(\theta, d(x))$ is the loss function. The risk function should always be deterministic and a function of θ or a constant
- Randomized Decision Rule: A RDR $d(a, x)$ is, for each x, a probability distribution on A, that is $d : A \times X \rightarrow [0, 1]$ and $d(a, x) \equiv d(a|X) =$ probability of action a when $X = x$ is observed
  - The Loss Function of a RDR is defined as $L(\theta, d(.|x)) = E_{d(.|x)}[L(\theta, a)] = \int L(\theta, a)p(a|x)da$. Loss function will always be a function of θ and x
  - The Frequentist Risk Function of a RDR is defined as $R(\theta, d) = E_{X|\theta}[L(\theta, d(.|x))] = \int_X \int_A L(\theta, a)p(a|x)p(x|\theta)dadx$ (if cts) and $= \sum_{i=1}^{k} L(\theta, a_i)\{\sum_{j=1}^{m} d(a_i, x_j)P_\theta(X = x_j)\}$ (if discrete). Should always be a function of θ or a constant or ∞
  - An example of a RDR is hypothesis testing, and RDRs are needed when the random variable of interest is discrete
- In general, the frequentist risk is equal to the expected value of the loss function
  - In estimation problems, under squared error loss the risk of an estimator $\hat{\theta}$ of θ is equal to the MSE. In symbols, $R(\theta, \hat{\theta}) = E_\theta[(\theta - \hat{\theta})^2] = Var_\theta(\hat{\theta}) + bias(\hat{\theta})^2$
  - A rule with constant frequentist risk is called an equalizer rule
- (In)Admissibility: A decision rule d is inadmissible if there is a rule d' such that $R(\theta, d') \leq R(\theta, d)$ for all θ and $R(\theta, d') < R(\theta, d)$ for some θ. A decision rule is admissible if it is not inadmissible.
  - Admissible rules are not necessarily optimal; many admissible rules may exist
  - Risk functions of admissible rules may cross

- Optimal Rules: A decision rule d is said to optimal if d is admissible and is equal to any other admissible rule. So, optimal rules consist of unique admissible rules, and optimal rules don't always exist
- Minimaxity: A decision rule is minimax if $\inf_{d\in D}\{\sup_{\theta\in\Theta} R(\theta,d)\} = \sup_{\theta\in\Theta} R(\theta,d_M)$ where the right side of the equation is called the minimax value of the problem and d_M is the minimax rule
  - That is, a rule d is minimax if it has the "best (smallest) worst case (highest risk) scenario among all randomized rules". Want to avoid the worst case scenario at all costs, even if that means choosing a rule that has higher average risk
- Prior Distribution: A probability distribution Λ over Θ is called a prior distribution.
  - A proper prior will integrate over its full domain to 1; Bayes Rules are based on proper priors
  - An improper prior is a prior for which $\int_\Theta \lambda(\theta)d\lambda = \infty$; proper posterior distributions can often be obtained with improper priors
- Posterior Distribution: The posterior density of θ can be written $p(\theta|x) = \frac{p(x|\theta)\lambda(\theta)}{\int p(x|\theta)\lambda(\theta)d\theta} = \frac{p(x|\theta)\lambda(\theta)}{p(x)} \propto p(x|\theta)\lambda(\theta)$ since x is given and so the marginal of x has no dependence on θ
  - The posterior expected loss is $\int L(\theta,d(x))p(\theta|x)d\theta$
- Bayes Risk: For any given prior Λ and $d\in D$, the Bayes Risk of d with respect to Λ is $\mathcal{R}(\Lambda,d) = E_X[E_{\theta|X}[L(\theta,d_\Lambda)|X]] = \int_\Theta R(\theta,d)d\Lambda(\theta)$ (if cts) and $= \sum_{i=1}^{l} R(\theta_i,d)\lambda_i$ (if discrete)
  - Bayes risk is always a number!
- A Bayesian Decision Rule (Bayes Rule) with respect to Λ, d_Λ is any rule satisfying $\mathcal{R}(\Lambda,d_\Lambda) = inf_{d\in D}\mathcal{R}(\Lambda,d) =$ Bayes Risk
  - In estimation problems, we call the Bayes rule the Bayes estimator
  - The idea of this rule is to minimize the average risk
- Common Loss Functions:
  - Quadratic Loss: $L(\theta,a) = k(\theta-a)^2$ for some $k = k(\theta)$ (if $k(\theta)=1$ then have Squared Error Loss)
    - Bayes (and generalized Bayes) estimator of θ under this loss function is the posterior mean of θ
  - 0-1 Loss: $L(\theta,a) = \begin{cases} 1, \ if\ \theta\notin\Theta_0 \\ 0, \ if\ \theta\in\Theta_0 \end{cases}$
    - Under 0-1 loss in hypothesis testing, the risks are just the probability of type I and type II errors
    - Bayes estimator of θ under 0-1 loss is the posterior mode of θ
  - Absolute Error Loss: $L(\theta,a) = k|\theta-a|$ for some $k = k(\theta)$ and where $|a| = \sqrt{a'a}$
    - Bayes estimator of θ under this loss function is the posterior median of θ
- Conjugate Prior: If the posterior distribution $p(\theta|x)$ has the same distribution as the prior $\lambda(\theta)$ then $\lambda(\theta)$ is called a conjugate prior (although parameters will change to include information from the x's). Some common examples are
  - $X|\theta\sim Pois(\theta),\ \theta\sim\Gamma \Rightarrow \theta|X\sim\Gamma$
  - $X|\theta\sim N(\theta,\sigma^2),\ \theta\sim N \Rightarrow \theta|X\sim N$
  - $X|\theta\sim Bin(n,\theta),\ \theta\sim Beta \Rightarrow \theta|X\sim Beta$
  - $X|\theta\sim Multi_k(n,\theta),\ \theta\sim Dirichlet \Rightarrow \theta|X\sim Dirichlet$
- Least Favorable Prior: A prior Λ_0 for which $\mathcal{R}(\Lambda,d_\Lambda)$ is maximized is called a least favorable prior, i.e. $\mathcal{R}(\Lambda_0,d_{\Lambda_0}) = sup_\Lambda\mathcal{R}(\Lambda,d_\Lambda)$
- Generalized Bayes Rule: Only difference between a Bayes rule and a generalized Bayes rule is that a GBR has an improper prior distribution on θ

- If log likelihood of a random variable belongs to the exponential family and $\tilde{\theta}_n$ is the Bayes estimator w.r.t. the prior λ under squared error loss, then $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N(0, I^{-1}(\theta_0))$ where $I^{-1}(\theta_0)$ is the CRLB. This tells us that the Bayes estimator $\tilde{\theta_n}$ is $\sqrt{n}$-consistent and asymptotically efficient
- Bayes rules, minimax rules, and admissible rules are always on the boundary of the risk body
- Finding Bayes Rules
  - The Bayes rule d_Λ is the rule which minimizes (w.r.t. d(x)) the posterior expected loss; Bayes rule = $arg\min_{d(x)} E_{\theta|X}[L(\theta, d(x))] = arg\min_{d(x)} \int L(\theta, d(x))p(\theta|x)d\theta$
  - If we don't know the form of the prior, must use the definition to get the Bayes rule
  - If X is a continuous random variable with exponential family density and Θ is closed, then any admissible estimator is a generalized Bayes rule
  - If doing estimation with weighted squared error loss, then $d_\Lambda = \frac{E[k(\theta)\cdot\theta|X]}{E[k(\theta)|X]}$ (Note that when $k(\theta) = 1$, this is the posterior mean of θ); d_Λ is a Bayes or Generalized Bayes rule depending on the prior
  - An admissible rule is a Bayes rule for some prior Λ
- Finding Minimax Rules
  - Use definition
  - If d_Λ is Bayes (w.r.t. Λ) and has constant risk (w.r.t. θ) then d_Λ is minimax
  - If d_Λ satisfies $\mathcal{R}(\Lambda, d_\Lambda) = \sup_\theta R(\theta, d_\lambda)$ then d_Λ is minimax and Λ is the least favorable prior
    - Works also if $\{\Lambda_k\}$ is a sequence of prior distributions with Bayes risks approaching the max risk (use when the supremum over θ of the frequentist risk is ∞)
  - If $\{\Lambda_k\}$ is a sequence of proper priors, then d is minimax if $R(\theta, d) \leq lim_{k\to\infty}\mathcal{R}(\Lambda_k, d_{\Lambda_k}) < \infty\ \forall\theta$
    - Good method to find a minimax rule is to guess a least favorable prior whose limit is an improper prior, derive the generalized Bayes rule for the improper prior, and often this rule will be minimax
    - use when the supremum over θ of the frequentist risk is ∞
  - If we have that an estimator is minimax in one parameter space, and we consider a larger parameter space (containing the first) that has the same max risk as the smaller space, then the estimator is still minimax in the larger space
- Proving (In)Admissibility
  - Use definition - directly compare the risks of two rules d and d'
  - In a no-data problem, $R(\theta, d) = L(\theta, d)$ and so we examine the loss matrix to determine (in)admissibility
  - Unique minimax estimator is admissible
  - Any unique Bayes estimator with finite Bayes risk is admissible
  - If $X_1, ..., X_n$ are k-dim random vectors such that i.i.d. $N_k(\theta, I)$ under squared error loss then $\bar{X}$ is admissible if $k \leq 2$ and inadmissible if $k \geq 3$ (James Stein estimator beats this, but it is also inadmissible)
  - If X is a continuous random variable with exponential family density and Θ is closed, then not generalized Bayes $\Rightarrow$ not admissible
  - Most common technique for proving inadmissibility of a generalized Bayes rule is to construct an improved estimator (often a shrinkage estimator)

Hypothesis Testing

- Size and Power of a Test:
  - A test is size α if $\sup_{\theta\in\Theta_0} E_\theta[\phi(x)] = \alpha$
  - The power function of a test ϕ is $\beta(\theta) = E_\theta[\phi(x)]$ where θ ranges over the whole sample space, i.e. $\theta \in \Theta_0 \cup \Theta_1$
  - A test ϕ_0 is uniformly most powerful (UMP) of size α if it has size α and $E_\theta[\phi_0(x)] \geq E_\theta[\phi(x)]$ for all $\theta \in \Theta_1$ and all ϕ in the class of size α tests

  - If the rejection region for the test depends on the value(s) specified in the alternative hypothesis, no UMP test exists
  - If there are nuisance parameters present, (most likely) no UMP test exists
  - Any test is size and power consistent if both probabilities of error (α and $1-\beta$) shrink to zero as $n \to \infty$ (all NP tests are size and power consistent)
  - The power function $\beta_\phi(\theta)$ is continuous in θ for all ϕ for any exponential family

- Neyman-Pearson Lemma: Have desities p_0 and p_1 (can be anything). Let $0 \leq \alpha \leq 1$. Then there exists a constant k and a critical function ϕ of the form $\phi(x) = \begin{cases} 1, \ if \ p_1(x) > kp_0(x) \\ \gamma, \ if \ p_1(x) = kp_0(x) \\ 0, \ if \ p_1(x) < kp_0(x) \end{cases}$ such that $E_0[\phi(x)] = \alpha$, where the subscript "0" here indicates that the expectation is taken under H_0. This is a most powerful size α test of p_0 vs p_1.
  - The constant k will depend on α
  - When X is continuous, generally $p_1(x) = kp_0(x)$ only occurs on a set of measure zero
  - Use this whenever we have a simple vs. simple hypothesis
  - if $\gamma \neq 0$ then we say $\gamma(x) = \frac{\alpha - P_0(p_1 > kp_0)}{P_0(p_1 = kp_0)}$
  - The usual problem setup is as follows:
    1. Find sufficient statistic for your parameter from the distribution of your random variable (or joint distribution of your random sample)
    2. We can plug in this $T(x)$ to $\phi(x)$ in place of $\frac{p_1(x)}{p_0(x)}$
    3. Write out $\phi(x)$
    4. Solve for k using distribution of your sufficient statistic $T(x)$ (with the given expectation relating it to α). This is simple with continuous distributions; with discrete, need concrete values of α, θ_0, n and need to find k with distribution of $T(x)$, make a table of several k's, etc.
    5. If critical function is randomized, solve for γ; plug in k and use the expectation

- Monotone Likelihood Ratio: If the family of densities $\{p_\theta : \theta \in [\theta_0, \theta_1] \subset R\}$ is such that $\frac{p_{\theta'}(x)}{p_\theta(x)}$ is nondecreasing in $T(x) \ \forall \ \theta < \theta'$, then the family has the MLR property.
  - Have it: Binomial, Poisson, Unif[0,θ], Hypergeometric, Gamma, Beta, Negative Binomial, Geometric, Normal, Exponential, non-central t, χ^2, F, and all exponential family
  - Don't have it: Cauchy
  - If X has density $p_\theta(x)$ with MLR in $T(x)$, then there exists a UMP size α test for type (1) hypotheses; namely,
    * For $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, the UMP size α test is $\phi(x) = \begin{cases} 1, \ if \ T(x) > k \\ \gamma, \ if \ T(x) = k \\ 0, \ if \ T(x) < k \end{cases}$ with $E_{\theta_0}[\phi(x)] = \alpha$
    * For $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$, the UMP size α test is $\phi(x) = \begin{cases} 1, \ if \ T(x) < k \\ \gamma, \ if \ T(x) = k \\ 0, \ if \ T(x) > k \end{cases}$ with $E_{\theta_0}[\phi(x)] = \alpha$

- Generalized N-P Lemma ?

- SOB, Neyman Structure, and UMPU
  - ϕ is unbiased if $\beta_\phi(\theta) \geq \alpha \ \forall \ \theta \in \Theta_1$ and $\beta_\phi(\theta) \leq \alpha \ \forall \ \theta \in \Theta_0$
  - ϕ is similar on the boundary (SOB) if $\beta_\phi(\theta) = \alpha \ \forall \ \theta \in \Theta_B$ (SOB tests only care about power on the boundary of the parameter space)
  - A UMPU test of size α is a test ϕ_0 for which $E_\theta[\phi_0] \geq E_\theta[\phi]$ (i.e. $\beta_{\phi_0}[\theta] \geq \beta_\phi[\theta]$) for all $\theta \in \Theta_1$, for all unbiased size α tests ϕ

- UMP $\subset$ UMPU ; Unbiased tests $\subset$ SOB tests
- UMP among all SOB tests + continuous power function $\beta(\theta) \Rightarrow$ UMPU
- Let T be sufficient for $P_B \equiv \{P_\theta : \theta \in \Theta_B\}$ and let $P^T \equiv \{P_\theta^T : \theta \in \Theta_B\}$. A test function ϕ has Neyman Structure with respect to T if $E[\phi(x)|T] = \alpha$ (i.e. ϕ has N.S. if the expectation of $\phi|T$ is on the boundary α)
- Neyman Structure $\Rightarrow$ SOB (with respect to T)
- To use these facts:
  1. Check that $E[\phi(x)|T] = \alpha$ when $\theta \in \Theta_B$
  2. If true, ϕ has N.S. and is therefore SOB
  3. Can find UMP N.S. test with simple hypothesis using N.P. Lemma
  4. This gives UMP SOB, then if $\beta(\theta)$ is continuous we have a UMPU test

- Types of Hypotheses Being Tested:
  1. $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$
  2. $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs $H_1 : \theta_1 < \theta < \theta_2$
  3. $H_0 : \theta_1 \leq \theta \leq \theta_2$ vs. $H_1 : \theta < \theta_1$ or $\theta > \theta_2$
  4. $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

- If the vector of random variables $X = (X_1, ..., X_n)$ belong to a one parameter exponential family and $T(x)$ is a CSS for θ, then tests of hypotheses (1) - (4) are as follows:
  1. The test $\phi(T(x)) = \begin{cases} 1, \ if \ T(x) > k \\ \gamma, \ if \ T(x) = k \\ 0, \ if \ T(x) < k \end{cases}$ with $E_{\theta_0}[\phi(T)] = \alpha$ is UMP level α for hypothesis (1)
  2. The test $\phi(T(x)) = \begin{cases} 1, \ if \ k_1 < T(x) < k_2 \\ \gamma_i, \ if \ T(x) = k_i, i = 1, 2 \\ 0, \ if \ else \end{cases}$ with $E_{\theta_i}[\phi(T)] = \alpha, \ i = 1, 2$, is UMP lvl α for (2)
  3. The test $\phi(T(x)) = \begin{cases} 1, \ if \ T(x) < k_1 \ or \ T(x) > k_2 \\ \gamma_i, \ if \ T(x) = k_i, i = 1, 2 \\ 0, \ if \ else \end{cases}$ with $E_{\theta_i}[\phi(T)] = \alpha, \ i = 1, 2$, is UMPU level α for (3).
     A UMP test for (3) does not exist.
  4. The test $\phi(T(x)) = \begin{cases} 1, \ if \ T(x) < k_1 \ or \ T(x) > k_2 \\ \gamma_i, \ if \ T(x) = k_i, i = 1, 2 \\ 0, \ if \ else \end{cases}$ with $E_{\theta_0}[\phi(T)] = \alpha$ and $E_{\theta_0}[T\phi(T)] = \alpha E_{\theta_0}[T]$ is UMPU level α for (4). A UMP test for (4) does not exist.

- If the vector of random variables $X = (X_1, ..., X_n)$ has density belonging to the multiparameter exponential family with k+1 dimensions $\Theta = \{(\theta, \xi_1, ..., \xi_k)\}$ where the ξ_i are nuisance parameters, $U(x)$ is the CSS for θ and $T_i(x)$ are CSS for the nuisance parameters, then the following are UMPU tests for (1) - (4) respectively:
  1. The test $\phi(x) = \begin{cases} 1, \ if \ U > c(t) \\ \gamma(t), \ if \ U = c(t) \\ 0, \ if \ U < c(t) \end{cases}$ with $E_{\theta_0}[\phi(U)|T = t] = \alpha$.
     If $V \equiv h(U, T)$ is increasing in U for each fixed t and is independent of T on Θ_B, then
     $$\phi(x) = \begin{cases} 1, \ if \ V > c \\ \gamma, \ if \ V = c \\ 0, \ if \ V < c \end{cases}$$
     is the unconditional UMPU test.

2. The test $\phi(x) = \begin{cases} 1,\ if\ c_1(t) < U < c_2(t) \\ \gamma_i,\ if\ U = c_i(t) \\ 0,\ if\ else \end{cases}$ with $E_{\theta_i}[\phi(U)|T=t] = \alpha$, i=1,2.

   If can find a pivotal quantity V that is monotone in U and V is independent of T on the boundary, then can replace U with its expression in terms of V in the test and equation above and the test becomes unconditional.

3. The test $\phi(x) = \begin{cases} 1,\ if\ U < c_1(t)\ or\ U > c_2(t) \\ \gamma_i,\ if\ U = c_i(t) \\ 0,\ if\ else \end{cases}$ with $E_{\theta_0}[\phi(U)|T=t] = \alpha$.

   If can find a pivotal quantity V that is monotone in U and V is independent of T on the boundary, then can replace U with its expression in terms of V in the test and equation above and the test becomes unconditional.

4. The test $\phi(x) = \begin{cases} 1,\ if\ U < c_1(t)\ or\ U > c_2(t) \\ \gamma_i,\ if\ U = c_i(t) \\ 0,\ if\ else \end{cases}$ with $E_{\theta_0}[\phi(U)|T=t] = \alpha$ and $E_{\theta_0}[u\phi(U)|T=t] = \alpha E_{\theta_0}[U|T=t]$.

   If $V \equiv h(U,T) = a(t)U + b(t)$ with $a(t) > 0$, and if V is independent of T on the boundary, then replacing U with $(V - b(t))/a(t)$ in the test and equations above makes the test unconditional.

- Likelihood Ratio Test (LRT): Suppose $X_1, ..., X_n$ are i.i.d. each with density $p(x|\theta)$. The LRT is defined by $\Lambda = \frac{\sup_{\theta\in\Theta_0} p_n(x|\theta)}{\sup_{\theta\in\Theta_0\cup\Theta_1} p_n(x|\theta)}$ (where $p_n(x|\theta) = \prod_{i=1}^n p(x|\theta)$ is the joint density) for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$.

  So, the LRT size α test is $\phi(x) = \begin{cases} 1,\ if\ \Lambda < k \\ \gamma,\ if\ \Lambda = k \\ 0,\ if\ \Lambda > k \end{cases}$, where k is chosen such that $\alpha = \sup_{\theta\in\Theta_0} E[\phi(x)]$

  - LRT always exists and is a good alternative test when UMP(U) test doesn't exist or is too difficult to find (e.g. when have many nuisance parameters)
  - LRT is often equivalent to the UMP(U) test when the UMP(U) test exists (can test this by finding expressions for each and comparing their rejection regions)
  - The idea is to find the values of the parameters (θ, ξ) that maximize the likelihood function under H_0 (numerator) and the whole space (denominator). Plug in these MLE's and reduce as far as possible (remember that any constants w.r.t. X_i can be absorbed by k). Ideally, will know the distribution of this reduced form of Λ and then we can solve for k
  - If X_i have distribution in the one-parameter exponential family and $Q(\theta)$ (from exponential family form) is monotone in θ then the LRT coincides with the UMP(U) test
  - The NP test coincides with the LRT when we have a simple vs. simple hypothesis
  - Remember that we only need the distribution of the test statistic (reduced form of Λ) under H_0 to solve for k
  - When finding MLE's under H_0 choose boundary values for parameters involved in H_0, and find MLE's of other parameters with the boundary values plugged in for the parameters we already have

- Wald and Score Tests: If we wish to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, then we can use the Wald and Score tests (see 762 general hypothesis testing section)

  - All three of $W_n,\ SC_n$ and $-2log(\Lambda_n) \to_{H_0} \chi^2_r$ where $r = dim(\Theta) - dim(\Theta_0)$ = difference between number of parameters free to vary in whole and null spaces

- Bayesian Hypothesis Testing: The quantity used to test hypotheses in the Bayesian framework is called the Bayes factor, which is the Bayesian analog of the LRT. If we have $H_0 : \theta \in \Theta_0$ vs $H_0 : \theta \in \Theta_1$ then the Bayes factor in favor of H_0 is $B = \frac{p(x|H_0)}{p(x|H_1)} = \frac{\int_{\Theta_0} p(x|\theta,H_0)\lambda(\theta|H_0)d\theta}{\int_{\Theta_1} p(x|\theta,H_1)\lambda(\theta|H_1)d\theta}$ where $\lambda(\theta|H_i)$ is the prior distribution under H_i

  - B can't be used with an improper prior for θ (and will usually be given the prior under H_0 and H_1)
  - B doesn't require models to be nested, whereas the LRT does
  - In simple vs. simple case, B = NP test = LRT
  - The larger the value of B, the stronger the evidence in support of H_0

  - If have simple hypothesis, then just plug in the given value of θ to the density
  - Advantage of B is that it easily accommodates strange hypotheses that are not handled well in the frequentist setting
  - Posterior probability of $H_0 = \frac{rB}{1+rB}$ where $r = \frac{p(H_0)}{p(H_1)} = \frac{p(H_0)}{1-p(H_0)}$

- G-Invariant Tests: A test $\phi(x)$ is invariant under $G \iff \phi(g(x)) = \phi(x)\ \forall g \in G$ and $x \in X$. A function $T : X \to R^k$ is GMI if T is invariant under G and $T(x_1) = T(x_2) \Rightarrow x_2 = g(x_1)$. The idea is to create a test that won't change its conclusion under different types of transformations
  - A group G is a set equipped with an operation that satisfies 4 properties: Closure, Associativity, Identity, and Inverse
  - If T is a GMI, then $\phi(x) = h(T(x))\ \forall\ x \in X$ is invariant for some h, i.e. all invariant tests should be a function of $T(x)$ (the test statistic)
  - How to find a UMP G-invariant size α test of H_0 vs H_1
    1. Will be given a group G such that H_0 and H_1 are invariant under it (should check)
    2. Write out joint likelihood of random sample(s) and get SS for all unknown parameters
    3. Build a function $T(x)$ from SS's such that T is unchanged in G^*, i.e. T is G^*MI (check 2 conditions from definition)
    4. Multiply T by whatever constants necessary to get it to a distribution you recognize
    5. Transform the expression of your H_0 in the same manner you transformed the SS's of the parameters in H_0 to get δ
    6. For these $H_0' : \delta \leq \delta_0$ and $H_1' : \delta > \delta_0$, the UMP G-invariant size α test of H_0 vs H_1 is

$$\psi(T) = \begin{cases} 1, \ if\ T > c \\ \gamma, \ if\ T = c \\ 0, \ if\ T < c \end{cases} \qquad with\ E_{\delta_0}[\psi(T)] = \alpha$$

- Pitman Efficiency: When we have 2 tests T_1, T_2 that are both size α and consistent, we use Pitman efficiency to compare them by examining power under a sequence of local alternatives that approach $\theta_0 : \theta_0 + n^{-1/2}c_n$ where $c_n \to c$ and n here is the sample size. P.E. is the limiting ratio of sample sizes that produces equal asymptotic power against the same sequence of alternatives
  - Compute $\frac{N_2}{N_1} \to \frac{(\mu_1/\sigma_1)^2}{(\mu_2/\sigma_2)^2} = e_{1,2}$ by setting the asymptotic powers of the two tests equal to each other and setting the alternatives of the two tests equal to each other, then solve for N_1 and N_2. Here μ_i and $\sigma_i,\ i = 1, 2$ are the limiting mean and SD of test i (assuming the distribution of the test statistic is Normal asymptotically). If $e_{1,2} > 1$ then test 1 is more efficient (because it requires smaller sample size to achieve the same power)

- Confidence Regions: Can be found with 2 common methods, either by using pivotal quantities or by inverting the acceptance region of a test
  - Confidence bound $\iff$ 1-sided alternative $H_1 \iff$ 1-sided confidence set
  - Confidence interval $\iff$ 2-sided alternative $H_1 \iff$ 2-sided confidence set
  - A pivotal quantity for θ must meet 2 requirements: the expression must depend on θ and the data, and the distribution of the pivotal quantity must be independent of θ
    * It is desirable to construct our pivotal quantity from our SS $T(x)$ for θ, because we can usually easily obtain its distribution, which we will need in order to identify the appropriate quantiles a and b of $P(a < P.Q. < b) = 1 - \alpha$
    * If want joint confidence region of 2 parameters, ideally we can find 2 pivotal quantities (one dependent on one parameter and another dependent on the other parameter only or both). If the pivots are independent (can use Basu's to show), then the joint confidence region may be defined by these two pivots
  - Once we have an optimal α-level hypothesis test of $H_0 : \theta = \theta_0$ vs some H_1 (specify H_1 to get desired CR structure), can always put the acceptance region in terms of θ to get an optimal confidence region; when we have a UMP(U) test, can invert the acceptance region to get a UMA(U) confidence set

    - ∗ If we have a test with acceptance region $A(\theta)$, then the corresponding CR is $CR = \{\theta : X \in A(\theta)\}$. The acceptance region will depend on some statistic $T(X)$; find its distribution and use it to determine a and b as above
    - ∗ The expression for the acceptance region will have θ_0 in it, but we can swap θ in for θ_0 because θ_0 is just an unrestricted specification of θ

- Highest Posterior Density Regions and Credible Sets: The Bayesian CI is called the HPD region.
    - Let $p(\theta|x)$ denote the posterior density of θ. A region R in the parameter space of θ is a HPD region of content 1-α if $P(\theta \in R|x) = 1 - \alpha$ and for $\theta_1 \in R$ and $\theta_2 \notin R$, $p(\theta_1|x) \geq p(\theta_2|x)$
    - A credible set is a posterior region that is constructed by removing the upper and lower $\alpha/2$ percentiles of the posterior distribution
    - HPD region and credible set are the same when the posterior distribution is symmetric (e.g. Normal, t)

Resampling Methods

- In denial

High-Dimensional Regression

- that these will be covered