PhD Course and Research Review

Fall 2024
Version 2

Miles Woollacott

PhD Student in Statistics at North Carolina State University
miles.woollacott@gmail.com

Note: This document is likely not free from typos or other errors. Please email the author with any changes necessary for this document.

Contents

1 ST 701: Statistical Theory I

Instructor: Dr. Luo Xiao
Semester: Fall 2023
Main Textbook: Casella and Berger, *Statistical Inference*

1.1 Probability

Return to Table of Contents

- The **sample space**, denoted as $\boldsymbol{\zeta}$, is the set of all possible outcomes of an experiment.
  - An **event** is any subset of ζ.
- The **complement** of set A is $A^c = \{b \in \zeta : b \notin A\}$.
- **DeMorgan's law** states that $(A \cap B)^c = A^c \cup B^c$, and $(A \cup B)^c = A^c \cap B^c$.
- Two sets A and B are **disjoint**, or **mutually exclusive**, if $A \cap B = \varnothing$.
- Two *disjoint* sets A and B form a **partition** of C if $A \cup B = C$.
- A **probability function** takes in events from ζ as input, and outputs a probability.
  - $0 \leq P(A) \leq 1$ for all $A \in \zeta$.
  - $P(\zeta) = 1$.
  - If A_i are mutually exclusive for all $i \in \{1, \ldots, n\}$, then $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- The **Bonferroni inequality** states that $P(A \cap B) \geq P(A) + P(B) - 1$.
- $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$.
- The **fundamental theorem of counting** says that a job consisting of k separate tasks can be done in $\prod_{i=1}^k n_i$ ways, where n_i is the number of ways the ith task can be done.
- The **conditional probability** of A given B is $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
  - **Bayes' formula** is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Events A and B are **independent** iff $P(A|B) = A$ (or $P(B|A) = B$), or $P(A \cap B) = P(A)P(B)$.

1.2 Distributions

Return to Table of Contents

- A **random variable**, or **RV**, is a function of the sample space.
- A function f is **right-continuous** at point c if $\lim_{x \to c^+} f(x) = f(c)$.
- A function f is **non-decreasing** if $f(x_1) \leq f(x_2)$ for $x_1 < x_2$.
- A function f is **increasing** if $f(x_1) < f(x_2)$ for $x_1 < x_2$.
- A function f is **monotone** if f is either increasing or decreasing over its entire support.
- The **cumulative distribution function**, or **CDF**, of RV X is $F_X(x) = P(X \leq x)$ for $x \in \mathbb{R}$.
  - F_X must be right-continuous and non-decreasing.
- RVs X and Y are **identically distributed** if $F_X(a) = F_Y(a)$ for all $a \in \mathbb{R}$.
- The **probability mass function**, or **PMF**, of a discrete RV X is $f_X(x) = P(X = x)$.
- The **probability density function**, or **PDF**, or a continuous RV X is $\frac{d}{dx} F_X(x)$.
- If g is increasing, then $F_Y(y) = F_X(g^{-1}(y))$.
  - If g is decreasing, then $F_Y(y) = 1 - F_X(g^{-1}(y))$.

- If g is monotone, then $f_Y(y) = f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|$.

- The CDF of a continuous RV follows a $U(0,1)$ distribution.
- Suppose F is a CDF, and $Y \sim U(0,1)$. Then, $F_X^{-1}(Y) = F_X(x)$.
- Suppose there exists partitions of X, called $A_1, \ldots, A_p$, such that $g(x)$ is monotone on each A_i, $g(x) = g_i(x)$ for $x \in A_i$, and $\{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for all A_i. Then, $f_Y(y) = \sum_{i=1}^p f_X(g_i^{-1}(y)) \left|\frac{d}{dy} g_i^{-1}(y)\right|$.
  Example: Suppose $Z \sim \mathcal{N}(0,1)$, and apply the transformation $Y = Z^2$.

  $g(z) = z^2$ is monotone for $z < 0$ and $z > 0$. Define $A_0 = \{0\}$, $A_1 = (-\infty, 0)$, and $A_2 = (0, \infty)$, with $g_1(z) = g_2(z) = z^2$, $g_1^{-1}(y) = -\sqrt{y}$, and $g_2^{-1}(y) = \sqrt{y}$.

$$\begin{aligned} f_Y(y) =& \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left|-\frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \left|\frac{1}{2\sqrt{y}}\right| \\ =& \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2},\ y > 0. \ \blacksquare \end{aligned}$$

1.3 Moments and Expectations

Return to Table of Contents

- The **expected value** of RV Y is $\mathbb{E}(Y) = \int_{\zeta_Y} y f_Y(y) dy$ if continuous, or $\sum_{\zeta_Y} y f_Y(y)$ if discrete.
  - If $\mathbb{E}(X^2)$ exists, then $\mathbb{E}(X-b)^2$ is minimized at $b = \mathbb{E}(X)$.
- **Markov's inequality**: $P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.
- **Chebyshev's inequality**: $P(|X| \geq a) \leq \frac{\mathbb{E}(X^2)}{a^2}$ for $a > 0$.
- **Holder's inequality**: Let X, Y be two RVs, and p and q satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then $|\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}$.
- **Jensen's inequality**: Suppose a function f is convex. Then, $\mathbb{E}[f(X)] \geq f[\mathbb{E}(X)]$.
- **Stein's Lemma**: $\mathbb{E}[g(X)(X-\mu)] = \sigma^2 \mathbb{E}[g'(X)]$, where $X \sim \mathcal{N}(\mu, \sigma^2)$.
- The nth **moment** of an RV X is $\mathbb{E}(X^n)$.
- The nth **central moment** of an RV X is $\mathbb{E}[(X - \mathbb{E}(X))^n]$.
- The **variance** of an RV X is $\mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.
  - $Var(a + bX) = b^2 Var(X)$.
- The **moment generating function**, or **MGF**, of an RV X is $M_X(t) = \mathbb{E}(e^{tX})$.
  - $M_{(aX+b)}(t) = e^{tb} M_X(at)$.
  - If $M_X(t) < \infty$ for all t in an open interval containing zero, then $\mathbb{E}(X^n) = \frac{d^n}{dx^n} M_X(t)|_{t=0}$.
  - If $M_X(t) = M_Y(t) < \infty$ for all t in an open interval containing zero, then X and Y are identically distributed.
- A family of PDFs or PMFs is an **exponential family** if it can be rewritten as $f(x|\boldsymbol{\theta}) = h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^p w_i(\boldsymbol{\theta}) t_i(x)\right)$.
  - $\eta := w(\theta)$ is the **natural parameterization**.
    - The **natural parameter space** is the region(s) η is defined on.
  - Can be reparameterized to be $f(x|\theta) = h(x) c^*(\eta) \exp\left(\sum_{i=1}^p \eta_i t_i(x)\right)$.
    - An exponential family is **full-rank** iff $\eta(\Theta)$ contains an open set.
    - An exponential family is **curved** if it is not full-rank.
  - $\mathbb{E}\left(\sum_{i=1}^p \frac{\partial}{\partial \theta_j} w_i(\boldsymbol{\theta}) t_i(X)\right) = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta})$.
  - $Var\left(\sum_{i=1}^p \frac{\partial}{\partial \theta_j} w_i(\boldsymbol{\theta}) t_i(X)\right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - \mathbb{E}\left(\sum_{i=1}^p \frac{\partial^2}{\partial \theta_j^2} w_i(\boldsymbol{\theta}) t_i(X)\right)$.
- A family of PDFs and PMFs is called a **location and scale family** if it has the form $\frac{1}{\sigma} f_X\left(\frac{x-\mu}{\sigma}\right)$, indexed by $\mu \in \mathbb{R}$ and $\sigma > 0$.

1.4 Bivariate Random Variables

Return to Table of Contents

- The **marginal PDF** of X is $f_X(x) = \int_y f_{X,Y}(x,y)dy$.
- The **conditional PDF** of $Y|X$ is $f_{Y|X}(y|X=x) = \frac{f(x,y)}{f_X(x)}$.
- X and Y are **independent** iff $f(x,y) = f_X(x)f_Y(y)$.
  - $X \perp Y$ iff $f(x,y) = g(x)h(y)$.
  - $X \perp Y$ iff $M_{X+Y}(t) = M_X(t)M_Y(t)$.
- Suppose $U = g_1(x,y)$ and $V = g_2(x,y)$, where $(g_1, g_2) : \zeta_{X,Y} \to \zeta_{U,V}$ is bijective. The **Jacobian** is $J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{pmatrix}$.
- Under proper conditions, $f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) \cdot |\det(J)|$.
  Example: Suppose $f(x,y) = \frac{1}{4}e^{-\frac{x+y}{2}}$ for $x > 0$, $y > 0$. Find the PDF of $Z = X - Y$.

  Let $U = Y$. Therefore, $Y = U$, and $X = Z + U$. $J = \begin{pmatrix} \frac{\partial x}{\partial z} & \frac{\partial y}{\partial z} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, so $\det(J) = 1$.

$$\begin{aligned} f_{Z,U}(z,u) =& f_{X,Y}(z+u,u) \cdot |\det(J)| \\ =& \frac{1}{4}e^{-\frac{z+u+u}{2}} \cdot 1 = \frac{1}{4}e^{-\frac{z}{2}-u}; \end{aligned}$$

  Note that $x > 0$ and $y > 0$ means that $z + u > 0$ and $u > 0$, so if $z < 0$, then $u > -z$.

  Case 1: $z < 0$.

$$\begin{aligned} f_Z(z) &= \int_{-z}^{\infty} \frac{1}{4}e^{-\frac{z}{2}-u}du \\ &= \left[\frac{1}{4}e^{-z/2}e^{-u}\right]\Big|_{-z}^{\infty} = \frac{1}{4}e^{z/2}. \end{aligned}$$

  Case 2: $z > 0$.

$$\begin{aligned} f_Z(z) &= \int_{0}^{\infty} \frac{1}{4}e^{-\frac{z}{2}-u}du \\ &= \left[\frac{1}{4}e^{-z/2}e^{-u}\right]\Big|_{0}^{\infty} = \frac{1}{4}e^{-z/2}. \end{aligned}$$

  This means that $f_Z(z) = \frac{1}{4}e^{-|z|/2}$ for $z \in \mathbb{R}$. ■
- If $X \perp Y$, then $U = h(X) \perp V = g(Y)$.
- $\mathbb{E}(Y) = \mathbb{E}\left[\mathbb{E}(Y|X)\right]$.
- $M_Y(t) = \mathbb{E}\left[\mathbb{E}(e^{tY}|X)\right]$.
- $Var(Y) = \mathbb{E}\left[Var(Y|X)\right] + Var\left[\mathbb{E}(Y|X)\right]$.
- $Cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$.
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$.

1.5 Statistics and Order Statistics

Return to Table of Contents

- **Statistics** are functions of random variables.
- **Sample variance** is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
  - $\mathbb{E}(S^2) = \sigma^2$.
  - If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$:
    - $\bar{X} \perp S^2$.
    - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- If $X \sim F_{p,q}$, then $\frac{1}{X} \sim F_{q,p}$.
- If $X \sim t_q$, then $X^2 \sim F_{1,q}$.
- If $X \sim F_{p,q}$, then $\frac{pX/q}{1+pX/q} \sim$Beta$(p/2, q/2)$.
- The PDF of $X_{(j)}$ is $f_{X_{(j)}} = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}$.
- The joint PDF of $X_{(i)}$ and $X_{(j)}$ is

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v)[F_X(u)]^{i-1}[F_X(v)-F_X(u)]^{j-1-i}[1-F_X(v)]^{n-j}.$$

1.6 Convergence

Return to Table of Contents

- An estimator a is **consistent** if $a \xrightarrow{\text{p}} \mathbb{E}(X)$.
- A sequence of RVs $X_1, X_2, \ldots$ **converges in probability** to RV X if for every $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$, or $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$, denoted as $X_n \xrightarrow{\text{p}} X$.
- **Weak law of large numbers**: Let $X_1, X_2, \ldots, X_n$ be iid RVs with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then, $\bar{X}_n \xrightarrow{\text{p}} \mu$.
- X_n **converges in distribution to** X if $F_{X_n}(x) \to F_X(x)$ for all x where $F_X(x)$ is continuous.
  - Equivalent to showing $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for some bounded and continuous f.
  - If $M_{X_n}(t) \to M_X(t)$ for all t in an open neighborhood with 0, then $X_n \xrightarrow{\text{d}} X$.
  - If $X_n \xrightarrow{\text{p}} X$, then $X_n \xrightarrow{\text{d}} X$.
  - $X_n \xrightarrow{\text{p}} \mu$ iff $X_n \xrightarrow{\text{d}} \mu$.
- **Central limit theorem**, or **CLT**: Suppose $X_1, X_2, \ldots \overset{\text{iid}}{\sim} D(\mu, \sigma^2)$, where $\sigma^2 < \infty$. Then, $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$.
- **First-order delta method**: Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$. If $g'(\mu) \neq 0$, then $\sqrt{n}\left(g(X_n) - g(\mu)\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2[g'(\mu)]^2\right)$.
  - If $g'(\mu) = 0$ but $g''(\mu) \neq 0$, then $n\left[g(X_n) - g(\mu)\right] \xrightarrow{\text{d}} \frac{\sigma^2}{2} g''(\mu)\chi^2_1$.

  Example: Suppose visits to the NCSU statistics department website follows a Poisson process with rate equal to 0.5 visits/minute. Denote by X the time (in minutes) from the last visit to the nth ($n \geq 1$) visit.

  a. Show that X has a χ^2_m distribution, where $m = 2n$.

  b. Use the CLT to show that $\frac{X-m}{\sqrt{2m}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

  c. Use (b) to show that $\sqrt{2X} - \sqrt{2m} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

a. Let $N_t \sim$Pois(λt) represent the number of visits of from $[0,t]$. Also define N_1 as the probability of observing a single view.

$$1 - F_{T_1}(t) = P(T_1 > t) = P(N_t = 0) = \frac{e^{-t/2}(-t/2)^0}{(0)!} = e^{-t/2}, \ \therefore T_1 \sim \text{Exp}(2);$$

We assume independence for the times between visiting the website. X now becomes the sum of i.i.d. N_1's, which means that $X \sim$Gamma$(n,2) \equiv \chi^2_{2n} = \chi^2_m$.

b. Similarly to the previous part, let $Y_i \overset{\text{iid}}{\sim}$Exp(2), so $X = \sum_{i=1}^n Y_i$. By the CLT, $\sqrt{n}(\bar{Y} - 2) \xrightarrow{\text{d}} \mathcal{N}(0,4)$.

$$\begin{aligned}
\sqrt{n}(\bar{Y} - 2) &\xrightarrow{\text{d}} \mathcal{N}(0,4) \\
\frac{\sqrt{n}}{2}\left(\frac{1}{n}\sum_{i=1}^n Y_i - 2\right) &\xrightarrow{\text{d}} \mathcal{N}(0,1) \\
\frac{1}{2\sqrt{n}}(X - 2n) &\xrightarrow{\text{d}} \mathcal{N}(0,1) \\
\frac{1}{\sqrt{2m}}(X - m) &\xrightarrow{\text{d}} \mathcal{N}(0,1).
\end{aligned}$$

c.

$$\sqrt{2X} - \sqrt{2m} = \cdots = \sqrt{n}(\sqrt{2\bar{Y}} - \sqrt{2 \cdot 2});$$

From part b), we know that $\bar{Y}$ converges in distribution. Therefore, we will attempt the Delta Method. $g(x) = \sqrt{2X}$, so $g'(x) = \frac{1}{\sqrt{2X}}$, which at $\mu = \frac{1}{2}$, $\neq 0$. Therefore, by the Delta Method,

$$\begin{aligned}
\sqrt{n}(\sqrt{2\bar{Y}} - \sqrt{2 \cdot 2}) &\xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2[g'(\mu)]^2\right); \\
\sqrt{n}(\sqrt{2\bar{Y}} - \sqrt{2 \cdot 2}) &\xrightarrow{\text{d}} \mathcal{N}\left(0, 4\left[\frac{1}{\sqrt{2(2)}}\right]^2\right); \\
(\sqrt{2n\bar{Y}} - \sqrt{2 \cdot 2n}) &\xrightarrow{\text{d}} \mathcal{N}(0,1); \\
\sqrt{2X} - \sqrt{2m} &\xrightarrow{\text{d}} \mathcal{N}(0,1). \ \blacksquare
\end{aligned}$$

- **Slutsky's theorem**: Suppose $X_n \xrightarrow{\text{d}} X$, and $Y_n \xrightarrow{\text{p}} a$. Then,
  - $Y_n X_n \xrightarrow{\text{d}} aX$.
  - $X_n + Y_n \xrightarrow{\text{d}} X + a$.
  - $\frac{X_n}{Y_n} \xrightarrow{\text{d}} \frac{X}{a}$ if $a \neq 0$.
  - $\frac{Y_n}{X_n} \xrightarrow{\text{d}} \frac{a}{X}$ if $P(X = 0) = 0$.
- **Continuous mapping theorem**: Suppose g is a continuous function, and $X_n \xrightarrow{\text{d}} X$. Then, $g(X_n) \xrightarrow{\text{d}} g(X)$.

Suppose that $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$. Define $S_n = \sum_{i=1}^n U_i$ and $V_n \prod_{i=1}^n U_i$.

a. Find the PDF of V_n.
b. Determine $\mathbb{E}\left(\frac{U_1}{S_n}\right)$.
c. Show that $(V_n)^{-1/S_n} \xrightarrow{\text{p}} c$, and find c.
d. Compute $\lim_{n\to\infty} P\left(\frac{-\log(V_n)}{S_n} \geq 2\right)$.

a. Define $V_i := -\log(U_i)$, and $W := \sum_{i=1}^n V_i$.

$$\begin{aligned}
F_{V_i}(v) = P(V_i \leq v) &= P(-\log U_i \leq v) = 1 - F_{U_i}(e^{-v}) = 1 - e^{-v}; \\
f_{V_i}(v) &= e^{-v} \sim \text{Exp}(1), \text{ so } W \sim \text{Gamma}(n,1); \\
F_{V_n}(v) = P(V_n \leq v) &= P\left(\sum_{i=1}^n -\log(U_i) \leq -\log(v)\right) = F_W(-\log(v)); \\
f_{V_n}(v) &= f_W(-\log(v)) \cdot \frac{1}{v} \\
&= \frac{1}{\Gamma(n)(1)^n}(-\log(v))^{n-1} e^{-(-\log(v))/1} \cdot \frac{1}{v} \\
&= \frac{1}{\Gamma(n)}(-\log(v))^{n-1}(v)\frac{1}{v} = \frac{1}{\Gamma(n)}(-\log(v))^{n-1}.
\end{aligned}$$

b. Note that $\mathbb{E}\left(\frac{U_1}{S_n}\right) = \cdots = \mathbb{E}\left(\frac{U_n}{S_n}\right)$, since U_i are identically distributed. Define $X_i := \frac{U_i}{S_n}$.

$$\begin{aligned}\mathbb{E}(X_1) =& \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \frac{\mathbb{E}\left(\sum_{i=1}^n X_i\right)}{n} \\ =& \frac{\mathbb{E}\left(\frac{\sum_{i=1}^n U_i}{S_n}\right)}{n} = \frac{\mathbb{E}\left(\frac{S_n}{S_n}\right)}{n} = \frac{1}{n}.\end{aligned}$$

c. If $\log\left((V_n)^{-1/S_n}\right) \xrightarrow{\text{p}} a$, then by the continuous mapping theorem, $(V_n)^{-1/S_n} \xrightarrow{\text{p}} e^a$.

$$\begin{aligned}\log\left((V_n)^{-1/S_n}\right) =& \frac{-\frac{1}{n}\log(V_n)}{\frac{1}{n}S_n}; \\ -\frac{1}{n}\log(V_n) = -\frac{1}{n}\sum_{i=1}^n \log(U_i) &\xrightarrow{\text{p}} -E[\log(U_i)] = 1 \text{ by WLLN}; \\ \frac{1}{n}S_n \xrightarrow{\text{p}} \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2}\right) &= \frac{1}{2} \text{ also by WLLN}; \\ \frac{-\log(V_n)}{S_n} \xrightarrow{\text{d}} &\frac{1}{1/2} = 2;\end{aligned}$$

Since $\frac{-\log(V_n)}{S_n} \xrightarrow{\text{d}} c$ for some constant c, $\frac{-\log(V_n)}{S_n} \xrightarrow{\text{p}} c$ as well, so $\frac{-\log(V_n)}{S_n} \xrightarrow{\text{p}} 2$, so $c = 2$.

d. From the previous part, we found that $\frac{-\log(V_n)}{S_n} \xrightarrow{\text{p}} 2$. The probability statement is the definition of convergence in probability, so $\lim_{n\to\infty} P\left(\frac{-\log(V_n)}{S_n} \geq 2\right) = 1$. ■

2 ST 702: Statistical Theory II

Instructor: Dr. Ryan Martin
Semester: Spring 2024
Main Textbook: Casella and Berger, *Statistical Inference*

2.1 Consistency and Sufficiency

Return to Table of Contents

- An **estimator** of ϕ is a function $\hat{\phi}_n = \hat{\phi}(X^n)$ of our data.
- $\hat{\phi}_n$ is **consistent** for $\phi = \phi(\theta)$ if $\hat{\phi}_n \xrightarrow{\text{p}} \phi(\theta)$.
- $\hat{\phi}_n$ is $\boldsymbol{r_n}$-consistent for ϕ if $\lim_{n\to\infty} P(|\hat{\phi}_n - \phi(\theta)| > M_n r_n) = 0$, where $r_n \to 0$, and M_n is an arbitrary sequence where $M_n \to \infty$.
  - We often don't care about the precise values of M_n.
- A function $Q_\theta(X^n)$ is a **pivot** if its distribution doesn't depend on θ.
  - **Location-scale problems** exist in the form $X = \mu + \sigma Z$, where Z is a pivot.
- Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta$. A statistic $T(X^n)$ is **sufficient** if $(X^n | T(X^n) = t)$ is a pivot.
  - A statistic T is sufficient iff there exists functions g_θ and h such that $P_\theta(x^n) = g_\theta\{T(x^n)\}h(x^n)$.
  - Sufficient statistics are not unique.
  - If T is a vector, then the vector as a whole is sufficient for $\boldsymbol{\theta}$ (rather than individual elements of T being sufficient for individual elements of $\boldsymbol{\theta}$).
  - If P_θ is an exponential family, then $T = \sum_{i=1}^n X_i$ is sufficient.
- A statistic T is **minimal sufficient** if it is a function of every other sufficient statistic.
  - If T is sufficient and $\left[\frac{P_\theta(x^n)}{P_\theta(y^n)} \text{ is constant in } \theta \iff T(x^n) = T(y^n)\right]$, then T is minimal sufficient.
  - If P_θ is a full-rank exponential family, then $T = \sum_{i=1}^n X_i$ is minimal sufficient.
- A statistic U is **ancillary** if its distribution doesn't depend on θ.
- A statistic T is **complete** if $E_\theta[f(T)] = 0 \ \forall\theta \implies f \equiv 0$.
  - A complete statistic doesn't contain any ancillary features.
  - If T is complete and sufficient, then it is minimal sufficient.
  - If a minimal sufficient statistic exists, then complete statistics are also minimal sufficient.
  - If P_θ is a full-rank exponential family, then $T = \sum_{i=1}^n X_i$ is complete and sufficient.
- **Basu's Theorem**: If $T = T(X^n)$ is complete, sufficient and $U = U(X^n)$ is ancillary, then $T \perp U$.
  Example: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \theta^2)$, where $\theta > 0$. Calculate $\mathbb{E}\left(\frac{X_1^2}{\sum_{i=1}^n X_i^2}\right)$.

  Define $N := X_1^2$ and $D := \sum_{i=1}^n X_i^2$. If we can show that $\frac{N}{D}$ is ancillary, and that D is complete, then by Basu's theorem,

$$\mathbb{E}_\theta(N) = \mathbb{E}_\theta\left(\frac{N}{D} \cdot D\right) = \mathbb{E}_\theta\left(\frac{N}{D}\right)\mathbb{E}_\theta(D) \longrightarrow \mathbb{E}_\theta\left(\frac{N}{D}\right) = \frac{\mathbb{E}_\theta(N)}{\mathbb{E}_\theta(D)}.$$

  N/D is ancillary: use location-scale family. $X_i = \theta Z_i$ for $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

$$\frac{N}{D} = \frac{X_1^2}{\sum_{i=1}^n X_i^2} = \frac{\theta^2 Z_1}{\theta^2 \sum_{i=1}^n Z_i^2} = \frac{Z_1}{\sum_{i=1}^n Z_i^2};$$

  Since this distribution doesn't depend on θ, N/D must be ancillary.
  D is complete; since P_θ is a full-rank exponential family, this follows naturally.
  Therefore, by Basu's Theorem,

$$\mathbb{E}_\theta\left(\frac{N}{D}\right) = \frac{\mathbb{E}_\theta(N)}{\mathbb{E}_\theta(D)} = \frac{\theta^2}{n\theta^2} = \frac{1}{n}. \ \blacksquare$$

- **Regularity conditions**:
  - $\theta \to P_\theta(x)$ is differentiable for all x.
  - $\theta \to \int g(x)P_\theta(x)dx$ can be differentiated under the integral sign.
  - Support of P_θ does not depend on θ.
- The **score function** of $X \sim P(\theta)$ is $S_X(\theta) = \frac{\partial}{\partial\theta} \log P_\theta(X)$.
  - $\mathbb{E}_\theta[S_X(\theta)] = 0$ for all θ.
- The **Fisher information** of $X \sim p_\theta$ is $I_X(\theta) = Var_\theta[S_X(\theta)] = \mathbb{E}_\theta[S_X(\theta)^2]$.
  - $I_{X^n}(\theta) = nI_{X_i}(\theta)$.
  - $I_{T(X^n)}(\theta) = I_X(\theta)$ if T is sufficient.
  - If P_θ is exponential family, then $I_X(\theta) = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}S_\theta(X)\right]$.
  - $I_{X^n}(\theta) = \mathbb{E}_\theta[I_{W|U}(\theta)]$, where U is ancillary, and W is not necessarily complete.
  - The **observed Fisher information** at θ^* is $J_n(\vartheta) = -\frac{\partial^2}{\partial\vartheta^2}\ell(\vartheta)\big|_{\vartheta=\theta^*}$.
- **Cramer-Rao lower bound**, or **CRLB**: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta$, and that the regularity conditions hold. Also suppose that T is an unbiased estimator of ϕ. Then, $Var_\theta[T(X^n)] \geq \frac{\dot{\phi}(\theta)^2}{I_{X^n}(\theta)}$.
  - **Attainment theorem**: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, where $f(x|\theta)$ satisfies the regularity conditions. If $W(X^n)$ is an unbiased estimator of $\tau(\theta)$, then $W(X^n)$ attains the CRLB iff $a(\theta)[W(x^n) - \tau(\theta)] = \frac{\partial}{\partial\theta}\ell_n(\vartheta)$ for some $a(\theta)$.
- **Sufficiency principle**: If two datasets have the same minimal sufficient statistics, then the same inferences for θ should be drawn.
- **Conditionality principle**: Experiments that were not performed are not relevant to statistical analysis, and should be ignored.
- **Likelihood principle**: Formed from sufficiency and conditionality principles.

2.2 Estimation

Return to Table of Contents

- **Method of moments estimation**, or **MOM**, uses moments to estimate θ using our data.
  - $\theta := (g(\mu_1), \ldots, g(\mu_k))$ for some g, where μ_i is the ith moment.
    - If g is continuous, then MOM is consistent.
    - If g is differentiable, then MOM is $n^{1/2}$-consistent.
- The **likelihood function** is $L_n(\vartheta) = P_\vartheta(X^n) = \prod_{i=1}^n P_\vartheta(X_i)$.
  - The **log-likelihood** is $\ell_n(\vartheta) = \log L_n(\vartheta)$.
- **Asymptotic efficiency conditions**:
  - The support of P_θ doesn't depend on θ.
  - $P_\theta(x)$ is twice continuously differentiable in θ for most of x.
  - We can interchange expectations and derivatives w.r.t. $P_\theta(x)$.
  - We can use Taylor approximations with low error.
- The **maximum likelihood estimate**, or **MLE**, is $\hat{\theta}_n = \hat{\theta}_{\text{MLE}} = \arg\max_\vartheta L_n(\vartheta) = \arg\max_\vartheta \ell_n(\vartheta)$.
  - Might not be unique.
  - The MLE of $\phi(\theta)$ is $\phi(\hat{\theta}_n)$.
  - If $L_n(\vartheta)$ is smooth, then we can find MLE with calculus.
    - Be sure to verify calculated MLE is a maximum by taking the second derivative.
  - The **likelihood equation** is $\nabla\ell_n(\theta) = 0$.

- MLEs are consistent (under efficiency conditions); that is, $g(\hat{\theta}_n) \xrightarrow{\text{p}} g(\theta)$ for continuous g.
- MLEs are asymptotically Normal, unbiased, and efficient. That is, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{d}} \mathcal{N}(0, I(\theta)^{-1})$, and $Var(\hat{\theta}_n)$ achieves the CRLB.
  - If $\phi(\theta)$ is smooth, then $\sqrt{n}(\hat{\phi}_n - \phi) \xrightarrow{\text{d}} \mathcal{N}(0, \underbrace{\dot{\phi}(\theta)^T I(\theta)^{-1} \dot{\phi}(\theta)}_{=:V^\phi(\theta)})$.
    - We would need to estimate $V^\phi(\theta)$. Directly using the MLE is volatile for small n. Bootstrapping is okay. Could also use $\hat{V}_n^\phi \overset{\text{set}}{=} \hat{\dot{\phi}}_n' [J_n(\hat{\theta}_n)]^{-1} \hat{\dot{\phi}}_n$, which accommodates conditioning on an ancillary statistic.
- MLE does not satisfy the likelihood principle.

2.3 Hypothesis Tests and CIs

Return to Table of Contents

- A **point-null hypothesis**, or **simple hypothesis**, is where $H_0 : \theta = \theta_0$, where $\theta_0 \in \mathbb{R}$.
- A **composite hypothesis** is where $H_0 : \theta \in \Theta_0$.
- A ***p*-value function** is $p(x^n) = P_\vartheta\{T_\vartheta(X^n) \geq T_\vartheta(x^n)\}$, where large $T_\vartheta(x^n)$ signifies incompatibility between θ and x^n.
  - $T_\vartheta(x^n)$ is a constant, whereas $T_\vartheta(X^n)$ is an RV.
  - p-value functions measure **plausibility**, which low values indicate we should reject H_o.
  - $\sup_{\theta \in \Theta_0} P_\theta\{p_{\Theta_0}(X) \leq \alpha\} \leq \sup_{\theta_0 \in \Theta_0} P_{\theta_0}\{p_{\theta_0}(X) \leq \alpha\} = \alpha$.
  - $p_A(x^n) = \sup_{\vartheta \in A} p_\vartheta(x^n)$ for $A \subseteq \Theta_0$.
- A **hypothesis test** is a function $\delta : \zeta \to \{0, 1\}$ such that $\delta(X^n) = \begin{cases} 1, \text{ reject } H_0 \\ 0, \text{ fail to reject } H_0 \end{cases}$.
  - The **size** of a test δ is $\sup_{\theta \in A} P_\theta\{\delta(X^n) = 1\} = \sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.
    - A **level-α test** satisfies $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.
- The **power** of test at θ is denoted as $\beta(\theta) = P_\theta(X \in RR)$.
- Type I Errors and Type II Errors compete against one another, so we have to impose constraints.
- A **set estimator** of ϕ is a function $C : \zeta \to 2^{\phi(\Theta)}$, where $2^{\phi(\Theta)}$ is a subset of $\phi(\Theta)$.
  - Values in $C(X^n)$ are plausible values based on our data.
- A set estimator is a **$100(1-\alpha)\%$ confidence set** if the coverage probability is at least $1 - \alpha$.
  - If $\delta_{\theta_0}^\alpha$ is a size-α test of $H_0 : \theta = \theta_0$, then $C^\alpha(X^n) = \{\theta_0 : \delta_{\theta_0}^\alpha(X^n) = 0\}$ is a $100(1-\alpha)\%$ confidence set.
  - A **uniformly most accurate confidence set**, or **UMA confidence set**, is a $100(1-\alpha)\%$ confidence set that minimizes the probability of false coverage, compared to all other $100(1-\alpha)\%$ confidence sets.
    - A UMP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ yields a UMA lower confidence bound.
  - A $100(1-\alpha)\%$ confidence set $C(x^n)$ is **unbiased** if $P_\theta(\theta' \in C(X^n)) \leq 1 - \alpha$ for all $\theta' \neq \theta$.
- The **relative likelihood** is $\lambda(x^n) = \frac{L_n(\vartheta)}{L_n(\hat{\theta}_n)}$.
  - The **likelihood ratio test**, or **LRT**, rejects $H_0 : \theta \in \Theta_0$ iff $\lambda(x^n)$ is small.
    - $\lambda^*(T(x^n)) = \lambda(x^n)$, where λ^* is the LRT based on sufficient statistic T.
  - $\lambda(x^n)$ is a function of a minimal sufficient statistic.
  - A p-value function for $\lambda(x^n)$ would be $P_\vartheta\{\lambda(X^n) \leq \lambda(x^n)\}$.
    - Recall that incompatibility is measured by small values of $\lambda(x^n)$, which is why we use $\geq$ instead of $\leq$.
  - **Wilk's Theorem**: $-2 \log \lambda(X^n) \xrightarrow{\text{d}} \chi^2_{\dim(\theta)}$.
    - $\lambda(x^n)$ is an approximate pivot!

∗ Using the p-value function, we would reject H_0 if $-2\log\lambda(\Theta) > \chi^2_{1-\alpha,p}$.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(\theta_1, \theta_2)$, where $-\infty < \theta_1 < \theta_2 < \infty$.

a. Find the asymptotic limit of $n^{-1}\log\left(\prod_{i=1}^n U_i\right)$, where $U_i = \frac{(X_i - \theta_1)}{\theta_2 - \theta_1}$.
b. Find the MLE of (θ_1, θ_2).
c. Show that $n(X_{(1)} - \theta_1)$ converges in distribution. Find this limiting distribution.
d. Find the MLE of (θ_1, θ_2) under $H_0 : \theta_1 = -\theta_2$.
e. Derive the LRT for H_0.

a.

$$\begin{aligned} n^{-1}\log\left(\prod_{i=1}^n U_i\right) =& \frac{1}{n}\sum_{i=1}^n \log(U_i) \\ \overset{\text{i.i.d.}}{=}& \frac{1}{n}\left[n\log U_1\right] = \log U_1 \xrightarrow{\text{p}} \mathbb{E}[\log U_1]; \\ \mathbb{E}[\log U_1] =& \int_0^1 \log u du = -1. \end{aligned}$$

b.

$$\begin{aligned} L_n(\theta_1, \theta_2) =& \prod_{i=1}^n \frac{1}{\theta_2 - \theta_1} = (\theta_2 - \theta_1)^{-n}\mathbb{I}(\theta_1 < X_i < \theta_2) \\ =& (\theta_2 - \theta_1)^{-n}\mathbb{I}(X_{(1)} > \theta_1)\mathbb{I}(X_{(n)} < \theta_2); \end{aligned}$$

$L_n(\theta_1, \theta_2)$ is larger when θ_1 is closer to θ_2. From this, the MLE of (θ_1, θ_2) is $(X_{(1)}, X_{(n)})$.

c. Let $Y_n = n(X_{(1)} - \theta_1)$; Note that $X_i - \theta_1 \sim U(0, \theta_2 - \theta_1)$.

$$\begin{aligned} F_{Y_n}(y) = P\left(n(X_{(1)} - \theta_1) \le y\right) =& P\left(X_{(1)} - \theta_1 \le \frac{y}{n}\right) = 1 - P\left(X_{(1)} - \theta_1 \ge \frac{y}{n}\right) \\ =& 1 - \left[P\left(X_i - \theta_1 \ge \frac{y}{n}\right)\right]^n = 1 - \left[1 - P\left(X_i - \theta_1 \le \frac{y}{n}\right)\right]^n \\ =& 1 - \left[1 - F_{X_i}\left(\frac{y}{n}\right)\right]^n = 1 - \left[1 - \frac{y/n}{\theta_2 - \theta_1}\right]^n \\ =& 1 - \left[1 - \frac{1}{n}(y(\theta_2 - \theta_1)^{-1})\right]^n; \end{aligned}$$

As $n \to \infty$, $1 - \left[1 - \frac{1}{n}(y(\theta_2 - \theta_1)^{-1})\right]^n \to 1 - \exp\left\{y(\theta_2 - \theta_1)^{-1}\right\}$.

d.

$$\begin{aligned} L_n(\theta_1, \theta_2) =& (\theta_2 - \theta_1)^{-n}\mathbb{I}(X_{(1)} > \theta_1)\mathbb{I}(X_{(n)} < \theta_2) \\ \overset{H_0}{=}& (\theta_2 + \theta_2)^{-n}\mathbb{I}(X_{(1)} > -\theta_2)\mathbb{I}(X_{(n)} < \theta_2) \\ =& (2\theta_2)^{-n}\mathbb{I}(-X_{(1)} < \theta_2)\mathbb{I}(X_{(n)} < \theta_2) \\ =& (2\theta_2)^{-n}\mathbb{I}(\theta_2 > \max(-X_{(1)}, X_{(n)})); \end{aligned}$$

This means that $\hat{\theta}_2 = \max(-X_{(1)}, X_{(n)})$.

e.

$$\begin{aligned} \lambda(x^n) =& \frac{L_n((\hat{\theta}_2)_{H_0}, (\hat{\theta}_2)_{H_0})}{L_n(\hat{\theta}_1, \hat{\theta}_2)} = \frac{(2(\hat{\theta}_2)_{H_0})^{-n}\mathbb{I}((\hat{\theta}_2)_{H_0} > \max(-X_{(1)}, X_{(n)}))}{(\hat{\theta}_2 - \hat{\theta}_1)^{-n}\mathbb{I}(X_{(1)} > \hat{\theta}_1)\mathbb{I}(X_{(n)} < \hat{\theta}_2)} \\ =& \left(\frac{X_{(n)} - X_{(1)}}{2 \cdot \max(-X_{(1)}, X_{(n)})}\right)^n = \begin{cases} \left(\frac{X_{(n)} - X_{(1)}}{2\cdot(-X_{(1)})}\right)^n, & -X_{(1)} < X_{(n)} \\ \left(\frac{X_{(n)} - X_{(1)}}{2\cdot X_{(n)}}\right)^n, & \text{o.w.} \end{cases} = \begin{cases} \left(\frac{1}{2} - \frac{X_{(n)}}{2X_{(1)}}\right)^n, & -X_{(1)} < X_{(n)} \\ \left(\frac{1}{2} - \frac{X_{(1)}}{2X_{(n)}}\right)^n, & \text{o.w.} \end{cases} \blacksquare \end{aligned}$$

- A **Wald pivot** is $\frac{\hat{\phi} - \phi}{\sqrt{Var(\hat{\phi})}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.
  - If $\hat{\phi} = \hat{\phi}_n$, then $Var(\hat{\phi}_n) \equiv \frac{1}{J_n(\vartheta)}$.
- A **score pivot** is $\frac{S_\theta(X^n)}{I_{X^n}(\theta)} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.
- The **loss function** is $L(\theta, a)$, where a is some action.
  - The **action space** is $\mathbb{A}$.

- A **decision rule** is $\delta : X^n \to \mathbb{A}$.
  - Minimize expected loss with the **risk function** $R(\theta, \delta) = \mathbb{E}_\theta \{L(\theta, \delta(X^n))\}$.
    - **MSE** is a risk function, where $MSE_\delta = \mathbb{E}_\theta\{(\delta(X^n) - \theta)^2\} = [\theta - \mathbb{E}_\theta(\delta)]^2 + Var(\delta)$.
  - A decision rule is **inadmissable** if there exists another decision rule $\tilde{\delta}$ such that $R(\theta, \tilde{\delta}) \leq R(\theta, \delta)$ for all θ.
    - If δ is inadmissible due to $\tilde{\delta}$, then $\tilde{\delta}$ **dominates** δ.
- **Rao-Blackwell theorem**: Suppose we have a convex loss function, and T is sufficient. Then, $\delta^{RB} = \mathbb{E}\{\delta(X^n)|T\}$ dominates δ.
  - If δ is unbiased, then δ^{RB} is also unbiased, but with smaller variance.
- δ is the **uniformly minimum variance unbiased estimator**, or **UMVUE**, for ϕ if δ is unbiased and $Var_\theta(\delta) < Var_\theta(\tilde{\delta})$ for all θ and unbiased $\tilde{\delta}$.
  - **Lehmann-Scheffe**: If T is complete and sufficient, then $h(T)$ is the UMVUE for ϕ.
  - If an unbiased estimator exists, then Rao-Blackwell guarantees the UMVUE exists.

  Example: Let $\phi(\theta) = \theta(\theta + 1)$, where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Pois(\theta)$. Find the UMVUE of ϕ.

  Since P_θ is an exponential family, then $T = \sum_{i=1}^n X_i$ is complete and sufficient. Choose $\delta(X^n) = \frac{1}{n}\sum_{i=1}^n X_i^2$ as a naive estimator.

$$\begin{aligned}
P_\theta\{X_1 = x_1 | T = t\} =& \frac{P_\theta\{X_1 = x_1, T = t\}}{P_\theta(T = t)} \\
=& \frac{P_\theta\{X_1 = x_1, \sum_{i=2}^n X_i = t - x_1\}}{P_\theta(T = t)} \\
=& \frac{P_\theta\{X_1 = x_1\} P_\theta\{\sum_{i=2}^n X_i = t - x_1\}}{P_\theta(T = t)} \\
=& \frac{\left[\frac{e^{-\theta}\theta^{x_1}}{x_1!}\right]\left[\frac{e^{-(n-1)\theta}[(n-1)\theta]^{t-x_1}}{(t-x_1)!}\right]}{\left[\frac{e^{-n\theta}(n\theta)^t}{(t)!}\right]} \\
=& \frac{t!}{x_1!(t-x_1)!}\left(\frac{1}{n}\right)^{x_1}\left(1 - \frac{1}{n}\right)^{t-x_1} \sim \text{Bin}\left(t, \frac{1}{n}\right).
\end{aligned}$$

  Therefore, $\delta^{RB}(T) = \mathbb{E}(X_1^2|T) = Var(X_1^2|T) + \mathbb{E}(X_1|T)^2 = \frac{T}{n}\left(1 - \frac{1}{n}\right) + \left(\frac{T}{n}\right)^2$. ■
- A **test** maps the sample space to an action. In other words, $\delta : \zeta \to \{0, 1\}$.
  - Using 0-1 loss, the risk function is $R(\theta, \delta) = \begin{cases} P_\theta\{\delta(X^n) = 1\}, \theta \in \Theta_0 \text{ (Type I Error)} \\ P_\theta\{\delta(X^n) = 0\}, \theta \notin \Theta_0 \text{ (Type II Error)} \end{cases}$.
    - There is no δ that globally minimizes risk. We must impose a constraint, which is often by controlling the Type I Error rate by setting it to be $\alpha \in (0, 1)$.
- $\beta^*(\theta)$ is a **uniformly most powerful test**, or **UMP**, at size α if $\beta^*(\theta) \geq \beta(\theta)$ for all $\theta \in \Theta_0^c$ and $\beta(\theta)$ that is a power function with the same level.
  - Suppose T is sufficient, and $g(t|\theta_i)$ is the PDF of T w.r.t θ_i for $i \in \{0, 1\}$. Any test based on T is a UMP level-α test if it satisfies $t \in RR$ if $g(t|\theta_1) > k \cdot g(t|\theta_0)$, $t \in RR^c$ if $g(t|\theta_1) < k \cdot g(t|\theta_0)$ for some nonnegative k, and the test is size-α.
  - A **uniformly most powerful and unbiased test**, or **UMPU test**, is a UMP test within the class of unbiased tests.
- A model P_θ has the **monotone likelihood ratio property**, or **MLR property**, w.r.t. statistic T if $t \to \frac{g_{\theta_1}(t)}{g_{\theta_0}(t)}$ is monotone for any θ_0, θ_1.
- **Neyman-Pearson lemma**: Given $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, there exists a unique UMP size α test $\delta_\alpha^*(X^n) = \mathbb{I}\left(\frac{L_n(\theta_1)}{L_n(\theta_0)} > k_\alpha\right)$, where k_α is such that $P_{\theta_0}\left(\frac{L_n(\theta_1)}{L_n(\theta_0)} > k_\alpha\right) = \alpha$.
  - $\delta_\alpha^*(X^n)$ is an indicator of the rejection region.
  - **Karlin-Rubin theorem**: If a model has the MLR property, then the Neyman-Pearson test that is UMP for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ is also UMP for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$ (or $H_1 : \theta < \theta_1$).
  - There is no UMP test for two-sided alternatives without constraints.

2.4 Introduction to Bayesian Inference

Return to Table of Contents

- The axiom of Bayesian statistics is that all uncertanties are quantified with probability.
  - Unknown parameters are treated as RVs.
- A **sampling distribution** is $f(x|\Theta)$.
- The **prior distribution** is $\pi(\theta)$.
- The **posterior distribution** is $\pi(\theta|X) = f(x|\Theta)\pi(\theta)$.
- A **conjugate distribution** is when the prior's distribution is the same as the posterior.
- We can estimate ϕ with $\hat{\phi}_{Bayes} = \mathbb{E}_{\pi(\theta|X)}[\phi]$.
  - With squared error loss, the risk of the Bayes estimator is the expected value of the posterior.
- A $\mathbf{100(1-\alpha)\%}$ **credible interval** defines bounds l and u such that $Q_n(l \leq \Theta \leq u) = 1 - \alpha$.
- If we have a known prior distribution, then the likelihood principle is satisfied.
  - If we don't know the prior, we can use **Jeffrey's prior** $q_J(\theta) \propto \sqrt{\det |I(\theta)|}$.

3 ST 703: Statistical Methods I

Instructor: Dr. Jacqueline Hughes-Oliver
Semester: Fall 2023
Main Textbook: Rao, *Statistical Research Methods in the Life Sciences*

3.1 Hypothesis Tests and CIs

Return to Table of Contents

- **Satterthwaite's approximation for ν**: $\nu \approx \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$.
- The **power** of a test is $\text{Power}(\theta) = P(\text{reject } H_o|\theta)$.
- The **significance level**, denoted as $\boldsymbol{\alpha}$, is $\alpha = \sup_{\theta\in\Theta_0}\text{Power}(\theta)$.
- Confidence intervals for μ: Suppose $Y_i \overset{\text{iid}}{\sim} D$.
  - If σ is known:
    - If D is Normal, then use $\bar{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.
    - If n is large, then use $\bar{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ (approximate).
    - Otherwise, use nonparametric methods.
  - If σ is unknown:
    - If D is Normal, then use $\bar{y} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$.
    - If n is large, then use $\bar{y} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ (approximate).
    - Otherwise, use nonparametric methods.
  - Decreasing n or α, or using t instead of z, results in narrower intervals.
- Hypothesis tests for μ:
  - σ known: $Z = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
  - σ unknown, but large sample size: Approximate version of above case.
  - σ unknown, Normality: $T = \frac{\bar{Y}-\mu_0}{S/\sqrt{n}} \sim t_\alpha$.
- Confidence intervals for p:
  - **Wald** CI: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
    - MOE=0 when $\hat{p} = 0$ or 1.
    - Interval can include values outside $[0,1]$.
    - Has erratic coverage probabilities.
  - **Wilson** CI: $\frac{\hat{p}+z_{\alpha/2}^2/(2n)}{1+z_{\alpha/2}^2/n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})/n+z_{\alpha/2}^2/(4n^2)}{1+z_{\alpha/2}^2/n}}$.
  - **Agresti-Coull** CI: $\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$, where $\tilde{p} = \frac{\hat{X}+z_{\alpha/2}^2/2}{\tilde{n}}$, and $\tilde{n} = n + z_{\alpha/2}^2$.
  - **Clopper-Pearson** CI: $\begin{cases} \left[0, 1-(\alpha/2)^{1/n}\right), & x = 0 \\ \left((\alpha/2)^{1/n}, 1\right], & x = n \end{cases}$.
- Hypothesis tests for p:
  - Large-sample approximate Rao test:
    - $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim \mathcal{N}(0,1)$.

    Let $p_a = \sqrt{\frac{p_0(1-p_0)}{n}}$ and $p_b = \sqrt{\frac{p'(1-p')}{n}}$.

H_1	Power (p')	Sample size needed
$p > p_0$	$1 - \Phi\left(\frac{p_0-p'+z_\alpha p_a}{p_b}\right)$	$\left[\frac{z_\alpha p_a+z_\beta p_b}{p'-p_0}\right]^2$
$p < p_0$	$\Phi\left(\frac{p_0-p'-z_\alpha p_a}{p_b}\right)$	$\left[\frac{z_\alpha p_a+z_\beta p_b}{p'-p_0}\right]^2$
$p \neq p_0$	$1 - \Phi\left(\frac{p_0-p'+z_{\alpha/2} p_a}{p_b}\right) + \Phi\left(\frac{p_0-p'-z_{\alpha/2} p_a}{p_b}\right)$	$\left[\frac{z_{\alpha/2} p_a+z_\beta p_b}{p'-p_0}\right]^2$

- Confidence intervals for $\mu_1 - \mu_2$: Suppose $Y_{i1} \overset{\text{iid}}{\sim} D_1$, and $Y_{j2} \overset{\text{iid}}{\sim} D_2$.
  - If D_1, D_2 are Normal, and σ_1 and σ_2 are both unknown, then use $(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \sim t_\nu$, where ν is approximated using Satterthwaite's approximation.
  - If D_1, D_2 are not Normal, but both sample sizes are large, then use $(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \sim \mathcal{N}(0,1)$ (approximate).
  - If D_1, D_2 are Normal, and $\sigma_1 = \sigma_2$ ($S_1 \approx S_2$) are both unknown, then $(\bar{y}_1 - \bar{y}_2) \pm \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sim \mathcal{N}(0,1)$.
- Hypothesis test for $\mu_1 - \mu_2$:
  - σ_1, σ_2 known: $Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$.
  - σ_1, σ_2 unknown, but large samples: Approximate version of above case.
  - σ_1, σ_2 unknown, Normality:
    - $T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$.
    - Power and sample size done computationally.
  - $\sigma_1 = \sigma_2$ unknown, Normality:
    - $T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$.
    - Power and sample size done computationally.
- Confidence intervals for $p_1 - p_2$:
  - Paired data: $\frac{B-C}{n} \pm z_{\alpha/2}\frac{\sqrt{B+C-\frac{1}{n}(B-C)^2}}{n}$.
    - B is the number of observations where the first trial is a success, and the second a failure.
    - C is the number of observations where the second trial is a success, and the first a failure.
  - Non-paired data: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.
- Hypothesis tests for $p_1 - p_2$:
  - Independent data, $\Delta_0 \neq 0$:
    - $Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \mathcal{N}(0,1)$.
    - Power and sample size done computationally.
  - Independent data, $\Delta_0 = 0$:
    - $Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1)$, where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$.
    - Power and sample size done computationally.
  - Paired data:
    - $Z = \frac{B-C-\Delta_0}{\sqrt{\frac{B+C-n\Delta_0^2}{n}}} \overset{\Delta_0=0}{=} \frac{B-C}{\sqrt{B+C}} \sim \mathcal{N}(0,1)$.
    - Power and sample size done computationally.

3.2 ANOVA Model

Return to Table of Contents

- **ANOVA models** compare values of means across different groups.
- The **2-sample pooled *t*-test** is the simplest ANOVA model.
  - Used to compare the means from two independent Normal samples.
  - Test statistic is $T = \frac{\bar{y}_{1+} - \bar{y}_{2+}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$, where $s_p^2 = \frac{\left[\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1+})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2+})^2\right]}{n_1+n_2-2}$.
    - Can also use $T^2 \sim F_{1,n_1+n_2-2}$.

- Extending the 2-sample pooled t-test to p groups:
  - Used to compare the means from p independent Normal samples.
  - Test statistic is $F = \frac{\sum_{i=1}^{p} n_i(\bar{y}_{i+}-\bar{y}_{++})^2}{(p-1)s_p^2} \sim F_{p-1,\sum_{i=1}^{p} n_i - p}$, where $s_p^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i+})^2}{\sum_{i=1}^{p} n_i - p}$.

Source	*df*	SSq
Model	$p-1$	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{i+}-\bar{y}_{++})^2$
Error	$n-p$	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i+})^2$
Total	$n-1$	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{++})^2$

- The **one-way ANOVA model**, or **one-way classification model**, is in the form $Y_{ij} = \mu + \tau_i + E_{ij}$, where $E_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, and $\tau_i = \mu_i - \mu$.
  - Omit one column so X is full-rank, the τ_i that is eliminated is the **reference group**.
    - We now estimate $\mu + \tau_i$ and $\tau_j - \tau_i$ instead of μ, τ_j.

3.3 Multiple Comparisons

Return to Table of Contents

- $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T\right)$.
  - If $\boldsymbol{A}$ is only one row, then $t = \frac{\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m}}{\sqrt{MSE\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T}} \sim t_{df_E}$.
  - If $\boldsymbol{A}$ has $k > 1$ independent rows, then $F = \frac{Q/k}{MSE} \sim F_{k,df_E}$.
    - $Q = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m})^T\left[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m}) = SSE_R - SSE_F$.
    - Can only test for $H_1 : \boldsymbol{A}\boldsymbol{\beta} \neq \boldsymbol{m}$.
    - Simultaneously tests k linear hypotheses (not one-at-a-time).
- **Completely randomized design**, or **CRD**, assigns n_i units to the ith treatment, where $i = 1, \ldots, t$, and t is fixed.
  - **Balanced CRD** is when $n_1 = \cdots = n_p \equiv n$, so $N = nt$.
- A **contrast of means** for linear combination $\theta = \sum_{i=1}^{p} c_i\mu_i$ is when $\sum_{i=1}^{p} c_i = 0$.
  - The **contrast sum of squares** for a single contrast is $SS(\hat{\theta}) = \frac{\hat{\theta}}{Var(\hat{\theta})} = \frac{\left(\sum_{i=1}^{p} c_i\hat{\mu}_i\right)^2}{\sum_{i=1}^{p}\frac{c_i^2}{n_i}}$.
  - $F = \frac{SS(\hat{\theta})}{MSE} \sim F_{1,df_E}$ lets us test for a single contrast.
  - Two contrasts are **orthogonal** if $\sum_{i=1}^{p}\frac{c_i d_i}{n_i} = 0$.
  - Under the one-way classification model, there exists a set of $p-1$ mutually orthogonal contrasts such that $SSR = \sum_{i=1}^{p-1} SS(\theta_i)$.
- **Scheffe**: compare $|\hat{\theta}|$ to $SE(\hat{\theta})\sqrt{(p-1)F_{(p-1),df_E}}$.
  - Is very conservative (can result in low power).
  - FWE $\leq \alpha$.
  - Can investigate any number of linear hypotheses (doesn't depend on s, which is the number of tests).
- **Fisher**: compare p-values to α.
  - Is too lenient when $s > 1$.
- **Bonferroni**: compare $|\hat{\theta}|$ to $t_{\alpha/(2s),df_E} \cdot SE(\hat{\theta})$.
  - Could also compare p-values to $\frac{\alpha}{s}$.
  - Also controls $FWE \leq \alpha$.

- **Tukey-Kramer**: compare $|\hat{\theta}|$ to $q_{t,df_E,\alpha}\frac{SE(\hat{\theta})}{\sqrt{2}}$.
  - Only useful for pairwise comparisons.
  - $Q_{t,\nu} = \frac{W_{(t)}-W_{(1)}}{\hat{\sigma}_\nu}$, where $W_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- A **simultaneous confidence coefficient** is a set of k CIs such that the probability that all of the intervals contain the true values is $1-\alpha$.
  - Can convert our rejection regions defined as $|\hat{\theta}_j| > a \cdot SE(\hat{\theta})$ into $\hat{\theta} \pm a \cdot SE(\hat{\theta}_j)$.
- When s is large, shift towards accounting for **false discovery rate**, or **FDR**, which is $P\left(\frac{\text{falsely reject } H_0}{\text{reject } H_0}\right)$.
  - **Benjamini-Hochberg**: reject each test where p-value$\leq \max\left\{p_{(j)} : p_{(j)} \leq \alpha\frac{j}{k}, 1 \leq j \leq k\right\}$.
- **Unadjusted means** do not account for the value of the covariate within each group.
- **Adjusted means** are estimated mean responses at a common reference value of the covariates.
  - Assumes the covariate term does not interact with the main effects.
- An **ANCOVA model** has the form $Y = \mu_e(x_1, \ldots, x_r) + \mu_c(z_1, \ldots, z_s) + E$, where $E \sim \mathcal{N}(0, \sigma^2)$.
  - The estimated adjusted mean response at $(x_1, \ldots, x_r)$ is $\hat{\mu}_e(x_1, \ldots, x_r) + \hat{\mu}_c(z_1, \ldots, z_s)$.
- **Lack-of-fit testing** tests how a model compares to the most complicated model possible.
  - Very similar to a nested F-test.
  - $F = \frac{(SSE_R - SSE_{\text{pure error}})}{(t-1-q)MSE_{\text{pure error}}} \sim F_{t-1-q, df_{E_{\text{pure error}}}}$, where q is the order of the model.
- Sample sizes needed to detect $1-\beta \leq \text{Power}\sum_{i=1}^p \tau_i^2$ is $1-\beta \overset{\text{set}}{\leq} P\left(F_{t-1,N-t}(\gamma) > F_{t-1,N-t,\alpha} | \sum_{i=1}^p \tau_i^2\right)$ (assuming equal sample sizes).
  - $\gamma = \frac{1}{\sigma^2}\sum_{i=1}^p \tau_i^2 n_i$ is the ncp.
  - Power increases as ncp and/or sample size increases, and as the variance decreases.
- **Randomized complete blocked design**, or **RCBD**, uses $N = rt$ units that are divided into r blocks of t units each.
  - Eliminates the effect of confounding factors in studies.
  - Model is $Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$.
  - Has smaller MSE and df_E than one-way ANOVA.
  - Assumes no interactions between blocks and treatments.

Source	*df*	SSq
Treatment	$t-1$	$\sum_{i=1}^t \sum_{j=1}^r (\bar{y}_{i+} - \bar{y}_{++})^2$
Block	$r-1$	$\sum_{i=1}^t \sum_{j=1}^r (\bar{y}_{+j} - \bar{y}_{++})^2$
Error	$(t-1)(r-1)$	$\sum_{i=1}^t \sum_{j=1}^r (y_{ij} + \bar{y}_{i+} + \bar{y}_{+j} - \bar{y}_{++})^2$
Total	$rt-1$	$\sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{++})^2$

3.4 Two-Way Classificiation Models

Return to Table of Contents

- **Balanced designs** have the same number of sample in each treatment combination.
- **Complete designs** have at least one observation in each treatment combination.
- **Simple effects** are contrasts with only two nonzero coefficients.
- **Interaction effects** are differences of simple effects.

- **Main effects** are averages or sums of simple effects.
- A **two-way classification model** assigns the responses according to two covariate terms.
  - An $\boldsymbol{a \times b}$ **factorial design** has a levels of treatment A, and b levels of treatment B.
  - Model is $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$, where $E_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

Source	*df*	SSq
A	$a-1$	$\sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2$
B	$b-1$	$\sum_i \sum_j \sum_k (\bar{y}_{+j+} - \bar{y}_{+++})^2$
AB	$(a-1)(b-1)$	$\sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2$
Error	$N-ab$	$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij+})^2$
Total	$N-1$	$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2$

  - If the design is balanced, then contrasts for main effects are orthogonal.
    - If not balanced but complete, then contrasts might not be orthogonal.
    - If not complete, then contrasts are not estimable.
  - Always test for the interaction effect first.
  - A contrast follows the form $\theta = \boldsymbol{c}^T \boldsymbol{\mu}$, where $\boldsymbol{\mu}^T = \Big(\alpha_1, \ldots, \alpha_a, |\beta_1, \ldots, \beta_b, |(\alpha\beta)_{11}, \ldots, (\alpha\beta)_{ab}\Big)$.
    - The simple effect of β_j is defined as $\theta_{AB_j} = \mathbb{E}(\bar{Y}_{ij+} - \bar{Y}_{kj+})$.
  - $\mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

3.5 Mixed Effects Models

Return to Table of Contents

- A **random effect** T_i is a random variable representing the level of a treatment.
  - Useful when we have too many combinations to sample from.
- The **one-way random effects model** is $Y_{ij} = \mu + T_i + E_{ij}$, where $E_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, $T_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_T^2)$, and $T_i \perp E_{ij}$.

Source	*df*	$E(MSq)$ (Random)	$E(MSq)$ (Fixed)
Model	$t-1$	$\sigma^2 + n_0\sigma_T^2$	$\sigma^2 + \psi_T^2 n_0$
Error	$N-t$	σ^2	σ^2

, $\psi_T^2 = \frac{1}{n_0(t-1)} \sum_{i=1}^t n_i \tau_i^2$.

  - σ_T^2 is a measure of the variability of the effects among the treatments.
  - $Var(Y_{ij}) = \sigma^2 + \sigma_T^2$.
  - $Cov(Y_{ij}, Y_{i\ell}) = \sigma_T^2$, and $Cov(Y_{ij}, Y_{k\ell}) = 0$ for $i \neq k$.
  - We need to estimate σ_T^2.
    - MOM estimate is $\hat{\sigma}_T^2 = \frac{MSR - \hat{\sigma}^2}{n_0}$, where $n_0 = \frac{1}{t-1}\left(N - \frac{\sum_{i=1}^t n_i}{N}\right)$.
    - Maximum likelihood is also an option, but tends to underestimate.
    - REML is an option that performs similarly to MOM.
  - CI for μ is $\bar{Y}_{++} \pm t_{n-1,\alpha/2}\sqrt{\frac{MSR}{nt}}$.
  - CI for σ^2 is $\left(\frac{(N-t)MSE}{\chi^2_{N-t,\alpha/2}}, \frac{(N-t)MSE}{\chi^2_{N-t,1-\alpha/2}},\right)$.
  - CI for σ_T^2 is $\left(\frac{\hat{\nu}\hat{\sigma}_T^2}{\chi^2_{\hat{\nu},\alpha/2}}, \frac{\hat{\nu}\hat{\sigma}_T^2}{\chi^2_{\hat{\nu},1-\alpha/2}}\right)$, where $\hat{\nu} = \frac{(n\hat{\sigma}_T^2)^2}{\frac{MSR^2}{t-1} + \frac{MSE^2}{N-t}}$.

- The **coefficient of variation**, or **CV**, is $CV = \frac{\sqrt{Var(Y_{ij})}}{|E(Y_{ij})|} = \frac{\sqrt{\sigma^2+\sigma_T^2}}{|\mu|}$.
- **Satterthwaite's approximation for linear combinations**: $\hat{df} = \frac{\left(\sum_{i=1}^k c_i MS_i\right)^2}{\sum_{i=1}^k \frac{(c_i MS_i)^2}{df_i}}$.
- **Crossed factors** have every possible combination of factors.

Source	*df*	$E(MSq)$ (A, B fix.)	$E(MSq)$ (A, B rand.)	$E(MSq)$ (A fix., B rand.)
A	$a-1$	$nb\psi_A^2 + \sigma^2$	$nb\sigma_A^2 + n\sigma_{AB}^2 + \sigma^2$	$nb\psi_A^2 + n\sigma_{\alpha B}^2 + \sigma^2$
B	$b-1$	$na\psi_B^2 + \sigma^2$	$na\sigma_B^2 + n\sigma_{AB}^2 + \sigma^2$	$na\sigma_B^2 + n\sigma_{\alpha B}^2 + \sigma^2$
AB	$(a-1)(b-1)$	$n\psi_{AB}^2 + \sigma^2$	$n\sigma_{AB}^2 + \sigma^2$	$n\sigma_{\alpha B}^2 + \sigma^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

  - Assumes $n_{ij} = n$.
  - $\psi_A^2 = \frac{1}{a-1}\sum_{i=1}^a \alpha_i^2$, $\psi_B^2 = \frac{1}{b-1}\sum_{i=1}^b \beta_i^2$, $\psi_{AB}^2 = \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2$.
- B is **nested** in A if possible levels of B change on the value of A.
  - Model is $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk}$.

Source	*df*	$E(MSq)$ (A, B fix.)	$E(MSq)$ (A, B rand.)	$E(MSq)$ (A fix., B rand.)
A	$a-1$	$nb\psi_A^2 + \sigma^2$	$nb\sigma_A^2 + n\sigma_{B(A)}^2 + \sigma^2$	$nb\psi_A^2 + n\sigma_{B(A)}^2 + \sigma^2$
$B(A)$	$a(b-1)$	$n\psi_{B(A)}^2 + \sigma^2$	$n\sigma_{B(A)}^2 + \sigma^2$	$n\sigma_{B(A)}^2 + \sigma^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

    * Assumes $n_{ij} = n$.
    * $\psi_A^2 = \frac{1}{a-1}\sum_{i=1}^a \alpha_i^2$, $\psi_{B(A)}^2 = \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2$.
  - Interactions are not defined for nested models.

3.6 Repeated Measures Designs

Return to Table of Contents

- **Repeated measures designs** are defined by multiple observations per experimental unit.
  - Leads to correlation between responses for experimental units.
  - **Longitudinal study** arises from repeated observations over time.
  - **Subsampling studies** partition an experimental unit to create multiple observational units without additional intervention.
  - **Split-plot studies** partition an experimental unit into multiple observational units, where additional factors are then applied.
    * Factor A is the **between-plot factor**, factor B is the **within-plot factor**.
    * Useful when whole-plot factor is hard to change.
- The **split-plot model** with fixed treatment effects is $Y_{ijk} = \mu + \alpha_i + S_{k(i)} + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$, where $S_{k(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_s^2) \perp E_{ijk}$.
  - $S_{k(i)}$ is the **whole-plot error**, or the error of kth replicate of ith level of A.
  - E_{ijk} is the **split-plot error**, or the error of jth level of B in kth replicate of ith level of A.

Source	df	$E(MSq)$ $(n_i = n)$	F stat	Projection
A	$a-1$	$bn\psi_A^2 + b\sigma_s^2 + \sigma^2$	$\frac{MSA}{MSS(A)}$	$\boldsymbol{P_{X_A}} - \boldsymbol{P_1}$
$S(A)$	$\sum_i n_i - a$	$b\sigma_s^2 + \sigma^2$	$\frac{MSS(A)}{MSE}$	$\boldsymbol{P_Z} - \boldsymbol{P_{X_A}}$
B	$b-1$	$an\psi_B^2 + \sigma^2$	$\frac{MSB}{MSE}$	$\boldsymbol{P_{X_B}} - \boldsymbol{P_1}$
AB	$(a-1)(b-1)$	$n\psi_{AB}^2 + \sigma^2$	$\frac{MSAB}{MSE}$	$\boldsymbol{P_{X_{AB}}} - \boldsymbol{P_{X_A}} - \boldsymbol{P_{X_B}} + \boldsymbol{P_{X_1}}$
Error	$(\sum_i n_i - a)(b-1)$	σ^2		$\boldsymbol{I_N} - \boldsymbol{P_Z} + \boldsymbol{P_{X_A}} - \boldsymbol{P_{X_{AB}}}$
Total	$b\sum_i n_i - 1$			

– Estimate stuff (assumes $n_i = n$, fixed):

Comparison	Estimate	Variance	SE	df
A_i vs. A_j	$\bar{Y}_{i++} - \bar{Y}_{j++}$	$\frac{2}{bn}(b\sigma_s^2 + \sigma^2)$	$\sqrt{\frac{2}{bn}MSS(A)}$	$a(n-1)$
B_i vs. B_j	$\bar{Y}_{+i+} - \bar{Y}_{+j+}$	$\frac{2}{an}\sigma^2$	$\sqrt{\frac{2}{an}MSE}$	$a(n-1)(b-1)$
A_i and B_j vs. B_k	$\bar{Y}_{ij+} - \bar{Y}_{ik+}$	$\frac{2}{n}\sigma^2$	$\sqrt{\frac{2}{n}MSE}$	$a(n-1)(b-1)$
A_i, B_j vs. A_k, B_j	$\bar{Y}_{ij+} - \bar{Y}_{kj+}$	$\frac{2}{n}(\sigma_s^2 + \sigma^2)$	$\sqrt{\frac{2}{n}[MSSA + (b-1)MSE]}$	Satterthwaite
A_i, B_j vs. A_k, B_ℓ	$\bar{Y}_{ij+} - \bar{Y}_{k\ell+}$	$\frac{2}{n}(\sigma_s^2 + \sigma^2)$	$\sqrt{\frac{2}{n}[MSSA + (b-1)MSE]}$	Satterthwaite

Example: Three southern experiment stations are selected to study the effects of aeration on weed abundance in four species of grass. Separately at each station, four fields are randomized to species. Three sections of each field are randomized to three levels of aeration: none, once/year and twice/year. Weed counts are measured on each section. A partial ANOVA table is given below. Assume any effects involving station are random and that random effects are independent and normally distributed about 0.

Source	df	SSQ	MSQ	$EMSQ$
species		228.0		
station		151.3		
station×species		135.6		
aerate		296.4		
aerate×species		40.0		
Error		304.3		
Total		1155.4		

a. Complete the ANOVA table.

b. Report two F-tests and associated degrees of freedom for a test of the main effect of species and also for the main effect of aeration.

c. Report the standard errors (don't need to estimate variance components) of each of the following contrasts among treatment means:

   i. the difference between two species, averaging over aeration,

   ii. the species-specific aeration effect: the difference between aerating once and aerating twice, for a given species.

d. Report an unbiased estimate of the variance component for station.

a. First, we handle degrees of freedom. Species has 4 levels, station and aerate have 3, so their degrees of freedom is 3, 2, and 2, respectively. 4*3*3=36, so $df_{Total} = 36 - 1 = 35$. For the interaction, multiply the degrees of freedom for the main effects. $df_{Error} = 35 - (3 + 2 + 6 + 2 + 6) = 16$. Sum of squares is the MSQ divided by their respective degrees of freedom. For EMSQ, $\mathbb{E}$(MSE) is always σ^2. Every other EMSQ inherits this σ^2. For $\mathbb{E}$(MSaerate×species), count up the number of levels of station, which is 3, and multiply by ψ^2_{ASp}, since this term is fixed. For $\mathbb{E}$(MSaerate), similarly count up the number of levels of combinations of station and species, which is 12. Similar logic follows for $\mathbb{E}$(MSstation×species), but since it is random, we use σ^2_{StSp}, which is inherited by the main effects. Using the same logic as before, our final table is

Source	*df*	*SSQ*	*MSQ*	*EMSQ*
species	3	228.0	76	$\sigma^2 + 3\sigma^2_{StSp} + (3*3)\psi^2_{Sp}$
station	2	151.3	75.65	$\sigma^2 + 3\sigma^2_{StSp} + (4*3)\sigma^2_{St}$
station×species	6	135.6	22.6	$\sigma^2 + 3\sigma^2_{StSp}$
aerate	2	296.4	148.2	$\sigma^2 + (4*3)\psi^2_{A}$
aerate×species	6	40.0	6.6667	$\sigma^2 + 3\psi^2_{ASp}$
Error	16	304.3	19.0188	σ^2
Total	35	1155.4	$(\cdot)$	$(\cdot)$

b. $F_{species} = \frac{MSspecies}{MSstation\times species} = \frac{76}{22.6} = 3.3628 \overset{H_0}{\sim} F_{3,6}$;
$F_{species} = \frac{MSaerate}{MSE} = \frac{148.2}{19.0188} = 7.7923 \overset{H_0}{\sim} F_{2,16}$.

c. This is a split-plot model. Define $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + E_{ijk}$, where α, B, and γ refer to the species, station, and aerate effects, respectively.

i. Define $\hat{\theta}_1 := \bar{y}_{i..} - \bar{y}_{j..}$.

$$\begin{aligned}
Var(\hat{\theta}_1) =& Var\Big[(\mu + \alpha_i + B_+ + \overline{(\alpha B)}_{i+} + \bar{\gamma}_k + \overline{(\alpha\gamma)}_{i+} + \bar{E}_{i++}) - (\mu + \alpha_j + B_+ \\
&+ \overline{(\alpha B)}_{j+} + \bar{\gamma}_k + \overline{(\alpha\gamma)}_{j+} + \bar{E}_{j++})\Big] \\
=& Var\left[\alpha_i + \overline{(\alpha B)}_{i+} + \overline{(\alpha\gamma)}_{i+} + \bar{E}_{i++} - \alpha_j - \overline{(\alpha B)}_{j+} - \overline{(\alpha\gamma)}_{j+} - \bar{E}_{j++}\right] \\
\overset{\perp}{=}& Var(\alpha_i) + Var(\alpha_j) + Var(\overline{(\alpha B)}_{i+}) + Var(\overline{(\alpha B)}_{j+}) \\
&+ Var(\overline{(\alpha\gamma)}_{i+}) + Var(\overline{(\alpha\gamma)}_{j+}) + Var(\bar{E}_{i++}) + Var(\bar{E}_{j++}) \\
\overset{\text{i.d.}}{=}& 2Var(\alpha_i) + \frac{2}{j}Var((\alpha B)_{i\cdot}) + \frac{2}{k}Var((\alpha\gamma)_{i\cdot}) + \frac{2}{j*k}Var(E_{i\cdot\cdot}) \\
=& 2(0) + \frac{2}{3}\sigma^2_{St} + \frac{2}{3}(0) + \frac{2}{9}\sigma^2 = \frac{2}{3}\sigma^2_{St} + \frac{2}{9}\sigma^2; \\
SE(\hat{\theta}_1) =& \sqrt{\frac{2}{3}\sigma^2_{St} + \frac{2}{9}\sigma^2}.
\end{aligned}$$

ii. Define $\hat{\theta}_2 := \bar{y}_{j.2} - \bar{y}_{j.3}$.

$$\begin{aligned}
Var(\hat{\theta}_2) =& Var\left[\gamma_2 + \overline{(\alpha\gamma)}_{+2} + \bar{E}_{j+2} - \gamma_3 - \overline{(\alpha\gamma)}_{j3} - \bar{E}_{j+3}\right] \\
\overset{\text{i.i.d.}}{=}& 2Var(\gamma_i) + \frac{2}{i}Var((\alpha\gamma)_{ji}) + \frac{2}{i}Var(E_{j.i}) \\
=& 2(0) + \frac{2}{4}(0) + \frac{2}{3}\sigma^2 = \frac{2}{3}\sigma^2;\ SE(\hat{\theta}_2) = \sqrt{\frac{2}{3}\sigma^2}.
\end{aligned}$$

d.

$$MSStation = \hat{\sigma}^2 + 3\hat{\sigma}^2_{StSp} + 12\hat{\sigma}^2_{St};\ \hat{\sigma}^2_{St} = \frac{1}{12}\left[MSStation - \hat{\sigma}^2 - 3\hat{\sigma}^2_{StSp}\right]$$
$$= \frac{1}{12}\left[75.65 - MSE - 3\frac{MSStation \times Species - MSE}{3}\right]$$
$$= \frac{1}{12}(75.65 - MSStation \times Species) = \frac{1}{12}(75.65 - 22.6) = 4.4208. \blacksquare$$

Example: *(Note: We believe there is something incorrect with this problem, but we don't know what yet)* Suppose we want to study the effect of four types of fertilizers and two types of irrigation systems on yield of corn. A total of six fields are prepared for the experiment. First, each of the two irrigation systems is applied to three fields at random. Each of the fields are then divided into four sections, and the four types of fertilizers are applied in a random order.

a. Let us first focus on the irrigation effect only (that is, using the average yield of each field as the response). Complete the following ANOVA table. Show all your calculations.

Source	*df*	SSQ	MSQ	*F*
Irrigation		195.51		
Error				
Total	5	389.35		

b. Now suppose we conduct an analysis suitable for a Completely Randomized Design with two factors. Complete the following ANOVA table. Show all your calculations.

Source	*df*	SSQ	MSQ	*F*
Irrigation				
Fertilizer		266.01		
Irrigation×Fertilizer		62.79		
Error				
Total	23	2038.72		

c. Finally, consider an analysis for a split-plot design with irrigation as the whole-plot factor and fertilizer as the split-plot factor. Complete the following ANOVA table. Show all your calculations.

Source	*df*	SSQ	MSQ	*F*
Irrigation				
Whole-Plot Error				
Fertilizer				
Irrigation×Fertilizer				
Split-Plot Error				
Total	23			

d. Provide a clear argument as to which of the three analyses presented above is appropriate for analyzing all factorial effects.

a. First, $df_{Irrigation} = 2-1 = 1$, since there are two types of irrigation. This means that $df_{Error} = 5-1 = 4$. Similarly, $SSE = 389.35 - 195.51 = 193.84$. $MSIrrigation = \frac{SSIrrigation}{df_{Irrigation}} = 195.51$, similarly for MSE. Lastly, $F = \frac{MSIrrigation}{MSE} = 4.0345$. The resulting table is

Source	*df*	SSQ	MSQ	*F*
Irrigation	1	195.51	195.51	4.0345
Error	4	193.84	48.46	$(\cdot)$
Total	5	389.35	$(\cdot)$	$(\cdot)$

b. The only thing done differently than the strategies in part a) is $SSIrrigation$. In part a), $SSIrrigation = 3\sum_{i=1}^{2}(\bar{Y}_{i++} - \bar{Y}_{+++})^2$, but now with a CRD, $SSIrrigation = 4\sum_{i=1}^{2}(\bar{Y}_{i++} - \bar{Y}_{+++})^2$, which we can easily solve to get $SSIrrigation = 260.68$. Since all effects are fixed, the F-statistic uses MSE in the denominator. The final table is

Source	*df*	SSQ	MSQ	*F*
Irrigation	1	260.68	260.68	2.88
Fertilizer	3	266.01	88.67	0.93
Irrigation×Fertilizer	3	62.79	20.93	0.23
Error	16	1449.24	90.5775	$(\cdot)$
Total	23	2038.72	$(\cdot)$	$(\cdot)$

c. The whole-plot SSQ is equal to the $SSIrrigation$ from part a), and SSE is the same as in part b). Note that the denominator for the F-test for irrigation is the whole-plot error (which can be seen with EMSQ). The final table is

Source	*df*	SSQ	MSQ	*F*
Irrigation	1	65.17	65.17	1.3335
Whole-Plot Error	4	195.51	48.87	$(\cdot)$
Fertilizer	3	266.01	88.67	0.7342
Irrigation×Fertilizer	3	62.79	20.93	0.1733
Split-Plot Error	12	1449.24	120.77	$(\cdot)$
Total	23	2038.72	$(\cdot)$	$(\cdot)$

d. We need all appropriate terms to be accounted for in our model, in order to actually determine the importance of effects. Based on the context of the problem, the irrigation technique is a hard-to-measure effect, which means that split-plot is an appropriate model, so we use the split-plot analysis from part c). ■

4 ST 704: Statistical Methods II

Instructor: Dr. Erin Schliep (with Dr. Jacqueline Hughes-Oliver)
Semester: Spring 2024
Main Textbook: Faraway, *Extending the Linear Model with R*

4.1 Linear Regression

Return to Table of Contents

- The **fitted linear model**, or **estimated linear model**, is $\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_i X_{ij}$.
  - X_i are assumed to be constants.
  - Y_i are independent, and assumed to be functions of X_i.
  - β_i ($i \neq 0$) is the average increase in Y, given a unit increase in X_i, with other X_j values held constant.
  - The **rate of change** for X_i is $\frac{\partial}{\partial X_i}\left[\sum_{j=1}^{p} \hat{\beta}_j X_j\right]$.
- **Ordinary least squares regression**, or **OLS regression**, minimizes $\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|_2^2$ w.r.t. $\hat{\boldsymbol{\beta}}$.
  - Equivalent to $\min_{\hat{\beta}} \|\boldsymbol{e}\|_2^2$.
  - Under OLS, $\hat{\boldsymbol{Y}} \perp \boldsymbol{e}$.
  - Assumptions:
    - $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.
    - If β_0 is included, then $\sum_{i=1}^{n} e_i = 0$.
    - If β_0 is included, then $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.
  - If the Gauss-Markov assumptions are satisfied, then OLS is the BLUE.
  - If $e_i \sim \mathcal{N}$, then OLS is the MVUE.
  - Normality assumption is often violated in practice, but is still a useful approximation.
- For a linear model, our usual goal is inference on $\boldsymbol{A\beta}$.
  - The **estimated mean response** for OLS is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T\boldsymbol{Y} =: \boldsymbol{PY}$.
  - Examining $\boldsymbol{P}$ can give us the influence of individual observations of $\hat{\boldsymbol{Y}}$.
  - $\boldsymbol{P}_{ii}$ is the **leverage** of the ith observation.
    - $P_{ii} = \sum_{j=1}^{n} P_{ij}^2$.
    - $\frac{1}{n} \leq P_{ii} \leq 1$.
    - Large P_{ii} indicates larger influence on fit.
      - If $P_{ii} = 1$, then $\hat{Y}_i = Y_i$.
      - If $P_{ii} = 0$, then $\hat{Y}_i = 0$.
    - $\boldsymbol{P}$ is a projection matrix.
  - $\boldsymbol{A}\hat{\boldsymbol{\beta}} = \boldsymbol{CPY}$, where $\boldsymbol{A} = \boldsymbol{CX}$.
  - $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T)$.
  - Estimation: $\hat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\boldsymbol{X}_0\boldsymbol{\beta}, \sigma^2\boldsymbol{X}_0(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}_0)$.
  - Prediction: $\hat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\boldsymbol{X}_0\boldsymbol{\beta}, \sigma^2\boldsymbol{I} + \sigma^2\boldsymbol{X}_0(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}_0)$.
  - $\hat{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{P})$.
  - $\boldsymbol{e} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{I} - \boldsymbol{P}))$.
- Assumption issues in regression:
  - $\boldsymbol{X}$ is observed with error.
    - Estimators are usually biased towards zero.
  - The mean model is misspecified. Includes things like omitting important predictors, biased estimators, $\hat{\sigma}^2$ is too big, non-additive model, or a nonlinear relationship is more appropriate.
    - Plot of $\hat{\boldsymbol{Y}}$ versus $\boldsymbol{e}$ should show no trend.

    - If a predictor is omitted, then $\bar{\boldsymbol{e}} \neq 0$.
    - If there are multiple predictors, then use partial residual plots.
        - The **partial residual** for X_j is $e^* = e + \hat{\beta}_j X_j$.
        - If X_j is relevant, then the residuals of a model fit without X_j should not be uncorrelated with X_j.
  - Suppose the true relationship is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$, but we fit $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.
    - If $\boldsymbol{Z} \in \text{col}(\boldsymbol{X})$, then $\mathbb{E}(\hat{\boldsymbol{Y}}) \neq \boldsymbol{X\beta}$, and $\mathbb{E}(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$.
    - If the columns of $\boldsymbol{Z}$ are orthogonal to $\boldsymbol{X}$, then $\mathbb{E}(\hat{\boldsymbol{Y}}) \neq \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$.
    - This means that estimating $\boldsymbol{\beta}$ and estimating $\mathbb{E}(\boldsymbol{Y})$ might have different requirements, and different consequences depending on the model.
  - Errors are not uncorrelated.
    - $Cov(\hat{\boldsymbol{Y}}, \boldsymbol{e}) \neq 0$.
    - $Var(\hat{\boldsymbol{\beta}})$ is not minimal.
    - Detect correlation with the **Durbin-Watson test**.
        - $d = 2(1 - \hat{\rho})$, where $\hat{\rho} = \widehat{Corr}(e_i, e_{i-1})$.
  - $Var(\boldsymbol{e})$ is not constant.
    - Standard error of estimates are different than what is specified.
        - HTs and CIs are no longer valid.
    - Could transform variables, or use a different model.
        - The **Box-Cox transformation** family is $Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda-1}}{\lambda \dot{Y}^{\lambda-1}}, \lambda \neq 0 \\ \dot{Y} \log(Y_i), \lambda = 0 \end{cases}$. Choose λ such that $SSE^{(\lambda)} \leq SSE^{(\lambda)}_{\min}\left[1 + \frac{t^2_{df_E, \alpha/2}}{df_E}\right]$.
        - If $Var(Y) \propto [\mathbb{E}(Y)]^{2k}$, then choose Y^{1-k}, where $Y^0 = \log(Y)$.
  - $\boldsymbol{\epsilon}$ does not follow a Normal distribution.
    - Actually not too horrible if violated, since expectation/variances of estimators don't change, and F-tests are robust to this assumption.
    - HTs and CIs need a large sample size so asymptotic Normality holds.
    - Look for a nonlinear pattern in a QQ-Plot.
        - A "J" shape means a right-skewed distribution.
        - If the line doesn't go through the origin, then we are missing an important predictor.
        - Theory says we need $n \geq 5$ to be sufficient, but recommended $n \geq 30$.
    - The **studentized residual** is $r_i = \frac{e_i}{\sqrt{MSE(1-P_{ii})}}$.
    - **Jackknife**, or **LOOCV**: see how much the ith observation impacts the estimates.
        - $r_i^* = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE_{(i)}(1-P_{ii})}}$, where $MSE_{(i)} = \frac{(n-p)(MSE)^2 - \frac{r_i^2}{1-P_{ii}}}{n-p-1}$.
        - QQ-Plots plot r_i^* against the quantiles from the Normal distribution.

- SLR equations:
  - Unbiased estimator of σ^2 is $MSE = s^2 = \frac{SSE}{n-2}$.
  - $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = r_{XY}\frac{S_Y}{S_X} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.
  - $SE(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.
  - $SE(\hat{\beta}_1) = s\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

- Estimation vs. prediction:
  - Estimator will be the fitted value for both.
  - Estimate $\mathbb{E}(Y)$ at $X = x_0$: $SE(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.
  - Predict Y at $X = x_0$: $SE(Y - \hat{Y}_0) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

- MLR equations:
  - $SSE = \|\boldsymbol{e}\|_2^2$.
  - $s^2 = \frac{SSE}{n-rank(\boldsymbol{X})} = \frac{SSE}{n-(p+1)}$.
  - $Var(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
  - $\mathbb{E}(Y|x_0) = \sigma^2\boldsymbol{x_0'}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}$.
  - $R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST}$.
- Model A is **nested** in model B if model A can be obtained by constraining model B.
  - A is referred to as the **reduced model**, whereas B is the **full model**.
  - $R(\beta_{q+1}, \ldots, \beta_p|\beta_0, \ldots, \beta_q)$ is the **extra sum of squares** due to model B from model A.
  - A **nested *F*-test** determines if the full model is necessary.
    - Test statistic is $F = \frac{(SSE_R - SSE_F)/(p-q)}{MSE_F} = \frac{(SSR_R - SSR_F)/(p-q)}{MSE_F}$.
    - Test statistic follows $F_{df_{E_F}}^{p-q}$.
- **Sequential sum of squares**, or **type I sum of squares**, adds one variable to the model at a time to measure the change in sum of squares.
  - Order matters!
- **Partial sum of squares**, or **type III sum of squares**, is the change in sum of squares with all other predictors in the model.
  - Order does not matter.
  - Is equal to sequential sum of squares when $X'X$ is diagonal.
- A model is **additive** with respect to a set of variables if we can group the model by the variables.
  - Models with interaction terms are not additive.

4.2 Model Assessment

Return to Table of Contents

- **Internal validation** determines which model and variables best explain the sample data.
  - Could result in overfitting the data.
  - Relative importance of variables can vary from the population and our sample.
  - Could use SSE, R^2, R_a^2 to choose the model.
    - MSE is not necessarily monotone.
    - Choose the simplest reasonable model.
  - **Akaike information criterion**, or **AIC**, is $AIC = n\log(SSE) - n\log(n) + 2k$.
    - Smaller values are better.
  - **Bayesian information criterion**, or **BIC**, is $BIC = n\log(SSE) - n\log(n) + k\log(n)$.
    - Smaller values are better.
  - **Mallow's C_p** is $C_p = \frac{SSE}{\sigma_F^2} + 2(p+1) - n$.
    - An adequate model has $C_p \approx p+1$.
    - An inadequate model has $C_p > p+1$.
- **External validation** determines which model and variables best predict data outside of our sample data.
  - Requires two independent and representative datasets.
  - Criteria for external validation: suppose $Y_{n+1}, \ldots, Y_{n+m}$ is the test set, with mean $\bar{Y}$.
    - $R_{pred}^2 = 1 - \frac{\sum_{i=n+1}^{n+m}(Y_i - \hat{Y}_i)^2}{\sum_{i=n+1}^{n+m}(Y_i - \bar{Y}_i)^2}$.
    - $MSE_{pred} = \frac{1}{m}\sum_{i=n+1}^{n+m}(Y_i - \hat{Y}_i)^2$.
    - $Corr(Y, \hat{Y})^2$.

- When we don't have two independent and representative datasets, we partition our one dataset into a training and test set.
  - ***K*-fold cross-validation**, or ***K*-fold CV**, partitions dataset into K folds, and iteratively uses the ith fold as the test set.
    - The best model that results in the smallest $\overline{CV} = \frac{1}{K}\sum_{k=1}^{K} CV_k$ likely overfits our data, so we instead use the smallest model such that $\overline{CV}_* < \overline{CV} + SE(\overline{CV})$.
- Inference is affected by the model we select, along with the selection process we use.
  - Selection is heavily affected by noise, especially when $p \approx n$.
- **All-subset regression** considers all 2^{p-1} models.
  - May not even be possible, especially for larger p.
- **Forward selection** starts with a base model, and adds in the single best predictor one-at-a-time until no new predictor adds much to the model.
  - Once a predictor is added, it cannot be removed.
- **Backwards elimination** starts with the most complex model, and removes predictors one-at-a-time until no predictors should be removed.
  - Once a predictor is removed, it cannot be re-added.
- **Stepwise selection** starts with the base model, and adds/removes predictors one-at-a-time until no noticeable change.

4.3 Biased Regression and Dimension Reduction

Return to Table of Contents

- We now want to fit a regression model such that SSE is minimized, but also places a penalty on $\hat{\boldsymbol{\beta}}$ in the form of λ.
- **Ridge regression** is $\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_i X_{ji}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$, where $\lambda \geq 0$.
  - Center/scale predictors beforehand.
    - Shrinkage applies to the partial slopes, not the intercept.
    - Scaling impacts estimates and choice of λ.
  - We balance minimizing SSE with making the length of the slope vector close to zero.
  - A larger λ shrinks the $\hat{\boldsymbol{\beta}}$ vector closer to zero.
  - Handles collinearity by shrinking elements of $\hat{\boldsymbol{\beta}}$ closer to zero faster.
  - $\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.
    - Is linear in $\boldsymbol{Y}$.
    - $Bias(\hat{\boldsymbol{\beta}}^{ridge}) = -\lambda(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}$, so larger λ means more bias.
    - $Var(\hat{\boldsymbol{\beta}}^{ridge}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$, so larger λ means less variance.
  - Shrinkage is proportional, so $\hat{\beta}_j^{ridge} = \frac{n}{n+\lambda}\hat{\beta}_j$ for orthogonal $\boldsymbol{X}^T\boldsymbol{X}$.
  - Never shrinks coefficients exactly to zero.
  - Choose λ with CV.
- **Lasso regression** is $\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_i X_{ji}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$, where $\lambda \geq 0$.
  - Center/scale predictors beforehand.
  - Has no closed-form matrix expression.
  - Is nonlinear in $\boldsymbol{Y}$.
  - Shrinkage is soft-thresholded, so $\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$.
  - Can shrink coefficients to zero, so can be a variable-selection technique.
    - Choice of zeroed coefficients might be arbitrary.

  - Choose λ with CV.
- **Elastic net regression** chooses β_0 and $\boldsymbol{\beta}$ that minimizes $SSE + \lambda \left[\frac{(1-\alpha)}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right]$.
  - Is a combination of Ridge and LASSO.
  - α_1 and α_2 must be specified beforehand.
  - Choose λ with CV.
- **Dimension reduction** projects predictors from $\mathbb{R}^p$ to $\mathbb{R}^g$, where $g << p$.
  - Performs regression on transformed predictors.
  - Does not perform variable selection.
  - Could improve interpretation using new variables.
  - Choose number of components with CV.
- If $\boldsymbol{X}$ has near-redundancies, then we convert the $\boldsymbol{X}$-space into the $\boldsymbol{W}$-space of orthogonal columns.
  - Center/scale predictors beforehand (convention).
  - **Scores** are the columns of $\boldsymbol{W}$, which are linear combinations of $\boldsymbol{X}$.
    * Scores are ordered by relevance.
    * We drop irrelevant scores to get $\boldsymbol{W}_{(g)}$-space.
- **Principal components regression**, or **PCR**, obtains the $\boldsymbol{W}$-space using the eigen-decomposition of $\boldsymbol{X}^T\boldsymbol{X}$.
  - Is unsupervised, meaning it does not use $\boldsymbol{Y}$.
  - $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{V}$, where $\boldsymbol{W} \in \mathbb{R}^{n \times p}$, $\boldsymbol{V}$ are corresponding eigenvectors corresponding to eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.
    * The columns of $\boldsymbol{W}$ are known as **principal components**.
      · kth component is considered "irrelevant" if $\sqrt{\frac{\lambda_1}{\lambda_k}} > 10$.
    * $\boldsymbol{X}v_1$ explains most of the variation in the $\boldsymbol{X}$-space.
    * $\boldsymbol{X}v_1$ and $\boldsymbol{X}v_2$ explains $\left(\frac{\lambda_1+\lambda_2}{\sum_{i=1}^p \lambda_i} \right) 100\%$ of the variation in the $\boldsymbol{X}$-space.
- **Partial least squares**, or **PLS**, is a supervised dimension reduction approach.
  - $\boldsymbol{W}$-space seeks highest level of variation in $\boldsymbol{X}$-space and strong correlation with $\boldsymbol{Y}$.
  - Is more algorithmic than theoretical.

4.4 GLMs

Return to Table of Contents

- Recall that in linear models, we want to estimate β and σ^2, and that $\mathbb{E}(\boldsymbol{Y}) = x'\boldsymbol{\beta} =: \boldsymbol{\eta}$.
- A **generalized linear model**, or **GLM**, is defined such that $g\left[\mathbb{E}(Y_i)\right] = \eta_i$, where g is known as the **link function**.
  - $E(Y_i) = g^{-1}(\eta_i)$, where g^{-1} is called the **inverse link function**.
  - $Y_1, \ldots, Y_n$ are now iid exponential family with dispersion parameter ϕ.
    * We now estimate β and ϕ.
- Properties of the link function:
  - Must be invertible (thus also monotone).
  - Must be able to map the mean response to an additive model.
  - Ensures a range restriction on the mean response.
  - Distributions in the exponential family have a natural parameterization.
  - Any suitable link function may be paired with any distribution in the exponential family.
- With GLMs, we want to estimate β and ϕ, perform inference on β (which requires a standard error), estimate the mean response $g^{-1}(x_i^T\boldsymbol{\beta})$, and determine model fit.

- The **exponential family** with natural parameter $\theta = \theta(\mu)$, dispersion parameter ϕ has PDF $f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}$ for some functions a, b, c.
  - For example, $N(\mu,\sigma^2)$ looks like $\exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]\right\}$, where $\theta = \mu$, $a(\phi) = \sigma^2$, $b(\theta) = \frac{\mu^2}{2}$, $w_i = 1$, and $c = -\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$.
  - The **canonical link function** is the link function g such that $g(\mu_i) = \theta_i$.
  - $b'(\boldsymbol{\theta})$ is the **mean function**; that is, $b'(\boldsymbol{\theta}) = \mathbb{E}(Y_i)$.
  - $b''(\boldsymbol{\theta})$ is the **variance function**; that is, $b''(\boldsymbol{\theta}) = a(\phi)Var(Y_i)$.
- $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}, a(\phi)[\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F}]^{-1}\right)$, where $\boldsymbol{F} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'}$, $V_{ii} = Var(\mu_i)$ (0 o.w.).
  - $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, [I(\boldsymbol{\beta})]^{-1})$, where $I(\boldsymbol{\beta})_{ij} = -\mathbb{E}\left[\left\{\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right\}_{ij}\right]$ under regularity conditions.
  - $T_W = (\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^T[a(\phi)\boldsymbol{L}(\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \chi^2_q$.
  - $T_{LR} = 2(\ell(\hat{\boldsymbol{\beta}}_F) - \ell(\hat{\boldsymbol{\beta}}_R)) \sim \chi^2_q$, where ℓ is the log-likelihood.
- **Deviance** for model M is $D^*(y;\hat{\boldsymbol{\mu}}) = 2\phi\left\{\ell(\boldsymbol{y};\tilde{\boldsymbol{\mu}}) - \ell(\boldsymbol{y};\hat{\boldsymbol{\mu}})\right\}$.
  - The **saturated model** "fits the data perfectly," where $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$.
  - Measures how well a chosen model fits our data, compared to the saturated model.
  - The **scaled deviance** is $\frac{D^*(y;\hat{\boldsymbol{\mu}})}{\phi}$.
    - If Y_i approximately follows a Normal distribution with a roughly identity link function ($\theta_i = \mu_i$), then $\frac{D^*(y;\hat{\boldsymbol{\mu}})}{\phi} \approx \chi^2_{n-p}$.
      - Approximation does not improve when n increases!
- Using MOM, $\hat{\phi} = \frac{D^*(y;\hat{\boldsymbol{\mu}})}{n-p}$.
  - If $\hat{\phi}$ is large, then we might be missing important predictors, overdispersion may be present, or Y_i are not uncorrelated.
  - Over-reporting the value of $\hat{\phi}$ will lead to larger SEs than anticipated (and vice versa).
- $AIC = -2\ell(\hat{\boldsymbol{\mu}}) + 2p$ and $BIC = -2\ell(\hat{\boldsymbol{\mu}}) + p\log(n)$ are also used for model selection.
- Residuals used for diagnostics:
  - **Pearson residual**: $\mathbb{X}^2_i = \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}$.
  - **Deviance residual**: $r_{D_i} = \text{sign}(y - \hat{\mu})\sqrt{d_i}$.
    - The **standardized deviance residual** is $r_{s,D_i} = \frac{r_{D_i}}{\hat{\phi}(1 - h_{ii}^{GLM})^{1/2}}$.
- We often plot $\hat{\eta}$ against fitted values for diagnostics.
- **Logistic regression** models the probability of an observation belonging to a class.
  - Uses the logit link, which is $\log\left(\frac{p(x)}{1-p(x)}\right)$.
  - Assumes independent $Y_i \sim \frac{1}{n_i}Bin(n_i, p_i)$.
  - **Even odds** are when $Odds \approx 1$, which means that $p(x) \approx 0.5$.
  - If $Odds \approx \beta_0$, then odds don't change with x.
  - $\left[\frac{\left(\frac{p(x+1)}{1-p(x+1)}\right)}{\left(\frac{p(x)}{1-p(x)}\right)}\right] = e^{\sum_{i=1}^p \beta_i}$ are the odds ratio for increasing all of x by 1.
    - Odds increase multiplicatively by $e^{\sum_{i=1}^p \beta_i}$ for unit increase in x.
  - We often use MLE to estimate $\boldsymbol{\beta}$.
    - $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i\eta_i - \log(1 + e_i^\eta)]$.
    - ℓ is nonlinear w.r.t. $\boldsymbol{\beta}$, so we must obtain estimates computationally.
    - If there exists a lot of separation, then estimates will have trouble converging.
  - We use asymptotic intervals and tests for $\hat{\boldsymbol{\beta}}$.

- The **likelihood displacement** diagnostic is $LD_i = 2\left\{\ell_M(\hat{\boldsymbol{\theta}};\boldsymbol{y}) - \ell_M(\hat{\boldsymbol{\theta}}_{(-i)};\boldsymbol{y})\right\}$, where $\hat{\boldsymbol{\theta}}_{(-i)}$ is MLE with the ith observation excluded.

- **Poisson regression** models count data.
  - Uses a log link ($\log(\lambda_i)$), with inverse link $e^{\boldsymbol{x}'\boldsymbol{\beta}}$.
  - $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i - e^{x_i^T\beta} - \log(y_i!)\right]$.
  - The variance is a function of the mean.
  - If overdispersion is present, use negative binomial regression instead.

4.5 Mixed Models

Return to Table of Contents

- **Restricted maximum likelihood estimation**, or **REML estimation**, is used to estimate σ^2 without worrying about $\boldsymbol{\beta}$ by zeroing out the mean.
  - Estimate σ^2 using ML for $\boldsymbol{KY} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{KK'})$, where $\boldsymbol{K}$ is positive-definite.
  - $\ell_{REML}(\sigma^2;\boldsymbol{y}) = c - \frac{1}{2}\log\left|\sigma^2\boldsymbol{KK'}\right| - \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{K}^T(\boldsymbol{KK'})^{-1}\boldsymbol{Ky}$.
    - If $\boldsymbol{K}$ is $n-p$ independent rows of $(\boldsymbol{I}-\boldsymbol{P}_X)$, then $\hat{\sigma}^2_{REML}$ is maximized at $\frac{1}{n-p}\left\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|_2^2$.

- In a **mixed model**, we allow for random coefficients.
  - Previously, $\boldsymbol{\beta}$ was a vector of fixed parameters.
  - $\boldsymbol{\epsilon} \sim \mathcal{N}(0, V)$, where V is positive-definite.

- The **classical linear mixed model** is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}$, where $\boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G})$ is a vector of our random effects, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{R})$ and $Cov(\boldsymbol{\alpha}, \boldsymbol{\epsilon}) = 0$.
  - We need to estimate $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\delta})$, where $\boldsymbol{\delta} = (\boldsymbol{G}, \boldsymbol{R})$.
  - The **marginal model** is $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \boldsymbol{ZGZ'} + \boldsymbol{R})$, where $Var(\boldsymbol{\delta}) = \boldsymbol{ZGZ'} + \boldsymbol{R}$.
  - The **subject-specific model** is $\boldsymbol{Y}|\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{Z\alpha}, \boldsymbol{R})$.
  - $\ell(\boldsymbol{\beta}, \boldsymbol{\delta};\boldsymbol{y}) = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})$.
    - If $\boldsymbol{\delta}$ is known, then $\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$.
    - $\hat{\boldsymbol{\beta}}_{ML} \sim \mathcal{N}(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1})$.
      - If $\hat{\boldsymbol{\beta}}_{ML}$ is computed with $\hat{\boldsymbol{V}}$ instead of $\boldsymbol{V}$, then distribution is approximate, but is still consistent if $\hat{\boldsymbol{V}}$ is consistent.
    - Could also obtain $\hat{\boldsymbol{\beta}}_{ML}$ first to then obtain $\hat{\boldsymbol{\delta}}_{ML}$.
      - Usually biased, but asymptotically Normal.
  - $\ell(\boldsymbol{\delta};\boldsymbol{y})_{REML} = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})$.
    - Maximize numerically to get $\hat{\boldsymbol{\delta}}_{REML}$, where $\hat{\boldsymbol{\beta}}_{ML}$ is obtained by then using $\hat{\boldsymbol{\delta}}_{REML}$.
      - Is less biased than $\hat{\boldsymbol{\delta}}_{ML}$, and is asymptotically Normal.
  - With $\boldsymbol{V}$ known, $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T)$.
    - A $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm z_{\alpha/2}\sqrt{\left[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right]_{jj}}$.
    - $(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})^T\left[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m}) \sim \chi^2_{rank(\boldsymbol{A})}$.
  - With $\boldsymbol{V} = \sigma^2\boldsymbol{D}$, where $\boldsymbol{D}$ known, $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T)$.
    - A $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm t_{\alpha/2,n-rank(\boldsymbol{X})}\sqrt{\hat{\sigma}^2_{REML}\left[(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\right]_{jj}}$.
    - $\frac{n-rank(\boldsymbol{X})}{\sigma^2}\hat{\sigma}^2_{REML} \sim \chi^2_{n-rank(\boldsymbol{X})}$, independent of $\hat{\boldsymbol{\beta}}$.
    - $\frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})^T\left[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})/rank(\boldsymbol{A})}{\hat{\sigma}^2_{REML}} \sim F_{rank(\boldsymbol{A}),n-rank(\boldsymbol{X})}$.
  - With $\boldsymbol{V} = \sigma^2\boldsymbol{D}$, where both are unknown, $\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1})$.
    - $(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})^T\left[\hat{\sigma}^2_{REML}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m}) \to \chi^2_{rank(\boldsymbol{A})}$ (same with using σ^2).

    * $\frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})^T\left[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^T\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})/rank(\boldsymbol{A})}{\hat{\sigma}^2_{REML}}$ does not converge to an F distribution!
        · In practice, F distribution is okay. Use Satterthwaite for *df* adjustment.

- With $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^T$, $\hat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta}) = diag\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X}, \frac{n}{2\sigma^4}\right)$.
  - Suppose we want to test $H_0 : h(\boldsymbol{\theta}) = \boldsymbol{0}$ vs. $H_1 : h(\boldsymbol{\theta}) \neq \boldsymbol{0}$, where $h(\boldsymbol{\theta}) \in \mathbb{R}^r$.
    * $T_W = h(\hat{\boldsymbol{\theta}}_n)^T \left[H(\hat{\boldsymbol{\theta}}_n)I(\hat{\boldsymbol{\theta}}_n)^{-1}H(\hat{\boldsymbol{\theta}}_n)^T\right]^{-1} h(\hat{\boldsymbol{\theta}}_n) \sim \chi^2_r$, where $H(\boldsymbol{\theta}) = \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$.
    * $T_S = S(\hat{\boldsymbol{\theta}}_0)^T I(\hat{\boldsymbol{\theta}}_0)^{-1} S(\hat{\boldsymbol{\theta}}_0) \sim \chi^2_r$.
  - Tests using REML typically reduce bias.
- Prediction for marginal model is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, whereas for subject-specific model, it is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{G}}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.
  - $\hat{\boldsymbol{\alpha}}$ is the conditional mean of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$.
  - If $\boldsymbol{G}$ and $\boldsymbol{R}$ are known, then this is the BLUP.
- The **random intercepts, random slope model** lets us treat β_0 as a random effect.
- A **two-level LMM** has the form $\boldsymbol{Y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{1,ij}\boldsymbol{\alpha}_i + \boldsymbol{Z}_{2,ij}\boldsymbol{\alpha}_{ij} + \boldsymbol{\epsilon}_{ij}$, where RVs are independent.
- Following an LMM form of $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha}$, a **generalized linear mixed model** has the form $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{\alpha}) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha})$.
  - $Var(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{A}^{1/2}\boldsymbol{R}\boldsymbol{A}^{1/2}$, where $\boldsymbol{A} = diag(w_1 h(\mu_1), \ldots, w_n h(\mu_n))$, and $\boldsymbol{R} = \phi\boldsymbol{I}$ typically.
- **Newton-Raphson**: $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)} + f(\hat{\boldsymbol{\beta}}^{(i)}) + F(\hat{\boldsymbol{\beta}}^{(i)})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(i)})$.

5 ST 705: Linear Models and Variance Components

Instructor: Dr. Jonathan Williams
Semester: Spring 2024
Main Textbook: Monahan, *A Primer on Linear Models*

5.1 Linear Algebra Review

Return to Table of Contents

- λ is an **eigenvalue** of matrix $\boldsymbol{X}$ if it satisfies $\boldsymbol{Xv} = \lambda\boldsymbol{v}$, where $\boldsymbol{v} \neq \boldsymbol{0}$ is the respective **eigenvector.**
- Two vectors are **orthogonal** if their inner product is zero.
- A matrix is **orthogonal** if $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$.
- The **Euclidean norm**, or $\boldsymbol{\ell_2}$ **norm**, is $\|\boldsymbol{x}\|_2 = (\langle x, x\rangle)^{1/2}$.
- The **Frobenius norm** is $\|\boldsymbol{A}\|_F = [\text{tr}(\boldsymbol{A}^T\boldsymbol{A})]^{1/2}$.
- $|\langle \boldsymbol{a}, \boldsymbol{b}\rangle| \leq \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|$.
- $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$.
- The **matrix product** of $\boldsymbol{A}$ and $\boldsymbol{B}$ is $(\boldsymbol{AB})_{ij} = \sum_{k=1}^{p} \boldsymbol{A}_{ik}\boldsymbol{B}_{kj}$.
- An **orthonormal matrix** has mutually orthogonal and unit length columns.
- The **rank** of a matrix is the number of linearly independent rows or columns.
  - $rank(\boldsymbol{AB}) \leq \min\{rank(\boldsymbol{A}), rank(\boldsymbol{B})\}$.
  - $p = rank(\text{null}(\boldsymbol{A})) + rank(\text{col}(\boldsymbol{A}))$.
- $\boldsymbol{A}^\perp$ is the **orthogonal complement** to $\boldsymbol{A}$ is defined as $\boldsymbol{A}^\perp := \{\boldsymbol{x} \in \boldsymbol{A} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0\ \forall \boldsymbol{y} \in \boldsymbol{A}^\perp\}$.
  - Suppose $\boldsymbol{S} \subseteq \boldsymbol{A}$. Then, for every $\boldsymbol{y} \in \boldsymbol{A}$, there exists a unique $\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{z}$ for $\boldsymbol{u} \in \boldsymbol{S}$, $\boldsymbol{z} \in \boldsymbol{S}^\perp$.
  - col($\boldsymbol{A}$) and null($\boldsymbol{A}^T$) are orthogonal complements in $\mathbb{R}^p$.
- If $\boldsymbol{Bx} = \boldsymbol{Cx}$ for all $\boldsymbol{x}$, then $\boldsymbol{B} = \boldsymbol{C}$.
- If $\boldsymbol{AB} = \boldsymbol{AC}$ for full-rank $\boldsymbol{A}$, then $\boldsymbol{B} = \boldsymbol{C}$.
- $\boldsymbol{X}^T\boldsymbol{XA} = \boldsymbol{X}^T\boldsymbol{XB}$ iff $\boldsymbol{XA} = \boldsymbol{XB}$.
- A system of equations $\boldsymbol{Ax} = \boldsymbol{c}$ is **consistent** iff there exists a solution $\boldsymbol{x}^*$ such that $\boldsymbol{Ax}^* = \boldsymbol{c}$.
  - $\boldsymbol{Ax} = \boldsymbol{c}$ is consistent iff $\boldsymbol{c} \in$ col($\boldsymbol{A}$).
  - Suppose $\boldsymbol{Ax} = \boldsymbol{c}$ is consistent. Let $\boldsymbol{G}$ be a generalized inverse of $\boldsymbol{A}$. $\tilde{\boldsymbol{x}}$ is a solution to $\boldsymbol{Ax} = \boldsymbol{c}$ iff $\tilde{\boldsymbol{x}} = \boldsymbol{Gc} + (\boldsymbol{I} - \boldsymbol{GA})\boldsymbol{z}$ for some $\boldsymbol{z}$.
- $\boldsymbol{X}$ is **idempotent** if $\boldsymbol{XX} = \boldsymbol{X}$.
  - If $\boldsymbol{X}$ is idempotent, then $rank(\boldsymbol{X}) = \text{tr}(\boldsymbol{X})$.
  - If $\boldsymbol{X}$ is idempotent, then the eigenvalues of $\boldsymbol{X}$ are 0 or 1.
- $(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T$ is a generalized inverse for $\boldsymbol{X}$.
- A square matrix $\boldsymbol{P}$ is a **projection** onto vector space S iff $\boldsymbol{P}$ is idempotent, $\boldsymbol{Px} \in S$ for some $\boldsymbol{x}$, and $\boldsymbol{Pz} = \boldsymbol{z}$ for all $\boldsymbol{z} \in S$.
  - $\boldsymbol{AA}^g$ is a projection onto col($\boldsymbol{A}$).
  - $(\boldsymbol{I} - \boldsymbol{A}^g\boldsymbol{A})$ is a projection onto null($\boldsymbol{A}$).
  - $\boldsymbol{P}$ is unique if it is symmetric.
  - If $\boldsymbol{P}$ is symmetric and projects onto S, then $\boldsymbol{I} - \boldsymbol{P}$ projects onto $S^\perp$.
- $\boldsymbol{P}_X := \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T$ is the **symmetric projection matrix** of $\boldsymbol{X}$.
  - If col($\boldsymbol{X}$) = col($\boldsymbol{W}$), then $\boldsymbol{P}_X = \boldsymbol{P}_W$.

  - Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{W} \in \mathbb{R}^{n\times q}$. If $\text{col}(\boldsymbol{W}) \subseteq \text{col}(\boldsymbol{X})$, then $\boldsymbol{P}_X - \boldsymbol{P}_W$ is the projection onto $\text{col}\{(\boldsymbol{I} - \boldsymbol{P}_W)\boldsymbol{X}\}$.

- $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- $\det(c\boldsymbol{A}) = c^p \det(\boldsymbol{A})$ for square $\boldsymbol{A}$.
- The **spectral decomposition** of square $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{QDQ}'$, where $\boldsymbol{D}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{Q}$ is an orthonormal matrix of eigenvectors of A.
- A matrix is **nonnegative-definite** if $\boldsymbol{X}^T\boldsymbol{Ax} \geq 0$ for all $\boldsymbol{x}$.
  - If $\boldsymbol{A}$ is non-singular, then it is **positive-definite**.
- **Cholesky decomposition**: $\boldsymbol{A}$ is positive-definite iff there exists a non-singular, lower-triangular matrix $\boldsymbol{L}$ such that $\boldsymbol{A} = \boldsymbol{LL}^T$.
- A square matrix $\boldsymbol{A}$ is **diagonalizable** if there exists a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{AD}$.
- If $\text{col}(\boldsymbol{X}) = \text{col}(\boldsymbol{W})$, then $\exists \boldsymbol{S}, \boldsymbol{T}$ such that $\boldsymbol{X} = \boldsymbol{WS}$ and $\boldsymbol{W} = \boldsymbol{XT}$.
- $\text{null}(\boldsymbol{X}^T\boldsymbol{X}) = \text{null}(\boldsymbol{X})$.
- $\text{col}(\boldsymbol{X}^T\boldsymbol{X}) = \text{col}(\boldsymbol{X}^T)$.
- $rank(\boldsymbol{X}^T\boldsymbol{X}) = rank(\boldsymbol{X})$.
- If $rank(\boldsymbol{BC}) = rank(\boldsymbol{B})$, then $\text{col}(\boldsymbol{BC}) = \text{col}(\boldsymbol{B})$.
- $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.
- $\mathbb{E}[\boldsymbol{a}^T\boldsymbol{Y}] = \boldsymbol{a}^T\mathbb{E}[\boldsymbol{Y}]$, $\mathbb{E}[\boldsymbol{Y}^T\boldsymbol{a}] = \mathbb{E}[\boldsymbol{Y}^T]\boldsymbol{a}$.
- $Var(\boldsymbol{Y}) = \mathbb{E}\left[(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}))(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}))^T\right]$.
- $Var(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{a}^T Var(\boldsymbol{Y})\boldsymbol{a}$.
- $Cov(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T Cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{b}$.
- Trace trick: $\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X}) = \text{tr}\{\mathbb{E}(\boldsymbol{XX}^T)\}$.

5.2 The Normal Equations

Return to Table of Contents

- For $f : \mathbb{R}^p \to \mathbb{R}$, the **gradient** is $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left(\frac{\partial}{\partial x_1} f(\boldsymbol{x}) \quad \ldots \quad \frac{\partial}{\partial x_p} f(\boldsymbol{x})\right)^T$.
  - $\nabla_{\boldsymbol{b}}(\boldsymbol{a}^T\boldsymbol{b}) = \boldsymbol{a}$.
  - $\nabla_{\boldsymbol{b}}(\boldsymbol{b}^T\boldsymbol{Ab}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{b}$.
- The **sum of squares function** is $\boldsymbol{Q}(\beta) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$.
  - The **least squares solution** is $\arg\min_{\boldsymbol{\beta}} \boldsymbol{Q}(\boldsymbol{\beta})$.
- The **Normal equations** are $\{\boldsymbol{\beta} : \boldsymbol{X}^T\boldsymbol{X\beta} = \boldsymbol{X}^T\boldsymbol{y}\}$.
  - Equivalent to $\{\boldsymbol{\beta} : \boldsymbol{X\beta} = \boldsymbol{P}_X\boldsymbol{y}\}$.
  - Equivalent to $\{\boldsymbol{\beta} : \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T\boldsymbol{y} - [\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{z}\}$ for some $\boldsymbol{z}$.
  - $X\hat{\beta}$ is invariant to choice of $\hat{\beta}$ that solves the Normal equations.
- The **residual vector** is $\hat{\boldsymbol{e}} := \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.
  - $\hat{\boldsymbol{e}} \in \text{null}(\boldsymbol{X}^T)$.
  - The **sum of squared errors** is $SSE := \|\hat{\boldsymbol{e}}\|_2^2 = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}$.

- Two linear models $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$ and $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{e}$ are **reparameterizations** of each other if $\text{col}(\boldsymbol{X}) = \text{col}(\boldsymbol{W})$.
  - Suppose there are reparameterized design matrices $\boldsymbol{X}$ and $\boldsymbol{W}$. If $\hat{\boldsymbol{\gamma}}$ solves the Normal equations with $\boldsymbol{W}$, then $\hat{\boldsymbol{\beta}} := \boldsymbol{T}\hat{\boldsymbol{\gamma}}$ solves the Normal equations with $\boldsymbol{X}$, where $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{T}$.
- **Gram-Schmidt orthonormalization**: $\boldsymbol{u}_i := (\boldsymbol{I}_n - \sum_{j=1}^{i-1} \boldsymbol{P}_{u_j})\boldsymbol{x}_i = \boldsymbol{x}_i - \sum_{j=1}^{i-1} \frac{\langle \boldsymbol{u}_j, \boldsymbol{x}_i \rangle}{\|\boldsymbol{u}_j\|_2^2} \boldsymbol{u}_j$.
  - Constructs a set of orthonormal vectors from a set of linearly independent vectors.

5.3 Estimability

Return to Table of Contents

- An estimator $t(\boldsymbol{y})$ is **unbiased** for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ iff $\mathbb{E}[t(\boldsymbol{y})] = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.
- An estimator $t(\boldsymbol{y})$ is **linear** if $t(\boldsymbol{y}) = c + \boldsymbol{a}^T\boldsymbol{y}$ for constants c, $\boldsymbol{a}$.
- A function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is **estimable** iff there exists a linear unbiased estimator for it.
  - Under the linear mean model, $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable iff there exists $\boldsymbol{a} : \mathbb{E}(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.
  - $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda} \in \text{col}(\boldsymbol{X}^T)$.
    - Equivalently, $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda} \perp \text{null}(\boldsymbol{X})$.
  - $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable iff we can express $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ as a linear combination of $E(y_i)$.
  - If $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable, then the least squares estimator $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is invariant to the choice of $\hat{\boldsymbol{\beta}}$.
  - The least squares estimator $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ of an estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is a linear unbiased estimator of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$.
  - If $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable in the model with $\boldsymbol{X}$, and $\hat{\boldsymbol{c}}$ solves the Normal equations with $\boldsymbol{W}$, then $\boldsymbol{W}^T\boldsymbol{T}\hat{\boldsymbol{c}}$ is the least squares estimator of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$.
  - If $\boldsymbol{q}^T\boldsymbol{c}$ is estimable in the reparameterized model, then $\boldsymbol{q}^T\boldsymbol{S}\boldsymbol{b}$ is estimable in the original model with least squares estimator $\boldsymbol{q}^T\hat{\boldsymbol{c}}$, where $\hat{\boldsymbol{c}}$ solves the Normal equations with $\boldsymbol{W}$.
- Consider $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$. If $\text{col}(\boldsymbol{X}^T) \cap \text{col}(\boldsymbol{C}^T) = \{0\}$ and $rank(\boldsymbol{C}) = p - rank(\boldsymbol{X})$, then $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C})^{-1}$ exists, and is a generalized inverse for $\boldsymbol{X}^T\boldsymbol{X}$.
  - $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the unique solution to the Normal equations and $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$.
  - $\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{X}^T = \boldsymbol{0}$.
  - $\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T = \boldsymbol{I}$.
- The **restricted model** is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{P}^T\boldsymbol{\beta} = \boldsymbol{\delta}$.
  - The **restricted Normal equations**, or **RNEs**, are $\left\{ \boldsymbol{\beta} : \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{\delta} \end{pmatrix} \right\}$.
  - $\tilde{\boldsymbol{\beta}}$ solves the RNEs if $\boldsymbol{P}^T\tilde{\boldsymbol{\beta}} = \boldsymbol{\delta}$ and $Q(\boldsymbol{\beta}) = Q(\tilde{\boldsymbol{\beta}})$.
  - $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable in the restricted model if there exists c, $\boldsymbol{a}$ such that $\mathbb{E}(c + \boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$ that satisfy $\boldsymbol{P}^T\boldsymbol{\beta} = \boldsymbol{\delta}$.
  - $(c + \boldsymbol{a}^T\boldsymbol{y})$ is estimable for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ in the restricted model iff there exists a $\boldsymbol{d}$ such that $\boldsymbol{\lambda} = \boldsymbol{X}^T\boldsymbol{a} + \boldsymbol{P}\boldsymbol{d}$, and $c = \boldsymbol{d}^T\boldsymbol{\delta}$.
  - If $\hat{\boldsymbol{\beta}}_H$ is the first component of a solution to the RNEs, then $\hat{\boldsymbol{\beta}}_H$ minimizes $Q(\boldsymbol{\beta})$ over the restricted parameter space.
  - If $\hat{\boldsymbol{\beta}}_H$ is the first component of a solution to the RNEs, and $\tilde{\boldsymbol{\beta}}$ satisfies $\boldsymbol{P}^T\tilde{\boldsymbol{\beta}} = \boldsymbol{\delta}$, then $Q(\tilde{\boldsymbol{\beta}}) = Q(\hat{\boldsymbol{\beta}}_H)$ iff $\tilde{\boldsymbol{\beta}}$ is also a part of a solution to the RNEs.

5.4 Gauss-Markov/Aitken Theorem and Model Misspecification

Return to Table of Contents

- Suppose $\boldsymbol{z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\mathbb{E}(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} + \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma})$.
- $\boldsymbol{t}^T\boldsymbol{y}$ is the BLUE for $\mathbb{E}(\boldsymbol{t}^T\boldsymbol{y})$ iff $\boldsymbol{V}\boldsymbol{t} \in \text{col}(\boldsymbol{X})$ for known, positive-definite $\boldsymbol{V}$.
- The **Gauss-Markov model** follows $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$, and $Var(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$.
  - **Gauss-Markov theorem**: Under the Gauss-Markov assumptions, $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{OLS}$ is the BLUE for estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$.
  - Under the Gauss-Markov model, an unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{SSE}{N-r}$.
- The **Aitken equations** are $\{\boldsymbol{\beta} : \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}\}$.
  - $\hat{\sigma}^2_{GLS} = \frac{1}{N-r}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})$.
- The **Aitken model** follows $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$, and $Var(\boldsymbol{u}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known, positive-definite matrix.
  - **Aitken's theorem**: Under the Aitken model, $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{GLS}$ is the BLUE for estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$.
  - Decompose $\boldsymbol{V}$ into positive-definite $\boldsymbol{L}$ and $\boldsymbol{L}'$ using either spectral or Cholesky decomposition.
  - Under the Aitken assumptions, OLS estimators are BLUE for estimable functions if there exists $\boldsymbol{Q}$ such that $\boldsymbol{V}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}$.
  - Under the Aitken model, $\boldsymbol{t}^T\boldsymbol{y}$ is the BLUE for its expectation iff $\boldsymbol{V}\boldsymbol{t} \in \text{col}(\boldsymbol{X})$.
- Suppose we misspecify the model. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{u}$, where $\boldsymbol{\eta}$ are coefficients for missing terms.
  - The least squares estimates for $\boldsymbol{\beta}$ and σ^2 are biased!
    - $\text{Bias}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{OLS}) = \mathbb{E}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{OLS}) - \boldsymbol{\lambda}^T\boldsymbol{\beta} = \boldsymbol{\lambda}^T(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T\boldsymbol{\eta}$.
    - $\mathbb{E}(SSE) = \boldsymbol{\eta}^T(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta} + \sigma^2(N - r)$.
- Suppose we overfit our model with $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{u}$, where $\boldsymbol{X}\boldsymbol{\beta}_2$ is unnecessary.
  - Estimators are still unbiased.
  - Variance of $\hat{\boldsymbol{\beta}}_{OLS}$ increase.
  - Variance of $\hat{\sigma}^2$ only slightly increases (due to *df*).
- **Mean squared error** is $E\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \sigma^2\text{tr}\left\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right\}$ (if unbiased).

  Example: Suppose we have a table with K cells. The data consists of the counts for each cell, which are denoted by $N_1, \ldots, N_q$. Assume that the counts are mutually independent and are generated from Poisson distribution with $\mathbb{E}(N_k) = \mu_k$, $k = 1, \ldots, K$. Let $\boldsymbol{X}$ be a $K \times p$ matrix of rank p. We model the mean parameters using a log-linear model, i.e., we define $\boldsymbol{\eta} = (\log\mu_1, ..., \log\mu_K)^T$, and $\boldsymbol{\eta} = (\log\mu_1, \ldots, \log\mu_q)^T$, then we posit that

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

  where $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector. In other words, we posit that $\boldsymbol{\eta}$ is in $\text{col}(\boldsymbol{X})$, the column space of $\boldsymbol{X}$. For convenience, define the vectors $\boldsymbol{N} = (N_1, ..., N_K)^T$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q)^T$; also denote the vector of ones as j, and assume j is in $\text{col}(\boldsymbol{X})$. Show that $\arg\max_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}} \left[\boldsymbol{N}^T\boldsymbol{\eta} - \boldsymbol{j}^T\boldsymbol{\mu}\right]$, and to find the MLE of $\boldsymbol{\beta}$, which uses the constraint $\boldsymbol{\eta} \in \text{col}(\boldsymbol{X})$, $\hat{\boldsymbol{\mu}}$ must satisfy $\boldsymbol{X}^T\hat{\boldsymbol{\mu}} = \boldsymbol{X}^T\boldsymbol{N}$.

$$\begin{aligned}
L(\boldsymbol{\mu}) &= \prod_{k=1}^{K} \frac{e^{-\mu_k}(\mu_k)^{n_k}}{n_k!} = \prod_{k=1}^{K} \frac{e^{-\mu_k}}{n_k!} \exp\{n_k \log(\mu_k)\} \\
&= \prod_{k=1}^{K} \frac{1}{n_k!} \exp\left\{-\sum_{k=1}^{K}\mu_k + \sum_{k=1}^{K} n_k\eta_k\right\} \\
&= \prod_{k=1}^{K} \frac{1}{n_k!} \exp\left\{-\boldsymbol{j}^T\boldsymbol{\mu} + \boldsymbol{N}^T\boldsymbol{\eta}\right\}; \\
\ell(\boldsymbol{\mu}) &= c - \boldsymbol{j}^T\boldsymbol{\mu} + \boldsymbol{N}^T\boldsymbol{\eta}; \\
\arg\max_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}) &= \arg\max_{\boldsymbol{\mu}} \left[c - \boldsymbol{j}^T\boldsymbol{\mu} + \boldsymbol{N}^T\boldsymbol{\eta}\right] = \arg\max_{\boldsymbol{\mu}} \left[\boldsymbol{N}^T\boldsymbol{\eta} - \boldsymbol{j}^T\boldsymbol{\mu}\right].
\end{aligned}$$

$$\begin{aligned}
\ell(\boldsymbol{\beta}) \propto& \boldsymbol{N}^T\boldsymbol{\eta} - \boldsymbol{j}^T\boldsymbol{\mu} = \boldsymbol{N}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{j}^T\boldsymbol{X}\boldsymbol{\beta};\\
\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) =& \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{N}^T\boldsymbol{X}\boldsymbol{\beta} - \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{j}^T\boldsymbol{X}\boldsymbol{\beta}\\
=& \boldsymbol{N}^T\boldsymbol{X} - \sum_{k=1}^{K}\frac{\partial}{\partial\boldsymbol{\beta}}\exp\left\{x_k^T\boldsymbol{\beta}\right\}\\
=& \boldsymbol{N}^T\boldsymbol{X} - \sum_{k=1}^{K}x_k^T\exp\left\{x_k^T\boldsymbol{\beta}\right\} = \boldsymbol{N}^T\boldsymbol{X} - \sum_{k=1}^{K}x_k^T\mu_k\\
=& \boldsymbol{N}^T\boldsymbol{X} - \hat{\boldsymbol{\mu}}^T\boldsymbol{X} \overset{\text{set}}{=} 0 \implies \boldsymbol{X}^T\hat{\boldsymbol{\mu}} = \boldsymbol{X}^T\boldsymbol{N}. \blacksquare
\end{aligned}$$

5.5 Distributions/General Linear Hypotheses

Return to Table of Contents

- The **moment generating function**, or **MGF**, is $M_{\boldsymbol{X}}(\boldsymbol{t}) = E[e^{\boldsymbol{t}^T\boldsymbol{X}}]$.
  - Must be defined in an open region that contains the origin.
  - The CDFs of two RVs are equal iff the MGFs exist and are equal in an open region that contains the origin.
  - Two or more RVs are mutually independent iff we can express the joint MGF as the product of the marginal MGFs in an open interval containing the origin.
- $\boldsymbol{X}$ has the **multivariate Normal distribution** with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ iff its MGF has the form $\exp\left\{\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right\}$ in an open neighborhood containing the origin.
  - If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{Y} = a + \boldsymbol{B}\boldsymbol{X}$, then $\boldsymbol{Y} \sim \mathcal{N}_q(a + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.
  - Suppose $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is non-singular, then:
    - A nonsingular matrix $\boldsymbol{A}$ exists such that $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$.
    - $\boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$.
    - The PDF is defined as $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}$.
    - $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are jointly independent iff $\boldsymbol{\Sigma}_{ij} = \boldsymbol{0}$ for all $i \neq j$.
  - Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{Y}_1 = \boldsymbol{a}_1 + \boldsymbol{B}_1\boldsymbol{X}$, $\boldsymbol{Y}_2 = \boldsymbol{a}_2 + \boldsymbol{B}_2\boldsymbol{X}$. $\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2$ iff $\boldsymbol{B}_1\boldsymbol{\Sigma}\boldsymbol{B}_2' = \boldsymbol{0}$.
  - Suppose $\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{V}_{11} & \boldsymbol{V}_{12}\\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22}\end{pmatrix}\right)$.

$$(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\right).$$

- Let $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. $\boldsymbol{U} = \boldsymbol{Z}^T\boldsymbol{Z}$ has the **χ^2-distribution** with p degrees of freedom.
  - MGF is $M_{\boldsymbol{U}}(t) = (1 - 2t)^{-p/2}$.
  - PDF is $\frac{u^{(p-2)/2}e^{-u/2}}{\Gamma(p/2)2^{p/2}}$.
- Suppose $(\boldsymbol{U}|J = j) \sim \chi^2_{p+2j}$, where $\boldsymbol{J} \sim Pois(\phi)$. Then, $\boldsymbol{U}$ follows the **non-central χ^2-distribution** with degrees of freedom p and non-centrality parameter ϕ.
  - MGF is $M_{\boldsymbol{U}}(t) = (1 - 2t)^{-p/2}\exp\left\{\frac{2\phi t}{1-2t}\right\}$.
  - If $\boldsymbol{U} \sim \chi^2_p(\phi)$, then $\mathbb{E}(\boldsymbol{U}) = p + 2\phi$ and $Var(\boldsymbol{U}) = 2p + 8\phi$.
  - If $\boldsymbol{U}_i \sim \chi^2_{p_i}(\phi_i)$ are jointly independent, then $\sum_{i=1}^n \boldsymbol{U}_i \sim \chi^2_{\sum_{i=1}^n p_i}(\sum_{i=1}^n \phi_i)$.
  - If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, then $\boldsymbol{U} = \boldsymbol{X}^T\boldsymbol{X} \sim \chi^2_p\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu}\right)$.
  - If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is non-singular, then $\boldsymbol{U} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_p\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$.
- Suppose $\boldsymbol{U}_1 \sim \chi^2_{p_1}(\phi) \perp U_2 \sim \chi^2_{p_2}$. Then, $\frac{U_1/p_1}{U_2/p_2}$ follows the ***F*-distribution** with degrees of freedom p_1, p_2, and non-centrality parameter ϕ.
- Suppose $\boldsymbol{U} \sim \mathcal{N}(\mu, 1) \perp \boldsymbol{V} \sim \chi^2_k$. $\frac{\boldsymbol{U}}{\sqrt{\boldsymbol{V}/k}}$ follows the ***T*-distribution** with degrees of freedom k, and non-centrality parameter μ.

- If $\mu = 0$, then the PDF is $\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}}\left(1+\frac{t^2}{k}\right)^{-(k+1)/2}$.
- If $\boldsymbol{T} \sim t_k(\mu)$, then $\boldsymbol{T}^2 \sim \boldsymbol{F}_{1,k}\left(\frac{1}{2}\mu^2\right)$.

- If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, and $\boldsymbol{A}$ be symmetric and idempotent with rank s, then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}\right)$.
- Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be symmetric with rank s. If $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$, then $\boldsymbol{B}\boldsymbol{X} \perp \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
- Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be symmetric with rank r, and $\boldsymbol{B}$ be symmetric with rank s. If $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$, then $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} \perp \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
- Given the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{u} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$, the distribution of the BLUE of estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is $N(\boldsymbol{\lambda}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{\lambda}^T(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{\Lambda})$.
  - The BLUE is independent of $\frac{SSE}{\sigma^2}$.
  - The unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{SSE}{N-r}$.
  - $T(\boldsymbol{y}) = (\boldsymbol{y}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{y})$ is a complete and sufficient statistic.
    - $(\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{y})$ is also minimal sufficient.
  - The least squares estimator of an estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ has the smallest variance of any estimator for its expectation.
  - $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is the MLE for an estimable $\boldsymbol{\lambda}^T\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}$ solves the Normal equations.
  - $(\hat{\boldsymbol{\beta}}, SSE/N)$ is an MLE of $(\boldsymbol{\beta}, \sigma^2)$, where $\hat{\boldsymbol{\beta}}$ solves the Normal equations.
- The general linear hypothesis $H_0 : \boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m}$ is **testable** iff $\boldsymbol{K} \in \mathbb{R}^{q\times s}$ has full-column rank, and each column of $\boldsymbol{K}^T\boldsymbol{\beta}$ is estimable.
  - We can test $H_0 : \boldsymbol{\beta} \in \text{col}(\boldsymbol{B})$ by constructing basis vectors for $\text{col}(\boldsymbol{B})^\perp$, and setting $\boldsymbol{m} = \boldsymbol{0}$.
  - If $\boldsymbol{K}^T\boldsymbol{\beta}$ is estimable, then $\boldsymbol{H} = \boldsymbol{K}^T(\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{K}$ is non-singular.
    - A result is $(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m}) \sim \chi^2_s\left(\frac{1}{2}(\boldsymbol{K}^T\boldsymbol{\beta} - \boldsymbol{m})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{K}^T\boldsymbol{\beta} - \boldsymbol{m})\right)$.
    - Therefore, $F = \frac{(\boldsymbol{K}^T\hat{\boldsymbol{\beta}}-\boldsymbol{m})^T\boldsymbol{H}^{-1}(\boldsymbol{K}^T\hat{\boldsymbol{\beta}}-\boldsymbol{m})/s}{SSE/(N-r)} \sim F_{s,N-r}\left(\frac{1}{2}(\boldsymbol{K}^T\boldsymbol{\beta} - \boldsymbol{m})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{K}^T\boldsymbol{\beta} - \boldsymbol{m})\right)$.
      - Note that σ^2 in the above term was cancelled out.
      - $r = rank(\boldsymbol{X})$.

Example: An experiment randomizes $n = 11$ units to 9 combinations of two factors, x_1 and x_2, which populate the 2nd and 3rd column of the design matrix $\boldsymbol{X}$. $\boldsymbol{X}^T\boldsymbol{X}$, and $\boldsymbol{X}^T\boldsymbol{y}$ are given below:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 11 & & \\ & 34 & \\ & & 34 \end{bmatrix}, \boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 220 \\ 34 \\ -68 \end{bmatrix}.$$

Suppose $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Consider a linear regression model of the form

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}.$$

a. Report the least squares estimate of Y given $x_1 = x_2 = 2$.
b. Noting that $\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = 96$, report a standard error for the estimate in part (a).
c. Conduct a test of $H_0 : \beta_1 + \beta_2 = 0$ at a significance level of $\alpha = 0.05$.
d. Find the values of $\hat{Y}_L$ and $\hat{Y}_H$ such that

$$0.95 = P(\hat{Y}_L < Y < \hat{Y}_H),$$

where $x_1 = x_2 = 2$.

a.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS} =& (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ =& \begin{bmatrix} 1/11 & & \\ & 1/34 & \\ & & 1/34 \end{bmatrix}\begin{bmatrix} 220 \\ 34 \\ -68 \end{bmatrix} = \begin{bmatrix} 20 \\ 1 \\ -2 \end{bmatrix};\end{aligned}$$

$$\mathbb{E}(Y|x_1 = 2, x_2 = 2) = 20 + 1(2) - 2(2) = 18.$$

b. $\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = SSE = 96$; also note that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ij} = 0$ for $i \neq j$, so the estimates of the β components are uncorrelated.

$$\begin{aligned}
Var\left[\mathbb{E}(Y|x_1 = 2, x_2 = 2)\right] =& Var(\hat{\beta}_0 + 2\hat{\beta}_1 + 2\hat{\beta}_2) \\
\overset{\text{uncorrelated}}{=}& Var(\hat{\beta}_0) + 4Var(\hat{\beta}_1) + 4Var(\hat{\beta}_2) \\
=& \hat{\sigma}^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{00} + 4(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{11} + 4(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{22}\right] \\
=& \frac{96}{11-3}\left[\frac{1}{11} + 4\cdot\frac{1}{34} + 4\cdot\frac{1}{34}\right] = 3.9144; \\
SE\left[\mathbb{E}(Y|x_1 = 2, x_2 = 2)\right] =& \sqrt{Var\left[\mathbb{E}(Y|x_1 = 2, x_2 = 2)\right]} = 1.9785.
\end{aligned}$$

c. Construct linear hypotheses; $\boldsymbol{K}^T = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$, with $\boldsymbol{m} = 0$; $\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m} = -1$.

$$\begin{aligned}
F^* =& \frac{(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m})^T\boldsymbol{H}^{-1}(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m})/s}{MSE} \\
=& \frac{(-1)^T[\boldsymbol{K}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(-1)/1}{96/(11-3)} \\
=& \frac{\left(\frac{1}{17}\right)^{-1}}{12} = 1.4167 \overset{H_0}{\sim} F_{1,8};
\end{aligned}$$

p-value= $P(F_{1,8} > F^*) > 0.05$, therefore we fail to reject H_0.

d. The problem is essentially asking for a 95% prediction interval.

$$\begin{aligned}
\text{CI} =& \hat{y} \pm t_{df_E, 0.025}\sqrt{\hat{\sigma}^2\left(1 + Var(E(Y|x_1 = 2, x_2 = 2))\right)} \\
=& 18 \pm 2.306\sqrt{12(1 + 1.9785^2)} = (\underbrace{8.75}_{\hat{Y}_L}, \underbrace{27.25}_{\hat{Y}_H}). \ \blacksquare
\end{aligned}$$

- Under the Normal Gauss-Markov assumptions, suppose now we want to carry out a **likelihood ratio test**, or **LRT** for an estimable $\boldsymbol{K}^T\boldsymbol{\beta}$.
  - The parameter space under $H_0 : \boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m}$ is $\Omega_0 = \{(\boldsymbol{\beta}, \sigma^2) : \boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m}, \sigma^2 > 0\}$.
  - The union of the parameter space under H_0 and $H_1 : \boldsymbol{K}^T\boldsymbol{\beta} \neq \boldsymbol{m}$ is $\Omega = \{(\boldsymbol{\beta}, \sigma^2) : \boldsymbol{\beta} \in \mathbb{R}^p, \sigma^2 > 0\}$.
  - The **likelihood ratio** is $\phi(\boldsymbol{y}) = \frac{\max_{\Omega_0} L(\boldsymbol{\beta}, \sigma^2)}{\max_{\Omega} L(\boldsymbol{\beta}, \sigma^2)}$, rejecting when $\phi(\boldsymbol{y}) < c$ for some c.
    - Finding c is tricky in this form, but we can use MLE and algebra to get that

$$\phi(\boldsymbol{y}) = \frac{[Q(\hat{\boldsymbol{\beta}}_H) - Q(\hat{\boldsymbol{\beta}})]/s}{Q(\hat{\boldsymbol{\beta}})/(N-r)} > \frac{N-r}{s}(c^{-2/N} - 1).$$

- If $\boldsymbol{K}^T\boldsymbol{\beta}$ is a set of linearly independent estimable functions, and $\hat{\boldsymbol{\beta}}_H$ is a part of a solution to the RNEs with constraint $\boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m}$, then $Q(\hat{\boldsymbol{\beta}}_H) - Q(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}}) = (\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m})^T\boldsymbol{H}^{-1}(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m})$.
- If $\boldsymbol{K}^T\boldsymbol{\beta}$ is a set of linearly independent estimable functions, and $\hat{\boldsymbol{\beta}}$ is a solution to the Normal equations, then we can find $\hat{\boldsymbol{\beta}}_H$, a part of a solution to the RNEs with constraint $\boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m}$, by solving for $\boldsymbol{\beta}$ in

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{K}\boldsymbol{H}^{-1}(\boldsymbol{K}^T\hat{\boldsymbol{\beta}} - \boldsymbol{m}).$$

- $\boldsymbol{P}^T\boldsymbol{\beta}$ is **jointly nonestimable** if no linear combination of $\boldsymbol{P}^T\boldsymbol{\beta}$ is estimable.
- If $\boldsymbol{P}^T\boldsymbol{\beta}$ is a set of linearly independent, jointly nonestimable functions, and $\hat{\boldsymbol{\beta}}_H$ is a part of a solution to the RNEs with constraint $\boldsymbol{P}^T\boldsymbol{\beta} = \boldsymbol{\delta}$, then $Q(\hat{\boldsymbol{\beta}}_H) = Q(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{0}$, where $\hat{\boldsymbol{\theta}}$ is the Lagrange multiplier.
- Define $\tau_j := \boldsymbol{\lambda}_j^T\boldsymbol{\beta}$. We can then construct one-at-a-time CIs $\hat{\tau}_j \pm t_{N-r,\alpha/2}\sqrt{\hat{\sigma}^2\boldsymbol{H}_{jj}}$.
  - **Bonferroni method**: Replace $t_{N-r,\alpha/2}$ with $t_{N-r,\alpha(2s)}$, where s is the number of intervals.
    - Number of intervals needs to be specified in advance.
  - **Scheffé method**: Construct a CI for any linear combination $\boldsymbol{u}'\tau : \boldsymbol{u}'\hat{\tau} \pm \sqrt{\hat{\sigma}^2 s F_{s,N-r,\alpha}\boldsymbol{u}'\boldsymbol{H}\boldsymbol{u}}$.
    - Number of intervals does not need to be specified in advance.
    - Intervals are often larger than other methods.

- **Tukey method**: Let Z_i be iid $\mathcal{N}(0,1)$ RVs for $i \in \{1,\dots,k\}$, and let $U \sim \chi^2_v \perp Z_i$. Then, $Q = \frac{Z_{(k)}-Z_{(1)}}{\sqrt{U/v}}$. Then, $(\bar{y}_i - \bar{y}_j) \pm \frac{\hat{\sigma}}{\sqrt{n}} q^*_{a,n(a-1)}$ is the CI.
  - Use only with balanced, one-way ANOVA models, and testing for pairwise differences.
  - If $|\tau_i - \tau_j| \le h$ for all i, j, and $\sum_i u_i = 0$ (a contrast), then $|\sum_i u_i \tau_i| \le h \cdot \frac{1}{2}\sum_i |u_i|$.
    - Lets us extend the Tukey intervals to cover all contrasts.
  - **Tukey-Kramer method** extends the Tukey method to unbalanced designs,

$$(\bar{y}_i - \bar{y}_j) \pm q^*_{a,N-a}\sqrt{\hat{\sigma}^2 \cdot \frac{n_i^{-1} + n_j^{-1}}{2}}.$$

- Two parameter vectors $\theta^{(1)}$ and $\theta^{(2)}$ are **observationally equivalent**, denoted $\theta^{(1)} \sim \theta^{(2)}$, iff the distribution of the response is the same for both parameter vectors.
  - $\theta^{(1)} \sim \theta^{(2)}$ if there does not exist an A such that $P(\boldsymbol{y} \in A|\theta^{(1)}) \neq P(\boldsymbol{y} \in A|\theta^{(2)})$.
  - A function $g(\theta)$ is an **identifying function** iff $g(\theta^{(1)}) = g(\theta^{(2)})$ iff $\theta^{(1)} \sim \theta^{(2)}$.
  - A function $g(\theta)$ is **identified** iff $\theta^{(1)} \sim \theta^{(2)} \implies g(\theta^{(1)}) = g(\theta^{(2)})$.
    - If $g(\theta^{(1)}) \neq g(\theta^{(2)})$, then the distributions are different.
- A family of distributions $F(\boldsymbol{y}|\theta)$ is a **location family** with location parameter θ if $F(\boldsymbol{y}|\theta) = F_0(\boldsymbol{y} - \theta)$ for some distribution F_0.

5.6 Cochran's Theorem

Return to Table of Contents

- A $p \times p$ symmetric matrix $\boldsymbol{A}$ is idempotent with rank s iff there exists a $p \times s$ matrix $\boldsymbol{G}$ with orthonormal columns such that $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}'$.
- If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is non-singular, and $\boldsymbol{A}$ be symmetric such that $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent with rank s, then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}\right)$.
  - For $\boldsymbol{y} \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_N)$, we can set $\boldsymbol{A} = \frac{1}{\sigma^2}(\boldsymbol{I} - \boldsymbol{P}_X)$ to get that $\frac{SSE}{\sigma^2} \sim \chi^2_{N-r}$.
- **Cochran's theorem**: Suppose $\boldsymbol{y} \sim \mathcal{N}_N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_n)$, and let $\boldsymbol{A}_i$ be symmetric, idempotent matrices with rank s_i. If $\sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_N$, then $\frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{A}_i\boldsymbol{y}$ are independently distributed as $\chi^2_{s_i}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}^T\boldsymbol{A}_i\boldsymbol{\mu}\right)$, and $\sum_{i=1}^k s_i = N$.
- ANOVA table for SSQ: Define $\boldsymbol{X}^*_j := [\boldsymbol{X}_0|\boldsymbol{X}_1|\dots|\boldsymbol{X}_j]$, and $R(\boldsymbol{b}_j, \dots) = R(\boldsymbol{b}_0, \dots, \boldsymbol{b}_j) = \boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{X}^*_j}\boldsymbol{y}$.

Source	*df*	Projection	SSQ	ncp
$\boldsymbol{b}_0$	$r(\boldsymbol{X}_0)$	$\boldsymbol{P}_{\boldsymbol{X}_0}$	$R(\boldsymbol{b}_0)$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{P}_{\boldsymbol{X}_0}(\boldsymbol{X}\boldsymbol{b})$
$\boldsymbol{b}_1$ after $\boldsymbol{b}_0$	$r(\boldsymbol{X}^*_1) - r(\boldsymbol{X}_0)$	$\boldsymbol{P}_{\boldsymbol{X}^*_1} - \boldsymbol{P}_{\boldsymbol{X}_0}$	$R(\boldsymbol{b}_0, \boldsymbol{b}_1) - R(\boldsymbol{b}_0)$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{P}_{\boldsymbol{X}^*_1} - \boldsymbol{P}_{\boldsymbol{X}_0})(\boldsymbol{X}\boldsymbol{b})$
...				
$\boldsymbol{b}_j$ after $\boldsymbol{b}_0, \dots, \boldsymbol{b}_{j-1}$	$r(\boldsymbol{X}^*_j) - r(\boldsymbol{X}^*_{j-1})$	$\boldsymbol{P}_{\boldsymbol{X}^*_j} - \boldsymbol{P}_{\boldsymbol{X}^*_{j-1}}$	$R(\boldsymbol{b}_j, \dots) - R(\boldsymbol{b}_{j-1}, \dots)$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{P}_{\boldsymbol{X}^*_j} - \boldsymbol{P}_{\boldsymbol{X}^*_{j-1}})(\boldsymbol{X}\boldsymbol{b})$
...				
$\boldsymbol{b}_k$ after $\boldsymbol{b}_0, \dots, \boldsymbol{b}_{k-1}$	$r(\boldsymbol{X}^*_k) - r(\boldsymbol{X}^*_{k-1})$	$\boldsymbol{P}_{\boldsymbol{X}^*_k} - \boldsymbol{P}_{\boldsymbol{X}^*_{k-1}}$	$R(\boldsymbol{b}_k, \dots) - R(\boldsymbol{b}_{k-1}, \dots)$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}^*_{k-1}})(\boldsymbol{X}\boldsymbol{b})$
Error	$N - r(\boldsymbol{X})$	$\boldsymbol{I} - \boldsymbol{P}_X$	$\boldsymbol{y}^T\boldsymbol{y} - R(\boldsymbol{b})$	0
Total	N	$\boldsymbol{I}$	$\boldsymbol{y}^T\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{X}\boldsymbol{b})$

- $\frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{A}_j\boldsymbol{y} \sim \chi^2_{r_j}\left(\frac{(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_j(\boldsymbol{X}\boldsymbol{b})}{2\sigma^2}\right)$, where $A_j := \begin{cases} \boldsymbol{P}_{\boldsymbol{X}_0},\ j = 0 \\ \boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}^*_{k-1}},\ j = k \\ \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}},\ j = k+1 \\ \boldsymbol{P}_{\boldsymbol{X}^*_j} - \boldsymbol{P}_{\boldsymbol{X}^*_{j-1}}, \text{ otherwise} \end{cases}$ with rank r_j.

5.7 Variance Component Estimation

Return to Table of Contents

- Moving from fixed effects to random effects is that SSM, SSA, and SSE may no longer be mutually independent.
  - Since the conditional distribution of SSE (given α_i) does not depend on α_i, SSE is still independent of α_i, and remains a central χ^2 (when divided by σ^2).
- For a balanced one-way ANOVA model with a random effect, $\frac{SSA}{\sigma^2+n\sigma_A^2} \sim \chi^2_{a-1}$.
- For two-way models, $\frac{SS\text{Source}}{E(SS\text{Source})}$ still forms an independent central $\chi^2_{df_{\text{Source}}}$ distribution.
- The SSq decomposition for split plots is different than a two-factor with interaction, because the correlation structure is different, and this can be thought of as a mixed, crossed model with two variance components.

6 ST 740: Bayesian Statistical Inference

Instructor: Dr. Sujit Ghosh
Semester: Fall 2024
Main Textbook: Ghosh and Reich, *Bayestian Statistical Methods*

6.1 Basics of Bayesian Inference

Return to Table of Contents

- **Bayes' Rule**: $f(x|y) = \frac{f(y|x)f(x)}{f(y)}$.
- In Bayesian statistics, we treat the parameter(s) θ as a random variable.
  - This means that θ has a distribution associated with it.
  - A Bayesian asks what is the probability of the hypothesis, given our data.
- **Prior Distribution**, or $\boldsymbol{\pi(\theta)}$: Represents our uncertainty about the parameters of interest before we observe the data.
  - If we have some knowledge about θ, then we should choose a corresponding prior.
  - Unless the prior information is null (ie. when $\pi(\theta) = c$), additional "Bayes learning" is gained in addition to "data learning."
- **Likelihood Function**, or $\boldsymbol{f(Y|\theta)}$: Links the data with the parameter of interest.
- **Marginal Density**, or $\boldsymbol{m(Y)}$: $m(\underline{Y}) = \int f(\underline{Y}|\theta)\pi(\theta)d\theta$.
  - Ensures that the posterior distribution is a valid distribution (integrates to one).
  - Does not depend on the parameter of interest.
- **Posterior Distribution**, or $\boldsymbol{\pi(\theta|Y)}$: $\pi(\theta|\underline{Y}) = \frac{f(\underline{Y}|\theta)\pi(\theta)}{m(\underline{Y})} \propto f(\underline{Y}|\theta)\pi(\theta)$.
  - Quantifies the uncertainty of the parameters of interest after we observe some data and account for prior knowledge.
  - Requires $m(\underline{Y}) > 0$.
  - The posterior distribution depends on data only through sufficient statistics.
  - The posterior of a function $\eta = \eta(\theta)$ can be obtained by usual transformation methods (CDF method, Jacobian, etc.).
- **Sequential Bayesian Learning**: Define $\underline{Y}_q := (Y_1 \dots Y_q)$ for $q \in \{1, 2, \dots, \}$.

$$\pi_k(\theta|\underline{Y}_k) = \frac{f(y_k|\underline{Y}_{k-1};\theta)\pi_{k-1}(\theta|\underline{Y}_{k-1})}{\int f(y_k|\underline{Y}_{k-1};\theta)\pi_{k-1}(\theta|\underline{Y}_{k-1})d\theta}.$$

  - The posterior distribution based on $\underline{Y}_{k-1}$ becomes the prior for the following posterior distribution based on $\underline{Y}_k$.
  - The prior variances decrease as k increases.
    - This means that the posterior uncertainty about θ decreases as we collect more data.
- **Bayes Estimator**: $\hat{\theta}_{Bayes} = E_\pi\left[L(\theta, \hat{\theta})\right]$, where π relates to the posterior.
  - $\hat{\theta}_{Bayes} = E_\theta\left[\pi(\theta|\underline{Y})\right]$ (mean of the posterior) under squared error loss.
  - Is biased under squared error loss.
  - Bayes estimators exist under mild conditions on the loss functions.
  - Bayes estimators are unique for strictly convex functions.
  - Depends on the choice of parameterization.
    - **Intrinsic Losses**, or $\boldsymbol{L(\theta, d)}$: $K(f(\cdot|\theta), f(\cdot|d))$, where K is some distance measure between $f(\cdot|\theta)$ and $f(\cdot|d)$.
      - Ex. **Entropy Loss**: $L(\theta, d) = \mathbb{E}\left[\log\left(\frac{f(x|\theta)}{f(x|d)}\right)\right]$.
    - Intrinsic loss functions let us obtain parameterization-invariant Bayes estimators.

- **Maximum a Posteriori Estimator**, or **MAP Estimator**: $\hat{\theta}_{MAP} = \arg\max_\theta \log(\pi(\theta|\underline{Y}))$.
  - Is the mode of the posterior distribution.
- **100(1 − α)% Credible Interval**: Any interval (l, u) such that $P(l < \theta < U|\underline{Y}) = 1 - \alpha$.
  - There are an infinite number of these intervals.
    - One choice of interval is the equal-tailed interval, where l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the posterior.
    - **Highest Posterior Density Interval**, or **HPD Interval**: A $100(1-\alpha)\%$ credible interval (l, u) for a given posterior that minimizes $u - l$.
  - If (l, u) is a $100(1 - \alpha)\%$ credible interval, then given our prior and observed data, we are 95% sure that θ is between l and u.
    - Is less nuanced than the definition of a confidence interval used in frequentist estimation.
    - We relate the probability to the posterior distribution whereas in CIs we relate to the likelihoods.
- We can also use posterior probabilities to conduct hypothesis tests.
  - We calculate the probabilities of θ being under H_0 and H_1, and we can thus compare these probabilities.
  - Reject H_0 iff $\int_{\Theta_0} K(\theta; X)d\theta < \alpha \int_{\Theta_1} K(\theta; X)d\theta$, where $K(\theta; X)$ is some distance measure.
- While graphical summaries are helpful for posterior densities in low dimensions, we cannot use plots in a helpful for high-dimensional parameter spaces without making some adjustments.
  - One strategy is to marginalize out other parameters one-at-a-time, and provide univariate summaries for each parameter.
    - This can still not be very helpful for very large dimensions.
  - Another option is MCMC (discussed later).
- **Posterior Predictive Distribution**, or **PPD**: The distribution of an outcome of Y, given the data.
  - $Y^*|\underline{Y} \sim f^*(Y^*|\underline{Y}) = \int f(Y^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\underline{Y})d\boldsymbol{\theta}$, where f is the likelihood.
  - In a parametric model, PPD accounts for uncertainty in the model parameters.
  - $Var(Y^*|\underline{Y}) \geq Var(Y^*|\underline{Y}, \boldsymbol{\theta})$.

6.2 Bayesian Inference

Return to Table of Contents

- **Decision Theoretical Framework**, or **DTF**: Consists of a sample space, a parameter space Θ, and a decision space $\mathcal{D}$.
  - **Loss Function**, or $\boldsymbol{L(\theta, \delta)}$: Evaluates the penalty associated with the decision δ when the parameter takes the value θ.
    - Is often some distance metric, like a norm.
    - Determination of the loss function is often awkward in practice.
  - For most problems, $\mathcal{D} = g(\Theta)$ for some arbitrary function g.
  - The Bayesian DTF is based on the rigorous determination of the sampling distribution, prior distribution, and loss function.
    - It is generally impossible to uniformly minimize $L(\theta, \delta)$ when θ is unknown.
- **Frequentist Risk**: $R(\theta, \delta) = \mathbb{E}[L(\theta, \delta(\boldsymbol{x}))] = \int L(\theta, \delta(x)) f(x|\theta)dx$.
- **Bayesian Risk**: $r(\theta, \delta) = \mathbb{E}_\pi[L(\theta, \delta(x)] = \int\int L(\theta, \delta(x)) f(x|\theta)\pi(\theta)dxd\theta$.
  - Is the expected loss with respect to the posterior distribution of θ.
  - Loss functions depend on the true value of θ, and so we can't evaluate it in real data analysis.
  - **Bayes Estimator**: The estimator $\hat{\theta}(\underline{Y})$ that minimizes Bayesian risk.
    - **Generalized Bayes Estimator**: The estimator that minimizes the posterior expected loss with an improper prior.
    - Bayes estimators exist under somewhat loose conditions on the loss functions.

    - ∗ Bayes estimators are unique for strictly convex loss functions (wrt δ).
    - ∗ Depends on the choice of parameterization (as opposed to MLE).
    - ∗ **Intrinsic Losses**: $L(\theta, \delta) = K\left(f(\cdot|\theta), f(\cdot|\delta)\right)$, where K is a distance metric.
  - Under squared error loss, the posterior mean minimizes Bayesian risk.
  - Under absolute loss, the posterior median minimizes Bayesian risk.
  - Given a prior, the Bayes risk can compare estimators without additional assumptions (as opposed to frequentist risk).

- A common approach to estimate estimator performance is

$$MSE[\hat{\theta}(\underline{Y})] = Bias[\hat{\theta}(\underline{Y})]^2 + Var[\hat{\theta}(\underline{Y})].$$

  - This calculation depends on θ and n, which could result in some complications.
  - Adding prior information can reduce variance, but may result in an increase in bias if this information is erroneous.

- We can use frequentist methods to evaluate coverage probability of credible sets.

- The Bayesian CLT means that any choice of (a reasonable) prior will lead to the same conclusions, and that the posterior will converge to the true value.

- We can use Monte Carlo to simulate complicated sampling distributions.

- A sequence of posterior densities $p(\cdot|\underline{x})$ is **consistent** at $\theta_0 \in \Theta$ if, for every neighborhood $\mathcal{N}$ of θ_0, the posterior probability $P\left(\theta \in \mathcal{N}|\underline{x}\right) \overset{\text{a.s.}}{\to} 1$, given $x_i \overset{\text{iid}}{\sim} f(x|\theta_0)$.
  - As $n \to \infty$, the posterior distribution concentrates all its mass around the 'true' θ_0.
  - To define neighborhoods, we need some distance metric D (such as norms).
  - We can establish posterior consistency if we can show that, $\forall \epsilon > 0$, $P\left(D(\theta, \theta_0) \geq \epsilon|\underline{x}\right) \overset{\text{a.s.}}{\to} 0$.
  - We assume that $P(\theta \in \mathcal{N}) > 0$ for almost all subsets $\mathcal{N} \subseteq \Theta$.

- Under a set of suitable regularity conditions, one can show that if $x_i \overset{\text{iid}}{\sim} f(x|\theta_0)$, then

$$\lim_{n\to\infty} \left\{ \sup_{t\in\Theta} |P\left[\sqrt{n}(\theta - \hat{\theta}) \preceq t|\underline{x}\right] - \phi(t, 0, I(\theta_0)^{-1})| \right\} = 0.$$

  In other words,

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\text{d}}{\to} \mathcal{N}_d\left(0, I(\theta)^{-1}\right)$$

  - This means that, roughly speaking, $\pi(\theta|\underline{Y}) \dot{\sim} \mathcal{N}\left(\hat{\theta}, (nI(\hat{\theta}))^{-1}\right)$.
    - ∗ Assumes that $\pi(\theta)$ is continuous and positive on Θ.
  - This means that Bayes estimates are asymptotically efficient.
  - This also means that Bayesian and frequentist inferences merge asymptotically.
  - This establishes that the posterior is asymptotically insensitive to the prior distribution (given a semi-appropriate choice of prior). In other words, for two given priors $\pi_1(\theta)$ and $\pi_2(\theta)$,

$$\int |\pi_1(\theta|\underline{x}) - \pi_2(\theta|\underline{x})| d\theta \overset{\text{a.s.}}{\to} 0.$$

6.3 Prior Distributions

Return to Table of Contents

- There is no sense of an "optimal" prior.

- **Conjugate**: A prior and likelihood pair are conjugate if the posterior distribution is the same family as the prior.
  - Updating the posterior only changes the posterior parameters.
  - Conjugate priors are not unique (for instance, the beta prior is conjugate for the binomial and negative binomial likelihoods).

– This doesn't happen often, so conjugate priors often don't exist.
– Conjugate priors have somewhat easy interpretations of the prior/data's impact on the posterior.
– Conjugate priors are mainly used in limited information cases, since they only call for the determination of a few parameters.
– A conjugate family by no means minimizes or maximizes prior information.
– Is often easier to find a conjugate family if the sampling model $f(x|\theta)$ has a sufficient statistic of constant dimension.
  * If a family of distributions has a sufficient statistic of constant dimension whose support doesn't depend on θ, then the family is exponential.
– **Exponential Family**: $f(x|\theta) = h(x)\exp\{\eta(\theta)'T(x) - b(\theta)\}$, and the support doesn't depend on θ.
  * The conjugate prior for a canonical exponential family is $\pi(\eta|\mu,\lambda) \propto \exp\{\eta'\mu - \lambda\psi(\eta)\}$.
  * The posterior is exponential family $\equiv$ the likelihood is exponential family.
  * If η has a natural conjugate prior, then $\mathbb{E}[\mathbb{E}(y|\eta)] = \mathbb{E}\left[\frac{\partial\psi(\eta)}{\partial\eta}\right] = \frac{\mu}{\lambda}$, and $\mathbb{E}\left[\frac{\partial\psi(\eta)}{\partial\eta}|\bar{y}\right] = \frac{\mu+n\bar{y}}{\lambda+n}$.

Example: Suppose $Y_i|\theta \overset{\text{iid}}{\sim} f(y|\eta) = \exp\{y\eta - \psi(\eta)\}h(y)$ for $i \in \{1,\dots,n\}$. We are interested in estimating $\mu = \psi'(\eta)$ under squared error loss.

1. Obtain the MLE for μ. Is it unbiased?
   Use the invariance property of the MLE. That is, $\hat{\mu} = \psi'(\hat{\eta})$.

$$\begin{aligned}
L(\eta) =& \prod_{i=1}^{n}\exp\{Y_i\eta - \psi(\eta)\}h(Y_i)\\
=& \exp\left\{\eta\sum_{i=1}^{n}Y_i - n\psi(\eta)\right\}\prod_{i=1}^{n}h(Y_i);\\
\ell(\eta) =& \eta\sum_{i=1}^{n}Y_i - n\psi(\eta) + c\\
=& \eta n\bar{Y} - n\psi(\eta) + c;\\
\ell'(\eta) =& n\bar{Y} - n\psi'(\eta) \overset{\text{set}}{=} 0 \implies \hat{\eta} = (\psi')^{-1}(\bar{Y});\\
\hat{\mu} = \psi'(\hat{\eta}) =& \psi'\left((\psi')^{-1}(\bar{Y})\right) = \bar{Y}.
\end{aligned}$$

$$\mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}(Y_i|\eta)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\theta) = \frac{1}{n}(n\mu) = \mu.$$

   Therefore, $\hat{\mu}$ is unbiased for μ.
2. Find the class of natural conjugate priors for η.

$$\begin{aligned}
\pi(\eta|Y) \propto& \exp\{y\eta - \psi(\eta)\}\cdot\exp\{\eta\theta - \lambda\psi(\eta)\}\\
=& \exp\{(y+\theta)\eta - (\lambda+1)\psi(\eta)\}h(Y);
\end{aligned}$$

   Thus, the exponential family is the class of priors for η.
3. Obtain the Bayes estimator of μ under squared error loss, using a conjugate prior.

$$\begin{aligned}
\pi(\eta|y) =& K(\eta,y,\theta,\lambda)\exp\{(\theta+n\bar{y})\eta - (\lambda+n)\psi(\eta)\}\\
d\pi(\eta|y) =& \pi(\eta|y)\{(\theta+n\bar{y})\eta - (\lambda+n)\psi'(\eta)\}d\eta\\
\int d\pi(\eta|y)d\eta =& \int \pi(\eta|y)\{(\theta+n\bar{y})\eta - (\lambda+n)\psi'(\eta)\}d\eta\\
d\left(\int\pi(\eta|y)d\eta\right) =& \int\pi(\eta|y)(\theta+n\bar{y})\eta d\eta - \int\pi(\eta|y)(\lambda+n)\psi'(\eta)d\eta\\
\frac{\partial}{\partial\eta}(1) =& (\theta+n\bar{y})(1) - (\lambda+n)\int\pi(\eta|y)\psi'(\eta)d\eta\\
0 =& (\theta+n\bar{y}) - (\lambda+n)\mathbb{E}[\psi'(\eta)];\\
\hat{\mu}_{Bayes} = \mathbb{E}[\psi'(\eta)] =& \frac{\theta+n\bar{y}}{\lambda+n}.
\end{aligned}$$

4. Is there any (proper) conjugate prior for which the Bayes estimator is unbiased?

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_{Bayes}\right] =& \mathbb{E}\left[\frac{\theta + n\bar{Y}}{\lambda + n}\right] \\ =& \frac{\theta}{\lambda+n} + \frac{1}{\lambda+n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}(Y_i|\eta)\right] \\ =& \frac{1}{\lambda+n}(\theta + n\mu).\end{aligned}$$

For this estimator to be unbiased, $\theta = \lambda = 0$. However, since $\lambda > 0$, the Bayes estimator will never be unbiased for μ.

5. Obtain the Bayes estimator of μ under weighted squared error loss, using a conjugate prior,

$$L(\eta, \delta) = e^{\psi(\eta)}(\eta - \delta)^2.$$

Is this unbiased for any value of θ and λ?

$$\begin{aligned}\hat{\mu}_{Bayes} = E_\pi\left[L((\mu,\delta))\right] &= \int_\eta e^{\psi(\eta)}(\eta-\delta)^2 \prod_{i=1}^{n} h(Y_i)\exp\left\{y\eta - \psi(\eta)\right\} \times h(\eta)\exp\left\{\eta\theta - \lambda\psi(\eta)\right\} d\eta \\ &= \int_\eta (\eta-\delta)^2 h(\eta)\prod_{i=1}^{n} h(Y_i)\exp\left\{(n\bar{y}+\theta)\eta - (\lambda+n-1)\psi(\eta)\right\} d\eta;\end{aligned}$$

This is equal to the ordinary squared error loss under an exponential family. Using the formula given above,

$$\hat{\mu}_{Bayes} = \frac{n\bar{y}+\theta}{\lambda+n-1}.$$

This can be unbiased when $\theta = 0$ and $\lambda = 1$. ■

- **Natural Conjugate Prior**: Let $X_i \overset{\text{iid}}{\sim} f(x|\theta)$. The natural conjugate prior is given by $\pi(\theta) \propto \prod_{j=1}^{m} f(x_j^o|\theta)$, where x_i^o and m are fixed parameters of the prior distribution such that m is large enough such that the corresponding integral is finite.
  - Useful when the sampling distribution has no sufficient statistic of finite dimension.
  - Corresponds to an update of a flat prior θ for m virtual observations $x_1^o, \ldots, x_m^o$ from $f(x|\theta)$.
  - Mixtures of natural conjugate priors can approximate any prior distribution.
    Example: Normal distribution with fixed variance.

$$\begin{aligned}f(y|\mu) &\propto \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}; \\ \pi(\mu|\underline{y}^o) &\propto \exp\left\{-\sum_{j=1}^{m}\frac{(y_j^o-\mu)^2}{2\sigma^2}\right\} \propto \exp\left\{-\frac{m(\mu-\bar{y}^o)^2}{\sigma^2}\right\}; \\ \theta &\sim N\left(\bar{y}^o, \frac{\sigma^2}{m}\right). \blacksquare\end{aligned}$$

Example: Poisson counts. Suppose $Y|\theta \sim \text{Pois}(\theta) \implies f(Y|\theta) \propto e^{-\theta}\theta^y$.
Using the definition of a natural conjugate prior, choose

$$\pi(\theta|\underline{y}^o, m) \propto \prod_{j=1}^{m} f(y_j^o|\theta) = \theta^{\sum_{j=1}^{m} y_j^o} e^{-m\theta}.$$

Since θ is continuous, using kernel matching, we know then that $\pi(\theta) \sim \text{Gamma}\left(\sum_{j=1}^{m} y_j^o + 1, m\right)$. ■

- **Mixture Prior**: $\pi(\boldsymbol{\theta}) = \sum_{i=1}^{k} q_i \pi_i(\boldsymbol{\theta})$, where $\sum_{i=1}^{k} q_i = 1$, and $q_i \geq 0$.
  - Restricting the prior to a parametric family limits how accurately prior uncertainty can be expressed (for instance, the Normal family is conjugate, but is symmetric and unimodal, which may not accurately reflect our data).
  - Forms a mixture posterior $\pi(\boldsymbol{\theta}|\underline{Y}) \propto \sum_{i=1}^{k} Q_i \pi_i(\boldsymbol{\theta}|\underline{Y})$, where Q_i are more weights.

* $Q_i \neq q_i$ necessarily.
* The posterior is a mixture of the same family of distributions as the prior, so it is conjugate.

- **Improper Prior**: A prior distribution that is non-negative, but does not have a finite integral over the parameter's support.
  - Any prior from a common family of distributions is proper.
  - Any proper prior distribution leads to a proper posterior distribution.
  - It is only okay to use an improper prior if the resulting posterior is proper.
    Example: Considering the following model,

$$f(Y_i|\theta_i) \overset{\perp}{\sim} \mathcal{N}(\theta_i, 1), \text{ and } f(\theta_i|\mu,\sigma^2) \overset{\text{iid}}{\sim} N(\mu,\sigma^2), \text{ and } \pi(\mu,\sigma^2) \propto \frac{1}{\sigma^2}.$$

    Is $\pi(\mu,\sigma^2)$ proper? Is $\pi(\mu,\sigma^2)$ proper using Jeffrey's prior?

$$\begin{aligned}
f(Y_i,\theta_i|\mu,\sigma^2) =& f(Y_i|\theta_i,\mu,\sigma^2)f(\theta_i|\mu,\sigma^2) = N(Y_i|\theta_i,1)N(\theta_i|\mu,\sigma^2);\\
f(Y_i|\mu,\sigma^2) =& \int_{\theta_i} N(Y_i|\theta_i,1)N(\theta_i|\mu,\sigma^2)d\theta_i \sim \mathcal{N}(Y_i|\mu,\sigma^2+1);\\
\int_{\sigma^2}\int_{\mu}\pi(\mu,\sigma^2|\underline{Y}) &\propto \int_{\sigma^2}\int_{\mu}\frac{1}{\sigma^2}\prod_{i=1}^{n}N(Y_i|\mu,\sigma^2+1)d\mu d\sigma^2\\
&\propto \int_0^\infty\int_\mu \frac{1}{\sigma^2}(\sigma^2+1)^{-n/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\sum_{i=1}^{n}(Y_i-\mu)^2\right\}d\mu d\sigma^2\\
&= \int_0^\infty\int_\mu \frac{1}{\sigma^2}(\sigma^2+1)^{-n/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\mu-\bar{Y})^2\right]\right\}d\mu d\sigma^2\\
&\propto \int_0^\infty \frac{1}{\sigma^2}(\sigma^2+1)^{-n/2+1/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}d\sigma^2\\
&> \int_0^1 \frac{1}{\sigma^2}(\sigma^2+1)^{-n/2+1/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}d\sigma^2;
\end{aligned}$$

    First, $\int_0^1 \frac{1}{\sigma^2}d\sigma^2 = \infty$, so we have an improper prior when $\sigma^2 \in (0,1)$. In this range, $\exp\left\{-\frac{\sum_{i=1}^n(Y_i-\bar{Y})^2}{2(\sigma^2+1)}\right\} > \exp\left\{-\frac{\sum_{i=1}^n(Y_i-\bar{Y})^2}{2}\right\}$, and $(\sigma^2+1)^{-n/2+1/2} \geq 2^{-n/2+1/2}$. Since we have bounded the posterior below by a positive function of $\underline{Y}$, the posterior is improper.
    Jeffrey's prior leads to a constant, so $\pi(\mu,\sigma^2) \propto 1$. Therefore,

$$\int_{\sigma^2}\int_{\mu}\pi(\mu,\sigma^2|\underline{Y}) \propto \int_0^\infty (\sigma^2+1)^{-n/2+1/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}d\sigma^2,$$

    following the same steps as above. Define $\tau^2 := \sigma^2+1$.

$$\begin{aligned}
&\int_0^\infty (\sigma^2+1)^{-n/2+1/2}\exp\left\{-\frac{1}{2(\sigma^2+1)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}d\sigma^2\\
&\quad= \int_{-1}^\infty (\tau^2)^{-n/2+1/2}\exp\left\{-\frac{1}{2(\tau^2)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\}d\tau^2. \ \blacksquare
\end{aligned}$$

  - An improper prior leads to an improper marginal of $\underline{Y}$.
  - If $L(\theta;\underline{Y}) \geq L_0(\underline{Y})$ for some $L_0(\underline{Y}) > 0$ for all $\boldsymbol{\theta}$, then any improper prior yields an improper posterior.
    Example: Suppose $f(Y|\theta) = p\phi(y) + (1-p)\phi(y-\theta)$ for some $p \in (0,1)$.

$$p\phi(y) + (1-p)\phi(y-\theta) \geq p\phi(y) > 0. \ \blacksquare$$

- In the absence of prior information, selecting the prior might be a nuisance to be avoided.
- **Objective Priors**, or **Non-Informative Priors**: Priors that are systemically and objectively formulated.
  - Often provide posterior estimates that are very similar to MLEs.

- **Jeffreys' Prior**: $\pi(\theta) \propto \sqrt{I(\theta)}$.
  - In the multivariate case, $\pi(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{\theta})|}$.
  - An objective prior that is invariant to reparameterization.
  - $I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2} f(\underline{Y}|\theta)\right)$.
    - In the multivariate case, $I(\boldsymbol{\theta})_{i,j} = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\underline{Y}|\boldsymbol{\theta})\right)$.
  - Is often an improper prior.
  - Jeffreys' prior for location-scale families: Suppose $f(x|\mu,\tau) = \tau \cdot g(\tau(x-\mu))$, where $\mu \in \mathbb{R}$, and $\tau > 0$.
    - $\pi(\mu) \propto 1$ when τ is known.
    - $\pi(\tau) \propto \frac{1}{\tau}$ when μ is known.
    - $\pi(\mu,\tau) \propto 1$ when both are unknown.
- **Bernardo's Reference Prior**: $K_n(\pi) = E[K(\pi|\boldsymbol{x}_n)]$.
  - Maximizes the expected difference between the prior and posterior with KL divergence.
  - Is a computationally challenging optimization problem.
  - Ensures that the prior does not overwhelm the data.
  - Distinguishes between parameters of interest and nuisance parameters.
- **Maximum Entropy Prior**: A prior that maximizes $\mathcal{E}(\pi|\pi_0) = \mathbb{E}_\pi\left[\log\left(\frac{\pi_0(\theta)}{\pi(\theta)}\right)\right] = \int \log\left(\frac{\pi_0(\theta)}{\pi(\theta)}\right)\pi(\theta)d\theta$.
  - Maximizes the negative KL-divergence between π and π_0.
- **Empirical Bayes**: Uses the data to select the priors.
  - Inspects the data to select values of nuisance parameters in the prior, and then performs a Bayesian analysis as though the nuisance parameters were known all along.
  - Uses the data twice, which ignores uncertainty about the nuisance parameter.
    - Empirical Bayes analysis on $\boldsymbol{\theta}$ thus will have artificially narrower posterior distributions.
    - Still useful in higher dimensions, when uncertainty in the nuisance parameters is negligible.
- **Penalized Complexity Prior**: An uninformative prior that puts little weight on a base model with a lot of parameters.
  - Is designed to prevent overfitting caused by using a model that is too complex.
  - Uses the KL divergence between the priors for the full and base models.
  - Is not objective.
  - Provides a systemic way to set priors for high-dimensional models.

6.4 MCMC and Computational Methods

Return to Table of Contents

- Often, Bayesian methods require us to perform complex integration, and find complex posterior densities.
  - A lot of the time, analytical or closed-form expressions are extremely difficult or impossible to compute using ordinary methods.
- **Deterministic Methods**: Methods that yield the exact quantity of interest (or yield a "close enough" value every time, within some margin of error).
- **Maximum a Posteriori Estimator**, or **MAP Estimator**: $\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \log[\pi(\boldsymbol{\theta}|\underline{Y})]$.
  - Avoids integration of the posterior, as opposed to the ordinary Bayes estimator.
  - Can be obtained directly, or computationally using gradient ascent.
- **Deterministic Numerical Integration**: Finite integral approximation techniques.
  - Trapezoid rule, midpoint rule, and Simson's rule from Calc II all are examples.
  - `integrate` uses this for one dimension, and `cubature` in higher dimensions in R.

  - Accuracy is good in low dimensions, but suffers as the dimensionality increases (commonly known as the curse of dimensionality).
    * Typically, the rate of accuracy is $N^{-4/d}$, where d is the number of dimensions.
  - Greatly suffers in computing time for higher dimensions (typically > 8).

- Earlier, we discussed the Bayesian CLT, which uses differentiation instead of integration.
  - Suffers in performance for smaller sample sizes.
  - Also assumes Normality, which may not accurately reflect the posterior distribution.

- Using SLLN and CLT, if we can generate iid samples from the posterior density $\pi(\boldsymbol{\theta}|\underline{X})$, then we can approximate $\int \eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\underline{X})d\boldsymbol{\theta}$ by Monte Carlo estimate $\hat{\eta}_M$.
  - **Monte Carlo Standard Error**: $\sqrt{\frac{\hat{V}_M}{M}} = O(N^{-1/2})$, regardless of d.
  - How we choose to sample the $\boldsymbol{\theta}$ values will be discussed shortly.

- Advantages of MCMC:
  - Addresses the curse of dimensionality by evaluating the functions at randomly chosen points.
  - MCMC works for non-smooth functions.
  - Convergence is guaranteed (eventually!).

- **Rejection Sampling**: We "accept" an alternate candidate value of $\boldsymbol{\theta}$ by some criterion.
  - Suppose $K(\boldsymbol{\theta}; X) \leq M(X)g(\boldsymbol{\theta}; X)$, where $M(X) > 0$ is a constant wrt $\boldsymbol{\theta}$, $g(\boldsymbol{\theta}; X)$ is a known density function, and $K(\boldsymbol{\theta}; X)$ is the posterior kernel. Then, to simulate $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|X) = c(X)K(\boldsymbol{\theta}; X)$, it is sufficient to generate $\boldsymbol{\theta} \sim g$ and $u \sim U(0, M(X) \cdot g(\boldsymbol{\theta}; X))$ until $0 < u < K(\boldsymbol{\theta}; X)$.
  - The probability of acceptance of the rejection sampling is $\frac{1}{M}\int K(\boldsymbol{\theta}; X)d\boldsymbol{\theta}$.
    * Computational efficiency depends on the choice of the bracketing density. If $g(\boldsymbol{\theta}; X)$ is too high, then we won't often accept new samples.
    * The smaller M is, the greater the computational efficiency of the algorithm.

- **Transition Kernel Density**, or **TKD**, $\boldsymbol{T(\theta', \theta)}$: A density function such that $T(\boldsymbol{\theta}', \cdot)$ is a probability density for each $\boldsymbol{\theta}'$, and $T(\cdot, \boldsymbol{\theta})$ is a measurable function for each $\boldsymbol{\theta}$.
  - We do not have independent samples! The samples depend directly on previous samples.
  - Some people like to drop samples in increments to mitigate the correlated data, but it doesn't seem to matter much for the sampling techniques covered here.
  - MCMC samples $\boldsymbol{\theta}^{(l)} \sim T(\boldsymbol{\theta}^{(l-1)})$.
  - **Symmetric**: $T(\boldsymbol{\theta}', \boldsymbol{\theta}) = T(\boldsymbol{\theta}, \boldsymbol{\theta}')$.

- **Markov Chain**: A sequence of RVs $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots\}$ with TKD $T(\cdot, \cdot)$ such that $P(\boldsymbol{\theta}^{(l)} \in A|\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(l-1)}) = \int_A T(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{\theta})d\boldsymbol{\theta}$ for all l.
  - $(\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(l-1)}) \equiv (\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^{(l-1)})$.

- **Stationary Density**: $p(\cdot)$ is stationary for TKD $T(\cdot)$ if $p(\boldsymbol{\theta}) = \int T(\boldsymbol{\theta}', \boldsymbol{\theta})p(\boldsymbol{\theta}')d\boldsymbol{\theta}'$.
  - If $\boldsymbol{\theta}^{(l-1)} \sim p(\cdot)$, then $\boldsymbol{\theta}^{(l)} \sim p(\cdot)$.
  - $\pi(\boldsymbol{\theta}|\underline{X})$ is the stationary distribution of the resulting Markov chain, so the densities in MCMC are stationary by construction.

- **Resolvant**: $T_\epsilon(\boldsymbol{\theta}', \boldsymbol{\theta}) := (1 - \epsilon)\sum_{l=1}^{\infty} \epsilon^l T_l(\boldsymbol{\theta}', \boldsymbol{\theta})$.

- **Irreducible**: An MCMC $\{\boldsymbol{\theta}^{(l)}; l = 0, 1, \dots\}$ with TKD $T(\boldsymbol{\theta}', \boldsymbol{\theta})$ such that $T_\epsilon(\boldsymbol{\theta}', \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}', \boldsymbol{\theta}$, for some $\epsilon \in (0, 1)$.

- In MCMC, $(\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}')$ is closer to invariant $p(\cdot)$ than $(\boldsymbol{\theta}^{(l-1)}|\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}')$.

- If the TKD has $p(\cdot)$ as the invariant density and satisfies the strong drift condition, and if $|g(\boldsymbol{\theta})| \leq M$ for all $\boldsymbol{\theta}$, then posterior moments converge geometrically fast.
  - **Strong Drift Condition**: $T(\boldsymbol{\theta}', \boldsymbol{\theta}) \geq (1 - \rho)p(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}', \boldsymbol{\theta}$, and for some $\rho \in (0, 1)$.

  - We can often relax the strong drift condition.

- $T_\epsilon(\cdot,\cdot)$ shares the same invariant density $p(\cdot)$ as $T(\cdot,\cdot)$.
  - $\int p(\boldsymbol{\theta}')T_\epsilon(\boldsymbol{\theta}',\boldsymbol{\theta})d\boldsymbol{\theta}' = p(\boldsymbol{\theta})$.
- **Detailed Balance Condition**, or **DBC**: $p(\boldsymbol{\theta}')T(\boldsymbol{\theta}',\boldsymbol{\theta}) = p(\boldsymbol{\theta})T(\boldsymbol{\theta},\boldsymbol{\theta}')$.
  - If T satisfies the DBC, then so does T_ϵ.
  - The new TKD obtained by the resolvant T_ϵ enjoys the same properties as T.
  - If $\epsilon > \frac{1}{1+\rho}$, then the Markov chain with the transition kernel function $T_\epsilon(\cdot,\cdot)$ converges at a faster rate than $T(\cdot,\cdot)$.
- **CLT for Markov Chains**: If the Markov chain $\{\boldsymbol{\theta}^{(l)}; l = 0, 1, \dots\}$ is irreducible, aperiodic and reversible with invariant density $p(\cdot)$, then

$$\frac{1}{\sqrt{N}}\sum_{l=1}^{N}\left(g(\boldsymbol{\theta}^{(l)}) - E_p\left[g(\boldsymbol{\theta})\right]\right) \xrightarrow{\text{d}} N\left(0, \gamma_g^2\right),$$

  where $\gamma_g^2 = Var_p[g(\boldsymbol{\theta})] + 2\sum_{k=1}^{\infty} Cov_p\left(g(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(l)})\right)$ is finite.
  - Negative autocorrelations benefits γ_g^2.
  - To prove irreducibility and aperiodicity conditions, $\exists A$ such that $\inf_{\boldsymbol{\theta},\boldsymbol{\theta}'\in A} T(\boldsymbol{\theta}',\boldsymbol{\theta}) > 0$, and $\sum_{l=1}^{\infty}\int_A T_l(\boldsymbol{\theta},\boldsymbol{\theta}')p_0(\boldsymbol{\theta}')d\boldsymbol{\theta}' > 0\ \forall\boldsymbol{\theta}$.
- We can use lags to determine rate of convergence.
  - Some autocorrelation is expected (recall the dependence), but too much could mean slow convergence.
  - Usually, inspecting the first lag is sufficient.
  - Typically, overly-complex models, or models with poor starting values, converge much more slowly.
- **Metropolis-Hastings**: For $l \in \{1, 2, \dots\}$, and given $\boldsymbol{\theta}^{(l-1)}$ and TKD $T_0(\boldsymbol{\theta}',\boldsymbol{\theta})$:
  1. Draw $\boldsymbol{\theta}^* \sim T_0(\boldsymbol{\theta}^{(l-1)}, \cdot)$.
  2. Draw $u \in U(0,1)$, and calculate the acceptance probability, $\rho(\boldsymbol{\theta}',\boldsymbol{\theta}) = \min\left\{\frac{K(\boldsymbol{\theta})T_0(\boldsymbol{\theta},\boldsymbol{\theta}')}{K(\boldsymbol{\theta}')T_0(\boldsymbol{\theta}',\boldsymbol{\theta})}, 1\right\}$.
  3. Set $\boldsymbol{\theta}^{(l)} = \begin{cases}\boldsymbol{\theta}^*, & u \le \rho(\boldsymbol{\theta}^{(l-1)},\boldsymbol{\theta}^*)\\ \boldsymbol{\theta}^{(l-1)}, & \text{otherwise}\end{cases}$.
     - If the TKD is symmetric, then $\rho(\boldsymbol{\theta}',\boldsymbol{\theta}) = \min\left\{\frac{K(\boldsymbol{\theta})}{K(\boldsymbol{\theta}')}, 1\right\}$.
     - We often work with log densities and $\log u$, in order to mitigate rounding errors.
     - A good acceptance rate for Metropolis-Hastings is about 40%.
  4. A common choice of a candidate distribution is a random-walk Gaussian distribution, where $\theta_j^*|\theta_j^{(l-1)} \sim \mathcal{N}(\theta_j^{(l-1)}, c_j^2)$, $c_j > 0$.
     - Does not require knowledge about the form of the posterior.
     - Simplifies the acceptance ratio.
     - Will be suboptimal if the Gaussian distribution does not closely approximate the posterior.
     - c_j is chosen to tune the acceptance rate.
     - c_j may vary during the burn-in, but must be fixed for the kept iterations.
     - If we are rejecting too many candidates, decrease c_j, and vice versa.
- **Gibbs Sampler**: For $l = 1, 2, \dots$, given $\boldsymbol{\theta}^{(l-1)} = (\theta_1^{(l-1)}, \dots, \theta_d^{(l-1)})$:
  - **Systematic Gibbs**: Draw $\theta_j^{(l)} \sim p(\theta_1^{(l)}, \dots, \theta_{j-1}^{(l)}, \theta_{j+1}^{(l-1)}, \dots, \theta_d^{(l-1)})$.
  - **Random Scan Gibbs**: Randomly shuffle $\boldsymbol{\theta}$ at each iteration, and then apply systematic Gibbs to the shuffled $\boldsymbol{\theta}$.
  - Random scan satisfies DBC, systematic does not.
  - TKDs for each marginal satisfy DBC.

- Under regularity conditions, the Gibbs sampler converges geometrically, and the rate of convergence is related to correlation between variables.
- Can be viewed as a special case of Metropolis-Hastings, where ρ is always unity.
- Reduces the problem of sampling from a complicated, multivariate distribution to sampling from several simpler, univariate distributions.
- Requires full conditional posteriors for each parameter, which might not be feasible.
- Can perform poorly for parameters with a strong posterior dependence (there is strong correlation between parameters).
  - Can update dependent parameters in chunks, or blocks.
- **Metropolis-Within-Gibbs**: For cases with multiple parameters, we can use Gibbs sampling for parameters which have nice conditional posteriors, and use Metropolis-Hastings for the rest.

- **Slice Sampler**: Follows Gibbs sampling logic. For $l = 1, 2, \ldots,$
  1. Draw $u^{(l)} \sim U(0, p(\boldsymbol{\theta}^{(l-1)})$.
  2. Draw $\boldsymbol{\theta}^{(l)} \sim U(S^{(l)})$, where $S^{(l)} = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}) \geq u^{(l)}\}$.
     - This step is often challenging. If $K(\boldsymbol{\theta}; X) = \prod_{m-1}^{M} h_m(\boldsymbol{\theta})$, we could replace this step by introducing M auxiliary variables $u_1, \ldots, u_M$.
       (a) Draw $u_m^{(l)} \sim U(0, h_m(\boldsymbol{\theta}^{(l-1)}))$ for $m \in \{1, \ldots, M\}$.
       (b) Draw $\boldsymbol{\theta}^{(l)} \sim U(S^{(l)})$, where $S^{(l)} = \bigcap_{m=1}^{M} \{\boldsymbol{\theta} : h_m(\boldsymbol{\theta}) \geq u_m^{(l)}\}$.
     - Convergence is slower if we use the auxiliary variables.
  - $\boldsymbol{\theta} \sim \pi(\cdot)$ is equivalent to generating $(\boldsymbol{\theta}, u) \sim U(S)$, where $S = \{(\boldsymbol{\theta}, u) : p(\boldsymbol{\theta}) \geq u\}$.
  - Dr. Ghosh seems to like this one.

- **Independence Sampler**: The proposal density $T_0(\boldsymbol{\theta}', \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$.
  1. Draw $\boldsymbol{\theta}^* \sim p^*(\cdot)$, where $p^*(\cdot)$ is a density with a heavier tail than $p(\cdot)$.
  2. Draw $u = U(0, 1)$.
  3. Calculate $w(\boldsymbol{\theta}^*) = \frac{K(\boldsymbol{\theta}^*)}{p^*(\boldsymbol{\theta}^*)}$, and $w(\boldsymbol{\theta}^{(l-1)}) = \frac{K(\boldsymbol{\theta}^{(l-1)})}{p^*(\boldsymbol{\theta}^{(l-1)})}$.
  4. Set $\boldsymbol{\theta}^{(l)} = \begin{cases} \boldsymbol{\theta}^*, \, u \leq \min\left\{\frac{w(\boldsymbol{\theta}^*)}{w(\boldsymbol{\theta}^{(l-1)})}, 1\right\} \\ \boldsymbol{\theta}^{(l-1)}, \text{ otherwise} \end{cases}$.
  - $w(\cdot)$ is the importance ratio.

6.5 Bayesian Linear Models

Return to Table of Contents

- Recall that a linear model is of the form $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}_i$ is a vector of covariates (can be fixed or random), $\boldsymbol{\epsilon} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $\underline{Y} \in \mathbb{R}^n$, and $\boldsymbol{\beta} \in \mathbb{R}^p$.
  - σ^2 may be known or unknown.
  - Bayesian linear models often place priors on $\boldsymbol{\beta}$ and σ^2.
  - Bayesian linear regression performs similarly to frequentist linear regression when $n >> p$.
- The Bayesian linear regression model assumes that the mean is a linear combination of the covariates, and that the observations are independent Normal RVs.
  - Later sections will deal with some models that don't immediately uphold these assumptions.
- A lot of the distributional results follow from elementary linear model theory.
- If σ^2 is unknown, then a common choice of prior is either an improper prior or an Inverse Gamma prior.
  - The `dnorm` command in JAGS uses $\tau^2 := \frac{1}{\sigma^2}$ in lieu of the variance parameterization.
- A typical choice of prior for β_i is a univariate Gaussian distribution centered at zero.

- Assuming $\underline{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$ and $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Omega})$, the resulting posterior is

$$\boldsymbol{\beta}|\underline{Y}, \sigma^2 \sim \mathcal{N}\left(\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right).$$

  - If σ^2 is constant, then a flat prior for β_j results in a proper posterior under the least squares conditions, and

$$\boldsymbol{\beta}|\underline{Y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_{OLS}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right).$$

    - If σ^2 is unknown, then $\boldsymbol{\beta}|\underline{Y} \sim t_n$.
- **Bayesian Lasso**: $\beta_i|\sigma^2 \sim$LaPlace(μ, σ).
  - Favors β values closer to (and including) zero, just like ordinary LASSO regression.
  - In JAGS, `ddexp` gets us the density from the LaPlace distribution.
  - Is useful when p is large, but most of the covariates are noise.
  - Can implement in JAGS using Gibbs or Metropolis sampling.
- Choosing a univariate Gaussian prior can counteract collinearity.
- If the variance term of the Gaussian prior is too small, then the resulting β_j estimates will end up biased, even with MCMC.

- When $p > n$, proper priors are required.
  - We cannot use improper priors (such as Jeffrey's priors).
- While we can use JAGS for prediction, it often impacts runtime.
  - It is instead recommended you do predictions/construct PPDs after running the JAGS model.
- We can use ordinary GLM theory to fit exponential family models under the Bayesian framework.
  - Define $\eta_i := \sum_{j=1}^p X_{ij}\beta_j$. $g(\theta_i) = \eta_i$ is the link function, where θ_i is the parameter in the likelihood for the response (such as $\mathbb{E}(Y_i) = \theta_i$, or $Var(Y_i) = \theta_i$).
- The standard linear regression model assumes the same regression model applies to all observations.
  - This assumption does not apply for random effects. Using a random effects model alleviates this assumption, and we can fit this model under the Bayesian framework.
    - Since, in random effects models, we assume $A_i \overset{\text{iid}}{\sim} (\theta, \tau^2)$, we can have MCMC sample from this distribution, while placing priors on both θ and τ^2.
    - In random effects models, we are often interested in testing whether or not $\tau^2 = 0$. The Inverse Gamma prior assigns zero probability to this outcome, so we often use a Half-Cauchy prior.
  - The random slopes model is another case, where $Y_{ij}|\boldsymbol{A}_j \overset{\perp}{\sim} \mathcal{N}(A_{i1} + A_{i2}X_j, \sigma^2)$, and $\boldsymbol{A}_i \overset{\text{iid}}{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Omega})$.
    - This model assumes random intercepts and random slopes.
- The MLR model assumes $Y_i|\boldsymbol{X}_i$ is linear in $\boldsymbol{X}_i$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
  - Minor model violations are not an issue, but multiple or major model violations can be a problem with respect to model misspecification.
  - The linearity assumption can be relaxed with nonparametric regression. That is, $Y_i = g(\boldsymbol{X}_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
    - Fully nonparametric models have an infinite number of parameters (essentially, g is some arbitrary smooth function), whereas semi-parametric models specify g with a finite number of parameters (ex. g is some polynomial function of some specified order, or there are B number of splines).
  - Heteroskedastic models relaxes the constant variance assumption, where $Var(\epsilon_i) = g(\boldsymbol{X}_i)$.
    - Since σ_i^2 should be positive, $g : \mathbb{R}^p \to [0, \infty)$.
  - Non-Gaussian error models relax the Gaussian error assumption, where $Y_i|g_i \sim \mathcal{N}\left(\sum_{j=1}^p X_{ij}\beta_j + \theta_{g_i}\right)$, where $g_i \in \{1, \ldots, K\}$ is the cluster label for observation i with probability $P(g_i = k) = \pi_k$.
    - Useful for models with heavy tails on the errors.
  - For correlated data, $\underline{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ must be specified correctly.
    - We usually use Metropolis-Hastings to sample with the correlation parameters.

    - We can account for uncertainty in the correlation parameters in prediction/inference under the Bayesian framework.

- Similar to frequentist analysis, we should verify model assumptions with QQ-plots, variable plots, etc.
- We could use cross-validation to compare out-of-sample model performance.
  - Bayesian prediction is based on the PPD of the out-of-sample data.
- **Bayes Factor**, or **BF**: $BF = \frac{P(\mathcal{M}_2|\underline{Y})/P(\mathcal{M}_1|\underline{Y})}{P(\mathcal{M}_2)/P(\mathcal{M}_1)} \overset{\text{Equal Priors}}{=} \frac{\int f(\underline{Y}|\boldsymbol{\theta};\mathcal{M}_2)\pi(\boldsymbol{\theta}|\mathcal{M}_2)d\boldsymbol{\theta}}{\int f(\underline{Y}|\boldsymbol{\theta};\mathcal{M}_1)\pi(\boldsymbol{\theta}|\mathcal{M}_1)d\boldsymbol{\theta}}$.
  - Doesn't require models to be nested, but it makes calculations easier.
  - Quantifies the data's support of the models.
  - If $BF > 10$, then there is strong evidence in favor of $\mathcal{M}_2$ over $\mathcal{M}_1$ If $BF > 100$, then the evidence is decisively in favor of $\mathcal{M}_2$.
  - $P(\mathcal{M} = \mathcal{M}_j|\underline{Y}) = \int p(\boldsymbol{\theta}, \mathcal{M} = \mathcal{M}_j|\underline{Y})d\boldsymbol{\theta}$.
  - BF cannot be used with improper priors, since the marginal distribution of $\mathcal{M}$ is not defined.
  - BF is very sensitive to the choice of hyperparameters.
  - Can behave erratically with uninformative priors.
  - The model with the lowest BF is closest to the true model.
- **Posterior Predictive Checks**: Compute some relevant set of test statistics (i.e. sample mean, skewness, accuracy) for each iteration in MCMC, and compare to the actual statistic in the data.
  - To compare the MCMC distribution of the test statistics, we compute a Bayesian p-value, where values near 0 or 1 indicate inadequacy of the model.
  - There is no generic choice of the test statistic(s), and the model may perform better or worse on some test statistics.
- **Model Averaging**: Let $\pi_j = P(\mathcal{M}_j)$. Then, $\pi(M_j|x) \propto m(x|\mathcal{M}_j)\pi_j = \pi_j \int f(x|\theta, \mathcal{M}_j)\pi(\theta|\mathcal{M}_j)d\theta$.
  - Converts prior model probabilities into posterior model probabilities.
  - $m(x|\mathcal{M}_j)$ is the prior predictive distribution under $\mathcal{M}_j$.
  - Could either use MAP ($j_0 = \arg\max_j \pi(\mathcal{M}_j|x)$ or the median of j to select the best model.
- **Stochastic Search Variable Selection**: Suppose $Y_i|\boldsymbol{\beta}, \sigma^2 \overset{\perp}{\sim} \mathcal{N}\left(\sum_{j=1}^p X_{ij}\beta_j, \sigma^2\right)$, where $\beta_j = \gamma_j\delta_j$, $\gamma_j \sim$Ber(q), and $\delta_j \sim \mathcal{N}(0, \tau^2\sigma^2)$.
  - β_j in this example follows the spike-and-slab prior.
  - Stochastically obtains the model that explains the data the best.
  - We can now approximate inclusion probabilities with MCMC.
  - If the number of potential models is large, we need a lot of iterations to truly find a good subset.

7 ST 793: Advanced Statistical Inference

Instructor: Dr. Ana-Maria Staicu
Semester: Fall 2024
Main Textbook: Boos and Stefanski, *Essential Statistical Inference*

7.1 Likelihood Functions

Return to Table of Contents

- **Parametric Statistical Models**: A family of distributions specified by a finite number of parameters.
  - **Non-Parametric Statistical Models**: A family of distributions specified by an infinite number of parameters of interest.
  - **Semi-Parametric Statistical Models**: A family of distributions specified by a finite number of parameters of interest, and and infinite number of nuisance parameters.
- **Exponential Family Models**: $Y \sim EF(\theta, \phi)$ if $f(y;\theta = \exp\left\{\frac{T(y)g(\theta)-b(\theta)}{a(\phi)}\right\} h(y;\phi)$, and the support doesn't depend on θ.
  - $\mathbb{E}(T(y)) = b'(\theta)$, and $Var(T(y)) = b''(\theta)a(\phi)$.
  - $T(y)$ is a sufficient statistic when we have an EF sample.
    - Recall that if T is sufficient, then $Y|T$ doesn't depend on θ.
- **Likelihood Function**: $L(\theta|\underline{y}) = f_{joint}(\underline{y};\theta)$.
  - A function of the parameters which is equal to the PDF of $\underline{Y}$.
  - Selecting a θ that maximizes $L(\theta)$ maximizes the probability to observe the sample we did.
- **Generalized Linear Model**, or **GLM**: Assumes $(Y_i|\underline{X}_i)$, but y_i is discrete. Contains 3 components:
  1. $Y_i \overset{\perp}{\sim} EF(\theta_i, \phi)$.
  2. The **linear predictor** $\eta_i = X_i'\beta$ exists.
  3. The **link function** $g(E(Y_i)) = \eta_i$ also exists, and is known and monotone (invertible).
     - **Canonical Link Function**: When g is defined such that $\theta_i = x_i'\beta$.
       - $\theta_i = g(b'(\theta_i))$.
       - $b'(\theta_i)$ is always monotone.
- **Example** Constructing various likelihoods:
  - $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. $L(\theta) = \prod_{i=1}^n f(y_i;\theta)$.
  - $Y_1, \ldots, Y_n \overset{\perp}{\sim} f_i(y;\theta)$. $L(\theta) = \prod_{i=1}^n f_i(y_i;\theta)$.
  - $\underline{Y} \sim f_{joint}(\underline{y};\theta)$. $L(\theta) = f_{joint}(\underline{y};\theta)$.
  - Mixture model: Suppose $Y_i \overset{\text{iid}}{\sim} g(y;\theta,p) := pI(y=0) + (1-p)f(y;\theta)$. In other words, Y is a mixture of a point mass at zero, and some continuous distribution $f(y;\theta)$.

$$L(\theta,p) = \prod_{i=1}^n p^{I(Y_i=0)} \left[(1-p)f(y_i;\theta)\right]^{I(Y_i\neq 0)}.$$

  - Truncated data: Suppose $Y_i \overset{\text{iid}}{\sim} f(y;\theta)$. We observe $\underline{y}$ such that each element of $\underline{y}$ is greater than some constant L.

$$\begin{aligned} F_{obs}(y;\theta) = P(Y \leq y;\theta) =& P(Y \leq y | Y \geq L;\theta) \\ =& \frac{P(L \leq Y \leq y;\theta)}{P(Y \geq L;\theta)} = \frac{F(y;\theta) - F_Y(L;\theta)}{1 - F_Y(L;\theta)}; \end{aligned}$$

    Therefore, $f_{obs}(y;\theta) = \frac{f(y;\theta)}{1-F_Y(L;\theta)}$, and $L(\theta) = \prod_{i=1}^n f_{obs}(y;\theta)$.

– Left-Censored data: Assume IID samples. Denote T_i as the survival time for the ith subject. Therefore, $T_i \overset{\text{iid}}{\sim} f(t;\theta)$. We observe $Y_i = \max\{T_i, C_i\}$, where C_i is the time we start our study. Denote f_C as the PDF of C_i. We also observe $\delta_i = I(t_i \geq c_i)$ (we know which units survived).
If $\delta_i = 1$, then $y_i = t_i$, with probability $f_T(y_i;\theta)F_C(y_i)$. Otherwise, $y_i = c_i$, with probability $f_C(y_i)F_T(y_i;\theta)$. Therefore,

$$L(\theta) = \prod_{i=1}^{n} \left[f_T(y_i;\theta)F_C(y_i)\right]^{\delta_i} \left[f_C(y_i)F_T(y_i;\theta)\right]^{1-\delta_i} \propto \prod_{i=1}^{n} \left[f_T(y_i;\theta)\right]^{\delta_i} \left[F_T(y_i;\theta)\right]^{1-\delta_i}.$$

– Right-Censored data: Assume IID samples. Denote T_i as the survival time for the ith subject. Therefore, $T_i \overset{\text{iid}}{\sim} f(t;\theta)$. We observe $Y_i = \min\{T_i, C_i\}$, where C_i is the time we start our study. Denote f_C as the PDF of C_i. We also observe $\delta_i = \mathbb{I}(t_i \leq c_i)$ (we know which units survived).
If $\delta_i = 1$, then $y_i = t_i$, with probability $f_T(y_i;\theta)[1 - F_C(y_i)]$. Otherwise, $y_i = c_i$, with probability $f_C(y_i)[1 - F_T(y_i;\theta)]$. Therefore,

$$L(\theta) = \prod_{i=1}^{n} \left[f_T(y_i;\theta)(1 - F_C(y_i))\right]^{\delta_i} \left[f_C(y_i)(1 - F_T(y_i;\theta))\right]^{1-\delta_i} \propto \prod_{i=1}^{n} \left[f_T(y_i;\theta)\right]^{\delta_i} \left[1 - F_T(y_i;\theta)\right]^{1-\delta_i}.$$

– Regression model: Suppose we observe $(y_i, \underline{x}_i)$, where $Y_i := X_i^T\beta + \epsilon$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, where σ^2 and X_i are known.

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} f(y_i;\beta,\sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right\} \\ &\propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right\}. \end{aligned}$$

– Regression model: Suppose we observe $(y_i, \underline{x}_i)$, where $Y_i := X_i'\beta + \epsilon$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, where σ^2 is known, but X_i is unknown. Suppose $X_i \sim f(\cdot;\tau^2)$.

$$\begin{aligned} L(\beta, \sigma^2, \tau^2) &= \prod_{i=1}^{n} f(y_i|x_i;\beta,\sigma^2) f(x_i;\tau^2); \\ L(\beta, \sigma^2) &\propto \prod_{i=1}^{n} f(y_i|x_i;\beta,\sigma^2). \end{aligned}$$

– Logistic regression: $Y_i \overset{\perp}{\sim} Ber(p_i)$ with logit link. This is a GLM, so $\log\left(\frac{p_i}{1-p_i}\right) = X_i^T\beta \implies p_i = \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}} =: g(x_i;\beta)$.

$$L(\beta) = \prod_{i=1}^{n} f(Y_i|x_i;\beta) = \prod_{i=1}^{n} \left(\frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}}\right)^{Y_i} \left(1 - \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}}\right)^{1-Y_i}.$$

– Poisson regression: $Y_i \overset{\perp}{\sim} Pois(\lambda_i)$, with log link. $\lambda_i = \exp\{x_i^T\beta\}$.

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left\{y_i x_i^T\beta - e^{x_i^T\beta}\right\}}{y_i!} \propto \exp\left\{\beta^T \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} e^{x_i^T\beta}\right\}.$$

– Accelerated failure time model: Denote T_i as the time to an event, and X_i is a set of covariates. Uses a log link, so

$$\log(T_i) = x_i^T\beta + \sigma\epsilon_i,\ \epsilon_i \overset{\perp}{\sim} f(\cdot), \text{ with } \mathbb{E}(\epsilon_i) = 0$$

Observe $Y_i = \min\{\log(T_i), \log(C_i)\}$, where C_i is censored time, and $\delta_i = \mathbb{I}(\log(T_i) \leq \log(C_i))$, and $\underline{X}_i$.

$$L(\boldsymbol{\beta}, \sigma | \{Y_i, \delta_i, \boldsymbol{x}_i\}_{i=1}^n) = \prod_{i=1}^{n} \left[\frac{1}{\sigma} f_\epsilon(r_i)\right]^{\delta_i} \left[1 - F_\epsilon(r_i)\right]^{1-\delta_i}, \text{ where } r_i = \frac{Y_i - \boldsymbol{x}_i^T\beta}{\sigma}. \blacksquare$$

- **Strong Likelihood Principle**: We only need the likelihood function.

- We don't care about the data generating function at all.
- **Weak Likelihood Principle**: Suppose we have the data generating model $f(y;\theta)$, and $L(\theta|y_1) \equiv L(\theta|y_2)$ for all $\theta \in \Theta$. Then, any conclusions about θ from y_1 are the same as from y_2.

- For this section, assume $\underline{\theta} = (\theta_1', \theta_2')'$, where θ_1 are the parameters of interest.
- **Pseudo Likelihood**: Maximizes a function that is similar to the log-likelihood.
  - Produces estimates that could behave erratically wrt efficiency and unbiasedness.
  - Pseudo likelihoods are not proper likelihoods, which do have nice properties.
- **Profile Likelihood**: Maximize elements of θ_1 sequentially.
  - **Example**: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown. We first maximize μ to get $\hat{\mu} = \bar{Y}$, then maximize for σ^2 using $\hat{\mu}$.
  - Is a pseudo likelihood.
- **Integrated Pseudo-Likelihood**: $L_\pi(\theta_1) = \int L(\theta_1, \theta_2)\pi(\theta_2)d\theta_2$, where $\pi(\theta_2)$ is a weight function.
  - Replaces maximization with integration.
  - Is a pseudo-likelihood function.
- Suppose there exists a one-to-one transformation from Y to statistics (V, W).

$$f_Y(y;\theta_1,\theta_2) = f_{W,V}(w,v;\theta_1,\theta_2) = f_{W|V}(w|v;\theta_1,\theta_2)f_V(v;\theta_1) = f_{W|V}(w|v;\theta_1)f_V(v;\theta_1,\theta_2).$$

- **Marginal Likelihood**: The likelihood derived from $f_V(V;\theta_1)$ above, where V is ancillary for θ_2.
  - Is a proper likelihood function.
  - Is useful for location-scale families.
  - This likelihood is not unique wrt V.
  - We don't necessarily need to find a W. This would be only if we wanted to gain insight about the loss of information for only using V.
  - There isn't a general approach for finding marginal likelihoods.
- **Conditional Likelihood**: The likelihood derived from $f_{W|V}(w|v;\theta_1)$ above, where $W|V$ is ancillary.
  - Could be more useful for EF.
  - V might not exist for a given distribution.
  - This is also not unique.
  - V is sufficient for θ_2.
  - We could still lose information.
  - Is a pseudo likelihood.
  - **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown. Calculate the conditional likelihood of σ^2.

$$\begin{aligned}
f_{joint}(\underline{y};\mu,\sigma^2) =& \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\} \\
=&(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{Y}-\mu)^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]\right\} \\
=&(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{Y}-\mu)^2+\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]\right\};
\end{aligned}$$

By the Factorization theorem, $(V, W) = \left(\bar{Y}, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)$ is sufficient for (μ, σ). The transformation $\underline{Y} \to (V, W)$ is one-to-one. Define $g_{(V,W)}\left((\cdot,\cdot);\mu,\sigma^2\right) := f_{joint}(\cdot)$. The previous claim leads to $g_{(V,W)}\left((\cdot,\cdot);\mu,\sigma^2\right) = f_{(V,W)}\left((\cdot,\cdot);\mu,\sigma^2\right)$. In addition, since $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma^2)$, $V = \bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. So,

$$\begin{aligned}
f_{\underline{Y}|V}(\underline{y}|v;\sigma^2) =& \frac{f_{V,W}((v,w);\mu,\sigma^2)}{f_V(v;\mu,\sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{n}{2\sigma^2}\left[w+(v-\mu)^2\right]\right\}}{(2\pi\sigma^2/n)^{-1/2}\exp\left\{-\frac{n}{2\sigma^2}(v-\mu)^2\right\}} \\
=& \frac{(2\pi\sigma^2)^{-(n-1)}}{\sqrt{n}}\exp\left\{-\frac{n}{2\sigma^2}w\right\},
\end{aligned}$$

This is a conditional likelihood that does not depend on μ. Maximizing this likelihood returns an unbiased estimator for σ^2. ■

- **Maximum Likelihood Estimator**: Any $\hat{\boldsymbol{\theta}}_n$ that maximizes $L(\boldsymbol{\theta}; \underline{Y})$.
  - For any function $g(\boldsymbol{\theta})$, the corresponding MLE is $g(\hat{\boldsymbol{\theta}}_n)$.
- **Score Vector**: $S(\underline{Y}; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} & \dots & \frac{\partial \ell}{\partial \theta_b} \end{pmatrix}^T$.
  - Has the same dimension as $\hat{\theta}$.
  - **Likelihood Equations**: $S(\boldsymbol{\theta}) = \mathbf{0}$.
  - If ℓ is continuously differentiable, then $\hat{\boldsymbol{\theta}}_n$ satisfies the likelihood equations.
  - $\mathbb{E}[S(\underline{Y}; \boldsymbol{\theta})] = 0$.
  - $\frac{1}{\sqrt{n}} S(\underline{Y}; \boldsymbol{\theta}) \xrightarrow{\text{d}} \mathcal{N}_b(\mathbf{0}, I(\boldsymbol{\theta}))$.
- **Fisher Information Matrix**: $I(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]$, and $I_n(\boldsymbol{\theta}) = Var[S(\underline{Y}; \theta)] = \mathbb{E}\left[\frac{\partial \ell}{\partial \boldsymbol{\theta}} \frac{\partial \ell}{\partial \boldsymbol{\theta}'}\right]$.
  - For EF model, $I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{\theta}} S_{\boldsymbol{\theta}}(Y)\right]$.
  - Is the information that an RV contains about $\boldsymbol{\theta}$.
  - $I_n(\boldsymbol{\theta}) = n \cdot I(\boldsymbol{\theta})$.
  - $I_{i,j}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial f(Y;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial f(Y;\boldsymbol{\theta})}{\partial \theta_j'}\right]$.
  - $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\text{d}} \mathcal{N}\left(\mathbf{0}, I(\boldsymbol{\theta})^{-1}\right)$, and $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}$.
- **Average Fisher Information**: $\bar{I}(\underline{Y}, \boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^n \left[-\frac{\partial^2 \log f_i(Y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$.
  - Is the average expected information in a sample of independent data points.
  - $I(\boldsymbol{\theta}) = \mathbb{E}[\bar{I}(\underline{Y}, \boldsymbol{\theta})]$.
- **Total Fisher Information**: $I_T(\underline{Y}, \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ell(\boldsymbol{\theta}|\underline{Y})$, and $I_T(\boldsymbol{\theta}) = -E\left[I_T(\underline{Y}, \boldsymbol{\theta})\right]$.
  - For n iid data points, $I_T(\boldsymbol{\theta}) = n\bar{I}(\boldsymbol{\theta}) = nI(\boldsymbol{\theta})$.
  - **Example**: Consider a GLM for our observed data $(\boldsymbol{X}_i, Y_i)$, where $Y_i \sim EF(\theta_i, \phi)$ with PDF $f(y; \theta_i, \phi) = h(y_i; \phi) \exp\left\{\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)}\right\}$, and canonical link function $g(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta}$. Construct $L(\boldsymbol{\beta}, \phi)$, and find $\bar{I}(\boldsymbol{\beta})$. Since $\theta_i = \boldsymbol{X}_i'\boldsymbol{\beta}$, the likelihood is

$$L\left(\boldsymbol{\beta}, \phi | (Y_i, X_i)_{i=1}^n\right) \overset{\perp}{=} \exp\left\{\sum_{i=1}^n \left[\frac{Y_i \boldsymbol{X}_i'\boldsymbol{\beta} - b(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi)}\right]\right\} \prod_{i=1}^n h(Y_i; \phi).$$

The score function is

$$\begin{aligned} S(\underline{Y}; \boldsymbol{\beta}, \phi) =& \frac{\partial}{\partial \boldsymbol{\beta}'} \ell\left(\boldsymbol{\beta}, \phi | (Y_i, X_i)_{i=1}^n\right) = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}'} \left\{\frac{Y_i \boldsymbol{X}_i'\boldsymbol{\beta} - b(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi)} + \log h(Y_i; \phi)\right\} \\ =& \sum_{i=1}^n \left[\frac{Y_i - b'(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi)}\right] \boldsymbol{X}_i = \sum_{i=1}^n \left[\frac{Y_i - b'(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi) b''(\boldsymbol{X}_i'\boldsymbol{\beta})}\right] b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i; \end{aligned}$$

Define $\boldsymbol{D}_i := \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i$. We also know that $\mu_i = b'(\boldsymbol{X}_i'\boldsymbol{\beta})$. Therefore,

$$S(\underline{Y}; \boldsymbol{\beta}, \phi) = \sum_{i=1}^n \left[\frac{Y_i - b'(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi) b''(\boldsymbol{X}_i'\boldsymbol{\beta})}\right] b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i = \sum_{i=1}^n \boldsymbol{D}_i \left[\frac{Y_i - \mu_i}{Var(Y_i)}\right].$$

$$\begin{aligned} I_T(\boldsymbol{\beta}) =& -\mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}} S(\underline{Y}; \boldsymbol{\beta}, \phi)\right] = -\mathbb{E}\left[\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}} \left[\frac{Y_i - b'(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi) b''(\boldsymbol{X}_i'\boldsymbol{\beta})}\right] b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i\right] \\ =& -\mathbb{E}\left[-\sum_{i=1}^n \frac{b''(\boldsymbol{X}_i'\boldsymbol{\beta})}{a_i(\phi)} \boldsymbol{X}_i \boldsymbol{X}_i'\right] = \mathbb{E}\left[\sum_{i=1}^n \frac{b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i \left(b''(\boldsymbol{X}_i'\boldsymbol{\beta}) \boldsymbol{X}_i\right)'}{a_i(\phi) b''(\boldsymbol{X}_i'\boldsymbol{\beta})}\right] = \mathbb{E}\left[\sum_{i=1}^n \frac{\boldsymbol{D}_i \boldsymbol{D}_i'}{Var(Y_i)}\right]; \\ \bar{I}(\boldsymbol{\beta}) =& \frac{1}{n} I_T(\boldsymbol{\beta}). \blacksquare \end{aligned}$$

	Data Type		
	iid	inid	General
$L(\boldsymbol{\theta}\|\boldsymbol{Y})$	$\prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta})$	$\prod_{i=1}^{n} f_i(Y_i;\boldsymbol{\theta})$	$f(\boldsymbol{Y};\boldsymbol{\theta})$
$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}\|\boldsymbol{Y})$	$\sum_{i=1}^{n} \log f(Y_i;\boldsymbol{\theta})$	$\sum_{i=1}^{n} \log f_i(Y_i;\boldsymbol{\theta})$	$\log f(\boldsymbol{Y};\boldsymbol{\theta})$
$\boldsymbol{S}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}^T}\ell(\boldsymbol{\theta})$	$\sum_{i=1}^{n} \boldsymbol{s}(Y_i,\boldsymbol{\theta})$	$\sum_{i=1}^{n} \boldsymbol{s}_i(Y_i,\boldsymbol{\theta})$	$\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(\boldsymbol{Y};\boldsymbol{\theta})$
$\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{Y},\boldsymbol{\theta}) = -\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{S}(\boldsymbol{\theta})$	$-\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\theta}} \boldsymbol{s}(Y_i,\boldsymbol{\theta})$	$-\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\theta}} \boldsymbol{s}_i(Y_i,\boldsymbol{\theta})$	$-\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{S}(\boldsymbol{\theta})$
$\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\theta}) = E\{\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{Y},\boldsymbol{\theta})\}$	$n\boldsymbol{I}(\boldsymbol{\theta})$	$n\overline{\boldsymbol{I}}(\boldsymbol{\theta})$	$\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\theta})$
$\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{Y},\boldsymbol{\theta})$	$\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta})$	$\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta})$	–
$\overline{\boldsymbol{I}}(\boldsymbol{\theta}) = E\{\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta})\}$	$\boldsymbol{I}(\boldsymbol{\theta})$	$\overline{\boldsymbol{I}}(\boldsymbol{\theta})$	–
$\overline{\boldsymbol{I}}^*(\boldsymbol{Y},\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}(Y_i,\boldsymbol{\theta})\boldsymbol{s}(Y_i,\boldsymbol{\theta})^T$	$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i(Y_i,\boldsymbol{\theta})\boldsymbol{s}_i(Y_i,\boldsymbol{\theta})^T$	–
$\overline{\boldsymbol{I}}^*(\boldsymbol{\theta}) = E\{\overline{\boldsymbol{I}}^*(\boldsymbol{Y},\boldsymbol{\theta})\}$	$\boldsymbol{I}(\boldsymbol{\theta})$	$\overline{\boldsymbol{I}}(\boldsymbol{\theta})$	–

- Whenever parameters are added to a model, the diagonal elements of $I(\boldsymbol{\theta})^{-1}$ are always greater than or equal to the corresponding elements of the simpler model.
- **Example**: Consider a dose-response situation with k dose levels. At the ith dose d_i we observe $(Y_{ij}, n_{ij}, j = 1,\ldots,m_i)$, where n_{ij} are fixed constants. We often assume $Y_{ij} \overset{\perp}{\sim} \mathrm{Bin}\big(n_{ij}, F(\boldsymbol{x}_i^T\boldsymbol{\beta})\big)$, where F is some distribution function, $\boldsymbol{x}_i^T = (1, d_i)$ or $\boldsymbol{x}_i^T = (1, d_i, d_i^2)$.

$$\ell(\boldsymbol{\beta}) = \mathrm{c} + \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[Y_{ij}\log\{F(\boldsymbol{x}_i^T\boldsymbol{\beta})\} + (n_{ij} - Y_{ij})\log\{1 - F(\boldsymbol{x}_i^T\boldsymbol{\beta})\}\right].$$

Derive the score function, $I_T(\underline{Y},\boldsymbol{\beta})$, and $\boldsymbol{I}_T(\boldsymbol{\beta})$. Also find $I_T(\underline{Y},\boldsymbol{\beta})$ when $F(x) = (1+e^{-x})^{-1}$.
Define $p_i(\boldsymbol{\beta}) = F(\boldsymbol{x}_i^T\boldsymbol{\beta})$.

$$\begin{aligned} S(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{Y_{ij}}{p_i(\boldsymbol{\beta})}p_i'(\boldsymbol{\beta})\boldsymbol{x}_i - \frac{n_{ij}-Y_{ij}}{1-p_i(\boldsymbol{\beta})}p_i'(\boldsymbol{\beta})\boldsymbol{x}_i\right] = \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{Y_{ij}}{p_i(\boldsymbol{\beta})} - \frac{n_{ij}-Y_{ij}}{1-p_i(\boldsymbol{\beta})}\right]p_i'(\boldsymbol{\beta})\boldsymbol{x}_i \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{Y_{ij} - p_i(\boldsymbol{\beta})Y_{ij} - n_{ij}p_i(\boldsymbol{\beta}) + p_i(\boldsymbol{\beta})Y_{ij}}{p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]}\right]p_i'(\boldsymbol{\beta})\boldsymbol{x}_i = \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{Y_{ij} - n_{ij}p_i(\boldsymbol{\beta})}{p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]}\right]p_i'(\boldsymbol{\beta})\boldsymbol{x}_i. \end{aligned}$$

$$\begin{aligned} I_T(\boldsymbol{Y},\boldsymbol{\beta}) &= -\frac{\partial}{\partial\boldsymbol{\beta}}S(\boldsymbol{\beta}) = -\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{Y_{ij}p_i'(\boldsymbol{\beta})}{p_i(\boldsymbol{\beta})} - \frac{n_{ij}-Y_{ij}}{1-p_i(\boldsymbol{\beta})}p_i'(\boldsymbol{\beta})\right]\boldsymbol{x}_i \\ &= -\sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[\frac{p_i(\boldsymbol{\beta})Y_{ij}p_i''(\boldsymbol{\beta}) - Y_{ij}p_i'(\boldsymbol{\beta})p_i'(\boldsymbol{\beta})}{p_i(\boldsymbol{\beta})^2} + \frac{(n_{ij}-Y_{ij})\{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})p_i'(\boldsymbol{\beta})\}}{[1-p_i(\boldsymbol{\beta})]^2}\right]\boldsymbol{x}_i\boldsymbol{x}_i^T \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left[Y_{ij}\left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2} - \frac{p_i(\boldsymbol{\beta})p_i''(\boldsymbol{\beta}) - p_i'(\boldsymbol{\beta})^2}{p_i(\boldsymbol{\beta})^2}\right)\right. \\ &\quad \left. - n_{ij}\left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2}\right)\right]\boldsymbol{x}_i\boldsymbol{x}_i^T; \end{aligned}$$

Note that, since $Y_{ij} \sim \text{Bin}(n_{ij}, p_i(\boldsymbol{\beta}))$, $E(Y_{ij}) = n_{ij}p_i(\boldsymbol{\beta})$.

$$\begin{aligned}
I_T(\boldsymbol{\beta}) =& \mathbb{E}\left[I_T(\boldsymbol{Y}, \boldsymbol{\beta})\right] \\
=& \mathbb{E}\Bigg\{ \sum_{i=1}^{k}\sum_{j=1}^{m_i} \Bigg[Y_{ij} \left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2} - \frac{p_i(\boldsymbol{\beta})p_i''(\boldsymbol{\beta}) - p_i'(\boldsymbol{\beta})^2}{p_i(\boldsymbol{\beta})^2} \right) \\
& - n_{ij} \left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2} \right) \Bigg] \boldsymbol{x}_i \boldsymbol{x}_i^T \Bigg\} \\
=& \sum_{i=1}^{k}\sum_{j=1}^{m_i} \Bigg[n_{ij}p_i(\boldsymbol{\beta}) \left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2} - \frac{p_i(\boldsymbol{\beta})p_i''(\boldsymbol{\beta}) - p_i'(\boldsymbol{\beta})^2}{p_i(\boldsymbol{\beta})^2} \right) \\
& - n_{ij} \left(\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{[1-p_i(\boldsymbol{\beta})]^2} \right) \Bigg] \boldsymbol{x}_i \boldsymbol{x}_i^T \\
=& \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij} \left[\frac{[1-p_i(\boldsymbol{\beta})]p_i''(\boldsymbol{\beta}) + p_i'(\boldsymbol{\beta})^2}{1-p_i(\boldsymbol{\beta})} - \frac{p_i(\boldsymbol{\beta})p_i''(\boldsymbol{\beta}) - p_i'(\boldsymbol{\beta})^2}{p_i(\boldsymbol{\beta})} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T \\
=& \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij} \left[p_i''(\boldsymbol{\beta}) + \frac{p_i'(\boldsymbol{\beta})^2}{1-p_i(\boldsymbol{\beta})} - p_i''(\boldsymbol{\beta}) + \frac{p_i'(\boldsymbol{\beta})^2}{p_i(\boldsymbol{\beta})} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T \\
=& \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij}p_i'(\boldsymbol{\beta})^2 \left[\frac{1}{1-p_i(\boldsymbol{\beta})} + \frac{1}{p_i(\boldsymbol{\beta})} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_{i=1}^{k} p_i'(\boldsymbol{\beta})^2 \left[\frac{1}{p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \sum_{j=1}^{m_i} n_{ij}.
\end{aligned}$$

$$F'(x) = -\frac{1}{(1+e^{-x})^2}(-e^{-x}) = -\frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1}{(1+e^{-x})} \right) \left(1 - \frac{1}{(1+e^{-x})} \right) = F(x)[1-F(x)].$$

This means that $p_i'(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]$.

$$\begin{aligned}
\sum_{i=1}^{k} p_i'(\boldsymbol{\beta})^2 \left[\frac{1}{p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \sum_{j=1}^{m_i} n_{ij} &\to \sum_{i=1}^{k} p_i(\boldsymbol{\beta})^2 [1-p_i(\boldsymbol{\beta})]^2 \left[\frac{1}{p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})]} \right] \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \sum_{j=1}^{m_i} n_{ij} \\
&= \sum_{i=1}^{k} p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})] \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \sum_{j=1}^{m_i} n_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij}p_i(\boldsymbol{\beta})[1-p_i(\boldsymbol{\beta})] \boldsymbol{x}_i \boldsymbol{x}_i^T. \ \blacksquare
\end{aligned}$$

- **Orthogonal**: θ_1 and θ_2 are orthogonal if $I(\boldsymbol{\theta})_{1,2} = 0$.
  - **Asymptotically Independent**: If θ_1 and θ_2 are orthogonal, then $\hat{\theta}_1 \perp \hat{\theta}_2$ asymptotically.
- Methods to maximize likelihoods:
  - Directly, using the pseudo or exact likelihoods described above.
  - **Newton-Raphson**: Uses Taylor series approximations to find roots of $S(\hat{\boldsymbol{\theta}})$. In other words,

$$S(\hat{\boldsymbol{\theta}}) \approx S(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} + I^{-1}(Y; \boldsymbol{\theta}) S(\boldsymbol{\theta}).$$

    - Is iterative.
    - Measure convergence using distance metrics (ex. norms).
    - Requires smooth likelihood functions ($\ell''(\boldsymbol{\theta})$ exists), and invertible $I(\boldsymbol{\theta})$.
    - Performance depends on good starting points.
    - Approximating FI costs quadratic convergence rate.
    - **Fisher Scoring**: Replace $I(Y; \boldsymbol{\theta})$ with $I_T(\boldsymbol{\theta})$.
      - Now, $I_T(\boldsymbol{\theta})$ must be invertible near $\hat{\boldsymbol{\theta}}_n$.
  - **Expectation Maximization**, or **EM**: Suppose $Y_i \overset{\text{iid}}{\sim} f(y; \boldsymbol{\theta})$, with likelihood $L(\boldsymbol{\theta})$. We introduce latent RVs $\underline{Z}$ such that $L_C(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Z_i, Y_i; \boldsymbol{\theta})$ is iterative, but with two steps:
    1. Expectation step: $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\nu)}, Y) = \mathbb{E}_{Z|Y;\boldsymbol{\theta}^{(\nu)}}[\ell_C(\boldsymbol{\theta})]$.
    2. Maximization step: $\boldsymbol{\theta}^{(\nu+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\nu)}, y)$.

    - Considers the model as incomplete, and embeds it into a model that becomes complete by introducing additional variables.

* Convergence is not guaranteed, but we do guarantee $\ell(\boldsymbol{\theta}^{(\nu+1)}) \geq \ell(\boldsymbol{\theta}^{(\nu)})$., with equality if Q and f are the same.
* If ℓ is bounded, then $\ell(\boldsymbol{\theta}^{(\nu)}) \to a$ for some constant a.
  - With some additional conditions, $a = \hat{\boldsymbol{\theta}}_n$.
* If the M-step leads to analytical expressions for updates, EM converges very fast.

– **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} f(y;\boldsymbol{\theta},\boldsymbol{p}) = \sum_{j=1}^{3} p_j f_j(y;\boldsymbol{\theta}_j)$, where $\sum_{j=1}^{3} p_j = 1$. Construct the log-likelihood ordinarily, then describe the steps for EM.

$$\ell(\boldsymbol{\theta},\boldsymbol{p};\underline{Y}) \overset{\text{iid}}{=} \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{3} p_j f_j(y;\boldsymbol{\theta}_j)\right\};$$

This is difficult to maximize directly due to the sum term in the logarithm.
Define $(Z_{i1}, Z_{i2}, Z_{i3})' \overset{\text{iid}}{\sim}\text{MultNom}(1, p_1, p_2, p_3)$; notice that $f(Y_i|Z_{ij}) = f_j(Y_i;\boldsymbol{\theta}_j)^{Z_{ij}}$. Therefore, $f(Y_i, Z_{ij}) \propto (p_j f_j(Y_i;\boldsymbol{\theta}_j))^{Z_{ij}}$, and

$$\ell(\boldsymbol{\theta},\boldsymbol{p}) = \sum_{i=1}^{n}\sum_{j=1}^{3}\left[Z_{ij}\left(\log(p_j) + \log f_j(Y_i;\boldsymbol{\theta}_j)\right)\right],$$

which is much easier to maximize directly.
Expectation Step: $Q\left(\boldsymbol{\theta},\boldsymbol{p},\boldsymbol{\theta}^{(\nu)},\boldsymbol{p}^{(\nu)},\underline{Y}\right) = \mathbb{E}_{\boldsymbol{\theta}^{(\nu)},\boldsymbol{p}^{(\nu)}}[\ell(\boldsymbol{\theta},\boldsymbol{p})|Y]$.
Define $w_{ij}^{(\nu)} := \mathbb{E}_{\boldsymbol{\theta}^{(\nu)},\boldsymbol{p}^{(\nu)}}(Z_{ij}|Y_i) = \frac{p_j^{(\nu)} f_j(Y_i;\boldsymbol{\theta}^{(\nu)})}{\sum_{k=1}^{3} p_k^{(\nu)} f_k(Y_i;\boldsymbol{\theta}^{(\nu)})}$. Thus,

$$Q\left(\boldsymbol{\theta},\boldsymbol{p},\boldsymbol{\theta}^{(\nu)},\boldsymbol{p}^{(\nu)},\underline{Y}\right) = \sum_{i=1}^{n}\sum_{j=1}^{3}\left[w_{ij}^{(\nu)}\left(\log(p_j) + \log f_j(Y_i;\boldsymbol{\theta}_j)\right)\right].$$

Maximization Step: $(\boldsymbol{\theta}^{(\nu+1)},\boldsymbol{p}^{(\nu+1)}) = \arg\max_{(\boldsymbol{\theta},\boldsymbol{p})} Q\left(\boldsymbol{\theta},\boldsymbol{p},\boldsymbol{\theta}^{(\nu)},\boldsymbol{p}^{(\nu)},\underline{Y}\right)$. ■

– **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma^2)$, but we lost the sign of the last few observations, so we observe $\underline{Y}' = Y_1,\ldots,Y_q,|Y_{q+1}|,\ldots,|Y_n| = |Y_i|\mathbb{I}(Y_i > 0) - |Y_i|\mathbb{I}(Y_i \leq 0)$. Suppose we know $w_i(\mu,\sigma) = \mathbb{E}\left[\mathbb{I}(Y_i > 0)||Y_i|\right] = P(Y_i > 0||Y_i|)$. Give the likelihood for the data, then describe EM, calculating the M-step for μ.
For $i = 1,\ldots,q$, this is the ordinary Normal density, since we have all of the data. For $i > q$, the density of Y_i is actually a folded Normal distribution, with means μ and variance σ^2.

$$\begin{aligned}
L(\mu,\sigma^2|\underline{Y}) =& \prod_{i=1}^{q} f(Y_i;\mu,\sigma^2)\prod_{i=q+1}^{n} f(Y_i;\mu,\sigma^2)\\
=& \prod_{i=1}^{q}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}\\
&\times\prod_{i=q+1}^{n}(2\pi\sigma^2)^{-1/2}\left[\exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}+\exp\left\{-\frac{1}{2\sigma^2}(Y_i+\mu)^2\right\}\right]\\
=&(2\pi\sigma^2)^{-n/2}\left[\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2\right\}+\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{q}(Y_i-\mu)^2\right)\left(\sum_{i=q+1}^{n}(Y_i+\mu)^2\right)\right\}\right].
\end{aligned}$$

Define $Z_i := 2\mathbb{I}(Y_i > 0) - 1$. This means that $Y_i = |Y_i|Z_i$.

$$\begin{aligned}
L_C(\mu,\sigma^2|\underline{Y}') =& f_{joint}(\underline{Y}') \overset{\perp}{=} \prod_{i=1}^{q}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}\prod_{i=q+1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(|Y_i|Z_i-\mu)^2\right\}\\
=&(2\pi\sigma^2)^{-n/2}\exp\left\{-\sum_{i=1}^{q}\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}\exp\left\{-\sum_{i=q+1}^{n}\frac{1}{2\sigma^2}(|Y_i|Z_i-\mu)^2\right\};
\end{aligned}$$

$$\begin{aligned}
\ell_C(\mu,\sigma^2|\underline{Y}') =& \cdots = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{q}\frac{1}{2\sigma^2}(Y_i-\mu)^2 - \sum_{i=q+1}^{n}\frac{1}{2\sigma^2}(|Y_i|Z_i-\mu)^2\\
=& \cdots = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}Y_i^2 + n\mu^2 + \sum_{i=1}^{q}(-2\mu Y_i) + \sum_{i=q+1}^{n}(-2\mu|Y_i|Z_i)\right].
\end{aligned}$$

We first note that

$$\mathbb{E}\left(Z_i \middle| |Y_i|, \mu, \sigma^2\right) = 2\mathbb{E}\left[\mathbb{I}_{Y_i>0}|Y_i|\right] - 1 = 2w_i - 1,$$

where $w_i = w_i(\mu^{(\nu)}, (\sigma^2)^{(\nu)}))$. At the M step, we maximize Q (calculated below) with respect the parameters of interest.

$$\begin{aligned}
Q(\mu, \sigma^2; \mu^{(\nu)}, (\sigma^2)^{(\nu)}) &= \mathbb{E}_{\mu^{(\nu)}, (\sigma^2)^{(\nu)}}\left[\ell_C(\mu^{(\nu)}, (\sigma^2)^{(\nu)} | \underline{Y}')\right] \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} Y_i^2 + n\mu^2 + \sum_{i=1}^{q}(-2\mu Y_i) + \sum_{i=q+1}^{n}\left(-2\mu|Y_i| \cdot \mathbb{E}_{\mu^{(\nu)}, (\sigma^2)^{(\nu)}}\left(Z_i \middle| |Y_i|, \mu, \sigma^2\right)\right)\right] \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} Y_i^2 + n\mu^2 + \sum_{i=1}^{q}(-2\mu Y_i) + \sum_{i=q+1}^{n}\left(-2\mu|Y_i| \cdot (2w_i - 1)\right)\right].
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mu} Q(\mu, \sigma^2; \mu^{(\nu)}, (\sigma^2)^{(\nu)}) &= \cdots = 0 - \frac{1}{2\sigma^2}\left[2n\mu - 2\sum_{i=1}^{q} Y_i - 2\sum_{i=q+1}^{n}|Y_i|(2w_i - 1)\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{q} Y_i + \sum_{i=q+1}^{n}|Y_i|(2w_i - 1)\right] - \frac{n\mu}{\sigma^2} \overset{\text{set}}{=} 0 \\
\implies \mu^{(\nu+1)} &= \frac{1}{n}\left[\sum_{i=1}^{q} Y_i + \sum_{i=q+1}^{n}|Y_i|(2w_i - 1)\right].
\end{aligned}$$

Verify that this is a maximizer with the second derivative.

$$\frac{\partial^2}{\partial\mu^2} Q(\mu, \sigma^2; \mu^{(\nu)}, (\sigma^2)^{(\nu)}) = -\frac{n}{\sigma^2} < 0. \ \blacksquare$$

Example: Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta) = \theta\exp\{-\theta y\}$ for $y \geq 0$. We observe $\max\{c, Y_i\}$. Suppose the first m terms equal c, whereas the remaining $(n-m)$ values are Y_i. Write down the likelihood for the observed data. Consider EM. Write down the complete log-likelihood. At the $(\nu+1)$th iteration, obtain Q. Give the formula of $\theta^{(\nu+1)}$.
The data is $\{Y_i^{obs}, \delta_i\}_{i=1}^n$, where $Y_i^{obs} = \delta_i Y_i + (1-\delta_i)c$, where $\delta_i = \mathbb{I}(Y_i > c)$. Notice that

$$P(Y_i \leq c) = \int_0^c f(y;\theta)dy = 1 - \exp\{-c\theta\} = 1 - P(\delta_i = 1).$$

We know that $\delta_i = 0$ for all $i \in \{1, \ldots, m\}$, and $\delta_i = 1$ for $i \in \{m+1, \ldots, n\}$. Therefore,

$$L(\theta; Y_i^{obs}) = \prod_{i=1}^{n}\left[f(Y_i;\theta)\right]^{\delta_i}\left[F_Y(c)\right]^{1-\delta_i} = \theta^{n-m}[1 - \exp\{-c\theta\}]^m \exp\left[-\theta\sum_{i=1}^{n} Y_i\right].$$

The complete data is simply $Y_1, \ldots, Y_n$, so

$$\ell_C(\theta) = \sum_{i=1}^{n}\log f(Y_i;\theta) = n\log(\theta) - \theta\sum_{i=1}^{n} Y_i.$$

$$\begin{aligned}
Q(\theta, \theta^{(\nu)}; \{Y_i\}_{i=1}^n) &= \mathbb{E}_{\theta^{(\nu)}}[\ell_C(\theta) | \{Y_i^{obs}, \delta_i\}_{i=1}^n] = n\log(\theta) - \theta\mathbb{E}_{\theta^{(\nu)}}\left[\sum_{i=1}^{n} Y_i | \{Y_i^{obs}, \delta_i\}_{i=1}^n\right] \\
&= n\log(\theta) - \theta\sum_{i=m+1}^{n} Y_i - m\theta\mathbb{E}_{\theta^{(\nu)}}[Y_i | Y_i \leq c];
\end{aligned}$$

$$\begin{aligned}
P(Y_i \leq y | Y_i \leq c) &= \frac{P(Y_i \leq y)}{P(Y_i \leq c)}\mathbb{I}(y < c); \\
f(y | y < c; \theta) &= \frac{\theta\exp\{-y\theta\}}{1 - \exp\{-c\theta\}}\mathbb{I}(y < c); \\
\mathbb{E}_{\theta^{(\nu)}}[Y_i | Y_i \leq c] &= \frac{1}{1 - \exp\{-c\theta^{(\nu)}}\int_0^c \theta^{(\nu)} y\exp -y\theta^{(\nu)} dy = \frac{1}{\theta^{(\nu)}}\left[1 - \frac{c\theta^{(\nu)}}{\exp\{c\theta^{(\nu)}\} - 1}\right]; \\
Q(\theta; \theta^{(\nu)}) &= n\log(\theta) - \theta\sum_{i=m+1}^{n} Y_i - m\frac{\theta}{\theta^{(\nu)}}\left[1 - \frac{c\theta^{(\nu)}}{\exp\{c\theta^{(\nu)}\} - 1}\right].
\end{aligned}$$

$$\theta^{(\nu+1)} = \arg\max_{\theta} Q(\theta;\theta^{(\nu)};\{Y_i\}_{i=1}^n);$$

$$\frac{\partial}{\partial\theta}Q(\theta;\theta^{(\nu)};\{Y_i\}_{i=1}^n) = \frac{n}{\theta} - \sum_{i=m+1}^{n} Y_i - \frac{m}{\theta^{(\nu)}}\left[1 - \frac{c\theta^{(\nu)}}{\exp\{c\theta^{(\nu)}\}-1}\right] \overset{\text{set}}{=} 0$$

$$\implies \theta^{(\nu+1)} = n\left[\sum_{i=m+1}^{n} Y_i + \frac{m}{\theta^{(\nu)}}\left(1 - \frac{c\theta^{(\nu)}}{\exp\{c\theta^{(\nu)}\}-1}\right)\right]^{-1}. \blacksquare$$

– **Example**: Suppose that $T_1,\ldots,T_n \overset{\perp}{\sim} f_i(t;\beta,\theta) = \beta x_i - \frac{t}{\theta}\exp\{\beta x_i\} - \log\theta$. Denote by $(\hat{\beta}_n\ \hat{\theta}_n)^T$ the MLE. Derive $\ell(\beta,\theta)$. Then, state the consistency result for the MLE, in the context of this problem. Now, assume $S(\beta,\theta) = \begin{pmatrix}\frac{\partial\ell}{\partial\beta}\\ \frac{\partial\ell}{\partial\theta}\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n x_i - \frac{1}{\theta}\sum_{i=1}^n T_i x_i e^{\beta x_i}\\ \frac{1}{\theta^2}\sum_{i=1}^n T_i e^{\beta x_i} - \frac{n}{\theta}\end{pmatrix}$. Calculate the FI matrix for the entire sample (the Total Information). You can use $\mathbb{E}[T_i] = \theta\exp\{-\beta x_i\}$ without proof. Lastly, assume that $\frac{1}{n}\sum_{i=1}^n \begin{pmatrix}x_i\\ x_i^2\end{pmatrix} \to \begin{pmatrix}0\\ b\end{pmatrix}$ as $n\to\infty$. Give the asymptotic distribution of the MLE.

$$L(\beta,\theta) = \prod_{i=1}^{n}\left[\beta x_i - \frac{t}{\theta}\exp\{\beta x_i\} - \log\theta\right] \to \ell(\beta,\theta) = \beta\sum_{i=1}^{n} x_i - \frac{1}{\theta}\sum_{i=1}^{n} T_i\exp\{\beta x_i\} - n\log\theta.$$

The consistency result is that $\begin{pmatrix}\hat{\beta}_n\\ \hat{\theta}_n\end{pmatrix} \xrightarrow{\text{p}} \begin{pmatrix}\beta\\ \theta\end{pmatrix}$.

$$I_T(\beta,\theta) = -\,\mathbb{E}\left[\frac{\partial}{\partial(\beta,\theta)}S(\beta,\theta)\right] = \mathbb{E}\begin{bmatrix}\frac{1}{\theta}\sum_{i=1}^n T_i x_i^2 e^{-\beta x_i} & -\frac{1}{\theta^2}\sum_{i=1}^n T_i x_i e^{-\beta x_i}\\ -\frac{1}{\theta^2}\sum_{i=1}^n T_i e^{\beta x_i} & \frac{2}{\theta^3}\sum_{i=1}^n T_i e^{\beta x_i} - \frac{n}{\theta^2}\end{bmatrix}$$

$$= \begin{bmatrix}\frac{1}{\theta}\sum_{i=1}^n \theta e^{-\beta x_i} x_i^2 e^{-\beta x_i} & -\frac{1}{\theta^2}\sum_{i=1}^n \theta e^{-\beta x_i} x_i e^{-\beta x_i}\\ -\frac{1}{\theta^2}\sum_{i=1}^n \theta e^{-\beta x_i} x_i e^{-\beta x_i} & \frac{2}{\theta^3}\sum_{i=1}^n \theta e^{-\beta x_i} e^{-\beta x_i} - \frac{n}{\theta^2}\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^n x_i^2 & -\frac{1}{\theta}\sum_{i=1}^n x_i\\ -\frac{1}{\theta}\sum_{i=1}^n x_i & \frac{n}{\theta^2}\end{bmatrix}.$$

Using the CLT for the MLE, we know that

$$\sqrt{n}\left(\begin{pmatrix}\hat{\beta}_n\\ \hat{\theta}_n\end{pmatrix} \to \begin{pmatrix}\beta\\ \theta\end{pmatrix}\right) \xrightarrow{\text{d}} \mathcal{N}\left(\mathbf{0}, I(\beta,\theta)^{-1}\right).$$

Using the consistency result,

$$\frac{1}{n}I_T(\beta,\theta) \to \begin{pmatrix}b & 0\\ 0 & \sigma^{-2}\end{pmatrix} = I(\beta,\theta). \text{ Therefore, } \begin{pmatrix}\hat{\beta}_n\\ \hat{\theta}_n\end{pmatrix} \sim AN\left(\mathbf{0}, \begin{pmatrix}1/(nb) & 0\\ 0 & \theta^2/n\end{pmatrix}\right). \blacksquare$$

7.2 Asymptotics

Return to Table of Contents

- **Almost Sure Convergence**: For $Y_1,\ldots,Y_n \overset{\perp}{\sim} f_n(\cdot;\boldsymbol{\theta})$, $Y_n \xrightarrow{\text{a.s.}} Y$ if $P\left(\lim_{n\to\infty} Y_n = Y\right) = 1$.
  - For vectors: same definition.
  - Vector convergence $\Leftrightarrow$ individual convergence.
  - $\underline{Y}_n \xrightarrow{\text{a.s.}} \underline{Y} \implies \prod_{i=1}^k Y_{ni} \xrightarrow{\text{a.s.}} \prod_{i=1}^k Y_i$ and $\sum_{i=1}^k Y_{ni} \xrightarrow{\text{a.s.}} \sum_{i=1}^k Y_i$.
- **Convergence in Probability**: For $Y_1,\ldots,Y_n \overset{\perp}{\sim} f_n(\cdot;\boldsymbol{\theta})$, $Y_n \xrightarrow{\text{p}} Y$ if $\lim_{n\to\infty} P\left(|Y_n - Y| > \epsilon\right) = 0\ \forall\epsilon > 0$.

- $\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{\text{p}}$.
- For vectors: $\lim_{n\to\infty} P\left(\|\underline{Y}_n - \underline{Y}\| < \epsilon\right) = 1 \ \forall \epsilon > 0$.
- Vector convergence $\Leftrightarrow$ individual convergence.
- **Example**: Consider the linear regression setting

$$Y_i = \alpha + \beta x_i + \epsilon_i,\ i \in \{1, \ldots, n\},$$

where x_i are known constants, and $\epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$. The least squares estimator has the representation

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Assume $\sum_{i=1}^n (x_i - \bar{x})^2 \to \infty$ as $n \to \infty$. Show that $\hat{\beta} \xrightarrow{\text{p}} \beta$ as $n \to \infty$.
Use Markov's inequality.

$$\begin{aligned} P\left(|\hat{\beta} - \beta| \geq \epsilon\right) \leq & \frac{\mathbb{E}\left[(\hat{\beta} - \beta)^2\right]}{\epsilon^2}; \\ \mathbb{E}\left[(\hat{\beta} - \beta)^2\right] = & \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \mathbb{E}(\epsilon_i^2) + \sum_{i=1}^n \sum_{j\neq i} (x_i - \bar{x})(x_j - \bar{x})\mathbb{E}(\epsilon_i \epsilon_j)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \\ = & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \to 0. \end{aligned}$$

Note that $\epsilon_i \perp \epsilon_j$, and the given assumption. With this in mind, $\hat{\beta} \xrightarrow{\text{p}} \beta$ as $n \to \infty$. ■

- **Convergence in Distribution**: For $Y_1, \ldots, Y_n \overset{\perp}{\sim} f_n(\cdot; \boldsymbol{\theta})$, $Y_n \xrightarrow{\text{d}} Y$ if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$ for all y.
  - $\xrightarrow{\text{p}} \Rightarrow \xrightarrow{\text{d}}$, with equality if converging to a constant.
  - For vectors: $\lim_{n\to\infty} F_{Y_n}(\underline{y}) = F_Y(\underline{y})$.
  - Individual convergence $\to$ vector convergence.
- **Uniform Convergence**: F_n converges uniformly to F is $\forall \epsilon > 0$, $\exists N : \forall n \geq N$, $\sup_y |F_{Y_n}(y) - F_Y(y)| < \epsilon$.
- **Markov's Inequality**: $P(|X| > a) \leq \frac{\mathbb{E}(|X|^r)}{a^r}$ for $r, a > 0$.
- **Chebyshev's Inequality**: If $\mathbb{E}(X) = 0$, then $P\left(|X - \mathbb{E}(X)| > a\right) \leq \frac{Var(X)}{a^2}$.
- **WLLN**: For $Y_i \overset{\text{iid}}{\sim} f(\cdot; \boldsymbol{\theta})$, $\bar{Y} \xrightarrow{\text{p}} \mathbb{E}(Y_i)$.
- **SLLN**: For $Y_i \overset{\text{iid}}{\sim} f(\cdot; \boldsymbol{\theta})$, $\bar{Y} \xrightarrow{\text{a.s.}} \mathbb{E}(Y_i)$.
- **CLT**: Suppose $Y_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, where both parameters are finite. Then, $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$.
  - For vectors: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \Sigma)$, where Σ is positive-definite. Then, $\sqrt{n}(\underline{X}_n - \mu) \xrightarrow{\text{d}} \mathcal{N}_K(0, \Sigma)$.
- **Continuous Mapping Theorem**, or **CMT**: Suppose g is a continuous function. Then, $\underline{Y}_n \xrightarrow{\text{a.s.}} \underline{Y} \implies g(\underline{Y}_n) \xrightarrow{\text{a.s.}} g(\underline{Y})$. Works also for $\xrightarrow{\text{p}}$ and $\xrightarrow{\text{d}}$.
- **Slutsky's Theorem**: Suppose $\underline{Y}_n \xrightarrow{\text{d}} \underline{Y}$:
  - If $X_n \xrightarrow{\text{p}} a$ and $Z_n \xrightarrow{\text{p}} b$, then $X_n \underline{Y}_n + Z_n \xrightarrow{\text{d}} a\underline{Y} + b$.
  - If $X_n \xrightarrow{\text{p}} A \in \mathbb{R}^{m\times k}$, and $Z_n \xrightarrow{\text{p}} B \in \mathbb{R}^{m\times 1}$, then $X_n \underline{Y}_n + Z_n \xrightarrow{\text{d}} A\underline{Y} + B$.
  - Note that we need $\xrightarrow{\text{p}}$ for at least one term! Counterexample includes $\boldsymbol{X}_n, \boldsymbol{Y}_n \overset{\perp}{\sim} \mathcal{N}(0, 1)$. $\boldsymbol{X}_n - \boldsymbol{Y}_n \sim \mathcal{N}(0, 2)$, which is not $\mathbf{0}$.
  - **Example**: Suppose $\boldsymbol{Y}_n \in \mathbb{R}^k \xrightarrow{\text{d}} \boldsymbol{Y}$. Also suppose that $\boldsymbol{Z}_n \in \mathbb{R}^{k\times k} \xrightarrow{\text{p}} C$. Show that $\boldsymbol{Y}_n \boldsymbol{Z}_n \boldsymbol{Y}_n \xrightarrow{\text{d}} \boldsymbol{Y}^T C \boldsymbol{Y}$ as $n \to \infty$.
    Using Slutsky's Theorem, we know that $(\boldsymbol{Y}_n, \boldsymbol{Z}_n) \xrightarrow{\text{d}} (\boldsymbol{Y}, C)$ as $n \to \infty$. Applying CMT, we get that

$$g(\boldsymbol{Y}_n, \boldsymbol{Z}_n) = \boldsymbol{Y}_n^T \boldsymbol{Z}_n \boldsymbol{Y}_n \xrightarrow{\text{d}} g(\boldsymbol{Y}, C) = \boldsymbol{Y}^T C \boldsymbol{Y}.$$

– **Example**: Consider iid samples $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ with respective means μ_1 and μ_2, and with equal, unknown variance σ^2. For testing $H_0 : \mu_1 = \mu_2$, we use

$$t_p = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}}, \text{ where } s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}.$$

Assume that $\lambda_{m,n} = \frac{m}{m+n} \to \lambda > 0$ as $\min\{m,n\} \to \infty$. Using this, prove that $t_p \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

$$\begin{aligned} t_p =& \frac{1}{s_p}\left[(\bar{X}_m - \mu)\sqrt{\frac{nm}{(n+m)}} - (\bar{Y}_n - \mu)\sqrt{\frac{nm}{(n+m)}}\right] \\ =& \frac{1}{s_p}\left[(\bar{X}_m - \mu)\sqrt{m}\sqrt{1-\lambda_{m,n}} - (\bar{Y}_n - \mu)\sqrt{n}\sqrt{\lambda_{m,n}}\right]; \end{aligned}$$

Using the CLT, we know that $\sqrt{m}(\bar{X}_m - \mu) \xrightarrow{\text{d}} \mathcal{N}(0,\sigma^2)$, and $\sqrt{n}(\bar{Y}_n\mu) \xrightarrow{\text{d}} \mathcal{N}(0,\sigma^2)$. Using Slutsky's Theorem, we get that $(\bar{X}_m - \mu)\sqrt{m}\sqrt{1-\lambda_{m,n}} \xrightarrow{\text{d}} \mathcal{N}\left(0,(1-\lambda)\sigma^2\right)$, and $(\bar{Y}_n - \mu)\sqrt{n}\sqrt{\lambda_{m,n}} \xrightarrow{\text{d}} \mathcal{N}(0,\lambda\sigma^2)$. When combined with independent samples, we get that

$$(\bar{X}_m - \mu)\sqrt{m}\sqrt{1-\lambda_{m,n}} - (\bar{Y}_n - \mu)\sqrt{n}\sqrt{\lambda_{m,n}} \xrightarrow{\text{d}} \mathcal{N}(0,\sigma^2).$$

We can rewrite s_p to form

$$s_p = \sqrt{\lambda_{m-1,n-1}s_X^2 + (1-\lambda_{m-1,n-1})s_Y^2}.$$

We know that s_X^2 and s_Y^2 are consistent. When combined with the convergence of $\lambda_{m-1,n-1}$, $s_p \xrightarrow{\text{p}} \sigma$ as $\min\{m,n\} \to \infty$ by continuity. Applying Slutsky's theorem, we get that $t_p \xrightarrow{\text{d}} \mathcal{N}(0,1)$. ■

- **Cramer-Wold Theorem**: $\underline{Y}_n \xrightarrow{\text{d}} \underline{Y} \Leftrightarrow \forall t \in \mathbb{R}^k$, $t^T\underline{Y}_n \xrightarrow{\text{d}} t^T\underline{Y}$.
- $Y_n = O_p(1)$ if Y_n is bounded in probability. That is, $\forall \epsilon > 0$, $\exists M_\epsilon > 0$ and $n_0 \geq 1$ such that $\forall n \geq n_0$, $P(\|Y_n\| < M_\epsilon) > 1 - \epsilon$.
  - $O_p(1) \equiv Y_n$ is a tight sequence.
  - $Y_n \xrightarrow{\text{d}} Y \Leftrightarrow Y_n = O_p(1)$.
  - $Y_n = O_p(X_n)$ is $Y_n = X_n \cdot O_p(1)$.
- $Y_n = o_p(1)$ if $Y_n \xrightarrow{\text{p}} 0$.
  - $Y_n = o_p(X_n)$ if $Y_n = X_n \cdot o_p(1)$, or $\frac{Y_n}{X_n} \xrightarrow{\text{p}} 0$.
  - Suppose $X_n = o_p(a_n)$ and $Y_n = o_p(b_n)$:
    - $X_nY_n = o_p(a_nb_n)$.
    - $X_n + Y_n = o_p(\max\{a_n, b_n\})$.
  - Suppose $Y_n = O_p(X_n)$, and $X_n = o_p(1)$. Then, $Y_n = o_p(1)$.
  - Suppose $X_n = o_p(1)$, and $Y_n = (1+X_n)^{-1}$. Then, $Y_n = O_p(1)$.
- If $\underline{Y}_n \xrightarrow{\text{p}} 0$, and $g : \mathbb{R}^k \to \mathbb{R}$, such that $g(0) = 0$. If $g(h) = o(\|h\|^p)$ for some p, then $g(\underline{Y}) = o_p(\|\underline{Y}_n\|^p)$. Same applies for $O_p(\cdot)$.
- **Asymptotic Normal**: $Y_n \sim AN(\mu_n, \sigma_n^2)$ if $\frac{Y_n - \mu_n}{\sigma_n^2} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.
  - μ_n can be random.
  - If $Y_n \sim AN(\mu_n, \sigma_n^2)$, then μ_n is not necessarily the mean of Y_n.
  - Suppose $Y_n \sim AN(\mu_n, \sigma_n^2)$. Then, $Y_n \xrightarrow{\text{p}} \mu \Leftrightarrow \lim_{n\to\infty} \sigma_n^2 = 0$.
  - For vectors: $\underline{Y}_n \sim AN(\underline{\mu}_n, \Sigma_n)$ if $\forall c \in \mathbb{R}^k$ such that $c^T\Sigma_n c > 0\ \forall n \geq n_0$, then $c^T\underline{Y}_n \sim AN(c^T\underline{\mu}_n, c^T\Sigma_n c)$.
  - If $\underline{Y}_n \sim AN(\underline{\mu}_n, \Sigma_n)$, then $\underline{Y}_n \xrightarrow{\text{p}} \underline{\mu} \Leftrightarrow \Sigma_n \to \mathbf{0}$.
  - $\underline{Y}_n \sim AN(\underline{\mu}_n, b_n^2\Sigma_n)$ for positive-definite Σ if $\frac{\underline{Y}_n - \underline{\mu}_n}{b_n} \xrightarrow{\text{d}} \mathcal{N}(0,\Sigma)$.

- **First Order Delta Method**: Suppose $Y_n \xrightarrow{\text{p}} \theta$, and g is continuous at θ. Suppose we also know that $\frac{Y_n-\theta}{b_n} \xrightarrow{\text{d}} Y$, or $Y_n = \theta + b_n O_p(1)$. If g is differentiable, and $g'(\theta) \neq 0$, then $\frac{g(Y_n)-g(\theta)}{b_n} \xrightarrow{\text{d}} g'(\theta)Y$, or $g(Y_n) \approx g(\theta) + b_n O_p(1)$.
  - Is a Taylor expansion about θ.
  - If $Y \sim \mathcal{N}(0, \sigma^2)$, then $Y_n \sim AN(\theta, b_n^2\sigma^2)$, so $g(Y_n) \sim AN\left(g(\theta), b_n^2\sigma^2[g'(\theta)^2]\right)$.
  - If g is continuous at θ, then $b_n = o_p(1)$.
  - If $g : \mathbb{R}^k \to \mathbb{R}$ is differentiable, and $g'(\theta) \neq 0$, and Y_n is a sequence such that $Y_n \sim AN(\theta, b_n^2\Sigma)$, then $g(Y_n) \sim AN\left(g(\theta), b_n^2 g'(\theta)\Sigma g'(\theta)^T\right)$.
    - This applies piecewise, where for $D(\theta) = \begin{pmatrix} g_1(\theta) & \dots & g_m(\theta) \end{pmatrix}^T$, $g(Y_n) \sim AN\left(g(\theta), b_n^2 D(\theta)\Sigma D(\theta)^T\right)$.
  - **Example**: Suppose $X_1, X_2 \overset{\perp}{\sim} Bin(n_i, p_i)$. We are interested in the odds ratio $\theta = \frac{p_1(1-p_2)}{p_2(1-p_1)}$. $\hat{\theta}_n$ uses $\hat{p}_i = \frac{X_i}{n_i}$, where $n = n_1 + n_2$. Show that $Var\left[\log(\hat{\theta}_n)\right] \doteq \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}$.

    $\log(\hat{\theta}_n) = \log(\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log(\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right)$. Represent X_i as $\sum_{j=1}^{n_i} X_{ij}$, where $X_{ij} \overset{\perp}{\sim} Ber(p_i)$. By the CLT,

    $$\sqrt{n_1}(\hat{p}_1 - p_1) \xrightarrow{\text{d}} N(0, p_1(1-p_1)), \text{ and } \sqrt{n_2}(\hat{p}_2 - p_2) \xrightarrow{\text{d}} N(0, p_2(1-p_2)).$$

    Consider $g(x) = \log\left(\frac{x}{1-x}\right)$. Then, $g'(x) = \frac{1}{x(1-x)} \neq 0$ for p_1 and p_2. Therefore, by the Delta Theorem,

    $$\sqrt{n}_i \left[\log(\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) - \log(\left(\frac{p_i}{1-p_i}\right)\right] \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{1}{p_i(1-p_i)}\right).$$

    This means that the asymptotic variance of $\log(\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$ is $\frac{1}{n_i p_i(1-p_i)}$. With independent samples,

    $$Var(\hat{\theta}_n) = \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}. \blacksquare$$

  - **Example**: Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be iid pairs such that $E(X) = \mu_1$, $E(Y) = \mu_2$, $Var(X) = \sigma_1^2$, $Var(Y) = \sigma_2^2$, and $Cov(X, Y) = \sigma_{12}$. Determine the asymptotic distribution of $(\bar{X}, \bar{Y})^T$. Then, suppose that $\mu_1 = \mu_2 = 0$, and define $T := \overline{XY}$. Show that $nT \xrightarrow{\text{d}} Q$ for some RV Q. Then, suppose $\mu_1 = 0$ and $\mu_2 \neq 0$. Show that $\sqrt{n}T \xrightarrow{\text{d}} R$ for some RV R.

    Following the CLT for both $\bar{X}$ and $\bar{Y}$, $\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \sim AN\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$.

    By magic, $\overline{XY} = \frac{1}{4}(\bar{X} + \bar{Y})^2 - \frac{1}{4}(\bar{X} - \bar{Y})^2$. Define $g(X, Y) = X + Y$ and $h(X, Y) = X - Y$. By the Delta Theorem,

    $$\sqrt{n}(\bar{X} + \bar{Y}) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_1^2 + \sigma_2^2 + \sigma_{12}), \text{ and } \sqrt{n}(\bar{X} - \bar{Y}) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_1^2 + \sigma_2^2 - \sigma_{12}).$$

    Therefore,

    $$\frac{n}{4}(\bar{X} + \bar{Y})^2 \xrightarrow{\text{d}} \frac{\sigma_1^2 + \sigma_2^2 + \sigma_{12}}{4}\chi_1^2, \text{ and } \frac{n}{4}(\bar{X} - \bar{Y})^2 \xrightarrow{\text{d}} \frac{\sigma_1^2 + \sigma_2^2 - \sigma_{12}}{4}\chi_1^2.$$

    Therefore, $nT \xrightarrow{\text{d}} Q := A - B + C$, where $A, B \overset{\text{iid}}{\sim} \frac{\sigma_1^2+\sigma_2^2}{4}\chi_1^2$, and $C \sim \sigma_{12}\chi_2^2$.
    By the CLT,

    $$\sqrt{n}(\bar{X} - 0) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_1^2), \text{ and } \sqrt{n}(\bar{Y} - \mu_2) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_2^2).$$

    Define $g(X, Y) = XY$. Then, $g'(X, Y) = (Y, X)$, and $g'(\mu_1, \mu_2) \neq \mathbf{0}$. Therefore, by the Delta Theorem,

    $$\sqrt{n}(T - 0) \xrightarrow{\text{d}} R := \mathcal{N}\left(0, g'(0, \mu_2)\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} g'(0, \mu_2)'\right) = \mathcal{N}(0, \mu_2^2\sigma_1^2). \blacksquare$$

- **Approximation by Averages Approximation**, or **Bahadur Approximation**, or **ABAR**: Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} f$. Statistic T_n based on $\underline{X}$ admits an ABAR if T_n can be decomposed into

$$T_n = \underbrace{T_\infty}_{T_n \xrightarrow{\text{p}} T_\infty} + \frac{1}{n}\sum_{i=1}^{n} \underbrace{h_T(X_i)}_{E(h_T(X_i))=0, Var(h_T(X_i))<\infty} + \underbrace{R_n}_{o_p(n^{-1/2})}.$$

- $\frac{1}{n}\sum_{i=1}^n h_T(X_i) = O_p(n^{-1/2})$.
- $T_\infty = O_p(1)$.
- If T_n admits an ABAR, and $\sqrt{n}R_n = o_p(1)$ then $\sqrt{n}(T_n - T_\infty) \xrightarrow{\text{d}} \mathcal{N}(0, Var(h_T(X_1)))$.

- **Example**: ABAR for $\hat{\eta}_{0.75}$.

$$\hat{\eta}_{0.75} = \eta_{0.75} + \frac{1}{n}\sum_{i=1}^n \left(\frac{\frac{3}{4} - \mathbb{I}(X_i \leq \eta_{0.75})}{F'(\eta_{0.75})} \right) + R_{n,0.75}. \blacksquare$$

- **Example**: ABAR for σ^2.

$$s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \sigma^2 + \frac{1}{n}\sum_{i=1}^n \left[(X_i - \mu)^2 - \sigma^2\right] - (\mu - \bar{X})^2. \blacksquare$$

- **ABAR for Vectors**: Suppose $\underline{T}_n \in \mathbb{R}^{k\times 1} \xrightarrow{\text{p}} \underline{T}_\infty \in \mathbb{R}^{k\times 1}$. Denote $T_{n\ell}$ as the ℓth component of T_n. If each component of T_n admits and ABAR, then $\sqrt{n}(\underline{T}_n - \underline{T}_\infty) \xrightarrow{\text{d}} \mathcal{N}_k(\mathbf{0}, \Sigma)$, where $\Sigma = Cov\begin{pmatrix} h_{T_1}(X_i) & \dots & h_{T_k}(X_i) \end{pmatrix}^T$.

- **ABAR for Functions of Statistics**: Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} f$, and T_n admits an ABAR. Suppose $g : \mathbb{R}^k \to \mathbb{R}$ is differentiable near T_θ, then $g(T_n)$ admits an ABAR with

$$g(T_n) = g(T_\infty) + \frac{1}{n}\sum_{i=1}^n \triangledown g(T_\infty) h_T(X_i) + \triangledown g(T_\infty) R_n.$$

- **ABAR for Means of Functions Involving Statistics**: Suppose T_n admits an ABAR, and define $Q_n := \frac{1}{n}\sum_{i=1}^n g(T_n, X_i)$. Then, if:

  1. $Var(g(T_\infty, X_i)) < \infty$.
  2. $\mathbb{E}\left[g'_T(T_\infty, X_i)\right] < \infty$.
  3. $\exists M(X) : \forall T^*$ near T_∞, $|g''_{TT}(T^*, X_i)| < M(X)$, where $\mathbb{E}(M(X)) < \infty$.

  then Q_n admits an ABAR of

$$Q_n = Q_\infty + \frac{1}{n}\sum_{i=1}^n h_Q(X_i) + R_n^*,$$

  where $h_Q(X_i) = g(T_\infty, X_i) - \mathbb{E}\left[g(T_\infty, X_i)\right] + \mathbb{E}\left[g'_T(T_\infty, X_i)\right] h_T(X_i)$ and $Q_\infty = \mathbb{E}\left[g(T_\infty, X_i)\right]$.

- **Example**: ABAR of $\hat{\mu}_3$, where $\mu_3 := \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^3$.
  $g(x, t) = (x - t)^3$. Check conditions:

  1. $Var[(X_i - \mu)^3] < \infty$.
  2. $\mathbb{E}[(X_i - \mu)^2] < \infty$.
  3. $\exists M(X) : |X_i - T^*| < M(X)$ for all T^* near μ.

  Therefore, the ABAR for $\hat{\mu}_3$ is

$$Q_n = \mu_3 + \frac{1}{n}\sum_{i=1}^n \left[(X_i - \mu)^3 - \mu_3 - 3\sigma^2(X_i - \mu)\right] + o_p(n^{-1/2}). \blacksquare$$

- **Strong Consistency**: $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$.
- **Weak Consistency**: $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$.
- There are two approaches for proving consistency of the MLE:
  - **Cramer**: $\hat{\theta}_n$ solves the likelihood equations, $\frac{\partial \ell}{\partial \theta} = \mathbf{0}$.
  - **Wald**: $\hat{\theta}_n$ is a maximizer of $\ell_n(\theta)$.
- **Cramer Consistency of MLE (Univariate)**: Define three conditions:

  (A) Identifiability: $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \implies F(y; \boldsymbol{\theta}_1) \neq F(y; \boldsymbol{\theta}_2)$.

- Different values of $\boldsymbol{\theta}$ uniquely define different distributions in the parametric family.
- Mixture models where $p_i \approx 0$ may not satisfy this.

(B) $|\bar{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| = |\mathbb{E}_{\boldsymbol{\theta}_0} \log f(y_1; \boldsymbol{\theta})| = \left| \int \log f(y; \boldsymbol{\theta}) dF(y; \boldsymbol{\theta}_0) \right| < \infty$ for all $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$.

- Used for SLLN to hold.

(C) $\log f(y; \boldsymbol{\theta})$ has a continuous derivative wrt $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ for each y in the support of $F(y; \boldsymbol{\theta}_0)$.

- Guarantees that $\hat{\boldsymbol{\theta}}_n$ is a solution to the likelihood equations.

If $Y_1, \ldots, Y_n$ are iid, then there exists a strongly consistent solution of the likelihood equations.

- Proof uses the fact that $\bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0) - \bar{\ell}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) > 0$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

- **Cramer Consistency of MLE (Vector)**: Using the previous three conditions, define one more condition:

(D) Uniform SLLN: $\exists h(y) : |\log f(y; \theta)| < h(y)$ for all $y \in Supp(Y)$, and $\boldsymbol{\theta}$ in a compact neighborhood of $\boldsymbol{\theta}_0$ where $\mathbb{E}_{\boldsymbol{\theta}_0}[h(Y)] < \infty$.

- In the univariate case, strong consistency relies on exactly two solutions to $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| = \delta$, which is not true in higher dimensions.
- Ensures SLLN holds near $\boldsymbol{\theta}_0$.

Then, there exists a strongly consistent solution of the likelihood equations.

- **Wald Consistency of MLE (Vector)**: Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y; \boldsymbol{\theta}_0)$. Assume conditions (A), (B), and (D). Also assume that f is continuous wrt $\boldsymbol{\theta}$ for all y, and that $\boldsymbol{\Theta}$ is compact. If $\hat{\boldsymbol{\theta}}_n$ maximizes $\ell_n(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \boldsymbol{\theta}$.

- **Asymptotic Normality of MLE (Univariate)**: Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y; \theta_0)$, where θ_0 is an interior point of Θ, and $f(y; \theta)$ satisfies the following conditions:

A1: Identifiability: $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies that $F(y; \boldsymbol{\theta}_1) \neq F(y; \boldsymbol{\theta}_2)$ for at least one y.

A2: For each $\boldsymbol{\theta} \in \Theta$, $F(y; \boldsymbol{\theta})$ has the same support not depending on $\boldsymbol{\theta}$.

A3: $\forall \boldsymbol{\theta} \in \Theta$, the first three partial derivatives of $\log f(y; \boldsymbol{\theta})$ wrt $\boldsymbol{\theta}$ exist for y in the support of $F(y; \boldsymbol{\theta})$.

A4: For each $\boldsymbol{\theta}_0 \in \Theta$, there exists a function $M(y; \boldsymbol{\theta}_0)$ such that in a neighborhood of $\boldsymbol{\theta}_0$ and for all $j, k, l \in \{1, \ldots, b\}$, $\left| \frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log f(y; \boldsymbol{\theta}) \right| \leq M(Y)$ for all y and where $\mathbb{E}_{\boldsymbol{\theta}_o}(M(Y)) < \infty$.

A5: Correct model specification: $\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log f(Y; \theta) \right] = 0$, so $I(\theta) = \left[\left(\frac{\partial}{\partial \theta} \log f(Y; \theta) \right)^2 \right] = - \left[\frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) \right]$, and $0 < I(\theta) < \infty$.

Then, if $\hat{\boldsymbol{\theta}}_n$ is the solution to the likelihood equations, and $\hat{\boldsymbol{\theta}}_n \overset{\text{p}}{\to} \boldsymbol{\theta}_0$, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \overset{\text{d}}{\to} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}_0))$.

Example: Consider the density $f(y; \sigma) = \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\}$. Verify that the conditions are upheld.

A1:

$$F(y; \sigma) = \int_0^y \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{2}{\sigma^2} \int_0^{\sqrt{u}} y \exp\left\{ -\frac{u}{\sigma^2} \right\} \frac{1}{2y} du = 1 - \exp\left\{ -\frac{y^2}{\sigma^2} \right\}.$$

If $\sigma_1 \neq \sigma_2$ and $y = 1$, then $F(1; \sigma_1) \neq F(2; \sigma_2)$.

A2: The support of $F(y; \sigma)$ does not depend on σ.

A3: $\ell(f(y; \sigma)) = \log\left(\frac{2y}{\sigma^2}\right) - \frac{y^2}{\sigma^2}$. $\ell^{(3)}(y; \sigma) = -\frac{4}{\sigma^3} + 24\frac{y^2}{\sigma^5}$. Up to three derivatives exist for $y > 0$.

A4: Define $M(Y)$ first.

$$\begin{aligned} \left| -\frac{4}{\sigma_0^3} + 24\frac{y^2}{\sigma_0^5} \right| = \left| -4\sigma_0^{-3} + 24y^2\sigma_0^{-5} \right| \leq & | -4\sigma_0^{-3}| + |24y^2\sigma_0^{-5}| \\ = & 4\sigma_0^{-3} + 24y^2\sigma_0^{-5} =: M(Y); \end{aligned}$$

$$\int_0^\infty M(Y) dF(y; \sigma_0) dy = \int_0^\infty M(Y) f(y; \sigma_0) dy = (4\sigma_0^{-3} + 24y^2\sigma_0^{-5}) \left(2\frac{y}{\sigma_0} \exp\left\{ -\frac{y^2}{\sigma_0^2} \right\} \right).$$

A5:

$$\mathbb{E}\left(-\frac{2}{\sigma} + 2\frac{y^2}{\sigma^3} \right) = -\frac{2}{\sigma} + \frac{2}{\sigma^3} \int_0^\infty 2\frac{y^3}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = -\frac{2}{\sigma} + \frac{2}{\sigma^3} \cdot \sigma^2 = 0.$$

$$\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \sigma} \right)^2 \right] = \int_0^\infty \left[-\frac{2}{\sigma} + 2\frac{y^2}{\sigma^3} \right] \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{4}{\sigma^2}, \text{ and}$$

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \int_0^\infty \left[\frac{2}{\sigma^2} + 6\frac{y^2}{\sigma^4} \right] \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{4}{\sigma^2}. \blacksquare$$

7.3 Test Statistics and Confidence Intervals

Return to Table of Contents

Setup for this section:

- Assume $Y_i \overset{\text{iid}}{\sim} f(y;\boldsymbol{\theta})$.
- For scalars, $\theta \in \mathbb{R}$. For vector cases, $\boldsymbol{\theta} \in \mathbb{R}^r$. For partitioned cases, $\boldsymbol{\theta}_1 \in \mathbb{R}^r$, and $\boldsymbol{\theta} \in \mathbb{R}^b$.
- $h(\boldsymbol{\theta})$ is some function of H_o. Unless otherwise specified, $H_o : h(\boldsymbol{\theta}) = \mathbf{0}$.
  - $H(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^T} h(\boldsymbol{\theta})$.
- We are testing two-sided H_a's.
- Unless other specified, $T_* \overset{\text{d}}{\to} \chi^2_{\dim(\boldsymbol{\theta})}$ when H_o is true.
- **Wald Statistic for Scalars**: $T_W = (\hat{\theta}_n - \theta_0)^2 I_T(\hat{\theta}_n)$.
- **Wald Statistic for Vectors**: $T_W = (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I_T(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$.
- **Wald Statistic for Partitioned Hypotheses**: $T_W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^T \left[I_T^{11}(\hat{\boldsymbol{\theta}}_n)\right]^{-1} (\hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$, where $\boldsymbol{\theta}^T := (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$, $\left[I_T^{11}(\hat{\boldsymbol{\theta}}_n)\right]^{-1} = I_{T,11}(\hat{\boldsymbol{\theta}}_n) - I_{T,12}(\hat{\boldsymbol{\theta}}_n) I_{T,22}^{-1}(\hat{\boldsymbol{\theta}}_n) I_{T,21}(\hat{\boldsymbol{\theta}}_n)$, and we are testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs. $H_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$.
  - Using the CLT,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\boldsymbol{\theta}}_1\\ \hat{\boldsymbol{\theta}}_2\end{pmatrix} - \begin{pmatrix}\boldsymbol{\theta}_{10}\\ \boldsymbol{\theta}_{20}\end{pmatrix}\right) \overset{\text{d}}{\to} \mathcal{N}_b\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix} I^{11}(\boldsymbol{\theta}_0) & I^{12}(\boldsymbol{\theta}_0)\\ I^{21}(\boldsymbol{\theta}_0) & I^{22}(\boldsymbol{\theta}_0)\end{pmatrix}\right)$$

- **Wald Statistic for Functions**: $T_W = n h(\hat{\boldsymbol{\theta}})^T \left\{H(\hat{\boldsymbol{\theta}}) I^{-1}(\hat{\boldsymbol{\theta}}) H(\hat{\boldsymbol{\theta}})\right\}^{-1} h(\hat{\boldsymbol{\theta}})$.
  -

$$\sqrt{n}\left(h(\hat{\boldsymbol{\theta}}) - \mathbf{0}\right) \overset{\text{d}}{\to} \mathcal{N}_{\dim(h(\boldsymbol{\theta}))}\left(0, H(\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0) H(\boldsymbol{\theta}_0)^T\right).$$

- Notes about Wald Statistics:
  - Wald statistics are not parameterization-invariant (ex. using a $\text{Bin}(n,p)$ will yield different results than a $\text{Bin}(n, 1-p)$).
  - T_W is not invariant to choice of $h(\boldsymbol{\theta})$.
  - T_W standardizes the distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.
  - T_W rejects too often for smaller sample sizes, and has lower power than other test statistics discussed.
  - As n increases, regardless of $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\|$, T_W is more likely to reject H_o.
- **Example**: ZIP model. $Y_i \overset{\text{iid}}{\sim} f(y;\lambda)$, where $P(Y_1 = y) = \begin{cases} p + (1-p) f(0;\lambda) & , y = 0 \\ (1-p) f(y;\lambda) & , y = 1, 2, \ldots \end{cases}$. Obtain T_W for $H_0 : p = 0$ (the Poisson model is adequate).
  Define $\pi := p + (1-p)e^{-\lambda}$.

$$\ell(p,\lambda) = \log\left\{\prod_{i=1}^n \pi^{\mathbb{I}(Y_i=0)} \left[(1-p) f(y;\lambda)\right]^{\mathbb{I}(Y_i \neq 0)}\right\} = \sum_{i=1}^n \left\{\mathbb{I}(Y_i = 0)\log(\pi) + \mathbb{I}(Y_i \neq 0)\log\left[(1-p) f(y;\lambda)\right]\right\};$$

$I_T(p,\lambda) = -E\left[\frac{\partial^2}{\partial(p,\lambda)} \ell(p,\lambda)\right]$ is tedious to calculate, yielding

$$I_T(p,\lambda) = \begin{bmatrix} \frac{n(1-e^{-\lambda})}{\pi(1-p)} & \frac{-ne^{-\lambda}}{\pi} \\ \frac{-ne^{-\lambda}}{\pi} & n\left[\frac{1-p}{\lambda} - \frac{p(1-p)e^{-\lambda}}{\pi}\right] \end{bmatrix}.$$

$$T_W = (\hat{p} - p_0)^2 \hat{I}_T^{11}(\hat{p},\hat{\lambda}) = \hat{p}^2 (\hat{I}_{T,11} - \hat{I}_{T,12}^2 \hat{I}_{T,22}^{-1}).$$

We would yield a different test statistic if we tested for the equivalent hypothesis $H_0 : \pi = e^{-\lambda}$. ■

- **Score Statistic for Scalars**: $T_S = S(\theta_0)^2 I_T(\theta_0)^{-1}$
- **Score Statistic for Vectors**: $T_S = S(\boldsymbol{\theta}_0)^T [I_T(\boldsymbol{\theta}_0)]^{-1} S(\boldsymbol{\theta}_0)$.
- **Score Statistic for Partitioned Hypotheses**: $T_S = S_1(\tilde{\boldsymbol{\theta}})^T [\tilde{I}_{T,11} - \tilde{I}_{T,12}\tilde{I}_{T,22}^{-1}\tilde{I}_{T,21}]^{-1} S_1(\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}}^T = \left(\boldsymbol{\theta}_{10}^T, \arg\max_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \ell(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2)^T\right)$, $\tilde{I}_T = I_T(\tilde{\boldsymbol{\theta}})$, and we are testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs. $H_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$.
- **Score Statistic for Functions**: $T_S = \frac{1}{n} S(\tilde{\boldsymbol{\theta}})^T I^{-1}(\tilde{\boldsymbol{\theta}}) S(\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, h(\boldsymbol{\theta})=0} \ell(\boldsymbol{\theta})$.
- Notes about Score Statistics:
  - Score statistics are parameterization-invariant.
  - Does not use the MLE.
  - T_S standardizes $\ell'(\boldsymbol{\theta})$.
- **Example**: Goodness of fit test. Let $Y \sim$Multinomial$(n; p_1, \ldots, p_K)$ where $\sum_{i=1}^k p_i = 1$. Assume we want to test $H_0 : p_1 = p_{10}, \ldots, p_K = p_{K0}$ vs. $H_a : p_l \neq p_{l0}$ for at least one $l = 1, \ldots, K$. Write down the score test for testing this hypothesis and specify its asymptotic null distribution. Comment on the similarity between this test and the goodness-of-fit test.

  Note that, since $\underline{Y} \sim \text{MultNom}\,(n; p_1, \ldots, p_k)$, $Y_1, \ldots, Y_k \overset{\text{iid}}{\sim} \text{MultNom}\,(1; p_1, \ldots, p_k)$. In addition, only $k-1$ parameters are free, since $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Denote $p_- = \begin{pmatrix} p_1 & \ldots & p_{k-1} \end{pmatrix}' \in (0,1)^{k-1}$.

$$L(p_1, \ldots, p_k) \propto \prod_{i=1}^k p_i^{y_i} \to \ell(p_1, \ldots, p_k) = c + \sum_{i=1}^k y_i \log p_i;$$

$$\ell(p_-) = c + \sum_{i=1}^{k-1} y_i \log p_i + y_k \log\left(1 - \sum_{i=1}^{k-1} p_i\right);\ S(p_-) = \begin{pmatrix} \frac{\partial \ell(p_-)}{\partial p_1} \\ \vdots \\ \frac{\partial \ell(p_-)}{\partial p_{k-1}} \end{pmatrix} = \begin{pmatrix} \frac{y_1}{p_1} - \frac{y_k}{p_k} \\ \vdots \\ \frac{y_{k-1}}{p_{k-1}} - \frac{y_k}{p_k} \end{pmatrix};$$

$$I_T(p_-) = -\,\mathbb{E}\left[\frac{\partial^2}{\partial p_- \partial p_-'} \ell(p_-)\right] = \mathbb{E}\begin{bmatrix} \frac{Y_1}{p_1^2} + \frac{Y_k}{p_k^2} & & \frac{Y_k}{p_k^2} \\ & \ddots & \\ \frac{Y_k}{p_k^2} & & \frac{Y_{k-1}}{p_{k-1}^2} + \frac{Y_k}{p_k^2} \end{bmatrix} = n\begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_k} & & \frac{1}{p_k} \\ & \ddots & \\ \frac{1}{p_k} & & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{bmatrix};$$

$$I_T(p_-)^{-1} = \frac{1}{n}\left[\text{diag}(p_1, \ldots, p_{k-1}) - p_- p_-'\right];$$

$$T_S = S(p_0)' [I_T(p_0)]^{-1} S(p_0) = \begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix}' [I_T(p_0)]^{-1} \begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix}$$

$$= \frac{1}{n}\begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix}' \text{diag}(p_{10}, \ldots, p_{(k-1)0}) \begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix} - \frac{1}{n}\begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix}' p_- p_-' \begin{pmatrix} \frac{y_1}{p_{10}} - \frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}} - \frac{y_k}{p_{k0}} \end{pmatrix}$$

$$
\begin{aligned}
=&\frac{1}{n}\text{diag}\begin{pmatrix} p_{10}\left(\frac{y_1}{p_{10}}-\frac{y_k}{p_{k0}}\right) \\ \vdots \\ p_{(k-1)0}\left(\frac{y_{k-1}}{p_{(k-1)0}}-\frac{y_k}{p_{k0}}\right)\end{pmatrix}'\begin{pmatrix} \frac{y_1}{p_{10}}-\frac{y_k}{p_{k0}} \\ \vdots \\ \frac{y_{k-1}}{p_{(k-1)0}}-\frac{y_k}{p_{k0}}\end{pmatrix}-\left[\frac{1}{n}\sum_{i=1}^{k-1}p_{i0}\left(\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}\right)\right]^2 \\
=&\frac{1}{n}\sum_{i=1}^{k-1}p_{i0}\left(\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}\right)^2-\left[\frac{1}{n}\sum_{i=1}^{k-1}p_{i0}\left(\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}\right)\right]^2=\frac{1}{n}\sum_{i=1}^{k}p_{i0}\left(\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}\right)^2-\left[\frac{1}{n}\sum_{i=1}^{k}p_{i0}\left(\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}\right)\right]^2 \\
=&\frac{1}{n}\sum_{i=1}^{k}\left[\frac{y_i}{p_{i0}}-\frac{y_k}{p_{k0}}-\left(n-\frac{y_k}{p_{k0}}\right)\right]^2=\sum_{i=1}^{k}\frac{(y_i-np_{i0})^2}{np_{i0}}.
\end{aligned}
$$

Under H_0, $T_S \sim \chi^2_{k-1}$, which is the same null distribution as T_{GOF}. This is actually equivalent to the GOF test, where y_i are the observed values, and np_{i0} is the expected value. ■

- **LRT Statistic for Scalars**: $T_{LR} = 2\{\ell(\hat{\theta}_n) - \ell(\theta_0)\}$.
- **LRT Statistic for Vectors**: $T_{LR} = 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\}$.
- **LRT Statistic for Partitioned Hypotheses**: $T_{LR} = 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\tilde{\boldsymbol{\theta}})\}$, where $\tilde{\boldsymbol{\theta}}^T = \left(\boldsymbol{\theta}_{10}^T, \arg\max_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \ell(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2)^T\right)$, and we are testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs. $H_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$.
- **LRT Statistic for Functions**: $T_{LR} = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\}$, where $\tilde{\boldsymbol{\theta}}^T = \left(\boldsymbol{\theta}_{10}^T, \arg\max_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \ell(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2)^T\right)$, and we are testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs. $H_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$.
- Notes about LRT Statistics:
  - LRT statistics are parameterization-invariant.
  - T_{LR} standardizes the distance between $\ell(\hat{\boldsymbol{\theta}})$ and $\ell(\boldsymbol{\theta}_0)$.
- **$100(1-\alpha)\%$ Confidence Region**: $C_{1-\alpha} = \left\{\theta \in \Theta : T_n(\boldsymbol{\theta}_0) < T^*_{\infty,\alpha}\right\}$.
  - $T_n(\boldsymbol{\theta}_0) = T_n$ denotes any test statistic to test $H_o : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, that depends on n.
  - $T_n \xrightarrow{\text{d}} T_\infty$.
  - $T^*_{\infty,\alpha}$ is the critical value of the tail probability α under T_∞.
- Non-standard situations for when the asymptotic distribution of $T_* \neq \chi^2$:
  1. Identifiability assumption is violated (ex. mixture models).
  2. Support depends on $\boldsymbol{\theta}$ (ex. uniform distribution).
  3. Dimensionality of $\boldsymbol{\theta}$ increases with sample size.
     - Consistency of MLE is not guaranteed.
  4. Testing a value near the boundaries of $\boldsymbol{\Theta}$.
     - $\ell(\boldsymbol{\theta})$ might not have a maximizer in the region.
     - T_S might be okay, the other two might not work due to their dependence on the MLE.
  5. Ordered data.
- **Example**: Suppose $Y_{1i} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, 1)$ with sample size n_1 and $Y_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, 1)$ with sample size n_2. For $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 < \mu_2$, find T_{LR} and the testing procedure at $\alpha = 0.05$. Under H_a, the MLEs are the usual ones if $\bar{Y}_1 \leq \bar{Y}_2$, but $\hat{\mu}_1 = \hat{\mu}_2 = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{n_1+n_2}$ if $\bar{Y}_1 > \bar{Y}_2$.

$$
\begin{aligned}
L(\mu_1, \mu_2) &= \prod_{i=1}^{n_1}(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}(Y_{1i}-\mu_1)^2\right\}\prod_{i=1}^{n_2}(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}(Y_{2i}-\mu_2)^2\right\} \\
&\propto \exp\left\{-\sum_{i=1}^{n_1}\frac{1}{2}(Y_{1i}-\mu_1)^2-\sum_{i=1}^{n_2}\frac{1}{2}(Y_{2i}-\mu_2)^2\right\}; \\
\ell(\mu_1, \mu_2) &= -\sum_{i=1}^{n_1}\frac{1}{2}(Y_{1i}-\mu_1)^2-\sum_{i=1}^{n_2}\frac{1}{2}(Y_{2i}-\mu_2)^2.
\end{aligned}
$$

Under H_0, $\mu_1 = \mu_2 = \mu$. We now find the MLE of μ under H_0.

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell(\mu) =& \frac{\partial}{\partial\mu}\left[-\sum_{i=1}^{n_1}\frac{1}{2}(Y_{1i}-\mu)^2 - \sum_{i=1}^{n_2}\frac{1}{2}(Y_{2i}-\mu)^2\right] \\ =& \sum_{i=1}^{n_1}(Y_{1i}-\mu) + \sum_{i=1}^{n_2}(Y_{2i}-\mu) = (n_1+n_2)\mu + \sum_{i=1}^{n_1}Y_{1i} + \sum_{i=1}^{n_2}Y_{2i} \stackrel{\text{set}}{=} 0 \\ \implies \hat{\mu}_N =& \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{n_1+n_2}.\end{aligned}$$

$$\begin{aligned}\ell(\hat{\mu}_N) =& -\sum_{i=1}^{n_1}\frac{1}{2}\left(Y_{1i}-\hat{\mu}_N\right)^2 - \sum_{i=1}^{n_2}\frac{1}{2}\left(Y_{2i}-\hat{\mu}_N\right)^2 \\ =& -\sum_{i=1}^{n_1}\frac{1}{2}\left(Y_{1i}-\bar{Y}_1\right)^2 - \frac{n_1}{2}\left(\bar{Y}_1-\hat{\mu}_N\right)^2 - \sum_{i=1}^{n_2}\frac{1}{2}\left(Y_{2i}-\bar{Y}_2\right)^2 - \frac{n_2}{2}\left(\bar{Y}_2-\hat{\mu}_N\right)^2.\end{aligned}$$

If $\bar{Y}_1 \leq \bar{Y}_2$, then using the same trick,

$$\ell(\hat{\mu}_1,\hat{\mu}_2) = -\sum_{i=1}^{n_1}\frac{1}{2}\left(Y_{1i}-\bar{Y}_1\right)^2 - \frac{n_1}{2}\left(\bar{Y}_1-\hat{\mu}_1\right)^2 - \sum_{i=1}^{n_2}\frac{1}{2}\left(Y_{2i}-\bar{Y}_2\right)^2 - \frac{n_2}{2}\left(\bar{Y}_2-\hat{\mu}_2\right)^2.$$

When $\hat{Y}_1 > \hat{Y}_2$, $\ell(\hat{\mu}_1,\hat{\mu}_2) = \ell(\hat{\mu}_N)$. Under this case, then $T_{LR} = 0$ trivially, but when $\bar{Y}_1 \leq \bar{Y}_2$, then

$$\begin{aligned}T_{LR} =& -2\{\ell(\hat{\mu}_N) - \ell(\hat{\mu}_1,\hat{\mu}_2)\} \\ =& -2\left[-\sum_{i=1}^{n_1}\left(Y_{1i}-\bar{Y}_1\right)^2 - \frac{n_1}{2}\left(\bar{Y}_1-\hat{\mu}_N\right)^2 - \sum_{i=1}^{n_2}\left(Y_{2i}-\bar{Y}_2\right)^2 - \frac{n_2}{2}\left(\bar{Y}_2-\hat{\mu}_N\right)^2\right. \\ & \left. - \left(-\sum_{i=1}^{n_1}\left(Y_{1i}-\bar{Y}_1\right)^2 - \frac{n_1}{2}\left(\bar{Y}_1-\hat{\mu}_1\right)^2 - \sum_{i=1}^{n_2}\left(Y_{2i}-\bar{Y}_2\right)^2 - \frac{n_2}{2}\left(\bar{Y}_2-\hat{\mu}_2\right)^2\right)\right] \\ =& n_1\left(\bar{Y}_1-\hat{\mu}_N\right)^2 + n_2\left(\bar{Y}_2-\hat{\mu}_N\right)^2 - n_1\left(\bar{Y}_1-\hat{\mu}_1\right)^2 - n_2\left(\bar{Y}_2-\hat{\mu}_2\right)^2 \\ =& n_1\left(\bar{Y}_1-\hat{\mu}_N\right)^2 + n_2\left(\bar{Y}_2-\hat{\mu}_N\right)^2 - n_1\left(\bar{Y}_1-\bar{Y}_1\right)^2 - n_2\left(\bar{Y}_2-\bar{Y}_2\right)^2 \\ =& n_1\left(\bar{Y}_1-\hat{\mu}_N\right)^2 + n_2\left(\bar{Y}_2-\hat{\mu}_N\right)^2.\end{aligned}$$

We need the asymptotic distribution of T_{LR} in order to perform the rest of the hypothesis test. Consider the case when $T_{LR} \neq 0$.

$$\begin{aligned}n_1\left(\bar{Y}_1-\hat{\mu}_N\right)^2 + n_2\left(\bar{Y}_2-\hat{\mu}_N\right)^2 =& n_1\left(\bar{Y}_1 - \frac{n_1\bar{Y}_1+n_2\bar{Y}_2}{n_1+n_2}\right)^2 + n_2\left(\bar{Y}_2 - \frac{n_1\bar{Y}_1+n_2\bar{Y}_2}{n_1+n_2}\right)^2 \\ =& \frac{n_1\left(n_1\bar{Y}_1+n_2\bar{Y}_1-n_1\bar{Y}_1-n_2\bar{Y}_2\right)^2 + n_2\left(n_1\bar{Y}_2+n_2\bar{Y}_2-n_1\bar{Y}_1-n_2\bar{Y}_2\right)^2}{(n_1+n_2)^2} \\ =& \frac{n_1\left(n_2\bar{Y}_1-n_2\bar{Y}_2\right)^2 + n_2\left(n_1\bar{Y}_2-n_1\bar{Y}_1\right)^2}{(n_1+n_2)^2} = \frac{n_1n_2^2\left(\bar{Y}_1-\bar{Y}_2\right)^2 + n_1^2n_2\left(\bar{Y}_2-\bar{Y}_1\right)^2}{(n_1+n_2)^2} \\ =& \frac{n_1n_2(n_1+n_2)\left(\bar{Y}_1-\bar{Y}_2\right)^2}{(n_1+n_2)^2} = \frac{\left(\bar{Y}_1-\bar{Y}_2\right)^2}{\frac{1}{n_1}+\frac{1}{n_2}}.\end{aligned}$$

The asymptotic distribution of T_{LR} is a mixture model with two components. The first component is point mass at zero, with probability $p = \frac{1}{2}$ (since, under H_0, $P(\bar{Y}_1 < \bar{Y}_2) = \frac{1}{2}$). By the CLT, $\bar{Y}_1 \sim AN\left(\mu, \frac{1}{n_1}\right)$, and $\bar{Y}_2 \sim AN\left(\mu, \frac{1}{n_2}\right)$ under H_0. When combined with independence, $\bar{Y}_1 - \bar{Y}_2 \sim AN\left(0, \frac{1}{n_1}+\frac{1}{n_2}\right)$. This means that the second component of $T_{LR} \dot{\sim} \chi^2_1$. Regardless, we would reject H_o if $T_{LR} > X_{1-\alpha}$, where X_a is the ath quantile of the mixture density, and α is the significance level. ■

7.4 Misspecified Models and M-Estimation

Return to Table of Contents

- Suppose $Y_i \stackrel{\text{iid}}{\sim} f(y)$, but we don't know f. We also suppose that $\boldsymbol{\theta}$ is some summary of the distribution (mean, variance, etc.). We use a *working model* $g(y;\boldsymbol{\theta})$, where $f \neq g$ necessarily.

- **Estimand**: The true value, $\boldsymbol{\theta}$, that we want to estimate with $\hat{\boldsymbol{\theta}}_n$.
- ***M*-Estimator**: Any $\hat{\boldsymbol{\theta}}_n$ that solves $\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta}) = \mathbf{0}$.
  - $\boldsymbol{\psi}$ is a known or given system of equations that doesn't depend on n, and is a function of y and $\boldsymbol{\theta}$.
  - If $\boldsymbol{\psi}$ has other parameters other than $\hat{\boldsymbol{\theta}}_n$, we need more equations to estimate them.
  - **Partial *M*-Estimator**: $\hat{\boldsymbol{\theta}}_n$ needs additional equations in $\boldsymbol{\psi}$ to become an M-estimator.
  - If $\hat{\boldsymbol{\theta}}_n$ is an M-estimator with $\boldsymbol{\psi}$, then $\boldsymbol{\psi}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta}) = 0$.
  - $\boldsymbol{\psi}_n(\boldsymbol{\theta}) \xrightarrow{\text{p}} \underline{\boldsymbol{\psi}}(\boldsymbol{\theta}) := \mathbb{E}_{Y_i}[\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})]$.
  - $\boldsymbol{\psi}$ functions may not be the same for the same estimators, but the results discussed later should yield equivalent results.
  - If $\boldsymbol{\psi}$ is not smooth, switch the derivative and expectation for A, so $A(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}}\left[\mathbb{E}\left(\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})\right)\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.
- **Weak Consistency Theorem**: Suppose Y_i is iid, and assume:

  C1. Uniform LLN: $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{\psi}_n(\theta) - \boldsymbol{\psi}(\boldsymbol{\theta})\| \xrightarrow{\text{p}} \mathbf{0}$.

  C2. Unique minimum: If $\boldsymbol{\theta}_0$ solves $\underline{\boldsymbol{\psi}}(\boldsymbol{\theta}) = \mathbf{0}$, then $\forall \epsilon > 0$, $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{\left\|\underline{\boldsymbol{\psi}}(\boldsymbol{\theta}); \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon\right\|\right\} > 0$.

  Then, if $\hat{\boldsymbol{\theta}}_n$ solves $\underline{\boldsymbol{\psi}}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, then $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}$.
  - If we only have independent samples, then $\underline{\boldsymbol{\psi}}(\boldsymbol{\theta}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y_i}[\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})]$.
- **Asymptotic Distribution of *M*-Estimator**: Suppose Y_i is iid, with $\hat{\boldsymbol{\theta}}_n$ is an M-estimator with associated $\boldsymbol{\psi}$ function. Also assume regularity assumptions for Y_i and $\boldsymbol{\psi}$, and $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_0$, where $\underline{\boldsymbol{\psi}}(\boldsymbol{\theta}) = 0$. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\text{d}} \mathcal{N}_b\left(0, V(\boldsymbol{\theta}_0)\right), \text{ where } V(\boldsymbol{\theta}_0) = A(\boldsymbol{\theta}_0)^{-1}B(\boldsymbol{\theta}_0)A^T(\boldsymbol{\theta}_0)^{-1},$$
$$A(\boldsymbol{\theta}_0) = -\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_0)\right], \text{ and } B(\boldsymbol{\theta}_0) = \mathbb{E}\left[\boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_0)^T\right].$$

  - In practice, use $A_n(\underline{Y};\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}'(Y_i;\hat{\boldsymbol{\theta}}_n)$ and $B_n(\underline{Y};\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\hat{\boldsymbol{\theta}}_n)\boldsymbol{\psi}(Y_i;\hat{\boldsymbol{\theta}}_n)^T$.
  - If model is correctly specified, $V(\boldsymbol{\theta}_0) = I^{-1}(\boldsymbol{\theta}_0)$.
- **Example**: Suppose $Y_i \overset{\perp}{=} \exp(X_i\beta) + e_i$, where $e_i \overset{\text{iid}}{\sim} (0,\sigma^2) \perp X_i$, and $X_i, \beta \in \mathbb{R}$. We estimate β with $\arg\min_\beta \sum_{i=1}^n [Y_i - \exp(X_i\beta)]^2$, and σ^2 with $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left[Y_i - \exp(X_i\hat{\beta})\right]^2$. Construct $\boldsymbol{\psi}$ for $\hat{\sigma}^2$.
  $\hat{\sigma}^2$ will be a partial M-estimator, since there is a β term that also must be estimated. Then, determine the asymptotic distribution of $\hat{\sigma}^2$, and derive an estimator of its asymptotic variance.

$$\frac{\partial}{\partial\beta}\sum_{i=1}^n [Y_i - \exp(X_i\beta)]^2 \propto \sum_{i=1}^n [Y_i - \exp(X_i\beta)]\exp(X_i\beta) \overset{\text{set}}{=} 0.$$

  We also note that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left[Y_i - \exp(X_i\hat{\beta})\right]^2 \equiv \frac{1}{n}\sum_{i=1}^n \left\{\left[Y_i - \exp(X_i\hat{\beta})\right]^2 - \hat{\sigma}^2\right\} = 0$. Combining these results, we get that $\boldsymbol{\psi}(\beta,\sigma^2)^T = \left([Y_i - \exp(X_i\beta)]\exp(X_i\beta), \left[Y_i - \exp(X_i\hat{\beta})\right]^2 - \sigma^2\right)$.
  By M-estimation theory, $\hat{\sigma}^2 \sim AN\left(\sigma^2, \left[A(\beta,\sigma^2)^{-1}B(\beta,\sigma^2)A^T(\beta,\sigma^2)^{-1}\right]\Big|_{22}\right)$.
  Use $\hat{A}_n = -\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial(\beta,\sigma^2)}\boldsymbol{\psi}(Y_i, X_i, \hat{\beta}, \hat{\sigma}^2)$, and $\hat{B}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i, X_i, \hat{\beta}, \hat{\sigma}^2)\boldsymbol{\psi}(Y_i, X_i, \hat{\beta}, \hat{\sigma}^2)^T$. ■
- **Example**: Let $Y_1, \ldots, Y_n$ be IID from a distribution with finite fourth moment. Use the framework of M-estimation to study theoretical properties (consistency and asymptotic behavior) of the coefficient of variation, $\frac{s_n}{\bar{Y}_n}$, where $s_n^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y}_n)^2}{n-1}$.
  The CV, θ, will be a partial M-estimator. Define $\boldsymbol{\psi}^T(y, \mu_Y, \sigma, \theta) = \left(y - \mu_Y, (y-\mu_Y)^2 - \sigma^2, \sigma - \mu_Y\theta\right)$, where the third element forms the ratio of question, $\theta \overset{\text{set}}{=} \frac{\sigma}{\mu_Y}$. We now derive the second equation.

$$\sigma^2 \overset{\text{set}}{=} \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_Y)^2 \longrightarrow \frac{1}{n}\sum_{i=1}^n \left[(Y_i - \mu_Y)^2 - \sigma^2\right] = 0.$$

  We therefore have an M-estimator with additional arguments added to $\boldsymbol{\psi}$. Using the properties of M-estimators, under the regularity assumptions, $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$ ($\hat{\theta}_n$ being the MLE of θ), and

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, V_{33}(\theta_0)\right), \text{ where } V_{33}(\theta_0) = \left[A(\theta_0)^{-1}B(\theta_0)^{-1}[A(\theta_0)^{-1}]^T\right]_{33}.$$

$$A(\boldsymbol{\theta}_0) = -\mathbb{E}\left[\boldsymbol{\psi}'(X_i, Y_i; \boldsymbol{\theta}_0)\right] = \mathbb{E}\begin{bmatrix} 1 & 0 & 0 \\ -2(Y_i - \mu_{Y0}) & 2\sigma_0 & 0 \\ \theta_0 & -1 & \sigma_0^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2\sigma_0 & 0 \\ \theta_0 & -1 & \sigma_0^2 \end{bmatrix}; A(\boldsymbol{\theta}_0)^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2\sigma_0} & 0 \\ -\frac{\sigma_0}{\mu_{Y0}^2} & \frac{1}{2\mu_{Y0}\sigma_0} & \frac{1}{\mu_{Y0}} \end{bmatrix};$$

Define μ_i as the ith central moment of Y.

$$\begin{aligned} B(\boldsymbol{\theta}_0) =& \mathbb{E}\left[\boldsymbol{\psi}(X_i, Y_i; \boldsymbol{\theta}_0)\boldsymbol{\psi}^T(X_i, Y_i; \boldsymbol{\theta}_0)\right] \\ =& \mathbb{E}\begin{bmatrix} (Y_i - \mu_{Y0})^2 & (y - \mu_{Y0})^3 - (Y_i - \mu_{Y0})\sigma_0 & (Y_i - \mu_{Y0})(\sigma_0 - \mu_{Y0}\theta_0) \\ (y - \mu_{Y0})^3 - (Y_i - \mu_{Y0})\sigma_0 & \left[(Y_i - \mu_{Y0})^2 - \sigma_0^2\right]^2 & \left[(Y_i - \mu_{Y0})^2 - \sigma_0^2\right](\sigma_0 - \mu_{Y0}\theta_0) \\ (Y_i - \mu_{Y0})(\sigma_0 - \mu_{Y0}\theta_0) & \left[(Y_i - \mu_{Y0})^2 - \sigma_0^2\right](\sigma_0 - \mu_{Y0}\theta_0) & (\sigma_0 - \mu_{Y0}\theta_0)^2 \end{bmatrix} \\ =& \begin{bmatrix} \sigma_0^2 & \mu_3 - (0)\sigma_0 & (0)(\sigma_0 - \mu_{Y0}\theta_0) \\ \mu_3 - (0)\sigma_0 & \mu_4 - 2(\sigma_0^4 - \sigma_0^4) + \sigma_0^4 & (0)(\sigma_0 - \mu_{Y0}\theta_0) \\ (0)(\sigma_0 - \mu_{Y0}\theta_0) & (0)(\sigma_0 - \mu_{Y0}\theta_0) & (\sigma_0 - \sigma_0)^2 \end{bmatrix} = \begin{bmatrix} \sigma_0^2 & \mu_3 & 0 \\ \mu_3 & \mu_4 - \sigma_0^4 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \end{aligned}$$

$$\begin{aligned} V(\boldsymbol{\theta}_0) =& A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0)[A(\boldsymbol{\theta}_0)^{-1}]^T \\ =& \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2\sigma_0} & 0 \\ -\frac{\sigma_0}{\mu_{Y0}^2} & \frac{1}{2\mu_{Y0}\sigma_0} & \frac{1}{\mu_{Y0}} \end{bmatrix} \begin{bmatrix} \sigma_0^2 & \mu_3 & 0 \\ \mu_3 & \mu_4 - \sigma_0^4 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -\frac{\sigma_0}{\mu_{Y0}^2} \\ 0 & \frac{1}{2\sigma_0} & \frac{1}{2\mu_{Y0}\sigma_0} \\ 0 & 0 & \frac{1}{\mu_{Y0}} \end{bmatrix} \\ =& \begin{bmatrix} \sigma_0^2 & \mu_3 & 0 \\ \frac{\mu_3}{2\sigma_0} & \frac{\mu_4 - \sigma_0^4}{2\sigma_0} & 0 \\ \frac{\mu_3}{2\mu_{Y0}\sigma_0} - \frac{\mu_3\sigma_0}{\mu_{Y0}^2} & \frac{\mu_4 - \sigma_0^4}{2\mu_{Y0}\sigma_0} - \frac{\mu_3}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -\frac{\sigma_0}{\mu_{Y0}^2} \\ 0 & \frac{1}{2\sigma_0} & \frac{1}{2\mu_{Y0}\sigma_0} \\ 0 & 0 & \frac{1}{\mu_{Y0}} \end{bmatrix}; \\ V_{33}(\boldsymbol{\theta}_0) =& \frac{\mu_4 - \sigma_0^4}{2\mu_{Y0}^2\sigma_0^2} + \frac{\mu_3\sigma_0^2}{\mu_{Y0}^4} - \frac{\mu_3}{2\mu_{Y0}^3} - \frac{\mu_3}{4\mu_{Y0}^3\sigma_0}. \blacksquare \end{aligned}$$

- Under H_0, assuming regularity assumptions for $\boldsymbol{\psi}$ and g, and $I(\boldsymbol{\theta})$ is continuous,

$$T_W, T_S, \text{ and } T_{LR} \xrightarrow{\text{d}} \sum_{\ell=1}^{r} \lambda_\ell Z_\ell^2,$$

  where λ_i is the ith eigenvalue of $[I^{11}(\boldsymbol{\theta}_g)]^{-1}V_{11}(\boldsymbol{\theta}_g)$, $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and $\boldsymbol{\theta}_g$ solves $\mathbb{E}_Y[\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})] = \mathbf{0}$.

  - Under H_0, the first components of $\boldsymbol{\theta}_g = \boldsymbol{\theta}_{10}$, so $\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \xrightarrow{\text{d}} \mathcal{N}_r\left(\mathbf{0}, V_{11}(\boldsymbol{\theta}_g)\right)$.

- **Generalized Wald Statistic**: $T_{GW} = n(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^T[V_{11}(\hat{\boldsymbol{\theta}})]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$, where $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}_g$ is an M-estimator, and $\boldsymbol{\theta}_g$ solves $\mathbb{E}[\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})] = \mathbf{0}$.

  - Under assumptions for $\boldsymbol{\psi}$ and g, $V_{11}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{p}} (V_g)_{11}$, and $(V_g)_{11}^{-1}$ exits, then under H_0, $T_{GW} \xrightarrow{\text{d}} \chi_r^2$.
  - If the model is specified correctly, then $T_{GW} \to T_W$.
  - We can replace $V(\hat{\boldsymbol{\theta}})_{11}^{-1}$ with a consistent estimator of $(V_g)_{11}^{-1}$ (assumptions still needed).
  - For functions, $T_{GW} = nh(\hat{\boldsymbol{\theta}})^T\left\{H(\hat{\boldsymbol{\theta}})\hat{V}H(\hat{\boldsymbol{\theta}})^T\right\}^{-1}h(\hat{\boldsymbol{\theta}})$.

- **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} f(y;\boldsymbol{\theta}_0)$, where f satisfies the regularity conditions. Consider testing H_0 : $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ vs. H_a : $\boldsymbol{h}(\boldsymbol{\theta}) \neq \mathbf{0}$, where $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{h}'(\boldsymbol{\theta})$ is continuous and not identically $\mathbf{0}$. Also suppose $I(\boldsymbol{\theta})$ is continuous. Prove that $T_W = n\boldsymbol{h}(\hat{\boldsymbol{\theta}})^T\{\boldsymbol{H}(\hat{\boldsymbol{\theta}})I(\hat{\boldsymbol{\theta}})^{-1}H(\hat{\boldsymbol{\theta}})^T\}^{-1}\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{d}} \chi_r^2$ under H_0. Then, assume

$Y_i \sim g(y)$ actually, but $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}_0$ and $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ still. Which asymptotic results still hold?
We know that $\hat{\boldsymbol{\theta}} \sim AN(\boldsymbol{\theta}_0, \frac{1}{n} I(\boldsymbol{\theta}_0)^{-1})$. Since $\boldsymbol{H}(\boldsymbol{\theta})$ is not identically zero, by the Delta method,

$$\sqrt{n}\{\boldsymbol{H}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)\}^{-1/2} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}, I_r).$$

By continuity, $\boldsymbol{H}(\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\theta}})^T \xrightarrow{\text{p}} \boldsymbol{H}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)^T$, so by Slutsky's theorem,

$$\sqrt{n}\{\boldsymbol{H}(\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\theta}})^T\}^{-1/2} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}, I_r).$$

Therefore, under H_0, $T_W = n\boldsymbol{h}(\hat{\boldsymbol{\theta}})^T \{\boldsymbol{H}(\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\theta}})^T\}^{-1} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{d}} \chi^2_r$.
If the model is misspecified, $Var(\hat{\boldsymbol{\theta}})$ changes. We are now in the M-estimation framework, where $\boldsymbol{\psi} = f'(y; \boldsymbol{\theta})$, so $Var(\hat{\boldsymbol{\theta}}) = \frac{1}{n} A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) A^T(\boldsymbol{\theta}_0)^{-1}$. T_W also no longer converges to χ^2_R. Under H_0, $\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}, \boldsymbol{H}(\boldsymbol{\theta}_0) A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) A^T(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)^T)$. Introduce $Z \sim \mathcal{N}(0, \boldsymbol{H}(\boldsymbol{\theta}_0) A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) A^T(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)^T)$, so by continuity,

$$\boldsymbol{H}(\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\theta}})^T \xrightarrow{\text{p}} \boldsymbol{H}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)^T.$$

Applying Slutsky's theorem,

$$T_W \xrightarrow{\text{d}} Z^T \{\boldsymbol{H}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0)^T\}^{-1} Z. \ \blacksquare$$

- **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} g(y)$, and we are interested in estimating $\eta_{0.25}$ and $\eta_{0.75}$. Find a bivariate $\boldsymbol{\psi}$ such that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \eta_{0.25}, \eta_{0.75}) = \boldsymbol{c}_n$, where $|c_{n1}|$ and $|c_{n2}| \leq \frac{1}{n}$. Then, specify the asymptotic distribution of $(\hat{\eta}_{0.25}, \hat{\eta}_{0.75})^T$, and suggest an empirical-based estimator for the asymptotic covariance. Derive T_{GW} for testing $H_0 : \eta_{0.25} = \eta_{0.75}$. Calculate the asymptotic covariance of $(\hat{\eta}_{0.25}, \hat{\eta}_{0.75})^T$ when $g(y) = \lambda e^{-\lambda y}$ for $y > 0$. Lastly, for $a_1, a_2 > 0$ such that $a_1 + a_2 = 1$, show that $a_1 \frac{\log(4/3)}{\hat{\eta}_{0.25}} + a_2 \frac{\log(4)}{\hat{\eta}_{0.75}} \sim AN(\lambda, \frac{1}{n} V(\lambda))$ for some function $V(\lambda)$, and find a_1, a_2 that minimize $V(\lambda)$.

Define $\boldsymbol{\psi}(Y_i; \eta_{0.25}, \eta_{0.75}) = \begin{pmatrix} 0.25 - \mathbb{I}(Y_i \leq \eta_{0.25}) \\ 0.75 - \mathbb{I}(Y_i \leq \eta_{0.75}) \end{pmatrix}$. Thus, $(\hat{\eta}_{0.25}, \hat{\eta}_{0.75})^T$ solves $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i; \eta_{0.25}, \eta_{0.75}) = \boldsymbol{c}_n$, where $c_{n1} = \frac{\lceil (0.25)n \rceil - 0.25n}{n}$, similarly for c_{n2}. Using the asymptotic distribution for M-estimators,

$$\begin{pmatrix} \hat{\eta}_{0.25} \\ \hat{\eta}_{0.75} \end{pmatrix} \sim AN\left(\begin{pmatrix} \eta_{0.25} \\ \eta_{0.75} \end{pmatrix}, \frac{1}{n} A(\eta_{0.25}, \eta_{0.75})^{-1} B(\eta_{0.25}, \eta_{0.75}) A^T(\eta_{0.25}, \eta_{0.75})^{-1} \right).$$

$$A(\eta_{0.25}, \eta_{0.75}) = -\mathbb{E}\left[\frac{\partial \boldsymbol{\psi}(Y_i, \eta_{0.25}, \eta_{0.75})}{d(\eta_{0.25}, \eta_{0.75})^T} \right] = -\frac{\partial}{\partial \theta_0} \begin{pmatrix} 0.25 - \mathbb{I}(Y_i \leq \hat{\eta}_{0.25}) \\ 0.75 - \mathbb{I}(Y_i \leq \hat{\eta}_{0.75}) \end{pmatrix} = \begin{pmatrix} g(\eta_{0.25}) & 0 \\ 0 & g(\eta_{0.75}) \end{pmatrix}.$$

$$B(\eta_{0.25}, \eta_{0.75}) = \mathbb{E}[\boldsymbol{\psi}\boldsymbol{\psi}^T] = \begin{pmatrix} \frac{3}{16} & \frac{1}{16} \\ \frac{1}{16} & \frac{3}{16} \end{pmatrix}.$$

To estimate the asymptotic covariance, use $\hat{\eta}_{0.25}$ and $\hat{\eta}_{0.75}$ in lieu of $\eta_{0.25}$ and $\eta_{0.75}$, respectively.
For simplicity, let $h := h(\hat{\eta}_{0.25}, \hat{\eta}_{0.75}) = \eta_{0.25} - \eta_{0.75}$. Same for H, A, and B. $T_{GW} = \frac{nh^2}{HA^{-1}B\{A^{-1}\}^T H^T}$.
Using the fact that $H = (1, -1)$, we get that $T_{GW} = \frac{n(\hat{\eta}_{0.25} - \hat{\eta}_{0.75})^2}{\frac{3/16}{\hat{g}^2(\hat{\eta}_{0.25})} - \frac{2/16}{\hat{g}(\hat{\eta}_{0.25})\hat{g}(\hat{\eta}_{0.75})} + \frac{3/16}{\hat{g}^2(\hat{\eta}_{0.75})}} \xrightarrow{\text{d}} \chi^2_1$.
If $g(y) = \lambda e^{-\lambda y}$, then using CDFs, $1 - e^{-\lambda \eta_p} = p \Rightarrow \eta_p = -\frac{1}{\lambda}(1 - p)$. Hence, $g(\eta_p) = \lambda(1 - p)$. This yields an asymptotic covariance of $V(\eta) = \begin{pmatrix} (3\lambda^2)^{-1} & (3\lambda^2)^{-1} \\ (3\lambda^2)^{-1} & 3\lambda^{-2} \end{pmatrix}$.
Consider $h(x, y) = a_1 \frac{\log(4/3}{x} + a_2 \frac{\log(4)}{y} \Rightarrow h'(x, y) = -\lambda^2 \left(\frac{a_1}{\log(4/3)}, \frac{a_2}{\log(4)} \right)^T \neq 0$. By the Delta method, $h(x, y) \xrightarrow{\text{d}} \mathcal{N}(\lambda, \frac{1}{n} V(\lambda))$, where

$$V(\lambda) = h'(\eta_{0.25}, \eta_{0.75})^T \begin{pmatrix} (3\lambda^2)^{-1} & (3\lambda^2)^{-1} \\ (3\lambda^2)^{-1} & 3\lambda^{-2} \end{pmatrix} h'(\eta_{0.25}, \eta_{0.75}) = \frac{\lambda^2}{3} \left\{ \left(\frac{1 - a_2}{\log(4/3)} + \frac{a_2}{\log(4)} \right)^2 + 8 \left(\frac{a_2}{\log(4)} \right)^2 \right\}.$$

Using the fact that $a_1 = 1 - a_2$, we take the derivative of $V(\lambda)$ wrt a_2, and solve accordingly. ■

- **Generalized Score Statistic**: $T_{GS} = \frac{1}{n}\left[\sum_{i=1}^n \boldsymbol{\psi}_1(Y_i;\tilde{\boldsymbol{\theta}})\right]^T V_{\boldsymbol{\psi}_1}^{-1}(\tilde{\boldsymbol{\theta}})\left[\sum_{i=1}^n \boldsymbol{\psi}_1(Y_i;\tilde{\boldsymbol{\theta}})\right]$, where $\hat{\boldsymbol{\theta}}$ is an M-estimator, and $\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\hat{\boldsymbol{\theta}}) = \mathbf{0}$. This uses $\boldsymbol{\psi}^T =: (\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T)$, and define $\tilde{\boldsymbol{\theta}}$ by $\sum_{i=1}^n \boldsymbol{\psi}_2(Y_i;\tilde{\boldsymbol{\theta}}) = \mathbf{0}$, where $\tilde{\boldsymbol{\theta}}^T = (\boldsymbol{\theta}_{10}^T, \tilde{\boldsymbol{\theta}}_2^T)$. Lastly, $V_{\boldsymbol{\psi}_1}(\tilde{\boldsymbol{\theta}})$ is the asymptotic covariance of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\psi}_1(Y_i;\hat{\boldsymbol{\theta}})$.
  - $V_{\boldsymbol{\psi}_1} = \boldsymbol{B}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{21} - \boldsymbol{B}_{12}\{\boldsymbol{A}_{22}^{-1}\}^T\boldsymbol{A}_{12}^T + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{22}\{\boldsymbol{A}_{22}^{-1}\}^T\boldsymbol{A}_{12}^T$.
    - In one-dimensional cases, or when we have a completely specified hypothesis $H_0 : \boldsymbol{\theta}_g = \boldsymbol{\theta}_0$, $\boldsymbol{A}_{12} = 0$, so this simplifies to $V_{\boldsymbol{\psi}} = \boldsymbol{B}_{11} = \boldsymbol{B}$.
  - Using ABAR, we can show that $\tilde{\boldsymbol{\theta}}_g = \boldsymbol{\theta}_{2g} + \frac{1}{n}\sum_{i=1}^n A_{22}^{-1}(\boldsymbol{\theta}_g)\boldsymbol{\psi}_2(Y_i;\boldsymbol{\theta}_g) + o_p(n^{-1/2})$.
  - Assume regularity assumptions for $\boldsymbol{\psi}$ and g, and that $V_{\boldsymbol{\psi}_1}(\tilde{\boldsymbol{\theta}}) \xrightarrow{\text{p}} V_{\boldsymbol{\psi}_1}(\boldsymbol{\theta}_g)$, where $V_{\boldsymbol{\psi}_1}(\boldsymbol{\theta}_g)$ is invertible. Then, $T_{GS} \xrightarrow{\text{d}} \chi_r^2$ under H_0.
  - If we only have independent data, then $V_{\boldsymbol{\psi}_1}(\boldsymbol{\theta}_g) = \lim_{n\to\infty} V_{\boldsymbol{\psi}_1}^n(\boldsymbol{\theta}_g)$, and $A(\boldsymbol{\theta}_g) = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_g)\right]$, similarly for $B(\boldsymbol{\theta}_g)$.
  - T_{GS} is invariant to reparameterization when we use empirical estimates.

- **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, and we are interested in testing $H_0 : \sigma^2 = \sigma_0^2$. Obtain T_{GS}.

  Use M-estimation. Define $\boldsymbol{\psi}(Y_i;\sigma^2,\mu) = \begin{pmatrix} (Y_i-\mu)^2 - \sigma^2 \\ Y_i - \mu \end{pmatrix}$. Define $\boldsymbol{\theta}^T = (\sigma^2, \mu)$. $V_{\boldsymbol{\psi}_1}(\tilde{\boldsymbol{\theta}}) = A(\tilde{\boldsymbol{\theta}})^{-1}B(\tilde{\boldsymbol{\theta}})A(\tilde{\boldsymbol{\theta}})^{-1}$.

$$A(\tilde{\boldsymbol{\theta}}) = -E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(Y_i;\tilde{\boldsymbol{\theta}})\right] = \begin{bmatrix} 1 & 2\mathbb{E}(Y_i - \tilde{\mu}) \\ 0 & 1 \end{bmatrix} = I_2.$$

$$B(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \mathbb{E}[(Y_i-\tilde{\mu})^4] - 2\sigma_0^2\mathbb{E}[(Y_i-\tilde{\mu})^2] + \sigma_0^4 & (\cdot) \\ \mathbb{E}[(Y_i-\tilde{\mu})^3] - \sigma_0^2\mathbb{E}[Y_i - \tilde{\mu}] & \mathbb{E}[(Y_i-\tilde{\mu})^2] \end{bmatrix} = \begin{bmatrix} \mu_4 - 2\sigma_0^2 Var(Y) + \sigma_0^2 & \mu_3 \\ \mu_3 & Var(Y) \end{bmatrix}.$$

$$V_{\boldsymbol{\psi}_1}(\tilde{\boldsymbol{\theta}}) = (1,0)B(\tilde{\boldsymbol{\theta}})(1,0)^T = \mu_4 - 2\sigma_0^2 Var(Y) + \sigma_0^2.$$

$$\hat{V}_{\boldsymbol{\psi}_1}(\tilde{\boldsymbol{\theta}}) = \mu_4 - 2\sigma_0^2 s_n^2 + \sigma_0^2.$$

$$T_{GS} = \frac{1}{n}\cdot\frac{[\sum(Y_i - \bar{Y})]^2 - \sigma_0^2}{\mu_4 - 2\sigma_0^2 s_n^2 + \sigma_0^2} \xrightarrow{\text{d}} \chi_1^2.$$

  Use MOM estimates for all remaining quantities. ■

- **Example**: Suppose we have data $X_1, \ldots, X_n$ that are iid. The sign test for H_0 : median$= 0$ is to count the number of X's above 0, say Y, and compare Y to a $\text{Bin}\left(n, \frac{1}{2}\right)$ distribution. Derive $\boldsymbol{\psi}$, T_{GW} and T_{GS}, and comment on which statistic should be used.

  The sample median is equal to $\hat{\eta}_{0.5}$. We know that this satisfies $\sum_{i=1}^n\left[\frac{1}{2} - \mathbb{I}(X_i \le \hat{\eta}_{0.5})\right] = c_n$, so we get that $\boldsymbol{\psi}(X_i;\eta_{0.5}) = \frac{1}{2} - \mathbb{I}(X_i \le \eta_{0.5})$. Simple calculations yield $B(\eta_0) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$, and $V(\eta_0) = \frac{1}{4f^2(\eta_0)}$. Since we are in the one-dimensional case, $V_{\boldsymbol{\psi}_1}(\tilde{\eta}_0) = B(\eta_0) = \frac{1}{4}$.

$$T_{GW} = n(\hat{\eta}_{0.5} - \eta_0)^T[V_{11}(\hat{\eta}_{0.5})]^{-1}(\hat{\eta}_{0.5} - \eta_0) = n\frac{(\hat{\eta}_{0.5} - 0)^2}{1/[4f^2(\hat{\eta}_{0.5})]} = 4n\hat{\eta}_{0.5}f^2(\hat{\eta}_{0.5}).$$

$$T_{GS} = \frac{1}{n}\left[\sum_{i=1}^n \boldsymbol{\psi}_1(X_i;\tilde{\boldsymbol{\theta}})\right]^T V_{\boldsymbol{\psi}_1}^{-1}(\tilde{\boldsymbol{\theta}})\left[\sum_{i=1}^n \boldsymbol{\psi}_1(X_i;\tilde{\boldsymbol{\theta}})\right] = \frac{1}{n}\left[\sum_{i=1}^n \boldsymbol{\psi}(X_i;\tilde{\boldsymbol{\theta}})\right]^2\left(\frac{1}{1/4}\right) = \frac{4}{n}\left[\frac{n}{2} - \sum_{i=1}^n \mathbb{I}(X_i \le 0)\right]^2.$$

  T_{GS} is preferred over T_{GW}, since we need to know (or estimate) f. If X_i is discontinuous, then estimating f may be challenging. ■

- **Quadratic Inference Function**, or **QIF**:

$$T_{QIF} = Q^2(\boldsymbol{\theta}_0) - Q(\hat{\boldsymbol{\theta}}) = Q^2(\boldsymbol{\theta}_0) = \frac{1}{n}\left[\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_0)^T\right]\{B_n(Y;\hat{\boldsymbol{\theta}})\}^{-1}\left[\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta}_0)\right],$$

  where $Q^2(\boldsymbol{\theta}) = \bar{\boldsymbol{\psi}}(Y;\boldsymbol{\theta})^T\hat{C}_{\boldsymbol{\theta}}\bar{\boldsymbol{\psi}}(Y;\boldsymbol{\theta})$, $\bar{\boldsymbol{\psi}}(Y;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta})$, $\hat{C}_{\boldsymbol{\theta}}$ is the estimator of $Cov\{\bar{\boldsymbol{\psi}}(Y;\boldsymbol{\theta})\}$, and $\hat{\boldsymbol{\theta}}$ is a minimizer of $Q(\boldsymbol{\theta})$ and M-estimator of $\boldsymbol{\theta}$.

- Is the generalized LRT statistic.
- $Q(\boldsymbol{\theta}) \geq 0$, so $\hat{\boldsymbol{\theta}}$ minimizes $Q^2(\boldsymbol{\theta})$.
- Under H_0, $T_{QIF} \xrightarrow{\text{d}} \chi^2_b$.
- In practice, $\hat{C}_{\boldsymbol{\theta}}$ is MOM estimator of $Cov\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta})\right) = \frac{1}{n}B_n(\underline{Y};\boldsymbol{\theta}) = \frac{1}{n^2}\sum_{i=1}^n \boldsymbol{\psi}(Y_i;\boldsymbol{\theta})\boldsymbol{\psi}(Y_i;\boldsymbol{\theta})^T$.
  - If we use $B_n(Y;\boldsymbol{\theta}_0)$, then $T_{QIF} = T_{GS}$.
- When testing all of the parameters, $T_{QIF} \equiv T_{GS}$.
- For partitioned hypotheses, use $T_{QIF} = \min_{\boldsymbol{\theta}_2} Q^2(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2)$.

- **Example**: Suppose $Y_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. We are interested in testing H_0: $\mu = \mu_0$.
$\bar{Y}$ is an M-estimator, with corresponding $\boldsymbol{\psi}(Y_i;\mu) = Y_i - \mu$.

$$T_{QIF} = \frac{\bar{\boldsymbol{\psi}}(Y;\mu_0)^2}{\hat{C}_\mu} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_0)}{\frac{1}{n}B_n(Y;\mu_0)} = \frac{\bar{Y} - \mu_0}{n^{-2}\sum_{i=1}^n (Y_i - \mu_0)^2} = \frac{n(\bar{Y} - \mu_0)}{s_n^2 + (\bar{Y} - \mu_0)^2} \equiv T_{GS}. \blacksquare$$

7.5 Monte Carlo

Return to Table of Contents

- **Monte Carlo Methods**: Use random simulation to obtain empirical estimates of quantities of interest.
- Monte Carlo also lets us investigate effectiveness of statistic performance in finite sample sizes.
- **Example**: Refer to the table below.

Distribution	$\widehat{\text{Bias}}(\hat{\theta})$	$\widehat{\text{Var}}(\hat{\theta})$	$\hat{\mathbb{E}}\{\widehat{\text{AVar}}(\hat{\theta})\}$
Normal	0.02	1.47	1.36
LaPlace	0.05	1.37	1.25
t_5	0.03	1.28	1.17

These estimates were calculated from $N = 1000$ Monte Carlo samples. Envision 1000 independent rows of data, where in each row we have the basic estimator $\hat{\theta}$ and an estimate of its asymptotic variance, $\widehat{\text{AVar}}(\hat{\theta})$ calculated from *one Monte Carlo sample*. $\widehat{\text{Bias}}(\hat{\theta})$ is the average of the $\hat{\theta}$ values minus the true parameter value, θ. $\widehat{\text{Var}}(\hat{\theta})$ is the sample variance based on the values of $\hat{\theta}$. The last column is the average of the $\widehat{\text{AVar}}(\hat{\theta})$ values. Give an expression for the MC SE of each estimate. Usually, we want to know if the entries in the 3rd column are close to the entries in the 4th column. However, that is made difficult because the entries in those columns are correlated. So, also suggest a simple way to combine those estimates to get another column that is easy to use for that purpose.
$\widehat{\text{Bias}}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^N (\hat{\theta}_i - \theta)$, $\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^N (\hat{\theta}_i - \bar{\hat{\theta}})^2$, and $\hat{\mathbb{E}}[\widehat{\text{Var}}(\hat{\theta})] = \frac{1}{N}\sum_{i=1}^N \widehat{\text{Var}}_i(\hat{\theta})$. Taking the square root of these expressions yields the SEs, plugging in the estimates provided in the table.
The new column would be a T_W that tests for whether or not the variance terms are equal. We use M-estimators, with a corresponding $\boldsymbol{\psi} = \begin{pmatrix} \hat{\theta}_i - \theta_1 \\ (\hat{\theta}_i - \theta_1)^2 - \theta_2 \\ \widehat{\text{Var}}_i(\theta) - \theta_3 \end{pmatrix}$, in order to get the variance estimates. $\blacksquare$

- Always consider which factors you want to study in the simulation study (which includes sample size).
- If possible, save every estimate at every iteration.
- When coding, start with a low number of simulations to make sure the code functions correctly.
- Track seed number, and organize code.
- This whole section is basically just variance calculations.

- **Example**: Bias estimation. The estimated bias is $\widehat{\text{Bias}}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i - \theta_0$. Suppose we have an estimated variance of this estimator. Find the minimum MC sample size N such that the precision of the bias is within two decimal places, and then for arbitrary d.

$$Var\left[\text{Bias}(\hat{\theta})\right] = Var\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i - \theta_0\right] = Var\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i\right] = \frac{1}{N}Var(\hat{\theta}).$$

  So, $SD\left[\text{Bias}(\hat{\theta})\right] = \frac{1}{\sqrt{N}}SD(\hat{\theta})$. We want the SD of the bias to be within 0.01. We double the SD to reflect the double-sided nature of the Bias. In other words, if $SD\left[\text{Bias}(\hat{\theta})\right] = 0.005$, then the range of likely bias values is $2 \cdot 0.005 = 0.01$. This does not apply to an arbitrary precision, d. $2 \cdot SD\left[\text{Bias}(\hat{\theta})\right] < 0.01 \implies N > Var(\hat{\theta}) \cdot \left(\frac{1}{0.005}\right)^2$. Thus, $N_{\min,0.01} = \left\lceil Var(\hat{\theta}) \cdot \left(\frac{1}{0.005}\right)^2 \right\rceil$. For arbitrary d, $N_{\min,d} = \left\lceil \frac{Var(\hat{\theta})}{d^2} \right\rceil$.

  - We could obtain an estimate for $Var\left[\text{Bias}(\hat{\theta})\right]$ based on previous simulation experiments, in order obtain a numerical answer for $N_{\min}$. ■

- **Example**: Variance estimation. Find the minimum MC sample size N such that $s^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$ is within a desired precision. Assume we have a preliminary estimate of the variance of the estimator.
  The ABAR for s^2_{N-1} is $s^2_{N-1} = \sigma^2 + \frac{1}{N}\sum_{i=1}^{N}\left[(X_i - \mu)^2 - \sigma^2\right] + o_p(n^{-1/2})$. Using this result, it stands that $s^2_{N-1} \sim AN\left(\sigma^2, \frac{1}{N}(\mu_4 - \sigma^4)\right)$.
  As in the previous example, $\frac{1}{\sqrt{N}}\sqrt{\mu_4 - \sigma^4} < d \implies N > \frac{\mu_4 - \sigma^4}{d^2}$, so $N_{\min} = \left\lceil \frac{\mu_4 - \sigma^4}{d^2} \right\rceil$. ■

- **Example**: Confidence intervals. What is the minimum MC sample size N such that the error in coverage probability is within some precision?
  The actual coverage probability is $AC = \frac{1}{N}\sum_{b=1}^{N}\mathbb{I}\left(\theta \in \hat{\theta}_b \pm z^*_{\alpha/2}\sqrt{Var(\hat{\theta}_b)}\right) = \frac{1}{N}\sum_{b=1}^{N}\mathbb{I}\left(\left|\frac{\theta - \hat{\theta}_b}{\sqrt{Var(\hat{\theta})}}\right| < z^*_{\alpha/2}\right)$.
  Note that these are now Bernoulli variables. In other words, $\mathbb{I}\left(\left|\frac{\theta - \hat{\theta}_b}{\sqrt{Var(\hat{\theta})}}\right| < z^*_{\alpha/2}\right) \sim \text{Ber}(1-\alpha)$. AC is thus the sum of iid Bernoulli RVs, so $\mathbb{E}(AC) = 1 - \alpha$, and $Var(AC) = \frac{\alpha(1-\alpha)}{N}$.
  As before, $\sqrt{\frac{\alpha(1-\alpha)}{N}} < d \implies N_{\min} = \left\lceil \frac{\alpha(1-\alpha)}{d^2} \right\rceil$. ■

- **Example**: Power estimation. Given $RR = \{T > c_\alpha\}$, determine the minimum MC sample size N such that $\hat{\beta}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(T_i > c_\alpha)$ is within some precision.
  This is very similar to the previous example, where $\hat{\beta}(\theta) \sim \text{Bin}\left(N, \beta(\theta)\right)$, and $\beta(\theta)$ is the true power. $\sqrt{\frac{\beta(\theta)(1-\beta(\theta))}{N}} < d \implies N_{\min} = \left\lceil \frac{\beta(\theta)(1-\beta(\theta))}{d^2} \right\rceil$.
  We technically can provide a more numerical bound. Since $\beta(\theta) \in (0,1)$, $\beta(\theta)(1-\beta(\theta))$ is maximized at $\frac{1}{4}$, in which case $N_{\min} = \left\lceil \frac{1}{4d^2} \right\rceil$. ■

- In any Monte Carlo study, report a measure of the standard deviation in the results.

- One option is to calculate the ratio $R_N = \frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}^2_{i,n}}{s^2_{N-1}}$.

  - If $R_N > 1$, then $\hat{\sigma}^2_n$ is 7% too large on average. If $R_N < 1$, then $\hat{\sigma}^2_n$ is too small on average.
  - Whether or not $R_N > 1$ or $R_N < 1$ significantly, we need to know the SE of R_N. Using ABAR and Delta Method, we obtain

$$Var(R_N) \simeq \frac{1}{N} \cdot \frac{\sigma^4_{aN}}{\sigma^4_n}\left\{\text{Kurt}(\hat{\theta}) - 1 - \frac{2Cov\left([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2, \hat{\sigma}^2_n\right)}{\sigma^2_n\sigma^2_{aN}} + \frac{Var(\hat{\sigma}^2_n)}{\sigma^4_{aN}}\right\}.$$

    This lets us compute $SE(R_N)$.

    * Uses the fact that $s^2_{N-1} \approx s^2_N$.
    * $\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}^2_{aNi} = \sigma^2_{aN} + \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}^2_{aNi} - \sigma^2_{aN})$ (mean of ABARs).
    * $s^2_N = \sigma^2_N + \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{\theta}^2_{iN} - \theta)^2 - \sigma^2_N\right] + o_p(n^{-1/2})$, where our h function is akin to $\boldsymbol{\psi}$.

* $\sqrt{N}\left(\begin{pmatrix}\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}^2_{aNi}\\ s^2_N\end{pmatrix}-\begin{pmatrix}\sigma^2_{aN}\\ \sigma^2_N\end{pmatrix}\right)\xrightarrow{\text{d}}\mathcal{N}_2\left(\mathbf{0},\begin{pmatrix}Var(\hat{\sigma}^2_{aNi}) & (\cdot)\\ Cov\left(\hat{\sigma}^2_{aNi},(\hat{\theta}_{aN}-\theta)^2\right) & \mu_4-\sigma^4_N\end{pmatrix}\right)$. Applying the Delta method with $f(x,y)=\frac{x}{y}$,

$$\sqrt{N}\left(R_N-\frac{\sigma^2_{aN}}{\sigma^2_N}\right)\xrightarrow{\text{d}}\mathcal{N}\left(0,Var(R_N)\right).$$

- MSE combines measures of bias and variance.
  - MSE often underestimates variance. This means CIs might suffer from undercoverage.
- If we wish to compare two estimators, we would conduct a paired t-test.
  - We will likely have correlated estimators, so we take $Var(\hat{\theta}_1-\hat{\theta}_2)$.
  - We can also compare which estimator is better by which one has a smaller variance, by once again taking the ratio $\frac{s^2_{1,N-1}}{s^2_{2,N-1}}$, which has asymptotic variance (via ABAR and Delta method)

$$\frac{1}{N}\cdot\frac{\sigma^4_{1,n}}{\sigma^2_{2,n}}\left\{\text{Kurt}(\hat{\theta}_1)+\text{Kurt}(\hat{\theta}_2)-2-2\cdot\frac{Cov(\hat{\theta}_1,\hat{\theta}_2)}{\sigma^2_{1,n}\sigma^2_{2,n}}\right\}.$$

- Tips for presenting results:
  - Use graphs whenever possible.
  - Provide standard errors of estimates whenever possible.
  - Unless absolutely necessary, don't include more than two decimal places.
    * You definitely don't need more decimal places than that of the SE.

8 Random Math Stuff

Return to Table of Contents

$\binom{n}{r} = \frac{n!}{r!(n-r)!}$	$(x+z)^n = \sum_{y=0}^{n} \binom{n}{y} x^y z^{n-y}$	$e^{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$	$\sum_{y=0}^{\infty} ap^y = \frac{a}{1-p}$
$\underbrace{\left(1+\frac{a_n}{n}\right)^n \to e^a}_{a_1,\ldots,a_n \to a}$	$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$	$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \ldots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \ldots & a_{mn}\boldsymbol{B} \end{bmatrix}$	$(\sum_{i=1}^{p} x_i)^n = \underbrace{\sum \frac{n!}{\prod_{i=1}^{p} k_i} \prod_{j=1}^{p} x_j^{k_j}}_{\sum_{i=1}^{p} k_i = n,\ k_i \geq 0}$
$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt$	$\Gamma(a+1) = a\Gamma(a),\ a > 0$	$\Gamma(n+1) = n!,\ n \in \mathbb{Z}$	$\Gamma(1/2) = \sqrt{\pi}$
$\sum_{i=0}^{\infty} \frac{f^{(i)}(a)(x-a)^i}{i!}$	$\sum_{r=0}^{n} a^r = \frac{1-a^{n+1}}{1-a}.$		

9 Distributions

Return to Table of Contents

Note: Parameterizations may vary. I used the parameterizations as in *Casella and Berger*.

Discrete Distributions

Name	PMF	Support	$E(X)$	$Var(X)$	MGF
Bernoulli	$p^x(1-p)^{1-x}$	$x \in \{0,1\}$	p	$p(1-p)$	$(1-p)+pe^t$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$x \in \{0,1,\dots,n\}$	np	$np(1-p)$	$[pe^t+(1-p)]^n$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$x \in \{0,1,\dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Geometric	$p(1-p)^{x-1}$	$x \in \{1,2,\dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
NegBin	$\binom{r+x-1}{x}p^r(1-p)^x$	$x \in \{0,1,\dots\}$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{p}{1-(1-p)e^t}\right)^r$
HyperGeom	$\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$	$x \in \{0,1,\dots,K\}$	$\frac{KM}{N}$	$\frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$	DNE
Multinomial	$n!\prod_{i=1}^k \frac{p_i^{x_i}}{x_i!}$	$x_i : \sum_{i=1}^k x_i = n$	$E(X_i)=np_i$	$np_i(1-p_i)$	$\left(\sum_{i=1}^k p_i e^{t_i}\right)^n$

Continuous Distributions

Name	PDF	Support	$E(X)$	$Var(X)$	MGF or $E(X^n)$
Uniform	$\frac{1}{b-a}$	$x \in [a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$x \in [0,1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
Exp.	$\frac{1}{\beta}e^{-x/\beta}$	$x \geq 0$	β	β^2	$\frac{1}{1-\beta t}$
Gamma	$\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$	$x \geq 0$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta t}\right)^\alpha$
Normal	$\frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma^2}}$	$x \in \mathbb{R}$	μ	σ^2	$e^{\mu t+\frac{t^2\sigma^2}{2}}$
Weibull	$\frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta}$	$x \geq 0$	$\beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right)$	$\beta^{2/\gamma}\left[\Gamma\left(1+\frac{2}{\gamma}\right)-\Gamma\left(1+\frac{1}{\gamma}\right)^2\right]$	$E(X^n)=\beta^{n/\gamma}\Gamma\left(1+\frac{n}{\gamma}\right)$
Cauchy	$\frac{1}{\pi\sigma}\frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}$	$x \in \mathbb{R}$	DNE	DNE	Neither DNE
GEV	$F=\begin{cases}\exp\left\{-e^{(x-\mu)/\sigma}\right\}, & \xi=0\\ e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}}, & \xi\neq 0\end{cases}$	$x \in \mathbb{R}$	$\begin{cases}\mu+\sigma\gamma, & \xi=0\\ \mu+\frac{g_1(\sigma-1)}{\xi}, & \xi<1\end{cases}$	$\begin{cases}\frac{\pi^2\sigma^2}{6}, & \xi=0\\ \sigma^2\frac{g_2-g_1^2}{\xi}, & \xi<\frac{1}{2}\end{cases}$	Non-trivial
Log N	$\frac{\exp\left\{\frac{(\log(x)-\mu)^2}{-2\sigma^2}\right\}}{x\sqrt{2\pi\sigma^2}}$	$x \geq 0$	$e^{\mu+\frac{1}{2}\sigma^2}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$	$E(X^n)=e^{n\mu+\frac{1}{2}n^2\sigma^2}$
Bivar N	$\frac{\exp\left\{-\frac{1}{2(1-\rho^2)}(*)\right\}}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}}$	$\begin{pmatrix}x_1\\x_2\end{pmatrix}\in\mathbb{R}^2$			
$\boldsymbol{\chi^2}$	$\frac{x^{p/2-1}e^{-x/2}}{\Gamma(p/2)2^{p/2}}$	$x \geq 0$	p	$2p$	$\left(\frac{1}{1-2t}\right)^{p/2}$
$\boldsymbol{F}$	$\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)(\nu_1/\nu_2)^{\nu_1/2}(x)^{(\nu_1-2)/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}$	$x \geq 0$	$\frac{\nu_2}{\nu_2-2}$, $\nu_2>2$	$2\left(\frac{\nu_2}{\nu_2-2}\right)^2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)}$, $\nu_2>4$	$E(X^n)=\frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(\nu_1/\nu_2)^n}$, $n<\frac{\nu_2}{2}$
$\boldsymbol{T}$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$x \in \mathbb{R}$	0, $\nu>1$	$\frac{\nu}{\nu-2}$, $\nu>2$	$E(X^n)=\begin{cases}\frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{\nu-n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\nu^{-n/2}}, & n<\nu\\ 0, & n<\nu\end{cases}$

(*) $\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2$

9.1 Equivalences

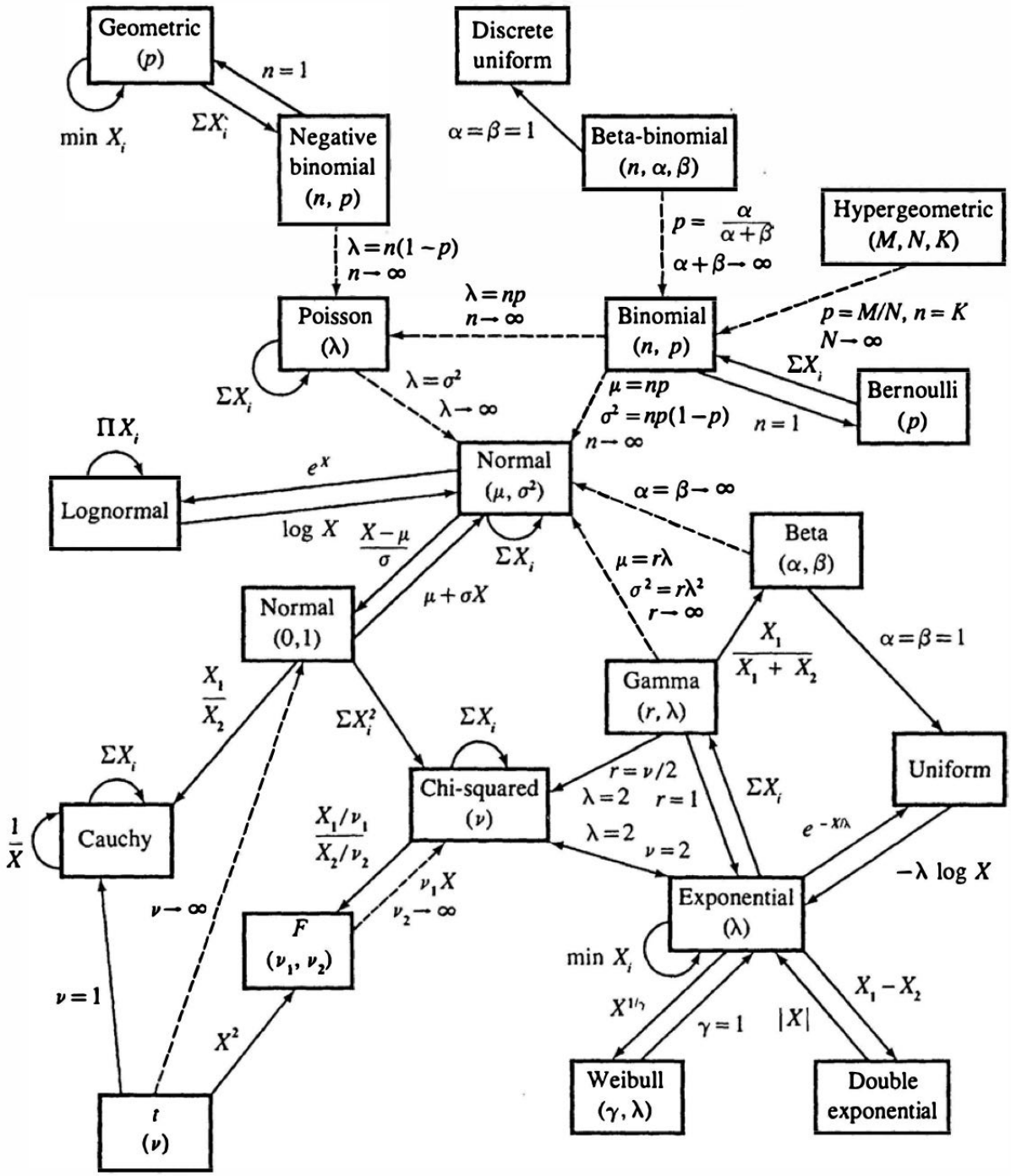
Return to Table of Contents

- $Bin(1,p)=Ber(p)$.
- $NegBin(1,p)=Geom(p)$.
- $MN(n,(p,1-p))=Bin(n,p)$.
- $\text{Gamma}(1,\beta)=Exp(\beta)$.
- $\text{Gamma}\left(\frac{p}{2},2\right)=\chi^2_p$.
- $\text{Weibull}(1,\beta)=Exp(\beta)$.

- $\frac{X}{Y} \sim$Cauchy$(0,1)$, $X \perp Y \sim \mathcal{N}(0,1)$.
- If $X \sim$Exp(1), then $\mu - \sigma \log(X) \sim$GEV$(\mu, \sigma, 0)$.
- If $X \sim$Weibull(μ, σ), then $\left[1 - \sigma \log\left(\frac{X}{\sigma}\right)\right] \sim$GEV$(\mu, \sigma, 0)$.
- If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0,1)$, then $X_{(k)} \sim$Beta$(k, n+1-k)$.
- If $X_1, \ldots, X_n \overset{\perp}{\sim}$Pois$(\lambda_i)$, then $(\underline{X} | n = \sum_{i=1}^n X_i) \sim MN(\sum_{i=1}^n X_i, \underline{\pi})$, where $\pi_j = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$.
- $Cauchy(\mu, \sigma) = t_1(\mu, \sigma)$.
- If $X \sim$Weibull$\left(\lambda, \frac{1}{2}\right)$, then $\sqrt{X} \sim$Exp$\left(\frac{1}{\sqrt{\lambda}}\right)$.
- If $X_1, \ldots, X_n \overset{\text{iid}}{\sim}Bin(m, p_i)$, then $(X_1, \ldots, X_{n-1} | X_n = x_n) \sim MN\left(m - x_n, \left[\frac{p}{1-p_n}\right]^{\underline{x}}\right)$.
- If $X \sim U(0,1)$, then $-\log(X) \sim$Exp(1).
- If $X_1, \ldots, X_n \overset{\text{iid}}{\sim}Exp(\lambda)$, then $\sum_{i=1}^n X_i \sim$Gamma(n, λ).
- If $X_1, \ldots, X_n \overset{\perp}{\sim}$Pois$(\lambda_i)$, then $\sum_{i=1}^n X_i \sim$Pois$(\sum_{i=1}^n \lambda_i)$.

Memorylessness

- **Discrete case**: $P(X > m+n | X \geq m) = P(X > n)$.
  - Geometric distribution is memoryless.
- **Continuous case**: $P(X > m+n | X > m) = P(X > n)$.
  - Exponential distribution is memoryless.



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Source: Casella and Berger, *Statistical Inference*.