PhD Qualifying Exam Review

Fall 2024
Version 5

Miles Woollacott

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

Note: This document is likely not free from typos. Please email the author with any observed typos or other general issues with this document.

Contents

1 ST 701: Statistical Theory I

Instructor: Dr. Luo Xiao
Semester: Fall 2023

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 $i \in \{1, \dots, n\}$, 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 $A_1, \dots, A_p$, 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 $Z \sim N(0,1)$, 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 $E(Y) = \int_{\zeta_Y} y f_Y(y)dy$ if continuous, or $\sum_{\zeta_Y} y f_Y(y)$ if discrete.
  - If $E(X^2)$ exists, then $E(X-b)^2$ is minimized at $b = E(X)$.
- **Markov's inequality**: $P(X \geq a) \leq \frac{E(X)}{a}$.
- **Chebyshev's inequality**: $P(|X| \geq a) \leq \frac{E(X^2)}{a^2}$ 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 $|E(XY)| \leq E|XY| \leq (E|X|^p)^{1/p}(E|Y|^q)^{1/q}$.
- **Jensen's inequality**: Suppose a function f is convex. Then, $E(f(X)) \geq f(E(X))$.
- **Stein's Lemma**: $E(g(X)(X-\mu)) = \sigma^2 E(g'(X))$, where $X \sim N(\mu, \sigma^2)$.
- The nth **moment** of an RV X is $E(X^n)$.
- The nth **central moment** of an RV X is $E[(X - E(X))^n]$.
- The **variance** of an RV X is $E[(X-E(X))^2] = E(X^2) - E(X)^2$.
  - $Var(a + bX) = b^2 Var(X)$.
- The **moment generating function**, or **MGF**, of an RV X is $M_X(t) = E(e^{tX})$.
  - $M_{(aX+b)}(t) = e^{tb}M_X(at)$.
  - If $M_X(t) < \infty$ for all t in an open interval containing zero, then $E(X^n) = \frac{d^n}{dx^n}M_X(t)|_{t=0}$.
  - 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.
  - $E\left(\sum_{i=1}^p \frac{\partial}{\partial\theta_j}w_i(\boldsymbol{\theta})t_i(X)\right) = -\frac{\partial}{\partial\theta_j}\log c(\boldsymbol{\theta})$.
  - $Var\left(\sum_{i=1}^p \frac{\partial}{\partial\theta_j}w_i(\boldsymbol{\theta})t_i(X)\right) = -\frac{\partial^2}{\partial\theta_j^2}\log c(\boldsymbol{\theta}) - E\left(\sum_{i=1}^p \frac{\partial^2}{\partial\theta_j^2}w_i(\boldsymbol{\theta})t_i(X)\right)$.
- 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: $z < 0$.

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

  Case 2: $z > 0$.

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

  This means that $f_Z(z) = \frac{1}{4}e^{-|z|/2}$ for $z \in \mathbb{R}$. ■

- If $X \perp Y$, then $U = h(X) \perp V = g(Y)$.
- $E(Y) = E(E(Y|X))$.
- $M_Y(t) = E\left[E(e^{tY}|X)\right]$.
- $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$.
- $Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y)$.
- $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$.
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$.

1.5 Statistics and Order Statistics

Return to Table of Contents

- **Statistics** are functions of random variables.
- **Sample variance** is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
  - $E(S^2) = \sigma^2$.
  - If $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$:
    - $\bar{X} \perp S^2$.
    - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- 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 $\frac{pX/q}{1+pX/q}$ ~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 $a \xrightarrow{\text{p}} E(X)$.
- 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 $E(X_i) = \mu$ 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 $\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$ 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 $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} D(\mu, \sigma^2)$, where $\sigma^2 < \infty$. Then, $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2)$.
- **First-order delta method**: Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2)$. If $g'(\mu) \neq 0$, then $\sqrt{n}\left(g(X_n) - g(\mu)\right) \xrightarrow{\text{d}} N\left(0, \sigma^2[g'(\mu)]^2\right)$.
  - 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 $\frac{X-m}{\sqrt{2m}} \xrightarrow{\text{d}} N(0, 1)$.

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

a. Let N_t ~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 ~Gamma$(n,2) \equiv \chi^2_{2n} = \chi^2_m$.

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

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

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,

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

- **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 $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} U(0,1)$. 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 $E\left(\frac{U_1}{S_n}\right)$.

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

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

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

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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$. 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 $\left[\frac{P_\theta(x^n)}{P_\theta(y^n)} \text{ is constant in } \theta \Leftrightarrow T(x^n) = T(y^n)\right]$, 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \theta^2)$, where $\theta > 0$. Calculate $E\left(\frac{X_1^2}{\sum_{i=1}^n X_i^2}\right)$.

  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,

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

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

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

  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,

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

- **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 $S_X(\theta) = \frac{\partial}{\partial\theta}\log P_\theta(X)$.
  - $E_\theta[S_X(\theta)] = 0$ for all θ.
- The **Fisher information** of $X \sim p_\theta$ is $I_X(\theta) = Var_\theta[S_X(\theta)] = E_\theta[S_X(\theta)^2]$.
  - $I_{X^n}(\theta) = nI_{X_i}(\theta)$.
  - $I_{T(X^n)}(\theta) = I_X(\theta)$ if T is sufficient.
  - If P_θ is exponential family, then $I_X(\theta) = -E_\theta\left[\frac{\partial}{\partial\theta}S_\theta(X)\right]$.
  - $I_{X^n}(\theta) = E_\theta[I_{W|U}(\theta)]$, 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$, 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, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{d}} N(0, I(\theta)^{-1})$, and $Var(\hat{\theta}_n)$ achieves the CRLB.
    - If $\phi(\theta)$ is smooth, then $\sqrt{n}(\hat{\phi}_n - \phi) \xrightarrow{\text{d}} N(0, \underbrace{\dot{\phi}(\theta)' I(\theta)^{-1} \dot{\phi}(\theta)}_{=:V^{\phi}(\theta)})$.
      - We would need to estimate $V^{\phi}(\theta)$. Directly using the MLE is volatile for small n. Bootstrapping is okay. Could also use $\hat{V}_n^{\phi} \overset{\text{set}}{=} \dot{\hat{\phi}}_n'[J_n(\hat{\theta}_n)]^{-1}\dot{\hat{\phi}}_n$, 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-$\boldsymbol{\alpha}$ 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 − $\boldsymbol{\alpha}$)% confidence set** if the coverage probability is at least $1 - \alpha$.
  - If $\delta^{\alpha}_{\theta_0}$ is a size-α test of $H_0 : \theta = \theta_0$, then $C^{\alpha}(X^n) = \{\theta_0 : \delta^{\alpha}_{\theta_0}(X^n) = 0\}$ 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 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} U(\theta_1, \theta_2)$, 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.

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

b.

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

$L_n(\theta_1, \theta_2)$ 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.

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

This means that $\hat\theta_2 = \max(-X_{(1)}, X_{(n)})$.

e.

$$\begin{aligned} \lambda(x^n) =& \frac{L_n((\hat\theta_2)_{H_0}, (\hat\theta_2)_{H_0})}{L_n(\hat\theta_1, \hat\theta_2)} \\ =& \frac{(2(\hat\theta_2)_{H_0})^{-n} I((\hat\theta_2)_{H_0} > \max(-X_{(1)}, X_{(n)}))}{(\hat\theta_2 - \hat\theta_1)^{-n} I(X_{(1)} > \hat\theta_1) I(X_{(n)} < \hat\theta_2)} \\ =& \left(\frac{X_{(n)} - X_{(1)}}{2\cdot\max(-X_{(1)}, X_{(n)})}\right)^n = \begin{cases} \left(\frac{X_{(n)}-X_{(1)}}{2\cdot(-X_{(1)})}\right)^n, & -X_{(1)} < X_{(n)} \\ \left(\frac{X_{(n)}-X_{(1)}}{2\cdot X_{(n)}}\right)^n, & \text{o.w.} \end{cases} = \begin{cases} \left(\frac{1}{2} - \frac{X_{(n)}}{2X_{(1)}}\right)^n, & -X_{(1)} < X_{(n)} \\ \left(\frac{1}{2} - \frac{X_{(1)}}{2X_{(n)}}\right)^n, & \text{o.w.} \end{cases} \quad \blacksquare \end{aligned}$$

- A **Wald pivot** is $\frac{\hat\phi - \phi}{\sqrt{Var(\hat\phi)}} \xrightarrow{\text{d}} N(0,1)$.
  - If $\hat\phi = \hat\phi_n$, then $Var(\hat\phi_n) \equiv \frac{1}{J_n(\vartheta)}$.
- A **score pivot** is $\frac{S_\theta(X^n)}{I_{X^n}(\theta)} \xrightarrow{\text{d}} N(0,1)$.
- 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** $R(\theta, \delta) = E_\theta \left\{ L(\theta, \delta(X^n)) \right\}$.
    - **MSE** is a risk function, where $MSE_\delta = E_\theta\{(\delta(X^n) - \theta)^2\} = [\theta - E_\theta(\delta)]^2 + Var(\delta)$.
  - 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, $\delta^{RB} = E\{\delta(X^n)|T\}$ 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Pois(\theta)$. 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, $\delta^{RB}(T) = E(X_1^2|T) = Var(X_1^2|T) + E(X_1|T)^2 = \frac{T}{n}\left(1 - \frac{1}{n}\right) + \left(\frac{T}{n}\right)^2$. ■

- 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 $\delta_\alpha^*(X^n) = I\left(\frac{L_n(\theta_1)}{L_n(\theta_0)} > k_\alpha\right)$, 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 $\hat{\phi}_{Bayes} = E_{\pi(\theta|X)}[\phi]$.
  - 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

3.1 Hypothesis Tests and CIs

Return to Table of Contents

- **Satterthwaite's approximation for $\boldsymbol{\nu}$**: $\nu \approx \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$.
- 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 $Y_i \overset{\text{i.i.d.}}{\sim} D$.
  - 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: $Z = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$.
  - 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:
    - ∗ $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$.

      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 $Y_{i1}\overset{\text{i.i.d.}}{\sim}D_1$, and $Y_{j2}\overset{\text{i.i.d.}}{\sim}D_2$.
  - 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 $(\bar{y}_1-\bar{y}_2)\pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\sim N(0,1)$ (approximate).
  - If D_1, D_2 are Normal, and $\sigma_1=\sigma_2$ ($S_1\approx S_2$) are both unknown, then $(\bar{y}_1-\bar{y}_2)\pm\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sim N(0,1)$.
- Hypothesis test for $\mu_1-\mu_2$:
  - σ_1, σ_2 known: $Z=\frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$.
  - 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, $\Delta_0\neq 0$:
    - $Z=\frac{(\hat{p}_1-\hat{p}_2)-\Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\sim N(0,1)$.
    - Power and sample size done computationally.
  - Independent data, $\Delta_0=0$:
    - $Z=\frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim N(0,1)$, 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:
    - $Z=\frac{B-C-\Delta_0}{\frac{\sqrt{B+C-n\Delta_0^2}}{n}}\overset{\Delta_0=0}{=}\frac{B-C}{\sqrt{B+C}}\sim N(0,1)$.
    - 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 $E_{ij} \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$, 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

- $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right)$.
  - If $\boldsymbol{A}$ is only one row, then $t = \frac{\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m}}{\sqrt{MSE\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'}} \sim t_{df_E}$.
  - If $\boldsymbol{A}$ has $k > 1$ independent rows, then $F = \frac{Q/k}{MSE} \sim F_{k,df_E}$.
    - * $Q = (\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m}) = SSE_R - SSE_F$.
    - * 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 $W_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$.
- 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 $E \sim N(0, \sigma^2)$.
  - 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 $1-\beta \leq$ 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 $E_{ijk} \sim N(0, \sigma^2)$.

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 $\theta = \boldsymbol{c}'\boldsymbol{\mu}$, where $\boldsymbol{\mu}' = \Big(\alpha_1, \ldots, \alpha_a, |\beta_1, \ldots, \beta_b, |(\alpha\beta)_{11}, \ldots, (\alpha\beta)_{ab}\Big)$.
    - The simple effect of β_j is defined as $\theta_{AB_j} = E(\bar{Y}_{ij+} - \bar{Y}_{kj+})$.
  - $E(Y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

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 $E_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $T_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_T^2)$, 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 $S_{k(i)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_s^2) \perp E_{ijk}$.
  - $S_{k(i)}$ 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	$a-1$	$bn\psi_A^2 + b\sigma_s^2 + \sigma^2$	$\frac{MSA}{MSS(A)}$	$\boldsymbol{P}_{\boldsymbol{X}_A} - \boldsymbol{P}_{\boldsymbol{1}}$
$S(A)$	$\sum_i n_i - a$	$b\sigma_s^2 + \sigma^2$	$\frac{MSS(A)}{MSE}$	$\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{X}_A}$
B	$b-1$	$an\psi_B^2 + \sigma^2$	$\frac{MSB}{MSE}$	$\boldsymbol{P}_{\boldsymbol{X}_B} - \boldsymbol{P}_{\boldsymbol{1}}$
AB	$(a-1)(b-1)$	$n\psi_{AB}^2 + \sigma^2$	$\frac{MSAB}{MSE}$	$\boldsymbol{P}_{\boldsymbol{X}_{AB}} - \boldsymbol{P}_{\boldsymbol{X}_A} - \boldsymbol{P}_{\boldsymbol{X}_B} + \boldsymbol{P}_{\boldsymbol{X}_{\boldsymbol{1}}}$
Error	$(\sum_i n_i - a)(b-1)$	σ^2		$\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{Z}} + \boldsymbol{P}_{\boldsymbol{X}_A} - \boldsymbol{P}_{\boldsymbol{X}_{AB}}$
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, E(MSE) is always σ^2. Every other EMSQ inherits this σ^2. For 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 E(MSaerate), similarly count up the number of levels of combinations of station and species, which is 12. Similar logic follows for 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 $\hat{\theta}_1 := \bar{y}_{i..} - \bar{y}_{j..}$.

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

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

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

d.

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

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

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 $\left\|Y - X\hat{\beta}\right\|_2^2$ w.r.t. $\hat{\beta}$.
  - Equivalent to $\min_{\hat{\beta}} \|\boldsymbol{e}\|_2^2$.
  - Under OLS, $\hat{\boldsymbol{Y}} \perp \boldsymbol{e}$.
  - Assumptions:
    - $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 \boldsymbol{I})$.
    - 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 $e_i \sim N$, 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 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}'\boldsymbol{Y} =: \boldsymbol{PY}$.
  - 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 $\boldsymbol{A} = \boldsymbol{CX}$.
  - $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}')$.
  - Estimation: $\hat{\boldsymbol{Y}}_0 \sim N(\boldsymbol{X}_0\boldsymbol{\beta}, \sigma^2\boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}_0)$.
  - Prediction: $\hat{\boldsymbol{Y}}_0 \sim N(\boldsymbol{X}_0\boldsymbol{\beta}, \sigma^2\boldsymbol{I} + \sigma^2\boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}_0)$.
  - $\hat{\boldsymbol{Y}} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{P})$.
  - $\boldsymbol{e} \sim N(0, \sigma^2(\boldsymbol{I} - \boldsymbol{P}))$.
- 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 $E(\hat{\boldsymbol{Y}}) \neq \boldsymbol{X\beta}$, and $E(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$.
    - If the columns of $\boldsymbol{Z}$ are orthogonal to $\boldsymbol{X}$, then $E(\hat{\boldsymbol{Y}}) \neq \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$.
    - This means that estimating $\boldsymbol{\beta}$ and estimating $E(\boldsymbol{Y})$ 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 $Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda-1}}{\lambda Y^{\lambda-1}}, \ \lambda \neq 0 \\ Y \log(Y_i), \ \lambda = 0 \end{cases}$. Choose λ such that $SSE^{(\lambda)} \leq SSE^{(\lambda)}_{\min} \left[1 + \frac{t^2_{df_E, \alpha/2}}{df_E}\right]$.
        - If $Var(Y) \propto [E(Y)]^{2k}$, 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 $E(Y)$ 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:

  - $SSE = \|\boldsymbol{e}\|_2^2$.
  - $s^2 = \frac{SSE}{n-rank(\boldsymbol{X})} = \frac{SSE}{n-(p+1)}$.
  - $Var(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
  - $E(Y|x_0) = \sigma^2 \boldsymbol{x_0}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_0}$.
  - $R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST}$.

- 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 $\boldsymbol{F}$-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 $\boldsymbol{C_p}$** 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.
  - $\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X'Y}$.
    - Is linear in $\boldsymbol{Y}$.
    - $Bias(\hat{\boldsymbol{\beta}}^{ridge}) = -\lambda(\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}$, so larger λ means more bias.
    - $Var(\hat{\boldsymbol{\beta}}^{ridge}) = \sigma^2(\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X'X})(\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}$, so larger λ means less variance.
  - Shrinkage is proportional, so $\hat{\beta}_j^{ridge} = \frac{n}{n+\lambda}\hat{\beta}_j$ for orthogonal $X'X$.
  - 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 $\boldsymbol{X}'\boldsymbol{X}$.
  - 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 $E(\boldsymbol{Y}) = x'\boldsymbol{\beta} =: \boldsymbol{\eta}$.

- A **generalized linear model**, or **GLM**, is defined such that $g(E(Y_i)) = \eta_i$, 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 $g^{-1}(x_i'\boldsymbol{\beta})$, 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, $b'(\boldsymbol{\theta}) = E(Y_i)$.
    - $b''(\boldsymbol{\theta})$ is the **variance function**; that is, $b''(\boldsymbol{\theta}) = a(\phi)Var(Y_i)$.
- $\hat{\boldsymbol{\beta}} \dot{\sim} N\left(\boldsymbol{\beta}, a(\phi)[\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F}]^{-1} \right)$, where $\boldsymbol{F} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'}$, $V_{ii} = Var(\mu_i)$ (0 o.w.).
    - $\hat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, [I(\boldsymbol{\beta})]^{-1})$, where $I(\boldsymbol{\beta})_{ij} = -E\left[\left\{ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j} \right\}_{ij} \right]$ under regularity conditions.
    - $T_W = (\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[a(\phi)\boldsymbol{L}(\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F})^{-1}\boldsymbol{L}']^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \chi^2_q$.
    - $T_{LR} = 2(\ell(\hat{\boldsymbol{\beta}}_F) - \ell(\hat{\boldsymbol{\beta}}_R)) \sim \chi^2_q$, 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 $r_{s,D_i} = \frac{r_{D_i}}{\hat{\phi}(1 - h^{GLM}_{ii})^{1/2}}$.
- 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 $\boldsymbol{\beta}$.
        - $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i\eta_i - \log(1 + e^{\eta}_i)]$.
        - 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 $e^{\boldsymbol{x}'\boldsymbol{\beta}}$.
    - $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left[y_i - e^{x_i'\beta} - \log(y_i!)\right]$.
    - 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 $\boldsymbol{KY} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{KK}')$, where $\boldsymbol{K}$ is positive-definite.
    - $\ell_{REML}(\sigma^2;\boldsymbol{y}) = c - \frac{1}{2}\log\left|\sigma^2\boldsymbol{KK}'\right| - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{K}'(\boldsymbol{KK}')^{-1}\boldsymbol{Ky}$.
        - 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.
    - $\boldsymbol{\epsilon} \sim N(0, V)$, where V is positive-definite.
- The **classical linear mixed model** is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}$, where $\boldsymbol{\alpha} \sim N(0, \boldsymbol{G})$ is a vector of our random effects, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{R})$ 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 $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ}' + \boldsymbol{R})$, where $Var(\boldsymbol{\delta}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$.
    - The **subject-specific model** is $\boldsymbol{Y}|\boldsymbol{\alpha} \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\alpha}, \boldsymbol{R})$.
    - $\ell(\boldsymbol{\beta}, \boldsymbol{\delta}; \boldsymbol{y}) = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$.
        - If $\boldsymbol{\delta}$ is known, then $\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$.
        - $\hat{\boldsymbol{\beta}}_{ML} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1})$.
            - 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.
    - $\ell(\boldsymbol{\delta}; \boldsymbol{y})_{REML} = c - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$.
        - 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, $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}')$.
        - A $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm z_{\alpha/2}\sqrt{[(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}]_{jj}}$.
        - $(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m}) \sim \chi^2_{rank(\boldsymbol{A})}$.
    - With $\boldsymbol{V} = \sigma^2\boldsymbol{D}$, where $\boldsymbol{D}$ known, $\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{A\beta}, \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}')$.
        - A $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm t_{\alpha/2, n-rank(\boldsymbol{X})}\sqrt{\hat{\sigma}^2_{REML}\left[(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\right]_{jj}}$.
        - $\frac{n-rank(\boldsymbol{X})}{\sigma^2}\hat{\sigma}^2_{REML} \sim \chi^2_{n-rank(\boldsymbol{X})}$, independent of $\hat{\boldsymbol{\beta}}$.
        - $\frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{m})/rank(\boldsymbol{A})}{\hat{\sigma}^2_{REML}} \sim F_{rank(\boldsymbol{A}), n-rank(\boldsymbol{X})}$.
    - With $\boldsymbol{V} = \sigma^2\boldsymbol{D}$, where both are unknown, $\hat{\boldsymbol{\beta}} \dot\sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1})$.
        - $(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\left[\hat{\sigma}^2_{REML}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{m}) \to \chi^2_{rank(\boldsymbol{A})}$ (same with using σ^2).

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

- With $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$, $\hat{\boldsymbol{\theta}} \dot{\sim} N(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta}) = diag\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}, \frac{n}{2\sigma^4}\right)$.
  - Suppose we want to test $H_0 : h(\boldsymbol{\theta}) = \boldsymbol{0}$ vs. $H_1 : h(\boldsymbol{\theta}) \neq \boldsymbol{0}$, where $h(\boldsymbol{\theta}) \in \mathbb{R}^r$.
    - $T_W = h(\hat{\boldsymbol{\theta}}_n)' \left[H(\hat{\boldsymbol{\theta}}_n)I(\hat{\boldsymbol{\theta}}_n)^{-1}H(\hat{\boldsymbol{\theta}}_n)'\right]^{-1} h(\hat{\boldsymbol{\theta}}_n) \sim \chi^2_r$, where $H(\boldsymbol{\theta}) = \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$.
    - $T_S = S(\hat{\boldsymbol{\theta}}_0)'I(\hat{\boldsymbol{\theta}}_0)^{-1}S(\hat{\boldsymbol{\theta}}_0) \sim \chi^2_r$.
  - 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 $E(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha}$, a **generalized linear mixed model** has the form $E(\boldsymbol{Y}|\boldsymbol{\alpha}) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha})$.
  - $Var(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{A}^{1/2}\boldsymbol{R}\boldsymbol{A}^{1/2}$, 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

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 $\boldsymbol{A}^{-1} = \boldsymbol{A}'$.
- The **Euclidean norm**, or **$\boldsymbol{\ell_2}$ norm**, is $\|\boldsymbol{x}\|_2 = (\langle x, x \rangle)^{1/2}$.
- The **Frobenius norm** is $\|\boldsymbol{A}\|_F = [\text{tr}(\boldsymbol{A}'\boldsymbol{A})]^{1/2}$.
- $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq \|\boldsymbol{a}\| \, \|\boldsymbol{b}\|$.
- $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$.
- 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 $\boldsymbol{u} \in \boldsymbol{S}$, $\boldsymbol{z} \in \boldsymbol{S}^{\perp}$.
  - $\text{col}(\boldsymbol{A})$ and $\text{null}(\boldsymbol{A}')$ 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 $\boldsymbol{B} = \boldsymbol{C}$.
- $\boldsymbol{X}'\boldsymbol{XA} = \boldsymbol{X}'\boldsymbol{XB}$ 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 $\boldsymbol{c} \in \text{col}(\boldsymbol{A})$.
  - 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.
- $(\boldsymbol{X}'\boldsymbol{X})^g \boldsymbol{X}'$ 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 $\text{col}(\boldsymbol{A})$.
  - $(\boldsymbol{I} - \boldsymbol{A}^g\boldsymbol{A})$ is a projection onto $\text{null}(\boldsymbol{A})$.
  - $\boldsymbol{P}$ is unique if it is symmetric.
  - If $\boldsymbol{P}$ is symmetric and projects onto S, then $\boldsymbol{I} - \boldsymbol{P}$ projects onto $S^{\perp}$.
- $\boldsymbol{P}_X := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^g \boldsymbol{X}'$ is the **symmetric projection matrix** of $\boldsymbol{X}$.
  - If $\text{col}(\boldsymbol{X}) = \text{col}(\boldsymbol{W})$, 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 $\boldsymbol{x'Ax} \geq 0$ 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 $\boldsymbol{A} = \boldsymbol{LL'}$.
- 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 $\boldsymbol{W} = \boldsymbol{XT}$.
- $\text{null}(\boldsymbol{X'X}) = \text{null}(\boldsymbol{X})$.
- $\text{col}(\boldsymbol{X'X}) = \text{col}(\boldsymbol{X'})$.
- $rank(\boldsymbol{X'X}) = rank(\boldsymbol{X})$.
- If $rank(\boldsymbol{BC}) = rank(\boldsymbol{B})$, then $\text{col}(\boldsymbol{BC}) = \text{col}(\boldsymbol{B})$.
- $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.
- $E[\boldsymbol{a'Y}] = \boldsymbol{a'}E[\boldsymbol{Y}]$, $E[\boldsymbol{Y'a}] = E[\boldsymbol{Y'}]\boldsymbol{a}$.
- $Var(\boldsymbol{Y}) = E\left[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))'\right]$.
- $Var(\boldsymbol{a'Y}) = \boldsymbol{a'}Var(\boldsymbol{Y})\boldsymbol{a}$.
- $Cov(\boldsymbol{a'X}, \boldsymbol{b'Y}) = \boldsymbol{a'}Cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{b}$.
- Trace trick: $E(\boldsymbol{X'X}) = \text{tr}\{E(\boldsymbol{XX'})\}$.

5.2 The Normal Equations

Return to Table of Contents

- For $f : \mathbb{R}^p \to \mathbb{R}$, the **gradient** is $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(\boldsymbol{x}) & \dots & \frac{\partial}{\partial x_p} f(\boldsymbol{x}) \end{pmatrix}'$.
  - $\nabla_{\boldsymbol{b}}(\boldsymbol{a'b}) = \boldsymbol{a}$.
  - $\nabla_{\boldsymbol{b}}(\boldsymbol{b'Ab}) = (\boldsymbol{A} + \boldsymbol{A'})\boldsymbol{b}$.
- 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 $\{\boldsymbol{\beta} : \boldsymbol{X'X\beta} = \boldsymbol{X'y}\}$.
  - Equivalent to $\{\boldsymbol{\beta} : \boldsymbol{X\beta} = \boldsymbol{P}_X\boldsymbol{y}\}$.
  - Equivalent to $\{\boldsymbol{\beta} : \boldsymbol{\beta} = (\boldsymbol{X'X})^g\boldsymbol{X'y} - [\boldsymbol{I}_p - (\boldsymbol{X'X})^g\boldsymbol{X'X}]\boldsymbol{z}\}$ for some $\boldsymbol{z}$.
  - $X\hat{\beta}$ is invariant to choice of $\hat{\beta}$ that solves the Normal equations.
- The **residual vector** is $\hat{\boldsymbol{e}} := \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.
  - $\hat{\boldsymbol{e}} \in \text{null}(\boldsymbol{X'})$.
  - The **sum of squared errors** is $SSE := \|\hat{e}\|_2^2 = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}$.

- 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 $\boldsymbol{\lambda}'\boldsymbol{\beta}$ iff $E(t(\boldsymbol{y})) = \boldsymbol{\lambda}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.
- An estimator $t(\boldsymbol{y})$ is **linear** if $t(\boldsymbol{y}) = c + \boldsymbol{a}'\boldsymbol{y}$ for constants c, $\boldsymbol{a}$.
- A function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is **estimable** iff there exists a linear unbiased estimator for it.
  - Under the linear mean model, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff there exists $\boldsymbol{a} : E(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$.
  - $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda} \in \text{col}(\boldsymbol{X}')$.
    - Equivalently, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{\lambda} \perp \text{null}(\boldsymbol{X})$.
  - $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable iff we can express $\boldsymbol{\lambda}'\boldsymbol{\beta}$ as a linear combination of $E(y_i)$.
  - If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then the least squares estimator $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ is invariant to the choice of $\hat{\boldsymbol{\beta}}$.
  - The least squares estimator $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ of an estimable $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is a linear unbiased estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.
  - If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the model with $\boldsymbol{X}$, and $\hat{\boldsymbol{c}}$ solves the Normal equations with $\boldsymbol{W}$, then $\boldsymbol{W}'\boldsymbol{T}\hat{\boldsymbol{c}}$ is the least squares estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$.
  - If $\boldsymbol{q}'\boldsymbol{c}$ is estimable in the reparameterized model, then $\boldsymbol{q}'\boldsymbol{S}\boldsymbol{b}$ is estimable in the original model with least squares estimator $\boldsymbol{q}'\hat{\boldsymbol{c}}$, where $\hat{\boldsymbol{c}}$ solves the Normal equations with $\boldsymbol{W}$.
- Consider $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}'\boldsymbol{C})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$. If $\text{col}(\boldsymbol{X}') \cap \text{col}(\boldsymbol{C}') = \{0\}$ and $rank(\boldsymbol{C}) = p - rank(\boldsymbol{X})$, then $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}'\boldsymbol{C})^{-1}$ exists, and is a generalized inverse for $\boldsymbol{X}'\boldsymbol{X}$.
  - $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the unique solution to the Normal equations and $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$.
  - $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{X}' = \boldsymbol{0}$.
  - $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}' = \boldsymbol{I}$.
- The **restricted model** is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$.
  - The **restricted Normal equations**, or **RNEs**, are $\left\{\beta : \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{P} \\ \boldsymbol{P}' & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{\delta} \end{pmatrix} \right\}$.
  - $\tilde{\boldsymbol{\beta}}$ solves the RNEs if $\boldsymbol{P}'\tilde{\boldsymbol{\beta}} = \boldsymbol{\delta}$ and $Q(\boldsymbol{\beta}) = Q(\tilde{\boldsymbol{\beta}})$.
  - $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable in the restricted model if there exists c, $\boldsymbol{a}$ such that $E(c + \boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$ that satisfy $\boldsymbol{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$.
  - $(c + \boldsymbol{a}'\boldsymbol{y})$ is estimable for $\boldsymbol{\lambda}'\boldsymbol{\beta}$ in the restricted model iff there exists a $\boldsymbol{d}$ such that $\boldsymbol{\lambda} = \boldsymbol{X}'\boldsymbol{a} + \boldsymbol{P}\boldsymbol{d}$, and $c = \boldsymbol{d}'\boldsymbol{\delta}$.
  - 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 $\boldsymbol{P}'\tilde{\boldsymbol{\beta}} = \boldsymbol{\delta}$, 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 $\boldsymbol{z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $E(\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma})$.
- $\boldsymbol{t}'\boldsymbol{y}$ is the BLUE for $E(\boldsymbol{t}'\boldsymbol{y})$ 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 $E(\boldsymbol{u}) = \boldsymbol{0}$, and $Var(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$.
  - **Gauss-Markov theorem**: Under the Gauss-Markov assumptions, $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}_{OLS}$ is the BLUE for estimable $\boldsymbol{\lambda}'\boldsymbol{\beta}$.
  - Under the Gauss-Markov model, an unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{SSE}{N-r}$.
- The **Aitken equations** are $\{\boldsymbol{\beta} : \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}\}$.
  - $\hat{\sigma}^2_{GLS} = \frac{1}{N-r}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})$.
- The **Aitken model** follows $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $E(\boldsymbol{u}) = \boldsymbol{0}$, and $Var(\boldsymbol{u}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known, positive-definite matrix.
  - **Aitken's theorem**: Under the Aitken model, $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}_{GLS}$ is the BLUE for estimable $\boldsymbol{\lambda}'\boldsymbol{\beta}$.
  - 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, $\boldsymbol{t}'\boldsymbol{y}$ 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!
    - $\text{Bias}(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}_{OLS}) = E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}_{OLS}) - \boldsymbol{\lambda}'\boldsymbol{\beta} = \boldsymbol{\lambda}'(\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}'\boldsymbol{\eta}$.
    - $E(SSE) = \boldsymbol{\eta}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta} + \sigma^2(N - r)$.
- 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 $E\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \sigma^2\text{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}$ (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 $E(N_k) = \mu_k$, $k = 1, \ldots, K$. 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 $\boldsymbol{\eta} = (\log\mu_1, ..., \log\mu_K)'$, and $\boldsymbol{\eta} = (\log\mu_1, \ldots, \log\mu_q)'$, 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 col($\boldsymbol{X}$), the column space of $\boldsymbol{X}$. For convenience, define the vectors $\boldsymbol{N} = (N_1, ..., N_K)'$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q)'$; also denote the vector of ones as j, and assume j is in col($\boldsymbol{X}$). Show that $\arg\max_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}} [\boldsymbol{N}'\boldsymbol{\eta} - \boldsymbol{j}'\boldsymbol{\mu}]$, and to find the MLE of $\boldsymbol{\beta}$, which uses the constraint $\boldsymbol{\eta} \in \text{col}(\boldsymbol{X})$, $\hat{\boldsymbol{\mu}}$ must satisfy $\boldsymbol{X}'\hat{\boldsymbol{\mu}} = \boldsymbol{X}'\boldsymbol{N}$.

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

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

5.5 Distributions/General Linear Hypotheses

Return to Table of Contents

- The **moment generating function**, or **MGF**, is $M_{\boldsymbol{X}}(\boldsymbol{t}) = E[e^{\boldsymbol{t}'\boldsymbol{X}}]$.
  - 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 $\exp\left\{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right\}$ in an open neighborhood containing the origin.
  - If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{Y} = a + \boldsymbol{B}\boldsymbol{X}$, then $\boldsymbol{Y} \sim N_q(a + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.
  - Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is non-singular, then:
    - A nonsingular matrix $\boldsymbol{A}$ exists such that $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$.
    - $\boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$.
    - The PDF is defined as $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}$.
    - $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are jointly independent iff $\boldsymbol{\Sigma}_{ij} = \boldsymbol{0}$ for all $i \neq j$.
  - Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{Y}_1 = \boldsymbol{a}_1 + \boldsymbol{B}_1\boldsymbol{X}$, $\boldsymbol{Y}_2 = \boldsymbol{a}_2 + \boldsymbol{B}_2\boldsymbol{X}$. $\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2$ iff $\boldsymbol{B}_1\boldsymbol{\Sigma}\boldsymbol{B}_2' = \boldsymbol{0}$.
  - Suppose $\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix} \sim N_p\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{V}_{11} & \boldsymbol{V}_{12}\\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22}\end{pmatrix}\right)$.

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

- Let $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. $\boldsymbol{U} = \boldsymbol{Z}'\boldsymbol{Z}$ has the $\boldsymbol{\chi}^2$**-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** $\boldsymbol{\chi}^2$**-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 $E(\boldsymbol{U}) = p + 2\phi$ 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 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, then $\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{X} \sim \chi^2_p\left(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}\right)$.
  - If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is non-singular, then $\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_p\left(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$.
- 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 $\boldsymbol{F}$**-distribution** with degrees of freedom p_1, p_2, and non-centrality parameter ϕ.
- Suppose $\boldsymbol{U} \sim N(\mu, 1) \perp \boldsymbol{V} \sim \chi^2_k$. $\frac{\boldsymbol{U}}{\sqrt{\boldsymbol{V}/k}}$ follows the $\boldsymbol{T}$**-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 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, and $\boldsymbol{A}$ be symmetric and idempotent with rank s, then $\boldsymbol{X'AX} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A\mu}\right)$.
- Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be symmetric with rank s. If $\boldsymbol{B\Sigma A} = \boldsymbol{0}$, then $\boldsymbol{BX} \perp \boldsymbol{X'AX}$.
- Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be symmetric with rank r, and $\boldsymbol{B}$ be symmetric with rank s. If $\boldsymbol{B\Sigma A} = \boldsymbol{0}$, then $\boldsymbol{X'BX} \perp \boldsymbol{X'AX}$.
- Given the linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, where $\boldsymbol{u} \sim N_N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$, the distribution of the BLUE of estimable $\boldsymbol{\Lambda}'\beta$ is $N(\boldsymbol{\Lambda'\beta}, \sigma^2\boldsymbol{\Lambda}'(\boldsymbol{X'X})^g\boldsymbol{\Lambda})$.
  - The BLUE is independent of $\frac{SSE}{\sigma^2}$.
  - The unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{SSE}{N-r}$.
  - $T(\boldsymbol{y}) = (\boldsymbol{y'y}, \boldsymbol{X'y})$ is a complete and sufficient statistic.
    - $(\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}, \boldsymbol{X'y})$ is also minimal sufficient.
  - The least squares estimator of an estimable $\boldsymbol{\Lambda'\beta}$ has the smallest variance of any estimator for its expectation.
  - $\boldsymbol{\Lambda}'\hat{\boldsymbol{\beta}}$ is the MLE for an estimable $\boldsymbol{\Lambda'\beta}$, 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 $H_0 : \boldsymbol{K'\beta} = \boldsymbol{m}$ is **testable** iff $\boldsymbol{K} \in \mathbb{R}^{q\times s}$ has full-column rank, and each column of $\boldsymbol{K'\beta}$ 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 $\boldsymbol{K'\beta}$ is estimable, then $\boldsymbol{H} = \boldsymbol{K}'(\boldsymbol{X'X})^g\boldsymbol{K}$ is non-singular.
    - A result is $(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'(\sigma^2\boldsymbol{H})^{-1}(K'\hat{\boldsymbol{\beta}} - \boldsymbol{m}) \sim \chi^2_s\left(\frac{1}{2}(\boldsymbol{K'\beta} - \boldsymbol{m})'(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{K'\beta} - \boldsymbol{m})\right)$.
    - Therefore, $F = \frac{(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\boldsymbol{H}^{-1}(K'\hat{\boldsymbol{\beta}} - \boldsymbol{m})/s}{SSE/(N-r)} \sim F_{s,N-r}\left(\frac{1}{2}(\boldsymbol{K'\beta} - \boldsymbol{m})'(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{K'\beta} - \boldsymbol{m})\right)$.
      - 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 $\boldsymbol{X}$. $\boldsymbol{X'X}$, and $\boldsymbol{X'y}$ are given below:

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

Suppose $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. 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. Using the fact that $\boldsymbol{y'y} - \boldsymbol{y'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = 96$, 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.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS} =& (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \\ =& \begin{bmatrix} 1/11 & & \\ & 1/34 & \\ & & 1/34 \end{bmatrix}\begin{bmatrix} 220 \\ 34 \\ -68 \end{bmatrix} = \begin{bmatrix} 20 \\ 1 \\ -2 \end{bmatrix}; \\ E(Y|x_1 = 2, x_2 = 2) =& 20 + 1(2) - 2(2) = 18.\end{aligned}$$

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

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

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

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

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 $\boldsymbol{K}'\boldsymbol{\beta}$.
  - The parameter space under $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ is $\Omega_0 = \{(\boldsymbol{\beta}, \sigma^2) : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}, \sigma^2 > 0\}$.
  - The union of the parameter space under H_0 and $H_1 : \boldsymbol{K}'\boldsymbol{\beta} \neq \boldsymbol{m}$ 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 $\boldsymbol{K}'\boldsymbol{\beta}$ is a set of linearly independent estimable functions, and $\hat{\boldsymbol{\beta}}_H$ is a part of a solution to the RNEs with constraint $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$, then $Q(\hat{\boldsymbol{\beta}}_H) - Q(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}}) = (\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\boldsymbol{H}^{-1}(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})$.

- If $\boldsymbol{K}'\boldsymbol{\beta}$ 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 $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$, by solving for $\boldsymbol{\beta}$ in

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

- $\boldsymbol{P}'\boldsymbol{\beta}$ is **jointly nonestimable** if no linear combination of $\boldsymbol{P}'\boldsymbol{\beta}$ is estimable.

- If $\boldsymbol{P}'\boldsymbol{\beta}$ 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 $\boldsymbol{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$, 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 $\tau_j := \boldsymbol{\lambda}'_j\boldsymbol{\beta}$. 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 $\boldsymbol{u}'\tau : \boldsymbol{u}'\hat{\tau} \pm \sqrt{\hat{\sigma}^2 sF_{s,N-r,\alpha}\boldsymbol{u}'\boldsymbol{H}\boldsymbol{u}}$.
    - 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 $N(0,1)$ RVs for $i \in \{1, \ldots, k\}$, 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 $|\tau_i - \tau_j| \leq h$ for all i, j, and $\sum_i u_i = 0$ (a contrast), then $|\sum_i u_i \tau_i| \leq h \cdot \frac{1}{2} \sum_i |u_i|$.
    - 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 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is non-singular, and $\boldsymbol{A}$ be symmetric such that $\boldsymbol{A\Sigma}$ is idempotent with rank s, then $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} \sim \chi^2_s\left(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}\right)$.
  - For $\boldsymbol{y} \sim N_N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_N)$, 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 $\boldsymbol{y} \sim N_N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, 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 $\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{A}_i\boldsymbol{y}$ are independently distributed as $\chi^2_{s_i}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}'\boldsymbol{A}_i\boldsymbol{\mu}\right)$, and $\sum_{i=1}^k s_i = N$.
- ANOVA table for SSQ: Define $\boldsymbol{X}_j^* := [\boldsymbol{X}_0 | \boldsymbol{X}_1 | \ldots | \boldsymbol{X}_j]$, and $R(\boldsymbol{b}_j, \ldots) = R(\boldsymbol{b}_0, \ldots, \boldsymbol{b}_j) = \boldsymbol{y}'\boldsymbol{P}_{\boldsymbol{X}_j^*}\boldsymbol{y}$.

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

- $\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{A}_j\boldsymbol{y} \sim \chi^2_{r_j}\left(\frac{(\boldsymbol{Xb})'\boldsymbol{A}_j(\boldsymbol{Xb})}{2\sigma^2}\right)$, 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 Random Math Stuff

Return to Table of Contents

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

7 Distributions

Return to Table of Contents

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*	$\frac{\exp\left\{-\frac{1}{2(1-\rho^2)}(*)\right\}}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}}$	$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$			
χ^2	$\frac{x^{p/2-1}e^{-x/2}}{\Gamma(p/2)2^{p/2}}$	$x \geq 0$	p	$2p$	$\left(\frac{1}{1-2t}\right)^{p/2}$
F	$\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)(\nu_1/\nu_2)^{\nu_1/2}(x)^{(\nu_1-2)/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}x\right)^{(\nu_1+\nu_2)/2}}$	$x \geq 0$	$\frac{\nu_2}{\nu_2-2}, \ \nu_2 > 2$	$2\left(\frac{\nu_2}{\nu_2-2}\right)^2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)}, \ \nu_2 > 4$	$E(X^n) = \frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(\nu_1/\nu_2)^n}, n < \frac{\nu_2}{2}$
T	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$x \in \mathbb{R}$	$0, \ \nu > 1$	$\frac{\nu}{\nu-2}, \ \nu > 2$	$E(X^n) = \begin{cases} \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{\nu-n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\nu^{-n/2}}, \ n > \nu \\ 0, \ n < \nu \end{cases}$

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

7.1 Equivalences

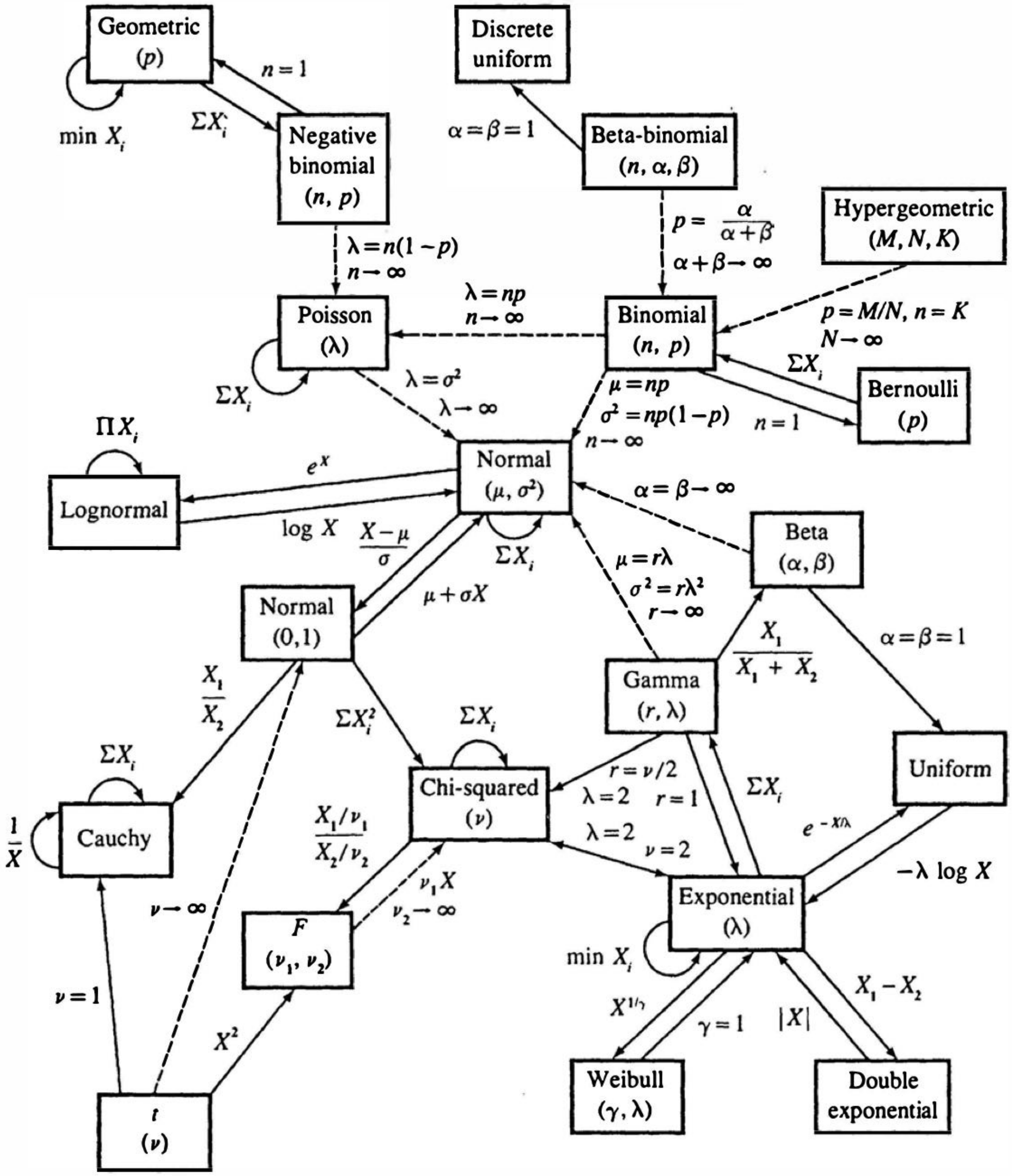
Return to Table of Contents

- $Bin(1, p) = Ber(p)$.
- $NegBin(1, p) = Geom(p)$.
- $MN(n, (p, 1-p)) = Bin(n, p)$.
- Gamma$(1, \beta) = Exp(\beta)$.
- Gamma$\left(\frac{p}{2}, 2\right) = \chi_p^2$.
- Weibull$(1, \beta) = Exp(\beta)$.
- $\frac{X}{Y}$ ~Cauchy$(0, 1)$, $X \perp Y \sim N(0, 1)$.

- If X ~Exp(1), then $\mu - \sigma \log(X)$ ~GEV$(\mu, \sigma, 0)$.
- If X ~Weibull(μ, σ), then $\left[1 - \sigma \log\left(\frac{X}{\sigma}\right)\right]$ ~GEV$(\mu, \sigma, 0)$.
- If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U(0,1)$, then $X_{(k)}$ ~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 ~Weibull$\big(\lambda, \frac{1}{2}\big)$, then $\sqrt{X}$ ~Exp$\left(\frac{1}{\sqrt{\lambda}}\right)$.
- If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}Bin(m, p_i)$, then $(X_1, \ldots, X_{n-1}|X_n = x_n)$ ~MN$\left(m - x_n, \left[\frac{\underline{p}}{1-p_n}\right]^{\underline{x}}\right)$.
- If $X \sim U(0,1)$, then $-\log(X)$ ~Exp(1).
- If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}Exp(\lambda)$, then $\sum_{i=1}^n X_i$ ~Gamma(n, λ).
- If $X_1, \ldots, X_n \overset{\perp}{\sim}$Pois$(\lambda_i)$, then $\sum_{i=1}^n X_i$ ~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*.